# ExtremeXOS User Guide

for Version 16.2

121154-01 Rev 01
March 2020

## Legal Notice

Extreme Networks, Inc. reserves the right to make changes in specifications and other information contained in this document and its website without prior notice. The reader should in all cases consult representatives of Extreme Networks to determine whether any such changes have been made.

The hardware, firmware, software or any specifications described or referred to in this document are subject to change without notice.

## Trademarks

Extreme Networks and the Extreme Networks logo are trademarks or registered trademarks of Extreme Networks, Inc. in the United States and/or other countries.

All other names (including any product names) mentioned in this document are the property of their respective owners and may be trademarks or registered trademarks of their respective companies/owners.

For additional information on Extreme Networks trademarks, please see:
www.extremenetworks.com/company/legal/trademarks

## Software Licensing

Some software files have been licensed under certain open source or third-party licenses. End-user license agreements and open source declarations can be found at:
www.extremenetworks.com/support/policies/software-licensing

## Support

For product support, phone the Global Technical Assistance Center (GTAC) at 1-800-998-2408 (toll-free in U.S. and Canada) or +1-408-579-2826. For the support phone number in other countries, visit: http://www.extremenetworks.com/support/contact/

For product documentation online, visit: https://www.extremenetworks.com/documentation/

# Introduction to the ExtremeXOS User Guide



This guide is intended for use by network administrators who are responsible for installing and setting up network equipment. In addition to comprehensive conceptual information about each feature of our software, you will also find detailed configuration material, helpful examples, and troubleshooting information. Also included are supported platforms and recommended best practices for optimal software performance.

Note
If the information in the release notes shipped with your switch differs from the information in this guide, follow the release notes.

## Conventions

This section discusses the conventions used in this guide.

### Text Conventions

The following tables list text conventions that are used throughout this guide.

**Table 1: Notice Icons**

| Icon | Notice Type | Alerts you to... |
|---|---|---|
| | General Notice | Helpful tips and notices for using the product. |
| | Note | Important features or instructions. |
| | Caution | Risk of personal injury, system damage, or loss of data. |

**Table 1: Notice Icons (continued)**

| Icon | Notice Type | Alerts you to... |
|---|---|---|
|  | Warning | Risk of severe personal injury. |
| ***New!*** | New Content | Displayed next to new content. This is searchable text within the PDF. |

**Table 2: Text Conventions**

| Convention | Description |
|---|---|
| `Screen displays` | This typeface indicates command syntax, or represents information as it appears on the screen. |
| The words **enter** and **type** | When you see the word "enter" in this guide, you must type something, and then press the Return or Enter key. Do not press the Return or Enter key when an instruction simply says "type." |
| **[Key]** names | Key names are written with brackets, such as **[Return]** or **[Esc]**. If you must press two or more keys simultaneously, the key names are linked with a plus sign (+). Example: Press **[Ctrl]**+**[Alt]**+**[Del]** |
| *Words in italicized type* | Italics emphasize a point or denote new terms at the place where they are defined in the text. Italics are also used when referring to publication titles. |

## VLAN Option Formatting in Commands

For commands with a *`vlan_list`* option, the input into this option must not contain spaces.

*Example*

The `enable stpd auto-bind` command *`vlan_list`* input should be entered as:

```
enable stpd "s0" auto-bind vlan 10,20-30
```

Not as:

```
enable stpd "s0" auto-bind vlan 10, 20-30
```

## Platform-Dependent Conventions

Unless otherwise noted, all information applies to all platforms supported by ExtremeXOS software, which are the following:

- ExtremeSwitching® switches
- Summit® switches
- SummitStack™

When a feature or feature implementation applies to specific platforms, the specific platform is noted in the heading for the section describing that implementation in the ExtremeXOS command documentation (see the Extreme Documentation page at www.extremenetworks.com/documentation/). In many cases, although the command is available on all platforms, each platform

uses specific keywords. These keywords specific to each platform are shown in the Syntax Description and discussed in the Usage Guidelines sections.

### Terminology

When features, functionality, or operation is specific to a switch family, such as ExtremeSecurity or Summit™, the family name is used. Explanations about features and operations that are the same across all product families simply refer to the product as the *switch*.

## Related Publications

### ExtremeXOS Publications

- *ACL Solutions Guide*
- *ExtremeXOS 16.2 Command Reference Guide*
- *ExtremeXOS 16.2 EMS Messages Catalog*
- *ExtremeXOS 16.2 Feature License Requirements*
- *ExtremeXOS 16.2 User Guide*
- *ExtremeXOS OpenFlow User Guide*
- *ExtremeXOS Quick Guide*
- *ExtremeXOS Legacy CLI Quick Reference Guide*
- *ExtremeXOS Release Notes*
- *Extreme Hardware/Software Compatibility and Recommendation Matrices*
- *Switch Configuration with Chalet for ExtremeXOS 16.2 and Earlier*
- *Using AVB with Extreme Switches*

### Open Source Declarations

Some software files have been licensed under certain open source licenses. More information is available at: www.extremenetworks.com/support/policies/software-licensing/.

## Providing Feedback to Us

We are always striving to improve our documentation and help you work better, so we want to hear from you! We welcome all feedback but especially want to know about:

- Content errors or confusing or conflicting information.
- Ideas for improvements to our documentation so you can find the information you need faster.
- Broken links or usability issues.

If you would like to provide feedback to the Extreme Networks Information Development team about this document, please contact us using our short online feedback form. You can also email us directly at documentation@extremenetworks.com.

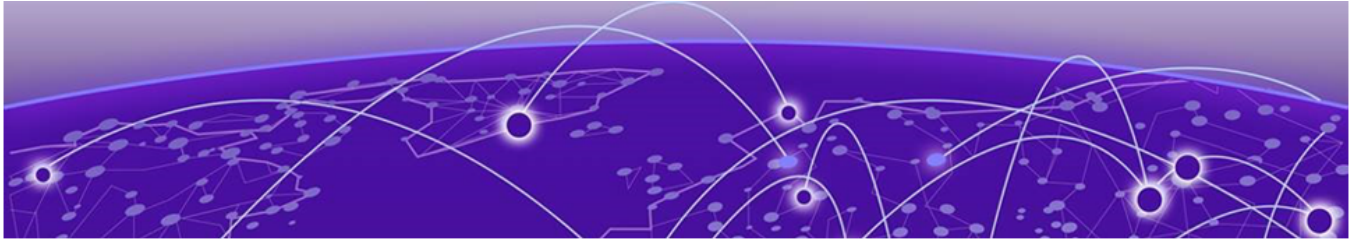
## Getting Help

If you require assistance, contact Extreme Networks using one of the following methods:

- GTAC (Global Technical Assistance Center) for Immediate Support
  - Phone: 1-800-998-2408 (toll-free in U.S. and Canada) or +1 408-579-2826. For the support phone number in your country, visit: www.extremenetworks.com/support/contact
  - Email: support@extremenetworks.com. To expedite your message, enter the product name or model number in the subject line.
- Extreme Portal — Search the GTAC knowledge base, manage support cases and service contracts, download software, and obtain product licensing, training, and certifications.
- The Hub — A forum for Extreme Networks customers to connect with one another, answer questions, and share ideas and feedback. This community is monitored by Extreme Networks employees, but is not intended to replace specific guidance from GTAC.

Before contacting Extreme Networks for technical support, have the following information ready:

- Your Extreme Networks service contract number and/or serial numbers for all involved Extreme Networks products
- A description of the failure
- A description of any action(s) already taken to resolve the problem
- A description of your network environment (such as layout, cable type, other relevant environmental information)
- Network load at the time of trouble (if known)
- The device history (for example, if you have returned the device before, or if this is a recurring problem)
- Any related RMA (Return Material Authorization) numbers



# Getting Started



This section is intended to help you learn about your ExtremeXOS software. Information about your product, software version requirements and navigation, common commands, and password management, along with other helpful software orientation information can be found in this chapter.

## Product Overview

The following table lists the Extreme Networks products that run the ExtremeXOS software.

**Table 3: ExtremeXOS Switches**

| Switch Series | Switches |
|---|---|
| BlackDiamond X8 Series | BlackDiamond X8, BlackDiamond X8-100G4X, BDXA-G48X, BDXA-G48T |
| BlackDiamond 8800 Series | BlackDiamond 8810, BlackDiamond 8806 |
| Cell Site Routers | E4G-200<br>E4G-400 |

**Table 3: ExtremeXOS Switches (continued)**

| Switch Series | Switches |
|---|---|
| Summit X430 Series | Summit X430-24t<br>Summit X430-48t<br>Summit X430-8p<br>Summit X430-24p |
| Summit X440 Series | Summit X440-8t<br>Summit X440-8p<br>Summit X440-24t<br>Summit X440-24p<br>Summit X440-24tDC<br>Summit X440-48tDC<br>Summit X440-24t-10G<br>Summit X440-24p-10G<br>Summit X440-48t<br>Summit X440-48p<br>Summit X440-48t-10G<br>Summit X440-48p-10G<br>Summit X440-24x<br>Summit X440-24x-10G |
| Summit X450-G2 Series | Summit X450-G2-24t-10GE4<br>Summit X450-G2-24p-10GE4<br>Summit X450-G2-48t-10GE4<br>Summit X450-G2-48p-10GE4<br>Summit X450-G2-24t-GE4<br>Summit X450-G2-24p-GE4<br>Summit X450-G2-48t-GE4<br>Summit X450-G2-48p-GE4 |
| Summit X460 Series | Summit X460-24x<br>Summit X460-24t<br>Summit X460-24p<br>Summit X460-48x<br>Summit X460-48t<br>Summit X460-48P<br>Summit X460-G2-24t-10GE4<br>Summit X460-G2-48t-10GE4<br>Summit X460-G2-24p-10GE4<br>Summit X460-G2-48p-10GE4<br>Summit X460-G2-24x-10GE4<br>Summit X460-G2-48x-10GE4<br>Summit X460-G2-24t-GE4<br>Summit X460-G2-48t-GE4<br>Summit X460-G2-24p-GE4<br>Summit X460-G2-48p-GE4 |
| Summit X480 Series | Summit X480-24x<br>Summit X480-48x<br>Summit X480-48t |
| Summit X670 | Summit X670-48x<br>Summit X670V-48x<br>Summit X670V-48t<br>Summit X670-G2-48x-4q<br>Summit X670-G2-72x |

**Table 3: ExtremeXOS Switches (continued)**

| Switch Series | Switches |
|---|---|
| Summit X770 | Summit X770-32q |
| SummitStack | All Summit family switches, except the Summit X430 series. |

## Software Required

This section identifies the software version required for each switch that runs ExtremeXOS software.

Note
The features available on each switch are determined by the installed feature license and optional feature packs. For more information, see the *Feature License Requirements* document.

The following table lists the BlackDiamond 8000 series modules and the ExtremeXOS software version required to support each module.

**Table 4: BlackDiamond 8000 Series Switch Modules and Required Software**

| Module Series Name | Modules | Minimum ExtremeXOS Software Version |
|---|---|---|
| MSMs | MSM-48c<br>8900-MSM128<br>8800-MSM96 | ExtremeXOS 12.1<br>ExtremeXOS 12.3<br>ExtremeXOS 16.1.3 |
| c-series | G24Xc<br>G48Xc<br>10G4Xc<br>10G8Xc<br>G48Tc<br>S-10G1Xc<br>S-10G2Xc<br>S-G8Xc | ExtremeXOS 12.1<br>ExtremeXOS 12.1<br>ExtremeXOS 12.1<br>ExtremeXOS 12.1<br>ExtremeXOS 12.1<br>ExtremeXOS 12.1<br>ExtremeXOS 12.5.3<br>ExtremeXOS 12.1 |
| | 8900-G96T-c<br>8900-10G24X-c | ExtremeXOS 12.3<br>ExtremeXOS 12.3 |
| xl-series | 8900-G48X-xl 8900-G48T-xl<br>8900-10G8X-xl | ExtremeXOS 12.4 |
| xm-series | 8900-40G6X-xm | ExtremeXOS 12.6 |

The following guidelines provide additional information on the BlackDiamond 8000 series modules described in the previous table:

- The term BlackDiamond 8000 series modules refers to all BlackDiamond 8800 and 8900 series modules. Beginning with the ExtremeXOS 12.5 release, it does not include other modules formerly listed as original-series modules.
- Module names that are not preceded with 8900 are BlackDiamond 8800 series modules.
- The c-series, e-series, xl-series, and xm-series names are used to distinguish between groups of modules that support different feature sets.

The following table lists the Summit family switches that run ExtremeXOS software and the minimum ExtremeXOS software version required.

**Table 5: Summit Family Switches and Required Software**

| Switch Series | Switches | Minimum ExtremeXOS Software Version |
|---|---|---|
| Summit X430 Series | Summit X430-24t<br>Summit X430-48t | ExtremeXOS 15.3.2 |
| | Summit X430-8p<br>Summit X430-24p | ExtremeXOS 15.5.2 |
| Summit X440 Series | Summit X440-8t<br>Summit X440-8p<br>Summit X440-24t<br>Summit X440-24p<br>Summit X440-24t-10G<br>Summit X440-24p-10G<br>Summit X440-24tDC<br>Summit X440-48tDC<br>Summit X440-48t<br>Summit X440-48p<br>Summit X440-48t-10G<br>Summit X440-48p-10G<br>Summit X440-24x<br>Summit X440-24x-10G | ExtremeXOS 15.1<br>ExtremeXOS 15.3<br>ExtremeXOS 15.3<br>ExtremeXOS 15.2<br>ExtremeXOS 15.3<br>ExtremeXOS 15.3 |
| Summit X450-G2 Series | Summit X450-G2-24t-10GE4<br>Summit X450-G2-24p-10GE4<br>Summit X450-G2-48t-10GE4<br>Summit X450-G2-48p-10GE4<br>Summit X450-G2-24t-GE4<br>Summit X450-G2-24p-GE4<br>Summit X450-G2-48t-GE4<br>Summit X450-G2-48p-GE4 | ExtremeXOS 16.1. |
| Summit X460 Series | Summit X460-24x<br>Summit X460-24t<br>Summit X460-24p<br>Summit X460-48x<br>Summit X460-48t<br>Summit X460-48p | ExtremeXOS 12.5 |
| | Summit X460-G2-24t-10GE4<br>Summit X460-G2-48t-10GE4<br>Summit X460-G2-24p-10GE4<br>Summit X460-G2-48p-10GE4<br>Summit X460-G2-24x-10GE4<br>Summit X460-G2-48x-10GE4<br>Summit X460-G2-24t-GE4<br>Summit X460-G2-48t-GE4<br>Summit X460-G2-24p-GE4<br>Summit X460-G2-48p-GE4 | ExtremeXOS 15.6 |
| Summit X480 Series | Summit X480-24x<br>Summit X480-48x<br>Summit X480-48t | ExtremeXOS 12.4 |

**Table 5: Summit Family Switches and Required Software (continued)**

| Switch Series | Switches | Minimum ExtremeXOS Software Version |
|---|---|---|
| Summit X670 | Summit X670-48x<br>Summit X670V-48x<br>Summit X670V-48t | ExtremeXOS 12.6<br>ExtremeXOS 15.2.2 |
| | Summit X670G2-48x-4q<br>Summit X670G2-72x | ExtremeXOS 15.6 |
| Summit X770 | Summit X770-32q | ExtremeXOS 15.4 |
| SummitStack | Summit family switches except the Summit X430 series | ExtremeXOS 12.0 |

The table above lists the current Summit Family Switches.

Stacking-capable switches are a combination of up to eight Summit family switches that are connected by stacking cables.

## Simple Switch Configuration with Chalet

Chalet is a web-based user interface for setting up and viewing information about a switch. Chalet removes the need to know and remember commands in a CLI environment. Viewable on desktop and mobile with a quick login and intuitive navigation, Chalet features an Quick Setup mode for configuring a switch in a few simple steps. Basic data surrounding port utilization, power, and Quality of Service (QoS) are available, and more advanced users can configure multiple VLANs, create Access Control Lists (ACLs), and configure Audio Video Bridging (AVB).

Chalet is packaged with ExtremeXOS release 15.7.1 and later for all platforms, so there's nothing extra to download or install. Chalet can be launched in any modern web browser and does not depend on any outside resources to work, including Java Applets, Adobe Flash, or dedicated mobile applications.

Chalet helps you interact with the switch outside of a CLI environment and allows you to easily:

- Configure the switch for the first time without the use of a console cable.
- View status and details of the switch and its slots and ports.
- Analyze power efficiency of power supplies, fans, and PoE ports.
- Create VLANs and ACL policies.
- Enable and disable multiple features, including QoS, AVB, auto-negotiation, and flooding.
- View recent system events.
- View device topology (stacked switches only).
- Manage users, including defining global and individual security policies.

Refer to the Chalet user guide (*Switch Configuration with Chalet for ExtremeXOS 16.x and Earlier*) for instructions on setting up, logging in, configuring, and monitoring your switch.

## Zero Touch Provisioning (Auto Configuration)

Zero Touch Provisioning enables switches “just out of the box” to automatically gain a management IP address and configuration without serial cables and manual configuration.

ZTP provides:

- Management port IP connectivity using an IPv4 link-local IP address
- *DHCP (Dynamic Host Configuration Protocol)* client to contact a DHCP server for:
  - Assigned IP address
  - ExtremeXOS image update
  - Configuration or script file
  - ExtremeManagement trap address

## IPv4 Link-Local Address

Link-Local addressing (subnet 169.254.x.x) allows a host device to automatically and predictably derive a non-routable IP address for IP communication over Ethernet links.

By configuring the Ethernet management port, 'just out of the box', with an IP address, a user can connect a laptop directly to the management Ethernet port. If the laptop is not configured with a fixed IP address, it tries to get an IP address from a DHCP server. If it cannot, it assigns its own Link-Local address putting the switch and the laptop on the same subnet. The laptop can then use telnet or a web browser to access the switch removing the need for the serial cable.

The IPv4 address format is used to make it simple for a user to determine the switch's IP address. The formula is to use the lower 2 bytes of the MAC address as the last two numbers in the Link-Local IPv4 address.

MAC address: 00:04:96:97:E9:EE

Link-Local IP address is:

- 169.254.233.238 or 0xa9fee9ee

Web browsers accept a hexadecimal value as an IPv4 address. (Microsoft IE displays the URL with the number dot notation 169.254.233.239.)

The web URL is http:// 0xa9fee9ee or just 0xa9fee9ee

The user documentation directs the customer to access the web browser by typing 0xa9fe followed by the last two number/letter groups in the MAC address found on the switch label. No hexadecimal translation is required.

With this information, a user can connect the Ethernet port directly from a laptop to this switch using the temporary Link-Local address. You can communicate via web or telnet to perform the initial switch configuration, if needed, and no longer needs a serial cable to configure a switch.

## DHCP Parameters

If a DHCP server is available, ZTP tries to contact it alternating between the default *VLAN (Virtual LAN)* and the management ethernet port. The DHCP server can provide:

- IP Address
- Gateway

- option43 parameters
- option125 paramters.

If an IP address is provided by a DHCP server on the management port, it replaces the Link-Local management IPv4 address.

If a TFTP server IP address is provided along with the name of a config file, ZTP downloads the config file to the switch. The switch reboots to activate the config file.

For .xos image files, ZTP executes the EXOS `download image` command to update the switch software. The switch does not reboot after the download image command completes.

## Option43

Option43 processing does not require an NMS. If a switch receives option43 as part of the DHCP response, it uses the TFTP protocol to transfer files from the specified TFTP server IP address.

Option43 parameters may contain:

- TFTP Server to Contact
- Config file to be loaded or script to be run (.xsf or .py)
- Policy files (.pol)
- EXOS image file to be downloaded (.xos)
- EXOS xmond file to be downloaded (.xmod)
- *SNMP (Simple Network Management Protocol)* trap receiver address for Extreme MIB traps

Multiple file names may be specified in option43. The file names can be either relative path names or a full URL with the IP address of the TFTP server. If relative path names are specified, the TFTP IP address is also required.

File name examples assuming a TFTP server is present with the IP address 10.10.10.1:

- exos/summitX-15.7.1.1.xos (specify the IP address in sub option 100)
- tftp://10.10.10.1/exos/summitX-15.7.1.1.xos (sub option 100 is not required)

Once all of the files specified in option43 have been transferred to the switch, the switch reboots.

## ExtremeXOS Image Update

Using ZTP, you can setup a DHCP/TFTP server and connect switches directly to it, possibly via an L2 switch. Switches can then update themselves with an ExtremeXOS generally available software image before being installed into a live network. The following figure shows one possible method of upgrading switches by connecting them to an L2 switch. This approach upgrades the switches before being deployed into a network.

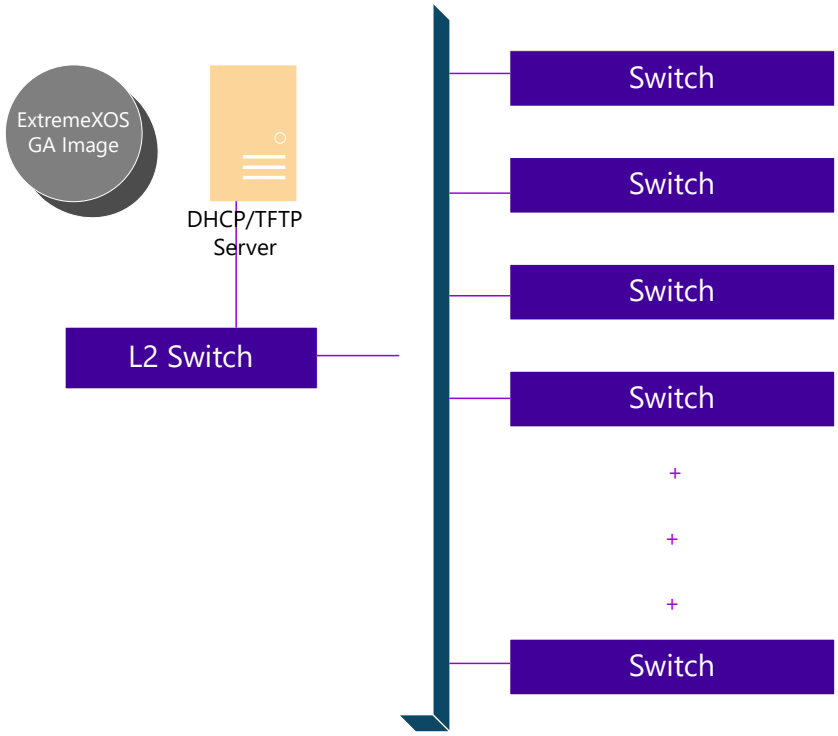



**Figure 1: ZTP DHCP/TFTP Server Setup**

## Option125

Option125 depends on ExtremeManagement being present for initial switch configuration and software upgrades.

Option125 parameters contain the ExtremeManagement trap address.

Specifying option125 in the DHCP response causes the switch to issue a etsysConfigMgmtReadyNotification trap to the ExtremeManagement NMS. NetSight then discovers the switch information via SNMP and can, optionally, send a series of commands to the switch to download files or configure the switch.

# Logging in to the Switch

Perform the following tasks to log in to the switch.

1. The initial login prompt appears as follows:

   ```
   (Pending-AAA) login:
   ```

   At this point, the failsafe account is now available, but the normal AAA login security is not. (For additional information on using the failsafe account, refer to Failsafe Accounts on page 30.)
2. Wait for the following message to appear:

   ```
   Authentication Service (AAA) on the master node is now available for login.
   ```

   At this point, the normal AAA login security is available.

3. Press **[Enter]**.

   Whether or not you press **[Enter]**, once you see the login prompt, you can perform a normal login. (See Default Accounts on page 29.)

   The following prompt appears: `login`

## Understanding the Command Syntax

This section describes the steps to take when you enter a command.

ExtremeXOS command syntax is described in detail in the *ExtremeXOS 16.2 User Guide*. Some commands are also included in this guide in order to describe how to use ExtremeXOS software features. However, only a subset of commands are described here, and in some cases only a subset of the options that a command supports. You should consider the *ExtremeXOS 16.2 User Guide* as the definitive source for information on ExtremeXOS commands.

You can enter configuration commands at the # prompt. At the > prompt, you can enter only monitoring commands, not configuration commands. When you log in as admin (which has read and write access), you see the # prompt. When you log in as user (which has only read access), you will see the > prompt. When the switch is booting up, you may see the > command prompt. When the bootup process is complete, the # prompt is displayed.

When you enter a command at the prompt, ensure that you have the appropriate privilege level.

Most configuration commands require you to have the administrator privilege level. For more information on setting CLI privilege levels, see the *ExtremeXOS 16.2 User Guide*.

### Using the CLI

This section describes how to use the CLI to issue commands.

1. At the prompt, enter the command name.

   If the command does not include a parameter or values, skip to step 3. If the command requires more information, continue to step 2.

2. If the command includes a parameter, enter the parameter name and values.

   The value part of the command specifies how you want the parameter to be set. Values include numerics, strings, or addresses, depending on the parameter.

3. After entering the complete command, press **[Enter]**.

Note

If an asterisk (*) appears in front of the command line prompt, it indicates that you have pending configuration changes that have not been saved. For more information on saving configuration changes, see Software Upgrade and Boot Options on page 1522.

### Syntax Helper

The CLI has a built-in syntax helper. If you are unsure of the complete syntax for a particular command, enter as much of the command as possible, and then press **[Tab]** or `?`. The syntax helper provides a list of options for the command, and places the cursor at the end of that portion of the command you already entered.

If you enter an invalid command, the syntax helper notifies you of your error, and indicates where the error is located.

If the command is one where the next option is a named component (such as a *VLAN*, access profile, or route map), the syntax helper also lists any currently configured names that might be used as the next option. In situations where this list is very long, the syntax helper lists only one line of names, followed by an ellipsis (...) to indicate that there are more names that can be displayed.

The syntax helper also provides assistance if you have entered an incorrect command.

## Object Names

You must provide all named components within a category of the switch configuration (such as *VLAN*) a unique *object name*.

Object names must begin with an alphabetical character, and may contain alphanumeric characters and underscores ( _ ), but they cannot contain spaces. The maximum allowed length for a name is 32 characters. User-created object names for the following modules are not case-sensitive: access list, account, CFM, EAPS, *ESRP (Extreme Standby Router Protocol)*, flow-redirect, meter, *MSDP (Multicast Source Discovery Protocol)*, Network Login, PVLAN, protocol, *SNMP*, SSHD2, *STP (Spanning Tree Protocol)*, tunnel, UPM, VLAN, VMAN, etc.

Object names can be reused across categories (for example, *STPD (Spanning Tree Domain)* and VLAN names). If the software encounters any ambiguity in the components within your command, it generates a message requesting that you clarify the object you specified.

Note

If you use the same name across categories, we recommend that you specify the identifying keyword as well as the actual name. If you do not use the keyword, the system may return an error message.

### *Reserved Keywords*

Keywords such as **vlan**, **stp**, and other second-level keywords are reserved and you cannot use them as object names. This restriction only applies to the specific word e.g (**vlan**); you can use expanded versions e.g (**vlan2**) of the word.

A complete list of the reserved keywords for ExtremeXOS 12.4.2 and later software is displayed in the following table. Any keyword that is not on this list can be used as an object name.

**Table 6: Reserved Keywords**

| Reserved Keywords | | | | |
|---|---|---|---|---|
| aaa | elsm | IPv6 | pim | sys-health-check |
| access-list | ems | ipv6acl | policy | syslog |
| account | epm | irdp | ports | sys-recovery-level |
| accounts | esrp | isid | power | tacacs |
| all | fabric | isis | primary | tacacs-accounting |
| bandwidth | failover | jumbo-frame | private-vlan | tacacs-authorization |
| banner | failsafe-account | jumbo-frame-size | process | tech |
| bfd | fans | l2stats | protocol | telnet |
| bgp | fdb | l2vpn | put | telnetd |
| bootp | fdbentry | lacp | qosprofile | temperature |
| bootprelay | firmware | learning | qosscheduler | tftpd |
| brm | flood-group | learning-domain | radius | thttpd |
| bvlan | flooding | license | radius-accounting | time |
| cancel | flow-control | license-info | rip | timeout |
| cfgmgr | flow-redirect | licenses | ripng | timezone |
| cfm | forwarding | lldp | rmon | tos |
| checkpoint-data | from | log | router-discovery | traffic |
| clear-flow | get | loopback-mode | rtmgr | trusted-ports |
| cli | hal | mac | safe-default-script | trusted-servers |
| cli-config-logging | hclag | mac-binding | script | ttl |
| clipaging | heartbeat | mac-lockdown-timeout | secondary | tunnel |
| configuration | icmp | management | session | udp |
| configure | identity-management | mcast | sflow | udp-echo-server |
| continuous | idletimeout | memory | sharing | udp-profile |
| count | idmgr | memorycard | show | update |
| counters | igmp | meter | slot | upm |
| cpu-monitoring | image | mirroring | slot-poll-interval | var |
| cvlan | ingress | mld | smartredundancy | version |
| debug | inline-power | mpls | snmp | virtual-router |
| debug-mode | internal-memory | mrinfo | snmpv3 | vlan |
| devmgr | interval | msdp | sntp-client | vman |
| dhcp | iob-debug-level | msgsrv | source | vpls |
| dhcp-client | iparp | msm | ssl | vr |
| dhcp-server | ipconfig | msm-failover | stacking | vrrp |
| diagnostics | ipforwarding | mstp | stacking-support | watchdog |
| diffserv | ipmc | mtrace | stack-topology | web |
| dns-client | ipmcforwarding | multiple-response-timeout | start-size | xmlc |
| dont-fragment | ipmroute | mvr | stp | xmld |
| dos-protect | ip-mtu | neighbor-discovery | stpd | xml-mode |
| dot1ag | ip-option | netlogin | subvlan-proxy- arp | xml-notification |
| dot1p | iproute | nettools | svlan | |
| dot1q | ip-security | node | switch | |
| ds | ipstats | nodemgr | switch-mode | |
| eaps | ipv4 | odometers | | |
| edp | IPv4 | ospf | | |
| egress | | ospfv3 | | |

**Table 6: Reserved Keywords (continued)**

| Reserved Keywords | | | | |
|---|---|---|---|---|
| **elrp**<br>**elrp-client** | **ipv6** | | | |

## Abbreviated Syntax

Abbreviated syntax is the shortest unambiguous allowable abbreviation of a command or parameter. Typically, this is the first three letters of the command.

When using abbreviated syntax, you must enter enough characters to make the command unambiguous and distinguishable to the switch. If you do not enter enough letters to allow the switch to determine which command you mean, the syntax helper provides a list of the options based on the portion of the command you have entered.

## Command Shortcuts

### *Component Name Shortcut*

Components are typically named using the `create` command. When you enter a command to configure a named component, you do not need to use the keyword of the component.

Note
True only for some named components.

For example, you can create a *VLAN* named `engineering`:

```
create vlan engineering
```

After you have created the name for the VLAN, you can eliminate the keyword **`vlan`** from all other commands that require the name to be entered. For example:

```
configure engineering delete port 1:3,4:6
```

Instead of entering the command:

```
configure vlan engineering delete port 1:3,4:6
```

### *Using VLAN IDs or Lists Instead of Names*

You can use *VLAN* IDs or vlan_lists in just about every case where you would use VLAN names. Referring to a VLAN by VID and specifying lists of VIDs is a useful shortcut that can greatly reduce the number of commands required to configure the switch.

For example, this shortcut allows you to create 100 VLANs with a single command. The following command creates VLANs named VLAN_0100 ... VLAN_0199.

```
create vlan 100-199
```

In another example, If you want to add two ports to four tagged VLANs, previously, this required four commands:

```
configure vlan red add ports 2, 10 tagged
configure vlan blue add ports 2, 10 tagged
configure vlan green add ports 2, 10 tagged
configure vlan orange add ports 2, 10 tagged
```

Assuming the tags for the VLANs are 100, 200, 300, 400 respectively this configuration can be accomplished with a single command:

```
configure vlan 100,200,300,400 add ports 2, 10 tagged
```

Note

commands enhanced with VID list support operate in a “best effort” fashion. If one of the VIDs in a VID list do not exist the command is still executed for all of the VIDs in the list that do exist. No error or warning is displayed for the invalid VIDs unless all of the specified VIDs are in valid.

## Symbols

You may see a variety of symbols shown as part of the command syntax.

These symbols explain how to enter the command, and you do not type them as part of the command itself. The following table summarizes command syntax symbols you may see throughout this guide.

Note

ExtremeXOS software does not support the ampersand (&), left angle bracket (<), or right angle bracket (>), because they are reserved characters with special meaning in XML.

**Table 7: Command Syntax Symbols**

| Symbol | Description |
|---|---|
| square brackets [ ] | Enclose a required value or list of required arguments. One or more values or arguments can be specified. For example, in the syntax<br>`disable port [port_list \| all]`<br>you must specify either specific ports or all for all ports when entering the command. Do not type the square brackets. |
| vertical bar \| | Separates mutually exclusive items in a list, one of which must be entered. For example, in the syntax<br>`configure snmp add community [readonly \| readwrite] alphanumeric_string`<br>you must specify either the read or write community string in the command. Do not type the vertical bar. |
| braces { } | Enclose an optional value or a list of optional arguments. One or more values or arguments can be specified. For example, in the syntax<br>`reboot {time month day year hour min sec } {cancel} {msm slot_id} {slot slot-number \| node-address node-address \| stack-topology {as-standby} }`<br>You can specify either a particular date and time combination, or the keyword **cancel** to cancel a previously scheduled reboot. (In this command, if you do not specify an argument, the command will prompt, asking if you want to reboot the switch now.) Do not type the braces. |

# Port Numbering

The ExtremeXOS software runs on both stand-alone and modular switches, and the port numbering scheme is slightly different on each.

Note

The keyword **all** acts on all possible ports; it continues on all ports even if one port in the sequence fails.

## Stand-alone Switch Numerical Ranges

On Summit family switches, the port number is simply noted by the physical port number.

Separate the port numbers by a dash to enter a range of contiguous numbers, and separate the numbers by a comma to enter a range of non-contiguous numbers:

- x-y—Specifies a contiguous series of ports on a stand-alone switch.
- x,y—Specifies a non-contiguous series of ports on a stand-alone switch.
- x-y,a,d—Specifies a contiguous series of ports and a non-contiguous series of ports on a stand-alone switch.

## Modular Switch and SummitStack Numerical Ranges

On a modular switches and SummitStack switches, the port number is a combination of the slot number and the port number.

The nomenclature for the port number is as follows: `slot:port`

For example, if an I/O module that has a total of four ports is installed in slot 2 of the chassis, the following ports are valid:

- 2:1
- 2:2
- 2:3
- 2:4

You can also use wildcard combinations (*) to specify multiple modular slot and port combinations.

The following wildcard combinations are allowed:

- slot:*—Specifies all ports on a particular I/O module.
- slot:x-slot:y—Specifies a contiguous series of ports on a particular I/O module.
- slot:x-y—Specifies a contiguous series of ports on a particular I/O module.
- slota:x-slotb:y—Specifies a contiguous series of ports that begin on one I/O module or SummitStack node and end on another node.

## Stacking Port Numerical Ranges

On a SummitStack, a stacking port number is a combination of the slot number and the stacking port number shown near the connector on the back of the Summit family switch.

`slot:port`

These numbers are context-specific. For example, while the front-panel port 2:1 on a Summit X440 is a 10/100/1000 Ethernet port, the stacking port 2:1 is a 10Gb port on the rear panel of the X440 that has been marked as "Stacking Port 1." When no context is given, port 2:1 refers to a front-panel port on the Summit family switch (the 10Gb ports on, for example, a XGM2-2xn option card are considered front-panel ports in this context).

The use of wildcards and ranges for stacking ports is the same as described in Modular Switch and SummitStack Numerical Ranges on page 180.

## Line-Editing Keys

The following table describes the line-editing keys available using the CLI.

**Table 8: Line-Editing Keys**

| Key(s) | Description |
|---|---|
| Left arrow or **[Ctrl] + B** | Moves the cursor one character to the left. |
| Right arrow or **[Ctrl] + F** | Moves the cursor one character to the right. |
| **[Ctrl] + H** or Backspace | Deletes character to left of cursor and shifts remainder of line to left. |
| **[Delete]** or **[Ctrl] + D** | Deletes character under cursor and shifts remainder of line to left. |
| **[Ctrl] + K** | Deletes characters from under cursor to end of line. |
| **[Insert]** | Toggles on and off. When toggled on, inserts text and shifts previous text to right. |
| **[Ctrl] + A** | Moves cursor to first character in line. |
| **[Ctrl] + E** | Moves cursor to last character in line. |
| **[Ctrl] + L** | Clears screen and moves cursor to beginning of line. |
| **[Ctrl] + P** or Up arrow | Displays previous command in command history buffer and places cursor at end of command. |
| **[Ctrl] + N** or Down arrow | Displays next command in command history buffer and places cursor at end of command. |
| **[Ctrl] + U** | Clears all characters typed from cursor to beginning of line. |
| **[Ctrl] + W** | Deletes previous word. |
| **[Ctrl] + C** | Interrupts the current CLI command execution. |

## Viewing Command History

The ExtremeXOS software stores the commands you enter. You can display a list of these commands you have entered by typing the `history` command.

## Common Commands

This section discusses common commands you can use to manage the switch.

Commands specific to a particular feature may also be described in other chapters of this guide. For a detailed description of the commands and their options, see the *ExtremeXOS 16.2 Command Reference Guide*.

**Table 9: Common Commands**

| Command | Description |
|---|---|
| `clear session [history \| sessId \| all]` | Terminates a Telnet or SSH2 session from the switch. |
| `configure account` | Configures a user account password.<br>Passwords can have a minimum of 0 character and can have a maximum of 32 characters. Passwords are case-sensitive. User names are not case-sensitive. |
| `configure banner` | Configures the banner string. You can configure a banner to be displayed before login or after login. You can enter up to 24 rows of 79-column text that is displayed before the login prompt of each session. |
| `configure ports port_list {medium [copper \| fiber]} auto off speed speed duplex [half \| full]` | Manually configures the port speed and duplex setting of one or more ports on a switch. |
| `configure slot slot module module_type` | Configures a slot for a particular I/O module card.<br>**Note:** This command is available only on modular switches. |
| `configure ssh2 key {pregenerated}` | Generates the SSH2 host key.<br>You must install the SSH software module in addition to the base image to run SSH. |
| `configure sys-recovery-level [all \| none]` | Configures a recovery option for instances where an exception occurs in ExtremeXOS software. |
| `configure time month day year hour min sec` | Configures the system date and time. The format is as follows:<br>`mm dd yyyy hh mm ss`<br>The time uses a 24-hour clock format. You cannot set the year earlier than 2003 or past 2036. |
| `configure timezone` | Configures the time zone information to the configured offset from GMT time. The format of **`GMT_offset`** is ± minutes from GMT time. The **`autodst`** and **`noautodst`** options enable and disable automatic Daylight Saving Time change based on the North American standard.<br>Additional options are described in the GMT Offsets topic. |
| `configure [ { vlan } vlan_name \| vlan vlan_id ] ipaddress [ipaddress {ipNetmask } \| ipv6-link-local \| {eui64} ipv6_address_mask]` | Configures an IP address and subnet mask for a *VLAN*. |

**Table 9: Common Commands (continued)**

| Command | Description |
|---|---|
| `create account` | Creates a user account. This command is available to admin-level users and to users with *RADIUS (Remote Authentication Dial In User Service)* command authorization. The username is between 1 and 32 characters and is not case-sensitive. The password is between 0 and 32 characters and is case-sensitive. |
| `create vlan [ { vlan } vlan_name \| vlan vlan_list ] {description vlan-description } {vr name}` | Creates a VLAN. |
| `delete account name` | Deletes a user account. |
| `delete vlan [ { vlan } vlan_name \| vlan vlan_list ]` | Deletes a VLAN. |
| `disable bootp vlan [vlan \| all]` | Disables BOOTP for one or more VLANs. |
| `disable cli prompting` | Disables CLI prompting for the session. |
| `disable cli-config-logging` | Disables logging of CLI commands to the Syslog. |
| `disable clipaging` | Disables pausing of the screen display when a show command output reaches the end of the page. |
| `disable idletimeout` | Disables the timer that disconnects all sessions. After being disabled, console sessions remain open until the switch is rebooted or until you log off. Telnet sessions remain open until you close the Telnet client. SSH2 sessions time out after 61 minutes of inactivity. |
| `disable port [port_list \| all]` | Disables one or more ports on the switch. |
| `disable ssh2` | Disables SSH2 Telnet access to the switch. |
| `disable telnet` | Disables Telnet access to the switch. |
| `enable bootp vlan [vlan \| all]` | Enables BOOTP for one or more VLANs. |
| `enable cli-config-logging` | Enables the logging of CLI configuration commands to the Syslog for auditing purposes. The default setting is disabled. |
| `enable clipaging` | Enables pausing of the screen display when show command output reaches the end of the page. The default setting is enabled. |
| `enable idletimeout` | Enables a timer that disconnects all sessions (Telnet, SSH2, and console) after 20 minutes of inactivity. The default setting is enabled. |
| `enable license {software} key` | Enables a particular software feature license. Specify *key* as an integer. The command `unconfigure switch {all}` does not clear licensing information. This license cannot be disabled once it is enabled on the switch. |

**Table 9: Common Commands (continued)**

| Command | Description |
|---|---|
| `enable ssh2 {access-profile [access_profile \| none]} {port tcp_port_number} {vr [vr_name \| all \| default]}` | Enables SSH2 sessions. By default, SSH2 is disabled. When enabled, SSH2 uses TCP port number 22. |
| `enable telnet` | Enables Telnet access to the switch. By default, Telnet uses TCP port number 23. |
| `history` | Displays the commands entered on the switch. |
| `show banner {after-login \| before-login}` | Displays the user-configured banner. |
| `unconfigure switch {all}` or `unconfigure switch erase` **`all`** \| **`nvram`** | Resets all switch parameters (with the exception of defined user accounts, and date and time information) to the factory defaults. If you specify the keyword **`all`**, the switch erases the currently selected configuration image in flash memory and reboots. As a result, all parameters are reset to default settings. |

## Using Safe Defaults Mode

When you take your switch from the box and set it up for the first time, you set the safe defaults mode. You should use the safe defaults mode, which disables Telnet and *SNMP*. All ports are enabled in the factory default setting; you can choose to have all unconfigured ports disabled on reboot using the interactive questions.

After you connect to the console port of the switch, or after you run `unconfigure switch {all}` or `configure safe-default-script`, you can change management access to your device to enhance security.

1. Connect the console and log in to the switch.

   You are prompted with an interactive script that specifically asks if you want to disable Telnet and SNMP.

2. Follow the prompts and set your access preferences.

   Note
   In ExtremeXOS 16.1 and later, an enhanced security mode was added as an option to the startup script. If this is selected, all default SNMP users and communities will be deleted.

   ```
   This switch currently has all management methods enabled for
   convenience reasons. Please answer these questions about the security
   settings you would like to use. You may quit and accept the default
   settings by entering 'q' at any time. The switch offers an enhanced
   security mode. Would you like to read more, and have the choice to
   enable this enhanced security mode? [y/N/q]: Yes Enhanced security
   mode configures the following defaults: * Disable Telnet server. *
   Disable HTTP server. * Disable SNMP server. * Remove all factory
   default SNMP users & community names. * Remove all factory default
   login accounts. * Force creation of a new admin (read-write) account.
   * Force setting of failsafe username & password. * Lockout accounts
   for 5 minutes after 3 consecutive login failures. * Plaintext password
   entry will not be allowed. * Generate an event when the logging memory
   buffer exceeds 90% of capacity. * Only admin privilege accounts are
   permitted to run "show log". * Only admin privilege accounts are
   permitted to run "show diagnostics". Would you like to use this
   enhanced security mode? [Y/n/q]:
   ```

3. Reboot the switch.

# Configuring Management Access

## Account Access Levels

ExtremeXOS software supports two levels of management: user and administrator .

In addition to the management levels, you can optionally use an external *RADIUS* server to provide CLI command authorization checking for each command. For more information on RADIUS, see Security on page 859.

Note
CLI commands are sent to the RADIUS server unencrypted. Sensitive information entered into the CLI could be seen by either internal or external third parties.

### *User Accounts*

A user-level account has viewing access to all manageable parameters. Users cannot access:

- User account database
- *SNMP* community strings

A person with a user-level account can use the `ping` command to test device reachability and change the password assigned to the account name.

If you have logged on with user privileges, the command line prompt ends with a (>) sign. For example:
`BD-1.2 >`

*Administrator Accounts*

Administrator-level accounts can view and change all switch parameters.

With this privilege level, you can also add and delete users, as well as change the password associated with any account name. To erase the password, use the `unconfigure switch all` command.

An administrator can disconnect a management session that has been established by way of a Telnet connection. If this occurs, the user logged on through the Telnet connection is notified that the session has been terminated.

If you log on with administrator privileges, the command line prompt ends with a pound or hash (#) sign.

For example: `BD-1.18 #`

*Lawful Intercept Account*

The Lawful Intercept account can log in to a session and execute lawful intercept commands on the switch. The commands provide for configuration consists of dynamic ACLs and a mirror-to port to direct traffic to a separate device for analysis. The lawful intercept login session, session-related events, and the ACLs and mirror instance are not visible to, or modifiable by, any other user (administrative or otherwise).

No lawful intercept configuration is saved in the configuration file, and it must be reconfigured in the case of a system reboot.

Other important feature information:

- An administrative user can create and delete a single local account having the lawful intercept privilege and user privileges, but not administrative privileges, and can set its initial password.
- The lawful intercept user is required to change the password (for the single lawful intercept-privileged account) upon logging in for the first time.
- The password for the lawful intercept account can only be changed by the lawful intercept user and cannot be changed by an administrative user.
- The `show accounts` command displays the existence of the lawful intercept account, but does not display any related statistics.
- The `show configuration` command does not display the lawful intercept account.
- The `show session {{detail} {sessID}} {history}` command does not display any lawful intercept user information. The *EMS (Event Management System)* events normally associated with logging in and out are suppressed, and do not occur relative to logging in and out of the lawful intercept account.
- The EMS events normally associated with the `enable cli-config-logging` command are suppressed, and do not occur relative to a lawful intercept user session.
- The lawful intercept user can create and delete non-permanent dynamic ACLs with the mirror action only. The lawful intercept user cannot create or delete any other ACLs.
- The `show access-list` command does not display any Lawful Intercept user-created ACLs to a non-lawful intercept user.

- The lawful intercept user-created ACLs are not accessible for any use by a non-lawful intercept user (specifically through the `configure access-list add` or `configure access-list delete` commands).
- The lawful intercept user can only create or delete one (non-permanent) mirror instance with which to bind the lawful intercept user-created ACLs and specify the mirror-to port.

## Configure Banners

You can add a banner to give users helpful information before or after logging in. You can configure the following types of CLI session banners:

- A banner for a session that displays before login.
- A banner for a session that displays after login.

When no optional parameters are specified, the command configures a banner for a CLI session that displays before login. A CLI banner can have a maximum size of 24 rows with 79 columns of text.

- To add a banner to your switch:

  Issue the `configure banner` command. When you specify the **`acknowledge`** parameter, users must press a key to get the login prompt.

  This configures the banner string to be displayed for CLI screens.
- To clear a configured banner:

  Use the `unconfigure banner { after-login | before-login }` command.
- To disable the acknowledgement feature (which forces the user to press a key before the login screen displays):

  Issue the `configure banner` command, omitting the **`acknowledge`** parameter.
- To display the banners that are configured on the switch:

  Issue the `show banner { after-login | before-login }` command.

## Startup Screen and Prompt Text

Once you log into the switch, the system displays the startup screen.

```
login: admin
password:

ExtremeXOS
Copyright (C) 1996-2015 Extreme Networks.
All rights reserved.
This product is protected by one or more US
patents listed at
http://www.extremenetworks.com/patents
along with their foreign counterparts.
================================================================================

Press the <tab> or '?' key at any time for completions.
Remember to save your configuration changes.

* <switchname>.1 #
```

You must have an administrator-level account to change the text of the prompt. The prompt text is taken from the *SNMP* sysname setting.

The number that follows the period after the switch name indicates the sequential line of the specific command or line for this CLI session.

If an asterisk (*) appears in front of the command line prompt, it indicates that you have outstanding configuration changes that have not been saved.

For example: `* BD-1.19 #`

If you have logged on with administrator capabilities, the command line prompt ends with a (#) sign.

For example: `BD-1.18 #`

If you have logged on with user capabilities, the command line prompt ends with a (>) sign.

For example: `BD-1.2 >`

Using the system recovery commands (refer to Getting Started for information on system recovery), you can configure either one or more specified slots on a modular switch or the entire stand-alone switch to shut down in case of an error. If you have configured this feature and a hardware error is detected, the system displays an explanatory message on the startup screen. The message is slightly different, depending on whether you are working on a modular switch or a stand-alone switch.

The following sample shows the startup screen if any of the slots in a modular switch are shut down as a result of the system recovery:

```
configuration: login: admin
password:
ExtremeXOS Copyright (C) 2000-2006 Extreme Networks. All rights reserved.
Protected by US Patent Nos: 6,678,248; 6,104,700; 6,766,482; 6,618,388;
6,034,957; 6,859,438; 6,912,592; 6,954,436; 6,977,891; 6,980,550; 6,981,174;
7,003,705; 7,01 2,082.
================================================================================
Press the <tab> or '?' key at any time for completions.
Remember to save your configuration changes.
The I/O modules in the following slots are shut down: 1,3
Use the "clear sys-recovery-level" command to restore I/O modules !
BD-8810.1 #
```

When an exclamation point (!) appears in front of the command line prompt, it indicates that one or more slots or the entire stand-alone switch are shut down as a result of your system recovery configuration and a switch error. (Refer to Setting the System Recovery Level on page 447 and Understanding the System Health Checker on page 444 for complete information on system recovery and system health check features.)

The following sample shows the startup screen if a stand-alone switch is shut down as a result of the system recovery configuration:

```
login: admin
password:
ExtremeXOS Copyright (C) 2000-2006 Extreme Networks. All rights reserved.
Protected by US Patent Nos: 6,678,248; 6,104,700; 6,766,482; 6,618,388; 6,034,957;
6,859,438; 6,912,592; 6,954,436; 6,977,891; 6,980,550; 6,981,174; 7,003,705; 7,01 2,082.
================================================================================
Press the <tab> or '?' key at any time for completions.
Remember to save your configuration changes.
All switch ports have been shut down.
Use the "clear sys-recovery-level" command to restore all ports.
switch #
```

## Default Accounts

By default, the switch is configured with two accounts. ExtremeXOS 15.7.1 added the ability to disable all default accounts ("admin" and "user").

**Table 10: Default Accounts**

| Account Name | Access Level |
|---|---|
| admin | This user can access and change all manageable parameters. However, the user may not delete all admin accounts. |
| user | This user can view (but not change) all manageable parameters, with the following exceptions:<br>• This user cannot view the user account database.<br>• This user cannot view the *SNMP* community strings. |

## Creating a Management Account

An account can be disabled or enabled locally using read/write access. Even all administrative privileged accounts and user privileged accounts can be disabled. Lawful-Intercept account will be disabled under user privileged option. When all administrative accounts will be disabled locally, a warning will be shown to use failsafe account, if necessary.

1. Log in to the switch as admin.
2. At the password prompt, press **[Enter]**, or enter the password that you have configured for the admin account.
3. Run the `create account [admin | user] account-name {encrypted password}` command to add a new user.

   If you do not specify a password or the keyword **`encrypted`**, you are prompted for one. Passwords are case-sensitive.

   If you do not want a password associated with the specified account, press **[Enter]** twice.

   User-created account names are not case-sensitive.

### *Viewing Accounts*

You can view all accounts. To view the accounts that have been created, you must have administrator privileges. Run the `show accounts` command.

### *Deleting an Account*

You can remove accounts that should no longer exist, but you must have administrator privileges. To delete an account, run the `delete account` command.

## Authenticating Management Sessions through the Local Database

You can use a local database on each switch to authenticate management sessions. An account can be disabled or enabled locally using read/write access. Even all administrative privileged accounts and user privileged accounts can be disabled. Lawful-Intercept account will be disabled under user privileged option. When all administrative accounts will be disabled locally, a warning will be shown to use failsafe account, if necessary.

This enable/disable command affects the following North Bound Interfaces (NBIs) in mgmt-access realm:

- console
- TELNET
- SSH
- HTTP
- XML

The local database stores user names and passwords and helps to ensure that any configuration changes to the switch can be done only by authorized users.

You can increase authentication security using Secure Shell 2 (SSH2). SSH2 provides encryption for management sessions. For information about SSH2, see #unique_63.

## Failsafe Accounts

The failsafe account is last possible method to access your switch.

This account is never displayed by the `show accounts` command, but it is always present on the switch. To display whether the user configured a username and password for the failsafe account, or to show the configured connection-type access restrictions, use the following command: `show failsafe account`.

The failsafe account has admin access level.

To configure the account name and password for the failsafe account, use the following command:

`configure failsafe-account {[deny | permit] [all | control | serial | ssh {vr vr-name} | telnet {vr vr-name}]}`

When you use the command with no parameters, you are prompted for the failsafe account name and prompted twice to specify the password for the account.

For example:

```
BD-8810.1 # configure failsafe-account
enter failsafe user name: blue5green
enter failsafe password:
enter password again:
BD-10808.2
```

When you use the command with the permit or deny parameter, the connection-type access restrictions are altered as specified. For example:

```
BD-8810.1 # configure failsafe-account deny all
BD-8810.2 # configure failsafe-account permit serial
```

The failsafe account is immediately saved to NVRAM. On a modular switch, the failsafe account is saved to both MSM/MMs' NVRAMs if both are present. On a SummitStack, the failsafe account is saved in the NVRAM of every node in the active topology.

Note

On a SummitStack, when the `synchronize stacking {node-address` *`node-address`* `| slot` *`slot-number`* `}` command is used, the failsafe account is transferred from the current node to the specified nodes in the stack topology.

You do not need to provide the existing failsafe account information to change it.

Note

The information that you use to configure the failsafe account cannot be recovered by Extreme Networks. Technical support cannot retrieve passwords or account names for this account. Protect this information carefully.

### Accessing the Switch using Failsafe Account

You can access your switch using the failsafe account.

1. Connect to the switch using one of the (configured) permitted connection types.
2. At the switch login prompt, carefully enter the failsafe account name.

   If you enter an erroneous account name, you cannot re-enter the correct name. In that case, press **[Enter]** until you get a login prompt and then try again.
3. When prompted, enter the password.

## Managing Passwords

When you first access the switch, you have a default account.

You configure a password for your default account. As you create other accounts (see Creating a Management Account on page 29), you configure passwords for those accounts.

The software allows you to apply additional security to the passwords. You can enforce a specific format and minimum length for the password. Additionally, you can age out the password, prevent a user from employing a previously used password, and lock users out of the account after three consecutive failed login attempts.

You can change the password to an encrypted password after you create an account.

## Applying a Password to the Default Account

Default accounts do not have passwords assigned to them. Passwords can have a minimum of zero and a maximum of 32 characters. (If you specify the format of passwords using the `configure account password-policy char-validation` command, the minimum is eight characters.)

Note
Passwords are case-sensitive. User-created account names are not case-sensitive.

1. Log in to the switch using the name admin or user.
2. At the password prompt, press **[Enter]**.
3. Add a default admin password of green to the admin account or blue to the user account.

```
configure account admin green
configure account user blue
```

Note
If you forget your password while logged out of the CLI, you can use the bootloader to reinstall a default switch configuration, which allows access to the switch without a password. Note that this process reconfigures all switch settings back to the initial default configuration.

## Applying Security to Passwords

You can increase the security of your system by enforcing password restrictions, which will make it more difficult for unauthorized users to access your system. You can specify that each password must include at least two characters of each of the following four character types:

- Upper-case A-Z
- Lower-case a-z
- 0-9
- !, @, #, $, %, ^, *, (, )

You can enforce a minimum length for the password and set a maximum time limit, after which the password will not be accepted.

By default, the system terminates a session after the user has three consecutive failed login attempts.

The user may then launch another session (which would also terminate after three consecutive failed login attempts). To increase security, you can lock users out of the system entirely after three failed consecutive login attempts.

After the user's account is locked out (using the configure account password-policy lockout-on-login-failures command), it must be re-enabled by an administrator.

- To set character requirements for the password, use the following command:
  `configure account [all | name] password-policy char-validation [none | all-char-groups]`
- To set a minimum length for the password, use the following command:
  `configure account [all | name] password-policy min-length [num_characters | none]`

- To age out the password after a specified time, use the following command:

  `configure account [all | name] password-policy max-age [num_days | none]`

- To block users from employing previously used passwords, use the following command:

  `configure account [all | name] password-policy history [num_passwords | none]`

- To disable an account after three consecutive failed login attempts, use the following command:

  `configure account [all | name] password-policy lockout-on-login-failures [on | off]`

  Note

  If you are not working on SSH, you can configure the number of failed logins that trigger lockout, using the `configure cli max-failed-logins num-of-logins` command. (This command also sets the number of failed logins that terminate the particular session.)

- To re-enable a locked-out account, use the following command:

  `clear account [all | name] lockout`

  Selecting the **all** option affects the setting of all existing and future new accounts.

*Hash Algorithm for Account Passwords*

As of ExtremeXOS 16.1, SHA-256 is hash for local passwords. This hash is visible in both the XML config file (.cfg) and ASCII (.xsf). All existing users (created with older software) are still recognized and their *MD5 (Message-Digest algorithm 5)* hashes can be verified.

However, if a new user is created or a password is changed, it will use the SHA-256 hash. After a downgrade, older software will not be able to validate users with the SHA-256 hash. Upgrading will not automatically change the hash for existing users.

*Removal of Cleartext Passwords*

As of ExtremeXOS 16.1, all passwords, secrets, or keys are not shown in the clear. The software does not emit those passwords in any commands or display them as part of the device configuration.

A new mode of operation for the CLI requires prompting (with no echo) for all passwords, secrets, or keys. The command `configure cli password prompting-only` controls this mode.

Each CLI command with password arguments is modified to use the new mode (designated with flags="prompting-only" in the CLI syntax attribute specification). Then prompting must be handled by that command.

## Timed Lockout

As of ExtremeXOS 16.1, this feature adds the option to disable an account for a configurable period of time after consecutive failed logins. After the configured duration elapses a disabled account is re-enabled automatically. The configurable period of lockout time ranges from 1 minute to 1 hour. The configurable number of the consecutive failed attempts ranges from 1 to 10.

Prior to ExtremeXOS 16.1, the failsafe account was never locked out. Also, an admin acccount could only be locked out only if there is at least one other admin account that is not locked out. The intent is to prevent ensure the box is not ever completely locked out.

This feature augments this behavior in two ways:

- The failsafe account can now be locked out provided that the lockout is timed.
- All admin accounts can now be locked out provided that at least one is timed.

The feature applies to Telnet/SSH/Console/Https/Http.

### Displaying Passwords

To display the accounts and any applied password security, use the following command:

- To display accounts and passwords, use the following command:

  ```
  show accounts password-policy
  ```

- To display which accounts can be locked out, use the following command:

  ```
  show accounts
  ```

## Accessing Both MSM/MM Console Ports--Modular Switches Only

You can access either the primary or the backup MSM/MM regardless of which console port you are connected to by running:

```
telnet msm [a | b]
```

## Accessing an Active Node in a SummitStack

You can access any active node in a SummitStack from any other active node in the active topology by running:

```
telnet slot slot-number
```

## Domain Name Service Client Services

The Domain Name Service (DNS) client in ExtremeXOS software augments the following commands to allow them to accept either IP addresses or host names.

- telnet
- download bootrom
- download image
- ping
- traceroute
- configure radius server client-ip
- configure tacacs server client-ip
- create cfm domain dns md-level

The DNS client can resolve host names to both IPv4 and IPv6 addresses. In addition, you can use the nslookup utility to return the IP address of a host name.

Use the following command to specify up to eight DNS servers for use by the DNS client:

`configure dns-client add`

Use the following command to specify a default domain for use when a host name is used without a domain.

`configure dns-client default-domain`

For example, if you specify the domain xyz-inc.com as the default domain, then a command such as `ping accounting1` is taken as if it had been entered `ping accounting1.xyz-inc.com`.

## Checking Basic Connectivity

To check basic connectivity to your switch, use the `ping` and `traceroute` commands.

### Ping

The ping command enables you to send *ICMP (Internet Control Message Protocol)* echo messages to a remote IP device.

The ping command is available for both the user and administrator privilege levels.

`ping {vr` *`vr-name`*`}` {**`continuous`**|**`count`**|**`dont-fragment`**|**`interval`**|**`start-size`**|**`tos`**|**`ttl`**|**`udp`**} {**`mac`**|**`mpls`**|**`ipv4`**|**`ipv6`**} {**`from`**|**`with`**}

**Table 11: Ping Command Parameters**

| Parameter | Description |
|---|---|
| *`count`* | Specifies the number of ping requests to send. |
| *`start-size`* | Specifies the size, in bytes, of the packet to be sent, or the starting size if incremental packets are to be sent. |
| **`continuous`** | Specifies that UDP or ICMP echo messages are to be sent continuously. This option can be interrupted by pressing **[Ctrl]** + **C**. |
| *`end-size`* | Specifies an end size for packets to be sent. |
| **`udp`** | Specifies that the ping request should use UDP instead of ICMP. |
| **`dont-fragment`** | Sets the IP to not fragment the bit. |
| *`ttl`* | Sets the TTL value. |
| *`tos`* | Sets the TOS value. |
| *`interval`* | Sets the time interval between sending out ping requests. |
| *`vrid`* | Specifies the *virtual router (VR)* name to use for sending out the echo message. If not specified, *VR-Default* is used.<br><br>Note: User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document. |
| *`ipv4`* | Specifies IPv4 transport. |

**Table 11: Ping Command Parameters (continued)**

| Parameter | Description |
|---|---|
| *ipv6* | Specifies IPv6 transport.<br><br>Note: If you are contacting an IPv6 link local address, you must specify the *VLAN* you are sending the message from: `ping ipv6 link-local address %vlan_name host` . |
| *host* | Specifies a host name or IP address (either v4 or v6). |
| *from* | Uses the specified source address. If not specified, the address of the transmitting interface is used. |
| **with record-route** | Sets the traceroute information. |

If a ping request fails, the switch stops sending the request after three attempts. Press **[Ctrl]** + **C** to interrupt a ping request earlier. The statistics are tabulated after the ping is interrupted or stops.

Use the ipv6 variable to ping an IPv6 host by generating an ICMPv6 echo request message and sending the message to the specified address. If you are contacting an IPv6 link local address, you must specify the VLAN that you are sending the message from, as shown in the following example (you must include the % sign):

`ping ipv6 link-local address %vlan_name host`

## Traceroute

The `traceroute` command enables you to trace the path between the switch and a destination endstation.

`traceroute {vr vrid} {ipv4 host} {ipv6 host} {ttl number} {from from} {[port port] | icmp}`

vr

The name of the *VR*.

ipv4/ipv6

The transport.

from

Uses the specified source address in the *ICMP* packet. If not specified, the address of the transmitting interface is used.

host

The host of the destination endstation. To use the hostname, you must first configure DNS.

ttl

Configures the switch to trace the hops until the time-to-live has been exceeded for the switch.

port

Uses the specified UDP port number.

icmp

Uses ICMP echo messages to trace the routed path.

## Displaying Switch Information

You can display basic information about the switch by running the `show switch` command.

### Filtering the Output of Show Commands

The output from many show commands can be long and complicated, sometimes containing more information than you need at a given time.

The filter output display feature allows you to extract the output information from a show command that fits your needs.

The feature is a restricted version of a UNIX/Linux feature that uses a "pipe" character to direct the output of one command to be used as input for the next command.

It provides support for “piping” show command output to the display filter using the vertical bar (|) operator. (In the following command, it is the first vertical bar.) The display filter displays the output based on the specified filter keyword option and the text pattern entered. By selecting different filter options you can include or exclude all output that matches the pattern. You can also exclude all output until a line matches the pattern and then include all output beginning with that line.

In ExtremeXOS software, the resulting command is as follows:

`show` *`specific show command syntax`* | {**`include`** | **`exclude`** | **`begin`** } *`regexp`*

The following describes the command syntax:

| | |
|---|---|
| **`show`** *`specific show command syntax`* | State the command. For example: show ports. (This is followed by the vertical bar (|) when used as the pipe character.) |
| **`include`** | Display the lines that match the regular expression. |
| **`exclude`** | Do not display the lines that match the regular expression. |
| **`begin`** | Display all the lines starting with the first line that matches the regular expression. |
| *`regexp`* | The regular expression to match. Regular expressions are case-sensitive. Special characters in regular expressions such as [ ], ?, and * have special significance to the Linux shell and it is therefore common to specify your regular expression in quotes to protect it from the shell. |

*Flow control*

To display the status of “flow control” on the ports of a BlackDiamond 8810 switch, use the following command:

```
show ports 2:1-2 information detail | include "(Port | Flow Control)"
```

The output would resemble the following:

```
Port: 2:1
            Flow Control:   Rx-Pause: Enabled       Tx-Pause: Disabled
Priority Flow Control: Disabled
```

```
Port: 2:2
            Flow Control:   Rx-Pause: Enabled       Tx-Pause: Disabled
Priority Flow Control: Disabled
```

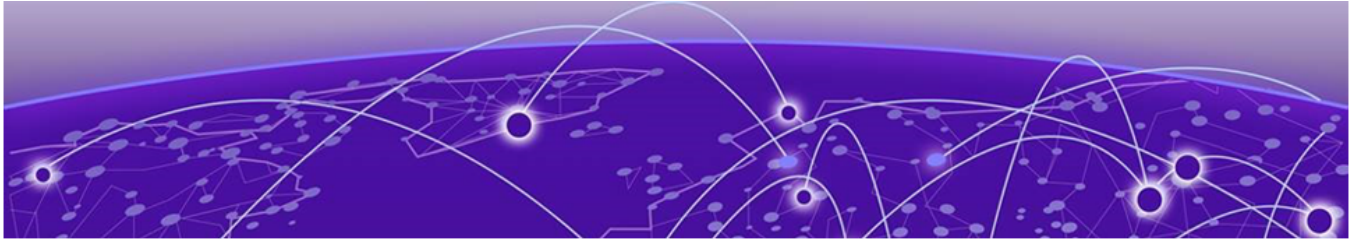
If the specified show command outputs a refreshed display, using the output display filter terminates the display without refreshing and a message is displayed to that effect.

This command is supported on most of the ExtremeXOS show commands. A few commands, for example, show tech-support, are not implemented in such a way as to make piping (filtering) possible.

The following table shows a summary of special characters.

**Table 12: Definition of Regular Expression Characters**

| Operator Type | Examples | Description |
|---|---|---|
| Literal characters match a character exactly | a A y 6 % @ | Letters, digits and many special characters match exactly |
| | \$ \^ \+ \\ \? | Precede other special characters with a \ to cancel their regex special meaning |
| | \n \t \r | Literal new line, tab, return |
| Anchors and assertions | ^ | Starts with |
| | $ | Ends with |
| Character groups any one character from the group | [aAeEiou] | Any character listed from [ to ] |
| | [^aAeEiou] | Any character except aAeEio or u |
| | [a-fA-F0-9] | Any hex character (0 to 9 or a to f) |
| | . | Any character at all |
| Counts apply to previous element | + | One or more ("some") |
| | * | Zero or more ("perhaps some") |
| | ? | Zero or one ("perhaps a") |
| Alternation | \| | Either, or |



# Managing the Switch



This chapter provides information about how to use your ExtremeXOS switch. Included you will find information about the ExtremeXOS Shell, system redundancy, power supply management, user authentication, Telnet, and hitless failover support, as well as *SNMP (Simple Network Management Protocol)* and *SNTP (Simple Network Time Protocol)* usage information.

## EXOS Switch Management Overview

This chapter describes how to use ExtremeXOS to manage the switch. It also provides details on how to perform the following various basic switch functions:

- Access the command line interface (CLI) by connecting a terminal (or workstation with terminal-emulation software) to the console port.
- Access the switch remotely using TCP/IP through one of the switch ports, or through the dedicated 10/100 unshielded twisted pair (UTP) Ethernet management port. Remote access includes:
  - Telnet using the CLI interface
  - Secure Shell (SSH2) using the CLI interface
  - *SNMP* access using Ridgeline™ or another SNMP manager

- Download software updates and upgrades. For more information, see Software Upgrade and Boot Options.

The switch supports the following number of concurrent user sessions:

- One console session—Two console sessions are available if two management modules are installed
- Eight shell sessions
- Eight Telnet sessions
- Eight Trivial File Transfer Protocol (TFTP) sessions
- Eight SSH2 sessions
- Six XML sessions

## Understanding the ExtremeXOS Shell

When you log in to ExtremeXOS from a terminal, a shell prompt is displayed.

At the prompt, input the commands you want to execute on the switch. After the switch processes and executes a command, the results are displayed on your terminal.

The shell supports ANSI, VT100, and XTERM terminal emulation and adjusts to the correct terminal type and window size. In addition, the shell supports UNIX-style page view for page-by-page command output capability.

By default, up to eight active shell sessions can access the switch concurrently; however, you can change the number of simultaneous, active shell sessions supported by the switch. You can configure up to 16 active shell sessions. Configurable shell sessions include both Telnet and SSH connections (not console CLI connections). If only eight active shell sessions can access the switch, a combination of eight Telnet and SSH connections can access the switch even though Telnet and SSH each support eight connections. For example, if you have six Telnet sessions and two SSH sessions, no one else can access the switch until a connection is terminated or you access the switch through the console.

If you configure a new limit, only new incoming shell sessions are affected. If you decrease the limit and the current number of sessions already exceeds the new maximum, the switch refuses only new incoming connections until the number of shell session drops below the new limit. Already connected shell sessions are not disconnected as a result of decreasing the limit.

Configure the number of shell sessions accepted by the switch, use the following command:

```
configure cli max-sessions
```

For more information about the line-editing keys that you can use with the ExtremXOS shell, see #unique_100.

## Using the Console Interface

You can access the switch as needed through the command line interface.

The switch is accessible using the following connectors:

- BlackDiamond X8 series: RJ-45 port for use with a rollover cable.

- BlackDiamond 8800 series and all Summit switches: 9-pin, RS-232 ports.

On a modular switch, the console port is located on the front of the management module (MSM/MM). On a stand-alone switch, the console port is located on the front panel.

Note

For more information on the console port pinouts, see the hardware installation guide for your switch.

After the connection is established, you will see the switch prompt and can now log in.

## Using the 10/100 or 10/100/1000 Ethernet Management Port

The management module provides a dedicated 10/100 Mbps or 10/100/1000 Mbps Ethernet management port. This port provides dedicated remote access to the switch using TCP/IP. It supports the following management methods:

- Telnet/SSH2 using the CLI interface
- *SNMP* access using ExtremeManagement, Ridgeline or another SNMP manager

The switch uses the Ethernet management port only for host operation, not for switching or routing. The TCP/IP configuration for the management port is done using the same syntax as used for *VLAN (Virtual LAN)* configuration. The VLAN management comes preconfigured with only the management port as a member. The management port is a member of the *virtual router (VR)* *VR-Mgmt*.

When you configure the IP address for the VLAN management, the address gets assigned to the primary MSM/MM. You can connect to the management port on the primary MSM/MM for any switch configuration. The management port on the backup MSM/MM is available only when failover occurs. If failover occurs, the primary MSM/MM relinquishes its role, the backup MSM/MM takes over, and VLAN management on the new primary MSM/MM acquires the IP address of the previous primary MSM/MM.

On a SummitStack, the master node is accessed using the management port primary IP address for other platforms. The primary IP address is acquired by the backup node when it becomes the master node due to a failover. You can also directly access any node in the stack using its alternate IP address if the node's management port is connected to your network.

- To configure the IP address and subnet mask for the VLAN mgmt, use the following command:

  `configure vlan mgmt ipaddress ip_address /subnet_mask`

- To configure the default gateway (you must specify VR-Mgmt for the management port and VLAN mgmt), use the following command:

  `configure iproute add default gateway { metric } {multicast | multicast-only | unicast | unicast-only} {vr vrname}`

  The following example configuration sets the management port IP address to 192.168.1.50, mask length of 25, and configures the gateway to use 192.168.1.1:

```
configure vlan mgmt ipaddress 192.168.1.50/25
configure iproute add default 192.168.1.1 vr vr-mgmt
```

For more information see Logging into a Stack on page 144.

# Using ExtremeManagement or Ridgeline to Manage the Network

Our NMSs are powerful yet easy-to-use application suites that facilitate the management of a network of Extreme Networks switches, as well as selected third-party switches.

These products offer a comprehensive set of network management tools that are easy to use from a client workstation running client software, or from a workstation configured with a web browser and the Java plug-in.

# Authenticating Users

ExtremeXOS provides three methods to authenticate users who log in to the switch: *RADIUS (Remote Authentication Dial In User Service)* client, TACACS+, and a local database of accounts and passwords.

Note

You cannot configure RADIUS and TACACS+ at the same time.

## RADIUS Client

Remote Authentication Dial In User Service (*RADIUS*, RFC 2865) is a mechanism for authenticating and centrally administrating access to network nodes.

The ExtremeXOS RADIUS client implementation allows authentication for Telnet or console access to the switch. For detailed information about RADIUS and configuring a RADIUS client, see Security on page 859.

## TACACS+

Terminal Access Controller Access Control System Plus (TACACS+) is a mechanism for providing authentication, authorization, and accounting on a central server, similar in function to the *RADIUS* client.

The ExtremeXOS version of TACACS+ is used to authenticate prospective users who are attempting to administer the switch. TACACS+ is used to communicate between the switch and an authentication database.

For detailed information about TACACS+ and configuring TACACS+, see Security on page 859.

## Management Accounts

ExtremeXOS supports two levels of management accounts (local database of accounts and passwords): user and administrator.

A user level account can view but not change all manageable parameters, with the exception of the user account database and *SNMP* community strings. An administrator level account can view and change all manageable parameters.

For detailed information about configuring management accounts, see Configuring Management Access on page 25.

# Using Telnet

ExtremeXOS supports the Telnet Protocol based on RFC 854.

Telnet allows interactive remote access to a device and is based on a client/server model. ExtremeXOS uses Telnet to connect to other devices from the switch (client) and to allow incoming connections for switch management using the CLI (server).

## Starting the Telnet Client

Ensure that the IP parameters described in Configuring Switch IP Parameters on page 44 are set up and then start an outgoing Telnet session.

Telnet is enabled and uses *VR-Mgmt* by default.

Note

Maximize the Telnet screen so that it correctly displays screens that automatically update.

1. Use Telnet to establish a connection to the switch.
2. Specify the IP address or host name of the device that you want to connect to.
   Check the user manual supplied with the Telnet facility if you are unsure of how to do this.
   After the connection is established, you see the switch prompt and you can log in. The same is true if you use the switch to connect to another host. From the CLI, you must specify the IP address or host name of the device that you want to connect to.
3. If the host is accessible and you are allowed access, you may log in.

For more information about using the Telnet client on the switch, see Connect to Another Host Using Telnet on page 43.

## About the Telnet Server

Any workstation with a Telnet facility should be able to communicate with the switch over a TCP/IP network using VT100 terminal emulation.

Up to eight active Telnet sessions can access the switch concurrently. If you enable the idle timer using the `enable idletimeout` command, the Telnet connection times out after 20 minutes of inactivity by default. If a connection to a Telnet session is lost inadvertently, the switch terminates the session within two hours.

The switch accepts IPv6 connections.

For information about the Telnet server on the switch, see the following sections:

- Configuring Telnet Access to the Switch on page 46
- Disconnecting a Telnet Session on page 47

## Connect to Another Host Using Telnet

You can Telnet from the current CLI session to another host. You can use Telnet to access either the primary or the backup MSM/MM regardless of which console port you are connected to. For more information see Starting the Telnet Client on page 43.

Run `telnet {vr vr_name} [host_name | remote_ip] {port}`.

User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document.

If the TCP port number is not specified, the Telnet session defaults to port 23. If the *VR* name is not specified, the Telnet session defaults to *VR-Mgmt*. Only VT100 emulation is supported.

## Configuring Switch IP Parameters

To manage the switch by way of a Telnet connection or by using an *SNMP* Network Manager, you must first configure the switch IP parameters.

### *Using a BOOTP or DHCP Server*

The switch contains a BOOTP and *DHCP (Dynamic Host Configuration Protocol)* client, so if you have a BOOTP or DHCP server in your IP network, you can have it assign IP addresses to the switch. This is more likely to be desirable on the switch's *VLAN* mgmt than it is on any other VLANs.

If you are using IP and you have a Bootstrap Protocol (BOOTP) server set up correctly on your network, you must provide the following information to the BOOTP server:

- Switch Media Access Control (MAC) address, found on the rear label of the switch
- IP address
- Subnet address mask (optional)

The switch does not retain IP addresses assigned by BOOTP or DHCP through a power cycle, even if the configuration has been saved. To retain the IP address through a power cycle, you must configure the IP address of the VLAN using the CLI or Telnet.

If you need the switch's MAC address to configure your BOOTP or DHCP server, you can find it on the rear label of the switch. Note that all VLANs configured to use BOOTP or DHCP use the same MAC address to get their IP address, so you cannot configure the BOOTP or DHCP server to assign multiple specific IP addresses to a switch depending solely on the MAC address.

- To enable the BOOTP or DHCP client per VLAN, use the following command:

  ```
  enable bootp vlan [ vlan_name | all]
  enable dhcp vlan [ vlan_name | all]
  ```

- To disable the BOOTP or DHCP client per VLAN, use the following command:

  ```
  disable bootp vlan [ vlan_name | all]
  disable dhcp vlan [ vlan_name | all]
  ```

- To view the current state of the BOOTP or DHCP client, use the following command:

  ```
  show dhcp-client state
  ```

**Note**

The ExtremeXOS DHCP client will discard the DHCP OFFER if the lease time is less than or equal to two seconds.

*Manually Configuring the IP Settings*

If you are using IP without a BOOTP server, you must enter the IP parameters for the switch in order for the SNMP Network Manager or Telnet software to communicate with the device.

1. Assign IP parameters to the switch.
   a. Log in to the switch with administrator privileges using the console interface.
   b. Assign an IP address and subnet mask to a VLAN.
   c. The switch comes configured with a default VLAN named default. To use Telnet or an SNMP Network Manager, you must have at least one VLAN on the switch, and that VLAN must be assigned an IP address and subnet mask. IP addresses are always assigned to each VLAN. The switch can be assigned multiple IP addresses (one for each VLAN).

      Note
      For information on creating and configuring VLANs, see VLANs on page 502.

2. Manually configure the IP settings.
   a. Connect a terminal or workstation running terminal emulation software to the console port, as detailed in Using the Console Interface on page 40.
   b. At your terminal, press **[Enter]** one or more times until you see the login prompt.
   c. At the login prompt, enter your user name and password. The user name is not case-sensitive; the password is case-sensitive. Ensure that you have entered a user name and password with administrator privileges.

      If you are logging in for the first time, use the default user name *admin* to log in with administrator privileges. For example: `login: admin`

      Administrator capabilities enable you to access all switch functions. The default user names have no passwords assigned.

      If you have been assigned a user name and password with administrator privileges, enter them at the login prompt.
   d. Enter the password when prompted.

      When you have successfully logged in to the switch, the command line prompt displays the name of the switch.
   e. Assign an IP address and subnetwork mask for the default VLAN by using the following command:

      `configure {vlan} vlan_name ipaddress [ipaddress {ipNetmask} | ipv6-link-local | {eui64} ipv6_address_mask]`

      For example:

      ```
      configure vlan default ipaddress 123.45.67.8 255.255.255.0
      ```

      The changes take effect immediately.

      Note
      As a general rule, when configuring any IP addresses for the switch, you can express a subnet mask by using dotted decimal notation or by using classless inter domain routing notation (CIDR). CIDR uses a forward slash plus the number of bits in the subnet mask. Using CIDR notation, the command identical to the previous example is: `configure vlan default ipaddress 123.45.67.8/24`.

3. Configure the default route for the switch using the following command:

   `configure iproute add default` *gateway* {*metric*} {**multicast | multicast-only | unicast | unicast-only**} {**vr** *vrname*}

   For example:

   ```
   configure iproute add default 123.45.67.1
   ```

4. Save your configuration changes so that they will be in effect after the next switch reboot.

   If you want to save your changes to the currently booted configuration, use the following command: `save`

   ExtremeXOS allows you to select or create a configuration file name of your choice to save the configuration to.

   a. If you want to save your changes to an existing or new configuration file, use the following command:

      `save configuration` {**primary** | **secondary** | *existing-config* | *new-config* | *as-script*}

5. When you are finished using the facility, log out of the switch by typing: `logout` or `quit`.

## Configuring Telnet Access to the Switch

By default, Telnet services are enabled on the switch and all virtual routers listen for incoming Telnet requests. The switch accepts IPv6 connections.

User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document..

The safe defaults mode runs an interactive script that allows you to enable or disable *SNMP*, Telnet, and switch ports. When you set up your switch for the first time, you must connect to the console port to access the switch. After logging in to the switch, you will enter into the safe defaults mode. Although SNMP, Telnet, and switch ports are enabled by default, the script prompts you to confirm those settings.

If you choose to keep the default setting for Telnet—the default setting is enabled—the switch returns the following interactive script:

```
Since you have chosen less secure management methods, please remember to increase the
security of your network by taking the following actions:
* change your admin password
* change your SNMP public and private strings
* consider using SNMPv3 to secure network management traffic
```

For more detailed information about safe defaults mode, see Using Safe Defaults Mode on page 24.

- To configure the *VR* from which you receive a Telnet request, use the following command:

  `configure telnet vr [`**all** | **default** | *vr_name*]

- To change the default, use the following command:

  `configure telnet port [`*portno* | **default**]

  The range for the port number is 1–65535. The following TCP port numbers are reserved and cannot be used for Telnet connections: 22, 80, and 1023. If you attempt to configure a reserved port, the switch displays an error message.

*Viewing Telnet Information*

To display the status of Telnet, including the current TCP port, the *VR* used to establish a Telnet session, and whether ACLs are controlling Telnet access, run the `show management` command.

*Disabling and Enabling Telnet*

Note
You must be logged in as an administrator to configure the virtual router(s) used by Telnet and to enable or disable Telnet.

- You can choose to disable Telnet by using the following command:

  ```
  disable telnet
  ```

- To re-enable Telnet on the switch, use the following command:

  ```
  enable telnet
  ```

## Disconnecting a Telnet Session

A person with an administrator level account can disconnect a Telnet management session.

1. Log in to the switch with administrator privileges.
2. Determine the session number of the session you want to terminate.

   ```
   show session {{detail} {sessID}} {history}
   ```

3. Terminate the session.

   ```
   clear session [history |sessId | all]
   ```

The user logged in by way of the Telnet connection is notified that the session has been terminated.

## Access Profile Logging for Telnet

By default, Telnet services are enabled on the switch.

The access profile logging feature allows you to use an *ACL (Access Control List)* policy file or dynamic ACL rules to control access to Telnet services on the switch. When access profile logging is enabled for Telnet, the switch logs messages and increments counters when packets are denied access to Telnet. No messages are logged for permitted access.

You can manage Telnet access using one (not both) of the following methods:

- Create and apply an ACL policy file.
- Define and apply individual ACL rules.

One advantage of ACL policy files is that you can copy the file and use it on other switches. One advantage to applying individual ACL rules is that you can enter the rules at the CLI command prompt, which can be easier than opening, editing, and saving a policy file.

### *ACL Match Conditions and Actions*

The ACLs section describes how to create *ACL* policies and rules using match conditions and actions. Access profile logging supports the following match conditions and actions:

- Match conditions
  - Source-address—IPv4 and IPv6
- Actions
  - Permit
  - Deny

If the ACL is created with more match conditions or actions, only those listed above are used for validating the packets. All other conditions and actions are ignored.

The source-address field allows you to identify an IPv4 address, IPv6 address, or subnet mask for which access is either permitted or denied.

If the *SNMP* traffic does not match any of the rules, the default behavior is deny.

### *Limitations*

Access profile logging for Telnet has the following limitations:

- Either policy files or *ACL* rules can be associated with Telnet, but not both at the same time.
- Only source-address match is supported.
- Access-lists that are associated with one or more applications cannot be directly deleted. They must be unconfigured from the application first and then deleted from the CLI.
- Default counter support is added only for ACL rules and not for policy files.

### *Managing ACL Policies for Telnet*

The ACLs section describes how to create *ACL* policy files.

1. To configure Telnet to use an ACL policy, use the following command:

   `configure telnet access-profile` *`profile_name`*
2. To configure Telnet to remove a previously configured ACL policy, use the following command:

   `configure telnet access-profile none`

**Note**

Do not also apply the policy to the access list. Applying a policy to both an access profile and an access list is neither necessary nor recommended.

### *Managing ACL Rules for Telnet*

Before you can assign an *ACL* rule to Telnet, you must create a dynamic ACL rule as described in ACLs.

1. To add or delete a rule for Telnet access, use the following command:

   `configure telnet access-profile [` *`access_profile`* `|` `[[`**`add`** *`rule`* `]` `[`**`first`** `|` `[[`**`before`** `|` **`after`**`]` *`previous_rule`*`]]]` `|` **`delete`** *`rule`* `|` **`none`** `]`
2. To display the access-list permit and deny statistics for an application, use the following command:

   `show access-list counters process [`**`snmp`** `|` **`telnet`** `|` **`ssh2`** `|` **`http`**`]`

### *Misconfiguration Error Messages*

The following messages can appear during configuration of policies or rules for the *SNMP* service:

| | |
|---|---|
| `Rule <rule> is already applied` | A rule with the same name is already applied to this service. |
| `Please remove the policy <policy> already configured, and then add rule <rule>` | A policy file is already associated with the service. You must remove the policy before you can add a rule. |
| `Rule <previous_rule> is not already applied` | The specified rule has not been applied to the service, so you cannot add a rule in relation to that rule. |
| `Rule <rule> is not applied` | The specified rule has not been applied to the service, so you cannot remove the rule from the service. |
| `Error: Please remove previously configured rule(s) before configuring policy <policy>` | A policy or one or more ACL rules are configured for the service. You must delete the remove the policy or rules from the service before you can add a policy. |

### *Sample ACL Policies*

The following are sample policies that you can apply to restrict Telnet access.

In the following example named MyAccessProfile.pol, the switch permits connections from the subnet 10.203.133.0 /24 and denies connections from all other addresses:

```
MyAccessProfile.pol
entry  AllowTheseSubnets {
    if  {
        source-address 10.203.133.0 /24;
    } then {
        permit;
    }
}
```

In the following example named MyAccessProfile.pol, the switch permits connections from the subnets 10.203.133.0 /24 or 10.203.135.0/24 and denies connections from all other addresses:

```
MyAccessProfile.pol
entry  AllowTheseSubnets {
    if match any {
        source-address 10.203.133.0 /24;
        source-address 10.203.135.0 /24;
    } then {
        permit;
    }
}
```

In the following example named MyAccessProfile_2.pol, the switch does not permit connections from the subnet 10.203.133.0 /24 but accepts connections from all other addresses:

```
MyAccessProfile_2.pol
entry dontAllowTheseSubnets {
    if {
        source-address 10.203.133.0 /24;
    } then {
        deny;
    }
```

```
}
entry  AllowTheRest {
    if  {
        ; #none specified
    } then {
        permit;
    }
}
```

In the following example named MyAccessProfile_2.pol, the switch does not permit connections from the subnets 10.203.133.0/24 or 10.203.135.0 /24 but accepts connections from all other addresses:

```
MyAccessProfile_2.pol
entry dontAllowTheseSubnets {
    if match any {
        source-address 10.203.133.0 /24;
        source-address 10.203.135.0 /24;
    } then {
        deny;
    }
}
entry  AllowTheRest {
    if  {
        ; #none specified
    } then {
        permit;
    }
}
```

## Using Secure Shell 2

Secure Shell 2 (SSH2) is a feature of the ExtremeXOS software that allows you to encrypt session data between a network administrator using SSH2 client software and the switch or send encrypted data from the switch to an SSH2 client on a remote system. Configuration, image, public key, and policy files can be transferred to the switch using the Secure Copy Protocol 2 (SCP2).

Note

Starting with ExtremeXOS 16.2, SSH2 no longer requires an xmod. SSH2 is included in the ExtremeXOS main software image.

The ExtremeXOS SSH2 switch application works with the following clients: Putty, SSH2 (version 2.x or later) from SSH Communication Security, and OpenSSH (version 2.5 or later).

The switch accepts IPv6 connections.

Up to eight active SSH2 sessions can run on the switch concurrently. If you enable the idle timer using the `enable idletimeout` command, the SSH2 connection times out after 20 minutes of inactivity by default. If you disable the idle timer using the `disable idletimeout` command, the SSH2 connection times out after 61 minutes of inactivity. If a connection to an SSH2 session is lost inadvertently, the switch terminates the session within 61 minutes.

For detailed information about SSH2, see Security on page 859.

## Access Profile Logging for SSH2

The access profile logging feature allows you to use an *ACL* policy file or dynamic ACL rules to control access to SSH2 services on the switch.

When access profile logging is enabled for SSH2, the switch logs messages and increments counters when packets are denied access to SSH2. No messages are logged for permitted access.

You can manage SSH2 access using one (not both) of the following methods:

- Create and apply an ACL policy file
- Define and apply individual ACL rules

One advantage of ACL policy files is that you can copy the file and use it on other switches. One advantage to applying individual ACL rules is that you can enter the rules at the CLI command prompt, which can be easier than opening, editing, and saving a policy file.

### *ACL Match Conditions and Actions*

The ACLs section describes how to create *ACL* policies and rules using match conditions and actions. Access profile logging supports the following match conditions and actions:

- Match conditions
  - Source-address—IPv4 and IPv6
- Actions
  - Permit
  - Deny

If the ACL is created with more match conditions or actions, only those listed above are used for validating the packets. All other conditions and actions are ignored.

The source-address field allows you to identify an IPv4 address, IPv6 address, or subnet mask for which access is either permitted or denied.

If the *SNMP* traffic does not match any of the rules, the default behavior is deny.

### *Limitations*

Access profile logging for SSH2 has the following limitations:

- Either policy files or *ACL*s can be associated with SSH2, but not both at the same time.
- Only source-address match is supported.
- Access-lists that are associated with one or more applications cannot be directly deleted. They must be unconfigured from the application first and then deleted from the CLI.
- Default counter support is added only for dynamic ACL rules and not for policy files.

### *Managing ACL Policies for SSH2*

The ACLs section describes how to create *ACL* policy files.

- To configure SSH2 to use an ACL policy, use the following command:

  `configure ssh2 access-profile profile_name`
- To configure SSH2 to remove a previously configured ACL policy, use the following command:

  `configure ssh2 access-profile none`

### *Managing ACL Rules for SSH2*

Before you can assign an *ACL* rule to HTTP, you must create a dynamic ACL rule as described in ACLs.

- To add or delete a rule for SSH2 access, use the following command:

  `configure ssh2 access-profile [ access_profile | [[add rule ] [first | [[before | after] previous_rule]]] | delete rule | none ]`

- To display the access-list permit and deny statistics for an application, use the following command:

  `show access-list counters process [snmp | telnet | ssh2 | http]`

### *Misconfiguration Error Messages*

The following messages can appear during configuration of policies or rules for the *SNMP* service:

| | |
|---|---|
| `Rule <rule> is already applied` | A rule with the same name is already applied to this service. |
| `Please remove the policy <policy> already configured, and then add rule <rule>` | A policy file is already associated with the service. You must remove the policy before you can add a rule. |
| `Rule <previous_rule> is not already applied` | The specified rule has not been applied to the service, so you cannot add a rule in relation to that rule. |
| `Rule <rule> is not applied` | The specified rule has not been applied to the service, so you cannot remove the rule from the service. |
| `Error: Please remove previously configured rule(s) before configuring policy <policy>` | A policy or one or more ACL rules are configured for the service. You must delete the remove the policy or rules from the service before you can add a policy. |

## Using the Trivial File Transfer Protocol

ExtremeXOS supports the Trivial File Transfer Protocol (TFTP) based on RFC 1350.

TFTP is a method used to transfer files from one network device to another. The ExtremeXOS TFTP client is a command line application used to contact an external TFTP server on the network. For example, ExtremeXOS uses TFTP to download software image files, switch configuration files, and ACLs from a server on the network to the switch.

Up to eight active TFTP sessions can run on the switch concurrently.

We recommend using a TFTP server that supports blocksize negotiation (as described in RFC 2348, TFTP Blocksize Option), to enable faster file downloads and larger file downloads.

Note
For better functionality, minimum block-size of 64 bytes is recommended.

For additional information about TFTP, see the following chapters:

- For information about downloading software image files, BootROM files, and switch configurations, see Software Upgrade and Boot Options on page 1522.
- For information about downloading *ACL* (and other) policy files, see Policy Manager on page 635.

- For information about using TFTP to transfer files to and from the switch, see Managing the Switch on page 39.
- For information about configuring core dump files and transferring the core dump files stored on your switch, see Troubleshooting on page 1557.

## TFTP Block-size Configuration

ExtremeXOS supports the TFTP client block-size option configuration based on RFC 2348, which can range from 8 octets to 65000 octets. The block-size refers to data octets and does not include TFTP header. This feature added the user configuration option for data bock-size:

- in the generic commands that are used for downloading/uploading image/configuration/log/core file etc.
- ranging from 24 bytes to 65000 bytes, taking into consideration the local/remote file name size and the current busy box TFTP client support limits.
- to support larger TFTP data packets exceeding normal MTU especially on a front-panel port in case of in-band management, enable jumbo frames on that port. Please refer "Jumbo Frames" for the usage and its functional restrictions that affect TFTP data packet transfers.

If you do not specify block-size, the default is 1400 bytes.

## Connecting to Another Host Using TFTP

You can TFTP from the current CLI session to another host to transfer files.

1. Run the `tftp` command:

   `tftp [host-name | ip-address] {-v vr_name} [-g | -p] [{-l [ local-file-internal | local-file-memcard | local-file} {-r remote-file} | {-r remote-file} {-l [ local-file-internal | local-file-memcard | local-file]}]`

   Note

   User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document.

   The TFTP session defaults to port 69. If you do not specify a *VR*, *VR-Mgmt* is used.

   For example, to connect to a remote TFTP server with an IP address of 10.123.45.67 and “get” or retrieve an ExtremeXOS configuration file named XOS1.cfg from that host, use the following command:

   ```
   tftp 10.123.45.67 -g -r XOS1.cfg
   ```

   When you “get” the file through TFTP, the switch saves the file to the primary MSM/MM. If the switch detects a backup MSM/MM in the running state, the file is replicated to the backup MSM/MM.

2. To view the files you retrieved, enter the `ls` command at the command prompt.

   In addition to the `tftp` command, the following two commands are available for transferring files to and from the switch:

   ```
   tftp get [host-name | ip-address] {-vr vr_name} [{ local-file-internal
   | local-file-memcard | local_file} {remote_file} | {remote_file}
   {[ local-file-internal | local-file-memcard | local_file]}] {force-
   overwrite}
   ```

   By default, if you transfer a file with a name that already exists on the system, the switch prompts you to overwrite the existing file. For more information, see the `tftp get` command.

## Understanding System Redundancy

With Modular Switches and SummitStack if you install two MSMs/MM or nodes in the chassis, or if you configure two master-capable nodes in a SummitStack, one assumes the role of primary (also called "master") and the other assumes the role of backup.

The primary MSM/MM or node provides all of the switch management functions including bringing up and programming the I/O modules, running the bridging and routing protocols, and configuring the switch. The primary MSM/MM or node also synchronizes the backup MSM/MM or node in case it needs to take over the management functions if the primary MSM/MM or node fails.

For SummitStack, a node can be a redundant primary node if it has been configured to be master-capable.

To configure master capability on one or all nodes in a SummitStack, use one of the following commands:

```
configure stacking [node-address node-address | slot slot-number]
alternate-ip-address [ipaddress netmask | ipNetmask] gateway
configure stacking redundancy [none | minimal | maximal]
```

### Node Election

Node election is based on leader election between the MSMs/MMs installed in the chassis, or master-capable nodes present in a SummitStack.

By default, the MSM/MM installed in slot A or the SummitStack node in slot 1 has primary status. Each node uses health information about itself together with a user configured priority value to compute its node role election priority. Nodes exchange their node role election priorities. During the node election process, the node with the highest node role election priority becomes the master or primary node, and the node with the second highest node role election priority becomes the backup node. All other nodes (if any) remain in STANDBY state.

The primary node runs the switch management functions, and the backup node is fully prepared to become the primary node if the primary fails. In SummitStack, nodes that remain in STANDBY state (called Standby nodes) program their port hardware based on instructions received from the primary. Standby nodes configured to be master-capable elect a new backup node from among themselves after a failover has occurred.

### *Determining the Primary Node*

The following parameters determine the primary node:

**Node state**

The node state must be STANDBY to participate in leader election and be selected as primary. If the node is in the INIT, DOWN, or FAIL states, it cannot participate in leader election. For more information about the node states, see Viewing Node Status on page 58.

**Configuration priority**

This is a user assigned priority. The configured priority is compared only after the node meets the minimum thresholds in each category for it to be healthy. Required processes and devices must not fail.

**Software health**

This represents the percent of processes available.

**Health of secondary hardware components**

This represents the health of the switch components, such as power supplies, fans, and so forth.

**Slot ID**

The MSM/MM slot where the node is installed (MSM-A or MSM-B), or the slot number configured on a stack node.

### *Configuring the Node Priority on a Modular Switch*

If you do not configure any priorities, MSM-A has a higher priority than MSM-B. By default, the priority is 0 and the node priority range is 1–100. The higher the value, the higher the priority.

To configure the priority of an MSM/MM node, use the following command:

```
configure node slot slot_id priority node_pri
```

For the *slot_id* parameter, enter `A` for the MSM/MM installed in slot A or `B` for the MSM/MM installed in slot B.

### *Configuring the Node Priority on a SummitStack*

If you do not configure any priorities, slot 1 has the highest priority, slot 2 the second highest priority, and so forth in order of increasing slot number. You may also use the factory assigned MAC address as the node-address value. By default the priority is "automatic" and the node-pri value is any number between 1 and 100. The higher the value, the higher the priority.

Configure the priority of a node in a SummitStack using the following command:

```
configure stacking {node-address node-address | slot slot-number}
priority [node-pri | automatic]
```

### *Relinquishing Primary Status*

Before relinquishing primary status and initiating failover, review the section Synchronizing Nodes--Modular Switches and SummitStack Only on page 1550 to confirm that your platform and both installed MSMs/MMs or master-capable nodes are running software that supports the `synchronize` command.

You can cause the primary to failover to the backup, thereby relinquishing its primary status.

1. Use the `show switch {detail}` command on the primary or the backup node to confirm that the nodes are synchronized and have identical software and switch configurations before failover.

   A node may not be synchronized because checkpointing did not occur, incompatible software is running on the primary and backup, or the backup is down.

   If the nodes are not synchronized and both nodes are running a version of ExtremeXOS that supports synchronization, proceed to step 2.

   If the nodes are synchronized, proceed to step 3 on page 56.

   The output displays the status of the nodes, with the primary node showing MASTER and the backup node showing BACKUP (InSync).
2. If the nodes are not synchronized because of incompatible software, use the `synchronize` command to ensure that the backup has the same software in flash as the primary.

   The `synchronize` command:

   - Reboots the backup node to prepare it for synchronizing with the primary node.
   - Copies both the primary and secondary software images.
   - Copies both the primary and secondary configurations.
   - Reboots the backup node after replication is complete.

   After you confirm the nodes are synchronized, proceed to step 3.
3. If the nodes are synchronized, use the `run failover {force}` command to initiate failover from the primary node to the backup node.

   The backup node then becomes the primary node and the original primary node reboots.

## Replicating Data Between Nodes

ExtremeXOS replicates configuration and run-time information between the primary node and the backup node so that the system can recover if the primary fails. This method of replicating data is known as checkpointing. Checkpointing is the process of automatically copying the active state from the primary to the backup, which allows for state recovery if the primary fails.

Replicating data consists of the following three steps:

- Configuration synchronization—Relays current and saved configuration information from the primary to the backup.
- Bulk checkpoint—Ensures that each individual application running on the system is synchronized with the backup.
- Dynamic checkpoint—Checkpoints any new state changes from the primary to the backup.

To monitor the checkpointing status, use the following command:

`show checkpoint-data {process}`

Data is not replicated from the primary to the standby nodes.

### *Relaying Configuration Information*

To facilitate a failover from the primary node to the backup node, the primary transfers its active configuration to the backup.

Relaying configuration information is the first level of checkpointing . During the initial switch boot-up, the primary's configuration takes effect. During the initialization of a node, its configuration is read from the local flash. After the primary and backup nodes have been elected, the primary transfers its current active configuration to the backup. After the primary and backup nodes are synchronized, any configuration change you make to the primary is relayed to the backup and incorporated into the backup's configuration copy.

Note

To ensure that all of the configuration commands in the backup's flash are updated, issue the `save` command after you make any changes. On a SummitStack, the save configuration command will normally save the primary node's configuration file to all active nodes in the SummitStack.

If a failover occurs, the backup node continues to use the primary's active configuration. If the backup determines that it does not have the primary's active configuration because a run-time synchronization did not happen, the switch or SummitStack reboots. Because the backup always uses the primary's active configuration, the active configuration remains in effect regardless of the number of failovers.

Note

If you issue the `reboot` command before you save your configuration changes, the switch prompts you to save your changes. To keep your configuration changes, save them before you reboot the switch.

### *Bulk Checkpointing*

Bulk checkpointing causes the primary and backup run-time states to be synchronized. Since ExtremeXOS runs a series of applications, an application starts checkpointing only after all of the applications it depends on have transferred their run-time states to the backup MSM/MM node.

After one application completes bulk checkpointing, the next application proceeds with its bulk checkpointing.

- To monitor the checkpointing status, use the `show checkpoint-data {`*`process`*`}` command.
- To see if bulk checkpointing is complete (that is, to see if the backup node is fully synchronized [in sync] with the primary node), use the `show switch {detail}` command.

If a failover occurs before bulk checkpointing is complete, the switch or SummitStack reboots. However, once bulk checkpointing is complete, failover is possible without a switch or SummitStack reboot.

### *Dynamic Checkpointing*

After an application transfers its saved state to the backup node, dynamic checkpointing requires that any new configuration information or state changes that occur on the primary be immediately relayed to the backup.

This ensures that the backup has the most up-to-date and accurate information.

### *Viewing Checkpoint Statistics*

View and check the status of one or more processes being copied from the primary to the backup node.

Run `show checkpoint-data {`*`process`*`}`.

This command is also helpful in debugging synchronization problems that occur at run time.

The output displays, in percentages, the amount of copying completed by each process and the traffic statistics between the process on both the primary and the backup nodes.

## Viewing Node Status

ExtremeXOS allows you to view node statistical information. Each node in a modular switch, or stackable switch in a SummitStack installed in your system is self-sufficient and runs the ExtremeXOS management applications. By reviewing this output, you can see the general health of the system along with other node parameters.

Run `show node {detail}`.

In a SummitStack, the `show stacking` command shows the node roles of active nodes.

### *Node Status Collected*

The following table provides descriptions of node states.

**Table 13: Node States**

| Node State | Description |
|---|---|
| BACKUP | In the backup state, this node becomes the primary node if the primary fails or enters the DOWN state. The backup node also receives the checkpoint state data from the primary. |
| DOWN | In the down state, the node is not available to participate in leader election. The node enters this state during any user action, other than a failure, that makes the node unavailable for management. Examples of user actions are:<br>• Upgrading the softwareRebooting the system using the `reboot` command.<br>• Initiating an MSM/MM failover using the `run failover` command.<br>• Synchronizing the MSM/MM software and configuration in non-volatile storage using the `synchronize` command. |
| FAIL | In the fail state, the node has failed and needs to be restarted or repaired. The node reaches this state if the system has a hardware or software failure. |
| INIT | In the initial state, the node is being initialized. A node stays in this state when it is coming up and remains in this state until it has been fully initialized. Being fully initialized means that all of the hardware has been initialized correctly and there are no diagnostic faults. |
| MASTER | In the primary (master) state, the node is responsible for all switch management functions. |
| STANDBY | In the standby state, leader election occurs—the primary and backup nodes are elected. The priority of the node is only significant in the standby state. In SummitStack, there can be more than two master-capable nodes. All such nodes that do not get elected either master or backup remain in standby state. |

## Understanding Hitless Failover Support

With Modular Switches and SummitStack the term *hitless failover* has slightly different meanings on a modular chassis and a SummitStack.

On a modular chassis, MSMs/MMs do not directly control customer ports; such ports are directly controlled by separate processors. However, a SummitStack node has customer ports that are under the control of its single central processor. When a modular chassis MSM/MM failover occurs, all of the ports in the chassis are under the control of separate processors which can communicate with the backup MSM/MM, so all ports continue to function. In a SummitStack, failure of the primary node results in all ports that require that node's processor for normal operation going down. The remaining SummitStack nodes' ports continue to function normally. Aside from this difference, hitless failover is the same on modular chassis and SummitStack.

As described in the section Understanding System Redundancy on page 54, if you install two MSMs/MMs (nodes) in a chassis or if you configure at least two master-capable nodes in a SummitStack, one assumes the role of primary and the other assumes the role of backup.

The primary node provides all of the switch management functions including bringing up and programming the I/O modules or other (standby) nodes in the SummitStack, running the bridging and routing protocols, and configuring the switch. The primary node also synchronizes the backup node in case it needs to take over the management functions if the primary node fails.

The configuration is one of the most important pieces of information checkpointed to the backup node. Each component of the system needs to checkpoint whatever runtime data is necessary to allow the backup node to take over as the primary node if a failover occurs, including the protocols and the hardware-dependent layers. For more information about checkpointing data and relaying configuration information, see Replicating Data Between Nodes on page 56.

Not all protocols support hitless failover; see the following table for a detailed list of protocols and their support. Layer 3 forwarding tables are maintained for pre-existing flows, but subsequent behavior depends on the routing protocols used. Static Layer 3 configurations and routes are hitless. You must configure *OSPF (Open Shortest Path First)* graceful restart for OSPF routes to be maintained, and you must configure *BGP (Border Gateway Protocol)* graceful restart for BGP routes to be maintained. For more information about OSPF, see OSPF on page 1341 and for more information about BGP, see BGP on page 1389. For routing protocols that do not support hitless failover, the new primary node removes and re-adds the routes.

### Protocol Support for Hitless Failover

The following table summarizes the protocol support for hitless failover. Unless otherwise noted, the behavior is the same for all modular switches.

If a protocol indicates support for hitless failover, additional information is also available in that particular chapter. For example, for information about network login support of hitless failover, see Network Login on page 756.

**Table 14: Protocol Support for Hitless Failover**

| Protocol | Behavior | Hitless |
|---|---|---|
| Bootstrap Protocol Relay | All bootprelay statistics (including option 82 statistics) are available on the backup node also | Yes |
| *BGP* | If you configure BGP graceful restart, by default the route manager does not delete BGP routes until 120 seconds after failover occurs. There is no traffic interruption. However, after BGP comes up after restart, BGP re-establishes sessions with its neighbors and relearns routes from all of them. This causes an increase in control traffic onto the network.<br>If you do not configure graceful restart, the route manager deletes all BGP routes 1 second after the failover occurs, which results in a traffic interruption in addition to the increased control traffic. | Yes |
| Connectivity Fault Management (IEEE 802.1ag) | An ExtremeXOS process running on the active MSM/MM should continuously send the MEP state changes to the backup. Replicating the protocol packets from an active MSM/MM to a backup may be a huge overhead if CCMs are to be initiated/ received in the CPU and if the CCM interval is in the order of milliseconds.<br>RMEP timeout does not occur on a remote node during the hitless failover.<br>RMEP expiry time on the new master node in case of double failures, when the RMEP expiry timer is already in progress, is as follows:<br>RMEP Expiry Time = elapsed expiry time on the master node + 3.5 * ccmIntervaltime + MSM convergence time. | Yes |
| Dynamic Host Configuration Protocol client | The IP addresses learned on all *DHCP* enabled VLANs are retained on the backup node after failover. | Yes |
| Dynamic Host Configuration Protocol server | A DHCP server continues to maintain the IP addresses assigned to various clients and the lease times even after failover. When a failover happens, all the clients work as earlier. | Yes |

**Table 14: Protocol Support for Hitless Failover (continued)**

| Protocol | Behavior | Hitless |
|---|---|---|
| Ethernet Automatic Protection Switching (EAPS) | The primary node replicates all EAPS BPDUs to the backup, which allows the backup to be aware of the state of the EAPS domain. Since both primary and backup nodes receive EAPS BPDUs, each node maintains equivalent EAPS states.<br>By knowing the state of the EAPS domain, the EAPS process running on the backup node can quickly recover after a primary node failover.<br>Although both primary and backup nodes receive EAPS BPDUs, only the primary transmits EAPS BPDUs to neighboring switches and actively participates in EAPS. | Yes |
| *EDP (Extreme Discovery Protocol)* | EDP does not checkpoint protocol data units (PDUs) or states, so the backup node does not have the neighbor's information. If the backup node becomes the primary node, and starts receiving PDUs, the new primary learns about its neighbors. | No |
| Extreme Loop Recovery Protocol (ELRP) | If you use ELRP as a standalone tool, hitless failover support is not needed since the you initiate the loop detection. If you use ELRP in conjunction with *ESRP (Extreme Standby Router Protocol)*, ELRP does not interfere with the hitless failover support provided by ESRP.<br>Although there is no hitless failover support in ELRP itself, ELRP does not affect the network behavior if a failover occurs. | No |
| Extreme Standby Router Protocol (ESRP) | If failover occurs on the ESRP MASTER switch, it sends a hello packet with the HOLD bit set. On receiving this packet, the ESRP SLAVE switch freezes all further state transitions. The MASTER switch keeps sending hellos with the HOLD bit set on every hello interval. When the MASTER is done with its failover, it sends another hello with the HOLD bit reset. The SLAVE switch resumes normal processing. (If no packet is received with the HOLD bit reset, the SLAVE timeouts after a certain time interval and resumes normal processing.)<br>Failover on the ESRP SLAVE switch is of no importance because it is the SLAVE switch. | Yes |
| Intermediate System-Intermediate System (IS-IS) | If you configure IS-IS graceful restart, there is no traffic interruption. However, after IS-IS comes up after restart, IS-IS re-establishes sessions with its neighbors and relearns Link State Packets (LSPs) from all of the neighbors. This causes an increase in network control traffic.<br>If you do not configure graceful restart, the route manager deletes all IS-IS routes one second after the failover occurs, which results in a traffic interruption and increased control traffic. IS-IS for IPv6 does not support hitless restart . | IS-IS (IPv4) Yes<br>IS-IS (IPv6) No |

**Table 14: Protocol Support for Hitless Failover (continued)**

| Protocol | Behavior | Hitless |
|---|---|---|
| Link Aggregation Control Protocol (LACP) | If the backup node becomes the primary node, there is no traffic disruption. | Yes |
| *LLDP (Link Layer Discovery Protocol)* | LLDP is more of a tool than a protocol, so there is no hitless failover support. LLDP is similar to EDP, but there is also a MIB interface to query the information learned. After a failover, it takes 30 seconds or greater before the MIB database is fully populated again. | No |
| *MSDP (Multicast Source Discovery Protocol)* | If the active MSM/MM fails, the MSDP process loses all state information and the standby MSM/MM becomes active. However, the failover from the active MSM/MM to the standby MSM/MM causes MSDP to lose all state information and dynamic data, so it is not a hitless failover. | No |
| *MLAG (Multi-switch Link Aggregation Group)* | All MLAG user configuration is executed on both master and backup nodes. Both nodes open listening health-check and checkpoint listening sockets on the respective well-known ports. All *FDB (forwarding database)* entries and IPMC group/cache information that were received through ISC checkpointing is synchronized to the backup node.<br>After failover, the TCP session, which is handled by the failed master, tears down and there is a new session with the MLAG peer switch. After the failover, the FDB & McMgr processes trigger bulk checkpointing of all its entries to the MLAG peer upon receiving ISC up notification. | Yes |

**Table 14: Protocol Support for Hitless Failover (continued)**

| Protocol | Behavior | Hitless |
|---|---|---|
| Network Login | 802.1X Authentication—Authenticated clients continue to remain authenticated after failover. However, one second after failover, all authenticated clients are forced to re-authenticate themselves.<br>Information about unauthenticated clients is not checkpointed, so any such clients that were in the process of being authenticated at the instant of failover must go through the authentication process again from the beginning after failover.<br>MAC-Based Authentication—Authenticated clients continue to remain authenticated after failover so the failover is transparent to them. Information about unauthenticated clients is not checkpointed so any such clients that were in the process of being authenticated at the instant of failover must go through the authentication process again from the beginning after failover.<br>In the case of MAC-Based authentication, the authentication process is very short with only a single packet being sent to the switch so it is expected to be transparent to the client stations.<br>Web-Based Authentication—Web-based Netlogin users continue to be authenticated after a failover. | Yes<br>Yes<br>Yes |
| *OSPF* | If you configure OSPF graceful restart, there is no traffic interruption. However, after OSPF comes up after restart, OSPF re-establishes sessions with its neighbors and relearns Link State Advertisements (LSAs) from all of the neighbors. This causes an increase in control traffic onto the network.<br>If you do not configure graceful restart, the route manager deletes all OSPF routes one second after the failover occurs, which results in a traffic interruption in addition to the increased control traffic. | Yes |
| *OSPFv3 (Open Shortest Path First version 3)* | OSPFv3 does not support graceful restart, so the route manager deletes all OSPFv3 routes one second after the failover occurs. This results in a traffic interruption.<br>After OSPFv3 comes up on the new primary node, it relearns the routes from its neighbors. This causes an increase in control traffic onto the network. | No |
| *PoE (Power over Ethernet)* | The PoE configuration is checkpointed to the backup node. This ensures that, if the backup takes over, all ports currently powered stay powered after the failover and the configured power policies are still in place.<br>This behavior is applicable only on the BlackDiamond 8800 series switches and SummitStack. | Yes |

**Table 14: Protocol Support for Hitless Failover (continued)**

| Protocol | Behavior | Hitless |
|---|---|---|
| Protocol Independent Multicast (PIM) | After a failover, all hardware and software caches are cleared and learning from the hardware is restarted. This causes a traffic interruption since it is the same as if the switch rebooted for all Layer 3 multicast traffic. | No |
| Routing Information Protocol (*RIP (Routing Information Protocol)*) | RIP does not support graceful restart, so the route manager deletes all RIP routes one second after the failover occurs. This results in a traffic interruption as well as an increase in control traffic as RIP re-establishes its database. | No |
| *RIPng (Routing Information Protocol Next Generation)* | RIPng does not support graceful restart, so the route manager deletes all RIPng routes one second after the failover occurs. This results in a traffic interruption.<br>After RIPng comes up on the new primary node, it relearns the routes from its neighbors. This causes an increase in control traffic onto the network. | No |
| Simple Network Time Protocol Client | *SNTP* client will keep the backup node updated about the last server from which a valid update was received, the time at which the last update was received, whether the SNTP time is currently good or not and all other statistics. | Yes |
| *STP (Spanning Tree Protocol)* | STP supports hitless failover including catastrophic failure of the primary node without interruption. There should be no discernible network event external to the switch. The protocol runs in lock step on both master and backup nodes and the backup node is a hot spare that can take over at any time with no impact on the network. | Yes |
| Virtual Router Redundancy Protocol (*VRRP (Virtual Router Redundancy Protocol)*) | VRRP supports hitless failover. The primary node replicates VRRP PDUs to the backup, which allows the primary and backup nodes to run VRRP in parallel. Although both nodes receive VRRP PDUs, only the primary transmits VRRP PDUs to neighboring switches and participates in VRRP. | Yes |

## Platform Support for Hitless Failover

The following table lists when each platform and management module began supporting hitless failover for a specific protocol.

Hitless failover requires a switch with two MSMs/MMs installed.

Remember, as described in the following table, not all protocols support hitless failover. If you are running an earlier version of ExtremeXOS than that listed in the ExtremeXOS version column, the switch does not support hitless failover for that protocol.

**Table 15: Platform Support for Hitless Failover**

| | | | |
|---|---|---|---|
| BlackDiamond 8800 series switches | MSM-48c | *BGP* graceful restart<br>EAPS<br>*ESRP*<br>LACP<br>*MLAG*<br>Network login<br>*OSPF* graceful restart<br>*PoE*<br>*STP*<br>*VRRP*<br>IS-IS graceful restart | 12.1<br>12.1<br>12.1<br>12.1<br>12.5<br>12.1<br>12.1<br>12.1<br>12.1<br>12.1<br>12.1 |
| | 8900-MSM128 | BGP graceful restart<br>EAPS<br>ESRP<br>LACP<br>MLAG<br>Network login<br>OSPF graceful restart<br>PoE<br>STP<br>VRRP<br>IS-IS graceful restart | 12.3<br>12.3<br>12.3<br>12.3<br>12.5<br>12.3<br>12.3<br>12.3<br>12.3<br>12.3<br>12.3 |
| BlackDiamond X8 switch | MM | All applicable | 15.1 |
| SummitStack | Any Summit family switch except the Summit X430 series. | BGP graceful restart | 12.0 |
| | (features available depend on license level) | EAPS<br>ESRP<br>LACP<br>MLAG<br>Network login<br>OSPF graceful restart<br>STP<br>VRRP<br>IS-IS graceful restart | 12.0<br>12.0<br>12.0<br>12.5<br>12.0<br>12.0<br>12.0<br>12.0<br>12.1 |

## Hitless Failover Caveats

This section describes the caveats for hitless failover.

Check the latest version of the ExtremeXOS release notes for additional information.

### *Caveat for BlackDiamond 8800 Series Switches Only*

The following summary describes the hitless failover caveat for BlackDiamond 8800 series switches:

I/O modules not yet in the Operational state are turned off and the card state machine is restarted to bring them to the Operational state. This results in a delay in the I/O module becoming Operational.

*Caveats for a SummitStack*

The following describes the hitless failover caveats for a SummitStack:

- All customer ports and the stacking links connected to the failed primary node will go down. In the recommended stack ring configuration, the stack becomes a daisy chain until the failed node restarts or is replaced.
- A brief traffic interruption (less than 50 milliseconds) can occur when the traffic on the ring is rerouted because the active topology becomes a daisy chain.
- Since the SummitStack can contain more than two master-capable nodes, it is possible to immediately elect a new backup node. If a new backup node is elected, when the original primary node restarts, it will become a standby node.
- To simulate the behavior of a chassis, a MAC address of one of the nodes is designated as the seed to form a stack MAC address. When a failover occurs, the SummitStack continues to be identified with this address.
- During an *OSPF* graceful restart, the SummitStack successfully restores the original link state database only if the OSPF network remains stable during the restart period. If the failed primary node provided interfaces to OSPF networks, the link state database restoration is prematurely terminated, and reconvergence occurs in the OSPF network due to the failover. See OSPF on page 1341 for a description of OSPF and the graceful restart function.
- During a *BGP* graceful restart, the SummitStack successfully restores the BGP routing table only if the BGP network remains stable during the restart period. If a receiving speaker detected the need for a routing change due to the failure of links on the failed primary node, it deletes any previous updates it received from the restarting speaker (the SummitStack) before the restart occurred. Consequently, reconvergence occurs in the BGP network due to the failover. See BGP on page 1389 for a description of BGP and its graceful restart function.

# Understanding Power Supply Management

## Using Power Supplies—Modular Switches Only

ExtremeXOS monitors and manages power consumption on the switch by periodically checking the power supply units (PSUs) and testing them for failures.

To determine the health of the PSU, ExtremeXOS checks the voltage, current, and temperature of the PSU.

The power management capability of ExtremeXOS:

- Protects the system from overload conditions.
- Monitors all installed PSUs, even installed PSUs that are disabled.
- Enables and disables PSUs as required.
- Powers up or down I/O and/or Fabric modules based on available power and required power resources.

- Logs power resource changes, including power budget, total available power, redundancy, and so on.
- Detects and isolates faulty PSUs.

The switch includes two power supply controllers that collect data from the installed PSUs and report the results to the MSM/MM modules. When you first power on the switch, the power supply controllers enable a PSU. As part of the power management function, the power controller disables the PSU if an unsafe condition arises. For more information about the power supply controller, refer to the hardware documentation which listed in Related Publications on page 5.

If you have a BlackDiamond 8000 series *PoE* I/O module installed in a BlackDiamond 8800 series switch, there are specific power budget requirements and configurations associated with PoE that are not described in this section. For more detailed information about PoE, see PoE on page 416.

ExtremeXOS includes support for the 600/900 W AC PSU for the BlackDiamond 8806 switch.

You can mix existing 700/1200 W AC PSUs and 600/900 W AC PSUs in the same chassis; however, you must be running ExtremeXOS 11.6 or later to support the 600/900 W AC PSUs. If you install the 600/900 W AC PSU in a chassis other than the BlackDiamond 8806, ExtremeXOS provides enough power to boot-up the chassis, display a warning message in the log, and disable the PSU. If this occurs, you see a message similar to the following:

```
<Warn:HAL.Sys.Warning>MSM-A:Power supply in slot 6 is not supported and is being disabled.
```

When a combination of 700/1200 W AC PSUs and 600/900 W AC PSUs are powered on in the same BlackDiamond 8806 chassis, all 700/1200 W AC PSUs are budgeted “down” to match the lower powered 600/900 W AC output values to avoid PSU shutdown. For more information about the 600/900 W AC PSU, refer to the hardware documentation listed in Related Publications on page 5.

### *Initial System Boot Up*

When ExtremeXOS boots up, it reads and analyzes the installed I/O modules (BlackDiamond 8800 and X8) and Fabric modules (BlackDiamond X8 series only).

ExtremeXOS prioritizes the powering up of modules as follows (see the following figure):

- BlackDiamond X8: Fabric modules are considered first for power up from the lowest numbered slot to the highest numbered slot, based on their power requirements and the available system power. I/O modules are then given priority from lowest numbered slot to highest numbered slot.
- BlackDiamond 8800 series: I/O modules are considered for power up from the lowest numbered slot to the highest numbered slot, based on their power requirements and the available system power.

If the system does not have enough power, some modules are not powered up.

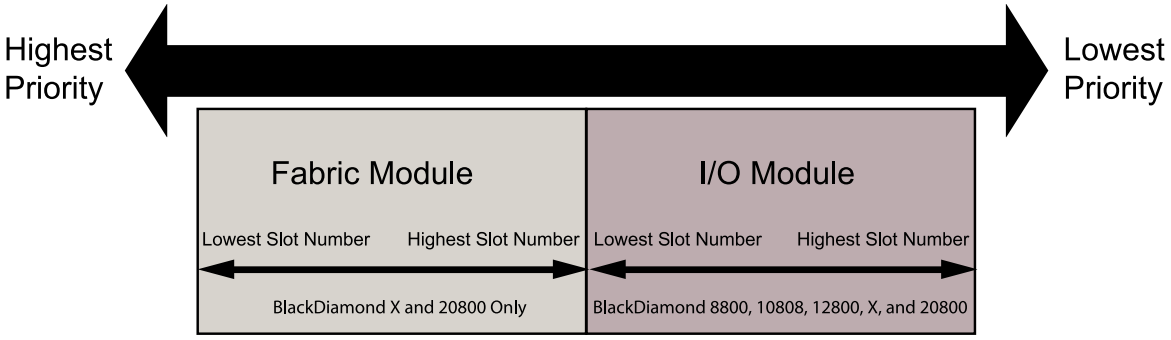



**Figure 2: I/O and Fabric Module Power Priority**

For example, ExtremeXOS:

- Collects information about the PSUs installed to determine how many are running and how much power each can supply.
- Checks for PSU failures.
- Calculates the number of Fabric (BlackDiamond X8 only) and I/O modules to power up based on the available power budget and the power requirements of each I/O module, including *PoE* requirements for the BlackDiamond 8000 series PoE I/O module.
- Reserves the amount of power required to power up a second MSM/MM if only one MSM/MM is installed.
- Reserves the amount of power required to power all fans and chassis components.
- Calculates the current power surplus or shortfall.
- Logs and sends *SNMP* traps for transitions in the overall system power status, including whether the available amount of power is:
  - Redundant or N+1—Power from a single PSU can be lost and no I/O or Fabric (BlackDiamond X8 only) modules are powered down.
  - Sufficient, but not redundant—Power from a single PSU is lost, and one or more I/O modules (and then Fabric modules, for BlackDiamond X8 only) are powered down.
  - Insufficient—One or more modules are not powered up due to a shortfall of available power.

  By reading the PSU information, ExtremeXOS determines the power status and the total amount of power available to the system. The total power available determines which I/O and Fabric (BlackDiamond X8 series only) modules can be powered up.

*Power Redundancy*

In simple terms, power redundancy (N+1) protects the system from shutting down.

With redundancy, if the output of one PSU is lost for any reason, the system remains fully powered. In this scenario, N is the minimum number of power supplies needed to keep the system fully powered and the system has N+1 PSUs powered.

If the system power status is not redundant, the removal of one PSU, the loss of power to one PSU, or a degradation of input voltage results in insufficient power to keep all of the I/O and Fabric (BlackDiamond X8 series only) modules powered up. If there is not enough power, ExtremeXOS powers down the modules as follows:

- BlackDiamond X8: I/O modules from the highest numbered slot to lowest numbered slot are powered down, and then Fabric modules from the highest numbered slot to lowest numbered slot

are powered down until the switch has enough power to continue operation (see Figure 2 on page 68).

- BlackDiamond 8800 series: I/O modules from the highest numbered slot to lowest numbered slot are powered down until the switch has enough power to continue operation (see Figure 2 on page 68).

If you install or provide power to a new PSU, modules powered down due to earlier insufficient power are considered for power up from the lowest slot number to the highest slot number, based on the module's power requirements (see Figure 2 on page 68).

Whenever the system experiences a change in power redundancy, including a change in the total available power, degraded input voltage, or a return to redundant power, the switch sends messages to the syslog.

### *Power Management Guidelines*

The following list describes some key issues to remember when identifying your power needs and installing PSUs:

- If you disable a slot, the module installed in that slot is always powered down regardless of the number of PSUs installed.
- If a switch has PSUs with a mix of both 220V AC and 110V AC inputs, ExtremeXOS maximizes system power by automatically taking one of two possible actions:
  - If all PSUs are enabled, then all PSUs must be budgeted at 110V AC to prevent overload of PSUs with 110V AC inputs.

    OR
  - If the PSUs with 110V AC inputs are disabled, then the PSUs with 220V AC inputs can be budgeted with a higher output per PSU.

ExtremeXOS computes the total available power using both methods and automatically uses the PSU configuration that provides the greatest amount of power to the switch. The following table and the following table list combinations where ExtremeXOS maximizes system power by disabling the PSUs with 110V AC inputs. This can be overridden if desired, as described in Overriding Automatic Power Supply Management on page 70.

**Table 16: BlackDiamond 8800 Series PSU Combinations Where 110V PSUs Are Disabled**

| Number of PSUs with 220V AC Inputs | Number of PSUs with 110V AC Inputs |
|---|---|
| 2 | 1 |
| 3 | 1 |
| 3 | 2 |
| 4 | 1 |

**Table 16: BlackDiamond 8800 Series PSU Combinations Where 110V PSUs Are Disabled (continued)**

| Number of PSUs with 220V AC Inputs | Number of PSUs with 110V AC Inputs |
|---|---|
| 4 | 2 |
| 5 | 1 |

**Table 17: BlackDiamond X8 Series PSU Combinations Where 110V PSUs Are Disabled**

| Number of PSUs with 220V AC Inputs | Number of PSUs with 110V AC Inputs |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 3 | 2 |
| 4 | 1 |
| 4 | 2 |
| 4 | 3 |
| 5 | 1 |
| 5 | 2 |
| 5 | 3 |
| 6 | 1 |
| 6 | 2 |
| 7 | 1 |

For all other combinations of 220V AC and 110V AC PSUs, ExtremeXOS maximizes system power by enabling all PSUs and budgeting each PSU at 110V AC.

BlackDiamond 8806 switch only—When a combination of 700/1200 W AC PSUs and 600/900 W AC PSUs are powered on in the same BlackDiamond 8806 chassis, all 700/1200 W AC PSUs are budgeted "down" to match the lower powered 600/900 W AC output values to avoid PSU shutdown.

### *Overriding Automatic Power Supply Management*

Perform this task if the combination of AC inputs represents one of those listed in the following table. You can override automatic power supply management to enable a PSU with 110V AC inputs that ExtremeXOS disables if the need arises, such as for a planned maintenance of 220V AC circuits.

Note
If you override automatic power supply management, you may reduce the available power and cause one or more I/O modules to power down.

- To turn on a disabled PSU, use the following command:
  `configure power supply` *`ps_num`* `on`
- To resume using automatic power supply management on a PSU, use the following command:
  `configure power supply` *`ps_num`***`auto`**

The setting for each PSU is stored as part of the switch configuration.

- To display power supply status and power budget information, use the following commands:

```
show power
```

```
show power budget
```

*Power Visualization*

Power visualization periodically polls for input power usage. The poll interval is configurable. Whenever the power is increased or decreased by the configured threshold power value, then a specified action is initiated (e.g., a trap, log, or trap-and-log). The configurable parameters are:

- input power usage poll interval (in seconds)
- change action (log, trap, or log-and-trap)
- change threshold (power value in watts)

In the stacking case, the master periodically polls the power usage of all the PSUs in the stack and sends the log or trap or both, depending on the specified change action. Configuration commands are synchronized between Master and backup.

If the change-action is configured as trap or log-and-trap then the power usage trap is sent to the configured *SNMP* servers.

To configure power visualization, use the following command:

```
configure power monitor poll-interval [off | seconds] change-action
[none | [log | log-and-trap | trap] change-threshold watts]
```

Note that the default poll interval is 60 seconds, and the default change action is none (input power usage values are only estimates).

## Using Power Supplies—Summit Family Switches Only

On Summit family switches, ExtremeXOS reports when the PSU has power or has failed.

The Summit family switches support an internal power supply with a range of 90V to 240V AC power as well as an external redundant power supply. The Extreme Networks External Power System (EPS) allows you to add a redundant power supply to the Summit family switches to protect against a power supply failure. The EPS consists of a tray or module that holds the EPS power supplies.

Note

When an EPS-T tray with two EPS-160 PSUs is connected to a Summit family switch, the internal power supply will show as failed.

On non-*PoE* Summit switches, if you experience an internal PSU failure and do not have an external PSU installed, the switch powers down. If you experience a PSU failure and have an external PSU installed, the switch uses the external PSU to maintain power to the switch.

On PoE Summit switches, there are specific power budget requirements and configurations associated with PoE that are not described in this section. The PoE Summit switches respond to internal and external PSU failures based on your PoE configurations. For more information about configuring PoE on the Summit PoE switches, see PoE on page 416.

For more information about Summit family switches and EPS, refer to the hardware documentation listed in Extreme Networks Documentation.

## Using Power Supplies—SummitStack Only

Since the nodes have their own power supplies and since they cannot be shared, management is the same as it is for standalone Summit family switches.

The only difference is that the power management commands have been centralized so that they can be issued from the primary node.

## Displaying Power Supply Information

Display the status of the currently installed power supplies on all switches.

1. Run `show power {ps_num} {detail}`.

   The **detail** option of this command shows power usage parameters on stacking and standalone Summit switches.
2. To view the system power status and the amount of available and required power, use the following command:

   ```
   show power budget
   ```

   On modular switches, these commands provide additional power supply information.
3. To display the status of the currently installed power supply controllers on modular switches, use the following command:

   ```
   show power {ps_num}
   ```

# Using Motion Detectors

On the Summit X670 switch, there is a motion detection system that controls whether the port LEDs are turned on or off. When the motion detector is enabled, the LEDs are turned on only when motion is detected. You can also configure the time in seconds that the LEDs stay on after motion is detected. When the motion detector is disabled, the LED are always turned on.

1. To configure the motion detector, use the following command:

   ```
   configure power led motion-detector [disable | enable {timeout seconds}]
   ```
2. To show the status and timeout setting of the motion detector, use the following command:

   ```
   show power led motion-detector
   ```

# Using the Network Time Protocol

Network Time Protocol (NTP) is used for synchronizing time on devices across a network with variable latency (time delay).

NTP provides a coordinated Universal Time Clock (UTC), the primary time standard by which the world regulates clocks and time. UTC is used by devices that rely on having a highly accurate, universally accepted time, and can synchronize computer clock times to a fraction of a millisecond. In a networked

environment, having a universal time can be crucial. For example, the stock exchange and air traffic control use NTP to ensure accurate, timely data.

NTP uses a hierarchical, semi-layered system of levels of clock sources called a *stratum*. Each stratum is assigned a layer number starting with 0 (zero), with 0 meaning the least amount of delay. The stratum number defines the distance, or number of NTP hops away, from the reference clock. The lower the number, the closer the switch is to the reference clock. The stratum also serves to prevent cyclical dependencies in the hierarchy.

*SNTP*, as the name would suggest, is a simplified version of NTP that uses the same protocol, but without many of the complex synchronization algorithms used by NTP. SNTP is suited for use in smaller, less complex networks. For more information about SNTP see the section, Using the Simple Network Time Protocol on page 96.

## Limitations

The Extreme Networks implementation of NTP includes the following limitations:

- *SNTP* cannot be enabled at the same time NTP is enabled.
- The NTP multicast delivery mechanism is not supported.
- The NTP autokey security mechanism is not supported.
- The broadcast client option cannot be enabled on a per-*VLAN* basis.
- NTP is not supported on the Summit X430.

## Configuring the NTP Server/Client

An NTP server provides clock information to NTP or *SNTP* clients. You can configure an NTP server as an NTP client to receive clock information from more reliable external NTP servers or a local clock. You can also build a hierarchical time distribution topology by using TCP/IP. The switch can work as both an NTP client and server at the same time to build a hierarchical clock distribution tree. This hierarchical structure eliminates the need for a centralized clock server and provides a highly available clock tree with minimal network load and overhead.

- To configure an NTP server, use the following command:

  ```
  configure ntp [server | peer] add [ip_address | host_name] {key keyid}
  {option [burst | initial-burst]}
  configure ntp restrict-list [add | delete] network {mask} [permit |
  deny]
  ```

- To delete an NTP server, use the following command:

  ```
  configure ntp [server | peer] delete [ip_address | host_name]
  ```

- To display NTP server or client information, use the following commands:

  ```
  show ntp
  show ntp association [{ip_address} | {host_name}]
  show ntp restrict-list {user | system }
  show ntp sys-info
  ```

## Managing NTP Peer Support

An NTP peer is a member of a group of NTP servers. Normally, an NTP peer is used to synchronize clock information among a group of servers that serve as mutual backups for each other. Typically, core switches are configured as NTP peers, and an NTP server is configured as a core switch to an NTP client, aggregation switch, or edge switch. An NTP client can choose the most reliable clock from all servers that have a peer relationship with the client.

- To configure an NTP peer, use the following command:

  ```
  configure ntp [server | peer] add [ip_address | host_name] {key keyid}
  {option [burst | initial-burst]}
  ```

- To delete an NTP peer, use the following command:

  ```
  configure ntp [server | peer] delete [ip_address | host_name]
  ```

- To display an NTP peer, use the following command:

  ```
  show ntp
  show ntp association [{ip_address} | {host_name}] statistics
  show ntp sys-info
  ```

## Managing NTP Local Clock Support

A local clock serves as backup to distribute clock information internally when reliable external clock sources are not reachable. Assign a higher stratum value to the local clock to ensure that it is not selected when an external reliable clock source with a lower stratum number exists.

- To configure a local clock, use the following command:

  ```
  configure ntp local-clock stratum stratum_number
  ```

- To delete a local clock, use the following command:

  ```
  configure ntp local-clock none
  ```

- To display local clock information, use the following command:

  ```
  show ntp association [{ip_address} | {host_name}]
  show ntp association [{ip_address} | {host_name}] statistics
  ```

## Managing NTP Broadcast Server Support

An NTP broadcast server sends periodic time updates to a broadcast address in a LAN. When a broadcast client is configured for NTP, that client can receive time information from the broadcasted NTP packets. Using broadcast packets can greatly reduce the NTP traffic on a network, especially in a network with many NTP clients.

To ensure that NTP broadcast clients get clock information from the correct NTP broadcast servers, with minimized risks from malicious NTP broadcast attacks, configure RSA Data Security, Inc. *MD5 (Message-Digest algorithm 5)* Message-Digest Algorithm authentication on both the NTP broadcast server and NTP clients.

- To configure an NTP broadcast server over a *VLAN* where NTP broadcast service is provided, use the following command:

  ```
  enable ntp {vlan} vlan-name broadcast-server {key keyid}
  ```

- To delete an NTP broadcast server over a VLAN where NTP broadcast service is enabled, use the following command:

  ```
  disable ntp {vlan} vlan-name broadcast-server
  ```

- To display an NTP broadcast server, use the following command:

  ```
  show ntp server
  show ntp vlan { vlan-name }
  ```

## Managing NTP Broadcast Client Support

An NTP client listens for NTP packets from an NTP broadcast server. To listen for network broadcast messages, enable an NTP broadcast client. This option is global (it cannot be enabled on a per-*VLAN* basis).

- To configure an NTP broadcast client, use the following command:

  ```
  enable ntp broadcast-client
  ```

- To delete an NTP broadcast client, use the following command:

  ```
  disable ntp broadcast-client
  ```

- To display an NTP broadcast client, use the following command:

  ```
  show ntp sys-info
  ```

## Managing NTP Authentication

To prevent false time information from unauthorized servers, enable NTP authentication to allow an authenticated server and client to exchange time information. The currently supported authentication method is the RSA Data Security, Inc. *MD5* Message-Digest Algorithm. First, enable NTP authentication globally on the switch. Then create an NTP authentication key configured as trusted, to check the encryption key against the key on the receiving device before an NTP packet is sent. After configuration is complete, an NTP server, peer, and broadcast server can use NTP authenticated service.

- To enable or disable NTP authentication globally on the switch, use the following command:

  ```
  enable ntp authentication
  disable ntp authentication
  ```

- To create or delete an RSA Data Security, Inc. MD5 Message-Digest Algorithm key for NTP authentication, use the following command:

  ```
  create ntp key keyid md5 key_string
  delete ntp key [keyid | all]
  ```

- To configure an RSA Data Security, Inc. MD5 Message-Digest Algorithm key as trusted or not trusted, use the following command:

  ```
  configure ntp key keyid [trusted | not-trusted]
  ```

- To display RSA Data Security, Inc. MD5 Message-Digest Algorithm authentication, use the following command:

  ```
  show ntp key
  ```

## VR Configuration Support

NTPD accepts/creates connection for only interfaces from one VR at a time. VR for NTP is configurable. NTP needs to be disabled globally before changing VR for NTP, and should be enabled again to take change into effect.

To configure VR for NTP, use the following command:

```
configure ntp vr vr_name
```

To configure ntp over vlan, use the following command:

```
enable ntp vlan vlan-name [broadcast-server | key keyid]
```

To configure ntp in all *VLAN*s for the configured VR, use the following command:

```
enable ntp all
```

By default NTP uses *VR-Default*. The VR configuration can be seen in `show ntp` command output.

Note
All present NTP VLAN configurations are deleted on changing VR.

## NTP Configuration Example

In the example shown in the following figure, SW#1 synchronizes its clock from the 0-3.us.pool.ntp.org timer server, and provides the synchronized clock information to SW#2 as a unicast message, and to SW#3 as a broadcast message.

SW#2 configures SW#1 as a time server using a normal unicast message. It also has a local clock (127.127.1.1) with a stratum level of 10. SW#3 is configured as broadcast client without specific server information. For security purposes, SW#2 and SW#3 use RSA Data Security, Inc. *MD5* Message-Digest Algorithm authentication with a key index of 100.

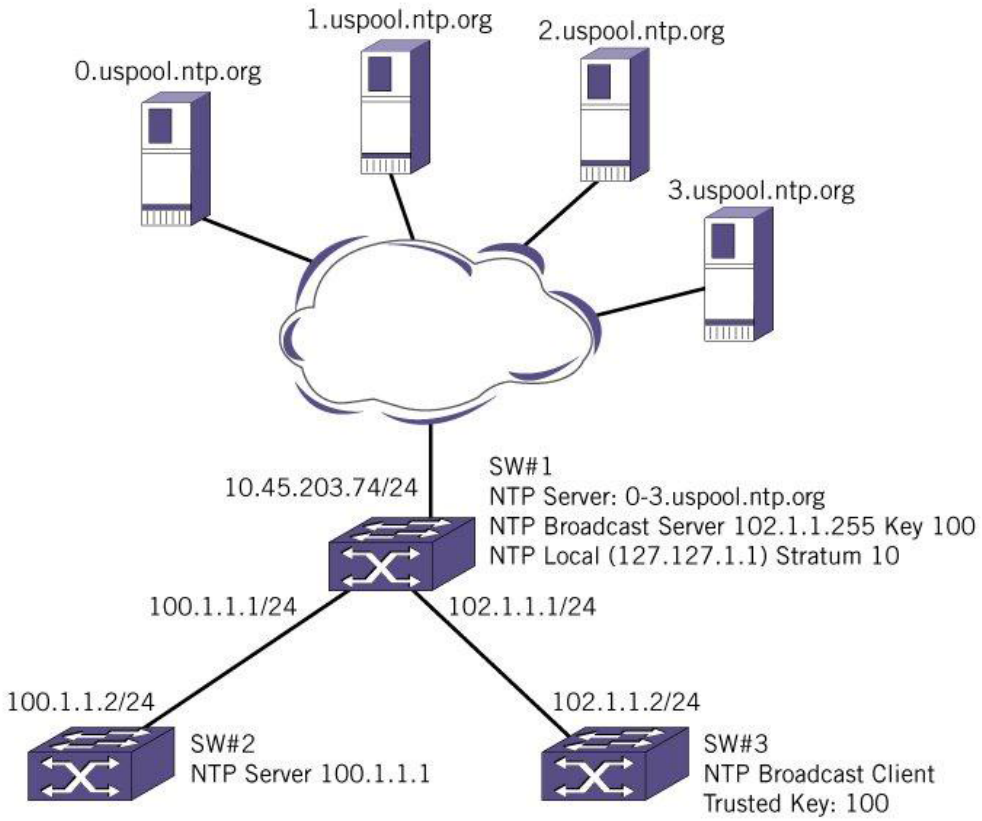



**Figure 3: NTP Configuration Example**

*SW#1 Configuration*

```
create vlan internet
create vlan toSW2
create vlan toSW3
config vlan internet add port 1
config vlan toSW2 add port 2
config vlan toSW3 add port 3
config vlan internet ipaddress 10.45.203.74/24
config vlan toSW2 ipaddress 100.1.1.1/24
config vlan toSW3 ipaddress 102.1.1.1/24
config iproute add default 10.45.203.1 vr vr-default
enable ntp
create ntp key 100  md5 EXTREME
config ntp key 100 trusted
enable ntp vlan internet
enable ntp vlan toSW2
enable ntp vlan toSW3
enable ntp vlan toSW3 broadcast-server key 100
config ntp server add 0.us.pool.ntp.org
config ntp server add 1.us.pool.ntp.org
config ntp server add 2.us.pool.ntp.org
config ntp server add 3.us.pool.ntp.org
config ntp local-clock stratum 10
```

*SW#2 Configuration*

```
create vlan toSW1
config vlan toSW1 add port 1
```

```
config vlan toSW1 ipaddress 100.1.1.2/24
enable ntp
enable ntp vlan toSW1
config ntp server add 100.1.1.1
```

*SW#3 Configuration*

```
create vlan toSW1
config vlan toSW1 add port 1
config vlan toSW1 ipaddress 102.1.1.2/24
enable ntp
enable ntp broadcast-client
create ntp key 100 md5 EXTREME
config ntp key 100 trusted
enable ntp vlan toSW1
```

# Using the Simple Network Management Protocol

Any network manager program running the *SNMP* can manage the switch if the Management Information Base (MIB) is installed correctly on the management station.

Each network manager program provides its own user interface to the management facilities.

Note

When using a network manager program to create a *VLAN*, we do not support the SNMP createAndWait operation. To create a VLAN with SNMP, use the createAndGo operation.

createAndGo is one of six values in the RowStatus column of SMIv2 tables. createAndGo is supplied by a manager wishing to create a new instance of a conceptual row and have its status automatically set to active in order to make it available for use by the managed device

The following sections describe how to get started if you want to use an SNMP manager. It assumes you are already familiar with SNMP management.

Note

Perform a save operation if you make any configurations using SNMP mibs. If you do not save, some of the configurations may not survive when you reboot.

## Enabling and Disabling SNMPv1/v2c and SNMPv3

ExtremeXOS can concurrently support SNMPv1/v2c and SNMPv3. The default is both types of *SNMP* enabled. Network managers can access the device with either SNMPv1/v2c methods or SNMPv3.

- To allow support for all SNMP access, or SNMPv1/v2c access only, or SNMPv3 access only, use the following command:

  `enable snmp access {`**`snmp-v1v2c`** | **`snmpv3`**`}`

- To prevent support for all SNMP access, or SNMPv1/v2c access only, or SNMPv3 access only, use the following command:

  `disable snmp access {`**`snmp-v1v2c`** | **`snmpv3`**`}`

Most of the commands that support SNMPv1/v2c use the keyword **`snmp`**; most of the commands that support SNMPv3 use the keyword **`snmpv3`**.

After a switch reboot, all slots must be in the "Operational" state before SNMP can manage and access the slots. To verify the current state of the slot, use the `show slot` command.

### *Understanding Safe Defaults Mode and SNMP*

The safe defaults mode runs an interactive script that allows you to enable or disable *SNMP*, Telnet, and switch ports.

When you set up your switch for the first time, you must connect to the console port to access the switch. After logging in to the switch, you enter safe defaults mode. Although SNMP, Telnet, and switch ports are enabled by default, the script prompts you to confirm those settings.

If you choose to keep the default setting for SNMP—the default setting is enabled—the switch returns the following interactive script:

```
Since you have chosen less secure management methods, please remember to increase the
security of your network by taking the following actions:
* change your admin password
* change your SNMP public and private strings
* consider using SNMPv3 to secure network management traffic
```

For more detailed information about safe defaults mode, see Using Safe Defaults Mode on page 24.

### *Enabling and Disabling SNMP Access on Virtual Routers*

Beginning with ExtremeXOS 12.4.2 software, you can enable and disable *SNMP* access on any or all VRs. By default, SNMP access is enabled on all VRs.

When SNMP access is disabled on a VR, incoming SNMP requests are dropped and the following message is logged:

```
SNMP is currently disabled on VR <vr_name> Hence dropping the SNMP requests on this VR.
```

SNMP access for a VR has global SNMP status that includes all SNMPv1v2c, SNMPv3 default users and default group status. However, trap receiver configuration and trap enabling/disabling are independent of global SNMP access and are still forwarded on a VR that is disabled for SNMP access.

- Enable SNMP access on a VR:

  `enable snmp access vr [vr_name | all]`
- Disable SNMP access on a VR, use the following command:

  `disable snmp access vr [vr_name | all]`
- Display the SNMP configuration and statistics on a VR:

  `show snmp {vr} vr_name`

## Accessing Switch Agents

To access the *SNMP* agent residing in the switch, at least one *VLAN* must have an assigned IP address. ExtremeXOS supports either IPv4 or IPv6 addresses to manage the switch.

By default, SNMP access and SNMPv1/v2c traps are enabled. SNMP access and SNMP traps can be disabled and enabled independently—you can disable SNMP access but still allow SNMP traps to be sent, or vice versa.

## Return-to-Normal SNMP Notifications

This feature implements two new *SNMP* notifications that indicate that an alert condition has "returned-to-normal". The first notification addresses CPU utilization. Currently, EXOS allows you to monitor the CPU utilization and history for all of the processes running on the switch. When this function is enabled, a CPU threshold value is used to flag a process in the system that has exceeded the threshold. A SNMP notification is generated for processes exceeding that threshold. When a process' cpu utilization falls back below the configured threshold, this feature adds support to generated a new "return-to-normal" notification.

The second notification is a "return-to-normal" message that corresponds to a previously generated overheat notification. The overheat notification indicates that the on board temperature sensor has reported a overheat condition. When the on board temperature sensor reports the clearing of an overheat condition, the new "return-to-normal" notification is generated.

## Supported MIBs

Standard MIBs supported by the switch. In addition to private MIBs, the switch supports the standard MIBs listed in Supported Standards, Protocols, and MIBs on page 1594.

## Configuring SNMPv1/v2c Settings

The following SNMPv1/v2c parameters can be configured on the switch:

- Authorized trap receivers—An authorized trap receiver can be one or more network management stations on your network. The switch sends SNMPv1/v2c traps to all configured trap receivers. You can specify a community string and UDP port individually for each trap receiver. All community strings must also be added to the switch using the `configure snmp add community` command.

  To configure a trap receiver on a switch, use the following command:

  ```
  configure snmp add trapreceiver [ip_address | ipv6_address] community
  [[hex hex_community_name] | community_name] {port port_number} {from
  [src_ip_address | src_ipv6_address]} {vr vr_name} {mode trap_mode}
  ```

  To configure the notification type (trap/inform), use the following command specifying trap as the type:

  ```
  configure snmpv3 add notify [[hex hex_notify_name] | notify_name] tag
  [[hex hex_tag] | tag] {type [trap | inform]}{volatile}
  ```

  To delete a trap receiver on a switch, use the following command:

  ```
  configure snmp delete trapreceiver [[ip_address | ipv6_address]
  {port_number} | all]
  ```

  Entries in the trap receiver list can also be created, modified, and deleted using the RMON2 trapDestTable MIB table, as described in RFC 2021.

- SNMP INFORM—*SNMP* INFORM allows for confirmation of a message delivery. When an SNMP manager receives an INFORM message from an SNMP agent, it sends a confirmation response back to the agent. If the message has not been received and therefore no response is returned, the

INFORM message is resent. You can configure the number of attempts to make and the interval between attempts.

To configure the notification type (trap/inform), use the following command specifying inform as the type:

```
configure snmpv3 add notify [[hex hex_notify_name] | notify_name] tag
[[hex hex_tag] | tag] {type [trap | inform]}{volatile}
```

To configure the number of SNMP INFORM notification retries, use the following command:

```
configure snmpv3 target-addr [[hex hex_addr_name] | addr_name] retry
retry_count
```

To configure the SNMP INFORM timeout interval, use the following command:

```
configure snmpv3 target-addr [[hex hex_addr_name] | addr_name] timeout
timeout_val
```

- Community strings—The community strings allow a simple method of authentication between the switch and the remote network manager. There are two types of community strings on the switch:
  - Read community strings provide read-only access to the switch. The default read-only community string is public.
  - Read-write community strings provide read- and-write access to the switch. The default read-write community string is private.

  To store and display the SNMP community string in encrypted format, use the following command:

  ```
  configure snmpv3 add community [[hex hex_community_index] |
  community_index] [encrypted name community_name | name [[hex
  hex_community_name] | community_name] {store-encrypted} ] user [[hex
  hex_user_name] | user_name] {tag [[hex transport_tag] |
  transport_tag]} {volatile}
  ```

  Note
  SNMP community string name can contain special characters.

- System contact (optional)—The system contact is a text field that enables you to enter the name of the person(s) responsible for managing the switch.
- System name (optional)—The system name enables you to enter a name that you have assigned to this switch. The default name is the model name of the switch (for example, BD-1.2).
- System location (optional)—Using the system location field, you can enter the location of the switch.

## Displaying SNMP Settings

To view *SNMP* settings configured on the switch, use the following command:

```
show management
```

This command displays the following information:

- Enable/disable state for Telnet and SNMP access
- Login statistics

- Enable/disable state for idle timeouts
- Maximum number of CLI sessions
- SNMP community strings
- SNMP notification type (trap or INFORM)
- SNMP trap receiver list
- SNMP trap receiver source IP address
- SNMP statistics counter
- SSH access states of enabled, disabled, and module not loaded
- CLI configuration logging
- SNMP access states of v1, v2c disabled and v3 enabled
- Enable/disable state for Remote Monitoring (RMON)
- Access-profile usage configured via ACLs for additional Telnet, SSH2 security, SNMP, and HTTP(s)
- CLI scripting settings
- Enable/disable state
- Error message setting
- Persistence mode
- Dropped SNMP packet counter

## ExtremeXOS SNMP Notification Log

*SNMP* traps and informs are two methods that an Network Element (NE) uses to notify an NMS about an autonomous events that occur on the NE. SNMP traps are unacknowledged notifications, while SNMP informs are acknowledged by the recipient (typically a management station). Sometimes SNMP notifications sent by an NE fail to be received by the management station. Typically, when a failure clears, the management station will re-sync its view of the NE state with the actual NE state.

Re-syncs are costly because the management station has to read the entire state of the NE, even if the changes in the state of the NE during the downtime are minimal. To reduce the need a full re-sync, EXOS adds a notification log to the NE. The log is populated with notifications sent by the NE to management stations. After a network or management station failure, the management station can read the log to see what events occurred during the downtime, thus eliminating the need for a full re-sync.

### *SNMP Notification Logs Overview*

This feature offers users, or a management station, the ability to define multiple *SNMP* notification logs that keep track of the SNMP notifications (either an SNMP trap, or an SNMP inform) sent by the NE to management stations. A notification log has a name and a notification filter profile associated with it. The name is used to uniquely identify the log, and the filter profile defines which notifications generated by the NE are added to the log, and which notifications are not.

A log is also associated with the security credentials (SNMP user name, SNMP security model, and SNMP security level) that are used to create the log. Notifications that are added to a log are restricted to the notifications that can be accessed using these security credentials. You can also create a default log (a null-named log). The default log does not have security credentials associated with it, so it does not implement any access checks.

A notification log is limited in the number of notifications that it can store by a global entry limit, and a log entry limit, both of which can be changed. The global entry limit specifies the number of notifications that are present in all logs combined, while the log entry limit specifies the number of entries that are present for a specific log. You can also let the system manage the log entry limit, in which case the log can use all available free space within the limit specified by the global entry limit.

You can also enable aging of log entries by configuring an age-out period for them. When enabled, log entries that are older than the specified period are removed from the log.

The information stored in a log for each notification entry includes the following:

- The value of system up time at which the notification was generated.
- The date and time at which the notification was generated.
- The context for the notification.
- The object ID of the notification.
- The list of var-binds that is present in the notification.

After the SNMP agent is restarted, the value of system up time when the notification was generated is reset to 0 for all entries that are present in the log. This serves as an indication to log viewers that the SNMP agent restarted.

#### Enabling and Disabling SNMP Notification Logs

To enable SNMP notification logging, create an entry in nlmConfigLogTable. After you create an entry, you can control the administrative status of the entry through both nlmConfigLogAdminStatus and nlmConfigLogEntryStatus MIB objects. You can view the operational status of the log using the nlmConfigLogOperStatus MIB object. You must associate an existing filter profile with the log for it to become operational.

#### Log Size Limits

You can set the maximum number of notification that can be logged at both the system level, and the individual log level. These limits are controlled through the nlmConfigGlobalEntryLimit and nlmConfigLogEntryLimit MIB objects, respectively. The sum of the values of nlmConfigLogEntryLimit for all entries cannot exceed nlmConfigGlobalEntryLimit.

If you try to set the value of nlmConfigLogEntryLimit so the sum of the values of nlmConfigLogEntryLimit for all entries exceeds the nlmConfigGlobalEntryLimit, the set operation is rejected. Similarly if you try to reduce the value of nlmConfigGlobalEntryLimit so that sum of the values of the nlmConfigLogEntryLimit for all entries exceeds the new value for nlmConfigGlobalEntryLimit, the operation is rejected. You can also set the nlmConfigLogEntryLimit to 0 (system-managed). If the entry limit for a log is set to 0, the log can use all available free space within the limit specified by nlmConfigGlobalEntryLimit.

#### Aging

You can specify an age limit in minutes for notifications in the log through the nlmConfigGlobalAgeOut MIB object. When a notification entry grows older than the specified age limit, the notification entry is deleted. You can disable aging of log entries by setting the value of this object to 0.

#### Access Control

When a named log is enabled moving the nlmConfigLogEntryStatus object of the log to the active state, the NE associates the security credentials used to perform that operation with the log. A notification may be added to the log only if the notification and the var-binds in the notification can be accessed using these security credentials. Access control does not apply to the default log (null-named log). The default log is not associated with any security credentials, so notifications are added to the default log without any access control restrictions.

### *Benefits and Limitations*

#### Benefits

The ExtremeXOS *SNMP* Notification Log feature has the following benefits:

- Ability to create multiple SNMP notification logs.
- Ability to restrict SNMP notifications that are added to a log.
- Ability to age log entries.
- Ability to limit the maximum number of entries in a log.
- Ability to control the feature through both CLI and SNMP.

#### Limitations

- No capability to query log entries by time duration (for example, list log entries from the last hour).
- The notification log name “default” is reserved to represent the default log in CLI. You cannot create a notification log with the name "default".
- Aging out of entries may occur sooner, or later, than the global age out period that you specify if the current time of the NE is changed.
- Notification log statistics (but not entries) are lost on a restart of the SNMP Master process.
- Notification log statistics (but not entries) are lost on failover.
- Notification log entries and statistics are lost if the NE is rebooted.

### *Logging Operation*

This section discusses operation of the notification-log feature when a notification is generated by the NE.

#### Logging

When a notification is generated by a NE, it is added to each log that exists in nlmConfigLogTable and satisfies the following conditions:

- The notification log is enabled and active.
- The security credentials associated with the log permit acces to the notification and all the var-binds contained in the notification. This condition does not apply to the default log as it is not associated with any security credentials.
- The filter associated with the log exists and is active and does not filter out the notification.

Before adding a notification to a log, the NE makes sure that the log size limits are not exceeded by this addition in the following manner:

- For system managed logs (i.e. nlmConfigLogEntryLimit is set to 0):
  - If the total number of entries in all logs combined is equal to the global entry limit (nlmConfigGlobalEntryLimit), then the oldest entry from the system managed log with the largest number of entries is removed before adding the new notification to the log.
- For logs with user defined size limits (i.e. nlmConfigLogEntryLimit is set to a value greater than 0):
  - If the number of entries in the log is equal to the entry limit of the log (nlmConfigLogEntryLimit), the oldest notification is removed from the log before adding the new notification to the log.
  - If the number of entries in the log is less than the entry limit of the log (nlmConfigLogEntryLimit), but the total number of entries in all logs combined is equal to the global entry limit (nlmConfigGlobalEntryLimit), then the oldest entry from the system managed log with the largest number of entries is removed before adding the new notification to the log.

#### Aging

Periodically (every minute), the notification log module calculates the difference between the current time and the time the notification entry is added to a log for each notification entry in each log. If the time difference is greater than the global age out period, the entry is removed from the log. Aging in this manner imposes a limitation that entries may be aged out sooner or later than the actual global age out period if the current time of the NE is changed (for example, to DST changes). Implementing age out accurately consumes 4 additional bytes of memory per notification entry.

#### Statistics

In ExtremeXOS Release 15.5, the following *SNMP* Notification Logs statistics are available:

- Total number of notifications that have been logged since the NE last restarted.
- Total number of notifications that have been removed due to size constraints since the NE last restarted.
- Per log number of notifications logged since the NE last restarted.
- Per log number of notifications removed due to size constraints since the NE last restarted.

### *Configuration Examples*

The following sections provide various examples of the *SNMP* Notification Log feature.

#### Log all notifications

The following example illustrates how to log all notifications sent by a switch, and retain them for as long as possible. However, to reduce memory usage, you might want to limit the number of notifications in all logs to 5000 entries:

```
configure snmp notification-log global-entry-limit 5000
```

Disable aging of notification entries.

```
configure snmp notification-log global-age-out none
```

Create the default log. Because you want to log all notifications, the default log can be used instead of a named log, because it does not impose any security checks.

```
configure snmp add notification-log default
```

Create a filter that accepts all notifications.

```
configure snmpv3 add filter "all" subtree 1 type included
```

Attach the filter to the log.

```
configure snmp notification-log "default" filter-profile-name "all"
```

View the configuration, status and entries of the default log.

```
show snmp notification-log "default"
```

View entry number 1 of the default log in detail.

```
show snmp notification-log "default" entry 1
```

#### Log all notifications using security

The following example illustrates how to log all notifications that are visible to the SNMP user “monitor” when using the security mode ‘USM’, and the security level ‘privacy’.

Create the log and associate it with the security credentials of the user “monitor”.

```
configure snmp add notification-log "monitor-log" user "monitor" sec-model usm sec-level
priv
```

Create a filter including only all traps.

```
configure snmpv3 add filter "all" subtree 1 type included
```

Attach the filter to the log.

```
configure snmp notification-log "monitor-log" filter-profile-name "all"
```

View the configuration, status and entries of “monitor-log”.

```
show snmp notification-log "monitor-log"
```

View entry number 1 of “monitor-log” log in detail.

```
show snmp notification-log "monitor-log" entry 1
```

#### NMS logs all link status change notifications

The following example illustrates the configuration for when an NMS wants to log all link status change notifications. The NMS queries the log every hour, and wants to age out the log entries every two hours. Additionally, to ensure that link status events are not replaced by other events, the NMS wants to reserve 1000 entries for this log.

- Create a notification filter profile including both linkUp and linkDown OIDs.

  snmpNotifyFilterMask.11."link-status".1.3.6.1.6.3.1.1.5.3 = "H

  snmpNotifyFilterType.11."link-status".1.3.6.1.6.3.1.1.5.3 = include

  snmpNotifyFilterStorageType.11."link-status".1.3.6.1.6.3.1.1.5.3 = nonVolatile

  snmpNotifyFilterRowStatus.11."link-status".1.3.6.1.6.3.1.1.5.3 = createAndGo

  snmpNotifyFilterMask.11."link-status".1.3.6.1.6.3.1.1.5.4 = "H

  snmpNotifyFilterType.11."link-status".1.3.6.1.6.3.1.1.5.4 = include

  snmpNotifyFilterStorageType.11."link-status".1.3.6.1.6.3.1.1.5.4 = nonVolatile

  snmpNotifyFilterRowStatus.11."link-status".1.3.6.1.6.3.1.1.5.4 = createAndGo

- Create a named log for link status notifications, attach the profile created above, and set its entry limit to 1000. The SNMP operation of creating this entry must be performed using security credentials that have access to the linkUp and linkDown notifications.

  nlmConfigLogFilterName.5."links" = "link-status"

nlmConfigLogEntryLimit.5."links" = 1000

nlmConfigLogAdminStatus.5."links" = enabled

nlmConfigLogStorageType.5."links" = nonVolatile

nlmConfigLogEntryStatus.5."links" = createAndGo

- Set the global age-out to 120 minutes.

  nlmConfigGlobalAgeOut.0 = 120

- To view the log contents, the NMS must query nlmLogTable and nlmLogVariableTable.

## SNMPv3

SNMPv3 is an enhanced standard for SNMP that improves the security and privacy of SNMP access to managed devices and provides sophisticated control of access to the device MIB. The prior standard versions of SNMP, SNMPv1, and SNMPv2c, provided no privacy and little security.

**Note**
If you downgrade from ExtremeXOS 15.6 to an earlier version, the SNMPv3 users do not work if the configuration was saved in 15.6. The SNMPv3 users must be manually created again.

The following RFCs provide the foundation for the Extreme Networks implementation of SNMPv3:

- RFC 3410, Introduction to version 3 of the Internet-standard Network Management Framework, provides an overview of SNMPv3.
- RFC 3411, An Architecture for Describing SNMP Management Frameworks, talks about SNMP architecture, especially the architecture for security and administration.
- RFC 3412, Message Processing and Dispatching for the Simple Network Management Protocol (SNMP), talks about the message processing models and dispatching that can be a part of an SNMP engine.
- RFC 3413, SNMPv3 Applications, talks about the different types of applications that can be associated with an SNMPv3 engine.
- RFC 3414, The User-Based Security Model for Version 3 of the Simple Network Management Protocol (SNMPv3), describes the User-Based Security Model (USM).
- RFC 3415, View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP), talks about VACM as a way to access the MIB.
- RFC 3826, The Advanced Encryption Standard (AES) Cipher Algorithm in the SNMP User-based Security Model.

**Note**
3DES, AES 192 and AES 256 bit encryption are proprietary implementations and may not work with some SNMP Managers.

The SNMPv3 standards for network management were driven primarily by the need for greater security and access control. The new standards use a modular design and model management information by cleanly defining a message processing (MP) subsystem, a security subsystem, and an access control subsystem.

The MP subsystem helps identify the MP model to be used when processing a received Protocol Data Unit (PDU), which are the packets used by SNMP for communication.

The MP layer helps in implementing a multilingual agent, so that various versions of SNMP can coexist simultaneously in the same network.

The security subsystem features the use of various authentication and privacy protocols with various timeliness checking and engine clock synchronization schemes.

SNMPv3 is designed to be secure against:

- Modification of information, where an in-transit message is altered.
- Masquerades, where an unauthorized entity assumes the identity of an authorized entity.
- Message stream modification, where packets are delayed and/or replayed.
- Disclosure, where packet exchanges are sniffed (examined) and information is learned about the contents.

The access control subsystem provides the ability to configure whether access to a managed object in a local MIB is allowed for a remote principal. The access control scheme allows you to define access policies based on MIB views, groups, and multiple security levels.

In addition, the SNMPv3 target and notification MIBs provide a more procedural approach for generating and filtering of notifications.

SNMPv3 objects are stored in non-volatile memory unless specifically assigned to volatile storage. Objects defined as permanent cannot be deleted.

Note

In SNMPv3, many objects can be identified by a human-readable string or by a string of hexadecimal octets. In many commands, you can use either a character string, or a colon-separated string of hexadecimal octets to specify objects. To indicate hexadecimal octets, use the keyword hex in the command.

## Message Processing

A particular network manager may require messages that conform to a particular version of *SNMP*. The choice of the SNMPv1, SNMPv2c, or SNMPv3 MP model can be configured for each network manager as its target address is configured. The selection of the MP model is configured with the **mp-model** keyword in the following command:

```
configure snmpv3 add target-params [[hex hex_param_name] | param_name ]
user [[hex hex_user_name] | user_name ] mp-model [snmpv1 | snmpv2c |
snmpv3] sec-model [snmpv1 | snmpv2c | usm] {sec-level [noauth |
authnopriv | priv]} {volatile}
```

## SNMPv3 Security

In SNMPv3 the User-Based Security Model (USM) for *SNMP* was introduced. USM deals with security related aspects like authentication, encryption of SNMP messages, and defining users and their various access security levels. This standard also encompasses protection against message delay and message replay.

### *USM Timeliness Mechanisms*

An Extreme Networks switch has one SNMPv3 engine, identified by its snmpEngineID. The first four octets are fixed to 80:00:07:7C, which represents the Extreme Networks vendor ID. By default, the additional octets for the snmpEngineID are generated from the device MAC address.

Every SNMPv3 engine necessarily maintains two objects: SNMPEngineBoots, which is the number of reboots the agent has experienced and SNMPEngineTime, which is the local time since the engine reboot. The engine has a local copy of these objects and the latestReceivedEngineTime for every authoritative engine it wants to communicate with. Comparing these objects with the values received in messages and then applying certain rules to decide upon the message validity accomplish protection against message delay or message replay.

In a chassis, the snmpEngineID is generated using the MAC address of the MSM/MM with which the switch boots first. In a SummitStack, the MAC address chosen for the snmpEngineID is the configured stack MAC address.

### *Configuring USM Timeliness Mechanism*

Configure the snmpEngineID and SNMPEngineBoots from the command line. The snmpEngineID can be configured from the command line, but when the snmpEngineID is changed, default users revert back to their original passwords/keys, and non-default users are removed from the device.

SNMPEngineBoots can be set to any desired value but will latch on its maximum, 2147483647.

1. To set the snmpEngineID, use the following command:

   `configure snmpv3 engine-id` *`hex_engine_id`*

2. To set the SNMPEngineBoots, use the following command:

   `configure snmpv3 engine-boots` *`(1-2147483647)`*

### *Users, Groups, and Security*

SNMPv3 controls access and security using the concepts of users, groups, security models, and security levels.

Users are created by specifying a user name. Depending on whether the user will be using authentication and/or privacy, you would also specify an authentication protocol ( RSA Data Security, Inc. *MD5* Message-Digest Algorithm or SHA) with password or key, and/or privacy (DES, 3DES, AES) password or key.

Before using the AES, 3DES users, you must install the SSH module and restart the snmpMaster process. Refer to Installing a Modular Software Package on page 1535 for information on installing the SSH module.

### *Managing Users*

Users are created by specifying a user name. Enabling the SNMPv3 default-user access allows an end user to access the MIBs using SNMPv3 default-user.By disabling default-users access, the end-user is not able to access the switch/MIBs using SNMPv3 default-user.

By disabling default-users access, the end-user is not able to access the switch/MIBs using SNMPv3 default-user.

- To create a user, use the following command:

  `configure snmpv3 add user [[hex hex_user_name] | user_name] {authentication [md5 | sha] [hex hex_auth_password | auth_password]} {privacy {des | 3des | aes {128 | 192 | 256}} [[hex hex_priv_password] | priv_password]} }{volatile}`

  A number of default users are initially available. These user names are: admin, initial, initialmd5, initialsha, initialmd5Priv, initialshaPriv. The default password for admin is *password*. For the other default users, the default password is the user name.
- To display information about a user, or all users, use the following command:

  `show snmpv3 user {[[hex hex_user_name] | user_name]}`
- To enable default-user, use the following command:

  `enable snmpv3 [default-group | default-user]`
- To disable default-user, use the following command:

  `disable snmpv3 [default-group | default-user]`
- To delete a user, use the following command:

  `configure snmpv3 delete user [all-non-defaults | [[hex hex_user_name] | user_name]]`

Note

The SNMPv3 specifications describe the concept of a security name. In the ExtremeXOS implementation, the user name and security name are identical. In this manual, both terms are used to refer to the same thing.

### *Managing Groups*

Groups are used to manage access for the MIB. You use groups to define the security model, the security level, and the portion of the MIB that members of the group can read or write.

The security model and security level are discussed in Security Models and Levels on page 91. The view names associated with a group define a subset of the MIB (subtree) that can be accessed by members of the group. The read view defines the subtree that can be read, write view defines the subtree that can be written to, and notify view defines the subtree that notifications can originate from. MIB views are discussed in Setting SNMPv3 MIB Access Control on page 92.

A number of default groups are already defined. These groups are: admin, initial, v1v2c_ro, v1v2c_rw.

Enabling SNMPv3 default-group access activates the access to an SNMPv3 default group and the user-created SNMPv3-user part of default group.

Disabling SNMPv3 default-group access removes access to default-users and user-created users who are part of the default-group.

The user-created authenticated SNMPv3 users (who are part of a user-created group) are able to access the switch.

- To underscore the access function of groups, groups are defined using the following command:

  `configure snmpv3 add access [[hex hex_group_name] | group_name] {sec-model [snmpv1 | snmpv2c | usm]} {sec-level [noauth | authnopriv | priv]} {read-view [[hex hex_read_view_name] | read_view_name]} {write-view [[hex hex_write_view_name]] | write_view_name]} {notify-view [[hex hex_notify_view_nam]] | notify_view_name]} {volatile}`

- To display information about the access configuration of a group or all groups, use the following command:

  `show snmpv3 access {[[hex hex_group_name] | group_name]}`

- To enable default-group, use the following command:

  `enable snmpv3 default-group`

- To disable a default-group, use the following command:

  `disable snmpv3 default-group`

- To associate users with groups, use the following command:

  `configure snmpv3 add group [[hex hex_group_name] | group_name] user [[hex hex_user_name] | user_name] {sec-model [snmpv1| snmpv2c | usm]} {volatile}`

- To show which users are associated with a group, use the following command:

  `show snmpv3 group {[[hex hex_group_name] | group_name] {user [[hex hex_user_name] | user_name]}}`

- To delete a group, use the following command:

  `configure snmpv3 delete access [all-non-defaults | {[[hex hex_group_name] | group_name] {sec-model [snmpv1 | snmpv2c | usm] sec-level [noauth | authnopriv | priv]}}]`

  When you delete a group, you do not remove the association between the group and users of the group.

- To delete the association between a user and a group, use the following command:

  `configure snmpv3 delete group {[[hex hex_group_name] | group_name]} user [all-non-defaults | {[[hex hex_user_name] | user_name] {sec-model [snmpv1|snmpv2c|usm]}}]`

*Security Models and Levels*

For compatibility, SNMPv3 supports three security models:

- SNMPv1—no security
- SNMPv2c—community strings-based security
- SNMPv3—USM security

The default is USM. You can select the security model based on your network manager.

The three security levels supported by USM are:

- noAuthnoPriv—No authentication, no privacy. This is the case with existing SNMPv1/v2c agents.

- AuthnoPriv—Authentication, no privacy. Messages are tested only for authentication.
- AuthPriv—Authentication, privacy. This represents the highest level of security and requires every message exchange to pass the authentication and encryption tests.

When a user is created, an authentication method is selected, and the authentication and privacy passwords or keys are entered.

When RSA Data Security, Inc. *MD5* Message-Digest Algorithm authentication is specified, HMAC-MD5-96 is used to achieve authentication with a 16-octet key, which generates a 128-bit authorization code. This authorization code is inserted in the msgAuthenticationParameters field of SNMPv3 PDUs when the security level is specified as either AuthnoPriv or AuthPriv. Specifying SHA authentication uses the HMAC-SHA protocol with a 20-octet key for authentication.

For privacy, the user can select any one of the following supported privacy protocols: DES, 3DES, AES 128/192/256. In the case of DES, a 16-octet key is provided as input to DES-CBS encryption protocol which generates an encrypted PDU to be transmitted. DES uses bytes 1-7 to make a 56 bit key. This key (encrypted itself) is placed in msgPrivacyParameters of SNMPv3 PDUs when the security level is specified as AuthPriv.

The *SNMP* Context Name should be set to the Virtual Router name for which the information is requested. If the Context Name is not set the switch will retrieve the information for "*VR-Default*.". If the SNMP request is targeted for the protocols running per *VR* (see Adding and Deleting Routing Protocols on page 630), then the contextName should be set to the exact virtual-Router for which the information is requested. List of protocols running per Virtual Router:

- *BGP*
- *OSPF*
- PIM
- *RIP*
- *OSPFv3*
- RIPNG
- *MPLS (Multiprotocol Label Switching)*
- ISIS

## Setting SNMPv3 MIB Access Control

SNMPv3 provides a fine-grained mechanism for defining which parts of the MIB can be accessed. This is referred to as the View-Based Access Control Model (VACM).

MIB views represent the basic building blocks of VACM. They are used to define a subset of the information in the MIB. Access to read, to write, and to generate notifications is based on the relationship between a MIB view and an access group. The users of the access group can then read, write, or receive notifications from the part of the MIB defined in the MIB view as configured in the access group.

A view name, a MIB subtree/mask, and an inclusion or exclusion define every MIB view. For example, there is a System group defined under the MIB-2 tree. The Object Identifier (OID) for MIB-2 is 1.3.6.1.2, and the System group is defined as MIB-2.1.1, or directly as 1.3.6.1.2.1.1.

When you create the MIB view, you can choose to include the MIB subtree/mask or to exclude the MIB subtree/mask.

In addition to the user-created MIB views, there are three default views: defaultUserView, defaultAdminView, and defaultNotifyView.

MIB views that are used by security groups cannot be deleted.

- To define a MIB view which includes only the System group, use the following subtree/mask combination:

  1.3.6.1.2.1.1/1.1.1.1.1.1.1.0

  The mask can also be expressed in hex notation (used in the ExtremeXOS CLI):

  1.3.6.1.2.1.1/fe
- To define a view that includes the entire MIB-2, use the following subtree/mask:

  1.3.6.1.2.1.1/1.1.1.1.1.0.0.0

  which, in the CLI, is:

  1.3.6.1.2.1.1/f8
- To create a MIB view, use the following command:

  `configure snmpv3 add mib-view [[hex hex_view_name] | view_name] subtree object_identifier {subtree_mask} {type [included | excluded]} {volatile}`

  After the view has been created, you can repeatedly use the `configure snmpv3 add mib-view` command to include and/or exclude MIB subtree/mask combinations to precisely define the items you want to control access to.
- To show MIB views, use the following command:

  `show snmpv3 mib-view {[[hex hex_view_name] | view_name] {subtree object_identifier}}`
- To delete a MIB view, use the following command:

  `configure snmpv3 delete mib-view [all-non-defaults | {[[hex hex_view_name] | view_name] {subtree object_identifier}}]`

## SNMPv3 Notification

SNMPv3 can use either SNMPv1 traps or SNMPv2c notifications to send information from an agent to the network manager.

The terms *trap* and *notification* are used interchangeably in this context. Notifications are messages sent from an agent to the network manager, typically in response to some state change on the agent system. With SNMPv3, you can define precisely which traps you want sent, to which receiver by defining filter profiles to use for the notification receivers.

To configure notifications, you configure a target address for the target that receives the notification, a target parameters name, and a list of notification tags. The target parameters specify the security and MP models to use for the notifications to the target. The target parameters name also points to the filter

profile used to filter the notifications. Finally, the notification tags are added to a notification table so that any target addresses using that tag will receive notifications.

### *Configuring Target Addresses*

A target address is similar to the earlier concept of a trap receiver.

- To configure a target address, use the following command:

  `configure snmpv3 add target-addr [[hex hex_addr_name] | addr_name] param [[hex hex_param_name] |param_name ] ipaddress [ ip_address | ipv4-with-mask ip_and_tmask ] | [ ipv6_address | ipv6-with-mask ipv6_and_tmask ]] {transport-port port_number} {from [src_ip_address | src_ipv6_address]} {vr vr_name} {tag-list [tag_list | hex hex_tag_list]} {volatile}`

  In configuring the target address you supply an address name that identifies the target address, a parameters name that indicates the MP model and security for the messages sent to that target address, and the IP address and port for the receiver. The parameters name also is used to indicate the filter profile used for notifications.

  The **from** option sets the source IP address in the notification packets.

  The **tag-list** option allows you to associate a list of tags with the target address. The tag defaultNotify is set by default. Tags are discussed in Managing Notification Tags on page 94.
- To display target addresses, use the following command:

  `show snmpv3 target-addr {[[hex hex_addr_name] | addr_name]}`
- To delete a single target address or all target addresses, use the following command:

  `configure snmpv3 delete target-addr [{[[hex hex_addr_name] | addr_name]} | all]`

### *Managing Notification Tags*

When you create a target address, either you associate a list of notification tags with the target or by default, the defaultNotify tag is associated with the target. When the system generates notifications, only those targets associated with tags currently in the standard MIB table, called snmpNotifyTable, are notified.

**Note**

This notification entry can be deleted via CLI and also via MIB. If this is deleted, then this can result in the traps not being sent for trap receivers which do not have the tag-list value mentioned explicitly.

- To add an entry to the table, use the following command:

  `configure snmpv3 add notify [[hex hex_notify_name] | notify_name] tag [[hex hex_tag] | tag] {type [trap | inform]}{volatile}`

  Any targets associated with tags in the snmpNotifyTable are notified, based on the filter profile associated with the target.
- To display the notifications that are set, use the following command:

  `show snmpv3 notify {[[hex hex_notify_name] | notify_name]}`

- To delete an entry from the snmpNotifyTable, use the following command:

  ```
  configure snmpv3 delete notify [{[[hex hex_notify_name] |
  notify_name]} | all-non-defaults]
  ```

*Configuring Notifications*

Because the target parameters name points to a number of objects used for notifications, configure the target parameter name entry first.

You can then configure the target address, filter profiles and filters, and any necessary notification tags.

## Access Profile Logging for SNMP

The access profile logging feature allows you to use an *ACL* policy file or dynamic ACL rules to control access to *SNMP* services on the switch.

When access profile logging is enabled for SNMP, the switch logs messages and increments counters when packets are denied access to SNMP. No messages are logged for permitted access.

You can manage SNMP access using one (not both) of the following methods:

- Create and apply an ACL policy file.
- Define and apply individual ACL rules.

One advantage of ACL policy files is that you can copy the file and use it on other switches. One advantage to applying individual ACL rules is that you can enter the rules at the CLI command prompt, which can be easier than opening, editing, and saving a policy file.

*ACL Match Conditions and Actions*

The ACLs section describes how to create *ACL* policies and rules using match conditions and actions. Access profile logging supports the following match conditions and actions:

- Match conditions
  - Source-address—IPv4 and IPv6
- Actions
  - Permit
  - Deny

If the ACL is created with more match conditions or actions, only those listed above are used for validating the packets. All other conditions and actions are ignored.

The source-address field allows you to identify an IPv4 address, IPv6 address, or subnet mask for which access is either permitted or denied.

If the *SNMP* traffic does not match any of the rules, the default behavior is deny.

*Limitations*

Access profile logging for *SNMP* has the following limitations:

- Either policy files or *ACL* rules can be associated with SNMP, but not both at the same time.
- Only source-address match is supported.

- Access-lists that are associated with one or more applications (SNMP or Telnet, for example) cannot be directly deleted. They must be unconfigured from the application first and then deleted from the CLI.
- Default counter support is added only for ACL rules and not for policy files.

### *Managing ACL Policies for SNMP*

The ACLs section describes how to create *ACL* policy files.

- Configure *SNMP* to use an ACL policy using one of the following commands:

  ```
  configure snmp access-profile profile_name
  configure snmp access-profile profile_name readonly
  configure snmp access-profile profile_name readwrite
  ```

  By default, SNMP supports the **readwrite** option. However, you can specify the **readonly** or **readwrite** option to change the current configuration.
- To configure SNMP to remove a previously configured ACL policy, use the following command:

  ```
  configure snmp access-profile none
  ```

### *Managing ACL Rules for SNMP*

Before you can assign an *ACL* rule to *SNMP*, you must create a dynamic ACL rule as described in ACLs.

- To add or delete a rule for SNMP access, use the following command:

  ```
  configure snmp access-profile [ access_profile {readonly | readwrite}
  | [[add rule ] [first | [[before | after] previous_rule]]] | delete
  rule | none ]
  ```
- To display the access-list permit and deny statistics for an application, use the following command:

  ```
  show access-list counters process [snmp | telnet | ssh2 | http]
  ```

### *Misconfiguration Error Messages*

The following messages can appear during configuration of policies or rules for the *SNMP* service:

| | |
|---|---|
| `Rule <rule> is already applied` | A rule with the same name is already applied to this service. |
| `Please remove the policy <policy> already configured, and then add rule <rule>` | A policy file is already associated with the service. You must remove the policy before you can add a rule. |
| `Rule <previous_rule> is not already applied` | The specified rule has not been applied to the service, so you cannot add a rule in relation to that rule. |
| `Rule <rule> is not applied` | The specified rule has not been applied to the service, so you cannot remove the rule from the service. |
| `Error: Please remove previously configured rule(s) before configuring policy <policy>` | A policy or one or more ACL rules are configured for the service. You must delete the remove the policy or rules from the service before you can add a policy. |

## Using the Simple Network Time Protocol

ExtremeXOS supports the client portion of the *SNTP* Version 3 based on RFC1769.

SNTP can be used by the switch to update and synchronize its internal clock from a Network Time Protocol (NTP) server. After SNTP has been enabled, the switch sends out a periodic query to the indicated NTP server, or the switch listens to broadcast NTP updates. In addition, the switch supports the configured setting for Greenwich Mean time (GMT) offset and the use of Daylight Saving Time.

## Configuring and Using SNTP

To use *SNTP*:

1. Identify the host(s) that are configured as NTP server(s). Additionally, identify the preferred method for obtaining NTP updates. The options are for the NTP server to send out broadcasts, or for switches using NTP to query the NTP server(s) directly. A combination of both methods is possible. You must identify the method that should be used for the switch being configured.
2. Configure the Greenwich Mean Time (GMT) offset and Daylight Saving Time preference. The command syntax to configure GMT offset and usage of Daylight Saving Time is as follows:

   `configure timezone {name tz_name} GMT_offset {autodst {name dst_timezone_ID} {dst_offset} begins [every floatingday | on absoluteday] {at time_of_day_hour time_of_day_minutes} {ends [every floatingday | on absoluteday] {at time_of_day_hour time_of_day_minutes}}}`

   By default beginning in 2007, Daylight Saving Time is assumed to begin on the second Sunday in March at 2:00 AM, and end the first Sunday in November at 2:00 AM and to be offset from standard time by one hour.

   a. If this is the case in your time zone, you can set up automatic daylight saving adjustment with the command:

      `configure timezone GMT_offset autodst`

   b. If your time zone uses starting and ending dates and times that differ from the default, you can specify the starting and ending date and time in terms of a floating day, as follows:

      ```
      configure timezone name MET 60 autodst name MDT begins every last sunday march
      at 1 30 ends every last sunday october at 1 30
      ```

   c. You can also specify a specific date and time, as shown in the following command:

      ```
      configure timezone name NZST 720 autodst name NZDT 60 begins every first sunday
      october
      at 2 00 ends on 3 16 2004 at 2 00
      ```

      The optional time zone IDs are used to identify the time zone in display commands such as `show switch {detail}`.

      The following table describes the time zone command options in detail.

**Table 18: Time Zone Configuration Command Options**

| | |
|---|---|
| tz_name | Specifies an optional name for this timezone specification. May be up to six characters in length. The default is an empty string. |
| GMT_offset | Specifies a Greenwich Mean Time (GMT) offset, in + or - minutes. |
| autodst | Enables automatic Daylight Saving Time. |
| dst_timezone_ID | Specifies an optional name for this Daylight Saving Time specification. May be up to six characters in length. The default is an empty string. |

**Table 18: Time Zone Configuration Command Options (continued)**

| | |
|---|---|
| dst_offset | Specifies an offset from standard time, in minutes. Value is from 1–60. The default is 60 minutes. |
| floatingday | Specifies the day, week, and month of the year to begin or end Daylight Saving Time each year. Format is *week day month* where:<br>• *week* is specified as [first \| second \| third \| fourth \| last]<br>• *day* is specified as [sunday \| monday \| tuesday \| wednesday \| thursday \| friday \| saturday]<br>• *month* is specified as [january \| february \| march \| april \| may \| june \| july \| august \| september \| october \| november \| december]<br>Default for beginning is second sunday march; default for ending is first sunday november. |
| absoluteday | Specifies a specific day of a specific year on which to begin or end DST. Format is *month day year* where:<br>• *month* is specified as 1–12<br>• *day* is specified as 1–31<br>• *year* is specified as 1970–2035<br>The year must be the same for the begin and end dates. |
| time_of_day_hour | Specifies the time of day to begin or end Daylight Saving Time. May be specified as an hour (0–23). The default is 2. |
| time_of_day_minutes | Specify the minute to begin or end Daylight Saving Time. May be specified as a minute (0–59). |
| noautodst | Disables automatic Daylight Saving Time. |

3. Automatic Daylight Saving Time changes can be enabled or disabled.

   The default setting is enabled. To disable automatic Daylight Saving Time, use the command:

   `configure timezone {name tz_name} GMT_offset noautodst`

4. Enable the SNTP client using the following command:

   `enable sntp-client`

   After SNTP has been enabled, the switch sends out a periodic query to the NTP servers defined in the next step (if configured) or listens to broadcast NTP updates from the network. The network time information is automatically saved into the onboard real-time clock.

5. If you would like this switch to use a directed query to the NTP server, configure the switch to use the NTP server(s). An NTP server can be an IPv4 address or an IPv6 address or a hostname. If the switch listens to NTP broadcasts, skip this step. To configure the switch to use a directed query, use the following command:

   `configure sntp-client [primary | secondary] host-name-or-ip {vr vr_name}`

   The following two examples use an IPv6 address as an NTP server and a hostname as an NTP server:

   ```
   configure sntp-client primary fd98:d3e2:f0fe:0:54ae:34ff:fecc:892
   configure sntp-client primary ntpserver.mydomain.com
   ```

   NTP queries are first sent to the primary server. If the primary server does not respond within one second, or if it is not synchronized, the switch queries the secondary server (if one is configured). If the switch cannot obtain the time, it restarts the query process. Otherwise, the switch waits for the sntp-client update interval before querying again.

6. Optionally, the interval for which the SNTP client updates the real-time clock of the switch can be changed using the following command:

   ```
   configure sntp-client update-interval update-interval
   ```

   The default sntp-client update-interval value is 64 seconds.
7. Verify the configuration.
   a. `show sntp-client`

      This command provides configuration and statistics associated with SNTP and its connectivity to the NTP server.
   b. `show switch {detail}`

      This command indicates the GMT offset, the Daylight Saving Time configuration and status, and the current local time.

NTP updates are distributed using GMT time.

To properly display the local time in logs and other time-stamp information, the switch should be configured with the appropriate GMT offset to GMT based on geographical location.

### *GMT Offsets*

The following table lists offsets for GMT.

**Table 19: Greenwich Mean Time Offsets**

| GMT Offset in Hours | GMT Offset in Minutes | Common Time Zone References | Cities |
|---|---|---|---|
| +0:00 | +0 | GMT - Greenwich Mean<br>UT or UTC - Universal (Coordinated)<br>WET - Western European | London, England; Dublin, Ireland; Edinburgh, Scotland; Lisbon, Portugal; Reykjavik, Iceland; Casablanca, Morocco |
| -1:00 | -60 | WAT - West Africa | Cape Verde Islands |
| -2:00 | -120 | AT - Azores | Azores |
| -3:00 | -180 | | Brasilia, Brazil; Buenos Aires, Argentina; Georgetown, Guyana |
| -4:00 | -240 | AST - Atlantic Standard | Caracas; La Paz |
| -5:00 | -300 | EST - Eastern Standard | Bogota, Columbia; Lima, Peru; New York, NY, Trevor City, MI USA |
| -6:00 | -360 | CST - Central Standard | Mexico City, Mexico |
| -7:00 | -420 | MST - Mountain Standard | Saskatchewan, Canada |
| -8:00 | -480 | PST - Pacific Standard | Los Angeles, CA, Santa Clara, CA, Seattle, WA USA |
| -9:00 | -540 | YST - Yukon Standard | |
| -10:00 | -600 | AHST - Alaska-Hawaii Standard<br>CAT - Central Alaska<br>HST - Hawaii Standard | |
| -11:00 | -660 | NT - Nome | |

**Table 19: Greenwich Mean Time Offsets (continued)**

| GMT Offset in Hours | GMT Offset in Minutes | Common Time Zone References | Cities |
|---|---|---|---|
| -12:00 | -720 | IDLW - International Date Line West | |
| +1:00 | +60 | CET - Central European<br>FWT - French Winter<br>MET - Middle European<br>MEWT - Middle European Winter<br>SWT - Swedish Winter | Paris France; Berlin, Germany; Amsterdam, The Netherlands; Brussels, Belgium; Vienna, Austria; Madrid, Spain; Rome, Italy; Bern, Switzerland; Stockholm, Sweden; Oslo, Norway |
| + 2:00 | +120 | EET - Eastern European, Russia Zone 1 | Athens, Greece; Helsinki, Finland; Istanbul, Turkey; Jerusalem, Israel; Harare, Zimbabwe |
| +3:00 | +180 | BT - Baghdad, Russia Zone 2 | Kuwait; Nairobi, Kenya; Riyadh, Saudi Arabia; Moscow, Russia; Tehran, Iran |
| +4:00 | +240 | ZP4 - Russia Zone 3 | Abu Dhabi, UAE; Muscat; Tblisi; Volgograd; Kabul |
| +5:00 | +300 | ZP5 - Russia Zone 4 | |
| +5:30 | +330 | IST - India Standard Time | New Delhi, Pune, Allahabad, India |
| +6:00 | +360 | ZP6 - Russia Zone 5 | |
| +7:00 | +420 | WAST - West Australian Standard | |
| +8:00 | +480 | CCT - China Coast, Russia Zone 7 | |
| +9:00 | +540 | JST - Japan Standard, Russia Zone 8 | |
| +10:00 | +600 | EAST - East Australian Standard<br>GST - Guam Standard<br>Russia Zone 9 | |
| +11:00 | +660 | | |
| +12:00 | +720 | IDLE - International Date Line East<br>NZST - New Zealand Standard<br>NZT - New Zealand | Wellington, New Zealand; Fiji, Marshall Islands |

## SNTP Example

In this example, the switch queries a specific NTP server and a backup NTP server.

The switch is located in Cupertino, California, and an update occurs every 20 minutes. The commands to configure the switch are as follows:

```
configure timezone -480 autodst
configure sntp-client update-interval 1200
enable sntp-client
configure sntp-client primary 10.0.1.1
configure sntp-client secondary 10.0.1.2
```

# Using Zero Touch Provisioning (Auto Provisioning) on Edge Switches

Auto provisioning allows you to configure certain parameters on a switch automatically using a *DHCP* and TFTP server.

This process can make an Extreme Networks switch ready to do the initial provisioning without any manual intervention, resulting in time saving and efficiency.

The parameters that an auto-provision capable switch can obtain from a DHCP server and apply are as follows:

- IP address
- Gateway
- TFTP server to contact
- Configuration file to be loaded

A switch enabled with auto provision can be identified as follows:

- A warning message for the console and each Telnet session is displayed as follows:

  ```
  Note: This switch has Auto-Provision enabled to obtain configuration
  remotely. Commands should be limited to:
  show auto-provision
  show log
  Any changes to this configuration will be discarded at the next reboot
  if auto provisioning sends a ".cfg" file.
  ```

- The shell prompt reads as follows: `(auto-provision) SummitX #`
- The status is shown in the `show auto-provision` command.

The DHCP server can be any that provides the needed functionality.

To obtain the desired parameters, the following DHCP options are used:

- Option 43 - vendor-encapsulated-options
- Option 60 - vendor-class-identifier. Extreme Networks switches use “Switch-type” as the option 60 parameter. You must configure this option on your DHCP server to provide the required attributes based on the specific model.

Following is a sample Linux DHCP configuration:

```
option space EXTREME;
option EXTREME.tftp-server-ip code 100 = ip-address;
option EXTREME.config-file-name code 101 = text;
option EXTREME.snmp-trap-ip code 102 = ip-address;
class "Edge-without-POE" {
match if (option vendor-class-identifier = "XSummit");
vendor-option-space EXTREME;
option EXTREME.tftp-server-ip 10.120.89.80;
option EXTREME.config-file-name "XSummit_edge.cfg";
option EXTREME.snmp-trap-ip 10.120.91.89;
}
class "Edge-SummitX-POE" {
match if (option vendor-class-identifier = "XSummit");
vendor-option-space EXTREME;
option EXTREME.tftp-server-ip 10.120.89.80;
option EXTREME.config-file-name "xSummit_edge.xsf";
```

```
option EXTREME.snmp-trap-ip 10.120.91.89;
}
subnet 10.127.8.0  netmask 255.255.255.0 {
option routers                  10.127.8.254;
option domain-name-servers      10.127.8.1;
option subnet-mask              255.255.255.0;
pool {
deny dynamic bootp clients;
range 10.127.8.170 10.127.8.190;
allow members of "Edge-without-POE";
allow members of "Edge-SummitX-POE";
}
}
```

## Auto-provisioning Process

The auto-provisioning process is first initiated through the default *VLAN* (bound to *VR-Default*).

After three unsuccessful attempts to reach the network, the switch waits for 15 seconds before it switches over to the Mgmt VLAN (bound to *VR-Mgmt*). It continues this process until it reaches the network.

Delay in the auto-provisioning process results from the following configuration problems:

- The *DHCP* server may not be reachable.
- The configuration file has an invalid file extension. Only .cfg or .xsf is accepted.
- The TFTP server is unreachable.
- The configuration file name does not exist in the TFTP server.

You can use the `show log` command to view the exact problem reported.

An *SNMP* trap is sent out for these conditions when the SNMP-Trap-IP code (code 102) is configured in the DHCP server. The SNMP trap is not sent out when the DHCP server is unreachable.

When these conditions occur, the switch continues to retry to reach the network and remains in an “In Progress” state.

When there is a system error or internal problem, the switch moves to an auto-provision “Failed” state. The switch does not retry the auto-provisioning process once it has reached the “Failed” state.

Once the network is reached, the switch receives and configures the IP address and the gateway. The switch then executes the configuration file (.cfg or .xsf file), sends the trap to inform the user of the successful auto-provisioning (only when the SNMP-Trap-IP code, code 102, is configured), and reboots for the new configuration to take effect.

Following is the mandatory DHCP option configuration used for auto provision to work:

Standard Option:

1. IP address
2. Subnet mask
3. Gateway

Option 60:

1. Vendor identifier option

Option 43:

1. TFTP server IP address
2. Configuration file name

Optional DHCP option

1. SNMP trap receiver IP address

**Note**

The file uploaded to the TFTP server using the upload configuration command is an .xsf file extension configuration. An .cfg file extension configuration is created using the tftp put command.

Configuration changes made to the switch when auto provisioning is in progress will be appended if auto provisioning uses an .xsf file extension configuration, and it will be discarded if auto provisioning uses a .cfg file extension configuration.

## Auto-provisioning Configuration

- Auto provisioning is enabled by default. It can be restarted by clearing the switches configuration using `unconfig sw all` and rebooting the switch.

  **Note**

  Auto provisioning is not enabled on the *VLAN* (Mgmt or Default) if the IP address is already configured.

- To disable auto provision, use the following command:

  ```
  disable auto-provision
  ```

  When the `disable auto-provision` command is issued, the following message is displayed:

  ```
  This setting will take effect at the next reboot of this switch.
  ```

- To display the current state of auto provision on the switch, use the following command:

  ```
  show auto-provision
  ```

# Access Profile Logging for HTTP/HTTPS

The access profile logging feature allows you to use an *ACL* policy file or dynamic ACL rules to control access to Hypertext Transfer Protocol (HTTP) services on the switch.

When access profile logging is enabled for HTTP, the switch logs messages and increments counters when packets are denied access to HTTP. No messages are logged for permitted access.

**Note**

For more information on ExtremeXOS software support for HTTP, see Hypertext Transfer Protocol on page 936.

You can manage HTTP access using one (not both) of the following methods:

- Create and apply an ACL policy file
- Define and apply individual ACL rules

One advantage of ACL policy files is that you can copy the file and use it on other switches. One advantage to applying individual ACL rules is that you can enter the rules at the CLI command prompt, which can be easier than opening, editing, and saving a policy file.

## ACL Match Conditions and Actions

The ACLs section describes how to create *ACL* policies and rules using match conditions and actions. Access profile logging supports the following match conditions and actions:

- Match conditions
  - Source-address—IPv4 and IPv6
- Actions
  - Permit
  - Deny

If the ACL is created with more match conditions or actions, only those listed above are used for validating the packets. All other conditions and actions are ignored.

The source-address field allows you to identify an IPv4 address, IPv6 address, or subnet mask for which access is either permitted or denied.

If the *SNMP* traffic does not match any of the rules, the default behavior is deny.

## Limitations

Access profile logging for HTTP/HTTPS has the following limitations:

- Policy file support is not available for HTTP and HTTPS.
- Only source-address match is supported.
- Access-lists that are associated with one or more applications cannot be directly deleted. They must be unconfigured from the application first and then deleted from the CLI.

## Managing ACL Rules for HTTP

Before you can assign an *ACL* rule to HTTP, you must create a dynamic ACL rule as described in ACLs.

- To add or delete a rule for HTTP access, use the following command:

  `configure web http access-profile [[[`**`add`** *`rule`* `] [`**`first`** `| [[`**`before`** `| `**`after`**`] `*`previous_rule`*`]]] | `**`delete`** *`rule`* `| `**`none`** `]`
- To display the access-list permit and deny statistics for an application, use the following command:

  `show access-list counters process [`**`snmp`** `| `**`telnet`** `| `**`ssh2`** `| `**`http`**`]`

## Misconfiguration Error Messages

The following messages can appear during configuration of policies or rules for the *SNMP* service:

| | |
|---|---|
| `Rule <rule> is already applied` | A rule with the same name is already applied to this service. |
| `Please remove the policy <policy> already configured, and then add rule <rule>` | A policy file is already associated with the service. You must remove the policy before you can add a rule. |
| `Rule <previous_rule> is not already applied` | The specified rule has not been applied to the service, so you cannot add a rule in relation to that rule. |
| `Rule <rule> is not applied` | The specified rule has not been applied to the service, so you cannot remove the rule from the service. |
| `Error: Please remove previously configured rule(s) before configuring policy <policy>` | A policy or one or more ACL rules are configured for the service. You must delete the remove the policy or rules from the service before you can add a policy. |

# Managing the ExtremeXOS Software



The ExtremeXOS software platform is a distributed software architecture.

The distributed architecture consists of separate binary images organized into discrete software modules with messaging between them. The software and system infrastructure subsystem form the basic framework of how the ExtremeXOS applications interact with each other, including the system startup sequence, memory allocation, and error events handling. Redundancy and data replication is a built-in mechanism of ExtremeXOS. The system infrastructure provides basic redundancy support and libraries for all of the ExtremeXOS applications.

Note
For information about downloading and upgrading a new software image, saving configuration changes, and upgrading the BootROM, see Software Upgrade and Boot Options on page 1522.

Like any advanced operating system, ExtremeXOS gives you the tools to manage your switch and create your network configurations.

With the introduction of ExtremeXOS, the following enhancements and functionality have been added to the switch operating system:

- File system administration
- Configuration file management
- Process control
- Memory protection

### File system administration

With the enhanced file system, you can move, copy, and delete files from the switch. The file system structure allows you to keep, save, rename, and maintain multiple copies of configuration files on the switch. In addition, you can manage other entities of the switch such as policies and access control lists (ACLs).

### Configuration file management

With the enhanced configuration file management, you can oversee and manage multiple configuration files on your switch. In addition, you can upload, download, modify, and name configuration files used by the switch.

Process control

With process control, you can stop and start processes, restart failed processes, and update the software for a specific process or set of processes.

Memory protection

With memory protection, each function can be bundled into a single application module running as a memory protected process under real-time scheduling. In essence, ExtremeXOS protects each process from every other process in the system. If one process experiences a memory fault, that process cannot affect the memory space of another process.

The following sections describe in more detail how to manage the ExtremeXOS software.

## Using the ExtremeXOS File System

The file system in ExtremeXOS is the structure by which files are organized, stored, and named.

The switch can store multiple user-defined configuration and policy files, each with its own name. Using a series of commands, you can manage the files on your system. For example, you can rename or copy a configuration file on the switch, display a comprehensive list of the configuration and policy files on the switch, or delete a policy file from the switch.

Note

Filenames are case-sensitive. For information on filename restrictions, refer to the specific command in the *ExtremeXOS 16.2 Command Reference Guide*.

You can also download configuration and policy files from the switch to a network Trivial File Transfer Protocol (TFTP) server using TFTP. For detailed information about downloading switch configurations, see Software Upgrade and Boot Options on page 1522. For detailed information about downloading policies and ACLs, see ACLs on page 640.

With guidance from Extreme Networks Technical Support personnel, you can configure the switch to capture core dump files, which contain debugging information that is useful in troubleshooting situations. For more information about configuring core dump files and managing the core dump files stored on your switch, see Troubleshooting on page 1557.

### Moving or Renaming Files on the Switch

To move or rename an existing configuration, policy, or if configured, core dump file in the system. XML-formatted configuration files have a .cfg file extension. The switch runs only .cfg files. ASCII-formatted configuration files have an .xsf file extension. See Uploading ASCII-Formatted Configuration Files on page 1545 for more information. Policy files have a .pol file extension.

When you rename a file, make sure the renamed file uses the same file extension as the original file. If you change the file extensions, the file may be unrecognized by the system. For example, if you have an existing configuration file named test.cfg, the new filename must include the .cfg file extension.

1. Run the `mv` command.

```
mv test.cfg megset.cfg
Rename config test.cfg to config megtest.cfg on switch? (y/n)
```

2. Enter `y` to rename the file on your system. Enter `n` to cancel this process and keep the existing filename.

   If you attempt to rename an active configuration file (the configuration currently selected the boot the switch), the switch displays an error similar to the following:

   ```
   Error: Cannot rename current selected active configuration.
   ```

For more information about configuring core dump files and managing the core dump files stored on your switch, see Troubleshooting on page 1557.

## Copying Files on the Switch

The copy function allows you to make a copy of an existing file before you alter or edit the file. By making a copy, you can easily go back to the original file if needed.

XML-formatted configuration files have a .cfg file extension. The switch runs only .cfg files. ASCII-formatted configuration files have an .xsf file extension. See Uploading ASCII-Formatted Configuration Files on page 1545 for more information. Policy files have a .pol file extension.

When you copy a configuration or policy file from the system, make sure you specify the appropriate file extension. For example, if you want to copy a policy file, specify the filename and .pol.

1. Copy an existing configuration or policy file on your switch.

   ```
   "cp test.cfg test1.cfg
   cp test.pol test1.pol
   Copy config test.cfg to config test1.cfg on switch? (y/n)
   ```

2. Enter `y` to copy the file. Enter `n` to cancel this process and not copy the file.

   When you enter `y`, the switch copies the file with the new name and keeps a backup of the original file with the original name. After the switch copies the file, use the `ls` command to display a complete list of files.

   Note

   If you make a copy of a file, such as a core dump file, you can easily compare new information with the old file if needed.

For more information about configuring the storage of core dump files, see Troubleshooting on page 1557.

## Displaying Files on the Switch

To display a list of the configuration, policy, or if configured, core dump files stored on your switch.

Run `ls {`*`file_name`*`}`

When you do not specify a parameter, this command lists all of the files in the current directory stored on your switch.

If you do specify a parameter you can refer to a specific directory to view all of the files in that directory.

Output from this command includes the file size, date and time the file was last modified, and the file name.

For more information about configuring core dump files and managing the core dump files stored on your switch, see Troubleshooting.

## Transferring Files to and from the Switch

TFTP allows you to transfer files between a TFTP server and the following switch storage areas: local file system, internal memory card, compact flash card, and USB 2.0 storage device.

- Download a file from a TFTP server to the switch, using the `tftp` or `tftp get` commands.

  `tftp [ ip-address | host-name ] { -v vr_name } [ -g ] [ { -l local-file | } { -r remote-file } | { -r remote-file } { -l local-file } ]`

  `tftp get [ ip-address | host-name] { vr vr_name } remote-file {local-file} {force-overwrite}`

  **Note**
  By default, if you transfer a file with a name that already exists on the system, the switch prompts you to overwrite the existing file. For more information, see the `tftp get` command.

- To upload a file from the switch to a TFTP server, use the `tftp` or `tftp put` commands:

  `tftp [ ip-address | host-name ] { -v vr_name } [ -p ] [ { -l local-file | } { -r remote-file } | { -r remote-file } { -l local-file } ]`

  `tftp put [ ip-address | host-name] {vr vr_name} local-file { remote-file}`

For more information about TFTP, see Managing the ExtremeXOS Software. For detailed information about downloading software image files, BootROM files, and switch configurations, see Software Upgrade and Boot Options on page 1522. For more information about configuring core dump files and managing the core dump files stored on your switch, see Troubleshooting on page 1557.

## Deleting Files from the Switch

To delete a configuration, policy, or if configured, core dump file from your system, use the following command:

`rm file_name`

When you delete a configuration or policy file from the system, make sure you specify the appropriate file extension. For example, when you want to delete a policy file, specify the filename and .pol. After you delete a file, it is unavailable to the system.

When you delete a file from the switch, a message similar to the following appears:

```
Remove testpolicy.pol from switch? (y/n)
```

Enter `y` to remove the file from your system. Enter `n` to cancel the process and keep the file on your system.

If you attempt to delete an active configuration file (the configuration currently selected to boot the switch), the switch displays an error similar to the following:

```
Error: Cannot remove current selected active configuration.
```

For more information about configuring core dump files and managing the core dump files stored on your switch, see Troubleshooting on page 1557.

## Managing the Configuration File

The configuration is the customized set of parameters that you have selected to run on the switch. The following table describes some of the key areas of configuration file management in ExtremeXOS.

**Table 20: Configuration File Management**

| Task | Behavior |
|---|---|
| Configuration file database | ExtremeXOS supports saving a configuration file into any named file and supports more than two saved configurations.<br>For example, you can download a configuration file from a network TFTP server and save that file as primary, secondary, or with a user-defined name. You also select where to save the configuration: primary or secondary partition, or another space.<br>The file names primary and secondary exist for backward compatibility with ExtremeWare®. |
| Downloading configuration files | ExtremeXOS uses the `tftp` and `tftp get` commands to download configuration files from the network TFTP server to the switch.<br>For more information about downloading configuration files, see Using TFTP to Download the Configuration on page 1549. |
| Uploading configuration files | ExtremeXOS uses the `tftp` and `tftp put` commands to upload configuration files from the switch to the network TFTP server.<br>For more information about uploading configuration files, see Using TFTP to Upload the Configuration on page 1548. |
| Managing configuration files, including listing, copying, deleting, and renaming | The following commands allow you to manage configuration files:<br>• ls: Lists all of the configuration files in the system<br>• cp: Makes a copy of an existing configuration file in the system<br>• rm: Removes/deletes an existing configuration file from the system<br>• mv: Renames an existing configuration file |
| Configuration file types | • XML-formatted configuration file<br>• ASCII-formatted configuration file |
| XML-formatted configuration file | ExtremeXOS configuration files are saved in Extensible Markup Language (XML) format. Use the `show configuration` command to view on the CLI your currently running switch configuration. |

**Table 20: Configuration File Management (continued)**

| Task | Behavior |
|---|---|
| ASCII-formatted configuration file | You can upload your current configuration in ASCII format to a network TFTP server. The uploaded ASCII file retains the CLI format.<br>To view your configuration in ASCII format, save the configuration with the .xsf file extension (known as the XOS CLI script file). This saves the XML-based configuration in an ASCII format readable by a text editor.<br>ExtremeXOS uses the `upload configuration` command to upload the ASCII-formatted configuration file from the switch to the network TFTP server.<br>ExtremeXOS uses the `tftp` and `tftp get` commands to download configuration files from the network TFTP server to the switch.<br>For more information about ASCII-formatted configuration files, see Uploading ASCII-Formatted Configuration Files on page 1545. |
| XML configuration mode | Indicated by (xml) at the front of the switch prompt. Do not use; instead, run `disable xml-mode` to disable this mode. |
| Displaying configuration files | You can also see a complete list of configuration files by entering the `ls` command followed by the **[Tab]** key. |

For more information about saving, uploading, and downloading configuration files, see Save the Configuration on page 1547.

## Managing ExtremeXOS Processes

ExtremeXOS consists of a number of cooperating processes running on the switch.

With process control, under certain conditions, you can stop and start processes, restart failed processes, examine information about the processes, and update the software for a specific process or set of processes.

### Displaying Process Information

This procedure shows you how to display information about the processes in the system.

Run `show process {`*name*`} {`**detail**`} {`**description**`} {`**slot** *slotid*`}` .

Where the following is true:

- name: Specifies the name of the process.
- **detail**: Specifies more detailed process information, including memory usage statistics, process ID information, and process statistics.
- **description**: Describes the name of all of the processes or the specified process running on the switch.
- slotid: On a modular chassis, specifies the slot number of the MSM/MM. A specifies the MSM/MM installed in slot A. B specifies the MSM/MM installed in slot B. On a SummitStack, specifies the target node's slot number. The number is a value from 1 to 8. (This parameter is available only on modular switches and SummitStack.)

The `show process {name} {detail} {description} {slot slotid}` and `show process slot slotid` commands display the following information in a tabular format:

- Card: The name of the module where the process is running (modular switches only).
- Process Name: The name of the process.
- Version: The version number of the process. Options are:
  - Version number: A series of numbers that identify the version number of the process. This is helpful to ensure that you have version-compatible processes and if you experience a problem.
  - Not Started: The process has not been started. This can be caused by not having the appropriate license or for not starting the process.
- Restart: The number of times the process has been restarted. This number increments by one each time a process stops and restarts.
- State: The current state of the process. Options are:
  - No License: The process requires a license level that you do not have. For example, you have not upgraded to that license, or the license is not available for your platform.
  - Ready: The process is running.
  - Stopped: The process has been stopped.
- Start Time: The current start time of the process. Options are:
  - Day/Month/Date/Time/Year: The date and time the process began. If a process terminates and restarts, the start time is also updated.
  - Not Started: The process has not been started. This can be caused by not having the appropriate license or for not starting the process.

When you specify the **detail** keyword, more specific and detailed process information is displayed.

The `show process detail` and `show process slot slotid detail` commands display the following information in a multi-tabular format:

- Detailed process information
- Memory usage configurations
- Recovery policies
- Process statistics
- Resource usage

## Stopping a Process

If recommended by Extreme Networks Technical Support personnel, you can stop a running process. You can also use a single command to stop and restart a running process during a software upgrade on the switch.

By using the single command, there is less process disruption and it takes less time to stop and restart the process.

- To stop a running process, use the following command:

  `terminate process name [forceful | graceful] {msm slot}`

  In a SummitStack:

  `terminate process name [forceful | graceful] {slot slot}`

Where the following is true:

- name: Specifies the name of the process.
- forceful: Specifies that the software quickly terminate a process. Unlike the **graceful** option, the process is immediately shutdown without any of the normal process cleanup.
- graceful: Specifies that the process shutdown gracefully by closing all opened connections, notifying peers on the network, and other types of process cleanup.
- slot: For a modular chassis, specifies the slot number of the MSM/MM. A specifies the MSM/MM installed in slot A. B specifies the MSM/MM installed in slot B. On a SummitStack, specifies the target node's slot number. The number is a value from 1 to 8. (This parameter is available only on modular switches and SummitStack.)

**Note**

Do not terminate a process that was installed since the last reboot unless you have saved your configuration. If you have installed a software module and you terminate the newly installed process without saving your configuration, your module may not be loaded when you attempt to restart the process with the `start process` command.

- To preserve a process's configuration during a terminate and (re)start cycle, save your switch configuration before terminating the process. Do not save the configuration or change the configuration during the process terminate and re(start) cycle. If you save the configuration after terminating a process, and before the process (re)starts, the configuration for that process is lost.
- To stop and restart a process during a software upgrade, use the following command:

  `restart process [class cname | name {msm slot}]`

  Where the following is true:

  - cname: Specifies that the software terminates and restarts all instances of the process associated with a specific routing protocol on all VRs.
  - name: Specifies the name of the process.

## Starting a Process

- To start a process, use the following command:

  `start process name {msm slot}`

  In a SummitStack:

  `start process name {slot slot}`

  Where the following is true:

  - name: Specifies the name of the process.
  - slot: For a modular chassis, specifies the slot number of the MSM/MM. A specifies the MSM/MM installed in slot A. B specifies the MSM/MM installed in slot B. On a SummitStack, specifies the slot number of the target node. The number is a value from 1 to 8. (This parameter is available only on modular switches and SummitStack.)

    You are unable to start a process that is already running.If you try to start a currently running process, for example telnetd, an error message similar to the following appears:

```
Error: Process telnetd already exists!
```

**Note**

After you stop a process, do not change the configuration on the switch until you start the process again. A new process loads the configuration that was saved prior to stopping the process. Changes made between a process termination and a process start are lost. Else, error messages can result when you start the new process.

As described in the section, Stopping a Process on page 112, you can use a single command, rather than multiple commands, to stop and restart a running process.

- Stop and restart a process during a software upgrade.

  `restart process [class cname | name {msm slot}]`

  In a SummitStack:

  `restart process [class cname | name {slot slot}]`

For more detailed information, see the *ExtremeXOS 16.2 Command Reference Guide*.

## Understanding Memory Protection

ExtremeXOS provides memory management capabilities.

With ExtremeXOS, each process runs in a protected memory space. This infrastructure prevents one process from overwriting or corrupting the memory space of another process. For example, if one process experiences a loop condition, is under some type of attack, or is experiencing some type of problem, that process cannot take over or overwrite another processes' memory space.

Memory protection increases the robustness of the system. By isolating and having separate memory space for each individual process, you can more easily identify the process or processes that experience a problem.

To display the current system memory and that of the specified process, use the following command: `show memory process name {slot slotid}`

Where the following is true:

- name: Specifies the name of the process.
- slot: On a modular chassis, specifies the slot number of the MSM/MM. A specifies the MSM/MM installed in slot A. B specifies the MSM/MM installed in slot B. On a SummitStack, specifies the slot number of the target node. The number is a value from 1 to 8. (This parameter is available only on modular switches and SummitStack.)

The `show memory process` command displays the following information in a tabular format:

- System memory information (both total and free)
- Current memory used by the individual processes

The current memory statistics for the individual process also includes the following:

- The module (MSM A or MSM B) and the slot number of the MSM/MM (modular switches only)
- The name of the process

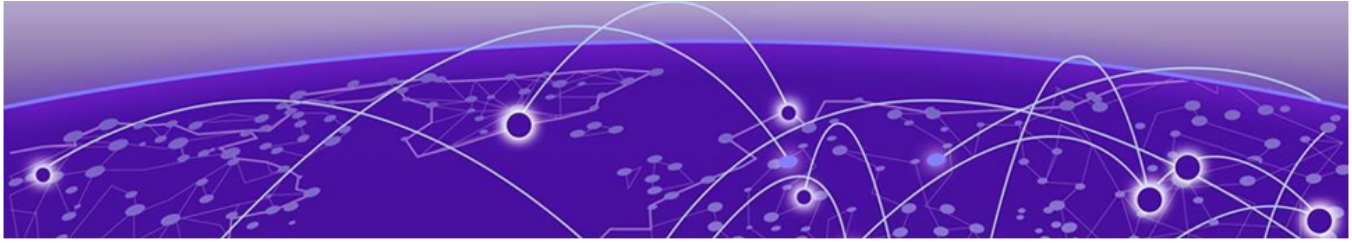
You can also use the `show memory {slot [slotid | a | b]}` command to view the system memory and the memory used by the individual processes, even for all processes on all MSMs/MMs installed in modular switches. The **slot** parameter is available only on modular switches and SummitStack.

In general, the *free* memory count for an MSM/MM or Summit family switch decreases when one or more running processes experiences an increase in memory usage. If you have not made any system configuration changes, and you observe a continued decrease in free memory, this might indicate a memory leak.

The information from these commands may be useful for your technical support representative if you experience a problem.

The following is sample truncated output from a Summit family switch:

```
CPU Utilization Statistics - Monitored every 25 seconds
-------------------------------------------------------------------------
Process         5   10   30    1     5    30   1     Max       Total
              secs secs secs min  mins  mins hour          User/System
              util util util util util util util    util    CPU Usage
              (%)  (%)  (%)  (%)   (%)  (%)  (%)     (%)       (secs)
System        n/a  n/a  0.0  0.9   0.1  0.2  0.5                34.6
-------------------------------------------------------------------------

aaa            n/a  n/a  0.0  0.0  0.0  0.0  0.0  1.8  1.72      0.78

acl            n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  0.40      0.24

bgp            n/a  n/a  0.0  0.0  0.0  0.0  0.0  12.6 11.18     2.21

cfgmgr         n/a  n/a  0.0  0.0  0.0  0.0  0.8  39.8 4743.92   3575.79

cli            n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  0.59      0.42

devmgr         n/a  n/a  0.0  0.0  0.0  0.0  0.0  19.5 74.44     24.52

dirser         n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  0.0       0.0

dosprotect     n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  0.8       0.12

eaps           n/a  n/a  0.0  0.0  0.0  0.0  0.1  5.5  36.40     15.41

edp            n/a  n/a  0.0  0.0  0.0  0.0  0.0  11.1 10.92     3.97

elrp           n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  0.49      0.44

ems            n/a  n/a  0.0  0.0  0.0  0.0  0.0  0.0  1.19      1.29

epm            n/a  n/a  0.0  0.0  0.0  0.0  0.0  30.7 48.74     32.93

esrp           n/a  n/a  0.0  0.0  0.0  0.0  0.0  2.7  0.82      0.45

etmon          n/a  n/a  0.0  0.0  0.0  0.0  0.5  30.5 4865.78   873.87

...
```



# Configuring Stacked Switches



A stack consists of a group of up to eight switches that are connected to form a ring. The stack offers the combined port capacity of the individual switches. But it operates as if it were a single switch, making network administration easier.

Stacking is facilitated by the SummitStack feature – part of the ExtremeXOS Edge license.

This chapter contains information about configuring a stack, maintaining the stack configuration, and troubleshooting.

Refer to the Stacking chapter in the *Extreme Switching and Summit Switches: Hardware Installation Guide for ExtremeXOS 16.x or Earlier* for descriptions of the supported configurations for stacking, the considerations for planning a stack, and the steps for setting up the hardware. We recommend that you read that chapter before installing the switches that will make up the stack.

## Introduction to Stacking

Using the SummitStack feature – part of the ExtremeXOS Edge license – a stack can combine switches from different series, provided that every switch in the stack:

- Runs in the same partition (primary or secondary).
- Runs the same version of ExtremeXOS.
- Includes support for stacking.

The stack operates as if it were a single switch with a single IP address and a single point of authentication. One switch – called the master switch – is responsible for running network protocols and managing the stack. The master runs ExtremeXOS software and maintains all the software tables for all the switches in the stack.

All switches in the stack, including the master switch, are called nodes. Figure 4 shows four nodes in a stack, connected to each other by SummitStack cables.

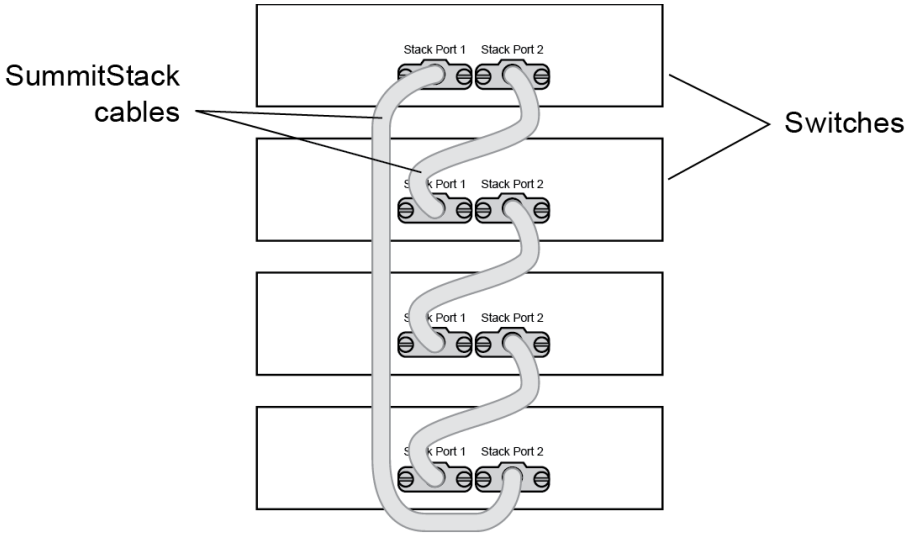



**Figure 4: Switches Connected to Form a Stack**

The following sections introduce you to the basic principles of stacking and provide recommendations for creating stacks.

More information to answer your questions about stacking and help you plan your configuration is available on the Extreme Networks GTAC Knowledge Base.

## Building Basic Stacks

A stack can be created in either of two ways:

- In *native stacking*, switches are connected using either designated Ethernet data ports or dedicated stacking connectors.
- In *alternate stacking*, switches are connected using 10-Gbps Ethernet data ports that have been configured for stacking. These ports are located either on the switch itself or on option cards installed on the rear of the switch.

When planning and building your stack, be sure to follow port compatibility and cabling recommendations as described in this chapter.

See *Extreme Switching and Summit Switches: Hardware Installation Guide for ExtremeXOS 16.x or Earlier* for information about which switch series can be combined to form a stack.

### *Slot Numbers in Stacks*

A switch stack can be thought of as a virtual chassis. Each switch (node) operates as if it were occupying a slot in a chassis and is controlled by the master. The high-speed stacking links function like the backplane links of a chassis.

Each switch in the stack is assigned a “slot number” during the initial software configuration of the stack. Starting at the switch with the console connection, numbers are assigned in numerical order following the physical path of the connected stacking cables. For example, if you follow the cabling recommendations presented in *Extreme Switching and Summit Switches: Hardware Installation Guide for ExtremeXOS 16.x or Earlier* and configure a vertical stack from the console on the switch at the top of the physical stack, the switches will be assigned slot numbers 1 through 8 from the top down.

Some stackable switches have a seven-segment LED, called the stack number indicator on the front panel. (See Figure 5.) When a stack is operating, the indicator displays the switch's slot number. This LED does not light on switches that are not currently operating as part of a stack.

The top half of the number blinks if the switch is the master, and the bottom half blinks if it is the backup. If the LED is steadily lit, the switch is a standby. If the LED is off the switch is not configured as a member of a stack.

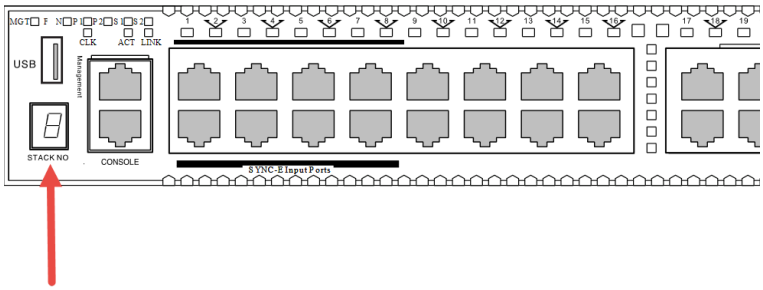



**Figure 5: Position of the Stack Number Indicator (X460-G2 Switch Shown)**

In addition to the Stack Number Indicator, each stacking port has an LED. The LED is steady green if the link is OK, blinking green if traffic is present, and off if no signal is present.

A quick way to verify that the cable connections match the software configuration is to check the stack number indicator on each switch. If the slot numbers do not line up in the order you arranged the switches, this might indicate that the stacking cable setup differs from what you intended when you configured the software. In this case, reconnect the cables in the correct order and perform the software configuration again.

### *Master/Backup Switch Redundancy*

When your stack is operational, one switch is the master switch, responsible for running network protocols and managing the stack.

To provide recovery in case of a break in the stack connections, you can configure redundancy by designating a backup switch to take over as master if the master switch fails. When you perform the initial software configuration of the stack, the "easy setup" configuration option automatically configures redundancy, with slot 1 as the master and slot 2 as the backup. You can also configure additional switches as "master-capable," meaning they can become a stack master in case the initial backup switch fails.

When assigning the master and backup roles in mixed stacks, consider the feature scalability and the speed of each switch model. The easy setup configuration process selects master and backup switches, based on capability and speed, in the following order:

1. Summit X670-G2
2. Summit X460-G2
3. Summit X770
4. Summit X450-G2
5. ExtremeSwitching X440-G2 and X620

For example, in a stack that combines Summit X460-G2 or X670-G2 switches with other switch models, an X460-G2 or X670-G2 switch might provide more memory and more features than other switches in

the stack. Consider these differences when selecting a master node, selecting a backup node, and configuring failover operation.

**Note**
We recommend that the master and backup roles be assigned to switches from the same series. For example, if the master node is an X460-G2 switch, the backup node should also be an X460-G2 switch. Similarly, if the master node is an X670-G2 series switch, the backup node should also be an X670-G2 switch.

**Note**
ExtremeSwitching X690 and X870 switches can be stacked with each other, but they cannot be stacked with other switch models.

When easy setup compares two switches that have the same capability, the lower slot number takes precedence.

We recommend that you follow the same ranking hierarchy when you plan the physical placement of the switches in the stack.

## SummitStack Topologies

Figure 6 presents a graphical representation of a stack and some of the terms that describe stack conditions.

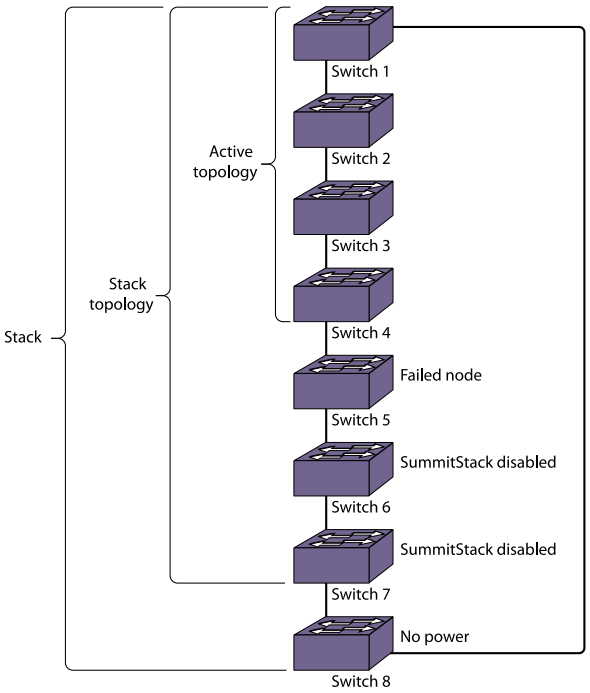



**Figure 6: Example of a Stack, Showing the Active Topology and the Stack Topology**

A stack is the collection of all switches, or nodes, that are cabled together to form one virtual switch using the ExtremeXOS SummitStack feature.

The maximum cable length supported between switches depends on the types of switches in your stack, the installed option cards, and the configured stacking ports. For more information, see *Extreme Switching and Summit Switches: Hardware Installation Guide for ExtremeXOS 16.x or Earlier*.

A stack topology is the set of contiguous nodes that are powered up and communicating with each other. In the example shown, Switch 8 is not part of the stack topology because it is not powered up.

An active topology is the set of contiguous nodes that are active. An active node is powered up, configured for stack operation, and communicating with the other active nodes.

Switch 5 in the example has failed, and stacking is disabled on Switch 6 and Switch 7. Switch 8 has no power, so the active topology includes switches: Switch 1, Switch 2, Switch 3, and Switch 4.

For more information on SummitStack terminology, see SummitStack Terms on page 125.

### *Ring Topology*

SummitStack nodes should be connected to each other in a ring topology. In a ring topology, one link is used to connect to a node and the other link is used to connect to another node. The result forms a physical ring connection. This topology is highly recommended for normal operation.

Figure 7 represents a maximal ring topology of eight active nodes.

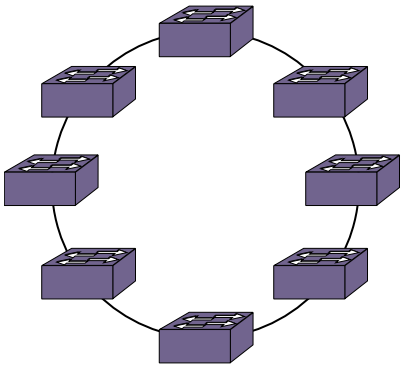


**Figure 7: Graphical Representation of a Ring Topology**

Figure 8 shows what the same ring topology would look in actual practice. Each switch in the rack is connected to the switch above it and the switch below it. To complete the ring, a longer cable connects Switch 1 with Switch 8.

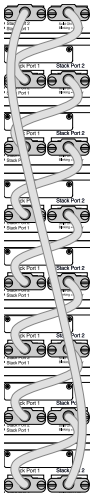


**Figure 8: Summit Family Switches in a Ring Topology**

Note that, while a physical ring connection may be present, a ring active topology exists only when all nodes in the stack are active.

### *Daisy Chain Topology*

Stackable switches can be connected in a daisy-chain topology. This is a ring topology with one of the links disconnected, inoperative, or disabled. A daisy chain can be created when a link fails or a node reboots in a ring topology, but the daisy chain topology is not recommended for normal operation.

In Figure 9, the nodes delineated as the active topology are operating in a daisy-chain configuration, even though there is physically a ring connection in the stack.

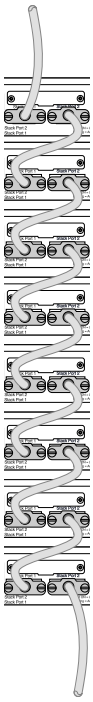


**Figure 9: Daisy-Chain Topology**

You might need to use a daisy chain topology while adding a new node, removing a node, or joining two stacks.

If you are using a daisy chain topology, the possibility of a dual master situation increases. Before you create a daisy chain topology, read Managing a Dual Master Situation on page 168.

## Use Ethernet Ports for Stacking (SummitStack-V Feature)

On many Extreme Networks switches, you can reconfigure one or two 10-Gbps Ethernet data ports to operate as stacking ports.

This feature, known as *SummitStack-V* or *alternate stacking*, means that you can use less expensive cables to connect the switches in a stack. Because copper and fiber Ethernet ports support longer cable distances, you can also extend the physical distance between stack nodes – connecting, for example, switches on different floors in a building or in different buildings on a campus.

The SummitStack-V feature means that you can stack switches that have no dedicated (or *native*) stacking ports but that do have at least two Ethernet ports. The ports can be configured to support either data communications or the stacking protocol. When configured to support stacking, they are called alternate stacking ports to distinguish them from the native stacking ports that use custom cables.

A single stack can use both native stacking ports and alternate stacking ports. On one switch, for example, you can use a native stacking port to connect to a switch in the same rack, and you can use an alternate stacking port to connect to a switch on a different floor.

Note
When you connect distant nodes using alternate stacking ports, be sure to run the cables over physically different pathways to reduce the likelihood of a cut affecting multiple links.

On each switch model, only specific data ports can be used as alternate stacking ports. The alternate stacking ports must be 10-Gbps Ethernet ports, either on the front panel of the switch or on installed port option cards or versatile interface modules at the rear of the switch. Switch models that do not have native stacking ports can still use alternate stacking if they have 10-Gbps Ethernet ports.

Alternate stacking ports on different switches must be directly connected, with no intervening switch connections. This is because alternate stacking ports use the proprietary ExtremeXOS stacking protocol, not the standard Ethernet protocol.

Table 21 lists the data ports that can be used as native and alternate stacking ports for each switch model.

When the stacking-support option is enabled (with the `enable stacking-support` command), data communication stops on the physical data ports that are designated for alternate stacking. Then, when stacking is enabled (with the `enable stacking` command), those ports – listed in the

Alternate Stacking Ports column of Table 21 – operate using the stacking protocol for the logical stacking ports.

**Table 21: Native and Alternate Stacking Ports**

| Switch Model | Type or location of Native Stacking Ports | Alternate Stacking Ports | Location of Alternate Stacking Ports |
|---|---|---|---|
| X440-24t-10G<br>X440-24x-10G<br>X440-24p-10G | None | 25,26 | Front panel |
| X440-48t-10G<br>X440-48p-10G | None | 49,50 | Front panel |
| X450a-24t<br>X450a-24tDC<br>X450a-24x<br>X450a-24xDC<br>X450e-24t<br>X450e-24p | Fixed (rear panel) | 25,26 | XGM2-2xf or<br>XGM2-2xn or<br>XGM2-2sf or<br>XGM2-2bt |
| X450a-48t<br>X450a-48tDC<br>X450e-48t<br>X450e-48p | Fixed (rear panel) | 49,50 | XGM2-2xf or<br>XGM2-2xn or<br>XGM2-2sf or<br>XGM2-2bt |
| X450-G2-24t-10GE4<br>X450-G2-24p-10GE4 | Fixed (rear panel) | 27,28 | Front panel |
| X450-G2-48t-10GE4<br>X450-G2-48p-10GE4 | Fixed (rear panel) | 51,52 | Front panel |
| X460-24t<br>X460-24x<br>X460-24p | SummitStack module or<br>SummitStack-V80 module | S1,S2<br>(29,30) | XGM3-2sf or<br>XGM3S-2sf or<br>XGM3S-2xf |
| X460-48t<br>X460-48p | SummitStack module or<br>SummitStack-V80 module | S1,S2<br>(53,54) | XGM3-2sf or<br>XGM3S-2sf or<br>XGM3S-2xf |
| X460-48x | SummitStack module or<br>SummitStack-V80 module | S1,S2<br>(49,50) | XGM3S-2sf or<br>XGM3S-2xf |
| X460-G2-24t-GE4<br>X460-G2-24p-GE4 | VIM-2ss or VIM-2q | 33,34 | VIM-2t or VIM-2x |
| X460-G2-48t-GE4<br>X460-G2-48p-GE4 | VIM-2ss or VIM-2q | 53,54 | VIM-2t or VIM-2x |
| X460-G2-24t-10GE4<br>X460-G2-24x-10GE4<br>X460-G2-24p-10GE4 | VIM-2ss or VIM-2q | 31,32 | Front panel |
| X460-G2-48t-10GE4<br>X460-G2-48x-10GE4<br>X460-G2-48p-10GE4 | VIM-2ss or VIM-2q | 51,52 | Front panel |

**Table 21: Native and Alternate Stacking Ports (continued)**

| Switch Model | Type or location of Native Stacking Ports | Alternate Stacking Ports | Location of Alternate Stacking Ports |
|---|---|---|---|
| X480-24x | None: VIM has only data ports | S3,S4 (29,30) | VIM2-10G4X module |
| | None: No installed VIM | 25,26 | Front panel |
| | VIM2-SummitStack module<br>VIM2-SummitStack128 module<br>VIM2-SummitStack-V80 module<br>VIM3-40G4X | 25,26 | Front panel |
| X480-48t<br>X480-48x | None: VIM has only data ports | S3,S4 (51,52) | VIM2-10G4X module |
| X670-48x | None | 47,48 | Front panel |
| X670V-48t<br>X670V-48x | VIM4-40G4X | 47,48 | Front panel |
| X670-G2-48x-4q | Ports 49,53,57,61 | 47,48 | Front panel |
| X670-G2-72x | None | 71,72 | Front panel |
| X770-32q | Ports 101,102,103,104 | 103,104 | Front panel |

## Available Stacking Methods

Each ExtremeSwitching and Summit switch model can use various methods of stacking.

Table 22 shows the switch models that can participate in each stacking method.

**Table 22: Summit Stacking by Stacking Method**

| Stacking Method | Speed per Link (HDX) | Cable Type and Lengths | Switch Models |
|---|---|---|---|
| SummitStack | 10 Gbps | 0.5 m, 1.5 m, 3.0 m, 5.0 m,<br>20Gb Stacking Cable | Summit X440, X460, X460-G2, X480 |
| SummitStack-V | 10 Gbps | 0.5 m - 40 km<br>SFP+, XENPAK (with SR, LR, and ER) | Summit X440<br>Summit X450-G2 (10G models)<br>Summit X460 (with XGM3-2sf, 2xsf),<br>Summit X460-G2 (1G models with VIM-2x, VIM-2t)<br>Summit X460-G2 (10G models)<br>Summit X480 (VIM2,3)<br>Summit X670 and X670V (ports 47 and 48),<br>Summit X670-G2<br>Summit X770 (ports 103,104) |
| SummitStack-V80 | 20 Gbps | 0.5 m - 100 m<br>QSFP+ only | Summit X460 (SSv80)<br>Summit X480 (VIM2,3)<br>Summit X670V (VIM4-40G4X)<br>Summit X670-G2-48x-4q (ports 57, 61) |

**Table 22: Summit Stacking by Stacking Method (continued)**

| Stacking Method | Speed per Link (HDX) | Cable Type and Lengths | Switch Models |
|---|---|---|---|
| SummitStack-V84 | 21 Gbps | 0.5 m - 5 m<br>QSFP+ passive copper | Summit X450-G2 (rear panel 21G stacking ports) |
| SummitStack-V160 | 40 Gbps | 0.5 m - 100 m<br>QSFP+ only | X460-G2 (VIM-2q)<br>Summit X480 (VIM3)<br>Summit X670 (VIM4)<br>Summit X670-G2-48x-4q (ports 57, 61)<br>Summit X770 (ports 103 and 104) |
| SummitStack-V320 | 80 Gbps | 0.5 m - 100 m<br>QSFP+ only | Summit X480 (VIM3)<br>Summit X670 (VIM4)<br>Summit X670-G2-48x-4q (ports 49, 53, 57, 61)<br>Summit X770-32q (ports 101 and 103, and 102 and 104) |
| SummitStack128 | 32 Gbps | 0.5 m, 1.5 m, 3.0m | Summit X480 (VIM2-SS128) |

For more details about the stacking methods that are available for each switch series, see the *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

Note

Because all switches in the stack must run the same version of ExtremeXOS, it is not possible to stack switches that require ExtremeXOS version 21, for example the X440-G2 and the X620, with switches that are incompatible with ExtremeXOS version 21, for example the X440 and the X460.

## SummitStack Terms

The following table describes the terms used for the SummitStack feature. These terms are listed in the recommended reading sequence.

**Table 23: List of Stacking Terms**

| Term | Description |
|---|---|
| Stackable switch | A Summit family switch that provides two stacking ports and can participate in a stack. |
| Stacking port | A physical interface of a stackable switch that is used to allow the connection of a stacking link. Stacking ports are point-to-point links that are dedicated for the purpose of forming a stack. |
| Native port | A stacking port that can be used only for connections between stacked switches, not for data connections. |
| Alternate port | A port that can be used for either stack connections or data connections. |

1 Combined over paired ports

2 The VIM2-SS128 module can be used for stacking X480 switches. It can also stack with SS256 with a conversion cable.

**Table 23: List of Stacking Terms (continued)**

| Term | Description |
|---|---|
| Stacking link | A cable that connects a stacking port of one stackable switch to a stacking port of another stackable switch, plus the stacking ports themselves. |
| Node | A switch that runs the ExtremeXOS operating system and is part of a stack. Synonymous with *stackable switch*. |
| Stack | A set of stackable switches and their connected stacking links made with the intentions that: (1) all switches are reachable through their common connections; (2) a single stackable switch can manage the entire stack; and (3) configurable entities such as VLANs and link trunk groups can have members on multiple stackable switches. A stack consists of all connected nodes regardless of the state of the nodes. |
| Stack topology | A contiguously connected set of nodes in a stack that are currently communicating with one another. All nodes that appear in the `show stacking` command display are present in the stack topology. |
| Stack path | A data path that is formed over the stacking links for the purpose of determining the set of nodes that are present in the stack topology and their locations in the stack. Every node is always present in a stack path whether or not stacking is enabled on the node. |
| Control path | A data path that is formed over the stacking links that is dedicated to carrying control traffic, such as commands to program hardware or software image data for software upgrade. A node must join the control path to fully operate in the stack. A node that is disabled for stacking does not join the control path, but does communicate over the stack path. |
| Active node | A node that has joined the control path. The active node can forward the control path messages or can process them. It can also forward data traffic. Only an active node can appear as a card inserted into a slot when the `show slot {`*`slot`* `{`**`detail`**`} \| `**`detail`**` }` command is executed on the master node of the stack. |
| Active topology | A contiguous set of active nodes in a stack topology plus the set of stacking links that connect them. When an active topology consists of more than one node, each node in the active topology is directly and physically connected to at least one other node in the active topology. Thus, the active topology is a set of physically contiguous active nodes within a stack topology. |
| Candidate node | A node that is a potential member of an active topology, or an active node that is already a member of an active topology. A candidate node may or may not be an active mode – that is, it may or may not have joined the control path. |
| Node role | The role that each active node plays in the stack – either master (or primary), backup, or standby. |
| Master node | The node that is elected as the master (or primary) node in the stack. The master node runs all of the configured control protocols such as OSPF, RIP, Spanning Tree, and EAPS.<br>The master node controls all of its own data ports as well as all data ports on the backup and standby nodes. To accomplish this, the master node issues specific programming commands over the control path to the backup and standby nodes. |

**Table 23: List of Stacking Terms (continued)**

| Term | Description |
|---|---|
| Backup node | The node assigned to take over the role of master if the master node fails. The master node keeps the backup node's databases synchronized with its own databases in preparation for such an event.<br>If and when the master node fails, the backup node becomes the master node and begins operating with the databases it has previously received. In this way, all other nodes in the stack can continue operating. |
| Standby node | A node that is prepared to become a backup node in the event that the backup node becomes the master node. When a backup node becomes a master node, the new master node synchronizes all of its databases to the new backup node.<br>When a node operates in a standby role, most databases are not synchronized – except those few that directly relate to hardware programming. |
| Acquired node | A standby or backup node that is acquired by a master node. This means that the master node has used its databases to program the hardware of the standby or backup node. The standby or backup node has acted as a hardware programming proxy, accepting the instructions of the master node to do so.<br>An acquired backup node maintains the databases needed to reflect why the hardware is programmed as it is. However, a standby node does not.<br>An acquired node can be re-acquired (without a reboot) by the backup node only when the backup node becomes the master node, and only when both the backup and standby nodes were already acquired by the same master node at the time of its failure. |
| Data ports | The set of ports on a stackable switch that are available for connection to your data networks. Such ports can be members of a user-configured *VLAN (Virtual LAN)* or trunk group. They can be used for Layer 2 and 3 forwarding of user data traffic, for mirroring, or other features you can configure. Data ports are different from stacking ports. |
| Failover | The process of changing the backup node to the master node when the original master node has failed.<br>When a master node fails, if a backup node is present, and if that node has completed its initial synchronization with the master node, then the backup node assumes the role of master node. The standby nodes continue their operation and their data ports do not fail. |
| Hitless failover | A failover in which all data ports in the stack, except those of the failing master node, continue normal operation when the master node fails. |
| Hitless upgrade | An operation in which the software image is upgraded, and the new image begins executing, without interrupting data traffic and without forcing any network reconvergence. This ExtremeXOS software version does not support hitless upgrade for a stack. |
| Node address | The unique MAC address that is factory-assigned to each node. |
| Node role election | The process that determines the role for each node. The election takes place during initial stack startup and elects one master node and one backup node. An election also takes place after a master node failover, when a new backup node is elected from the remaining standby nodes. |

**Table 23: List of Stacking Terms (continued)**

| Term | Description |
|---|---|
| Node role election priority | A priority assigned to each node, to bee used in node role election. The node with the highest node role election priority during a role election becomes the master node. The node with the second highest node role election priority becomes the backup. |
| Operational node | A node that has achieved operational state as a card in a slot. The operational state can be displayed using the `show slot {slot {detail} \| detail }` command. |
| System uptime | The amount of time that has passed since the last node role election. You can display the system uptime by entering the `show switch {detail }` command on the master node. |
| Stack segment | A collection of nodes that form a stack topology. The term is useful when a stack is severed. Each severed portion of the stack is referred to as a stack segment. |
| Stack state | A state assigned by the stack to a node. You can display the stack state by entering the `show stacking` command. |
| Easy Setup | A procedure that automatically configures the essential stacking parameters on every node for initial stack deployment, and then automatically reboots the stack to put the parameters into effect. The choice to run Easy Setup is offered when you run the `enable stacking {node-address node-address}` command and the essential stacking parameters are unconfigured or inconsistent. It can also be invoked directly by running the `configure stacking easy-setup` command. |

# Preparing to Configure a Stack

The following topics contain background information to help you configure your stack so that it functions as effectively as possible:



## About Stacking Node Roles, Redundancy, and Failover

ExtremeXOS supports control plane redundancy and hitless failover.

A stack supports control plane redundancy and hitless failover. Hitless failover is supported to the extent that the failing master node and all of its ports are operationally lost, including the loss of supplied power on any *PoE (Power over Ethernet)* ports that the node provided, but all other nodes and their provided ports continue to operate. After the failover, the backup node becomes the master node.

At failover time, a new backup node is selected from the remaining standby nodes that are configured to be master capable. All operational databases are then synchronized from the new master node to the

new backup node. Another hitless failover is possible only after the initial synchronization to the new backup node has completed. This can be seen using the `show switch {detail}` command on the master node and noting that the new backup node is In Sync.

When a backup node transitions to the master node role, it activates the Management IP interface that is common to the whole stack. If you have correctly configured an alternate management IP address, the IP address remains reachable.

When a standby node is acquired by a master node, the standby node learns the identity of its backup node. The master node synchronizes a minimal subset of its databases with the standby nodes.

When a standby node loses contact with both its acquiring master and backup nodes, it reboots.

A master node that detects the loss of an acquired standby node indicates that the slot the standby node occupied is now empty and flushes its dynamic databases of all information previously learned about the lost standby node.

A backup node restarts if the backup node has not completed its initial synchronization with the master node before the master node is lost. When a backup node transitions to the master node role and detects that the master node has not already synchronized a minimal subset of its databases with a standby node, the standby node is restarted.

### *Reboot or Failure of a Non-Master Node*

If a backup node fails, a standby node configured as master-capable is elected as the new backup. This new backup node is then synchronized to the databases of the master node.

For all non-master nodes, a node that reboots or is power-cycled loses all of its connections to all networks for the duration of the reboot cycle. Any *PoE* ports that were providing power prior to the event do not supply power.

When a non-master node fails, the master node marks the related slot as Empty. All other nodes exclude the failed node from the control path and any customer-configured VLANs, trunk group ports, mirroring ports, and so forth.

## Stack Configuration Parameters, Configuration Files, and Port Numbering

The stacking configurations are stored in the NVRAM of each node. Some of these configurations take effect only during the next node restart.

**Table 24: Stacking Configuration Items, Time of Effect and Default Value**

| Configuration Item | Takes Effect | Default Value |
|---|---|---|
| Stacking Mode | at boot time | Disabled |
| Slot Number | at boot time | 1 |
| Master-Capable | at boot time | Yes |
| License Restriction | at boot time | Not configured |
| Priority | at the next master election | Automatic |
| Alternate IP Address | immediately | Not configured |

**Table 24: Stacking Configuration Items, Time of Effect and Default Value (continued)**

| Configuration Item | Takes Effect | Default Value |
|---|---|---|
| Stack MAC | at boot time | Not configured |
| Stacking protocol | at boot time | Standard |

Note

Summit Series X770, X460-G2, X670-G2 and X450-G2 switches support the Enhanced Stacking protocol only. They do not support the Standard stacking protocol.

Stacking parameters, such as mode, slot number, etc., can be configured from a single unit in the stack topology. You can change the stacking-specific configuration even when a node is not in stacking mode but is connected to the stack. The target node for the configuration must be powered on and running a version of ExtremeXOS that supports stacking. Further, the node need not be in stacking mode and can be in any node role.

Most ExtremeXOS configuration parameters are not stored in NVRAM, but are instead stored in a configuration file. Configurations stored in NVRAM are those that are needed when the configuration file is not available. The configuration file chosen for the stack is the one selected on the master node that is first elected after a stack restart.

The data (non-stacking) port numbers, in the existing configuration files (which were created when not in stacking mode), are simple integer quantities. On a stack, the data port numbers are expressed as slot:port; where the slot is an integer representing the slot and port is an integer representing the port. For example, 1:2. The configuration file contains an indication that it was created on a stackable switch in stacking mode. The indication is the stacking platform ID.

Thus, when in stacking mode, the ports are referenced in the configuration file with the slot:port notation and when not in stacking mode, the ports are referenced as simple integers.

When the stack restarts, if a switch becomes the master and its selected configuration file was not created in stacking mode, the configuration file is de-selected, and the stack completes its restart using a default configuration. In addition, if the previously selected file was named with one of the default names (primary.cfg or secondary.cfg), the file is renamed to old_non_stack.cfg.

Similarly, if a switch is configured not to operate in stacking mode and the selected configuration file was created in stacking mode, the configuration file is de-selected, and the switch boots with a default configuration. In addition, if the file was named with one of the default names (primary.cfg or secondary.cfg), the file is renamed to old_non_stack.cfg.

The renamed file replaces any file that exists with the same name; the existing file is deleted.

## QoS in Stacking

Each stack uses *QoS (Quality of Service)* on the stacking links to prioritize the following traffic within the stack:

- Stack topology control packets
- ExtremeXOS control packets

- Data packets

For stack performance and reliability, the priority of control packets is elevated over that of data packets.

This is done to prevent control packet loss and avoid the timed retries that can lower performance. It is also done to prevent unneeded stack topology changes that can occur if enough stack topology information packets are lost. For these reasons, the SummitStack feature reserves one QoS profile to provide higher priority to control packets. The following sections describe the differences in QoS while using it in stack.

### *QoS Profile Restrictions*

In stacking mode, *CoS (Class of Service)* level 6 (hardware queue 6) is reserved for stacking, so you cannot create quality profile QP7.

Because QP7 cannot be created, you cannot use hardware queue 6 to assign CoS level 6 to a packet. However, you can assign packets received with 802.1p priority 6 to a *QoS* profile using the technique described in Processing of Packets Received With 802.1p Priority 6 on page 131.

Note

This restriction is applicable only when the stackable switch is operating in stacking mode.

### *QoS Scheduler Operation*

In stacking mode, the *QoS* scheduler operation is different for the stacking ports and the data ports.

The scheduler for the data ports operates the same as for standalone Summit family switches and is managed with the following command:

```
configure qosscheduler [default |strict-priority | weighted-round-robin
| weighteddeficit- round-robin]
```

The scheduler for the stacking ports is defined by the software when the stack is configured, and it cannot be modified. For all switches, the scheduler is set to strict-priority for the stacking ports, and meters are used to elevate the queue 6 priority above the priority of the other queues. This is the only scheduling method for stack ports.

### *Processing of Packets Received With 802.1p Priority 6*

By default, 802.1p examination is turned on.

Priority 7 is mapped to *QoS* profile QP8, and priorities 6 through 0 are mapped to QoS profile QP1. You can create other QoS profiles and can change this mapping as needed. Since you cannot create QP7 in stacking mode, 802.1p examination always maps packets with priority 6 to other QoS levels. However, you can use an *ACL (Access Control List)* rule entry to set the 802.1p egress value to 6 without affecting the QoS profile assignment as shown in this example:

```
entry VoIPinSummitStack { if { IP-TOS 46; } then { replace-dot1p-value 6; } }
```

### *Effects on 802.1p Examination*

You can turn off 802.1p examination.

When stacking is enabled, the examination remains turned on for priority 6. However, the examination happens at a lower precedence than that of all other traffic groupings.

The mapping you have configured for priority 6 remains in effect, and changes accordingly if you subsequently change the mapping.

When stacking is not enabled, all 802.1p examination is disabled when the feature is turned off.

*Effects on DiffServ Examination*

When DiffServ examination and 802.1p examination are both turned off, the 802.1p examination for packets arriving at 802.1p priority level 6 remains on at the lowered precedence.

In addition, the examination is adjusted to apply to all packets. The actual priority levels that are used for such packets are the defaults (QP1), or the values last configured using the following command:

```
configure dot1p type dot1p_priority {qosprofile} qosprofile
```

*Effects on Port QoS and VLAN QoS*

Port *QoS* and *VLAN* QoS have a higher precedence than the 802.1p priority examination performed when the 802.1p examination feature is turned off, and is therefore unaffected.

## Stacking Link Overcommitment

The stack is formed by each node supplying a pair of full-duplex, logical stacking ports. Each node can operate on a stack with full duplex throughput up to the limits in found in the *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

Even though two links are available, the links might not be fully utilized. For example, suppose there is a ring of eight nodes and the nodes are numbered clockwise from 1 to 8. The stacking port limit in this example is 10 Gbps in each direction for a total stack throughput of 20 Gbps for each port, or 40 Gbps total. Suppose node 1 wants to send 10 Gbps of unicast traffic to each of node 2 and node 3. The shortest path topology forces all traffic from node 1 over the link to node 2. Traffic from node 1 to node 3 passes through node 2. Thus, there is only 10 Gbps link available. However, if node 1 wanted to send 10 Gbps to node 2 and node 8, there would be 20 Gbps available because both links connected to node 1 would be used.

In a ring of eight nodes, between any two nodes (with one exception), only one link is used. If the devices provide 48 1-Gbps Ethernet ports, the overcommitment ratio between two such nodes is approximately 5:1. The exception is if there is an equal distance between the nodes. In this case, if both nodes are 48-port nodes, the nodes are grouped into two groups of 24 ports (by the hardware architecture), and thus it is possible to use both directions around the stack.

## Log Messages from Stack Nodes

Each node can generate log messages through the usual logging mechanism.

On backup and standby nodes, a log target and related filter is automatically installed. The log target is the master node. The filter allows all messages that have a log level of warning, error, or critical to be saved in the log file of the master node.

If the master node changes, the log target is updated on all the remaining nodes. You can also log in to any node in the active topology and see the complete log of the node.

## Configuring a Stack

Before configuring a new stack, do the following:

- Ensure that every switch, or node, in the stack is running on the same partition (primary or secondary).
- Ensure that every switch in the stack is running the same version of ExtremeXOS.
- Ensure that every master-capable switch in the stack is running the same version *and patch level* of ExtremeXOS.

Note
New switches are master-capable by default. To turn off master capability for a switch, use the following command: `configure stacking node-address address master-capability off`

Because stacks can consist of switches of different series and different models, ExtremeXOS does not restrict configuration settings based on the capabilities of any particular node in the stack. Therefore, you are responsible for ensuring that your configuration settings are appropriate for all switches in the stack.

Follow these steps to configure a new stack. Some of the steps include references where you can find additional information.

1. Physically connect the switches (stack nodes) using their stacking ports or alternate stacking ports.

   Instructions for physically setting up the stack are provided in *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

   Note
   To complete the cabling, you must first install any option cards you plan to use.

2. Power on the switches.
3. For each switch on which the stacking ports are not already enabled, issue the command `enable stacking-support`.

   Then reboot the switch.

   Note
   On new switches, stacking-support is disabled by default.

4. Configure all switches in the stack that will use the SummitStack-V, SummitStack-V80, SummitStack-V160, SummitStack-V320, or Multiprotocol Label Switching (MPLS) features.
   a. Configure switches that will use alternate stacking ports as described in #unique_306.
   b. Reboot the switches whose configurations you changed.

c. For each switch that will use the MPLS, SummitStack-V80, SummitStack-V160, or SummitStack-V320 features, issue the command `configure stacking protocol enhanced`.

   If the stack will use MPLS, it must contain only Summit X460-G2, X620, X670-G2, and X770 switches.

d. Reboot the switches whose configurations you changed.

5. Log in, using the console port, to the switch that will be the master.
6. Issue the command `show stacking stack-ports` to verify that the stacking ports are set up properly.

   If any ports display a state other than `Operational`, look for the following potential problems:

   - A `No-Neighbor` state might indicate a switch for which `configure stacking protocol enhanced` has not been set.
   - A `Link Down` state might indicate a connection problem. Check the physical connections for those ports. Also verify that both ports are using the same stacking technology, for example SummitStack-V80.

   The following example shows connection problems for the two ports on the switch in slot 2:

```
* switch1 > > show stacking stack-ports
Stack Topology is a Ring
Slot Port Select Node MAC Address  Port State  Flags Speed
---- ---- ------ ----------------- ----------- ----- -----
*1   1    Native 00:04:96:52:57:ab Operational C-      20G
*1   2    Native 00:04:96:52:57:ab Operational C-      20G
 2   1    Native 00:04:96:7e:00:6e Link Down   C-      20G
 2   2    Native 00:04:96:7e:00:6e No-Neighbor C-      20G
 3   1    Native 00:04:96:51:ea:18 Operational C-      20G
 3   2    Native 00:04:96:51:ea:18 Operational CB      20G
 4   1    Native 00:04:96:36:52:61 Operational CB      20G
 4   2    Native 00:04:96:36:52:61 Operational C-      20G
 5   1    Native 00:04:96:52:57:b8 Operational C-      20G
 5   2    Native 00:04:96:52:57:b8 Operational C-      20G
* - Indicates this node
Flags:  (C) Control path is active, (B) Port is Blocked
```

7. If you are using switches that were used previously in other stacks, issue the command `show stacking configuration`.

   All switches must have stacking-support enabled, but stacking disabled, before you run Easy Setup.

   Example:

```
* switch1 > show stacking configuration
Stack MAC in use: <unknown>
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
------------------ --- --- ---- ------------------ --------------- --------- ---
*00:04:96:52:57:ab 1   1   5    <none>             <none>          CcEe--iNn --
 00:04:96:7e:00:6e 2   2   1    <none>             <none>          --Ee--iNn --
 00:04:96:51:ea:18 3   3   2    <none>             <none>          --Ee--iNn --
 00:04:96:36:52:61 4   4   3    <none>             <none>          --Ee--iNn --
 00:04:96:52:57:b8 5   5   4    <none>             <none>          CcEe--iNn --
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
        (E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
        (M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
        (i) Stack MACs configured and in use are not the same or unknown,
        (N) Enhanced protocol is in use, (n) Enhanced protocol is configured,
        (-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
        (c) Core, (a) Advanced edge, or (e) Edge configured,
        (-) Not in use or not configured
```

   If a switch is enabled for stacking (shown by a capital letter `E` in the Flags column), issue the command `disable stacking node-address` *`mac_address`*. Then reboot the switch.

8. If necessary, configure a license level restriction.

   See Managing Licenses on a Stack on page 146.

9. From the node that will be the master, issue the command `enable stacking`.

   Enter `y` if you receive the following prompt:

   ```
   You have not yet configured all required stacking parameters. Would
   you like to perform an easy setup for stacking operation? (y/N)
   ```

   Entering `y` invokes Easy Setup. All of the switches reboot automatically and form a stack with a master node and a backup node. The rest of the switches in the new stack become standby nodes.

   Easy Setup configures all other required stacking parameters for every switch in the stack.

   **Note**

   To bypass Easy Setup (not recommended), respond `n` and follow the steps in Manually Configuring a Stack on page 136.

10. Verify the configuration.

    Follow the instructions in Verifying the Configuration on page 142.

11. Save the ExtremeXOS configuration to every active node in the stack.

    On the master node, issue the command `save configuration` *`config_name`*, where *`config_name`* is a descriptive name for this configuration.

    The stacking-specific configuration parameters are saved in a file called *`config_name`*`.cfg` to the NVRAM of each node.

The stack is ready to use.

## Manually Configuring a Stack

We recommend that you configure your stack using Easy Setup, as described in step 9 on page 135.

However, instead of running Easy Setup, you can configure the stack parameters manually. After performing step 1 on page 133 through step 8 on page 135, perform the following steps as needed:

1. Assign slot numbers to all switches in the stack.

   See Configuring Slot Numbers on page 136.
2. Configure node priorities on each slot.

   When the stack boots up, the node priority determines which node will be the master and which node will be the backup. Node priorities can be from 1 to 99, the lowest numbered slot having the highest priority.

   See Configuring the Master, Backup, and Standby Roles on page 137.
3. If any nodes are running a different ExtremeXOS version and patch level than the master, disable master capability for those nodes.

   See Configuring Master-Capability on page 155.
4. Assign a MAC address to the stack.

   See Assigning a MAC Address to a Stack on page 139.
5. Configure a failsafe account for the stack.

   See Failsafe Accounts on page 30.
6. Optionally, set a command prompt for the stack.

   Issue the command `configure snmp sysName` *`stack_name`*.

   If you do not define your own command prompt, the default command prompt looks similar to

   `* Slot-6 Stack.9 #`, where:

   - * indicates a changed and unsaved ExtremeXOS configuration
   - 9 is a sequence number indicating the 9th command to be entered since login
   - # indicates that your are logged into the master node (other nodes display the > symbol)
7. When you have performed all desired configuration steps, reboot the stack.
8. Verify the configuration.

   Follow the instructions in Verifying the Configuration on page 142.
9. Save the ExtremeXOS configuration to every active node in the stack.

   On the master node, issue the command `save configuration` *`config_name`*, where *`config_name`* is a descriptive name for this configuration.

   The stacking-specific configuration parameters are saved in a file called *`config_name`*`.cfg` to the NVRAM of each node.

### *Configuring Slot Numbers*

When you configure a stack manually, each node in the stack must be assigned a unique slot number. You can assign the slot number only through configuration; the stack does not dynamically assign a slot

number. The available slot numbers are 1 through 8. You can specify a slot number for each node manually, or you can have the system assign the slot numbers using a single command.

Note
Slot numbers take effect only after a restart. If you change a slot number, the unit continues to operate with the slot number it was last restarted with.

- To manually add a slot number to a node, use the following command:

  `configure stacking node-address` *`mac_address`* `slot-number` *`slot_number`*

  where *`mac_address`* is the node's MAC address.
- To configure the system to choose slot numbers for all nodes, enter the command:

  `configure stacking slot-number automatic`

  Automatic slot number assignment is performed in the order of appearance of the nodes in the `show stacking` display. In the case of a ring topology, the first node in the display is the intended master node into which you have logged in.
- Use the `show stacking` or `show stacking configuration` command to view the ordering and the assigned slot numbers.

Note
A node that boots in standalone mode does not use a slot number.

### *Configuring the Master, Backup, and Standby Roles*

Each stack has a master node, and it might have a backup node and multiple standby nodes.

The role of each stack node is determined by:

- The switch model number.
- The configured priority value.
- The configuration of the **`master-capability`** option (see Configuring Master-Capability on page 155).

Some switch models have greater CPU processing capability, more memory, and support additional features – thus making them more suitable for the role of master node.

To support the additional capabilities in a stack that includes multiple Summit switch models, the most capable switch automatically becomes the master node. For this release, the ranking of Summit switch models is as follows:

- X670-G2 (most capable)
- X460-G2
- X770
- X450-G2
- X460
- X440 (least capable)

If the stack configuration includes switches that are more capable than others, the stack will try to select the most-capable backup node.

If a switch with reduced capabilities serves as the backup node for a switch with greater capabilities, that switch might not be able to support the stack as a master node if a failover occurs (for example, the less-capable switch might not have enough processing power or table space to run efficiently as the master node). If your configuration needs to support automatic failover, we recommend that if a stack contains mixed model numbers, one of the following configurations should be used:

- Identical, most-capable switches available to become the master and backup nodes.
- The master-capability option is turned off for all less-capable switches.

When all master-capable nodes in a stack have the same model number, the node with the highest node role election priority becomes the master as a result of the first node role election, and the node with the second highest node role election priority becomes the backup node. All other nodes become standby nodes. See Node Election on page 54 for more information.

During subsequent node role elections that occur when a master node fails, the node priority configuration helps determine the node that becomes the replacement backup node.

Node priority configuration takes effect at the next node role election. A change in node priority configuration does not cause a new election. Once an active topology has elected a master node, that node retains the master node role until it fails or loses a dual master resolution.

You can configure one of the following election priority algorithms:

- Priority algorithm: If any node has a numeric priority value configured.
- Automatic algorithm: If all nodes participating in node role election have the automatic priority value configured.

The priority algorithm is selected if any node has a numeric priority value configured. You can specify an integer priority value between 1 and 100. The higher the value, the greater the node role election priority. If any node participating in a role election has a priority value configured, all nodes use the priority algorithm. A node configured with the automatic algorithm uses a priority value of zero (the lowest priority) in the priority algorithm if another node has a priority value configured.

The automatic algorithm is selected if no node participating in a role election has a numeric priority value configured. In automatic mode, the stack determines the highest role election priority based on factors such as available processing power, maintenance level of ExtremeXOS, and so forth.

In both algorithms, if the highest computed node role election priority is shared among multiple nodes, the slot number is used to adjust the node role election priority. A numerically lower slot number results in a higher role election priority than a numerically higher slot number. If you wish to use the slot number as the sole determining factor in node role election priority calculation, you should configure every node with the same priority value, and not automatic.

Note

The automatic priority algorithm may change in future ExtremeXOS releases.

Nodes that are configured as not master-capable do not participate in node role election. Priority configuration is not relevant on such nodes.

A dual master resolution does not use the configured node priority in most cases. Instead, it uses the oldest time that a node became a master in the current active topology.

### *Assigning a MAC Address to a Stack*

Each stack must use a single MAC address. When the master node fails over to the backup node, the backup node must continue to use the same MAC address that the master node was using.

Each stackable switch is assigned a single unique MAC address during production. By default, no stack MAC address is configured. You can choose any node to supply its factory assigned MAC address to form the stack MAC address.

**Note**
This task is not necessary when you configure the stack using Easy Setup. With Easy Setup, the MAC address is assigned by default.

When you assign a MAC address to a stack, one of the stackable switches is designated as the node whose factory-assigned MAC address is used to form the stack MAC address. Once this is done, all nodes receive and store this formed MAC address in their own NVRAM. Whenever the stack boots up, this MAC address is used, regardless of which node is the master.

When new nodes are added to the stack, the new nodes must be configured with the stack MAC address. The easiest way to do this is to use the `synchronize stacking {`**`node-address`** *`node_address`* | **`slot`** *`slot_number`*`}` command.

Before being stored as the stack MAC address, the chosen node's factory-assigned MAC address is converted to a locally administered MAC address. This prevents duplicate MAC address problems which lead to dual master conditions. The chosen MAC address is put into effect only at node boot time. If the address needs to be changed on a single node, rebooting that node results in usage of the same address stack-wide.

If you do not configure the stack MAC address or it is not the same on all nodes, a warning message appears in the log.

Each node operates with whatever address is available: the configured stack MAC address or the node's factory-assigned MAC address. If a master node fails over to the backup node, and the backup node's address is different than the one the former master node was using, the address is inconsistent with the addresses programmed into the packet forwarding hardware. The MAC address related to the management IP address changes to the one in use by the new master, but no gratuitous ARP requests are sent. In this case, it takes some time for hosts on the management network to flush the related ARP entry.

**Note**
If the node whose MAC address is chosen was removed from the stack with the intention of using the node elsewhere in the network, and that node is selected to supply the stack MAC in its new stack, the stack MAC of the original stack must be reconfigured to prevent a duplicate MAC address in the network.

To assign a MAC address to a stack, follow these steps.

1. Use the `show stacking configuration` command to display the stack MAC address configuration.

```
Slot-1 stack.3 # show stacking configuration
Stack MAC in use: 00:04:96:26:6a:f1
Node               Slot         Alternate          Alternate
```

```
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
------------------ --- --- ---- ------------------ --------------- --------- ---
*00:04:96:26:6a:f1 1   1   11   10.127.4.131/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:93 2   2   Auto 10.127.4.132/24    10.127.4.254    CcEeMm--- Aa
00:04:96:27:c8:c7 3   3   Auto 10.127.4.133/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:5f:4f 4   4   4    10.127.4.139/24    10.127.4.254    CcEeMm--- Aa
00:04:96:1f:a5:43 5   5   Auto 10.127.4.135/24    10.127.4.254    CcEeMm--- Aa
00:04:96:28:01:8f 6   6   6    10.127.4.136/24    10.127.4.254    CcEeMm--- Aa
00:04:96:20:b2:5c 7   7   Auto 10.127.4.137/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:92 8   8   Auto 10.127.4.138/24    10.127.4.254    CcEeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

The MAC Address column displays the factory MAC address for the node. The stack MAC address configuration information appears in the last three positions of the Flags column. As shown in the key at the bottom of the command display, the stack MAC configuration is displayed with the letters capital M, lower-case m, and lower-case i. If the flags read ---, the stack MAC address needs to be configured. If the flags read Mm-, the stack MAC address is already configured and in use.

2. To configure the stack to use the MAC address of the master, log in to the master console and enter the configure stacking mac-address command.

   For example:

   ```
   Slot-1 stack.43 # configure stacking mac-address
   This command will take effect at the next reboot of the specified node(s).
   ```

   If you enter the `show stacking` command now, the stack MAC flags show --i, indicating that the stack MAC is configured but is not in use. After you restart the stack, the i disappears from the Flags column.

   a. To see if the stack MAC is consistently configured, enter the `show stacking` {**node-address** *node_address* | **slot** *slot_number*} **detail** command and compare all configured stack MAC addresses for equality. In this case, they should be equal.

3. To configure the stack to use a MAC address from a non-master node, log in to the master console and enter the `configure stacking` {**node-address** *node-address* | **slot** *slot-number*} **mac-address** command. For example:

   ```
   Slot-1 stack.43 # configure stacking slot 2 mac-address
   This command will take effect at the next reboot of the specified node(s).
   ```

4. Reboot the stack.
5. Verify the new stack MAC address using the `show stacking configuration` command.

   The following example is based on the previous example:

   ```
   Slot-1 stack.3 # show stacking configuration
   Stack MAC in use: 00:04:96:26:6a:f1
   Node               Slot         Alternate          Alternate
   MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
   ------------------ --- --- ---- ------------------ --------------- --------- ---
   *00:04:96:26:6a:f1 1   1   11   10.127.4.131/24    10.127.4.254    CcEeMm--- Aa
   00:04:96:26:6c:93 2   2   Auto 10.127.4.132/24    10.127.4.254    CcEeMm--- Aa
   00:04:96:27:c8:c7 3   3   Auto 10.127.4.133/24    10.127.4.254    CcEeMm--- Aa
   ```

```
00:04:96:26:5f:4f 4   4   4    10.127.4.139/24    10.127.4.254    CcEeMm--- Aa
00:04:96:1f:a5:43 5   5   Auto 10.127.4.135/24    10.127.4.254    CcEeMm--- Aa
00:04:96:28:01:8f 6   6   6    10.127.4.136/24    10.127.4.254    CcEeMm--- Aa
00:04:96:20:b2:5c 7   7   Auto 10.127.4.137/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:92 8   8   Auto 10.127.4.138/24    10.127.4.254    CcEeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

### *Configuring Stacking Port Operation with the VIM4-40G4X Option Card*

Configure port partitions, selecting between the native and alternate stack ports. When the VIM4-40G4X option card is installed, you can use the following ports for stacking:

- The native stacking ports on the VIM4-40G4X option card, which can be configured for an 80 Gbps, 160 Gbps, or 320 Gbps data rate. These ports can also be partitioned to operate as one 40 Gbps data port or four 10 Gbps data ports.
- The alternate stacking ports, which are listed in #unique_315/unique_315_Connect_42_TABLE_IP1_SNM_3W.

The stacking rate of 320Gbps can be used across a stack of X670 switches (equipped with VIM4-40G4X) or X480 switches (equipped with VIM3-40G4X). This solution uses two trunked 40G ports for 80 Gbps per stack port.

320G Stack port 1 is formed by trunking VIM4 ports S1 and S3. Similarly, 320G stack port 2 is formed by trunking VIM4 ports S2 and S4. Figure 10 shows VIM4 trunk connection in case of 320G stacking.

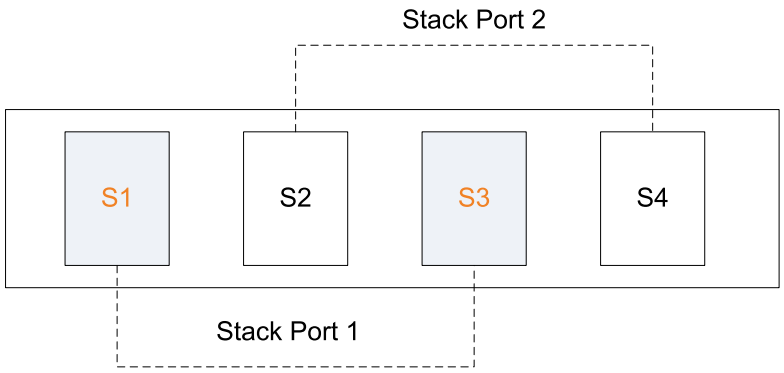



**Figure 10: VIM4 Connection for 320G Stacking**

- To select between the native and alternate stack ports, use the following command:

  `configure stacking-support` **`stack-port`** `[`*`stack-ports`* `|` **`all`**`]` **`selection`** `[`**`native`** `{`**`V80`** `|` **`V160`**`}` `|` **`V320`**`}` `|` **`alternate`**`]`
- To configure the port partition, use the following command:

  `configure ports [`*`port_list`* `|` **`all`**`]` **`partition`** `[`**`4x10G`** `|` **`1x40G`**`]`

After a configuration change, you must restart the switch to use the stacking ports.

## Verifying the Configuration

To verify that your stack is configured as you intended, issue any or all of the commands described here.

The `show slot` and `show stacking` commands contain stacking configuration information, including the state of the slot. These commands are also helpful when debugging stacking problems.

The `show slot` command shows the states of the nodes as they move from the empty to operational state.

- Use the `show slot` command and the following table to determine a slot state:

```
Slot-1 Stack.25 # show slot
Slots    Type                 Configured           State       Ports
------------------------------------------------------------------------
Slot-1   SummitX              SummitX              Operational   26
Slot-2   SummitX              SummitX              Operational   26
Slot-3   SummitX              SummitX              Operational   26
Slot-4   SummitX              SummitX              Operational   50
Slot-5   SummitX              SummitX              Operational   26
Slot-6   SummitX              SummitX              Operational   26
Slot-7   SummitX              SummitX              Operational   50
Slot-8   SummitX              SummitX              Operational   26
Slot-1 Stack.26 #
```

```
* Slot-1 Stack.1 # show stacking
Stack Topology is a Ring
Active Topology is a Ring
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
*00:04:96:26:60:DD  1     Active       Master   CA-
00:04:96:26:60:EE  2     Active       Backup   CA-
00:04:96:26:60:FF  3     Active       Standby  CA-
00:04:96:26:60:AA  4     Active       Standby  CA-
00:04:96:26:60:88  5     Active       Standby  CA-
00:04:96:26:60:99  6     Active       Standby  CA-
00:04:96:26:60:BB  7     Active       Standby  CA-
00:04:96:26:60:CC  8     Active       Standby  CA-
* - Indicates this node
Flags:  (C) Candidate for this active topology, (A) Active Node
(O) node may be in Other active topology
```

The asterisk (*) that precedes the node MAC address indicates the node to which you are logged in. The node MAC address is the address that is factory assigned to the stackable switch.

The slot number shown is the number currently in use by the related node. Because slot number configuration only takes effect during node initialization, a change in configured value alone does not cause a change to the slot number in use.

If a node role has not yet been determined, the node role indicates <none>. In a ring topology, the node on which this command is executed is always the first node displayed. In a daisy chain, the ends of the daisy chain are the first and last nodes displayed.

Even though the stack topology could be a ring, the active topology could be a daisy chain because it does not contain every node in the stack topology.

If the node on which this command is being executed is not active, the stacking topology is replaced with a line similar to this one:

```
This node is not in an Active Topology.
```

Note

It is possible for a node to be in stabilizing or waiting state and still be in the active topology.

- Use the `show stacking configuration` command to get a summary of the stacking configuration for all nodes in the stack:

```
Slot-1 Stack.2 # show stacking configuration
Stack MAC in use: 02:04:96:26:6b:ed
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------- --------- ---
*00:04:96:26:6b:ed 1   1   Auto <none>             <none>           CcEeMm--- --
00:04:96:34:d0:b8 2   2   Auto <none>            <none>          CcEeMm--- --
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

- Use the `show stacking {`**`node-address`** *`node_address`* | **`slot`** *`slot_number`*`}` **`detail`** command to get a full report from the stacking database:

```
Slot-1 Stack.33 # show stacking slot 1 detail
Stacking Node 00:04:96:26:60:DD information:
Current:
Stacking                  : Enabled
Role                      : Master
Priority                  : 2
Slot number               : 1
Stack state               : Active
Master capable            : Yes
Stacking protocol         : Enhanced
License level restriction : Advanced edge
In active topology?       : Yes
Factory MAC address       : 00:04:96:26:60:DD
Stack MAC address         : 02:04:96:26:60:DD
Alternate IP address      : 192.168.130.101/24
Alternate gateway         : 192.168.130.1
Stack port 1:
State                  : Operational
Blocked?               : No
Control path active?   : Yes
Stack port 2:
State                  : Operational
Blocked?               : No
Control path active?   : Yes
Configured:
Stacking                  : Enabled
Master capable            : Yes
```

```
Stacking protocol          : Enhanced
Slot number                : 1
Stack MAC address          : 02:04:96:26:60:DD
License level restriction : Edge
```

If you do not specify any node, the output is generated for all nodes in the stack topology. If the specified node does not exist, an error message appears.

The *slot* parameter is available only in stacking mode. The node-address parameter is always available.

Current information represents stacking states and configured values that are currently in effect. Configured information is that which takes effect at node reboot only.

The roles values are Master, Backup, Standby, and *none*. License level restrictions are Edge, Advanced Edge, or Core.

- To verify the stacking port states of each node in the stack topology use the command `show stacking stack-ports`.

# Managing an Operational Stack

The following topics describe common tasks for logging into an operational stack and managing it.

## Logging into a Stack

You can log into any node in a stack, but you can control more stack features when you log into the master. The following guidelines describe the options available to you when you log into different nodes:

- On master nodes, all features supported by the switch license operate correctly.
- On backup nodes, most show commands show correct data for the active stack. For example, `show vlan {virtual-router vr-name}` shows all configured VLANs.
- On all non-master nodes, most of the configuration commands are rejected. However, the failsafe account, enable license, and stacking configuration commands work on any node.
- On standby nodes, most show commands do not show correct data for the current stack operation. However, the `show switch {detail}`, `show licenses`, and all `show stacking` commands show correct data.
- If a node is connected to the stack and stacking is not enabled, you can still configure stacking features on that node.

The login security that is configured on the master node applies when logging into any node in the active topology. This includes any active node that is present in a slot. A node that is disabled for stacking is its own master, and uses its own security configuration.

You can log in to a SummitStack node using the following methods:

- Console connection to any node
- Management connection to the master
- Management connection to a standby node
- Telnet session over the stack from any active node to any other node in the same active topology

### *Logging in through the Console Port*

You can use the console port on any switch to manage the stack.

If you connect to the master node, you can configure and manage the stack. If you connect to a non-master node, you can view node status and configure only a few options from the node to which you are connected. However, you can use the Telnet feature to connect to another node and manage that node as if you were connected to it (see Logging Into a Node from Another Node on page 145).

### *Logging in from the Management Network*

The management network is an Ethernet network to which the management port of each switch connects.

The primary management IP address is assigned to the master node. You can use a terminal emulation program and this IP address to connect to the master for configuration and management.

The alternate management IP addresses allow you to connect to individual nodes from your management network. During normal operation, you connect to the stack using the primary management IP address. However, if the stack is split, you can use the alternate management IP address to connect to the other half of the stack. For more information, see Configuring an Alternate IP Address and Gateway on page 150.

After you log in to a master or standby node through the management network, you can Telnet to any other node and control that node as if you were directly connected to it. For more information, see Logging Into a Node from Another Node on page 145.

### *Logging Into a Node from Another Node*

You may log into any node in the active topology from any other node in the same active topology. If you do not know the slot number of the node to which you want to connect, enter the `show slot` command. You can Telnet to any switch that appears in the `show slot` command display.

Note

If the node to which you want to connect does not appear in the `show slot {`*`slot`* `{`**`detail`**`} | `**`detail`**` }` command display, you can connect to the node through the its console port or management port.

You have the most control over the stack when you log in to the master.

1. Determine which node is the master using the command `show stacking`.
2. Telnet to another node using the command `telnet slot` *`slot-number`* .
3. When prompted, log in normally.

   The switches must be active in the stack for this command to function.

   The `telnet slot` *`slot-number`* command accepts a slot number in stacking mode. When the telnet program accepts a connection from another node in the stack, it performs security validation. The master node validates all login security information (except for the failsafe account), regardless of the node into which you are attempting to establish a login. If you are not able to log in using your user credentials, use the failsafe account to log in.

## Managing Licenses on a Stack

The SummitStack feature is not licensed separately. You can use the SummitStack feature with an Edge license.

The rules for licensing are as follows:

- The effective license level of all master-capable nodes must be the same.
- If you set a license level for the stack, then each node must be at that license level or higher.
- If you do not set a license level for the stack, the license level the stack uses is the effective license level of the node that is elected master at startup.

**Note**

If the stack is using the Advanced Edge license and you attempt to add a master-capable node that is using an Edge license, the node does not become operational. In response to the `show slot {slot {detail} | detail }` command, the node displays as Failed with a License Mismatch reason.

- Master-capable nodes continuously monitor the effective license level of the master node.
- Nodes with higher license levels than other nodes can be restricted to operate at a lower or effective license level.

### *Viewing Switch Licenses and License Restrictions*

1. To view the current license information for a node, log into that node and enter the `show licenses` command.

   The command display is similar to the following:

```
Slot-1 Stack.1 # show licenses
Enabled License Level:
Advanced Edge
Enabled Feature Packs:
None
Effective License Level:
Advanced Edge
Slot-1 Stack.2 #
```

   The Enabled License Level is the purchased license level. This is the maximum level at which this node can operate without purchasing a license level upgrade.

   The Effective License Level is the operating license level. If a license level restriction is configured for this node, the effective license level may be lower than the enabled license level. All master-capable switches must be operated at the same effective license level.

2. On the master node, run `show stacking configuration` to view the license level restrictions configured for all nodes in a stack.

```
Slot-1 Stack.33 # show stacking configuration
Stack MAC in use: 02:04:96:26:60:DD
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------- --------- ---
*00:04:96:26:60:DD 1   1   Auto 192.168.130.101/24 192.168.130.1   CcEeMm--- Aa
00:04:96:26:60:EE 2   2   Auto 192.168.130.102/24 192.168.130.1   CcEeMm--- Aa
00:04:96:26:60:FF 3   3   Auto 192.168.130.103/24 192.168.130.1   --EeMm--- Aa
00:04:96:26:60:AA 4   4   Auto 192.168.130.104/24 192.168.130.1   --EeMm--- Aa
00:04:96:26:60:88 5   5   Auto 192.168.130.105/24 192.168.130.1   --EeMm--- Aa
00:04:96:26:60:99 6   6   Auto 192.168.130.106/24 192.168.130.1   --EeMm--- Aa
00:04:96:26:60:BB 7   7   Auto 192.168.130.107/24 192.168.130.1   --EeMm--- Aa
```

```
00:04:96:26:60:CC 8   8   Auto 192.168.130.108/24 192.168.130.1   --EeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

License level restrictions appear in the Lic column. The license level restriction in use appears first, represented by a capital letter as shown in the display legend. The configured license level restriction appears second, represented by a lower-case letter. When the letters in the Lic column are different, for example Ae, the node is configured with a different license level restriction than the one that is currently in use.

To put the configured license level restriction into effect, you must reboot the node.

### *Enabling a Switch License*

All nodes must have a purchased license level at least equal to the license level of the master node in order to become operational in the stack. The purchased license level of a node can be enabled only after you log in to that node (see Logging Into a Node from Another Node on page 145).

For instructions on enabling a license on a node, see Software Upgrade and Boot Options on page 1522.

### *Restricting a Switch License Level*

If the master-capable nodes in a stack have different license levels and you want to operate a stack at the minimum license level, you can apply a license level restriction. The restriction is stored in the NVRAM of each master-capable node. It forces the node to reduce its license level below its purchased level at node restart time for the life of the restart. This reduced license level is called the effective license level and can be displayed by entering the `show licenses` command on the node you want to evaluate.

To restrict a master-capable node to operate at a license level that is lower than the one purchased for the node, use the command:

`configure stacking {`**`node-address`** *`node-address`* | **`slot`** *`slot-number`*`}`
**`license-level`** `[`**`core`** | **`advanced-edge`** | **`edge`**`]`.

In the following example, node 7 is restricted to operate at the Edge license level:

```
* switch # configure stacking slot 7 license-level edge
This command will take effect at the next reboot of the specified node(s).
```

You must reboot the master-capable node for the command to take effect.

The command restricts the specified node to operate at the specified license level. The specified license level must match the effective license level of all master-capable nodes in the stack. To avoid stack reboots when future license level upgrades are purchased, during initial deployment you should purchase the same license level for every master-capable node in the stack, and the license level restriction should not be configured.

### *Upgrading Stack Licenses*

You can purchase license level upgrades for Summit family switches. All master-capable switches in a stack must run the same license level. If the license you want to run is not available for a specific Summit switch, you cannot use that switch and that license level as a master-capable switch. For example, if you want to upgrade to the core license, your master-capable nodes must be Summit family switches that support the core license.

**Note**

For information on which switches support which licenses, see the *Feature License Requirements* document. That document also lists which switches support the SummitStack feature.

To upgrade the licenses for the switches in a stack, follow these steps.

1. Log in to the master node.
2. Enter the `show stacking` command and note the role (master, backup, or standby) of each node in the stack.
3. Enter the `show stacking configuration` command and note any nodes that are configured with a license level restriction.

   See Viewing Switch Licenses and License Restrictions on page 146 for more information.
4. Install the required license level in each master-capable node (backup and standby nodes) by logging into each node (`telnet slot` *`slot-number`* ) and entering the command:

   `enable license {`**`software`**`}` *`key`*
5. Enter the license key given to you by Extreme Networks when you purchased the upgrade.
6. Use the commands in Step 4 to install the required license level on the master node.
7. If any nodes are configured with a license level restriction that is lower than the intended operating license level of the stack, log into the master node and remove the stack license level restriction using the command:

   `unconfigure stacking license-level`

   This command removes the restriction on all nodes.
8. If you removed a license level restriction, reboot the stack to put the license level restriction removal into effect using the command:

   `reboot {[`**`time`** *`mon day year hour min sec`*`] | `**`cancel`**`} {`**`slot`** *`slot-number`* `| `**`node-address`** *`node-address`* `| `**`stack-topology`** `{`**`as-standby`**`}}`
9. Verify that all master-capable nodes are operating at the intended license level.

   On the master node, run `show licenses` and `show slot {`*`slot`* `{`**`detail`**`} | `**`detail`** `}`.

   If no slot shows as Failed, then all master-capable nodes are operating at the effective license level shown for the master node.

## Upgrading ExtremeXOS on a Stack

The following topics describe how to upgrade the ExtremeXOS software and the bootrom on the switches in a stack.

*Upgrading the Software on all Active Nodes*

You can centrally upgrade the software on all active nodes in a stack. To upgrade all nodes in the stack, all nodes must be running an ExtremeXOS release that supports stacking (ExtremeXOS release 12.0 or later).

1. Download a new ExtremeXOS software release and install it on all nodes on the active topology using the command: `download image [[hostname | ipaddress] filename {{vr} vrname}] {partition}`

   Note

   ExtremeXOS offers the ability to synchronize a non-master slot from the master individually, but does not support the upgrade of the entire stack with one image download to the master.

2. If necessary, use the `use image {partition} {primary | secondary}` command to select the image partition (primary or secondary) into which the software was saved.
3. Restart all nodes in the new release using `reboot {[time mon day year hour min sec] | cancel}`

   For example:

   ```
   download image [[hostname | ipaddress] filename {{vr} vrname}] {primary | secondary}
   use image {partition} [primary | secondary]
   reboot
   ```

4. Before you upgrade a stack, make sure that the active image partition is same across all nodes.

   To determine the active partition selected on all nodes and the ExtremeXOS versions installed in each partition, use the `show slot detail` command. You can install the image only on the alternate image partition and not on the active image partition.
5. To run the upgraded software, you must reboot the stack after installation with the image partition that received the software being selected.
6. If the active partition is different on some nodes, the action you take depends on what is stored in both partitions:

   If both primary and secondary partitions have the same ExtremeXOS release, you may use the following commands to cause a node to use the same active image as the rest of the stack:

   ```
   use image {primary | secondary} slot slot-number
   reboot slot slot-number
   ```

   If you are using the primary image on your master node and some other node primary image does not contain the same ExtremeXOS version as your master node's primary image, you may use the command: `synchronize slot slotid` to cause the node to contain the same ExtremeXOS versions in both partitions as it is on the master node, and to reboot the node into the primary partition.

   Hitless upgrade is not supported in a stack.

*Upgrading the Software on a Single Node*

To upgrade the software on a single active node:

1. Enter the following commands to download an image to a node:

   ```
   download image [[hostname | ipaddress] filename {{vr} vrname}]
   {primary | secondary} slot slot number

   use image {partition} [primary | secondary] slot slotid

   reboot slot slot number
   ```

   The slot number is the one in use by the active node that is to be upgraded. Be sure that you keep the same image versions on all the other nodes as you have on the master node.

   Alternatively, if your master node has the same image versions in its partitions that you want installed in the node to be upgraded, you can use the command `synchronize {slot slotid}` to upgrade both images and select the desired image.

2. You can upgrade the image on an active node even if the node shows as Failed when using the `show slot` command.

*Upgrading the Bootrom*

The SummitStack feature does not require a bootrom upgrade. You should not upgrade the bootrom of any node unless there are specific reasons to do so. However, the SummitStack feature does allow centralized bootrom upgrade.

You can download and install the bootrom to a specific slot using the slot parameter. The slot parameter is available only on stackable switches in the active stack topology. For information on upgrading the bootrom, see Software Upgrade and Boot Options on page 1522.

If you do not provide a slot number, the stack attempts to download the bootrom image and install it on all stackable switches in the active topology.

## Configuring an Alternate IP Address and Gateway

The stack has a primary IP address and subnetwork. These values are assigned using the `configure vlan mgmt ipaddress` command. There can also be static or default routes associated with the stack.

For each node in the stack, you can configure an alternate management IP address, subnetwork mask, and default gateway. The alternate IP address is restricted to being a member of the primary IP subnetwork that is configured on the management *VLAN*, and thus the alternate IP subnetwork must exactly match the primary IP management subnetwork. A subnetwork match is exact if the subnetwork portion of the IP addresses match exactly. For example, 10.11.12.1/24 and 10.11.12.2/24 are an exact subnetwork match (because both represent the subnet 10.11.12.0/24).

Standby nodes always install their configured alternate management IP address and gateway on the management interface. A standby node does not have the ability to verify whether the configured alternate IP address matches the primary management IP subnetwork of the stack.

The backup and master nodes have the ability to verify the configured alternate IP address. The master and backup nodes compare the primary IP subnetwork information to the alternate IP subnetwork. If there is a match, the backup node installs the primary IP management subnetwork's default routes and installs only the alternate management IP address (not the primary IP address). The master node installs both the configured management subnetwork with specific IP address and the alternate IP address. In this case, the alternate gateway is not used, expecting that primary management routes are configured or will be configured. In either case, if the alternate IP subnetwork does not match the configured management subnetwork, the alternate IP address is not installed on the management interface.

Each node in the stack normally installs its alternate IP address on the management subnetwork. When an ARP request for the alternate IP address is satisfied, the stackable switch supplies its factory-assigned MAC address and not the stack MAC address. Only the master node installs the primary IP address. An ARP request for the configured management IP address returns the configured stacking MAC address. Because of the above behavior, all nodes are reachable over their management ports even during a dual master condition. The VLAN used is the management VLAN (VID 4095) and is untagged.

The alternate gateway is only installed on a master or backup node when the primary management IP subnetwork is not configured. Once the primary IP subnetwork is installed, the alternate gateway is removed. The alternate gateway is always installed on a standby node.

If a dual master condition occurs because a stack has been severed, the alternate IP addresses and associated MAC addresses are unique, and it is possible to use telnet or ssh to reach any node. Any node on the segment with the incorrect master can then be used to reboot the entire stack segment into standby mode if you want to rejoin the stack segments later.

If a node is operating in stacking mode, the alternate management IP address configuration takes effect immediately.

Note

Only IPv4 alternate management IP addresses are supported in this release.

To configure an alternate IP address and gateway, follow these steps.

1. View the alternate IP address configuration using the `show stacking configuration` command:

```
Slot-1 stacK.13 # show stacking configuration
Stack MAC in use: 00:04:96:26:6a:f1
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
------------------ --- --- ---- ------------------ --------------- --------- ---
*00:04:96:26:6a:f1 1   1   11   <none>             <none>          CcEeMm--- Aa
00:04:96:26:6c:93 2   2   Auto <none>             <none>          CcEeMm--- Aa
00:04:96:27:c8:c7 3   3   Auto <none>             <none>          CcEeMm--- Aa
00:04:96:26:5f:4f 4   4   4    <none>             <none>          CcEeMm--- Aa
00:04:96:1f:a5:43 5   5   Auto <none>             <none>          CcEeMm--- Aa
00:04:96:28:01:8f 6   6   6    <none>             <none>          CcEeMm--- Aa
00:04:96:20:b2:5c 7   7   Auto <none>             <none>          CcEeMm--- Aa
00:04:96:26:6c:92 8   8   Auto <none>             <none>          CcEeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
```

```
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

In the example above, no alternate IP address or alternate gateway is configured.

2. If you have a continuous block of IP addresses to assign to the stack, enter the `configure stacking alternate-ip-address [`*ipaddress netmask* | *ipNetmask*`] *gateway* **automatic**` command.

   For example:

```
Slot-1 Stack.14 # configure stacking alternate-ip-address 10.127.4.131/24 10.127.4.254
automatic
Slot-1 Stack.15 # show stacking configuration
Stack MAC in use: 00:04:96:26:6a:f1
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
------------------ --- --- ---- ------------------ --------------- --------- ---
*00:04:96:26:6a:f1 1   1   11   10.127.4.131/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:93 2   2   Auto 10.127.4.132/24    10.127.4.254    CcEeMm--- Aa
00:04:96:27:c8:c7 3   3   Auto 10.127.4.133/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:5f:4f 4   4   4    10.127.4.134/24    10.127.4.254    CcEeMm--- Aa
00:04:96:1f:a5:43 5   5   Auto 10.127.4.135/24    10.127.4.254    CcEeMm--- Aa
00:04:96:28:01:8f 6   6   6    10.127.4.136/24    10.127.4.254    CcEeMm--- Aa
00:04:96:20:b2:5c 7   7   Auto 10.127.4.137/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:92 8   8   Auto 10.127.4.138/24    10.127.4.254    CcEeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

3. If you do not have a continuous block of IP addresses for the stack, assign an alternate IP address and gateway to each node using the `configure stacking [`**node-address** *node-address* | **slot** *slot_number*`] **alternate-ip-address** [`*ipaddress netmask* | *ipNetmask*`] *gateway*` command.

   For example:

```
Slot-1 Stack.18 # configure stacking slot 4 alternate-ip-address 10.127.4.139/24
10.127.4.254
```

**Note**

If you try to assign an alternate IP address and gateway to a node that is already configured with these parameters, an error message appears. To remove an existing configuration so you can change the alternate IP address and gateway, enter the `unconfigure stacking {`**node-address** *node_address* | **slot** *slot_number*`} **alternate-ip-address**` command.

4. Enter the `show stacking configuration` command to verify that the alternate IP address and gateway is configured as intended for each node.

```
Slot-1 Stack.19 # show stacking configuration
Stack MAC in use: 00:04:96:26:6a:f1
```

```
Node              Slot         Alternate          Alternate
MAC Address       Cfg Cur Prio Mgmt IP / Mask     Gateway         Flags     Lic
----------------- --- --- ---- ------------------ --------------- --------- ---
*00:04:96:26:6a:f1 1   1   11   10.127.4.131/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:93 2   2   Auto 10.127.4.132/24    10.127.4.254    CcEeMm--- Aa
00:04:96:27:c8:c7 3   3   Auto 10.127.4.133/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:5f:4f 4   4   4    10.127.4.139/24    10.127.4.254    CcEeMm--- Aa
00:04:96:1f:a5:43 5   5   Auto 10.127.4.135/24    10.127.4.254    CcEeMm--- Aa
00:04:96:28:01:8f 6   6   6    10.127.4.136/24    10.127.4.254    CcEeMm--- Aa
00:04:96:20:b2:5c 7   7   Auto 10.127.4.137/24    10.127.4.254    CcEeMm--- Aa
00:04:96:26:6c:92 8   8   Auto 10.127.4.138/24    10.127.4.254    CcEeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
```

## Viewing the Alternate IP Address

To view the alternate IP address for a node, run `show vlan mgmt` or `show ipconfig mgmt`.

### *show vlan mgmt Command*

The `show vlan mgmt` command shows the alternate management IP address as applied to the management *VLAN* on the local unit. This allows you to see how the configured alternate management IP address has been applied.

The `show vlan mgmt` command displays the following information:

```
Slot-1 Stack.1 # show vlan "Mgmt"
VLAN Interface with name Mgmt created by user
    Admin State:         Enabled     Tagging:   802.1Q Tag 4095
    Description:         Management VLAN
    Virtual router:      VR-Mgmt
    IPv4 Forwarding:     Disabled
    IPv4 MC Forwarding:  Disabled
    IPv6 Forwarding:     Disabled
    IPv6 MC Forwarding:  Disabled
    IPv6:                None
    STPD:                None
    Protocol:            Match all unfiltered protocols
    Loopback:            Disabled
    NetLogin:            Disabled
    OpenFlow:            Disabled
    TRILL:               Disabled
    QosProfile:          None configured
    Flood Rate Limit QosProfile:       None configured
    Ports:   1.           (Number of active ports=1)
     Untag: Mgmt-port on Mgmt-1 is active
```

For the management VLAN, a secondary address cannot be configured and so the Secondary IP line does not appear.

The Alternate IP line shows one of the following:

- The configured alternate management IP address if it has been activated.

- <none> if it has not been configured.
- Mismatch if it has been configured but does not exactly match the Primary IP subnet.

*show ipconfig mgmt Command*

The `show ipconfig mgmt` command shows the configured alternate management IP address as applied to the management *VLAN* on the local unit. This allows you to see how the configured alternate management IP address has been applied.

The Multinetted VLAN indication always appears as NO. The alternate IP address is restricted to the same subnet as the primary subnet configured for the management IP interface. As a result, only a single subnet is used with the possibility of multiple station addresses. Further, you cannot configure a secondary IP address on the management VLAN.

The `show ip config mgmt` command displays the following information:

```
Slot1 Stack.36 # show ipconfig MgmtRouter Interface on VLAN Mgmt is enabled and up.
inet 10.66.4.74/24  broadcast 10.66.4.255  Mtu 1500
Alternate IP Address: 10.66.4.75/24
Flags:
AddrMaskRly  NO     BOOTP Host  NO   DirBcstHwFwd  NO     Fwd Bcast  NO
IgnoreBcast  NO     IP Fwding  NO    IPmc Fwd  NO         Multinetted VLAN  NO
IRDP Advert  NO     SendParam YES    SendPortUn YES       Send Redir YES
SendTimxceed YES    SendUnreach YES  TimeStampRly  NO     VRRP  NO
```

For the management VLAN, a secondary address cannot be configured and so the Secondary IP line does not appear.

The Alternate IP Address line shows one of the following:

- The configured alternate management IP address if it has been activated.
- <none> if it has not been configured.
- Mismatch if it has been configured but does not exactly match the Primary IP subnet.

## Viewing Stacking Port Statistics

To view the status of any stacking port, run any of the following variations:

`show ports stack-ports` *`stacking-port-list`* `utilization {`**`bandwidth`** `|` **`bytes`** `|` **`packets`**`}`
`show ports {`**`stack-ports`** *`stacking-port-list`* `|` *`port_list`*`} statistics {`**`norefresh`**`}`
`show ports {`*`port_list`* `|` **`stack-ports`** *`stacking-port-list`*`} rxerrors {`**`norefresh`**`}`
`show ports {`**`stack-ports`** *`stacking-port-list`* `|` *`port_list`*`} txerrors {`**`norefresh`**`}`

The commands accept stacking port ranges that span multiple nodes. For example, both *`port-list`* and *`stacking-port-list`* might be expressed as `1:1-3:2`.

There are no stacking port configuration options.

See *ExtremeXOS 16.2 Command Reference Guide* for details about these commands.

Note
There is no way to disable a stacking port. Stacking ports are always enabled.

## Configuring Master-Capability

Each node is configurable to be master-capable or not. This means that a node can either be allowed to take on any node role, or be restricted to executing the standby node role only. The default is that a node can take on any role. The restriction is used to avoid the dual master condition. A master-capability configuration change takes effect at the next restart.

You can configure one or more nodes to be allowed to operate either as a master or a backup.

The `configure stacking master-capability` command allows you to set the master-capability of specific nodes, while `configure stacking redundancy` allows you to set the master-capability on all nodes in the stack.

The commands do not allow you to disable master-capability on all nodes in a stack topology.

Note
If the entire stack is restarted in stacking mode without any node having master capability, you need to know the failsafe account and password to log into any node in the stack. If you do not know the failsafe account information, you might need to rescue the stack. See Rescuing a Stack that has No Master-Capable Node on page 172.

You can use any of the following commands to configure the master-capability:

- `configure stacking [node-address node_address | slot slot_number] master-capability [on | off]`
- `configure stacking redundancy [none | minimal | maximal]`

## Rebooting a Stack

You can reboot a stack by entering the command `reboot` from the master. You can:

- Reboot all the nodes in the stack topology
- Reboot a specific node
- Reboot all nodes in the active topology
- Move a node to a standby node
- Reboot the stack topology so that every node comes up in standby role

- To reboot all nodes in the active topology, enter the following command from a master node login session: `reboot`
- To reboot all the nodes in the stack topology, enter: `reboot stack-topology`
- To reboot a specific node, enter: `reboot node-address node-address`
- To reboot an active node from another active node, enter: `reboot slot slot-number`

# Changing the Stack Configuration

The following topics describe common tasks for changing the configuration of a stack that was defined previously.

## Adding a Node to a Stack

Adding a new switch, or node, to an active stack topology is similar to bringing up a new stack.

Note

If the node being added is actually a replacement node for one that was previously removed, see Replacing a Node with the Same Switch Type on page 158 or Replacing a Node with a Different Switch Type on page 160.

Review the general and model-specific configuration guidelines for the node you are installing. These guidelines are described in *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

The examples in the following procedure assume that your current stack has six nodes and you are adding a new node at slot 7.

To add a node to a stack, follow these steps.

1. Before connecting the new node to the stack, prepare it as follows:
   a. With the power off, install any required option cards as described in *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.
   b. Power on the new node.
   c. Use the `show switch` command to verify that the new node's software is compatible with the stack:
      - The new node, must run the same ExtremeXOS version as the stack. Install the correct version if necessary.
      - The ExtremeXOS software must be booted on the same image (primary or secondary) as the stack. If the new node is booted on a different image, change the image before you continue.
   d. Use the `enable stacking` command to enable stacking. Then decline the Easy Setup option.
   e. Configure a unique slot number for the new node (see Configuring Slot Numbers on page 136).

      Select the lowest slot number that is not already in use in the stack. In the following example, the new node has MAC address 00:04:96:26:6c:92 and is assigned to slot 7.

      ```
      Switch.3 # configure stacking node-address 00:04:96:26:6c:92 slot-number 7
      This command will take effect at the next reboot of the specified node(s).
      ```

   f. Configure the node's master-capability to correspond to the role it should have in the stack (see Configuring Master-Capability on page 155).
   g. If the new node will operate as a master-capable node, use the `show licenses` command to verify that the enabled license level is at the same level as the master-capable nodes in the stack.

      If necessary, configure the license-level restriction of the new node to be same as the other master-capable nodes in the stack (see Managing Licenses on a Stack on page 146).
   h. Configure the node role priority to correspond to the priority it should have in the stack (see Configuring the Master, Backup, and Standby Roles on page 137).

i. Configure an alternate IP address and gateway (see Configuring an Alternate IP Address and Gateway on page 150).

j. If the new node is a Summit X670V switch with a VIM4-40G4X option card, configure the option card ports as described in Configuring Stacking Port Operation with the VIM4-40G4X Option Card on page 141.

k. If the new node will use the SummitStack-V feature, configure the alternate stack ports as described in #unique_306.

l. If the stack will use *MPLS (Multiprotocol Label Switching)*, enter the command `configure stacking protocol enhanced`.

**Note**

To use MPLS, the stack can contain only Summit X460-G2, X670-G2, and X770 switches, and all switches must use the enhanced stacking protocol.

2. Connect the stacking cables to the new node.

   The connections should be made such that the new node appears in the natural position in the stack and in the slot. The following example adds a new node that becomes slot 7.

   - Break the connection between slot 6 port 2 and slot 1 port 1.
   - Connect slot 7 port 1 (the new node) to slot 6 port 2.
   - Connect slot 7 port 2 (the new node) to slot 1 port 1.

   For more information about cabling, , see *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

3. Reboot the new node.
4. At the stack master node, enter the command `synchronize stacking node-address` *`node-address`*.
5. Reboot the new node by entering the command `reboot node-address` *`node-address`*].
6. (Optional) Run the `show stacking` and `show slot` commands, as shown in the following example, to verify that the configuration is what you want.

```
Slot-1 Stack.14 # show stacking
Stack Topology is a Ring
Active Topology is a Ring
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
*00:04:96:26:6a:f1  1     Active       Master   CA-
00:04:96:26:6c:93   2     Active       Standby  CA-
00:04:96:26:5f:4f   3     Active       Backup   CA-
00:04:96:1f:a5:43   4     Active       Standby  CA-
00:04:96:28:01:8f   5     Active       Standby  CA-
00:04:96:20:b2:5c   6     Active       Standby  CA-
00:04:96:26:6c:92   7     Active       Standby  CA-
* - Indicates this node
Flags:  (C) Candidate for this active topology, (A) Active Node
(O) node may be in Other active topology
#
Slot-1 stack.15 # show slot
Slots    Type                Configured          State       Ports
-------------------------------------------------------------------
Slot-1   X670V-48x           X670V-48x           Operational   64
Slot-2   X480-48t(SSV80)     X480-48t(SSV80)     Operational   48
Slot-3   X480-48t(40G4X)     X480-48t(40G4X)     Operational   64
Slot-4   X460-48p            X460-48p            Operational   58
```

```
Slot-5   X670V-48x            X670V-48x            Operational   64
Slot-6   X670V-48x            X670V-48x            Operational   64
Slot-7   X670V-48x            X670V-48x            Operational   64
Slot-8                                             Empty         0
```

## Replacing a Node with the Same Switch Type

When you replace a node with the same switch type, for example when you replace an X460-G2-24t-GE4 switch with another X460-G2-24t-GE4 switch, you can continue to use the same stack configuration.

If you are replacing a node with a different switch type, you must change the stack configuration before the new node can operate. Follow the procedure in Replacing a Node with a Different Switch Type on page 160.

**Note**

Summit X440, X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 switches (configured using `configure stacking easy-setup`) use the enhanced stacking protocol by default. Summit X450-G2, X460-G2, X670-G2, and X770 switches support only enhanced mode. When you are replacing an X440, X460, X480, or X670 switch configured with the enhanced stacking protocol, be sure to add `configure stacking protocol enhanced` before joining the switch to the active stack topology.

To replace a node with an identical switch type, follow these steps:

1. Use the `show switch`, `show licenses`, and `show stacking configuration` commands to display configuration information for the node that is being replaced.

   Note the following attributes of the node you are replacing:

   - ExtremeXOS software version
   - Partition on which the switch is booted
   - Effective license level for the stack
   - Slot number
   - Stacking protocol: standard or enhanced
   - Master-capable feature configuration
   - Node priority
   - Alternate gateway IP address
2. Remove the stacking cables from the node that is being replaced.
3. Replace the node with another switch of the same type.
4. Before connecting the new (replacement) switch to the stack, prepare the switch as follows:
   a. Review the attributes needed for the node you are installing, as listed in step 1.
   b. With the power off, install any required option cards as described in *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.
   c. Power on the new node.

d. Use the `show switch` command to verify that the new node's software is compatible with the stack:
   - The new node, must run the same ExtremeXOS version as the stack. Install the correct version if necessary.
   - The ExtremeXOS software must be booted on the same image (primary or secondary) as the stack. If the new node is booted on a different image, change the image before you continue.

e. Use the `enable stacking` command to enable stacking. Then decline the Easy Setup option.

f. Configure the slot number for the replacement node using the slot number noted in step 1 (see Configuring Slot Numbers on page 136).

g. If the replaced node was using the enhanced stacking protocol, use the `configure stacking protocol` command to select that protocol.

h. Configure the node's master-capability to correspond to the role it should have in the stack (see Configuring Master-Capability on page 155).

i. If the new node will operate as a master-capable node, use the `show licenses` command to verify that the enabled license level is at the same level as the master-capable nodes in the stack.

   If necessary, configure the license-level restriction of the new node to be same as the other master-capable nodes in the stack (see Managing Licenses on a Stack on page 146).

j. Configure the node role priority to correspond to the priority it should have in the stack (see Configuring the Master, Backup, and Standby Roles on page 137).

k. Configure an alternate IP address and gateway (see Configuring an Alternate IP Address and Gateway on page 150).

l. If the new node is a Summit X670V switch with a VIM4-40G4X option card, configure the option card ports as described in Configuring Stacking Port Operation with the VIM4-40G4X Option Card on page 141.

m. If the new node will use the SummitStack-V feature, configure the alternate stack ports as described in #unique_306.

n. If the stack will use *MPLS*, enter the command `configure stacking protocol enhanced`.

**Note**
To use MPLS, the stack can contain only Summit X460-G2, X670-G2, and X770 switches, and all switches must use the enhanced stacking protocol.

5. Connect the stacking cables and reboot the node. The new node will join the stack topology.

   For cabling instructions, see the *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

6. At the stack master node, enter `synchronize stacking`.

**Note**
If the master node was replaced, log into another stack node before entering this command.

7. Reboot the new node by entering the command: `reboot slot [slot-number | node-address node-address]`.

   Note
   If the master node was replaced, reboot the stack by entering the reboot command at the master node.

8. (Optional) Run the `show stacking configuration` command and verify that the configuration is what you want.

   Note
   To verify that the new node became operational, enter the `show slot {slot {detail} | detail }` command. If the slot shows a Mismatch state, the node was replaced with a different type of switch (see Replacing a Node with a Different Switch Type on page 160).

## Replacing a Node with a Different Switch Type

When you replace a node with a different switch type, you cannot continue to use the same stack configuration. The slot configuration for the replaced node must change to reflect the new switch type.

Note
If you are replacing a node with the same switch type, you can continue to use the existing stack configuration. For more information, see Replacing a Node with the Same Switch Type on page 158.

To replace a node with a different switch type, follow these steps:

1. Use the `show switch`, `show licenses`, and `show stacking configuration` commands to display configuration information for the node to be replaced.

   Note the following about the node you are replacing:

   - ExtremeXOS software version
   - Partition on which the switch is booted
   - Effective license level for the stack
   - Slot number
   - Stacking protocol: standard or enhanced
   - Master-capable feature configuration
   - Node priority
   - Alternate gateway IP address

2. Enter the `unconfigure slot slot` command to remove the configuration for the node to be replaced.

   All configuration parameters (except for the related node's NVRAM-based configurations such as stacking parameters, image to be used, and failsafe account) for the slot are erased.

3. Remove the stacking cables from the node that is being replaced.
4. Add the new node to the stack following the procedure in Adding a Node to a Stack on page 156.

## Merging Two Stacks

You can join or merge two stacks to create one larger stack. However, the maximum number of nodes in an active topology is eight.

The operation performed when two stack segments are joined together depends on the following factors:

- Whether a slot number is duplicated
- Whether both stacks have master nodes
- The states of the nodes in each stack

If the nodes are configured with stacking enabled, one of the following occurs:

- If two segments are joined, both have operational masters, and at least one of the nodes in one of the stacks duplicates a slot number of a node in the other stack, the join is allowed. The link that has just connected the two stacks shows as Inhibited. This prevents accidental stack joins. In this condition, the nodes on the joined segment can still be reconfigured centrally for stacking.
- If two segments are joined, both have operational masters, and all nodes have assigned slot numbers that are unique in both stacks, the dual master situation is automatically resolved.
- If two segments are joined, there are no duplicate slot numbers, one of the segments has a master and a backup node, and the other segment does not have either a master or a backup node, the nodes in this segment are acquired by the master node. These nodes become standby nodes in the stack.

The nodes that are not configured for stacking do not attempt to join the active topology but join the stack anyway.

Any nodes enabled for stacking that are isolated between nodes (that are not enabled for stacking) attempt to form an isolated active topology.

If one of the nodes that is not configured for stacking is then configured for stacking and restarted, the behavior is as if two active stacks were joined.

### *Example: Merging Two Stacks*

This example demonstrates how to join two stacks, named StackA and StackB. The joined stack assumes the name StackA. Here are displays taken from the original StackA:

```
Slot-1 StackA.8 # show stacking
Stack Topology is a Ring
Active Topology is a Ring
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
*00:04:96:26:60:DD  1     Active       Master   CA-
00:04:96:26:60:EE  2     Active       Backup   CA-
00:04:96:26:60:FF  3     Active       Standby  CA-
(*) Indicates This Node
Flags: (C) Candidate for this active topology, (A) Active node,
(O) node may be in Other active topology
Slot-1 StackA.9 # show stacking configuration
Stack MAC in use: 02:04:96:26:60:DD
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------  --------- ---
*00:04:96:26:60:DD 1   1   Auto 192.168.130.101/24 192.168.130.1    CcEeMm--- Aa
```

```
00:04:96:26:60:EE 2   2   Auto 192.168.130.102/24 192.168.130.1    CcEeMm--- Aa
00:04:96:26:60:FF 3   3   Auto 192.168.130.103/24 192.168.130.1    --EeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
Slot-1 StackA.10 # show stacking stack-ports
Stack Topology is a Ring
Slot Port Select Node MAC Address  Port State  Flags Speed
---- ---- ------ ----------------- ----------- ----- -----
*1   1    Native 00:04:96:26:60:DD Operational CB      10G
*1   2    Native 00:04:96:26:60:DD Operational C-      10G
2   1    Native 00:04:96:26:60:EE Operational C-      10G
2   2    Native 00:04:96:26:60:EE Operational C-      10G
3   1    Native 00:04:96:26:60:FF Operational C-      10G
3   2    Native 00:04:96:26:60:FF Operational CB      10G
* - Indicates this node
Flags:  (C) Control path is active, (B) Port is Blocked
Slot-1 StackA.3 # show slot
Slots    Type                 Configured           State       Ports
-------------------------------------------------------------------
Slot-1   SummitX              SummitX              Operational   26
Slot-2   SummitX              SummitX              Operational   26
Slot-3   SummitX              SummitX              Operational   26
Slot-4                                             Empty          0
Slot-5                                             Empty          0
Slot-6                                             Empty          0
Slot-7                                             Empty          0
Slot-8                                             Empty          0
Slot-1 StackA.4 #
```

Here are displays taken from StackB:

```
Slot-1 StackB.3 # show stacking
Stack Topology is a Ring
Active Topology is a Ring
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
00:04:96:26:60:AA  1     Active       Master   CA-
00:04:96:26:60:88  2     Active       Backup   CA-
00:04:96:26:60:99  3     Active       Standby  CA-
(*) Indicates This Node
Flags: (C) Candidate for this active topology, (A) Active node,
(O) node may be in Other active topology
Slot-1 StackB.4 # show stacking configuration
Stack MAC in use: 02:04:96:26:60:AA
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------  --------- ---
*00:04:96:26:60:AA 1   1   Auto 192.168.131.101/24 192.168.131.1    CcEeMm--- Aa
00:04:96:26:60:88 2   2   Auto 192.168.131.102/24 192.168.131.1    CcEeMm--- Aa
00:04:96:26:60:99 3   3   Auto 192.168.131.103/24 192.168.131.1    --EeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
```

```
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured
Slot-1 StackB.5 # show stacking stack-ports
Stack Topology is a Ring
Slot Port Select Node MAC Address  Port State  Flags Speed
---- ---- ------ ----------------- ----------- ----- -----
1   1    Native 00:04:96:26:60:AA Operational C-      10G
1   2    Native 00:04:96:26:60:AA Operational CB      10G
2   1    Native 00:04:96:26:60:88 Operational CB      10G
2   2    Native 00:04:96:26:60:88 Operational C-      10G
3   1    Native 00:04:96:26:60:99 Operational C-      10G
3   2    Native 00:04:96:26:60:99 Operational C-      10G
* - Indicates this node
Flags:  (C) Control path is active, (B) Port is Blocked
Slot-1 StackB.6 # show slot
Slots    Type                Configured           State       Ports
-------------------------------------------------------------------
Slot-1   SummitX             SummitX              Operational   26
Slot-2   SummitX             SummitX              Operational   26
Slot-3   SummitX             SummitX              Operational   26
Slot-4                                            Empty          0
Slot-5                                            Empty          0
Slot-6                                            Empty          0
Slot-7                                            Empty          0
Slot-8                                            Empty          0
```

1. Form the new stack. Assuming both stacks are rings, break one link in each stack as follows:
   a. For StackA, break the link between node 00:04:96:26:60:FF port 2 and node 00:04:96:26:60:DD port 1.
   b. For StackB, break the link between node 00:04:96:26:60:99 port 2 and node 00:04:96:26:60:AA port 1.
2. Connect the broken links between the two stacks to form a ring as follows:
   a. Connect node 00:04:96:26:60:FF port 2 to node 00:04:96:26:60:AA port 1.
   b. Connect node 00:04:96:26:60:99 port 2 to node 00:04:96:26:60:DD port 1.

      Because both stacks are active stacks and have duplicate slot numbers, the links between the two stacks are in Inhibited state.
3. Assume that the master node of stackA is to be the master node of the joined stack. Log into the intended master node.
4. Verify the details of the new stack using the following commands: `show stacking`, `show stacking configuration`, and `show stacking stack-ports`.

```
Slot-1 StackA.11 # show stacking
Stack Topology is a Ring
Active Topology is a Daisy-Chain
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
*00:04:96:26:60:DD  1     Active       Master   CA-
00:04:96:26:60:EE  2     Active       Backup   CA-
00:04:96:26:60:FF  3     Active       Standby  CA-
00:04:96:26:60:AA  1     Active       Master   --O
00:04:96:26:60:88  2     Active       Backup   --O
00:04:96:26:60:99  3     Active       Standby  --O
(*) Indicates This Node
```

```
Flags: (C) Candidate for this active topology, (A) Active node,
(O) node may be in Other active topology

Slot-1 StackA.12 # show stacking configuration
Stack MAC in use: 02:04:96:26:60:DD
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------  --------- ---
*00:04:96:26:60:DD 1   1   Auto 192.168.130.101/24 192.168.130.1    CcEeMm--- Aa
00:04:96:26:60:EE 2   2   Auto 192.168.130.102/24 192.168.130.1    CcEeMm--- Aa
00:04:96:26:60:FF 3   3   Auto 192.168.130.103/24 192.168.130.1    --EeMm--- Aa
00:04:96:26:60:AA 1   1   Auto 192.168.131.101/24 192.168.131.1    CcEe--i-- --
00:04:96:26:60:88 2   2   Auto 192.168.131.102/24 192.168.131.1    CcEe--i-- --
00:04:96:26:60:99 3   3   Auto 192.168.131.103/24 192.168.131.1    --Ee--i-- --
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured

Slot-1 StackA.13 # show stacking stack-ports
Stack Topology is a Ring
Slot Port Select Node MAC Address  Port State  Flags Speed
---- ---- ------ ----------------- ----------- ----- -----
*1   1    Native 00:04:96:26:60:DD Inhibited   --      10G
*1   2    Native 00:04:96:26:60:DD Operational C-      10G
2   1    Native 00:04:96:26:60:EE Operational C-      10G
2   2    Native 00:04:96:26:60:EE Operational C-      10G
3   1    Native 00:04:96:26:60:FF Operational C-      10G
3   2    Native 00:04:96:26:60:FF Inhibited   --      10G
1   1    Native 00:04:96:26:60:AA Inhibited   --      10G
1   2    Native 00:04:96:26:60:AA Operational C-      10G
2   1    Native 00:04:96:26:60:88 Operational C-      10G
2   2    Native 00:04:96:26:60:88 Operational C-      10G
3   1    Native 00:04:96:26:60:99 Operational C-      10G
3   2    Native 00:04:96:26:60:99 Inhibited   --      10G
* - Indicates this node
Flags:  (C) Control path is active, (B) Port is Blocked
Slot-1 StackA.14 #
```

5. Configure the nodes so that they all have unique slot numbers.

   Because the slot numbers configured for the first three nodes in your stack are consistent with automatic slot assignment, you may perform automatic slot assignment using `configure stacking slot-number automatic`.

6. Configure the stack MAC address using the `configure stacking mac-address` command.
7. Configure stacking redundancy so that only slots 1 and 2 are master-capable with the `configure stacking redundancy minimal` command.
8. Configure new alternate IP addresses for nodes from original StackB.

   Assume that the block of addresses allocated to StackA can be extended, and use the automatic form of the command as follows: `configure stacking alternate-ip-address 192.168.130.101/24 192.168.130.1 automatic`

9. For master-capable nodes, configure a license restriction to be the minimum of the two original values on all master-capable nodes.

   Alternatively, you may purchase license upgrades from Extreme if necessary. In this case, use the command: `configure stacking license-level edge`

10. Either reboot the entire stack topology using the reboot stack-topology command, or individually reboot the three nodes formerly from StackB. The latter requires the following commands:

```
reboot node 00:04:96:26:60:99
reboot node 00:04:96:26:60:88
reboot node 00:04:96:26:60:AA
```

    Reboot the nodes in this order: standby nodes first, backup node next, and master node last. Because none of these nodes is master-capable, there is no temporary dual master situation as a result of these separate node reboots.

11. When the rebooted nodes come back up, run the following commands to see the resulting stack. You can verify that the joined stack came up as expected – that is, all nodes have unique slot numbers, a common stack MAC address, and so forth:

```
Slot-1 StackA.11 # show stacking
Stack Topology is a Ring
Active Topology is a Ring
Node MAC Address    Slot  Stack State  Role     Flags
------------------  ----  -----------  -------  ---
*00:04:96:26:60:DD  1     Active       Master   CA-
00:04:96:26:60:EE   2     Active       Backup   CA-
00:04:96:26:60:FF   3     Active       Standby  CA-
00:04:96:26:60:AA   4     Active       Standby  CA-
00:04:96:26:60:88   5     Active       Standby  CA-
00:04:96:26:60:99   6     Active       Standby  CA-
(*) Indicates This Node
Flags: (C) Candidate for this active topology, (A) Active node,
(O) node may be in Other active topology

Slot-1 StackA.12 # show stacking configuration
Stack MAC in use: 02:04:96:26:60:DD
Node               Slot         Alternate          Alternate
MAC Address        Cfg Cur Prio Mgmt IP / Mask     Gateway          Flags     Lic
------------------ --- --- ---- ------------------ ---------------  --------- ---
*00:04:96:26:60:DD 1   1   Auto 192.168.130.101/24 192.168.130.1    CcEeMm--- Aa
00:04:96:26:60:EE 2   2   Auto 192.168.130.102/24 192.168.130.1    CcEeMm--- Aa
00:04:96:26:60:FF 3   3   Auto 192.168.130.103/24 192.168.130.1    --EeMm--- Aa
00:04:96:26:60:AA 4   4   Auto 192.168.130.104/24 192.168.130.1    --EeMm--- Aa
00:04:96:26:60:88 5   5   Auto 192.168.130.105/24 192.168.130.1    --EeMm--- Aa
00:04:96:26:60:99 6   6   Auto 192.168.130.106/24 192.168.130.1    --EeMm--- Aa
* - Indicates this node
Flags:  (C) master-Capable in use, (c) master-capable is configured,
(E) Stacking is currently Enabled, (e) Stacking is configured Enabled,
(M) Stack MAC in use, (m) Stack MACs configured and in use are the same,
(N)  Stack link protocol Enhanced in use, (n) Stack link protocol Enhanced configured,
(i) Stack MACs configured and in use are not the same or unknown,
(-) Not in use or not configured
License level restrictions: (C) Core, (A) Advanced edge, or (E) Edge in use,
(c) Core, (a) Advanced edge, or (e) Edge configured,
(-) Not in use or not configured

Slot-1 StackA.13 # show stacking stack-ports
Stack Topology is a Ring
Slot Port Select Node MAC Address  Port State  Flags Speed
---- ---- ------ ----------------- ----------- ----- -----
*1   1    Native 00:04:96:26:60:DD Operational C-      10G
*1   2    Native 00:04:96:26:60:DD Operational C-      10G
2   1    Native 00:04:96:26:60:EE Operational C-      10G
2   2    Native 00:04:96:26:60:EE Operational C-      10G
3   1    Native 00:04:96:26:60:FF Operational C-      10G
3   2    Native 00:04:96:26:60:FF Operational C-      10G
4   1    Native 00:04:96:26:60:AA Operational C-      10G
4   2    Native 00:04:96:26:60:AA Operational CB      10G
5   1    Native 00:04:96:26:60:88 Operational CB      10G
```

```
5   2    Native 00:04:96:26:60:88 Operational C-      10G
6   1    Native 00:04:96:26:60:99 Operational C-      10G
6   2    Native 00:04:96:26:60:99 Operational C-      10G
* - Indicates this node
Flags:  (C) Control path is active, (B) Port is Blocked

Slot-1 StackA.14 #
Slot-1 StackA.3 # show slot
Slots    Type                Configured          State       Ports
-------------------------------------------------------------------
Slot-1   SummitX             SummitX             Operational   26
Slot-2   SummitX             SummitX             Operational   26
Slot-3   SummitX             SummitX             Operational   26
Slot-4   SummitX                                 Operational   50
Slot-5   SummitX                                 Operational   26
Slot-6   SummitX                                 Operational   26
Slot-7                                           Empty          0
Slot-8                                           Empty          0
```

12. Configure the new slots in VLANs, IP subnetworks, and so forth as required.

## Removing a Node from a Stack

To remove a node from a stack, follow these steps:

1. Determine if the target node to be removed is using the SummitStack-V feature by issuing the `show stacking stack-ports` command.

   Examine the Select column to see whether the target node is using alternate (non-native) stacking ports.
2. If the target node is using alternate stacking ports, do the following:
   a. Log into the target node and issue the command: `unconfigure stacking-support`
   b. Log out of the target node.

   **Note**
   Do not reboot the target node at this time.
3. Log into the master node.
4. Delete the target node stacking configuration by entering the following command:

   `unconfigure stacking {`**`node-address`** *`node_address`* | **`slot`** *`slot_number`*`}`
5. Reboot the target node by entering the following command:

   `reboot [`**`node`** *`node-address`* | **`slot`** *`slot-number`*`]`

   When the node reboots, it detects that the configuration file selected is a stacking configuration file (see Stack Configuration Parameters, Configuration Files, and Port Numbering on page 129). It de-selects the configuration file and uses the factory defaults.
6. Disconnect the node from the stack, and redeploy it as needed.

## Dismantling a Stack

To dismantle a stack and use the Summit switches in stand-alone mode, do the following:

1. Determine if the stack is using the SummitStack-V feature by issuing the `show stacking stack-ports` command.

   Examine the Select column to see whether any nodes are using alternate (non-native) stacking ports.
2. For every non-master node in the stack that is using alternate stacking ports, log into the node and issue the `unconfigure stacking-support` command.

   **Note**

   If a node is a member of the active topology, node login can be accomplished from the master node using the `telnet slot` *`slot-number`* command. Otherwise you will need access to the node's console port, or you can log in through a management network.

   Do not reboot any switches. It is not necessary to unconfigure stacking-support on the master node.
3. When the stacking-support option has been removed from all non-master stack nodes, log into the master node and issue the `unconfigure switch all` command.

   After this command is entered, the configuration file is deselected, all stacking parameters are reset to factory defaults, and all nodes in the active topology reboot. In effect, this sets all nodes back to the factory default configuration, thus allowing each switch to be redeployed individually.

# Troubleshooting a Stack

Use this section to diagnose and troubleshoot problems with your stacked switches.

The `show stacking`, `show stacking configuration`, `show stacking-support`, and `show stacking stack-ports` commands can help you identify switches and stack ports that are improperly configured, not properly cabled, or powered down.

The commands can help you spot common problems like the following.

## Incorrect Software Version

If a node appears in the stack as expected but does not appear to be operating as configured, use the `show slot {`*`slot`* `{`**`detail`**`} | `**`detail`**` }` command to see if there is a license mismatch or if the node is running an incorrect software version. For more information, see Managing Licenses on a Stack on page 146.

**Note**

If a correctly cabled and powered-on node does not appear in the stack, the node might be running an version that is earlier than 12.0. Upgrade the version using the same procedure you would use if the node was not part of the stack.

## Stacking Not Enabled

If the `show stacking` command displays the status as Disabled for any node in the stack, use the `enable stacking` command to enable stacking on that node. You can issue the command from the

master node, and you can reboot the disabled node from the master node to activate the slot number configuration.

## Choice of Master Node

If the switch with the highest priority was not elected master, it might be because the stack nodes nodes were powered up at different times. Reboot all nodes in the stack simultaneously.

The following topics contain information for troubleshooting problems related to the choice of the master node:

- Managing a Dual Master Situation on page 168
- Connecting to a Stack with No Master on page 171
- Rescuing a Stack that has No Master-Capable Node on page 172

## Choice of Backup and Standby Nodes

About five minutes after a master node takes control of the stack, you might see one of the following messages:

```
Warning: The Backup stack node is not as powerful or as capable
as the Master stack node. This configuration is not recommended
for successful use of the failover feature.
Notice: There are Standby stack nodes which are more powerful and more capable
than the Master and/or Backup stack nodes. This configuration is not recommended
for optimal stack performance. We recommend that you reconfigure the stacking
master-capability and/or priority parameters to allow the higher performing and
more capable nodes to become Master and/or Backup stack nodes.
```

In each case, to optimize your use of the failover feature, follow the guidelines in Configuring the Master, Backup, and Standby Roles on page 137.

## Loss of Saved Files

Note that saved files (backup configurations, script, etc.) are lost on a node that becomes a non-master when implementing stacking. Disabling or deleting the stacking configuration does not restore the files.

Refer to the following topics for help with troubleshooting other problems in your stack.

## Managing a Dual Master Situation

If a daisy chain is broken, or if a ring is broken in two places, it is possible to form two separate active stack topologies. This results in a dual master situation, as shown in the following example .

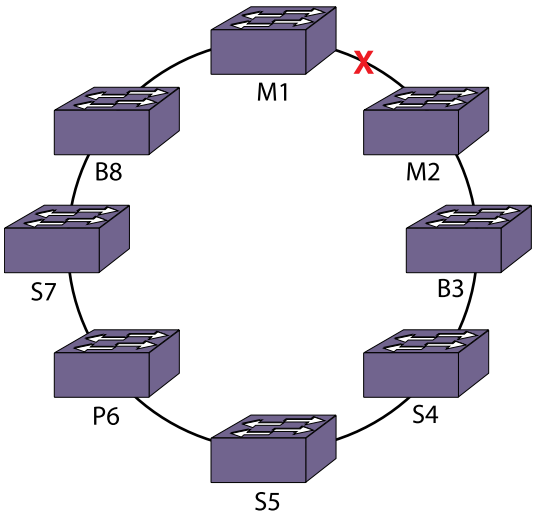



**Figure 11: Example of a Split Stack that Results in a Dual Master Situation**

| | |
|---|---|
| P6 | Node 6 is powered off |
| M | Master nodes |
| B | Backup nodes |
| S | Standby nodes |
| X | Indicates the broken link |

In the example, a link was broken while a node in the ring was powered off. The broken link formerly connected the original master (M1) and backup (M2) nodes of a single active topology.

All nodes in the stack except the powered-off node are in the active topology and all nodes are configured to be master-capable. Nodes 1, 7 and 8 form an active topology and nodes 2, 3, 4, and 5 form another active topology. Node M2 immediately transitions from backup to master node role. Nodes B8 and B3 are elected in their respective active topologies as backup nodes.

If the backup node is on one stack and the master node is on the other, the backup node becomes a master node because the situation is similar to that of master failure. Because both stacks are configured to operate as a single stack, there is confusion in your networks. For example, all of the switch's configured IP addresses appear to be duplicated. The management IP address also appears to be duplicated since that address applies to the entire original stack.

To help mitigate the dual master problem, you can configure master-capability so as to prevent some nodes in the stack from operating in backup or master node roles. In addition, you can force all nodes in the (broken) stack topology to restart and come up as not master-capable for the life of that restart. The `save configuration {primary | secondary | existing-config | new-config}` command saves the configuration on all nodes in the active topology.

Standby nodes that exist in a severed stack segment that does not contain either the original master or backup node do not attempt to become the master node. Instead, these nodes reboot. After rebooting, however, a master election process occurs among the nodes on this broken segment, resulting in a dual master situation.

Dual master conditions are also possible when two non-adjacent nodes in a ring or a single (middle) node in a daisy chain reboot.

For a period of time, a rebooting node does not advertise itself to its neighbors, resulting in temporary stacking link failures. This can cause node isolation, and the nodes that are isolated perform as a severed stack segment depending on the circumstances of the severance:

- If the backup node is on the broken portion, it becomes a (dual) master.
- If the backup node is on the same portion as the master, all nodes on the (other) broken portion reboot.

When the rebooting nodes have sufficiently recovered, or when a severed stack is rejoined, the dual master condition is resolved, resulting in the reboot of one of the master nodes. All standby and backup nodes that had been acquired by the losing master node also reboot.

You can avoid a dual master possibility during configuration by:

- Configuring the stack in a ring topology.
- Avoiding having too many master-capable nodes when configuring larger stacks.
- Placing the master-capable nodes that provide stack redundancy in such a way that broken stacking links are unlikely.

### *Eliminating a Dual Master Situation Manually*

To eliminate the dual master situation, you need to know all the nodes that are supposed to be in the stack. You might lose the management connectivity to the master node because the other master node duplicates the stack's primary management IP address and stack MAC address.

Note

The following procedure is necessary only if you cannot reconnect the severed link in a timely manner. If you can reconnect, the dual master condition resolves itself. The formerly broken portion of the stack reboots and the nodes come up as standby nodes.

1. If you lose the management connectivity, log into the master node using its alternate management IP address.
2. Use the `show stacking` command to determine which nodes have been lost from the stack.

   You should already know which nodes are expected to be part of the stack.
3. Log into any node in the severed segment you wish to deactivate, either through its console port or through the management interface using the alternate management IP address.
4. Issue the `show stacking` command to find out whether the broken segment has indeed elected a new master node.

5. Using the `reboot stack-topology as-standby` command, reboot the broken segment.

   This forces all nodes in the segment to come up as standby nodes.

   If you have unsaved configuration changes, take care when selecting the stack segment to be rebooted. You should reboot the segment that has the smaller System UpTime.

   If you know the node that was master of the unbroken stack, you can reboot the stack segment that does not contain this master node. Otherwise, determine the System UpTime shown by each master node.

   If the System UpTimes of both masters are the same, you can reboot either segment without loss of unsaved configuration changes. If the System UpTimes of both masters differ, you must reboot the segment with the smaller System UpTime.

*Automatic Resolution of the Dual Master Situation*

When two stack segments are connected together and no slot number is duplicated on either segment, it is assumed that this is a severed stack rejoin.

It is possible that each stack segment has its own master. Resolution of the dual master situation should generally be in favor of the original stack segment's master node. This is because the original stack segment may still retain the unsaved configuration. If the severed segment was restarted before electing a new master node, the unsaved configuration is lost on that segment.

The master election is done using the System UpTime. The master election process collects the System UpTime information of the nodes. If a failover occurs, the System UpTime is inherited by the new master node, and the new master node continues to increase it as time passes. Thus the System UpTime is the time since a master was first elected on a segment. When the stack is broken and both master and backup nodes are on the same segment, the severed segment always has the smaller System UpTime.

If a stack severance results in the master and backup nodes being on different segments, both have the same System UpTime. In this case, the master is elected using the normal node role election method.

## Connecting to a Stack with No Master

If an entire stack has no master node because the stack has been rebooted in standby only mode, you can log in to a node by using the failsafe account.

If a new node has been added to the stack since the stack failsafe account was configured, logging in to that node requires knowledge of the failsafe account information that is already configured into that node's NVRAM.

If you do not know the failsafe account and you still want to log in to the stack, you have to:

1. Join the stack to another segment that has a master node to which the you have access.
2. Manually restart the stack to clear the as-standby condition if the reboot stack-topology as-standby command was previously used.
3. Use the procedure described in Rescuing a Stack that has No Master-Capable Node on page 172.

## Rescuing a Stack that has No Master-Capable Node

It is possible for all nodes in a stack to have master-capability set to OFF. For example, if a stack was operating with no redundancy (for example, with one master-capable node) and the master node failed, all other nodes in the stack restart as standby nodes and there is no master node.

Another example is the case where you dismantle a stack before using the `unconfigure stacking` command or the `unconfigure switch all` command. In this case, the individual switches are configured for stacking, are not master-capable, and are isolated from a stack master.

In this situation, the only security information available is the failsafe account. If you know the failsafe user name and password, you can log into any node and reconfigure master-capability or redundancy. However, if you do not know the failsafe account information, there is another way you can change the configuration.

The procedure described here generally is not needed if another master-capable node is expected to rejoin the stack. If this procedure is used, it is possible that the new master will duplicate the master that is expected to rejoin later.

To assign a new master-capable node, follow these steps.

1. At the login prompt, enter the following special login ID exactly as displayed below (all uppercase letters) and press **[Enter]**:
   `REBOOT AS MASTER-CAPABLE`

   The following message displays:
   `Node reboot initiated with master-capability turned on.`

   This node then sets an internal indicator that is preserved across the reboot. While restarting, the node notices and resets this indicator, ignores the node master-capability configuration, and becomes a master node.

   Because the `configuring anycast RP` command saves the configuration file to all nodes, the node that just rebooted as master-capable should have access to the security information that was configured for the stack. If a *RADIUS (Remote Authentication Dial In User Service)* server is needed, the selected node requires a network connection for authentication.

   The special login ID described here is available only if all of the following conditions are met:

   - The node supports the SummitStack feature.
   - Stacking mode is active on the node.
   - All nodes in the active topology have master-capability turned off.
   - There is no master node in the active topology.

   If the above conditions are met, five minutes after starting the node and every five minutes after that, the following message displays on the console:

   ```
   Warning: the stack has no Master node and all active nodes are operating
   with master-capability turned off. If you wish to reconfigure, you may log
   in using the failsafe account. Alternatively, you may use the special login
   REBOOT AS MASTER-CAPABLE with no password to force a reboot of a node with
   master-capability temporarily turned on.
   ```

   Using the special login ID does not alter the master-capability configuration permanently. If you restart a node that has been restarted with the special login ID, that node restarts using its configured master-capability, unless you again use the special login ID to restart.

   When a node has been rebooted using the special login ID, it becomes a master node. While the node is a master, the special login ID is not recognized, even though the entire stack is still configured as not master-capable. To get the special login ID to be recognized, the node must be rebooted again.
2. If a node was intentionally separated from the stack without first being unconfigured, its security configuration might be unusable. In this case, perform the following steps:
   a. Connect to the node's console port.
   b. Reboot the node using the special `REBOOT AS MASTER-CAPABLE` login ID described in step 1.
   c. During the reboot, enter the bootrom program by waiting until you see the message `Starting Default Bootloader ...` and then pressing and holding the space bar until the bootrom prompt displays.
   d. Force the switch to boot up with a default configuration by entering the following commands at the bootrom prompt:

      ```
      config none
      boot
      ```

The switch boots up in stacking mode operating as a master-capable switch. You can then log in using the default admin account with no password.

## Daisy Chain

If you issue the **show stacking** command and see that the stack topology is a daisy chain, you must resolve the problem before the stack will function properly.

Here is a sample CLI output from **show stacking** showing that the stack is in a daisy chain.

```
 # show stacking
Stack Topology is a Daisy-Chain
This node is not in an Active Topology
Node MAC Address     Slot  Stack State  Role     Flags
------------------   ----  -----------  -------   ---
*00:04:96:7d:fc:9a    -    Disabled      Master   ---
```

To solve a daisy-chain problem, follow these steps:

1. Ensure that each stack node meets the following conditions:
2. Wait for all stack nodes to finish booting and get through AAA login.
3. Enter `show stacking` to verify that all stack nodes are connected in a ring topology.
4. Check that the stacking port LEDs show as active on all nodes.
5. Reconnect the stacking cables.

## Broken Stack (Isolated Nodes)

When a stack setup is not properly unconfigured before the stacking cables are removed, former stack nodes can become isolated.

A common symptom of this problem is that either or both ends of a stacking link show that the state is No Neighbor. This can mean that the port at either end is configured incorrectly or it can indicate a mismatch in the stacking protocol configured. Some configuration errors can produce the No Neighbor state at one end and the Link Down state at the other end.

Here is an example of how an isolated node would look when you log in to it:

```
 (pending-AAA) login:
Warning: the stack has no Master node and all active nodes are operating with
master-capability turned off. If you wish to reconfigure, you may log in using
the failsafe account. Alternatively, you may use the special login
REBOOT AS MASTER-CAPABLE
with no password to force a reboot of a node with master-capability temporarily
turned on.
```

If the `show stacking` command shows the stack state for a slot as Active with the “O” flag set, the node associated with that slot might be isolated from the rest of the stack by a failed node.

To fix an isolated node problem, follow these steps:

1. Verify the physical connections between each port and the rest of the stack.
2. Following the prompt in the sample console session, log in by entering `reboot at master-capable`

   The switch reboots as a master-capable switch.

3. Enter `unconfigure stacking` to unconfigure the stack.

   This command takes effect only once. If the switch is powered off or rebooted you must reissue the **`reboot at master-capable`** command.

The switch is now available for use in a new stack.

## Failed Stack

Even if the **`enable stacking`** command executes without any warnings or prompts, it is still possible for a stack to fail.

When a stack fails, verify that all of the following conditions are met.

A common sign of a failed stack is one of the non-master nodes being stuck at `(pending-AAA)` for more than five minutes. You can attempt to communicate with the master node (the switch on which the stack was enabled) and determine what went wrong by checking the state of each slot in the stack, as shown in this example CLI session:

```
Slot-1 Stack.1 # show slot
Slots    Type               Configured State      Ports
-------------------------------------------------------
Slot-1   X440-48p-10G       Operational            50
Slot-2   X440-24x-10G       Failed                 26
Slot-3   Empty 0
Slot-4   Empty 0
Slot-5   Empty 0
Slot-6   Empty 0
Slot-7   Empty 0
Slot-8   Empty 0
```

In this example, Slot-2 is in the failed state. To determine why it failed, enter the `show slot` command – in this example, `show slot 2` – to get detailed information about the state of the failed slot:

```
Slot-1 Stack.1 # show slot 2
Slot-2 information:
State: Failed
Download %: 0
Last Error: License Mismatch
Restart count: 1 (limit 5)
Serial number: N/A(0) N/A(0)
Hw Module Type: X440-24x-10G
Configured Type:
Ports available: 26
Recovery Mode: Reset
Node MAC: 00:00:00:00:00:00
Current State: FAIL (License Mismatch)
Image Selected:
Image Booted:
Primary ver:
Secondary ver:
Config Selected: NONE
```

In this example, the highlighted line in the output reveals the cause of the failure: a license mismatch in Slot-2.

## Failed Stack Node

If the `show stacking` command shows the stack state for a slot as Failed, check the following things:

- Does the `show stacking stack-ports` command show a port state as Inhibited? If so, the problem might be a duplicate slot number. If more than one node is using the same slot number, change the slot number on one of the affected nodes to a unique slot number.
- Is the affected node isolated by other nodes for which the stack state is listed as Disabled? If so, you need to enable stacking on the disabled nodes.
- Enter the show slot detail command. If the command displays License Mismatch, either upgrade the node license, or configure a license level restriction so that all master-capable nodes are at the same effective license level.
- Enter the `show slot detail` command. If the command displays `Incompatible EXOS Version`, log into the master node and use the `synchronize slot` command to update the failed node.

## License Mismatch

A stack might have a failed node even though the **`show licenses`** command shows that all nodes are running the same license and/or feature packs.

To fix a license mismatch, do the following:

1. At the master node's console, enter `show slot` *`slot_number`* repeatedly until you find the slot node that is failing.

   Note
   When you are creating a stack, a blinking orange MGMT light usually indicates a license mistmatch on a master-capable standby node.

2. Enter `unconfigure stacking` at the master node to unconfigure the stack.
3. After making a connection to the failed switch's slot, update the license information for that node:
   a. Enter `clear license-info`
   b. Enter `reboot`
   c. Obtain the correct license information for the failed switch, then enter `# enable license` *`license_key`* (where *`license_key`* has the format *`xxxx-xxxx-xxxx-xxxx-xxxx`*)

      For more information about licensing for stack nodes, see Managing Licenses on a Stack on page 146.
4. Enter `show stacking` and `enable stacking` to re-enable the stack.

   Note
   Before EXOS version 12.5, all switches in the stack had to have the upgrade/feature license installed. After EXOS version 12.5, only the first two switches in the stack (the master and master-capable standby) must have the upgrade/feature license installed..

## Stacking Link Failure

A stacking link is said to be failed when one of the following happens:

- The stacking link is physically disconnected.
- The neighbor on a link stops transmitting topology information.
- The link goes down while a node restarts or when it is powered off.

Based on the stacking topology, the stack behavior changes.

### Ring Topology

All traffic paths that were directed through the failed link are redirected. All nodes converge on the new (daisy chain) topology that results from the link break. The Topology Protocol that determines the stack topology immediately informs other nodes that a link has failed. Each node starts the process of redirecting traffic paths.

### Daisy Chain

A stacking link failure means a severed stack. The Topology Protocol reports the loss of all nodes in the severed portion. Depending on master capability configuration and the original location of the backup node, the severed portion may or may not elect a new master node. If it does, the dual master condition may be in effect.

The `show slot {`*`slot`* `{`**`detail`**`} | `**`detail`**` }` command displays the slots that contain active nodes that are in the severed portion as Empty.

## Understanding Stacking Traps

Every stack generates traps that provide status information about switches and stacking ports.

The stack routinely generates the following traps:

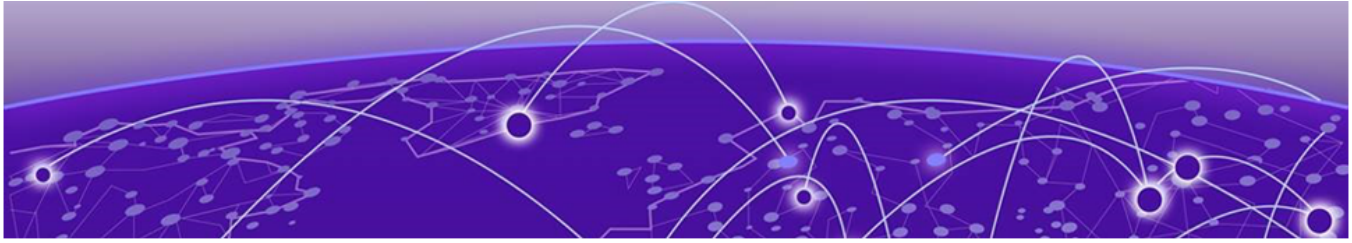
- `extremeStackMemberStatusChanged`
- `extremeStackMemberSlotId`: the slot ID
- `extremeStackMemberOperStatus`: the slot state of the switch

When an overheat condition is detected on an active node, the stack generates the following trap when the node reaches a steady state:

- `extremeStackMemberOverheat`

When a member is added to or deleted from the stack, or any time the status of a stacking port changes, the change is indicated by means of the following traps:

- `extremeStackingPortStatusChanged`
- `IfIndex`: Interface Index of the port
- `extremeStackingPortRemoteMac`: MAC Address of the remote switch attached to this port
- `extremeStackingPortLinkSpeed`: the port's link speed, for example, 100, or 1000 Mbps
- `extremeStackingPortLinkStatus`: the status of the link



# Configuring Slots and Ports on a Switch



This chapter describes the processes for enabling, disabling and configuring individual and multiple ports and displaying port statistics, and configuring slots on modular switches.

## Configuring Slots on Modular Switches

This section describes configuring slots on modular switches, which are the BlackDiamond X8 switches, BlackDiamond 8800 series switches, and SummitStack. In a SummitStack, a slot number is assigned to a node through configuration and stored in the node's NVRAM. It takes effect only when the node restarts. In the following descriptions, the phrase inserted into a slot in a SummitStack means that the node has become active, and because of its configured slot value it appears to be present in a slot when the show slot command is run. The relationship of a node and a slot does not change if the SummitStack is rewired. The term module refers to a Summit family switch that may be present in the stack as an active node.

If a slot has not been configured for a particular type of module, then any type of module is accepted in that slot, and a default port and *VLAN (Virtual LAN)* configuration is automatically generated.

After any port on the module has been configured (for example, a VLAN association, a VLAN tag configuration, or port parameters), all the port information and the module type for that slot must be

saved to non-volatile storage. Otherwise, if the modular switch or SummitStack is rebooted or the module is removed from the slot, the port, VLAN, and module configuration information is not saved.

Note

For information on saving the configuration, see Software Upgrade and Boot Options on page 1522.

You can also preconfigure the slot before inserting the module. This allows you to begin configuring the module and ports before installing the module in the chassis or activating the related node in the SummitStack.

If a slot is configured for one type of module, and a different type of module is inserted, the inserted module is put into a mismatch state and is not brought online.

All configuration information related to the slot and the ports on the module is erased. If a module is present when you issue this command, the module is reset to default settings.

You can configure the number of times that a slot can be restarted on a failure before it is shut down.

- To configure the modular switch or a SummitStack with the type of input/output (I/O) module that is installed in each slot, use the following command:

  `configure slot` *`slot`* **`module`** *`module_type`*
- Use the new module type in a slot, the slot configuration must be cleared or configured for the new module type. To clear the slot of a previously assigned module type, use the following command:

  `clear slot` *`slot`*
- To display information about a particular slot, use the following command:

  `show slot {`*`slot`*`} {`**`detail`**`}`

  Information displayed includes:

  - Module type, part number and serial number
  - Current state (power down, operational, diagnostic, mismatch)
  - Port information

  If no slot is specified, information for all slots is displayed.

  All slots on the modular switches are enabled by default.
- To disable a slot, use the following command:

  `disable slot`
- To re-enable slot, use the following command:

  `enable slot`
- On the BlackDiamond X8 switch, the command to disable a fabric slot is:

  `disable slot` *`FM-1`* `|` *`FM-2`* `|` *`FM-3`* `|` *`FM-4`* `{`**`offline`**`}`

  When a fabric slot is disabled, it is powered off and the bandwidth it provides is unavailable. Disabling an active fabric slot reroutes the switch fabric traffic before powering off the inserted FM blade. Thus, if there are four active fabric modules when one is disabled, there should be no traffic loss.

- On the BlackDiamond X8 switch, the command to enable a fabric slot is:

  `enable slot` *`FM-1 | FM-2 | FM-3 | FM-4`*
- To set the restart-limit, use the following command:

  `configure slot` *`slot_number`* **`restart-limit`** *`num_restarts`*

## Configuring Ports on a Switch

Note

A port can belong to multiple virtual routers (VRs). See Virtual Routers on page 624 for more information on VRs.

### Port Numbering

ExtremeXOS runs on both stand-alone and modular switches, and the port numbering scheme is slightly different on each. There are also special considerations for mobile backhaul routers.

#### *Stand-alone Switch Numerical Ranges*

On a stand-alone switch, such as a Summit family switch, the port number is simply noted by the physical port number, as shown below:

```
5
```

Separate the port numbers by a dash to enter a range of contiguous numbers, and separate the numbers by a comma to enter a range of noncontiguous numbers:

- x-y: Specifies a contiguous series of ports on a stand-alone switch
- x,y: Specifies a noncontiguous series of ports on a stand-alone switch
- x-y,a,d: Specifies a contiguous series of ports and a series of noncontiguous ports on a stand-alone switch

#### *Modular Switch and SummitStack Numerical Ranges*

On a modular switch and SummitStack, the port number is a combination of the slot number and the port number. The nomenclature for the port number is as follows:

```
slot:port
```

For example, if an I/O module that has a total of four ports is installed in slot 2 of the chassis, the following ports are valid:

- 2:1
- 2:2
- 2:3
- 2:4

You can also use wildcard combinations (*) to specify multiple modular slot and port combinations.

The following wildcard combinations are allowed:

- slot:*: Specifies all ports on a particular I/O module or stack node.
- slot:x-slot:y: Specifies a contiguous series of ports on multiple I/O modules or stack nodes.

- slot:x-y: Specifies a contiguous series of ports on a particular I/O module or stack node.
- slota:x-slotb:y: Specifies a contiguous series of ports that begin on one I/O module or stack node and end on another I/O module or stack node.

*Mobile Backhaul Routers*

Mobile backhaul routers include the E4G-200 and E4G-400.

Commands operating on a *port_list* for mobile backhaul routers all use the keyword **tdm**. When the **tdm** keyword is present, the *port_list* is expanded to include only time division multiplexing (TDM) ports, omitting any Ethernet ports occurring within the *port_list* range. Existing CLI commands without the **tdm** keyword continue to work as usual without any change, and these commands omit any TDM ports that may lie within the*port_list* range.

## Enabling and Disabling Switch Ports

By default, all ports are enabled. You have the flexibility to receive or not to receive *SNMP (Simple Network Management Protocol)* trap messages when a port transitions between up and down.

- To enable or disable one or more ports on a switch, use the following commands:

  `enable port [port_list | all]`

  `disable port [port_list | all]`

  For example, to disable slot 7, ports 3, 5, and 12 through 15 on a modular switch or SummitStack, enter:

  ```
  disable port 7:3,7:5,7:12-7:15
  ```

- To receive these SNMP trap messages, use the following command:

  `enable snmp traps port-up-down ports [port_list | all]`

- To stop receiving these messages, use the following command:

  `disable snmp traps port-up-down ports [port_list | all]`

Refer to Displaying Port Information on page 303 for information on displaying link status.

## Configuring Switch Port Speed and Duplex Setting

**Note**

Refer to Displaying Port Information on page 303 for information on displaying port speed, duplex, autonegotiation, and flow control settings.

ExtremeXOS supports the following port types:

- 10 Gbps ports
- 40 Gbps ports
- 100 Gbps ports
- 10/100/1000 Mbps copper ports
- 10/100/1000 SFPs
- 10/100/1000 Mbps copper ports with *PoE (Power over Ethernet)*—only on the G48Tc, G48Te2, and 8900-G48T-xl with PoE daughter card modules installed in the BlackDiamond 8800 series switch, and the Summit X440-24p, X460-24p, X460-48p, X460G2-24p, and X460G2-48p switches

- 1 Gbps small form factor pluggable (SFP) fiber ports
- 100 FX SFPs, which must have their speed configured to 100 Mbps
- Wide area network (WAN) PHY port—only on the Summit X480 series switches
- 10 Gbps stacking ports (Summit family switches only)
- 10 Gbps small Form Factor pluggable+ (SFP+) fiber ports. These should be configured to 10 Gbps auto off if an SFP+ optic is inserted; they should be configured to 1G auto on (or auto off) if 1G SFP optic is inserted.

Note

Stacking ports always use the same type of connector and copper PHY, which are built in to the Summit family switches. You cannot configure stacking port parameters such as port speed, duplex, and link fault signal. You also cannot configure data port features such as VLANs and link aggregation. Stacking links provide the same type of switch fabric that is provided in a BlackDiamond 8800 series switch or BlackDiamond X8 series switch.

Autonegotiation determines the port speed and duplex setting for each port (except 10 and 40 Gbps ports). You can manually configure the duplex setting and the speed of 10/100/1000 Mbps ports.

The 10/100/1000 Mbps ports can connect to either 10BASE-T, 100BASE-T, or 1000BASE-T networks. By default, the ports autonegotiate port speed. You can also configure each port for a particular speed (either 10 Mbps or 100 Mbps).

Note

With autonegotiation turned off, you cannot set the speed to 1000 Mbps.

In general, SFP gigabit Ethernet ports are statically set to 1 Gbps, and their speed cannot be modified.

However, there are two SFPs supported by Extreme that can have a configured speed:

- 100 FX SFPs, which must have their speed configured to 100 Mbps.
- 100FX/1000LX SFPs, which can be configured at either speed (available only on the BlackDiamond 8800 series switches, the BlackDiamond 12800 series switches, and the Summit family switches).

The 10 Gbps ports always run at full duplex and 10 Gbps. The 40 Gbps ports always run at full duplex and 40 Gbps.

ExtremeXOS allows you to specify the medium as copper or fiber when configuring Summit switches with combination ports. If the medium is not specified for combination ports then the configuration is applied to the current primary medium. The current primary medium is displayed in the Media Primary column of the `show ports configuration` command output.

Note

For switches that do not support half-duplex, the copper switch ports must have auto negotiation disabled and full duplex enabled when connecting 10/100/1000 Mbps devices that do not auto negotiate. If the switch attempts and fails to auto negotiate with its partner, it will fail to link up. A non-negotiating connected device must also be manually configured for full duplex or packet loss and port errors will occur each time it detects a collision.

To configure port speed and duplex setting , use the following command:

```
configure ports port_list {medium [copper | fiber]} auto off speed speed
duplex [half | full]
```

To configure the system to autonegotiate, use the following command:

```
configure ports port_list {medium [copper|fiber]} auto on {[{speed
speed} {duplex [half | full]}] | [{duplex [half | full]} {speed speed}]}
```

Note

The keyword medium is used to select the configuration medium for combination ports. If port_list contains any non-combination ports, the command is rejected.

When upgrading a switch running ExtremeXOS 12.3 or earlier software to ExtremeXOS 12.4 or later, saved configurations from combo ports (copper or fiber) are applied only to combo ports fiber medium. When downgrading from ExtremeXOS 12.4 or later to ExtremeXOS 12.3 or earlier, saved configurations from combo ports (copper or fiber) are silently ignored. Therefore, you need to reconfigure combo ports during such an upgrade or downgrade.

ExtremeXOS does not support turning off autonegotiation on the management port.

*Support for Autonegotiation on Various Ports*

The following table lists the support for autonegotiation, speed, and duplex setting for the various types of ports.

| Port | Autonegotiation | Speed | Duplex |
|---|---|---|---|
| 100 Gbps | Off | 100000 Mbps | Full duplex |
| 10 Gbps | Off | 10000 Mbps | Full duplex |
| 40 Gbps | Off | 40000 Mbps | Full duplex |
| 1 Gbps fiber SFP | On (default)Off | 1000 Mbps | Full duplex |
| 100 FX SFP | On (default)Off | 100 Mbps | Full duplex |
| 10/100/1000 Mbps | On (default)Off | 10 Mbps100 Mbps | Full/half duplexFull/half duplex |
| 10/100 Mbps | On (default)Off | 10 Mbps100 Mbps | Full/half duplexFull/half duplex |
| 10 Gbps SFP+ | Off | 10000 Mbps | Full duplex |

Note

The following products do not support half-duplex operation: Summit X450-G2, X460-G2, BDXA-10G48T, and BDXA-G48T.

Flow control on Gigabit Ethernet ports is enabled or disabled as part of autonegotiation (see IEEE 802.3x). If autonegotiation is set to Off on the ports, flow control is disabled. When autonegotiation is On, flow control is enabled.

With Extreme Networks devices, the 1 Gbps ports and the 10 Gbps ports implement flow control as follows:

- 1 Gbps ports
  - Autonegotiation enabled
    - Advertise support for pause frames
    - Respond to pause frames
  - Autonegotiation disabled
    - Do not advertise support for pause frames
    - Do not respond to pause frames
- 10 Gbps ports for the Summit X460, X460G2, X480, X670, X670G2, and X770 series switches, SummitStack, and on modules for the BlackDiamond X8 series switches and the BlackDiamond 8800 series switch:
  - Autonegotiation always disabled
  - Do not advertise support for pause frames
  - Respond to pause frames

## Configuring Extended Port Description

ExtremeXOS provides a configurable per-port "display-string" parameter that is displayed on each of the `show port` CLI commands, exposed through the *SNMP* ifAlias element, and accessible via the XML port.xsd API. This existing field is enhanced to allow up to 255 characters with much less stringent syntax limitations. Some characters are still not permitted, as they have special meanings. These characters include the following: <, >, ?, &. This new field is accessible through the `show port info detail` command, and is also accessible through the SNMP ifAlias object of IfXTable from IF-MIB (RFC 2233), and the XML API.

You can always configure a 255-character-long string regardless of the configured value of ifAlias size. Its value only affects the SNMP behavior.

Use the following commands to configure up to 255 characters associated with a port:

`config port` *`port_list`* **`description-string`** *`string`*

Use the following command to unconfigure the description-string setting:

`unconfig port` *`port_list`* **`description-string`**

Use the following command to control the accessible string size (default 64, per MIB) for the SNMP ifAlias object:

`config snmp ifmib ifalias size [`**`default`**` | `**`extended`**`]`

If you choose extended size option, the following warning will be displayed:

```
Warning: Changing the size to [extended] requires the use of increased 255 chars long
ifAlias object of ifXtable from IF-MIB(RFC 2233)
```

## Partitioning High Bandwidth Ports

The 40G ports on BlackDiamond X8 switches, BlackDiamond 8900-40G6X-xm modules, and Summit X670, X670G2, and X770 switches can be partitioned into 4x10G and 1x40G modes. The 100G ports on BlackDiamond X8 switches can be partitioned into 10x10G or 1x100G modes.

To partition the ports, enter the following command:

`configure ports [port_list | all] partition [4x10G | 1x40G | 1x100G | 10x10G]`

After you make a configuration change, you must do one of the following to apply the change:

- For BlackDiamond X8 series switches and BlackDiamond 8900-40G6X-xm modules, you can disable and then enable the affected slot, which applies the change without affecting other modules.
- For BlackDiamond X8 series switches, BlackDiamond 8900-40G6X-xm modules and Summit X670, X670G2, and X770 switches you can reboot the switch.

Note

Because of the nature of these ports at the physical layer level, the 10G side may show a remote or local linkup. A configuration change is not applied until the affected slot is disabled and enabled or the switch is rebooted.

## Flow Control

*IEEE 802.3x Flow Control*

With Summit Family Switches, BlackDiamond X8 Series Switches and BlackDiamond 8800 Series Switches only and with autonegotiation enabled, Summit family switches, BlackDiamond X8 switches, and BlackDiamond 8800 series switches advertise the ability to support pause frames.

This includes receiving, reacting to (stopping transmission), and transmitting pause frames. However, the switch does not actually transmit pause frames unless it is configured to do so, as described below.

IEEE 802.3x flow control provides the ability to configure different modes in the default behaviors. Ports can be configured to transmit pause frames when congestion is detected, and the behavior of reacting to received pause frames can be disabled.

**TX**

You can configure ports to transmit link-layer pause frames upon detecting congestion. The goal of IEEE 802.3x is to backpressure the ultimate traffic source to eliminate or significantly reduce the amount of traffic loss through the network. This is also called lossless switching mode.

The following limitations apply to the TX flow control feature:

- Flow control is applied on an ingress port basis, which means that a single stream ingressing a port and destined to a congested port can stop the transmission of other data streams ingressing the same port that are destined to other ports.
- High volume packets destined to the CPU can cause flow control to trigger. This includes protocol packets such as, *EDP (Extreme Discovery Protocol)*, EAPS, *VRRP (Virtual Router Redundancy Protocol)*, and *OSPF (Open Shortest Path First)*.

- When flow control is applied to the fabric ports, there can be a performance limitation. For example, a single 1G port being congested could backpressure a high-speed fabric port and reduce its effective throughput significantly.

To configure a port to allow the transmission of IEEE 802.3x pause frames, use the following command:

```
enable flow-control tx-pause ports port_list|all
```

Note
To enable TX flow-control, RX flow-control must first be enabled. If you attempt to enable TX flow-control with RX flow-control disabled, an error message is displayed.

To configure a port to return to the default behavior of not transmitting pause frames, use the following command:

```
disable flow-control tx-pause ports
```

**RX**

You can configure the switch to disable the default behavior of responding to received pause frames. Disabling rx-pause processing avoids dropping packets in the switch and allows for better overall network performance in some scenarios where protocols such as TCP handle the retransmission of dropped packets by the remote partner.

To configure a port to disable the processing of IEEE 802.3x pause frames, use the following command:

```
disable flow-control rx-pause ports port-list | all
```

Note
To disable RX flow-control, TX flow-control must first be disabled. If you attempt to disable RX flow-control with TX flow-control enabled, an error message is displayed.

To configure a port to return to the default behavior of enabling the processing of pause frames, use the following command:

```
enable flow-control rx-pause ports port-list | all
```

### *IEEE 802.1Qbb Priority Flow Control*

In BlackDiamond X8 Series Switches, BlackDiamond 8900-10G24X-c and 8900-40G6X-xm Modules and Summit X460, X670, and X770 Switches, priority flow control (PFC) as defined in the IEEE 802.1Qbb standard is an extension of IEEE 802.3x flow control, which is discussed in IEEE 802.3x Flow Control on page 185.

When buffer congestion is detected, IEEE 802.3x flow control allows the communicating device to pause all traffic on the port, whereas IEEE 802.1Qbb allows the device to pause just a portion of the traffic while allowing other traffic on the same port to continue.

For PFC, when an ingress port detects congestion, it generates a MAC control packet to the connected partner with an indication of which traffic priority to pause and an associated time for the pause to remain in effect. The recipient of the PFC packet then stops transmission on the priority indicated in the control packet and starts a timer indicating when traffic can resume.

Traffic can resume in two ways:

- On the transmitting side, when the timer expires, traffic on that priority can resume.
- On the receiving side, once congestion is relieved, another PFC packet is generated to un-pause the priority so that traffic can resume.

#### Limitations

The following limitations are associated with this feature:

- In this release, PFC must be explicitly configured by the user.
- In order to support the signaling of congestion across the fabric, an enhanced fabric mode is required. This enhanced mode is not available on some older models of Summits and BlackDiamond 8000 series modules (see the following supported platforms section). Also, this enhanced mode reduces the effective bandwidth on the fabric by a small amount (less than 5%). The BlackDiamond 8900-10G24X-c becomes slightly more blocking and the BlackDiamond 8900-10G8X-xl card is no longer non-blocking when this enhanced mode is configured.
- The fabric flow control packets take up some small amount of bandwidth on the fabric ports.
- On Summit X670 and X670V switches, the PFC feature does not support fabric flow control messages on alternate stack ports or SummitStack-V80 native stack ports.

#### Supported Platforms

PFC is currently supported only on 10G ports and on specific models of the following newer platforms indicated by the part number:

- BlackDiamond X8 series switches
- BlackDiamond 8900-10G24X-c modules (manufacturing number 800397-00)
- BlackDiamond 8900-40G6X-xm modules, 40G ports and 10G ports when in 4x10 partition mode
- Summit X450-G2 switches
- Summit X460 and X460-G2 switches, 10/40G ports
- Summit X670 switches, 10G ports
- Summit X670V and X670-G2 switches, 10G and 40G ports
- Summit X770-32q, 10G and 40G ports

To verify that your switch or module supports PFC, use the `show version` command. If you attempt to enable PFC on unsupported ports, an error message is displayed. (See Abnormal Configuration Examples on page 190.)

#### Setting the Priorities

Priority is established for reception of PFC packets with a *QoS (Quality of Service)* profile value on the ExtremeXOS switch and for transmission with a priority value added to the PFC packet.

- QoS profile—Ingress traffic is associated with a QoS profile for assignment to one of eight hardware queues in the system that define how the traffic flows with respect to bandwidth, priority, and other parameters. By default, there are two QoS profiles (QP1 and QP8) defined in these supported platforms and PFC works with this default. To segregate the ingress traffic with more granularity, you will want to define other QoS profiles. The traffic that will be paused on reception of the PFC packet is associated with the hardware queue of the QoS profile that you specify.

  The QoS profile is also used to configure the fabric ports.

- Priority—When the hardware transmits a PFC packet, it uses the priority bits in the *VLAN* header on the ingress packet to determine the priority to pause, if the ingress packet is tagged. You can specify this transmit priority independently from the QoS profile to associate it with the reception of a PFC packet thus giving flexibility in the configuration of the network. For untagged ingress packets, the hardware queue determines the priority in the transmitted PFC packet. (For additional information, see QoS Profiles on page 737.

It is suggested that the priority in the VLAN header match the QoS profile priority when traffic ingresses at the edge of the network so that the traffic can be more easily controlled as it traverses through the network.

### *Fabric Port Configuration*

This feature also configures the fabric between ingress and egress ports to optimize PFC behavior.

When the ingress and egress ports are located on separate BlackDiamond I/O modules or different nodes in a SummitStack, note that some older systems do not support the enhanced fabric mode required for PFC. The following applies:

- For BlackDiamond 8800 switches, the BlackDiamond 8900-MSM128 is needed. If other MSMs are installed, a log message is issued indicating that system performance for PFC will not be optimal.
- In a SummitStack, PFC cannot be enabled until the following command is executed:

  ```
  configure stacking protocol enhanced
  ```

The fabric can be set up to support the flow control messages only in the following switches:

- Summit X450-G2
- Summit X460
- Summit X460-G2
- Summit X480
- Summit X670
- Summit X670-G2
- Summit X770

If any other Summit switch attempts to join the stack after the initial configuration of PFC, it is not allowed to join.

If your situation does not respond well to having flow control enabled on the fabric links, you can turn off flow control in the fabric by using the following command:

```
configure forwarding flow-control fabric [auto | off]
```

### *Configuring Priority Flow Control*

With PFC, it is expected that both RX and TX be enabled or disabled.

- To enable PFC, use the following command:

  ```
  enable flow-control [tx-pause {priority priority} | rx-pause
  {qosprofile qosprofile}] ports [all | port_list]
  ```

- To disable PFC, use the following command:

  ```
  disable flow-control [tx-pause {priority priority} | rx-pause
  {qosprofile qosprofile}] ports [all | port_list]
  ```

### Example

The network needs to transport FCoE (Fiber Channel over Ethernet) traffic which is intermixed with other more typical LAN traffic on the same Ethernet network. FCoE needs a lossless transport and PFC can be used to enable this. You define *QoS* profiles for all eight traffic priorities. At the network level, it is decided that FCoE traffic will be assigned to priority 3 (which corresponds to QP4) and the remaining traffic is assigned to one or more other priorities. For this example, it is also assumed that the priority bits in the incoming packets are also 3.

One mechanism that can be used for this classification is the use of Access Control Lists (ACLs) that match on the FCoE ethertypes (0x8906 and 0x8914) using the ethernet-type qualifier with an action of QoS profile QP4 for both rules. Other traffic can be classified to other priorities. Once this configuration is applied, FCoE is now separated from the other Ethernet traffic and is assigned a priority of 3 through the switch.

PFC is enabled at the ports that will see FCoE traffic, in this case, ports 1:1, 2:3, and 6:5.

Since FCoE is assigned to QP4, you would enable the receive PFC for QoS profile to be QP4 and, in this example, would also enable PFC with a transmit priority of 3. The enable commands would then read as follows:

```
enable flow-control tx-pause priority 3 ports 1:1,2:3,6:5
enable flow-control rx-pause qosprofile qp4 ports 1:1,2:3,6:5
```

The `show port flow-control rx-pause no-refresh` CLI command displays the following information. With the no-refresh option the display shows the following information:

```
X770-32Q-J4-U7.85 # show ports 1,5,9 flow-control rx-pauses no-refresh
Flow Control Frames Received
Port   Pause   PFC0   PFC1   PFC2   PFC3   PFC4   PFC5   PFC6   PFC7
        Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs
================================================================================
1          -      -      -      - 1234567 1234567    - 1234567     -
5          -      -      -      - 1234567 1234567    - 1234567     -
16:104 1234567    -      -      -       -       -    -       -     -
================================================================================
">" Name truncated, "-" rx-pause not enabled, "." Counter not available
```

Without the no-refresh option the display refreshes until interrupted by the console operator pressing `Escape`:

```
X770-32Q-J4-U7.85 # show ports 1,5,9 flow-control rx-pauses
Flow Control Frame Monitor                    Sat Aug 18 19:35:12 2012
Port   Pause   PFC0   PFC1   PFC2   PFC3   PFC4   PFC5   PFC6   PFC7
        Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs   Rcvs
================================================================================
1          -      -      -      - 1234567 1234567    - 1234567     -
5          -      -      -      - 1234567 1234567    - 1234567     -
16:104 1234567    -      -      -       -       -    -       -     -
================================================================================
">" Name truncated, "-" rx-pause not enabled, "." Counter not available
Spacebar->Toggle screen 0->Clear counters U->Pageup D->Pagedown ESC->exit
```

The new `show port flow-control tx-pause no-refresh` CLI command displays the following information. With the no-refresh option the display shows the following information:

```
X770-32Q-J4-U7.85 # show ports 1,5,9 flow-control tx-pauses no-refresh
Flow Control Frames Transmitted
Port   Pause   PFC0   PFC1   PFC2   PFC3   PFC4   PFC5   PFC6   PFC7
       Xmts    Xmts   Xmts   Xmts   Xmts   Xmts   Xmts   Xmts   Xmts
================================================================================
1         -       -      -      - 1234567 1234567    - 1234567    -
5         -       -      -      - 1234567 1234567    - 1234567    -
16:104 1234567    -      -      -       -       -    -       -    -
================================================================================
">" Name truncated, "-" tx-pause not enabled, "." Counter not available
```

Without the no-refresh option the display refreshes until interrupted by the console operator pressing the escape key:

```
X770-32Q-J4-U7.85 # show ports 1,5,9 flow-control tx-pauses
Flow Control Frame Monitor Sat Aug 18 19:35:12 2012
Port   Pause   PFC0   PFC1   PFC2   PFC3   PFC4   PFC5   PFC6   PFC7
        Xmts   Xmts   Xmts   Xmts   Xmts   Xmts   Xmts   Xmts   Xmts
================================================================================
1          -      -      -      -  1234567 1234567   -  1234567    -
5          -      -      -      -  1234567 1234567   -  1234567    -
16:104 1234567    -      -      -       -       -   -       -    -
================================================================================
">" Name truncated, "-" tx-pause not enabled, "." Counter not available
Spacebar->Toggle screen 0->Clear counters U->Pageup D->Pagedown ESC->exit
```

Once this configuration is complete, if a switch ingress port detects congestion, it will send PFC packets to the remote link partner and will respond to PFC packets from the remote link partner by stopping transmit.

### *Abnormal Configuration Examples*

Examples of abnormal configuration scenarios.

- If you attempt to configure PFC on a port that does not support it, an error message similar to the following is issued and you will be informed that PFC cannot be configured on that port:

  ```
  BD8810.1# enable flow-control tx-pause priority 3 port 1:1
  Error: Port 1:1 does not support Priority Flow Control.
  ```

- If you attempt to configure PFC on a port in a system that has older MSM models, the PFC configuration will succeed as long as the user port supports it, but a log message will be issued indicating that overall PFC operation is not optimal.

  ```
  01/22/2010 14:14:37.88 <Warn:HAL.VLAN.PFCSubopt> MSM-A: Priority Flow Control is
  enabled but system behavior will not be optimal. Older modules in the system cannot be
  programmed for fabric flow control.
  ```

- When PFC is enabled on a port, IEEE 802.3x will be disabled. If, after enabling PFC, you try to modify RX or TX pause parameters, an error message similar to the following will be issued explaining the dependency on PFC configuration:

  ```
  BD8810.1# enable flow-control tx-pause port 1:1
  Error: Priority Flow Control is currently enabled on port 1:1 and is mutually
  exclusive with TX and RX pause configuration. TX and RX pause configuration cannot be
  done until PFC is disabled on this port.
  ```

- When PFC configuration is attempted on older versions of BlackDiamond 8900-10G24X-c modules that do not support PFC, as described in the following conditions, the switch will attempt the configuration.
  - If you try to configure PFC on older BlackDiamond 8900-10G24X-c modules that do not support PFC.
  - If a BlackDiamond 8900-10G24X-c module that supports PFC is replaced with a version that does not support PFC.
  - If a slot is preconfigured as an 8900-10G24X-c module, PFC is configured, and a version of the module that does not support PFC is inserted.

  Under any of these conditions, the scenario is flagged and the following log message is issued to alert you to the misconfiguration:

  ```
  01/22/2010 14:14:37.88 <Warn:HAL.VLAN.PFCUnsuprt> MSM-A: Port 4:1 is on an older model
  of the 8900-10G24X-c and does not support Priority Flow Control. 8900-10G24X-c 41632B,
  and VIM-10G8X 17012B are new models that support PFC.
  ```

- If you try to configure PFC on a port in a SummitStack before you have configured the SummitStack for enhanced mode, the following error message is issued:

  ```
  Slot-1 Stack.7 # enable flow-control rx-pause qosprofile qp1 port 1:1
  Error: The stack is not configured for enhanced stacking mode. Issue the command
  "configure stacking protocol enhanced" to enable this mode and retry the PFC
  configuration.
  ```

- On Summit X670 and X670V switches, if you try to configure PFC on alternate stack ports or SummitStack-V80 native stack ports, the following error message is issued:

  ```
  07/18/2011 10:42:07.60 <Warn:HAL.Port.FabFlowCtrlUnsuprt> Slot-1: Slot 3 does not
  support fabric flow control messages on alternate stack ports or V80 native stack ports.
  ```

### *Turning off Autonegotiation on a Gigabit Ethernet Port*

In certain interoperability situations, you need to turn autonegotiation off on a fiber gigabit Ethernet port. Although a gigabit Ethernet port runs only at full duplex, you must specify the duplex setting. The 10 Gbps ports do not autonegotiate; they always run at full duplex and 10 Gbps speed.

The following example turns autonegotiation off for port 1 (a 1 Gbps Ethernet port) on a module located in slot 1 of a modular switch:

```
configure ports 1:1 auto off speed 1000 duplex full
```

### *Running Link Fault Signal*

The 10 Gbps ports support the Link Fault Signal (LFS) function. This function, which is always enabled, monitors the 10 Gbps ports and indicates either a remote fault or a local fault. The system then stops transmitting or receiving traffic from that link. After the fault has been alleviated, the system puts the link back up and the traffic automatically resumes.

The Extreme Networks implementation of LFS conforms to the IEEE standard 802.3ae-2002.

Note

To display the part number of the module, use the `show slot` *`slot_number`* command. (All the modules on the BlackDiamond 8800 series switch support LFS.)

Although the physical link remains up, all Layer 2 and above traffic stops.

The system sends LinkDown and LinkUp traps when these events occur. Additionally, the system writes one or more information messages to the syslog, as shown in the following example for a BlackDiamond 8800 series switch:

```
09/09/2004 14:59:08.03 <Info:vlan.dbg.info> MSM-A: Port 4:3 link up at
10 Gbps speed and full-duplex
09/09/2004 14:59:08.02 <Info:hal.sys.info> MSM-A: 4:3 - remote fault recovered.
09/09/2004 14:59:05.56 <Info:vlan.dbg.info> MSM-A: Port 4:3 link down
due to remote fault
09/09/2004 14:59:05.56 <Info:hal.sys.info> MSM-A: 4:3 - remote fault.
09/09/2004 15:14:12.22 <Info:hal.sys.info> MSM-A: 4:3 - local fault
recovered.
09/09/2004 15:14:11.35 <Info:vlan.dbg.info> MSM-A: Port 4:3 link up at
10 Gbps speed and full-duplex
09/09/2004 15:13:33.56 <Info:vlan.dbg.info> MSM-A: Port 4:3 link down
due to local fault
09/09/2004 15:13:33.56 <Info:hal.sys.info> MSM-A: 4:3 - local fault.
09/09/2004 15:13:33.49 <Info:vlan.dbg.info> MSM-A: Port 4:3 link down
due to local fault
```

In Summit series switches, on disabling the 10 Gbps ports, the following message is logged to the syslog:

```
08/26/2008 06:05:29.29 Port 1 link down - Local fault
```

Note

A link down or up event may trigger Spanning Tree Protocol topology changes or transitions.

*Turn off Autopolarity*

The autopolarity feature allows the system to detect and respond to the Ethernet cable type (straight-through or crossover cable) used to make the connection to the switch port or an endstation. Summit Family Switches, SummitStack, and BlackDiamond 8800 Series Switches only.

This feature applies only to the 10/100/1000 BASE-T ports on the switch and copper medium on Summit combination ports.

When the autopolarity feature is enabled, the system causes the Ethernet link to come up regardless of the cable type connected to the port. When the autopolarity feature is disabled, you need a crossover cable to connect other networking equipment and a straight-through cable to connect to endstations. The autopolarity feature is enabled by default.

Under certain conditions, you might opt to turn autopolarity off on one or more ports.

- To disable or enable autopolarity detection, use the following command:

  `configure ports` *`port_list`* **`auto-polarity`** [**`off`** | **`on`**]

  Where the following is true:

  - port_list—Specifies one or more ports on the switch
  - off—Disables the autopolarity detection feature on the specified ports
  - on—Enables the autopolarity detection feature on the specified ports

  The following example turns autopolarity off for ports 5 to 7 on a Summit family switch:

  ```
  configure ports 5-7 auto-polarity off
  ```

- When autopolarity is disabled on one or more Ethernet ports, you can verify that status using the command:

  ```
  show ports information detail
  ```

## IPFIX

*IP Flow Information Export protocol*

For BlackDiamond 8900 G96Tc, G48T-xl, G48X-xl, and 10G8X-xl Modules, BlackDiamond X8 100G4X modules, Summit X460, X460-G2 and X480 switches, the IP Flow Information Export (IPFIX) protocol (created by the IETF) is a standard way to capture information about traffic flows passing through network elements in a data network.

The protocol consists of a metering process, an exporting process, and a collecting process. This section discusses the metering and exporting processes; the collecting process is not defined by the standard and therefore is outside the scope of this document. The IPFIX protocol is a rival, but complementary, protocol to sFlow.

The Extreme Networks switch contains various metering processes that gather information about flows through different ports, or observation points, on the switch. This information includes: the ingress and egress interfaces, the link state, IPFIX state, flow count, byte count, packet count, flow record count and premature exports. The metering process then sends the information to the exporting process in the switch which handles communication, using TCP, UDP, or SCTP transport protocols, over the network to a collecting process.

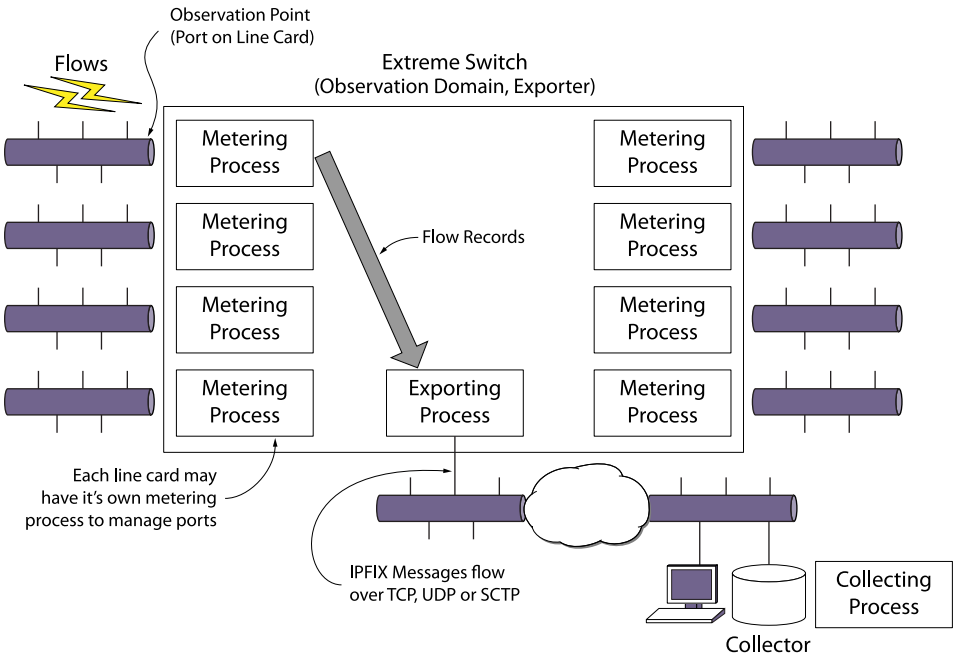



**Figure 12: IPFIX Processes**

**Limitations**

This feature has the following limitations:

- The flow key definition is limited to the L2, L3, and L4 header fields the hardware provides.
- There is a 8K flow limit per port—4K for ingress and 4K for egress for platforms X460-48t/x/p and X480.
- For other Summit platforms (such as the E4G-400, X460-24t/x/p, and X460-G2), the limit is 4K flows per port—2K for ingress and 2K for egress.
- For BDX8 100G4X, the limit is 4K flows per port—2K for ingress and 2K for egress.

*Enabling IPFIX*

- To enable IPFIX on a port and provide a check to ensure that the port being enabled has hardware support for IPFIX, use the following command:

  `enable ip-fix ports [`*`port_list`* `|` **`all`**`] {`**`ipv4`** `|` **`ipv6`** `|` **`non-ip`** `|` **`all_traffic`**`}`

  If the port does not support IPFIX, an error message is displayed.
- To disable an enabled port, use the following command:

  `disable ip-fix ports [`*`port_list`* `|` **`all`**`]`
- To enable or disable IPFIX globally and override an individual port enable, use the following command:

  `[`**`enable`** `|` **`disable`**`]` **`ip-fix`**

*Configuring IPFIX Flow Key Masks*

Flow keys define what data in the packet header identifies a unique flow to the hardware. On each port, there is a flow key for IPv4, IPv6, and non-IP traffic type data. Following are the flow keys together with the size of the field:

IPv4:

- Source IP Address (32)
- Destination IP Address (32)
- L4 Source Port (16)
- L4 Destination Port (16)
- L4 Protocol (8)
- TOS (DSCP +ECN) (8)

IPv6:

- Source IP Address (128)
- Destination IP Address (128)
- L4 Source Port (16)
- L4 Destination Port (16)
- Next Header (8)
- IPv6 Flow Label (20)
- TOS (DSCP +ECN) (8)

Non-IP:

- Source MAC Address (48)
- Destination MAC Address (48)
- *VLAN* ID (12)
- VLAN Priority (3)
- Ethertype (16)
- VLAN Tagged (1)

By default, IPFIX uses all the above listed flow keys and all bits. You can override this on a global basis and specify exactly which keys to use. The template that specifies the structure of the information that is communicated from the exporter to the collector will then contain only those specified keys.

To specify the flow keys to use for each of the three traffic types, use the following commands:

```
configure ip-fix flow-key ipv4 {src-ip} {src-port} {dest-ip} {dest-port}
{protocol} {tos}
```

```
configure ip-fix flow-key ipv6 {src-ip} {src-port} {dest-ip} {dest-port}
{next-hdr} {tos} {flow-label}
```

```
configure ip-fix flow-key nonip {src-mac} {dest-mac} {ethertype} {vlan-
id} {priority} {tagged}
```

To reset to the all keys default, use the following command:

```
unconfigure ip-fix flow-key
```

You can then define masks for the IPv4 and IPv6 source and destination address fields on a per port basis.

Use the following commands: `configure ip-fix ports` *port_list* **flow-key ipv6 mask** [**source** | **destination**] **ipaddress** *value*

**Example**

You can use the flow keys and masks to minimize the information sent to the collector and aggregate certain types of flows.

A common use of the non-default values may be to see all traffic from a user only instead of each individual flow. For example, in the case of IPv4:

```
configure ip-fix flow-key ipv4 src-ip dest-ip
```

Then, by configuring the mask on a port, the aggregation could be further restricted to meter only individual subnets.

For example, with a 255.255.255.0 mask:

```
configure ip-fix ports 3:1 flow-key ipv4 mask source ipaddress 255.255.255.0
configure ip-fix ports 3:1 flow-key ipv4 mask destination ipaddress 255.255.255.0
```

To unconfigure the masks, use the following command:

```
unconfigure ip-fix ports port_list flow-key mask
```

*Configuring IPFIX Parameters on a Port*

These are optional commands; when not configured, the defaults are used.

- To configure whether to meter on ingress and/or egress ports, use the following command:

  ```
  configure ip-fix ports port_list [ingress | egress | ingress-and-egress]
  ```

  (The default is ingress.)
- To configure whether to meter all, dropped only, or non-dropped only records, use the following command:

  ```
  configure ip-fix ports port_list record [all | dropped-only | non-dropped]
  ```

  (The default is all.)
- To unconfigure these IPFIX settings on a port or group of ports, use the following command:

  ```
  unconfigure ip-fix ports port_list
  ```

  This restores the configuration to the defaults for those ports. It does not enable or disable IPFIX.

*Configuring Domain IDs*

Observation points are aggregated into observation domains. The entire switch operates as one domain. The IPFIX protocol contains an observation domain ID in the flow records that are sent to the collector. The collector can use the domain to correlate records to their origin. How this field is used is up to a given collector.

- To configure a domain ID, use the following command:

  ```
  configure ip-fix domain domain_id
  ```

*Configuring a Collector*

To export flow records using the IPFIX protocol, you must first configure a collector. Only a single collector is allowed. You can specify the source IP address and VR to use when sending from the switch to a given collector. When not specified, the system defaults to the switch IP address the traffic exits.

- To specify the source IP address to be used in IPFIX packets, use the following command:

  ```
  configure ip-fix source ip-address ipaddress {vr vrname}
  ```
- To reset back to the default of using the switch IP, use the following command:

  ```
  unconfigure ip-fix source ip-address
  ```
- To specify the IP address, port number, transport protocol and VR for a collector, use the following command:

  ```
  configure ip-fix ip-address ipaddress {protocol [sctp | tcp | udp]} {L4-port portno} {vr vrname}
  ```
- To unconfigure this, use the following command:

  ```
  unconfigure ip-fix ip-address
  ```

*Unconfiguring IPFIX*

- To unconfigure IPFIX completely, use the following command. This removes all port and collector configuration and disables all IPFIX ports.

  ```
  unconfigure ip-fix
  ```

*Displaying IPFIX Information*

You can view information about IPFIX information on ports.

- To display the global state, the collector information and the ports that are enabled for IPFIX, use the following command:

  ```
  show ip-fix
  ```

- To display information about per port metering, use the following command without the **tag** option:

  ```
  show ports {port_list } ip-fix {no-refresh | port-number |refresh}
  ```

- To show whether IPFIX is enabled on a specific port together with port IPFIX configuration, use the following command without the **mgmt** or **tag** options:

  ```
  show port {mgmt | port_list | tag tag} information {detail}
  ```

## WAN PHY OAM

*Summit X480 Series Switches Only*

You can configure WAN PHY OAM on the Summit X480 series switches whether or not they are included in a SummitStack.

The WAN-PHY OAM feature is a subset of the SONET/SDH overhead function and the WAN PHY interface is defined in IEEE 802.3ae.

Summit X480 series switches are WAN-PHY capable on 10G XFP ports. XFP ports can operate in both LAN and WAN modes. For such ports, the WAN PHY configuration commands that are shown in the following section, are available only after setting the ports to "WAN PHY" mode using the command:

```
configure ports port_list mode {lan | wan-phy}
```

*Configuring WAN PHY OAM Parameters*

The following are configurable WAN PHY OAM parameters.

- Framing—either SONET or SDH; default is SONET.
- Clock source—either internal or line; default is line.
- J0 section trace string—16-character string; default is the IEEE default value, which has no string representation.
- J1 path trace string—16-character string; default is the IEEE default value, which has no string representation.
- Loopback—line, internal, or off; the default is off
- To set the framing, use the following command:

  ```
  configure ports port_list wan-phy framing [sonet | sdh]
  ```

- To choose the clock source, use the following command:

  ```
  configure ports port_list wan-phy clocking [line | internal]
  ```

- To set a section trace ID, use the following command:

  `configure ports` *port_list* **wan-phy trace-section** *id_string*
- To set a path trace ID, use the following command:

  `configure ports` *port_list* **wan-phy trace-path** *id_string*
- To set a WAN PHY port to loopback, use the following command:

  On Summit X480 series switches:

  `configure ports` *port_list* **wan-phy loopback** {**off** | **internal** | **line**}
- To reset the configuration parameters of a WAN PHY port to default values, use the following command:

  `unconfigure ports` [*port_list* | **all**] **wan-phy**

*Displaying WAN PHY OAM Information*

- To display information on the WAN PHY ports, use the following commands:

  `show port` {**mgmt** | *port_list* | **tag** *tag*} **information** {**detail**}

  `show ports` {*port_list* | **tag** *tag*} **wan-phy configuration**

  `show ports` {*port_list* | **tag** *tag*} **wan-phy errors** {**no-refresh**}

  `show ports` {*port_list* | **tag** *tag*} **wan-phy events** {**no-refresh**}

  `show ports` {*port_list* | **tag** *tag*} **wan-phy overhead** {**no-refresh**}

## Configuring Switching Mode--Cut-through Switching

*Default Switching Mode*

All platforms use store-and-forward by default. The platforms are also capable of supporting cut-through forwarding to reduce latency. Store-and-forward switching requires the complete receipt of a packet prior to transmitting it out the interface. The packet is stored in its entirety in packet memory and can be validated via the frame CRC by the switch prior to forwarding it to the next hop.

On the BlackDiamond 8900 series modules, you can configure the switch to a cut-through switching mode. Cut-through switching allows the switch to begin transmitting a packet before its entire contents have been received thereby reducing the overall forwarding latency for large packet sizes.

Of the BlackDiamond 8900 series modules, only the 8900-10G24X-c and 8900-MSM128 fully support cut-through switching mode. The BlackDiamond 8900-G96T-c has partial support; it can operate only the switching fabric in cut-through mode.

For the Summit X770, both 40G and 10G ports support store-and-forward switching mode. On the X770, cut-through switching mode is only supported on 40G ports, and is not supported on 10G ports.

Summit X670-G2, BDXB-100G4X, and BDXB-100G4X-XL also support cut-through switching.

The following limitations apply to the cut-through switching feature:

- Error packets may be forwarded when using cut-through mode. These packets need to be detected and discarded by one of the downstream switches, routers, or the ultimate end station.

In some circumstances, store-and-forward is automatically used. Following are examples:

- Cut-through mode cannot be achieved when switching a packet internally from a low-speed front-panel port (1G or 10G) to a higher-speed fabric port. In this case, store-and-forward switching will automatically be used. However, cut-through switching can be used when switching between equal speed ports or from a higher-speed interface to a lower-speed interface.
- Store-and-forward is used for packets that are switched to multiple egress ports in scenarios such as *VLAN* flooding and multicast.
- Store-and-forward is used whenever the egress interface is congested including when *QoS* rate shaping is in effect.

### *Configuring Switching Mode*

You can change or view the default switching mode.

- To configure the switching mode, use the following command:

  `configure forwarding switching-mode [cut-through | store-and-forward]`

- To display the switching mode settings, use the following command:

  `show forwarding configuration`

## SyncE

**Note**
SyncE is supported only on X460G2 platforms and cell site routers.

### *E4G-200 and E4G-400 Cell Site Routers*

Synchronous Ethernet (SyncE) is defined in ITU-T recommendations G.8262/G.8264.

This feature provides the capability for the hardware to synchronize the clock time that is used for data transmission to a reference clock. This primary reference clock (PRC) comes from a base station controller (BSC).

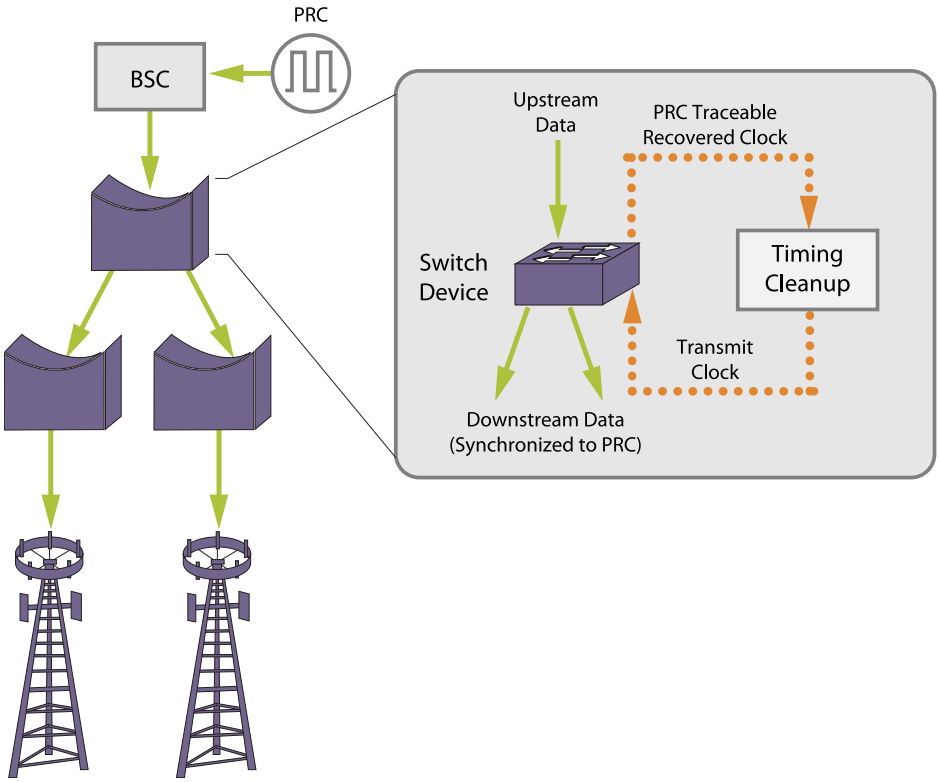



**Figure 13: SyncE Structure**

On the switch, one port is configured to be the source for the master interface clock. A second port can be configured to be the source for a backup reference clock should the master be disconnected or fail. Up to two ports can be specified as a clock source. Data transmission for all other ports are synchronized to the master interface clock. If the master port fails, clock accuracy is maintained. When the ExtremeXOS software detects the failure, it enables the secondary port for the clock. If, at any time, the master port comes back up, it again becomes the source of the primary clock still with accuracy maintained.

It is not necessary for data from the clock master or backup ports to be sent over the other interfaces to maintain synchronization. Only the transmission timing is affected.

The Ethernet Synchronization Messaging Channel (ESMC) is defined by ITU-T for synchronous Ethernet links. ESMC PDUs guide hardware to pick primary clock source and send ESMC messages downstream with clock accuracy details for systems to synchronize.

### *Limitations and Requirements*

SyncE is supported on 100 Mbps / 1 Gbps ports, and it is also available on E4G-400 XGM 10G Ethernet ports, if present.

For synchronous Ethernet (SyncE), the following ports are supported on each platform:

- E4G-200: All Ethernet ports
- E4G-400: All Ethernet ports including XGMS 10G ports if present

### *Clocking Subsystem Selection for E4G-200 and E4G-400*

The E4G-200 and E4G-400 have clock sources beyond SyncE. The clock that drives all of the ports on a switch may be selected from:

- SyncE
- PTP: An optional 1588v2 module
- TDM: An optional module which has multiple T1/E1 interfaces for TDM/Ethernet interworking
- BITS: Building Integrated Timing Supply. A connector capable of receiving a timing signal provided by other building equipment

### *SyncE for E4G Stacking*

The network timing clock can be distributed across different nodes in a stack using 10G alternate stacking links.

Clock distribution on a stack requires a specific configuration:

- All nodes in a stack must be SyncE capable.
- All nodes in a stack must support SyncE on stacking links.

Currently, only E4G-400 with an XGM3S card in slot A is capable of supporting SyncE for stacking.

The E4G-400 can use any stacking module used by the X460 series. However, the native stacking modules cannot carry network timing signals throughout the stack. Only the XGM3S plugin modules have that capability. If clock distribution is desired in an E4G-400 stack, alternate stacking must be used with an XGM3S module in slot A.

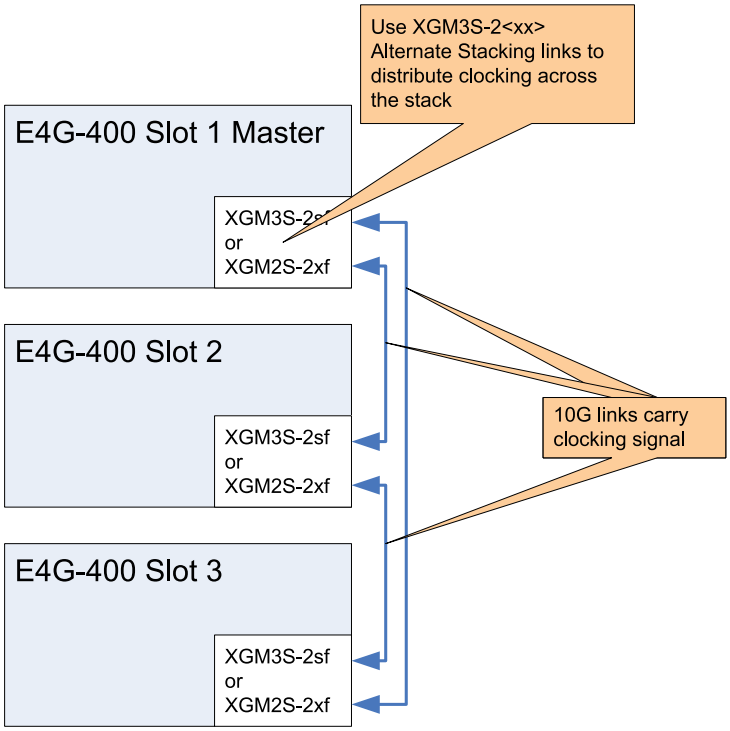



**Figure 14: E4G-400 Stack Clocking**

**SyncE for E4G Stacking Limitations**

- Currently SyncE is only supported on stacks with EG4-400 with XGM3S-2SF or XGM3-2XF cards in Slot A configured as alternate stacking.
- SyncE CLI commands are only available if all nodes in the stack have stacking ports capable of SyncE distribution.
- If SyncE is configured on stackng and a new node not capable of SyncE is added to the stack, an error message will be logged as not capable and the node will be allowed to join. This will break the SyncE, so the user must be careful when adding a new node into the SyncE stack.

*Configuring SyncE*

A link flap occurs in the following scenarios:

- Link is configured as clock source via the command configure network-clock sync-e clock-source source-1/source-2.
- Link is unconfigured for clock source via the command:

  `unconfigure network-clock sync-e ports` *port*.
- When a valid input clock is selected via the port configured as clock source.
- When a valid input clock becomes unavailable via the port configured as clock source.

- To enable SyncE on ports, use the following command:

  `enable network-clock sync-e` **`port`** [*port_list* | **`all`**]
- To disable SyncE on ports, use the following command:

  `disable network-clock sync-e` **`port`** [*port_list* | **`all`**]
- To configure SyncE on ports, use the following command:

  `configure network-clock sync-e` [**`source-1`** | **`source-2`**] **`port`** *port*
- To unconfigure SyncE on ports, use the following command:

  `unconfigure network-clock sync-e` [**`port`** *port*]
- To display the configuration and port state, use the following command:

  `show network-clock sync-e` **`port`** [*port_list*] {**`details`**}
- To display SyncE as part of the port configuration, use the following command:

  `show port` {**`mgmt`** | *port_list* | **`tag`** *tag*} **`information`** {**`detail`**}
- To configure the input network clock source, use the following command:

  `configure network-clock clock-source input` {[**`sync-e`** | **`ptp`** | **`tdm`** | [**`bits-rj45`** | **`bits-bnc`**] {**`quality-level`** *value*}] | **`region`** [**`E1`** | **`T1`**]}
- To configure the output network clock source, use the following command:

  `configure network-clock clock-source output` {**`bits-bnc-1`** [**`1pps`** | **`8KHz`**] **`bits-bnc-2`** [**`E1`** | **`T1`** | **`10MHz`**]}
- To display the configured network clock source information, use the following command:

  `show network-clock` **`clock-source`**

## TDM PWE and TDM Timing

*Introduction*

Time-Division Multiplexed circuits can be transported through pseudowires (TDM PWE) using tunnels based on Ethernet, IP/UDP or *MPLS (Multiprotocol Label Switching)*. A pseudowire is an emulation of point-to-point circuit over a Packet Switching Network (typically Ethernet). It emulates the operation of a “transparent wire” carrying the service. This method of transporting TDM circuits over a Packet Switching Network is also known as Circuit Emulation Service (CES).

This feature is available only on the E4G-400 and E4G-200 cell site routers.

- Ethernet Pseudowires: When the service being carried over the “wire” is Ethernet, it is referred as Ethernet Pseudo-wires. L2VPN is an example of Ethernet Pseudowires.
- TDM Pseudowires: When the service being carried over the “wire” is TDM, it is referred as TDM Pseudowires.
- Cell Site: This is the Radio Access Customer Network Edge, and refers to that part of the Mobile network that includes 2G (T1/E1 Connectivity), 3G and 4G radio towers.
- Cell Site Router: The Cell-Site Router backhauls the traffic from the radio towers over the Ethernet network. Several 2G, 3G, and 4G radio towers can be connected to the Ethernet mobile backhaul at the same time through the Cell Site Router.
- Cell Site Aggregation Router: This router aggregates multiple Ethernet links from various Cell Site Routers as well as the T1/E1 links (that are co-located with it), and transports them over the Mobile core. It is likely that Cell Site Aggregation routers are connected to each other through multiple synchronous Gigabit Ethernet rings.
- Base Station Control: Terminates TDM pseudowires and hand-off cell site (TDM/ATM) traffic to BSC/RNC devices.

The figure below shows the components involved in supporting TDM pseudowires.

TDM pseudowires can be realized as structure-agnostic transport, or SAToP (RFC4553), and structure-aware transport (RFC5086) of TDM circuits. The components involved in both the types of pseudowires are mostly similar. In the following figure, PW#1 and PW#2 are realized using structure-aware transport of TDM circuits, while PW#3 is realized using structure-agnostic transport of TDM circuit.

- SAToP: This is a pseudowire encapsulation of TDM bit streams (T1/E1) without any cognizance of the structure of the TDM bit-streams. The entire frame received over the T1/E1 port is treated as data and sent over the pseudowire. This method has the following advantages:
  - Low overhead.
  - Lower end-to-end delay.
- CESoPSN: In this method, there is a structure awareness of the TDM bit streams (signals), meaning the data that is encapsulated is NXDS0. This method has the benefit of lower packetization delay when transporting several timeslots. CESoP supports channel-associated signaling (CAS) for TDM interfaces.

*Packet Encapsulation Formats Supported by ExtremeXOS*

The packet encapsulation formats of the different pseudowire transports supported by ExtremeXOS are shown below:

- MEF-8 (Ethernet) based encapsulation

  Note
  The Ethertype used for MEF-8 encapsulation is 0x88D8.

- IP/UDP-based encapsulation (RFC 4553 and RFC 5086)
- MPLS-based encapsulation (RFC 4385 and RFC 5287)

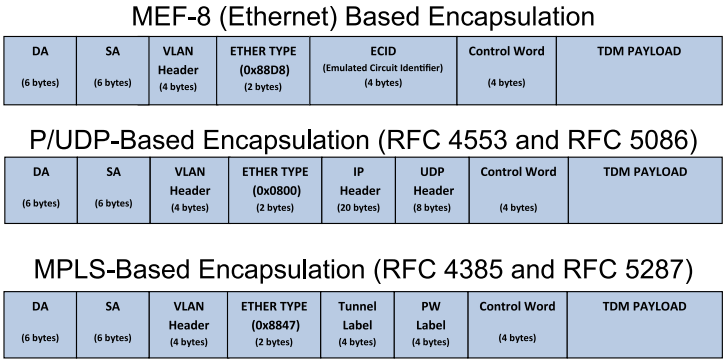



**Figure 15: Packet Encapsulation Formats Supported by ExtremeXOS**

*Configuring TDM Hierarchy*

The switch boots up in the E1 hierarchy by default (the TDM ports are configured to operate in E1 mode). For T1 mode of operation, the hierarchy must be configured, followed by the save and reboot of the switch. After reboot, the switch boots up in T1 hierarchy based on the configuration saved and the TDM ports operate in T1 mode.

Other TDM configurations can be performed after setting up the switch in the correct hierarchy.

Note
For a TDM line where TDM services and/or CES pseudowires have been configured, and the hierarchy need to be changed, we recommend that you first remove or reset all of the CES pseudowires, TDM services, and TDM line configurations before you configure the TDM hierarchy.

- To configure the TDM hierarchy for the switch (T1 or E1), use the following command:

  ```
  configure tdm hierarchy [t1 | e1]
  ```

*Understanding TDM Ports Numbering*

ExtremeXOS supports 16 TDM ports on E4G-200 and E4G-400 cell site routers.

The TDM ports are numbered from 1 to 16 in the face-plate of the switch. However, when the TDM ports are configured using the ExtremeXOS CLI, the TDM ports are numbered sequentially after the Ethernet ports. The following table shows the port number mapping in E4G-200 and E4G-400 cell site routers.

**Table 25: TDM Port Number Mapping for E4G-200 and E4G-400**

| Cell Site Router | Module | Panel TDM Port Numbers | TDM Port Numbers in ExtremeXOS CLI |
|---|---|---|---|
| Note that the panel TDM port numbers are different from the TDM port numbers used for the configuration. | | | |
| E4G-400 | E4G-B16T1E1 | 1–16 | 35-50 |
| E4G-200 | E4G-F16T1E1 | 1–16 | 13–28 |

In E4G-200, the port number 13 in the ExtremeXOS CLI refers to the TDM port 1 in the face-plate. Similarly, port 14 in the ExtremeXOS CLI refers to TDM port 2 in the face-plate and so on.

In E4G-400, the port number 35 in the ExtremeXOS CLI refers to the TDM port 1 in the face-plate. Similarly, port 36 in the ExtremeXOS CLI refers to TDM port 2 in the face-plate, and so on.

### *Examples of TDM Ports Numbering*

Enable TDM ports numbering in E4G-200 and E4G-400.

- To enable TDM port 2 in E4G-400 using the port number 36, use the following command:

```
enable port 36 tdm
```

- Enable TDM port 5 in E4G-200 using the port number 17, use the following command

```
enable port 17 tdm
```

Note

**tdm** indicates that the port number in the enable/disable/configure port commands is a TDM port.

### *Configuring TDM Ports*

- To configure the framing used on TDM ports, use the following command:

  `configure ports` *`port_list`* **`tdm framing`** `[`**`d4`** | **`esf`** | `[`**`basic`** | **`mf`**`]` `{`**`crc4`**`}` | **`unframed`**`]`

- To configure the line coding scheme to be used on TDM ports, use the following command:

  `configure ports` *`port_list`* **`tdm line-coding`** `[`**`b8zs`** | **`hdb3`** | **`ami`**`]`

- To configure the cable length and receiver gain to be used on TDM ports, use the following command:

  `configure ports` *`port_list`* **`tdm cable-length`** `[` **`short-haul`** `[`**`110`** | **`220`** | **`330`** | **`440`** | **`550`** | **`660`**`]` | **`long-haul line-build-out`** `[`**`0db`** | **`75db`** | **`150db`** | **`225db`**`]]`

- To configure the local and network loopback mode for TDM ports, use the following commands to enable and disable loopback:

  `enable ports` *`port_list`* **`tdm loopback`** `[`**`local`** | **`network`** `[`**`line`** | **`payload`**`]]`

```
disable ports port_list tdm loopback [local | network [line |
payload]]
```

- To configure or clear a display string for TDM ports, use the following commands:

```
configure ports port_list tdm display-string string
```

```
unconfigure ports port_list tdm display-string
```

- To enable or disable TDM ports, use the following commands:

```
enable ports [port_list | all] tdm
```

```
disable ports [port_list | all] tdm
```

- To configure the transmit clock source for TDM ports, use the following command:

```
configure ports port_list tdm clock-source [line | network | [adaptive
| differential] ces ces_name]
```

- To configure the recovered clock and quality level for TDM ports, use the following command:

```
configure ports port_list tdm recovered-clock {quality-level value}
```

- To unconfigure the recovered clock for TDM ports, use the following command:

```
unconfigure ports port_list tdm recovered-clock
```

- To configure the idle code to be used on TDM ports, use the following command:

```
configure tdm service circuit service_name seized-code seized_code
```

- To configure signaling on TDM ports, use the following command:

```
configure ports port_list tdm signaling [bit-oriented | robbed-bit |
none]
```

Note

A given TDM port cannot belong to more than one TDM service when the port is in unframed mode.

*Configuring TDM Services*

- To create or delete a TDM service, use the following commands:

```
create tdm service circuit service_name
```

```
delete tdm service circuit [service_name | all]
```

- To add a port or port/time-slot to a TDM service, use the following command:

```
configure tdm service circuit service_name add port port {time-slots
[time_slot_list | all]}
```

- To delete a port from a TDM service, use the following command:

- To configure the idle code and seized code, use the following command:

  `configure tdm service circuit service_name **seized-code** seized_code`

- To configure the trunk-conditioning value for alarm conditions, use the following command:

  `configure **tdm service circuit** service_name **trunk-conditioning** trunk_conditioning`

Note

- A given {TDM port, time-slot} combination cannot belong to more than one TDM service.
- A TDM service can belong to only one TDM pseudowire
- In the framed mode of operation on E1 hierarchy, timeslot 1 cannot be added to a TDM service. Additionally, if TDM port is configured as multiframed, timeslot 17 cannot be added to a TDM service.
- Time-slots from different TDM ports cannot belong to the same TDM service.

*Configuring and Managing CES Pseudowires*

Use the following commands to configure Circuit Emulation Service (CES) pseudowires.

- To create or delete a CES pseudowire, use the following commands:

  `create ces ces_name **psn** [**mef8** | **udp** | **mpls**]]`

  `delete ces [ces_name | **all**]`

- To enable or disable the administrative status of a CES pseudowire, use the following commands:

  `enable ces [ces_name | **all**]`

  `disable ces [ces_name | **all**]`

- To manually add an IPv4 peer (far-end) for a CES pseudowire, use the following command:

  `configure ces ces_name **add peer ipaddress** ipaddress [**fec-id-type pseudo-wire** pw_id {**lsp** lsp_name} | **udp-port local** src_udp_port **remote** dst_udp_port **vlan** vlan_name]`

- To manually add an Ethernet (MEF-8) peer (far-end) for a CES pseudowire, use the following command:

  `configure ces ces_name **add peer mac-address** mac_address **ecid local** tx_ecid **remote** rx_ecid **vlan** vlan_name`

- To delete a peer of a CES pseudowire, use the following command:

  `configure ces ces_name **delete peer** [**ipaddress** ipaddress | **mac-address** mac_address]`

- To add or delete a TDM service on a CES pseudowire, use the following command:

  `configure ces ces_name **add service** service_name`

  `configure ces ces_name **delete service**`

- To configure the jitter-buffer value for a CES pseudowire, use the following command:

  `configure ces ces_name **jitter-buffer** min_jbf {**max** max_jbf}`

- To configure the payload-size value for a CES pseudowire, use the following command:

  `configure ces ces_name **payload-size** bytes`

- To configure the QoS profile for a CES pseudowire, use the following command:

  `configure ces ces_name **qosprofile** qosprofile`

- To configure the filler pattern for a CES pseudowire, use the following command:

  `configure ces ces_name filler-pattern byte_value`
- To configure Loss of Packet State (LOPS) on a CES pseudowire, use the following command:

  `configure ces ces_name lops-threshold [entry num_packets_for_entry {exit num_packets_for_exit} | exit num_packets_for_exit`
- To configure time-to-live (TTL) on a CES pseudowire, use the following command:

  `configure ces ces_name ttl ttl_value`
- To enable or disable the CES pseudowire peer, use the following command:

  `[enable | disable] ces ces_name peer ipaddress ipaddress`
- To configure DSCP value on a CES pseudowire, use the following command:

  `configure ces ces_name dscp dscp_value`

Note

- Payload size can be reconfigured only after disabling the TDM pseudowire.
- TDM service can be removed from a TDM pseudowire only after the Peer Configuration of the TDM pseudowire is removed.
- The CES pseudowire configured for recovering clock cannot be deleted when it is configured as the clock source for the TDM port. Change the TDM port transmit clock source before deleting the pseudowire.

*Displaying TDM PW Configurations*

- To display TDM port information, use the following command:

  `show ports {port_list} tdm information {detail}`
- To display TDM port configuration information, use the following command:

  `show ports {port_list} tdm configuration {no-refresh}`
- To display the TDM port alarms, use the following command:

  `show ports {port_list} tdm alarms {no-refresh}`
- To display TDM service interface information, use the following command:

  `show tdm service {circuit} {service_name}`
- To display CES pseudowire parameters, use the following command:

  `show ces {ces_name} {detail}`
- To display CES peer information, use the following command:

  `show ces peer [ipaddress ipaddress | mac-address mac_address]`
- To display TDM port information, use the following command:

  `show ports {port_list} tdm {no-refresh}`
- To display TDM hierarchy information, use the following command:

  `show tdm hierarchy`
- To display CES clock recovery information, use the following command:

  `show ces {ces_name} clock-recovery`

### *TDM Port and PW Statistics*

You can display errors in TDM ports and CES pseudowires.

- To display specified TDM port error counters, use the following command:

  `show ports {port_list} tdm errors {near-end} {total | intervals | current {no-refresh}}`

- To display specified CES pseudowire error counters, use the following command:

  `show ces {ces_name} errors {total | intervals | day-intervals | current {no-refresh}}`

### *Understanding Adaptive Clock Recovery*

The clock to drive TDM ports can be recovered from a TDM pseudowire using the Adaptive Clock Recovery (ACR) algorithm.

ACR recovers the TDM service clock based on the packet arrival rate and typically employed when no other clock is available in the network to achieve synchronization.

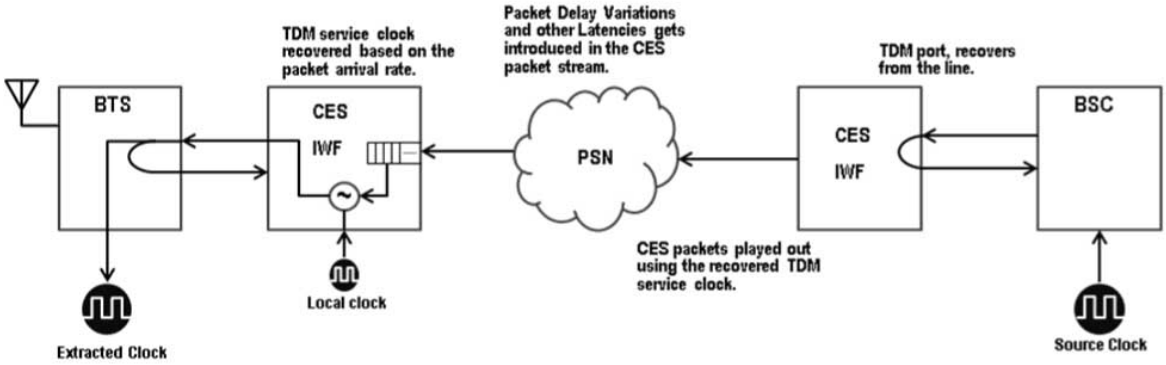



**Figure 16: Adaptive Clock Recovery**

The adaptive clock recovery uses techniques to filter out the Packet Delay Variations (PDV) introduced in the packet stream by the PSN and recovers the TDM service clock.

The Wander and the Jitter budgets are defined by G.8261 deployment cases (case 1-a, 2-a, 2-b). Network deployments that differ from above cases require deriving the budgets based on the deployment model.

Limitations:

- Only one TDM port can be timed using the clock recovered from the pseudowire.
- The pseudowire can only time the TDM port that is attached as a service circuit.
- The clock recovered from the pseudowire cannot be used as a system clock source for synchronization. This implies that the pseudowire recovered clock cannot be carried through Sync-E or PTP or BITS.
- When configuring a SAToP pseudowire for clock recovery, the TDM payload bytes carried in pseudowire should be a multiple of 32.
- Adaptive clock recovery cannot filter out the low frequency wander introduced by the 'beating effect'.

*Understanding TDM Transmit Clock Configuration*

The TDM transmit clock is configured using the clock-source command. The TDM line can be configured to use one of the following clock sources for transmit:

*Terms*

Line

The clock recovered from the received TDM stream on the TDM port is used as a transmit clock source on the same TDM port.

Adaptive

The transmit clock source for the TDM port is the clock recovered from the PSN pseudowire packets. The transmit clock is adaptively recovering clock from the pseudowire packet arrival rate.

Network clock:

The transmit clock source for the TDM port is the common synchronized clock in the system. The system clock could be synchronized to one of the following clock sources: SyncE, 1588v2, BITS or a clock recovered from the TDM port.

*Understanding TDM Port Alarms*

The alarm events from the TDM port that are detected and the alarm response transmitted on the TDM port are listed in the following section.

The alarm response on the TDM port/time-slot(s) depends on the port or time-slot(s) configuration state. The port or time-slot is said to be in disconnected or in idle configuration state when they are not added to a TDM service. The port or time-slot(s) are said to be connected if they are part of a TDM service. In idle state, depending on the framing configuration on the port, the alarm response would vary. The alarms generated and the alarm events detected are logged.

**TDM Port Alarms in Unframed mode**

AIS Alarm Generation

The TDM ports generate an Alarm Indication Signal (AIS) alarm by default on the ports that are not connected to a service. On ports that are connected to a service, the AIS alarm is generated to indicate pseudowire faults.


**Figure 17: AIS Alarm**

E4G switches do not detect AIS alarm events in unframed mode of operation.

LOS Alarm Generation

The TDM ports generate Loss of Signal (LOS) alarm on the ports that are administratively disabled. The alarm is cleared when the port is enabled.


**Figure 18: LOS Alarm Generation**

LOS Alarm Response

The TDM ports detect a Loss of Signal alarm event. No specific data is played out as a response for this alarm event. However, when the port is not a part of TDM service, the preset idle pattern of all ones (or AIS) is played out. If the port is connected to a TDM service bound to a CES pseudowire, the TDM data from the remote end of the CES pseudo-wire, is played out, facilitating the tunneling of alarm event response from the remote TDM CE node.


**Figure 19: LOS Alarm Response**

**TDM Port Alarms in Framed Mode**

Default Line State

The TDM ports send out a preset idle pattern of 0xFF on all timeslot(s) that carry TDM data. If signaling multiframe is configured on the TDM port (mf in E1 hierarchy and d4 or esf in T1 hierarchy), a configurable idle code is played out on the signaling channel/bits.


**Figure 20: Default Line State**

However, the idle pattern is not played out on certain special timeslots, as listed in the following table below.

**Table 26: Idle Pattern on Timeslots**

| | |
|---|---|
| E1 Hierarchy | |
| Timeslot - 1 | Carries frame alignment signal, CRC and remote alarm information. |
| Timeslot - 16 (in frame 1/16) | Carries signaling multiframe alignment signal, spare bits and multiframe alarms. Applicable only if port handles signaling multiframe. |
| T1 Hierarchy | |
| F-bits | Carries framing alignment signal information. In case of Extended Super Frame formats, carries data link and CRC-6 information. |

Note that the idle pattern playout does not indicate the presence or generation of alarms and is presented here for information purposes.

### LOS/LOF/AIS Alarm Response

The Loss of Signal, Loss of Frame and Alarm Indication Signal events are detected and a Remote Alarm Indication is played out as alarm response. The Framed modes in E1 and T1 hierarchy have specific bits in the frame formats for indicating the remote TDM CE interface about the faults.

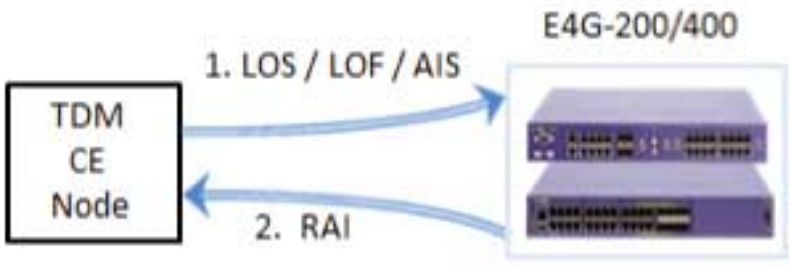



**Figure 21: LOS/LOF/AIS Alarm Response**

The following framing types configured on the CE node and on the E4G node are considered as incompatible in the E4G node. This would result in detection of Loss of Frame alarm. The Loss of Signaling Multiframe and the Loss of CRC Multiframe are detected as Loss of Frame alarm events.

**Table 27: Framing Types**

| Hierarchy | Framing in E4G node | Framing in Remote CE node |
|---|---|---|
| E1 | Basic | Unframed |
| | Signaling multiframe | Unframed<br>Basic |
| | CRC4 enabled | CRC4 disabled |
| T1 | Super frame (D4) | Unframed<br>Extended super frame (ESF) |
| | Extended super frame (ESF) | Unframed<br>Super frame (D4) |

### LOS Alarm Generation

The TDM ports generate Loss of Signal alarm on the ports that are administratively disabled. The alarm is cleared when the port is enabled.


**Figure 22: LOS Alarm Generation**

*TDM Port Alarms and Remedies*

The following table shows the TDM port alarm conditions detected and generated in different configuration setting, with and without the port being part of a TDM service bound to a CES pseudowire with suggested remedies.

**Table 28: TDM Port Alarms and Remedies**

| Alarm | Description | Remedy |
|---|---|---|
| LOS | This condition occurs on the TDM port when the local end of the TDM port is in Loss of Signal state. The mismatch in the configured hierarchy, cable length or line gain parameters results in the Loss of Signal state in the local end of the TDM port. | The hierarchy configuration and the interface parameters such as cable length or line gain needs to be reviewed. If no configuration deviations are observed, the transmit clocking option in the remote end requires to be reviewed to isolate the possibility of using an unavailable clock. |
| LOF | This condition occurs on the framed TDM port when the local end of the TDM port is in Loss of Frame state. The mismatch in the transmitted framing format in the local end and the configured framing format in the remote end results in the Loss of Frame state in the local end of the TDM port. | The framing configuration in the local and remote end of the TDM port needs to be reviewed. If no configuration deviations are observed, the fault due to unstable clock can be isolated by performing loopback tests on the local and/or remote end of the TDM port. |
| TxRAI | This condition occurs on the framed TDM port due to either of the two cases:<br>When there is a mismatch in the configured and received framing format. In this case, the transmission of remote alarm indication is triggered by the Loss of Frame state in the local end of the TDM port. In the presence of CES pseudowires on the TDM port, when the remote end of the CES pseudowire sends an RDI (Remote defect indicator) signal, RAI is transmitted on the TDM port. | If there are CES pseudo-wires defined on the port, the pseudowire remote fault can be referred to. In the presence of attachment circuit Tx fault, no action is required. If there are no CES pseudo-wires defined, the framing configuration on the TDM port needs to be reviewed. Additionally, if no configuration deviations are observed, the fault due to unstable clock can be isolated by performing loopback tests on the local and/or remote end of the TDM port.<br>This condition, if occurs on the unframed TDM port, can be cleared by administratively disabling and enabling the TDM port.<br>The following framing configuration in local/remote would cause RAI to be generated from the local end: |

**Table 28: TDM Port Alarms and Remedies (continued)**

| Alarm | Description | Remedy |
|---|---|---|
| RxRAI | This condition occurs on the framed TDM port due to either of the two cases:<br>The Loss of Frame state in the remote end of the TDM port. The application associated with the remote end of the TDM port tunnels the alarm indication to the local end. The mismatch in the transmitted framing format in the local end and the configured framing format in the remote end results in the Loss of Frame state in the remote end of the TDM port. | If this condition occurs due to the Loss of Frame state in the remote end of the TDM port, the framing configuration on the TDM port needs to be reviewed. Additionally, if no configuration deviations are observed, the fault due to unstable clock can be isolated by performing loopback tests on the local and/or remote end of the TDM port. If this condition occurs due to the application associated with the remote end of the TDM port, no action to be taken. This condition, if occurs on the unframed TDM port, can be cleared by administratively disabling and enabling the TDM port. |
| TxAIS | This condition occurs on the unframed TDM port due to either of the two cases:<br>The AIS is transmitted on the TDM port by default in the absence of loopback or CES pseudowire configuration. In the presence of CES pseudowires on the TDM port, the AIS is transmitted to indicate the remote end pseudowire faults, namely, local end loss of packet state and remote end attachment circuit fault. | If no CES pseudowire is configured on the TDM port, no action is required. If CES pseudowires are configured on the TDM port, the pseudowire fault information should be referred to for the remote end fault indication. This condition, if occurs on a framed TDM port or occurs on an unframed TDM port with no remote end fault indication in the CES pseudowire, can be cleared by administratively disabling and enabling the TDM port. |
| RxAIS | This condition occurs on the framed TDM port when the remote end of the TDM port transmits an AIS alarm indication. | This condition requires no action to be performed. If CES pseudo-wires are present on the TDM port, this condition is signaled as local attachment circuit Rx fault. |

*Understanding TDM CES Pseudowire Alarms*

The CES pseudowires transport the alarm events detected on the service interface and the alarm events triggered on the PSN transport using the LRM bits in the pseudowire control word. The significance and the usage of the LRM bits are covered by RFC4553 for SAToP pseudowires and RFC5086 for CESoP pseudowires. The end-to-end alarm handling between two E4G units for SAToP and CESoP pseudowires are discussed below. The alarms generated and the alarm events detected in the CES pseudowires are logged.

#### CES Alarms in SAToP Pseudowires

TDM Service LOS Alarm

The Loss of Signal alarm event in the TDM service attached to the SAToP pseudowire is handled end-to-end as shown in the following figure.

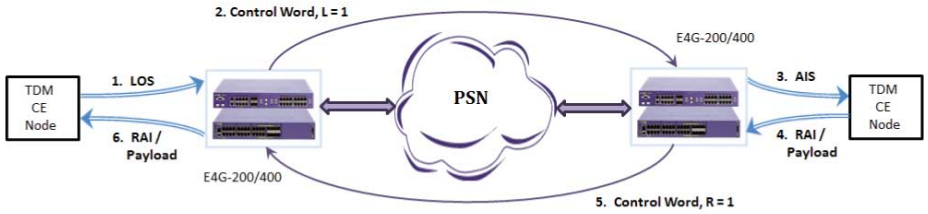



**Figure 23: SAToP Alarm Handling: TDM Service LOS Alarm**

1. The Loss of Signal alarm event from the TDM service is detected by the local end E4G node.
2. The local end E4G node notifies the alarm condition to the remote end of the CES pseudowire by setting the L-bit in the TDM pseudowire control word.
3. The remote E4G node, upon receiving the CES pseudowire with L-bit, ignores the TDM payload carried in the packet and plays out Alarm Indication Signal to the remote TDM CE node.
4. The remote TDM CE node sends a response to the Alarm Indication Signal, which could be a specific pattern in case of unframed services, for example, an all ones pattern.
5. The remote E4G node sends the alarm response with R-bit set, indicating the packet loss caused due to dropping of packets with L-bit set.
6. The local E4G node receives the alarm response packets with R-bit set and forwards the alarm response data to the local TDM CE node.

### TDM Service AIS Alarm

The Alarm Indication Signal alarm from the TDM service is not detected by the E4G switch. This alarm is carried transparently to the remote TDM CE node and the alarm response is carried back transparently to the local TDM CE node as pictured. The CES pseudowire control word is not updated to reflect the presence of this alarm condition.


**Figure 24: SAToP Alarm Handling: TDM Service AIS Alarm**

### PSN Loss of Packet State

The CES pseudowire packets carry the TDM service payload at a constant rate depending on the payload size. The replay of TDM service payload at the remote end of the CES pseudowire is done based on the sequence number in the CES pseudowire control word. Due to the variable nature of the packet switched network, the CES pseudowire streams get dropped in the intermediate nodes. Under this scenario, the remote end of the CES pseudowire is said to be in LOSY state. The LOSY state of the CES pseudowire is indicated to the peer by setting the R-bit in the CES pseudowire control word.

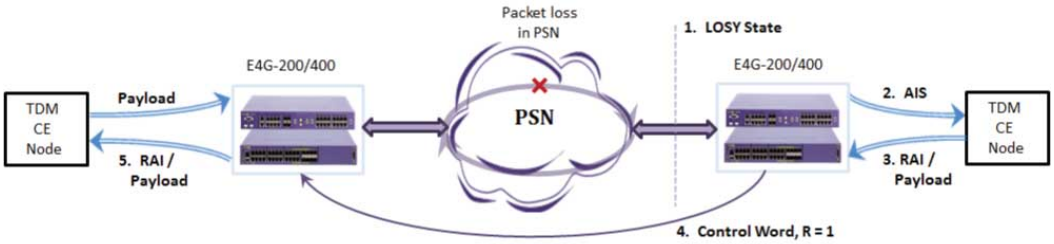



**Figure 25: SAToP Alarm Handling: PSN Loss of Packet State**

The R-bit in the control word is set on the CES pseudowire packets from remote E4G node in LOSY state to the local E4G node, regardless of the RAI pattern received from its local TDM CE node.

#### CES Alarms in CESoP Pseudowires

The handling of CES alarms in CESoP pseudowires are more involved due to association of one or more timeslots to a TDM service and hence multiple services originating from a single TDM port with disjoint timeslots. On alarm conditions, the configured trunk condition code for data channels is played out. For signaling channels, the configured seized code pattern is played out.

### TDM Service LOS/LOF/AIS Alarm

The Loss of Signal, Loss of Frame and Alarm Indication Signal events in the TDM service attached to the CESoP pseudowire is handled end-to-end as shown in the following figure.

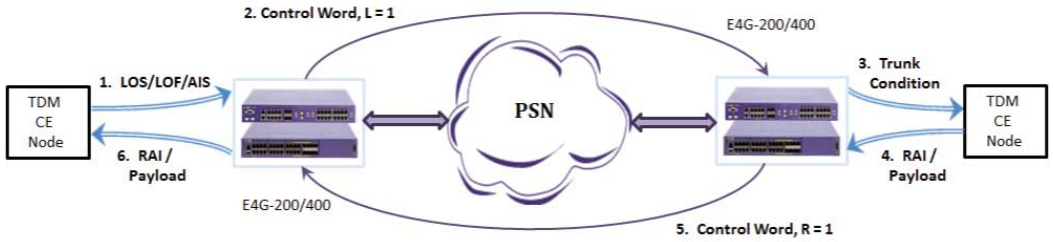



**Figure 26: CESoP Alarm Handling: TDM Service LOS/LOF/AIS Alarm**

The alarm handling sequence is similar to SAToP pseudowires, with an exception that the alarm is indicated only on the specific TDM service bound to the CES pseudowires. For instance, if the TDM service has 10 timeslots bound to the CES pseudowire, the alarm is indicated by the remote E4G node by playing out the configurable trunk conditioning pattern on those 10 timeslots in the TDM service. If signaling multiframe mode is configured on the TDM port, the configurable seized code pattern is played on the signaling bits.

### TDM Service RAI Alarm

The CESoP pseudowires indicates the remote E4G node of the Remote Alarm Indication (or Remote Defect Identifier) alarms detected on the TDM service attached to local TDM CE node. The M-bit in the CES pseudowire control word is set to indicate the detected alarm. The remote E4G node sets the RAI indication on the TDM port attached to its local TDM CE node in addition to playing out the TDM payload received. The following figure shows the alarm handling sequence.

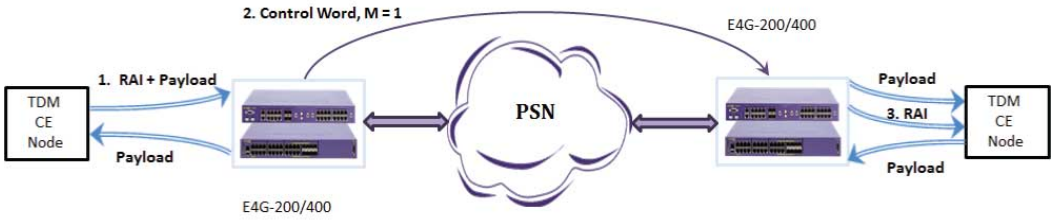



**Figure 27: CESoP Alarm Handling: TDM Service RAI Alarm**

### PSN Loss of Packet State

The CESoP pseudowires handle the LOSY state due to loss of CES pseudowire packets in the PSN, in the similar way as handled by SAToP pseudowires. The configured trunk conditioning code and seized code is played on the timeslots connected to the TDM service instead of AIS. The following figure shows the alarm handling sequence.

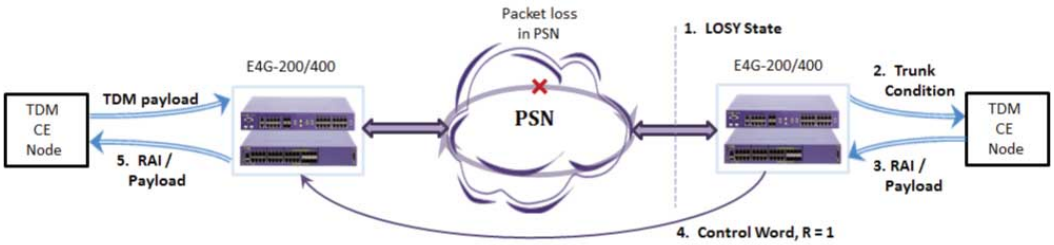



**Figure 28: CESoP Alarm Handling: PSN Loss of Packet State**

*CES Pseudowire Alarms and Remedies*

The following table lists the CES pseudowire alarm conditions detected and generated in the E4G node with the suggested remedies.

**Table 29: CES Pseudowire Alarms and Remedies**

| Alarm | Description | Remedy |
|---|---|---|
| Local L-bit (Local attachment circuit Rx fault) | This condition occurs when the TDM port in attachment circuit of the CES pseudowire is in failure state or is administratively disabled. When the TDM port is in LOS, AIS or LOF condition, the local end of the CES pseudowire carries the L-bit in the control word as an indication of the local attachment circuit alarm to the remote end of the pseudowire. This condition can be induced by disabling the TDM port administratively. This condition applies for both SAToP and CESoP pseudowires. | The alarms associated with the TDM port in the attachment circuit of the CES pseudowire can be referred. The L-bit condition is cleared when the associated TDM port is restored from the failure state. |
| Local R-bit (Local pseudo-wire LOSY state) | This condition occurs when the local end of the CES pseudowire enters the LOSY state. The CES pseudowire enters LOSY state when the packets from the remote end of the CES pseudowire are lost in the transit. The local end of the CES pseudowire carries the R-bit in the control word as an indication of the remote pseudowire fault. The CES pseudo-wire stream from the remote end, carrying L-bit would result in LOSY state in the local end of the CES pseudowire. Administratively disabling the CES pseudo-wire in the local end would result in the LOSY state in the remote end of the CES pseudo-wire. This condition applies for both SAToP and CESoP pseudowires. | If this condition occurs and the CES pseudowire is not disabled, the reachability of the CES pseudowire peer needs to be checked. If the peer is reachable, occurrence of remote end attachment circuit fault (remote L-bit condition) could be referred to. The R-bit condition is cleared when the remote CES pseudowire packets are received. |
| Local M-bit (Local attachment circuit Tx fault) | This condition occurs when the TDM port in the attachment circuit of the CES pseudowire receives remote alarm indication (Rx RAI). The local end of the CES pseudowire carries M-bit in the control word to indicate the remote end of the CES pseudowire, of the reception of RAI in the local attachment circuit. This condition applies for CESoP pseudo-wires only. | If this condition occurs, the alarms associated with the TDM port in the attachment circuit of the CES pseudowire can be referred. The M-bit condition is cleared when the associated TDM port stops receiving RAI indication. |

**Table 29: CES Pseudowire Alarms and Remedies (continued)**

| Alarm | Description | Remedy |
|---|---|---|
| Remote L-bit (Remote attachment circuit Rx fault) | This condition occurs when the TDM port in attachment circuit of the remote CES pseudowire is in failure state or is administratively disabled. The remote end of the CES pseudowire carries the L-bit in the control word to the local end of the CES pseudowire as an indication of the attachment circuit alarm in the remote end of the CES pseudo-wire. This condition can be induced by disabling the TDM port in the remote end of the CES pseudowire administratively. This condition applies for both SAToP and CESoP pseudowires. This condition causes AIS or trunk conditioning pattern to be transmitted on the local attachment circuit. | If this condition occurs, the TDM port/attachment circuit alarms in the remote end of the CES pseudowire can be referred. The L-bit condition is cleared when the TDM port in the remote end of CES pseudowire is restored of the failure state. |
| Remote R-bit (Remote attachment circuit LOSY state) | This condition occurs when the remote end of the CES pseudowire enters the LOSY state. When the packets from the local end of the CES pseudowire are lost in transit the remote end of the CES pseudowire enters LOSY state. This condition also occurs when the local end of the CES pseudowire carries L-bit to indicate the remote end of the CES pseudo-wire, of the fault in the local attachment circuit. When the local end of the CES pseudo-wire is administratively disabled, the remote end of the CES pseudo-wire enters LOSY state. This condition applies for both SAToP and CESoP pseudo-wires. | If this condition occurs and the remote end of CES pseudowire is not disabled, occurrence of the local end attachment circuit fault (local L-bit condition) could be referred to. If local end of CES pseudowire does not carry L-bit, the reachability of the local peer from the remote end of the CES pseudo-wire needs to be checked. The R-bit condition is cleared when the local CES pseudo-wire packets are received in the remote end. |
| Remote M-bit (Remote attachment circuit Tx fault) | This condition occurs when the TDM port in the attachment circuit of the remote CES pseudowire receives remote alarm indication (Rx RAI). The remote end of the CES pseudowire carries M-bit in the control word to the local end as an indication of the RAI reception by the attachment circuit. This condition applies for CESoP pseudo-wires only. This condition causes RAI to be transmitted on the local attachment circuit. | If this condition occurs, the alarms associated with the TDM port in the attachment circuit of the remote CES pseudo-wire can be referred. The M-bit condition is cleared when the associated TDM port stops receiving RAI indication |

*Management Information Base (MIB) Support*

The following TDM pseudowire related MIBs are supported in ExtremeXOS:

- Read-only support for RFC5604—Managed objects for TDM over Packet Switched Networks (PSNs).
- Read-only support for RFC5601—PW MIB.
- Read-only support for RFC2494—Definitions of managed objects for the DS0 and DS0 Bundle Interface Type.
- Read-only support for RFC4805—Definitions of managed objects for DS1, J1, E1, DS2 and E2 Interface Types.

*TDM PW Configurations Examples*

Examples of TDM PW configurations.

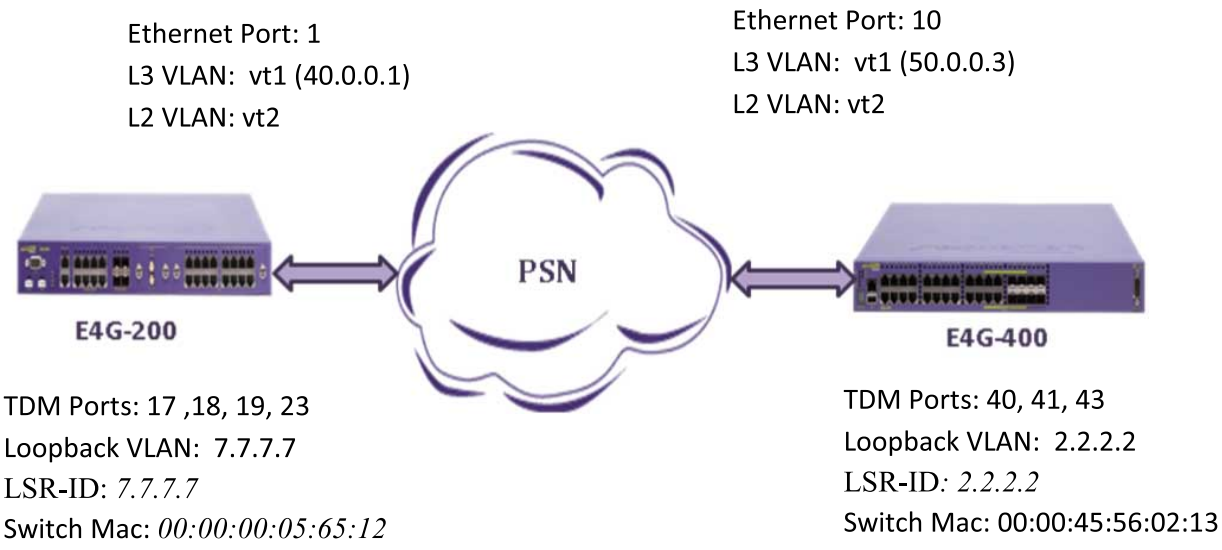



**Figure 29: TDM PW Configuration Example**

*Configuring TDM UDP SAToP Pseudowire*

1. Create TDM Service circuit.

   On the left E4G-200 Switch:

   ```
   create tdm service circuit "udp-satop-s1"
   configure tdm service circuit "udp-satop-s1" add port 18
   ```

   On the right E4G-400 Switch:

   ```
   create tdm service circuit "udp-satop-s1"
   configure tdm service circuit "udp-satop-s1" add port 41
   ```

2. Create CES and add TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create ces udp-ces1 psn udp
   configure ces udp-ces1 add service udp-satop-s1
   ```

   On the right E4G-400 Switch:

   ```
   create ces udp-ces1 psn udp
   configure ces udp-ces1 add service udp-satop-s1
   ```

3. Configure the loopback vlan.

   On the left E4G-200 Switch:

   ```
   create vlan "lpbk"
   enable loopback-mode vlan lpbk
   configure vlan lpbk ipaddress 7.7.7.7 255.255.255.255
   enable ipforwarding vlan lpbk
   ```

On the right E4G-400 Switch:

```
create vlan "lpbk"
enable loopback-mode vlan lpbk
configure vlan lpbk ipaddress 2.2.2.2 255.255.255.255
enable ipforwarding vlan lpbk
```

4. Configure the L3 transport vlan to reach the PW peer.

   On the left E4G-200 Switch:

```
create vlan "vt1"
configure vlan vt1 tag 30
configure vlan vt1 add ports 1 tagged
configure vlan vt1 ipaddress 40.0.0.1 255.255.255.0
enable ipforwarding vlan vt1
```

   On the right E4G-400 Switch:

```
create vlan "vt1"
configure vlan vt1 tag 20
configure vlan vt1 add ports 10 tagged
configure vlan vt1 ipaddress 50.0.0.3 255.255.255.0
enable ipforwarding vlan vt1
```

5. Add peer to the CES.

   On the left E4G-200 Switch:

```
configure ces udp-ces1 add peer ipaddress 2.2.2.2 udp-port local 10000 remote 10000
vlan lpbk
```

   On the right E4G-400 Switch

```
configure ces udp-ces1 add peer ipaddress 7.7.7.7 udp-port local 10000 remote 10000
vlan lpbk
```

*Configuring TDM UDP CESoP Pseudowire*

1. Configure TDM Port Framing mode.

   On the left E4G-200 Switch:

```
configure ports 17 tdm framing mf
```

   On the Right E4G-400 Switch:

```
configure ports 40 tdm framing mf
```

2. Create TDM Service Circuit.

   On the left E4G-200 Switch:

```
create tdm service circuit "udp-cesop-s2"
configure tdm service circuit "udp-cesop-s2" add port 17 time-slots 2-4
```

   On the Right E4G-400 Switch:

```
create tdm service circuit "udp-cesop-s2"
configure tdm service circuit "udp-cesop-s2" add port 40 time-slots 2-4
```

3. Configure the loopback vlan.

   On the left E4G-200 Switch:

```
create vlan "lpbk"
enable loopback-mode vlan lpbk
configure vlan lpbk ipaddress 7.7.7.7 255.255.255.255
enable ipforwarding vlan lpbk
```

On the Right E4G-400 Switch:

```
create vlan "lpbk"
enable loopback-mode vlan lpbk
configure vlan lpbk ipaddress 2.2.2.2 255.255.255.255
enable ipforwarding vlan lpbk
```

4. Configure the L3 transport vlan to reach the PW peer.

   On the left E4G-200 Switch:

   ```
   create vlan "vt1"
   configure vlan vt1 tag 30
   configure vlan vt1 add ports 1 tagged
   configure vlan vt1 ipaddress 40.0.0.1 255.255.255.0
   enable ipforwarding vlan vt1
   ```

   On the Right E4G-400 Switch:

   ```
   create vlan "vt1"
   configure vlan vt1 tag 20
   configure vlan vt1 add ports 10 tagged
   configure vlan vt1 ipaddress 50.0.0.3 255.255.255.0
   enable ipforwarding vlan vt1
   ```

5. Create CES and add the TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create ces udp-ces2 psn udp
   configure ces udp-ces2 add service udp-cesop-s2
   ```

   On the Right E4G-400 Switch:

   ```
   create ces udp-ces2 psn udp
   configure ces udp-ces2 add service udp-cesop-s2
   ```

6. Add Peer to the CES.

   On the left E4G-200 Switch:

   ```
   configure ces udp-ces1 add peer ipaddress 2.2.2.2 udp-port local 10001 remote 10001
   vlan lpbk
   ```

   On the Right E4G-400 Switch:

   ```
   configure ces udp-ces1 add peer ipaddress 7.7.7.7 udp-port local 10001 remote 10001
   vlan lpbk
   ```

**Note**

A single loopback vlan is sufficient when configuring multiple pseudo-wires to the same peer and each PW is identified using the unique UDP port numbers configured. The recommended option is to use loopback vlan to specify source IP address to be used in TDM UDP PW. However, the user can also use the normal vlan instead of loopback vlan.

*Configuring TDM MEF-8 SAToP Pseudowwire*

1. Create TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create tdm service circuit "mef8-satop-s3"
   configure tdm service circuit "mef8-satop-s3" add port 19
   ```

   On the right E4G-400 Switch:

   ```
   create tdm service circuit "mef8-satop-s3"
   configure tdm service circuit "mef8-satop-s3" add port 42
   ```

2. Create CES and add the TDM Service circuit.

   On the left E4G-200 Switch:

   ```
   create ces mef8-ces3 psn mef8
   configure ces mef8-ces3 add service mef8-satop-s3
   ```

   On the right E4G-400 Switch:

   ```
   create ces mef8-ces3 psn mef8
   configure ces mef8-ces3 add service mef8-satop-s3
   ```

3. Configure the L2 transport *VLAN* to reach the PW peer.

   On the left E4G-200 Switch:

   ```
   create vlan "vt2"
   configure vlan vt2 tag 130
   configure vlan vt2 add ports 1 tagged
   ```

   On the right E4G-400 Switch:

   ```
   create vlan "vt2"
   configure vlan vt2 tag130
   configure vlan vt2 add ports 10 tagged
   ```

4. Add peer to the CES.

   On the left E4G-200 Switch:

   ```
   configure ces mef8-ces3 add peer mac-address 00:00:45:56:02:13 ecid local 1001 remote
   1001 vlan vt2
   ```

   On the right E4G-400 Switch:

   ```
   configure ces mef8-ces3 add peer mac-address 00:00:00:05:65:12 ecid local 1001 remote
   1001 vlan vt2
   ```

*Configuring TDM MEF-8CESoP PW*

1. Configure TDM Port framing mode.

   On the left E4G-200 Switch:

   ```
   configure ports 17 tdm framing mf
   ```

   On the right E4G-400 Switch:

   ```
   configure ports 40 tdm framing mf
   ```

2. Create TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create tdm service circuit "mef8-cesop-s4"
   configure tdm service circuit "mef8-cesop-s4" add port 17 time-slots 6-8
   ```

   On the right E4G-400 Switch:

   ```
   create tdm service circuit "mef8-cesop-s4"
   configure tdm service circuit "mef8-cesop-s4" add port 40 time-slots 6-8
   ```

3. Create CES and add the TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create ces mef8-ces4 psn mef8
   configure ces mef8-ces4 add service mef8-cesop-s4
   ```

   On the right E4G-400 Switch:

   ```
   create ces mef8-ces4 psn mef8
   configure ces mef8-ces4 add service mef8-cesop-s4
   ```

4. Configure the L2 transport *VLAN* to reach the PW peer.

   On the left E4G-200 Switch:

   ```
   create vlan "vt2"
   configure vlan vt2 tag 130
   configure vlan vt2 add ports 1 tagged
   ```

   On the Right E4G-400 Switch:

   ```
   create vlan "vt2"
   configure vlan vt2 tag 130
   configure vlan vt2 add ports 10 tagged
   ```

5. Add peer to the CES.

   On the left E4G-200 Switch:

   ```
   configure ces mef8-ces4 add peer mac-address 00:00:45:56:02:13 ecid local 1002 remote
   1002 vlan vt2
   ```

   On the right E4G-400 Switch:

   ```
   configure ces mef8-ces4 add peer mac-address 00:00:00:05:65:12 ecid local 1002 remote
   1002 vlan vt2
   ```

   **Note**
   IP address should not be configured on the transport vlan specified for TDM MEF-8 PW.

*Configuring MPLS TDM SaTOP Pseudowire*

1. Configure the loopback vlan.

   On the left E4G-200 Switch:

   ```
   create vlan "lpbk"
   enable loopback-mode vlan lpbk
   configure vlan lpbk ipaddress 7.7.7.7 255.255.255.255
   enable ipforwarding vlan lpbk
   ```

On the right E4G-400 Switch:

```
create vlan "lpbk"
enable loopback-mode vlan lpbk
onfigure vlan lpbk ipaddress 2.2.2.2 255.255.255.255
enable ipforwarding vlan lpbk
```

2. Configure the L3 transport vlan to reach the PW peer.

   On the left E4G-200 Switch:

```
create vlan "vt1"
configure vlan vt1 tag 30
configure vlan vt1 add ports 1 tagged
configure vlan vt1 ipaddress 40.0.0.1 255.255.255.0
enable ipforwarding vlan vt1
```

   On the right E4G-400 Switch:

```
create vlan "vt1"
configure vlan vt1 tag 20
configure vlan vt1 add ports 10 tagged
configure vlan vt1 ipaddress 50.0.0.3 255.255.255.0
enable ipforwarding vlan vt1
```

3. Configure *OSPF*.

   On the left E4G-200 Switch:

```
configure ospf routerid 7.7.7.7
enable ospf
configure ospf add vlan lpbk area 0.0.0.0
configure ospf add vlan vt1 area 0.0.0.0
```

   On the right E4G-400 Switch:

```
configure ospf routerid 2.2.2.2
enable ospf
configure ospf add vlan lpbk area 0.0.0.0
configure ospf add vlan vt1 area 0.0.0.0
```

4. Configure MPLS.

   On the left E4G-200 Switch:

```
configure mpls add vlan "lpbk"
enable mpls vlan "lpbk"
enable mpls ldp vlan "lpbk"
configure mpls add vlan "vt1"
enable mpls vlan "vt1"
enable mpls ldp vlan "vt1"
configure mpls ldp advertise direct all
configure mpls lsr-id 7.7.7.7
enable mpls protocol ldp
enable mpls
```

   On the right E4G-400 Switch:

```
configure mpls add vlan "lpbk"
enable mpls vlan "lpbk"
enable mpls ldp vlan "lpbk"
configure mpls add vlan "vt1"
enable mpls vlan "vt1"
enable mpls ldp vlan "vt1"
configure mpls lsr-id 2.2.2.2
enable mpls protocol ldp
enable mpls
```

5. Create TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create tdm service circuit "mpls-satop-s6"
   configure tdm service circuit "mpls-satop-s6" add port 23
   ```

   On the right E4G-400 Switch:

   ```
   create tdm service circuit "mpls-satop-s6"
   configure tdm service circuit "mpls-satop-s6" add port 43
   ```

6. Create CES and add TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create ces mpls-ces6 psn mpls
   configure ces mpls-ces6 add service mpls-satop-s6
   ```

   On the right E4G-400 Switch:

   ```
   create ces mpls-ces6 psn mpls
   configure ces mpls-ces6 add service mpls-satop-s6
   ```

7. Add peer to the CES.

   On the left E4G-200 Switch:

   ```
   configure ces mpls-ces6 add peer ipaddress 2.2.2.2 fec-id-type pseudo-wire 102
   ```

   On the right E4G-400 Switch:

   ```
   configure ces mpls-ces6 add peer ipaddress 7.7.7.7 fec-id-type pseudo-wire 102
   ```

*Configuring MPLS TDM CeSOP Pseudowire*

1. Follow steps 1–3 of Configuring MPLS TDM SaTOP Pseudowire on page 225.
2. Configure TDM Port Framing Mode.

   On the left E4G-200 Switch:

   ```
   configure ports 17 tdm framing mf
   ```

   On the right E4G-400 Switch:

   ```
   configure ports 40 tdm framing mf
   ```

3. Create TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create tdm service circuit "mpls-cesop-s5"
   configure tdm service circuit "mpls-cesop-s5" add port 17 time-slots 18-23
   ```

   On the right E4G-400 Switch:

   ```
   create tdm service circuit "mpls-cesop-s5"
   configure tdm service circuit "mpls-cesop-s5" add port 40 time-slots 18-23
   ```

4. Create CES and add TDM Service Circuit.

   On the left E4G-200 Switch:

   ```
   create ces mpls-ces5 psn mpls
   configure ces mpls-ces5 add service mpls-cesop-s5
   ```

   On the right E4G-400 Switch:

   ```
   create ces mpls-ces5 psn mpls
   configure ces mpls-ces5 add service mpls-cesop-s5
   ```

5. Add peer to the CES.

   On the left E4G-200 Switch:

   ```
   configure ces mpls-ces5 add peer ipaddress 2.2.2.2 fec-id-type pseudo-wire 101
   ```

   On the right E4G-400 Switch:

   ```
   configure ces mpls-ces5 add peer ipaddress 7.7.7.7 fec-id-type pseudo-wire 101
   ```

   Note
   You must configure a loopback vlan with MPLS lsr-id as its IP address.

## Using the Precision Time Protocol

IEEE1588v2 (also known as Precision Time Protocol, or PTP) is an industry-standard protocol that enables the precise transfer of frequency and time to synchronize clocks over packet-based Ethernet networks.

The locally available clock on each network device synchronizes with a grandmaster clock by exchanging timestamps that contain sub-nanosecond granularity. This allows them to deliver very high accuracy to ensure the stability of base station frequency and handovers. The timestamps between master and slave devices are exchanged through PTP event packets. The ExtremeXOS 1588v2 implementation uses the IPv4/UDP transport mechanism PTP packets.

Note
The Precision Time Protocol is currently available only on X770, X460G2, X670G2 switches, and cell site routers (E4G-200 and E4G-400). For these routers, accurate synchronization of base stations to nanoseconds accuracy is critical to minimize service disruptions and eliminate dropped connections as calls move between adjacent cells.

### Overview of PTP

The IEEE 1588v2 Precision Time Protocol (PTP) defines a packet-based time synchronization method that provides frequency, phase, and time-of-day information with nanoseconds level of accuracy.

PTP relies on the use of carefully time-stamped packets to synchronize one or more slave clocks to a master clock. Synchronous time information is distributed hierarchically, with a grandmaster clock at the root of the hierarchy.

The grandmaster provides the time reference for one or more slave devices. These slave devices can, in turn, act as master devices for further hierarchical layers of slave devices.

To determine the master-slave hierarchy, a Best Master Clock (BMC) algorithm is used. This algorithm determines which clock is the highest quality clock within a network. The clock elected by BMC (the master clock) then synchronizes all other clocks (slave clocks) in the network. If the BMC is removed from the network or is determined by the BMC algorithm to no longer be the highest quality clock, the algorithm then redefines the new BMC and adjusts all other clocks accordingly. No administrator input is needed for this readjustment because the algorithm provides a fault tolerant behavior.

Synchronizing time across a network requires two essential functions: the measurement of delays and the distribution of time information. Each node is responsible for independently determining the delays across the network links from it to its link partners. Once this is accomplished, periodic time synchronization messages may be sent from the grandmaster clock device to the slave clock devices. Link-based delays wander over time, so periodic delay measurements are required. Because these delays vary slowly, the period between link delay measurements is typically in the order of seconds.

A PTP network must have a grandmaster clock reference and a slave. Between a master and a slave, a PTP network may have multiple boundary clocks, transparent clocks, and non-PTP bridges.

The following figure illustrates a typical PTP network hierarchy.

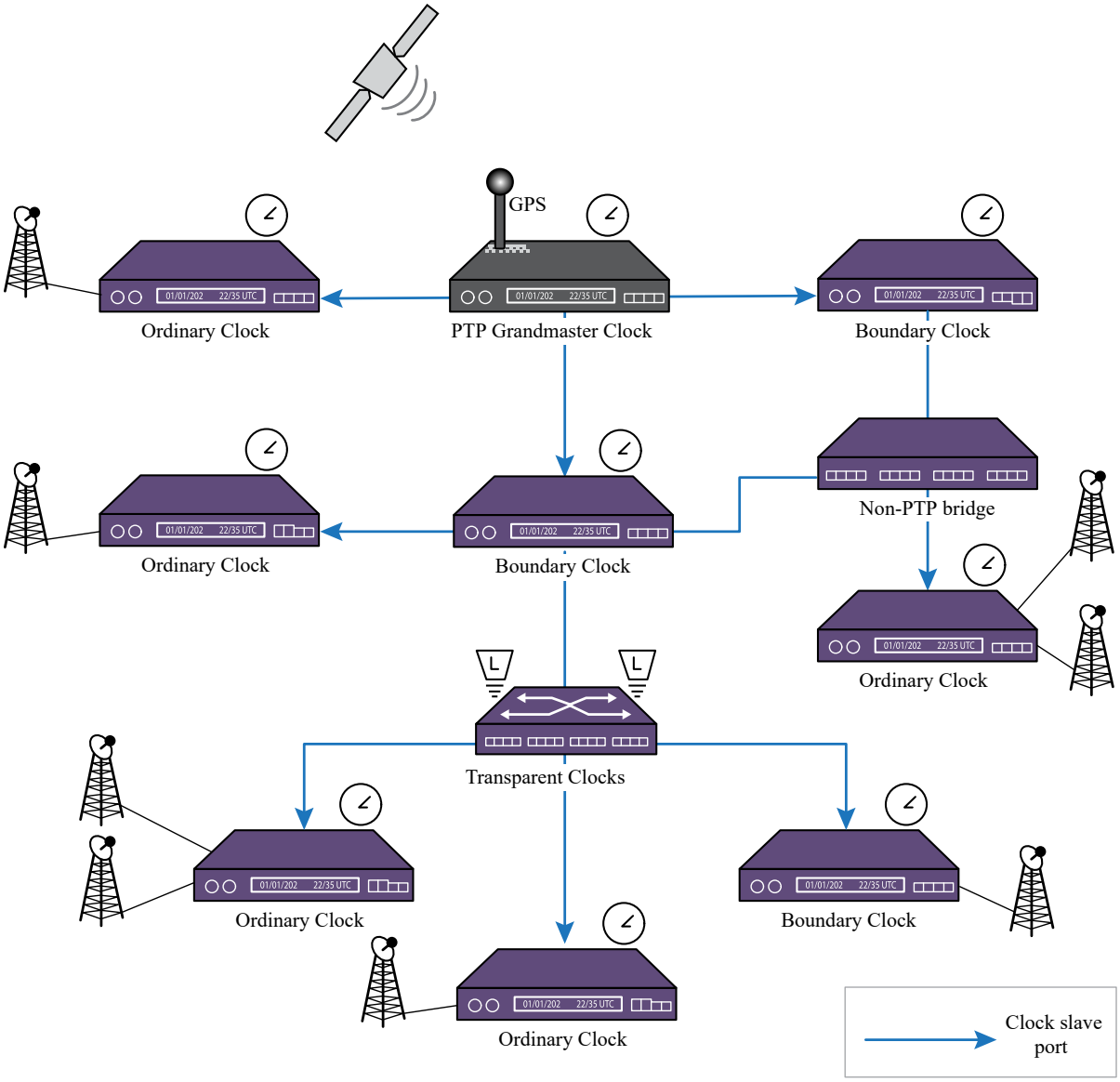



**Figure 30: PTP Network Hierarchy**

Ordinary clocks are devices with only one PTP port. The grandmaster clock is an ordinary clock acting as a master.

**Note**

A PTP port is a logical interface (*VLAN* / IP interface). A loopback VLAN is added as a clock port to PTP in ExtremeXOS.

Boundary clocks are switches with one or more PTP ports. One PTP port of a boundary clock can act as a slave to a master clock in the network, and the rest of the PTP ports can act as a master for the downstream nodes.

Transparent clocks correct the delays for PTP messages in the correction field.

End-to-end transparent clocks accumulate the residence time in the CorrectionField of the PTP messages. End-to-end transparent clocks do not participate directly in time synchronization with the

master clock. The CorrectionField of Sync, Delay Request, and Delay Response messages are updated by the end-to-end transparent clocks at each hop. The Signaling and Management messages are not updated by transparent clocks. In a typical setting, boundary and slave clocks are separated by one or more end-to-end transparent clocks that accumulates the residence time in the CorrectionField.

The residence time is defined as the delay between the reception and the transmission of packets through the device. The accumulated CorrectionField value is used by boundary or slave clocks for delay compensation in the time offset correction.

### *Basic Synchronization*

The following event flow describes basic synchronization of PTP:

1. The master sends a Sync message to the slave, notes the time (t1) it was sent, and embeds the t1 time in the message.
2. The slave receives the Sync message, and notes the time it is received (t2).
3. The master conveys to the slave the timestamp t1 either by (1) embedding the timestamp t1 in the Sync message (this requires hardware processing for highest accuracy and precision, and is called as one step clocking) or by (2) embedding the timestamp t1 in a Follow_Up message (this is called as two step clocking).
4. The slave sends a DelayReq message to the master, and notes the time (t3) it was sent.
5. The master receives the DelayReq message, and notes the time it is received (t4).
6. The master embeds the t4 timestamp in a DelayResp message to the slave.

   At the conclusion of this exchange of messages, the slave possesses all four timestamps. You can use these timestamps to compute the offset of the slave's clock with respect to the master, and the mean propagation time of messages between the two clocks.

   The formula to compute the offset is as follows:

   - master_to_slave_delay = offset + delay = t2 – t1
   - slave_to_master_delay = offset – delay = t4 – t3
   - one_way_delay = (master_to_slave_delay + slave_to_master_delay) / 2

   The calculation for the offset from the master is as follows: offset_from_master = t2 - t1 - one_way_delay.

**Note**
The above explanation for offset calculation is provided simply for understanding the basic PTP protocol. The actual calculations involve many more variables.

The computation of offset and propagation time assumes that the master-to-slave and slave-to-master propagation times are equal. Any asymmetry in propagation time introduces an error in the computed value of the clock offset.

### *End-to-End Transparent Clocks Between Master And Slave*

PTP defines the notion of End-to-end transparent clocks which do not participate in time synchronization with master clock.

Rather, they simply accumulate the residence time of PTP event messages such as Sync/DelayReq that transit the switch. The residence time is updated in the CorrectionField of these messages.

The transit delay in the link between the hops are not accounted in the CorrectionField by the End-to-end transparent clocks.

## PTP Slave Clock Adjustments

The following sections identify factors and components involved in Slave Clock adjustment. The actual values for the clock adjustment are determined from a Servo Algorithm described in .

### *PTP Slave Time Correction*

Each PTP network element maintains a PTP-free running time-of-day counter used as a basis for generating recovered clock signals and computing latencies, offsets and drift rates. The free-running counter runs on the local system clock, asynchronously to the clocks maintained by the other members of the PTP network. Correction factors are applied to the free-running clock in order to arrive at a local time value that is synchronized to the grand master clock. The PTP free-running time value consists of a 32-bit nanoseconds field, and a 48-bit seconds field. Every second the nanoseconds [31:0] field is rolled over and the 48-bit seconds [47:0] counter is incremented. The PTP network element uses the information calculated from the master clock's sync messages to perform local clock adjustments.

### *Drift Adjustment*

If the trend of slave offset values calculated from the Sync Messages continues to increase or decrease over time, the local reference clock that increments the free-running counter is operating at a rate slightly slower or faster than the master reference. You can make a drift adjustment to the free-running counter by slightly increasing or decreasing the rate at which the counter increments. Doing so locks the frequency of the counter to the master reference (syntonization). Syntonization is the adjustment of a clock signal to match the frequency, but not necessarily the phase, of another clock signal.

The drift adjustment calculation is done as follows:

Let, X = syncEventIngressTimestamp
Y = correctedMasterEventTimestamp

$$\text{drift Adjustment} = \frac{Xn - X0}{Yn - Y0}$$

Where, n = number of sync interval separating time stamps (n>0).

Syntonization may also be achieved using SyncE.

### *Offset Adjustment*

Once the drift rate has been measured and compensated for correctly, the slave clock offset should remain fairly constant at each Sync interval. Ideally, once an offset is computed and put in place, it is only rarely changed. The offset is applied to the local time value to synchronize the local time with the master's.

Ultimately, the slave clock uses the drift and offset adjusted counter to generate a usable clock signal externally. Through PTP, each slave free-running counter is both frequency and time-of-day locked to the master clock. The slave clock uses the free-running clock and phase information to generate a frequency and phase aligned clock signal traceable to the master clock.

*PTP Clock Servo Algorithm*

PTP as a protocol basically gives the timestamp values based on its messages and operation for calculating offset and drift adjustment. The problem is that the Slave clock cannot be simply corrected by setting the free running counter to a new value. If done this way, there would be time intermission or time back scenarios, and inaccurate synchronization. Thus the change to the slave clock should be brought in gradually while taking several other factors into account that affect the working of PTP:

- Packet Error (if there is an error in time stamp)
- Extended Packet Loss (an outage scenario)
- Packet Delay Variation (network load increase)

Timestamps provided by PTP are used as an input to the Servo Algorithm. This algorithm addresses all of above mentioned network impairments (PDV being the most critical one) and gives an output for the amount of adjustment that should be made to the local slave clock. 1588v2 Specification does not define the Servo part. The Servo Algorithm's goal is to achieve zero time difference between the master and the slave, and that the frequency of the slave clock and master clock are locked (meaning the ratio should be constant).

## Hybrid Networks

The Precise Time Protocol (PTP) synchronizes the network by transferring the master clock information in the form of timestamps in the PTP messages (Sync/FollowUp/DelayReq/DelayResp). In the slave clock, the clock offset is computed through the reception of PTP messages that carry master clock as timestamps.

In practice, a network could employ multiple synchronization methods in the same network. SyncE transfers the frequency of the reference clock through the ethernet's physical layer. The frequency recovered from the Synchronous Ethernet is highly accurate when compared to the frequency recovered through PTP messages. However, SyncE does not carry the Time-of-Day (TOD), or the Phase information of the clock as PTP does. Networks that employ SyncE and PTP for synchronization can leverage the accuracy of time transfer through PTP by using SyncE. Such Hybrid networks use SyncE for frequency transfer, and PTP for Phase/Time-of-Day transfer.

*PTP Time-source*

Each node in the Hybrid network recovers and transfers primary reference frequency using SyncE and phase/ToD using PTP. The PTP implementation usually includes a local reference clock at each boundary and ordinary clocks. The PTP protocol synchronizes this local reference clock to the Grandmaster clock by correcting its offset. In Hybrid networks, since frequency of the Primary reference is already available through clock recovered from SyncE, this recovered clock is used as a local reference clock. The timestamps received through the PTP event messages are processed and the Phase/ToD information is recovered.

The PTP protocol does not specify the source of time (frequency and/or Phase/ToD), or the method of clock recovery. This is left to the implementation. For boundary and ordinary slave clocks, the source of time is through PTP messages. However, since PTP messages carry timestamps that can be used to recover both frequency and phase/ToD, the implementations separate this aspect of recovery, and perform only phase/ToD recovery on the received PTP messages. The frequency recovery from the received PTP messages is not performed. The PTP implementation uses the recovered clock from SyncE as a frequency time-source.

## Supported PTP Features

The following PTP features are supported in this release:

- Ordinary Clock (slave only)
- Boundary Clock
- End-to-End Transparent Clock
- IPv4 Unicast-UDP transport
- PTP protocol 1-step and 2-step mode with end-to-end delay mechanism.
- Unicast static slaves and masters.

## Limitations of PTP

The following are the limitations of the current implementation of PTP:

- Layer 2 transport is not supported.
- IPv6-UDP transport is not supported.
- Multicast event messages are not supported.
- One-step timestamp functionality is not supported on 1G Fiber ports, 10G ports on CSR switches, and stacking ports.
- Peer-to-peer delay mechanism is not supported.
- PTP datasets are not maintained for end-to-end transparent clocks.
- Domain number cannot be assigned to end-to-end transparent clocks.
- Boundary clock does not support synchronizing clocks across multiple domains.
- Distributing clock frequency recovered from SyncE or from BITS or from a TDM port over PTP is not supported.
- Ordinary clock master (Grandmaster) mode is not supported.
- Synchronizing system time with the time recovered from PTP event messages is not supported.
- Time of Day (ToD) output and inputs are not supported.
- Unicast message negotiation on clock ports is not supported.
- PTP cannot be used if network clock source is configured as BITS.

## Configuring and Displaying PTP Clocks and Data Sets

*PTP Transparent clock*

A PTP Transparent clock updates the residence time of the PTP event packets that transit the switch. The switch supports End-to-End delay mechanism and accounts for the residence time for Sync and DelayReq packets in the switch.

- To create or delete an End-to-End Transparent clock, use the following command:

```
create network-clock ptp end-to-end-transparent
delete network-clock ptp end-to-end-transparent
```

- To add PTP capable front-panel ports for End-to-End Transparent clock, use the following command:

```
configure network-clock ptp end-to-end-transparent [add | delete] ports port_list {one-step}
```

- To display the End-to-End Transparent clock configuration, use the following command:

```
show network-clock ptp end-to-end-transparent ports port_list {detail}
```

- To enable/disable the End-to-End Transparent clock configuration on the front-panel ports, use the following commands:

```
enable network-clock ptp end-to-end-transparent ports port_list
disable network-clock ptp end-to-end-transparent ports port_list
```

*PTP Boundary/Ordinary Clocks*

A PTP Boundary or Ordinary clock synchronizes to the master clock through the reception of the PTP event packets. To reconfigure clocks to a different domain, the existing configuration must be deleted. The Boundary and Ordinary clocks can be configured to operate on a single PTP domain.

- To create or delete the Boundary or Ordinary clock, use the following commands:

  `create network-clock ptp [boundary | ordinary] {domain domain_number}`

  `delete network-clock ptp [boundary | ordinary]`

- Enable or disable the Boundary or Ordinary clock, use the following commands:

  `enable network-clock ptp [boundary | ordinary]`

  `disable network-clock ptp [boundary | ordinary]`

Note

After you enable a boundary clock, you cannot create an ordinary clock. However, you can delete the boundary clock instance and create a new one in order to change the domain number.

To create an ordinary clock instance in the switch that has the boundary clock instance enabled, delete the boundary clock instance, save the configuration and reboot the switch. After the reboot, you can create and enable the ordinary clock instance. Similarly, to create and enable a boundary clock in a switch that has an ordinary clock enabled, delete the ordinary clock instance, save the configuration and reboot the switch. After the reboot you can create and enable a boundary clock. The following message is displayed when you create the boundary clock instance in a device with no prior clock instances:

```
Warning: The ordinary clock cannot be created after enabling the boundary clock. A
delete followed by save and reboot are required to create the ordinary clock.
```

After you enable a boundary clock instance, if you delete the instance and try to create an ordinary clock instance, the above message is displayed as an error, and the ordinary clock instance is not created.

- To configure priority1 and priority2 values of the Boundary and Ordinary clock, use the following commands:

  `configure network-clock ptp [boundary | ordinary] priority1 priority`

  `configure network-clock ptp [boundary | ordinary] priority2 priority`

- To display the datasets such as Port, Time-properties and Parent of the Ordinary or Boundary clock, use the following commands:

  `show network-clock ptp ordinary {parent | port | time-property}`

  `show network-clock ptp boundary {parent | port | time-property}`

*PTP Boundary/Ordinary Clock Ports*

The Boundary and Ordinary clocks operate in 1-step protocol mode. An Ordinary clock can have at most one clock port in slave mode. A Boundary clock can have multiple clock ports in master or slave modes. Multiple unicast master or slave entries can be added to the clock ports.

- Add or remove a slave clock port to an Ordinary clock.

  `configure network-clock ptp ordinary add {vlan} vlan_name {one-step | two-step} slave-only`

  `configure network-clock ptp ordinary delete {vlan} {vlan_name}`

- Add or remove a clock port to the Boundary clock.

  `configure network-clock ptp boundary add {vlan} vlan_name {one-step | two-step} {master-only | slave-only}`

  `configure network-clock ptp boundary delete {vlan} vlan_name`

- Enable or disable a clock port in the Boundary or Ordinary clock.

  `enable network-clock ptp [boundary | ordinary] {{vlan} vlan_name}`

  `disable network-clock ptp [boundary | ordinary] {{vlan} vlan_name}`

- Display the clock ports added to the Boundary or Ordinary clock.

  `show network-clock ptp ordinary [{vlan} vlan_name | vlan all]`

  `show network-clock ptp boundary [{vlan} vlan_name | vlan all]`

  For Ordinary clocks, only unicast master entries can be added on the slave port.

  The query interval for unicast announce messages from the slave port is specified in log base 2.

- Add or remove the unicast master entries on a slave port of the Ordinary clock.

  `configure network-clock ptp ordinary add unicast-master ipv4_address {query-interval seconds_log_base_2} {vlan} vlan_name`

  `configure network-clock ptp ordinary delete unicast-master ipv4_address {vlan} vlan_name`

  The unicast master entries can be added to the slave port of the Boundary clock. The Boundary clock also support addition of unicast master entries on the port of type 'master or slave'.

- Add or remove unicast master entries on the port of the Boundary clock.

  `configure network-clock ptp boundary add unicast-master ipv4_address {query-interval vlan_name} {vlan} vlan_name`

  `configure network-clock ptp boundary delete unicast-master ipv4_address {vlan} vlan_name`

  The unicast slave entries can be added to the master port of the Boundary clock. Additionally, these entries can be added to the port of type 'master or slave'. The Ordinary clock do not support addition of unicast slave entries.

- Add or remove unicast slave entries on the port of the Boundary clock.

  `configure network-clock ptp boundary add unicast-slave ipv4_address {vlan} vlan_name`

  `configure network-clock ptp boundary delete unicast-slave ipv4_address {vlan} vlan_name`

- Display the unicast-master entries added to the Boundary or Ordinary clock port.

  ```
  show network-clock ptp boundary unicast-master [{vlan} vlan_name | vlan all]
  ```

  ```
  show network-clock ptp ordinary unicast-master [{vlan} vlan_name | vlan all]
  ```

- Display the unicast-slave entries added to the Boundary clock port.

  ```
  show network-clock ptp boundary unicast-slave [{vlan} vlan_name | vlan all]
  ```

- Display the PTP message counters for the peers added to Boundary or Ordinary clock port.

  ```
  show network-clock ptp boundary vlan [vlan_name {{ipv4_address} [unicast-master | unicast-slave]} | all] counters
  ```

  ```
  show network-clock ptp ordinary vlan [vlan_name {{ipv4_address} [unicast-master | unicast-slave]} | all] counters
  ```

- Clear the PTP message counters for the peers added to Boundary or Ordinary clock port.

  ```
  clear network-clock ptp boundary vlan [vlan_name {ipv4_address [unicast-master | unicast-slave]} | all] counters
  ```

  ```
  clear network-clock ptp ordinary vlan [vlan_name {ipv4_address [unicast-master | unicast-slave]} | all] counters
  ```

- The following properties can be configured on the clock ports added to the Boundary and Ordinary clocks:

  a. Sync message interval:

     ```
     configure network-clock ptp [boundary | ordinary] sync-interval seconds_log_base_2 {vlan} vlan_name
     ```

  b. DelayReq message interval:

     ```
     configure network-clock ptp [boundary | ordinary} delay-request-interval seconds_log_base_2 {vlan} vlan_name
     ```

  c. Announce message interval:

     ```
     configure network-clock ptp [boundary | ordinary} announce interval seconds_log_base_2 {vlan} vlan_name
     ```

  d. Announce message timeout period:

     ```
     configure network-clock ptp [boundary | ordinary] announce timeout seconds_log_base_2 {vlan} vlan_name
     ```

*PTP Clock Recovery State*

- To configure the PTP clock recovery state (servo state) in the system, use the following command:

  ```
  show network-clock ptp
  ```

  The clock recovery using PTP event messages undergoes the following servo state changes:

  - Warmup: The local reference clock is in warmup state. This state signifies that either clock recovery is not configured to use the PTP event messages or no clock recovery messages from the master have been received.
  - FastLoop: The local reference clock is being corrected and the correction is converging.
  - Bridge: The local reference clock correction has been interrupted due to changes in the clocking information in the received PTP event messages or loss of PTP event messages.

- Holdover: Prolonged loss of PTP event messages puts the local reference clock correction to the holdover state.
- Normal: The local reference clock correction has converged and the corrected clock is synchronous to the master clock information received in the PTP event messages.

## PTP Configuration Example

The following figure shows a sample configuration using the E4G-200/E4G-400 as a transparent clock, a boundary clock, and ordinary clock slaves.

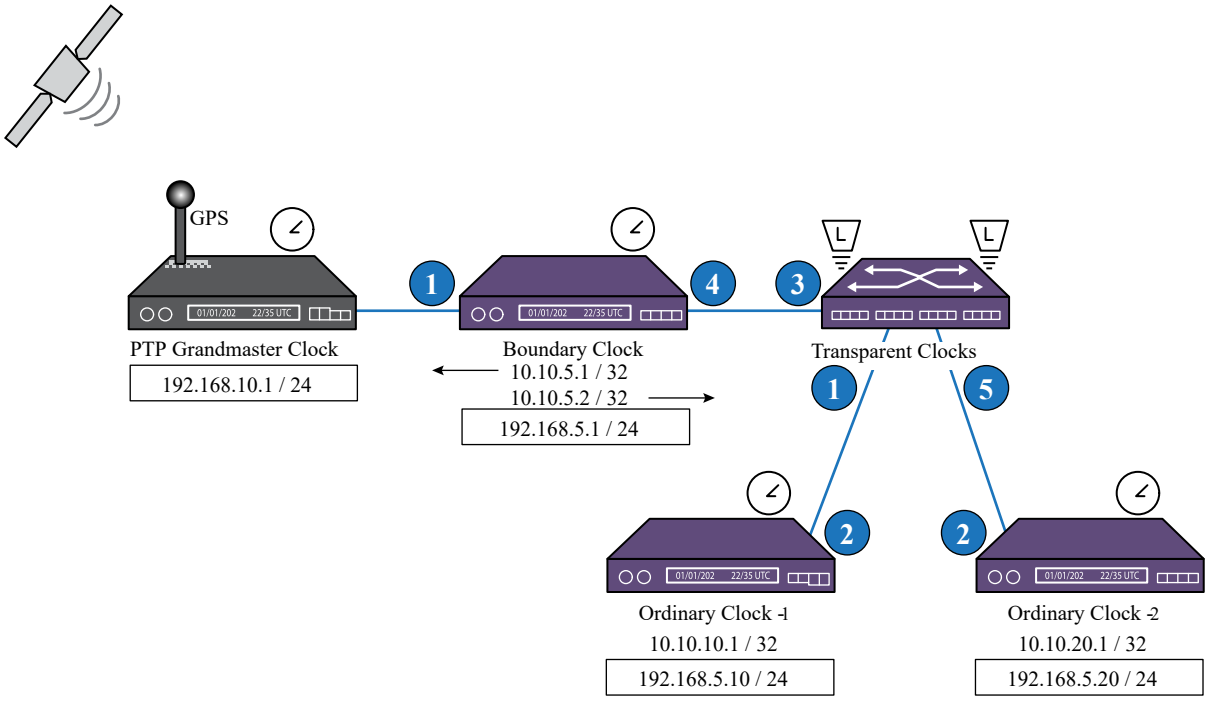



**Figure 31: PTP 1588v2 Configuration Example**

The IP address of the grandmaster and the IP address of the clock ports in each of the boundary/ordinary clocks are shown in the topology. The IP addresses that are not enclosed in the boxes are assigned to the clock ports added to boundary/ordinary clocks. The transparent clock node can be configured as L2 or L3. In the configuration example below, the transparent clock node is configured as L2.

Note

The grandmaster clock should be reachable from the boundary clock and vice versa. Similarly the ordinary clocks should be reachable from the boundary clock and vice versa. The configuration example below does not consider the provisioning methods used to achieve reachability between the switches, and only limits to the PTP 1588v2 and its associated configuration.

### *End-to-End Transparent Clock Configuration*

In this example, the transparent clock is configured as an L2 switch to transit the PTP event stream between boundary and ordinary clocks.

```
create vlan ptp_tc
```

```
configure vlan ptp_tc tag 100
configure vlan ptp_tc add port 1,3,5 tagged
create network-clock ptp end-to-end-transparent
configure network-clock ptp end-to-end-transparent add ports 1,3,5 one-step
```

**Note**

The transparent clocks accumulate the residence times on 1-step event messages by performing timestamp in ingress PHYs and in egress PHYs. For proper transparent clock operation, you must ensure in the configuration that the PTP events stream ingress and egress through physical ports that are PTP capable.

*Ordinary Clock Slave Configuration*

The ordinary clock node is configured to synchronize with the boundary clock node.

The master clock port's (loopback *VLAN*) IP address in the boundary clock node is added as "unicast-master" in the ordinary clock node.

**Note**

For PTP event messages originating from ordinary clocks (such as DelayReq), the ingress timestamp for updating the CorrectionField is done in the switch. So you must enable the End-to-End Transparent clock on all the egress ports. Ensure that you do not include the non-PTP capable ports in the configuration of possible egress ports through which the boundary is reachable.

*Ordinary Clock Slave Configuration (Node-1)*

```
create vlan lpbk
configure vlan lpbk tag 10
configure vlan lpbk ipaddress 10.10.10.1/32
enable loopback-mode lpbk
enable ipforwarding lpbk
create vlan ptp_slave
configure vlan ptp_slave tag 100
configure vlan ptp_slave add port 2 tagged
configure vlan ptp_slave ipaddress 192.168.5.10/24
enable ipforwarding ptp_slave
create network-clock ptp end-to-end-transparent
configure network-clock ptp end-to-end-transparent add port 2 one-step
create network-clock ptp ordinary
enable network-clock ptp ordinary
configure network-clock ptp ordinary add vlan lpbk one-step slave-only
configure network-clock ptp ordinary add unicast-master 10.10.5.2 lpbk
```

*Ordinary Clock Slave Configuration (Node-2)*

```
create vlan lpbk
configure vlan lpbk tag 20
configure vlan lpbk ipaddress 10.10.20.1/32
enable loopback-mode lpbk
enable ipforwarding lpbk
create vlan ptp_slave
configure vlan ptp_slave tag 100
configure vlan ptp_slave add port 2 tagged
configure vlan ptp_slave ipaddress 192.168.5.20/24
enable ipforwarding ptp_slave
create network-clock ptp end-to-end-transparent
configure network-clock ptp end-to-end-transparent add port 2 one-step
create network-clock ptp ordinary
```

```
enable network-clock ptp ordinary
configure network-clock ptp ordinary add vlan lpbk one-step slave-only
configure network-clock ptp ordinary add unicast-master 10.10.5.2 lpbk
```

*Boundary Clock Configuration*

The boundary clock node is configured to synchronize with the grandmaster node.

The grandmaster clock's IP address is added as "unicast-master" in the boundary clock node. The ptp_gm *VLAN*'s configuration depends on the grandmaster for properties such as tag, or IP.

Note

For boundary clocks, the End-to-End Transparent clock configuration must be applied on the egress ports through with the grandmaster and the ordinary clocks are reachable.

```
configure vlan lpbk-gm tag 51
configure vlan lpbk-gm ipaddress 10.10.5.1/32
enable loopback-mode lpbk-gm
enable ipforwarding lpbk-gm
create vlan lpbk-slaves
configure vlan lpbk-slaves tag 52
configure vlan lpbk-slaves ipaddress 10.10.5.2/32
enable loopback-mode lpbk-slaves
enable ipforwarding lpbk-slaves
create vlan ptp_gm
configure vlan ptp_gm tag 40
configure vlan ptp_gm add port 1 untagged
configure vlan ptp_slave ipaddress 192.168.10.5/24
enable ipforwarding ptp_gm
create vlan ptp_slaves
configure vlan ptp_slaves tag 100
configure vlan ptp_slaves add port 4 tagged
configure vlan ptp_slaves ipaddress 192.168.5.1/24
enable ipforwarding ptp_slaves
create network-clock ptp end-to-end-transparent
configure network-clock ptp end-to-end-transparent add port 1,4 one-step
create network-clock ptp boundary
enable network-clock ptp boundary
configure network-clock ptp boundary add vlan lpbk-gm one-step slave-only
configure network-clock ptp boundary add unicast-master 192.168.10.1 lpbk-gm
configure network-clock ptp boundary add vlan lpbk-slaves one-step master-only
configure network-clock ptp boundary add unicast-slave 10.10.10.1 lpbk-slaves
configure network-clock ptp boundary add unicast-slave 10.10.20.1 lpbk-slaves
```

# DWDM Optics Support

## BlackDiamond 8800 Series Switches and Summit X480 Switches

This feature allows you to configure a dense wavelength division multiplexing (DWDM) channel to a DWDM capable tunable XFP module on a port.

This provides the capability to multiplex 102x10G traffic over a single fiber. Below is a diagram of a DWDM network.

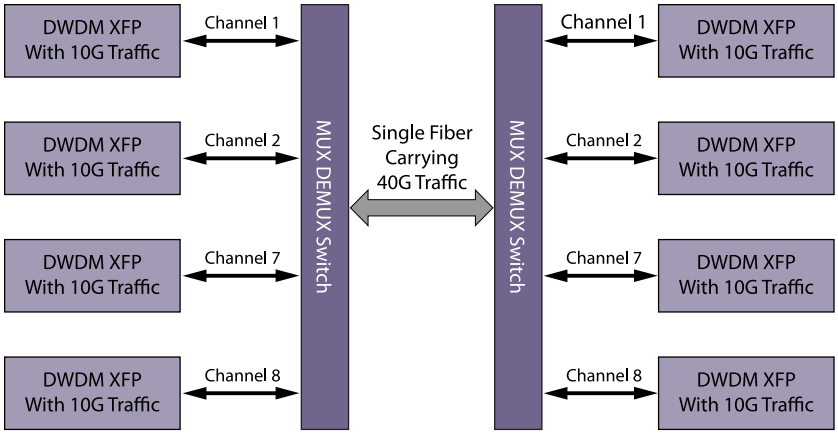



**Figure 32: Conceptual Diagram of a DWDM Network**

The feature is supported on BlackDiamond 8800 switches with 10G8Xc, 10G4Xc, 10G4Xa or 8900-10G8X-xl modules and S-10G1Xc option cards and Summit X480 switches with VIM2-10G4X modules.

For DWDM, there is no standard channel numbering specified by MSA. Extreme Networks devices support ITU standard channel numbers that range from 11 to 6150. The software can map these appropriately to the vendor-specific channels internally. For more information about the channel number and wavelength mapping, see the *Extreme Networks Pluggable Transceivers Installation Guide*.

## Limitations

This feature has the following limitations:

- Support exists only for Extreme Networks certified XFP modules.
- It may take 500-1000 ms to stabilize the channel once DWDM channel configuration is completed, meaning that the port loses its data transmission capability for that time frame. Links are dropped until the channel is stabilized. However, this is expected behavior when the physical medium is changed.
- When a tunable dense wavelength division multiplexing (TDWDM) XFP module is inserted, the software configures the default channel or the configured channel based on the existing configuration, and the link is likely to be dropped during this process.

## Configuring DWDM

- To configure DWDM specific channels on the port(s), use the following command:
  `configure port` **`all`** | *`port_list`* **`dwdm channel`** *`channel_number`*
- To configure the DWDM default channel 21 on the port(s), use the following command:
  `configure port` **`all`** | *`port_list`* **`dwdm channel none`**

## Displaying DWDM

- To display DWDM configuration, use the following command
  `show ports` {**`mgmt`** | *`port_list`* | **`tag`** *`tag`*} **`configuration`** {**`no-refresh`**}
  `show port` {**`mgmt`** | *`port_list`* | **`tag`** *`tag`*} **`information`** {**`detail`**}

- To display the channel scheme for mapping the DWDM wavelengths, use the following command

  `show dwdm channel-map { channel_first { - channel_last } } {port port_num}`

# Jumbo Frames

Jumbo frames are Ethernet frames that are larger than 1522 bytes, including four bytes used for the cyclic redundancy check (CRC).

Extreme products support switching and routing of jumbo frames at wire-speed on all ports. The configuration for jumbo frames is saved across reboots of the switch.

Jumbo frames are used between endstations that support larger frame sizes for more efficient transfers of bulk data. Both endstations involved in the transfer must be capable of supporting jumbo frames. The switch only performs IP fragmentation, or participates in maximum transmission unit (MTU) negotiation on behalf of devices that support jumbo frames.

## Guidelines for Jumbo Frames

You need jumbo frames when running the Extreme Networks VMAN implementation. If you are working on a BlackDiamond X8 series switch, BlackDiamond 8800 series switch, SummitStack, or a Summit family switch, you can enable and disable jumbo frames on individual ports before configuring VMANs. For more information on configuring VMANs, refer to VLANs on page 502.

Jumbo frame support is available on Summit family switches that are operating in a SummitStack. Refer to Displaying Port Information on page 303 for information on displaying jumbo frame status.

## Enabling Jumbo Frames per Port

You can enable jumbo frames per port on the following switches:

- BlackDiamond X8 series switches
- BlackDiamond 8000 c-, e-, xl-, and xm-series modules
- Summit family switches
- E4-G 200 and E4-G 400

When you configure VMANs on BlackDiamond X8, BlackDiamond 8800 series switches, SummitStack, and the Summit family switches, you can enable or disable jumbo frames for individual ports before configuring the VMANs.

## Enabling Jumbo Frames

Note

Some network interface cards (NICs) have a configured maximum MTU size that does not include the additional 4 bytes of CRC. Ensure that the NIC maximum MTU size is at or below the maximum MTU size configured on the switch. Frames that are larger than the MTU size configured on the switch are dropped at the ingress port.

To enable jumbo frame support, enable jumbo frames on the desired ports:

- To set the maximum jumbo frame size, use the following command:

  `configure jumbo-frame-size framesize`

  The jumbo frame size range is 1523 to 9216. This value describes the maximum size of the frame in transit (on the wire), and includes 4 bytes of CRC.

  Note
  For Tagged Packets the maximum frame size supported is 9220 bytes. It is *size*+ 4 bytes where *size* is mentioned in the `configure jumbo-frame-size size`" command.

- To set the MTU size for the *VLAN*, use the following command:

  `configure ip-mtu mtu vlan vlan_name`

- Next, enable support on the physical ports that will carry jumbo frames using the following command:

  `enable jumbo-frame ports [all | port_list]`

## Path MTU Discovery

BlackDiamond X8 series switches, BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X440, X460, X480, X670, and X770 series switches, whether or not included in a SummitStack, support path MTU discovery.

Using path MTU discovery, a source host assumes that the path MTU is the MTU of the first hop (which is known). The host sends all datagrams on that path with the “don’t fragment” (DF) bit set which restricts fragmentation. If any of the datagrams must be fragmented by an Extreme switch along the path, the Extreme switch discards the datagrams and returns an *ICMP (Internet Control Message Protocol)* Destination Unreachable message to the sending host, with a code meaning "fragmentation needed and DF set." When the source host receives the message (sometimes called a “Datagram Too Big” message), the source host reduces its assumed path MTU and retransmits the datagrams.

The path MTU discovery process ends when one of the following is true:

- The source host sets the path MTU low enough that its datagrams can be delivered without fragmentation.
- The source host does not set the DF bit in the datagram headers.

If it is willing to have datagrams fragmented, a source host can choose not to set the DF bit in datagram headers. Normally, the host continues to set DF in all datagrams, so that if the route changes and the new path MTU is lower, the host can perform path MTU discovery again.

## IP Fragmentation with Jumbo Frames

The BlackDiamond X8 series switches, BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X440, X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 series switches support fragmentation of IP packets. The above support is included whether or not the switches are present in a SummitStack.

ExtremeXOS supports the fragmenting of IP packets. If an IP packet originates in a local network that allows large packets and those packets traverse a network that limits packets to a smaller size, the packets are fragmented instead of discarded.

This feature is designed to be used in conjunction with jumbo frames. Frames that are fragmented are not processed at wire-speed within the switch fabric.

Note
Jumbo frame-to-jumbo frame fragmentation is not supported. Only jumbo frame-to-normal frame fragmentation is supported.

1. Enable jumbo frames on the incoming port.
2. Add the port to a *VLAN*.
3. Assign an IP address to the VLAN.
4. Enable ipforwarding on the VLAN.
5. Set the MTU size for the VLAN.

   `configure ip-mtu` *`mtu`* **`vlan`** *`vlan_name`*

   The ip-mtu value ranges between 1500 and 9194, with 1500 the default. However, CLI will allow the maximum limit up to 9216 considering port configuration such as tagging which influences L2 Header size. But the values greater than 9194 may lead to packet loss and hence not recommended.

Note
To set the MTU size greater than 1500, all ports in the VLAN must have jumbo frames enabled.

## IP Fragmentation within a VLAN

ExtremeXOS supports IP fragmentation within a *VLAN*. This feature does not require you to configure the MTU size.

1. Enable jumbo frames on the incoming port.
2. Add the port to a VLAN.
3. Assign an IP address to the VLAN.
4. Enable ipforwarding on the VLAN.

If you leave the MTU size configured to the default value, when you enable jumbo frame support on a port on the VLAN you will receive a warning that the ip-mtu size for the VLAN is not set at maximum jumbo frame size. You can ignore this warning if you want IP fragmentation within the VLAN, only. However, if you do not use jumbo frames, IP fragmentation can be used only for traffic that stays within the same VLAN. For traffic that is sent to other VLANs, to use IP fragmentation, all ports in the VLAN must be configured for jumbo frame support.

Note
IP fragmentation within a VLAN does not apply to Summit X440, X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 series switches (whether or not included in a SummitStack), and BlackDiamond 8000 c-, e-, xl-, and xm-series, and BlackDiamond X8 modules. The platforms that currently support fragmentation do so only for Layer 3 forwarding.

# Link Aggregation on the Switch

The link aggregation (also known as load sharing) feature allows you to increase bandwidth and availability by using a group of ports to carry traffic in parallel between switches.

Load sharing, link aggregation, and trunking are terms that have been used interchangeably in Extreme Networks documentation to refer to the same feature, which allows multiple physical ports to be aggregated into one logical port, or *LAG (Link Aggregation Group)*. Refer to IEEE 802.1AX for more information on this feature. The advantages to link aggregation include an increase in bandwidth and link redundancy.

## Link Aggregation Overview

Link aggregation, or load sharing, is disabled by default. Load sharing allows the switch to use multiple ports as a single logical port, or *LAG*.

Note
All ports in a LAG must be running at the same speed and duplex setting. Each port can belong to only one LAG.

For example, VLANs see the LAG as a single logical port. And, although you can only reference the master port of a LAG to a *STPD (Spanning Tree Domain)*, all the ports of the LAG actually belong to the specified STPD.

If a port in a load-sharing group (or LAG) fails, traffic is redistributed to the remaining ports in the LAG. If the failed port becomes active again, traffic is redistributed to include that port.

Note
Load sharing must be enabled on both ends of the link, or a network loop may result.

Link aggregation is most useful when:

- The egress bandwidth of traffic exceeds the capacity of a single link.
- Multiple links are used for network resiliency.

In both situations, the aggregation of separate physical links into a single logical link multiplies total link bandwidth in addition to providing resiliency against individual link failures.

In modular switches, ExtremeXOS supports LAGs across multiple modules, so resiliency is also provided against individual module failures.

The software supports control protocols across the LAGs, both static and dynamic. If you add the protocols (for example, EAPS, *ESRP (Extreme Standby Router Protocol)*, and so forth) to the port and then create a LAG on that port, you may experience a slight interruption in the protocol operation. To seamlessly add or delete bandwidth when running control protocols, we recommend that you create a LAG consisting of only one port. Then add your protocols to that port and add other ports as needed.

## Configurable Per Slot LAG Member Port Distribution

Rather than distribute to all active members in a *LAG*, you have two options to specify a subset of the active member ports as eligible for distribution on a per slot basis: “local slot distribution” and

"distribution port lists". The specific choice of configuration is described in the command line syntax as a **distribution-mode**. The choice of distribution mode is configurable per LAG. You may dynamically switch between distribution modes using the `configure sharing` **distribution-mode** command.

Note
This feature is not supported on standalone Summit series switches.

*Local Slot Distribution*

The "local-slot" distribution mode restricts distribution of unicast packets to the active LAG members on the same slot where the packet was received. If no active LAG members are present on the slot where the packet was received, all active LAG member ports are included in the distribution algorithm. The "local-slot" distribution mode is useful for reducing the fabric bandwidth load of a switch. Reducing fabric bandwidth may be especially important for a SummitStack, which has significantly less fabric (inter-slot) bandwidth available in comparison to chassis switches. In many SummitStack hardware configurations, the "local-slot" distribution mode may reduce the switching latency of some flows distributed to a LAG.

*Distribution Port Lists*

The "port-lists" distribution mode configures one or more LAG member ports to be eligible for unicast LAG distribution on each slot in a switch. If a slot does not have a distribution port list configured or if none of the configured member ports is active in the LAG, all active member ports are eligible for unicast distribution. The use of the "port-lists" distribution mode should be taken into consideration when adding ports to a LAG with the configure sharing command. Any newly added port on a LAG is not available for unicast distribution unless it is also added to the distribution port list of at least one slot.

*Limitations*

The distribution modes affect only the distribution of known unicast packets on a LAG. Non-unicast packets are distributed among all active members of a LAG.

## Link Aggregation and Software-Controlled Redundant Ports

If you are configuring software-controlled redundant ports and link aggregation together, the following rules apply:

- You must unconfigure the software-controlled redundant ports before either configuring or unconfiguring load sharing.
- The entire *LAG* must go down before the software-controlled redundant port takes effect.

## Dynamic Versus Static Load Sharing

ExtremeXOS software supports two broad categories of load sharing, or link aggregation.

*Dynamic Load Sharing*

Dynamic load sharing includes the Link Aggregation Control Protocol (LACP) and Health Check Link Aggregation. The Link Aggregation Control Protocol is used to dynamically determine if link aggregation is possible and then to automatically configure the aggregation. LACP is part of the IEEE

802.1AX standard and allows the switch to dynamically reconfigure the link aggregation groups (LAGs). The *LAG* is enabled only when LACP detects that the remote device is also using LACP and is able to join the LAG. Health Check Link Aggregation is used to create a link aggregation group that monitors a particular TCP/IP address and TCP port.

*Static Load Sharing*

Static load sharing is a grouping of ports specifically configured to load share. The switch ports at each end must be specifically configured as part of a load-sharing group.

Note
The platform-related load sharing algorithms apply to LACP (as well as static load sharing).

## Load-Sharing Algorithms

Load-sharing, or link aggregation, algorithms select an egress link for each packet forwarded to egress the *LAG*.

The ExtremeXOS software supports the following types of load sharing algorithms:

- Address based—The egress link is chosen based on packet contents.
- Port-based—The egress link is chosen based on the key assigned to the ingress port.

The ExtremeXOS software provides multiple addressed-based algorithms. For some types of traffic, the algorithm is fixed and cannot be changed. For other types of traffic, you can configure an algorithm. Algorithm selection is not intended for use in predictive traffic engineering.

The following sections describe the algorithm choices for different platforms:

- #unique_469
- BlackDiamond and SummitStack Link Aggregation Algorithms on page 248
- Link Aggregation Standard and Custom Algorithms on page 248

Note
Always reference the master logical port of the load-sharing group when configuring or viewing *VLAN*s. VLANs configured to use other ports in the LAG will have those ports deleted from the VLAN when link aggregation is enabled.

*Address-based Load Sharing*

All ExtremeXOS platforms support address-based load sharing.

The following are the types of traffic to which addressed-based algorithms apply and the traffic components used to select egress links:

- Layer 2 frames and non-IP traffic—The source and destination MAC addresses.
- IPv4 and IPv6 packets
  - L2 algorithm—Layer 2 source and destination MAC addresses. Available on SummitStack and all Summit family switches.
- Broadcast, multicast, and unknown unicast packets (not configurable)—Depends on traffic type:
  - IPv4 and IPv6 packets—The source and destination IP addresses.
  - Non-IP traffic—The source and destination MAC addresses.

You can control the field examined by the switch for address-based load sharing when the load-sharing group is created by using the following command:

```
enable sharing port grouping port_list {algorithm [address-based {L2 |
L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
```

#### BlackDiamond and SummitStack Link Aggregation Algorithms

The following are the types of traffic to which addressed-based algorithms apply and the traffic components used to select egress links:

- IPv4 and IPv6 packets—When no BlackDiamond 8900 series modules are installed in a modular switch or SummitStack, load sharing is based on the configured options supported on each platform:
  - L2 algorithm—Layer 2 source and destination MAC addresses. Available on BlackDiamond 8800 series switches, SummitStack, and all Summit family switches.
  - L3 algorithm—Layer 3 source and destination IP addresses. Available on BlackDiamond 8800 series switches, Summit family switches, and SummitStack.
  - L3_L4 algorithm—Layer 3 and Layer 4, the combined source and destination IP addresses and source and destination TCP and UDP port numbers. Available on Summit family switches, BlackDiamond 8000 a-, c-, and e-series modules.
- IPv4 and IPv6 packets—When BlackDiamond 8900 series modules are installed in a BlackDiamond 8800 series switch, load sharing on all other module or switch types is based on the combined source and destination IP addresses and source and destination TCP and UDP port numbers.
- Non-IP traffic—The source and destination MAC addresses.

Note

On platforms such as Summit X670, X670v, X480, X460, and Black Diamond 8900 series I/O modules, load sharing based on inner L3 fields in the *MPLS* terminated packet are not supported and the packets will be forwarded as per L2 hashing.

You can control the field examined by the switch for address-based load sharing when the load-sharing group is created by using the following command:

```
enable sharing port grouping port_list {algorithm [address-based {L2 |
L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
```

#### Link Aggregation Standard and Custom Algorithms

- BlackDiamond X8 Series Switches, BlackDiamond 8900-series modules (xl, xm, and c variants), but not the 8800 series (for example, G48Tc, G48Te2, etc.).
- SummitStack, and Summit X460, X-450-G2, X460-G2, X480, X670, X670-G2, X770, E4G-200, E4-G400 (but not X430 & X440) series switches support address-based load sharing.

  These platforms do support two types of algorithms:
  - Standard algorithms, which are also supported by other switch platforms.
  - Custom algorithms, which use newer switch hardware to offer additional options, including the ability to evaluate IP address information from the inner header of an IP-in-IP or GRE tunnel packet.

Standard Algorithms

The following are the types of traffic to which standard addressed-based algorithms apply and the traffic components used to select egress links:

- Layer 2 frames, Unknown unicast packet and non-IP traffic—The source and destination MAC addresses.
- IPv4 and IPv6 packets—Load sharing is based on the configured options supported on each platform:
  - L2 algorithm—Layer 2 source and destination MAC addresses.
  - L3 algorithm—Layer 3 source and destination IP addresses.
  - L3_L4 algorithm—Layer 3 and Layer 4, the combined source and destination IP addresses and source and destination TCP and UDP port numbers.
- *MPLS* packets—The source and destination MAC addresses.

Custom Algorithms

The following are the types of traffic to which custom addressed-based algorithms apply and the traffic components used to select egress links:

- Non-IP Layer 2—Uses the *VLAN* ID, the source and destination MAC addresses, and the ethertype.
- IPv4 packets—Uses IP address information from an IP header, or for tunneled packets, the custom algorithm always uses the inner header of an IP-in-IP or GRE tunnel packet. The configuration options are:
  - The source and destination IPv4 addresses and Layer 4 port numbers (default)
  - The source IP address only,
  - The destination IP address only
  - The source and destination IP addresses
- IPv6 packets—Uses the source and destination IPv6 addresses and Layer 4 port numbers.
- MPLS packets—Uses the top, second, and reserved labels and the source and destination IP addresses.

Note

In a switch having at least one *LAG* group with custom algorithm, the egress port for unknown unicast packets across all LAG groups in switch will be decided based on Layer 3 source and destination IP address.

The following command allows you to enable load sharing and select either a standard algorithm or specify that you want to use a custom algorithm:

```
enable sharing port grouping port_list {algorithm [address-based {L2 |
L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
```

If you choose the **custom** option when you enable load sharing, you can use the following command to select a custom load sharing algorithm:

```
configure sharing address-based custom [ipv4 [L3-and-L4 | source-only |
destination-only | source-and-destination] | hash-algorithm [xor |
crc-16]]
```

The **hash-algorithm** option controls how the source information (such as an IP address) is used to select the egress port. The **xor** hash algorithm guarantees that the same egress port is selected for

traffic distribution based on a pair of IP addresses, Layer 4 ports, or both, regardless of which is the source and which is the destination.

**Note**
Use of the *ACL (Access Control List)* redirect-port-no-sharing *port* action overrides any load-sharing algorithm hash that is generated based on the lookup results. For more information on this action, see LAG Port Selection on page 708.

### *Port-based Load Sharing*

A port-based load sharing key is used to determine the index of the selected aggregator port where the list of aggregator ports is organized as an array sorted by increasing front panel port number with a zero-based index. The index of the selected aggregator port is equal to the key value module of the number of ports in the aggregator.

**Note**
This feature is supported only on BlackDiamond X8, Summit X670, X770, X670-G2 and X460-G2 platforms.

index = key value % N

where N = the number of ports in the aggregator

The resulting behavior is that ports with a key value of 0 distribute to the lowest numbered port in an aggregator, ports with a key value of 1 distribute to the second lowest numbered port in an aggregator, etc.

#### Example

A port-based load sharing group contains aggregator ports 2,4,6 and 8. If the zero-based load sharing key 7 is assigned to port 1, then traffic received on port 1 and forwarded to the group will be transmitted on port 8 according to the following calculation:

7 (key) modulo 4 (number of ports in the aggregator) = 3 (index), which corresponds to port 8 which has zero-based index 3 in the sorted array of aggregator ports as shown in the following table:

**Table 30: Aggregator Ports Array**

| Index | Keys | Member port |
|---|---|---|
| 0 | 0, 4, 8, 12 | 2 |
| 1 | 1, 5, 9, 13 | 4 |
| 2 | 2, 6, 10, 14 | 6 |
| 3 | 3, 7, 11, 15 | 8 |

#### Port-based Load Sharing Limitations

When considering the selection of a *LAG* algorithm, even distribution over member ports is usually the goal. Full utilization of the LAG's bandwidth requires even distribution. In the absence of even distribution, a single member port may become oversubscribed while other member ports are undersubscribed resulting in traffic loss when the LAG, viewed as an aggregate of member ports, is undersubscribed. LAGs distribute best when the diversity of flows destined for the LAG is large relative

to the number of ports in the aggregator. For example, when many thousands of L2 flows are destined to a LAG using the “L2” algorithm, distribution on the LAG is typically good (even). Since the number of ports which will switch to a LAG is unlikely to be much larger (orders of magnitude) than the number of ports in the aggregator, extra care may be required from a network administrator when configuring and/or provisioning a switch using port-based LAGs.

## LACP

Note
LACP fails over hitlessly in the event of a failover to a duplicate MSM/MM in a modular switch.

You can run the Link Aggregation Control Protocol (LACP) on Extreme Networks devices. LACP enables dynamic load sharing and hot standby for link aggregation links, in accordance with the IEEE 802.3ad standard. All third-party devices supporting LACP run with Extreme Networks devices.

The addition of LACP provides the following enhancements to static load sharing, or link aggregation:

- Automatic configuration
- Rapid configuration and reconfiguration
- Deterministic behavior
- Low risk of duplication or misordering

After you enable load-sharing, the LACP protocol is enabled by default. You configure dynamic link aggregation by first assigning a primary, or logical, port to the group, or *LAG* and then specifying the other ports you want in the LAG.

LACP, using an automatically generated key, determines which links can aggregate. Each link can belong to only one LAG. LACP determines which links are available. The communicating systems negotiate priority for controlling the actions of the entire trunk (LAG), using LACP, based on the lowest system MAC number. You can override this automatic prioritization by configuring the system priority for each LAG.

After you enable and configure LACP, the system sends PDUs (LACPDUs) on the LAG ports. The LACPDUs inform the remote system of the identity of the sending system, the automatically generated key of the link, and the desired aggregation capabilities of the link. If a key from a particular system on a given link matches a key from that system on another link, those links are aggregatable. After the remote system exchanges LACPDUs with the LAG, the system determines the status of the ports and whether to send traffic on which ports.

Among those ports deemed aggregatable by LACP, the system uses those ports with the lowest port number as active ports; the remaining ports aggregatable to that LAG are put into standby status. Should an active link fail, the standby ports become active, also according to the lowest port number. (See Configuring LACP on page 259 for the number of active and standby LACP links supported per platform.)

All ports configured in a LAG begin in an unselected state. Based on the LACPDUs exchanged with the remote link, those ports that have a matching key are moved into a selected state. If there is no matching key, the ports in the LAG remain in the unselected state.

However,if more ports in the LAG are selected than the aggregator can handle because of the system hardware, those ports that fall out of the hardware's capability are moved into standby state. The lowest numbered ports are the first to be automatically added to the aggregator; the rest go to standby. As the name implies, these ports are available to join the aggregator if one of the selected ports should go offline.

You can configure the port priority to ensure the order that ports join the aggregator. However, that port must first be added to the LAG before you can configure the LACP settings. Again, if more than one port is configured with the same priority, the lowest-numbered port joins the aggregator first.

After the ports in the LAG move into the selected state, LACP uses the mux portion of the protocol to determine which ports join the aggregator and can collect and distribute traffic. A few seconds after a port is selected, it moves into the mux state of waiting, and then into the mux state of attached. The attached ports then send their own LACP sync messages announcing that they are ready to receive traffic.

The protocol keeps sending and receiving LACPDUs until both sides of the link have echoed back each other's information; the ends of the link are then considered synchronized. After the sync messages match up on each end, that port is moved into the aggregator (into the mux state of collecting-distributing) and is able to collect and distribute traffic.

The protocol then enables the aggregated link for traffic and monitors the status of the links for changes that may require reconfiguration. For example, if one of the links in a LAG goes down and there are standby links in that LAG, LACP automatically moves the standby port into selected mode and that port begins collecting and distributing traffic.

The marker protocol portion of LACP ensures that all traffic on a link has been received in the order in which it was sent and is used when links must be dynamically moved between aggregation groups. The Extreme Networks LACP implementation responds to marker frames but does not initiate these frames.

Note

Always verify the LACP configuration by issuing the `show ports sharing` command; look for the ports specified as being in the aggregator. You can also display the aggregator count by issuing the `show lacp lag` command.

You can configure additional parameters for the LACP protocol and the system sends certain *SNMP* traps in conjunction with LACP. The system sends a trap when a member port is added to or deleted from an aggregator.

The system now detects and blocks loopbacks; that is, the system does not allow a pair of ports that are in the same LAG but are connected to one another by the same link to select the same aggregator. If a loopback condition exists between two ports, they cannot aggregate. Ports with the same MAC address and the same admin key cannot aggregate; ports with the same MAC address and a different admin key can belong to the same LAG.

The system sends an error message if a LAG port is configured and up but still not attached to the aggregator or in operation within 60 seconds. Use the `show lacp member-port` *`port`* `detail` command to display the churn on both sides of the link. If the churn value is shown as True in the display, check your LACP configuration. The issue may be either on your end or on the partner link, but you should check your configuration. The display shows as True until the aggregator forms, and then it changes to False.

A LAG port moves to expired and then to the defaulted state when it fails to receive an LACPDU from its partner for a specified time. You can configure this timeout value as long, which is 90 seconds, or short, which is three seconds; the default is long. (In ExtremeXOS 11.3, the timeout value is not configurable and is set as long, or 90 seconds.) Use the `show lacp lag group-id` detail command to display the timeout value for the LAG.

There are two LACP activity modes: active and passive. In LACP active mode, the switch periodically sends LACPDUs; in passive mode, the switch sends LACPDUs only when it receives one from the other end of the link. The default is active mode. (In ExtremeXOS 11.3, the mode is not configurable; it is always active mode.) Use the `show lacp lag group-id detail` command to display the LACP mode for the LAG.

Note

One side of the link must be in active mode in order to pass traffic. If you configure your side in the passive mode, ensure that the partner link is in LACP active mode.

A LAG port moves into a defaulted state after the timeout value expires with no LACPDUs received for the other side of the link. You can configure whether you want this defaulted LAG port removed from the aggregator or added back into the aggregator. If you configure the LAG to remove ports that move into the default state, those ports are removed from the aggregator and the port state is set to Unselected. The default configuration for defaulted ports is to be removed, or deleted, from the aggregator. (In ExtremeXOS version 11.3, defaulted ports in the LAG are always removed from the aggregator; this is not configurable.)

Note

To force the LACP trunk to behave like a static sharing trunk, use the `configure sharing port` **`lacp defaulted-state-action`** [**`add`** | **`delete`**] command to add ports to the aggregator.

If you configure the LAG to add the defaulted port into the aggregator, the system takes inventory of the number of ports currently in the aggregator. If there are fewer ports in the aggregator than the maximum number allowed, the system adds the defaulted port to the aggregator (port set to selected and collecting-distributing). If the aggregator has the maximum ports, the system adds the defaulted port to the standby list (port set to standby). Use the `show lacp lag group-id` {**`detail`**} command to display the defaulted action set for the LAG.

Note

If the defaulted port is assigned to standby, that port automatically has a lower priority than any other port in the LAG (including those already in standby).

### *LACP Fallback*

Preboot Execution Environment (PXE) is a client/server environment that allows workstations to boot from the server before their full operating system is running. PXE images are too small to take advantage of LACP functionality, and therefore it is up to the administrator to statically configure the switch for correct connectivity. This also means that after the full operating is fully running the switch needs to be reconfigured back for LACP. The LACP fallback feature introduced in ExtremeXOS 21.1 allows for this process to be automated.

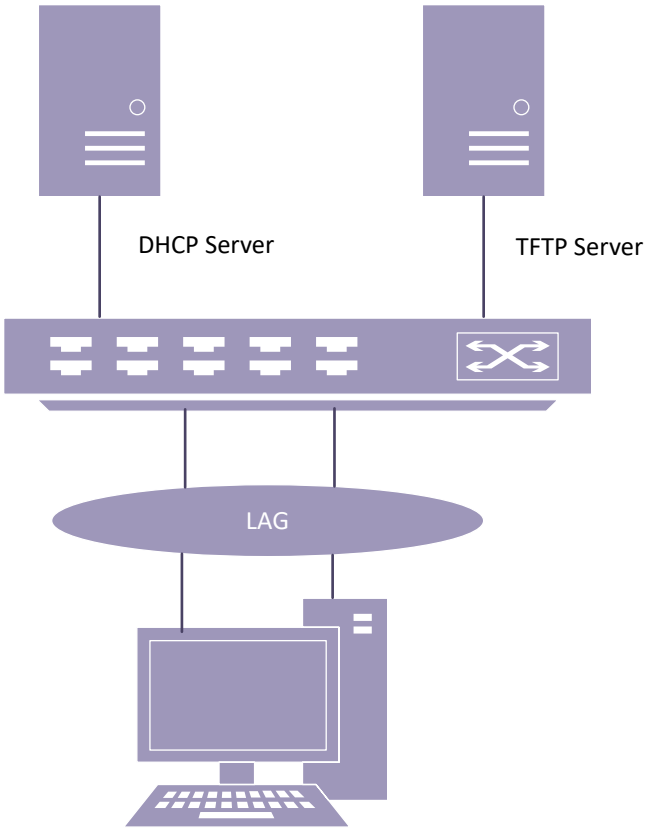



**Figure 33: LACP Fallback**

The LACP fallback feature allows you to select a single port which will be automatically added to the aggregator if LACP PDUs are not seen on any of the member ports within a specified period of time. If LACP PDUs are exchanged before this timeout expires, then the aggregator is formed using traditional means. If LACP PDUs are not received, then an active port with the lowest priority level is automatically added to the aggregator (enters fallback state). If ports have the same priority value then the lowest port number on the lowest slot number is chosen.

The selected port stays in fallback mode until fallback is disabled or until LACP PDUs start being received on any of the member ports, at which point the old aggregator is removed and a new one is selected based on information propagated in the LACP PDUs. The new fallback port may also be reelected if the existing fallback port changes state; such as port priority change, link bounce, or port enable/disable.

## Link Aggregation Minimum Active Links

The *LAG* Minimum Links feature allows you to configure a value for the minimum number of active links to keep the entire LAG up. For example, for a LAG consisting of 4 ports and the minimum links set to 2, at least 2 links must be up for the LAG to be up. When the LAG falls below 2 active links, the entire LAG

is brought down, and all applications using this LAG will be informed that the LAG is down. Currently, the implicit minimum link value is 1, which means if there is at least 1 link up the entire LAG will stay up.

To configure this ability, use the #unique_479 command.

Both static and LACP LAGs are supported. In the case of static LAG, we check if the number of active physical member port links is greater than or equal to the user-configured minimum link value. If so, the LAG remains up. If the number of active physical member port links falls below the configured minimum link value, the static LAG is brought DOWN and all applications receive a Link-Down message for this LAG. As soon as the number of Active physical member ports equals or exceed the configured minimum link value, the static LAG is brought UP and all applications receive a Link-Up message for this LAG.

In the case of LACP, we keep track of the how many member ports LACP has requested to be added to the LAG. After successfully negotiating with its peer, LACP will send a request to add a member port to the LAG. When the number of member ports LACP has requested to be added to the LAG drops below the configured minimum link value, the LACP LAG will be brought down and all applications will receive a Link-Down message for this LAG. As soon LACP has added enough member ports such that the total member ports equals or exceed the configured minimum link value, the LAG is brought up and all applications receive a Link-Up message for this LAG.

Both static and LACP LAGs can be used with *MLAG (Multi-switch Link Aggregation Group)* on either the ISC or the MLAG ports. We will verify that the minimum links feature behaves properly for static and LACP LAGs used by MLAG.

Note

For static LAGs, ensure that both ends of the LAG use the same value for minimum links. If the minimum links value differs, one side may see the LAG as down, where the other side may see the LAG as up.

## Health Check Link Aggregation

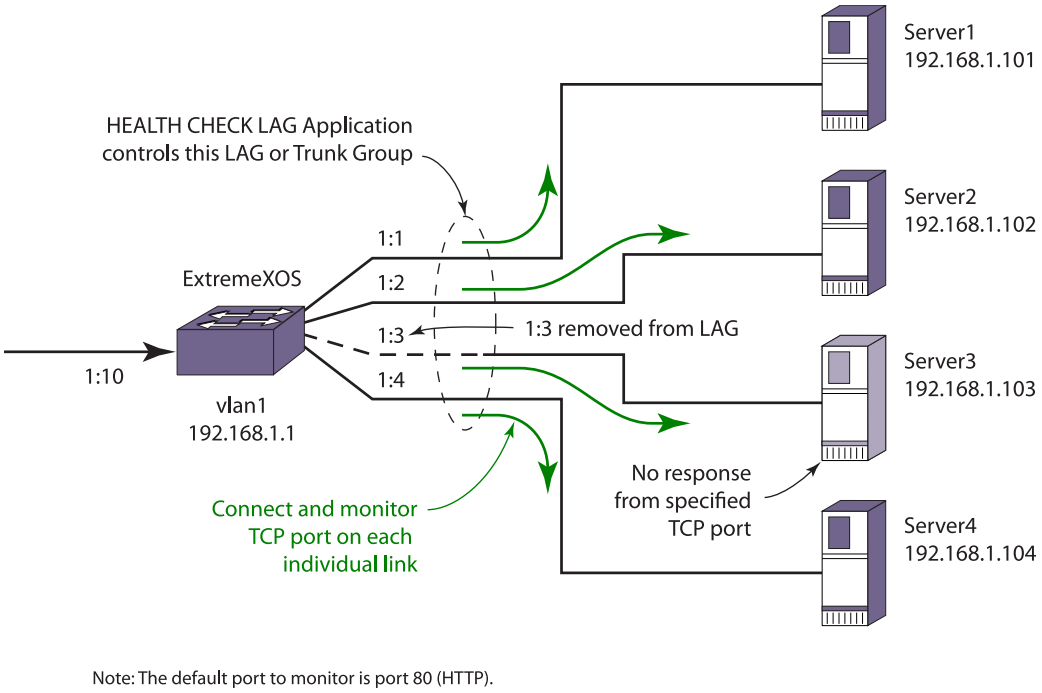
The Health Check *LAG* application allows you to create a link aggregation group where individual member links can monitor a particular TCP/IP address and TCP port.

When connectivity to the TCP/IP address and TCP port fails, the member link is removed from the link aggregation group.

Establishing the status of a TCP connectivity is based on standard TCP socket connections. As long as the switch can establish a TCP connection to the target switch and TCP port, the connection is considered up. The TCP connection will retry based on the configured frequency and miss settings.

A typical use case for this application is when a user wishes to connect each member link to a Security Server to validate traffic. Each member link of the Health Check LAG is connected to an individual Security Server. The LAG is added to a *VLAN* on the same subnet as the Security Server IP addresses they wish to monitor. Each member port is configured to monitor a particular IP address and TCP port. The Health Check LAG application attempts to do a TCP connect to each IP/TCP port through each member port. The Health Check LAG, by virtue of the sharing algorithm, will load balance traffic across the member links. If a TCP connection cannot be established through the member link, the port is removed from the aggregator and traffic through that particular link is redistributed to the other LAG member links.

The following figure displays an example of a Health Check LAG.




Note: The default port to monitor is port 80 (HTTP).

**Figure 34: Health Check LAG Example**

## Guidelines for Load Sharing

### *Load Sharing Guidelines*

For Summit Family Switches and SummitStack the following rules apply to load sharing.

- A static *LAG* can contain up to eight ports.
- An LACP LAG can contain twice the number of ports as a static LAG. The maximum number of selected links is the same as the limit for a static LAG. The remaining links are standby links.
- A Health Check LAG can contain the same number of ports as a static LAG.
- You can configure only the address-based load-sharing algorithm as described in the following sections:
  - #unique_469
  - BlackDiamond and SummitStack Link Aggregation Algorithms on page 248
  - Link Aggregation Standard and Custom Algorithms on page 248
- The maximum number of LAGs for Summit family switches is 128.

**Note**

See Configuring LACP on page 259 for the maximum number of links, selected and standby, per LACP.

The limits on the number of ports per LAG are different for X670. The following rules apply to load sharing on the X670.

- A static LAG can contain up to 32 ports when configured to use the custom address-based algorithm. For all other algorithms, a static LAG can contain up to 16 ports.
- An LACP LAG configured to use the custom address-based algorithm can contain up to 64 ports per LAG, which includes up to 32 selected links and 32 standby links. For all other algorithms, an LACP LAG can contain up to 32 ports per LAG, which includes up to 16 selected links and 16 standby links. A static LAG in a SummitStack consisting entirely of X670s can contain up to 64 ports when configured to use the custom address-based algorithm.

#### Guidelines for the Summit X460-G2, X670-G2, and X770 Switches

For the Summit X460-G2, X670-G2, and X770 switches, the following rules apply to load sharing.

- A static LAG can contain up to 32 ports when configured to use the L2,L3,L3_L4 or custom algorithm.
- For all the algorithms, LACP LAG can contain up to 64 ports per LAG, which includes up to 32 selected links and 32 standby links.
- A SummitStack consisting entirely of X770s can contain up to 64 ports for all algorithms.

For BlackDiamond X8 Series Switches and BlackDiamond 8800 Series Switches the following rules apply to load sharing.

BlackDiamond X8

- The maximum number of LAGs is 384.
- A static LAG can contain up to 64 ports.
- An LACP LAG can contain up to 128 links per LAG, which includes up to 64 selected links and 64 standby links.

BlackDiamond 8800

- A static LAG can contain up to 8 ports.
- An LACP LAG can contain up to 16 links per LAG, which includes up to 8 selected links and 8 standby links.
- A Health Check LAG can contain the same number of ports as a static LAG.
- You can configure only the address-based load-sharing algorithm as described in #unique_483.
- The maximum number of LAGs is 128. See Configure LACP for the maximum number of links, selected and standby, per LACP.

### *Load Sharing Rules and Restrictions for All Switches*

Additionally, the following rules apply to load sharing on all switches:

- The ports in the *LAG* do not need to be contiguous. The ports in the LAG can span multiple modules in a modular switch, or multiple stacked switches.
- A LAG that spans multiple modules must use ports that have the same maximum bandwidth capability, with one exception—you can mix media type on 1 Gbps ports.

- On both ingress and egress direction on BlackDiamond 8800 series switches and Summit family switches, when you configure an *ACL* to a LAG group, you must configure each of the member ports exclusively.

Note
Due to a hardware limitation, *MPLS* terminated traffic can not be load shared across member ports of BlackDiamond 8000 Series modules. Traffic will only be forwarded through the master port.

## Configuring Switch Load Sharing

To set up a switch for load sharing, or link aggregation, among ports, you must create a load-sharing group of ports, also known as a *LAG*.

Note
See Guidelines for Load Sharing on page 256 for specific information on load sharing for each specific device.

The first port in the load-sharing group is configured to be the master logical port. This is the reference port used in configuration commands and serves as the LAG group ID. It can be thought of as the logical port representing the entire port group.

All the ports in a load-sharing group must have the same exact configuration, including autonegotiation, duplex setting, *ESRP* host attach or don't-count, and so on. All the ports in a load-sharing group must also be of the same bandwidth class.

### *Creating and Deleting Load Sharing Groups*

To define a load-sharing group, or *LAG*, you assign a group of ports to a single, logical port number.

- To enable or disable a load-sharing group, use the following command

```
enable sharing port grouping port_list {algorithm [address-based {L2 |
L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
disable sharing port
```

Note
All ports that are designated for the LAG must be removed from all VLANs prior to configuring the LAG.

### *Adding and Deleting Ports in a Load-Sharing Group*

Ports can be added or deleted dynamically in a load-sharing group, or Load sharing group (*LAG*).

Note
When aggregator membership changes on a LAG, both redistribution of flows, as well as some traffic loss, may be expected.

- To add or delete ports from a load-sharing group, use the following commands:

```
configure sharing port add ports port_list
```

```
configure sharing port delete ports port_list
```

Note

See Configuring LACP on page 259 for the maximum number of links, selected and standby, per LACP.

### *Configuring the Load Sharing Algorithm*

For some traffic on selected platforms, you can configure the load sharing algorithm. This is described in Load-Sharing Algorithms on page 247.

- The commands for configuring load sharing algorithms are:

  On SummitStack and all Summit family switches:

  ```
  enable sharing port grouping port_list {algorithm [address-based {L2 | L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
  ```

  On BlackDiamond 8900 series modules and SummitStack:

  ```
  enable sharing port grouping port_list {algorithm [address-based {L2 | L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
  ```

  On all platforms:

  ```
  configure sharing address-based custom [ipv4 [L3-and-L4 | source-only | destination-only | source-and-destination] | hash-algorithm [xor | crc-16]]
  ```

### *Configuring LACP*

To configure LACP, you must, again, first create a *LAG*. The first port in the LAG serves as the logical port for the LAG. This is the reference port used in configuration commands. It can be thought of as the logical port representing the entire port group, and it serves as the LAG Group ID.

1. Create a LAG using:

   ```
   enable sharing port grouping port_list {algorithm [address-based {L2 | L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}
   ```

   The port you assign using the first parameter becomes the logical port for the link aggregation group and the LAG Group ID when using LACP. This logical port must also be included in the port list of the grouping itself.
2. If you want to override the default prioritization in LACP for a specified LAG, use:

   ```
   configure sharing port lacp system-priority priority
   ```

   This step is optional; LACP handles prioritization using system MAC addresses.
3. Add or delete ports to the LAG as desired using:

   ```
   configure sharing port add ports port_list
   ```
4. Override the ports selection for joining the LAG by configuring a priority for a port within a LAG by using the command:

   ```
   configure lacp member-port port priority port_priority
   ```

5. Change the expiry timer using:

   `configure sharing port lacp timeout [long | short]`

   The default value for the timeout is long, or 90 seconds.
6. Change the activity mode using:

   `configure sharing port lacp activity-mode [active | passive]`

   The default value for the activity mode is active.
7. Configure the action the switch takes for defaulted LAG ports.

   `configure sharing port lacp defaulted-state-action [add | delete]`

   The default value for defaulted LAG ports is delete the default ports.

Note

Always verify the LACP configuration by issuing the `show ports sharing` command; look for the ports listed as being in the aggregator.

### *Configuring Health Check Link Aggregation*

To configure Health Check link aggregation you must first create a *LAG*. One port in the LAG serves as the logical port for the LAG and is the reference port used in configuration commands.

When you create the LAG, no monitoring is initially configured. The LAG is created in the same way that a static LAG is created and if no monitoring is ever created, this LAG behaves like a static LAG.

1. Create a LAG using:

   `enable sharing port grouping port_list {algorithm [address-based {L2 | L3 | L3_L4 | custom} | port-based }]} {lacp | health-check}`

   The port you assign using the *port* parameter becomes the logical port for the link aggregation group and the LAG Group ID when using Health Check link aggregation. This logical port must also be included in the port list of the grouping itself.
2. Configure monitoring for each member port using:

   `configure sharing health-check member-port port add tcp-tracking IP Address {tcp-port TC Port frequency sec misses count}`

   If the TCP-port, frequency, or misses are not specified, the defaults described in the *ExtremeXOS 16.2 Command Reference Guide* are used.

3. Add the LAG to a *VLAN* whose subnet is the same as the configured tracking IP addresses.
   `configure vlan vlan add port lag port [tagged | untagged]`

   All of the tracking IP addresses must be in the same subnet in which the LAG belongs.

   Note
   *VLAN*s to which Health Check LAG ports are to be added must be configured in loopback mode. This is to prevent the VLAN interface from going down if all ports are removed from the Health Check LAG. In a normal LAG when all ports are removed from the aggregator, the trunk is considered DOWN. As a consequence, if this were the only port in the VLAN, the VLAN interface would be brought DOWN as well. In the Health Check LAG situation, this would cause the TCP monitoring to fail because the L3 vlan interface used by TCP monitoring would no longer send or receive TCP data.

### *Modifying Configured Health Check LAG*

The following commands are used to modify the configured Health Check *LAG*.

1. Delete the monitoring configuration for a member port using the following command:
   `configure sharing health-check member-port port delete tcp-tracking IP Address {tcp-port TC Port}`
2. Enable or disable monitoring for a member port in the Health Check LAG using the following command:
   `configure sharing health-check member-port port [disable | enable] tcp-tracking`

## Load-Sharing Examples

This section provides examples of how to define load sharing, or link aggregation, on stand-alone and modular switches, as well has defining dynamic link aggregation.

### *Load Sharing on a Stand-alone Switch*

The following example defines a static load-sharing group that contains ports 9 through 12, and uses the first port in the group as the master logical port 9:

```
enable sharing 9 grouping 9-12
```

In this example, logical port 9 represents physical ports 9 through 12.

When using load sharing, you should always reference the master logical port of the load-sharing group (port 9 in the previous example) when configuring or viewing VLANs; the logical port serves as the *LAG* Group ID. *VLAN*s configured to use other ports in the load-sharing group will have those ports deleted from the VLAN when load sharing becomes enabled.

### *Cross-Module Load Sharing on a Modular Switch or SummitStack*

The following example defines a static load-sharing group on modular switches that contains ports 9 through 12 on slot 3, ports 7 through 10 on slot 5, and uses port 7 in the slot 5 group as the primary logical port, or *LAG* Group ID:

```
enable sharing 5:7 grouping 3:9-3:12, 5:7-5:10
```

In this example, logical port 5:7 represents physical ports 3:9 through 3:12 and 5:7 through 5:10.

When using load sharing, you should always reference the LAG Group ID of the load-sharing group (port 5:7 in the previous example) when configuring or viewing *VLAN*s. VLANs configured to use other ports in the load-sharing group will have those ports deleted from the VLAN when load sharing becomes enabled.

Address-based load sharing can also span modules.

*Single-Module Load Sharing on a Modular Switch or SummitStack*

The following example defines a static load-sharing, or link aggregation, group that contains ports 9 through 12 on slot 3 and uses the first port as the master logical port 9, or *LAG* group ID:

```
enable sharing 3:9 grouping 3:9-3:12
```

In this example, logical port 3:9 represents physical ports 3:9 through 3:12.

*LACP Example*

The following configuration example:

- Creates a dynamic *LAG* with the logical port (LAG Group ID) of 10 that contains ports 10 through 12.
- Sets the system priority for that LAG to 3.
- Adds port 5 to the LAG.

```
enable sharing 10 grouping 10-12 lacp
configure sharing 10 lacp system-priority 3
configure sharing 10 add port 5
```

*Health Check LAG Example*

The following example creates a Health Check *LAG* of 4 ports:

```
create vlan v1
configure v1 ip 192.168.1.1/24
enable sharing 5 grouping 5-8 health-check
enable loopback-mode v1
configure v1 add port 5
configure sharing health-check member-port 5 add track-tcp 192.168.1.101 tcp-port 8080
configure sharing health-check member-port 6 add track-tcp 192.168.1.102 tcp-port 8080
configure sharing health-check member-port 7 add track-tcp 192.168.1.103 tcp-port 8080
configure sharing health-check member-port 8 add track-tcp 192.168.1.104 tcp-port 8080
```

## Displaying Switch Load Sharing

You can display static and dynamic load sharing. In the link aggregation displays, the types are shown by the following aggregation controls:

- Static link aggregation—static
- Link Aggregation Control Protocol—LACP
- Health check link aggregation—hlth-chk

- To verify your configuration, use the following command:

  `show ports sharing`

- Verify LACP configuration, use the following command:

  `show lacp`

- To display information for the specified *LAG*, use the following command:

  `show lacp lag group-id {detail}`
- To display LACP information for a port that is a member of a LAG, use the following command:

  `show lacp member-port port {detail}`

  Refer to Displaying Port Information on page 303 for information on displaying summary load-sharing information.
- To clear the counters, use the following command:

  `clear lacp counters`
- To display the LACP counters, use the following command:

  You can display the LACP counters for all member ports in the system by using:

  `show lacp counters`
- To display information for a health check LAG, use the following command:

  `show sharing health-check`

# MLAG

## BlackDiamond X8 Series Switches, BlackDiamond 8000 Series Modules, Summit Family Switches, and SummitStack

The *MLAG* feature allows you to combine ports on two switches to form a single logical connection to another network device. The other network device can be either a server or a switch that is separately configured with a regular *LAG* (or appropriate server port teaming) to form the port aggregation.

The following diagram displays the elements in a basic MLAG configuration:

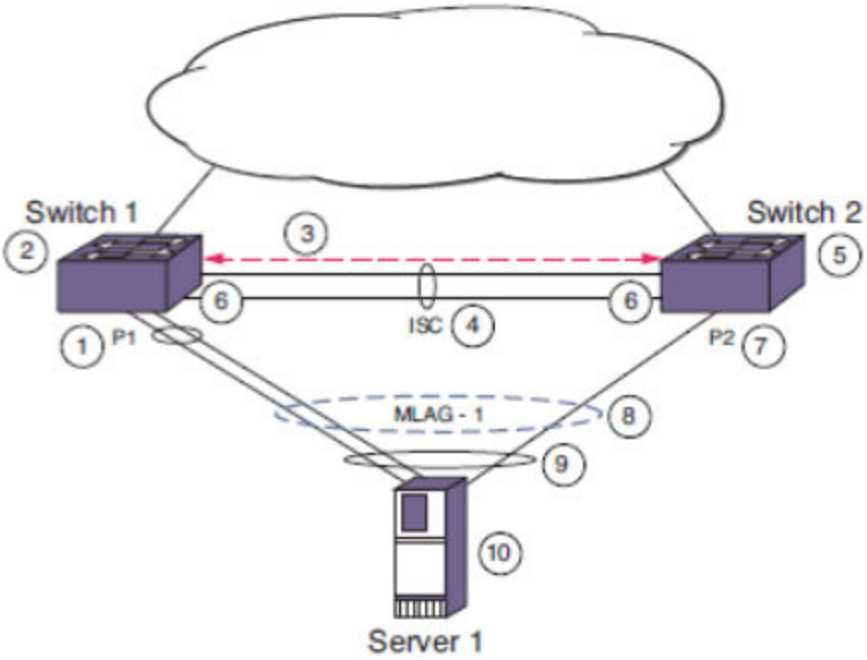



**Figure 35: MLAG Elements**

1. MLAG port that is a load-shared link. This port is the peer MLAG port for <Switch2:Port P2>.
2. MLAG peer switch for Switch 2.
3. Inter-switch connection (ISC or ISC *VLAN*) has only the ISCport as a member port on both MLAG peers.
4. ISC link that connects MLAG peers.
5. MLAG peer switch for Switch 1.
6. ISC ports.
7. MLAG port that is a non-bad-shared link. This port is the peer MLAG port for <Switch1:Port P1>.
8. MLAG group (MLAG-ID 1) that has two member ports (one load-shared and one non-load-shared member).
9. MLAG remote node sees the MLAG ports as regular load-shared link.
10. MLAG remote node - can be a server or a switch.

The operation of this feature requires two ExtremeXOS switches interconnected by an Inter-Switchconnection (ISC). The ISC is a normal, directly connected, Ethernet connection and it is recommended that you engineer reliability, redundancy where applicable, and higher bandwidth for the ISC connection. Logically aggregate ports on each of the two switches by assigning MLAG identifiers (MLAG-ID). Ports with the same MLAG-ID are combined to form a single logical network connection. Each MLAG can be comprised of a single link or a LAG on each switch. When an MLAG port is a LAG, the MLAG port state remains up until all ports in the LAG go down.

As long as at least one port in the LAG remains active, the MLAG port state remains active.When an MLAG port (a single port or all ports in a LAG) fails, any associated MAC *FDB (forwarding database)* entries are moved to the ISC, forcing traffic destined to the MLAG to be handled by the MLAG peer switch. Additionally, the MLAG peer switch is notified of the failure and changes its ISC blocking filter to allow transmission to the MLAG peer port. In order to reduce failure convergence time, you can configure MLAG to use ACLs for redirecting traffic via the "fast" convergence-control option.

Each of the two switches maintains the MLAG state for each of the MLAG ports and communicates with each other to learn the MLAG states, MAC FDB, and IP multicast FDB of the peer MLAG switch.

### *ISC Blocking Filters*

The ISC blocking filters are used to prevent looping and optimize bandwidth utilization.

When at least one *MLAG* peer port is active, the upper layer software initiates a block of traffic that ingresses the ISC port and needs to be forwarded to the local MLAG ports. This is considered to be the steady state condition. In normal steady state operation most network traffic does not traverse the ISC. All unicast packets destined to MLAG ports are sent to the local MLAG port only. However, flood and multicast traffic will traverse the ISC but will be dropped from MLAG peer port transmission by the ISC blocking filter mechanism. The ISC blocking filter matches all Layer 2 traffic received on the ISC and blocks transmission to all MLAG ports that have MLAG peer ports in the active state.

When there are no active MLAG peer ports, the upper layer software initiates an unblocking of traffic that ingresses the ISC port and needs to be forwarded to the local MLAG ports thus providing redundancy. This is considered to be the failed state.

### *Inter-Switch Communication*

#### Keep-alive Protocol

*MLAG* peers monitor the health of the ISC using a keep-alive protocol that periodically sends health-check messages. The frequency of these health-check hellos is configurable. When the MLAG switch stops receiving health check messages from the peer, it could be because of the following reasons:

- Failure of the ISC link when the remote peer is still active.
- The remote peer went down.

If the ISC link alone goes down when the remote peer is alive, both the MLAG peers forward the south-bound traffic, resulting in duplication of traffic. However, this does not result in traffic loops. This is because the remote node load shares to both the MLAG peers and does not forward the traffic received on one of the load shared member ports to other member ports of the same load shared group.

Starting in ExtremeXOS 15.5, health check messages can also be exchanged on an alternate path by separate configuration – typically the “Mgmt” *VLAN*. If the peer is alive when the ISC link alone goes down, one of the MLAG peers disables its MLAG ports to prevent duplicate south-bound traffic to the remote node. To reduce the amount of traffic on the alternate path, health check messages are initiated on the alternate path only when the ISC link goes down. When the ISC link is up, no health check messages are exchanged on the alternate path.

When the MLAG switch misses 3 consecutive health check messages from the peer, it declares that the MLAG peer is not reachable on the ISC link. It then starts sending out health check messages on the alternate path to check if the peer is alive. When the first health check message is received from the MLAG peer on the alternate path, it means that the peer is alive. In this scenario, one of the MLAG peers disables its MLAG ports to prevent duplication of south-bound traffic to the remote node.

Note
The MLAG switch having the lower IP address for the alternate path VLAN disables its ports.

When the ISC link comes up and the switch starts receiving health check messages on the ISC control VLAN, the ports that were disabled earlier have to be re-enabled. This action is not performed

immediately on the receipt of the first health check message on the ISC control VLAN. Instead the switch waits for 2 seconds before enabling the previously disabled ports. This is done to prevent frequent enabling and disabling of MLAG ports due to a faulty ISC link up event.

#### MLAG Status Checkpointing

Each switch sends its MLAG peer information about the configuration and status of MLAGs that are currently configured over the ISC link.

This information is checkpointed over a TCP connection that is established between the MLAG peers after the keep-alive protocol has been bootstrapped.

#### Authentication for Checkpoint Messages

The checkpoint messages exchanged between the MLAG peers over the TCP connection are sent in plain text and can be subjected to spoofing. Starting from EXOS 15.5 a provision is provided as part of this feature to secure the checkpoint connection against spoofing.

A key for authentication must be configured on both the MLAG peer switches. This key will be used in calculating the authentication digest for the TCP messages. TCP_MD5SIG socket option will be used for authentication and so only *MD5 (Message-Digest algorithm 5)* authentication is supported. The configured key will be used in setting up TCP_MD5SIG option on the checkpoint socket. The same key must be configured on both the MLAG peers. The checkpoint connection will not be established if different keys are configured on the MLAG peer switches.

### *PIM MLAG Support*

ExtremeXOS allows configuring PIM on both *MLAG* peers. PIM adjacencies between each MLAG vlan are established across the ISC link.

- The checkpoint PIM state between MLAG peers. This should include all MLAG egresses.
- You can verify that PIM functionality for MLAG is present by issuing the following show command:

  `show pim cache {{`**`detail`**`} | {`**`state-refresh`**`} {`**`mlag-peer-info`**`} {`*`group_addr {source_addr`*`}}}`
- Additionally, the existing show pim cache command displays ingress *VLAN* information for all MLAG peers.

  The output of the command is shown below:

```
Index    Dest Group      Source             InVlan   Origin
          [0001]  226.1.1.1       112.2.2.202 (S)    fifthteen Sparse
       Expires after 203 secs UpstNbr: 51.15.15.2
       RP: 61.2.2.2 via 51.15.15.2 in fifthteen
       Peer Ingress VLAN (ISC 1): 51.15.15.4/24 (Same)
       EgressIfList =  eight(0)(FW)(SM)(I) , five(0)(FW)(SM)(I)
       PrunedIfList =  ten(0)(SM)(AL)
```

## Support for More than One MLAG Peer

Beginning in ExtremeXOS 15.5, each *MLAG* switch can support the creation of two MLAG peers. The following topology is now supported:

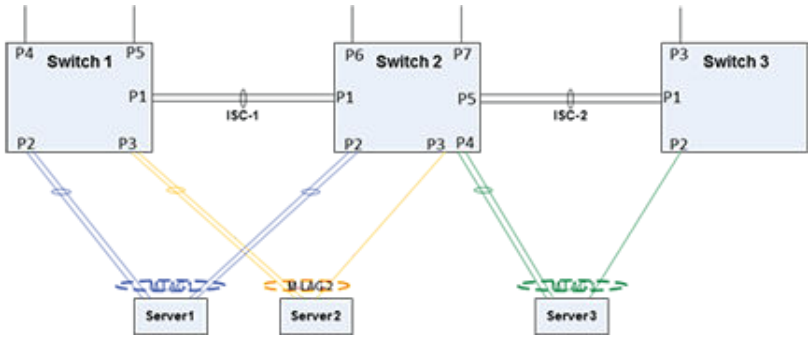



**Figure 36: Two MLAG Peer Topology**

All basic MLAG functionality and traffic forwarding rules that existed earlier apply with this topology.

Note

A port is an MLAG port only with respect to a particular MLAG peer switch. In the above example, the “green” port on switch-2 is an MLAG port with respect to switch-3 on ISC-2. It is not an MLAG port with respect to switch-1 on ISC-1. Similarly, on switch-2, “blue” and “orange” ports are MLAG ports with respect to switch-1 on ISC-1. They are not MLAG ports with respect to switch-3 on ISC-2.

*Traffic Flows*

In the steady state when all the ports are enabled, the *MLAG* blocking rule prevents traffic coming from the ISC port to be forwarded to its MLAG ports. So, in steady state, any unknown (unknown unicast/broadcast/multicast) traffic from Switch-1 to Switch-2 over ISC-1 will not be flooded to “blue” and “orange” MLAG ports on switch-2. However, since the “green” MLAG port on switch-2 does not correspond to ISC-1, the traffic from ISC-1 will be forwarded to “green” MLAG port as shown in the following illustration:

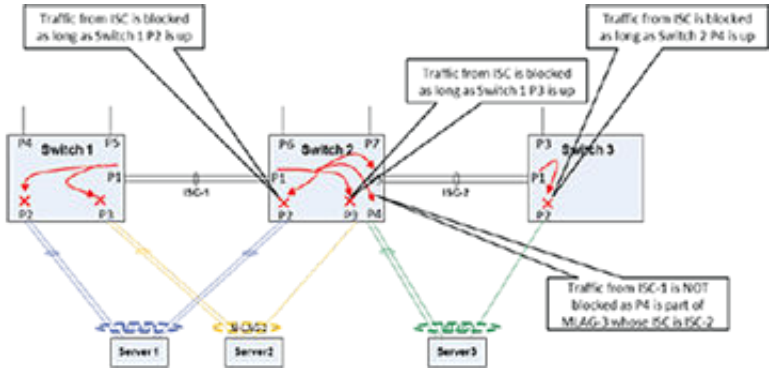



**Figure 37: Traffic Flow**

In steady state, the unknown traffic from Switch-3 to Switch-2 over ISC-2, will be blocked to “green” MLAG ports. However, this traffic will be sent to “blue” and “orange” MLAG ports.

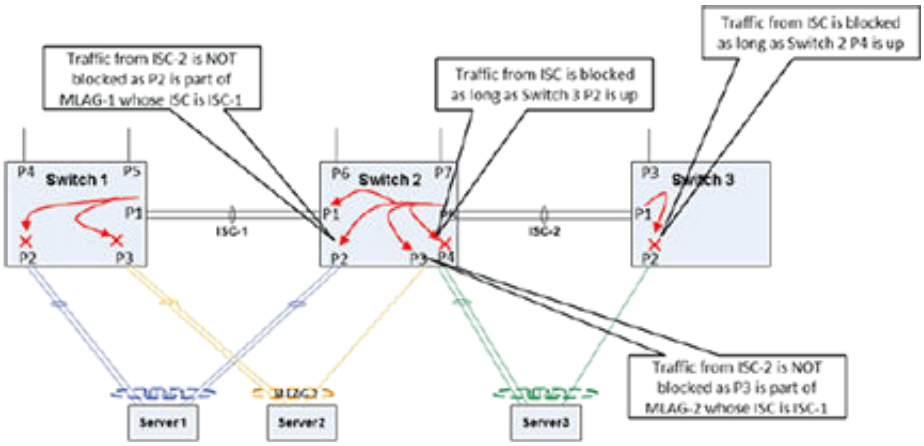



**Figure 38: Traffic Flow (Unknown Traffic Switch-3 to Switch-2)**

*MLAG Peer Port Failure*

In the case of an *MLAG* port failure, the blocking rule on the peer MLAG switch will be removed and the traffic coming on the ISC will be forwarded to the corresponding MLAG port.

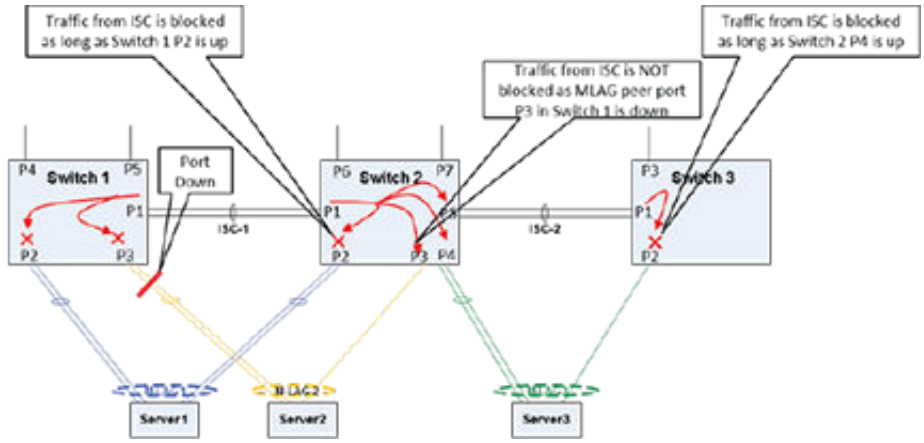



**Figure 39: Peer Port Failure**

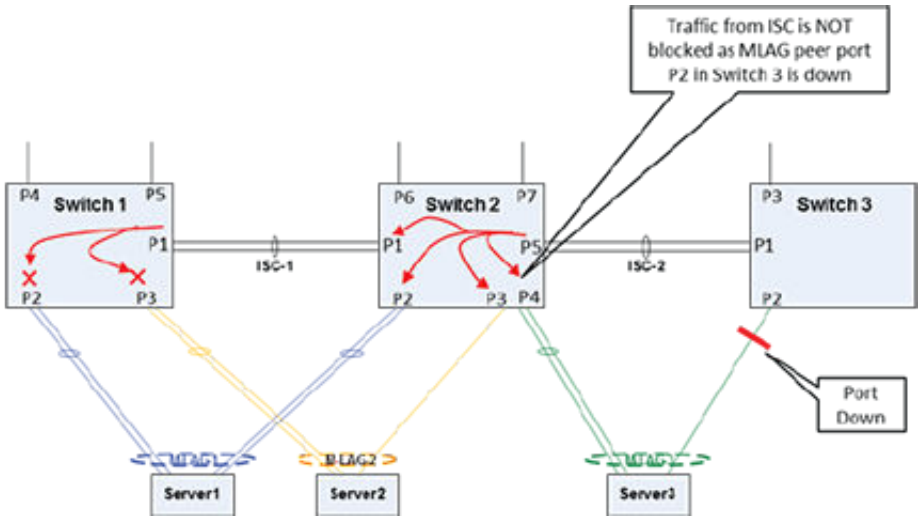



**Figure 40: Peer Port Failure (cont.)**

*FDB Checkpointing*

All *FDB* entries on VLANs having ISC port as a member will be checkpointed to the peer *MLAG* switch(es). FDB entries learned on an MLAG port will be checkpointed to the corresponding MLAG port on the peer switch. FDB entries on the non-MLAG, non-ISC port will be checkpointed to the corresponding ISC port on the peer switch.

*Layer-2 IP Multicast*

The receiver Rx1 sends an *IGMP (Internet Group Management Protocol)* report towards Server 1. Since server 1 is connected to both Switch 1 and Switch 2 through a *LAG*, let us assume that it selects a port towards Switch 2 to forward the report. This results in Switch 2 receiving a report on port P2. It adds the port to the group table. Switch 2 sends a checkpoint message to Switch 1 since Switch 1 is the *MLAG* peer for MLAG-1. But Switch 2 doesn't checkpoint the report to Switch 3 since Switch 3 is not an MLAG peer for MLAG-1. However, Switch 2 forwards the report towards Switch 3 over ISC. Switch 3 would learn this on ISC port. The group information tables on each of the switches will be as shown below.

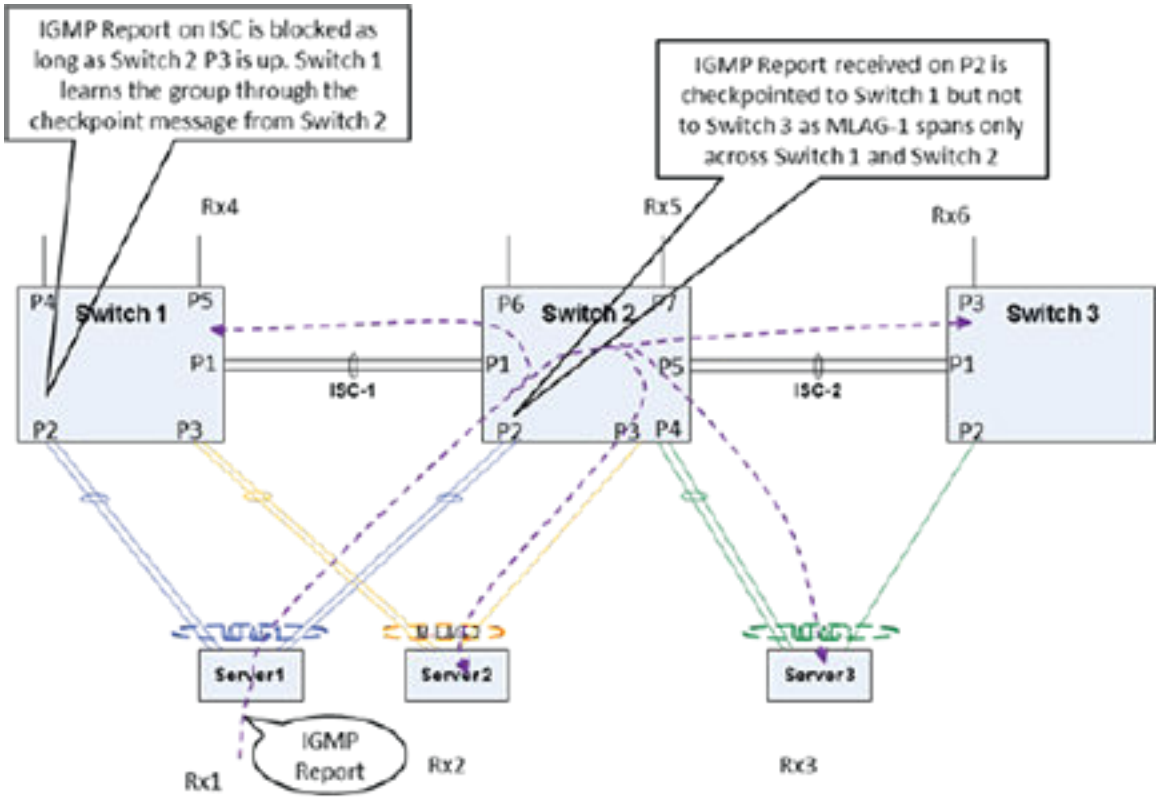



**Figure 41: Group Information Table per Switch**

**Table 31: Switch 1**

| Group | Port |
|---|---|
| G1 | P2 |

**Table 32: Switch 2**

| Group | Port |
|---|---|
| G1 | P2 |

**Table 33: Switch 3**

| Group | Port |
|---|---|
| G1 | P1 |

The following figure depicts a scenario where a multicast stream STR1 ingressing Switch 1 will reach Rx1 and Rx2 directly via P2 and P3 respectively. Similarly traffic ingressing Switch 2 will reach Rx3 through P4 directly.

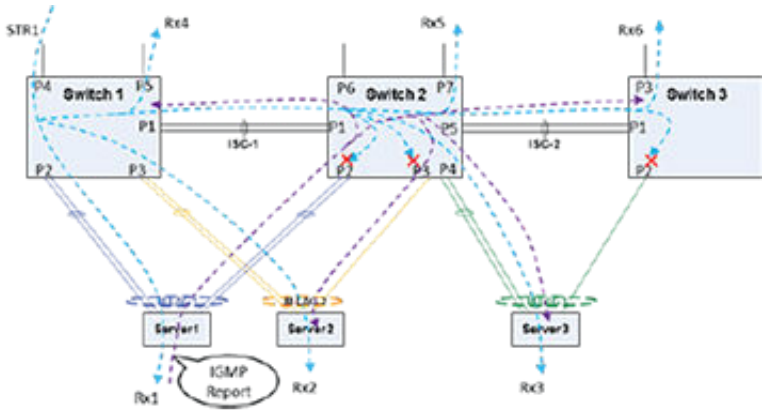



**Figure 42: Multicast Stream Ingressing Example**

The following figure depicts a traffic flow scenario when the local MLAG port is down. When the local MLAG port is down, IGMP group information received from the peer MLAG router will result in ISC being added as egress port. Now traffic stream STR1 ingressing Switch 1 will go over ISC-1 to Switch 2 where it gets forwarded towards Server 2 over P3.

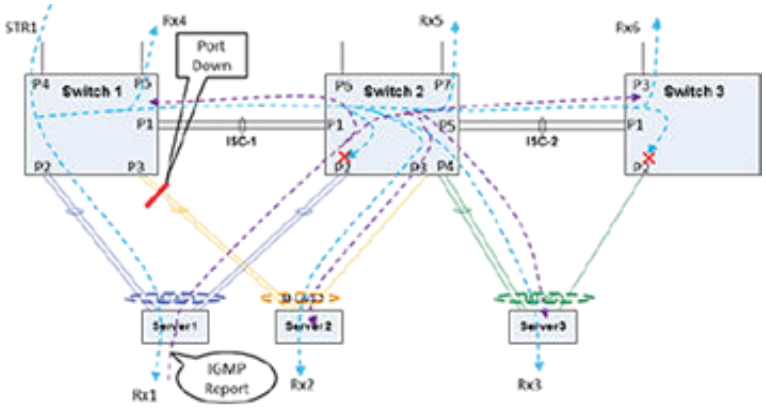



**Figure 43: Traffic Flow with Local MLAG Port Down**

The following figure illustrates a scenario where a multicast stream STR2 ingressing Switch 2 will reach Rx1, Rx2 and Rx3 directly via P2, P3 and P4 respectively.

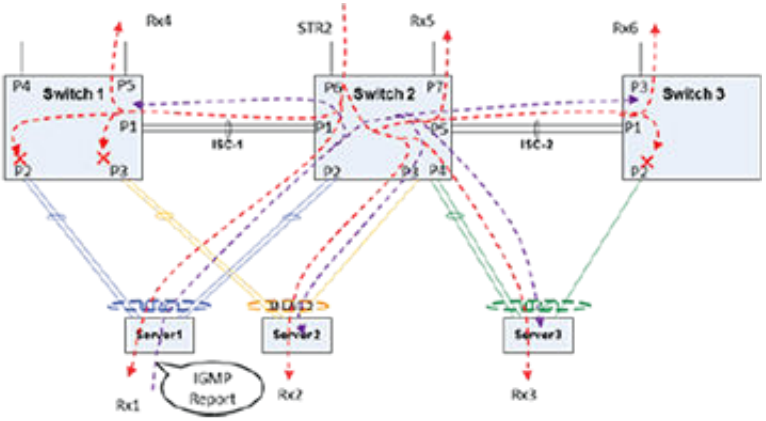



**Figure 44: Multicast Stream Ingressing Switch 2 Example**

*Layer 3 IP Multicast using PIM-SM*

Consider the following sample topology for *MLAG*:

```
DUT-1(core lic)======ISC vlan========DUT-2(core lic)

| |

+----------------DUT-3(edge lic)---------------+
```

DUT-1 and DUT-2 are MLAG peers, DUT-3 is a L2 switch whose uplink is a *LAG* up to the pair of MLAG switches.

- RP and BSR can be configured on same device along with the MLAG configuration, but we recommend to keep RP node away from MLAG peers. One MLAG peer will be Designated Router (DR) and another one will be elected as NON-DR for MLAG *VLAN*. DR node will send *,G and try to pull the traffic from RP, and Non-DR will not pull the traffic until DR is alive. If you config RP on Non-DR node then both MLAG peers will pull the traffic which triggesr the assert to avoid the traffic duplication. It is not recommended to setup RP on any VLAN on MLAG peers.
- It is recommended to avoid the assert operation since a small amount of traffic duplication happens during the assert operation. We can avoid assert in some scenario but not all the scenarios.
- DR priority configuration will help to make RP node as DR. The DR priority feature is available from 15.3.2 release onwards.
- It is recommended that for PIM-SM deployments, the RP is configured on loopback VLANs instead of regular VLANs. This ensures continuous connectivity to the RP without the needing active ports present in that respective VLAN.
- It is recommended that for PIM-SM deployments, route-to-source must exist and receivers should get the traffic from the SPT tree in the MLAG configurations. *,G forwarding in MLAG is not a recommended configuration.

## MLAG Limitations and Requirements

The *MLAG* feature has the following limitations:

- MLAG peer switches must be of the same platform family. The following MLAG peers are allowed:
  - BlackDiamond 8800 switches with BlackDiamond 8800 switches
  - BlackDiamond X8 switches with BlackDiamond X8 switches
  - Summit switches with Summit switches
  - SummitStack with SummitStack

  **Note**

  In the case of Summit standalone switches, it is strongly recommended that MLAG peer switches be of the same type, for example, Summit X480 switches with Summit X480 switches.

  In the case of SummitStack and BlackDiamond 8800 switches, we recommend that the MLAG ports be from slots of similar capability, for example, BlackDiamond 8900-G48X-xl to BlackDiamond 8900-G48T-xl modules.

- Layer 2 protocols such as EAPS or *STP (Spanning Tree Protocol)* will be configured to not allow the blocking of the ISC.

- The number of MLAG ports for each pair of switches is limited to 768.
- MLAG peers should run the same version of ExtremeXOS for proper functioning.
- *ESRP* cannot be configured in a MLAG environment with more than one peer.
- The MLAG peers in a multi peer setup cannot be looped however can be extended as a linear daisy chain.

### *MLAG Requirements*

The following table shows additional MLAG requirements that are specific to other protocols and features.

**Note**
To function properly, MLAG peers should run the same version of ExtremeXOS.

| Items | Impact |
|---|---|
| VLAN:Membership | You must add the respective port (or *LAG*) that is part of an MLAG to a *VLAN* on both MLAG peers.<br>The set of configured VLANs on [Switch1:P1] must be identical to the set of VLANs configured on [Switch2:P2].<br>You must add the ISC to every VLAN that has an MLAG link as a member port. |
| VMAN:Membership | The restrictions are the same as those for VLAN Membership. |
| VLAN:ISC | You must create a Layer 3 VLAN for control communication between MLAG peers.<br>You cannot enable IP forwarding on this VLAN.<br>The ISC is exclusively used for inter-MLAG peer control traffic and should not be provisioned to carry any user data traffic. Customer data traffic however can traverse the ISC port using other user VLANs. |
| VMAN:ISC | Although not recommended, a VMAN may be configured to carry Inter-MLAG peer traffic, |
| LAG:Load-Sharing Algorithm | It is recommended but not required that LAGs that form an MLAG be configured to use the same algorithm. |
| Ports:Flooding | To disable flooding on an MLAG, you must disable flooding on both ports (or LAGs) that form the MLAG. |
| Ports:Learning | To disable learning on an MLAG, you must disable learning on both ports (or LAGs) that form the MLAG. Learning is disabled by default on ISC ports. |
| *FDB*:Static & Blackhole entries | Configuration must be identical on both MLAG peers for entries that point to an MLAG port. |
| FDB:Limit learning | Learning limits are applicable to member ports of each peer. The limit on the MLAG is the sum of the configured value on each peer. |
| FDB:MAC Lockdown | This is supported but needs configuration on both peers. A switch can still receive checkpointed MAC addresses from its peer in the window between executing the lockdown command on both switches. |
| EAPS | MLAG ports cannot be configured to be EAPS ring ports.<br>Configuration of the ISC port as an EAPS blocked port is disallowed. |

| Items | Impact |
|---|---|
| STP | STP cannot be enabled on MLAG ports.<br>STP should not be enabled on the ports present in the remote node which connects to the MLAG ports.<br>You should ensure that the ISC port is never blocked by STP. |
| *VRRP* | VRRP must be enabled on Layer 3 VLANs that have MLAG member ports. |
| ESRP | MLAG and ISC ports must be added as ESRP host-attach ports. |
| *EDP/LLDP (Link Layer Discovery Protocol)* | There are no restrictions but the remote end of the MLAG will display different neighbors for different ports in the same LAG. |
| ELSM | ELSM is not to be configured on MLAG ports at either end of an MLAG. |
| Software-Redundant Ports | These are not to be configured on MLAG ports at either end of an MLAG. |
| Mirroring | Mirroring on local ports in an MLAG is supported. Mirroring of MLAG peer ports to a local port is not supported. |
| Routing Protocols | OSPFV2/OSPFV3 neighborship can be formed across an MLAG. |
| Multicast:IGMP | All timers related to *IGMP* must be identical on both the peers. |
| Multicast:PIM | PIM should be configured on both the MLAG peers, and the PIM timers must be identical.<br>MLAG functionality must not be enabled on PIM Intermediate routers. It should be enabled only on Last Hop (LHR) and First Hop (FHR) routers.<br>MLAG peer switches S1 and S2 perform Checkpoint PIM for S and G states. This should include all MLAG egresses.<br>To avoid traffic drops due to asserts, do not include ISC port in MLAG egresses if the ingress VLAN includes ISC port, and both the peers have the same ingress for the S, G cache. |
| Multicast:MVR | MVR should be enabled on only one of the MLAG peer switches.<br>MVR must not be enabled on MLAG VLANs. |
| Multicast:PIM Snooping | This is not supported. |
| Multicast:IPv6 | There are no restrictions. |
| CFM | There are no restrictions. |
| *MPLS*:General | MPLS cannot be enabled on VLANs having MLAG member ports. |
| MPLS:VPLS | VPLS must be configured for redundancy using ESRP. The ESRP master VLAN must include the ISC ports and the VPLS service VLAN ports as members.<br>Pseudowires cannot traverse an ISC link. You should not add the ISC port as a member to MPLS VLANs that can be used by LSPs that can carry Layer 2 VPN traffic terminating on MLAG peer switches. |
| ACLs | It is strongly recommended that configuration be identical across peers on MLAG ports. |
| *QoS* | It is strongly recommended that configuration be identical across peers on MLAG ports. |
| Netlogin | This is not supported. |

| Items | Impact |
|---|---|
| VLAN:PVLAN | If an MLAG port is a member of either a subscriber VLAN or a network VLAN, the ISC port needs to be added as a member of the network VLAN. Subscriber VLANs in a private VLAN cannot have overlapping MLAG ports as members. Configuring dedicated loopback ports for subscriber VLANs in a private VLAN that shares an MLAG port causes duplicate traffic to be sent to the remote node. |
| DAD | DAD detects duplicate IPv4 addresses configured on a VLAN that spans MLAG peer switches. This occurs only when the solicitation attempts to use the following command is more than one:<br>`configure ip dad attempts max_solicitations` |

## Configuring MLAGs

This section provides the commands used to configure MLAGs and display information about those configured.

- To create an *MLAG* peer switch association structure, use the following command:

  `create mlag peer peer_name`

- To delete a peer switch, use the following command:

  `delete mlag peer peer_name`

- To associate an MLAG peer structure with an MLAG peer switch IP address, use the following command:

  `configure mlag peer peer_name ipaddress peer_ip_address {vr VR}`

- To unconfigure the association, use the following command:

  `unconfigure mlag peer peer_name ipaddress`

- To configure the time interval between health check hello packets exchanged between MLAG peer switches, use the following command:

  `configure mlag peer peer_name interval msec`

- To unconfigure the time interval setting and reset the interval to the default of 1000 ms, use the following command:

  `unconfigure mlag peer peer_name interval`

- To bind a local port or *LAG* to an MLAG specified with an integer identifier, use the following command:

  `enable mlag port port peer peer_name id identifier`

- To disable a local port or LAG from an MLAG, use the following command:

  `disable mlag port port`

- To set a preference for having a fast convergence time or conserving access lists, use the following command:

  `configure mlag ports convergence-control [conserve-access-lists | fast]`

Note

Executing the `refresh policy` command with MLAG configuration may result in health check hellos not reaching the CPU. To avoid MLAG peer connectivity disruption, you can either execute the `disable access-list refresh blackhole` command or temporarily increase the peer hello interval to a large value (for instance, 10000 ms) and reset it back once refresh policy is complete.

## Displaying MLAG Information

- To display information about an *MLAG* peer, including MLAG peer switch state, MLAG group count, and health-check statistics, use the following commands:

  `show mlag peer {peer_name}`

- To display each MLAG group, including local port number, local port status, remote MLAG port state, MLAG peer name, MLAG peer status, local port failure count, remote MLAG port failure count, and MLAG peer failure count, use the following command:

  `show mlag ports {port_list}`

- To see if a port is part of an MLAG group, or an ISC port, use the following command without the **mgmt** and **tag** options:

  `show port {mgmt | port_list | tag tag} information {detail}`

## Example of MLAG Configuration

Below is an example of how to configure an *MLAG*. The following figure shows a finished MLAG network.

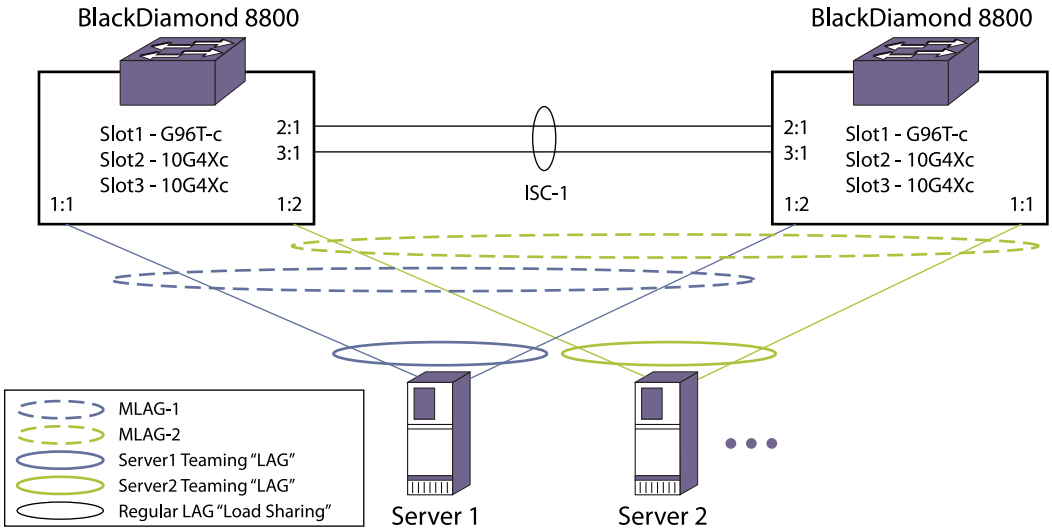



**Figure 45: Simple MLAG Configuration**

1. Create the Inter-Switch Connection (ISC).

   Description: The ISC provides an out-of-band IP communications path between the two MLAG peer switches to exchange keep-alive packets and to checkpoint various state information between switches.

   On the "Left" BlackDiamond 8800 switch:

   ```
   enable sharing 2:1 group 2:1,3:1
   create vlan isc
   config vlan isc tag 3000
   config vlan isc add port 2:1 tag
   config vlan isc ipaddress 1.1.1.1/24
   ```

   On the "Right" BlackDiamond 8800 switch:

   ```
   enable sharing 2:1 group 2:1,3:1
   create vlan isc
   config vlan isc tag 3000
   config vlan isc add port 2:1 tag
   config vlan isc ipaddress 1.1.1.2/24
   ```

2. Create the MLAG peer and associate the peer switch's IP address.

   Description: By creating an MLAG peer you associate a peer name that can be associated with the peer switch's IP address and other peer configuration properties. The peer is then bound to each individual MLAG port group.

   On the "left" BlackDiamond 8800 switch:

   ```
   create mlag peer "rightBD8K"
   config mlag peer "rightBD8K" ipaddress 1.1.1.2
   ```

   On the "right" BlackDiamond 8800 switch:

   ```
   create mlag peer "leftBD8K"
   config mlag peer "leftBD8K" ipaddress 1.1.1.1
   ```

3. Create the MLAG port groups.

   Description: Creates an MLAG port group by specifying the local switch's port, the MLAG peer switch, and an "mlag-id" which is used to reference the corresponding port on the MLAG peer switch. The specified local switch's port can be either a single port or a load share master port.

   On the “left” BlackDiamond 8800 switch:

```
enable mlag port 1:1 peer "rightBD8K" id 1
enable mlag port 1:2 peer "rightBD8K" id 2
```

   On the “right” BlackDiamond 8800 switch:

```
enable mlag port 1:2 peer "leftBD8K" id 1
enable mlag port 1:1 peer "leftBD8K" id 2
```

4. Verify MLAG peers and ports are operational.

   Description: After MLAG groups are configured, you can verify the connections via the show mlag peer and show mlag ports commands. Be sure to note the peer status, the Local Link State, and the Remote Link status.

   On the “left” BlackDiamond 8800 switch:

```
BD-8810.5 # show mlag peer
Multi-switch Link Aggregation Peers:
MLAG Peer         : leftBD8k
VLAN              : isc             Virtual Router    : VR-Default
Local IP Address  : 1.1.1.2         Peer IP Address   : 1.1.1.1
MLAG ports        : 2               Tx-Interval       : 1000 ms
Checkpoint Status : Up              Peer Tx-Interval  : 1000 ms
Rx-Hellos         : 184             Tx-Hellos         : 184
Rx-Checkpoint Msgs: 12              Tx-Checkpoint Msgs: 12
Rx-Hello Errors   : 0               Tx-Hello Errors   : 0
Hello Timeouts    : 1               Checkpoint Errors : 0
Up Time           : 0d:0h:0m:10s    Peer Conn.Failures: 1

BD-8810.3 # show mlag ports
Local
                                                 Local  Remote
MLAG          Local   Link     Remote                   Peer    Fail    Fail
Id            Port    State    Link    Peer             Status  Count   Count
================================================================================
1       1:1     A        Up          rightBD8K          Up       0       0
2       1:2     A        Up          rightBD8K          Up       0       0
================================================================================
Local Link State: A - Active, D - Disabled, R - Ready, NP - Port not present
Remote Link     : Up - One or more links are active on the remote switch,
Down - No links are active on the remote switch,
N/A - The peer has not communicated link state for this MLAG port
Number of Multi-switch Link Aggregation Groups  : 2
Convergence control                             : Fast
```

   On the “right” BlackDiamond 8800 switch:

```
BD-8810.3 # show mlag peer
Multi-switch Link Aggregation Peers:
MLAG Peer         : rightBD8k
VLAN              : isc             Virtual Router    : VR-Default
Local IP Address  : 1.1.1.1         Peer IP Address   : 1.1.1.2
MLAG ports        : 2               Tx-Interval       : 1000 ms
Checkpoint Status : Up              Peer Tx-Interval  : 1000 ms
Rx-Hellos         : 167             Tx-Hellos         : 167
Rx-Checkpoint Msgs: 12              Tx-Checkpoint Msgs: 12
Rx-Hello Errors   : 0               Tx-Hello Errors   : 0
```

```
Hello Timeouts    : 1               Checkpoint Errors : 0
Up Time           : 0d:0h:0m:7s     Peer Conn.Failures: 1

BD-8810.5 # show mlag ports
Local                                          Local Remote
MLAG        Local   Link     Remote            Peer    Fail    FailId
            Port    State    Link    Peer      Status  Count   Count
=========================================================================
2     1:1   A        Up              leftBD8K   Up     0       0
1     1:2   A        Up              leftBD8K   Up     0       0
=========================================================================
Local Link State: A - Active, D - Disabled, R - Ready, NP - Port not
present
Remote Link     : Up - One or more links are active on the remote switch,
Down - No links are active on the remote switch,
N/A - The peer has not communicated link state for this MLAGport
Number of Multi-switch Link Aggregation Groups  : 2
Convergence control                             : Fast
```

5. Add ISC port to *VLAN*.

   Description: The ISC port must be added as a member port for any VLAN that has MLAG member ports.

   ```
   create vlan "xyz"
   ```

   On the “left” BlackDiamond 8800 switch:

   ```
   configure vlan "xyz" add port 1:1, 2:1 tagged
   ```

   On the “right” BlackDiamond 8800 switch:

   ```
   configure vlan "xyz" add port 1:2, 2:1 tagged
   ```

   The previous figure, the example above, shows a basic MLAG network. The following figure shows a network with back-to-back aggregation. There is one MLAG configured on the BlackDiamond switches and three configured on the Summit switches.

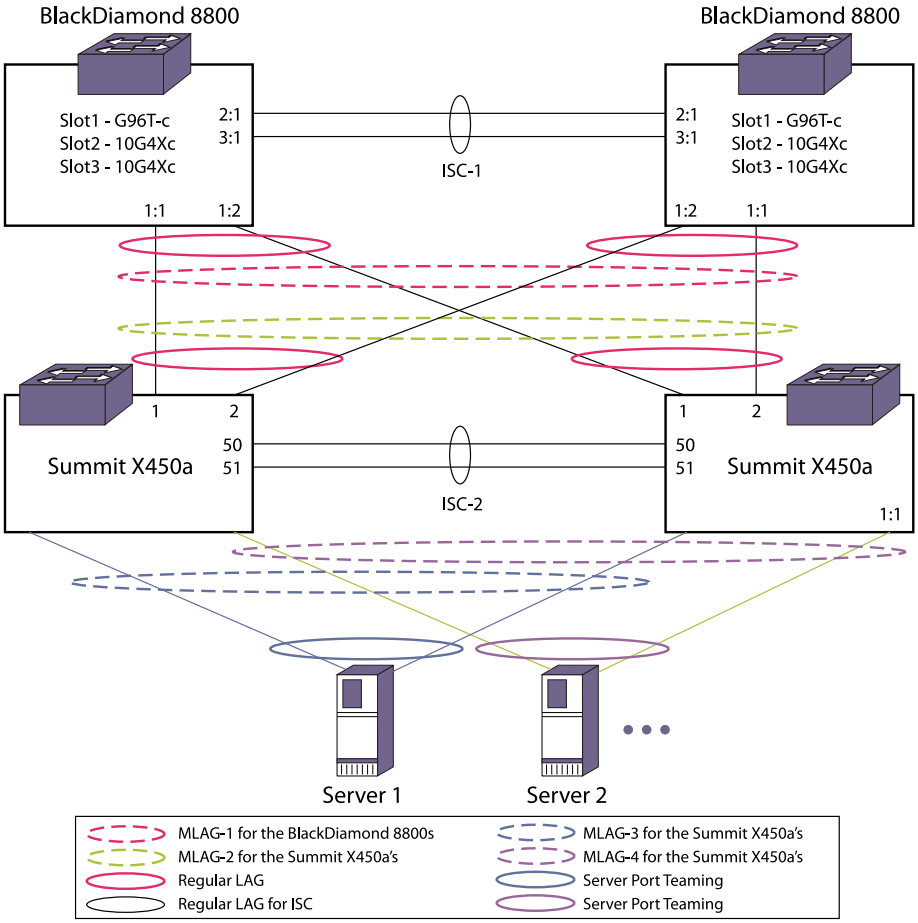



**Figure 46: Two-tier MLAG Network**

## MLAG-LACP

Beginning in EXOS 15.3, the EXOS *MLAG* feature supports Link Aggregation Control Protocol (LACP) over MLAG ports. To do this, all MLAG peer switches use a common MAC in the System Identifier

portion of the LACPDU transmitted over the MLAG ports. The following options and requirements are provided:

- The MLAG peer that has the highest IP address for the ISC control *VLAN* is considered the MLAG LACP master. The switch MAC of the MLAG LACP master is used as the System Identifier by all the MLAG peer switches in the LACPDUs transmitted over the MLAG ports. This is the default option.
- You can configure a common unicast MAC address for use on all the MLAG peer switches. This MAC address is used as the System Identifier by all the MLAG peer switches in the LACPDUs transmitted over the MLAG ports. This configuration is not checkpointed to the MLAG peers, and you must make sure that the same MAC address is configured on all the MLAG switches. Additionally, you must ensure that this address does not conflict with the switch MAC of the server node that teams with the MLAG peer switches.

Note

When LACP shared ports are configured as MLAG ports, a *LAG* ID change after MLAG peer reboot may result in MLAG ports being removed and re-added to the aggregator. To avoid the MLAG port flap, it is recommended to configure a common LACP MAC in both the MLAG peers using the `configure mlag peer` *`peer_name`* `lacp-mac` *`lacp_mac_address`* command.

### *LACPDU Transmission on MLAG Ports*

To prevent the server node from forming two separate aggregators to the *MLAG* peers (which could result in a loop), it is necessary that both MLAG peers transmit LACPDUs with the same System Identifier and Actor Key. The following points discuss how the System Identifier is determined:

- The MLAG peers must communicate at least once with each other to generate LACPUDs on MLAG ports. If the MLAG peers do not communicate with each other, no LACPDUs are sent out on the MLAG ports. The MLAG peers checkpoint their system MAC and the configured MAC (if configured) to the peer so that the LACP Operational MAC is determined.
- If no LACP MAC is configured on the MLAG peers, the LACP Operational MAC is the MAC address of the MLAG peer that has the highest IP address for ISC control *VLAN*.
- If a different LACP MAC address is configured on the MLAG peers, the configured MAC is not used. In this case, the LACP Operational MAC address is the MAC address of the MLAG peer that has the highest IP address for ISC control VLAN.
- The configured MAC address is only used when the same MAC is configured on both the MLAG peers.

### *Scalability Impact on Load Shared Groups*

When static load sharing is used for the *MLAG* ports, and if there is a single link connecting the server node and the MLAG peer switches, the port does not need to be configured as a load shared port on the MLAG peer switches. Configuring LACP on MLAG ports can reduce the number of load shared ports that can be configured in the system.

#### Configuration Guidelines

- LACP configuration for MLAG ports (system priority, LACP timeout, activity mode, etc.) should be the same on all the MLAG peer switches.
- We recommend that the server node has a lower System Aggregation Priority than the MLAG peers so that the server node chooses which of the ports can be aggregated. As an example, there are a maximum of 8 ports that can be aggregated together, and there are eight links from Peer1 to the

Server, and another eight links from Peer2 to the server node. When the server node has a lower System Aggregation Priority, it can choose which of the total available links can be aggregated together.

- If the Port Aggregation Priority is not configured for the load shared member ports, there is a chance that only the links from server node to one of the MLAG peer are aggregated together (based on the port numbers). In this instance, the links from the server node to the other MLAG peer are unused. To avoid this, you can configure the Port Aggregation Priority on the server node so that the number of active links to the MLAG peers is balanced.
- You must configure load sharing groups on all the MLAG ports even if they contain just one port.

Below are sample configurations.

**Configuration on Peer1**

```
create vlan "isc"
configure vlan isc tag 4000
enable sharing 5 grouping 5,10 lacp
configure vlan "isc" add ports 5 tagged
configure vlan "isc" ipaddress 1.1.1.1/8

create mlag peer "peer2"
configure mlag peer "peer2" ipaddress 1.1.1.2
configure mlag peer "peer2" lacp-mac 00:11:22:33:44:55

enable sharing 6 grouping 6,12 lacp
enable sharing 18 grouping 18 lacp
enable mlag port 6 peer "peer2" id 1
enable mlag port 18 peer "peer2" id 2
```

**Configuration on Peer2**

```
create vlan "isc"
configure vlan isc tag 4000
enable sharing 5 grouping 5,10 lacp
configure vlan "isc" add ports 5 tagged
configure vlan "isc" ipaddress 1.1.1.2/8

create mlag peer "peer1"
configure mlag peer "peer1" ipaddress 1.1.1.1
configure mlag peer "peer1" lacp-mac 00:11:22:33:44:55
enable sharing 20 grouping 20 lacp
enable sharing 6 grouping 6,15 lacp
enable mlag port 20 peer "peer1" id 1
enable mlag port 6 peer "peer1" id 2
```

**Configuration on Server Nodes (assumed to be Extreme Switches)**

```
enable sharing 1 grouping 1,2,3 lacp
configure sharing 1 lacp system-priority 100
configure lacp member-port 1 priority 10
configure lacp member-port 2 priority 20
configure lacp member-port 3 priority 15
```

Here is an illustration of the configuration:

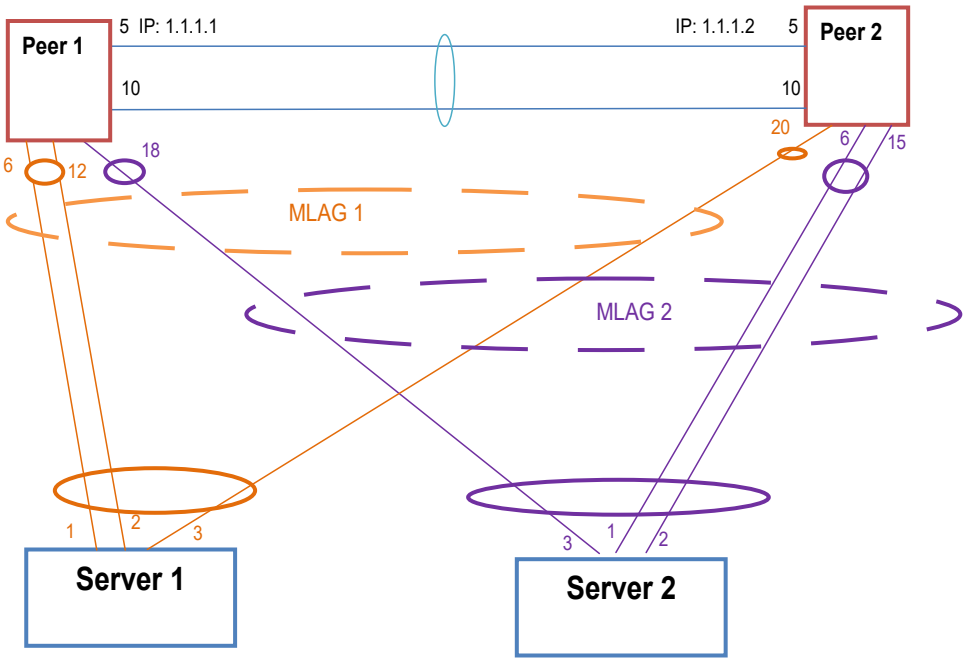



**Figure 47: LACP Ports in MLAG Configuration**

## Mirroring

Mirroring is a function on existing Extreme Networks switches that allows copies of packets to be replicated to additional ports without affecting the normal switching functionality. The mirrored data actually occupies fabric bandwidth, so it is very likely that normal forwarding and mirroring cannot both be done at line rate. In the most general terms, traffic ingressing and/or egressing an interface is mirrored. For ports, traffic ingressing and/or egressing a port can be mirrored (refer to the `configure mirror add` command). For VLANs and virtual ports, only traffic ingressing these interfaces are mirroring.

One of the common uses of the mirroring functionality is packet capture; for example sending copies of all packets that arrive on port P, vlan V, to a monitor port Q. Previous implementations of mirroring were limited to a single instance, where only one destination port (or port list) was allowed to be configured in the system. That implementation was also limited to 128 total sources of this traffic (also referred to as filters). Only *VLAN* and VLAN/port "filters" are currently implemented as filters.

ExtremeXOS 15.3 and above supports Multi Instanced Mirroring that expands the number of destinations allowed to match the hardware capabilities (current hardware allows for up to 4 ingress mirroring instances and two egress mirroring instances). A mirroring instance consists of a unique destination port, or port list, and the source filters associated with it. While the previous implementation allowed for sixteen sources, our current implementation allows for 128 per instance.

ExtremeXOS 16.2 has enhanced mirroring to support IPFIX flows to be mirrored as well. This is done by using the mirroring capabilities in ExtremeXOS along with IPFIX to provide additional information about flows that can be analyzed with our Extreme Application Analytics application. As mentioned earlier, IPFIX can collect statistics about flows that are recognized based on configured flow keys. However IPFIX cannot inspect packets deeper than the L4 (TCP) level as the deepest flow keys configurable are L4 Source and Destination Ports. Extreme Application Analytics however, can do deep packet

inspection beyond L4 if it is provided a copy of the packet payload. This enhancement provides the ability to mirror the first 15 packets of any IPFIX flow to a port where Extreme Application Analytics can receive a copy of the packet for deep packet inspection. As with mirroring, this allows you to configure multiple mirroring instances. This feature is supported on Summit X460 and X460-G2, and BlackDiamond X8 (40G12X-XL, 100G4X-XL, and 100G4X).

Note

You can have a maximum of 16 mirroring instances in the switch (including default mirroring instance) but only 4 can be active at a time as explained below:

- Four (4) ingress
- Three (3) ingress and one (1) egress
- Two (2) ingress and two (2) egress

The maximum possible combinations for mirroring instances include:

- 2 (ingress + egress)
- 1 (ingress + egress) + 2 ingress
- 1 (ingress + egress) + 1 egress + 1 ingress

In general, there are four hardware resource slots. Each single instance uses one slot, while one (ingress + egress) instance uses two slots. So, you can use of total four slots, but there can be no more then two egress instances.

Note

You can accomplish port mirroring using ACLs, or CLEAR-Flow. See ACLs on page 640 and CLEAR-Flow on page 948 for more information.

A virtual port is a combination of a VLAN and a port. The monitor port or ports can then be connected to a network analyzer or RMON probe for packet analysis. The system uses a traffic filter that copies a group of traffic to the monitor port(s). You can have only one monitor port or port list on the switch. This feature allows you to mirror multiple ports or VLANs to a monitor port, while preserving the ability of a single protocol analyzer to track and differentiate traffic within a broadcast domain (VLAN) and across broadcast domains (for example, across VLANs when routing).

Note

The mirroring filter limits discussed later do not apply when you are using ACLs or CLEAR-Flow.

Up to 128 mirroring filters can be configured across all active mirroring instances.

## Tagging of Mirrored packets

The following conditions describe tagging of mirrored packets:

- Untagged ingress mirrored traffic egresses the monitor port(s) untagged. Tagged ingress mirrored traffic egresses the monitor port tagged.
- Egress mirrored traffic always egresses the monitor port tagged.
- On Summit family switches, all traffic ingressing the monitor port or ports is tagged only if the ingress packet is tagged. If the packet arrives at the ingress port as untagged, the packet egresses the monitor port or ports as untagged.

## Guidelines for Mirroring

The guidelines for mirroring are hardware dependent. Find your hardware type in this section for your specific guidelines.

### *Summit Family Switches*

The traffic filter on Summit family switches can be defined based on one of the following criteria:

- Physical port—All data that traverses the port, regardless of *VLAN* configuration, is copied to the monitor port(s). You can specify which traffic the port mirrors:
  - Ingress—Mirrors traffic received at the port.
  - Egress—Mirrors traffic sent from the port.
  - Ingress and egress—Mirrors traffic either received at the port or sent from the port.

    If you omit the optional parameters, all traffic is forwarded; the default for port-based mirroring is ingress and egress.

    **Note**
    You can create an instance where the source is ingress only. When you add a source, pay attention to the monitor port.
- VLAN—All packets ingressing any port on a particular VLAN, regardless of the physical port configuration, is copied to the monitor port(s).
- Virtual port—All traffic ingressing the switch on a specific VLAN and port combination is copied to the monitor port(s).
- IPFIX—mirrors the first 15 packets of any IPFIX flow to a port where Extreme Application Analytics can receive a copy of the packet for deep packet inspection.
- Summit family switches support a maximum of 128 mirroring filters per instance.
- ExtremeXOS supports up to 16 monitor ports for one-to-many mirroring.
- Only traffic ingressing a VLAN can be monitored; you cannot specify ingressing or egressing traffic when mirroring VLAN traffic and a virtual port filter.
- In normal mirroring, a monitor port cannot be added to a load share group. In one-to-many mirroring, a monitor port list can be added to a load share group, but a loopback port cannot be used in a load share group.
- Two packets are mirrored when a packet encounters both an ingress and egress mirroring filter.
- The configuration of **`remote-tag`** does not require the creation of a VLAN with the same tag; on these platforms the existence of a VLAN with the same tag as a configured **`remote-tag`** is prevented. This combination is allowed so that an intermediate remote mirroring switch can configure remote mirroring using the same remote mirroring tag as other source switches in the network. Make sure that VLANs meant to carry normal user traffic are not configured with a tag used for remote mirroring.

  When a VLAN is created with **`remote-tag`**, that tag is locked and a normal VLAN cannot have that tag. The tag is unique across the switch. Similarly if you try to create a **`remote-tag`** VLAN where **`remote-tag`** already exists in a normal VLAN as a VLAN tag, you cannot use that tag and the VLAN creation fails.

*BlackDiamond X8, BlackDiamond 8800 Series Switches and SummitStack*

The traffic filter on BlackDiamond X8, BlackDiamond 8800 series switches and SummitStack can be defined based on one of the following criteria:

- Physical port—All data that traverses the port, regardless of *VLAN* configuration, is copied to the monitor port(s). You can specify which traffic the port mirrors:
  - Ingress—Mirrors traffic received at the port.
  - Egress—Mirrors traffic sent from the port.
  - Ingress and egress—Mirrors traffic either received at the port or sent from the port.

  Note
  You can create an instance where the source is ingress only. When you add a source, pay attention to the monitor port.

  If you omit the optional parameters, all traffic is forwarded; the default for port-based mirroring is ingress and egress.
- VLAN—All data to a particular VLAN, regardless of the physical port configuration, is copied to the monitor port(s).
- Virtual port—All data specific to a VLAN on a specific port is copied to the monitor port(s).
- BlackDiamond X8, BlackDiamond 8800 series switches, and SummitStack support a maximum of 128 mirroring filters per mirroring instance.
- ExtremeXOS supports up to 16 monitor ports for one-to-many mirroring.
- Only traffic ingressing a VLAN can be monitored; you cannot specify ingressing or egressing traffic when mirroring VLAN traffic.
- Ingress traffic is mirrored as it is received (on the wire).
- Egress mirrored traffic always egresses the monitor port tagged.
- Two packets are mirrored when a packet encounters both an ingress and egress mirroring filter.
- With a monitor port or ports on BlackDiamond X8 series switch, BlackDiamond 8000 series module, a Summit family switch, or a Summit family switch in a SummitStack, all ingress mirrored traffic egressing the monitor port or ports is tagged only if the ingress packet is tagged. If the packet arrived at the ingress port as untagged, the packet egresses the monitor port or ports as untagged.
- On BlackDiamond X8 Series Switches, CPU generated packets for link-based protocols (for example, *EDP* and LACP) are not egress mirrored. CPU generated PDUs on L2 protocol blocked ports are also not egress mirrored.
- The configuration of **remote-tag** does not require the creation of a VLAN with the same tag; on these platforms the existence of a VLAN with the same tag as a configured **remote-tag** is prevented. This combination is allowed so that an intermediate remote mirroring switch can configure remote mirroring using the same remote mirroring tag as other source switches in the network. Make sure that VLANs meant to carry normal user traffic are not configured with a tag used for remote mirroring.
- When a VLAN is created with **remote-tag**, that tag is locked and a normal VLAN cannot have that tag. The tag is unique across the switch. Similarly if you try to create a **remote-tag** VLAN where **remote-tag** already exists in a normal VLAN as a VLAN tag, you cannot use that tag and the VLAN creation fails.

## Mirroring Rules and Restrictions

This section summarizes the rules and restrictions for configuring mirroring:

- Each configured mirror instance that you configure is saved, regardless of its state.
- To change monitor ports you must first remove all the filters.
- You cannot mirror the monitor port.
- The mirroring configuration is removed only if the configuration matches the removed *VLAN* or slot. If you have a match you can do the following:
  - Delete a VLAN (for all VLAN-based filters).
  - Delete a port from a VLAN (for all VLAN-, port-based filters).
  - Unconfigure a slot (for all port-based filters on that slot).
- Any mirrored port can also be enabled for load sharing (or link aggregation); however, each individual port of the load-sharing group must be explicitly configured for mirroring.
- When traffic is modified by hardware on egress, egress mirrored packets may not be transmitted out of the monitor port as they egressed the port containing the egress mirroring filter. For example, an egress mirrored packet that undergoes VLAN translation is mirrored with the untranslated VLAN ID. In addition, IP multicast packets which are egress mirrored contain the source MAC address and VLAN ID of the unmodified packet.
- The monitor port is automatically removed from all VLANs; you cannot add it to a VLAN.
- You cannot use the management port in mirroring configurations.
- You cannot run ELSM and mirroring on the same port. If you attempt to enable mirroring on a port that is already enabled for ELSM, the switch returns a message similar to the following: `Error: Port mirroring cannot be enabled on an ELSM enabled port.`
- With one-to-many mirroring, you need to enable jumbo frame support in the mirror-to port and loopback port, if you need to mirror tagged packets of length 1519 to 1522.
- The loopback port is dedicated for mirroring, and cannot be used for other configurations. This is indicated through the glowing LED.
- Egress mirrored packets are always tagged when egressing the monitor port. If an egress mirrored packet is untagged on the egress mirrored port, the mirrored copy contains a tag with an internal VLAN ID.
- As traffic approaches line rate, mirroring rate may decrease. Since mirroring makes copies of traffic, the bandwidth available will be devoted mostly to regular traffic instead of mirrored traffic when the load is high.
- A mirror port cannot be a *LAG*.

## Mirroring Examples

- To create a named mirror instance:

  `create mirror` *`mirror_name`*

- To enable mirroring on multiple ports, use the following command:

  `configure mirror` *`mirror_name`* `to port-list` *`port-list`* **`loopback-port`** *`port`*

  The port-list is a list of monitor ports that transmit identical copies of mirrored packets. The loopback-port is an unused port that is required when you mirror to a port-list. The loopback-port is not available for switching user data traffic.

- Mirroring is disabled by default. To enable mirroring on a single port, use the following command:

  `enable mirror mirror name`
- To disable mirroring, use the following command:

  `disable mirror`

Note

When you change the mirroring configuration, the switch stops sending egress packets from the monitor port until the change is complete. The ingress mirroring traffic to the monitor port and regular traffic are not affected.

*BlackDiamond X8 Series Switches, BlackDiamond 8800 Series Switches, SummitStack, and Summit Family Switches*

- The following example selects slot 3, port 4 on a modular switch or SummitStack as the monitor port and sends all traffic received at slot 6, port 5 to the monitor port:

```
enable mirror to port 3:4
configure mirror add port 6:5 ingress
```

- The following example selects slot 3, port 4 on a modular switch or SummitStack as the monitor port and sends all traffic sent from slot 6, port 5 to the monitor port:

```
enable mirroring to port 3:4
configure mirror add port 6:5 egress
```

- The following example selects port 4 on a standalone switch as the monitor port and sends all traffic ingressing the *VLAN* red to the monitor port:

```
enable mirroring to port 4
configure mirror add vlan red
```

- The following example selects port 4 on a standalone switch as the monitor port and sends all traffic ingressing the VLAN red on port 5 to the monitor port:

```
enable mirroring to port 4
configure mirror add vlan red port 5
```

- The following example selects ports 5, 6, and 7 on slot 2 on a modular switch or SummitStack as the monitor ports and sends all traffic received at slot 6, port 5 to the monitor ports.

  Slot 3, port 1 is an unused port selected as the loopback port.

```
enable mirroring to port-list 2:5-2:7 loopback-port 3:1
configure mirror add port 6:5 ingress
```

### Verifying the Mirroring Configuration

The screen output resulting from the `show mirror` command lists the ports that are involved in mirroring and identifies the monitor port. The display differs slightly depending on the platform.

## Remote Mirroring

Remote mirroring enables the user to mirror traffic to remotely connected switches. Remote mirroring allows a network administrator to mirror traffic from several different remote switches to a port at a

centralized location. Remote mirroring is accomplished by reserving a dedicated *VLAN* throughout the network for carrying the mirrored traffic. You can enable remote mirroring on the following platforms:

- BlackDiamond X8 series switches
- BlackDiamond 8000 c-, e-, xl-, and xm-series modules
- Summit Family switches

The following figure shows a typical remote mirroring topology. Switch A is the source switch that contains ports, VLANs, and/or virtual ports to be remotely mirrored. Port 25 is the local monitor port on Switch A. Switch B is the intermediate switch. Switch C is the destination switch, which is connected to the network analyzer.

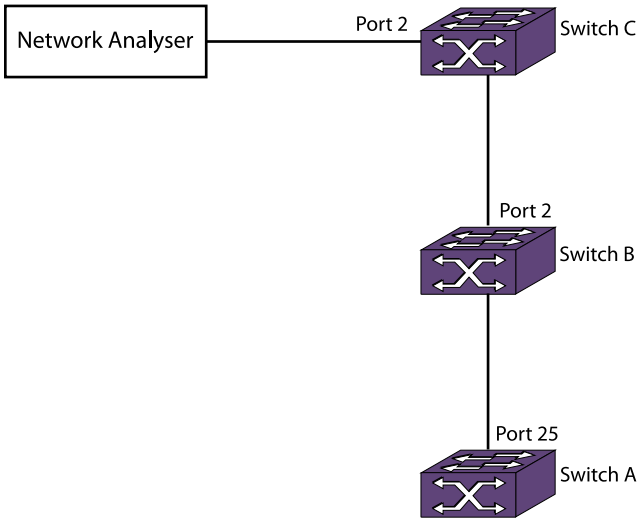



**Figure 48: Remote Mirroring Topology**

All the mirrored packets are tagged with the **remote-tag** specified by the source switch, whether the packet is already tagged or not. The intermediate switches forward the remote-tagged mirrored packets to the adjacent intermediate/destination switch, as these ports are added as tagged. The port connected to the network analyzer is added as untagged in the destination switch. This causes the destination switch to remove the remote-tag, and the mirrored packet reaches the network analyzer as the source switch sent it.

Unlike basic mirroring, remote mirroring does not remove VLAN membership from the local monitor port(s). This allows remote mirroring to use the existing network topology to transport remote mirrored packets to a destination switch.

## Configuration Details

This section describes in detail the configuration details for the topology shown in the following figure.

### *Configuring the Source Switch*

The **remote-tag** keyword followed by the tag is added in the command to enable mirroring. For example, you can establish ports 24 and 25 as monitor ports, from which any mirrored packets are transmitted with an additional *VLAN* tag containing a VLAN ID of 1000:

In the supported platforms of Summit family switches and BlackDiamond X8 and 8800 series switches, remote mirroring can also be enabled to a single port, without the **port-list** and **loopback-port** keywords.

- To establish ports 24 and 25 as monitor ports, follow the example:

```
enable mirror to port-list 24,25 loopback-port 1 remote-tag 1000
```

  The `show mirror` output displays the remote tag when remote mirroring is configured.

- To enable remote mirroring to port 25, follow the example:

```
enable mirror to port 25 remote-tag 1000
```

*Configuring the Intermediate Switch*

Reserve a *VLAN* with the remote-mirroring keyword in all the intermediate switches for remote mirroring. When you enable mirroring with **remote-tag 1000**, you need to reserve a VLAN with tag **1000** in all the intermediate switches for remote mirroring. The remote mirroring VLAN in the intermediate switches is used for carrying the mirroring traffic to the destination switch. The ports connecting the source and destination switches are added as tagged in the intermediate switches.

Another way to configure a remote mirroring VLAN is to create a normal VLAN and disable learning on the VLAN. *IGMP* snooping must be disabled on that VLAN for you to remotely mirror multicast packets through the switch.

- You may add the **remote-mirroring** keyword when you configure the tag to differentiate a normal VLAN from the remote mirroring VLAN.

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 1,2 tagged
```

  Using the **remote-mirroring** keyword automatically disables learning and IGMP snooping on the VLAN.

- You may use the following configuration for creating the remote mirroring VLAN:

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000
disable learning vlan remote_vlan
disable igmp snooping remote_vlan
```

*Configuring the Destination Switch*

Remote mirroring *VLAN* on the destination switch. The configuration on the destination switch is same as that of the intermediate switches, except that the port connected to the network analyzer is added as untagged whereas all the other ports connected to the switches are added as tagged.

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 1 tagged
configure vlan remote_vlan add ports 2 untagged
```

For a remote mirroring VLAN, the configured tag displayed by the show vlan output is remote tag instead of the normal tag.

## Remote Mirroring Guidelines

The following are guidelines for remote mirroring:

- Configurations of remote mirroring, which might cause protocol packets to be remotely mirrored, are not recommended. Since all packet types are mirrored when you configure remote mirroring, remotely mirrored protocol packets may have undesirable affects on intermediate and destination switches. Blocking *EDP* packets on a remote mirroring *VLAN* is one example of a case where you must perform an extra action to accommodate the remote mirroring of protocol packets.

  For EDP configuration on the remote mirroring VLAN, in the intermediate and destination switches you need to install *ACL* to block the EDP packets on the remote mirroring VLAN. Use the following commands for installation:

```
create access-list remote_edp " ethernet-destination-address 00:e0:2b:00:00:00 mask
ff:ff:ff:ff:ff:ff ;" "deny"
conf access-list add "remote_edp" first vlan "remote_vlan"
```

## Using Remote Mirroring with Redundancy Protocols

You can use remote mirroring with one-to-many mirroring to provide a redundant path from the source switch to the destination switch. Using EAPS or Spanning Tree can provide remote mirroring packets a redundant loop-free path through the network. You should perform the configuration of EAPS or Spanning Tree before adding mirroring filters on the source switch to prevent looping.

## Remote Mirroring with EAPS

In the following figure, the traffic from switch A is mirrored to the two ports 8:2 and 1:48 to connect to the destination switch. Using the configuration shown in the following figure, remote mirrored packets have a loop-free redundant path through the network using EAPS.

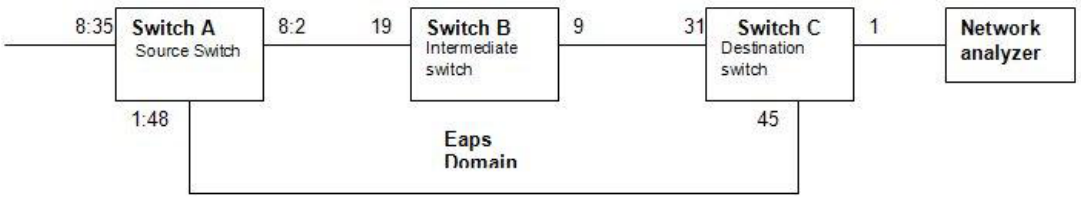



**Figure 49: Remote Mirroring with EAPS**

The configuration for the topology in the above figure is given in the following sections.

### *Switch A Configuration*

The configuration details for a BlackDiamond 8810 switch.

```
enable mirror to port-list 8:2,1:48 loopback-port 8:1 remote-tag 1000
configure mirror add port 8:35 create vlan eaps_control
configure vlan eaps_control tag 1001
configure vlan eaps_control add ports 8:2,1:48 tag
create eaps eaps1
configure eaps1 mode master
configure eaps1 primary port 8:2
configure eaps1 secondary port 1:48
configure eaps1 add control eaps_control
configure eaps1 add protected internalMirrorLoopback
enable eaps1
enable eaps
```

*Switch B Configuration*

The configuration details for a Summit X440 switch.

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 19,9 tag
create vlan eaps_control
configure vlan eaps_control tag 1001
configure vlan eaps_control add ports 19,9 tag
create eaps eaps1
configure eaps1 mode transit
configure eaps1 primary port 19
configure eaps1 secondary port 9
configure eaps1 add control eaps_control
configure eaps1 add protected remote_vlan
enable eaps1
enable eaps
```

*Switch C configuration*

The configuration details for a Summit X670 switch.

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 31,45 tag
configure vlan remote_vlan add ports 1
create vlan eaps_control
configure vlan eaps_control tag 1001
configure vlan eaps_control add ports 31,45 tag
create eaps eaps1
configure eaps1 mode transit
configure eaps1 primary port 31
configure eaps1 secondary port 45
configure eaps1 add control eaps_control
configure eaps1 add protected remote_vlan
enable eaps1
enable eaps
```

Note

The internalMirrorLoopback is an internal VMAN created when enabling mirroring to multiple ports. Depending on the platform, the internal *VLAN* or VMAN needs to be added as the protected VLAN in the source switch in order to block the ports for mirroring when EAPS is complete.

*Remote Mirroring With STP*

For the same topology shown in the following figure you can use *STP* instead of using EAPS. A sample configuration follows.

**Switch A Configuration**

```
enable mirror to port-list 8:2,1:48 loopback-port 8:1 remote-tag 1000
configure mirror add port 8:35
create vlan v1
configure vlan v1 tag 1001
configure vlan v1 add ports 8:2,1:48 tag
create stp stp1
configure stp1 mode dot1w
configure stp1 add v1 ports all
configure stp1 tag 1001
```

```
configure stp1 add vlan internalMirrorLoopback ports 8:2,1:48
enable stp1
enable stpd
```

**Switch B Configuration**

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 19,9 tag
create vlan v1
configure vlan v1 tag 1001
configure vlan v1 add ports 19,9 tag
create stp stp1
configure stp1 mode dot1w
configure stp1 add v1 ports all
configure stp1 tag 1001
configure stp1 add vlan remote_vlan ports all
enable stp1
enable stpd
```

**Switch C Configuration**

```
create vlan remote_vlan
configure vlan remote_vlan tag 1000 remote-mirroring
configure vlan remote_vlan add ports 31,45 tag
configure vlan remote_vlan add ports 1
create vlan v1
configure vlan v1 tag 1001
configure vlan v1 add ports 31,45 tag
create stp stp1
configure stp1 mode dot1w
configure stp1 add v1 ports all
configure stp1 tag 1001
configure stp1 add vlan remote_vlan ports 31,45
enable stp1
enable stpd
```

## Extreme Discovery Protocol

The *EDP* is used to gather information about neighbor Extreme Networks switches. EDP is used by the switches to exchange topology information. Information communicated using EDP includes:

- Switch MAC address (switch ID)
- Switch software version information
- Switch IP address
- Switch *VLAN* IP information
- Switch port number
- Switch configuration data: duplex and speed

EDP is enabled on all ports by default. EDP enabled ports advertise information about the Extreme Networks switch to other switches on the interface and receives advertisements from other Extreme

Networks switches. Information about other Extreme Networks switches is discarded after a timeout interval is reached without receiving another advertisement.

- To disable EDP on specified ports, use the following command:

  `disable edp ports [`*`ports`* `|` **`all`**`]`
- To enable EDP on specified ports, use the following command:

  `enable edp ports [`*`ports`* `|` **`all`**`]`
- To clear EDP counters on the switch, use the following command:

  `clear counters edp`

  This command clears the following counters for EDP protocol data units (PDUs) sent and received per EDP port:
  - Switch PDUs transmitted
  - VLAN PDUs transmitted
  - Transmit PDUs with errors
  - Switch PDUs received
  - VLAN PDUs received
  - Received PDUs with errors
- To view EDP port information on the switch, use the following command:

  `show edp`
- To view EDP information, use the following command:

  `show edp port` *`ports`* **`detail`**
- To configure the advertisement interval and the timeout interval, use the following command:

  `configure edp` **`advertisment-interval`** *`timer`* **`holddown-interval`** *`timeout`*

Refer to Displaying Port Information on page 303 for information on displaying EDP status.

## ExtremeXOS Cisco Discovery Protocol

Network elements exchange various information with neighboring elements that ranges from the switch IP/Mac addresses, switch names, versions of software, etc. This information is maintained in an Information Base, and management applications use it to create a topology map and manage the entire network. Vendor-specific discovery protocols are used to learn about network neighbor elements and their capabilities, and to exchange this information. Cisco Discovery Protocol (CDP) is a Cisco proprietary protocol that runs between direct connected network entities (routers, switches, remote access devices, IP telephones) to provide information sharing capabilities. ExtremeXOS supports CDP, and is called EXOS-CDP.

### Limitations

The ExtremeXOS support for CDP has the following limitations:

- *SNMP* for this feature is not supported.
- You should use this feature mainly for the network-endpoint devices, but you can also use it in network-network devices that have CDP support.
- When port access is controlled by net login, the port must be authorized before it receives CDP packets.

## Platform Support

This feature is supported on all ExtremeXOS platforms.

## ExtremeXOS CDP Support

The ExtremeXOS-CDP implementation gathers information about network neighbors that support the Cisco Discovery protocol. This includes edge devices (VoIP phones) in the network-edge domain.

EXOS-CDP runs on top of the controlled port of an 802 MAC client. If port access is controlled by IEEE 802.1X, the port must be authorized before you enable CDP protocol receive functionality. CDP also runs over an aggregated MAC client, and the CDP protocol information must run over all the physical MAC clients of the aggregated ports. The spanning tree state of a port does not affect the transmission of CDP PDUs.

Each CDP message contains information identifying the source port as a connection endpoint identifier. It also contains at least one network address that can be used by an NMS to reach a management agent on the device (through the indicated source port). Each CDP message contains a time-to-live value, which tells the recipient when to discard each element of learned topology information.

By default EXOS-CDP feature is disabled.

Both CDP version 1 and version 2 are supported in EXOS-CDP implementation.

### *CDP Packet Format*

The CDP control packets are encapsulated to the Sub-network Access Protocol (SNAP), and are sent as multicasts with Cisco -defined multicast MAC address 01:00:0C:CC:CC:CC.

The following table lists and describes the CDP packet fields.

| Field | Description |
|---|---|
| Version | Version of the CDP being used values 0x1 or 0x02 (In our Implementation 0x01 will be used in this release) |
| Time-to-Live (TTL) | TTL indicates the amount of time in seconds that a receiver should retain the information contained in this packet. |
| Check Sum | Standard IP checksum |
| Type | Indicates the type of the TLV |
| Length | Indicates the total length of the Type,Length,Value fields |
| Value | Indicates the value of the TLV |

### *Different Types of TLVs*

The following list identifies the different types of TLVs.

- Device ID Information TLVs

  This TLV is used to identify Device name in form of Character string. In EXOS-CDP implementation it will be the configurable value. System Mac address will be the default Device Id.

- Address Information TLVs

This TLV contains a number that indicates how many addresses are contained in the packet, followed by one entry for each address being advertised. The addresses advertised are the ones assigned to the interface on which the CDP message is sent. A device can advertise all addresses for a given protocol suite and, optionally, can advertise one or more loopback IP addresses. If the device can be managed by *SNMP*, the first entry in the address type/length/value is an address at which the device receives SNMP messages. Maximum of 32 IP address are supported.

The following table identifies and describes the various fields in the Address Information TLV.

| Field | Description |
|---|---|
| Protocol Type | Protocol type It can be one of the following values<br>1. NLPID<br>2. 802.2 format |
| Length | Length of the protocol field. |
| Protocol | One of the following values:<br>◦ 0x81—ISO CLNS (protocol type 3D 1)<br>◦ 0xCC—IP (protocol type 3D 1)<br>◦ 0xAAAA03 000000 0800—Pv6 (protocol type 3D 2)<br>◦ 0xAAAA03 000000 6003—DECNET Phase IV (protocol type 3D 2)<br>◦ 0xAAAA03 000000 809B—AppleTalk (protocol type 3D 2)<br>◦ 0xAAAA03 000000 8137—Novell IPX (protocol type 3D 2)<br>◦ 0xAAAA03 000000 80c4—Banyan VINES (protocol type 3D 2)<br>◦ 0xAAAA03 000000 0600— XNS (protocol type 3D 2)<br>◦ 0xAAAA03 000000 8019—Apollo Domain (protocol type 3D 2) |
| Address Length | Length of the address fields in bytes |
| Address | Address of the interface or the address of the system if addresses are not assigned to the interface. |

- Port ID TLVs

  The port ID type/length/value contains an ASCII character string that identifies the port on which the CDP message is sent. The type/length/value length determines the length of the string. In EXOS-CDP port description from the Pif structure is added in this TLV. If this Value is NULL then it will be the slot and port information.

- Capabilities TLVs

  The capability TLV describes the device's functional capability. It can be set to one of the bits listed below.

  The following table identifies and describes the various bits in the Capabilities TLV.

| Bit Description | Bit Description |
|---|---|
| 0x01 | Performs level 3 routing for at least one network layer protocol. |
| 0x02 | Performs level 2 transparent bridging. |
| 0x04 | Performs level 2 source-route bridging. A source-route bridge would set both this bit and bit 0x02. |

| Bit Description | Bit Description |
|---|---|
| 0x08 | Performs level 2 switching. The difference between this bit and bit 0x02 is that a switch does not run the *STP*. |
| 0x10 | Sends and receives packets for at least one network layer protocol. If the device is routing the protocol, this<br>Bit should not be set. |
| 0x20 | The bridge or switch does not forward *IGMP* Report packets on non-router ports. |
| 0x40 | Provides level 1 functionality. |

- Version TLV

  The version TLV contains a character string that provides information about the software release version that the device is running. In EXOS-CDP version will be software version.

- Platform TLV

  The platform TLV contains an ASCII character string that describes the hardware platform of the device. These platform TLV values are EXOS platform information that is the same seen when you issue the `show switch` command.

  Here are some of the possible string platform values

  - X440-24t-10G
  - X460-48t
  - BD-8810

- Native *VLAN* TLV

  The native VLAN TLV indicates, per interface, the assumed VLAN for untagged packets on the interface. CDP learns the native VLAN for an interface. These Native VLAN values are taken from VLAN Manager APIs. In ExtremeXOS 21.1 both receive and transmit of this TLV is supported. This native VLAN will be granted the value of the VLAN statistically added value. This value will be retrieved from VIPF with Client usage as VLAN_CLIENT_STATIC.

- Duplex TLV

  The duplex TLV indicates the duplex configuration of the interface. In ExtremeXOS 21.1 both receive and transmit of this TLV is supported.

- Voice VLAN Query TLV

  The voice VLAN query TLV indicates what VLAN ID is to be used for voice. This TLV is mostly used in IP Phones to configure the voice VLAN. IP Phone will query the Neighbor switch/Router for voice VLAN. The Switch/Router will replay the VLAN ID value by transmitting in CDP PDU with type voice vlan reply TLV. In ExtremeXOS 21.1 only receive functionality is supported.

- Voice VLAN Reply TLV

  The voice VLAN reply TLV indicates the voice VID for the VOIP VLAN Query. The switch/router will replay the VLAN ID value when it receives the voice vlan query TLV. In ExtremeXOS 21.1 only transmit functionality is supported. When the VOIP VLAN Query is received immediately a CDP PDU is sent with reply TLV filled.

- Power TLV

The power TLV indicates the amount of power consumed by remote devices. In ExtremeXOS 21.1 only receive functionality is supported.

- Power Request TLV

  The power request TLV indicates the power requested by the IP phone . Only the receive functionality of this TLV is supported.

- Power Available TLV

  The power available TLV indicates the power available in the powered source device. This information will be sent in the CDP PDU to the IP phone . Only the transmission of this TLV is supported.

- MTU TLV

  The MTU TLV indicates the size of the largest datagram that can be sent/received by a remote device. In ExtremeXOS 21.1 only receive functionality is supported.

- Trust Bit TLV

  The trust bit TLV offers the ability to specify whether the PC port of the phone is trusted. In ExtremeXOS 21.1 only transmit functionality is supported. This value is configured by CLI command to say whether the port is trusted/untrusted.

- Untrust QOS TLV

  The untrust QOS TLV offers the ability to specify whether the PC port of the phone is untrusted then what to mark the L2 packet cos bit. In ExtremeXOS 21.1 only transmit functionality is supported. This value is configured by user through CLI command.

- System Name TLV

  The system name TLV indicates the value of the remote device's sysName MIB object. In ExtremeXOS 21.1 both transmit and receive functionality is supported.

- System Object ID TLV

  The system object ID TLV indicates the value of the remote device's sysObjectID MIB. In ExtremeXOS 21.1 only receive functionality is supported.

- Management Address TLV

  The management address indicates the ipaddress for which the device accepts the SNMP messages. In EextremeXOS 21.1 both transmit and receive functionality is supported.

- Location TLV

  The location TLV indicates the physical location, which says either the postal pin address. In ExtremeXOS 21.1 both transmit and receive functionality is supported. The location value will be configured by `configure snmp syslocation` *`sysLocation`*

## ExtremeXOS CDP Protocol Operations

The EXOS CDP implementation performs the following tasks:

- Adds a filter for CDP Multicast MAC when the CDP protocol is enabled on port.
- Processes received CDP messages.

- Transmits CDP messages to peer when CDP is enabled.
- Maintains the current neighbor's information in the local database.

*Protocol Operations*

- Multicast Address

  This EXOS-CDP Implementation describes the operation of the protocol for reception of CDPPDUs on a single port, using a Multicast Mac address as the destination (01:00:0C:CC:CC:CC).

- CDP Initialization

  The following EXOS CDP related functions are created as part of the existing CDP process.

  - Upon system initialization, the appropriate default values are assigned to CDP global, port, and neighbor data structure.
  - Retrieval of non-volatile configuration values.

- CDP Frame Reception

  CDP frame reception consists of three phases: frame recognition, frame validation, and CDP neighbor updates.

  Frame Recognition is performed at the kernel level. A CDP raw socket is created and bound to the nexTx device. A filter to receive all the CDP packets with the Multicast MAC address of 01:00:0C:CC:CC:CC is attached during initialization, and when CDP is enabled in the port. Then a per port filter is added using expktInterface_t.

  Frames that are recognized as CDP frames are validated to determine whether they are properly constructed and contain the correct CDP packet format. Frames that pass validation criteria are used to update the contents of the neighbor entries in the database.

  If the neighbor entry does not exist in the CDP neighbor database, a new entry is created. If the neighbor entry already exists in the database, new information contained in the CDPPDU is used to replace the changed information of the existing entry in the neighbor database.

  One of the parameters received in an incoming CDPPDU is the TTL value. This determines how long the information is stored in the neighbor database before it is aged out and deleted. Aging out old data ensures that neighbor entries are purged that were originated by systems that are no longer neighbors, either because of system failure or system inactivation.

- CDP Port/Connection Failure

  If the port, or the connection to the remote system fails, the neighbor entity from the database is not deleted. It will be deleted after the associated TTL timer expires.

- Timers

  rx_timer -- A global rx_timer is defined for the CDP receive functionality. This timer is created when any one neighbor is added to the neighbor database. This timer is operated as a countdown timer, with a value of 15 seconds.

  The timers are started when at least one neighbor is created. Once the timer expires, it checks the neighbor database and deletes the neighbor entry when the hold time reaches 0.

tx_timer -- A tx_timer is defined for the CDP transmit functionality on each port. This timer is created when CDP is enabled on the port.

- CDPPDU Validation

  The following attributes are validated in the received CDP PDU:

  - The checksum must be re-calculated in order for the packet to be valid. Otherwise, the EXOS-cdp will reject the packet. Standard IP checksum mechanism is used for calculating the checksum
  - The receiving CDP packet length must be greater than, or equal to 4.
  - The packet version must be 1.

- CDP Frame Transmission

  The ExtremeXOS CDP module uses the same socket created during initialization for transmitting the CDPPDUs.

  Under normal circumstances, remote devices send CDP messages at a configured interval (default 1 minute). If you enable CDP between CDP Messages in Extreme Switches, the feature must wait for next CDPPDU to update in the neighbor database. To avoid this delay when CDP is enabled, a CDPPDUs is sent to peer.

  For every default configuration of message interval 60 seconds, CDP packets are transmitted on the CDP-enabled ports.This message interval is the configurable through CLI. The default value is 60 seconds, and the range for this message interval is 5 to 254 seconds

- Storing CDP Neighbors

  Advertisements received from the peer contain time-to-live information that indicates the length of time a receiving device holds CDP information before discarding it.

  A neighbor entry is discarded after three advertisements from the device are missed.

- CDP Packet Statistics

  CDP counters of total TX and RX CDP control packets are maintained on each port.

- CDP and Link Aggregation

  CDP will run on all member ports of the link aggregated ports. Transmit and receive of CDP control packets will be for each individual port. CDP will be enabled for each individual port of the link aggregated ports.

  For example, if ports 1,2,3 are part of link aggregation then the administrator needs to enable CDP on the Port 1,Port 2 ,Port 3 . Rx and Tx of CDP will be enabled for ports 1,2,3.

## Software-Controlled Redundant Port and Smart Redundancy

Using the software-controlled redundant port feature, you can back up a specified Ethernet port (primary) with a redundant, dedicated Ethernet port. Both ports must be on the same switch.

If the primary port fails, the switch will establish a link on the redundant port and the redundant port becomes active. Only one side of the link must be configured as redundant because the redundant port

link is held in standby state on both sides of the link. This feature provides very fast path or network redundancy.

**Note**
You cannot have any Layer 2 protocols configured on any of the *VLAN*s that are present on the ports.

The Smart Redundancy feature allows control over how the failover from a redundant port to the primary port is managed. If this feature is enabled, which is the default setting, the switch attempts to revert to the primary port as soon as it can be recovered. If the feature is disabled, the switch attempts only to recover the primary port to active if the redundant port fails.

A typical configuration of software-controlled redundant ports is a dual-homed implementation (shown in the following figure). This example maintains connectivity only if the link between switch A and switch B remains open; that link is outside the scope of the software-controlled port redundancy on switch C.

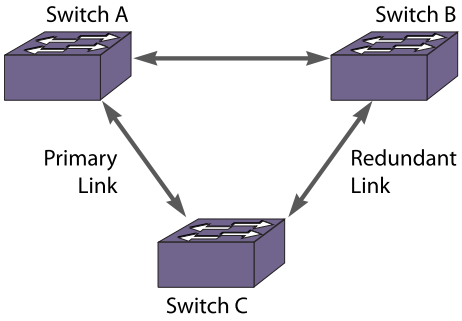



**Figure 50: Dual-Homed Implementation for Switch C**

In normal operation, the primary port is active and the software redundant switch (switch C in the following figure) blocks the redundant port for all traffic, thereby avoiding a loop in the network. If the switch detects that the primary port is down, the switch unblocks the redundant port and allows traffic to flow through that redundant port.

**Note**
The primary and redundant ports must have identical VLAN membership.

You configure the software-controlled redundant port feature either to have the redundant link always physically up but logically blocked, or to have the link always physically down. The default value is to have the link physically down, or Off.

By default, Smart Redundancy is always enabled. If you enable Smart Redundancy, the switch automatically fails over to the redundant port and returns traffic to the primary port after connectivity is restored on that port. If you do not want the automatic restoration of the primary link when it becomes active, disable Smart Redundancy.

## Guidelines for Software-Controlled Redundant Ports and Port Groups

Software-controlled redundant ports and port groups have the following guidelines and limitations:

- You cannot have any Layer 2 protocols configured on any of the *VLAN*s that are present on the ports. (You will see an error message if you attempt to configure software redundant ports on ports with VLANs running Layer 2 protocols.)
- The primary and redundant ports must have identical VLAN membership.
- The master port is the only port of a load-sharing group that can be configured as either a primary or redundant port. Also, all ports on the load-sharing group must fail before the software-controlled redundancy is triggered.
- You must disable the software redundancy on the master port before enabling or disabling load sharing.
- You can configure only one redundant port for each primary port.
- Recovery may be limited by *FDB* aging on the neighboring switch for unidirectional traffic. For bi-directional traffic, the recovery is immediate.

## Configuring Software-Controlled Redundant Ports

When provisioning software-controlled redundant ports, configure only one side of the link as redundant. To enable the software-controlled redundant port feature, the primary and redundant ports must have identical *VLAN* membership.

In the following figure only the ports on switch C would be configured as redundant.

- To configure a software-controlled redundant port, use the following command:
  `configure ports` *primaryPort* **redundant** *secondaryPort* {**link** [**on** | **off**]}

  The first port specified is the primary port. The second port specified is the redundant port.
- Unconfigure a software-controlled redundant port, use the following command and enter the primary port(s):
  `unconfigure ports` *port_list* **redundant**
- To configure the switch for the Smart Redundancy feature, use the following command:
  `enable smartredundancy` *port_list*
- To disable the Smart Redundancy feature, use the following command:
  `disable smartredundancy` *port_list*

## Verifying Software-Controlled Redundant Port Configurations

You can verify the software-controlled redundant port configuration by issuing a variety of commands.

- To display the redundant ports as well as which are active or members of load-sharing groups, use the following command:
  `show ports redundant`
- To display information on which ports are primary and redundant software-controlled redundancy ports, use the following command:
  `show port` {**mgmt** | *port_list* | **tag** *tag*} **information** {**detail**}

Refer to Displaying Port Information on page 303 for more information on the `show ports information` command.

# Configuring Automatic Failover for Combination Ports

## Summit Family Switches with Shared Copper/Fiber Gigabit Ports Only

On Summit family switches with shared copper/fiber gigabit ports, you configure automatic failover using the combination ports. These ports are called combination ports because either the fiber port or the copper port is active, but they are never active concurrently. These ports, also called redundant ports, are shared PHY copper and fiber ports.

If you plan to use the automatic failover feature, ensure that port settings are set correctly for autonegotiation.

Note
You may experience a brief episode of the link going down and recovering during the failover.

- To display the port type currently used as well as the preferred media setting, use the following command:

  `show port {mgmt | port_list | tag tag} information {detail}`

  Refer to Displaying Port Information on page 303 for more information on the `show ports information` command.

  Hardware determines when a link is lost and swaps the primary and redundant ports to maintain stability.

  After a failover occurs, the switch keeps or sticks with the current port assignment until there is another failure or until a user changes the assignment using the CLI.
- To change the uplink failover assignment, use the following command:

  `configure ports port_list preferred-medium [copper | fiber] {force}`

  The default preferred-medium is fiber. If using the **force** option, it disables automatic failover. If you force the preferred-medium to fiber and the fiber link goes away, the copper link is not used, even if available.

  Note
  For more information about combination ports on Summit family switches, refer to the *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier*.

# Displaying Port Information

You can display summary port configuration information. The following commands display summary port configuration information:

`show ports {mgmt | port_list | tag tag} configuration {no-refresh}`

`show port {mgmt | port_list | tag tag} information {detail}`

The `show ports configuration` command shows either summary configuration information on all the ports, or more detailed configuration information on specific ports. If you specify the **no-refresh** parameter, the system displays a snapshot of the data at the time you issue the command.

The `show ports information` command shows you either summary information on all the ports, or more detailed information on specific ports. The output from the command differs very slightly depending on the platform you are using.

- To display real-time port utilization information, use the following command:

  `show ports {`**mgmt** | *port_list* | **tag** *tag* | **stack-ports** *stacking-port_list*`}` **utilization** {**bandwidth** | **bytes** | **packet**}

When you use a parameter (packets, bytes, or bandwidth) with the above command, the display for the specified type shows a snapshot per port when you issue the command.

## DDMI

Digital Diagnostics Monitoring Interface (DDMI) is an optional feature which is implemented for debugging optical transceivers. DDMI provides the critical system information about 1G, 10G and 40G Optical transceiver modules.

Note

Not all transceivers support DDMI. DDMI is supported for any optic that supports DDMI. For XFP, SFP, and SFP+, there is a bit in the eeprom that designates if support is included in the module. For QSFP, the transceiver must be fiberQSFP to support DDMI.

Digital diagnostics monitor:

- Module temperature (in Celsius)
- Receiver power
- Transmitter bias current
- Transmitter power

The output of the physical value of each parameter is an analog voltage or current from the transimpedance amplifier, the laser driver, or the postamplifier. The interface allows access to device-operating parameters, and includes alarm and warning flags which alert end users when particular operating parameters are outside of a normal range. The interface is compliant with SFF-8472, "Digital Diagnostic Monitoring Interface for Optical Transceivers."

In the detailed version of the CLI command, all the threshold values for warnings and alarms are displayed for the lower and higher side of the threshold. These thresholds are not user configurable. The status value is determined based on the current value complying with the high/low thresholds for each parameter. The media type, vendor name, part number, and serial number of the optical module are also displayed.

The interface is an extension of the serial ID interface defined in the GBIC specification as well as the SFP MSA. Both specifications define a 256-byte memory map in EEPROM, which is accessible over a two-wire serial interface at the 8 bit address 1010000X (A0h). The digital diagnostic monitoring interface makes use of the 8 bit address 1010001X (A2h), so the originally defined serial ID memory map remains unchanged. The interface is identical to and fully backward compatible with both the GBIC Specification and the SFP Multi Source Agreement. The operating and diagnostics information is

monitored and reported by a Digital Diagnostics Transceiver Controller (DDTC), which is accessed via a 2-wire serial bus. For QSFPs, the eeprom uses the same address of A0, but different pages to access the data. The details of these pages can be found in the specification for QSFPs: SFF-8436: QSFP+ 10 Gbs 4X PLUGGABLE TRANSCEIVER.

As none of the DDMI information is notified through *EMS (Event Management System)* or any other method of notification, all information needs to be drawn by executing CLI commands. For 40G transceivers (QSFPs), the port may be partitioned in either 1x40G mode or 4x10G mode. As a result, when in 4X10G mode, the 4 channels of the QSFP will be shown and reported as ports since they are utilized individually. When in 1X40G mode, all 4 channels are shown as the port is now the aggregate of the 4 channels and 1 channel could bring down the entire 40G port.

## EXOS Port Description String

EXOS provides a configurable per-port “display-string” parameter that is displayed on each of the `show port` CLI commands, exposed through the *SNMP* ifAlias element, and accessible through the XML port.xsd API. Previously, the field was limited to 20 characters. This feature provides a new and separate per-port field call “description-string” that allows you to configure strings up to 255 characters.

Some characters are not permitted as they have special meanings. These characters include the following:

- “
- <
- >
- :
- <space>
- &.

The first character should by alphanumeric.

This new field is CLI accessible only through the `show port info detail` command, but you can also access it through the SNMP ifAlias object of IfXTable from IF-MIB (RFC 2233), and through the XML API.

### Configure Port decription-string

To configure up to 255 characters associated with a port, issue the command:

`configure port` *`port_list`* `description-string` *`string`*

If you configure a string longer than 64 characters, the following warning will be displayed:
`Note: Port description strings longer than 64 chars are only accessible via SNMP if the following command is issued: configure snmp ifmib ifalias size extended`

To control the accessible string size (default 64, per MIB) for the *SNMP* ifAlias object, issue the command:

`config snmp ifmib ifalias size [`**`default`**` | `**`extended`**`]`

If you choose the extended size option, the following warning will be displayed: `Warning: Changing the size to [extended] requires the use of increased 255 chars long ifAlias object of ifXtable from IF-MIB(RFC 2233)`

You can always configure a 255-character-long string regardless of the configured value of ifAlias size. Its value only affects the SNMP behavior.

## Port Isolation

The Port Isolation feature blocks accidental and intentional inter-communication between different customers residing on different physical ports. This feature provides a much simpler blocking mechanism without the use of *ACL* hardware. The fundamental requirements are as follows:

- Blocking Rules: All traffic types received on a isolation port is blocked from being forwarded through other 'isolation' ports.
- All traffic types received on an isolation port can be forwarded to any other port.
- All traffic types received on non-isolation ports are permitted to be forwarded to isolation ports.

There is no access-list hardware use. The blocking mechanism is a set of one or two table memories. These resources are not shared with other features, nor do they have any scaling limits that can be reached by configuring this feature. Port isolation can be configured in conjunction with other features, including VPLS, IDM, and *Extreme Network Virtualization (XNV)*. However, you cannot configure a mirror-to port to be an isolated port.

### Configuring Port Isolation

To enable isolation mode on a per-port basis, perform the following tasks:

1. Issue the `configure ports` *`port_list`* **`isolation`** [**`on`** | **`off`**] command.

   You can issue this command either on a single port or on a master port of a load share group. If you issue this command on a non-master port of a load share group, the command will fail. When a port load share group is formed, all of the member ports assume the same isolation setting as the master port.
2. Issue the `show port info detail` and `show port info` outputs to display the current isolation settings.

## Energy Efficient Ethernet

As part of the 802.az standard, Energy Efficient Ethernet (EEE) is targeted at saving energy in Ethernet networks for a select group of physical layer devices, or PHYs. The PHYs use EEE during idle periods to reduce power consumption. If you do not utilize EEE, the PHY is fully powered up even when there is no traffic being sent. Enabling EEE reduces significant power consumption on the switch. Within EXOS, a PHY/switch combination, or a PHY with autogrEEEN capability is provided to allow EEE to work. In a typical setup, the PHY and switch communicate when to enter or exit low power idle (LPI) mode.

AutoGrEEEn technology implements the EEE standard directly into PHYs, and enables the EEE mode when interfacing with non-EEE enabled MAC devices,. This allows you to make existing network equipment EEE-compliant by simply changing the PHY devices. EEE is currently only implemented for copper ports

Previously, most wireline communications protocols used continuous transmission, consuming power whether or not data was sent. EEE puts the PHY in an active mode only when real data is sent on the media. To save energy during gaps in the data stream, EEE uses a signaling protocol that allows a transmitter to indicate the data gap and to allow the link to go idle. This signaling protocol is also used to indicate that the link needs to resume after a pre-defined delay.

The EEE protocol uses an LPI signal that the transmitter sends to indicate that the link can go idle. After sending LPI for a period (Ts = time to sleep), the transmitter stops signaling altogether, and the link becomes quiescent. Periodically, the transmitter sends some signals so that the link does not remain quiescent for too long without a refresh. When the transmitter wishes to resume the fully functional link, it sends normal idle signals. After a pre-determined time (Tw = time to wake), the link is active and data transmission resumes.

The EEE protocol allows the link to be re-awakened at any time; there is no minimum or maximum sleep interval. This allows EEE to function effectively in the presence of unpredictable traffic. The default wake time is defined for each type of PHY, and is designed to be similar to the time taken to transmit a maximum length packet at the particular link speed.

The refresh signal serves the same purpose as the link pulse in traditional Ethernet. The heartbeat of the refresh signal ensures that both partners know that the link is present, and allows for immediate notification following a disconnection. The frequency of the refresh prevents any situation where one link partner is disconnected and another inserted without causing a link fail event. This maintains compatibility with security mechanisms that rely on continuous connectivity and require notification when a link is broken.

The maintenance of the link through refresh signals also allows higher layer applications to understand that the link is continuously present, preserving network stability. You can also use the refresh signal to test the channel, and create an opportunity for the receiver to adapt to changes in the channel characteristics. For high speed links, this is vital to support the rapid transition back to the full speed data transfer without sacrificing data integrity. The specific makeup of the refresh signal is designed for each PHY type to assist the adaptation for the medium supported.

## Supported Platforms

EEE is supported on the following Extreme Networks platforms:

- BD X – 10G48T.

  Note
  EEE is only supported at 10G on this card.

- Summit Series 670V-48T.

  Note
  EEE is only supported at 10G on this switch.

- Summit Series X440 – all copper ports will support EEE.
- Summit X450-G2 - supported on BASE-T ports only.
- Summit Series X460-G2 - all copper ports will support EEE.
- Summit series E4G400 – EEE is implemented through autogrEEEn.
- E4G200 – EEE is implemented through autogrEEEn.

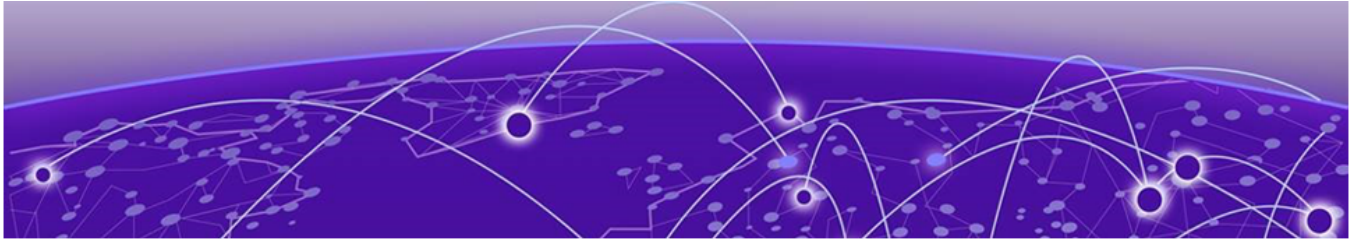
## Configuring Energy Efficient Ethernet

- To enable or disable the EEE feature on EXOS, use the following command:

  `configure port {`*`port_list`*`} eee enable [`**`on`**` | `**`off`**`]`

  The **`enable on`** specifies that the port advertises to its link partner that it is EEE capable at certain speeds. If both sides, during auto-negotiation, determine that they both have EEE on and are compatible speed wise, they will determine other parameters (how long it takes to come out of sleep time, how long it takes to wake up) and the link comes up. During periods of non-activity, the link will shut down parts of the port to save energy. This is called LPI for low power idle. When one side sees it must send something, it wakes up the remote and then transmits.
- To display the statistics of the EEE features, use the following command:

  `show port `*`port_list`*` `**`eee`**



# Universal Port



Universal Port is a flexible framework that enables automatic switch configuration in response to special events such as:

- User login and logoff
- Device connection to or disconnection from a port
- Time of day
- *EMS (Event Management System)* event messages

Note
The Universal Port feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

The primary component of the Universal Port feature is the profile, which is a special form of command script that runs when triggered by the events mentioned above.

Profiles execute commands and use variables as do the scripts described in Using CLI Scripting on page 354. The primary difference is that a profile can be executed manually or automatically in response to switch events.

Note
The term *profile* is distinct from the term *policy* because a policy is only one particular application of a profile.

Universal Port works with the following ExtremeXOS components and third-party products:

- ExtremeXOS Network Login
- ExtremeXOS LLDP
- ExtremeXOS CLI Scripting
- Status Monitoring and Statistics
- RADIUS servers
- Active directory services such as LDAP and Microsoft Active Directory

The following are some examples of how you can use Universal Port on a network:

- Automatically provision a VoIP phone and the attached switch port with appropriate *PoE (Power over Ethernet)* budget and *QoS (Quality of Service)* settings when the phone connects.
- Create security policies that can follow a user as the user roams around a campus. For example, an engineer can walk from Building 1 to Building 5, plug his PC into the network and be authenticated with the appropriate access rights and ACLs.
- Support separate authentication for VoIP phones and workstations on the same port.
- Create profile templates with variables so that you can re-use templates with different address ranges and parameters.
- Apply different security policies for different locations (for example, a restricted area).
- Disable wireless access after business hours.

Note

Special scripts can be run when the switch boots. For more information, see Using Autoconfigure and Autoexecute Files on page 1547.

## Profile Types

The ExtremeXOS software supports two types of profiles: static and dynamic.

### Static Profiles

Static profiles are so named because they are not triggered by dynamic system events. To trigger a static profile, you must enter a command at the switch prompt or run a script that contains the command to start a static profile. The following guidelines apply to static profiles:

- Static profiles are not limited to individual ports and can include system-wide configuration changes.
- Static profiles are not assigned to a port and are not specific to a device or a user.
- Changes made by static profiles are persistent. They are saved in the switch configuration and are preserved during system reboots.

Static profiles are typically used to establish default switch settings. Using scripts and variables, you can create static profiles that serve as templates for initializing switches or reconfiguring switches to manually respond to network or business events. These templates can simplify complex configuration tasks such as Netlogin.

## Dynamic Profiles

Dynamic profiles are so named because they are dynamically triggered by the following types of events:

- Device discovery and disconnect
- User- or standards-based authentication and logoff
- Time of day
- Switch events reported by the *EMS*

Dynamic profiles are event- or action-driven and do not require an administrator to start the profile.

Without dynamic profile support, IT personnel must be available when devices are added, moved, or changed so they can configure both the network port and the new device. These tasks typically take a long time, do not support mobility, and are often prone to human error.

When dynamic profiles are configured properly and a device connects to an edge port, a triggering event triggers a profile that runs a script to configure the port appropriately. The script can use system run-time variables and information gathered from tools such as Netlogin and *LLDP (Link Layer Discovery Protocol)* to customize the port configuration for the current network environment. For example, the profile can customize the port configuration based on the user ID or MAC address. Dynamic profiles allow you to automate the network response to a variety of network events.

Dynamic profiles create temporary states. For example, if a power outage causes the switch to restart, all ports return to the default configuration. When a triggering event such as a specific device connection occurs again, the profile is applied again. When the device is no longer connected, the disconnect event can trigger another profile to unconfigure the port.

The temporary state configured by a dynamic profile is configured by prepending the `configure cli mode non-persistent` command to the script. The temporary nature of profile configuration is critical for network security. Imagine a situation where a dynamic security profile is used. If the information granting access to specific network resources is saved in the configuration, the switch is restarted, and a user loses network connectivity on a secure port, the secure port still provides network access after the switch restarts. Anybody else can access network resources simply by connecting to that secure port.

Although the switch configuration returns to the default values after a restart, there is no automatic configuration rollback for dynamic profiles. For example, if a profile grants secure access to network resources at user login, the configuration is not automatically rolled back when the user logs off. To roll back the configuration at user log off, you must create another profile that responds to user log off events.

To support configuration rollback, the scripting feature allows you to save information used in dynamic profiles in variables. When a profile is activated and you want the option to roll back to the previous default setting, some information must be saved, such as the default *VLAN (Virtual LAN)* setting or the default configuration of a port. Essentially anything modified from the previous setting can be preserved for future use by the profile that rolls back the configuration.

There can be multiple profiles on a switch, but only one profile runs at a time. Data from a trigger event is used to select the appropriate profile, and that data can also be used to make decision points within a profile. A typical example is the use of a *RADIUS (Remote Authentication Dial In User Service)* server to specify a particular profile and then apply port-based policies to the user based on the user's location.

There is no profile hierarchy and no software validation to detect if a new profile conflicts with older profile. If two profiles conflict, the same profile might produce different results, depending on the events leading up to the profile trigger. When you create profiles, you must be familiar with all profiles on the switch and avoid creating profiles that conflict with each other.

## Dynamic Profile Trigger Types

The following sections describe the types of dynamic profile trigger types: device, user authentication, time, and event-management system.

### Device Triggers

Device triggers launch a profile when a device connects to or disconnects from a port.

The two types of device triggers are labeled device-detect and device-undetect in the software. Profiles that respond to these triggers are called device-detect profiles or device-undetect profiles.

Typically, a device-detect profile is used to configure a port for the device that has just connected.

Likewise, a device-undetect profile is used to return the port to a default configuration after a device disconnects. A variety of different devices can be connected to a port. When devices connect to the network, Universal Port helps provide the right configuration at the port.

Device triggers respond to the discovery protocols IEEE 802.1ab *LLDP* and ANSI/TIA-1057 LLDP-MED for Voice-over-IP (VoIP) phone extensions. A device-detect trigger occurs when an LLDP packet reaches a port that is assigned to a device-detect profile. A device-undetect trigger occurs when periodically transmitted LLDP packets are not received anymore. LLDP age-out occurs when a device has disconnected or an age-out time has been reached. LLDP must be enabled on ports that are configured for device-detect or device-undetect profiles. LLD P is described in LLDP Overview on page 369.

The combination of device triggers and LLDP enables the custom configuration of devices that connect to switch ports. For example, VoIP phones can send and receive information in addition to normal device identification information. The information sent through LLDP can be used to identify the maximum power draw of the device. The switch can then set the maximum allocated power for that port.

If the switch does not have enough *PoE* left, the switch can take action to lower the PoE loading and try again. The switch can also transmit additional VoIP files and call server configuration information to the phone so the phone can register itself and receive necessary software and configuration information.

There can only be one device-detect profile and one device-undetect profile per port. To distinguish between different connecting devices, you can use if-then-else statements in a profile along with detailed information provided through LLDP.

### User Authentication Triggers

User authentication triggers launch a profile when a user or an identified device logs in or out of the network using Network Login on page 756.

The network login feature does not permit any access beyond the port until the user or device is authenticated.

The two types of user authentication triggers are labeled user-authenticate and user-unauthenticated in the software. Profiles that respond to these triggers are called user-authenticate profiles or user-unauthenticated profiles. Typically, a user-authenticate profile is used to configure a port for a user and device that has just connected. Likewise, a user-unauthenticated profile is used to return the port to a default configuration after a user or device disconnects. Successful network login triggers the user-authenticate profile, and either an explicit logout, a session time out, or a disconnect triggers the user-unauthenticated profile.

Note

VoIP phones are also capable of being authenticated before being allowed on the network. The phone begins 802.1X authentication based on a personal username and password. This authentication step is available and supported by the latest firmware from vendors such as Avaya and Mitel.

Network login requires a *RADIUS* server for user or device authentication.

The RADIUS server provides the following features:

- Centralized database for network authentication
- Further centralization when connected to an LDAP or Active Directory database
- Dynamic switch configuration through Vendor Specific Attributes (VSAs)

VSAs are values that are passed from the RADIUS server to the switch after successful authentication. VSAs can be used by the switch to configure connection attributes such as security policy, *VLAN*, and location. For more information on RADIUS and VSAs, see Security on page 859.

The following sections introduce each of the network login event types that can trigger profiles:

- 802.1X Network Login
- MAC-Based Network Login
- Web-Based Network Login

### *802.1X Network Login*

Network login 802.1X requires 802.1X client software on the device to be authenticated.

At login, the user supplies a user name and password, which the switch passes to the RADIUS server for authentication. When the user passes authentication, the RADIUS server notifies the switch, and the user-authenticate profile is triggered.

One advantage of 802.1X network login is that it can uniquely identify a user. A disadvantage is that not all devices support 802.1X authentication. For more information, see Network Login on page 756.

### *MAC-Based Network Login*

MAC-based network login requires no additional software, and it does not require any interaction with the user.

When network login detects a device with a MAC address that is configured on the switch, the switch passes the MAC address and an optional password to the RADIUS server for authentication. When the

device passes authentication, the RADIUS server notifies the switch, and the user-authenticate profile is triggered.

One advantage of MAC-based network login is that it requires no special software. A disadvantage is that security is based on the MAC address of the client, so the network is more vulnerable to spoofing attacks. For more information, see Network Login on page 756.

Note

MAC-based authentication can also be used to identify devices. For example, an entire MAC address or some bits of the MAC address can identify a device and trigger switch port auto-configuration similar to the *LLDP*-based device detect event. The difference between MAC-based authentication and LLDP authentication is that MAC-based authentication does not provide information on the connected device. The advantage of MAC-based authentication is that it enables non-LLDP devices to trigger profiles.

### *Web-Based Network Login*

Web-based network login requires a *DHCP (Dynamic Host Configuration Protocol)* server and may require a DNS server.

At login, the user supplies a user name and password through a web browser client, which the switch passes to the RADIUS server for authentication. When the user passes authentication, the RADIUS server notifies the switch, and the user-authenticate profile is triggered.

Some advantages of web-based network login are that it can uniquely identify a user and it uses commonly available web client software. Some disadvantages are a lower level of security and the IP configuration requirement. For more information, see Network Login.

## Time Triggers

Time triggers launch a profile at a specific time of day or after a specified period of time.

For example, you can use time triggers to launch profiles at the following times:

- 6:00 p.m. every day
- One-time after 15 minutes
- One hour intervals

A profile that uses a time trigger is called a time-of-day profile. You might use a time trigger to launch a profile to disable guest *VLAN* access, shut down a wireless service, or power down a port after business hours. Time triggers enable profiles to perform timed backups for configurations, policies, statistics, and so forth. Anything that needs to happen on a regular basis or at a specific time can be incorporated into a time-of-day profile. Time-of-day profiles are not limited to non-persistent-capable CLI commands and can use any command in the ExtremeXOS CLI.

Unlike the device-detect and user-authenticate triggers, time triggers do not have an equivalent function to the device-undetect or user-unauthenticated triggers. If you need the ability to unconfigure changes made in a time-of-day profile, just create another time-of-day profile to make those changes.

## Event Management System Triggers

*EMS*-event triggers launch a profile when the EMS produces a message that conforms to a predefined definition that is configured on the switch. The ExtremeXOS EMS feature is described in CLEAR-Flow on page 948.

Profiles that respond to EMS-event triggers are called EMS-event profiles. Typically, an EMS-event profile is used to change the switch configuration in response to a switch or network event.

The EMS events that trigger Universal Port profiles are defined in EMS filters and can be specified in more detail with additional CLI commands.

You can create EMS filters that specify events as follows:

- Component.subcomponent
- Component.condition
- Component.subcomponent.condition

You can use the `show log components` command to display all the components and subcomponents for which you can filter events. If you specify a filter to take action on a component or subcomponent, any event related to that component triggers the profile. You can use the show log events all command to display all the conditions or events for which you can filter events. If you decide that you want to configure a profile to take action on an *ACL (Access Control List)* policy change, you can add a filter for the ACL.Policy.Change event.

You can further define an event that triggers a Universal Port profile by specifying an event severity level and text that must be present in an event message.

When a specified event occurs, event information is passed to the Universal Port profile in the form of variables, which can be used to modify the switch configuration.

EMS-triggered profiles allow you to configure responses for any EMS event listed in the `show log components` and `show log events all` commands. However, you must be careful to select the correct event and corresponding response for each profile. For example, if you attempt to create a Universal Port log target for a specific event (component.subcomponent.condition) and you accidentally specify a component (component), the profile is applied to all events related to that component. Using EMS-triggered profiles is similar to switch programming. They provide more control and therefore more opportunity for misconfiguration.

Unlike the device-detect and user-authenticate triggers, EMS-event triggers do not have an equivalent function to the device-undetect or user-unauthenticated triggers.

If you need the ability to unconfigure changes made in an EMS-event profile, just create another static or dynamic profile to make those changes.

# How Device-detect Profiles Work

Device-detect profiles enable dynamic port configuration without the use of a *RADIUS* server. Device-detect profiles and device-undetect profiles are triggered as described earlier in Device Triggers on page 312.

When a device connects to a port that has a device-detect profile configured, the switch runs the specified profile and stops. Only one device detect profile can be configured for a port, so the same profile runs each time a device is detected on the port. Only one device-undetect profile can be configured for a port, so the same profile is run each time the switch detects that all previously-connected devices are no longer connected.

## How User Authentication Profiles Work

User-authentication profiles can be assigned to user groups or individual users. Typically, a company creates profiles for groups such as software engineering, hardware engineering, marketing, sales, technical support, operations, and executive. These kinds of categories make profile management more streamlined and simple.

The authentication process starts when a switch receives an authentication request through network login. The authentication request can be for a specific user or a MAC address. A user name and password might be entered directly or by means of other security instruments, such as a smart card. A MAC address would be provided by *LLDP*, which would need to be operating on the ingress port. Network login enforces authentication before granting access to the network. All packets sent by a client on the port do not go beyond the port into the network until the user is authenticated through a *RADIUS* server.

The switch authenticates the user through a RADIUS server, which acts as a centralized authorization point for all network devices. The RADIUS server can contain the authentication database, or it can serve as a proxy for a directory service database, such as LDAP or Active Directory. The switch also supports optional backup authentication through the local switch database when a RADIUS server is unavailable.

The RADIUS server responds to the switch and either accepts or denies user authentication. When user authentication is accepted, the RADIUS server can also send Vendor Specific Attributes (VSAs) in the response. The VSAs can specify configuration data for the user such as the Universal Port profile to run for logon, a *VLAN* name, a user location, and a Universal Port profile to run for logout. Extreme Networks has defined vendor specific attributes that specify configuration settings and can include variables to be processed by the Universal Port profile. If profile information is not provided by the RADIUS server, the user-authenticate profile is used.

Profiles are stored and processed on the switch. When a user name or MAC address is authenticated, the switch places the appropriate port in forwarding mode and runs either a profile specified by the RADIUS server, or the profile defined for the authentication event. The profile configures the switch resources for the user and stops running until it is activated again.

When a user or MAC address is no longer active on the network, due to logoff, disconnect, or inactivity, user unauthentication begins.

To complete unauthentication, the switch stops forwarding on the appropriate port and does one of the following:

1. Runs an unauthenticate profile specified by the RADIUS server during authentication.
2. Runs an unauthenticate profile configured on the switch and assigned to the affected port.
3. Runs the authenticate profile initially used to authenticate the user .

The preferred unauthenticate profile is one specified by the RADIUS server during authentication. If no unauthenticate profiles are specified, the switch runs the authenticate profile used to authenticate the user or device.

## Profile Configuration Guidelines

You can configure both static and dynamic profiles using the CLI, ExtremeManagement, or the Ridgeline Universal Port Manager.

### Obtaining Profiles

You can write your own profiles.

You can obtain profiles from the Extreme Networks website, another Extreme Networks user or partner, or Extreme Networks professional services.

Sample profiles are listed in Sample Universal Port Configurations on page 332. The Universal Port Handset Provisioning Module is a collection of profiles and documentation that is available with other samples on the Extreme Networks website.

### Profile Rules

All profiles have the following restrictions:

- Maximum 5000 characters in a profile.
- Maximum 128 profiles on a switch.
- Profiles are stored as part of the switch configuration file.
- Copy and paste is the only method to transfer profile data using the CLI.
- Unless explicitly preceded with the command `configure cli mode persistent`, all non-persistent-capable commands operate in non-persistent mode when operating in dynamic profiles.
- Unless explicitly preceded with the command `configure cli mode non-persistent`, all non-persistent-capable commands operate in persistent mode when operating in static profiles.

Note

There is no profile hierarchy, which means users must verify there are no conflicting rules in static and dynamic profiles. This is a normal requirement for ACLs, and is standard when using policy files or dynamic ACLs.

When the switch is configured to allow non-persistent-capable commands to operate in non-persistent mode, the switch configuration can rollback to the configuration that preceded the entry of the non-persistent-capable commands. This rollback behavior enables ports to return to their initial state when a reboot or power cycle occurs.

### Multiple Profiles on the Same Port

Multiple Universal Port profiles can be created on a switch, but only one profile per event can be applied per port. Different profiles on the same port apply to different events—different authentication events for different devices or users.

You can configure multiple user profiles on a port or a group of ports. For instance, you might create user-authentication profiles for different groups of users, such as Engineering, Marketing, and Sales.

You can also configure a device-triggered profile on a port that supports one or more user profiles. However, you can configure only one device-triggered profile on a port.

## Supported Configuration Commands and Functions

Static and dynamic profiles support the full ExtremeXOS command set. They also support the built-in functions described in Using CLI Scripting on page 354.

Commands that are executed in persistent mode become part of the saved switch configuration that persists when the switch is rebooted. Commands that are executed in non-persistent mode configure temporary changes that are not saved in the switch configuration and do not persist when the switch is rebooted.

However, a subset of these commands operates by default in non-persistent mode when executed in a dynamic profile.

Most commands operate only in persistent mode. The subset of commands that operate in non-persistent mode are called non-persistent-capable commands. The Universal Port feature uses the non-persistent-capable commands to configure temporary changes that could create security issues if the switch were rebooted or reset. The use of non-persistent-capable commands in scripts and Universal Port profiles allows you to make temporary configuration changes without affecting the default configuration the next time the switch is started.

The following table shows the non-persistent capable CLI commands.

*Non-Persistent-Capable Configuration Commands*

### ACL Commands

```
configure access-list add dynamic_rule [ [[first | last] {priority
p_number} {zone zone} ] | [[before | after] rule] | [ priority p_number
{zone zone} ]] [ any | vlan vlanname | ports portlist ] {ingress |
egress}
```

```
configure access-list delete ruleName [ any | vlan vlanname | ports
portlist | all] {ingress | egress}
```

### LLDP Commands

```
configure lldp ports portlist [advertise | dont advertise | no-advertise
| dcbx] {all-tlvs | management-address | port-description | system-
capabilities | system-description | system-name | vendor-specific}
```

### Port Commands

```
disable port [port_list | all]
```

```
disable jumbo-frame ports [all | port_list]
```

```
enable port [port_list | all]

enable jumbo-frame ports [all | port_list]
```

### Power over Ethernet Commands

```
configure inline-power label string ports port_list

configure inline-power operator-limit milliwatts ports [all |port_list]

configure inline-power priority [critical | high | low] ports port_list

disable inline-power

disable inline-power ports [all | port_list]

disable inline-power slot slot

enable inline-power

enable inline-power ports [all | port_list]

enable inline-power slot slot

unconfigure inline-power priority ports [all | port_list]
```

### VLAN Commands

```
configure {vlan} vlan_name add ports [port_list | all] {tagged |
untagged} {{stpd} stpd_name} {dot1d | emistp | pvst-plus}}

configure ip-mtu mtu vlan vlan_name
```

### QOS/Rate-limiting Commands

```
configure ports port_list {qosprofile} qosprofile
```

### Show Commands

All show commands can be executed in non-persistent mode.

By default, all commands operate in persistent mode with the following exceptions:

- In Universal Port dynamic profiles, the non-persistent-capable commands operate in non-persistent mode unless preceded by the `configure cli mode persistent` command in the profile.
- In the CLI, CLI scripts, and static profiles, the non-persistent-capable commands operate in non-persistent mode only when preceded by the `configure cli mode non-persistent` command.

You can use the `configure cli mode persistent` command and the `configure cli mode non-persistent` command to change the mode of operation for non-persistent-capable commands multiple times within a script, profile, or configuration session.

## Universal Port Variables

Universal Port uses CLI Scripting variables to make system and trigger event information available to profiles. For more information, see Using CLI Scripting on page 354.

Variables allow you to create profiles and scripts that respond to the state of the switch as defined in the variables. When a profile is triggered, the system passes variables to the profile. You can also create and use variables of your own. User-defined variables are limited to the current context unless explicitly saved.

Note

You must enable CLI scripting before using variables or executing a script.

If you save variables (as described in Saving, Retrieving, and Deleting Session Variables on page 364), certain data from one profile can be reused in another profile for another event. For example, between login and logout events, the data necessary for the rollback of a port configuration can be shared.

The following sections describe the variables that are available to profiles:

- Common Variables
- User Profile Variables
- Device Detect Profile Variables
- Event Profile Variables

### *Common Variables*

The following table shows the variables that are always available for use by any script. These variables are set up for use before a script or profile is executed.

**Table 34: Common Variables**

| Variable Syntax | Definition |
|---|---|
| $STATUS | Status of last command execution. |
| $CLI.USER | Username for the user who is executing this CLI. |
| $CLI.SESSION_ID | An identifier for a session. This identifier is available for the roll-back event when a device or user times out. |
| $CLI.SESSION_TYPE | Type of session of the user. |
| $EVENT.NAME | This is the event that triggered this profile. |
| $EVENT.TIME | Time this event occurred. The time is in seconds since epoch. |
| $EVENT.TIMER_TYPE | Type of timer, which is periodic or non_periodic. |
| $EVENT.TIMER_NAME | Name of the timer that the Universal Port is invoking. |

**Table 34: Common Variables (continued)**

| Variable Syntax | Definition |
|---|---|
| $EVENT.TIMER_LATE_SECS | Time difference between when the timer fired and when the actual shell was run in seconds. |
| $EVENT.PROFILE | Name of the profile that is being run currently. |

*User Profile Variables*

The following table shows the variables available to user profiles.

**Table 35: User Profile Variables**

| Variable Syntax | Definition |
|---|---|
| $EVENT.USERNAME | Name of user authenticated. This is a string with the MAC address for MAC-based user-login. |
| $EVENT.NUMUSERS | Number of authenticated supplicants on this port after this event occurred.<br><br>Note:<br>For user-authenticated events, the initial value of this variable is 0. For user unauthenticated events, the initial value is 1. |
| $EVENT.USER_MAC | MAC address of the user. |
| $EVENT.USER_PORT | Port associated with this event. |
| $EVENT.USER_VLAN | *VLAN* associated with this event or user. |
| $EVENT.USER_ALL_VLANS | When a user is authenticated to multiple VLANs, this variable includes all VLANs for which the user is authenticated. |
| $EVENT.USER_IP | IP address of the user if applicable. Otherwise, this variable is blank. |

*Device Detect Profile Variables*

The following table shows the variables available to device detect profiles.

**Table 36: Device Profile Variables**

| Variable Syntax | Definition |
|---|---|
| $EVENT.DEVICE | Device identification string.<br>Possible values for EVENT.DEVICE are: AVAYA_PHONE, GEN_TEL_PHONE, ROUTER, BRIDGE, REPEATER, WLAN_ACCESS_PT, DOCSIS_CABLE_SER, STATION_ONLY and OTHER.<br>These strings correspond to the devices that the *LLDP* application recognizes and reports to the Universal Port management application. |
| $EVENT.DEVICE_IP | The IP address of the device (if available). Blank if not available. |
| $EVENT.DEVICE_MAC | The MAC address of the device (if available). Blank if not available. |

**Table 36: Device Profile Variables (continued)**

| Variable Syntax | Definition |
|---|---|
| $EVENT.DEVICE_POWER | The power of the device in milliwatts (if available). Blank if not available. |
| $EVENT.DEVICE_MANUFACTURER_NAME | The manufacturer of the device. |
| $EVENT.DEVICE_MODEL_NAME | Model name of the device. |
| $EVENT.USER_PORT | Port associated with the event. |

*Event Profile Variables*

The following table shows the variables available to event profiles.

**Table 37: Event Profile Variables**

| Variable Syntax | Definition |
|---|---|
| $EVENT.NAME | The event message. |
| $EVENT.LOG_DATE | The event date. |
| $EVENT.LOG_TIME | The event time. |
| $EVENT.LOG_COMPONENT_SUBCOMPONENT | The component and subcomponent affected by the event as it appears in the show log components command display. |
| $EVENT.LOG_EVENT | The event condition as it appears in the show log events command display. |
| $EVENT.LOG_FILTER_NAME | The *EMS* filter that triggered the profile. |
| $EVENT.LOG_SEVERITY | The event severity level defined in EMS. |
| $EVENT.LOG_MESSAGE | The event message with arguments listed in the format %1%. |
| $EVENT.LOG_PARAM_0 to $EVENT.LOG_PARAM_9 | Event arguments 0 to 9. |

## Collecting Information from Supplicants

A supplicant is a device such as a VoIP phone or workstation that connects to the switch port and requests network services.

As described in LLDP, LLDP is a protocol that can be used to collect information about device capabilities from attached devices or supplicants.

To use Universal Port with LLDP, you must enable LLDP on the port.

Note

Avaya and Extreme Networks have developed a series of extensions for submission to the standards consortium for inclusion in a later version of the LLDP-MED standard:

- Avaya Power conservation mode
- Avaya file server
- Avaya call server

The following is an example of information provided through LLDP about an IP phone:

```
LLDP Port 1 detected 1 neighbor
Neighbor: (5.1)192.168.10.168/00:04:0D:E9:AF:6B, age 7 seconds
- Chassis ID type: Network address (5); Address type: IPv4 (1)
Chassis ID     : 192.168.10.168
- Port ID type: MAC address (3)
Port ID     : 00:04:0D:E9:AF:6B
- Time To Live: 120 seconds
- System Name: "AVAE9AF6B"
- System Capabilities : "Bridge, Telephone"
Enabled Capabilities: "Bridge, Telephone"
- Management Address Subtype: IPv4 (1)
Management Address        : 192.168.10.168
Interface Number Subtype  : System Port Number (3)
Interface Number          : 1
Object ID String          : "1.3.6.1.4.1.6889.1.69.1.13"
- IEEE802.3 MAC/PHY Configuration/Status
Auto-negotiation       : Supported, Enabled (0x03)
Operational MAU Type   : 100BaseTXFD (16)
- MED Capabilities: "MED Capabilities, Network Policy, Inventory"
MED Device Type : Endpoint Class III (3)
- MED Network Policy
Application Type  : Voice (1)
Policy Flags      : Known Policy, Tagged (0x1)
VLAN ID           : 0
L2 Priority       : 6
DSCP Value        : 46
- MED Hardware Revision: "4625D01A"
- MED Firmware Revision: "b25d01a2_7.bin"
- MED Software Revision: "a25d01a2_7.bin"
- MED Serial Number: "061622014487"
- MED Manufacturer Name: "Avaya"
- MED Model Name: "4625"
- Avaya/Extreme Conservation Level Support
Current Conservation Level: 0
Typical Power Value       : 7.4 Watts
Maximum Power Value       : 9.8 Watts
Conservation Power Level  : 1=7.4W
- Avaya/Extreme Call Server(s): 192.168.10.204
- Avaya/Extreme IP Phone Address: 192.168.10.168 255.255.255.0
Default Gateway Address       : 192.168.10.254
- Avaya/Extreme CNA Server: 0.0.0.0
```

```
- Avaya/Extreme File Server(s): 192.168.10.194
- Avaya/Extreme IEEE 802.1q Framing: Tagged
```

**Note**

LLDP is tightly integrated with IEEE 802.1X authentication at edge ports. When used together, LLDP information from authenticated end point devices is trustable for automated configuration purposes. This tight integration between 802.1X and LLDP protects the network from automation attacks.

## Supplicant Configuration Parameters

As described in LLDP, *LLDP* is a protocol that can be used to configure attached devices or supplicants. The following LLDP parameters are configurable on the switch ports when device-detect profiles execute are:

- *VLAN* Name
- Port VLAN ID
- Power Conservation Mode
- Avaya File Server
- Avaya Call server
- 802.1Q Framing

## Universal Port Configuration Overview

Because Universal Port operates with multiple ExtremeXOS software features and can operate with multiple third-party products, Universal Port configuration can require more than just the creation of profiles and triggers.

No single overview procedure can cover all the possible Universal Port configurations. The following sections provide overviews of the common types of Universal Port configurations.

### Device-Detect Configurations

A Universal Port device-detect configuration requires only a switch and supplicants. If *PoE* devices will connect to the switch, the switch should support PoE. Supplicants should support *LLDP* in the applicable software or firmware.

**Note**

To support supplicant configuration, you might consider adding a *DHCP* server to your network.

Use the following procedure to configure Universal Port for device detection:

1. Create a device-detect profile as described in Creating and Configuring New Profiles on page 326.
2. Create a device-undetect profile as described in Creating and Configuring New Profiles on page 326.
3. Assign the device-detect profile to the edge ports as described in Configuring a Device Event Trigger on page 327.

4. Assign the device-undetect profile to the edge ports as described in Configuring a Device Event Trigger on page 327.
5. Verify that correct profiles are assigned to correct ports using the following command:

   `show upm event event-type`
6. Enable LLDP message advertisements on the ports that are configured for device-detect profiles as described in LLDP.
7. Test profile operation as described in Verifying a Universal Port Profile on page 330.

## User-Authentication Configurations

A Universal Port user-authenticate configuration requires specific components:

- An Extreme Networks switch, which might need to include *PoE* support.
- *RADIUS* server for user authentication and VSA transmission.
- Supplicants that support the authentication method you select. *LLDP* support is recommended, but is optional when MAC address authentication is used.

**Note**

To support supplicant configuration, you might consider adding a *DHCP* server to your network. For VoIP applications, you can use a TFTP server and a call server to provide for additional supplicant configuration.

Use the following procedure to configure Universal Port for user login:

1. Configure the RADIUS server as described in Security on page 859. The configuration should include the following:
   - User ID and password for RADIUS clients.
   - Extreme Networks custom VSAs.
   - Addition of the edge switch as a RADIUS client.
2. Create a user-authenticate profile as described in Creating and Configuring New Profiles on page 326.
3. Create a user-unauthenticate profile as described in Creating and Configuring New Profiles on page 326.
4. Assign the user-authenticate profile to the edge ports as described in Configuring a User Login or Logout Event Trigger on page 327.
5. Assign the user-unauthenticate profile to the edge ports as described in Configuring a User Login or Logout Event Trigger on page 327.
6. Configure network login on the edge switch as described in Network Login.
7. Configure the edge switch as a RADIUS client as described in Security.
8. Verify that correct profiles are assigned to correct ports by entering the following command:

   `show upm event event-type`
9. Enable LLDP message advertisements on the ports that are configured for device-detect profiles as described in LLDP Overview on page 369.
10. Test profile operation as described in Verifying a Universal Port Profile on page 330.

### Time-of-Day Configurations

To configure Universal Port to use a time-of-day profile, use the following procedure:

1. Create a profile as described in Creating and Configuring New Profiles on page 326.
2. Create and configure a timer as described in Configuring a Universal Port Timer on page 328.
3. Create the timer trigger and attach it to the profile as described in Configuring a Timer Trigger on page 328.

### EMS-Event Configurations

To configure Universal Port to use an *EMS*-event profile, use the following procedure:

1. Create the EMS-Event profile as described in Creating and Configuring New Profiles on page 326.
2. Create and configure an event filter to identify the trigger event as described in Creating an EMS Event Filter on page 328.
3. Create the event trigger and attach it to the profile and filter as described in Configuring an EMS Event Trigger on page 328.
4. Enable the event trigger as described in Enabling and Disabling an EMS Event Trigger on page 329.

## Using Universal Port in an LDAP or Active Directory Environment

The *RADIUS* server can operate in proxy mode with information stored in a central directory service such as LDAP or Active Directory.

This proxy mode is configured between the RADIUS server and the central directory service. Once configured, supplicants can be authenticated from the central directory service.

For more information, see the following:

- Setting Up Open LDAP on page 932.
- RADIUS server product documentation
- Product documentation for your central directory service

## Configuring Universal Port Profiles and Triggers

You can configure both static and dynamic profiles using the CLI, ExtremeManagement, or the Ridgeline Universal Port Manager.

Note

In the CLI, “upm” is used as an abbreviation for the Universal Port feature.

### Creating and Configuring New Profiles

When you create and configure a new profile, you are basically writing a script within a profile that can be triggered by system events. For more information on the rules, commands, and variables that apply to profiles, see Profile Configuration Guidelines on page 317.

1. Create and configure a new profile using the following command:

   ```
   configure upm profile profile-name maximum execution-time seconds
   ```

2. After you enter the command, the switch prompts you to add command statements to the profile as shown in the following example:

```
switch # create upm profile detect-voip
Start typing the profile and end with a . as the first and the only character on a
line.
Use - edit upm profile <name> - for block mode capability
create log message Starting_Script_DETECT-voip
set var callServer 192.168.10.204
set var fileServer 192.168.10.194
set var voiceVlan voice
set var CleanupProfile CleanPort
set var sendTraps false
#
.
switch #
```

The example above creates a log entry and sets some variables, but it is not complete. This example shows that after you enter the `create upm profile` command, you can enter system commands. When you have finished entering commands, you can exit the profile creation mode by typing the period character ( . ) at the start of a line and pressing **[Enter]**.

## Editing an Existing Profile

- Edit an existing profile.

  `edit upm profile` *`profile-name`*

## Configuring a Device Event Trigger

There are two types of device event triggers, which are named as follows in the CLI: device-detect and device-undetect. When you configure a device event trigger, you assign one of the two device event trigger types to a profile and specify the ports to which the triggered profile applies.

To configure a device event trigger, use the following command:

`configure upm event` *`upm-event`* `profile` *`profile-name`* `ports` *`port_list`*

Replace *`upm-event`* with one of the device event trigger types: device-detect or device-undetect.

## Configuring a User Login or Logout Event Trigger

There are two types of user event triggers, which are named as follows in the CLI: user-authenticate and user-unauthenticated. When you configure a user event trigger, you assign one of the two user event trigger types to a profile and specify the ports to which the triggered profile applies.

To configure a user event trigger, use the following command:

`configure upm event` *`upm-event`* `profile` *`profile-name`* `ports` *`port_list`*

Replace *`upm-event`* with one of the device event trigger types: user-authenticate or user-unauthenticated.

## Configuring a Universal Port Timer

To configure a Universal Port timer, you must complete two steps:

1. Create the timer by using the following command:

   ```
   create upm timer timer-name
   ```

2. Configure the timer by using the following commands:

   ```
   configure upm timer timer-name after time-in-secs {every seconds}
   ```

   ```
   configure upm timer timer-name at month day year hour min secs {every seconds}
   ```

## Configuring a Timer Trigger

When you configure a timer trigger, you assign a configured timer to a profile. When the configured time arrives, the switch executes the profile.

To configure a timer trigger, use the following command:

```
configure upm timer timerName profile profileName
```

Replace *timerName* with the timer name and *profileName* with the profile name.

## Creating an EMS Event Filter

An *EMS* event filter identifies an event that can be used to trigger a profile. To create an EMS event filter, use the following procedure:

1. Create a log filter to identify the event by using the following command:

   ```
   create log filter name {copy filter name}
   ```

2. Configure the log filter using the following commands:

   ```
   configure log filter name [add | delete] {exclude} events [event-condition | [all | event-component] {severity severity {only}}]
   ```

   ```
   configure log filter name [add | delete] {exclude} events [event-condition | [all | event-component] {severity severity {only}}] [match | strict-match] type value
   ```

## Configuring an EMS Event Trigger

When you configure an *EMS* event trigger, you identify an EMS filter that defines the event and a profile that runs when the event occurs. To configure an EMS event-triggered profile, use the following procedure:

1. Create a log target to receive the event notification by using the following command:

   ```
   create log target upm {upm_profile_name}
   ```

2. Configure the log target to specify a filter and any additional parameters that define the event by using the following commands:

   ```
   configure log target upm {upm_profile_name} filter filter-name
   {severity [[severity] {only}]}
   ```

   ```
   configure log target upm {upm_profile_name} match {any | regex}
   ```

### Enabling and Disabling an EMS Event Trigger

When you configure an EMS event trigger, it is disabled.

To enable an EMS event trigger or disable a previously enabled trigger, use the following commands:

```
enable log target upm {upm_profile_name}
```

```
disable log target upm {upm_profile_name}
```

### Unconfiguring a Timer

To unconfigure a timer, use the following command:

```
unconfigure upm timer timerName profile profileName
```

## Managing Profiles and Triggers

Use the following actions to manage profiles and triggers.

### Manually Executing a Static or Dynamic Profile

Profiles can be run from the CLI by configuring the system to run as it would when the trigger events happen. This facility is provided to allow you to test how the system behaves when the actual events happen. The actual configuration is applied to the switch when the profile is run.

Note

Variables are not validated for correct syntax.

To manually execute a profile, use the following command:

```
run upm profile profile-name {event event-name} {variables variable-
string}
```

Example:

```
run upm profile afterhours
```

If the **variables** keyword is not present, but an event variable is specified, you are prompted for various environment variables appropriate for the event, including the VSA string for user authentication.

## Displaying a Profile

To display a profile, use the following command:

```
show upm profile name
```

## Displaying Timers

Display a list of timers and associated timer information.

```
show upm timers
```

## Displaying Universal Port Events

Display events for trigger types. You can display a list of events that relate to one of the following trigger types:

- device-detect
- device-undetect
- user-authenticate
- user-unauthenticated

To display a list of Universal Port events for one of the above triggers, use the following command:

```
show upm event event-type
```

Replace *event-type* with one of the trigger types listed above.

## Displaying Profile History

To display a list of triggered events and associate event data, enter one of the following commands:

```
show upm history {profile profile-name | event upm-event | status [pass
| fail] | timer timer-name | detail}
```

```
show upm history exec-id number
```

## Verifying a Universal Port Profile

To verify a Universal Port profile configuration, trigger the profile and verify that it works properly. Trigger the profile based on the trigger type as follows:

- Device triggers–plug in the device
- Authentication triggers–authenticate a device or user
- Timer triggers–temporarily configure the timer for an approaching time
- *EMS* event triggers–reproduce the event to which the trigger responds

You can use the commands described in Managing Profiles and Triggers on page 329 to view information about the profile and how it behaves.

Because Universal Port works with multiple switch features, you might want to enter commands to examine the configuration of those features.

The following commands are an example of some of the commands that can provide additional information about profile operation.

Run:

```
show lldp
```

```
show lldp neighbors
```

```
show log {messages [memory-buffer | nvram]} {events {event-condition |
event-component]} {severity severity {only}} {starting [date date time
time | date date | time time]} {ending [date date time time | date date
| time time]} {match regex} {chronological}
```

```
show netlogin {port port_list [ {vlan} vlan_name | vlan vlan_list]}
{dot1x {detail}} {mac} {web-based}
```

## Handling Profile Execution Errors

To conserve resources, the switch stores only the last execution log for the profile that resulted in an error.

- To see a tabular display showing the complete history of the last 100 profiles, use the following command:

  ```
  show upm history {profile profile-name | event upm-event | status
  [pass | fail] | timer timer-name | detail}
  ```

  Use the **detail** keyword to display the actual executions that happened when the profile was run.

- To display a specific execution that was run, use the following command:

  ```
  show upm history exec-id number
  ```

  Select the exec-id number from the list in the tabular display.

## Disabling and Enabling a Profile

Universal Port profiles are automatically enabled when they are created.

To disable a profile or enable a previously disabled profile, use the following commands:

```
disable upm profile profile-name
```

```
enable upm profile profile-name
```

## Deleting a Profile

To delete a profile, use the following command:

```
delete upm profile profile-name
```

## Deleting a Timer

To delete a timer, use the following command:

```
delete upm timer timer-name
```

## Deleting an EMS Event Trigger

To delete an *EMS* event trigger, use the following command:

```
delete log target upm {upm_profile_name}
```

# Sample Universal Port Configurations

The following sections provide example configurations for a switch, profile, and policy file.

## Sample MAC Tracking Profile

The example in this section shows how to create a profile that takes action based on the MAC tracking feature. When the MAC tracking feature detects a MAC move in a *VLAN*, the MAC tracking feature generates an *EMS* log, which then triggers a profile.

**Note**

You can also use the Identity Management feature to configure ports in response to MAC device detection events. For more information, see CLEAR-Flow on page 948.

### *Switch Configuration*

The general switch configuration is as follows:

```
#Vlan  config
create vlan v1
configure v1 add ports 1:17-1:18
configure vlan v1 ipadd 192.168.10.1/24
#mac tracking config
create fdb mac-tracking entry 00:01:02:03:04:01
create fdb mac-tracking entry 00:01:02:03:04:02
create fdb mac-tracking entry 00:01:02:03:04:03
create fdb mac-tracking entry 00:01:02:03:04:04
create fdb mac-tracking entry 00:01:02:03:04:05
#Log filter configuration
create log filter macMoveFilter
configure log filter "macMoveFilter" add events "FDB.MACTracking.MACMove"
#Meter configuration for ingress /egress rate limit
create meter m1
configure meter m1 peak-rate 250 mbps
create meter m2
configure meter m2 peak-rate 500 mbps
```

### *MAC Tracking EMS Log Message*

The MAC tracking feature produces the following *EMS* log message and message parameters:

```
The MAC address %0% on VLAN '%1%' has moved from port %2% to port %3%"
```

```
EVENT.LOG_PARAM_1 "vlan name"
EVENT.LOG_PARAM_2 "source port"
EVENT.LOG_PARAM_3 "moved port"
```

*Profile Configuration*

The profile is configured as follows:

```
create upm profile macMove   ;# editor
enable cli scripting
create access-list dacl1 "source-address  192.168.10.0/24 " "permit ;count dacl1"
create access-list dacl2 "source-address  192.168.11.0/24 " "permit ;count dacl2"
create access-list dacl3 "source-address  192.168.15.0/24 " "deny ;count dacl3"
create access-list dacl4 "source-address  192.168.16.0/24 " "deny ;count dacl4"
create access-list dacl5 "source-address  192.168.17.0/24 " "deny ;count dacl5"
configure access-list add dacl1 first ports $(EVENT.LOG_PARAM_3)
configure access-list add dacl2 first ports $(EVENT.LOG_PARAM_3)
configure access-list add dacl3 first ports $(EVENT.LOG_PARAM_3)
configure access-list add dacl4 first ports $(EVENT.LOG_PARAM_3)
configure access-list add dacl5 first ports $(EVENT.LOG_PARAM_3)
conf access-list ingress_limit vlan v1
conf access-list ingress_limit ports $(EVENT.LOG_PARAM_3)
conf access-list egress_limit any   ;# enter . for SAVE/EXIT
log target configuration
create log target upm "macMove"
configure log target upm "macMove" filter "macMoveFilter"
enable log target upm "macMove"
```

*Policy File Configuration*

This example uses the following two policy files:

```
Ingress rate limit (ingress_limit.pol)
==================================
entry ingress {
if {
ethernet-source-address 00:AA:00:00:00:01;
ethernet-destination-address 00:BB:00:00:00:01;
} then {
Meter m1;
count c1;
}
}
Egress QoS (egress_limit.pol)
==================================
entry egress {
if {
ethernet-source-address 00:BB:00:00:00:01;
ethernet-destination-address 00:AA:00:00:00:01;
} then {
qosprofile qp2;
count c2;
}
}
```

*Console Logs*

The following show commands display the switch configuration:

```
* (debug) BD-12804.7 # show log con fil
Log Filter Name: DefaultFilter
I/                                                Severity
E Component   SubComponent Condition               CEWNISVD
- ----------- ------------ ---------------------- --------
```

```
I All                                                 ********
Log Filter Name: macMoveFilter
I/                                                    Severity
E Component   SubComponent Condition                  CEWNISVD
- ----------- ------------ ------------------------ --------
I FDB         MACTracking  MACMove                  ---N----
* (debug) BD-12804.14 # sh fdb mac-tracking configuration
SNMP trap notification    : Disabled
MAC address tracking table (5 entries):
00:01:02:03:04:01
00:01:02:03:04:02
00:01:02:03:04:03
00:01:02:03:04:04
00:01:02:03:04:05
* (debug) BD-12804.15 #
* (debug) BD-12804.27 # show meter
-------------------------------------------------------------------------------
Name   Committed Rate(Kbps) Peak Rate(Kbps)   Burst Size(Kb)
-------------------------------------------------------------------------------
m1                     --         250000          --
m2                     --         500000          --
Total number of Meter(s) : 2
* (debug) BD-12804.28 #
```

The following show commands display the switch status after a MAC address move:

```
==================================
(debug) BD-12804.7 # show log
05/14/2009 11:33:54.89 <Noti:ACL.Policy.bind> MSM-A:
Policy:bind:egress_limit:vlan:*:port:*:
05/14/2009 11:33:54.89 <Info:pm.config.loaded> MSM-A: Loaded Policy: egress_limit number
of entries 1
05/14/2009 11:33:54.89 <Info:pm.config.openingFile> MSM-A: Loading policy egress_limit
from file /config/egress_limit.pol
05/14/2009 11:33:54.89 <Noti:ACL.Policy.bind> MSM-A:
Policy:bind:ingress_limit:vlan:*:port:1:18:
05/14/2009 11:33:54.88 <Noti:ACL.Policy.bind> MSM-A:
Policy:bind:ingress_limit:vlan:v1:port:*:
05/14/2009 11:33:54.87 <Info:pm.config.loaded> MSM-A: Loaded Policy: ingress_limit number
of entries 1
05/14/2009 11:33:54.87 <Info:pm.config.openingFile> MSM-A: Loading policy ingress_limit
from file /config/ingress_limit.pol
05/14/2009 11:33:54.72 <Noti:UPM.Msg.upmMsgExshLaunch> MSM-A: Launched profile macMove
for the event log-message
A total of 8 log messages were displayed.
* (debug) BD-12804.8 # show upm history
-------------------------------------------------------------------------------
Exec   Event/              Profile          Port Status Time Launched
Id     Timer/ Log filter
-------------------------------------------------------------------------------
1      Log-Message(macMoveF macMove          --- Pass    2009-05-14 11:33:54
-------------------------------------------------------------------------------
Number of UPM Events in Queue for execution: 0
* (debug) BD-12804.9 # sh upm history detail
UPM Profile: macMove
Event:  Log-Message(macMoveFilter)
Profile Execution start time: 2009-05-14 11:33:54
Profile Execution Finish time: 2009-05-14 11:33:54
Execution Identifier: 1 Execution Status: Pass
Execution Information:
1 # enable cli scripting
2 # configure cli mode non-persistent
3 # set var EVENT.NAME LOG_MESSAGE
4 # set var EVENT.LOG_FILTER_NAME "macMoveFilter"
```

```
5 # set var EVENT.LOG_DATE "05/14/2009"
6 # set var EVENT.LOG_TIME "11:33:54.72"
7 # set var EVENT.LOG_COMPONENT_SUBCOMPONENT "FDB.MACTracking"
8 # set var EVENT.LOG_EVENT "MACMove"
9 # set var EVENT.LOG_SEVERITY "Notice"
10 # set var EVENT.LOG_MESSAGE "The MAC address %0% on VLAN '%1%' has moved from port %2%
to port %3%"
11 # set var EVENT.LOG_PARAM_0 "00:01:02:03:04:05"
12 # set var EVENT.LOG_PARAM_1 "v1"
13 # set var EVENT.LOG_PARAM_2 "1:17"
14 # set var EVENT.LOG_PARAM_3 "1:18"
15 # set var EVENT.PROFILE macMove
16 # enable cli scripting
17 # create access-list dacl1 "source-address  192.168.10.0/24 " "permit ;count dacl1"
18 # create access-list dacl2 "source-address  192.168.11.0/24 " "permit ;count dacl2"
19 # create access-list dacl3 "source-address  192.168.15.0/24 " "deny ;count dacl3"
20 # create access-list dacl4 "source-address  192.168.16.0/24 " "deny ;count dacl4"
21 # create access-list dacl5 "source-address  192.168.17.0/24 " "deny ;count dacl5"
22 # configure access-list add dacl1 first ports $(EVENT.LOG_PARAM_3)
done!
23 # configure access-list add dacl2 first ports $(EVENT.LOG_PARAM_3)
done!
24 # configure access-list add dacl3 first ports $(EVENT.LOG_PARAM_3)
done!
25 # configure access-list add dacl4 first ports $(EVENT.LOG_PARAM_3)
done!
26 # configure access-list add dacl5 first ports $(EVENT.LOG_PARAM_3)
done!
27 # conf access-list ingress_limit vlan v1
done!
28 # conf access-list ingress_limit ports $(EVENT.LOG_PARAM_3)
done!
29 # conf access-list egress_limit any
done!
--------------------------------------------------------------------------------
Number of UPM Events in Queue for execution: 0
* (debug) BD-12804.10 #
* (debug) BD-12804.7 # show fdb mac-tracking statistics
MAC Tracking Statistics      Thu May 14 11:41:10 2009
Add         Move        Delete
MAC Address        events      events      events
=====================================================
00:01:02:03:04:01           0           0           0
00:01:02:03:04:02           0           0           0
00:01:02:03:04:03           0           0           0
00:01:02:03:04:04           0           0           0
00:01:02:03:04:05           1           1           0
=====================================================
0->Clear Counters  U->page up  D->page down ESC->exit
(debug) BD-12804.5 # show access-list
Vlan Name    Port   Policy Name          Dir      Rules  Dyn Rules
=====================================================================
*            *      egress_limit         ingress  1      0
*            1:18   ingress_limit        ingress  1      5
v1           *      ingress_limit        ingress  1      0
* (debug) BD-12804.6 # show access-list dynamic
Dynamic Rules: ((*)- Rule is non-permanent )
(*)dacl1                          Bound to 1 interfaces for application Cli
(*)dacl2                          Bound to 1 interfaces for application Cli
(*)dacl3                          Bound to 1 interfaces for application Cli
(*)dacl4                          Bound to 1 interfaces for application Cli
(*)dacl5                          Bound to 1 interfaces for application Cli
(*)hclag_arp_0_4_96_1e_32_80      Bound to 0 interfaces for application HealthCheckLAG
* (debug) BD-12804.7 #
```

```
* (debug) BD-12804.7 #
================================================================================
```

## Universal Port Handset Provisioning Module Profiles

The Universal Port Handset Provisioning Module provides the following profiles:

- Device-Triggered Generic Profile
- Authentication-Triggered Avaya Profile

### *Device-Triggered Generic Profile*

This is a template for configuring network parameters for VoIP phone support without 802.1X authentication. The profile is triggered after an *LLDP* packet is detected on the port.

**Note**

The MetaData information is used by the Ridgeline to create a user-friendly interface to modify the variables. You can ignore the MetaData while using the CLI.

```
#*********************************
# Last Updated: April 11, 2007
# Tested Phones: Avaya 4610, 4620, 4625
# Requirements: LLDP capable devices
#*********************************
# @MetaDataStart
# @ScriptDescription "This is a template for configuring network parameters for VoIP
phones support LLDP but without authentication.  The module is triggered through the
detection of an LLDP packet on the port.   The following network side configuration is
done: enable SNMP traps, QOS assignment, adjust POE reservation values based on device
requirements, add the voiceVlan to the port as tagged. "
# @VariableFieldLabel "Voice VLAN name"
set var voicevlan voiceavaya
# @VariableFieldLabel "Send trap when LLDP event happens (true or false)"
set var sendTraps false
# @VariableFieldLabel "Set QoS Profile (true or false)"
set var setQuality false
# @MetaDataEnd
#
if (!$match($EVENT.NAME,DEVICE-DETECT)) then
create log message Starting_LLDP_Generic_Module_Config
# VoiceVLAN configuration
configure vlan $voicevlan add port $EVENT.USER_PORT tagged
#SNMP Trap
if (!$match($sendTraps,true)) then
create log message Config_SNMP_Traps
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
#Link Layer Discovery Protocol-Media Endpoint Discover
create log message Config_LLDP
configure lldp port $EVENT.USER_PORT advertise vendor-specific med capabilities
configure lldp port $EVENT.USER_PORT advertise vendor-specific dot1 vlan-name vlan
$voicevlan
configure lldp port $EVENT.USER_PORT advertise vendor-specific med policy application
voice vlan $voicevlan dscp 46
configure lldp port $EVENT.USER_PORT advertise vendor-specific med power-via-mdi
```

```
#Configure POE settings per device requirements
create log message Config_POE
configure inline-power operator-limit $EVENT.DEVICE_POWER ports $EVENT.USER_PORT
#QoS Profile
if (!$match($setQuality,true)) then
create log message Config_QOS
configure port $EVENT.USER_PORT qosprofile qp7
endif
endif
if (!$match($EVENT.NAME,DEVICE-UNDETECT) && $match($EVENT.DEVICE_IP,0.0.0.0)) then
create log message Starting_LLDP_Generic_UNATUH_Module_Config
if (!$match($sendTraps,true)) then
create log message UNConfig_SNMP_Traps
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
create log message UNConfig_LLDP
unconfig lldp port $EVENT.USER_PORT
if (!$match($setQuality,true)) then
create log message UNConfig_QOS
unconfig qosprofile ports $EVENT.USER_PORT
endif
unconfig inline-power operator-limit ports $EVENT.USER_PORT
endif
if (!$match($EVENT.NAME,DEVICE-UNDETECT) && !$match($EVENT.DEVICE_IP,0.0.0.0)) then
create log message DoNothing_0.0.0.0
create log message $EVENT.TIME
endif
create log message End_LLDP_Generic_Module_Config
```

### *Authentication-Triggered Generic Profile*

This profile has been created for phones that support an authentication protocol and assumes that the phone does not support *LLDP* and is provisioned using *DHCP* options.

This is a template for configuring network parameters for 802.1X authenticated devices. The module is triggered through successful authentication or unauthentication of the device.

```
#*******************************************
# Last Updated: April 11, 2007
# Tested Phones: Avaya 4610, 4620, 4625
# Requirements: 802.1X capable devices, netlogin configured and enabled on deployment
ports
#*******************************************
# @MetaDataStart
# @ScriptDescription "This is a template for configuring network parameters for 802.1X
authenticated devices.  The module is triggered through successful authentication of the
device.  The following network side configuration is done: QOS assignment and enables DOS
protection.  When used with IP phones, phone provisioning is done through DHCP options."
# @Description "VLAN name to add to port"
set var vlan1 voiceavaya
# @VariableFieldLabel "Set QoS Profile (yes or no)"
set var setQuality yes
# @Description "QoS Profile (0-100)"
set var lowbw 50
# @VariableFieldLabel "QoS MAX Bandwidth (0-100)"
set var highbw 100
# @VariableFieldLabel "Enable Denial of Service Protection (yes or no)"
set var dosprotection yes
# @MetaDataEnd
################################
# Start of USER-AUTHENTICATE block
```

```
#################################
if (!$match($EVENT.NAME,USER-AUTHENTICATED)) then
############
#QoS Profile
############
# Adds a QOS profile to the port
if (!$match($setQuality,yes)) then
create log message Config_QOS
configure port $EVENT.USER_PORT qosprofile qp7
configure qosprofile qp7 minbw $lowbw maxbw $highbw ports $EVENT.USER_PORT
endif
#
########################
#Security Configurations
########################
create log message Applying_Security_Limits
# enables Denial of Service Protection for the port
if (!$match($dosprotection,yes)) then
enable dos-protect
create log message DOS_enabled
endif
#
endif
################################
# End of USER-AUTHENTICATE block
################################
#
#
###################################
# Start of USER-UNAUTHENTICATE block
###################################
if (!$match($EVENT.NAME,USER-UNAUTHENTICATED)) then
create log message Starting_8021x_Generic_UNATUH_Module_Config
if (!$match($setQuality,yes)) then
create log message UNConfig_QOS
unconfig qosprofile ports $EVENT.USER_PORT
endif
unconfig inline-power operator-limit ports $EVENT.USER_PORT
endif
#################################
# End of USER-UNAUTHENTICATE block
#################################
create log message End_802_1x_Generic_Module_Config
```

*Authentication-Triggered Avaya Profile*

This script has been created for Avaya phones that support both 802.1X authentication and *LLDP*. Instead of using *DHCP* options, the phone is provisioned using LLDP parameters developed jointly by Extreme Networks and Avaya.

```
#*******************************
# Last Updated: April 11, 2007
# Tested Phones: SW4610, SW4620
# Requirements: 802.1X authentication server, VSA 203 and VSA 212 from authentiication
server. QP7 defined on the switch
#*******************************
# @MetaDataStart
# @ScriptDescription "This is a template for configuring LLDP capable Avaya phones using
the authentication trigger.  This module will provision the phone with the following
parameters: call server, file server,  dot1q, dscp, power.  Additionally the following
network side configuration is done: enable SNMP traps and QOS assignment"
# @VariableFieldLabel "Avaya phone call server IP address"
set var callserver 192.45.95.100
```

```
# @VariableFieldLabel "Avaya phone file server IP address"
set var fileserver 192.45.10.250
# @VariableFieldLabel "Send trap when LLDP event happens (true or false)"
set var sendTraps true
# @VariableFieldLabel "Set QoS Profile (true or false)"
set var setQuality true
# @MetaDataEnd
#
if (!$match($EVENT.NAME,USER-AUTHENTICATED)) then
create log message Starting_Avaya_VOIP_802.1X_AUTH_Module_Config
if (!$match($sendTraps,true)) then
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
enable lldp port $EVENT.USER_PORT
configure lldp port $EVENT.USER_PORT advertise vendor-specific dot1 vlan-name
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme call-server
$callserver
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme file-server
$fileserver
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme dot1q-
framing tag
if (!$match($setQuality,true)) then
configure port $EVENT.USER_PORT qosprofile qp7
endif
endif
#
if (!$match($EVENT.NAME,USER-UNAUTHENTICATED)) then
create log message Starting_Avaya_VOIP_802.1X_UNATUH_Module_Config
if (!$match($sendTraps,true)) then
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
disable lldp port $EVENT.USER_PORT
if (!$match($setQuality,true)) then
unconfig qosprofile ports $EVENT.USER_PORT
endif
endif
create log message End_Avaya_VOIP_802.1X_Module_Config
```

## Sample Static Profiles

The following configuration creates a profile and runs it statically:

```
* BD-10808.4 # Create upm profile p1
Enable port 1:1
.
* BD-10808.4 #run upm profile p1
* BD-10808.4 # show upm history exec 8006
UPM Profile: p1
Event:  User Request    , Time run: 2006-10-18 11:56:15
Execution Identifier: 8006      Execution Status: Pass
Execution Information:
1 # enable cli scripting
2 # set var EVENT.NAME USER-REQUEST
3 # set var EVENT.TIME 1161172575
```

```
4 # set var EVENT.PROFILE p1
5 # enable port 1:1
```

This profile creates and configures EAPS on the edge switch for connecting to the aggregation switch, creates specific VLANs and assigns tags, configures network login, and configures the *RADIUS* client component on the switch.

```
#*********************************************
# Last Updated: May 11, 2007
# Tested Devices: SummitX EXOS 12.0
# Description: This profile configures the switch with an EAPs ring, creates specified
# vlans, configure network login, RADIUS.
#*********************************************
# @MetaDataStart
# @ScriptDescription "This is a template for configuring network parameters for edge
Summit devices.  The profile will configure the listed features:  EAPs ring, Network
login, 802.1X, vlans, and default routes."
# @VariableFieldLabel "Create EAPs ring? (yes or no)"
set var yneaps yes
# @VariableFieldLabel "Name of EAPs domain"
set var eapsdomain upm-domain
# @VariableFieldLabel "Primary port number"
set var eapsprimary 23
# @VariableFieldLabel "Secondary port number"
set var eapssecondary 24
# @VariableFieldLabel "Name of EAPs control VLAN"
set var eapsctrl upm_ctrl
# @VariableFieldLabel "Tag for EAPs control VLAN"
set var eapsctrltag 4000
# @VariableFieldLabel "Create standard VLANs? (yes or no)"
set var ynvlan yes
# @VariableFieldLabel "Name of Voice vlan"
set var vvoice voice
# @VariableFieldLabel "Voice VLAN tag"
set var vvoicetag 10
# @VariableFieldLabel "Voice VLAN virtual router"
set var vvoicevr vr-default
# @VariableFieldLabel "Name of Security Video"
set var vidsec vidcam
# @VariableFieldLabel "Security Video VLAN tag"
set var vidsectag 40
# @VariableFieldLabel "Security Video VLAN virtual router"
set var vidsecvr vr-default
# @VariableFieldLabel "Name of Data vlan"
set var vdata datatraffic
# @VariableFieldLabel "Data VLAN tag"
set var vdatatag 11
# @VariableFieldLabel "Data VLAN virtual router"
set var vdatavr vr-default
# @VariableFieldLabel "Enable Network Login? (yes or no)"
set var ynnetlogin yes
# @VariableFieldLabel "RADIUS Server IP Address"
set var radserver 192.168.11.144
# @VariableFieldLabel "RADIUS Client IP Address"
set var radclient 192.168.11.221
# @VariableFieldLabel "RADIUS Server Shared Secret"
set var radsecret goextreme
# @VariableFieldLabel "Network Login port list"
set var netloginports 1-20
# @MetaDataEnd
##################################
# Start of EAPs Configuration block
##################################
if (!$match($yneaps,yes)) then
```

```
create log message Config_EAPs
config eaps config-warnings off
create eaps $eapsdomain
config eaps $eapsdomain mode transit
config eaps $eapsdomain primary port $eapsprimary
config eaps $eapsdomain secondary port $eapssecondary
create vlan $eapsctrl
config $eapsctrl tag $eapsctrltag
config $eapsctrl qosprofile qp8
config $eapsctrl add port $eapsprimary tagged
config $eapsctrl add port $eapssecondary tagged
config eaps $eapsdomain add control vlan $eapsctrl
enable eaps
enable eaps $eapsdomain
else
create log message EAPs_Not_Configured
endif
#############
#VLAN Config
#############
if (!$match($ynvlan,yes)) then
create log message CreateStandardVLANs
create vlan $vvoice vr $vvoicevr
config vlan $vvoice tag $vvoicetag
config vlan $vvoice add port $eapsprimary tagged
config vlan $vvoice add port $eapssecondary tagged
config eaps $eapsdomain add protected $vvoice
enable lldp ports $netloginports
create qosprofile qp5
config vlan $vvoice ipa 192.168.10.221
#
create vlan $vidsec vr $vidsecvr
config vlan $vidsec tag $vidsectag
config vlan $vidsec add port $eapsprimary tagged
config vlan $vidsec add port $eapssecondary tagged
config eaps $eapsdomain add protected $vidsec
config vlan $vidsec ipa 192.168.40.221
#
create vlan $vdata vr $vdatavr
config vlan $vdata tag $vdatatag
config vlan $vdata add port $eapsprimary tagged
config vlan $vdata add port $eapssecondary tagged
config eaps $eapsdomain add protected $vdata
config vlan $vdata ipa 192.168.11.221
#  config ipr add default 192.168.11.254 vr vr-default
else
create log message NoVLANsCreated
endif
#############
#RADIUS & Netlogin
#############
if (!$match($ynnetlogin,yes)) then
create log message ConfigNetlogin
#configure $vdata ipaddress 192.168.11.221
create vlan nvlan
config netlogin vlan nvlan
config default del po $netloginports
enable netlogin dot1x
enable netlogin mac
enable netlogin ports $netloginports dot1x mac
config netlogin ports $netloginports mode mac-based-vlans
config radius netlogin primary server $radserver client-ip $radclient vr VR-Default
config radius netlogin primary shared-secret $radsecret
enable radius netlogin
```

```
config netlogin add mac-list 00:19:5B:D3:e8:DD
else
create log message NoNetlogin
endif
```

## Sample Configuration with Device-Triggered Profiles

The following example demonstrates how to configure Universal Port for device detection:

```
# Create and configure the VLAN for the VoIP network.
#
switch 1 # create vlan voice
switch 2 # configure voice ipaddress 192.168.0.1/24
# Create the universal port profile for device-detect on the switch.
#
switch 3 # create upm profile detect-voip
Start typing the profile and end with a . as the first and the only character on a line.
Use - edit upm profile <name> - for block mode capability
create log message Starting_Script_DETECT-voip
set var callServer 192.168.10.204
set var fileServer 192.168.10.194
set var voiceVlan voice
set var CleanupProfile CleanPort
set var sendTraps false
#
create log message Starting_DETECT-VOIP_Port_$EVENT.USER_PORT
#*********************************************************
# adds the detected port to the device "unauthenticated" profile port list
#*********************************************************
create log message Updating_UnDetect_Port_List_Port_$EVENT.USER_PORT
configure upm event Device-UnDetect profile CleanupProfile ports $EVENT.USER_PORT
#*********************************************************
# adds the detected port to the proper VoIP vlan
#*********************************************************
configure $voiceVlan add port $EVENT.USER_PORT tag
#*********************************************************
# Configure the LLDP options that the phone needs
#*********************************************************
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme call-server
$callServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme file-server
$fileServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme dot1q-
framing tagged
configure lldp port $EVENT.USER_PORT advertise vendor-specific med capabilities
#configure lldp port $EVENT.USER_PORT advertise vendor-specific med policy application
voice vlan $voiceVlan dscp 46
#*********************************************************
# Configure the POE limits for the port based on the phone requirement
#*********************************************************
# If port is PoE capable, uncomment the following lines
#configure lldp port $EVENT.USER_PORT advertise vendor-specific med power-via-mdi
#configure inline-power operator-limit $EVENT.DEVICE_POWER ports $EVENT.USER_PORT
create log message Script_DETECT-phone_Finished_Port_$EVENT.USER_PORT
.
switch 4 #
# Create the universal port profile for device-undetect on the switch.
#
switch 5 # create upm profile clearports
Start typing the profile and end with a . as the first and the only character on a line.
Use - edit upm profile <name> - for block mode capability
create log message STARTING_UPM_Script_CLEARPORT_on_$EVENT.USER_PORT
```

```
#configure $voiceVlan delete port $EVENT.USER_PORT
unconfigure lldp port $EVENT.USER_PORT
create log message LLDP_Info_Cleared_on_$EVENT.USER_PORT
#unconfigure upm event device-undetect profile avaya-remove ports $EVENT.USER_PORT
unconfigure inline-power operator-limit ports $EVENT.USER_PORT
create log message POE_Settings_Cleared_on_$EVENT.USER_PORT
create log message FINISHED_UPM_Script_CLEARPORT_on_$EVENT.USER_PORT
.
* switch 5 #
#
# Assign the device-detect profile to the edge ports.
#
* switch 6 # config upm event device-detect profile detect-voip ports 1-10
#
# Assign the device-undetect profile to the edge ports.
#
* switch 7 # config upm event device-undetect profile clearports ports 1-10
* switch 8 #
#
# Verify that correct profiles are assigned to correct ports.
#
*  switch 9 # show upm profile
UPM Profile          Events               Flags Ports
=============================================================
clearports           Device-Undetect       e  1-10
detect-voip          Device-Detect         e  1-10
=============================================================
Number of UPM Profiles: 2
Number of UPM Events in Queue for execution: 0
Flags: d - disabled, e - enabled
Event name: log-message(Log filter name) - Truncated to 20 chars
#
# Enable LLDP message advertisements on the ports assigned to universal ports.
#
* switch 10 # enable lldp ports 1-10
```

## Sample Configuration with User-Triggered Profiles

This example demonstrates how to configure a *RADIUS* server and Universal Port for user login. The first part of the example shows the RADIUS server configuration. For more information on RADIUS server configuration, see Security on page 859.

```
# Configure the RADIUS server for the userID and password pair.
# For FreeRADIUS, edit the users file located at /etc/raddb/users as shown in the
# following lines.
#
#Sample entry of using an individual MAC addresses
00040D50CCC3    Auth-Type := EAP, User-Password == "00040D50CCC3"
Extreme-Security-Profile = "phone LOGOFF-PROFILE=clearport;",
Extreme-Netlogin-VLAN = voice
#Sample entry of using wildcard MAC addresses (OUI Method)
00040D000000    Auth-Type := EAP, User-Password == "1234"
Extreme-Security-Profile = "phone LOGOFF-PROFILE=clearport;",
Extreme-Netlogin-VLAN = voice
#Sample entry of using numeric UserID and password
10284    Auth-Type := EAP, User-Password == "1234"
Extreme-Security-Profile = "voip LOGOFF-PROFILE=voip",
Extreme-Netlogin-Vlan = voice
#Sample entry of using a text UserID and password
Sales    Auth-Type := EAP, User-Password == "Money"
Extreme-Security-Profile = "Sales-qos LOGOFF-PROFILE=Sales-qos",
```

```
Extreme-Netlogin-Vlan = v-sales
# Define the Extreme custom VSAs on RADIUS.
# For FreeRADIUS, edit the dictionary file located at //etc/raddb/dictionary to
# include the following details:
VENDOR          Extreme         1916
ATTRIBUTE       Extreme-CLI-Authorization       201     integer Extreme
ATTRIBUTE       Extreme-Shell-Command           202     string  Extreme
ATTRIBUTE       Extreme-Netlogin-Vlan           203     string  Extreme
ATTRIBUTE       Extreme-Netlogin-Url            204     string  Extreme
ATTRIBUTE       Extreme-Netlogin-Url-Desc       205     string  Extreme
ATTRIBUTE       Extreme-Netlogin-Only           206     integer Extreme
ATTRIBUTE       Extreme-User-Location           208     string  Extreme
ATTRIBUTE       Extreme-Netlogin-Vlan-Tag       209     integer Extreme
ATTRIBUTE       Extreme-Netlogin-Extended-Vlan  211     string  Extreme
ATTRIBUTE       Extreme-Security-Profile        212     string  Extreme
ATTRIBUTE       Extreme-CLI-Profile             213     string  Extreme

VALUE   Extreme-CLI-Authorization       Disabled        0
VALUE   Extreme-CLI-Authorization       Enabled         1
VALUE   Extreme-Netlogin-Only           Disabled        0
VALUE   Extreme-Netlogin-Only           Enabled         1
# End of Dictionary
# Add the switch as an authorized client of the RADIUS server.
# For FreeRADIUS, edit the file located at //etc/raddb/clients.conf to include the
# switches as details:
#
client   192.168.10.4 {
secret = purple
shortname = SummitX
# End of clients.conf
```

The rest of this example demonstrates the configuration that takes place at the ExtremeXOS switch:

```
# Create the universal port profile for user-authenticate:
* switch 1 # create upm profile phone
Start typing the profile and end with a . as the first and the only character on a line.
Use - edit upm profile <name> - for block mode capability
create log message Starting_Script_Phone
set var callServer 192.168.10.204
set var fileServer 192.168.10.194
set var voiceVlan voice
set var CleanupProfile CleanPort
set var sendTraps false
#
create log message Starting_AUTH-VOIP_Port_$EVENT.USER_PORT
#*****************************************************
# adds the detected port to the device "unauthenticated" profile port list
#*****************************************************
create log message Updating_Unauthenticated_Port_List_Port_$EVENT.USER_PORT
#*****************************************************
# Configure the LLDP options that the phone needs
#*****************************************************
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme call-server
$callServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme file-server
$fileServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme dot1q-
framing tagged
configure lldp port $EVENT.USER_PORT advertise vendor-specific med capabilities
# If port is PoE capable, uncomment the following lines
#create log message UPM_Script_A-Phone_Finished_Port_$EVENT.USER_PORT
.
switch 2 #
```

```
#
# Create the universal port profile for user-unauthenticate on the switch:
#
switch 1 # create upm profile clearport
Start typing the profile and end with a . as the first and the only character on a line.
Use - edit upm profile <name> - for block mode capability
create log message STARTING_Script_CLEARPORT_on_$EVENT.USER_PORT
unconfigure lldp port $EVENT.USER_PORT
create log message LLDP_Info_Cleared_on_$EVENT.USER_PORT
unconfigure inline-power operator-limit ports $EVENT.USER_PORT
create log message POE_Settings_Cleared_on_$EVENT.USER_PORT
create log message FINISHED_Script_CLEARPORT_on_$EVENT.USER_PORT
.
* switch 2 #
# Configure RADIUS on the edge switch.
#
* switch 4 # config radius primary server 192.168.11.144 client-ip 192.168.10.4 vr "VR-
Default"
* switch 5 # config radius primary shared-secret purple
# Configure Network Login on the edge switch.
#
For Network Login 802.1X, use the following command:
* switch 7 # create vlan nvlan
* switch 8 # config netlogin vlan nvlan
* switch 9 # enable netlogin dot1x
* switch 10 # enable netlogin ports 11-20 mode mac-based-vlans
* switch 11 # enable radius netlogin
#
# For Network Login MAC-based or OUI method, use the following command:
* switch 7 # create vlan nvlan
* switch 8 # config netlogin vlan nvlan
* switch 9 # enable netlogin mac
* switch 10 # config netlogin add mac-list 00:04:0D:00:00:00 24 1234
* switch 11 # enable radius netlogin
# Assign the user-authenticate profile to the edge port.
#
* switch 12 # configure upm event user-authenticate profile "phone" ports 11-20
* switch 13 #
# Assign the user-unauthenticate profile to the edge port.
#
* switch 14 # configure upm event user-unauthenticated profile "clearport" ports 11-20
* switch 15 #
# Check that the correct profiles are assigned to the correct ports.
#
* switch 16 # show upm profile
============================================================
UPM Profile          Events               Flags Ports
============================================================
phone                User-Authenticated      e  11-20
clearport          User-Unauthenticated      e  11-20
============================================================
Number of UPM Profiles: 5
Number of UPM Events in Queue for execution: 0
Flags: d - disabled, e - enabled
Event name: log-message(Log filter name) - Truncated to 20 chars
# Enable LLDP message advertisements on the ports.
#
* switch 17 # enable lldp ports 11-20
```

## Sample Timer-Triggered Profile

The following profile and timer configuration disables *PoE* on ports 1 to 20 everyday at 6:00 p.m.:

```
* switch 1 # create upm profile eveningpoe
Start typing the profile and end with a . as the first and the only character on a line.
Use - edit upm profile <name> - for block mode capability
create log message Starting_Evening
disable inline-power ports 1-20
.
*switch 2
*switch 3 # create upm timer night
*switch 4 # config upm timer night profile eveningpoe
*switch 5 # config upm timer night at 7 7 2007 19 00 00 every 86400
```

## Sample Profile with QoS Support

The example below can be used with a Summit family switch that supports *QoS* profiles qp1 and qp8.

When the user or phone logs in with a particular MAC address, the script configures the QoS profile configured by the user in the *RADIUS* server for the USER-AUTHENTICATED event. In this example, the user sets the QoS profile to be qp8.

You must configure network login, the RADIUS server, and Universal Port on the switch as part of the user log-in authentication process. For more information on configuring the RADIUS users file, see Security on page 859.

The following example is an entry in the RADIUS users file for the MAC address of the phone:

```
00040D9D12A9 Auth-Type := local, User-Password == "test"
Extreme-security-profile = "p1 QOS=\"QP8\";LOGOFF-PROFILE=p2;VLAN=\"voice-        test\";"
```

Below is the Universal Port profile configuration for this example:

```
Create upm profile p1
set var z1 $uppercase($EVENT.USER_MAC)
set var z2 $uppercase(00:04:0d:9d:12:a9)
#show var z1
#show var z2
if ($match($EVENT.NAME, USER-AUTHENTICATED) == 0) then
    if ($match($z1, $z2) == 0) then
        configure port $EVENT.USER_PORT qosprofile $QOS
    endif
endif
.
```

## Sample Event Profile

If not configured properly, the *STP (Spanning Tree Protocol)* can create loops in a network.

Should these loops develop, they can cause network degradation and eventually crash the network by duplicating too many Ethernet frames. By leveraging Universal Port and the Extreme Loop Recovery

Protocol (ELRP) as shown in example below, it is not only possible to detect and isolate the egress port, but it is also possible to disable the egress port to break loops.

Note

This example illustrates how to create an event profile that reconfigures the switch after an event. After this example was created, ELRP was updated with the capability to disable a port without the help of an event profile. For more information, see Using ELRP to Perform Loop Tests on page 1570.

When a loop is detected on ports where ELRP is enabled and configured, ELRP logs a message using the following format:

```
01/17/2008 08:08:04.46 <Warn:ELRP.Report.Message> [CLI:ksu:1] LOOP DETECTED : 436309
transmited, 64 received, ingress slot:port (1) egress slot:port (24)
```

To view more information on format of this command, enter the show log events command as shown in the following example:

```
* BD8810-Rack2.6 # show log events "ELRP.Report.Message" details
Comp    SubComp     Condition               Severity      Parameters
------- ----------- ----------------------- ------------- ----------
ELRP    Report      Message                 Warning        8 total
0 - string
1 - string
2 - number (32-bit
unsigned int)
3 - string
4 - number (32-bit
unsigned int)
5 - number (unsigned int)
6 - string
7 - string
"[%0%:%1%:%2%] %3% : %4% transmited, %5% received, ingress slot:port (%6%) egress
slot:port (%7%)"
```

In the example log statement, the *VLAN* ksu, the ports is all, and the interval is “1.”

If a loop is detected, we want to disable the egress port on which the ELRP control packets went out. In this example, we enable ELRP on all ports of a VLAN as follows:

```
configure elrp-client periodic vlan ports all interval 1
```

We want the profile to disable egress ports 1 and 24 (which have been configured for loop). If we enable ELRP on only one port, then the port alone would be disabled.

We observe that parameter 7 is the one we have to disable from the above log message and the details for that event.

*Configuring Universal Port Example*

The following procedure configures Universal Port to disable the egress port identified by parameter 7:

1. Create the profile and add the command to disable the egress port as follows:

   ```
   create upm profile disable_port_elrp

   disable port $EVENT.LOG_PARAM_7
   ```

2. Verify that the profile is created and enabled by entering the following command:

   ```
   show upm profile
   ```

3. Create the *EMS* configuration by entering the following commands:

   ```
   create log target upm disable_port_elrp

   create log filter f1

   configure log filter f1 add event ELRP.Report.Message match string
   "LOOP"

   enable log target upm "disable_port_elrp"

   configure log target upm "disable_port_elrp" filter f1
   ```

4. At this point, connect the ports 1 and 24 to form a loop.

   Two log messages will be logged when the loop is detected on ports 1 and 24 and ELRP is enabled both. This triggers the disable_port_elrp profile twice, and ports 1 and 24 should be disabled.

5. View the log.

   ```
   > show log
   01/17/2008 08:08:05.49 <Info:vlan.dbg.info> Port 24 link down
   01/17/2008 08:08:05.22 <Noti:UPM.Msg.upmMsgExshLaunch> Launched profile
   disable_port_elrp for the event log-message
   01/17/2008 08:08:04.69 <Info:vlan.dbg.info> Port 1 link down
   01/17/2008 08:08:04.46 <Noti:UPM.Msg.upmMsgExshLaunch> Launched profile
   disable_port_elrp for the event log-message
   01/17/2008 08:08:04.46 <Warn:ELRP.Report.Message> [CLI:ksu:1] LOOP DETECTED : 436309
   transmited, 64 received, ingress slot:port (1) egress slot:port (24)
   01/17/2008 08:08:04.46 <Warn:ELRP.Report.Message> [CLI:ksu:1] LOOP DETECTED : 436309
   transmited, 63 received, ingress slot:port (24) egress slot:port (1)
   01/17/2008 08:08:03.50 <Info:vlan.dbg.info> Port 24 link up at 1 Gbps speed and full-
   duplex
   ```

6. To view the profile execution history, enter the show upm history command.

   If you want to see the more details, enter the show upm history details command to see all the profiles or display information on a specific event by entering the exec-id.

7. To view the configuration, use the `show config upm` and `show config ems` commands.

## Sample Configuration for Generic VoIP LLDP

```
#*******************************
# Last Updated: March 20, 2007
# Tested Phones: Avaya 4610, 4620, 4625
# Requirements: LLDP capable devices
#*******************************
# @META_DATA_START
# @FileDescription "This is a template for configuring network parameters for VoIP phones
support LLDP but without 802.1X authentication.  The module is triggered through the
```

```
detection of an LLDP packet on the port.   The following network side configuration is
done: enable SNMP traps, QOS assignment, adjust POE reservation values based on device
requirements, add the voiceVlan to the port as tagged."
# @Description "Voice VLAN name"
set var voicevlan voice
# @Description "Send trap when LLDP event happens (true or false)"
set var sendTraps false
# @Description "Set QoS Profile (true or false)"
set var setQuality false
# @META_DATA_END
#
if (!$match($EVENT.NAME,DEVICE-DETECT)) then
create log message Starting_LLDP_Generic_Module_Config
# VoiceVLAN configuration
configure vlan $voicevlan add port $EVENT.USER_PORT tagged
#SNMP Trap
if (!$match($sendTraps,true)) then
create log message Config_SNMP_Traps
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
#Link Layer Discovery Protocol-Media Endpoint Discover
create log message Config_LLDP
configure lldp port $EVENT.USER_PORT advertise vendor-specific med capabilities
configure lldp port $EVENT.USER_PORT advertise vendor-specific dot1 vlan-name vlan
$voicevlan
configure lldp port $EVENT.USER_PORT advertise vendor-specific med policy application
voice vlan $voicevlan dscp 46
configure lldp port $EVENT.USER_PORT advertise vendor-specific med power-via-mdi
#Configure POE settings per device requirements
create log message Config_POE
configure inline-power operator-limit $EVENT.DEVICE_POWER ports $EVENT.USER_PORT
#QoS Profile
if (!$match($setQuality,true)) then
create log message Config_QOS
configure port $EVENT.USER_PORT qosprofile qp7
endif
endif
if (!$match($EVENT.NAME,DEVICE-UNDETECT) && $match($EVENT.DEVICE_IP,0.0.0.0)) then
create log message Starting_LLDP_Generic_UNATUH_Module_Config
if (!$match($sendTraps,true)) then
create log message UNConfig_SNMP_Traps
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
create log message UNConfig_LLDP
unconfig lldp port $EVENT.USER_PORT
if (!$match($setQuality,true)) then
create log message UNConfig_QOS
unconfig qosprofile ports $EVENT.USER_PORT
endif
unconfig inline-power operator-limit ports $EVENT.USER_PORT
endif
if (!$match($EVENT.NAME,DEVICE-UNDETECT) && !$match($EVENT.DEVICE_IP,0.0.0.0)) then
create log message DoNothing_0.0.0.0
create log message $EVENT.TIME
endif
create log message End_LLDP_Generic_Module_Config
```

## Sample Configuration for Generic VoIP 802.1X

```
#*********************************************
# Last Updated: April 6, 2007
# Tested Phones: Avaya 4610, 4620, 4625
# Requirements: 802.1X capable devices, netlogin configured and enabled on deployment
ports
#*********************************************
# @META_DATA_START
# @FileDescription "This is a template for configuring network parameters for 802.1X
authenticated devices.  The module is triggered through successful authentication of the
device.  The following network side configuration is done: QOS assignment and enables DOS
protection.  When used with IP phones, phone provisioning is done through DHCP options."
# @Description "VLAN name to add to port"
set var vlan1 voice
# @Description "Set QoS Profile (yes or no)"
set var setQuality yes
# @Description "QoS Profile (0-100)"
set var lowbw 50
# @Description "QoS MAX Bandwidth (0-100)"
set var highbw 100
# @Description "Enable Denial of Service Protection (yes or no)"
set var dosprotection yes
# @META_DATA_END
##################################
# Start of USER-AUTHENTICATE block
##################################
if (!$match($EVENT.NAME,USER-AUTHENTICATED)) then
#############
#QoS Profile
#############
# Adds a QOS profile to the port
if (!$match($setQuality,yes)) then
create log message Config_QOS
configure port $EVENT.USER_PORT qosprofile qp7
configure qosprofile qp7 minbw $lowbw maxbw $highbw ports $EVENT.USER_PORT
endif
#
########################
#Security Configurations
########################
create log message Applying_Security_Limits
# enables Denial of Service Protection for the port
if (!$match($dosprotection,yes)) then
enable dos-protect
create log message DOS_enabled
endif
#
endif
################################
# End of USER-AUTHENTICATE block
```

## Sample Configuration for Avaya VoIP 802.1X

```
#*******************************
# Last Updated: March 20, 2007
# Tested Phones: SW4610, SW4620
# Requirements: 802.1X authentication server, VSA 203 and VSA 212 from authentiication
server. QP7 defined on the switch#
*******************************
# @META_DATA_START
```

```
# @FileDescription "This is a template for configuring LLDP capable Avaya phones using
the authentication trigger.  This module will provision the phone with the following
parameters: call server, file server,  dot1q, dscp, power.  Additionally the following
network side configuration is done: enable SNMP traps and QOS assignment."
# @Description "Avaya phone call server IP address"
set var callserver 192.45.95.100
# @Description "Avaya phone file server IP address"
set var fileserver 192.45.10.250
# @Description "Send trap when LLDP event happens (true or false)"
set var sendTraps true
# @Description "Set QoS Profile (true or false)"
set var setQuality true
# @META_DATA_END
#
if (!$match($EVENT.NAME,USER-AUTHENTICATED)) then
create log message Starting_Avaya_VOIP_802.1X_AUTH_Module_Config
if (!$match($sendTraps,true)) then
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
enable lldp port $EVENT.USER_PORT
configure lldp port $EVENT.USER_PORT advertise vendor-specific dot1 vlan-name
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme call-server
$callserver
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme file-server
$fileserver
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme dot1q-
framing tag
if (!$match($setQuality,true)) then
configure port $EVENT.USER_PORT qosprofile qp7
endif
endif
#
if (!$match($EVENT.NAME,USER-UNAUTHENTICATED)) then
create log message Starting_Avaya_VOIP_802.1X_UNATUH_Module_Config
if (!$match($sendTraps,true)) then
enable snmp traps lldp ports $EVENT.USER_PORT
enable snmp traps lldp-med ports $EVENT.USER_PORT
else
disable snmp traps lldp ports $EVENT.USER_PORT
disable snmp traps lldp-med ports $EVENT.USER_PORT
endif
disable lldp port $EVENT.USER_PORT
if (!$match($setQuality,true)) then
unconfig qosprofile ports $EVENT.USER_PORT
endif
endif
create log message End_Avaya_VOIP_802.1X_Module_Config
Dynamic Security Policy
if (!$match($CLI_EVENT,USER-AUTHENTICATED) ) then
create access-list $(DEVICE_MAC)_192_168_1_0 "ethernet-source-address $DEVICE_MAC ;
destination-address 192.168.1.0/24 " "permit "
create access-list $(DEVICE_MAC)_192_168_2_0 "ethernet-source-address $DEVICE_MAC ;
destination-address 192.168.2.0/24 " "permit "
create access-list $(DEVICE_MAC)_192_168_3_0 "ethernet-source-address $DEVICE_MAC ;
destination-address 192.168.3.0/24 " "permit "
create access-list $(DEVICE_MAC)_smtp "ethernet-source-address $DEVICE_MAC ;
destination-address 192.168.100.125/32 ; protocol tcp ; destination-port 25" "permit "
create access-list $(DEVICE_MAC)_http "ethernet-source-address $DEVICE_MAC ; protocol
tcp ; destination-port 80" "permit "
create access-list $(DEVICE_MAC)_https "ethernet-source-address $DEVICE_MAC ; protocol
```

```
tcp ; destination-port 443" "permit "
create access-list $(DEVICE_MAC)_dhcp "protocol udp; destination-port 67" "permit"
create access-list $(DEVICE_MAC)_deny "destination-address 0.0.0.0/0" "deny "
configure access-list add $(DEVICE_MAC)_192_168_1_0 first port $USER_PORT
configure access-list add $(DEVICE_MAC)_192_168_2_0 first port $USER_PORT
configure access-list add $(DEVICE_MAC)_192_168_3_0 first port $USER_PORT
configure access-list add $(DEVICE_MAC)_smtp first port $USER_PORT
configure access-list add $(DEVICE_MAC)_http last port $USER_PORT
configure access-list add $(DEVICE_MAC)_https last port $USER_PORT
configure access-list add $(DEVICE_MAC)_dhcp first port $USER_PORT
configure access-list add $(DEVICE_MAC)_deny last port $USER_PORT
endif
if (!$match($CLI_EVENT,USER-UNAUTHENTICATED) ) then
# Clean up
configure access-list delete $(DEVICE_MAC)_192_168_1_0 ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_192_168_2_0 ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_192_168_3_0 ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_smtp ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_http ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_https ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_dhcp ports $USER_PORT
configure access-list delete $(DEVICE_MAC)_deny ports $USER_PORT
delete access-list $(DEVICE_MAC)_192_168_1_0
delete access-list $(DEVICE_MAC)_192_168_2_0
delete access-list $(DEVICE_MAC)_192_168_3_0
delete access-list $(DEVICE_MAC)_smtp
delete access-list $(DEVICE_MAC)_http
delete access-list $(DEVICE_MAC)_https
delete access-list $(DEVICE_MAC)_dhcp
delete access-list $(DEVICE_MAC)_deny
endif
```

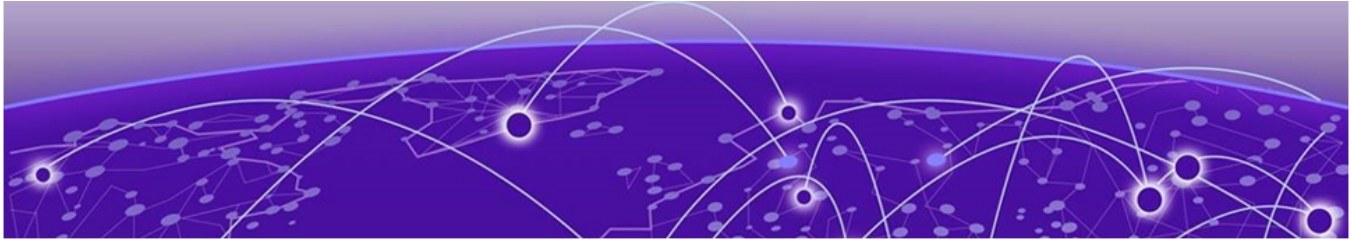
## Sample Configuration for a Video Camera

This template adds an *ACL* to an edge port when a video camera connects.

The profile configures and applies an ACL onto a switch port when a user authenticates. This ACL blocks a particular IP address from accessing the video camera and assigns the user to *QoS* profile 7.

```
#*********************************************
# Last Updated: March 9, 2007
# Tested Devices: Dlink DCS 1110
# Requirements: netlogin configured and enabled on deployment ports
#*********************************************
# @MetaDataStart
# @ScriptDescription "This is a template for configuring the switch for the right
environment for this webcam. It creates a dynamic access-list to restrict access"
# @Description "VLAN name to add to port"
# set var vlan1 voiceavaya
# @VariableFieldLabel "Set QoS Profile (yes or no)"
# set var setQuality yes
# @Description "QoS Profile (0-100)"
# set var lowbw 50
# @VariableFieldLabel "QoS MAX Bandwidth (0-100)"
# set var highbw 100
# @MetaDataEnd
################################
# Start of USER-AUTHENTICATE block
################################
if (!$match($EVENT.NAME,USER-AUTHENTICATED)) then
############
```

```
#QoS Profile
############
# Adds a QOS profile to the port
#   if (!$match($setQuality,yes)) then
#    create log message Config_QOS
#    configure port $EVENT.USER_PORT qosprofile qp7
#    configure qosprofile qp7 minbw $lowbw maxbw $highbw ports $EVENT.USER_PORT
#  endif
#
############
#ACL Section
############
# Adds an ACL to stop traffic to a particular address
create log message Config_ACL
create access-list webcamblock "destination-address 192.168.10.220/32" "deny"
configure access-list add webcamblock first port $EVENT.USER_PORT
#endif
#
endif
################################
# End of USER-AUTHENTICATE block
################################
#
#
####################################
# Start of USER-UNAUTHENTICATE block
####################################
if (!$match($EVENT.NAME,USER-UNAUTHENTICATED)) then
#   create log message Starting_8021x_Generic_UNATUH_Module_Config
#   if (!$match($setQuality,yes)) then
#     create log message UNConfig_QOS
#     unconfig qosprofile ports $EVENT.USER_PORT
#   endif
#   unconfigure inline-power operator-limit ports $EVENT.USER_PORT
#### remove acl
configure access-list delete webcamblock port $EVENT.USER_PORT
delete access-list webcamblock
endif
################################
# End of USER-UNAUTHENTICATE block
################################
create log message End_802_1x_Generic_Module_Config
```



# Using CLI Scripting



CLI-based scripting allows you to create a list of commands that you can execute manually with a single command or automatically when a special event occurs.

CLI-based scripting supports variables and functions, so that you can write scripts that operate unmodified on multiple switches and in different environments. It allows you to significantly automate switch management.

Note
Special scripts can be used to configure the switch when it boots. For more information, see Using Autoconfigure and Autoexecute Files on page 1547.

## Setting Up Scripts

### Enabling and Disabling CLI Scripting

The **permanent** option enables CLI scripting for new sessions only and makes the configuration change part of the permanent switch configuration so that the scripting configuration remains the same when the switch reboots. When the command is used without the permanent option, it enables CLI scripting for the current session only. If you do not include the permanent option, CLI scripting is disabled the next time the switch boots.

- To support scripting, including the testing of script-related commands, enable scripting using the following command. CLI scripting is disabled by default.

  ```
  enable cli scripting {permanent}
  ```

  Note
  CLI scripting cannot be enabled when CLI space auto completion is enabled with the enable cli space-completion command.

- To disable scripting, use the following command:

  ```
  disable cli scripting {permanent}
  ```

## Creating Scripts

There are two ways to create scripts.The method you choose depends on how you want to execute the script.

If you want to create a script file to configure a switch or a switch feature, and you plan to execute that script manually, you can create a script file.

If you want to create a script that is activated automatically when a device or user connects to or disconnects from a switch port, you should create the script with the Universal Port feature. The following sections provide more information on these options.

### *Creating a Script File*

A script file is an ASCII text file that you can create with any ASCII text editor program. The text file can contain CLI commands and can use the scripting options described in the following sections:

- Using Script Variables on page 358
- Using Special Characters in Scripts on page 360
- Using Operators on page 360
- Using Control Structures in Scripts on page 362
- Using Built-In Functions on page 362

You can move an ASCII script file to the switch using the same techniques described for managing ASCII configuration files in Software Upgrade and Boot Options on page 1522.

### *Creating Scripts for Use with the Universal Port Feature*

The Universal Port feature allows you to create dynamic profiles that are activated by a trigger event, such as a user or device connection to a switch port.

These dynamic profiles contain script commands and cause dynamic changes to the switch configuration to enforce a policy. The universal port profiles support all the scripting options listed in Creating Scripts for Use with the Universal Port Feature on page 355 for creating script files. For more information on entering script commands in a universal port profile, see Universal Port on page 309.

### *Python Scripting*

Introduced in ExtremeXOS 15.6, Python scripting provides the ability to customize your switch and add functionality quickly by downloading and running scripts without the need for engineering expertise. Python scripting is extended using the synonymous `load script` and `run script` commands.

#### Expect Functionality

It may be necessary to interact with ExtremeXOS using Expect scripting functionality. The Python community offers a pexpect.py module that can provide this capability. EXOS uses pexpect.py version 3.2. Documentation can be found here: https://pypi.python.org/pypi/pexpect/

The pexpect.py provides interfaces to run interactive shell commands and to process the results through expect like functions. The challenge with exsh.clicmd() is that it is a synchronous call and not a separate process. The exshexpect.py module was developed to provide a wrapper for pexpect.py that interfaces to exsh.clicmd().

Below is an example of using pexect together with the exshexpect module.

```
import pexpect

import exshexpect
```

Create a prompt for expect that will not match any command output.

```
exosPrompt = '::--:--::'
```

Create an expect object. Pass in the prompt and the back end function to call:

```
p = exshexpect.exshspawn(exosPrompt, exsh.clicmd)
```

Use sendline to send commands to the backend function (in this case exsh.clicmd) :

```
p.sendline('show fan') <-

idx = p.expect([exosPrompt, pexpect.EOF, pexpect.TIMEOUT])

print 'idx=',idx

print 'before->',p.before

print 'after->',p.after
```

Special case for passing in a command with an additional argument described in exsh.clicmd():

```
p.send('enable debug-mode\nC211-6BB4-E6F5-B579\n')
```

In the line above, use send() and include 'newline' terminations for the command and the additional parameter.

This is the same as calling:

```
exsh.clicmd('enable debug-mode', True, args='C211-6BB4-E6F5-B579')
```

but using expect to process any response.

```
idx = p.expect([exosPrompt, pexpect.EOF, pexpect.TIMEOUT])
```

#### Creating Sockets Using Python Scripts

Creating a socket in a Python script defaults to the management interface. You can change the VR when the socket is create by using the following example:

Example: changing the VR to *VR-Default* (VRID = 2) for an UDP socket:

```
import socket
EXTREME_SO_VRID = 37

udp_sock = socket.socket(socket.AF_INET, socket.SOCK_DGRAM)
try:
    f = open('/proc/self/ns_id', 'w')
    f.write('2\n')
    f.close()
except:
    udp_sock.setsockopt(socket.SOL_SOCKET, EXTREME_SO_VRID, 2)
```

```
# your program here
```

**Note**

Do not use os.system('echo 2 > /proc/self/ns_id') to do this because os.system() spawns a subprocess. The subprocess will switch to vr-default, and then exit. The calling process will still be using the same VR as before.

Finding the VR ID of any VR can be accomplished with the following example:

```
def vrNameToNumber(vrName):
    vrId = None
    with open('/proc/net/vr','r') as f:
      for l in f.readlines():
            rslt = l.split()
            if rslt[3] == vrName:
               vrId =  int(rslt[0]
               break
        return vrId
```

### Python Scripting Examples

Example 1

This is a single line script that illustrates the `exsh.clicmd ()` called from a Python script.

```
print exsh.clicmd('show switch', True)
```

In this example, the second parameter passed to `exsh.clicmd ()` is True. This returns the CLI display output that would have normally been sent to the user. In this case, the line prints the output anyway, which is shown below.

```
SysName:          X670V-48x
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:6D:12:A7
System Type:      X670V-48x

SysHealth check:  Enabled (Normal)
Recovery Mode:    All
System Watchdog:  Enabled

Current Time:     Tue Dec 17 01:49:10 2013
Timezone:         [Auto DST Disabled] GMT Offset: 0 minutes, name is UTC.
Boot Time:        Mon Dec 16 00:28:13 2013
Boot Count:       1559
Next Reboot:      None scheduled
System UpTime:    1 day 1 hour 20 minutes 57 seconds

Current State:    OPERATIONAL
Image Selected:   secondary
Image Booted:     secondary
Primary ver:      16.1.0.3
Secondary ver:    16.1.0.3
Config Selected:  primary.cfg
Config Booted:    primary.cfg
primary.cfg       Created by ExtremeXOS version 16.1.0.3
                  255914 bytes saved on Sun Dec  8 00:04:20 2013
```

Example 2

In this example, we create a number of *VLAN (Virtual LAN)*s in the form of *prefix tag* and add tagged port lists to each VLAN. There are a number of methods for collecting user input and validating the data type. This example uses a simple `while True:` loop until the input is the correct type. A more robust example would also apply range checking etc... to the vlan tag(s). E.g. Create vlans with tags from 1000 to 1099 The prefix is 'dave', so each vlan name will have a name like 'dave1000', 'dave1001' etc... The user provided port list will be added to each vlan name as tagged ports.

```
import sys
while True:
    prefix = raw_input('enter vlan name prefix:')
    if len(prefix):
        break

while True:
    reply = raw_input('beginning vlan tag:')
    if reply.isdigit():
        minTag = int(reply)
        break

while True:
   reply = raw_input('ending vlan tag:')
   if reply.isdigit():
       maxTag = int(reply)
       break

if minTag > maxTag:
    print 'Beginning vlan id cannot be greater than ending vlan id'
    sys.exit(-1)

while True:
    portlist = raw_input('port list:')
    if len(portlist):
        break

print 'Creating vlans from ',prefix+str(minTag),'to',prefix+str(maxTag)

for vlanId in range(minTag,maxTag + 1):
    vlanName = prefix + '{0:>04d}'.format(vlanId)
    cmd = 'create vlan '+vlanName + ' tag ' + str(vlanId)
    print exsh.clicmd(cmd, True)
    cmd = 'config vlan ' + vlanName + ' add port ' + portlist + ' tag'
    print exsh.clicmd(cmd, True)
```

**Note**

UPM variables can be used with Python scripting. They must be passed as arguments of the script.

## Using Script Variables

Variable names must be followed by white space or otherwise enclosed in parentheses.

For example: create vlan v($X)e where X is the variable. The variable created will persist through the session and will not get reset after disable/enable cli scripting.

The following table shows the predefined system variables that are always available for use by any script.

Predefined variables are automatically set up for use before a script or profile is executed.

Note

You must enable CLI scripting before using these variables or executing a script.

**Table 38: Predefined System Variables**

| Variable Syntax | Definition |
|---|---|
| $STATUS | Status of last command execution. Values -100 to 100 are reserved and automatically set by the software, but you can override the value with the command `return statusCode` . Common values are:<br>0—Successful command completion-53—Variable not found-57—WHILE depth exceeded-78—Script timeout |
| $CLI.USER | User who is executing this CLI. |
| $CLI.SESSION_TYPE | User session type. |
| $CLI.SCRIPT_TIME_REMAINING | When a script timeout value is configured with the `configure cli script timeout timeout` command, the system creates this variable, which returns the following values (in seconds):<br>If no script is running, this variable returns the configured timeout value.<br>If a script is aborted due to timeout, this variable returns the value 0.<br>If a script finishes execution (before the timeout value is reached), this variable returns the remaining time. |
| $CLI.SCRIPT_TIMEOUT | When a script timeout value is configured with the `configure cli script timeout timeout` command, the system creates this variable, which returns the current timeout value. If no script is running, this variable returns the configured timeout value. |

The following table shows the system variables that you must define before use.

**Table 39: System Variables that Must Be Created**

| Variable Syntax | Definition |
|---|---|
| $CLI.OUT | Output of last show command. The maximum size of this variable is 1 MB. This output can be used for operations such as match and regexp. For more information on these operations, see Using Built-In Functions on page 362.<br>You must define this variable before it is used. To define this variable, enter either of the following statements:<br>`set var CLI.OUT " "`<br>`set var CLI.OUT 0`<br>We recommend that you delete this variable after each use. |

### *Creating Variables*

You can create your own variables and change their values. When using variables, the following guidelines apply:

- Variable names are case insensitive and are limited to 32 characters.
- The variable name must be unique.
- A variable can be referenced as $X or $(X).
- If a variable already exists, it is overwritten. No error message is displayed.

- The variable expression can be a constant value, another variable, or a combination of the above with operators and functions. Operators are described in Using Operators on page 360, and functions are described in Using Built-In Functions on page 362.
- Only the `set var` command supports expression evaluation.
- If the variable name X contains special characters such as +-/*, then the variable needs to be enclosed in parentheses. For example: set var z ($(x) + 100).
- When you use a variable with a TCL special character in the $TCL function, the variable must be enclosed in braces. For example: set var x $TCL(string length ${CLI.USER}.
- For more information on TCL functions, see Using Built-In Functions on page 362.

**Note**
EXOS does not consider the dot/period character as a de-limiter. This is different from standard TCL behavior, where dot/period is considered as a de-limiter.

- To create a variable and set a value for it, or to change a variable value, use the following command:

  `set var varname _expression`

  The following examples show various ways to define or change variables:

  ```
  set var x 100

  set var x ($x + 2)

  set var y ($x - 100)
  ```

- To display all variables or a specified variables, use the following command:

  `show var {varname}`

## Using Special Characters in Scripts

The dollar sign ($) character and quote (") characters have special purposes in scripts.

The ($) indicates a variable, and the (") surrounds text strings. To use these characters as regular characters, precede the special character with a backslash character (\). For example:

```
set var variablename \$<varname>
set var CLI.USER "Robert \"Bob\" Smith"
```

Scripts also support quote characters within quotes.

## Using Operators

Operators allow you to manipulate variables and evaluate strings.

Some operators can be used in all numeric expressions, while others are restricted to integer or string expressions. The following table lists the operators supported and provides comments on when they can be used. The valid operators are listed in decreasing order of precedence.

**Table 40: Operators**

| Operator | Action | Comments |
|---|---|---|
| - | Unary minus | None of these operands can be applied to string operands, and the bit-wise NOT operand can be applied only to integers. |
| + | Unary plus | |
| ~ | Bit-wise NOT | |
| ! | Logical NOT | |
| * | Multiply | None of these operands can be applied to string operands, and the remainder operand can be applied only to integers. The remainder always has the same sign as the divisor and an absolute value smaller than the divisor. |
| / | Divide | |
| % | Remainder | |
| + | Add | These operands are valid for any numeric operations. |
| - | Subtract | |
| << | Left shift | These operands are valid for integer operands only. A right shift always propagates the sign bit. |
| >> | Right shift | |
| < | Boolean less | Each operator produces 1 if the condition is true, 0 otherwise. These operators can be applied to strings as well as numeric operands, in which case string comparison is used. |
| > | Boolean greater | |
| <= | Boolean less than or equal | |
| >= | Boolean greater than or equal | |
| == | Boolean equal | Each operator produces a zero or one result. These operators are valid for all operand types. |
| != | Boolean not equal | |
| & | Bit-wise AND | This operator is valid for integer operands only. |
| ^ | Bit-wise exclusive OR | This operator is valid for integer operands only. |
| \| | Bit-wise OR | This operator is valid for integer operands only. |
| && | Logical AND | This operator produces a result of 1 if both operands are non-zero. Otherwise, it produces a result of 0. This operator is valid for numeric operands only (integers or floating-point). |
| \|\| | Logical OR | This operator produces a result of 0 if both operands are zero. Otherwise, it produces a result of 1.This operator is valid for numeric operands only (integers or floating-point). |
| x?y:z | If-then-else (as in the C programming language) | If x evaluates to non-zero, then the result is the value of y. Otherwise the result is the value of z. The x operand must have a numeric value. |

## Using Control Structures in Scripts

The CLI supports the control structures described in the following sections.

### *Conditional Execution*

```
IF (<expression>) THEN
<statements>
ELSE
<statements>
ENDIF
```

The expression must be enclosed in parentheses.

### *Loop While Condition is TRUE*

```
WHILE (<expression>) DO
    <statements>
ENDWHILE
```

The expression must be enclosed in parentheses.

Nesting is supported up to five levels. The Ctrl-C key combination can be used to break out of any While loop(s).

The operators mentioned in Using Operators on page 360 can be used in an expression in the set var command or in an IF or WHILE condition.

If there is incorrect nesting of an IF condition or WHILE loop, an error message appears. If a user tries to type more than five WHILE loops or five IF conditions, an error message appears. Breaking out of any number of WHILE loops always clears the WHILE condition.

Comments can be inserted by using the number sign (#).

## Using Built-In Functions

Built in functions allow you to manipulate and evaluate the variables inside your script and the script output. The following table shows the built-in functions. The following table shows the built-in Tool Command Language (TCL) functions.

**Note**
ExtremeXOS uses TCL version 8.5.14.

**Table 41: Built-In Functions**

| Syntax | Function |
|---|---|
| $MATCH(string 1, string 2) | Compares the two strings string 1 and string 2. Returns 0 if string1 matches string2.<br>It returns -1,0, or 1, depending on whether string1 is less than, equal to, or greater than string2. |
| $READ(prompt) | Displays a prompt for user input and accepts input until the user presses [Return] or the session times out. Replace prompt with the prompt to display to the user. |
| $TCL(function args) | Calls a TCL built-in function (see the following table). Note that the software does not support the simultaneous operation of multiple TCL functions. For more information on TCL functions, go to http://www.tcl.tk/man/tcl8.3/TclCmd/contents.htm. |
| $UPPERCASE(string) | Returns the string uppercased. |
| $VAREXISTS(varname) | Returns zero if the specified variable does not exist. Returns non-zero if the specified variable does exist. |

**Table 42: Supported TCL Functions**

| Function | Function Type | Description |
|---|---|---|
| after | System related | Execute a command after a time delay. |
| binary | String handling | Insert and extract fields from binary strings. |
| clock | System related | Obtain and manipulate time. |
| concat | List handling | Join lists together. |
| expr | Math | Evaluate an expression. |
| join | List handling | Create a string by joining list elements together. |
| lindex | List handling | Retrieve an element from a list. |
| linsert | List handling | Insert elements into a list. |
| list | List handling | Create a list. |
| llength | List handling | Count the number of elements in a list. |
| lrange | List handling | Return one or more adjacent elements from a list. |
| lreplace | List handling | Replace elements in a list with new elements. |
| lsearch | List handling | See if a list contains a particular element. |
| lsort | List handling | Sort the elements of a list. |
| regexp | String handling | Match a regular expression against a string. |
| regsub | String handling | Perform substitutions based on regular expression pattern matching. |
| re_syntax | String handling | Syntax of TCL regular expressions. |

**Table 42: Supported TCL Functions (continued)**

| Function | Function Type | Description |
|---|---|---|
| split | List handling | Split a string into a proper TCL list. |
| string | String handling | Manipulate strings. |

For examples of scripts that use TCL functions, see CLI Scripting Examples on page 367.

## Control Script Configuration Persistence

When a script runs, the commands within the script can make persistent changes to the switch configuration, which are saved across reboots, or it can make non-persistent changes. Non-persistent configuration changes remain part of the switch configuration only until the switch reboots.

The default setting for scripts is non-persistent. To change the script configuration persistence setting, use the following command:

```
configure cli mode [persistent | non-persistent]
```

## Saving, Retrieving, and Deleting Session Variables

Session variables are the set of variables that are active for a particular session. For example, if a device is detected on a universal port and this triggers a profile (and the script commands within it), the variable values that were active when the profile started are replaced with the variable values defined in the profile. The first session is the device-undetected session, and the second session is the device-detected session. Each session has its own set of variables and values.

The software allows you to save session variables before replacing them. In the example above, this allows you to retrieve the earlier values when the port returns to the device-undetected state. Up to five variables can be saved or retrieved at a time. These variables are saved to system memory using a key, which is an ID for the saved variables. You are responsible for generating unique keys. When you want to retrieve the variables, you use the key to identify the variables to be retrieved.

- To save up to five session variables, use the following command:

  ```
  save var key key [var1 var2 …]
  ```

- To retrieve saved session variables, use the following command:

  ```
  load var key key [var1 var2 …]
  ```

- To delete saved session variables, use the following command:

  ```
  delete var key key
  ```

The variables saved by the save var command are saved using the specified key and are retrievable and restored in the context that this profile was applied. They are available to rollback events like user-unauthenticate and device-undetect.

Note

For modular switches and SummitStack, session variables are saved on the master MSM or master Summit switch. To repopulate session variables after a failover event, manually execute the appropriate script.

## Nesting Scripts

The ExtremeXOS software supports three levels of nested scripts. An error appears if you attempt to start a fourth-level script. The following example demonstrates nested scripts:

```
Load script x1
# Script x1 contents:
Cmd 1
Cmd 2
Load script x2
Cmd 3
Cmd 4
# Script x2 contents:
Cmd 1
Cmd 2
Load script x3
```

In the above example, Cmd x is a generic representation for any script command. Script x1 is the first level script and launches the second level script Script x2. Script x2 launches the third level script Script x3. If Script x3 contained a command to launch another script, the script would not launch the software would generate an error.

## Executing Scripts

You can execute scripts by loading a script file or through the Universal Port feature.

### *Execute a Script File*

Transfer the script file to the switch and use the load script *filename* {arg1} {arg2} ... {arg9} command.

### *Executing a Universal Port Script*

Universal port scripts are called profiles and are executed based on several types of trigger events, including device detection and undetection and user authentication and unauthentication.

For information on how to create profiles and configure the triggers, see Universal Port on page 309.

## Configuring Error Handling

The default error handling behavior is to ignore errors. You can change options within the scripts.

- To control script error handling, use the following command:
  `configure cli mode scripting [abort-on-error | ignore-error]`

## Aborting a Script

There are three ways to abort a script:

- Press **[Ctrl]** + **C** while the script is executing.
- Configure the switch to abort a script when an error occurs by using the following command:
  `configure cli mode scripting [abort-on-error | ignore-error]`
- Abort a script and store a status code in the $STATUS variable by using the following command:
  `return statusCode`

# Displaying CLI Scripting Information

You can use the information in the following sections to display CLI scripting information.

## Viewing CLI Scripting Status

The `show management` command displays whether or not CLI scripting is enabled, whether or not the configuration is persistent, and the CLI scripting error mode as shown in the following example:

```
show management
switch # show management
CLI idle timeout                  : Enabled (20 minutes)
CLI max number of login attempts  : 5
CLI max number of sessions        : 16
CLI paging                        : Enabled (this session only)
CLI space-completion              : Disabled (this session only)
CLI configuration logging         : Disabled
CLI scripting                     : Disabled (this session only)
CLI scripting error mode          : Ignore-Error (this session only)
CLI persistent mode               : Persistent (this session only)
CLI prompting                     : Enabled (this session only)
Telnet access                     : Enabled (tcp port 23 vr all)
: Access Profile : not set
SSH Access                        : ssh module not loaded.
Web access                        : Enabled (tcp port 80)
Total Read Only Communities       : 1
Total Read Write Communities      : 1
RMON                              : Disabled
SNMP access                       : Enabled
: Access Profile Name : not set
SNMP Traps                        : Enabled
SNMP v1/v2c TrapReceivers         :
Destination          Source IP Address     Flags
10.255.43.38 /10550                         2E
10.255.43.11 /10550                         2E
10.255.99.13 /10550                         2E
10.255.57.2 /10550                         2E
10.255.43.15 /10550                         2E
10.255.42.81 /10550                         2E
Flags:  Version: 1=v1 2=v2c
Mode: S=Standard E=Enhanced
SNMP stats:     InPkts 0       OutPkts   6       Errors 0       AuthErrors 0
Gets   0       GetNexts  0       Sets   0       Drops      0
SNMP traps:     Sent   6       AuthTraps Enabled
```

## Viewing CLI Scripting Variables

The show var command displays the currently defined variables and their values as shown in the following example:

```
Switch.4 # show var
----------------------------------------
Count : 3
----------------------------------------
--------------------------------------------------------------------
variableName                    variableValue
------------------------------  --------------------------------
CLI.SESSION_TYPE                serial
```

```
CLI.USER                          admin
STATUS                            0
-------------------------------------------------------------------
```

## Controlling CLI Script Output

When the load script command is entered, the software disables CLI scripting output until the script is complete, and then CLI scripting output is enabled.

When the CLI scripting output is disabled, the only script output displayed is the `show var {varname}` command and its output. When the CLI scripting output is enabled, all script commands and responses are displayed.

Use the `enable cli scripting output` and `disable cli scripting output` commands to control what a script displays when you are troubleshooting.

# CLI Scripting Examples

The following script creates 100 VLANS with IP Addresses from 10.1.1.1/16 to 10.100.1.1/16:

```
enable cli scripting
Set var count 1
while ($count < 101) do
Create vlan v$count
configure vlan v$count ipaddress 10.$(count).1.1/16
set var count ($count + 1)
endwhile
show vlan
```

The following script introduces a 60-second delay when executed:

```
set var temp $TCL(after [expr 60 *1000])
```

The following script displays the date and time:

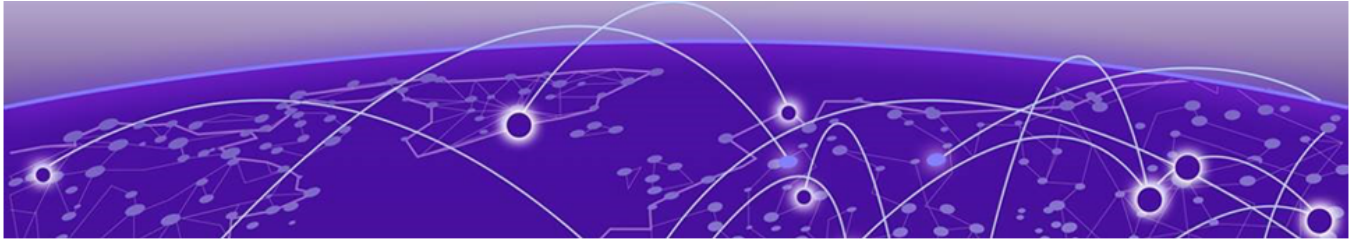
```
set var CLI.OUT " "
show switch
set var date $TCL(lrange ${CLI.OUT} 27 29)
set var year $TCL(lrange ${CLI.OUT} 31 31)
set var date $TCL(linsert $date 3 $year)
set var time $TCL(lrange ${CLI.OUT} 30 30)
show var date
show var time
```

The following script sorts the *FDB (forwarding database)* table in descending order:

```
set var CLI.OUT " "
show fdb
set var x1 $TCL(split ${CLI.OUT} "\n")
set var x2 $TCL(lsort -decreasing $x1)
set var output $TCL(join $x2 "\n")
show var output
```

The following script extracts the MAC address given the age of an FDB entry:

```
set var CLI.OUT " "
show fdb
set var input $TCL(split ${CLI.OUT} "\n")
set var y1 $TCL(lsearch -glob $input *age*)
set var y2 $TCL(lindex $input $y1)
set var y3 $TCL(split $y2 " ")
set var y4 $TCL(lindex $y3 0)
show var y4
```



# LLDP Overview



The *LLDP (Link Layer Discovery Protocol)* is defined by IEEE standard 802.1ab and provides a standard method for discovering physical network devices and their capabilities within a given network management domain.

LLDP-enabled network devices include repeaters, bridges, access points, routers, and wireless stations, and LLDP enables these devices to do the following:

- Advertise device information and capabilities to other devices in the management domain.
- Receive and store device information received from other network devices in the management domain.

LLDP-discovered information can be used to do the following:

- Discover information from all LLDP compatible devices in a multivendor environment.
- Trigger universal port profiles that can configure a switch port for a remote device (see Universal Port on page 309).
- Supply identity information that can be used for authentication and identity management (see CLEAR-Flow on page 948).
- Provide device information to *SNMP (Simple Network Management Protocol)* compatible Network Management Systems such as ExtremeManagement, which can present the information in inventory reports and topology maps.

The following sections provide additional information on LLDP support in the ExtremeXOS software.

## Supported Advertisements (TLVs)

*LLDP* defines a set of common advertisement messages. These are distributed in Type Length Value (TLV) format in an LLDP packet (see LLDP Packets on page 373).

The individual advertisements within the packet are called TLVs, and each TLV advertises device information or a device capability. Certain TLVs are mandatory and are always advertised after LLDP is enabled; optional TLVs are advertised only when so enabled by default or during configuration.

The following sections provide more information on TLVs.

## Mandatory TLVs

Mandatory TLVs are those TLVs that must be advertised (as defined in IEEE standard 802.1ab) when *LLDP* is enabled.

If you enable LLDP on a port, the mandatory TLVs are advertised and there is no command to disable this advertising.

**Table 43: Mandatory TLVs**

| Name | Comments |
|---|---|
| Chassis ID | The ExtremeXOS software advertises the system's MAC address in this TLV to uniquely identify the device.<br><br>Note:<br>*EDP (Extreme Discovery Protocol)* also uses the MAC address to identify the device. |
| Port ID | The ExtremeXOS software advertises the port ID in this TLV to uniquely identify the port that sends the TLVs. This port ID is the ifName object in the MIB. |
| Time to live (TTL) | This TLV indicates how long the record should be maintained in the LLDP database. The default value is 120 seconds (or 2 minutes). When a port is shutting down or LLDP is disabled on a port, this TLV is set to value 0, which indicates that the record for this port should be deleted from the database on other devices.<br>Although you cannot directly configure the TTL TLV, you can configure the transmit hold value, which is used to calculate the TTL TLV. (See Configuring LLDP Timers on page 377 for more information.) |
| End-of-LLDP PDU | The end-of-LLDPDU (LLDP protocol data unit) TLV marks the end of the data in each LLDP packet. |

## Optional TLVs

IEEE standard 802.1ab defines a set of optional TLVs, which are listed in the following table.

The system description TLV is advertised by default. All other optional TLVs must be configured to enable advertising. You can use the CLI to configure the optional TLVs, or you can use an *SNMP*-compatible network management system (NMS) such as ExtremeManagement or Ridgeline. For more information on the optional TLVs, see Configuring Optional TLV Advertisements on page 378.

**Table 44: Optional TLVs**

| Name | Comments |
|---|---|
| Port description | Advertises the display-string that is configured for the port. |
| System name | Advertises the device's configured system name, which is configured with the configure snmp sysname command. |

**Table 44: Optional TLVs (continued)**

| Name | Comments |
|---|---|
| System description | Advertises show version command information similar to the following:<br>`ExtremeXOS version 11.2.0.12 v1120b12 by release-manager on Fri Mar 18 16:01:08 PST 2005`<br>The default configuration advertises this optional TLV when *LLDP* is enabled. |
| System capabilities | Advertises bridge capabilities. If IP forwarding is enabled on at least one *VLAN (Virtual LAN)* in the switch, the software also advertises router capabilities. |
| Management address | Advertises the IP address of the management VLAN in the management address TLV. If the management VLAN does not have an assigned IP address, the management address TLV advertises the system's MAC address. |
| VLAN name | Advertises the name of a tagged VLAN for the port. You can configure this TLV multiple times to support multiple VLANs. |
| Port VLAN ID | Advertises the untagged VLAN on the port. |
| Port and protocol VLAN ID | Advertises a VLAN and whether the port supports protocol-based VLANs or not. You can configure this TLV multiple times to support multiple VLANs. |
| MAC/PHY configuration/ status | Advertises the autonegotiation and physical layer capabilities of the port. |
| Power via MDI | For Ethernet (*PoE (Power over Ethernet)*) or PoE+ ports, this TLV advertises the device type, power status, power class, and pin pairs used to supply power. |
| Link aggregation | Advertises information on the port's load-sharing (link aggregation) capabilities and status. |
| Maximum frame size | Advertises the maximum supported frame size for a port to its neighbors. When jumbo frames are not enabled on the specified port, the TLV advertises a value of 1518. If jumbo frames are enabled, the TLV advertises the configured value for the jumbo frames. |

## Avaya–Extreme Networks Optional TLVs

The software supports a set of TLVs that are proprietary to Avaya and Extreme Networks.

These TLVs advertise and receive information for Avaya voice over IP (VoIP) telephones, which include powered device (PD) information. Some TLVs are advertised by the switch, and some are advertised by the telephone. The switch starts advertising these proprietary TLVs when you enable LLDP and

configure the specified TLVs to be advertised. The switch receives the proprietary TLVs when *LLDP* is enabled; you do not have to configure the switch to receive individual TLVs.

**Table 45: Avaya–Extreme Networks TLVs**

| Name | Comments |
|---|---|
| *PoE* conservation level request | When the switch software needs to reduce power on a PoE-enabled port, this TLV advertises that request to the connected Avaya device. |
| Call server | The switch uses this TLV to advertise the IP addresses of up to eight call servers to an Avaya phone. |
| File server | The switch uses this TLV to advertise the IP addresses of up to four file servers to an Avaya phone. |
| 802.1Q framing | The switch uses this TLV to advertise information about Layer 2 priority tagging to an Avaya phone. |
| PoE Conservation level support | An Avaya phone uses this TLV to communicate the current power consumption level and current conservation level for the phone, including typical power value, maximum power value, and available conservation power levels. |
| IP phone address | An Avaya phone uses this TLV to communicate the IP address and mask configured in the phone, as well as a default gateway address. |
| CNA server | An Avaya phone uses this TLV to communicate the IP address of a Converged Network Analyzer (CNA). |

## LLDP MED Optional TLVs

*LLDP* Media Endpoint Discovery (MED) is an extension to LLDP that is published as standard ANSI/TIA-1057. LLDP MED TLVs advertise and receive information for endpoint devices, which can include powered device (PD) information. Some TLVs are advertised by the switch, and some are advertised by the endpoint device.

**Table 46: LLDP MED TLVs**

| Name | Comments |
|---|---|
| LLDP MED capabilities | The switch uses this TLV to advertise the switch LLDP MED capabilities to endpoint devices.<br>This TLV must be configured for advertisement before any other LLDP MED TLVs can be configured for advertisement, and advertisement for all other MED TLVs, must be disabled before advertisement for this TLV can be disabled. |
| Network policy | The switch uses this TLV to advertise network policy information for specific applications to endpoint devices.<br><br>Note:<br>Network policies cannot be configured by *SNMP*-based management programs. |
| Location ID | The switch uses this TLV to advertise a location ID to an endpoint device. |

**Table 46: LLDP MED TLVs (continued)**

| Name | Comments |
|---|---|
| Extended power via MDI | The switch uses this TLV to advertise a power information and settings to an endpoint device on a *PoE*-capable port. |
| Hardware revision | An endpoint device uses this TLV to advertise the configured hardware revision to the switch. |
| Firmware revision | An endpoint device uses this TLV to advertise the configured firmware revision to the switch. |
| Software revision | An endpoint device uses this TLV to advertise the configured software revision to the switch. |
| Serial number | An endpoint device uses this TLV to advertise the configured serial number to the switch. |
| Manufacturer name | An endpoint device uses this TLV to advertise the configured manufacturer name to the switch. |
| Model name | An endpoint device uses this TLV to advertise the configured model name to the switch. |
| Asset ID | An endpoint device uses this TLV to advertise the configured asset ID to the switch. |

You must enable the LLDP-MED capabilities TLV for advertising before you configure any other LLDP MED TLVs for advertising. Likewise, when disabling LLDP MED TLV advertisement, you can disable the LLDP-MED capabilities TLV only after you have disabled advertisement for all other LLDP MED TLVs.

After the LLDP-MED capabilities TLV is configured for advertising, the switch can receive LLDP MED TLVs from endpoints; you do not have to configure the switch to receive individual TLVs.

The switch advertises LLDP MED TLVs only after the switch receives an LLDP MED TLV from an endpoint, and the switch only advertises on ports from which an LLDP MED TLV has been received. This approach prevents the switch from advertising LLDP MED TLVs to another switch, and it prevents the wasted bandwidth and processing resources that would otherwise occur.

The LLDP MED protocol extension introduces a new feature called MED fast start, which is automatically enabled when the LLDP MED capabilities TLV is configured for advertisement.

When a new LLDP MED-capable endpoint is detected, the switch advertises the configured LLDP MED TLVs every one second for the configured number of times (called the repeat count). This speeds up the initial exchange of LLDP MED capabilities. After the repeat count is reached, the LLDP MED TLVs are advertised at the configured interval.

Note

The fast-start feature is automatically enabled, at the default level of 3, when you enable the LLDP MED capabilities TLV on the port.

## LLDP Packets

*LLDP* packets transport TLVs to other network devices (the following figure).

The LLDP packet contains the destination multicast address, the source MAC address, the LLDP EtherType, the LLDPDU data (which contains the TLVs), and a frame check sequence (FCS). The LLDP multicast address is defined as 01:80:C2:00:00:0E, and the EtherType is defined as 0x88CC.

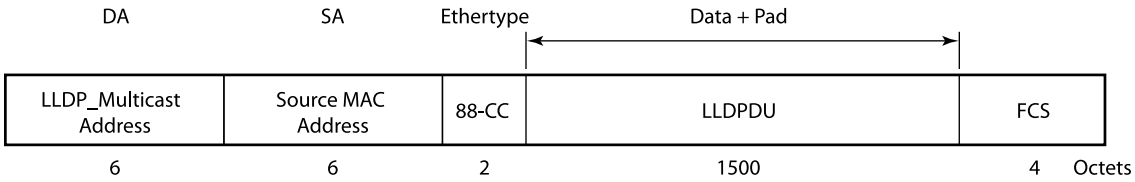



**Figure 51: LLDP Packet Format**

The following characteristics apply to LLDP packets:

- They are IEEE 802.3 Ethernet frames.
- The frames are sent as untagged frames.
- The frames are sent with a link-local-assigned multicast address as the destination address.
- The *STP (Spanning Tree Protocol)* state of the port does not affect the transmission of LLDP frames.

The length of the packet cannot exceed 1500 bytes. As you add TLVs, you increase the length of the LLDP frame. When you reach 1500 bytes, the remaining TLVs are dropped. We recommend that you advertise information regarding only one or two VLANs on the LLDP port, to avoid dropped TLVs.

If the system drops TLVs because of exceeded length, the system logs a message to the *EMS (Event Management System)* and the `show lldp statistics` command shows this information under the Tx Length Exceeded field.

Note

The LLDPDU maximum size is 1500 bytes, even with jumbo frames enabled. TLVs that exceed this limit are dropped.

## Transmitting LLDP Messages

You can configure each port to transmit *LLDP* messages, receive LLDP messages, or both.

When configured to transmit LLDP messages, the LLDP agent running on the switch passes serially through the list of ports that are enabled for LLDP. For each LLDP-enabled port, the switch periodically sends out an untagged LLDP packet that contains the mandatory LLDP TLVs as well as the optional

TLVs that are configured for advertisement. These TLVs are advertised to all neighbors attached to the same network. LLDP agents cannot solicit information from other agents by way of this protocol.

Note

When both IEEE 802.1X and LLDP are enabled on the same port, LLDP packets are not sent until one or more clients authenticate a port.

Also, LLDP MED TLVs are advertised only after an LLDP MED TLV is received on a port that is configured for LLDP MED. (See LLDP MED Optional TLVs on page 372.)

The source information for TLVs is obtained from memory objects such as standard MIBs or from system management information. If the TLV source information changes at any time, the LLDP agent is notified. The agent then sends an update with the new values, which is referred to as a triggered update. If the information for multiple elements changes in a short period, the changes are bundled together and sent as a single update to reduce network load.

## Receiving LLDP Messages

You can configure each port to transmit *LLDP* messages, receive LLDP messages, or both.

After you configure a port to receive TLVs, all LLDP TLVs are received (even if the LLDP MED capabilities TLV is not enabled). Each port can store LLDP information for a maximum of four neighbors.

Note

When both IEEE 802.1X and LLDP are enabled on the same port, incoming LLDP packets are accepted only when one or more clients are authenticated.

When configured to receive LLDP messages, the TLVs received at a port are stored in a standard Management Information Base (MIB), which makes it possible for the information to be accessed by an *SNMP*-compatible NMS such as ExtremeManagement or Ridgeline. Unrecognized TLVs are also stored on the switch, in order of TLV type. TLV information is purged after the configured timeout interval, unless it is refreshed by the remote LLDP agent. You can also manually clear the LLDP information for one or all ports with the `clear lldp neighbors` command.

- To display TLV information received from LLDP neighbors, use the following command:

  ```
  show lldp neighbors detailed
  ```

  You must use the detailed variable to display all TLV information.

## LLDP Management

You can manage *LLDP* using the CLI and/or an *SNMP*-compatible NMS such as ExtremeManagement or Ridgeline. LLDP works concurrently with the *EDP*. It also works independently; you do not have to run EDP to use LLDP.

You can use the `save configuration` command to save LLDP configurations across reboots.

The switch logs *EMS* messages regarding LLDP, including when optional TLVs exceed the 1500-byte limit and when more than four neighbors are detected on a port.

After you enable LLDP, you can enable LLDP-specific SNMP traps; the traps are disabled by default. After you enable LLDP-specific traps, the switch sends all LLDP traps to the configured trap receivers. You can configure the period between SNMP notifications; the default interval is five seconds.

You can configure an optional TLV to advertise or not advertise the switch management address information to the neighbors on a port. When enabled, this TLV sends out the IPv4 address configured on the management *VLAN*. If you have not configured an IPv4 address on the management VLAN, the software advertises the system's MAC address. LLDP does not send IPv6 addresses in this field.

# Configuring and Managing LLDP

The following sections describe how to configure *LLDP* on the switch.

## Configuration Overview

You configure *LLDP* per port, and each port can store received information for a maximum of four neighbors.

Note

LLDP runs with link aggregation.

You can configure LLDP per port.

1. Enable LLDP on the desired port(s) as described in Enable and Disable LLDP.
2. If you want to change any default timer values, see Configure LLDP Timers.
3. Enable the *SNMP* traps and configure the notification interval as described in Configure SNMP for LLDP.
4. Configure any optional TLV advertisements as described in Configuring Optional TLV Advertisements.

Note

By default, an LLDP-enabled port advertises the optional system description TLV. By default, all other optional TLVs are not advertised.

## Enable and Disable LLDP

*LLDP* is enabled on all ports by default. When you enable LLDP on one or more ports, you select whether the ports will only transmit LLDP messages, only receive the messages, or both transmit and receive LLDP messages.

After you enable LLDP, the following TLVs are automatically added to the LLDPDU:

- Chassis ID
- Port ID
- TTL
- System description
- End of LLDPDU

All of these, except the system description, are mandated by the 802.1ab standard and cannot be configured. For information on changing the system description TLV advertisement, see System Description TLV on page 379.

- Enable LLDP.

  ```
  enable lldp ports [all | port_list] {receive-only | transmit-only}
  ```

- Disable LLDP.

  ```
  disable lldp ports [all | port_list] {receive-only | transmit-only}
  ```

## Configuring LLDP Timers

The *LLDP* timers apply to the entire switch and are not configurable by port. After you enable LLDP, LLDP timers control the time periods for the transmission and storage of LLDP TLVs as follows:

- Reinitialization period (default is two seconds).
- Delay between LLDP transmissions (default is two seconds)—applies to triggered updates, or updates that are initiated by a change in the topology.
- Transmit interval (default is 30 seconds)—applies to messages sent periodically as part of protocol.
- Time-to-live (TTL) value (default is two minutes)—time that the information remains in the recipient's LLDP database.

**Note**

Once the LLDP MED TLVs begin transmitting (after detecting LLDP MED TLVs from a connected endpoint), those TLVs are also controlled by these timers.

When LLDP is disabled or if the link goes down, LLDP is reinitialized. The reinitialize delay is the number of seconds the port waits to restart the LLDP state machine; the default is two seconds.

The time between triggered update LLDP messages is referred to as the transmit delay, and the default value is two seconds. You can change the default transmit delay value to a specified number of seconds or to be automatically calculated by multiplying the transmit interval by 0.25.

Each LLDP message contains a TTL value. The receiving LLDP agent discards all LLDP messages that surpass the TTL value; the default value is 120 seconds. The TTL is calculated by multiplying the transmit interval value and the transmit hold value; the default transmit hold value is four.

- To change the default reinitialize delay period, use the following command:

  ```
  configure lldp reinitialize-delay seconds
  ```

  LLDP messages are transmitted at a set interval; this interval has a default value of every 30 seconds.

- To change this default value, use the following command:

  ```
  configure lldp transmit-interval seconds
  ```

- To change the value for the transmit delay, use the following command:

  ```
  configure lldp transmit-delay [ auto | seconds]
  ```

- To change the default transmit hold value, use the following command:

  ```
  configure lldp transmit-hold hold
  ```

## Configuring SNMP for LLDP

By default, *SNMP* *LLDP* traps are disabled on all ports. The default value for the interval between SNMP LLDP trap notifications is 5 seconds.

- To enable LLDP SNMP traps on one or more ports, use the following command:
  `enable snmp traps lldp {ports [all | port_list]}`

  The traps are only sent for those ports that are both enabled for LLDP and have LLDP traps enabled.
- To disable the LLDP SNMP traps on one or more ports, use the following command:
  `disable snmp traps lldp {ports [all | port_list]}`
- To change the default value for the interval for the entire switch, use the following command:
  `configure lldp snmp-notification-interval seconds`

Note

If you want to send traps for LLDP MED, you must configure it separately. Use the `enable snmp traps lldp-med {ports [all | port_list]}` command to enable these traps.

## Configuring Optional TLV Advertisements

By default, the ExtremeXOS software advertises the mandatory *LLDP* TLVs (which are not configurable) and the optional system description TLV. For details, see Supported Advertisements (TLVs) on page 369). All other optional TLVs are not advertised.

You can choose to advertise or not advertise any optional TLV, but be aware that the total LLDPDU length, which includes the mandatory TLVs, cannot exceed 1500 bytes. Optional TLVs that cause the LLDPDU length to exceed the 1500-byte limit are dropped. You can see if the switch has dropped TLVs by referring to the *EMS* log or by issuing the `show lldp statistics` command.

The following sections describe configuration for the following types of optional TLVs.

### *Configuring Standards-Based TLVs*

This section describes the following optional standards-based TLVs.

#### Port description TLV

The port description TLV advertises the ifDescr MIB object, which is the ASCII string you configure.

The string can be configured using the `configure ports display-string` command.

If you have not configured this parameter, the TLV carries an empty string.

To control advertisement of the port description TLV, use the following command:

`configure lldp ports [all | port_list] [advertise | no-advertise] port-description`

#### System name TLV

The system name TLV advertises the configured system name for the switch. This is the sysName as defined in RFC 3418, which you can define using the `configure snmp sysname` command.

To control advertisement of the system name TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
system-name
```

#### System Description TLV

By default, the ExtremeXOS software advertises this TLV whenever you enable *LLDP* on a port, but you can disable advertisement. This TLV advertises show version command information similar to the following in the system description TLV:

```
ExtremeXOS version 11.2.0.12 v1120b12 by release-manager
on Fri Mar 18 16:01:08 PST 2005
```

To control advertisement of the system description TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
system-description
```

#### System Capabilities TLV

The system capabilities TLV advertises the capabilities of the switch and which of these capabilities are enabled. When so configured, the ExtremeXOS software advertises bridge capabilities. If IP forwarding is enabled on at least one *VLAN* in the switch, the software also advertises router capabilities.

To control advertisement of the system capabilities TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
system-capabilities
```

#### Management Address TLV

The management address TLV advertises the IP address of the management *VLAN*. If the management VLAN does not have an assigned IP address, the management address TLV advertises the system's MAC address. *LLDP* does not recognize IPv6 addresses in this field.

To control advertisement of the management address TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
management-address
```

Note

The ExtremeXOS software sends only one management address TLV.

#### VLAN Name TLV

The *VLAN* name TLV advertises a VLAN name for one or more VLANs on the port. You can advertise this TLV for tagged and untagged VLANs. When you enable this TLV for tagged VLANs, the TLV advertises the IEEE 802.1Q tag for that VLAN. For untagged VLANs, the internal tag is advertised.

If you do not specify a VLAN when you configure this TLV, the switch advertises all VLAN names on the specified ports. You can choose to advertise one or more VLANs for the specified ports by specifying the name of a VLAN in the configuration command. You can repeat the command to specify multiple VLANs.

To control advertisement of the port VLAN Name TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific dot1 vlan-name {vlan [all | vlan_name]}
```

**Note**

Because each VLAN name requires 32 bits and the LLDPDU cannot exceed 1500 bytes, we recommend that you configure each port to advertise no more than one or two specific VLANs. Optional TLVs that cause the LLDPDU length to exceed the 1500-byte limit are dropped.

#### Port VLAN ID TLV

The port *VLAN* ID advertises the untagged VLAN on that port. Thus, only one port VLAN ID TLV can exist in the LLDPDU. If you configure this TLV and there is no untagged VLAN on the specified port, this TLV is not included in the LLDPDU.

- To control advertisement of the port VLAN ID TLV, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific dot1 port-vlan-ID
  ```

#### Port and Protocol VLAN ID TLV

When configured for advertisement, this TLV advertises whether the specified *VLAN*s on the specified ports support protocol-based VLANs or not.

If you do not specify a VLAN when you configure this TLV, the switch advertises protocol-based VLAN support for all VLAN names on the specified ports. You can choose to advertise support for one or more VLANs for the specified ports by specifying the name of a VLAN in the configuration command. You can repeat the configuration command to specify multiple VLANs.

Because all VLANs on Extreme Networks switches support protocol-based VLANs, the switch always advertises support for protocol-based VLANs for all VLANs for which this TLV is advertised. If no protocol-based VLANs are configured on the port, the switch sets the VLAN ID value to 0.

- To control advertisement of the port and protocol VLAN ID TLV, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific dot1 port-protocol-vlan-ID {vlan [all | vlan_name]}
  ```

**Note**

Because a TLV is advertised for every VLAN that is advertised, and because the LLDPDU cannot exceed 1500 bytes, we recommend that you advertise this VLAN capability only for those VLANs that require it. Optional TLVs that cause the LLDPDU length to exceed the 1500-byte limit are dropped.

#### MAC/PHY Configuration/Status TLV

When configured for advertisement, this TLV advertises the autonegotiation and physical layer capabilities of the port. The switch adds information about the port speed, duplex setting, bit rate, physical interface, and autonegotiation support and status.

- To control advertisement of the port and protocol MAC/PHY configuration/status TLV, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific dot3 mac-phy
  ```

#### Power Via MDI TLV

The device type field contains a binary value that represents whether the *LLDP*-compatible device transmitting the LLDPDU is a power sourcing entity (PSE) or power device (PD), as listed in the following table.

**Table 47: Power Management TLV Device Information**

| Value | Power source |
|---|---|
| 0 | PSE device |
| 1 | PD device |
| 2-3 | Reserved |

Control advertisement of the power via MIDI TLV.

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific dot3 power-via-mdi {with-classification}
```

Refer to for Configuring Avaya-Extreme TLVs on page 382 and Configuring LLDP MED TLVs on page 383 more information on power-related TLVs.

#### Link Aggregation TLV

When configured for advertisement, this TLV advertises information on the port's load-sharing (link aggregation) capabilities and status.

To control advertisement of the link aggregation TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific dot3 link-aggregation
```

#### Maximum frame size TLV

When configured for advertisement, this TLV advertises the maximum supported frame size for a port to its neighbors. When jumbo frames are not enabled on the specified port, the TLV advertises a value of 1518. If jumbo frames are enabled, the TLV advertises the configured value for the jumbo frames.

To control advertisement of the maximum frame size TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific dot3 max-frame-size
```

*Configuring Avaya-Extreme TLVs*

This section describes the following optional proprietary Avaya-Extreme Networks TLVs that you can configure the switch to transmit.

Note

You can display the values for these TLVs using the `show lldp neighbors detailed` command.

**PoE Conservation Level Request TLV**

When configured for advertisement, this TLV advertises a request to the connected PD to go into a certain power conservation level or go to the maximum conservation level. This *LLDP* TLV is sent out only on PoE-capable Ethernet ports.

By default, the requested conservation value on this proprietary LLDP TLV is 0, which is no power conservation. You can change this level temporarily using a network station or *SNMP* with the MIB; this change is not saved across a reboot.

- To control advertisement of the *PoE* conservation level request TLV, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific avaya-extreme poe-conservation-request
  ```

**Call Server TLV**

When configured for advertisement, this TLV advertises the IP addresses of up to eight call servers. Avaya phones use this addressing information to access call servers.

- To control advertisement of the call server TLV and define call server addresses, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific avaya-extreme call-server ip_address_1 {ip_address_2
  {ip_address_3 {ip_address_4 {ip_address_5 {ip_address_6 {ip_address_7
  {ip_address_8}}}}}}
  ```

**File Server TLV**

When configured for advertisement, this TLV advertises the IP addresses of up to 4 file servers. Avaya phones use this address information to access file servers.

- To control advertisement of the file server TLV and define file server addresses, use the following command:

  ```
  configure lldp ports [all | port_list] [advertise | no-advertise]
  vendor-specific avaya-extreme file-server ip_address_1 {ip_address_2
  {ip_address_3 {ip_address_4}}}
  ```

#### 802.1Q Framing TLV

When configured for advertisement, this TLV advertises information about Layer 2 priority tagging for Avaya phones.

- To control advertisement of the 802.1Q framing TLV, use the following command:
  `configure lldp ports [all | port_list] [advertise | no-advertise] vendor-specific avaya-extreme dot1q-framing [tagged | untagged | auto]`

> **Note**
> For this command to work, you must have previously enabled both the `configure lldp ports vendor-specific med capabilities` and the `configure lldp ports vendor-specific med policy application` commands. (See Configuring LLDP MED TLVs on page 383 for complete information.)

*Configuring LLDP MED TLVs*

> **Note**
> After you enable an *LLDP* MED TLV, the switch waits until it detects a MED-capable device before it begins transmitting the configured LLDP MED TLVs.
> You must configure the LLDP MED capabilities TLV to advertise before any of the other LLDP MED TLVs can be configured. Also, this TLV must be set to no-advertise after all other MED TLVs are set to no-advertise.

This approach assures that network connectivity devices advertise LLDP MED TLVs only to end-devices and not to other network connectivity devices.

The following sections describe LLDP MED TLVs and features.

> **Note**
> You can display the values for these TLVs using the `show lldp neighbors detailed` command.

#### LLDP MED capabilities TLV

This TLV advertises the *LLDP* MED capabilities of the switch and must be enabled before any of the other LLDP MED TLVs can be enabled.

To enable configuration and transmission of any other LLDP MED TLV and to determine the LLDP MED capabilities of endpoint devices, use the following command:

`configure lldp ports [all | port_list] [advertise | no-advertise] vendor-specific med capabilities`

#### LLDP MED Fast-Start Feature

The *LLDP* MED fast-start feature allows you to increase the learning speed of the switch for LLDP MED TLVs. The fast-start feature is automatically enabled once you enable the LLDP MED capabilities TLV.

By default, the switch sends out the LLDPDU every second for up to the default repeat count, which is 3. Once the repeat count is reached, the configured transmit interval value is used between LLDPDUs. You can configure the repeat count to any number in the range of 1 to 10.

To configure the LLDP fast-start feature, use the following command:

```
configure lldp med fast-start repeat-count count
```

#### Network policy TLV

This TLV advertises which *VLAN* an endpoint should use for the specified application. You can configure only one instance of an application on each port, and you can configure a maximum of eight applications, each with its own DSCP value and/or priority tag.

To control advertisement and configure one or more network policy TLVs for a port, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific med policy application [voice | voice-signaling |guest-
voice | guest-voice-signaling | softphone-voice | video-conferencing |
streaming-video | video-signaling] vlan vlan_name dscp dscp_value
{priority-tagged}
```

#### Location identification TLV

This TLV advertises one of three different location identifiers for a port, each with a different format.

- Coordinate-based, using a 16-byte hexadecimal string.
- Civic-based, using a hexadecimal string with a minimum of six bytes.
- ECS ELIN, using a numerical string with a range of 10-25 characters.

To control advertisement and configure location information, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific med location-identification [coordinate-based hex_value
| civic-based hex_value | ecs-elin elin]
```

#### Extended power-via-MDI TLV

This TLV advertises fine-grained power requirement details, including the *PoE* settings and support. You can enable this TLV only on PoE-capable ports; the switch returns an error message if you attempt to transmit this *LLDP* TLV over a non-PoE-capable port.

To control advertisement for this TLV, use the following command:

```
configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific med power-via-mdi
```

#### SNMP Traps for LLDP MED

To receive *SNMP* traps on the *LLDP* MED, you must enable these separately from the other LLDP traps. (See Configuring SNMP for LLDP on page 378.)

- Enable the LLDP MED SNMP traps.

  ```
  enable snmp traps lldp-med {ports [all | port_list]}
  ```

- Disable the LLDP MED SNMP traps.

  ```
  disable snmp traps lldp-med {ports [all | port_list]}
  ```

### Clearing LLDP Neighbor Entries

- To remove the *LLDP* entries received on one or more ports from the switch database, use the following command:

  ```
  clear lldp neighbors [all | port port_list]
  ```

### Unconfiguring LLDP

- To unconfigure the *LLDP* timers, use the following command:

  ```
  unconfigure lldp
  ```

  This command returns the LLDP timers to default values; LLDP remains enabled, and all the configured TLVs are still advertised.
- To leave LLDP enabled, but reset the advertised TLVs to the five default TLVs, use the following command, and specify the affected ports:

  ```
  unconfigure lldp port [all | port_list]
  ```

## Displaying LLDP Information

The following sections describe how to display *LLDP* information for the switch. The system displays information on the LLDP status and statistical counters of the ports, as well as about the LLDP advertisements received and stored by the system.

You can display information on the LLDP port configuration and on the LLDP neighbors detected on the port.

### Displaying LLDP Port Configuration Information and Statistics

- To display *LLDP* port configuration information, use the following command:

  ```
  show lldp {port [all | port_list]} {detailed}
  ```
- To display the statistical counters related to the LLDP port, use the following command:

  ```
  show lldp {port [all | port_list]} statistics
  ```

### Display LLDP Information Collected from Neighbors

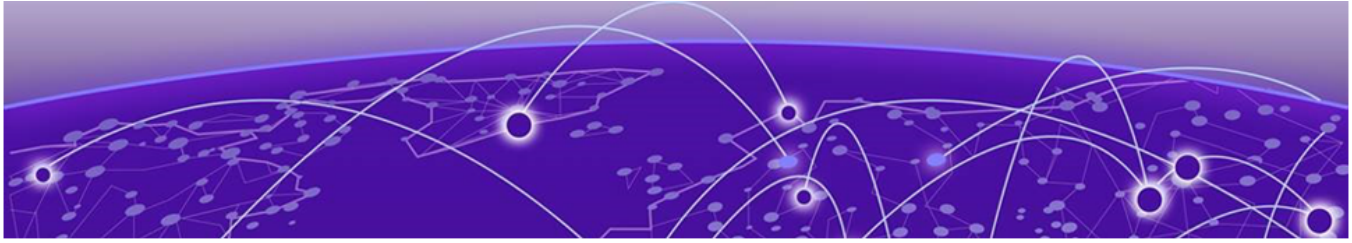
- Display information collected from *LLDP* neighbors.

  show lldp neighbors

  Note

  You must use the detailed option to display information on the proprietary Avaya-Extreme Networks TLVs and the LLDP MED TLVs.



# OAM



Operation, Administration, and Maintenance (OAM) includes functions used to detect network faults, measure network performance and distribute fault-related information, including CFM, Y.1731, EFM, and BFD.

## CFM

Connectivity Fault Management (CFM), discussed in the IEEE 802.1Q-2011 standard and originally specified in the IEEE 802.1ag-2007 standard, allows you to detect, verify, and isolate connectivity failures in virtual bridged LANs.

Part of this specification is a toolset to manually check connectivity, which is sometimes referred to as Layer 2 ping.

Note
The ExtremeXOS implementation of CFM is based on the IEEE 802.1Q-2011 standard.

There is no direct interaction between CFM and other Layer 2 protocols; however, blocked *STP (Spanning Tree Protocol)* ports are taken into consideration when forwarding CFM messages.

### CFM Overview

You can create hierarchical networks, or domains, and test connectivity within a domain by sending Layer 2 messages, known as Connectivity Check Messages (CCMs).

Note
Extreme Networks uses values defined in IEEE 802.1Q-2011 for the MAC addresses and Ethernet type for CFM.

The following figure shows an example of hierarchical CFM domains.

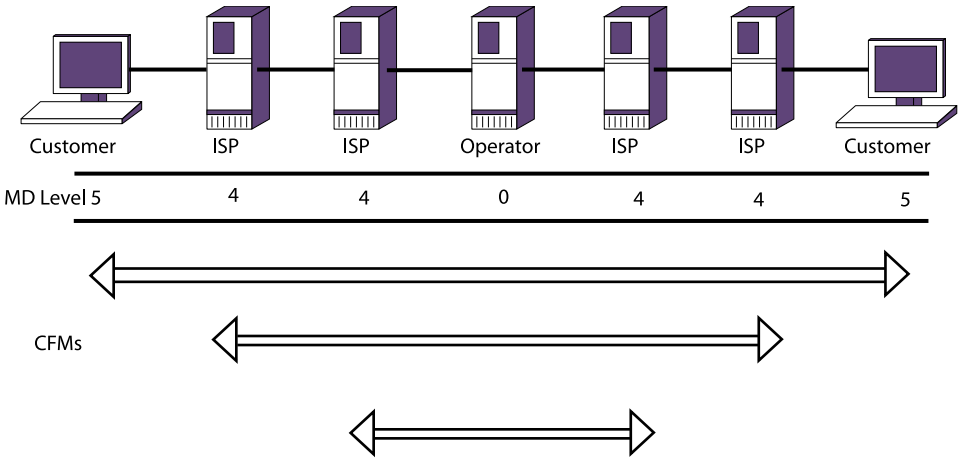



**Figure 52: CFM Hierarchical Domains Example**

**Note**
The arrows in the above figure indicate the span that CCM messages take, not the direction. (See Figure 53 on page 388 for more information on spans for CCM messages.) This has been removed until the missing xref can be fixed.

To achieve this hierarchical connectivity testing, create and configure the following entities:

- Maintenance domains, or domains
- Maintenance domain (MD) level; a unique hierarchical numeric value for each domain
- Maintenance associations (MAs)
- Maintenance points (MPs) and maintenance end points (MEPS), which are one of the following types:
  - UP MEPs
  - DOWN MEPs
- Maintenance intermediate points (MIPs)

**Note**
The CFM filter function (CFF) is no longer supported from ExtremeXOS 12.1. The functionality of CFF is implicitly performed by MEPs.

An UP MEP sends CFM frames toward the frame filtering entity, which forwards the frames to all other ports of a service instance other than the port on which the UP MEP is configured. This is similar to how the frame filtering entity forwards a normal data frame, taking into account the port's *STP* state. For an UP MEP, a CFM frame exits from a port if only if the STP state of the port is in the forwarding state.

A DOWN MEP sends CFM frames directly to the physical medium without considering the port STP state. For a DOWN MEP, a CFM frame exits from a port even if the port STP state is in blocking state.

The following figure shows the concept of UP and DOWN MEP at logical level:

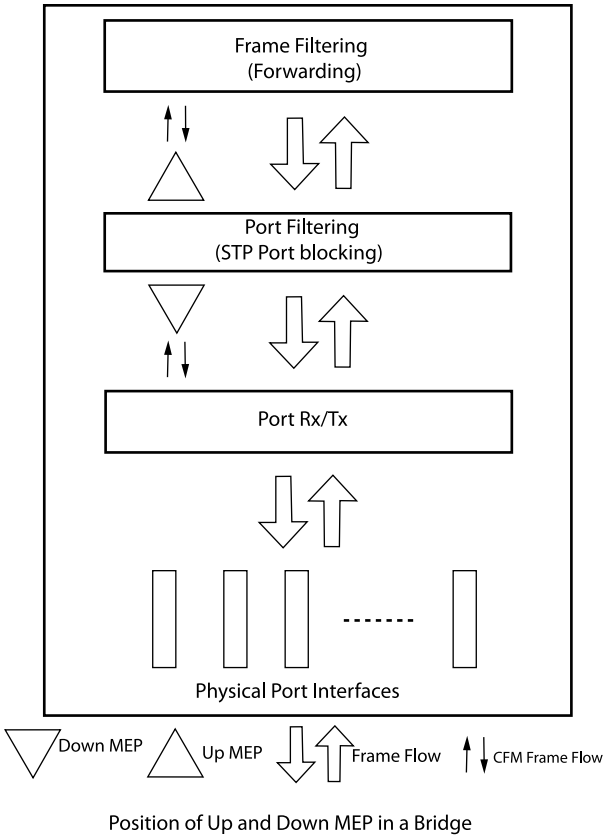



**Figure 53: CFM UP and DOWN MEP at the Logical Level**

You must have at least one MP on an intermediate switch in your domain. Ensure that you map and configure all ports in your domain carefully, especially the UP MEPs and the DOWN MEPs. If these are incorrectly configured, the CCMs are sent in the wrong direction in your network, and you will not be able to test the connectivity within the domain.

You can have up to eight domains on an Extreme Networks switch. A domain is the network or part of the network for which faults are to be managed; it is that section where you are monitoring Layer 2 connectivity. A domain is intended to be fully connected internally.

Note

Domains may cross VR boundaries; domains are not *virtual router (VR)*-aware.

You assign each domain an MD level, which functions in a hierarchy for forwarding CFM messages. The MD levels are from 0 to 7. The highest number is superior in the CFM hierarchy.

The IEEE standard 802.1Q-2011 recommends assigning different MD levels to different domains for different network users, as follows:

- 5 to 7 for end users
- 3 and 4 for Internet service providers (ISPs)
- 0 to 3 for operators (entities carrying the information for the ISPs)

All CFM messages with a superior MD level (numerically higher) pass throughout domains with an inferior MD level (numerically lower). CFM messages with an inferior MD level are not forwarded to

domains with a superior MD level. Refer to the following table for an illustration of domains with hierarchical MD levels.

**Table 48: MD Levels and Recommended Use**

| MD level | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Use | Operator | | | Service provider | | Customer | | |
| Superiority | Most inferior | < ----- Inferior / Superior ----- > | | | | | | Most superior |

Within a given domain, you create maintenance associations (MAs). Extreme Networks' implementation of CFM associates MAs with service instances (a service instance can be a *VLAN (Virtual LAN)*, VMAN, BVLAN, or SVLAN). All of the ports in that VLAN service instance are now in that MA and its associated domain. In general, you should configure one MIP on each intermediate switch in the domain and a MEP on every edge switch.

Each MA associates with one service instance, and a service instance can be associated with more than one MA. The MA is unique within the domain. One switch can have 8 domains, 128 ports, and 256 associations (see Supported Instances for CFM).

Note
You cannot associate the Management VLAN with an MA or a domain.

You assign the MPs to ports: UP MEPs, DOWN MEPs, and MIPs. These various MPs filter or forward the CFM messages to test the connectivity of your network.

Each configured MEP periodically sends out a Layer 2 multicast or unicast CCM message.

The destination MAC address in the CCM frame is from a multicast MAC address range that is reserved for CFM messages. Each MEP must have a MEP ID that is unique within the MA. The MEPs send the CCM messages differently, depending on the configuration, as follows:

- The DOWN MEPs sends out a single CCM message.
- The UP MEPs potentially sends the CCM message to all ports on the service instance (MA)—except the sending port—depending on the MPs configured on the outgoing ports.

Note
Ensure that you configured the UP and DOWN MEPs correctly, or the CCMs will flow in the wrong direction through the domain and not allow connectivity testing.

MIPs define intermediate points within a domain. MIPs relay the CCM messages to the next MIP or MEP in the domain.

You configure the time interval for each MEP to send a CCM. We recommend setting this interval for at least one second. Each MEP also makes a note of what port and what time it received a CCM. This information is stored in the CCM database.

Each CCM has a time-to-live (TTL) value also noted for that message. This TTL interval is 3.5 times the CCM transmission interval you configured on the switch that is originating the CCM. After the TTL expires, the connectivity is considered broken, and the system sends a message to the log. One

important result of the continual transmission of CCM frames is that the MAC address of the originating MEP is known to all MPs in the association.

Note
All MEPs in an MA must be configured with the same CCM transmission interval.

The MD values are from 0 to 7; in the hierarchy, the MD level of 0 is lowest and 7 is highest.

Not all combinations of MPs are allowed on the same port within an MA; only the following combinations can be on the same port within an MA:

- UP MEP and MIP
- DOWN MEP with neither UP MEP nor MIP

CFM protocol imposes the following MP restrictions within an MA on a switch:

- MA can have either up MEP or down MEP and not both.
- MA can have multiple Down MEPs.
- Only one Up MEP per MA.
- MA can have both up MEP and MIP.
- MA cannot have MIP if down MEP is present.
- Down MEPs on regular ports are created in hardware for all CCM intervals 3.3 msec–600000 sec on Summit X460, E4G-400, and E4G-200.
- Up MEPs and MEPs on *LAG (Link Aggregation Group)* ports are created in software with CCM intervals 100 msec–600000 sec on all platforms.
- Dynamic Remote MEP learning is not supported for the MEPs created in hardware. You must explicitly create static Remote MEPs.
- Sender-Id-IP Address cannot be configured for the MEPs created in hardware.
- Unicast CCM transmission is not supported by the MEPs created in hardware.
- Domain name format should be of string type to create any MEPs in hardware in that domain.
- The CCM transmission state is disabled by default for the MEPs created in hardware by the CFM user interface.
- The CCM transmission state is enabled by default for the MEPs created in hardware by CFM clients like *ERPS (Ethernet Ring Protection Switching)*.
- The hardware Remote MEP status appears in `show cfm detail`. It is also forwarded to the client if created by a client like ERPS.
- CFM objects like domain, association, MEP, Remote MEP created by a client are not saved by dot1ag.

Note
An MA can have an UP MEP in one switch and a DOWN MEP on another switch for the same MA.

## Ping and Traceroute

When operators see a connectivity fault message from CCM in the system log, they can send a loopback message (LBM) or a link trace message (LTM).

These are also referred to as a Layer 2 ping or a traceroute message. You can send with an LBM or an LTM only from an MEP (either UP or DOWN).

You can only send a ping from a MEP, and you ping to the unique system MAC address on the switch you are testing connectivity to. The operator sends out a unicast LBM, and the first MIP or MEP on the switch with the destination MAC address matching the embedded MAC address replies with an LBR (loopback reply).

You can only send a traceroute (LTM) from a MEP. You send the traceroute to the unique system MAC address on the switch to which you are testing connectivity. The system sends out an LTM to the special multicast address. If the destination address is not present in the *FDB (forwarding database)*, the LTM is flooded on all the ports in the MIP node. Each MIP in the path passes the frame only in the direction of the path and sends a link trace reply (LTR) message back to the originating with information that the LTM passed. The traceroute command displays every MIP along the path (see `traceroute mac port` ).

## Supported Instances for CFM

The following table displays the CFM support in ExtremeXOS.

**Table 49: ExtremeXOS CFM Support**

| Item | Limit | Notes |
|---|---|---|
| Domains | 8 | Per switch; one for each MD level |
| Associations (MAs) | 256 | Per switch |
| UP MEPs | 32 on Summit Family switches, SummitStack, E4G-200, E4G-400, BDx8, and BlackDiamond 8000 series modules. | Per switch |
| DOWN MEPs | 256 hardware-placed MEPs on Summit series X460, E4G-400, E4G-200 (non-load shared ports)<br>32 on Summit series X460, E4G-400, E4G-200 (load shared ports)<br>32 on other Summit family switches, BDx8, and BlackDiamond 8000 series | Per switch |
| MIPs | 32 on Summit Family switches, SummitStack, BDx8, and BlackDiamond 8000 series modules. | Per switch |

[3] RMEPs need to be explicitly configured for hardware MEPs. Unlike software MEPs, hardware MEPs do not support dynamic RMEP learning.

**Table 49: ExtremeXOS CFM Support (continued)**

| Item | Limit | Notes |
|---|---|---|
| Total CFM ports | 128 | Per switch; total number of all ports for all service instances assigned to an MA (see command for ports configured for CFM) |
| Entries in the CCM database | 2000 | Number of remote end points stored in a CCM database on each MEP; 64 end points per MEP (additional CCMs discarded after this limit is reached) |
| CFM Segments | 1000 | Maximum number of CFM segments for Y.1731. |
| CFM Groups | 1000 | Maximum number of CFM groups. |

**Note**

The total number of CFM ports is a guideline, not a limit enforced by the system.

*CFM Groups*

Loop detection protocols like EAPS/*ERPS* want to depend upon CFM to detect link status for faster failover recovery.

They register with LMEP and RMEP objects created by CFM in order to receive the link status event notifications to take the necessary action.

Currently LMEP is identified with domain, association, port, MEPId quadruples. And RMEP is identified with domain, association, LMEP, RMEPId quadruples. Each LMEP can be tied up to multiple RMEPs. So applications need to configure domain, association, LMEP and RMEPs through a client/server interface.

To simplify this, CFM provides a simple API to client applications to register/deregister CFM with a specified string name. The string name can be identified as a CFM group that binds an LMEP to multiple RMEPs. The group name is unique across the switch. Each application can create its own group for a required LMEP/RMEP combination.

You can associate a group to each LMEP created on a port. There exists a one-to-one relationship between LMEP-port-group. Whenever CFM stops receiving CCMs on this port, it informs a group DOWN event to registered clients like ERPS/EAPS. Whenever CFM starts receiving the CCMs again on this port, a group UP event is sent to registered clients.

## Configuring CFM

To configure CFM, create a maintenance domain and assign it a unique MD level. Next, associate MAs with the specified domain and assign MPs within that MA. Optionally, you can configure the transmission interval for the CCMs, destination MAC type for an MA and remote MEPs statically in an MA.

If a MEP fails to receive a CCM before the last advertised TTL value expires, the system logs a message. After the network administrator sees the system log message, he can send a Layer 2 ping and/or a traceroute message to isolate the fault.

Note

CFM does not use *ACL (Access Control List)*; there are no additional ACL entries present for CFM in the show access-list dynamic command output.

ExtremeXOS 15.5 provides support for transmitting and receiving ITU-T Y.1731 CCMs. The main difference between IEEE 802.1ag and ITU-T Y.1731 CCMs is between the MAID and MEG ID formats in CCMs:

- MAID ---- MA (format + length + name) with/without MD (format + length +name) details.
- MEG ID ---- MEG (format + length + name) without MD details.
- MA is referred to as MEG in Y.1731 and both are same.
- MA assumes four formats (string/integer/vpn-id/vlan-id) for its name.
- MEG assumes ICC format which is a combination of ICC (6 bytes) + organization specific UMC (6 bytes).
- To support Y.1731 CCMs, an additional name format for MEG name is added for association.

### *Creating Maintenance Domains*

You can create maintenance domains (MDs), or domains, and assign a unique MD level at that time. Available MD levels are numbered from 0 to 7. Higher numerical values are superior MD levels in the CFM hierarchy. Each switch can have a total of eight domains, each with a unique MD level.

You can name domains using any one of the following three formats:

- Simple string—Use an alphanumeric character string with a maximum of 43 characters.
- Domain name server (DNS) name—Use an alphanumeric character string with a maximum of 43 characters.
- MAC address plus 2-octet integer—Use a MAC address and a 2-octet integer. The display format is XX.XX.XX.XX.XX.XX.YYY, where X is the MAC address, and Y is the 2-octet integer. For example, a domain name in this format using 123 as the 16-bit unsigned integer appears as follows: 00:11:22:33:44:55.123.

Note

Whatever convention you choose, you must use that same format throughout the entire domain.

The CFM messages carry the domain name, so the name and naming format must be identical to be understood throughout the domain. You can, however, use different naming conventions on different domains on one switch (up to eight domains allowed on one switch). User-created CFM names are not case sensitive.

- To create a domain and assign an MD level using the DNS convention, use the following command:.

  `create cfm domain dns` *`name`* **`md-level`** *`level`*

- To create a domain and assign an MD level using the MAC address convention, use the following command:.

  `create cfm domain mac` *`mac-addr int`* **`md-level`** *`level`*

- To create a domain and assign an MD level using the string convention, use the following command :

  `create cfm domain` **`string`** *`str_name`* **`md-level`** *`level`*

- Although you assign an MD level to the domain when you create that domain, you can change the MD level on an existing domain by running:

  `configure cfm domain` *`domain_name`* **`md-level`** *`level`*

- To delete a domain, use the following command:

  `delete cfm domain` *`domain`*

*Creating and Associating MAs*

Within a given domain, you can associate maintenance associations (MAs). Extreme Networks' implementation of CFM associates MAs with service instances. All of the ports in that service instance are now in that MA and its associated domain.

Each MA associates with one service instance, and each service instance may associate with more than one MA; you can configure more than one MAs in any one domain.

Like the domains, ExtremeXOS supports multiple formats for naming the MA. The following formats are supported for naming the MAs:

- Character string
- 2-octet integer
- RFC 2685 VPN
- *VLAN* ID

- To add an MA to a domain using the character string format, use the following command:

  `configure cfm domain` *`domain_name`* `add association string` *`name`* `[vlan` *`vlan_name`*`|vman` *`vman_name`*`]`

- To add an MA to a domain using the 2-octet integer format, use the following command:

  `configure cfm domain` *`domain_name`* `add association integer` *`int`* `[vlan` *`vlan_name`*`|vman` *`vman_name`*`]`

- To add an MA to a domain using the RFC 2685 VPN ID format, use the following command:

  `configure cfm domain` *`domain_name`* `add association vpn-id oui` *`oui`* `index` *`index`* `[vlan` *`vlan_name`*`|vman` *`vman_name`*`]`

- To add an MA to a domain using the VLAN ID format, use the following command:

  `configure cfm domain` *`domain_name`* `add association vlan-id` *`vlanid`* `[vlan` *`vlan_name`*`|vman` *`vman_name`*`]`

- To delete an MA from a domain, use the following command:

  `configure cfm domain` *`domain_name`* `delete association` *`association_name`*

  In addition to supporting multicast destination MAC address for CCM and LTM frames specified by the 802.1ag standard, ExtremeXOS CFM supports the use of a unicast destination address for CCM and LTM frames.

  This allows the support of a CFM operation in a network where use of multicast address is prohibited.

- To configure the destination MAC address type for an MA, use the following command:

  ```
  configure cfm domain domain-name association association_name
  destination-mac-type [unicast | multicast]
  ```

- Use the following command to add a remote MEP to an MA statically:

  ```
  configure cfm domain domain-name association association_name add
  remote-mep mepid { mac-address mac_address }
  ```

  ExtremeXOS CFM supports configuring remote MEPs statically for CFM operation where dynamic discovery of MEPs in an MA using multicast address is prohibited.

- To delete a remote MEP from an MA, use the following command:

  ```
  configure cfm domain domain-name association association_name delete
  remote-mep mepid
  ```

- To configure a remote MEP MAC address, use the following command:

  ```
  configure cfm domain domain-name association association_name remote-
  mep mepid mac-address mac_address
  ```

*Creating MPs and the CCM Transmission Interval*

Within an MA, you configure the following MPs:

- Maintenance end points (MEPs), which are one of the following types:
  - UP MEPs—transmit CCMs, and maintain CCM database.
  - DOWN MEPs—transmit CCMs, and maintain CCM database.
- Maintenance intermediate points (MIPs)—pass CCMs through.

Each MEP must have an ID that is unique for that MEP throughout the MA.

- To configure UP and DOWN MEPs and its unique MEP ID, use the following command:

  ```
  configure cfm domain domain_name association association_name ports
  port_list add [ end-point [ up | down ] mepid group { group_name } ] |
  [ intermediate-point ]]
  ```

- To change the MEP ID on an existing MEP, use the following command:

  ```
  configure cfm domain domain_name association association_name ports
  port_list end-point [ up | down ] mepid { mepid }
  ```

- To delete UP and DOWN MEPs, use the following command:

  ```
  configure cfm domain domain_name association association_name ports
  port_list delete [[ end-point [ up | down ]] | intermediate-point ]
  ```

- To configure a MIP, use the following command:

  ```
  configure cfm domain domain_name association association_name ports
  port_list add [ end-point [ up | down ] mepid group { group_name }] |
  [ intermediate-point ]
  ```

- To delete a MIP, use the following command:

  ```
  configure cfm domain domain_name association association_name ports
  port_list delete [[ end-point [ up | down ]] | intermediate-point ]
  ```

- To configure the transmission interval for the MEP to send CCMs, use the following command:

  `configure cfm domain` *`domain_name`* **`association`** *`association_name`* **`ports`** *`port_list`* **`end-point`** [ **`up`** | **`down`** ] **`transmit-interval`** [ **`3`**| **`10`** | **`100`** | **`1000`** | **`10000`** | **`60000`** | **`600000`** ]

- To unconfigure the transmission interval for the MEP to send CCMs and return it to the default, use the following command:

  `unconfigure cfm domain` *`domain_name`* **`association`** *`association_name`* **`ports`** *`port_list`* **`end-point`** [ **`up`** | **`down`** ] **`transmit-interval`**

- To enable or disable a MEP, use the following command:

  `configure cfm domain` *`domain_name`* **`association`** *`association_name`* **`ports`** *`port_list`* **`end-point`** [ **`up`** | **`down`** ] [**`enable`** | **`disable`** ]

*Configuring EAPS for CFM Support*

#### Assigning MEP Group Names to New MEP

To assign MEP Group name when creating a MEP, use the following command:

`configure cfm domain` *`domain_name`* **`association`** *`association_name`* **`ports`** *`port_list`* `add [[` **`end-point`** [ **`up`** | **`down`** ] **`mepid group`** { *`group_name`* }] | [ **`intermediate-point`** ]]

#### Assign MEP Group Name to Existing MEP

To assign a MEP Group name to an existing MEP, use the following command:

`configure cfm domain` *`domain_name`* `association` *`association_name`* `ports` *`port_list`* `end-point [up|down] [add|delete] group` *`group_name`*

#### Add a RMEP to MEP Group

To add specific RMEPs for a MEP Group to monitor, use the following command:

`configure cfm group` *`group_name`* `[add|delete] rmep` *`mepid`*

*Monitoring CFM in EAPS*

#### Displaying MEP Groups

To display MEP groups, use the following command:

`show cfm groups {group_name}`

```
X480-48t.1 # sh cfm groups
Group : eapsCfmGrp1     Status : UP
Local MEP      : 11     port  : 41
Remote MEPs    : 10
Client(s)      : eaps
Domain         : MD1
Association    : MD1v2
Group : eapsCfmGrp2     Status : UP
Local MEP      : 12     port  : 31
Remote MEPs    : 13
Client(s)      : eaps
Domain         : MD1
Association    : MD1v2
```

*Executing Layer 2 Ping and Traceroute Messages*

If the system logs a missed CCM message, the operator can use Layer 2 ping and traceroute messages to isolate the fault. (See Ping and Traceroute for information on how each MP handles these messages.)

Note

You must have all the CFM parameters configured on your network before issuing the ping and traceroute messages.

- To send a Layer 2 ping, use the following command:

  `ping mac` *`mac`* `port` *`port`* `{domain}` *`domain_name`* `{association}` *`association_name`*

- To send a Link Trace Message (LTM) and receive information on the path, use the following command:

  `traceroute mac` *`mac`* `{up-end-point} port` *`port`* `{domain}` *`domain_name`* `{association}` *`association_name`* `{ttl` *`ttl`*`}`

## Displaying CFM

To verify your CFM configuration, you can display the current CFM configuration using the `show cfm` command.

The information this command displays includes the total ports configured for CFM, the domain names and MD levels, the MAs and associated service instances, and the UP and DOWN MEPs.

To display the CCM database for each MEP, use the `show cfm detail` command.

Counters are added for missed-hellos, which mainly help to constantly monitor the health of CFM session path and hence tweaking the configured interval times to get the best reliability with the shortest fail over times for radio and wireless networks that rely on CFM to detect any failures. These counters can be displayed using `show cfm session counters missed-hellos` command. The counters can be cleared using "clear counters cfm session missed-hellos" command.

## CFM Example

As shown in the following figure, this example assumes a simple network and assumes that CFM is configured on the access switches, as well as the necessary VMANs configured with the ports added. This example shows a VMAN associated with two maintenance domains and two different MAs. UP MEPs are configured for an MA with MD level 6 and DOWN MEPs are configured for an MA with MD level 3.

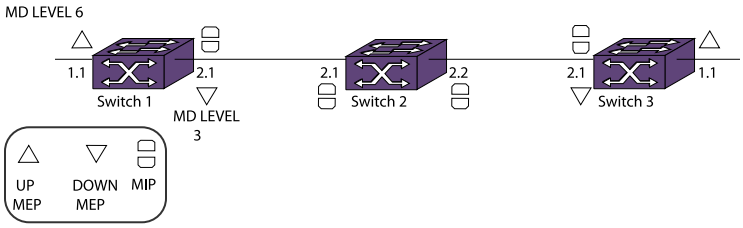



**Figure 54: CFM Configuration Example**

- Configure switch 1 for this example.

```
create cfm domain string cust-xyz-d6 md-level 6
configure cfm domain cust-xyz-d6 add association string cust-xyz-d6-m100 vman m100
configure cfm domain cust-xyz-d6 association cust-xyz-d6-m100 port 1:1 add end-point
up 10
configure cfm domain cust-xyz-d6 association cust-xyz-d6-m100 port 2:1 add
intermediate-point
create cfm domain string core-d3 md-level 3
configure cfm domain core-d3 add association string core-d3-m100 vman m100
configure cfm domain core-d3 association core-d3-m100 port 2:1 add end-point down 10
```

- Configure switch 2 for this example.

```
create cfm domain string core-d3 md-level 3
configure cfm domain core-d3 add association string core-d3-m100 vman m100
configure cfm domain core-d3 association core-d3-m100 port 2:1 add intermediate-point
configure cfm domain core-d3 association core-d3-m100 port 2:2 add intermediate-point
```

- Configure switch 3 for this example.

```
create cfm domain string cust-xyz-d6 md-level 6
configure cfm domain cust-xyz-d6 add association string cust-xyz-d6-m100 vman m100
configure cfm domain cust-xyz-d6 association cust-xyz-d6-m100 port 1:1 add end-point
up 20
configure cfm domain cust-xyz-d6 association cust-xyz-d6-m100 port 2:1 add
intermediate-point
create cfm domain string core-d3 md-level 3
configure cfm domain core-d3 add association string core-d3-m100 vman m100
configure cfm domain core-d3 association core-d3-m100 port 2:1 add end-point down 20
```

- To display the group database, use the following command:

```
show cfm groups
```

# Y.1731--Compliant Performance Monitoring

Compliant performance monitoring is based on the ITU-T Y.1731 standard and deals with the Ethernet Delay Measurement (ETH-DM) function and Ethernet Frame-Loss Measurement (ETH-LM).

## Frame-Delay Measurement

ExtremeXOS software supports:

- Two-way delay measurement—Delay Measurement Message (DMM) and Delay Measurement Reply (DMR).
- Continuous (proactive) measurement of frame delay and frame delay variation.
- On-demand measurement of frame delay and frame delay variation.

By default, DMM is not enabled. You must explicitly enable the DMM transmission for a CFM segment, either as continuous or on-demand mode.

A network interface is considered attached to a subnetwork. The term "segment" is used to refer to such a subnetwork, whether it be an Ethernet LAN, a ring, a WAN link, or even an SDH virtual circuit.

Frame-Delay measurement is done between two specific end points within an administrative domain. Frame delay and frame delay variation measurements are performed in a maintenance association end point (MEP) by sending and receiving periodic frames with ETH-DM information to and from the peer end point during the diagnostic interval.

When a CFM segment is enabled to generate frames with ETH-DM information, it periodically sends frames with ETH-DM information to its peer in the same maintenance association (MA) and expects to receive frames with ETH-DM information from its peer in the same MA.

The following list offers specific configuration information that is required by a peer to support ETH-DM:

- Maintenance domain (MD) level—The MD level at which the peer exists.
- Priority—The priority of the frames with ETH-DM information.
- Drop eligibility—Frames with ETH-DM information that are always marked as drop ineligible.
- Transmission rate.
- Total transmit interval.

A node transmits frames with ETH-DM information with the following information element:

- TxTimeStampf: Timestamp at the transmission time of the ETH-DM frame.
- RxTimeStampb: Timestamp at which the switch receives the DMR back.

  Whenever a valid DMM frame is received by the peer, a DMR frame is generated and transmitted to the requesting node.

- A DMM frame with a valid MD level and a destination MAC address equal to the receiving node's MAC address is considered to be a valid DMM frame. Every field in the DMM frame is copied to the DMR frame with the following exceptions:
  - The source and destination MAC addressed are swapped.
  - The OpCode field is changed from DMM to DMR.

The switch makes two-way frame delay variation measurements based on its ability to calculate the difference between two subsequent two-way frame delay measurements.

To allow a more precise two-way frame delay measurement, the peer replying to the frame with ETH-DM request information may include two additional timestamps in the ETH-DM reply information:

- RxTimeStampf—Timestamp at the time of receiving a frame with ETH-DM request information
- TxTimeStampb—Timestamp at the time of transmitting a frame with ETH-DM reply information

Here the frame delay is calculated by the peer that receives the DMR as follows:

- Frame Delay = (RxTimeStampb - TxTimeStampf) - (TxTimeStampb - RxTimeStampf)

The following figure describes the DMM and DMR message flows between two end points.

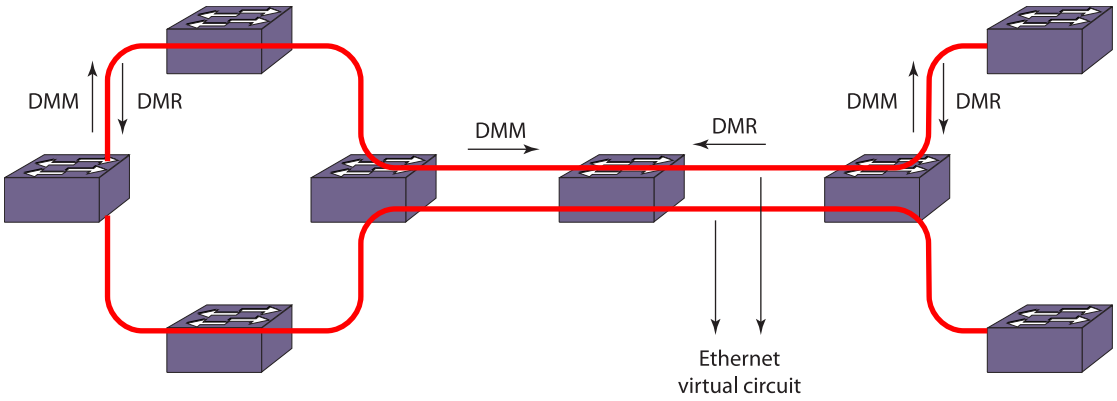



**Figure 55: Two-Way Frame Delay and Frame Delay Variance Measurement**

The PDUs used to measure frame delay and frame delay variation are the DMM and the DMR PDUs where DMM is initiated from a node as a request to its peer and DMR is the reply from the peer.

Note

When Summit X460, E4G-200 series switches are running EXOS 15.1 or later firmware, the down MEPs are performed in the hardware when configured on a normal port and the down MEPs are performed in the software when configured on a *LAG* port and Up MEPs are performed in the software for all the ports. When E4G-200 series switch running EXOS 15.1 or later firmware, the measurement (time stamping) of frame delay and loss measurements are performed in the hardware. On all other ExtremeXOS-based platforms, time stamping is always performed in the software.

If you try to enable the transmission for a CFM segment whose configuration is not complete, the trigger is rejected and an error message similar to the following is given:

```
ERROR: CFM Configuration is not complete for segment "s1" to start transmission
```

Note

A CFM segment without a domain and an association is considered to be an incomplete segment.

Upon enabling the transmission from a CFM segment, the segment transmits DMM frames, one at each transmit-interval which is configured through the CLI. If the user enables on-demand transmission, the segment transmits "X" number of DMMs and moves back to the disabled state, where "X" is the number of frames specified by the user through the CLI.

For continuous transmission, the segment continues to transmit DMM frames until stopped by the user. This transmission continues even after reboot for both continuous and on-demand mode. For on-demand transmission, the segment, which was enabled to transmit "X" number of frames, and is still transmitting, starts transmitting again "X" number of frames after reboot, or MSM failover, or process restart. The old statistics are not preserved for both continuous and on-demand mode for all the above three scenarios.

Upon transmitting a DMM, the segment is expected to get a reply from the destination within the specified time. If a reply is received after that time, that reply will be considered as a delayed one.

If a reply is not received within the transmit-interval, that is, between two subsequent DMM transmissions, then that frame is considered as lost. Once the percentage of the sum of lost and delayed

frames reaches the alarm threshold, an alarm is generated and the segment is moved to the alarming state. This state is maintained until the percentage of valid replies reaches the clear threshold. These alarm and clear states are maintained for a specified window, which holds a set of recent frames and their corresponding delays.

Various times are recorded at the segment level during the transmission of DMM frames.

- Start time—Time at which the segment started the current transmission.
- Min delay time—Time at which the minimum delay occurred in the current transmission window.
- Max delay time—Time at which the maximum delay occurred in the current transmission window.
- Alarm time—The recent alarm time, if any, during the current transmission.

The mean delay and delay variance for the current window is also measured whenever the user polls the segment statistics.

## Frame-Loss Measurement

Frame-loss is measured by sending and receiving frames with frame-loss information between peer maintenance end points (MEPs).

Frame-loss ratio is defined as a percentage of the number of service frames not delivered divided by the total number of service frames during a defined time interval, where the number of service frames not delivered is the difference between the number of service frames arriving at the ingress Ethernet flow point and the number of service frames delivered at the egress Ethernet flow point in a point-to-point Ethernet connection (see the following figure).

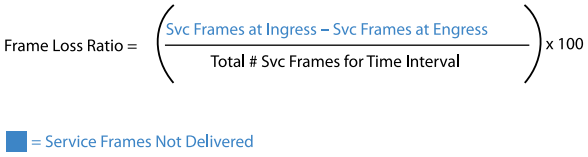



**Figure 56: Frame-Loss Ratio Formula**

To support frame-loss measurement, a MEP requires the following configuration information:

- Maintenance domain (MD) level—MD level at which the MEP exists.
- Frame-loss measurement transmission period—time interval when frame-loss measurement frames are sent.
- Priority—identifies the priority of the frames with frame-loss measurement information (configurable per operation).
- Drop eligibility—frames with frame-loss measurement information are always marked as drop ineligible (not necessarily configured).

A maintenance intermediate point (MIP) is transparent to frames with frame-loss measurement information. Therefore MIPs do not require any information to support frame-loss measurement functionality.

There are two frame-loss measurement methods:

- Dual-Ended Frame-Loss Measurement
- Single-Ended Frame-Loss Measurements

### *Dual-Ended Frame-Loss Measurement*

Dual-ended frame-loss measurement is a form of proactive OAM for performance monitoring and is useful for fault management.

**Note**

ExtremeXOS does not support dual-ended frame-loss measurement.

MEPs send periodic dual-ended frames with frame-loss measurement information to peer MEPs in a point-to-point MD. Each MEP terminates the dual-ended frames with frame-loss measurement information and makes the near-end and far-end loss measurements. Near-end frame loss refers to frame loss associated with ingress data frames, while far-end frame loss refers to frame loss associated with egress data frames. This function is used for performance monitoring at the same priority level as used for CCM.

The protocol data unit (PDU) for dual-ended frame-loss measurement information is Continuity Check Message (CCM).

### *Single-Ended Frame-Loss Measurements*

Single-ended frame-loss measurement facilitates on-demand OAM. MEPs carry out frame-loss measurments by sending frames to peer MEPs with frame-loss measurement request information and receiving frames with frame-loss measurement reply information.

The PDU for single-ended frame-loss measurement requests is Loss Measurement Message (LMM). The PDU for single-ended frame-loss measurement reply is Loss Measurement Reply (LMR). The following figure shows the transmission of LMM and LMR for frame-loss measurement.

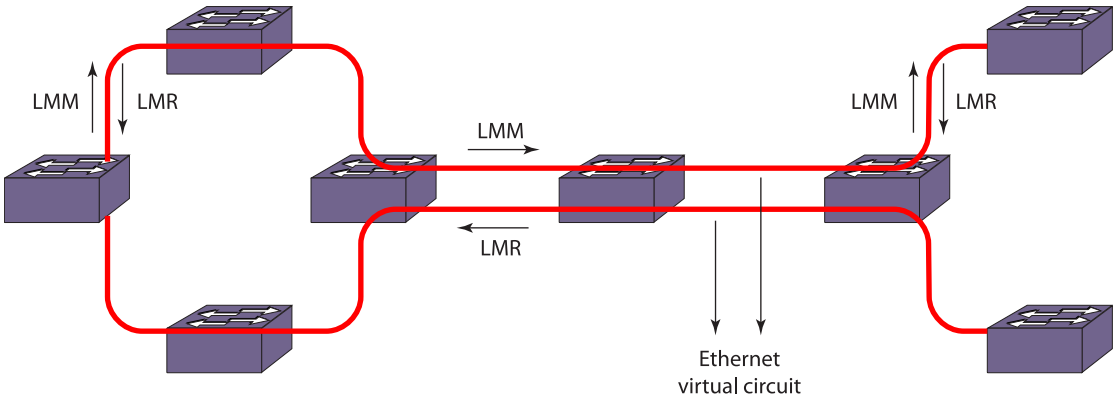



**Figure 57: Two-Way Frame-Loss Measurement**

A MEP maintains two local counters for each peer MEP it is monitoring for frame-loss:

- TxFCl—in-profile data frames transmitted to the peer MEP.
- RxFCl—in-profile data frames received from the peer MEP.

For an on-demand loss measurement, a MEP periodically transmits LMM frames with TxFCf (value of the local TxFCl counter at the time of LMM frame transmission). Upon receiving a valid LMM frame, a MEP sends an LMR frame to the requesting MEP. (Valid LMM frames have a valid MD level and a destination MAC address equal to the receiving MEP's MAC address.)

An LMR frame contains the following values:

- TxFCf—TxFCf value copied from the LMM frame.
- RxFCf—RxFCl value when the LMM frame was received.
- TxFCb—TxFCl value when the LMR frame was transmitted.

Upon receiving an LMR frame, a MEP uses the following values to make near-end and far-end loss measurements:

- Received LMR frame's TxFCf, RxFCf, and TxFCb values, and local counter RxFCl value at the time this LMR frame was received. These values are represented as TxFCf[tc], RxFCf[tc], TxFCb[tc], and RxFCl[tc]; where tc is the time the current reply frame was received.
- Previous LMR frame's TxFCf, RxFCf, and TxFCb values, and local counter RxFCl value at the time the previous LMR frame was received. These values are represented as TxFCf[tp], RxFCf[tp], TxFCb[tp], and RxFCl[tp],where tp is the time the previous reply frame was received.

Far-End Frame Loss = (TxFCf[tc] – TxFCf[tp]) – (RxFCf[tc] – RxFCf[tp])

Near-End Frame Loss = (TxFCb[tc] – TxFCb[tp]) – (RxFCl[tc] – RxFCl[tp])

*Availability Time and Severly Errored Seconds (SES)*

Frame loss is measured by sending and receiving frames with frame-loss information between peer MEPs.

Each MEP performs frame-loss measurements which contribute to unavailable time. Since a bidirectional service is defined as unavailable if either of the two directions is declared unavailable, frame-loss measurement must facilitate each MEP to perform near-end and far-end frame loss measurements.

Near-end frame loss refers to frame loss associated with ingress data frames, while far-end frame loss refers to frame loss associated with egress data frames. Both near-end and far-end frame loss measurements contribute to near-end severely errored seconds (near-end SES) and far-end severely errored seconds (far-end SES) respectively, which together contribute to unavailable time.

A period of unavailable time begins at the onset of x consecutive Severely Errored Seconds (SES) events. These x seconds are part of unavailable time. A new period of available time begins at the onset of x consecutive non-SES events. These x seconds are part of available time.

A SES is declared when, during one measurement period, the number of frames lost exceeds a threshold. ExtremeXOS logs the start and end time of the unavailable periods (see the following figure from ITU-T G.7710).

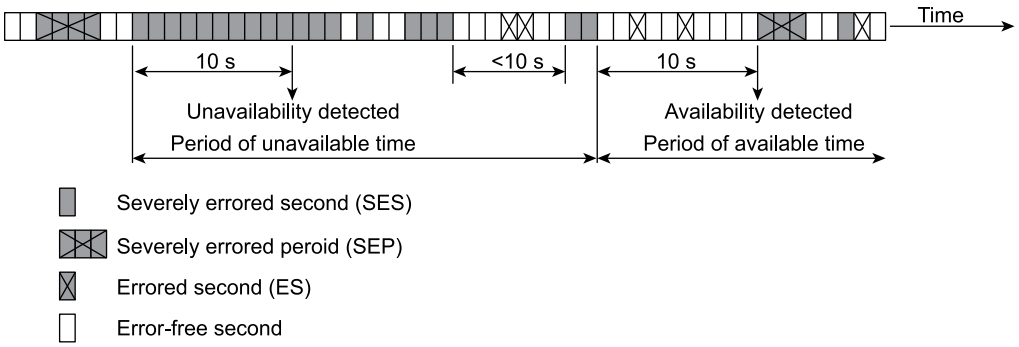



**Figure 58: SES**

## Configuring a CFM Segment

Use the following commands to configure a CFM segment.

Some of these commands are optional and, if not configured, the default values are used. The following table lists the default values for delay measurement for a CFM segment.

**Table 50: Default Values for Delay Measurement for a CFM Segment**

| Configuration | Default Values |
|---|---|
| Transmit interval | 10 seconds |
| Window | 60 frames |
| Timeout | 50 milliseconds |
| Alarm threshold | 10% |
| Clear threshold | 95% |
| Dot1p priority | 6 |

The following table lists the default values for loss measurement for a CFM segment.

**Table 51: Default Values for Loss Measurement for a CFM Segment**

| Configuration | Default Values |
|---|---|
| LMM Transmit interval | 90 seconds |
| Dot1p priority | 6 |
| Window | 1200 frames |
| SES threshold | 0.01 |
| Consecutive available count | 4 |

Note

The statistics for a particular transmission are preserved until the user triggers the transmission once again or if `clear counters cfm segment` is triggered from the CLI.

*Managing a CFM Segment*

You can create, delete, add CFM segments.

- To create a CFM segment, use the following command:

  `create cfm segment` *`segment_name`* `destination` *`mac_addr`* `{copy` *`segment_name_to_copy`*`}`

- To delete a CFM segment, use the following command:

  `delete cfm segment [`*`segment_name`* `| all]`

- To add a CFM domain to a CFM segment, use the following command:

  `configure cfm segment` *`segment_name`* `add domain` *`domain_name`* `association` *`association_name`*

- To delete a CFM domain from a CFM segment, use the following command:

  `configure cfm segment` *`segment_name`* `delete domain association`

- To configure the transmission interval between two consecutive DMM or two consecutive LMM frames, use the following command:

  `configure cfm segment` *`segment_name`* `{frame-delay | frame-loss} transmit-interval` *`interval`*

  The same transmit-interval is used for both delay and loss measurements.

- To get separate values for delay and loss measurements, use the following command:

  `configure cfm segment frame-delay/frame-loss transmit interval` *`interval`*

- To configure the dot1p priority of a DMM frame, use the following command:

  `configure cfm segment` *`segment_name`* `frame-delay dot1p` *`dot1p_priority`*

- To configure the dot1p priority of a LMM frame, use the following command:

  `configure cfm segment` *`segment_name`* `frame-loss dot1p` *`dot1p_priority`*

- To configure the dot1p priority of the CFM segment, use the following command:

  `configure cfm segment` *`segment_name`* `dot1p` *`dot1p_priority`*

  The same priority is used for both delay and loss measurements.

- To get separate values of priority for delay and loss measurements, use the following command:

  `configure cfm segment` *`segment_name`* `frame-delay dot1p` *`dot1p_priority`*

  `configure cfm segment` *`segment_name`* `frame-loss dot1p` *`dot1p_priority`*

- To configure the alarm and clear threshold value for CFM segment, use the following command:

  `configure cfm segment` *`segment_name`* `[alarm-threshold | clear-threshold]` *`value`*

- To configure the window size to be used for calculating the threshold values, use the following command:

  `configure cfm segment` *`segment_name`* `window` *`size`*

  The same window size is used for both delay and loss measurements.

- To get separate values of window size for delay and loss measurements, use the following:

  `configure cfm segment` *`segment_name`* `frame-loss window` *`window_size`*

```
configure cfm segment segment_name frame-delay window window_size
```

- To configure the window size of a DMM frame to be used for calculating the threshold values, use the following command:

  ```
  configure cfm segment segment_name frame-delay window window_size
  ```

- To configure the window size of a LMM frame to be used for calculating the threshold values, use the following command:

  ```
  configure cfm segment segment_name frame-loss window window_size
  ```

- To trigger DMM frames at the specified transmit interval, use the following command:

  ```
  enable cfm segment frame-delay measurement segment_name mep mep_id
  [continuous | count ] value
  ```

- To disable the transmission of the DMM frames for a particular CFM segment, use the following command:

  ```
  disable cfm segment frame-delay measurement segment_name mep mep_id
  ```

- To show the configuration and status of a specific CFM segment, use the following command:

  ```
  show cfm segment {segment_name}
  ```

- To show the configuration and status of a specific CFM segment doing delay measurement, use the following command:

  ```
  show cfm segment frame-delay {segment_name]}
  ```

- To show the configuration and status of a specific CFM segment doing loss measurement, use the following command:

  ```
  show cfm segment frame-loss {segment_name]}
  ```

- To display the frame delay statistics for the CFM segment, use the following command:

  ```
  show cfm segment frame-delay statistics {segment-name}
  ```

- To configure the timeout value for a CFM segment, use the following command:

  ```
  configure cfm segment segment_name timeout msec
  ```

- To add or delete the local MEP for a given CFM segment, use the following command:

  ```
  configure cfm segment segment_name frame-loss [add|delete] mep mep_id
  ```

- To set the percentage of frames lost in a measurement period so that it will be marked as SES (severely errored second), use the following command:

  ```
  configure cfm segment segment_name frame-loss ses-threshold percent
  ```

- To set the number of consecutive measurements used to determine the availability status of a CFM segment, use the following command:

  ```
  configure cfm segment segment_name frame-loss consecutive frames
  ```

- To start the transmission of LMM frames for the set transmit interval, use the following command:

  ```
  enable cfm segment frame-loss measurement segment_name mep mep_id
  [continuous | count frames]
  ```

  Note

  For the above command, If the the segment is not completely configured, frames are not transmitted and an error occurs.

- To stop the transmission of the LMM frames for a particular CFM segment, use the following command:

  `disable cfm segment frame-loss measurement segment_name mep mep_id`

- To display the frame loss or frame delay statistics for the CFM segment, use the following command:

  `show cfm segment {{segment_name} | {frame-delay {segment_name}} | {frame-loss {segment_name {mep mep_id}}}}`

Note

Frame-loss measurements are not supported on platforms where the *VLAN* packet statistics are not retrieved, and on up-meps.

### Clearing CFM Information

- To clear cfm segment counters, use the following commands:

  `clear counters cfm segment segment_name`

  `clear counters cfm segment all`

- To clear cfm segment counters specific to DMM, use the following command:

  `clear counters cfm segment segment_name frame-delay`

- To clear cfm segment counters specific to LMM, use the following commands:

  `clear counters cfm segment segment_name frame-loss`

  `clear counters cfm segment segment_name frame-loss mep mep_id`

## Y.1731 MIB Support

ExtremeXOS 15.5 supports Y.1731 performance measurement MIB defined by MEF - 36. The performance monitoring process is made up of a number of performance monitoring instances, known as performance monitoring (PM) sessions. A PM session can be initiated between two MEPs in a MEG and be defined as either a loss measurement (LM) session or delay measurement (DM) session. The LM session can be used to determine the performance metrics frame loss ratio, availability, and resiliency. The DM session can be used to determine the performance metrics Frame Delay.

The MIB is divided into a number of different object groupings: the PM MIB MEP objects, PM MIB loss measurement objects, PM MIB delay measurement objects, and SOAM PM notifications. The initial implementation of MIB supports GET operations for Frame Loss& Frame Delay.

### MIB Specific Data

- A measurement interval of 15 min to be supported
- Default message period/transmit-interval of LMMs is 1 sec (Min = 1sec in current CLI) * Default message period/transmit-interval of DMMs is 100msec (Min = 1 sec in current CLI)
- Repetition Time can be set to 0 which means that there is no gap between measurement intervals
- Number of History measurement intervals can be 2-1000, though it expects at least 32 measurement intervals to be stored and 96 are recommended.

- Both DM and LM sessions are MEP to MEP sessions. The index of all the DM/LM tables includes MD, MA, MEP table indices as well as the particular DM/LM session.
- Currently DM sessions are not MEP-to MEP based but only segment based sessions. To support DM tables in the MIB, changes are required in the current CLI & backend delay implementation.

## Limitations

- Currently we are storing a maximum of 1800 frames data for each LMM/DMM session. But to support at least 2 history and 1 current measurement intervals, we need to store 2700 frames (if message period is 1 sec, Repetition time is 0, measurement interval is 15 min) for each delay/loss session.
- Each frame's data is about 60 bytes for LMM and which takes about 44 MB of memory for 288 sessions

# EFM OAM--Unidirectional Link Fault Management

## Unidirectional Link Fault Management

With EFM OAM, certain physical layers can support a limited unidirectional capability.

The ability to operate a link in a unidirectional mode for diagnostic purposes supports the maintenance objective of failure detection and notification. Unidirectional OAM operation is not supported on some legacy links but is supported on newer links such as 100BASE-X PCS, 1000BASE-X PCS, and 10GbE RS. On technologies that support the feature, OAM PDUs can be transmitted across unidirectional links to indicate fault information. To the higher layers, the link is still failed in both directions, but to the OAM layer, some communication capabilities exist. The distinction between a unidirectional link and a normal link is shown in the following figure.

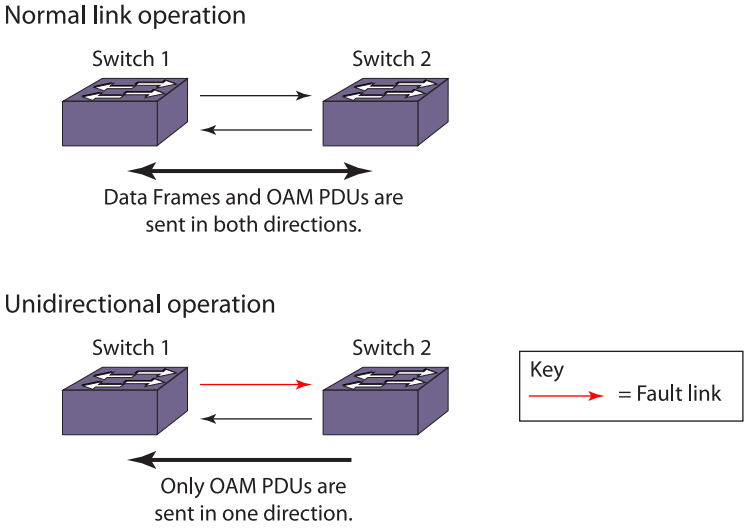



**Figure 59: Normal Link and Unidirectional Operation**

You can enable unidirectional link fault detection and notification on individual ports with CLI commands. This allows appropriate register settings to transmit OAM PDUs even on a link that has a slowly deteriorating quality receive path or no receive path at all. Then, when a link is not receiving a

signal from its peer at the physical layer (for example, if the peer's laser is malfunctioning), the local entity can set a flag in an OAM PDU to let the peer know that its transmit path is inoperable.

The operation of OAM on an Ethernet interface does not adversely affect data traffic because OAM is a slow protocol with very limited bandwidth potential, and it is not required for normal link operation. By utilizing the slow protocol MAC address, OAM frames are intercepted by the MAC sub layer and cannot propagate across multiple hops in an Ethernet network. This implementation assures that OAM PDUs affect only the operation of the OAM protocol itself and not user data traffic.

The IEEE 802.3ah standard defines fault notifications based on one-second timers. But by sending triggered OAM PDUs on detecting link down/local fault rather that waiting to send on periodic PDUs, failure detection is less than one second can be achieved, thereby accelerating fault recovery and network restoration.

EFM OAM uses standard length Ethernet frames within the normal frame length of 64 to 1518 bytes as PDUs for their operation. The following table describes the fields of OAM PDUs.

**Table 52: OAM PDU Fields**

| Field | Octets | Description | Value |
|---|---|---|---|
| Destination Address | 6 | Slow protocol multicast address | 01:80:C2:00:00:02 |
| Source Address | 6 | Port's individual MAC address | Switch MAC |
| Length/Type | 2 | Slow protocol type | 0x8809 |
| Subtype | 1 | Identifies specific slow protocol | 0x03 |
| Flags | 2 | Contains status bits | see the following figure |
| Code | 1 | Identifies OAM PDU type | 0x00 (Information TLV) |
| Data/Pad | 42-1496 | OAM PDU data | 0x00 (END of TLV) |
| FCS | 4 | Frame check sequence | |

## Configuring Unidirectional Link Fault Management

To configure unidirectional link fault management on a port or ports, use the following command:

```
enable ethernet oam ports [port_list | all] link-fault-management
```

To clear the counters on a configured port, use the following command:

```
clear ethernet oam {ports [port_list} counters
```

To unconfigure unidirectional link fault management, use the following command:

```
disable ethernet oam ports [port_list | all] link-fault-management
```

To display the Ethernet OAM settings, use the following command:

```
show ethernet oam {ports [port_list} {detail}
```

When configured, the following behavior on the port is observed:

- A log indicates that traffic on the port is blocked.

- All received traffic on that port is blocked except for Ethernet OAM PDUs.
- To higher layers, a failure is reported as a link down but OAM can use the link to send OAM traffic.

## Two-Way Active Measurement Protocol

The Two-Way Active Measurement Protocol, defined in RFC 5357, specifies an industry standard for measuring round-trip performance between two devices that support the TWAMP protocols. TWAMP defines two protocols; the TWAMP-Control protocol and the TWAMP-Test protocol. The TWAMP-Control protocol is used to setup test sessions. The test sessions use the TWAMP-Test protocol to transmit and reflect performance measurement packets. The TWAMP-Control protocol utilizes TCP for communication, while the TWAMP-Test protocol utilizes UDP.

Note
For ExtremeXOS 16.1, only the TWAMP-Test protocol is supported.

The test sessions are setup via the ‘Request-Session’ command message, sent from the Control-Client to the Server. The Server replies with an ‘Accept Session’ message, which indicates if the Server is capable of accommodating the request or not. The Control-Client may send several ‘Request-Session’ command messages to setup multiple test sessions. To begin the tests, the Control-Client transmits a ‘Start-Session’ command message. The Server replies with a ‘Start-Ack’ message. The Control-Client does not begin its test until it receives the ‘Start-Ack’ message. This allows the Server ample time to configure the test sessions. The Control-Client will stop the test sessions with a ‘Stop-Sessions’ message. The Server does not respond to this message.

The ExtremeXOS 16.1 implementation of TWAMP consists of the development of the TWAMP logical role of the Session-Reflector. The user is required to configure endpoints, which define the destination of TWAMP-Test packets generated by the client. An endpoint receiving a new TWAMP-test packet creates a test session consisting of the following four-part tuple; client IP address, client UDP port, endpoint IP address, and endpoint UDP port. The tuple does not include the VR because it requires the default VR for the first phase. A session timeout value, configured globally, determines the amount of time test sessions exist after the last reception of a TWAMP-Test packet. Test sessions are used to keep track of the session data, such as the sequence number. The Session-Reflector will still respond to TWAMP-Test packets that do not match an existing test session or if a new test session cannot be created due to lack of resources.

### TWAMP-Test Protocol

The TWAMP logical roles of the Session-Sender and Session-Reflector use the TWAMP-Test protocol. The Session-Sender, controlled by the Client, transmits TWAMP-Test packets to the Session-Reflector. The Session-Reflector processes the packet, copies the fields from that packet into the reply, and then transmits a reply packet back to the Session-Sender. The reply packet is larger than the request packet unless the Clients adds padding.

The implementation of the Session-Reflector relies on the user to create endpoints, which open up available ports for reflecting TWAMP-Test packets. The endpoint configuration acts as a control list to limit the number of locally opened sockets. The user may choose to use authentication or encryption to protect the transmitted data from nefarious entities. The shared keys used for the TWAMP-Test derive from the TWAMP 3-way handshake that establishes the control session. Both devices must configure matching shared keys.

# Bidirectional Forwarding Detection (BFD)

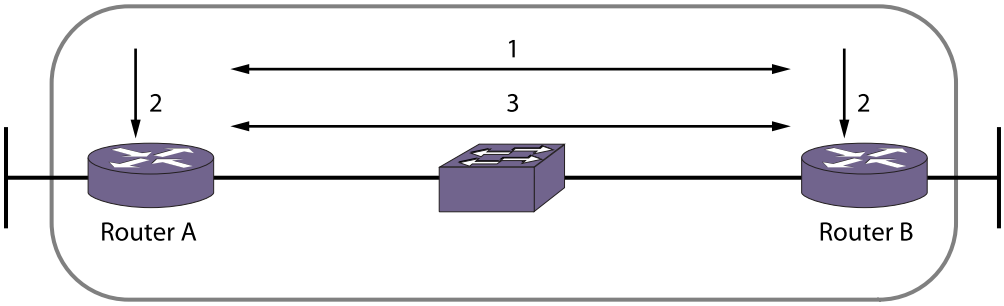
## BFD Overview

Bidirectional Forwarding Detection (BFD) is a hello protocol that provides the rapid detection of failures in the path and informs the clients (routing protocols) to initiate the route convergence.

It is independent of media, routing protocols, and data protocols. BFD helps in the separation of forwarding plane connectivity and control plane connectivity.

Different routing protocol hello mechanisms operate in variable rates of detection, but BFD detects the forwarding path failures at a uniform rate, thus allowing for easier network profiling and planning, and consistent and predictable re-convergence time.

The following figure shows a BFD topology.




1. Routing protocol (BFD client) discovers neighbor
2. Routing protocol informs BFD to create session with neighbor
3. BFD establishes session with neighbor

**Figure 60: BFD Topology**

The routing protocols first learn the neighbor and make entries in the forwarding table. Then protocols can register the neighbor address with BFD and ask to monitor the status of the path. BFD establishes the session with a remote BFD and monitors the path status.

You can configure detection multipliers and TX and RX intervals on a directly connected interface (*VLAN*).

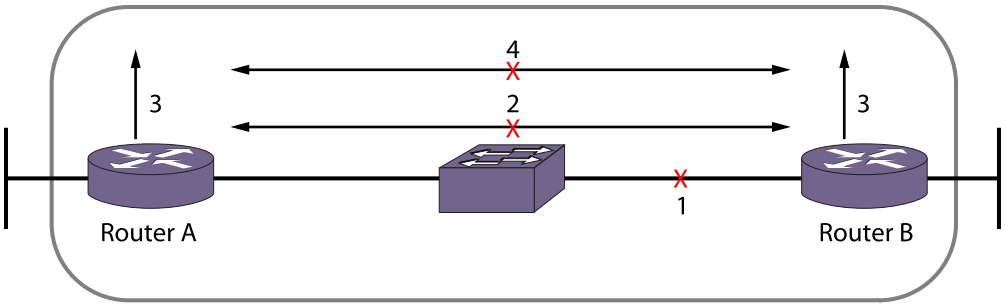
- The detection multiplier signifies the number of BFD packets the BFD server waits for after which a timeout is declared.
- The receive interval is the interval at which the BFD server is ready to receive packets.
- The transmit interval is the interval at which the BFD server is ready to transmit packets.

For example, when two nodes, A and B, initiate a BFD session between them, a negotiation about the receive and transmit intervals occurs.

The receive interval of node A is calculated as the maximum of the configured receive interval of node A and the configured transmit interval of node B. The same applies to node B.

If multiple clients ask for the same neighbor on the same interface, then a single BFD session is established between the peers.

The following figure shows the behavior when a failure occurs.




1. Link fails
2. BFD detects failure
3. BFD reports failure to routing protocol
4. Routing protocol takes necessary action for failure

**Figure 61: BFD Failure Detection**

BFD detects the failure first and then informs the registered clients about the neighbors.

BFD operates in an asynchronous mode in which systems periodically send BFD control packets to one another. If a number of those packets in a row are not received by the other system, the session is declared to be down.

Simple password authentication can be included in the control packet to avoid spoofing.

This feature is available on all platforms.

**Note**
BFD can be used to protect IPv4 & IPv6 static routes, OSPFv2 & *OSPFv3 (Open Shortest Path First version 3)* interfaces and *MPLS (Multiprotocol Label Switching)* interfaces. For more information, see Configuring Static Routes on page 1266, BFD for OSPF on page 1342, or refer to Managing the MPLS BFD Client on page 1212.

## Limitations

The following limitations apply to BFD in this release:

- Direct connection (single hop) networks only are supported.
- *OSPF (Open Shortest Path First)*, *MPLS* and static routes act as BFD clients.
- Hitless failover is supported.
- The echo function is not supported.
- BFD protocol has been implemented in software. The number of sessions handled by BFD at minimal timers (less than 100ms) varies depending on platform type and processing load (which is effected by other protocols being enabled, or other system conditions such as software forwarding).
- Hardware assist BFD is supported on standalone X460-G2 platforms and homogeneous X460-G2 Summit Stacking only.
- The following scaling limits apply to BFD hardware assist:
  - Maximum of 2047 BFD sessions.

- Maximum of 256 sessins with 3.3 ms transmit interval.
- Maximum if 800 sessions with 15 ms transmit interval.

## Configuring BFD

You can enable, disable, configure, and unconfigure BFD.

- To enable or disable BFD, use the following command:

  `[enable | disable] bfd vlan vlan_name`

- To configure the detection multipliers and TX and RX intervals, use the following command:

  `configure bfd vlan vlan_name [{detection-multiplier multiplier} {receive-interval rx_interval} {transmit-interval tx_interval}]`

- To specify either authentication using a simple password or no authentication, use the following command:

  `configure bfd vlan vlan_name authentication [none | simple-password {encrypted} password]]`

- To unconfigure BFD, use the following command:

  `unconfigure bfd vlan vlan_name`

## Displaying BFD Information

The following commands display information regarding BFD configuration and process.

- To display information on BFD sessions, use the following command:

  `show bfd`

- To display information on BFD global counters, use the following command:

  `show bfd counters`

- To display information on BFD session counters, use the following command:

  `show bfd session counters vr all`

- To display the configuration of a specific interface or those specific counters, use the following command:

  `show bfd vlan {vlan_name}`

- To display the counters of a specific interface, use the following command:

  `show bfd vlan {vlan_name} counters`

- To display the session status of a particular client, use the following command:

  `show bfd session client [mpls | ospf {ipv4 | ipv6} | static {ipv4 | ipv6}] {vr [vrname | all]}`

- To display the session status information for all VRs, use the following command:

  `show bfd session vr all`

- To display session status information in detail for all VRs, use the following command:

  `show bfd session {ipv4| ipv6} detail vr all`

## Clearing BFD Information

To clear global, session, or interface counters, use the following command:

```
clear counters bfd {session | interface}
```

## BFD MIB Table Support

ExtremeXOS Release 15.5 supports read-only for all BFD MIB tables, global objects, and supports BFD notifications as well. BFD-MIB implementation is based on draft-ietf-bfd-mib-14, and draft-ietf-bfd-tc-mib-02. Currently, the BFD MIB is kept under the enterprise MIB in EXOS implementation.

The SET operation is supported only for MIB object 'bfdSessNotificationsEnable' (to control up/down traps). The default value for this object is disabled state. No notification is sent in disabled state. Thus, the SET operation is also supported for this MIB object in order to control the emission of traps.

### *BFD Session Up/Down Traps*

BFD has two traps, one for notifying that the session moved to the UP state, and the other trap for notifying that the session moved to DOWN state. To reduce the number of traps sent to NMS, a single trap is generated to combine the status changes of multiple sessions if the sessions have contiguous session IDs and multiple sessions move to either the UP or DOWN state in the same window of time. However, status changes of different types (UP & DOWN), will not be mixed in single trap. The window of time to combine the traps can be configured using the CLI command `configure snmp traps batch-delay bfd`.

For example, if sessions with session IDs 1, 2, 3, 4, and 5 are moving to the UP state in the same window of time, then a single trap is sent with low range index 1 and high range index 5. As a second example, after all sessions moved to the UP state, session ID 2 goes DOWN and comes back UP before generating the first trap. In this case also, the first trap which is the UP trap, is set to include all sessions. Then, the second trap would be the DOWN trap for session ID 2, and finally the third trap would be the UP trap again for session ID 2. Thus, events are not missed or reordered.

NMS relates traps to sessions using only the session index which is provided in traps. It is necessary that the session index does not change until NMS retrieves session details via GET requests. To achieve this, the session will be retained for fifteen minutes after deletion is initiated by the BFD client (control protocol). During this period transmission and reception of BFD control packets will be stopped. If BFD protection is requested for the same destination again within this period, the same session index is reused. With this change, NMS can also have good history of the session to a particular destination.

#### Configuring SNMP Traps for BFD

To enable snmp traps for bfd:

```
enable snmp traps bfd {session-down | session-up}
```

To disable snmp traps for bfd:

```
disable snmp traps bfd {session-down | session-up}
```

To configure batch delay for sending the traps:

```
configure snmp traps batch-delay bfd {none | delay}
```

To display the configuration:

```
show snmp traps bfd
```

Note

*SNMP (Simple Network Management Protocol)* traps for BFD are disabled by default for both session-down and session-up.

**Configuration Example**


**Figure 62: BFD Configuration Example**

Consider the network segment like above, wherein two routers R1 and R2 are connected via an L2 switch. Following is the list of commands to configure BFD protection for static routes.

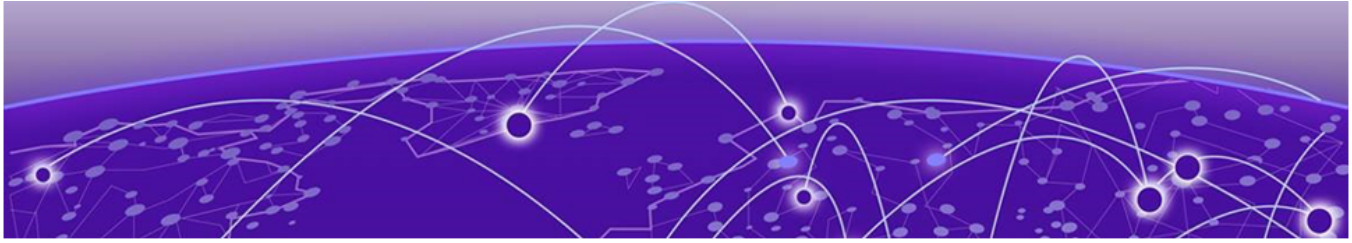
Router R1:

1. Create vlan and configure IP address.

   ```
   create vlan v1 tag 100

   configure vlan v1 add port 2 tagged

   configure vlan v1 ipaddress 10.0.0.1/24 2
   ```

2. Create BFD session to the next-hop which is being monitored.

   ```
   enable iproute bfd 10.0.0.2 vr VR-Default
   ```



# PoE



Power over Ethernet (PoE) is an effective method of supplying 48 VDC power to certain types of powered devices (PDs) through Category 5, Category 5E and Category 6 twisted pair Ethernet cables.

PDs include wireless access points, IP telephones, laptop computers, web cameras, and other devices. With PoE, a single Ethernet cable supplies power and the data connection, reducing costs associated with separate power cabling and supply.

The system supports hitless failover for PoE in a system with two Management Switch Fabric Modules (MSMs). Hitless failover means that if the primary MSM fails over to the backup MSM, all port currently powered will maintain power after the failover and all the power configurations remain active.

Similar failover support is available for a SummitStack. In a SummitStack, power is maintained across a failover on all PoE ports of non-primary nodes but is lost on all PoE ports of the failed primary node. Each Summit switch has its own PSU and the power budget for each Summit switch is determined by the internal/external PSUs connected to that Summit switch.

PoE+ supports higher power levels as defined by the IEEE 802.3at standard.

## Extreme Networks PoE Devices

Following is a list of the Extreme Networks devices that support *PoE (Power over Ethernet)* and the minimum required software:

- G48Tc module (with daughter card) for the BlackDiamond 8800 series switch—ExtremeXOS 12.1 and later
- G48Te2 module (with daughter card) for the BlackDiamond 8800 series switch—ExtremeXOS 12.1 and later
- 8900-G48T-xl module (with daughter card) for the BlackDiamond 8800 series switch—ExtremeXOS 12.4 and later

Following is a list of the Extreme Networks devices that support PoE+ and the minimum required software:

- Summit X430-8p switch—ExtremeXOS 15.5.2 and later
- Summit X430-24p switch -ExtremeXOS 15.5.2 and later
- Summit X440-8p switch—ExtremeXOS 15.1 and later
- Summit X440-24p switch—ExtremeXOS 15.1 and later
- Summit X440-48p switch—ExtremeXOS 15.1.2 and later
- Summit X450-G2-24p-GE4—ExtremeXOS 16.1 and later
- Summit X450-G2-48p-GE4—ExtremeXOS 16.1 and later
- Summit X450-G2-24p-10GE4—ExtremeXOS 16.1 and later
- Summit X450-G2-48p-10GE4—ExtremeXOS 16.1 and later
- Summit X460-24p switch—ExtremeXOS 12.5 and later
- Summit X460-48p switch—ExtremeXOS 12.5 and later
- Summit X460-G2-24p-10GE4—ExtremeXOS 15.6 and later
- Summit X460-G2-48p-10GE4—ExtremeXOS 15.6 and later
- Summit X460-G2-24p-GE4—ExtremeXOS 15.6 and later
- Summit X460-G2-48p-GE4—ExtremeXOS 15.6 and later

Note

PoE capability for the G48Tc and G48Te2 modules are available only with the addition of an optional PoE Daughter Module. See Adding an S-PoE Daughter Card to an Existing Configuration for more information.

## Summary of PoE Features

The ExtremeXOS implementation of *PoE* supports the following features:

- Configuration and control of the power distribution for PoE at the system, slot, and port levels
- Real-time discovery and classification of IEEE 802.3af-compliant PDs and many legacy devices
- Support for IEEE 802.3at-compliant PDs on PoE+ devices
- Monitor and control of port PoE fault conditions including exceeding configured class limits and power limits and short-circuit detection
- Support for configuring and monitoring PoE status at the system, slot, and port levels
- Management of an over-subscribed power budget
- Port LED control for indicating the link state
- Support for hitless failover in a chassis with two MSMs

For detailed information on using the PoE commands to configure, manage, and display PoE settings, refer to the section on PoE on page 416.

## Power Checking for PoE Module

*PoE* modules require more power than other I/O modules.

When a chassis containing a PoE module is booted or a new PoE module is inserted, the power drain is calculated. Before the PoE module is powered up, the chassis calculates the power budget and powers up the PoE module only if there is enough power. The chassis powers up as many I/O modules as possible with lower-numbered slots having priority.

Note

If your chassis has an inline power module and there is not enough power to supply the configured inline power for the slot, that slot will not power on; the slot will not function in data-only mode without enough power for inline power.

If a PoE module is inserted into a chassis, the chassis calculates the power budget and powers up the PoE module only if there is enough power. Installed modules are not affected. However, if you reboot the chassis, power checking proceeds as described in the previous paragraph. If there is now enough power, I/O modules that were not powered up previously are powered up.

If you lose power or the overall available power decreases, the system removes power to the I/O modules beginning with the highest numbered slots until enough power is available. Inline power reserved for a slot that is not used cannot be used by other PoE slots (inline power is not shared among PoE modules).

Before you install your PoE module, consult your sales team to determine the required power budget.

## Power Delivery

This section describes how the system provides power to the PDs.

Note

In Summit X440 (*PoE*-capable) switches, it is recommended that the power budget should not be increased abruptly by a large extent. When the power budget is suddenly increased by a large value, it exceeds the capabilities of the PSU, and the inline power state of the ports are disabled. If the inline power state of the ports is disabled due to increasing the load abruptly, a reboot should be done to recover from the situation.

### Enabling PoE to the Switch

*PoE* is enabled by default. Refer to Configure PoE for details about changing the configuration.

### Power Reserve Budget

*Summit X460-24p and X460-48p Switches Only*

Summit X460-24p and X460-48p switches have two removable internal PSUs, each capable of delivering 380 W of power.

When two PSUs are present, the total power budget is 760 W and PSU load-sharing is in effect. If one PSU fails or is removed then the power budget will drop to 380 W and port priority will be used to determine which ports remain powered up if usage was more than 380 W before the event.

*Summit X430-8p Switches Only*

The Summit X430-8p switches have one internal PSU capable of delivering 90 W of power (in standalone desktop configuration).

Note
The power budget is set to 60W during startup.

*Summit X430-24p Switches Only*

The Summit X430-24p have one internal PSU capable of delivering 370W of power.

*Summit X440-24p and X440-48p Switches Only*

The Summit X440-24p switches have one internal PSU capable of delivering 380 W of power.

*Summit X440-8p Switches Only*

The Summit X440-8p switches have one internal PSU capable of delivering 170 W of power.

*Summit X4560-G2 Switches Only*

The Summit X450-G2 non-PoE+ models support a fixed internal 156W AC/DC PSU, whereas X450-G2 PoE+ models support dual, hot-swappable 1100W/715W/350W AC/DC PSUs.

*Summit X460-G2 Switches Only*

The Summit X460-G2 supports two PSUs : 1100W and 715W. When 1100W is used, the switch is capable of delivering 850W of power. When 715W is used, the switch is capable of delivering 465W of power.

*Modular Switches Only*

On modular *PoE* switches, the power budget is provided on a per slot basis, not switchwide.

You can reserve power for each slot, or PoE module. Power reserved for a specific PoE module cannot be used by any other slot regardless of how much power is actually consumed on the specified slot. The default power budget reserved for each PoE module is 50 W. The minimum power you can assign to a slot is 37 W, or 0 W if the slot is disabled. The maximum possible for each slot is 768 W.

Note
We recommend that when using a modular switch you fully populate a single PoE module with PDs until the power usage is just below the usage threshold, instead of spacing PDs evenly across PoE modules.

If you disable a slot with a PoE module, the reserved power budget remains with that slot until you unconfigure or reconfigure the power budget. Also, you can reconfigure the reserved power budget for a PoE module without disabling the device first; you can also reconfigure dynamically. These settings are preserved across reboots and other power-cycling conditions.

The total of all reserved slot power budgets cannot be larger than the total available power to the switch. If the base module power requirements plus the reserved PoE power for all modules exceeds the

unallocated power in the system, the lowest numbered slots have priority in getting power and one or more modules in higher-numbered slots will be powered down.

Note

On modular switches, PoE modules are not powered-up at all, even in data-only mode, if the reserved PoE power cannot be allocated to that slot.

*Guard Band*

To reduce the chances of ports fluctuating between powered and non-powered states, newly inserted PDs are not powered when the actual delivered power for the module or switch is within a preset value below the configured inline power budget for that slot or switch.

This band is called the guard band and the value used is 20 W for Summit X430, X440-24p, X460-24p, X460-48p, X460-G2, X440-8p, and X440-48p switches. However, actual aggregate power can be delivered up to the configured inline power budget for the slot or switch (for example, when delivered power from ports increases or when the configured inline power budget for the slot is reduced).

If total consumed power is within the guard band, new ports are not powered up.

## PD Disconnect Precedence

*Summit X430, X440-24p, X460-24p, X460-48p, X460-G2, X440-8p, X440-48p and Modular PoE Switches Only*

After a PD is discovered and powered on a Summit X440-24p, X460-24p, X460-48p or a modular *PoE* switch, the actual power drain is continuously measured.

If the usage for power by PDs is within the guard band, the system begins denying power to PDs.

To supply power to all PDs on a modular switch, you can reconfigure the reserved power budget for the switch or slot, so that enough power is available to power all PDs. You reconfigure the reserved power budget dynamically; you do not have to disable the device to reconfigure the power budget.

You can configure the switch to handle a request for power that exceeds the power budget situation in one of two ways, called the disconnect precedence:

- Disconnect PDs according to the configured PoE port priority for each PD.
- Deny power to the next PD requesting power, regardless of that port's PoE priority.

On modular switches, this is a switchwide configuration that applies to each slot; you cannot configure this disconnect precedence per slot.

The default value is deny-port. So, if you do not change the default value and the switch's or slot's power is exceeded, the next PD requesting power is not connected (even if that port has a higher configured PoE port priority than those ports already receiving power). When you configure the deny-port value, the switch disregards the configured PoE port priority and port numbering.

When the switch is configured for lowest-priority mode, PDs are denied power based on the individual port's configured PoE priority. If the next PD requesting power is of a higher configured PoE priority than an already powered port, the lower-priority port is disconnected and the higher-priority port is powered.

## Port Disconnect or Fault

On modular *PoE* switches, when a port is disconnected, the power is removed from that port and can be used only by ports on the same slot. The power from the disconnected port is not redistributed to any other slot.

On all PoE devices, when a port enters a fault state because of a class violation or if you set the operator limit lower than the amount requested by the PD, the system removes power from that port. The power removed is, again, available only to other ports on the same slot or stand-alone switch; it cannot be redistributed to other slots on modular switches. The port stays in the fault state until you disable that port, disconnect the attached PD, or reconfigure the operator limit to be high enough to satisfy the PD requirements.

When a port is disconnected or otherwise moves into a fault state, *SNMP (Simple Network Management Protocol)* generates an event (after you configure SNMP and a log message is created).

## Port Power Cycling

You can set ports to experience a power-down, discover, or power-up cycle.

On the Summit X430, X440-24p, X440-48P, X440-8P, X460-24p, X460-48p, X460-G2, and modular *PoE* switches, this power-cycling occurs without returning the power to the slot's reserved power budget. This function allows you to reset PDs without losing their claim to the reserved power budget.

Ports are immediately depowered and repowered, maintaining current power allocations on modular switches.

## PoE Usage Threshold

The system generates an *SNMP* event when any slot or stand-alone switch has consumed a specified percentage of that slot's reserved power budget or of the entire power for the stand-alone switch.

The default value is 70%; you can configure this threshold to generate events from 1% to 99% consumption of the reserved power budget. You can also configure the system to log an *EMS (Event Management System)* message when the usage threshold is crossed (refer to Status Monitoring and Statistics for more information on EMS). On modular switches, this threshold percentage is set to be the same for each *PoE* slot; you cannot configure it differently for each PoE module.

On modular switches, although the threshold percentage of measured to budgeted power applies to all PoE modules, the threshold measurement applies only to the percentage per slot of measured power to budgeted power use; it does not apply to the amount of power used switchwide.

## Legacy Devices

ExtremeXOS software allows the use of non-standard PDs with the switch. These are PDs that do not comply with the IEEE 802.3af standard.

The system detects non-standard PDs using a capacitance measurement. You must enable the switch to detect legacy devices; the default value is disabled. You configure the detection of legacy *PoE* devices per slot.

Detecting a PD through capacitance is used only if the following two conditions are both met:

- Legacy PD detection is enabled.
- The system unsuccessfully attempted to discover the PD using the standard resistance measurement method.

## PoE Operator Limits

You can set the power limit that a PD can draw on the specified ports. For *PoE*, the range is 3000 to 16800 mW, and the default value is 15400 mW. For PoE+, the range is 3000 to 30000 mW, and the default value is 30000 mW.

If the measured power for a specified port exceeds the port's operator limit, the power is withdrawn from that port and the port moves into a fault state.

If you attempt to set an operator limit outside the accepted range, the system returns an error message.

# Configuring PoE

*PoE* supports a full set of configuration and monitoring commands that allows you to configure, manage, and display PoE settings at the system, slot, and port level. To enable inline power, or PoE, you must have a powered switch or chassis and module.

Note

On a modular switch, if your chassis has an inline power module and there is not enough power to supply a slot, that slot will not power on; the slot will not function in data-only mode without enough power for inline power.

To configure inline power, or PoE, you must accomplish the following tasks:

1. Enable inline power to the system, slot, and/or port.
2. On modular switches, reserve power to the switch or slot using a power budget.
3. On modular switches and Summit X430, X440-24p, X450-G2, X460-G2, X460-24p, and X460-48p switches, configure the disconnect precedence for the PDs in the case of excessive power demands.
4. Configure the threshold for initiating system alarms on power usage.

Additionally, you can configure the switch to use legacy PDs, apply specified PoE limits to ports, apply labels to PoE ports, and configure the switch to allow you to reset a PD without losing its power allocation.

Refer to the *ExtremeXOS 16.2 Command Reference Guide* for complete information on using the CLI commands.

## Enable Inline Power

- To enable inline power to the switch, slot, or port, use the following commands:

```
enable inline-power

enable inline-power slot slot
```

```
enable inline-power ports [all | port_list]
```

**Note**

On modular switches, if your chassis has an inline power module and there is not enough power to supply a slot, that slot will not power on; the slot will not function in data-only mode without enough power for inline power.

- To disable inline power to the switch, slot (on modular switches), or port, use the following commands:

  ```
  disable inline-power
  ```

  ```
  disable inline-power slot slot
  ```

  ```
  disable inline-power ports [all | port_list]
  ```

  Disabling the inline power to a PD immediately removes power from the PD. Inline power is enabled by default.
- Display the configuration for inline power.

  ```
  show inline-power
  ```

## Reserving Power

Summit X430, X440-24p, X450-G2, X460-24p, X460-48p, X460-G2 Switches or a Slot on Modular *PoE* Switches only:

On modular PoE switches, you reserve power for a given slot. The power reserved for a given slot cannot be used by any other PoE slots, even if the assigned power is not entirely used. To reallocate power among the slots, you must reconfigure each slot for the power budget you want; the power is not dynamically reallocated among PoE modules.

The Summit X440-24p, X440-8p and X440-48p have one internal PSU capable of 380 W of PoE power.

For Summit X460-24p and X460-48p switches, each internal PSU is capable of 380 W of PoE power. (Refer to the *ExtremeSwitching and Summit Switches: Hardware Installation Guide for Switches Using ExtremeXOS 16 or Earlier* for complete information on power availability with this optional unit.)

You do not have to disable the PoE devices to reconfigure the power budgets.

On modular switches, the default power budget is 50 W per slot, and the maximum is 768 W. The minimum reserved power budget you can configure is 37 W for an enabled slot. If inline power on the slot is disabled, you can configure a power budget of 0.

**Note**

We recommend that you fully populate a single PoE module with PDs until the power usage is just below the usage threshold, instead of spacing PDs evenly across PoE modules.

- To reset the power budget for a PoE module to the default value of 50 W, use the following command:

  ```
  unconfigure inline-power budget slot slot
  ```

- To display the reserved power budget for the PoE modules, use the following command:

  `show inline-power slot slot`

*Configuring Budget for Summit X430-8p*

```
Configure inline-power budget
90
Warning: Only stand-alone x430-8p unit can be
applied with more than 60 W POE power.
If more than 60 W POE power is used for a
rack-mounted x430-8p configuration, unit may get overheated.
Do you want to continue? (y/N)
```

## Setting the Disconnect Precedence

Summit X430, X440-24p, X450-G2, X460-24p, X460-48p, X460-G2, or Modular *PoE* Switches only.

**Note**

The switch generates an *SNMP* event if a PD goes offline, and the port's state moves from Power to Searching. You must configure SNMP to generate this event.

When the actual power used by the PDs on a switch or slot exceeds the power budgeted for that switch or slot, the switch refuses power to PDs. There are two methods used by the switch to refuse power to PDs, and whichever method is in place applies to all PoE slots in the switch. This is called the disconnect precedence method, and you configure one method for the entire switch.

The available disconnect precedence methods are:

- Deny port
- Lowest priority

The default value is deny port. Using this method, the switch simply denies power to the next PD requesting power from the slot, regardless of that port's PoE priority or port number.

Using the lowest priority method of disconnect precedence, the switch disconnects the PDs connected to ports configured with lower PoE priorities. (Refer to Configuring the PoE Port Priority for information on port priorities.)

When several ports have the same PoE priority, the lower port numbers have higher PoE priorities. That is, the switch withdraws power (or disconnects) those ports with the highest port number(s).

The system keeps dropping ports, using the algorithm you selected with the disconnect ports command, until the measured inline power for the slot is lower than the reserved inline power.

- To Configure the disconnect precedence for the switch, use the following command:

  `configure inline-power disconnect-precedence [deny-port | lowest-priority]`
- To return the disconnect precedence to the default value of deny port, use the following command:

  `unconfigure inline-power disconnect-precedence`
- To display the currently configured disconnect precedence, use the following command:

  `show inline-power`

*Configuring the PoE Port Priority*

Summit X440-24p, X460-24p, X460-48p, and Modular PoE Switches only.

You can configure the *PoE* port priority to be low, high, or critical. The default value is low.

If you configure the disconnect precedence as lowest priority and the PDs request power in excess of the switch's or slot's reserved power budget, the system allocates power to those ports with the highest priorities first.

If several ports have the same PoE priority, the lower port numbers have higher PoE priorities. That is, the switch withdraws power (or disconnects) those ports with the highest port number(s).

- To configure PoE port priority, use the following command:

  ```
  configure inline-power priority [critical | high | low] ports
  port_list
  ```

- To reset the port priority to the default value of low, use the following command:

  ```
  unconfigure inline-power priority ports [all | port_list]
  ```

- To display the PoE port priorities, use the following command:

  ```
  show inline-power configuration ports port_list
  ```

## Configuring the Usage Threshold

The system generates an *SNMP* event after a preset percentage of the reserved power for any slot or total power for a stand-alone switch is actually used by a connected PD. This preset percentage is called the usage threshold and is the percentage of the measured power to the budgeted power for each slot or total power for a stand-alone switch.

On modular switches, although the percentage of used to budgeted power is measured by each *PoE* module, you can set the threshold for sending the event for the entire switch. That is, after any PoE module passes the configured threshold, the system sends an event.

The default value for this usage threshold is 70%. You can configure the usage threshold to be any integer between 1% and 99%.

- To configure the threshold percentage of budgeted power used on a slot or the total power on a stand-alone switch that causes the system to generate an SNMP event and *EMS* message, use the following command:

  ```
  configure inline-power usage-threshold threshold
  ```

- To reset the threshold that causes the system to generate an SNMP event and EMS message per slot to 70% for measured power compared to budgeted power, use the following command:

  ```
  unconfigure inline-power usage-threshold
  ```

- To display the currently configured usage threshold, use the following command:

  ```
  show inline-power
  ```

## Configuring the Switch to Detect Legacy PDs

The *PoE* device can detect non-standard, legacy PDs, which do not conform to the IEEE 802.3af standard, using a capacitance measurement. However, you must specifically enable the switch to detect these non-standard PDs; the default value for this detection method is disabled.

This configuration applies to the entire switch; you cannot configure the detection method per slot.

The switch detects PDs through capacitance only if both of the following conditions are met:

- The legacy detection method is enabled.
- The switch unsuccessfully attempted to discover the PD using the standard resistance measurement method.

- To configure a switch to detect legacy, non-standard PDs for a specified slot, use the following command:

  `configure inline-power detection [802.3af-only | legacy-and-802.3af] slot slotnum`

- To configure the switch to detect legacy PDs on a switch, use the following command:

  `configure inline-power detection [802.3af-only | legacy-and-802.3af | bypass] ports port_list`

- To reset the switch to the default value which does not detect legacy PDs, on a specified slot use the following command:

  `unconfigure inline-power legacy slot slot`

- To reset the switch to the default value which does not detect legacy PDs, use the following command:

  `unconfigure inline-power legacy`

- To display the status of legacy detection, use the following command:

  `show inline-power`

## Configuring the Operator Limit

You can configure the maximum amount of power that the specified port can deliver to the connected PD, in milliwatts (mW). For *PoE*, the default value is 15400 mW, and the range is 3000 to 16800 mW. For PoE+, the default value is 30000 mW, and the range is 3000 to 30000 mW.

If the operator limit for a specified port is less than the power drawn by the legacy PD, the legacy PD is denied power.

- To set the operator limit on specified ports, which limits how much power a PD can draw from that port, use the following command:

  `configure inline-power operator-limit milliwatts ports [all | port_list]`

- To reset the operator limit to the default value of 15.4 W for PoE or 30 W for PoE+, use the following command:

  `unconfigure inline-power operator-limit ports [all |port_list]`

- To display the current operator limit on each port, use the following command:

  `show inline-power configuration ports port_list`

## Configuring PoE Port Labels

You can assign labels to a single or group of *PoE* ports using a string of up to 15 characters.

- To assign a label to PoE ports, use the following command:

  `configure inline-power label string ports port_list`

To rename a port or to return it to a blank label, reissue the command.

- To display the PoE port labels, use the following command:
  `show inline-power configuration ports port_list`

## Power Cycling Connected PDs

You can power cycle a connected PD without losing the power allocated to its port using the following command:

```
reset inline-power ports port_list
```

## Adding an S-PoE Daughter Card to an Existing Configuration

G48Tc and G48Te2 I/O Modules for the BlackDiamond 8800 Series Switches.

This section describes how to add an S-*PoE* daughter card to an EXOS configuration that has already been saved without PoE capabilities.

The example in this section uses the G48Te2 module.

The following output displays the results of the show slot command with slot 4 configured:

```
BD-8810.6 # show slot
Slots    Type                Configured          State       Ports  Flags
-------------------------------------------------------------------------------
Slot-1   G48Tc                G48Tc               Operational   48   MB
Slot-2                                           Empty          0
Slot-3                                           Empty          0
Slot-4   G48Te2              G48Te2              Operational   48   MB
Slot-5   G8Xc                 G8Xc                Operational    8   MB
Slot-6   10G1Xc               10G1Xc              Operational    1   MB
Slot-7   10G4X                10G4X               Operational    4   MB
Slot-8                                           Empty          0
Slot-9                                           Empty          0
Slot-10                                          Empty          0
MSM-A    MSM-48c                                 Operational    0
MSM-B    MSM-48c                                 Operational    0
Flags: M - Backplane link to Master is Active
B - Backplane link to Backup is also Active
D - Slot Disabled
I - Insufficient Power (refer to "show power budget")
```

To configure a module for the PoE daughter card, follow these steps:

1. Remove the G48Te2 module.
2. Attach the PoE daughter card to the G48Te2 module (as described in installation document provided with the daughter card).
3. Re-insert G48Te2 module with the PoE daughter card attached.

   The following output displays the results of the show slot command after the card is attached:

```
* BD-8810.20 # show slot
Slots    Type                Configured          State       Ports  Flags
-------------------------------------------------------------------------------
Slot-1   G48Tc                G48Tc               Operational   48   MB
Slot-2                                           Empty          0
```

```
Slot-3                                                  Empty          0
Slot-4   G48Te2(PoE)          G48Te2                    Operational   48   MB
Slot-5   G8Xc                 G8Xc                      Operational    8   MB
Slot-6   10G1Xc               10G1Xc                    Operational    1   MB
Slot-7   10G4X                10G4X                     Operational    4   MB
Slot-8                                                  Empty          0
Slot-9                                                  Empty          0
Slot-10                                                 Empty          0
MSM-A    MSM-48c                                        Operational    0
MSM-B    MSM-48c                                        Operational    0
Flags : M - Backplane link to Master is Active
        B - Backplane link to Backup is also Active
        D - Slot Disabled
        I - Insufficient Power (refer to "show power budget")
```

You can expect to see the following log messages generated by the system after you have attached the card:

```
<Warn:HAL.Card.Warning> MSM-A: Powering on mismatch card - cfg: G48Te2
actual: G48Te2(PoE)

<Warn:HAL.Card.Warning> MSM-B: Powering on mismatch card - cfg: G48Te2
actual: G48Te2(PoE)
```

4. Change the slot module type to include POE by executing the command configure slot 4 module G48Te2 (PoE).

**Note**

You must configure the slot as (PoE) before the power feature is accessible or enabled.

The following output displays the results of the show slot command after this command has been executed:

```
* BD-8810.20 # show slot
Slots    Type                 Configured           State       Ports  Flags
-------------------------------------------------------------------------------
Slot-1   G48Tc                G48Tc                Operational   48   MB
Slot-2                                             Empty          0
Slot-3                                             Empty          0
Slot-4   G48Te2(PoE)          G48Te2(PoE)          Operational   48   MB
Slot-5   G8Xc                 G8Xc                 Operational    8   MB
Slot-6   10G1Xc               10G1Xc               Operational    1   MB
Slot-7   10G4X                10G4X                Operational    4   MB
Slot-8                                             Empty          0
Slot-9                                             Empty          0
Slot-10                                            Empty          0
MSM-A    MSM-48c                                   Operational    0
MSM-B    MSM-48c                                   Operational    0
Flags : M - Backplane link to Master is Active
        B - Backplane link to Backup is also Active
        D - Slot Disabled
        I - Insufficient Power (refer to "show power budget")
```

5. Save the configuration by executing the `save configuration` command.

## Displaying PoE Settings and Statistics

You can display the *PoE* status, configuration, and statistics for the system, slot, and port levels.

## Clearing Statistics

You can clear the *PoE* statistics for specified ports or for all ports.

To clear the statistics and reset the counters to 0, enter the following command:

```
clear inline-power stats ports [all | port_list]
```

## Displaying System Power Information

You can display the status of the inline power for the system and, for additional information, display the power budget of the switch.

*Display System PoE Status*

- Display the *PoE* status for the switch.

  show inline-power

  The command provides status for the following areas:

  - Configured inline power status—The status of the inline power for the switch: enabled or disabled.
  - System power surplus—The surplus amount of power on the system, in watts, available for budgeting.
  - Redundant power surplus—The amount of power on the system, in watts, available for budgeting if one power supply is lost.
  - System power usage threshold—The configured power usage threshold for each slot, shown as a percentage of budgeted power. After this threshold has been passed on any slot, the system sends an *SNMP* event and logs a message.
  - Disconnect precedence—The method of denying power to PDs if the budgeted power on any slot is exceeded.
  - Legacy mode—The status of the legacy mode, which allows detection of non-standard PDs.

  The output indicates the following inline power status information for each slot:

  - Inline power status—The status of inline power. The status conditions are:
  - Enabled
  - Disabled
  - Firmware status—The operational status of the slot. The status conditions are:
  - Operational
  - Not operational
  - Disabled
  - Subsystem failure
  - Card not present
  - Slot disabled
  - Budgeted power—The amount of inline power, in watts, that is reserved and available to the slot.
  - Measured power—The amount of power, in watts, that is currently being used by the slot.

*Display System Power Data*

- Additionally, you can view the distribution of power, as well as currently required and allocated power, on the entire modular switch including the power supplies by using the following command:

  show power budget

## Displaying Slot PoE Information on Modular Switches

You can display *PoE* status and statistics per slot.

*Displaying Slot PoE Status*

To display *PoE* status for each slot, use the following command:

```
show inline-power slot slot
```

The command provides the following information:

- Inline power status—The status of inline power. The status conditions are:
  - Enabled
  - Disabled
- Firmware status—The operational status of the slot. The status conditions are:
  - Operational
  - Not operational
  - Disabled
  - Subsystem failure
  - Card not present
  - Slot disabled
- Budgeted power—The amount of power, in watts, that is available to the slot.
- Measured power—The amount of power, in watts, that is currently being used by the slot.

*Display Slot PoE Statistics*

To display the *PoE* statistics for each slot, use the following command:

```
show inline-power stats slot slot
```

The command provides the following information:

- Firmware status—Displays the firmware state:
  - Operational
  - Not operational
  - Disabled
  - Subsystem failure
  - Card not present
  - Slot disabled
- Firmware revision—Displays the revision number of the PoE firmware
- Total ports powered—Displays the number of ports powered on specified slot
- Total ports awaiting power—Displays the number of remaining ports in the slot that are not powered

- Total ports faulted—Displays the number of ports in a fault state
- Total ports disabled—Displays the number of ports in a disabled state

## Displaying PoE Status and Statistics on Stand-alone Switches

To display the *PoE* statistics for the switch, run the `show inline-power stats` command.

The command provides the following information:

- Firmware status—Displays the firmware state:
  - Operational
  - Not operational
  - Disabled
  - Subsystem failure
- Firmware revision—Displays the revision number of the PoE firmware.
- Total ports powered—Displays the number of ports powered on specified slot.
- Total ports awaiting power—Displays the number of remaining ports in the slot that are not powered.
- Total ports faulted—Displays the number of ports in a fault state.
- Total ports disabled—Displays the number of ports in a disabled state.

## Displaying Port PoE Information

You can display the *PoE* configuration, status, and statistics per port.

### *Displaying Port PoE Configuration*

To display *PoE* configuration for each port, use the following command:

`show inline-power configuration ports` *`port_list`*

This command provides the following information:

- Config—Indicates whether the port is enabled to provide inline power:
  - Enabled: The port can provide inline power.
  - Disabled: The port cannot provide inline power.
- Operator Limit—Displays the configured limit, in milliwatts, for inline power on the port.
- Label—Displays a text string, if any, associated with the port (15 characters maximum).

### *Display Port PoE Status*

To display the *PoE* status per port, run the `show inline-power info {detail} ports` *`port_list`* command. This command provides the following information:

- State—Displays the port power state:
  - Disabled
  - Searching
  - Delivering

  - Faulted
  - Disconnected
  - Other
  - Denied
- PD's power class—Displays the class type of the connected PD:
  - "-----": disabled or searching
  - "class0": class 0 device
  - "class1": class 1 device
  - "class2": class 2 device
  - "class3": class 3 device
  - "class4": class 4 device
- Volts—Displays the measured voltage. A value from 0 to 2 is valid for ports that are in a searching or discovered state.
- Curr—Displays the measured current, in milliamperes, drawn by the PD.
- Power—Displays the measured power, in watts, supplied to the PD.
- Fault—Displays the fault value:
  - None
  - UV/OV fault
  - UV/OV spike
  - Over current
  - Overload
  - Undefined
  - Underload
  - HW fault
  - Discovery resistance fail
  - Operator limit violation
  - Disconnect
  - Discovery resistance, A2D failure
  - Classify, A2D failure
  - Sample, A2D failure
  - Device fault, A2D failure
  - Force on error

The **`detail`** command lists all inline power information for the selected ports. Detail output displays the following information:

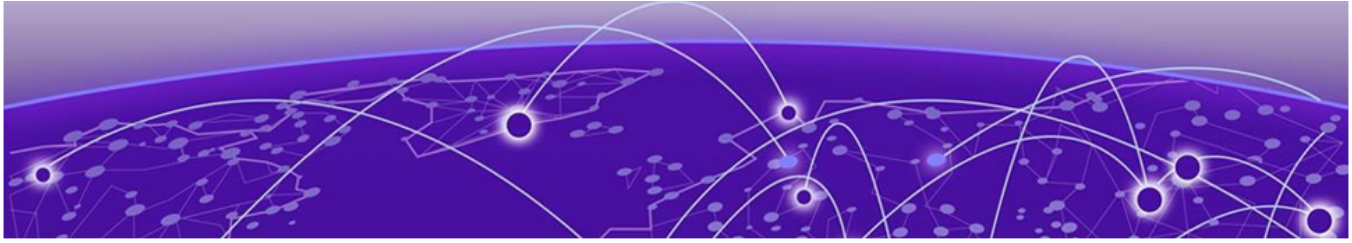
- Configured admin state
- Inline power state
- MIB detect status
- Label
- Operator limit
- PD class
- Max allowed power
- Measured power

- Line voltage
- Current
- Fault status
- Detailed status
- Priority

*Displaying Port PoE Statistics*

To display the *PoE* statistics for each port, run the `show inline-power stats ports port_list` command. The command provides the following information:

- State—Displays the port power state:
  - Disabled
  - Searching
  - Delivering
  - Faulted
  - Disconnected
  - Other
  - Denied
- PD's power class—Displays the class type of the connected PD:
  - "-----": disabled or searching
  - "class0": class 0 device
  - "class1": class 1 device
  - "class2": class 2 device
  - "class3": class 3 device
  - "class4": class 4 device
- Absent—Displays the number of times the port was disconnected.
- InvSig—Displays the number of times the port had an invalid signature.
- Denied—Displays the number of times the port was denied.
- Over-current—Displays the number of times the port entered an overcurrent state.
- Short—Displays the number of times the port entered undercurrent state.



# Status Monitoring and Statistics



Viewing statistics on a regular basis allows you to see how well your network is performing. If you keep simple daily records, you can see trends emerging and notice problems arising before they cause major network faults. In this way, statistics can help you get the best out of your network.

The status monitoring facility provides information about the switch.

This information may be useful for your technical support representative if you have a problem. ExtremeXOS software includes many command line interface (CLI) show commands that display information about different switch functions and facilities.

Note
For more information about show commands for a specific ExtremeXOS feature, see the appropriate chapter in this guide.

## Viewing Port Statistics

ExtremeXOS software provides a facility for viewing port statistical information. The summary information lists values for the current counter for each port on each operational module in the system. The switch automatically refreshes the display (this is the default behavior).

You can also display a snapshot of the real-time port statistics at the time you issue the command and view the output in a page-by-page mode. This setting is not saved; therefore, you must specify the **no-refresh** parameter each time you want a snapshot of the port statistics.

Values are displayed to nine digits of accuracy.

- View port statistics with the following command:

  `show ports {port_list | stack-ports stacking-port-list} statistics {no-refresh}`

  The switch collects the following port statistical information:

  - Link State—The current state of the link. Options are:
    - Active (A)—The link is present at this port.
    - Ready (R)—The port is ready to accept a link.
    - Loopback (L)—The port is configured for WANPHY loopback.
    - Not Present (NP)—The port is configured, but the module is not installed in the slot (modular switches only).
  - Transmitted Packet Count (TX Pkt Count)—The number of packets that have been successfully transmitted by the port.
  - Transmitted Byte Count (TX Byte Count)—The total number of data bytes successfully transmitted by the port.
  - Received Packet Count (RX Pkt Count)—The total number of good packets that have been received by the port.
  - Received Byte Count (RX Byte Count)—The total number of bytes that were received by the port, including bad or lost frames. This number includes bytes contained in the Frame Check Sequence (FCS), but excludes bytes in the preamble.
  - Received Broadcast (RX Bcast)—The total number of frames received by the port that are addressed to a broadcast address.
  - Received Multicast (RX Mcast)—The total number of frames received by the port that are addressed to a multicast address.
- View port statistics for SummitStack stacking ports with the following command:

  `show ports stack-ports {stacking-port-list} statistics {no-refresh}`

## Viewing Port Errors

The switch keeps track of errors for each port and automatically refreshes the display (this is the default behavior). You can also display a snapshot of the port errors at the time you issue the command and view the output in a page-by-page mode. This setting is not saved; therefore, you must specify the *no-refresh* parameter each time you want a snapshot of the port errors.

- View port transmit errors with the following command:

  `show ports {port_list | stack-ports stacking-port-list} txerrors {no-refresh}`

  The switch collects the following port transmit error information:

  - Port Number—The number of the port.

- Link State—The current state of the link. Options are:
  - Active (A)—The link is present at this port.
  - Ready (R)—The port is ready to accept a link.
  - Loopback (L)—The port is configured for WANPHY loopback.
  - Not Present (NP)—The port is configured, but the module is not installed in the slot (modular switches only).
- Transmit Collisions (TX Coll)—The total number of collisions seen by the port, regardless of whether a device connected to the port participated in any of the collisions.
- Transmit Late Collisions (TX Late Coll)—The total number of collisions that have occurred after the port's transmit window has expired.
- Transmit Deferred Frames (TX Deferred)—The total number of frames that were transmitted by the port after the first transmission attempt was deferred by other network traffic.
- Transmit Errored Frames (TX Errors)—The total number of frames that were not completely transmitted by the port because of network errors (such as late collisions or excessive collisions).
- Transmit Lost Frames (TX Lost)—The total number of transmit frames that do not get completely transmitted because of buffer problems (FIFO underflow).
- Transmit Parity Frames (TX Parity)—The bit summation has a parity mismatch.

- View port receive errors with the command:

  `show ports {`*`port_list`* `|` **`stack-ports`** *`stacking-port-list`*`}` **`rxerrors`** `{`**`no-refresh`**`}`

  The switch collects the following port receive error information:

  - Port Number
  - Link State—The current state of the link. Options are:
    - Active (A)—The link is present at this port.
    - Ready (R)—The port is ready to accept a link.
    - Not Present (NP)—The port is configured, but the module is not installed in the slot (modular switches only).
    - Loopback (L)—The port is in Loopback mode.
  - Receive Bad CRC Frames (RX CRC)—The total number of frames received by the port that were of the correct length but contained a bad FCS value.
  - Receive Oversize Frames (RX Over)—The total number of good frames received by the port greater than the supported maximum length of 1,522 bytes.
  - Receive Undersize Frames (RX Under)—The total number of frames received by the port that were less than 64 bytes long.
  - Receive Fragmented Frames (RX Frag)—The total number of frames received by the port that were of incorrect length and contained a bad FCS value.
  - Receive Jabber Frames (RX Jabber)—The total number of frames received by the port that were greater than the support maximum length and had a Cyclic Redundancy Check (CRC) error.
  - Receive Alignment Errors (RX Align)—The total number of frames received by the port with a CRC error and not containing an integral number of octets.
  - Receive Frames Lost (RX Lost)—The total number of frames received by the port that were lost because of buffer overflow in the switch.

- For SummitStack stacking ports, you can also view transmit and receive errors with the following commands:
  `show ports stack-ports` *stacking-port-list*`}`**`txerrors`** `{`**`no-refresh`**`}`
  `show ports stack-ports {`*stacking-port-list*`}`**`rxerrors`** `{`**`no-refresh`**`}`
  Information displayed is identical to the details displayed for non-stacking ports.

## Using the Port Monitoring Display Keys

The following table describes the keys used to control the displays that appear if you use any of the show ports commands without specifying the *no-refresh* parameter (this is the default behavior).

**Table 53: Port Monitoring Display Keys with Auto-Refresh Enabled**

| Key(s) | Description |
|---|---|
| U | Displays the previous page of ports. |
| D | Displays the next page of ports. |
| [Esc] | Exits from the screen. |
| 0 | Clears all counters. |
| [Space] | Cycles through the following screens:<br>• Packets per second<br>• Bytes per second<br>• Percentage of bandwidth<br><br>Note: Available only using the `show ports utilization` command. |

The following table describes the keys used to control the displays that appear if you use any of the show ports commands and specify the **`no-refresh`** parameter.

**Table 54: Port Monitoring Display Keys with Auto-Refresh Disabled**

| Key | Description |
|---|---|
| Q | Exits from the screen. |
| [Space] | Displays the next page of ports. |

## Viewing VLAN Statistics

### BlackDiamond 8000 Series modules, SummitStack and Summit Family Switches Only

ExtremeXOS software provides the facility for configurable *VLAN (Virtual LAN)* statistics gathering and display of packet and byte counters on the port level and on the VLAN level.

*Configuring VLAN Statistics*

- Configure the switch to start counting *VLAN* statistics with the commands:
  `clear counters`

`configure ports [port_list|all] monitor vlan vlan_name {rx-only | tx-only}`

- View VLAN statistics at the port level with the command:
  `show ports {port_list} vlan statistics {no-refresh}`

  The switch collects and displays the following statistics:

  - Port—The designated port.
  - VLAN—The associated VLANs.
  - Rx Frames Count—The total number of frames successfully received by the port on the designated VLAN.
  - Rx Byte Count—The total number of bytes that were received by the port on the designated VLAN.
  - Tx Frame Count—The total number of frames that were transmitted by the port on the designated VLAN.
  - Tx Byte Count—The total number of bytes that were transmitted by the port on the designated VLAN.

  Frame and byte counters are also displayed through *SNMP (Simple Network Management Protocol)*.
- View VLAN statistics at the VLAN level with the command:
  `show ports {port_list} vlan statistics {no-refresh}`

  The switch collects and displays the following statistics:

  - VLAN—The designated VLAN.
  - Rx Total Frames —The total number of frames successfully received by the port.
  - Rx Byte Count—The total number of bytes that were received by the port.
  - Tx Total Frames—The total number of frames that were transmitted by the port.
  - Tx Byte Count—The total number of bytes that were transmitted by the port.
- Stop counting VLAN statistics.
  `unconfigure ports [port_list |all] monitor vlan vlan_name`

  **Note**
  While using VLAN statistics on Summit family switches or BlackDiamond 8000 series modules, traffic also matching egress *ACL (Access Control List)* counters will cause the VLAN statistics Tx counter to not increment.

### *Guidelines and Limitations*

The following describes guidelines for this feature.

- Support for VMAN statistics are provided in the same manner as *VLAN* statistics. CLI commands use the same syntax as used with monitoring VLANs including the use of the **vlan** keyword in CLI commands. *SNMP* access uses the same MIB objects as used for VLAN statistics.

  Statistics for VLANs encapsulated within VMANs are not supported.
- Only BlackDiamond X8 and 8900 series modules, and Summit X460, X460G2, X480, X670, X670G2, and X770 switches provide support for both receive and transmit statistics.
- BlackDiamond 8000 series modules and Summit family switches provide support only for byte counters.

- All of the counters supported are 64 bit counters. No indication of counter rollover is supported.
- Packets originating from the switch's CPU or forwarded by the CPU may not be reflected in transmit statistics.

## Performing Switch Diagnostics

The switch provides a facility for running normal or extended diagnostics.

In simple terms, a normal routine performs a simple ASIC and packet loopback test on all ports, and an extended routine performs extensive ASIC, ASIC-memory, and packet loopback tests. By running and viewing the results from diagnostic tests, you can detect and troubleshoot any hardware issues.

On BlackDiamond X8 and 8800 series switches, you can run the diagnostic routine on Input/Output (I/O) modules, management modules (MSMs/MMs), or Fabric modules (BlackDiamond X8 series only) without affecting the operation of the rest of the modules. The module under test is taken offline while the diagnostic test is performed. Traffic to and from the ports on that module is temporarily unavailable. When the diagnostic test is complete, the module is reset and becomes operational again.

On Summit family switches, you run the diagnostic routine on the switch or on the stacking ports. Running the switch or stacking port diagnostic routine affects system operation; the switch is unavailable during the diagnostic test.

Note
The diagnostics image is not included in the SummitX480 image. You must use the diagnostics.xmod to update the diagnostics image.

Note
Before running diagnostics, you must power on the External Power Supply (EPS) when it is connected to the switch.

When you run diagnostics on an I/O module, MSM/MM module, Fabric module (BlackDiamond X8 series only), or a Summit family switch, the switch verifies that the:

- Registers can be written to and read from correctly.
- Memory addresses are accessed correctly.
- Application-Specific Integrated Circuit (ASICs) and Central Processing Unit (CPUs) operate as required.
- Data and control fabric connectivity is active (modular switches only).
- External ports can send and receive packets.
- Sensors, hardware controllers, and LEDs are working correctly.

Note
Before running slot diagnostics on a modular switch, you must have at least one MSM/MM installed in the chassis.

When you run diagnostics on the SummitStack stacking ports, the switch completes a hardware test to ensure that the stacking ports are operational.

## Running Diagnostics

*BlackDiamond X8 and 8800 Series Switches*

If you run the diagnostic routine on an I/O or Fabric module (BlackDiamond X8 series only), that module is taken offline while the diagnostic test is performed. Traffic to and from the ports on that I/O or Fabric (BlackDiamond X8 series only) module is temporarily unavailable. When the diagnostic test is complete, the module is reset and becomes operational again.

If you run diagnostics on an MSM/MM, that module is taken offline while the diagnostics test is performed. When the diagnostic test is complete, the MSM/MM reboots and becomes operational again.

If you run diagnostics on the primary MSM/MM, the backup MSM/MM assumes the role of the primary and takes over switch operation.

After the MSM/MM completes the diagnostic routine and reboots, you can initiate failover from the new primary MSM/MM to the original primary MSM/MM. Before initiating failover, confirm that both MSMs/MMs are synchronized using the `show switch` command. If the MSMs/MMs are synchronized, initiate failover using the `run msm-failover` command. For more detailed information about system redundancy and MSM/MM failover, see #unique_938.

Run diagnostics on one MSM/MM at a time. After you run the diagnostic routine on the first MSM/MM, use the `show switch` command to confirm that both MSMs/MMs are up, running, and synchronized before running diagnostics on the second MSM/MM.

After the switch runs the diagnostic routine, test results are saved in the module's EEPROM and messages are logged to the syslog.

- Run diagnostics on I/O, Fabric (BlackDiamond X8 series only), or MSM/MM modules with the following command:

  `run diagnostics [`**`extended`**` | `**`normal`**` | `**`stack-port`**`] {`**`slot`**` [`*`slot`*` | `**`A`**` | `**`B`**`]}`

  Where the following is true:

  - **`extended`**—Takes the switch fabric and ports offline and performs extensive ASIC, ASIC-memory, and packet loopback tests. Extended diagnostic tests take approximately 15 to 20 minutes to complete. The CPU is not tested. Console access is available during extended diagnostics.

  If you have a *PoE (Power over Ethernet)* module installed, the switch also performs an extended PoE test, which tests the functionality of the inline power adapter.

  - **`normal`**—Takes the switch fabric and ports offline and performs a simple ASIC and packet loopback test on all ports.

- *slot*—Specifies the slot number of an I/O or Fabric (BlackDiamond X8 series only) module. When the diagnostic test is complete, the system attempts to bring the module back online.

  Note

  BlackDiamond 8800 series switches—To run diagnostics on the management portion of the master MSM, specify slot A or B. If an I/O subsystem is present on the MSM, then that I/O subsystem will be non-operational until diagnostics are completed.

  BlackDiamond 8810 switch—If you run diagnostics on slots 5 and 6 with an MSM installed in those slots, the diagnostic routine tests the I/O subsystem of the MSM.

  BlackDiamond 8806 switch—If you run diagnostics on slots 3 and 4 with an MSM installed in those slots, the diagnostic routine tests the I/O subsystem of the MSM.

- **A | B**—Specifies the slot letter of the primary MSM. The diagnostic routine is performed when the system reboots. Both switch fabric and management ports are taken offline during diagnostics.

  Note

  BlackDiamond X8 and 8800 series switches do not allow you to run diagnostics on a module that has been disabled. Command line interface message: `Cannot run diags because I/O card is not Operational or Offline Current state is Down.`

*SummitStack or Summit Family Switches*

Diagnostics cannot be run on a SummitStack,. You need to disable stacking on the switch to be tested, reboot the switch before logging in, and then run the diagnostics. Once the diagnostics routine is complete, the switch reboots again. Log in, enable stacking mode, and reboot the switch again. Upon reboot, the switch rejoins the stack. You can then use the show diagnostics command to see the last diagnostic result of any or all switches in the SummitStack.

If you run the diagnostic routine on Summit family switches, the switch reboots and then performs the diagnostic test.

Note

The diagnostics image is not included in the SummitX480 image. You must use the diagnostics.xmod to update the diagnostics image.

During the test, traffic to and from the ports on the switch is temporarily unavailable. When the diagnostic test is complete, the switch reboots and becomes operational again.

To run the diagnostic routine on the stack ports, you need a dedicated stacking cable that connects stack port 1 to stack port 2, which are located at the rear of the switch. The stacking cable is available from Extreme Networks. The switch performs a hardware test to confirm that the stack ports are

operational; traffic to and from the ports on the switch is temporarily unavailable. This Bit Error Rate Test (BERT) provides an analysis of the number of bits transmitted in error.

Note

The stack ports diagnostic does not require a dedicated cable to be connected for the following SummitStack-V80 plugins: VIM2-SSV80 on X480 and SS-V80 on X460/E4G-400. For this hardware, the stack port diagnostic implements an internal loopback within the module.

After the switch runs the diagnostic routine, test results saved to the switch's EEPROM and messages are logged to the syslog.

- Run diagnostics on Summit family switches with the following command:

  `run diagnostics [extended | normal | stack-port] {slot [slot | A | B]}`

  Where the following is true:

  **extended**—Reboots the switch and performs extensive ASIC, ASIC-memory, and packet loopback tests. Extended diagnostic tests take a maximum of five minutes. The CPU is not tested.

  - **normal**—Reboots the switch and performs a simple ASIC and packet loopback test on all ports.
  - **stack-port**—Performs a BERT on the stacking ports and reboots the switch.

## Observing LED Behavior During a Diagnostic Test

Whether you run a diagnostic test on an I/O module, MSM/MM, or a Summit family switch, LED activity occurs during and immediately following the test.

The LED behavior described in this section relates only to the behavior associated with a diagnostic test. For more detailed information about all of the I/O module, MSM/MM, and switch LEDs, see the hardware documentation listed in Extreme Networks Documentation.

### *I/O Module LED Behavior--BlackDiamond 8800 Series Switches*

The following table describes the BlackDiamond 8800 series switch I/O module LED behavior during a diagnostic test.

**Table 55: BlackDiamond 8800 Series Switch I/O Module LED Behavior**

| LED | Color | Indicates |
|---|---|---|
| DIAG | Amber blinking | Diagnostic test in progress. |
| | Amber | Diagnostic failure has occurred. |
| | Green | Diagnostic test has passed. |
| Stat | Amber blinking | Configuration error, code version error, diagnostic failure, or other severe module error. |
| | Off | Diagnostic test in progress, or diagnostic failure has occurred. |

After the I/O module completes the diagnostic test, or the diagnostic test is terminated, the DIAG and the Status LEDs are reset. During normal operation, the DIAG LED is off and the Status LED blinks green.

*MSM LED Behavior--BlackDiamond 8800 Series Switches*

This section describes the MSM behavior during a diagnostic test.

After the I/O modules complete the diagnostic test, or the diagnostic test is terminated, the DIAG and the Status LEDs are reset. During normal operation, the DIAG LED is off and the Status LED blinks green. If you start another diagnostic test, the LED returns to blinking amber.

*I/O and Fabric Module LED Behavior--BlackDiamond X8 Switch*

The following table describes the BlackDiamond X8 switch I/O and fabric module LED behavior during a diagnostic test.

After the I/O or Fabric module completes the diagnostic test, or the diagnostic test is terminated, the DIAG and the Status LEDs are reset. During normal operation, the DIAG LED is off and the Status LED blinks green.

**Table 56: BlackDiamond X8 Series Switch I/O and Fabric Module LED Behavior**

| LED | Color | Indicates |
|---|---|---|
| DIAG | Amber blinking | Diagnostic test in progress |
| | Amber | Diagnostic failure has occurred |
| Status | Amber blinking | Configuration error, code version error, or other severe module error |
| | Green blinking | Normal operation |

*MM LED Behavior--BlackDiamond X8 Switch*

The following table describes the BlackDiamond X8 switch MM LED behavior during a diagnostic test.

After the MM completes the diagnostic test, or the diagnostic test is terminated, the SYS LED is reset. During normal operation, the status LED blinks green.

**Table 57: BlackDiamond X8 Series Switch MM LED Behavior**

| LED | Color | Indicates |
|---|---|---|
| SYS | Amber blinking | Diagnostic test in progress |
| | Amber | Diagnostic failure has occurred |

*LED Behavior--Summit Family Switches*

The following table describes the Summit family switches LED behavior during a diagnostic test.

**Table 58: Summit Family Switch LED Behavior**

| LED | Color | Indicates |
|---|---|---|
| MGMT | Green blinking | Normal operation is occurring. |
| | Amber blinking | Diagnostic test in progress. |

While diagnostic tests are running, the MGMT LED blinks amber. If a diagnostic test fails, the MGMT LED continues to blink amber. During normal operation, the MGMT LED blinks green.

### Displaying Diagnostic Test Results

- Display the status of the last diagnostic test run on the switch.

  `show diagnostics {[cr] | slot [slot | A | B]}`

  Note

  The **slot**, **A**, and **B** parameters are available only on modular switches.

## Using the System Health Checker

The system health checker is a useful tool to monitor the overall health of your system.

Depending on your platform, the software performs a proactive, preventive search for problems by polling and reporting the health of system components, including I/O and management module processes, power supplies, power supply controllers, and fans. By isolating faults to a specific module, backplane connection, control plane, or component, the system health checker notifies you of a possible hardware fault.

This section describes the system health check functionality of the following platforms:

- BlackDiamond X8 series switches
- BlackDiamond 8800 series switches
- Summit family switches

### Understanding the System Health Checker

*BlackDiamond 8800 and BlackDiamond X8 Series Switches Only*

On BlackDiamond 8800 and BlackDiamond X8 series switches, the system health checker tests the backplane, the CPUs on the MSM/MM modules, the I/O modules, the processes running on the switch, and the power supply controllers by periodically forwarding packets and checking for the validity of the forwarded packets.

Two modes of health checking are available: polling (also known as control plane health checking) and backplane diagnostic packets (also known as data plane health checking).

These methods are briefly described in the following:

- Polling is always enabled on the system and occurs every 5 seconds by default. The polling value is not a user-configured parameter. The system health checker polls the control plane health between MSM/MMs and I/O modules, monitors memory levels on the I/O module, monitors the health of the I/O module, and checks the health of applications and processes running on the I/O module. If the system health checker detects an error, you will be notified through the switch log.

  Here is an example from a BDX8:

  ```
  <Noti:HAL.SM.ChanDsbl> MM-A: Switch fabric channel 0 provided by Fabric slot FM-1 has
  been disabled due to failed connection with Slot-1
  ```

  If you find errors, refer to #unique_260.

- Backplane diagnostic packets are disabled by default. If you enable this feature, the system health checker tests the data link for a specific I/O module every 5 seconds by default. The MSM/MM sends and receives diagnostic packets from the I/O module to determine the state and connectivity.

  If you disable backplane diagnostics, the system health checker stops sending backplane diagnostic packets.

  To see any results, you need to monitor the log to see if there are any errors detected when the backplane diagnostic packets system health checker is enabled.

  No log messages generated means no errors found.

  Here is an example of an error found on BDX8:

  ```
  <Warn:HAL.Sys.Warning> MM-A: Sys-Health-Check DataPath FM : slot/unit/port/modid F1->
  1/0->0/ 1-> 0/36-> 4 rc=3
  ```

  If you find errors, refer to #unique_260.

For more information about enabling and configuring backplane diagnostics, see the following sections:

- Enabling Diagnostic Packets on the Switch--Modular Switches Only on page 445
- Clearing the Shutdown State on page 451

System health check errors are reported to the syslog. If you see an error, contact support (see to #unique_260).

*Summit Family Switches Only*

On Summit family switches, the system health checker polls and reads the switch fabric and CPU registers.

Unlike the modular platforms, only polling is available on Summit family switches. Polling is always enabled on the system and occurs in the background every 10 seconds; the polling value is not a user-configured parameter.

System health check errors are reported to the syslog. If you see an error, contact Extreme Networks Technical Support (see to #unique_260). There are no health checking tests related to the stacking links in a SummitStack.

## Enabling Diagnostic Packets on the Switch--Modular Switches Only

- Enable diagnostic packets with the command:

  `enable sys-health-check` **`slot`** *`slot`*

For BlackDiamond 8800 and BlackDiamond X8 series switches, the system health checker tests the data link by default.

The 10 Gbps links (BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules)—the system health checker tests every five seconds for the specified slot.

Note

Enabling backplane diagnostic packets increases CPU utilization and competes with network traffic for resources.

## Configuring Diagnostic Packets on the Switch--Modular Switches Only

- Configure the frequency of sending backplane diagnostic packets on a BlackDiamond 8800 series or BlackDiamond X8 switch with the command:

  ```
  configure sys-health-check interval interval
  ```

  **Note**
  We do not recommend configuring an interval of less than the default interval. Doing so can cause excessive CPU utilization.

## Disabling Diagnostic Packets on the Switch--Modular Switches Only

- Disable diagnostic packets with the command:

  ```
  disable sys-health-check slot slot
  ```

For BlackDiamond 8800 and BlackDiamond X8 series switches, the system health checker stops sending backplane diagnostic packets to the specified slot. Polling, which is the default system health checker, remains enabled.

## Displaying the System Health Check Setting--All Platforms

- Display the system health check setting, including polling and how ExtremeXOS software handles faults on the switch with the command:

  ```
  show switch
  ```

  As previously described, polling is always enabled on the switch.

  The system health check setting, displayed as SysHealth check, shows the polling setting and how ExtremeXOS handles faults. The polling setting appears as Enabled, and the fault handling setting appears in parenthesis next to the polling setting. For more information about the fault handling setting, see the following sections: Configuring Hardware Recovery--SummitStack and Summit Family Switches Only on page 449 and Configuring Module Recovery--Modular Switches Only on page 451.

*Example*

In the following truncated output from a BlackDiamond 8810 switch, the system health check setting appears as SysHealth check: Enabled (Normal):

```
SysName:          TechPubs Lab
SysName:          BD-8810Rack3
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:1F:A2:60
SysHealth check:  Enabled (Normal)
Recovery Mode:    None
System Watchdog:  Enabled
```

## System Health Check Examples: Diagnostics--Modular Switches Only

This section provides examples for using the system health checker on BlackDiamond 8800 and X8 series switches. For more detailed information about the system health check commands, see the *ExtremeXOS 16.2 Command Reference Guide*.

### *Example on the BlackDiamond 8800 and BlackDiamond X8 Series Switch*

This section describes a series of two examples for:

- Enabling and configuring backplane diagnostics
- Disabling backplane diagnostics

#### Enabling and Configuring Backplane Diagnostics

The following example:

- Enables backplane diagnostic packets on slot 3
- Configures backplane diagnostic packets to be sent every seven seconds

1. Enable backplane diagnostic packets on slot 3.

   ```
   enable sys-health-check slot 3
   ```

   When you enable backplane diagnostic packets on slot 3, the timer runs at the default rate of five seconds.
2. Configure backplane diagnostic packets to be sent every seven seconds.

   ```
   configure sys-health-check interval 7
   ```

   Note

   We do not recommend configuring an interval of less than five seconds. Doing this can cause excessive CPU utilization.

#### Disabling Backplane Diagnostics

- Building upon the example in Enabling and Configuring Backplane Diagnostics on page 447, the following example disables backplane diagnostics on slot 3:

  ```
  disable sys-health-check slot 3
  ```

  Backplane diagnostic packets are no longer sent, but the configured interval for sending backplane diagnostic packets remains at seven seconds. The next time you enable backplane diagnostic packets, the health checker sends the backplane diagnostics packets every 7 seconds.
- To return to the "default" setting of five seconds, configure the frequency of sending backplane diagnostic packets to 5 seconds using the following command:

  ```
  configure sys-health-check interval 5
  ```

# Setting the System Recovery Level

Depending on your switch model, you can configure the switch, MSM/MM, or I/O module to take action if a fault detection exception occurs.

The following sections describe how to set the software and hardware recovery levels on the switch, MSM/MM, and I/O modules.

Note
You configure MSM/MM and I/O module recovery only on BlackDiamond X8 and BlackDiamond 8800 series switches.

## Configuring Software Recovery

The default setting and behavior is all. Extreme Networks strongly recommends using the default setting.

- You can configure the system to either take no action or to automatically reboot the switch after a software task exception using the following command:

  ```
  configure sys-recovery-level [all | none]
  ```

  Where the following is true:

  - **all**—Configures ExtremeXOS to log an error to the syslog and automatically reboot the system after any software task exception.

  On modular switches, this command sets the recovery level only for the MSMs/MMs. The MSM/MM should reboot only if there is a software exception that occurs on the MSM/MM. The MSM/MM should not reboot if a software exception occurs on an I/O module.

  - **none**—Configures the system to take no action if a software task exception occurs. The system does not reboot, which can cause unexpected switch behavior. On a SummitStack, the sys-recovery-level setting applies to all active nodes.

    Note
    Use this parameter only under the guidance of Extreme Networks Technical Support personnel.

## Display the Software Recovery Setting

- Display the software recovery setting on the switch using the command:

  ```
  show switch
  ```

  This command displays general switch information, including the software recovery level.

*Example*

The following truncated output from a Summit series switch displays the software recovery setting (displayed as Recovery Mode):

```
SysName:          TechPubs Lab
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:1F:A4:0E
```

```
Recovery Mode:    All
System Watchdog:  Enabled
```

**Note**
All platforms display the software recovery setting as Recovery Mode.

## Configuring Hardware Recovery--SummitStack and Summit Family Switches Only

You can configure Summit family switches or SummitStack to take no action, automatically reboot, or shut down if the switch detects a fault.

- Configure how the switch recovers from problems on a stand-alone Summit family switch with the command:

  `configure sys-recovery-level switch [none | reset | shutdown]`

- Configure hardware recovery on a particular active node in the SummitStack with the command:

  `configure sys-recovery-level slot [all | slot_number] [none | reset | shutdown]`

  Where the following is true:

  - **none**—Configures the MSM/MM or I/O module to maintain its current state regardless of the detected fault. The offending MSM/MM or I/O module is not reset. ExtremeXOS logs fault and error messages to the syslog and notifies you that the errors are ignored. This does not guarantee that the module remains operational; however, the switch does not reboot the module.
  - **reset**—Configures the offending MSM/MM or I/O module to reset upon fault detection. ExtremeXOS logs fault, error, system reset, and system reboot messages to the syslog.
  - **shutdown**—Configures the switch to shut down all slots/modules configured for shutdown upon fault detection. On the modules configured for shutdown, all ports in the slot are taken offline in response to the reported errors; however, the MSMs/MMs remain operational for debugging purposes only. You must save the configuration, using the `save configuration` command, for it to take effect. ExtremeXOS logs fault, error, system reset, system reboot, and system shutdown messages to the syslog.

  The default setting is **reset**.

  You can configure how ExtremeXOS handles a detected fault depending on the sys-recovery-level setting. To configure how ExtremeXOS handles faults, use the `configure sys-health-check all level [normal | strict]` command.

- To view the system health check settings on the switch, use the `show switch` command as described in Displaying the System Health Check Setting--All Platforms on page 446.

### *Confirmation Messages Displayed*

If you configure the hardware recovery setting to either none (ignore) or shut down, the switch prompts you to confirm this action. The following is a sample shutdown message:

```
Are you sure you want to shutdown on errors? (y/n)
```

- Enter `y` to confirm this action and configure the hardware recovery level.
- Enter `n` or press **[Enter]** to cancel this action.

*Messages Displayed at the Startup Screen*

If you configure the shutdown feature and an error is detected, the system displays an explanatory message on the startup screen.

The following truncated sample output shows the startup screen if a stand-alone switch is shut down as a result of the hardware recovery configuration:

```
All switch ports have been shut down.
Use the "clear sys-recovery-level" command to restore all ports.
! SummitX.1 #
```

When an exclamation point (!) appears in front of the command line prompt, it indicates that the entire stand-alone switch is shut down as a result of your hardware recovery configuration and a switch error.

*Displaying the Hardware Recovery Setting*

- Display the hardware recovery setting with the command:

  `show switch`

  If you change the hardware recovery setting from the default (reset) to either none (ignore) or shutdown, the switch expands the Recovery Mode output to include a description of the hardware recovery mode.

  If you keep the default behavior or return to reset, the Recovery Mode output lists only the software recovery setting.

**Example**

The following truncated output from a Summit series switch displays the software recovery and hardware recovery settings (displayed as Recovery Mode):

```
SysName:          TechPubs Lab
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:1F:A5:71
Recovery Mode:    All
System Watchdog:  Enabled
```

To see the output of `show switch` command for a particular node other than the master, you should log into that node and run the `show switch` command.

If you configure the hardware recovery setting to none, the output displays “Ignore” to indicate that no corrective actions will occur on the switch. “Ignore” appears only if you configure the hardware recovery setting to none.

If you configure the hardware recovery setting to shut down, the output displays “Shutdown” to indicate that the switch will shut down if fault detection occurs. “Shutdown” appears only if you configure the hardware recovery setting to shut down.

If you configure the hardware recovery setting to reset, the output only displays the software recovery mode.

*Clearing the Shutdown State*

If you configure the switch to shut down upon detecting a fault, and the switch enters the shutdown state, you must explicitly clear the shutdown state and reboot for the switch to become functional.

1. Clear the shutdown state with the command:

   ```
   clear sys-recovery-level
   ```

   On a SummitStack, use the command:

   ```
   clear sys-recovery-level slot slot
   ```

   The switch prompts you to confirm this action. The following is a sample confirmation message:

   ```
   Are you sure you want to clear sys-recovery-level? (y/n)
   ```

2. Enter `y` to confirm this action and clear the shutdown state. Enter `n` or press **[Enter]** to cancel this action.

After you clear the shutdown state, use the `reboot` command to bring the switch and ports back online. After rebooting, the switch is operational.

## Configuring Module Recovery--Modular Switches Only

You can configure the MSMs/MMs or I/O modules installed in BlackDiamond X8 and 8800 series switches to take no action, take ports offline in response to errors, automatically reset, shutdown, or if dual MSMs/MMs are installed failover to the other MSM/MM if the switch detects a fault. This enhanced level of recovery detects faults in the ASICs as well as packet buses.

- Configure module recovery with the command:

  ```
  configure sys-recovery-level slot [all | slot_number] [none | reset | shutdown]
  ```

  Where the following is true:

  - **none**—Configures the MSM/MM or I/O module to maintain its current state regardless of the detected fault. The offending MSM/MM or I/O module is not reset. ExtremeXOS logs fault and error messages to the syslog and notifies you that the errors are ignored. This does not guarantee that the module remains operational; however, the switch does not reboot the module.
  - **reset**—Configures the offending MSM/MM or I/O module to reset upon fault detection. ExtremeXOS logs fault, error, system reset, and system reboot messages to the syslog.
  - **shutdown**—Configures the switch to shut down all slots/modules configured for shutdown upon fault detection. On the modules configured for shutdown, all ports in the slot are taken offline in response to the reported errors; however, the MSMs/MMs remain operational for debugging purposes only. You must save the configuration, using the `save configuration` command, for it to take effect. ExtremeXOS logs fault, error, system reset, system reboot, and system shutdown messages to the syslog.

  The default setting is **reset**.

  Depending on your configuration, the switch resets the offending MSM/MM or I/O module if a fault detection occurs. An offending MSM/MM is reset any number of times and is not permanently taken offline. On BlackDiamond X8 and 8800 series switches, an offending I/O module is reset a maximum of five times. After the maximum number of resets, the I/O module is permanently taken offline. For

more information, see Module Recovery Actions--BlackDiamond 8800 Series Switches and BlackDiamond X8 Series Switches Only on page 453.

- You can configure how ExtremeXOS handles a detected fault based on the configuration of the `configure sys-recovery-level slot [all | slot_number] [none | reset | shutdown]` command.
- To configure how ExtremeXOS handles faults, use the `configure sys-health-check all level [normal | strict]` command.
- To view the system health check settings on the switch, use the `show switch` command as described in Displaying the System Health Check Setting--All Platforms on page 446.

### *Confirmation Messages Displayed*

If you configure the hardware recovery setting to either none (ignore) or shutdown, the switch prompts you to confirm this action. The following is a sample shutdown message:

```
Are you sure you want to shutdown on errors? (y/n)
```

- Enter `y` to confirm this action and configure the hardware recovery level.
- Enter `n` or press **[Enter]** to cancel this action.

### *Understanding the Shut Down Recovery Mode*

You can configure the switch to shut down one or more I/O modules upon fault detection by specifying the **shutdown** option.

If you configure one or more slots to shut down and the switch detects a fault, all ports in all of the configured shut down slots are taken offline in response to the reported errors. (MSMs/MMs are available for debugging purposes only.)

Note

On the BlackDiamond 8800 and BlackDiamond X8 chassis, you must save the configuration before the “shutdown” configuration takes effect.

The affected I/O module remains in the shutdown state across additional reboots or power cycles until you explicitly clear the shutdown state. If a module enters the shutdown state, the module actually reboots and the show slot command displays the state of the slot as Initialized; however, the ports are shut down and taken offline. For more information about clearing the shutdown state, see Clearing the Shutdown State on page 451.

### *Messages Displayed at the Startup Screen*

The following truncated sample output shows the startup screen if any of the slots in a modular switch are shut down as a result of the system recovery configuration:

```
The I/O modules in the following slots are shut down: 1,3
Use the "clear sys-recovery-level" command to restore I/O modules
! BD-8810.1 #
```

When an exclamation point (!) appears in front of the command line prompt , it indicates that one or more slots shut down as a result of your system recovery configuration and a switch error.

*Module Recovery Actions--BlackDiamond 8800 Series Switches and BlackDiamond X8 Series Switches Only*

The following table describes the actions module recovery takes based on your module recovery setting.

For example, if you configure a module recovery setting of reset for an I/O module, the module is reset a maximum of five times before it is taken permanently offline.

From left to right, the columns display the following information:

- Module Recovery Setting—This is the parameter used by the configure sys-recovery-level slot command to distinguish the module recovery behavior.
- Hardware—This indicates the hardware that you may have installed in your switch.
- Action Taken—This describes the action the hardware takes based on the module recovery setting.

**Table 59: Module Recovery Actions for the BlackDiamond X8 and 8800 Series Switches**

| Module Recovery Setting | Hardware | Action Taken |
|---|---|---|
| none | | |
| | Single MSM | The MSM remains powered on in its current state.<br>This does not guarantee that the module remains operational; however, the switch does not reboot the module. |
| | Dual MSM | The MSM remains powered on in its current state.<br>This does not guarantee that the module remains operational; however, the switch does not reboot the module. |
| | I/O Module | The I/O module remains powered on in its current state. The switch sends error messages to the log and notifies you that the errors are ignored.<br>This does not guarantee that the module remains operational; however, the switch does not reboot the module. |
| reset | | |
| | Single MSM | Resets the MSM. |
| | Dual MSM | Resets the primary MSM and fails over to the backup MSM. |
| | I/O Module | Resets the I/O module a maximum of five times. After the fifth time, the I/O module is permanently taken offline. |
| shutdown | | |
| | Single MSM | The MSM is available for debugging purposes only (the I/O ports also go down); however, you must clear the shutdown state using the `clear sys-recovery-level` command for the MSM to become operational.<br>After you clear the shutdown state, you must reboot the switch.<br>For more information see, Clearing the Shutdown State on page 456. |

**Table 59: Module Recovery Actions for the BlackDiamond X8 and 8800 Series Switches (continued)**

| Module Recovery Setting | Hardware | Action Taken |
|---|---|---|
| | Dual MSM | The MSMs are available for debugging purposes only (the I/O ports also go down); however, you must clear the shutdown state using the `clear sys-recovery-level` command for the MSM to become operational.<br>After you clear the shutdown state, you must reboot the switch.<br>For more information see, Clearing the Shutdown State on page 456. |
| | I/O Module | Reboots the I/O module. When the module comes up, the ports remain inactive because you must clear the shutdown state using the `clear sys-recovery-level` command for the I/O module to become operational.<br>After you clear the shutdown state, you must reset each affected I/O module or reboot the switch.<br>For more information see, Clearing the Shutdown State on page 456. |

*Displaying the Module Recovery Setting*

- Display the module recovery setting with the command:
  `show slot`

  The `show slot` output includes the shutdown configuration. If you configure the module recovery setting to shut down, the output displays an “E” flag that indicates any errors detected on the slot disables all ports on the slot. The “E” flag appears only if you configure the module recovery setting to "shutdown."

  Note
  If you configure one or more slots for shut down and the switch detects a fault on one of those slots, all of the configured slots enter the shutdown state and remain in that state until explicitly cleared.

  If you configure the module recovery setting to none, the output displays an “e” flag that indicates no corrective actions will occur for the specified MSM/MM or I/O module. The “e” flag appears only if you configure the module recovery setting to none.

**Example**

Here's an example from a BlackDiamond 8810 with slot 2 configured for "shutdown":

```
Slots    Type                Configured          State       Ports  Flags
--------------------------------------------------------------------------------
Slot-1   8900-G96T-c         8900-G96T-c         Operational   96   MB
Slot-2   8900-10G24X-c       8900-10G24X-c       Operational   24   MB   E
Slot-3   8900-40G6X-xm       8900-40G6X-xm       Operational   24   MB
Slot-4   G48Xc               G48Xc               Operational   48   MB
Slot-5   G8Xc                G8Xc                Operational    8   MB
Slot-6                                           Empty          0
Slot-7   G48Te2(PoE)         G48Te2(PoE)         Operational   48   MB
Slot-8   G48Tc               G48Tc               Operational   48   MB
Slot-9   10G4Xc              10G4Xc              Operational    4   MB
```

```
Slot-10                                                 Empty          0
MSM-A    8900-MSM128                                    Operational    0
MSM-B    8900-MSM128                                    Operational    0
Flags : M - Backplane link to Master is Active
B - Backplane link to Backup is also Active
D - Slot Disabled
I - Insufficient Power (refer to "show power budget")
e - Errors on slot will be ignored (no corrective action initiated)
E - Errors on slot will disable all ports on slot
```

**Note**

In ExtremeXOS 11.4 and earlier, if you configure the module recovery setting to none, the output displays an “e” flag that indicates no corrective actions will occur for the specified MSM/MM or I/O module. The “e” flag appears only if you configure the module recovery setting to none.

*Displaying Detailed Module Recovery Information*

- Display the module recovery setting for a specific port on a module, including the current recovery mode with the command:

  `show slot {`*`slot`* `{`**`detail`**`} | `**`detail`** `}`

  In addition to the information displayed with `show slot`, this command displays the module recovery setting configured on the slot.

**Example**

The following truncated output displays the module recovery setting (displayed as Recovery Mode) for the specified slot:

Here is an example of show slot using the same slot 2 as the example above:

```
Slot-2 information:
State:               Operational
Download %:          100
Flags:               MB   E
Restart count:       0 (limit 5)
Serial number:       800264-00-01 0907G-00166
Hw Module Type:      8900-10G24X-c
SW Version:          15.2.0.26
SW Build:            v1520b26
Configured Type:     8900-10G24X-c
Ports available:     24
Recovery Mode:       Shutdown
Debug Data:          Peer=Operational
Flags : M - Backplane link to Master is Active
B - Backplane link to Backup is also Active
D - Slot Disabled
I - Insufficient Power (refer to "show power budget")
e - Errors on slot will be ignored (no corrective action initiated)
E - Errors on slot will disable all ports on slot
```

### *Clearing the Shutdown State*

If you configure one or more modules to shut down upon detecting a fault, and the switch enters the shutdown state, you must explicitly clear the shutdown state and reset the affected modules for the switch to become functional.

1. Clear the shutdown state witht he command:

   ```
   clear sys-recovery-level
   ```

   The switch prompts you to confirm this action. The following is a sample confirmation message:

   ```
   Are you sure you want to clear sys-recovery-level? (y/n)
   ```

2. Enter `y` to confirm this action and clear the shutdown state. Enter `n` or press **[Enter]** to cancel this action.
3. After using the `clear sys-recovery-level` command, you must reset each affected module.
4. If you configured only a few I/O modules to shutdown, reset each affected I/O module as follows:
   a. Disable the slot using the `disable slot` *`slot`* command.
   b. Re-enable the slot using the `enable slot` *`slot`* command.

   **Note**
   You must complete this procedure for each module that enters the shutdown state.

5. If you configured all I/O modules or one or more MSM/MMs to shutdown, use the `reboot` command to reboot the switch and reset all affected I/O modules.

After you clear the shutdown state and reset the affected module, each port is brought offline and then back online before the module and the entire system is operational.

### *Troubleshooting Module Failures*

If you experience an I/O module failure, use the following troubleshooting methods when you can bring the switch offline to solve or learn more about the problem:

- Restart the I/O module.
  a. Run `disable slot` *`slot`*.
  b. Run `enable slot` *`slot`*.

  The I/O module and its associated fail counter is reset. If the module does not restart, or you continue to experience I/O module failure, contact Extreme Networks Technical Support.
- Run diagnostics.
  a. Run `run diagnostics normal` *`slot`*.

  This runs diagnostics on the offending I/O module to ensure that you are not experiencing a hardware issue.
  b. If the module continues to enter the failed state, please contact Extreme Networks Technical Support.

  For more information about switch diagnostics, see Performing Switch Diagnostics on page 439.
- If you experience an MSM/MM failure, contact Extreme Networks Technical Support.

# Using ELSM

Extreme Link Status Monitoring (ELSM) is an Extreme Networks proprietary protocol that monitors network health by detecting CPU and remote link failures.

ELSM is available only on Extreme Networks devices and operates on a point-to-point basis. You can configure ELSM on the ports that connect to other network devices and on both sides of the peer connection.

## About ELSM

ELSM monitors network health by exchanging various hello messages between two ELSM peers.

ELSM uses an open-ended protocol, which means that an ELSM-enabled port expects to send and receive hello messages from its peer. The Layer 2 connection between ports determines the peer connection. Peers can be either directly connected or separated by one or more hubs. If there is a direct connection between peers, they are considered neighbors.

If ELSM detects a failure, the ELSM-enabled port responds by blocking traffic on that port. For example, if a peer stops receiving messages from its peer, ELSM brings down that connection by blocking all incoming and outgoing data traffic on the port and notifying applications that the link is down.

In some situations, a software or hardware fault may prevent the CPU from transmitting or receiving packets, thereby leading to the sudden failure of the CPU. If the CPU is unable to process or send packets, ELSM isolates the connections to the faulty switch from the rest of the network. If the switch fabric sends packets during a CPU failure, the switch may appear healthy when it is not. For example, if hardware forwarding is active and software forwarding experiences a failure, traffic forwarding may continue. Such failures can trigger control protocols such as *ESRP (Extreme Standby Router Protocol)* or Ethernet Automatic Protection Switching (EAPS) to select different devices to resume forwarding. This recovery action, combined with the CPU failure, can lead to loops in a Layer 2 network.

Configuring ELSM on Extreme Networks devices running ExtremeXOS is backward compatible with Extreme Networks devices running ExtremeWare.

## ELSM Hello Messages

ELSM uses two types of hello messages to communicate the health of the network to other ELSM ports. The following describes the hello messages:

- Hello+ — The ELSM-enabled port receives a hello message from its peer and no problem is detected.
- Hello- — The ELSM-enabled port has not received a hello message from its peer.

In addition to the ELSM port states described in the next section, ELSM has hello transmit states. The hello transmit states display the current state of transmitted ELSM hello messages and can be one of the following:

- HelloRx(+)—The ELSM-enabled port is up and receives Hello+ messages from its peer. The port remains in the HelloRx+ state and restarts the HelloRx timer each time it receives a Hello+ message. If the HelloRx timer expires, the hello transmit state enters HelloRx(-). The HelloRx timer is 6 * hello timer, which by default is 6 seconds.
- HelloRx(-)—The ELSM-enabled port either transitions from the initial ELSM state or is up but has not received hello messages because there is a problem with the link or the peer is missing.

For information about displaying ELSM hello messages and hello transmit states, see Displaying ELSM Information on page 464.

## ELSM Port States

Each ELSM-enabled port exists in one of the following states:

- Up—Indicates a healthy remote system and this port is receiving Hello+ messages from its peer.

  If an ELSM-enabled port enters the Up state, the up timer begins. Each time the port receives a Hello + message from its peer, the up timer restarts and the port remains in the Up state. The up timer is 6 * hello timer, which by default is 6 seconds.
- Down—Indicates that the port is down, blocked, or has not received Hello+ messages from its peer.

  If an ELSM-enabled port does not receive a hello message from its peer before the up timer expires, the port transitions to the Down state. When ELSM is down, data packets are neither forwarded nor transmitted out of that port.
- Down-Wait—Indicates a transitional state.

  If the port enters the Down state and later receives a Hello+ message from its peer, the port enters the Down-Wait state. If the number of Hello+ messages received is greater than or equal to the hold threshold (by default, two messages), the port transitions to the Up state. If the number of Hello+ messages received is less than the hold threshold, the port enters the Down state.
- Down-Stuck—Indicates that the port is down and requires user intervention.

  If the port repeatedly flaps between the Up and Down states, the port enters the Down-Stuck state. Depending on your configuration, there are two ways for a port to transition out of this state:

  By default, automatic restart is enabled, and the port automatically transitions out of this state. For more information, see the `enable elsm ports` *`port_list`* **`auto-restart`** command.

  If you disabled automatic restart, and the port enters the Down-Stuck state, you can clear the stuck state and enter the Down state by using one of the following commands:
  - `clear elsm ports` *`port_list`* **`auto-restart`**
  - `enable elsm ports` *`port_list`* **`auto-restart`**

**Note**

If you reboot the peer switch, its ELSM-enabled peer port may enter the Down-Stuck state. If this occurs, clear the stuck state using one of the following commands:

- `clear elsm ports` *`port_list`* **`auto-restart`**
- `enable elsm ports` *`port_list`* **`auto-restart`**

For information about displaying ELSM port states, see Displaying ELSM Information on page 464.

## Link States

The state of the link between ELSM-enabled (peer) ports is known as the *link state*. The link state can be one of the following:

- Ready—Indicates that the port is enabled but there is no physical link
- Active—Indicates that the port is enabled and the physical link is up

- View the state of the link between the peer ports using the commands:

  `show elsm ports` **`all`** | *`port_list`*

  `show ports` {*`port_list`*} **`information`** {**`detail`**}

  If you use the `show elsm ports` **`all`** | *`port_list`* command, the Link State row displays link state information.

  If you use the `show ports` {*`port_list`*} information command, the Link State column displays link state information.

  If you use the `show ports` {*`port_list`*} information command and specify the **`detail`** option, the ELSM Link State row displays ELSM link state information. For more information, see ELSM Link States on page 459.

For more information about these show commands, see Displaying ELSM Information on page 464.

## ELSM Link States

The state of the ELSM logical link is known as the ELSM link state. The ELSM link state can be one of the following:

- ELSM is enabled and the ELSM peer ports are up and communicating
- ELSM is enabled but the ELSM peer ports are not up or communicating
- ELSM is disabled

- View the current state of ELSM on the switch using the commands:
  - `show elsm`

  `show elsm ports` **`all`** | *`port_list`*

  `show ports` {*`port_list`*} **`information`** {**`detail`**}

  If you use the `show elsm` commands, the following terms display the ELSM link state:

  - Up—Indicates that ELSM is enabled and the ELSM peer ports are up and communicating; the ELSM link state is up. In the up state, the ELSM-enabled port sends and receives hello messages from its peer.
  - Down—Indicates that ELSM is enabled, but the ELSM peers are not communicating; the ELSM link state is down. In the down state, ELSM transitions the peer port on this device to the down state. ELSM blocks all incoming and outgoing switching traffic and all control traffic except ELSM PDUs.

  If ELSM is disabled, the switch does not display any information.

If you use the `show ports {port_list} information {detail}` command, the following columns display the current state of ELSM on the switch:

- Flags
  - L—Indicates that ELSM is enabled on the switch
  - - —Indicates that ELSM is disabled on the switch
- ELSM
  - up—Indicates that ELSM is enabled and the ELSM peer ports are up and communicating; the ELSM link state is up. In the up state, the ELSM-enabled port sends and receives hello messages from its peer.
  - dn—Indicates that ELSM is enabled, but the ELSM peers are not communicating; the ELSM link state is down. In the down state, ELSM transitions the peer port on this device to the down state. ELSM blocks all incoming and outgoing switching traffic and all control traffic except ELSM PDUs.
  - - —Indicates that ELSM is disabled on the switch.

If you specify the optional **detail** parameter, the following ELSM output is called out in written explanations versus displayed in a tabular format:

- ELSM Link State (displayed only if ELSM is enabled on the switch)
  - Up—Indicates that ELSM is enabled and the ELSM peer ports are up and communicating; the ELSM link state is up. In the up state, the ELSM-enabled port sends and receives hello messages from its peer.
  - Down—Indicates that ELSM is enabled, but the ELSM peers are not communicating; the ELSM link state is down. In the down state, ELSM transitions the peer port on this device to the down state. ELSM blocks all incoming and outgoing switching traffic and all control traffic except ELSM PDUs.
- ELSM
  - Enabled—Indicates that ELSM is enabled on the switch
  - Disabled—Indicates that ELSM is disabled on the switch

For more information about these show commands, see Displaying ELSM Information on page 464.

## ELSM Timers

To determine whether there is a CPU or link failure, ELSM requires timer expiration between the ELSM peers. Depending on the health of the network, the port enters different states when the timers expire. For more information about the ELSM port states, see ELSM Port States on page 458.

The following table describes the ELSM timers. Only the hello timer is user-configurable; all other timers are derived from the hello timer. This means that when you modify the hello timer, you also modify the values for down, up, and HelloRx timers.

**Table 60: ELSM Timers**

| Timer | Description |
|---|---|
| Hello | The ELSM hello timer is the only user-configurable timer and specifies the time in seconds between consecutive hello messages. The default value is 1 second, and the range is 100 milliseconds to 255 seconds. |
| Down | The ELSM down timer specifies the time it takes the ELSM port to cycle through the following states:<br>• Down—Indicates that the port is down, blocked, or has not received Hello+ messages from its peer.<br>• Down-Wait—Indicates a transitional state.Up—Indicates a healthy remote system and this port is receiving Hello+ messages from its peer.<br>By default, the down timer is (2 + hold threshold) * hello timer, which is 4 seconds. If the hold threshold is set to 2 and the hello timer is set to 1 second, it takes 4 seconds for the ELSM port receiving messages to cycle through the states.<br>After the down timer expires, the port checks the number of Hello+ messages against the hold threshold. If the number of Hello+ messages received is greater than or equal to the configured hold threshold, the ELSM receive port moves from the Down-Wait state to the Up state.<br>If the number of Hello+ messages received is less than the configured hold threshold, the ELSM receive port moves from the Down-Wait state back to the Down state and begins the process again. |
| Up | The ELSM up timer begins when the ELSM-enabled port enters the UP state. Each time the port receives a Hello+ message, the timer restarts.<br>Up timer is the UpTimer threshold * hello timer. It is configurable and the range is 3 to 60 seconds.<br>By default, the UpTimer threshold is 6. Therefore, the default up timer is 6 seconds (6*1). |
| HelloRx | The ELSM HelloRx timer specifies the time in which a hello message is expected. If the port does not receive a hello message from its peer, there is the possibility of a CPU or link failure.<br>By default the HelloRx timer 6 * hello timer, which is 6 seconds. |

## Configuring ELSM on a Switch

This section describes the commands used to configure ELSM on the switch.

Note

ELSM and mirroring are mutually exclusive. You can enable either ELSM, or mirroring, but not both.

Note

ELSM is not to be configured on *MLAG (Multi-switch Link Aggregation Group)* ports at either end of an MLAG.

### *Enabling ELSM*

ELSM works between two connected ports (peers), and each ELSM instance is based on a single port. The Layer 2 connection between the ports determines the peer. You can have a direct connection between the peers or hubs that separate peer ports. In the first instance, the peers are also considered neighbors. In the second instance, the peer is not considered a neighbor.

- Enable ELSM on one or more ports with the command:

  `enable elsm ports` *`port_list`*

When you enable ELSM on a port, ELSM immediately blocks the port and it enters the Down state. When the port detects an ELSM-enabled peer, the peer ports exchange ELSM hello messages. At this point, the ports enter the transitional Down-Wait state. If the port receives Hello+ messages from its peer and does not detect a problem, the peers enter the Up state. If a peer detects a problem or there is no peer port configured, the port enters the Down state.

For more information about the types of ELSM hello messages, see ELSM Hello Messages on page 457. For information about configuring the ELSM hello timer, see the next section.

### *Configuring the ELSM Hello Timer*

The ELSM hello timer is the only user-configurable timer and specifies the time in seconds between consecutive hello messages. The default value is 1 second, and the range is 1 to 128 seconds. Although other timers rely on the hold timer for their values, you do not explicitly configure the down, up, or HelloRx timers. If you modify the hello timer on one port, we recommend that you use the same hello timer value on its peer port.

A high hello timer value can increase the time it takes for the ELSM-enabled port to enter the Up state. The down timer is (2 + hold threshold) * hello timer. Assuming the default value of 2 for the hold threshold, configuring a hello timer of 128 seconds creates a down timer of (2 + 2) 128, or 512 seconds. In this scenario it would take 512 seconds for the port to transition from the Down to the Up state.

- Configure the ELSM hello timer with the command:

  `configure elsm ports` *`port_list`* **`hellotime`** *`hello_time`*

### *Configuring the ELSM Hold Threshold*

The ELSM hold threshold determines the number of Hello+ messages the ELSM peer port must receive to transition from the Down-Wait state to the Up state. For example, a threshold of 1 means the ELSM port must receive at least one Hello+ message to transition from the Down-Wait state to the Up state. The default is two messages, and the range is one to three messages.

After the down timer expires, the port checks the number of Hello+ messages against the hold threshold. If the number of Hello+ messages received is greater than or equal to the configured hold threshold, the ELSM receive port moves from the Down-Wait state to the Up state.

If the number of Hello+ messages received is less than the configured hold threshold, the ELSM receive port moves from the Down-Wait state back to the Down state and begins the process again.

If you modify the hold threshold on one port, we recommend that you use the same hold threshold value on its peer port.

You must configure the hold threshold on a per-port basis, not on a per-switch basis.

- Configure the ELSM hold threshold with the command:

  `configure elsm ports` *port_list* **hold-threshold** *hold_threshold*

*Configure Automatic Restart*

You must explicitly configure automatic restart on each ELSM-enabled port; this is not a global configuration. By default, ELSM automatic restart is enabled. If an ELSM-enabled port goes down, ELSM bypasses the Down-Stuck state and automatically transitions the down port to the Down state, regardless of the number of times the port goes up and down. For information about the port states, see ELSM Port States on page 458.

If you disable ELSM automatic restart, the ELSM-enabled port can transition between the following states multiple times: Up, Down, and Down-Wait. When the number of state transitions is greater than or equal to the sticky threshold, the port enters the Down-Stuck state. The ELSM sticky threshold specifies the number of times a port can transition between the Up and Down states. The sticky threshold is not user-configurable and has a default value of 1. That means a port can transition only one time from the Up state to the Down state. If the port attempts a subsequent transition from the Up state to the Down state, the port enters the Down-Stuck state.

The following user events clear or re-set the sticky threshold:

- Enabling automatic restart on the port using the following command:

  `enable elsm ports` *port_list* **auto-restart**
- Clearing the port that is in the Down-Stuck state using the following command:

  `clear elsm ports` *port_list* **auto-restart**
- Disabling and re-enabling the port using the following commands:
  - `disable ports [`*port_list*`|`**all**`]`
  - `enable ports [`*port_list*`|`**all**`]`
- Disabling ELSM on the port using the following command:

  `disable elsm ports` *port_list*

We recommend that you use the same automatic restart configuration on each peer port.

- If the port enters the Down-Stuck state, you can clear the stuck state and enter the Down state by using one of the following commands:

  `clear elsm ports` *port_list* **auto-restart**

  `enable elsm ports` *port_list* **auto-restart**

  If you use the `enable elsm ports` *port_list* **auto-restart** command, automatic restart is always enabled; you do not have to use the `clear elsm ports` *port_list* **auto-restart** command to clear the stuck state.
- Disable automatic restart with the command:

  `disable elsm ports` *port_list* **auto-restart**
- Re-enable automatic restart with the command:

  `enable elsm ports` *port_list* **auto-restart**

  We recommend that you use the same automatic restart configuration on each peer port.

*Disabling ELSM*

- Disable ELSM on one or more ports with the command:

  `disable elsm ports` *`port_list`*

When you disable ELSM on the specified ports, the ports no longer send ELSM hello messages to their peers and no longer maintain ELSM states.

## Displaying ELSM Information

- Display summary information for all of the ELSM-enabled ports on the switch with the command:

  `show elsm`

  This command displays in a tabular format the operational state of ELSM on the configured ports.

  If ports are configured for ELSM (ELSM is enabled), the switch displays the following information:

  - Port—The port number of the ELSM-enabled port.
  - ELSM State—The current state of ELSM on the port. For information about the port states, see ELSM Port States on page 458.
  - Hello time—The current value of the hello timer, which by default is 1 second. For information about configuring the hello timer, see Configuring the ELSM Hello Timer on page 462.

  If no ports are configured for ELSM (ELSM is disabled), the switch does not display any information.
- Display detailed information for one or more ELSM-enabled ports on the switch.

  `show elsm ports` **`all`** | *`port_list`*

  In addition to the port, ELSM state, and hello timer information, this command displays in a tabular format the following:

  - Link State—The state of the link between ELSM-enabled ports. For information about the link states, see Link States on page 459.
  - ELSM Link State—The current state of the ELSM logical link on the switch. For more information, see ELSM Link States on page 459.
  - Hello Transmit State—The current state of ELSM hello messages being transmitted.
  - Hold Threshold—The number of Hello+ messages required by the ELSM-enabled port to transition from the Down-Wait state to the Up state within the hold threshold.
  - UpTimer Threshold—The number of hello times that span without receiving Hello+ packets before a port changes its ELSM state from Up to Down.
  - Auto Restart—The current state of ELSM automatic restart on the port.
  - Sticky Threshold—The number of times a port can transition between the Up and Down states. The sticky threshold is not user-configurable and has a default value of 1.
  - Sticky Threshold Counter—The number of times the port transitions from the Up state to the Down state.
  - Down Timeout—The actual waiting time (msecs or secs) before a port changes its ELSM state from Down to Up after receiving the first Hello+ packet. It is equal to [Hello Time * (Hold Threshold+2)].
  - Up Timeout—The actual waiting time (msecs or secs) before a port changes its ELSM state from Up to Down after receiving the last Hello+ packets. It is equal to [Hello Time * UpTimer Threshold].

The remaining output displays counter information.

- Use the counter information to determine the health of the ELSM peers and how often ELSM has gone up or down.

  The counters are cumulative.

  - Rx Hello+—The number of Hello+ messages received by the port.
  - Rx Hello- —The number of Hello- messages received by the port.
  - Tx Hello+—The number of Hello+ messages sent by the port.
  - Tx Hello- —The number of Hello- messages sent by the port.
  - ELSM Up/Down Count—The number of times ELSM has been up or down.

- Display summary port information in a tabular format for one or more ELSM-enabled ports with the command:

  `show ports {port_list} information {detail}`

  This command displays the following ELSM information:

  - Flags
    - L—ELSM is enabled on the switch
    - - —ELSM is disabled on the switch
  - ELSM (For more information, see ELSM Link States on page 459.)
    - up—ELSM is enabled and the ELSM link state is up.
    - dn—ELSM is enabled and the ELSM link state is down.
    - - —ELSM is disabled on the switch.

- Display summary port information called out in written explanations versus displayed in a tabular format, use the **detail** parameter.

  `show ports {port_list} information {detail}`

  This command displays the following ELSM information:

  - ELSM Link State (displayed only if ELSM is enabled on the switch). For more information, see ELSM Link States on page 459.
    - Up—ELSM is enabled and the ELSM link state is up.
    - Down—ELSM is enabled and the ELSM link state is down.
  - Link State—The state of the link between ELSM-enabled ports. For information about the link states, see Link States on page 459.
  - ELSM
    - Enabled—ELSM is enabled on the switch.
    - Disabled—ELSM is disabled on the switch.

### *Clearing ELSM Counters*

Before clearing the ELSM counters, you should use the `show elsm` and `show elsm ports` commands to view the ELSM information.

- Clear only the ELSM-related counters gathered by the switch using the command:

  `clear elsm {ports port_list} counters`

  You can also use the `clear counters` command, which clears all of the counters on the device, including those associated with ELSM.

## Using ELSM with Layer 2 Control Protocols

You can use ELSM with Layer 2 control protocols such as *STP (Spanning Tree Protocol)*, *ESRP*, EAPS, and so on to improve the recovery of Layer 2 loops in the network.

ELSM detects remote link failures if the established link is through a Layer 2 transport cloud and the ELSM endpoints traverse the Layer 2 cloud in a point-to-point fashion, as shown in the following figure.

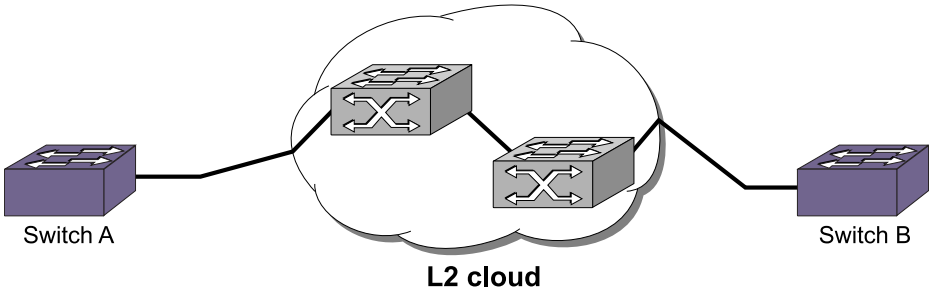



**Figure 63: ELSM Through a Layer 2 Cloud**

A sudden failure of the switch CPU may cause the hardware to continue forwarding traffic. For example, ESRP may select a second device for forwarding traffic thereby creating a Layer 2 loop in the network. If you configure ELSM and the CPU fails, ELSM closes the connection to the faulty device to prevent a loop.

## ELSM Configuration Example

The following example configures ELSM on two ports connected directly to each other and assumes the following:

Note
In the following sample configurations, any lines marked (Default) represent default settings and do not need to be explicitly configured.

*Switch A Configuration*

- ELSM-enabled port—Port 1
- Hello timer—2 seconds
- Hello threshold—2 hello messages

```
enable elsm ports 1
configure elsm ports 1 hellotime 2
configure elsm ports 1 hold-threshold 2 (Default)
```

*Switch B Configuration*

- ELSM-enabled port—Slot 2, port 1
- Hello timer—2 seconds
- Hello threshold—2 hello messages

```
enable elsm ports 2:1
configure elsm ports 2:1 hellotime 2
configure elsm ports 2:1 hold-threshold 2 (Default)
```

After you enable ELSM on the ports, the peers exchange hello messages with each other as displayed in the following figure.

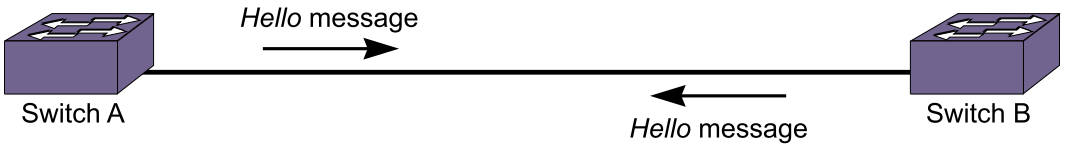



**Figure 64: Extreme Networks Switches with ELSM-Enabled Ports Exchanging Hello Messages**

To configure ELSM with *LAG (Link Aggregation Group)*, enable ELSM on each port member of the LAG, as shown in the following example:

```
enable sharing 7 grouping 7-9, 41 algorith address-based L2
enable elsm port 7
enable elsm port 8
enable elsm port 9
enable elsm port 41

X670-48x.23 # sh elsm
Port ELSM State Hello Time
==== ========== ==========
7    Up         1 (secs)
8    Down       1 (secs)
9    Down       1 (secs)
41   Up         1 (secs)
```

## Viewing Fan Information

You can view detailed information about the fans installed in your switch. Depending on your switch model, different information may be displayed.

- View detailed information about the health of the fans with the command: `show fans`

  The switch collects and displays the following fan information:

  - State—The current state of the fan. Options are:
    - Empty: There is no fan installed.
    - Failed: The fan failed.
    - Operational: The fan is installed and working normally.
  - NumFan—The number of fans in the fan tray.
  - Fan Name, displayed as Fan-1, Fan-2, and so on (modular switches also include a description of the location, for example, Upper or Upper-Right)—Specifies the individual state for each fan in a fan tray and its current speed in revolutions per minute (rpm).

  On modular switches, the output also includes the following information:

  - PartInfo—Information about the fan tray, including the:
    - Serial number—A collection of numbers and letters, that make up the serial number of the fan. This is the first series of numbers and letters in the display.
    - Part number—A collection of numbers and letters, that make up the part number of the fan. This is the second series of numbers and letters in the display.
    - Revision—The revision number of the fan.
  - Odometer—Specifies the power-on date and how long the fan tray has been operating since it was first powered-on.

## Fan Behavior on X440-24t Model Switches

Fan behavior on X440-24t model switches may vary based on the CPLD version of hardware.

CPLD version 3 maps to the following X440 models and corresponding 800k Revisions:

**Table 61: CPLD 3 with X440 and 800k Revisions**

| X440 Model | 800K Revision | CPLD 3 Version |
|---|---|---|
| Summit X440-24t | 800471-00 | Revision 6 or lower |
| Summit X440-24t-10G | 800475-00 | Revision 5 or lower |
| Summit X440-L2-24t | 800526-00 | Revision 1 or lower |

CPLD 5 version maps to the following X440 models and corresponding 800k Revisions:

**Table 62: CPLD 5 with X440 and 800k Revisions**

| X440 Model | 800K Revision | CPLD 5 Version |
|---|---|---|
| Summit X440-24t | 800471-00 | Revision 7 or higher |
| Summit X440-24t-10G | 800475-00 | Revision 6 or higher |
| Summit X440-L2-24t | 800526-00 | Revision 2 or higher |

Fan behavior on X440-24t switches with CPLD version 3:

- Fan will display speed as 11000 RPM if Fan status is good. Fan will display speed as 0 RPM if Fan status is bad.
- LED will be turned ON if Fan is ON and will get turned OFF if Fan is OFF.

Fan behavior on X440-24t switches with CPLD version 5:

- Fan will display speed as 11000 RPM if Fan is running. Fan will display speed as 0 RPM if Fan is not running.
- LED will be turned ON if Fan status is good and will get turned OFF if Fan is bad.

# Viewing the System Temperature

Depending on your switch model, you can view the temperature in Celsius of the I/O modules, management modules, power controllers, power supplies, and fan trays installed in your switch. In addition, depending on the software version running on your switch, additional or different temperature information might be displayed.

You can view the system temperature with the command: `show temperature`

## System Temperature Output

*Modular Switches and SummitStack Only*

On a modular switch, the output includes the current temperature and operating status of the I/O modules, management modules, and power controllers.

On a SummitStack, the output includes the current temperature and operating status of all active nodes and their option cards (if any).

The following output shows a sample display of the current temperature and operating status of the installed modules and power controllers:

```
BD-8810.7 # show temperature
Field Replaceable Units               Temp (C)   Status   Min  Normal  Max
-------------------------------------------------------------------------
Slot-1         : 8900-10G24X-c          37.00    Normal   -10    0-55  65
Slot-2         : 10G8Xc                 33.00    Normal   -10    0-50  60
Slot-3         :
Slot-4         :
Slot-5         : G8Xc                   31.50    Normal   -10    0-50  60
Slot-6         :
Slot-7         :
Slot-8         : 10G4Xa                 26.00    Normal   -10    0-50  60
Slot-9         :
Slot-10        : 8900-G96T-c            38.50    Normal   -10    0-55  65
MSM-A          : 8900-MSM128            31.00    Normal   -10    0-55  65
MSM-B          : 8900-MSM128            31.00    Normal   -10    0-55  65
PSUCTRL-1      :                        34.04    Normal   -10    0-50  60
PSUCTRL-2      :                        35.79    Normal   -10    0-50  60
```

The switch monitors the temperature of each component and generates a warning if the temperature exceeds the normal operating range. If the temperature exceeds the minimum/maximum limits, the switch shuts down the overheated module.

The following output shows a sample display from a SummitStack.

```
Slot-3 Stack.1 # show temperature
Field Replaceable Units               Temp (C)   Status   Min  Normal  Max
-------------------------------------------------------------------------
Slot-1         :
Slot-2         : SummitX                34.50    Normal   -10    0-54  59
Slot-3         : SummitX                36.50    Normal   -10    0-66  67
Slot-4         :
Slot-5         :
Slot-6         :
Slot-7         :
Slot-8         :
Slot-3 Stack.2 #
```

*Summit Family Switches Only*

On Summit family switches, the output includes the current temperature and operating status of the switch. The following shows a sample display from a Summit switch:

```
# show temperature
Field Replaceable Units               Temp (C)   Status   Min  Normal  Max
-------------------------------------------------------------------------
Switch         : SummitX                30.00    Normal   -10    0-52  57
```

The switch monitors its temperature and generates a warning if the temperature exceeds the normal operating range. If the temperature exceeds the maximum limit, the `show switch` output indicates the switch in an OPERATIONAL (Overheat) mode, and the `show temperature` output indicates an error state due to overheat.

### Power Supply Temperature--Modular Switches Only

- View the current temperature of the power supplies installed in BlackDiamond X8 or 8800 series switches with the command:

  `show power {ps_num} {detail}`

  The following is sample output of temperature information:

```
PowerSupply 1 information:
...
Temperature:    30.1 deg C
...
```

## Using the Event Management System/Logging

We use the general term *event* for any type of occurrence on a switch that could generate a log message or require an action.

For example, a link going down, a user logging in, a command entered on the command line, or the software executing a debugging statement, are all events that might generate a log message. The system for saving, displaying, and filtering events is called the *EMS (Event Management System)*. With EMS, you have many options about which events generate log messages, where the messages are sent, and how they are displayed.

Using EMS you can:

- Send event messages to a number of logging targets (for example, syslog host and NVRAM).
- Filter events per target, by:
  - Component, subcomponent, or specific condition (for example, *BGP (Border Gateway Protocol)* messages, IGMP.Snooping messages, or the IP.Forwarding.SlowPathDrop condition)
  - Match expression (for example, any messages containing the string “user5”)
  - Matching parameters (for example, only messages with source IP addresses in the 10.1.2.0/24 subnet)
  - Severity level (for example, only messages of severity critical, error, or warning)
- Change the format of event messages (for example, display the date as “12-May-2005” or “2005-05-12”).
- Display log messages in real time and filter the messages that are displayed, both on the console and from Telnet sessions.
- Display stored log messages from the memory buffer or NVRAM.
- Upload event logs stored in memory buffer or NVRAM to a TFTP server.
- Display counts of event occurrences, even those not included in filter.
- Display debug information using a consistent configuration method.

EMS supports IPv6 as a parameter for filtering events.

### Sending Event Messages to Log Targets

You can specify seven types of targets to receive log messages:

- Console display
- Current session (Telnet or console display)

- Memory buffer (can contain 200 to 20,000 messages)
- NVRAM (messages remain after reboot)
- Primary MSM/MM (for modular systems) or node (for SummitStack)
- Backup MSM/MM (for modular systems) or node (for SummitStack)
- Syslog server, supports IPv6 address family as of ExtremeXOS 16.2

The first six targets exist by default; but before enabling any syslog host, you must add the host's information to the switch using the `configure syslog` command. Extreme Networks ExtremeManagement or Ridgeline can be a syslog target.

By default, the memory buffer and NVRAM targets are already enabled and receive messages.

- To start sending messages to the targets, use the following command:
  `enable log target [console | memory-buffer | nvram | primary-msm | primary-node| backup-msm | backup-node| session | syslog [all | ipaddress | ipPort] {vr vr_name} [local0...local7]]]`

  After you enable this feature, the target receives the messages for which it is configured. See Target Configuration on page 472 for information on viewing the current configuration of a target. The memory buffer can contain only the configured number of messages, so the oldest message is lost when a new message arrives, when the buffer is full.
- To stop sending messages to the target, use the following command:
  `disable log target [console | memory-buffer | nvram | primary-msm | primary-node | backup-msm | backup-node | session | syslog [all | ipaddress | ipPort] {vr vr_name} [local0 ... local7]]]`

Refer to your UNIX documentation for more information about the syslog host facility.

### *Primary and Backup Systems--Modular Switches and SummitStack Only*

A system with dual MSMs/MMs (modular switches) or primary and backup nodes (SummitStack) keeps the two systems synchronized by executing the same commands on both.

However, the full data between the *EMS* servers is not synchronized. The reason for this design decision is to make sure that the control channel is not overloaded when a high number of log messages are generated.

To capture events generated by the primary node onto the backup node, two additional targets are shown in the target commands—one called primary-msm (modular switches) or primary-node (SummitStack) and one called backup-msm (modular switches) or backup-node (SummitStack). The first target is active only on the non-primary (backup) EMS server and is used to send matching events to the primary EMS server. The other target is active only on the primary EMS server and is used to send matching events to all other EMS servers.

If the condition for the backup target is met by a message generated on the primary node, the event is sent to the backup node. When the backup node receives the event, it detects if any of the local targets (NVRAM, memory, or console) are matched. If so that event gets processed. The session and syslog targets are disabled on the backup node, as they are handled on the primary. If the condition for the primary target is met by a message generated on the backup, the event is sent to the primary node.

Note that the backup target is active only on the primary node, and the primary target is active only on the backup node.

## Filtering Events Sent to Targets

Not all event messages are sent to every enabled target. Each target receives only the messages for which it is configured.

### *Target Configuration*

To specify the messages to send to an enabled target, you can set a message severity level, a filter name, and a match expression. These items determine which messages are sent to the target. You can also configure the format of the messages in the targets. For example, the console display target is configured to get messages of severity info and greater, the NVRAM target gets messages of severity warning and greater, and the memory buffer target gets messages of severity debug-data and greater. All the targets are associated by default with a filter named DefaultFilter that passes all events at or above the default severity threshold. All the targets are also associated with a default match expression that matches any messages (the expression that matches any message is displayed as Match : (none) from the command line). And finally, each target has a format associated with it.

- To display the current log configuration of the targets, use the following command:

  `show log configuration target {`**`console`** | **`memory-buffer`** | **`nvram`** | **`primary-msm`** | **`primary-node`** | **`backup-msm`** | **`backup-node`** | **`session`** | **`syslog`** {*`ipaddress`* | *`ipPort`* | **`vr`** *`vr_name`*} {[**`local0...local7`** ]}}

- To configure a target, you use specific commands for severity, filters, and formats, use the following command:

  `configure log target [`**`console`** | **`memory-buffer`** | **`nvram`** | **`primary-msm`** | **`primary-node`** | **`backup-msm`** | **`backup-node`** | **`session`** | **`syslog`** [**`all`** | *`ipaddress`* | *`ipPort`* {**`vr`** *`vr_name`*} [**`local0...local7`** ]]] {**`severity`** *`severity`* {**`only`**}}+

  In addition, you can configure the source IP address for a syslog target. Configuring the source IP address allows the management station or syslog server to identify from which switch it received the log messages.

- To configure the source IP address for a syslog target, use the following command.

  `configure log target syslog [`**`all`** | *`ipaddress`* | *`ipPort`*] {**`vr`** *`vr_name`*} {**`local0...local7`**} **`from`** *`source-ip-address`*

### *Severity*

Messages are issued with one of the following severity levels: Critical, Error, Warning, Notice, Info, Debug-Summary, Debug-Verbose, or Debug-Data, which are described in the table below.

When a message is sent to a syslog target, the severity is mapped to a corresponding syslog priority value (see RFC 3164).

The three severity levels for extended debugging—Debug-Summary, Debug-Verbose, and Debug-Data—require that log debug mode be enabled (which may cause a performance degradation).

See Displaying Debug Information on page 483 for more information about debugging.

**Table 63: Severity Levels Assigned by the Switch**

| Level | Description |
|---|---|
| Critical | A serious problem has been detected that is compromising the operation of the system; the system cannot function as expected unless the situation is remedied. The switch may need to be reset. |
| Error | A problem has been detected that is interfering with the normal operation of the system; the system is not functioning as expected. |
| Warning | An abnormal condition, not interfering with the normal operation of the system, has been detected that indicate that the system or the network in general may not be functioning as expected. |
| Notice | A normal but significant condition has been detected, which signals that the system is functioning as expected. |
| Info (Informational) | A normal but potentially interesting condition has been detected, which signals that the system is functioning as expected; this level simply provides potentially detailed information or confirmation. |
| Debug-Summary | A condition has been detected that may interest a developer seeking the reason underlying some system behavior. |
| Debug-Verbose | A condition has been detected that may interest a developer analyzing some system behavior at a more verbose level than provided by the debug summary information. |
| Debug-Data | A condition has been detected that may interest a developer inspecting the data underlying some system behavior. |

#### Configuring Severity Level

You can use more than one command to configure the severity level of the messages sent to a target.

- The most direct way to set the severity level of all the sent messages is to use the following command:

  `configure log target [`**`console`** | **`memory-buffer`** | **`nvram`** | **`primary-msm`** | **`primary-node`** | **`backup-msm`** | **`backup-node`** | **`session`** | **`syslog`** [**`all`** | *`ipaddress`* | *`ipPort`* {**`vr`** *`vr_name`*} [**`local0...local7`** ]]] {**`severity`** *`severity`* {**`only`**}}`

  When you specify a severity level, messages of that severity level and greater are sent to the target. If you want only those messages of the specified severity to be sent to the target, use the keyword **`only`**. For example, specifying **`severity warning`** will send warning, error, and critical messages to the target, but specifying **`severity warning only`** sends only warning messages.

- You can also use the following command to configure severity levels, which associate a filter with a target:

  `configure log target [`**`console`** | **`memory-buffer`** | **`primary-msm`** | **`primary-node`** | **`backup-msm`** | **`backup-node`** | **`nvram`** | **`session`** | **`syslog`** [**`all`** | *`ipaddress`* | *`ipPort`* {**`vr`** *`vr_name`*} [**`local0...local7`**]]] **`filter`** *`filter-name`*{**`severity`** *`severity`* {**`only`**}}`

When you specify a severity level as you associate a filter with a target, you further restrict the messages reaching that target. The filter may allow only certain categories of messages to pass. Only the messages that pass the filter and then pass the specified severity level reach the target.

Finally, you can specify the severity levels of messages that reach the target by associating a filter with a target. The filter can specify exactly which message it will pass. Constructing a filter is described in Filtering By Components and Conditions on page 475.

*Components and Conditions*

The event conditions detected by ExtremeXOS are organized into components and subcomponents.

- To get a listing of the components and subcomponents in your release of ExtremeXOS, use the following command:

  `show log components {event component } {version}`

  For example, to get a list of the components and subcomponents in your system, use the following command: `show log components`

  The following is partial output from this command:

```
Severity
Component           Title                                          Threshold
------------------- ---------------------------------------------- -------------
...
...
STP                 Spanning-Tree Protocol (STP)                   Error
InBPDU              STP In BPDU subcomponent                       Warning
OutBPDU             STP Out BPDU subcomponent                      Warning
System              STP System subcomponent                        Error
...
...
```

  The display above lists the components, subcomponents, and the severity threshold assigned to each. In *EMS*, you use a period (.) to separate component, subcomponent, and condition names. For example, you can refer to the InBPDU subcomponent of the *STP* component as STP.InBPDU. On the CLI, you can abbreviate or **[Tab]** complete any of these.

  A component or subcomponent typically has several conditions associated with it.

- To see the conditions associated with a component, use the following command:

  `show log events [event condition | [all | event component] {severity severity {only}}] {details}`

  For example, to see the conditions associated with the STP.InBPDU subcomponent, use the following command: `show log events stp.inbpdu`

  The following is sample output from this command:

```
Comp    SubComp     Condition              Severity      Parameters
------- ----------- ---------------------- ------------- ----------
STP     InBPDU      Drop                   Error          2 total
STP     InBPDU      Dump                   Debug-Data     3 total
STP     InBPDU      Trace                  Debug-Verbose  2 total
STP     InBPDU      Ign                    Debug-Summary  2 total
STP     InBPDU      Mismatch               Warning        2 total
```

The display above lists the five conditions contained in the STP.InBPDU component, the severity of the condition, and the number of parameters in the event message. In this example, the severities of the events in the STP.InBPDU subcomponent range from error to debug-summary.

When you use the **details** keyword, you see the message text associated with the conditions.

- For example, if you want to see the message text and the parameters for the event condition STP.InBPDU.Trace, use the following command: `show log events stp.inbpdu.trace details`

  The following is sample output from this command:

```
Comp    SubComp     Condition               Severity      Parameters
------- ----------- ----------------------- ------------- ----------
STP     InBPDU      Trace                   Debug-Verbose  2 total
0 - string
1 - string (printf)
Port=%0%: %1%
```

  The Comp heading shows the component name, the SubComp heading shows the subcomponent (if any), the Condition heading shows the event condition, the Severity heading shows the severity assigned to this condition, the Parameters heading shows the parameters for the condition, and the text string shows the message that the condition will generate. The parameters in the text string (for example, %0% and %1% above) will be replaced by the values of these parameters when the condition is encountered and displayed as the event message.

#### Filtering By Components and Conditions

You may want to send the messages that come from a specific component that makes up ExtremeXOS or to send the message generated by a specific condition. For example, you might want to send only those messages that come from the *STP* component, or send the message that occurs when the IP.Forwarding.SlowPathDrop condition occurs. Or you may want to exclude messages from a particular component or event. To do this, you construct a filter that passes only the items of interest, and you associate that filter with a target.

1. The first step is to create the filter using the `create log filter` command.

   You can create a filter from scratch, or copy another filter to use as a starting point. (It may be easiest to copy an existing filter and modify it.)
2. To create a filter, use the following command:

   `create log filter` *`name`* {**`copy`** *`filter_name`*}

   If you create a filter from scratch, that filter initially blocks all events until you add events (either the events from a component or a specific event condition) to pass. You might create a filter from scratch if you want to pass a small set of events and to block most events. If you want to exclude a small set of events, use the default filter that passes events at or above the default severity threshold (unless the filter has been modified), named DefaultFilter, that you can copy to use as a starting point for your filter.
3. After you create your filter, you configure filter items that include or exclude events from the filter.

   Included events are passed; excluded events are blocked.

4. To configure your filter, use the following command:

   `configure log filter` *name* [**add** | **delete**] {**exclude**} **events** [*event-condition* | [**all** | *event-component*] {**severity** *severity* {**only**}}]

   For example, if you create the filter myFilter from scratch, use the following command to include events:

   `configure log filter myFilter add events stp`

   All STP component events of at least the default threshold severity passes myFilter (for the STP component, the default severity threshold is error). You can further modify this filter by specifying additional conditions.

   For example, assume that myFilter is configured as before, and assume that you want to exclude the STP.CreatPortMsgFail event.

5. To add that condition, use the following command:

   `configure log filter myFilter add exclude events stp.creatportmsgfail`

6. You can also add events and subcomponents to the filter.

   For example, assume that myFilter is configured as before, and you want to include the STP.InBPDU subcomponent. To add that condition, use the following command:

   `configure log filter myFilter add events stp.inbpdu`

7. You can continue to modify this filter by adding more filter items.

   The filters process events by comparing the event with the most recently configured filter item first. If the event matches this filter item, the incident is either included or excluded, depending on whether the **exclude** keyword was used. If necessary, subsequent filter items on the list are compared. If the list of filter items is exhausted with no match, the event is excluded and is blocked by the filter.

8. To view the configuration of a filter, use the following command:

   `show log configuration filter {`*filter_name*`}`

   The following is sample output from this command (for the earlier filter):

```
Log Filter Name: myFilter
I/                                          Severity
E  Comp.   Sub-comp.   Condition              CEWNISVD
-  ------- ----------- ---------------------- --------
I  STP     InBPDU                             --------
E  STP                 CreatPortMsgFail       -E------
I  STP                                        --------
Include/Exclude: I - Include,  E - Exclude
Component Unreg: * - Component/Subcomponent is not currently registered
Severity Values: C - Critical,  E - Error,  W - Warning,  N - Notice,  I - Info
Debug Severity : S - Debug-Summary,  V - Debug-Verbose,  D - Debug-Data
+ - Debug Severities, but log debug-mode not enabled
If Match parameters present:
Parameter Flags: S - Source,  D - Destination, (as applicable)
I - Ingress,  E - Egress,  B - BGP
Parameter Types: Port - Physical Port list,  Slot - Physical Slot #
MAC  - MAC address,  IP - IP Address/netmask,  Mask - Netmask
VID  - Virtual LAN ID (tag),  VLAN - Virtual LAN name
L4   - Layer-4 Port #,  Num  - Number,  Str  - String
Nbr  - Neighbor, Rtr  - Routerid, EAPS - EAPS Domain
Proc - Process Name
```

```
Strict Match   : Y - every match parameter entered must be present in the event
N - match parameters need not be present in the event
```

The `show log configuration filter` command shows each filter item, in the order that it will be applied and whether it will be included or excluded. The above output shows the three filter items, one including events from the STP.InBPDU component, one excluding the event STP.CreatPortMsgFail, and the next including the remaining events from the STP component. The severity value is shown as "*", indicating that the component's default severity threshold controls which messages are passed. The Parameter(s) heading is empty for this filter because no match is configured for this filter. Matches are described in Matching Expressions on page 477.

Each time a filter item is added to or deleted from a given filter, the specified events are compared against the current configuration of the filter to try to logically simplify the configuration. Existing items will be replaced by logically simpler items if the new item enables rewriting the filter. If the new item is already included or excluded from the currently configured filter, the new item is not added to the filter.

### *Matching Expressions*

You can configure the switch so messages reaching the target match a specified match expression.

The message text is compared with the configured match expression to determine whether to pass the message on. To require that messages match a match expression, use the following command:

`configure log target [`**`console`** | **`memory-buffer`** | **`nvram`** | **`primary-msm`** | **`primary-node`**| **`backup-msm`** | **`backp-node`** | **`session`** | **`syslog`** [**`all`** | *`ipaddress`* | *`ipPort`* {**`vr`** *`vr_name`*}[**`local0 ... local7`**]]] **`match`** [**`any`** |*`match-expression`*]

The messages reaching the target will match the match-expression, a simple regular expression. The formatted text string that makes up the message is compared with the match expression and is passed to the target if it matches. This command does not affect the filter in place for the target, so the match expression is compared only with the messages that have already passed the target's filter. For more information on controlling the format of the messages, see Formatting Event Messages on page 480.

#### Simple Regular Expressions

A simple regular expression is a string of single characters including the dot character (.), which are optionally combined with quantifiers and constraints. A dot matches any single character, while other characters match only themselves (case is significant). Quantifiers include the star character (*) that matches zero or more occurrences of the immediately preceding token. Constraints include the caret character (^) that matches at the beginning of a message and the currency character ($) that matches at the end of a message. Bracket expressions are not supported. There are a number of sources

available on the Internet and in various language references describing the operation of regular expressions. The following table shows some examples of regular expressions.

**Table 64: Simple Regular Expressions**

| Regular Expression | Matches | Does Not Match |
|---|---|---|
| port | port 2:3<br>import cars<br>portable structure | poor<br>por<br>pot |
| ..ar | baar<br>bazaar<br>rebar | bar |
| port.*vlan | port 2:3 in vlan test<br>add ports to vlan<br>port/vlan | |
| myvlan$ | delete myvlan<br>error in myvlan | myvlan port 2:3<br>ports 2:4,3:4 myvlan link down |

*Matching Parameters*

Rather than using a text match, *EMS* allows you to filter more efficiently based on the parameter values of the message.

In addition to event components and conditions and severity levels, each filter item can also use parameter values to further limit which messages are passed or blocked. The process of creating, configuring, and using filters has already been described in Filtering By Components and Conditions on page 475, so this section describes matching parameters with a filter item.

To configure a parameter match filter item, use the following command:

`configure log filter` *`name`* [**`add`** | **`delete`**] {**`exclude`**} **`events`** [*`event-condition`* | [**`all`** | *`event-component`*] {**`severity`** *`severity`* {**`only`**}}] [**`match`** | **`strict-match`**] *`type value`*

Each event in ExtremeXOS is defined with a message format and zero or more parameter types.

The `show log events all` command can be used to display event definitions (the event text and parameter types). Only those parameter types that are applicable given the events and severity specified are exposed on the CLI. The syntax for the parameter types (represented by *`type`* in the command syntax above) is:

[address-family [**`ipv4-multicast`** | **`ipv4-unicast`** | **`ipv6-multicast`** | **`ipv6-unicast`**] | **`bgp-neighbor`** *`ip address`* | **`bgp-routerid`** *`ip address`* | **`eaps`** *`eaps domain name`* | {**`destination`** | **`source`**} [**`ipaddress`** *`ip address`* | **`L4-port`** *`L4-port`*| **`mac-address`** *`mac-address`*] | **`esrp`** *`esrp domain name`* | {**`egress`** | **`ingress`**} [**`slot`** *`slot number`* | **`ports`** *`portlist`*] | **`ipaddress`** *`ip address`* | **`L4-port`** *`L4-port`* | **`mac-address`** *`mac_address`* | **`netmask`** *`netmask`* | **`number`** *`number`* | **`port`** *`portlist`* | **`process`** *`process name`* | **`slot`** *`slotid`* | **`string`** *`exact string to be matched`* | **`vlan`** *`vlan name`* | **`vlan tag`** *`vlan tag`*]

You can specify the ipaddress type as IPv4 or IPv6, depending on the IP version.

The following examples show how to configure IPv4 addresses and IPv6 addresses:

IPv4 address

- To configure an IP address, with a mask of 32 assumed, use the following command:

```
configure log filter myFilter add events all match ipaddress 12.0.0.1
```

- To configure a range of IP addresses with a mask of 8, use the following command:

```
configure log filter myFilter add events all match ipaddress 12.0.0.0/8
```

IPv6 address

- To configure an IPv6 address, with a mask of 128 assumed, use the following command:

```
configure log filter myFilter add events all match ipaddress 3ffe::1
```

- To configure a range of IPv6 addresses with a mask of 16, use the following command:

```
configure log filter myFilter add events all match ipaddress 3ffe::/16
```

IPv6 scoped address

- IPv6 scoped addresses consist of an IPv6 address and a *VLAN*. The following examples identify a link local IPv6 address.

To configure a scoped IPv6 address, with a mask of 128 assumed, use the following command:

```
configure log filter myFilter add events all match ipaddress fe80::1%Default
```

To configure a range of scoped IPv6 addresses with a mask of 16, use the following command:

```
configure log filter myFilter add events all match ipaddress fe80::/16%Default
```

To configure a scoped IPv6 address with any VLAN, use the following command:

```
configure log filter myFilter add events all match ipaddress fe80::/16%*
```

To configure any scoped IPv6 address with a specific VLAN, use the following command:

```
configure log filter myFilter add events all match ipaddress fe80::/0%Default
```

Note

In the previous example, if you specify the VLAN name, it must be a full match; wild cards are not allowed.

The *value* depends on the parameter type specified.

As an example, an event may contain a physical port number, a source MAC address, and a destination MAC address. To allow only those *RADIUS (Remote Authentication Dial In User Service)* incidents, of severity notice and above, with a specific source MAC address, use the following command:

```
configure log filter myFilter add events aaa.radius.requestInit severity notice match
source mac-address 00:01:30:23:C1:00
```

The string type is used to match a specific string value of an event parameter, such as a user name. The exact string is matched with the given parameter and no regular expression is supported.

#### Match Versus Strict-Match

The **match** and **strict-match** keywords control the filter behavior for those incidents with event definition that does not contain all the parameters specified in a `configure log filter events match` command.

This is best explained with an example. Suppose an event in the XYZ component, named XYZ.event5, contains a physical port number, a source MAC address, but no destination MAC address. If you configure a filter to match a source MAC address and a destination MAC address, XYZ.event5 will match the filter when the source MAC address matches regardless of the destination MAC address because the event contains no destination MAC address. If you specify the **strict-match** keyword, then the filter will never match event XYZ.event5 because this event does not contain the destination MAC address.

In other words, if the **match** keyword is specified, an incident will pass a filter so long as all parameter values in the incident match those in the match criteria, but all parameter types in the match criteria need not be present in the event definition.

### *Formatting Event Messages*

Event messages are made up of a number of items. The individual items can be formatted; however, *EMS* does not allow you to vary the order of the items.

- Format the messages for a particular target.

  `configure log target format`

  Using the default format for the session target, an example log message might appear as:

  ```
  06/25/2004 22:49:10.63 <Info:dm.Info> MSM-A: PowerSupply:4 Powered On
  ```

  If you set the current session format using the following command:

  ```
  configure log target session format timestamp seconds date mm-dd-yyyy event-name
  component
  ```

  The same example would appear as:

  ```
  06/25/2004 22:49:10 <dm> PowerSupply:4 Powered On
  ```

- Provide some detailed information to technical support, set the current session format.

  ```
  configure log target session format timestamp hundredths date mmm-dd event-name
  condition process-name source-line
  ```

  The same example then appears as:

  ```
  Jun 25 22:49:10.63 <dm.info> devmgr: (dm.c:134) PowerSupply:4 Powered On
  ```

## Displaying Real-Time Log Messages

You can configure the system to maintain a running real-time display of log messages on the console display or on a (Telnet) session.

- Turn on the log display on the console with the command:

  `enable log target console`

  This setting may be saved to the FLASH configuration and is restored on boot-up (to the console display session).

- Turn on log display for the current session with the command:

  ```
  enable log target session
  ```

  This setting only affects the current session and is lost when you log off the session.

  The messages that are displayed depend on the configuration and format of the target. For information on message filtering, see Filtering Events Sent to Targets on page 472. For information on message formatting, see Formatting Event Messages on page 480.

## Displaying Event Logs

The log stored in the memory buffer and the NVRAM can be displayed on the current session (either the console display or Telnet).

- Display the log using the command:

  ```
  show log {messages [memory-buffer | nvram]} {events {event-condition |
  event-component]} {severity severity {only}} {starting [date date time
  time | date date | time time]} {ending [date date time time | date
  date | time time]} {match regex} {chronological}
  ```

  You can use many options to select those log entries of interest. You can select to display only those messages that conform to the specified:

  - Severity
  - Starting and ending date and time
  - Match expression

  The displayed messages can be formatted differently from the format configured for the targets, and you can choose to display the messages in order of newest to oldest or in chronological order (oldest to newest).

### *Log Buffer Threshold Alert*

ExtremeXOS 16.1 now supports the ability to generate syslog message or trigger a *SNMP* trap when the log memory buffer is filled up to 90% of configured log lines. SNMP/SNMP trap is not supported by ExtremeXOS *EMS* module at this point. The feature provides that an alert in the form of log message will be generated when the log memory buffer is filled up to a certain percentage threshold. A new command has been added to allow you to configure the memory buffer alert. The alert will be based on the percentage used of the memory buffer. By default, the alert is disabled.

The new command can be used to enable the alert and to specify the percentage threshold that triggers alerts. When the percentage threshold is set to a value in the allowed range, the memory buffer alert is enabled. When the percentage threshold is set to none, the memory buffer alert is disabled. The alert triggered will be a log message with Notice severity.

If the alert is enabled, a log message will be generated in the following three scenarios:

1. When the actual use of the memory buffer first exceeds the configured percentage threshold.
2. When the configured memory buffer size is changed and the actual use of the buffer exceeds the configured percentage threshold.
3. When the configured percentage threshold is changed and the actual use of the buffer exceeds the re-configured percentage threshold.

When you execute the command `clear log`, the memory buffer will be emptied. When the new logs re-fill up to the configured percentage threshold, the log message will be re-generated. Output of `show log configuration` will be modified as well to include the information of memory buffer alert configuration and current memory buffer usage.

## Uploading Event Logs

The log stored in the memory buffer and the NVRAM can be uploaded to a TFTP server.

- Use the following command to upload the log:

  `upload log ipaddress {vr vr_name} {block-size block_size} filename { messages [ memory-buffer | nvram ] { events { event-condition | event_component }}} { severity severity { only }} {match regex} {chronological}`

  You must specify the TFTP host and the filename to use in uploading the log.

  There are many options you can use to select the log entries of interest. You can select to upload only those messages that conform to the specified:

  - Severity
  - Match expression

  The uploaded messages can be formatted differently from the format configured for the targets, and you can choose to upload the messages in order of newest to oldest or in chronological order (oldest to newest).

## Displaying Counts of Event Occurrences

*EMS* adds the ability to count the number of occurrences of events. Even when an event is filtered from all log targets, the event is counted.

- Display the event counters.

  `show log counters {event condition | [all | event component]} {include | notified | occurred} {severity severity {only}}}`

  The system displays two counters. One counter displays the number of times an event has occurred, and the other displays the number of times that notification for the event was made to the system for further processing. Both counters reflect totals accumulated since reboot or since the counters were cleared using the `clear log counters` or `clear counters` command.

  The `show log counters` command also displays an included flag (the column titled In in the output). The included flag is set to Y(es) if one or more targets are receiving notifications of this event without regard to matching parameters.

  The keywords **include**, **notified**, and **occurred** display events only with non-zero counter values for the corresponding counter.

  The output of the command:

  ```
  show log counters stp.inbpdu severity debug-summary
  ```

is similar to the following:

```
Comp     SubComp     Condition               Severity      Occurred  In Notified
-------- ----------- ----------------------- ------------- --------  -- --------
STP      InBPDU      Drop                    Error                0  Y          0
STP      InBPDU      Ign                     Debug-Summary        0  N          0
STP      InBPDU      Mismatch                Warning              0  Y          0
Occurred  : # of times this event has occurred since last clear or reboot
Flags     : (*) Not all applications responded in time with there count values
In(cluded): Set to Y(es) if one or more targets filter includes this event
Notified  : # of times this event has occurred when 'Included' was Y(es)
```

The output of the command:

```
show log counters stp.inbpdu.drop
```

is similar to the following:

```
Comp     SubComp     Condition               Severity      Occurred  In Notified
-------- ----------- ----------------------- ------------- --------  -- --------
STP      InBPDU      Drop                    Error                0  Y          0
Occurred  : # of times this event has occurred since last clear or reboot
Flags     : (*) Not all applications responded in time with there count values
In(cluded): Set to Y(es) if one or more targets filter includes this event
Notified  : # of times this event has occurred when 'Included' was Y(es)
```

## Displaying Debug Information

By default, a switch does not generate events of severity Debug-Summary, Debug-Verbose, and Debug-Data unless the switch is in debug mode. Debug mode causes a performance penalty, so it should only be enabled for specific cases where it is needed.

- Place the switch in debug mode using the command:

  `enable log debug-mode`

  When the switch is in debug-mode, any filters configured for your targets still affect which messages are passed on or blocked.

## Logging Configuration Changes

ExtremeXOS allows you to record all configuration changes and their sources that are made using the CLI by way of telnet or the local console. The changes cause events that are logged to the target logs. Each log entry includes the user account name that performed the change and the source IP address of the client (if telnet was used). Configuration logging applies only to commands that result in a configuration change.

- Enable configuration logging with the command:

  `enable cli-config-logging`

- Disable configuration logging with the command:

  `disable cli-config-logging`

  CLI configuration logging is disabled by default.

# Using the XML Notification Client

## Introduction

This feature allows an event such as a configuration change, a fault, a change in status, the crossing of a threshold, or an external input to the system, to be sent as an asynchronous message or event notification to external web servers.

The only ExtremeXOS modules that support XML notification as targets are Identity Management and *EMS*.

XML notification does not provide any event filtering capability. However, in the case of EMS, the target Web server can be configured with log filters and conditions. The XML notification feature establishes and checks the connectivity with the web server only when an event is ready to be pushed. State transitions take place if required. Statistics are updated accordingly and can be monitored.

The Web servers must be configured and enabled using ExtremeXOS CLI with an IP address, port, protocol type, user authentication, session information, if any, and other web server configuration.

A maximum of four web servers can be configured at a time.

The XML schemas are defined using Web Services Description Language (WSDL) in the XML SDK.

XML Notification is supported on BlackDiamond 8000 series modules and Summit family switches.

## HTTP Client Interface

The event notifications are sent in XML format, using SOAP transport protocol and HTTP/HTTPS.

The XML notification client can communicate with the external HTTP/HTTPS server. HTTP/HTTPS client interfaces maintain persistent connections with external Web servers as long as the target is activated on the switch.

The HTTP URL format for the server is `http://<ip-address>:<port>/<service>`. The default HTTP port number 80 is used if a port number is not configured.

HTTP basic access authentication method (Base64 algorithm) is used to encrypt the user name and password when making a request. HTTP cookies are not supported in this release.

The SSH module must be installed on the ExtremeXOS switch to use the XML notification feature on HTTPS. Once the SSH module is installed, a server certificate should be created that can be used by the HTTPS server. Refer to the configuration guidelines of the HTTP server, to generate the secure certificate on the ExtremeXOS switch (see Secure Socket Layer on page 944 for more information).

## Configuring XML Notification

- Create a web server target on an XML client using the command:

  `create xml-notification target new-target url url {vr vr_name} {user [none | user]} {encrypted-auth encrypted-auth} {queue-size queue-size}`
- Configure a web server target on an XML client using the command:

  `configure xml-notification target target [url url {vr vr_name} | user [none | user] | [encrypted-auth encrypted-auth] | [queue-size queue-size]]`
- Add or delete an ExtremeXOS application to awWeb server target (*EMS* or Identity Management) using the command:

  `configure xml-notification target target [add | delete] module`
- Enable or disable web server target(s) using the command:

  `[enable|disable] xml-notification [all | target]`
- Delete a web server target on an XML client process using the command:

  `delete xml-notification target target`
- Unconfigure an XML client process using the command:

  `unconfigure xml-notification`
- Unconfigure and reset all statistics counters using the command:

  `clear counters xml-notification {all | target}`

## Displaying XML Notification

- Display the configuration of a web server target using the command:

  `show xml-notification configuration {target}`
- Display the connection status, enable status, and event statistics of the target web server using the command:

  `show xml-notification statistics {target}`
- Display information on the stored certificate using the command:

  `show ssl {detail}`

## Configuring Log Target in EMS

The following commands support the *EMS* XML target.

- Create a web server XML target using hte command:

  `create log target xml-notification [ target_name | xml_target_name ]`
- Configure the web server target with an EMS filter using the command:

  `configure log target xml-notification xml_target_name filter filter-name {severity [[severity] {only}]}`
- Enable the web server target using the command:

  `enable log target xml-notification xml_target_name`
- Disable the web server target using the command:

  `disable log target xml-notification xml_target_name`

- Delete the web server target XML target using the command:
  `delete log target xml-notification xml_target_name`
- Display XML target information using the command:
  `show log configuration target xml-notification {xml_target_name}`

## Examples

Below are examples of configuring web server targets in a XML Notification module.

*Scenario 1: Push filtered EMS events to external web server in a well-defined XML format.*

Create a Web Server Target test1, create a log target and a filter in EMS, and attach the filter to the web target. Enable the target in both EMS and XML-Notification module.

```
create XML-notification target test1 url http://10.255.129.22:8080/xos/webservice user
admin
create log target xml-notification "test1"
create log filter xmlc_filter_1
configure log filter "xmlc_filter_1" add events idmgr
configure log target xml-notification "test1" filter "xmlc_filter-1"
enable log target xml-notification "test1"
enable XML-notification test1
```

*Scenario 2: Push user identity events to the external web server without EMS module in a well defined (XML Schema) XML format.*

Create a web server target and attach an idmgr module. Idmgr modules use an XML-notification backend library to trigger events. In this case, no special filters are supported.

```
create xml-notification target test2 url http://10.255.129.22:8080/xos/webservice user
admin
configure xml-notification target test2 add module idmgr
enable xml-notification test2
```

*Scenario - 3: XMLC notifications using HTTPS*

Install the SSH module. If not installed, refer to Understanding System Redundancy on page 54 for details on SSH and Secure Socket Layer on page 944 for details on SSL.

```
configure ssl certificate privkeylen 1024 country us organization extreme common-name
name1
create xml-notification target test3 url https://10.120.91.64:8443/xos/webservice
configure xml-notification target "test3" user admin
configure xml-notification target "test3" add "ems"
enable xml-notification "test3"
```

# Using sFlow

sFlow is a technology for monitoring traffic in data networks containing switches and routers. It relies on statistical sampling of packets from high-speed networks, plus periodic gathering of thestatistics. A User Datagram Protocol (UDP) datagram format is defined to send the information to an external entity for analysis. sFlow consists of a Management Information Base (MIB), and a specification of the packet format for forwarding information to a remote agent. Details of sFlow specifications can be found in RFC 3176 and more information can be found at the following website: www.sflow.org.

The ExtremeXOS implementation is based on sFlow version 5, which is an improvement from the revision specified in RFC 3176.

Additionally, the switch hardware allows you to set the hardware sampling rate independently for each module on the switch, instead of requiring one global value for the entire switch. The switch software also allows you to set the individual port sampling rates, so that you can fine-tune sFlow statistics.

SFlow and mirroring are not mutually exclusive on BlackDiamond 8000 c-, e-, xl-, xm-series modules, and Summit family switches, whether or not they are included in a SummitStack. You can enable them simultaneously on the following platforms:

- BlackDiamond 8000 series modules
- BlackDiamond X8 modules
- Summit family switches

  For information on licensing, see the *Feature License Requirements* document.

However, you should be aware that the following limitations are present in the ExtremeXOS implementation:

- Generic port statistics are reported to the sFlow collector.
- Non-extended data.
- Only port-based sampling.
- There is no MIB support.

## Sampling Mechanisms

The following platforms support hardware-based sampling at a programmed interval:

- BlackDiamond 8000 series modules
- BlackDiamond X8 series modules
- Summit family switches

With hardware-based sampling, the data path for a packet that traverses the switch does not require processing by the CPU. Fast path packets are handled entirely by ASICs, and are forwarded at wire speed rate.

Both ingress and egress sFlow sampling can be enabled simultaneously on a port. The `enable sflow port` command provides an option to enable sFlow on ingress, or egress, or both directions. The default value is ingress. The sample-rate is maintained on a per-port basis, so a given port will have a same sample rate for ingress and egress traffic. Ingress and egress sFlows sample both unicast and multicast egress flows. The global enable/disable control of sFlow is common to both ingress and egress.

When sFlow sampling is enabled on a port, the sFlow agent samples the traffic on that port, processed in slow path and passed on to the collector. You can configure the rate at which the packets are sampled.

## Limitations of Egress Sflow

The following list identifies limitations of the egress sFlow feature:

- Due to the hardware limitation, destination port information is not supported for multicast traffic. The output interface index is populated as 0.
- Egress sFlow sampling does not support de-duplication of packets.
- For multicast traffic, the sampling rate and sample pool of the egress sFlow sampled datagram is populated as 0, because the source ID of the egress sampled multicast packet is unknown.
- For L3 unicast traffic, an unmodified packet is sampled and the destination port is supplied if the L3 traffic is a flow within single chip. When the egress port and ingress port are in different chips, then a modified packet is sampled and the destination ports are supplied. For L3 multicast traffic, an unmodified packet is sampled and the destination port is populated as zero.
- Packets dropped due to egress *ACL* will be sampled.
- In cases of unicast and multicast flooding, the packets are sampled before packet replication. If the ingress and member ports are in the same chip, then a single copy of the packet is sampled even though the egress sFlow is enabled on more than one member's ports. If the member ports are spread across different units, then packets are sampled on a per-chip basis.
- In flooding cases, the least configured sampling rate among the member ports on a port group is considered as a sample rate. Even if you configure different sample rates on a member ports, egress sampling is performed based on least configured sample rate among the member ports on a unit.

## sFlow Destination Port

Currently when a packet is sampled from the ingress traffic, the output interface index is populated as 0. This is applicable for both unicast and multicast traffic. As part of the egress sFlow, the output interface index is populated with the destination port information for unicast packets sampled in both the ingress and egress direction. Because of a hardware limitation, multicast packet samples still have an output interface index as 0.

The following table gives the expected ifIndex output based on traffic type and the sampling direction.

**Table 65: ifIndex Traffic Type Sampling Direction**

| Case # | Ingress/Egress Unicast/ Multicast | Scenario |
|---|---|---|
| 1 | Ingress/Unicast | sFlow sample includes both ingress and egress port (ifIndex). |
| 2 | Egress/Unicast | sFlow sample includes both ingress and egress port information. |
| 3 | Ingress/Multicast | Egress port information cannot be provided because of hardware limitation. Egress ifIndex will be 0 and ingress ifIndex will be supplied. |
| 4 | Egress/Multicast | Egress port information cannot be provided because of hardware limitation. Egress ifIndex will be 0 and ingress ifIndex will be supplied. |

## Sampling Mechanisms

The following platforms support hardware-based sampling at a programmed interval:

- BlackDiamond 8000 series modules
- BlackDiamond X8 series modules
- Summit family switches

With hardware-based sampling, the data path for a packet that traverses the switch does not require processing by the CPU. Fast path packets are handled entirely by ASICs and are forwarded at wire speed rate.

## Egress sFlow Sampling

Egress sFlow sampling functionality extends sampling to the egress traffic, both unicast and multicast streams. When egress sFlow sampling is enabled on a port, the sFlow agent samples the egress traffic on that port, and these sampled packets are processed by slow path passed on to the collector. You can configure the rate at which the packets are sampled.

Both ingress and egress sampling can be enabled simultaneously on a port. The sample-rate is maintained on a per-port basis, so a given port will have the same sample rate for ingress and egress traffic.

This feature supports the following configuration options:

- sFlow can sample the egress flow of a physical interface; in this case, the sFlow agent samples the packet from the egress flow of an interface.
- sFlow can sample both the ingress and egress flows of an interface; in this case the sFlow agent samples the packet from the ingress and egress flow of a configured interface.

Similar to existing ingress sFlow sampling, the egress sFlow sampling samples both unicast and multicast egress flows. The global enable/disable control of sFlow is common for both ingress and egress. When the global option is enabled, the port level sFlow parameters are applied to hardware.

*Limitations*

The following list identifies limitations of the egress sFlow feature:

- Due to the hardware limitation, destination port information is not supported for multicast traffic. The output interface index is populated as 0.
- The egress sFlow sampling does not support de-duplication of packets.
- For multicast traffic, the sampling rate, sample pool of the egress sFlow sampled datagram will be populated as 0, because the source ID of the egress sampled multicast packet is unknown.
- For L3 unicast traffic, an unmodified packet is sampled and the destination port is supplied if the L3 traffic is a flow within single chip. When the egress port and ingress port are in different chips, then a modified packet is sampled and the destination ports are supplied. For L3 multicast traffic, unmodified packet is sampled and destination port will be populated as zero.
- Packets dropped due to egress *ACL* will be sampled.
- In cases of unicast and multicast flooding, the packets are sampled before packet replication. If the ingress and member ports are in the same chip then a single copy of the packet is sampled even

though the egress sFlow is enabled on more than one member's ports. If the member ports are spread across different chips, then packets are sampled on a per-chip basis.

- In flooding cases, the least configured sampling rate among the member ports on a port group is considered as a sample rate. Even if you configure different sample rates on a member ports, egress sampling is performed based on least configured sample rate among the member ports on a unit.

## sFlow Destination Port

Currently when a packet is sampled from the ingress traffic, the output interface index is populated as 0. This is applicable for both unicast and multicast traffic. As part of the Egress sFlow feature, the output interface index is populated with the destination port information for unicast packets sampled in both ingress and egress direction. Because of a hardware limitation, multicast packet samples still have an output interface index as 0.

The following table illustrates the expected output of the ifIndex when based on traffic type, and sampling direction:

| Ingress/Egress, Unicast/ Multicast | Scenario |
|---|---|
| Ingress/Unicast | sFlow sample includes both ingress and egress port (ifIndex). |
| Egress/Unicast | sFlow sample includes both ingress and egress port information. |
| Ingress/Multicast | Egress port information cannot be provided because of hardware limitation. Egress if Index will be 0 and ingress if Index is supplied. |
| Egress/Multicast | Egress port information cannot be provided because of hardware limitation. Egress if Index will be 0 and ingress if Index is supplied. |

## Configuring sFlow

ExtremeXOS allows you to collect sFlow statistics on a per port basis.

An agent residing locally on the switch sends data to a collector that resides on another machine. You configure the local agent, the address of the remote collector, and the ports of interest for sFlow statistics gathering. You can also modify default values for how frequently on an average a sample is taken and the maximum number of samples allowed before throttling the sample gathering.

To configure sFlow on a switch:

- Configure the local agent
- Configure the addresses of the remote collectors
- Enable sFlow globally on the switch
- Enable sFlow on the desired ports

Optionally, you may also change the default values of the following items:

- How often the statistics are collected
- How frequently a sample is taken, globally or per port
- How many samples per second can be sent to the CPU

*Configuration Tasks*

Use the following commands to configure the sFlow feature:

- `enable sflow ports` **`all`** | *`port_list`* {**`ingress`** | **`egress`** | **`both`**}

  The keyword options enable you to configure sFlow types on a given set of ports. If you do not configure an sFlow type, then ingress sFlow sampling is enabled as the default configuration.

Use the following commands to display the type of sFlow configured on the physical interface, and various statistics about sFlow sampling:

- `show sflow configuration`
- `show sflow statistics`

  The following fields are displayed:

  - Received frames—Number of frames received on sFlow enabled ports.
  - Sampled Frames—Number of packets that have been sampled by sFlow.
  - Transmitted Frames—Number of UDP packets sent to remote collector(s).
  - Broadcast Frames—Number of broadcast frames received on sFlow enabled ports.
  - Multicast Frames—Number of multicast frames received on sFlow enabled ports.
  - Packet Drops—Number of samples dropped.

*Configuring the Local Agent*

The local agent is responsible for collecting the data from the samplers and sending that data to the remote collector as a series of UDP datagrams. The agent address is stored in the payload of the sFlow data, and is used by the sFlow collector to identify each agent uniquely. By default, the agent uses the management port IP address as its IP address.

- Change the agent IP address using the command:

  `configure sflow agent {`**`ipaddress`**`}` *`ipaddress`*

- Unconfigure the agent using the command:

  `unconfigure sflow agent`

*Configuring the Remote Collector Address*

You can specify up to four remote collectors to send the sFlow data to. Typically, you would configure the IP address of each collector. You may also specify a UDP port number different from the default value of 6343, and/or a *virtual router (VR)* different from the default of *VR-Mgmt*. When you configure a collector, the system creates a database entry for that collector that remains until the collector is unconfigured. All the configured collectors are displayed in the command: `show sflow {`**`configuration`**`}` .

- Configure the remote collector using the command:

  `configure sflow collector {`**`ipaddress`**`}` *`ipaddress`* `{`**`port`** *`udp-port-number`*`}` `{`**`vr`** *`vr_name`*`}`

- Unconfigure the remote collector and remove it from the database using the command:

  `unconfigure sflow collector {`**`ipaddress`**`}` *`ipaddress`* `{`**`port`** *`udp-port-number`*`}` `{`**`vr`** *`vr_name`*`}`

*Enabling sFlow Globally on the Switch*

Before the switch starts sampling packets for sFlow, you must enable sFlow globally on the switch.

- Enable sFlow globally using the command:
  `enable sflow`
- Disable sFlow globally using the command:
  `disable sflow`

  When you disable sFlow globally, the individual ports are also put into the disabled state. If you later enable the global sFlow state, individual ports return to their previous state.

*Enabling sFlow on the Desired Ports*

- Enable sFlow on specific ports using the command:
  `enable sflow` **`ports`** *`port_list`* {**`ingress`** | **`egress`** | **`both`** }

  The **`ingress`** option allows you to configure the sFlow type on a given set of ports. This option is configured on the port by default.

  You may enable and disable sFlow on ports irrespective of the global state of sFlow, but samples are not taken until both the port state and the global state are enabled.
- Disable sFlow on ports using the command:
  `disable sflow ports` *`port_list`*

## Additional sFlow Configuration Options

You can configure three global options to different values from the defaults. These options affect how frequently the sFlow data is sent to the remote collector, how frequently packets are sampled, and the maximum number of sFlow samples that could be processed in the CPU per second.

You can also configure how frequently packets are sampled per port.

*Polling Interval*

Each port counter is periodically polled to gather the statistics to send to the collector. If there is more than one counter to be polled, the polling is distributed in such a way that each counter is visited once during each polling interval, and the data flows are spaced in time. For example, assume that the polling interval is 20 seconds and there are 40 counters to poll. Two ports will be polled each second, until all 40 are polled.

- Configure the polling interval using the command:
  `configure sflow poll-interval` *`seconds`*

*Global Sampling Rate*

The global sampling rate is the rate set on newly enabled sFlow ports. Changing this rate does not affect currently enabled sFlow ports. The default sample rate is 8192, so sFlow samples one packet out of every 8192 received.

- Use the following command to configure the switch to use a different sampling rate.
  `configure sflow sample-rate` *`number`*

For example, if you set the sample rate number to 16384, the switch samples one out of every 16384 packets received. Higher numbers mean fewer samples and longer times between samples. If you set

the number too low, the number of samples can be very large, which increases the load on the switch. Do not configure the sample rate to a number lower than the default unless you are sure that the traffic rate on the source is low.

The Global Sampling Rate applies to the Summit X440, X450-G2, X460, X460-G2, X480, X670, X670G2, and X770 Series Switches and BlackDiamond 8000 c-, e-, xl-, xm-, and X8 Series Modules only.

Note
The minimum rate that these platforms sample is 1 out of every 256 packets. If you configure a rate to be less than 256, the switch automatically rounds up the sample rate to 256.

*Per Port Sampling Rate*

The per port sampling rate overrides the system-wide value.

The rate is rounded off to the next power of two, so if 400 is specified, the sample rate is configured as 512. The valid range is 256 to 536870912.

- Set the sampling rate on individual ports.

  `configure sflow ports` *`port_list`* **`sample-rate`** *`number`*

This configuration applies to Summit Family Switches, BlackDiamond 8000, c-, e-, xl-, and xm-Series Modules, and BlackDiamond X8 Series Switches only. All ports on the switch or the same I/O module are sampled individually.

*Maximum CPU Sample Limit*

A high number of samples can cause a heavy load on the switch CPU. To limit the load, there is a CPU throttling mechanism to protect the switch.

On a modular switch, whenever the limit is reached, the sample rate value is doubled on the slot from which the maximum number of samples are received. For ports on that slot that are sampled less frequently, the sampling rate is not changed; the sub-sampling factor is adjusted downward.

Note
Sflow sampling is limited to 2000 packets per second in the CPU on all Summit platforms. Any packets sent at a rate greater than 2000 pps are dropped.

On a stand-alone switch, whenever the limit is reached, the sample rate value is doubled on the ports from which the maximum number of samples are received. For ports that are sampled less frequently, the sampling rate is not changed; the sub-sampling factor is adjusted downward.

- Configure the maximum CPU sample limit.

  `configure sflow max-cpu-sample-limit` *`rate`*

*Unconfiguring sFlow*

- Reset the configured values for sFlow to their default values and remove from sFlow any configured collectors and ports using the following ocmmand:

  `unconfigure sflow`

## sFlow Configuration Example

In a service provider environment, you can configure sFlow to sample packets at the edge of the network to determine the hourly usage for each IP address in the data center.

You can capture web traffic, FTP traffic, mail traffic, and all bits of data that travel across service providers' edge routers to their customers' (end users') servers.

The example in this section assumes that you already have an sFlow data collector installed somewhere in your network. In many environments, the sFlow data collector is on a network PC.

The following sFlow configuration example performs the following tasks in a service provider environment:

- Configures the IP address of the sFlow data collector.

  **Note**
  In many environments, the sFlow data collector is not directly connected to the switch. Make sure to specify the VR used to forward traffic between the sFlow collector and the switch. In most cases the VR is *VR-Mgmt*.

- Configures the sampling rate on an edge port.
- Enables sFlow on the edge port.
- Enables sFlow globally on the switch.

```
configure sflow collector 10.127.11.88 vr vr-mgmt
configure sflow ports 5:21 sample-rate 8192
enable sflow ports 5:21 egress
enable sflow
```

Here is sample output for the configuration:

```
SFLOW Global Configuration
Global Status: enabled
Polling interval: 20
Sampling rate: 8192
Maximum cpu sample limit: 2000
SFLOW Configured Agent IP: 0.0.0.0
Operational Agent IP: 10.127.11.88
Collectors
SFLOW Port  Configuration Port Status  Sample-rate Subsampling Sflow-type
Config / Actual factor Ingress/Egress
5:21    enabled          8192 / 8192  1          Disabled / Enabled
```

## Displaying sFlow Information

- Display the current configuration of sFlow using the command:

  `show sflow {`**`configuration`**`}`

- Display the sFlow statistics using the command:

  `show sflow statistics`

# Using RMON

Using the Remote Monitoring (RMON) capabilities of the switch allows network administrators to improve system efficiency and reduce the load on the network.

Note

You can use the RMON features of the system only if you have an RMON management application and have enabled RMON on the switch.

## About RMON

RMON is the common abbreviation for the Remote Monitoring Management Information Base (MIB) system defined by the Internet Engineering Task Force (IETF) documents RFC 1757 and RFC 2021, which allows you to monitor LANs remotely.

A typical RMON setup consists of the following two components:

- RMON agent
- Management workstation

### *RMON Agent*

An RMON agent is an intelligent software agent that continually monitors port statistics and system variables.

The agent transfers the information to a management workstation on request, or when a predefined threshold is crossed.

Information collected by RMON includes Ethernet port statistics and history and the software version and hardware revision of the device. RMON generates alarms when threshold levels are met and then logs those events to the log. RMON can also send traps to the destination address configured by the management workstation. You can also use RMON to trigger a system reboot.

### *Management Workstation*

A management workstation communicates with the RMON agent and collects the statistics from it.

The workstation does not have to be on the same network as the RMON agent and can manage the agent by in-band or out-of-band connections.

If you enable RMON on the switch, you can use a management workstation to review port statistics and port history, no configuration of the management workstation is necessary. However, you must use a management workstation to configure the alarm and event entries.

## Supported RMON Groups of the Switch

The IETF defines nine groups of Ethernet RMON statistics. The switch supports the following four of these groups, as defined in RFC 1757:

- Statistics
- History

- Alarms
- Events

The switch also supports the following parameters for configuring the RMON agent and the trap destination table, as defined in RFC 2021:

- probeCapabilities
- probeSoftwareRev
- probeHardwareRev
- probeDateTime
- probeResetControl
- trapDestTable

The following sections describe the supported groups, the RMON probe configuration parameters, and the trap destination parameter in greater detail.

*Statistics*

The RMON Ethernet Statistics group provides traffic and error statistics showing packets, bytes, broadcasts, multicasts, and errors on an Ethernet port.

Information from the Statistics group is used to detect changes in traffic and error patterns in critical areas of the network.

*History*

The History group provides historical views of network performance by taking periodic samples of the counters supplied by the Statistics group.

The group features user-defined sample intervals and bucket counters for complete customization of trend analysis.

The group is useful for analysis of traffic patterns and trends on an Ethernet port, and for establishing baseline information indicating normal operating parameters.

*Alarms*

The Alarms group provides a versatile, general mechanism for setting threshold and sampling intervals to generate events on any RMON variable.

Both rising and falling thresholds are supported, and thresholds can be on the absolute value of a variable or its delta value.

Note

Creating an entry in the alarmTable does not validate the alarmVariable and does not generate a badValue error message.

Alarms inform you of a network performance problem and can trigger automated action responses through the Events group.

*Events*

The Events group creates entries in an event log and/or sends *SNMP* traps to the management workstation.

An event is triggered by an RMON alarm. The action taken can be configured to ignore it, to log the event, to send an SNMP trap to the receivers listed in the trap receiver table, or to both log and send a trap. The RMON traps are defined in RFC 1757 for rising and falling thresholds.

Effective use of the Events group saves you time. Rather than having to watch real-time graphs for important occurrences, you can depend on the Events group for notification. Through the SNMP traps, events can trigger other actions, which provides a mechanism for an automated response to certain occurrences.

*RMON Probe Configuration Parameters*

The RMON probe configuration parameters supported in ExtremeXOS are a subset of the probe configuration group as defined in RFC 2021. The probe configuration group controls and defines the operation of the RMON agent.

You can configure the following objects:

- probeCapabilities—View the RMON MIB groups supported on at least one interface by the probe.
- probeSoftwareRev—View the current software version of the monitored device.
- probeHardwareRev—View the current hardware version of the monitored device.
- probeDateTime—View the current date and time of the probe. For example, Friday December 31, 2004 at 1:30:15 PM EST is displayed as: 2004-12-31,13:30:15.0

  If the probe is aware of time zones, the display also includes the Greenwich Mean Time (GMT) offset. For example, Friday, December 31, 2004, 1:30:15 PM EST with the offset known is displayed as: 2004-12-31,13:30:15.0, -4.0

  If time information is unavailable or unknown, the time is not displayed.
- probeResetControl—Restart a managed device that is not running normally. Depending on your configuration, you can do one of the following:
  - Warm boot—A warm boot restarts the device using the current configuration saved in non-volatile memory.
  - Cold boot—A cold boot causes the device to reset the configuration parameters stored in non-volatile memory to the factory defaults and then restarts the device using the restored factory default configuration.

*trapDestTable*

The trapDestTable contains information about the configured trap receivers on the switch and stores this information in non-volatile memory.

To configure one or more trap receivers, see Using the Simple Network Management Protocol on page 78.

*Extreme-RtStats-MIB*

The extremeRtStatsTable provides the user with all the common measurement/monitoring attributes in a single table.

It includes measurements like utilization, error, and collision levels.

The extreme RtStatsUtilization variable gives an accurate measurement of segment utilization. ExtremeRtStatsCollisions is included in the segment utilization calculation for more accuracy. Collision statistics are collected periodically, and the segment utilization is calculated with a sampling interval of 5 minutes (300 seconds).

The extremeRtStatsTotalErrors variable is calculated by adding the following counters:

- extremeRtStatsCRCAlignErrors (receive errors)
- extremeRtStatsFragments (receive errors)
- extremeRtStatsJabbers (receive errors)
- extremeRtStatsCollisions (transmit errors)

## Configuring RMON

RMON requires one probe per LAN segment, and stand-alone RMON probes traditionally have been expensive. Therefore, the approach taken by Extreme Networks has been to build an inexpensive RMON probe into the agent of each system. This allows RMON to be widely deployed around the network without costing more than traditional network management. The switch accurately maintains RMON statistics at the maximum line rate of all of its ports.

By default, RMON is disabled. However, even in the disabled state, the switch collects etherStats and you can configure alarms and events.

RMON saves the history, alarm, and event configurations to the configuration file. Runtime data is not stored in the configuration file and is subsequently lost after a system restart.

- Enable or disable the collection of RMON statistics on the switch using the commands:

  ```
  enable rmon
  disable rmon
  ```

By enabling RMON, the switch begins the processes necessary for collecting switch statistics.

## Event Actions

The actions that you can define for each alarm are shown in the following table.

**Table 66: Event Actions**

| Action | High Threshold |
|---|---|
| no action | |
| log | Sends a log message. |
| log-and-trap | Sends both a log message and a trap to all trap receivers. |
| snmp-trap | Sends a trap to all trap receivers. |

To be notified of events using *SNMP* traps, you must configure one or more trap receivers, as described in Using the Simple Network Management Protocol on page 78.

### Display RMON Information

- View the status of RMON polling on the switch (the enable/disable state for RMON polling) using the command:

  ```
  show management
  ```

- View the RMON memory usage statistics for a specific RMON feature (for example, statistics, events, logs, history, or alarms) or for all features using the command:

  ```
  show rmon memory {detail | memoryType}
  ```

### SMON

SMON is the common abbreviation for the Switch Network Monitoring Management Information Base (MIB) system defined by the Internet Engineering Task Force (IETF) document RFC 2613.

SMON is a set of MIB extensions for RMON that allows monitoring of switching equipment from a *SNMP* Manager in greater detail. The supported MIB tables are described in Supported Standards, Protocols, and MIBs on page 1594; smonPrioStatsControlTable and smonPrioStatsTable cannot be supported due to hardware limitations.

Note

\When you delete all the mirroring filters through the portCopyConfigTable, the mirroring is disabled automatically.

## Monitoring CPU Utilization

You can monitor the CPU utilization and history for all of the processes running on the switch.

By viewing this history on a regular basis, you can see trends emerging and identify processes with peak utilization. Monitoring the workload of the CPU allows you to troubleshoot and identify suspect processes before they become a problem. By default, the switch monitors CPU utilization every five seconds. In addition, when CPU utilization of a process exceeds 90% of the regular operating basis, the switch logs an error message specifying the process name and the current CPU utilization for the process.

### Disabling CPU Monitoring

- Disable CPU monitoring using the command:

  ```
  disable cpu-monitoring
  ```

This command disables CPU monitoring on the switch; it does not clear the monitoring interval. Therefore, if you altered the monitoring interval, this command does not return the monitoring interval to five seconds. The next time you enable CPU monitoring, the switch uses the existing configured interval.

## Enable CPU Monitoring

By default, CPU monitoring is enabled and occurs every five seconds. The default CPU threshold value is 90%.

- Enable CPU monitoring using the command:

  `enable cpu-monitoring {interval seconds} {threshold percent}`

  Where the following is true:

  - *seconds*—Specifies the monitoring interval. The default interval is 5 seconds, and the range is 5 to 60 seconds. We recommend the default setting for most network environments.
  - **threshold**—Specifies the CPU threshold value. CPU usage is measured in percentages. The default is 90%, and the range is 0% to 100%.

## Displaying CPU Utilization History

- Display the CPU utilization history of one or more processes using the command:

  `show cpu-monitoring {process name} {slot slotid}`

  Where the following is true:

  - **name**—Specifies the name of the process.
  - **slot**—For a modular chassis, specifies the slot number of the MSM/MM. A specifies the MSM installed in slot A. B specifies the MSM installed in slot B. On a SummitStack, specifies the slot number of the target node. The number is a value from one (1) to eight (8). (This parameter is available only on modular switches and SummitStack.)

  Output from this command includes the following information:

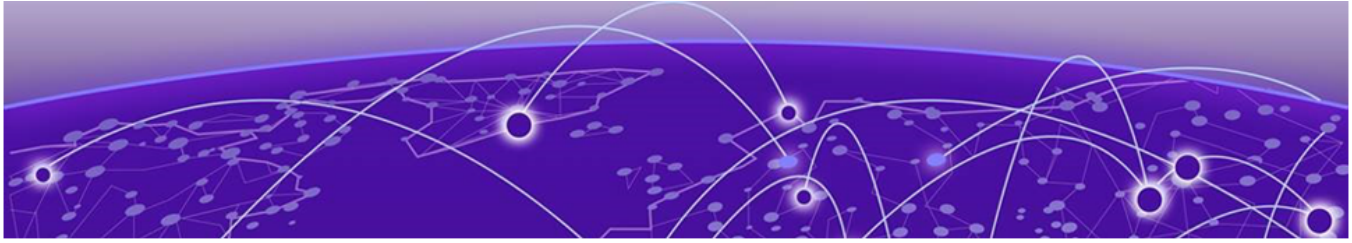
  - Card—The location (MSM A or MSM B) where the process is running on a modular switch.
  - Process—The name of the process.
  - Range of time (five seconds, ten seconds, and so forth)—The CPU utilization history of the process or the system. The CPU utilization history goes back only 1 hour.
  - Total User/System CPU Usage—The amount of time recorded in seconds that the process spends occupying CPU resources. The values are cumulative meaning that the values are displayed as long as the system is running. You can use this information for debugging purposes to see where the process spends the most amount of time: user context or system context.
- Clear the utilization history stored in the switch and reset the statistics to zero using the command:

  `clear cpu-monitoring {process name} {slot slotid}`

The following is sample truncated output from a modular switch:

```
#show cpu-monitoring
CPU Utilization Statistics - Monitored every 5 seconds
-------------------------------------------------------------------------------
Card   Process        5   10   30   1     5    30   1   Max       Total
                     secs secs secs min  mins mins hour User/System
                     util util util util util util util util     CPU Usage
                     (%)  (%)  (%)  (%)   (%)  (%)  (%)  (%)    (secs)
-------------------------------------------------------------------------------
MSM-A  System         0.0  0.0  0.1  0.0  0.0  0.0  0.0  0.9
MSM-B  System         0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
MSM-A  GNSS_cpuif     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0       0.0
MSM-A  GNSS_ctrlif    0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0       0.0
```

```
MSM-A  GNSS_esmi       0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_fabric     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_mac_10g    0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_pbusmux    0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_pktengine  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_pktif      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  GNSS_switch     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0        0.0
MSM-A  aaa             0.0  0.0  0.0  0.0  0.0  0.0  0.0  8.4  0.82       0.56
MSM-A  acl             0.0  0.0  0.0  0.0  0.0  0.0  0.0  7.5  0.37       0.33
MSM-A  bgp             0.0  0.0  0.0  0.0  0.0  0.0  0.0  5.2  0.27       0.42
MSM-A  cfgmgr          0.0  0.9  0.3  3.7  1.2  1.2  1.3  27.3 7.70       7.84
MSM-A  cli             0.0  0.0  0.0  48.3 9.6  2.5  2.1  48.3 0.51       0.37
MSM-A  devmgr          0.0  0.0  0.0  0.9  0.3  0.2  0.2  17.1 2.22       2.50
MSM-A  dirser          0.0  0.0  0.0  0.0  0.0  0.0  0.0  9.5  0.0        0.0
MSM-A  dosprotect      0.0  0.0  0.0  0.0  0.0  0.0  0.0  3.8  0.20       0.26
MSM-A  eaps            1.9  0.9  0.4  0.0  0.0  0.0  0.0  8.4  2.40       1.40
MSM-A  edp             0.0  0.0  0.0  0.0  0.0  0.0  0.0  10.2 0.99       0.47
MSM-A  elrp            0.0  0.0  0.0  0.0  0.0  0.0  0.0  8.4  0.44       0.28
MSM-A  ems             0.0  0.0  0.0  0.0  0.0  0.0  0.0  12.2 1.1        1.16
MSM-A  epm             0.0  0.0  0.0  0.9  0.1  0.2  0.2  4.7  2.6        4.18
MSM-A  esrp            0.0  0.0  0.0  0.0  0.0  0.0  0.0  7.5  0.44       0.36
MSM-A  etmon           0.9  0.4  0.6  1.2  1.1  1.0  1.0  23.3 21.84      7.24
...
```



# VLANs



This chapter contains information about configuring *VLAN (Virtual LAN)*s, displaying VLAN information, private VLANs, and VLAN translation. In addition, you can learn about the benefits and types of VLANs, along with valuable information about virtual routers.

## VLANs Overview

Setting up *VLAN*s on the switch eases many time-consuming tasks of network administration while increasing efficiency in network operations.

Note
The software supports using IPv6 addresses, in addition to IPv4 addresses. You can configure the VLAN with an IPv4 address, IPv6 address, or both. See IPv6 Unicast Routing on page 1294 for complete information on using IPv6 addresses.

The term *VLAN* is used to refer to a collection of devices that communicate as if they were on the same physical LAN.

Note
ExtremeXOS supports only 4,092 user-configurable VLANs. (VLAN 1 is the default VLAN, and 4,095 is the management VLAN, and you may not configure them.

Any set of ports (including all ports on the switch) is considered a VLAN. LAN segments are not restricted by the hardware that physically connects them. The segments are defined by flexible user groups that you create with the CLI.

Note
The system switches traffic within each VLAN using the Ethernet MAC address. The system routes traffic between two VLANs using the IP addresses.

## Benefits

Implementing *VLAN*s on your networks has the following advantages:

- VLANs help to control traffic—With traditional networks, broadcast traffic that is directed to all network devices, regardless of whether they require it, causes congestion. VLANs increase the efficiency of your network because each VLAN can be set up to contain only those devices that must communicate with each other.
- VLANs provide extra security—Devices within each VLAN can communicate only with member devices in the same VLAN. If a device in VLAN Marketing must communicate with devices in VLAN Sales, the traffic must cross a routing device.
- VLANs ease the change and movement of devices—With traditional networks, network administrators spend much of their time dealing with moves and changes. If users move to a different subnetwork, the addresses of each endstation must be updated manually.

## Virtual Routers and VLANs

The ExtremeXOS software supports *virtual router (VR)*s. Each port can belong to multiple virtual routers. Ports can belong to different *VLAN*s that are in different virtual routers.

Note
You can create virtual routers only on BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 switches.

If you do not specify a virtual router when you create a VLAN, the system creates that VLAN in the default virtual router (*VR-Default*). The management VLAN is always in the management virtual router (*VR-Mgmt*).

After you create virtual routers, the ExtremeXOS software allows you to designate one of these virtual routers as the domain in which all your subsequent configuration commands, including VLAN commands, are applied. After you create virtual routers, ensure that you are creating each VLAN in the desired virtual router domain. Also, ensure that you are in the correct virtual router domain before you begin modifying each VLAN.

For information on configuring and using virtual routers, see Virtual Routers on page 624.

## Types of VLANs

This section introduces the following types of VLANs:

- Port-Based VLANs
- Tagged VLANs
- Protocol-Based VLANs

Note
You can have netlogin dynamic VLANs and, on the Summit family of switches and BlackDiamond 8800 series switches only, netlogin MAC-based VLANs. See Network Login on page 756 for complete information on netlogin.

VLANs can be created according to the following criteria:

- Physical port
- IEEE 802.1Q tag
- Ethernet, LLC SAP, or LLC/SNAP Ethernet protocol type
- A combination of these criteria

*Port-Based VLANs*

In a port-based *VLAN*, a VLAN name is given to a group of one or more ports on the switch.

At boot-up, all ports are members of the port-based VLAN default. Before you can add any port to another port-based VLAN, you must remove it from the default VLAN, unless the new VLAN uses a protocol other than the default protocol any. An untagged port can be a member of only one port-based VLAN.

On the Extreme Networks switch in the following figure, ports 9 through 14 are part of VLAN Marketing; ports 25 through 29 are part of VLAN Sales; and ports 21 through 24 and 30 through 32 are in VLAN Finance.

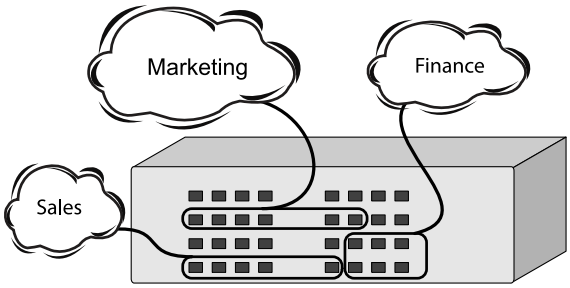



**Figure 65: Example of a Port-Based VLAN on an Extreme Networks Switch**

For the members of different IP VLANs to communicate, the traffic must be routed by the switch, even if the VLANs are physically part of the same I/O module. This means that each VLAN must be configured as a router interface with a unique IP address.

#### Spanning Switches with Port-Based VLANs

To create a port-based *VLAN* that spans two switches, you must do two things:

1. Assign the port on each switch to the VLAN.

2. Cable the two switches together using one port on each switch per VLAN.

   The following figure illustrates a single VLAN that spans a BlackDiamond switch and another Extreme Networks switch. All ports on the System 1 switch belong to VLAN Sales. Ports 1 through 29 on the system 2 switch also belong to VLAN Sales. The two switches are connected using slot 8, port 4 on System 1 (the BlackDiamond switch), and port 29 on system 2 (the other switch).

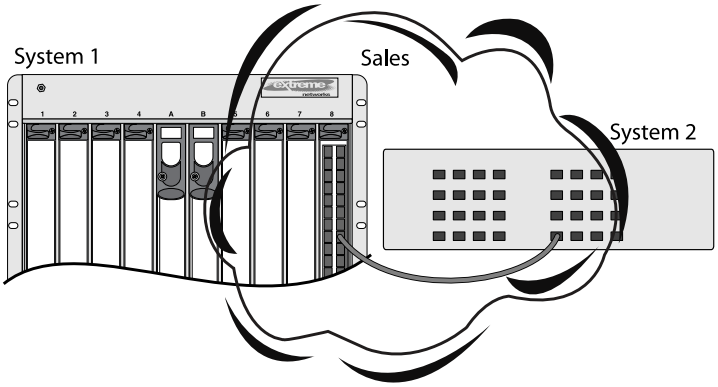



**Figure 66: Single Port-based VLAN Spanning Two Switches**

3. To create multiple VLANs that span two switches in a port-based VLAN, a port on System 1 must be cabled to a port on System 2 for each VLAN you want to have span across the switches.

   At least one port on each switch must be a member of the corresponding VLANs as well.

   The following figure illustrates two VLANs spanning two switches. On System 2, ports 25 through 29 are part of VLAN Accounting; ports 21 through 24 and ports 30 through 32 are part of VLAN Engineering. On System 1, all ports on slot 1 are part of VLAN Accounting; all ports on slot 8 are part of VLAN Engineering.

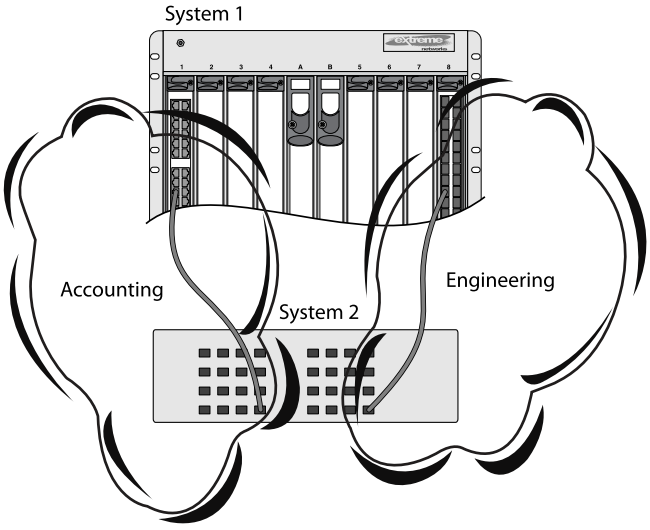



**Figure 67: Two Port-based VLANs Spanning Two Switches**

VLAN Accounting spans System 1 and System 2 by way of a connection between System 2, port 29 and System 1, slot 1, port 6. VLAN Engineering spans System 1 and System 2 by way of a connection between System 2, port 32, and System 1, slot 8, port 6.

4. Using this configuration, you can create multiple port-based VLANs that span multiple switches, in a daisy-chained fashion.

*Tagged VLANs*

Tagging is a process that inserts a marker (called a tag) into the Ethernet frame. The tag contains the identification number of a specific *VLAN*, called the VLANid (valid numbers are 1 to 4094).

Note
The use of 802.1Q tagged packets may lead to the appearance of packets slightly bigger than the current IEEE 802.3/Ethernet maximum of 1,518 bytes. This may affect packet error counters in other devices and may also lead to connectivity problems if non-802.1Q bridges or routers are placed in the path.

**Uses of Tagged VLANs**

Tagging is most commonly used to create VLANs that span switches.

The switch-to-switch connections are typically called *trunks*. Using tags, multiple VLANs can span multiple switches using one or more trunks. In a port-based VLAN, each VLAN requires its own pair of trunk ports, as shown in the following figure. Using tags, multiple VLANs can span two switches with a single trunk.

Another benefit of tagged VLANs is the ability to have a port be a member of multiple VLANs. This is particularly useful if you have a device (such as a server) that must belong to multiple VLANs. The device must have a Network Interface Card (NIC) that supports IEEE 802.1Q tagging.

A single port can be a member of only one port-based VLAN. All additional VLAN membership for the port must be accompanied by tags.

**Assigning a VLAN Tag**

Each VLAN may be assigned an 802.1Q VLAN tag. As ports are added to a VLAN with an 802.1Q tag defined, you decide whether each port uses tagging for that VLAN. The default mode of the switch is to have all ports assigned to the VLAN named default with an 802.1Q VLAN tag (VLANid) of 1 assigned.

Not all ports in the VLAN must be tagged. As traffic from a port is forwarded out of the switch, the switch determines (in real time) if each destination port should use tagged or untagged packet formats for that VLAN. The switch adds and strips tags, as required, by the port configuration for that VLAN.

Note
Packets arriving tagged with a VLANid that is not configured on a port are discarded.

The following figure illustrates the physical view of a network that uses tagged and untagged traffic.

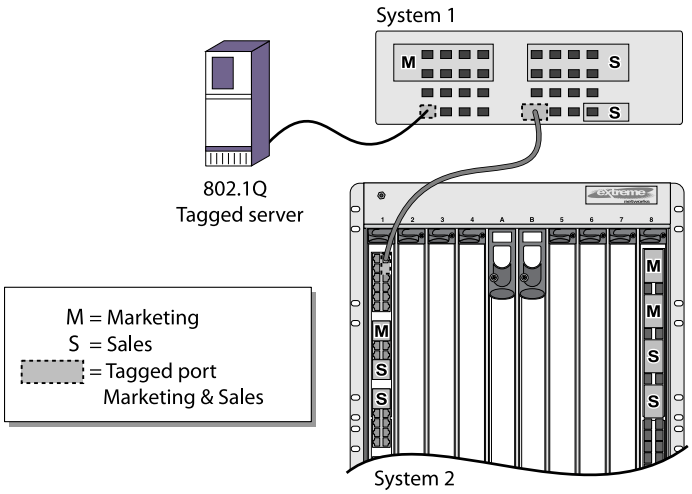



**Figure 68: Physical Diagram of Tagged and Untagged Traffic**

The following figure is a logical diagram of the same network.

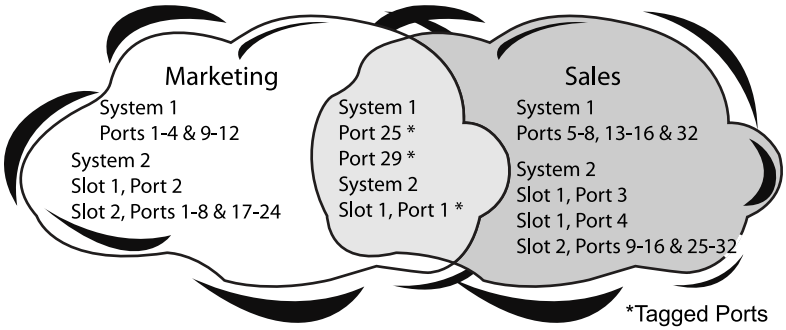



**Figure 69: Logical Diagram of Tagged and Untagged Traffic**

In the figures above:

- The trunk port on each switch carries traffic for both VLAN Marketing and VLAN Sales.
- The trunk port on each switch is tagged.
- The server connected to port 25 on System 1 has a NIC that supports 802.1Q tagging.
- The server connected to port 25 on System 1 is a member of both VLAN Marketing and VLAN Sales.
- All other stations use untagged traffic.

As data passes out of the switch, the switch determines if the destination port requires the frames to be tagged or untagged. All traffic coming from and going to the server is tagged. Traffic coming from and going to the trunk ports is tagged. The traffic that comes from and goes to the other stations on this network is not tagged.

#### Mixing Port-Based and Tagged VLANs

You can configure the switch using a combination of port-based and tagged VLANs. A given port can be a member of multiple VLANs, with the stipulation that only one of its VLANs uses untagged traffic.

In other words, a port can simultaneously be a member of one port-based VLAN and multiple tag-based VLANs.

**Note**
For the purposes of VLAN classification, packets arriving on a port with an 802.1Q tag containing a VLANid of 0 are treated as untagged.

### *Protocol-Based VLANs*

Protocol-based *VLAN*s enable you to define a packet filter that the switch uses as the matching criteria to determine if a particular packet belongs to a particular VLAN.

Protocol-based VLANs are most often used in situations where network segments contain hosts running multiple protocols. For example, in the following figure, the hosts are running both the IP and NetBIOS protocols.

The IP traffic has been divided into two IP subnets, 192.207.35.0 and 192.207.36.0. The subnets are internally routed by the switch. The subnets are assigned different VLAN names, Finance and Personnel, respectively. The remainder of the traffic belongs to the VLAN named MyCompany. All ports are members of the VLAN MyCompany.

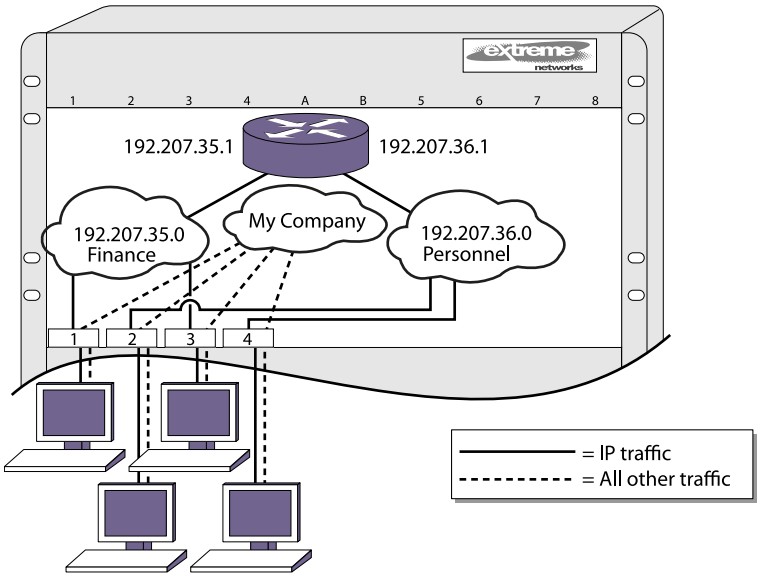



**Figure 70: Protocol-Based VLANs**

The following sections provide information on using protocol-based VLANs:

- Predefined Protocol Filters on page 509
- Defining Protocol Filters on page 509
- Configuring a VLAN to Use a Protocol Filter on page 510
- Deleting a Protocol Filter on page 510

#### Predefined Protocol Filters

The following protocol filters are predefined on the switch:

- IP (IPv4)
- IPv6 (11.2 IPv6)
- *MPLS (Multiprotocol Label Switching)*
- IPX
- NetBIOS
- DECNet
- IPX_8022
- IPX_SNAP
- AppleTalk

### Defining Protocol Filters

If necessary, you can define a customized protocol filter by specifying EtherType, Logical Link Control (LLC), or Subnetwork Access Protocol (SNAP). Up to six protocols can be part of a protocol filter. To define a protocol filter:

1. Create a protocol using the following command:

   `create protocol_name`

   For example: `create protocol fred`

   The protocol name can have a maximum of 32 characters.
2. Configure the protocol using the following command:

   `configure protocol_name add [etype | llc | snap] hex {[etype | llc | snap] hex}`

   Supported protocol types include:

   etype—EtherType

   The values for **etype** are four-digit hexadecimal numbers taken from a list maintained by the IEEE. This list can be found at the following URL: http://standards.ieee.org/regauth/ethertype/index.html.

   **Note**
   Protocol-based *VLAN* for Etype from 0x0000 to 0x05ff are not classifying as per filter. When traffic arrive with these Etypes, it is classifed to native VLAN rather protocol based vlan.

   llc—LLC Service Advertising Protocol (SAP)

   The values for **llc** are four-digit hexadecimal numbers that are created by concatenating a two-digit LLC Destination SAP (DSAP) and a two-digit LLC Source SAP (SSAP).

   snap—EtherType inside an IEEE SNAP packet encapsulation

   The values for **snap** are the same as the values for **etype**, described previously. For example:

   ```
   configure protocol fred add llc feff
   configure protocol fred add snap 9999
   ```

A maximum of 15 protocol filters, each containing a maximum of six protocols, can be defined. No more than seven protocols can be active and configured for use.

Note
For more information on SNAP for Ethernet protocol types, see TR 11802-5:1997 (ISO/IEC) [ANSI/IEEE std. 802.1H, 1997 Edition].

*Configuring a VLAN to Use a Protocol Filter*

To configure a *VLAN* to use a protocol filter, use the following command:.

```
configure {vlan} [vlan_name |vlan_list] protocol {filter}filter_name
```

**Deleting a Protocol Filter**

If a protocol filter is deleted from a *VLAN*, the VLAN is assigned a protocol filter of 'any'. You can continue to configure the VLAN. However, no traffic is forwarded to the VLAN until a protocol is assigned to it.

**Precedence of Tagged Packets Over Protocol Filters**

If a *VLAN* is configured to accept tagged packets on a particular port, incoming packets that match the tag configuration take precedence over any protocol filters associated with the VLAN.

## Default VLAN

The default switch configuration includes one default *VLAN* that has the following properties:

- The VLAN name is default.
- It contains all the ports on a new or initialized switch.
- The default VLAN is untagged on all ports. It has an internal VLANid of 1; this value is user-configurable.

## VLAN Names

*VLAN* names must conform to the guidelines listed in Object Names on page 16.

VLAN names can be specified using the **[Tab]** key for command completion. VLAN names are locally significant. That is, VLAN names used on one switch are only meaningful to that switch. If another switch is connected to it, the VLAN names have no significance to the other switch.

Note
We recommend that you use VLAN names consistently across your entire network.

You must use mutually exclusive names for the following:

- VLANs
- VMANs
- IPv6 tunnels
- SVLANs

- CVLANs
- BVLANs

## Configuring VLANs on the Switch

Refer to the following sections for instruction on configuring VLANs on a switch:



### VLAN Configuration Overview

The following procedure provides an overview of *VLAN* creation and configuration:

1. Create and name the VLAN.

   `create [ {vlan} vlan_name | vlan vlan_list] {tag tag } {description vlan-description } {vr name }`

2. If needed, assign an IP address and mask (if applicable) to the VLAN as described in Managing a VLAN IP Address on page 512.
3. If any ports in this VLAN will use a tag, assign a VLAN tag.

   `configure {vlan} vlan_name tag tag {remote-mirroring}`

4. Assign one or more ports to the VLAN.

   `configure {vlan} [vlan_name | vlan_list] add ports [port_list | all] {tagged tag | untagged} {{stpd} stpd_name} {dot1d | emistp | pvst-plus}}`

   As you add each port to the VLAN, decide if the port will use an 802.1Q tag.

5. For the management VLAN on the switch, configure the default IP route for *VR VR-Mgmt*.

   **Note**
   See IPv4 Unicast Routing on page 1243 for information on configuring default IP routes or adding secondary IP addresses to VLANs.

#### *VLAN ID and VLAN ID List Specification*

As of ExtremeXOS 16.1, you can now refer to a *VLAN* by VID as an alternative. Specifying lists of VIDs is a is a useful shortcut to all users that can reduce the number of commands required to configure the switch.

For example consider a case where two ports should be added to four tagged VLANs. Normally this would require four commands:

```
configure vlan red add ports 2,10 tagged

configure vlan blue add ports 2,10 tagged

configure vlan green add ports 2,10 tagged

configure vlan orange add ports 2,10 tagged
```

Assuming the tags for the VLANs are 100, 200, 300, 400 respectively this configuration can be accomplished with a single command:

```
configure vlan 100,200,300,400 add ports 2, 10 tagged
```

Note

Commands enhanced with VID list support operate in a “best effort” fashion. If one of the VIDs in a VID list do not exist the command is still executed for all of the VIDs in the list that do exist. No error or warning is displayed for the invalid VIDs unless all of the specified VIDs are in valid.

## Creating and Deleting VLANs

- To create a *VLAN*, use the following command:

  `create [ {vlan} vlan_name | vlan vlan_list] {tag tag } {description vlan-description } {vr name }`
- To delete a VLAN, use the following command:

  `delete vlan vlan_name | vlan_list`

## Managing a VLAN IP Address

Note

If you plan to use this *VLAN* as a control VLAN for an EAPS domain, do not assign an IP address to the VLAN.

- Configure an IP address and mask for a VLAN.

  Note

  IPv4 addresses with 31-bit prefixes can be configured on network VLANs and the management VLAN. Applications and protocols can use these IPv4 addresses.

  `configure [ {vlan} vlan_name | vlan vlan_id] ipaddress [ipaddress {netmask} | {ipNetmask} | ipv6-link-local | {eui64} ipv6_address_mask]`

  Note

  Each IP address and mask assigned to a VLAN must represent a unique IP subnet. You cannot configure the same IP subnet on different VLANs on the same *VR*.

  The software supports using IPv6 addresses, in addition to IPv4 addresses. You can configure the VLAN with an IPv4 address, IPv6 address, or both. See IPv6 Unicast Routing on page 1294 for complete information on using IPv6 addresses.

- Remove an IP address and mask for a VLAN.

  `unconfigure [ {vlan} vlan_name | vlan vlan_id]ipaddress {ipv6_address_mask}`

## Configuring a VLAN Tag

To configure a *VLAN*, use the following command:

`configure {vlan} vlan_name tag tag {remote-mirroring}`

## Adding and Removing Ports from a VLAN

- To add ports to a *VLAN*, use the following command:

  `configure {vlan} [vlan_name | vlan_list] add ports [port_list | all] {tagged tag | untagged} {{stpd} stpd_name} {dot1d | emistp | pvst-plus}}`

  The system returns the following message if the ports you are adding are already EAPS primary or EAPS secondary ports:

  ```
  WARNING: Make sure Vlan1 is protected by EAPS, Adding EAPS ring ports to a VLAN could cause a loop in the network.
  Do you really want to add these ports? (y/n)
  ```

- To remove ports from a VLAN, use the following command:

  `configure {vlan} [vlan_name | vlan_list] delete ports [all | port_list {tagged tag}]`

## Adding and Removing VLAN Descriptions

A *VLAN* description is a string of up to 64 characters that you can configure to describe the VLAN. It is displayed by several show vlan commands and can be read by using *SNMP (Simple Network Management Protocol)* to access the VLAN's ifAlias MIB object.

- To add a description to a VLAN, use the following command:

  `configure {vlan} vlan_name description [vlan-description | none]`

- To remove a description from a VLAN, use the following command

  `unconfigure {vlan} vlan_name description`

**Note**

You can add or remove multiple VLAN description in a single command using VLAN IDs or lists (see Using VLAN IDs or Lists Instead of Names on page 18).

## Renaming a VLAN

To rename an existing *VLAN*, use the following command:

`configure {vlan} vlan_name | vlan_id name name`

The following rules apply to renaming VLANs:

- You cannot change the name of the default VLAN.
- You cannot create a new VLAN named default.

## Enabling and Disabling VLANs

You can enable or disable individual *VLAN*s. The default setting is that all VLANs are enabled.

Consider the following guidelines before you disable a VLAN:

- Disabling a VLAN stops all traffic on all ports associated with the specified VLAN.
- You cannot disable any VLAN that is running any Layer 2 protocol traffic.

  When you attempt to disable a VLAN running Layer 2 protocol traffic (for example, the VLAN Accounting), the system returns a message similar to the following:

  ```
  VLAN accounting cannot be disabled because it is actively used by an L2 Protocol
  ```
- You can disable the default VLAN; ensure that this is necessary before disabling the default VLAN.
- You cannot disable the management VLAN.
- You cannot bind Layer 2 protocols to a disabled VLAN.
- You can add ports to and delete ports from a disabled VLAN.

1. Disable a VLAN by running:

   `disable vlan vlan_name | vlan_list`
2. After you have disabled a VLAN, re-enable that VLAN.

   `enable vlan vlan_name | vlan_list`

## VLAN Configuration Examples

**Note**

To add an untagged port to a *VLAN* you create, you must first delete that port from the default VLAN. If you attempt to add an untagged port to a VLAN before deleting it from the default VLAN, you see the following error message:

```
Error: Protocol conflict when adding untagged port 1:2. Either add this port as
tagged or assign another protocol to this VLAN.
```

The following modular switch example creates a port-based VLAN named accounting:

```
create vlan accounting
configure accounting ipaddress 132.15.121.1
configure default delete port 2:1-2:3,2:6,4:1,4:2
configure accounting add port 2:1-2:3,2:6,4:1,4:2
```

**Note**

Because VLAN names are unique, you do not need to enter the keyword **vlan** after you have created the unique VLAN name. You can use the VLAN name alone (unless you are also using this name for another category such as *STPD (Spanning Tree Domain)* or EAPS, in which case we recommend including the keyword **vlan**).

The following stand-alone switch example creates a port-based VLAN named development with an IPv6 address:

```
create vlan development
configure development ipaddress 2001:0DB8::8:800:200C:417A/64
configure default delete port 1-3
configure development add port 1-3
```

The following modular switch example creates a protocol-based VLAN named ipsales.

Slot 5, ports 6 through 8, and slot 6, ports 1, 3, and 4-6 are assigned to the VLAN. In this example, you can add untagged ports to a new VLAN without first deleting them from the default VLAN, because the new VLAN uses a protocol other than the default protocol.

```
create vlan ipsales
configure ipsales protocol ip
configure ipsales add port 5:6-5:8,6:1,6:3-6:6
```

The following modular switch example defines a protocol filter, myprotocol and applies it to the VLAN named myvlan. This is an example only, and has no real-world application.

```
create protocol myprotocol
configure protocol myprotocol add etype 0xf0f0
configure protocol myprotocol add etype 0xffff
create vlan myvlan
configure myvlan protocol myprotocol
```

To disable the protocol-based VLAN (or any VLAN) in the above example, use the following command:

```
disable vlan myprotocol
```

To re-enable the VLAN, use the following command:

```
enable vlan myprotocol
```

## Displaying VLAN Information

To display general *VLAN* settings and information, use the following commands:

- `show` **`vlan`** {**`virtual-router`** *`vr-name`*}
- `show` **`vlan`** *`vlan_name`* {**`ipv4`** | **`ipv6`**}
- `show` **`vlan`** [**`tag`** *`tag`* | **`detail`**] {**`ipv4`** | **`ipv6`**}
- `show vlan description`
- `show vlan` {*`vlan_name`* | *`vlan_list`*} **`statistics`** {**`no-refresh`** | **`refresh`** }

Note

To display IPv6 information, you must use either the show vlan detail command or show vlan command with the name of the specified VLAN.

To display the VLAN information for other ExtremeXOS software features, use the following commands:

- `show` {**`vlan`**} *`vlan_name`* **`dhcp-address-allocation`**
- `show` {**`vlan`**} *`vlan_name`* **`dhcp-config`**

- `show {vlan} vlan_name eaps`
- `show {vlan} vlan_name security`
- `show {vlan} {vlan_name | vlan_list} stpd`
- `show vid {vlan_list}`

You can display additional useful information on VLANs configured with IPv6 addresses by issuing the command:

```
show ipconfig ipv6 vlan vlan_name
```

To isplay protocol information, issue the command:

```
show protocol {name}
```

# Private VLANs

The following sections provide detailed information on private VLANs:


- PVLAN Overview on page 516
- Configuring PVLANs on page 524
- Displaying PVLAN Information on page 528
- PVLAN Configuration Example 1 on page 529
- PVLAN Configuration Example 2 on page 530


## PVLAN Overview

PVLANs offer the following features:

- *VLAN* translation
- VLAN isolation

**Note**
PVLAN features are supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

### *VLAN Translation in a PVLAN*

*VLAN* translation provides the ability to translate the 802.1Q tags for several VLANs into a single VLAN tag. VLAN translation is an optional component in a PVLAN.

VLAN translation allows you to aggregate Layer 2 VLAN traffic from multiple clients into a single uplink VLAN, improving VLAN scaling. The following figure shows an application of VLAN translation.

**Note**
The VLAN translation feature described in VLAN Translation on page 534 is provided for those who are already familiar with the ExtremeWare VLAN translation feature. If you have time to use the PVLAN implementation and do not have scripts that use the ExtremeWare commands, we suggest that you use the PVLAN feature, as it provides the same functionality with additional features.

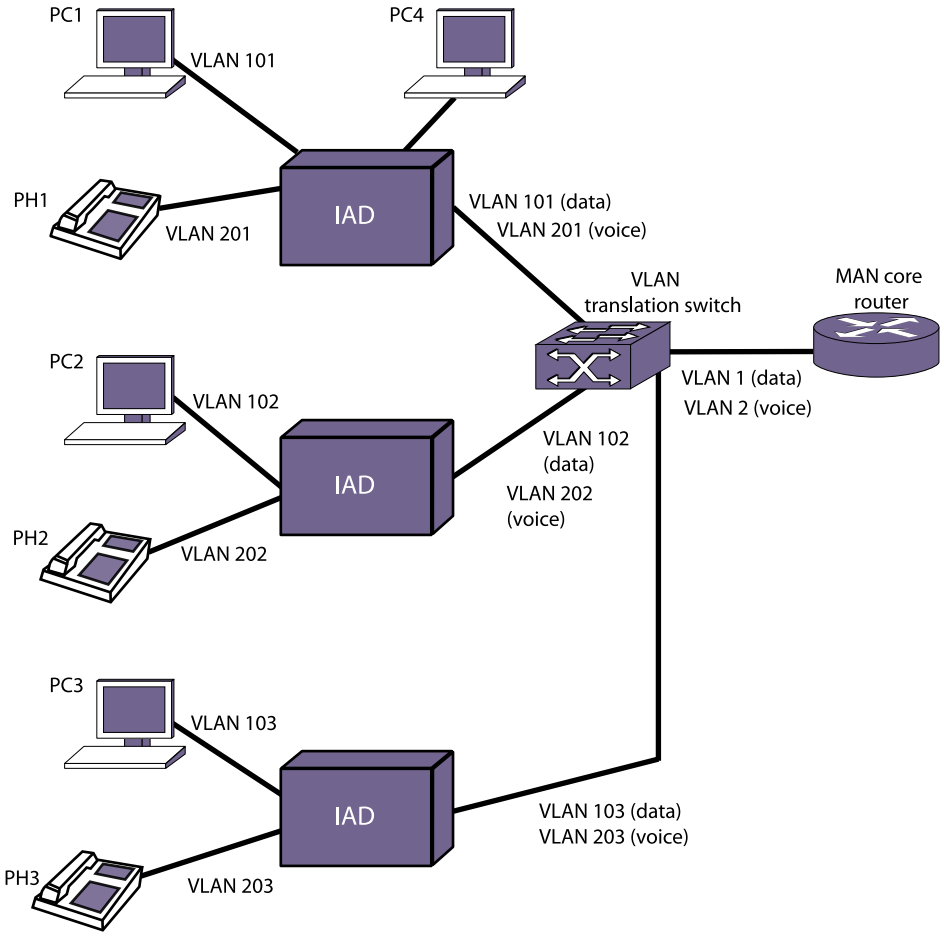



**Figure 71: VLAN Translation Application**

In the figure, VLANs 101, 102, and 103 are subscriber VLANS that carry data traffic while VLANs 201, 202, and 203 are subscriber VLANs that carry voice traffic. The voice and data traffic are combined on integrated access devices (IADs) that connect to the VLAN translation switch. Each of the three clusters of phones and PCs uses two VLANs to separate the voice and data traffic. As the traffic is combined, the six VLANs are translated into two network VLANs, VLAN1 and VLAN2. This simplifies administration, and scales much better for large installations.

Conceptually, this is very similar to Layer 3 VLAN aggregation (superVLANS and subVLANs).

The primary differences between these two features are:

- VLAN translation is strictly a Layer 2 feature.
- VLAN translation does not allow communication between the subscriber VLANs.

### *VLAN Isolation*

*VLAN* isolation provides Layer 2 isolation between the ports in a VLAN. The following figure shows an application of VLAN isolation.

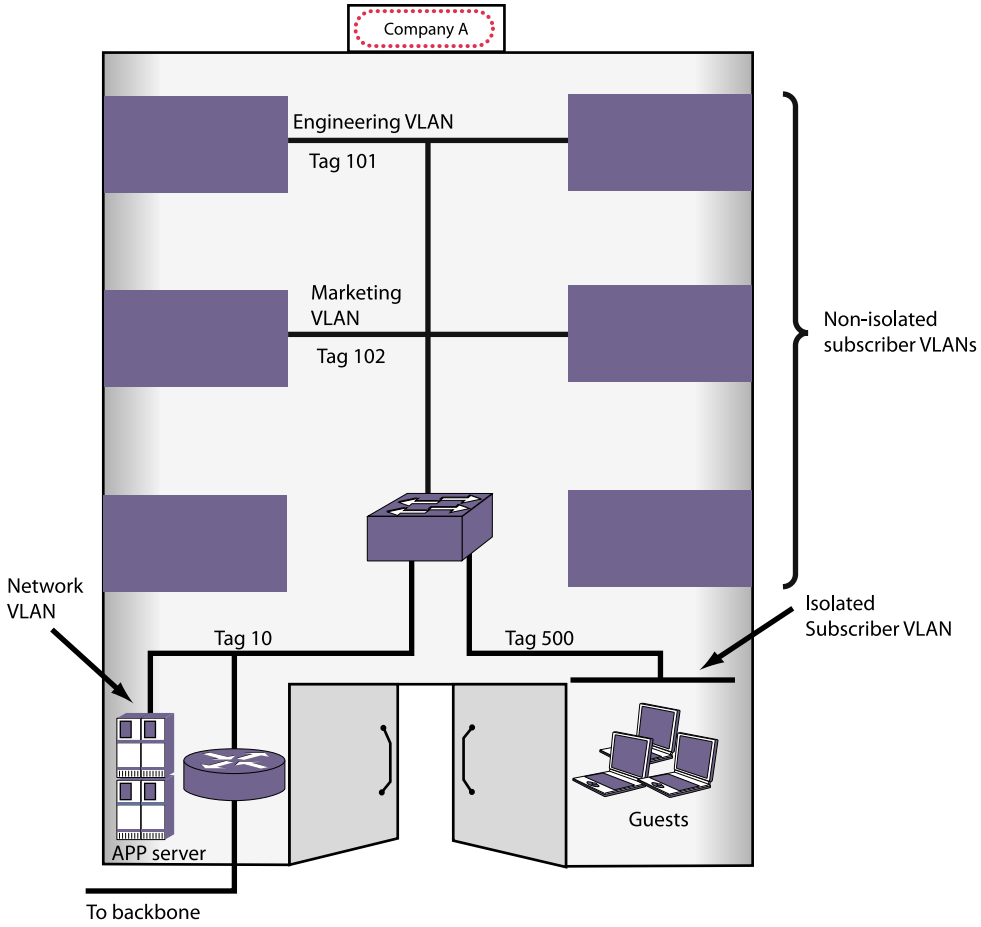



**Figure 72: VLAN Isolation Application**

In this figure, ports in the Guest VLAN have access to services on the network VLAN, but Guest VLAN ports cannot access other Guest VLAN ports over Layer 2 (or the Marketing or Engineering VLANs). This provides port-to-port security at Layer 2.

### *PVLAN Components*

The following figure shows the logical components that support PVLAN configuration in a switch.

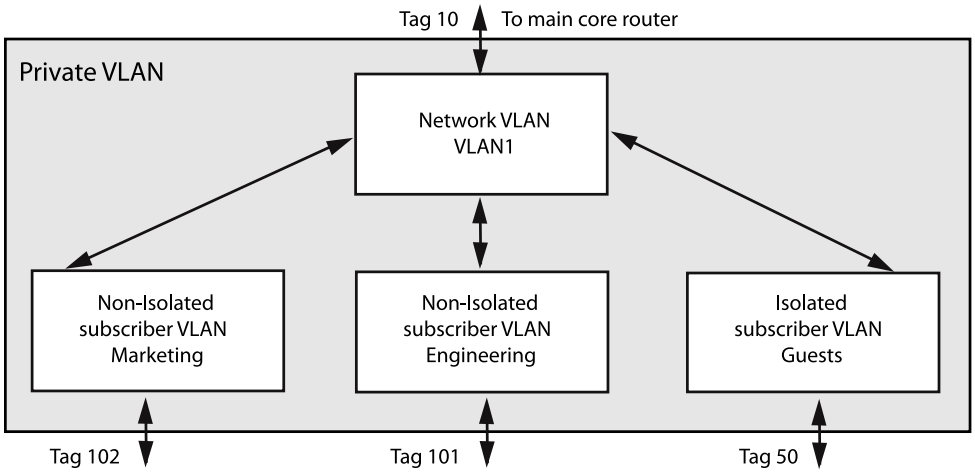



**Figure 73: Private VLAN Switch Components**

There is one network *VLAN* in each PVLAN. Ports within a network VLAN, called network ports, can communicate with all VLAN ports in the PVLAN. Network devices that connect to the network VLAN ports are considered to be on the network side of the switch.

The network VLAN aggregates the uplink traffic from the other VLANS, called subscriber VLANs, for egress communications on a network VLAN port. A network port can serve only one PVLAN, but it can serve one or more subscriber VLANs. Ingress communications on the network VLAN port are distributed to the appropriate subscriber VLANs for distribution to the appropriate ports. Devices that connect to subscriber VLAN ports are considered to be on the subscriber side of the switch.

Note
PVLAN network-tagged packets are allowed to ingress on subscriber VLAN ports.

Tag translation within the PVLAN is managed at the egress ports. To enable tag translation for uplink traffic from the subscriber VLANs, you must enable tag translation on the appropriate network VLAN port. Tag translation is automatically enabled on subscriber VLAN egress ports when the subscriber VLAN is created and the port is added to the VLAN as tagged. Egress traffic from a subscriber VLAN is always tagged with the subscriber VLAN tag when the port is configured as tagged.

A non-isolated subscriber VLAN is basically a standard VLAN that can participate in tag translation through the network VLAN when VLAN translation is enabled on the network VLAN port.

You can choose to not translate tags on a network VLAN port, but this is generally used only for extending a PVLAN to another switch. A non-isolated subscriber VLAN that does not use tag translation is functionally equivalent to a regular VLAN, so it is better to create non-isolated VLANs only when you plan to use tag translation.

Ports in a non-isolated VLAN can communicate with other ports in the same VLAN, ports in the network VLAN, and destinations on the network side of the switch. As with standard VLANs, non-isolated ports cannot communicate through Layer 2 with ports in other subscriber VLANs.

In the figure above, the Engineering and Marketing VLANs are configured as non-isolated subscriber VLANs, which means that they act just like traditional VLANs, and they can participate in tag translation when VLAN translation is enabled on a network VLAN port that leads to network side location.

VLAN isolation within the PVLAN is established by configuring a VLAN to be an isolated subscriber VLAN and adding ports to the isolated VLAN. Unlike normal VLANs, ports in an isolated VLAN cannot communicate with other ports in the same VLAN over Layer 2 or Layer 3. The ports in an isolated VLAN can, however, communicate with Layer 2 devices on the network side of the PVLAN through the network VLAN. When the network VLAN egress port is configured for tag translation, isolated VLAN ports also participate in uplink tag translation. When isolated subscriber VLAN ports are configured as tagged, egress packets are tagged with the isolated VLAN tag. As with standard VLANs and non-isolated VLANs, isolated ports cannot communicate through Layer 2 with ports in other subscriber VLANs.

### *PVLAN Support over Multiple Switches*

A PVLAN can span multiple switches. The following figure shows a PVLAN that is configured to operate on two switches.

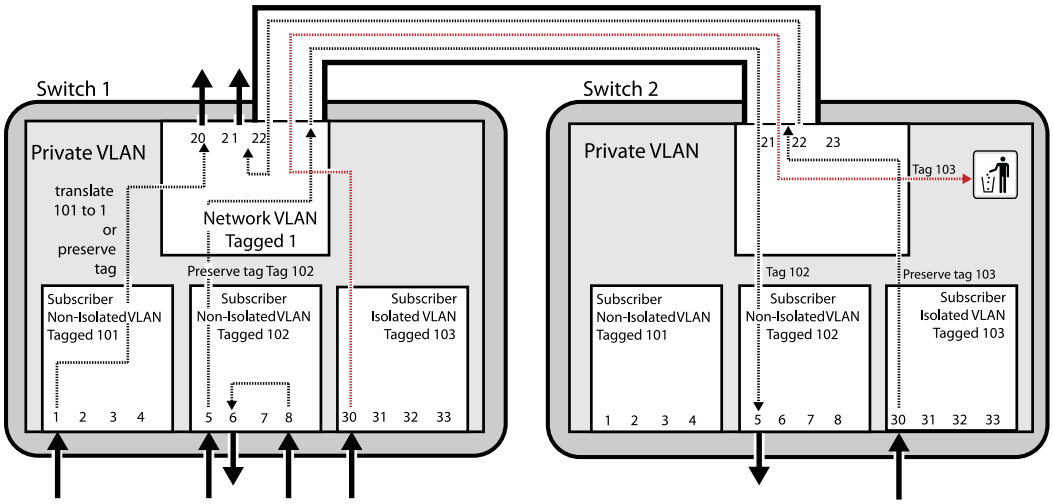



**Figure 74: Private VLAN Support on Multiple Switches**

A PVLAN can span many switches. For simplicity, the figure above shows only two switches, but you can extend the PVLAN to additional switches by adding connections between the network *VLAN*s in each switch. The ports that connect two PVLAN switches must be configured as regular tagged ports. The network and subscriber VLANs on each switch must be configured with the same tags.

Note

Although using the same VLAN names on all PVLAN switches might make switch management easier, there is no software requirement to match the VLAN names. Only the tags must match.

When a PVLAN is configured on multiple switches, the PVLAN switches function as one PVLAN switch. Subscriber VLAN ports can access the network VLAN ports on any of the PVLAN switches, and non-isolated VLAN ports can communicate with ports in the same VLAN that are located on a different physical switch. An isolated VLAN can span multiple switches and maintain isolation between the VLAN ports.

The network and subscriber VLANs can be extended to other switches that are not configured for the PVLAN (as described in Extending Network and Subscriber VLANs to Other Switches on page 521). The advantage to extending the PVLAN is that tag translation and VLAN isolation is supported on the additional switch or switches.

*Extending Network and Subscriber VLANs to Other Switches*

A network or subscriber *VLAN* can be extended to additional switches without a PVLAN configuration on the additional switches.

You might want to do this to connect to existing servers, switches, or other network devices. You probably do not want to use this approach to support clients, as tag translation and VLAN isolation are not supported unless the PVLAN is configured on all PVLAN switches as described in PVLAN Support over Multiple Switches on page 520.

The following figure illustrates PVLAN connections to switches outside the PVLAN.

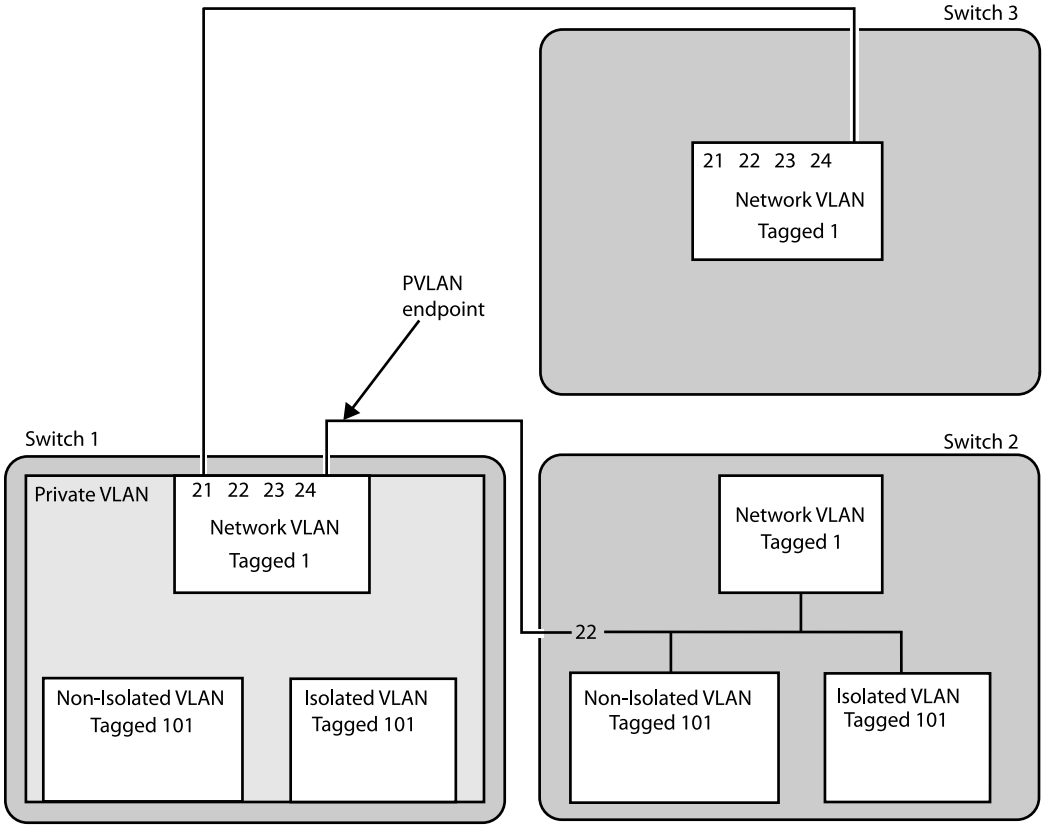



**Figure 75: Private VLAN Connections to Switches Outside the PVLAN**

In the above figure, Switch 1, Network VLAN Port 21 connects to a Switch 3 port that only supports the Network VLAN.

In this configuration, the Network VLAN Port 21 on Switch 1 is configured as "translated," which translates subscriber VLAN tags to the network VLAN tag for access to the Network VLAN extension on Switch 3. Switch 3, Port 24 is configured as tagged and only accepts traffic with the Network VLAN Tag. Switch 3 serves as an extension of the Network VLAN and can be used to connect to network devices such as servers or an internet gateway.

Switch 2, port 22 supports the Network, NonIsolated, and Isolated VLANs, but no PVLAN is configured.

Because port 22 supports multiple VLANs that are part of the PVLAN, and because these Switch 2 VLANs are not part of the PVLAN, Switch 1, port 24, must be configured as a PVLAN endpoint, which establishes the PVLAN boundary. Switch 2, port 22, is configured as a regular tagged VLAN port.

For most applications, it would be better to extend the PVLAN to Switch 2 so that the PVLAN features are available to the Switch 2 VLANs.

The configuration of Switch 2 behaves as follows:

- The Switch 2 NonIsolated VLAN ports can communicate with the NonIsolated VLAN ports on Switch 1, but they cannot participate in VLAN translation.
- The Switch 2 Isolated VLAN ports can communicate with other Switch 2 Isolated VLAN ports.
- The Switch 2 Isolated VLAN ports cannot participate in VLAN translation.
- The Switch 2 Isolated VLAN ports can receive broadcast and multicast info for the Isolated VLAN.
- Traffic is allowed from the Switch 1 Isolated VLAN ports to the Switch 2 Isolated VLAN ports.

### *MAC Address Management in a PVLAN*

Each device that connects to a PVLAN must have a unique MAC address within the PVLAN. Each MAC address learned in a PVLAN requires multiple *FDB (forwarding database)* entries. For example, each MAC address learned in a non-isolated subscriber *VLAN* requires two FDB entries, one for the subscriber VLAN and one for the network VLAN. The additional FDB entries for a PVLAN are marked with the P flag in the `show fdb` command display.

The following sections describe the FDB entries created for the PVLAN components and how to estimate the impact of a PVLAN on the FDB table:

- Non-Isolated Subscriber VLAN
- Isolated Subscriber VLAN
- Network VLAN
- Calculating the Total FDB Entries for a PVLAN

#### Non-Isolated Subscriber VLAN

When a MAC address is learned on a non-isolated subscriber VLAN port, two entries are added to the FDB table:

- MAC address, non-isolated subscriber VLAN tag, and the port number
- MAC address, network VLAN tag, port number, and a special flag for tag translation

The network VLAN entry is used when traffic comes in from the network ports destined for an non-isolated port.

#### Isolated Subscriber VLAN

When a new MAC address is learned on an isolated subscriber VLAN port, two entries are added to the FDB table:

- MAC address, isolated subscriber VLAN tag, port number, and a flag that indicates that the packet should be dropped
- MAC address, network VLAN tag, port number, and a special flag for tag translation

Ports in the isolated VLAN do not communicate with one another.

If a port in the isolated VLAN sends a packet to another port in the same VLAN that already has an entry in the FDB, that packet is dropped. You can verify the drop packet status of an FDB entry by using the `show fdb` command. The D flag indicates that packets destined for the listed address are dropped.

The network VLAN entry is used when traffic comes in from the network ports destined for an isolated port.

**Network VLAN**

When a new MAC address is learned on a network VLAN port, the following entry is added to the FDB table: MAC address, network VLAN tag, and port number.

For every subscriber VLAN belonging to this PVLAN, the following entry is added to the FDB table: MAC address, subscriber VLAN tag, and port number

**Calculating the Total FDB Entries for a PVLAN**

The following formula can be used to estimate the maximum number of FDB entries for a PVLAN:

FDBtotal = [(MACnon-iso + MACiso) * 2 + (MACnetwork * (VLANnon-iso + VLANiso + 1))]

The formula components are as follows:

- MACnon-iso = number of MAC addresses learned on all the non-isolated subscriber VLANs
- MACiso = number of MAC addresses learned on all the isolated subscriber VLANs
- MACnetwork = number of MAC addresses learned on the network VLAN
- VLANnon-iso = number of non-isolated subscriber VLANs
- VLANiso = number of isolated subscriber VLANs

Note
The formula above estimates the worst-case scenario for the maximum number of FDB entries for a single PVLAN. If the switch supports additional PVLANs, apply the formula to each PVLAN and add the totals for all PVLANs. If the switch also support standard VLANs, there will also be FDB entries for the standard VLANs.

### *Layer 3 Communications*

For PVLANs, the default switch configuration controls Layer 3 communications exactly as communications are controlled in Layer 2.

For example, Layer 3 communications is enabled between ports in a non-isolated subscriber *VLAN*, and disabled between ports in an isolated subscriber VLAN. Ports in a non-isolated subscriber VLAN cannot communicate with ports in other non-isolated subscriber VLANs.

You can enable Layer 3 communications between all ports in a PVLAN. For more information, see Managing Layer 3 Communications in a PVLAN on page 527.

### *PVLAN Limitations*

The Private *VLAN* feature has the following limitations:

- Requires more *FDB* entries than a standard VLAN.
- Within the same VR, VLAN tag duplication is not allowed.
- Within the same VR, VLAN name duplication is not allowed.
- Each MAC address learned in a PVLAN must be unique. A MAC address cannot exist in two or more VLANs that belong to the same PVLAN.
- MVR cannot be configured on PVLANs.

- A VMAN cannot be added to a PVLAN.
- A PBB network (BVLAN) cannot be added to a PVLAN.
- EAPS control VLANs cannot be either subscriber or network VLANs.
- For PVLAN with *STP (Spanning Tree Protocol)* implementation, irrespective of port translation configuration in the Network VLAN, it is recommended to add both the Network VLAN and all subscriber VLANs to the STP.
- For PVLAN with EAPS implementation, irrespective of port translation configuration in the Network VLAN, it is recommended to add both the Network VLAN and all subscriber VLANs to the EAPS ring.
- *ESRP (Extreme Standby Router Protocol)* can only be configured on network VLAN ports (and not on subscriber VLAN ports). To support ESRP on the network VLAN, you must add all of the VLANs in the PVLAN to ESRP.
- There is no *NetLogin* support to add ports as translate to the network VLAN, but the rest of NetLogin and the PVLAN features do not conflict.
- *IGMP (Internet Group Management Protocol)* snooping is performed across the entire PVLAN, spanning all the subscriber VLANs, following the PVLAN rules. For VLANs that are not part of a PVLAN, IGMP snooping operates as normal.
- PVLAN and VPLS are not supported on the same VLAN.
- When two switches are part of the same PVLAN, unicast and multicast traffic require a tagged trunk between them that preserves tags (no tag translation).
- Subscriber VLANs in a PVLAN cannot exchange multicast data with VLANs outside the PVLAN and with other PVLANs. However, the network VLAN can exchange multicast data with VLANs outside the PVLAN and with network VLANs in other PVLANs.

Note

A maximum of 80% of 4K VLANs can be added to a PVLAN. Adding more VLANS will display the following log error:

```
<Erro:HAL.VLAN.Error>Slot-<slot>: Failed to add egress vlan translation entry on
port <port> due to "Table full".
```

An additional limitation applies to BlackDiamond 8000 series modules and Summit family switches, whether or not they are included in a SummitStack. If two or more member VLANs have overlapping ports (where the same ports are assigned to both VLANs), each additional VLAN member with overlapping ports must have a dedicated loopback port. To state it another way, one of the VLAN members with overlapping ports does not require a dedicated loopback port, and the rest of the VLAN members do require a single, dedicated loopback port within each member VLAN.

Note

There is a limit to the number of unique source MAC addresses on the network VLAN of a PVLAN that the switch can manage. It is advised not to exceed the value shown in the Supported Limits table of the *ExtremeXOS Release Notes*.

## Configuring PVLANs

The following section describes how to configure a private *VLAN*.

### *Creating PVLANs*

To create a *VLAN*, you need to do the following:

1. Create the PVLAN.
2. Add one VLAN to the PVLAN as a network VLAN.
3. Add VLANs to the PVLAN as subscriber VLANs.

- To create a PVLAN, use the following command:

  `create private-vlan` *`name`* {**`vr`** *`vr_name`*}

- To add a network VLAN to the PVLAN, create and configure a tagged VLAN, and then use the following command to add that network VLAN:

  `configure private-vlan` *`name`* **`add network`** *`vlan_name`*

- To add a subscriber VLAN to the PVLAN, create and configure a tagged VLAN, and then use the following command to add that subscriber VLAN:

  `configure private-vlan` *`name`* **`add subscriber`** *`vlan_name`* {**`non-isolated`**} {**`loopback-port`** *`port`*}

  By default, this command adds an isolated subscriber VLAN. To create a non-isolated subscriber VLAN, you must include the **`non-isolated`** option.

### *Configuring Network VLAN Ports for VLAN Translation*

When subscriber *VLAN* traffic exits a network VLAN port, it can be untagged, tagged (with the subscriber VLAN tag), or translated (to the network VLAN tag).

Note

All traffic that exits a subscriber VLAN port uses the subscriber VLAN tag, unless the port is configured as untagged. There is no need to configure VLAN translation (from network to subscriber VLAN tag) on subscriber VLAN ports.

1. To configure network VLAN ports for VLAN translation, use the following command and specify the network VLAN and port numbers:

   `configure` {**`vlan`**} *`vlan_name`* | *`vlan_list`* **`add ports`** *`port_list`* **`private-vlan translated`**

2. If you want to later reconfigure a port that is configured for VLAN translation so that it does not translate tags, use the following command and specify either the **`tagged`** or the **`untagged`** option:

   `configure` {**`vlan`**} *`vlan_name`* | *`vlan_list`* **`add ports`** [*`port_list`* | **`all`**] {**`tagged`** | **`untagged`**} {{**`stpd`**} *`stpd_name`*} {**`dot1d`** | **`emistp`** | **`pvst-plus`**}}

### *Configuring Non-Isolated Subscriber VLAN Ports*

The process for configuring non-isolated *VLAN* ports requires two tasks:

- Add a VLAN to the PVLAN as a non-isolated subscriber VLAN.
- Assign ports to the non-isolated subscriber VLAN.

These tasks can be completed in any order, but they must both be completed before a port can participate in a PVLAN. When configuration is complete, all egress traffic from the port is translated to the VLAN tag for that non-isolated VLAN (unless the port is configured as untagged).

Note

To configure VLAN translation for network VLAN ports, see Configuring Network VLAN Ports for VLAN Translation on page 525.

- To add a non-isolated subscriber VLAN to the PVLAN, use the following command:
  `configure private-vlan` *`name`* **`add subscriber`** *`vlan_name`* **`non-isolated`**
- To add ports to a non-isolated VLAN (before or after it is added to the PVLAN), use the following command:
  `configure {`**`vlan`**`}` *`vlan_name`* `|` *`vlan_list`* **`add ports`** `[`*`port_list`* `|` **`all`**`]`
  `{`**`tagged`** `|` **`untagged`**`} {{`**`stpd`**`}` *`stpd_name`*`} {`**`dot1d`** `|` **`emistp`** `|` **`pvst-plus`**`}}`

If you specify the tagged option, egress traffic uses the non-isolated VLAN tag, regardless of the network translation configuration on any network port with which these ports communicate. Egress traffic from a non-isolated VLAN port never carries the network VLAN tag.

### *Configuring Isolated Subscriber VLAN Ports*

When a port is successfully added to an isolated *VLAN*, the port is isolated from other ports in the same VLAN, and all egress traffic from the port is translated to the VLAN tag for that VLAN (unless the port is configured as untagged).

Note

To configure VLAN translation for network VLAN ports, see Configuring Network VLAN Ports for VLAN Translation on page 525.

The process for configuring ports for VLAN isolation requires two tasks:

- Add a VLAN to the PVLAN as an isolated subscriber VLAN.
- Assign ports to the isolated subscriber VLAN.

These tasks can be completed in any order, but they must both be completed before a port can participate in an isolated VLAN.

- To add an isolated subscriber VLAN to the PVLAN, use the following command:
  `configure private-vlan` *`name`* **`add subscriber`** *`vlan_name`*
- To add ports to an isolated VLAN (before or after it is added to the PVLAN), use the following command:
  `configure {`**`vlan`**`}` *`vlan_name`* `|` *`vlan_list`* **`add ports`** `[`*`port_list`* `|` **`all`**`]`
  `{`**`tagged`** `|` **`untagged`**`} {{`**`stpd`**`}` *`stpd_name`*`} {`**`dot1d`** `|` **`emistp`** `|` **`pvst-plus`**`}}`

If you specify the tagged option, egress traffic uses the isolated VLAN tag, regardless of the network translation configuration on any network port with which these ports communicate. Egress traffic from an isolated VLAN port never carries the network VLAN tag.

### *Configuring a PVLAN on Multiple Switches*

To create a PVLAN that runs on multiple switches, you must configure the PVLAN on each switch and set up a connection between the network VLANs on each switch. The ports at each end of the connection must be configured as tagged ports that do not perform tag translation.

To configure these types of ports, use the following command:

```
configure {vlan} vlan_name add ports port_list tagged
```

*Configuring a Network or Subscriber VLAN Extension to Another Switch*

You can extend a network or subscriber *VLAN* to another switch without configuring a PVLAN on that switch. This configuration is introduced in Extending Network and Subscriber VLANs to Other Switches on page 521.

- To configure the port on the switch that is outside of the PVLAN, use the following command:

  ```
  configure {vlan} vlan_name add ports port_list tagged
  ```

*Adding a Loopback Port to a Subscriber VLAN*

BlackDiamond 8000 series modules and Summit family switches, whether or not included in a SummitStack, require a loopback port for certain configurations. If two or more subscriber *VLAN*s have overlapping ports (where the same ports are assigned to both VLANs), each of the subscriber VLANs with overlapping ports must have a dedicated loopback port.

The loopback port can be added when the subscriber VLAN is added to the PVLAN.

If you need to add a loopback port to an existing subscriber VLAN, use the following command:

```
configure {vlan} vlan_name vlan-translation add loopback-port port
```

*Managing Layer 3 Communications in a PVLAN*

The default configuration for Layer 3 PVLAN communications is described in Layer 3 Communications.

To enable Layer 3 communications between all ports in a PVLAN, use the following command:

```
configure iparp add proxy [ipNetmask | ip_addr {mask}] {vr vr_name} {mac
| vrrp} {always}
```

Specify the IP address or subnet specified for the network *VLAN* in the PVLAN. Use the **always** option to ensure that the switch will reply to ARP requests, regardless of the VLAN from which it originated.

*Delete PVLANs*

To delete an existing PVLAN, use the command:

```
delete private-vlan name
```

*Remove a VLAN from a PVLAN*

When you remove a *VLAN* from a PVLAN, you remove the association between a VLAN and the PVLAN. Both the VLAN and PVLAN exist after the removal.

To remove a network or subscriber VLAN from a PVLAN, use the following command:

```
configure private-vlan name delete [network | subscriber] vlan_name
```

*Deleting a Loopback Port from a Subscriber VLAN*

To delete a loopback port from a subscriber *VLAN*, use the command:

```
configure {vlan} vlan_name vlan-translation delete loopback-port
```

## Displaying PVLAN Information

This section describes how to display private *VLAN* information.

### *Displaying Information for all PVLANs*

To display information on all the PVLANs configured on a switch, use the command:

```
show private-vlan
```

### *Displaying Information for a Specific PVLAN*

To display information about a single PVLANs, use the command:

```
show {private-vlan} name
```

### *Displaying Information for a Network or Subscriber VLAN*

To display information about a network or subscriber *VLAN*, use the command:

```
show vlan {virtual-router vr-name}
```

The following flags provide PVLAN specific information:

**s flat**

Identifies a network VLAN port that the system added to a subscriber VLAN. All subscriber VLANs contain network VLAN ports that are marked with the s flag.

**L flag**

Identifies a subscriber VLAN port that is configured as a loopback port. Loopback ports are supported only on BlackDiamond 8000 series modules and Summit family switches, whether or not included in a SummitStack.

**t flag**

Identifies a tagged network VLAN port on which tag translation is enabled. The t flag only appears in the show vlan display for network VLANs.

**e flag**

Identifies a network VLAN port that is configured as an endpoint. The e flag only appears in the show vlan display for network VLANs.

### *Displaying PVLAN FDB Entries*

To view all *FDB* entries including those created for a PVLAN, use the command:

```
show fdb {blackhole {netlogin [all | mac-based-vlans]} | netlogin [all |
mac-based-vlans] | permanent {netlogin [all | mac-based-vlans]} |
mac_addr {netlogin [all | mac-based-vlans]} | ports port_list {netlogin
[all | mac-based-vlans]} | vlan vlan_name | vlan_list {netlogin [all |
mac-based-vlans]} | {{vpls} {vpls_name}}}
```

The P flag marks additional FDB entries for PVLANs.

## PVLAN Configuration Example 1

The following figure shows a PVLAN configuration example for a medical research lab.

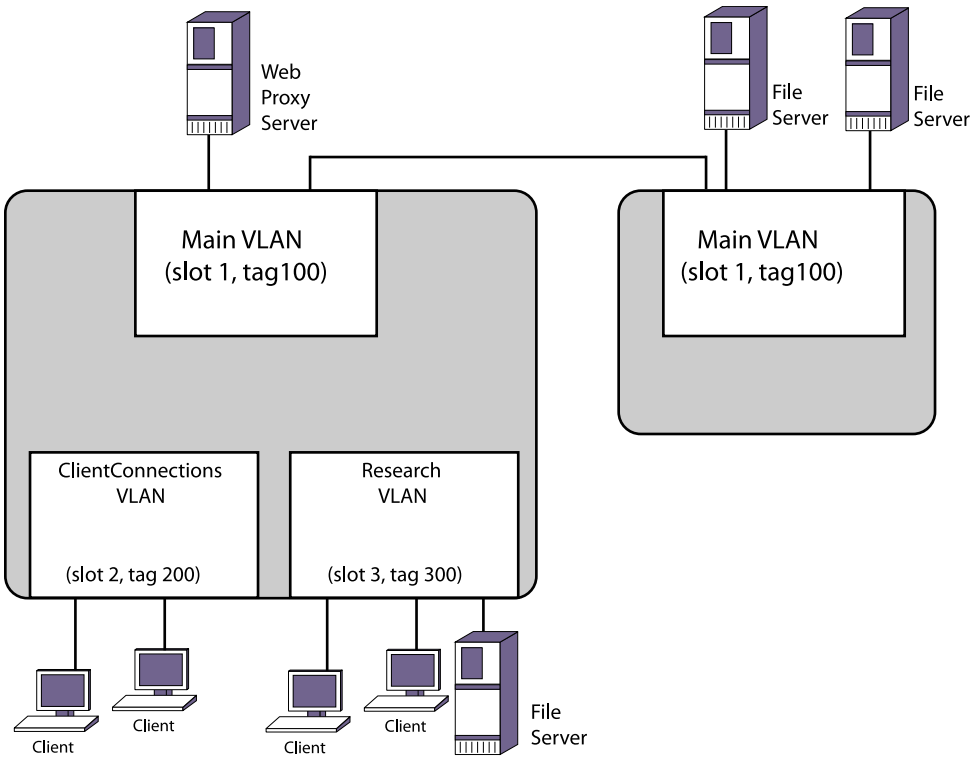



**Figure 76: PVLAN Configuration Example 1**

The medical research lab hosts lots of visiting clients. Each client has their own room, and the lab wants to grant them access to the internet through a local web proxy server but prevent them from accessing other visiting clients. There is a lab in the building where many research workstations are located. Workstations within the lab require access to other lab workstations, the internet, and file servers that are connected to a switch in another building. Visiting clients should not have access to the Research *VLAN* devices or the file servers on the remote switch.

The PVLAN in the following figure contains the following PVLAN components:

- Network VLAN named Main, which provides internet access through the proxy web server and access to file servers on the remote switch.
- Isolated subscriber VLAN named ClientConnections, which provides internet access for visiting clients and isolation from other visiting clients, the Research VLAN devices, and the remote file servers.
- Non-isolated subscriber VLAN named Research, which provides internet access and enables communications between Research VLAN devices and the remote file servers.

1. The first configuration step is to create and configure the VLANs on the local switch:

```
create vlan Main
configure vlan Main add port 1:*
configure vlan Main tag 100
create vlan ClientConnections
configure vlan ClientConnections add port 2:*
```

```
configure vlan ClientConnections tag 200
create vlan Research
configure vlan Research add port 3:*
configure vlan Research tag 300
```

2. The remote switch VLAN is configured as follows:

```
create vlan Main
configure vlan Main add port 1:*
configure vlan Main tag 100
```

3. The next step is to create the PVLAN on the local switch and configure each of the component VLANs for the proper role:

```
create private-vlan MedPrivate
configure private-vlan "MedPrivate" add network "Main"
configure private-vlan "MedPrivate" add subscriber "ClientConnections"
configure private-vlan "MedPrivate" add subscriber "Research" non-isolated
```

4. The final step is to configure VLAN translation on the local switch so that Research VLAN workstations can connect to the file servers on the remote switch:

```
configure Main add ports 1:1 private-vlan translated
```

5. To view the completed configuration, enter the show private-vlan command as follows:

```
show private-vlan
------------------------------------------------------------------------------------------
Name            VID  Protocol Addr        Flags                Proto  Ports  Virtual
Active router
/Total
------------------------------------------------------------------------------------------
MedPrivate                                                                    VR-Default
Network VLAN:
-main          100  ------------------------------------    ANY    2 /48  VR-Default
Non-Isolated Subscriber VLAN:
-Research      300  ------------------------------------    ANY    2 /96  VR-Default
Isolated Subscriber VLAN:
-ClientConnections 200  ---------------------------------    ANY    2 /52  VR-Default
```

## PVLAN Configuration Example 2

The following figure shows a PVLAN configuration example for a motel.

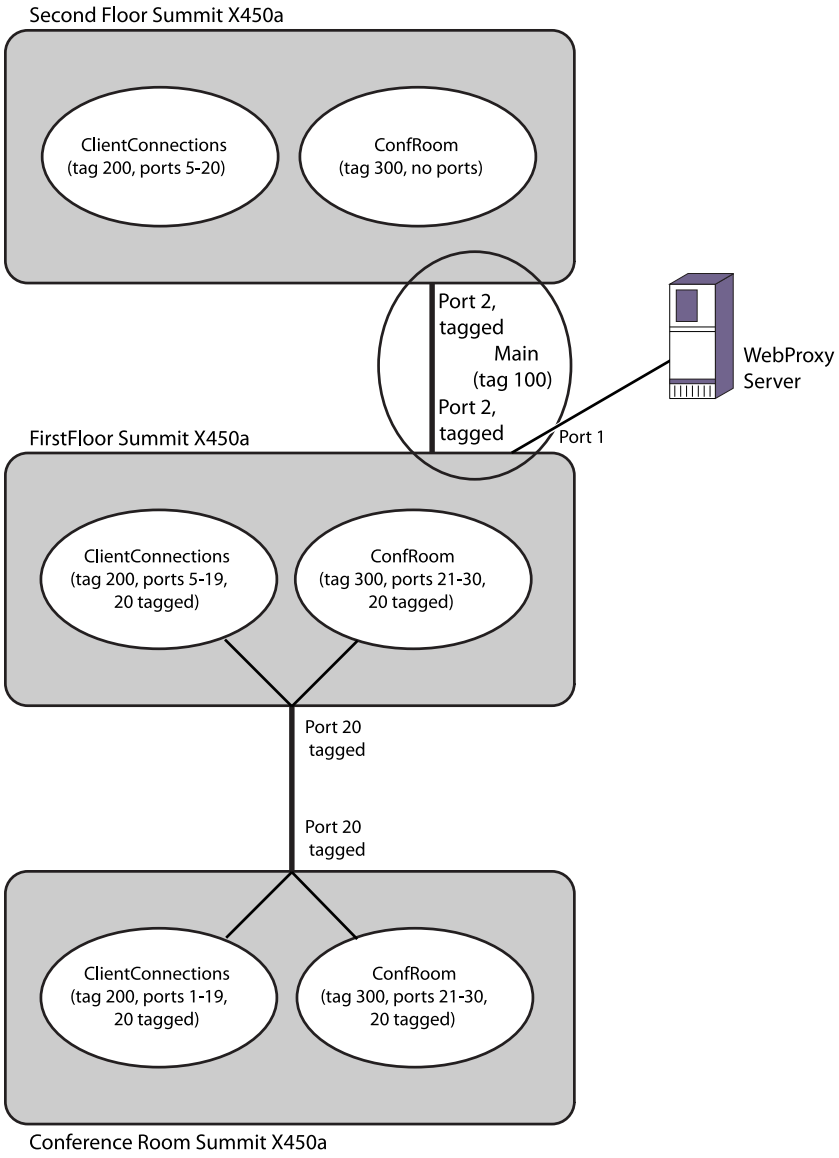



**Figure 77: PVLAN Configuration Example 2**

The motel example in the following figure has guest rooms, a conference room, and their web proxy server on the first floor, and guest rooms on the second floor. The motel has three Summit switches. There is one on the first floor in a closet, one on the first floor in the conference room, and one on the second floor.

The PVLAN in the following figure contains the following PVLAN components:

- A *VLAN* called Main that contains the web proxy server.
- A VLAN called ConfRoom that contains the ports for the conference room connections.
- A VLAN called ClientConnections that contains client PC connections for the guest rooms.

The goals for the motel network are as follows:

- Provide internet access for the ConfRoom and ClientConnections VLANs through the web proxy server.
- Prevent communications between the ConfRoom and ClientConnections VLANs.
- Enable communications between clients on the ClientConnections VLAN only within the conference room.
- Enable communications between devices on the ConfRoom VLAN.
- Prevent communications between the PCs in the ClientConnections VLAN that are not in the conference room.

Notice the following in the above figure:

- The Summit switches in the first floor closet and on the second floor contain the Main VLAN with a tag of 100. This VLAN is connected via a tagged port between the first and second floor switches.
- The Summit in the conference room does not contain the Main VLAN and cannot be a PVLAN member.
- All of the switches have the ClientConnections VLAN, and it uses VLAN tag 200.
- All of the switches have the ConfRoom VLAN, and it uses VLAN tag 300.
- The Conference Room Summit connects to the rest of the network through a tagged connection to the Summit in the first floor closet.
- Because the Summit in the first floor closet is a PVLAN member and uses the same port to support two subscriber VLANs, a loopback port is required in all subscriber VLANs, except the first configured subscriber VLAN (this applies to all switches and Summit family switches).

Note

The following examples contain comments that follow the CLI comment character (#). All text that follows this character is ignored by the switch and can be omitted from the switch configuration.

The following commands configure the Summit in the first floor closet:

```
# Create and configure the VLANs.
create vlan Main
configure vlan Main add port 1
configure vlan Main tag 100
configure vlan Main add port 2 tagged
create vlan ClientConnections
configure vlan ClientConnections tag 200
configure vlan ClientConnections add port 5-19
configure vlan ClientConnections add port 20 tagged
create vlan ConfRoom
configure vlan ConfRoom tag 300
configure vlan ConfRoom add port 21-30
configure vlan ConfRoom add port 20 tagged
# Create and configure the PVLAN named Motel.
create private-vlan Motel
configure private-vlan Motel add network Main
configure private-vlan Motel add subscriber ClientConnections # isolated subscriber VLAN
configure private-vlan "Motel" add subscriber "ConfRoom" non-isolated loopback-port 30
configure private-vlan Motel add subscriber ConfRoom non-isolated
# If you omit the loopback-port command, the above command produces the following error
message:
```

```
# Cannot add subscriber because another subscriber vlan is already present on the same
port, assign a loopback port when adding the subscriber vlan to the private vlan
# show vlan "ConfRoom"
VLAN Interface with name ConfRoom created by user
Admin State:    Enabled         Tagging:        802.1Q Tag 300
Virtual router: VR-Default
IPv6:           None
STPD:           None
Protocol:       Match all unfiltered protocols
Loopback:       Disabled
NetLogin:       Disabled
QosProfile:     None configured
Egress Rate Limit Designated Port: None configured
Private-VLAN Name:              Motel
VLAN Type in Private-VLAN:      Non-Isolated Subscriber
Ports:   13.       (Number of active ports=1)
Untag:      21,     22,     23,     24,     25,     26,     27,
28,     29
Tag:        1s,     2s,     20,   *30L
Flags:    (*) Active, (!) Disabled, (g) Load Sharing port
(b) Port blocked on the vlan, (m) Mac-Based port
(a) Egress traffic allowed for NetLogin
(u) Egress traffic unallowed for NetLogin
(t) Translate VLAN tag for Private-VLAN
(s) Private-VLAN System Port, (L) Loopback port
(x) VMAN Tag Translated port
(G) Multi-switch LAG Group port
# Note that the loopback port is flagged with an "L" and listed as a tagged port, and the
network VLAN ports are flagged with an "s" and listed as tagged ports.
```

The following commands configure the Summit on the second floor:

```
# create and configure the VLANs
create vlan Main
configure vlan Main tag 100
configure vlan Main add port 2 tagged
create vlan ClientConnections
configure vlan ClientConnections tag 200
configure vlan ClientConnections add port 5-20
create vlan ConfRoom
configure vlan ConfRoom tag 300
# Create and configure the PVLAN named Motel.
create private-vlan Motel
configure private-vlan Motel add network Main
configure private-vlan Motel add subscriber ClientConnections # isolated subscriber VLAN
configure private-vlan Motel add subscriber ConfRoom non-isolated
```

The following commands configure the Summit in the conference room:

```
# create and configure the VLANs
create vlan ClientConnections
configure vlan ClientConnections tag 200
configure vlan ClientConnections add port 1-19
configure vlan ClientConnections add port 20 tag
create vlan ConfRoom
configure vlan ConfRoom tag 300
configure vlan ConfRoom add port 21-30
configure vlan ConfRoom add port 20 tag
# The VLANs operate as extensions of the VLANs on the Summit in the first floor closet.
There is no PVLAN configuration on this switch.
```

# VLAN Translation

The *VLAN* translation feature described in this section provides the same VLAN translation functionality that is provided for PVLANs. This is described in VLAN Translation in a PVLAN on page 516.

The difference is that this feature is configured with different commands that are compatible with ExtremeWare.

**Note**

The VLAN translation feature described in this section is provided for those who are already familiar with the ExtremeWare VLAN translation commands. If you have not used this feature in ExtremeWare and do not use any scripts that use the ExtremeWare commands, we suggest that you use the Private VLAN feature described in Private VLANs on page 516, as it provides the same functionality with additional features.

The VLAN translation feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

The following figure shows how VLAN translation is configured in the switch.

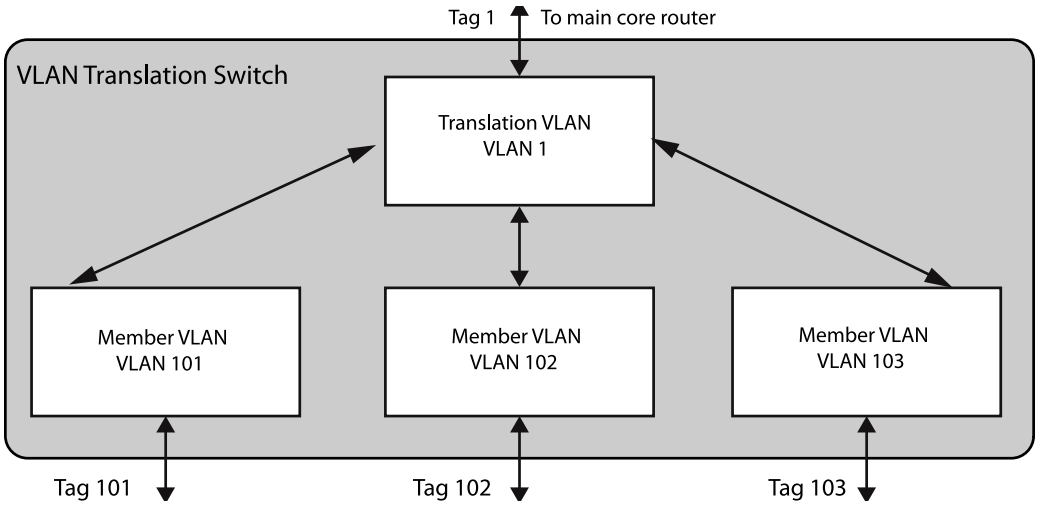



**Figure 78: VLAN Translation Switch Configuration**

In the above figure, VLAN1 is configured as a translation VLAN . The translation VLAN is equivalent to the network VLAN in the PVLAN implementation of VLAN translation.

VLANs 101, 102, and 103 are configured as member VLANs of translation VLAN1. The member VLANs are equivalent to the non-isolated subscriber VLANs in the PVLAN implementation of VLAN translation.

This configuration enables tag translation between the translation VLAN and the member VLANs. All member VLANs can communicate through the translation VLAN, but they cannot communicate through Layer 2 with each other.

## VLAN Translation Behavior

You should be aware of the behavior of unicast, broadcast, and multicast traffic when using *VLAN* translation.

*Unicast Traffic*

Traffic on the member *VLAN*s can be either tagged or untagged.

Traffic is switched locally between client devices on the same member VLAN as normal. Traffic cannot be switched between clients on separate member VLANs. Traffic from any member VLAN destined for the translation VLAN is switched and the VLAN tag is translated appropriately. Traffic from the translation VLAN destined for any member VLAN is switched and the VLAN tag is translated.

*Broadcast Behavior*

Broadcast traffic generated on a member *VLAN* is replicated in every other active port of that VLAN as normal.

In addition, the member VLAN traffic is replicated to every active port in the translation VLAN and the VLAN tag is translated appropriately. Broadcast traffic generated on the translation VLAN is replicated to every other active port in this VLAN as usual. The caveat in this scenario is that this traffic is also replicated to every active port in every member VLAN, with VLAN tag translation. In effect, the broadcast traffic from the translation VLAN leaks onto all member VLANs.

*Multicast Behavior*

*IGMP* snooping can be enabled on member and translation *VLAN*s so that multicast traffic can be monitored within the network.

IGMP snooping software examines all IGMP control traffic that enters the switch. IGMP control traffic received on a VLAN translation port is forwarded by the CPU to all other ports in the translation group. Software VLAN translation is performed on the packets which cross the translation boundary between member and translation VLANs. The snooping software detects ports joining and leaving multicast streams. When a VLAN translation port joins a multicast group, an *FDB* entry is installed only on receiving a data miss for that group. The FDB entry is added for the requested multicast address and contains a multicast PTAG. When a VLAN translation port leaves a multicast group, the port is removed from the multicast list. The last VLAN translation port to leave a multicast group causes the multicast FDB entry to be removed.

## VLAN Translation Limitations

The *VLAN* translation feature has the following limitations:

- Requires more *FDB* entries than a standard VLAN.
- Within the same VR, VLAN tag duplication is not allowed.
- Within the same VR, VLAN name duplication is not allowed.
- Each MAC address learned in the translation and member VLANs must be unique. A MAC address cannot exist in two or more VLANs that belong to the same VLAN translation domain.
- MVR cannot be configured on translation and member VLANs.
- A VMAN cannot be added to translation and member VLANs.
- A PBB network (BVLAN) cannot be added to translation and member VLANs.
- EAPS control VLANs cannot be either translation or member VLANs.
- EAPS can only be configured on translation VLAN ports (and not on member VLAN ports). To support EAPS on the network VLAN, you must add all of the translation and member VLANs to the EAPS ring.

- *STP* can only be configured on translation VLAN ports (and not on member VLAN ports). To support STP on the translation VLAN, you must add the translation VLAN and all of the member VLANs to STP.
- *ESRP* can only be configured on translation VLAN ports (and not on member VLAN ports). To support ESRP on the network VLAN, you must add the translation VLAN and all of the member VLANs to ESRP.
- There is no *NetLogin* support to add ports as translate to the translation VLAN, but the rest of NetLogin and the PVLAN feature do not conflict.
- *IGMP* snooping is performed across the entire VLAN translation domain, spanning all the member VLANs. For VLANs that are not part of a VLAN translation domain, IGMP snooping operates as normal.
- VLAN translation and VPLS are not supported on the same VLAN.
- Member VLANs in a VLAN translation domain cannot exchange multicast data with VLANs outside the VLAN translation domain. However, the translation VLAN can exchange multicast data with VLANs outside the VLAN translation domain and with translation VLANs in other VLAN translation domains.

*Interfaces*

Use the following information for selecting and configuring *VLAN* translation interfaces:

- A single physical port can be added to multiple member VLANs, using different VLAN tags.
- Member VLANs and translation VLANs can include both tagged and untagged ports.

## Configuring Translation VLANs

To create a translation *VLAN*, do the following:

1. Create the VLAN that will become the translation VLAN.
2. Add a tag and ports to the prospective translation VLAN.
3. Add member VLANs to the prospective translation VLAN.

A prospective translation VLAN becomes a translation VLAN when the first member VLAN is added to it.

- To add a member VLAN to a translation VLAN, use the following command:

  `configure {vlan} vlan_name vlan-translation add member-vlan member_vlan_name {loopback-port port}`
- To delete a member VLAN from a translation VLAN, use the following command:

  `configure {vlan} vlan_name vlan-translation delete member-vlan [member_vlan_name | all]`
- To view the translation VLAN participation status of a VLAN, use the following command:

  `show vlan {virtual-router vr-name}`

## Displaying Translation VLAN Information

This section describes how to display translation *VLAN* information.

*Displaying Information for a Translation or Member VLAN*

To display information about a translation or member *VLAN*, use the command:

```
show vlan {virtual-router vr-name}
```

*Displaying Translation VLAN FDB Entries*

To view all *FDB* entries including those created for a translation *VLAN*, use the command:

```
show fdb {blackhole {netlogin [all | mac-based-vlans]} | netlogin [all |
mac-based-vlans] | permanent {netlogin [all | mac-based-vlans]} |
mac_addr {netlogin [all | mac-based-vlans]} | ports port_list {netlogin
[all | mac-based-vlans]} | vlan vlan_name | vlan_list {netlogin [all |
mac-based-vlans]} | {{vpls} {vpls_name}}}
```

The T flag marks additional FDB entries for translation VLANs.

## VLAN Translation Configuration Examples

The following configuration examples show *VLAN* translation used in three scenarios:

- Basic VLAN Translation on page 537
- VLAN Translation with ESRP Redundancy on page 538
- VLAN Translation with STP Redundancy on page 540

*Basic VLAN Translation*

The example in the following figure configures a basic *VLAN* translation network. This network provides VLAN translation between four member VLANs and a single translation VLAN.

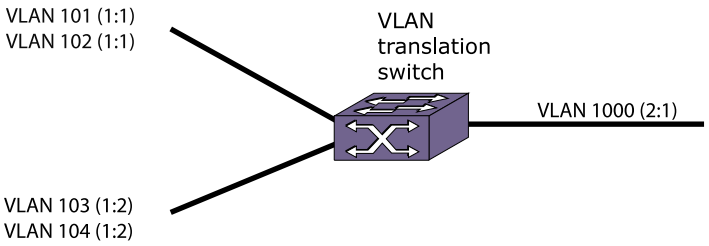



**Figure 79: VLAN Translation Configuration Example**

The following configuration commands create the member VLANs:

```
create vlan v101
configure v101 tag 101
configure v101 add ports 1:1 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:1 tagged
create vlan v103
configure v103 tag 103
configure v103 add ports 1:2 tagged
create vlan v104
configure v104 tag 104
configure v104 add ports 1:2 tagged
```

The following configuration commands create the translation VLAN and enable VLAN translation:

```
create vlan v1000
configure v1000 tag 1000
configure v1000 add ports 2:1 tagged
configure v1000 vlan-translation add member-vlan v101
configure v1000 vlan-translation add member-vlan v102
configure v1000 vlan-translation add member-vlan v103
configure v1000 vlan-translation add member-vlan v104
```

The following configuration commands create the translation VLAN and enable VLAN translation on BlackDiamond X8, BlackDiamond 8000 series modules, and Summit X440, X460, X480, X670, X670-G2, and X770 series switches:

```
create vlan v1000
configure v1000 tag 1000
configure v1000 add ports 2:1 tagged
configure v1000 vlan-translation add member-vlan v101
configure v1000 vlan-translation add member-vlan v102 loopback-port 1:23
configure v1000 vlan-translation add member-vlan v103
configure v1000 vlan-translation add member-vlan v104 loopback-port 1:24
```

### *VLAN Translation with ESRP Redundancy*

The example in the following figure configures a *VLAN* translation network with *ESRP* redundancy.

The SW2 and SW3 VLAN translation switches are protected by an ESRP control VLAN. The master ESRP switch performs the translation and provides the connectivity to the backbone. If a failure occurs, the slave ESRP switch takes over and begins performing the translation.

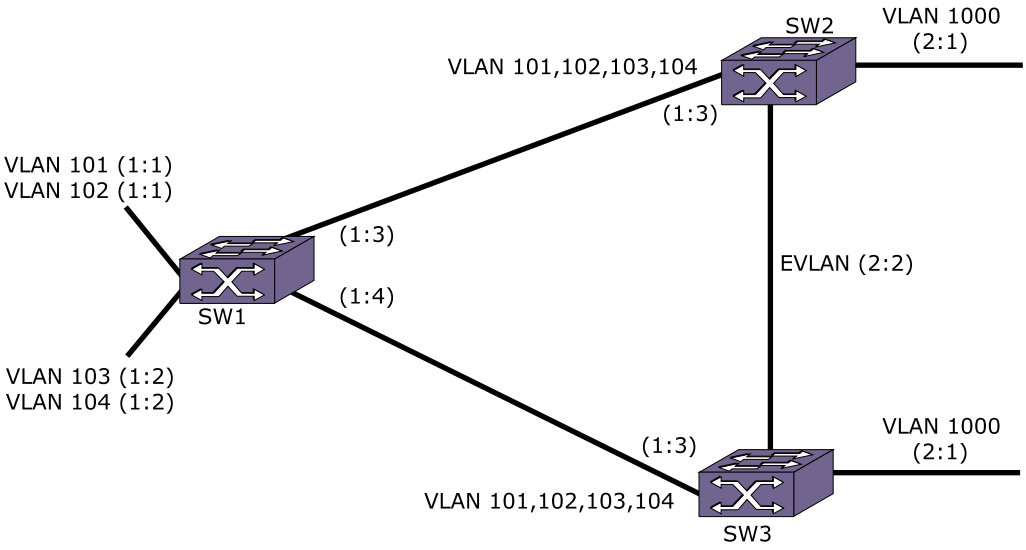



**Figure 80: ESRP Redundancy Configuration Example**

The following configuration commands create the member VLANs on SW1:

```
create vlan v101
configure v101 tag 101
```

```
configure v101 add ports 1:1 tagged
configure v101 add ports 1:3 tagged
configure v101 add ports 1:4 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:1 tagged
configure v102 add ports 1:3 tagged
configure v102 add ports 1:4 tagged
create vlan v103
configure v103 tag 103
configure v103 add ports 1:2 tagged
configure v103 add ports 1:3 tagged
configure v103 add ports 1:4 tagged
create vlan v104
configure v104 tag 104
configure v104 add ports 1:2 tagged
configure v104 add ports 1:3 tagged
configure v104 add ports 1:4 tagged
```

The configuration for SW2 and SW3 is identical for this example.

The following configuration commands create the member VLANs on SW2:

```
create vlan v101
configure v101 tag 101
configure v101 add ports 1:3 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:3 tagged
create vlan v103
configure v103 tag 103
configure v103 add ports 1:3 tagged
create vlan v104
configure v104 tag 104
configure v104 add ports 1:3 tagged
```

This set of configuration commands creates the translation VLANs and enables VLAN translation on SW2:

```
create vlan v1000
configure v1000 tag 1000
configure v1000 add ports 2:1 tagged
configure v1000 vlan-translation add member-vlan v101
configure v1000 vlan-translation add member-vlan v102
configure v1000 vlan-translation add member-vlan v103
configure v1000 vlan-translation add member-vlan v104
```

The final set of configuration commands creates the ESRP control VLAN and enables ESRP protection on the translation VLAN for SW2:

```
create vlan evlan
configure evlan add ports 2:2
enable esrp evlan
configure evlan add domain-member v1000
```

The following configuration commands create the translation VLAN and enable VLAN translation onVLANs that have overlapping ports:

```
configure v1000 vlan-translation add member-vlan v102 loopback-port 1:22
```

```
configure v1000 vlan-translation add member-vlan v103 loopback-port 1:23
configure v1000 vlan-translation add member-vlan v104 loopback-port 1:24
```

### *VLAN Translation with STP Redundancy*

The example in the following figure configures a *VLAN* translation network with redundant paths protected by *STP*.

Parallel paths exist from the member VLAN portion of the network to the translation switch. STP ensures that the main path for this traffic is active and the secondary path is blocked. If a failure occurs in the main path, the secondary paths are enabled.

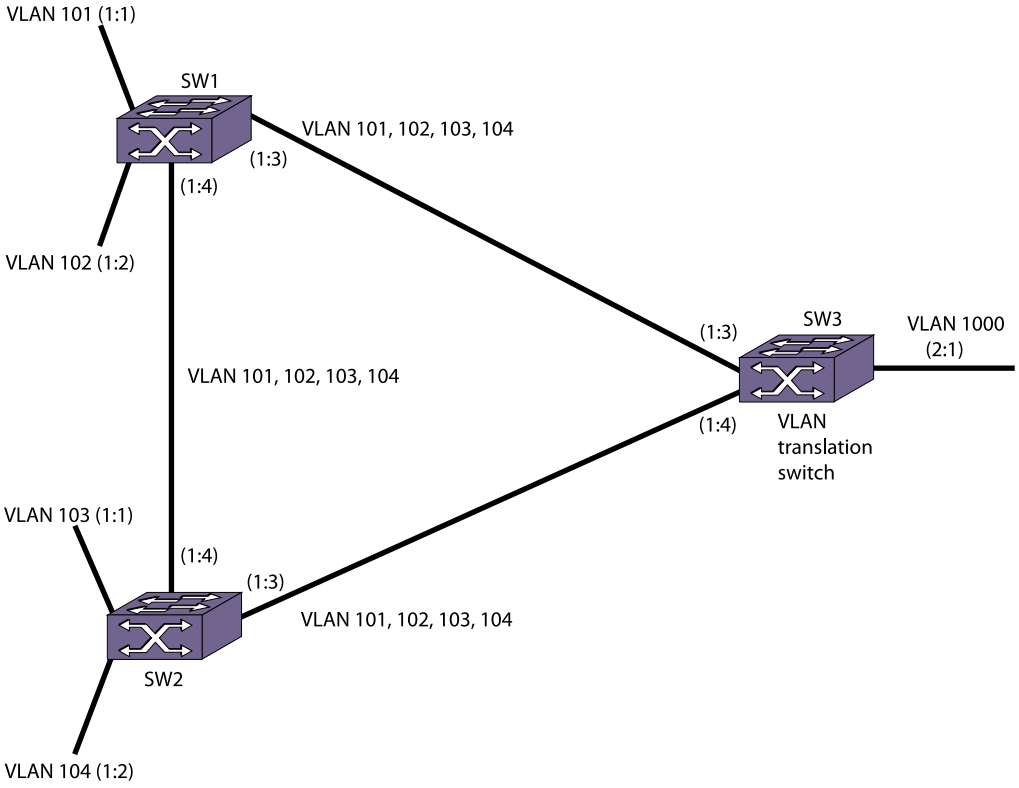



**Figure 81: STP Redundancy Configuration Example**

The following configuration commands create the member VLANs and enable STP on SW1:

```
create vlan v101
configure v101 tag 101
configure v101 add ports 1:1 tagged
configure v101 add ports 1:3 tagged
configure v101 add ports 1:4 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:2 tagged
configure v102 add ports 1:3 tagged
configure v102 add ports 1:4 tagged
create vlan v103
configure v103 tag 103
configure v103 add ports 1:3 tagged
configure v103 add ports 1:4 tagged
create vlan v104
configure v104 tag 104
```

```
configure v104 add ports 1:3 tagged
configure v104 add ports 1:4 tagged
create stpd stp1
configure stp1 tag 101
configure stp1 add vlan v101
configure stp1 add vlan v102
configure stp1 add vlan v103
configure stp1 add vlan v104
enable stpd stp1
```

These configuration commands create the member VLANs and enable STP on SW2:

```
create vlan v103
configure v103 tag 103
configure v103 add ports 1:1 tagged
configure v103 add ports 1:3 tagged
configure v103 add ports 1:4 tagged
create vlan v104
configure v104 tag 104
configure v104 add ports 1:2 tagged
configure v104 add ports 1:3 tagged
configure v104 add ports 1:4 tagged
create vlan v101
configure v101 tag 101
configure v101 add ports 1:3 tagged
configure v101 add ports 1:4 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:3 tagged
configure v102 add ports 1:4 tagged
create stpd stp1
configure stp1 tag 101
configure stp1 add vlan v101
configure stp1 add vlan v102
configure stp1 add vlan v103
configure stp1 add vlan v104
enable stpd stp1
```

This set of configuration commands creates the member VLANs and enables STP on SW3:

```
create vlan v101
configure v101 tag 101
configure v101 add ports 1:3 tagged
configure v101 add ports 1:4 tagged
create vlan v102
configure v102 tag 102
configure v102 add ports 1:3 tagged
configure v102 add ports 1:4 tagged
create vlan v103
configure v103 tag 103
configure v103 add ports 1:3 tagged
configure v103 add ports 1:4 tagged
create vlan v104
configure v104 tag 104
configure v104 add ports 1:3 tagged
configure v104 add ports 1:4 tagged
create stpd stp1
configure stp1 tag 101
configure stp1 add vlan v101
configure stp1 add vlan v102
configure stp1 add vlan v103
```

```
configure stp1 add vlan v104
enable stpd stp1
```

The final set of configuration commands creates the translation VLAN and enables VLAN translation on SW3:

```
create vlan v1000
configure v1000 tag 1000
configure v1000 add ports 2:1 tagged
configure v1000 vlan-translation add member-vlan v101
configure v1000 vlan-translation add member-vlan v102
configure v1000 vlan-translation add member-vlan v103
configure v1000 vlan-translation add member-vlan v104
```

The following configuration commands create the translation VLAN and enable VLAN translation on VLANs that have overlapping ports:

```
configure v1000 vlan-translation add member-vlan v102 loopback-port 1:22
configure v1000 vlan-translation add member-vlan v103 loopback-port 1:23
configure v1000 vlan-translation add member-vlan v104 loopback-port 1:24
```

## Port-Specific VLAN Tag

The Port-specific *VLAN* feature adds a layer of specificity between the port tag and the VLAN/VMAN tag: a port-specific VLAN tag. This feature adds the following functionality to the existing VLAN:

- Ability to associate a tag to a VLAN port. This tag is used as a filter to accept frames with matching VID. It is also used as the tag of the outgoing frames.
- Ability to add multiple VLAN ports on the same physical port as long as those VLAN ports are associated with different tags.
- Allows the existing untagged and tagged VLAN ports to be part of the VLAN.
- Ability to learn MAC address on port, tag and VLAN instead of only on the port.As a consequence of the previous point, ability to add static MAC address to port, tag and VLAN.
- Ability to specify limit-learning and MAC lockdown on a port, tag and VLAN, instead of only on the port.
- Rate limiting and counting of frames with matching VIDs is supported with the existing *ACL (Access Control List)*.

The Port-specific VLAN tag allows tagged VLAN ports to be configured with tag values. When the tag is not configured, it is implicit that the tag of the tagged port is the tag of the VLAN. We call the tag of the port the "port tag", and the tag of the VLAN the "base tag". The port tag is used to determine the eligibility of the frames allowed to be part of the VLAN. Once the frame is admitted to the VLAN port, the base tag is used. From a functional standpoint, the frame tag is rewritten to the base tag.

The base tag then is translated to the port tag for the outgoing frame.

Note

The port tag is equal to the base tag when the port tag is not specified, so the current VLAN behavior is preserved.

Untagged VLAN ports also have port tag, which is always the same as the base tag. Outgoing frames are untagged. The untagged VLAN port always has an implicit port tag thats's always equal to the base

tag. There can be only one untagged VLAN port on a physical port. It receives untagged frames, and tagged frames, and transmits only untagged frames.

A tagged VLAN port can have a port tag configured, or not. When not configured, the port tag is equal to the base tag. There can be more than one tagged VLAN port on a physical port. It receives tagged frames with tag equals to the port tag, and transmits tagged frames with port tag.

When the VLAN is assigned to L2VPN, the base tag is the tag that is carried by the pseudo-wire when the dot1q include is enabled. It can be viewed that VPLS PW port tag is equal to the base tag. To assign a VLAN with a port-specific tag to an L2VPN, use the existing `configure vpls vpls_name add service` **`vlan`** *`vlan_name`* command.

Since every tagged VLAN port has different VIDs, forwarding between them on the same physical port (hairpin switching) is possible. From the external traffic point of view, the frame tags are rewritten from the receive port tag to the transmit port tag. Since each port tag is a different VLAN port, a frame that has to be broadcasted to multiple VLAN ports is sent out multiple times with different tags when the VLAN ports are on the same physical port. Each port + port tag is an individual VLAN port.

MAC addresses are learned on the VLAN port. This means that the port in the *FDB* entry is the port + port tag. A unicast frame destined to a MAC address that is in the FDB is sent out of the associated VLAN port. As mentioned earlier, there is only one MAC addressed learned on the VLAN. If the MAC address is learned on a different port or a different tag, it is a MAC move. It is transmitted out of the physical port only on the associated VLAN port tagged with the port tag when the VLAN port is tagged.

When there are multiple tagged VLAN ports on the transmit port, only one frame with the right tag is transmitted. It is transmitted untagged on an untagged VLAN port. Accordingly, the static MAC address is configured on a VLAN port. This means that the port tag is specified when the tag is not equal to the base tag. The command to flush FDB does not need to change. But, a VLAN port-specific flush needs to be implemented to handle the case when a VLAN port is deleted. This flush is internal and not available through the CLI.

Per VLAN port (port + tag) rate limiting and accounting is achieve by the existing ACL. Use match condition `vlan-id` to match the port VID. You can use action `count` and `byte-count` for accounting. And you can use `show access-list counter` to view the counters. Action meter can be used for rate limiting. To create a meter, use the `create meter` command, and configure the committed rate and maximum burst size.

## Port-Specific Tags in L2VPN

You can assign a *VLAN* with port specific tag to VPLS/VPWS using the `configure vpls vsi add service vlan vl` command. Because this is a single VLAN, the base VID is used when dot1q include is enable. For example, when VLAN 100 that has ports on Ethernet port 1 with port tag 10 and 11 is assigned to L2VPN, the tag that is carried by the pseudo wire is 100. The configuration for this example is as follows:

```
create vlan exchange tag 100
config vlan exchange add ports 1 tagged 10
config vlan exchange add ports 1 tagged 11
config vpls vsi1 add service vlan exchange
```

Similarly, the following is an example for VPWS. There can only be a single VLAN port in the VLAN for assignment to VPWS to be successful:

```
create vlan exchange tag 100
config vlan exchange add ports 1 tagged 10
config l2vpn vpws pw1 add service vlan exchange
```

*VLAN Port State*

VLAN port state is the same as the state of the Ethernet port.

*ACLs*

You can use the existing match vlan-id *ACL* to accomplish counting and metering. You can assign the ACL to both ingress and egress port. The followings are the examples of such configuration. The port 3 tag is 4 and the port 4 tag is 5. These ACLs will match the frame vlan-ID, and the vlan-ID specified in the match criteria is independent of the port tag.

```
Content of acl.pol
entry tag_1 {
   if {
     vlan-id 4;
     } then {
       packet-count tag_1_num_frames;
       meter tag_1_meter;
     }
}
entry tag_2 {
   if {
     vlan-id 5;
     } then {
       byte-count tag_2_num_bytes;
       meter tag_2_meter;
     }
}
Content of acl_egress.pol
entry tag_1_egr {
   if {
     vlan-id 4;
     } then {
       packet-count tag_1_egr_num_frames;
       meter tag_1_egr_meter;
     }
}
entry tag_2_egr {
   if {
     vlan-id 5;
     } then {
       byte-count tag_2_egr_num_bytes;
       meter tag_2_egr_meter;
     }
}
```

## Configuring Port-Specific VLAN Tags

The following specific commands are modified by the port-specific *VLAN* tag:

- `clear fdb`: Only clears on physical port or VLAN, not on a vlan port.
- `delete fdb`: All or specific MAC address, or specific MAC address on a VLAN.
- `enable/disable flooding ports`: Only on physical port (applies to all VLAN ports).

- `enable/disable learning`: Only on physical port (applies to all VLAN ports on the same physical port), or on a VLAN (applies to all VLAN ports of the VLAN).
- `show fdb stats`: Only on physical port or VLAN, not on a VLAN port.

Use the following commands to configure Port-specific VLAN tags:

- To configure the port-specific tag, use the `configure ports` *port_list* {**tagged** *tag*} **vlan** *vlan_name* [**limit-learning** *number* {**action** [**blackhole** | **stop-learning**]} |**lock-learning** | **unlimited-learning** | **unlocklearning**] command.
- To specify the port tag when you need to put multiple vlans into a broadcast domain, use the `configure` {**vlan**} *vlan_name* `addports` [**port_list** | **all**] {**tagged**{*tag*} | **untagged**} {{**stpd**} *stpd_name*} {**dot1d** | **emistp** | **pvst-plus**}} command.
- To specify a port tag to delete a VLAN port that has a different tag from the VLAN tag, use the `configure` {**vlan**} *vlan_name* `deleteports` [**all** | **port_list** {**tagged** *tag*}] command.
- To display output of a vlan that has a port-specific tag, use the `show vlan` command.
- To display port info that has port-specific tag statistics, use the `show port` info detail command.
- To adds a permanent, static entry to the *FDB*, use the `create fdb` **mac_addr vlan** *vlan_name* [**ports** *port_list* {**tagged** *tag*} | **blackhole**] command.
- To show output where the port tag is displayed, use the `show fdb` command.

# VMAN (PBN)



The virtual metropolitan area network (VMAN) feature allows you to scale a Layer 2 network and avoid some of the management and bandwidth overhead required by Layer 3 networks.

Note
If a failover from MSM A to MSM B occurs, VMAN operation is not interrupted. The system has hitless failover—network traffic is not interrupted during a failover.

## VMAN Overview

The *Virtual MAN (VMAN)* feature is defined by the IEEE 802.1ad standard, which is an amendment to the IEEE 802.1Q *VLAN (Virtual LAN)* standard.

A VMAN is a virtual Metropolitan Area Network (MAN) that operates over a physical MAN or Provider Bridged Network (PBN). This feature allows a service provider to create VMAN instances within a MAN or PBN to support individual customers. Each VMAN supports tagged and untagged VLAN traffic for a customer, and this traffic is kept private from other customers that use VMANs on the same PBN.

The PBN uses Provider Bridges (PBs) to create a Layer 2 network that supports VMAN traffic. The VMAN technology is sometimes referred to as VLAN stacking or Q-in-Q.

Note
VMAN is an Extreme Networks term that became familiar to Extreme Networks customers before the 802.1ad standard was complete. The VMAN term is used in the ExtremeXOS software and also in this book to support customers who are familiar with this term. The PBN term is also used in this guide to establish the relationship between this industry standard technology and the Extreme Networks VMAN feature.

The following figure shows a VMAN, which spans the switches in a PBN.

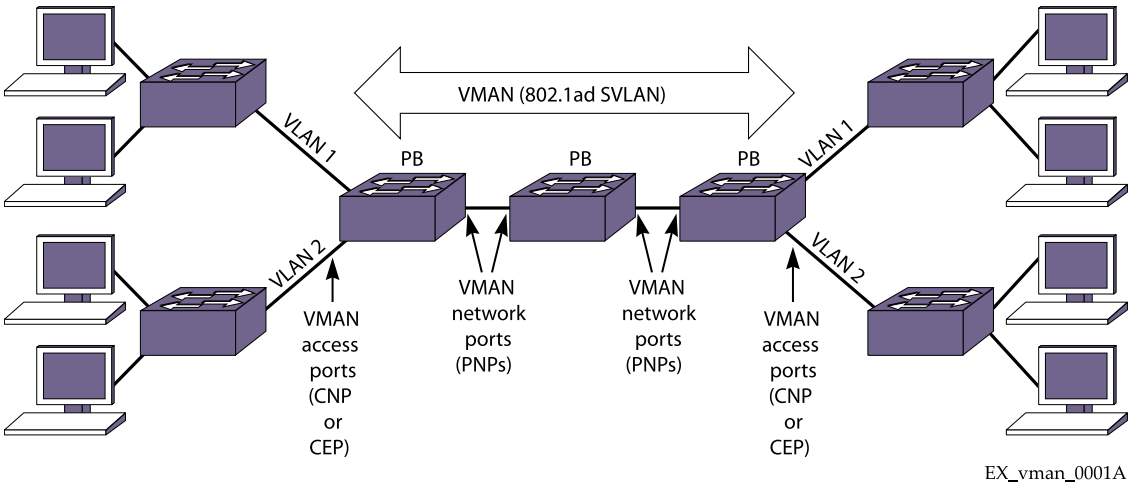



**Figure 82: VMAN**

The entry points to the VMAN are the access ports on the VMAN edge switches. Customer VLAN (CVLAN) traffic that is addressed to locations at other VMAN access ports enters the ingress access port, is switched through the VMAN, and exits the egress access port. If you do not configure any frame manipulation options, the CVLAN frames that exit the VMAN are identical to the frames that entered the VMAN.

VMAN access ports operate in the following roles:

- Customer Network Port (CNP)
- Customer Edge Port (CEP, which is also known as Selective Q-in-Q)

The CEP role, which is configured in software as a cep vman port, connects a VMAN to specific CVLANs based on the CVLAN CVID. The CNP role, which is configured as an untagged vman port, connects a VMAN to all other port traffic that is not already mapped to the port CEP role. These roles are described later.

All other VMAN ports (except the access ports) operate as VMAN network ports, which are also known as Provider Network Ports (PNPs) in the 802.1ad standard. The VMAN network ports connect the PBs that form the core of the VMAN. During configuration, the VMAN network ports are configured as tagged VMAN ports.

The following figure shows one VMAN, but a PBN can support multiple VMAN instances, which are sometimes called VMANs or Service VLANs (SVLANs). VMANs allow you to partition the PBN for customers in the same way that VLANs allow you to partition a Layer 2 network. For example, you can use different VMANs to support different customers on the PBN, and the PBN delivers customer traffic only to the PBN ports that are configured for appropriate VMAN.

A VMAN supports two tags in each Ethernet frame, instead of the single tag supported by a VLAN Ethernet frame. The inner tag is referred to as the customer tag (C-tag), and this optional tag is based on the CVLAN tag if the source VLAN is a tagged VLAN. The outer tag is referred to as the service tag (S-tag) or VMAN tag or SVLAN tag, and it is the tag that defines to which SVLAN a frame belongs. The following figure shows the frame manipulation that occurs at the VMAN edge switch.

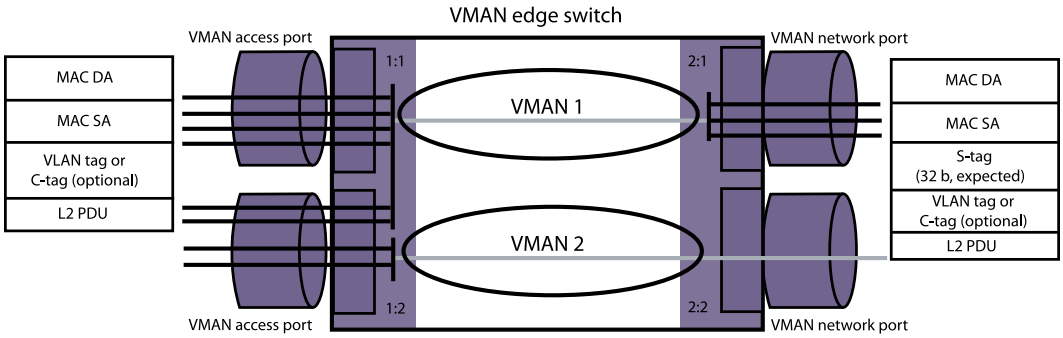



**Figure 83: Tag Usage at the VMAN Access Switch**

In the above figure, the switch accepts CVLAN frames on VMAN access ports 1:1 and 1:2. The switch then adds the S-tag to the frames and switches the frames to network ports 2:1 and 2:2. When the 802.1ad frames reach the PB egress port, the egress switch removes the S-tag, and the CVLAN traffic exits the egress access port in its original form.

When the switch in the figure above acts as the egress switch for a VMAN, VMAN frames arrive on network ports 2:1 and 2:2. The switch accepts only those frames with the correct S-tag, removes the S-tags, and switches those frames to access ports 1:1 and 1:2. Unless special configuration options are applied, the egress frames are identical to ingress CVLAN frames. (Configuration options are described in VMAN Configuration Options and Features on page 551.)

The following figure shows that the S-tags and C-tags used in VMAN frames contain more than just customer and service VLAN IDs.

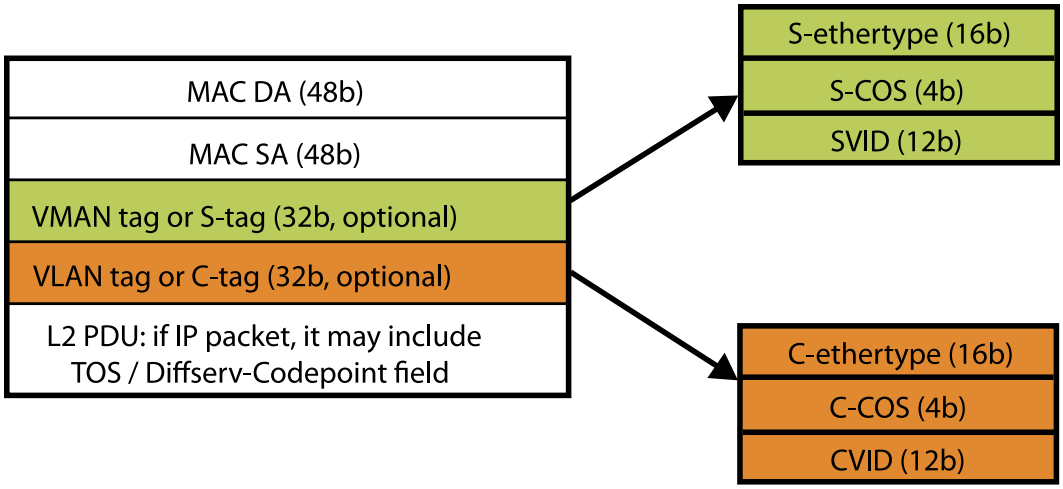



**Figure 84: S-tag and C-tag Components**

Each S-tag and C-tag contains an ethertype, a *CoS (Class of Service)*, and a SVLAN ID (SVID) or CVLAN ID (CVID). The ethertype is described in Secondary Ethertype Support on page 551, and the CoS is described in QoS Support on page 552.

The SVID is the VLAN tag you assign to a VMAN when you create it (see the `configure vman vman_name tag tag` command. The CVID represents the CVLAN tag for tagged VLAN traffic.

Switch ports support VMAN roles and features, which are described in the following sections:

- Customer Network Ports
- Customer Edge Ports
- CVID Translation
- CVID Egress Filtering

## Customer Network Ports

Customer Network Ports (CNPs) are edge switch ports that accept all tagged and untagged CVLAN traffic and route it over a single *VMAN*. A CNP is simpler to configure than a CEP, because it supports one VMAN on a physical port and requires no configuration of CVIDs. The VMAN service provider does not need to know anything about the CVLAN traffic in the VMAN. The service provider simply manages the VMAN, and the ingress CVLAN traffic is managed by the customer or another service provider. This separation of CVLAN and VMAN management reduces the dependence of the separate management teams on each other, and allows both management teams to make changes independent of the other.

Note

The CNP term is defined in the IEEE 802.1ad standard and is also called a port-based service interface. The CNP operation is similar to a MEF 13 UNI Type 1.2, and in releases before ExtremeXOS 12.6, CNPs were known as VMAN access ports or untagged vman ports. With the addition of CEPs, the term VMAN access port is now a generic term that refers to CNPs and CEPs.

A PBN can support up to 4094 VMANs, and each VMAN can support up to 4094 CVLANs. Because each CNP connects to only one VMAN, the maximum number of customer VMANs on an edge switch is equal to the total number of switch ports minus one, because at least one port is required to serve as the PNP (Provider Network Port).

## Customer Edge Ports

Each CEP supports the configuration of connections or mappings between individual CVLANs and multiple VMANs.

This provides the following benefits:

- Each physical port supports multiple customers (each connecting to a separate VMAN).
- Each switch supports many more customer VMANs using CEPs instead of CNPs.

To define the connections between CVLANs and SVLANs, each CEP uses a dedicated CVID map, which defines the supported CVIDs on the CEP and the destination VMAN for each CVID. For example, you can configure a CEP to forward traffic from five specific CVLANs to VMAN A and from ten other specific CVLANs to VMAN B. During VMAN configuration, certain ports are added to the VMAN as CEPs, and certain CVIDs on those ports are mapped to the VMAN. To enable customer use of a VMAN, service providers must communicate the enabled CVIDs to their customers. The customers must use those CVIDs to access the VMAN.

Note

The *CEP (Customer Edge Port)* term is defined in the IEEE 802.1ad standard and is also called a *C-tagged service interface*. The CEP operation is similar to a MEF 13 UNI Type 1.1.

## CVID Translation

To support CVLANs that are identified by different CVIDs on different CEPs, some switches support a feature called CVID translation, which translates the CVID received at the VMAN ingress to a different CVID for egress from the VMAN (see the following figure).

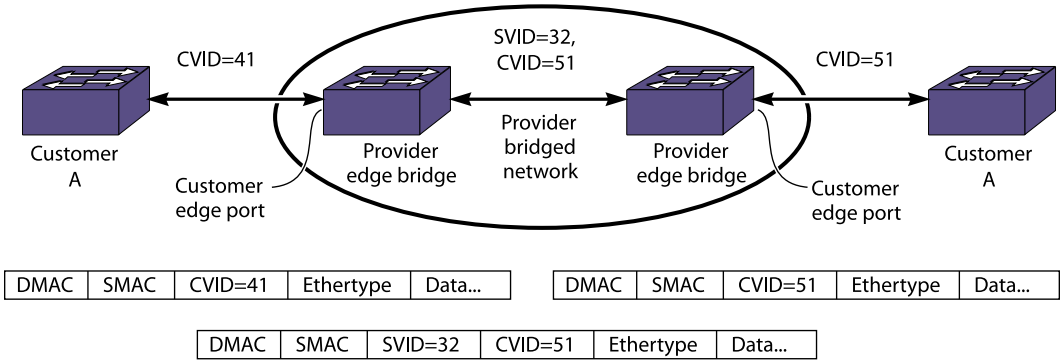



**Figure 85: CVID Translation**

You can use a CLI command to configure CVID translation for a single CVID or for a range of CVIDs, and you can enter multiple commands to define multiple CVIDs and ranges for translation. The commands can be applied to a single port or a list of ports, and after configuration, the configuration applied to a port is retained by that port.

Note

CVID translation is available only on the platforms listed for this feature in the *Feature License Requirements* document.

CVID translation can reduce the number of CVIDs that can be mapped to VMANs.

## CVID Egress Filtering

CVID egress filtering permits the egress from VMAN to CEP of only those frames that contain a CVID that has been mapped to the source VMAN; all other frames are blocked. For example, Customer Edge Port A in the following figure is configured to support CVIDs 10-29, and Customer Edge Port B is configured to support CVIDs 10-19. If CVID egress filtering is enabled on Customer Edge Port B, frames with CVIDs 20-29 will not be forwarded at the egress of Customer Edge Port B.

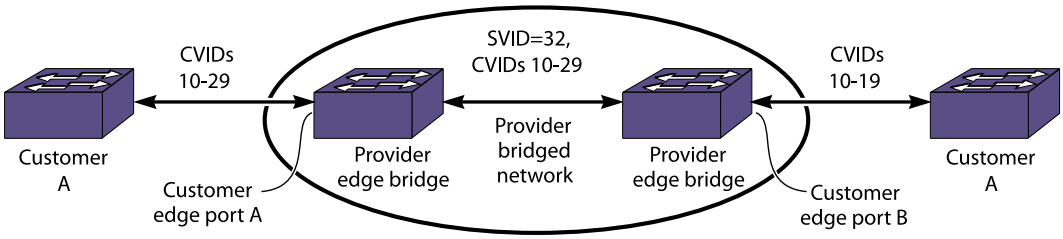



**Figure 86: CVID Egress Filtering**

You can enable CVID egress filtering for a single CEP or for all CEPs with a CLI command. You can also repeat the command to enable this feature on multiple CEPs.

Note
CVID egress filtering is available only on the platforms listed for this feature in the *Feature License Requirements* document.

When this feature is enabled, it reduces the maximum number of CVIDs that can be mapped to VMANs. The control of CVID egress filtering applies to fast-path forwarding. When frames are forwarded through software, CVID egress filtering is always enabled.

## VMAN Configuration Options and Features

As of ExtremeXOS 16.1, you can now referr to a VMAN by VID as an alternative. Specifying lists of VIDs is a is a useful shortcut to all users that can reduce the number of commands required to configure the switch.

Note
Commands enhanced with VID list support operate in a “best effort” fashion. If one of the VIDs in a VID list do not exist the command is still executed for all of the VIDs in the list that do exist. No error or warning is displayed for the invalid VIDs unless all of the specified VIDs are in valid.

### ACL Support

The ExtremeXOS software includes VMAN (PBN) *ACL (Access Control List)* support for controlling VMAN frames.

VMAN ACLs define a set of match conditions and modifiers that can be applied to VMAN frames. These conditions allow specific traffic flows to be identified, and the modifiers allow a translation to be performed on the frames.

### Secondary Ethertype Support

The C-tag and S-tag components that are added to all VMAN (PBN) frames include C-ethertype and S-ethertype components that specify an ethertype value for the customer *VLAN* and VMAN, respectively.

Note
This feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

The C-tag and S-tag components that are added to all VMAN (PBN) frames include C-ethertype and S-ethertype components that specify an ethertype value for the customer VLAN and VMAN, respectively. The I-tag used in PBBN frames also includes an ethertype value. When a VLAN or VMAN frame passes between two switches, the ethertype is checked for a match. If the ethertype does not match that of the receiving switch, the frame is discarded.

The default ethertype values are:

- VLAN port (802.1q frames): 0x8100

- Primary VMAN port (802.1ad frames): 0x88A8
- Secondary VMAN port (802.1ad frames): Not configured

The secondary ethertype support feature applies only to VMANs. The ethertype value for VLAN frames is standard and cannot be changed.

If your VMAN transits a third-party device (in other words, a device other than an Extreme Networks device), you must configure the ethertype value on the Extreme Networks device port to match the ethertype value on the third-party device to which it connects.

The secondary ethertype support feature allows you to define two ethertype values for VMAN frames and select either of the two values for each port.

For example, you can configure ports that connect to other Extreme Networks devices to use the default primary ethertype value, and you can configure ports that connect to other equipment to use the secondary ethertype value, which you can configure to match the requirements of that equipment.

When you create a VMAN, each VMAN port is automatically assigned the primary ethertype value. After you define a secondary ethertype value, you can configure a port to use the secondary ethertype value. If two switch ports in the same VMAN use different ethertype values, the switch substitutes the correct value at each port. For example, for VMAN edge switches and transit switches, the switch translates an ingress ethertype value to the network port ethertype value before forwarding. For egress traffic at VMAN edge switches, no translation is required because the switch removes the S-tag before switching packets to the egress port.

For BlackDiamond 8800 series switches, BlackDiamond X8, SummitStack, and the Summit family of switches, you can set the primary and secondary ethertypes to any value, provided that the two values are different.

## QoS Support

The VMAN (PBN) feature interoperates with many of the *QoS (Quality of Service)* and HQoS features supported in the ExtremeXOS software.

One of those features is egress queue selection, which is described in the next section. For more information on other QoS and HQoS features that work with VMANs, see Quality of Service on page 724.

## Egress Queue Selection

This feature examines the 802.1p value or Diffserv code point in a VMAN (PBN) S-tag and uses that value to direct the packet to the appropriate queue on the egress port.

Note
This feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

On some systems (listed in the *Feature License Requirements* document., you can configure this feature to examine the values in the C-tag or the S-tag. For instructions on configuring this feature, see Selecting the Tag used for Egress Queue Selection on page 557.

## VMAN Double Tag Support

The VMAN double tag feature adds an optional port CVID parameter to the existing untagged VMAN port configuration. When present, any untagged packet received on the port will be double tagged with the configured port CVID and SVID associated with the VMAN. Packets received with a single CVID on the same port will still have the SVID added. As double tagged packets are received from tagged VMAN ports and forwarded to untagged VMAN ports, the SVID associated with the VMAN is stripped. Additionally, the CVID associated with the configured Port CVID is also stripped in the same operation.

Much like the CVIDs configured as part of the CEP feature, the configured Port CVID is not represented by a *VLAN* within EXOS. The implication is that protocols and individual services cannot be applied to the Port CVID alone. Protocols and services are instead applied to the VMAN and/or port as the VMAN represents the true layer-2 broadcast domain. Much like regular untagged VMAN ports, MAC *FDB (forwarding database)* learning occurs on the VMAN, so duplicate MAC addresses received on multiple CVIDs that are mapped to the same VMAN can be problematic. Even when the additional Port CVID is configured, the port still has all of the attributes of a regular untagged VMAN port. This means that any single c-tagged packets received on the same port will have just the SVID associated with the VMAN added to the packet. Likewise, any egress packet with a CVID other than the configured Port CVID will have the SVID stripped.

### *Coexistence with Tagged VLANs Interfaces, CEP VMAN Interfaces, and Tagged VMAN Interfaces*

Since the port-cvid configuration still has the attributes of a regular untagged VMAN, all of the VLAN and VMAN exclusion and compatibility rules of a regular untagged VMAN port also apply. A list of these rules is contained in “EXOS Selective Q-in-Q.”

### *Protocol and Feature Interactions*

Because this feature leverages existing untagged VMAN port infrastructure, any protocol that works with a regular untagged VMAN port also works when the optional Port CVID is additionally configured. Protocols that locally originate control packets, such as *STP (Spanning Tree Protocol)* and ELRP which are used for loop prevention, transmit packets as natively untagged on the wire when the port is an untagged VMAN member. EXOS can also receive and process these untagged packets. This makes STP edge safeguard + BPDU guard or ELRP effective ways to detect and react to network loops on the device. However, because control packets are transmitted as untagged upstream, devices may need additional configuration support to properly detect remote loops not directly attached to the device. Other effective loop prevention mechanisms work without any interaction with untagged VMAN ports. For example, turning physical port auto-polarity off will prevent an accidental looped cable from becoming active. Likewise, storm-control rate limiting of broadcast and flood traffic can be applied in this environment to minimize the effects of a network loop.

In addition to detecting, preventing, and minimizing the effects of a network loop, user *ACL*s can be applied to gain visibility and control of L2, L3, and L4 match criteria, even with double tagged packets. All applicable ACL action modifiers are available in this environment. IP multicast pruning within a VMAN can be accomplished via normal *IGMP (Internet Group Management Protocol)* snooping. EXOS supports full IGMP snooping and IP multicast pruning of single tagged and double tagged packets. However, when an IP address is configured on the VMAN, the IGMP protocol engine will transmit single tagged packets on tagged VMAN ports or untagged packets on untagged VMAN ports. Therefore, upstream switch configuration and support may be necessary to properly propagate group memberships across the network.

# Configuration

## Configuring VMANs (PBNs)

*Guidelines for Configuring VMANs*

The following sections provide VMAN configuration guidelines for the supported platforms:

- Guidelines for All Platforms
- Guidelines for BlackDiamond X8 and 8000 Series Modules and Summit Family Switches

**Guidelines for All Platforms**

The following are VMAN configuration guidelines for all platforms:

- Duplicate customer MAC addresses that ingress from multiple VMAN access ports on the same VMAN can disrupt the port learning association process in the switch.
- VMAN names must conform to the guidelines described in Object Names on page 16.
- You must use mutually exclusive names for:
  - VLANs
  - VMANs
  - IPv6 tunnels
- VMAN ports can belong to load-sharing groups.

**Guidelines for BlackDiamond X8 and 8000 Series Modules and Summit Family Switches**

The following are VMAN configuration guidelines for BlackDiamond X8 and 8000 series modules, SummitStack, and Summit family switches:

- You can enable or disable jumbo frames before configuring VMANs. You can enable or disable jumbo frames on individual ports. See Configuring Ports on a Switch on page 180 for more information on configuring jumbo frames.
- *STP* operation on CVLAN components in a PEB as described in IEEE 802.1ad is not supported.

- The initial version of this feature does not implement an XML API.
- Multiple VMAN roles can be combined on one port with certain *VLAN* types as shown in the following table.

**Table 67: Port Support for Combined VMAN Roles and VLANs**

| Platform | Combined CNP, CEP, and Tagged VLAN [4] | Combined PNP, CNP, and CEP[a, b, 4] | Combined PNP and Tagged VLAN | Combined PNP and Untagged VLAN |
|---|---|---|---|---|
| Summit X440, X460, X460-G2, X480, X670, X670-G2, and X770, and E4G-200 and E4G-400 | X | X | X | X |
| BlackDiamond 8800 a-, c-, and e-series modules | X | X | X | X |
| BlackDiamond X8, 8900 c-, xl-, and xm-series modules | X | X | X | X |

**Note**
If you already configured VLANs and VMANs on the same module or stand-alone switch using ExtremeXOS 11.4, you cannot change the VMAN ethertype from 0X8100 without first removing either the VLAN or VMAN configuration.

*Procedure for Configuring VMANs*

This section describes the procedure for configuring VMANs. Before configuring VMANs, review Guidelines for Configuring VMANs on page 554. To configure a VMAN, complete the following procedure at each switch that needs to support the VMAN:

1. If you are configuring a BlackDiamond 8800 series switch, a SummitStack, or a Summit family switch, enable jumbo frames on the switch.

   **Note**
   Because the BlackDiamond 8800 series switches, SummitStack, and the Summit family of switches enable jumbo frames switch-wide, you must enable jumbo frames before configuring VMANs on these systems.

2. Create a VMAN using the command:

   ```
   create vman vman_name | vman_list vr vr_name
   ```

3. Assign a tag value to the VMAN using the command:.

   ```
   configure vman vman_name | vman_list tag tag_id
   ```

4 Subsets of this group are also supported. That is, any two of these items are supported.
5 When a CNP is combined with a CEP or tagged VLAN, any CVIDs not explicitly configured for a CEP or tagged VLAN are associated with the CNP.
6 A PNP (tagged VMAN) and a CNP (untagged VMAN) or CEP cannot be combined on a port for which the selected VMAN ethertype is 0x8100.
7 If the secondary VMAN ethertype is selected for the port, it must be set to 0x8100.

4. To configure PNP ports on a PEB or PB, use the following command with the **tagged** option:

   `configure vman vman_name | vman_idadd ports [ all | port_list ] {untagged { port-cvid port_cvid} | tagged}`

5. To configure CNP ports on a PEB, use the following command with the **untagged** option:

   `configure vman vman_name add ports [ all | port_list ] {untagged { port-cvid port_cvid} | tagged}`

   Note
   You must configure CNP ports as **untagged**, so that the S-tag is stripped from the frame on egress. If the **port-cvid** is configured, any untagged packet received on the port will be double tagged with the configured port CVID and the SVID associated with the VMAN. Packets received with a single CVID on the same port will still have the SVID added as usual. As double tagged packets are received from tagged VMAN ports and forwarded to untagged VMAN ports,the SVID associated with the VMAN is stripped. Additionally, the CVID associated with the configured port CVID is also stripped in the same operation.

6. To configure CEP ports on a PEB, do the following:
   a. Use the following command to establish a physical port as a CEP and configure CVID mapping and translation:

      `configure vman vman_name | vman_idadd ports port_list cep cvid cvid_range {translate cvid | cvid_range}}`

   b. Use the following commands to add or delete CVIDs for a CEP and manage CVID mapping and translation:

      `configure vman vman_name | vman_idports port_list add cvid {cvid | cvid_range} {translate cvid | cvid_range }`

      `configure vman vman_name ports port_list delete cvid {cvid | cvid_range }`

   c. Use the following commands to manage CVID egress filtering for a CEP:

      `enable vman cep egress filtering ports {port_list | all}`

      `disable vman cep egress filtering ports {port_list | all}`

7. Configure additional VMAN options as described in Configuring VMAN Options on page 556.
8. To configure a *VLAN* to use a VMAN, configure the VLAN on the switch port at the other end of the line leading to the VMAN access port.

## Configuring VMAN Options

### *Configuring the Ethertype for VMAN Ports*

The ethertype is a component of *VLAN* and VMAN frames. It is introduced in Secondary Ethertype Support on page 551.

Note
This feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

To configure the ethertype for VMAN (PBN) ports, do the following:

1. Configure the primary and secondary (if needed) VMAN ethertype values for the switch using the following command:

   `configure vman ethertype hex_value [primary | secondary]`

   By default, all VMAN ports use the primary ethertype value.
2. If you plan to use a secondary ethertype, select the secondary ethertype for the appropriate VMAN ports using the following command:

   `configure port port_list ethertype {primary | secondary}`

*Selecting the Tag used for Egress Queue Selection*

By default, switches that support the enabling and disabling of this feature use the 802.1p value in the S-tag to direct the packet to the queue on the egress port.

Note
This feature is supported and configurable only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

- Configure egress queue dot1p examination of the C-tag using:

  `enable dot1p examination inner-tag port [all | port_list]`
- Return to the default selection of using the 802.1p value in the S-tag using:

  `disable dot1p examination inner-tag ports [all | port_list]`

Note
See Quality of Service on page 724 for information on configuring and displaying the current 802.1p and DiffServ configuration for the S-tag 802.1p value. To enable dot1p examination for inner-tag, dot1p examination for outer-tag must be disabled using the command `disable dot1p examination ports [all | port_list]`

# Displaying Information

## Displaying VMAN Information

Use the following commands to display information on one or all VMANs.

**show** {**vman**} *vman_name* {**ipv4** | **ipv6**}

**show vman** [**tag** *tag_id* | **detail**] {**ipv4** | **ipv6**}

show {**vman**} *vman_name* **eaps**

Note
The display for the `show vman` command is different depending on the platform and configuration you are using. See the *ExtremeXOS 16.2 Command Reference Guide* for complete information on this command.

You can also display VMAN information, as well as all the VLANs, by issuing the `show ports information detail` command. And you can display the VMAN ethernet type and secondary etherType port_list by using the `show vman etherType` command

# Configuration Examples

## VMAN Example, BlackDiamond 8810

The following example shows the steps to configure a VMAN (PBN) on the BlackDiamond 8810 switch shown in the following figure.

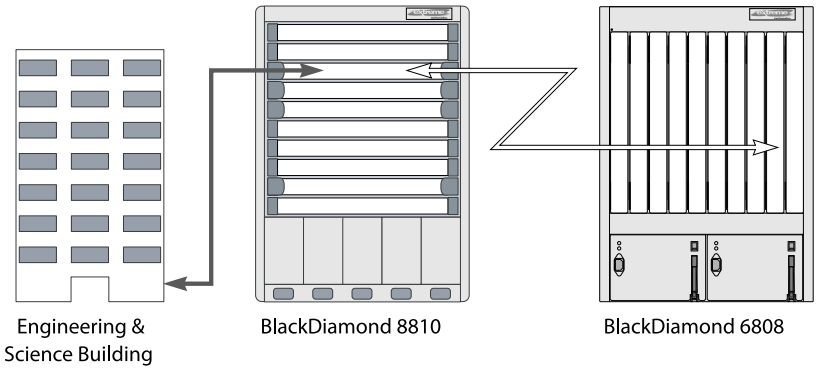



**Figure 87: Sample VMAN Configuration on BlackDiamond 8810 Switch**

The VMAN is configured from the building to port 1, slot 3 on the BlackDiamond 8810 switch and from port 2, slot 3 on the BlackDiamond 8810 switch to the BlackDiamond® 6808 switch:

```
enable jumbo frames
create vman vman_tunnel_1
configure vman vman_tunnel_1 tag 100
configure vman vman_tunnel_1 add port 3:1 untagged
configure vman vman_tunnel_1 add port 3:2 tagged
disable dot1p examination port 3:2
enable dot1p examination inner-tag port 3:2
```

The following example configuration demonstrates configuring IP multicast routing between VMANs and VLANs (when VMAN traffic is not double-tagged) on the BlackDiamond 8800 series switch and the Summit family of switches.

Using this configuration you can use a common uplink to carry both *VLAN* and VMAN traffic and to provide multicast services from a VMAN through a separate VLAN (notice that port 1:1 is in both a VLAN and a VMAN):

```
enable jumbo-frame ports all
configure vman ethertype 0x8100
create vlan mc_vlan
configure vlan mc_vlan tag 77
create vman vman1
configure vman vman1 tag 88
configure vlan vman1 ipaddress 10.0.0.1/24
configure vlan mc_vlan ipaddress 11.0.0.1/24
enable ipforwarding vman1
enable ipforwarding mc_vlan
```

```
enable ipmcforwarding vman1
enable ipmcforwarding mc_vlan
configure vlan mc_vlan add port 1:1 tag
configure vman vman1 add port 1:1 tag
configure vman vman1 add port 2:1, 2:2, 2:3
```

Note

*IGMP* reports can be received untagged on ports 2:1, 2:2, and 2:3. Tagged IP multicast data is received on mc_vlan port 1:1 and is routed using IP multicasting to vman1 ports that subscribe to the IGMP group.

IGMP snooping (Layer 2 IP multicasting forwarding) does not work on the VMAN ports because there is no double-tagged IP multicast cache lookup capability from port 1:1.

## VMAN CEP Example

The following configuration configures a VMAN CEP to support up to 10 customer VLANs for each of three VMANs.

```
create vman cust1
create vman cust2
create vman cust3
config vman cust1 tag 1000
config vman cust2 tag 1001
config vman cust3 tag 1002
config vman cust1 add port 22 tag
config vman cust2 add port 22 tag
config vman cust3 add port 23 tag
config vman cust1 add port 1 cep cvid 100 - 109
config vman cust2 add port 1 cep cvid 110 - 119
config vman cust3 add port 1 cep cvid 120 - 129
enable vman cep egress filtering ports 1
```

Port 1 serves as the CEP, and egress filtering is enabled on the port. Ports 22 and 23 serve as CNPs, providing the connection between the CEP port and the rest of each VMAN.

## Multiple VMAN Ethertype Example

The following figure shows a switch that is configured to support the primary ethertype on three ports and the secondary ethertype on a fourth port.

The primary VMAN (PBN) ethertype is changed from the default value, but that is not required.

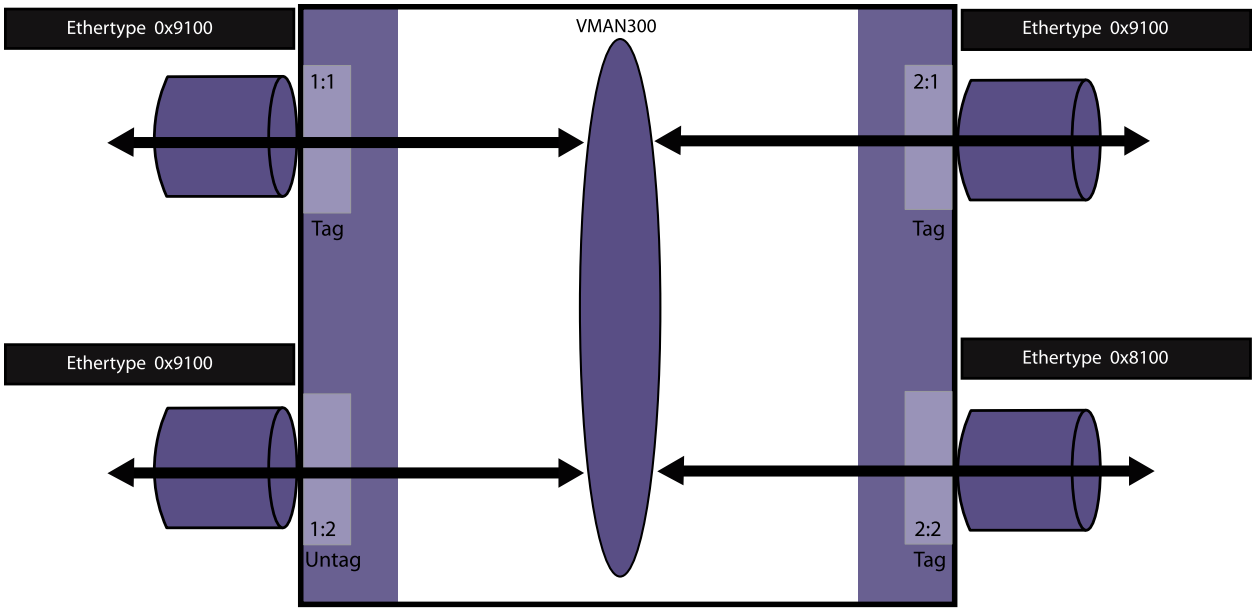
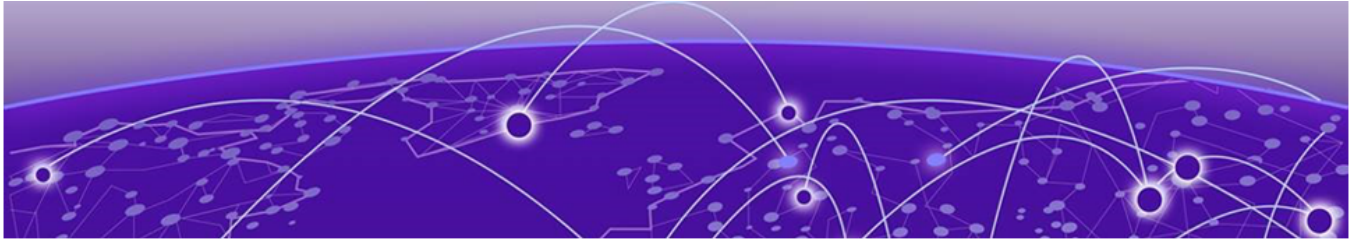



**Figure 88: Multiple VMAN Ethertype Example**

The following configuration commands accomplish what is shown in the figure above:

```
# configure vman ethertype 0x9100 primary
# configure vman ethertype 0x8100 secondary
#
# configure port 2:2 ethertype secondary
#
# create vman vman300
# configure vman vman300 tag 300
#
# configure vman vman300 add port 1:1, 2:1, 2:2 tagged
# configure vman vman300 add port 1:2 untagged
```



# FDB



The *FDB (forwarding database)* chapter is intended to help you learn about forwarding databases, adding and displaying entries and entry types, managing the FDB, and MAC-based security. This chapter also provides information about MAC Address tracking.

The switch maintains a forwarding database (FDB) of all MAC addresses received on all of its ports. It uses the information in this database to decide whether a frame should be forwarded or filtered.

Note
See the *ExtremeXOS 16.2 Command Reference Guide* for details of the commands related to the FDB.

## FDB Contents

Each *FDB* entry consists of:

- The MAC address of the device
- An identifier for the port and *VLAN (Virtual LAN)* on which it was received
- The age of the entry
- Flags

Frames destined for MAC addresses that are not in the FDB are flooded to all members of the VLAN.

## How FDB Entries Get Added

The MAC entries that are added to the *FDB* are learned in the following ways:

- Source MAC entries are learned from ingress packets on all platforms. This is Layer 2 learning.
- On all switches, MAC entries can be learned at the hardware level.
- Virtual MAC addresses embedded in the payload of IP ARP packets can be learned when this feature is enabled.

- Static entries can be entered using the command line interface (CLI).
- Dynamic entries can be modified using the CLI.
- Static entries for switch interfaces are added by the system upon switch boot-up.

The ability to learn MAC addresses can be enabled or disabled on a port-by-port basis. You can also limit the number of addresses that can be learned, or you can lock down the current entries and prevent additional MAC address learning.

BlackDiamond 8000 and BlackDiamond X8 series modules and Summit switches support different FDB table sizes.

On a BlackDiamond 8800 and BlackDiamond X8 series switch with a variety of modules or on a SummitStack with different Summit switch models, the FDB tables on some modules or switches can be filled before the tables on other modules or switches. In this situation, when a lower-capacity FDB table cannot accept FDB entries, a message appears that is similar to the following:

```
 HAL.FDB.Warning> MSM-A: FDB for vlanID1 mac 00:00:03:05:15:04 was not added to slot 3 -
table full.
```

Note

For information on increasing the FDB table size on BlackDiamond 8900 and BlackDiamond X8 xl-series modules, and Summit X480 switches, see Increasing the FDB Table Size on page 564. For information on FDB tables sizes, see the ExtremeXOS Release Notes.

## How FDB Entries Age Out

### Software Aging Platforms (All Platforms except Summit X480 and X460-G2, 100G4X, and 100G4X-xl and BD8900 xl-series cards)

When a MAC is learned on a *VLAN*, an *FDB* entry is created for this MAC VLAN combination.Once an FDB entry is created, the aging counter in the "show fdb" output increases from 0 to “polling inteval”.

The hardware is checked every “polling interval” seconds to see if there is traffic flow from the given FDB entry.If there is a traffic flow from this MAC, the entry is refreshed and aging counter is reset to 0.

If there is no traffic from that FDB entry during this polling interval, the age gets incremented till the configured age time (configured by `configure fdb agingtime` *`seconds`*).The entry gets removed when there is no traffic flow from this FDB entry when the age count reaches the configured age time.

Polling interval = FDB aging time/4 (subject to the minimum and maximum values being 10 and 60 seconds respectively).

### Hardware Aging Platforms (Only Summit X480 and X460-G2, 100G4X, and 100G4X-xl, and BD8900 xl-series cards)

Aging is controlled entirely by the hardware based on the traffic hit that happens for the individual FDB entry. The age from `show fdb` output is alway shown as 0. The entry will be removed when it ages out.

# FDB Entry Types

## Dynamic Entries

A dynamic entry is learned by the switch by examining packets to determine the source MAC address, *VLAN*, and port information.

The switch then creates or updates an *FDB* entry for that MAC address. Initially, all entries in the database are dynamic, except for certain entries created by the switch at boot-up.

Entries in the database are removed (aged-out) if, after a period of time (aging time), the device has not transmitted. This prevents the database from becoming full with obsolete entries by ensuring that when a device is removed from the network, its entry is deleted from the database.

The aging time is configurable, and the aging process operates on the supported platforms as follows:

- On all platforms, you can configure the aging time to 0, which prevents the automatic removal of all dynamic entries.
- On BlackDiamond X8 series switches, BlackDiamond 8000 a-, c-, e- and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X440, X460, X670, X670-G2, and X770 series switches, the aging process takes place in software and the aging time is configurable.
- On BlackDiamond 8900 xl-series and Summit X480 and X460-G2, 100G4X, and 100G4X-xl switches, the aging process takes place in hardware and the aging time is based on (but does not match) the configured software aging time.

For more information about setting the aging time, see Configuring the FDB Aging Time on page 566.

Note

If the FDB entry aging time is set to 0, all dynamically learned entries in the database are considered static, non-aging entries. This means that the entries do not age, but they are still deleted if the switch is reset.

Dynamic entries are flushed and relearned (updated) when any of the following take place:

- A VLAN is deleted.
- A VLAN identifier (VLANid) is changed.
- A port mode is changed (tagged/untagged).
- A port is deleted from a VLAN.
- A port is disabled.
- A port enters blocking state.
- A port goes down (link down).

A non-permanent dynamic entry is initially created when the switch identifies a new source MAC address that does not yet have an entry in the FDB. The entry can then be updated as the switch continues to encounter the address in the packets it examines. These entries are identified by the “d” flag in the `show fdb` command output.

## Static Entries

A static entry does not age and does not get updated through the learning process.

A static entry is considered permanent because it is retained in the database if the switch is reset or a power off/on cycle occurs. A static entry is maintained exactly as it was created. Conditions that cause dynamic entries to be updated, such as *VLAN* or port configuration changes, do not affect static entries.

To create a permanent static *FDB* entry, see Adding a Permanent Unicast Static Entry on page 565.

If a duplicate MAC address is learned on a port other than the port where the static entry is defined, all traffic from that MAC address is dropped. By default, the switch does not report duplicate addresses. However, you can configure the switch to report these duplicate addresses as described in Managing Reports of Duplicate MAC Addresses for Static Entries on page 567.

A locked static entry is an entry that was originally learned dynamically, but has been made static (locked) using the MAC address lock-down feature. It is identified by the "s," "p," and "l" flags in show fdb command output and can be deleted using the `delete fdb` command. See MAC Address Lockdown on page 874 for more information about this feature.

Note

Static FDB entries created on EAPS- or *STP (Spanning Tree Protocol)*-enabled ports forward traffic irrespective of the port state. Consequently, you should avoid such a configuration.

## Blackhole Entries

A blackhole entry configures the switch to discard packets with a specified MAC destination address.

Blackhole entries are useful as a security measure or in special circumstances where a specific source or destination address must be discarded. Blackhole entries can be created through the CLI, or they can be created by the switch when a port's learning limit has been exceeded.

Blackhole entries are treated like permanent entries in the event of a switch reset or power off/on cycle. Blackhole entries are never aged out of the database.

## Private VLAN Entries

A Private *VLAN* (PVLAN) creates special *FDB* entries. These are described in MAC Address Management in a PVLAN on page 522.

# Managing the FDB

## Increasing the FDB Table Size

BlackDiamond 8900 xl-series and BlackDiamond X8 xl-series modules and Summit X480 switches provide an additional table that can be configured to support additional *FDB* table entries with the following command:

`configure forwarding external-tables [` **`l3-only`** `{` **`ipv4`** `|` **`ipv4-and-ipv6`** `|` **`ipv6`**`} |` **`l2-only`** `|` **`acl-only`** `|` **`l2-and-l3`** `|` **`l2-and-l3-and-acl`** `|` **`l2-and-l3-and-ipmc`** `|` **`none`**`]`

Summit X670-G2, X460-G2, X770, 100G4X, and 100G4X-xl switches provides a Unified Forwarding Table that allows for flexible allocation of entries to L2 or L3. You can configure this table with the `configure forwarding internal-tables [` **`l2-and-l3`** `|` **`more`** `[` **`l2`** `|` **`l3-and-ipmc`**`] ]` command.

## Adding a Permanent Unicast Static Entry

To add a static entry use the following command:

`create fdb` *`mac_addr`* **`vlan`** *`vlan_name`* `[` **`ports`** *`port_list`* `|` **`blackhole`**`]`

The following example adds a permanent static entry to the *FDB*:

```
create fdb 00:E0:2B:12:34:56 vlan marketing port 3:4
```

The permanent entry has the following characteristics:

- MAC address is 00:E0:2B:12:34:56.
- *VLAN* name is marketing.
- Slot number for this device is 3 (only on modular switches).
- Port number for this device is 4.

On Summit family switches, BlackDiamond X8 series switches, and BlackDiamond 8000 series modules, you can specify multiple ports when you create a unicast static entry. However, all ports in the list must be on the same SummitStack switch, BlackDiamond X8 series switch or BlackDiamond 8000 series module. When the port list contains ports on different slots, the following error is generated:

```
Error: Multiple ports must be on the same slot for unicast MAC FDB entries.
```

Once the multiport static FDB entry is created, any ingress traffic with a destination MAC address matching the FDB entry is multicasted to each port in the specified list. On Summit family switches and BlackDiamond 8000 series modules, if the FDB entry is the next hop for an IP adjacency, unicast routing sends the packet to the first port in the list.

**Note**

When a multiport list is assigned to a unicast MAC address, load sharing is not supported on the ports in the multiport list.

Summit family switches, BlackDiamond X8 series switches, and BlackDiamond 8000 series modules do not support this multiport feature natively using the FDB table. Instead, for each FDB entry of this type, a series of system *ACL (Access Control List)*s have been installed which match the specified MAC address and VLAN ID, and override the egress port forwarding list with the supplied list of ports. Multiple ACLs per FDB are required to handle Layer 2 echo kill by installing a unique ACL per individual port in the list to send matching traffic to all other ports in the list.

User-configured ACLs take precedence over these FDB-generated ACL rules, and the total number of rules is determined by the platform.

The hardware ACL limitations for each platform are described in ACLs on page 640.

## Adding a Permanent Multicast Static Entry

On BlackDiamond X8 series switches, BlackDiamond 8000 series modules, SummitStack, and Summit family switches, you can create *FDB* entries to multicast MAC addresses (that is, 01:00:00:00:00:01) and list one or more ports.

Use the `create fdb` *`mac_addr`* **`vlan`** *`vlan_name`* [**`ports`** *`port_list`* | **`blackhole`**] command to enter the multicast FDB address. After traffic with a multicast MAC destination address enters the switch, that traffic is multicast to all ports on the list.

However, if the MAC address is in the IP multicast range (for example, 01:00:5e:XX:XX:XX), *IGMP (Internet Group Management Protocol)* snooping rules take precedence over the multicast static FDB entry. Of course, if you disable IGMP snooping on all VLANs, the static FDB entry forwards traffic.

## Configuring the FDB Aging Time

- To configure the aging time for dynamic *FDB* entries, use the following command:.

  `configure fdb agingtime` *`seconds`*

  If the aging time is set to 0, all aging entries in the database are defined as static, nonaging entries. This means the entries will not age out, but non-permanent static entries can be deleted if the switch is reset.
- To display the aging time, use the following command:.

  `show fdb`

Note

On BlackDiamond 8900 xl-series, Summit X480, X460-G2, 100G4X, and 100G4X-xl switches, FDB entries are aged in hardware, the aging time is always displayed as 000, and the h flag is set for entries that are hardware aged.

## Adding Virtual MAC Entries from IP ARP Packets

Generally, the *FDB* is programmed with the source MAC address of frames that contain an IP ARP payload. MAC entries present in the ARP payload as Sender-MAC are not learned. When IP ARP Sender-MAC learning is enabled, the switch learns both the source MAC address and the Sender-MAC from the ARP payload, and the switch programs these MAC addresses in the FDB.

This feature is useful when you want the switch to learn the Sender-MAC address for a redundant protocol, such as *VRRP (Virtual Router Redundancy Protocol)*. For example, if your network has a gateway with a virtual MAC address, the switch learns the system MAC address for the gateway. If you enable the IP ARP Sender-MAC learning feature, the switch also learns the virtual MAC address embedded in IP ARP packets for the gateway IP address.

- To enable the IP ARP sender-MAC learning feature, use the command:

  `enable learning iparp` **`sender-mac`**
- To view the configuration of this feature, use the command:

  `show iparp`
- To disable this feature, use the command:

  `disable learning iparp` **`sender-mac`**

## Managing Reports of Duplicate MAC Addresses for Static Entries

By default, if a MAC address that is a duplicate of a static MAC address entry is learned on another port (other than the port where the static MAC address is configured), traffic from the duplicate address is silently dropped.

- To enable or disable *EMS (Event Management System)* and *SNMP (Simple Network Management Protocol)* reporting of duplicate addresses for static entries, use the commands:

  ```
  enable fdb static-mac-move
  disable fdb static-mac-move
  ```

- To control the number of EMS and SNMP reports per second issued, use the commands:

  ```
  configure fdb static-mac-move packets count
  ```

- To display the configuration of this feature, use the commands:

  ```
  show fdb static-mac-move configuration
  ```

## Clearing FDB Entries

You can clear dynamic and permanent entries using different CLI commands. Clear dynamic *FDB* entries by targeting:

- Specified MAC addresses
- Specified ports
- Specified VLANs
- All blackhole entries

- To clear dynamic entries from the FDB, use the command:

  ```
  clear fdb {mac_addr | ports port_list | vlan vlan_name | blackhole}
  ```

- To clear permanent entries from the FDB, use the command:

  ```
  delete fdb [all | mac_address [vlan vlan_name ]
  ```

## Supporting Remote Mirroring

The remote mirroring feature copies select traffic from select ports and *VLAN*s and sends the copied traffic to a remote switch for analysis.

The mirrored traffic is sent using a VLAN that is configured for this purpose. For more information, see MLAG Limitations and Requirements on page 272.

Transit switches are the switches between the source switch where ports are mirrored and the destination switch where the mirrored traffic exits the network to a network analyzer or network storage device.

Because the mirrored traffic is an exact copy of the real traffic, a transit switch can learn the MAC addresses and make incorrect forwarding decisions.

# Displaying FDB Entries and Statistics

## Display FDB Entries

- Display *FDB* entries using the following command:

  **show fdb** {**blackhole** {**netlogin** [**all** | **mac-based-vlans**]} | **netlogin** [**all** | **mac-based-vlans**] | **permanent** {**netlogin** [**all** | **mac-based-vlans**]} | *mac_addr* {**netlogin** [**all** | **mac-based-vlans**]} | **ports** *port_list* {**netlogin** [**all** | **mac-based-vlans**]} | **vlan** *vlan_name* | *vlan_list* {**netlogin** [**all** | **mac-based-vlans**]} | {{**vpls**} {*vpls_name*}}}

  Note
  The MAC-based *VLAN* netlogin parameter applies only for Summit family switches and BlackDiamond 8800 series switches. See Network Login on page 756 for more information.

  With no options, this command displays all FDB entries. (The age parameter does not show on the display for the backup MSM/MM on modular switches; it does show on the display for the primary MSM/MM.)

## Display FDB Statistics

To display *FDB* statistics, use the command:

**show fdb stats** {{**ports** {**all** | *port_list*} | **vlan** {**all**} | {**vlan**} *vlan_name* | *vlan_list* } {**no-refresh**}}

With no options, this command displays summary FDB statistics.

# MAC-Based Security

MAC-based security allows you to control the way the *FDB* is learned and populated. By managing entries in the FDB, you can block and control packet flows on a per-address basis.

MAC-based security allows you to limit the number of dynamically-learned MAC addresses allowed per virtual port. You can also “lock” the FDB entries for a virtual port, so that the current entries will not change, and no additional addresses can be learned on the port.

You can also prioritize or stop packet flows based on the source MAC address of the ingress *VLAN* or the destination MAC address of the egress VLAN.

Note
For detailed information about MAC-based security, see Security on page 859.

## Managing MAC Address Learning

By default, MAC address learning is enabled on all ports. MAC addresses are added to the *FDB* as described in How FDB Entries Get Added on page 561.

When MAC address learning is disabled on a port, the switch no longer stores the source address information in the FDB. However, the switch can still examine the source MAC address for incoming packets and either forward or drop the packets based on this address. The source address examination serves as a preprocessor for packets. Forwarded packets are forwarded to other processes, not to other ports. For example, if the switch forwards a packet based on the source address, the packet can still be dropped based on the destination address or the egress flooding configuration.

When MAC address learning is disabled, the two supported behaviors are labeled as follows in the software:

- forward-packets
- drop-packets

The drop-packets behavior is supported on BlackDiamond 8000 series modules, SummitStack, and Summit family switches. When the drop-packets option is chosen, *EDP (Extreme Discovery Protocol)* packets are forwarded, and all unicast, multicast, and broadcast packets from a source address not in the FDB are dropped. No further processing occurs for dropped packets.

The disable learning forward-packets option saves switch resources (FDB space), however, it can consume network resources when egress flooding is enabled. When egress flooding is disabled or the drop-packet option is specified, disabling learning adds security by limiting access to only those devices listed in the FDB.

Note

When the forward-packet option is chosen,

- If unicast, multicast, and broadcast packet from a source address is not present in the FDB , the packets is flooded.
- If the destination MAC is present in the forwarding database, the packet is forwarded.

- To disable learning on specified ports, use the command:

  `disable learning {`**`drop-packets`** `|` **`forward-packets`**`} port [`*`port_list`* `|` **`all`**`]`

  Note

  The **drop-packets** and **forward-packets** options are available only on the BlackDiamond 8800 series switches, SummitStack, and the Summit family switches. If neither option is specified, the **drop-packets** behavior is selected.

- To enable learning on specified ports, use the command:

  `enable learning {`**`drop-packets`**`} ports [`**`all`** `|` *`port_list`*`]`

## Managing Egress Flooding

Egress flooding takes action on a packet based on the packet destination MAC address. By default, egress flooding is enabled, and any packet for which the destination address is not in the *FDB* is flooded to all ports except the ingress port.

You can enhance security and privacy as well as improve network performance by disabling Layer 2 egress flooding on a port, *VLAN*, or VMAN. This is particularly useful when you are working on an edge

device in the network. Limiting flooded egress packets to selected interfaces is also known as upstream forwarding.

**Note**
Disabling egress flooding can affect many protocols, such as IP and ARP.

The following figure illustrates a case where you want to disable Layer 2 egress flooding on specified ports to enhance security and network performance.

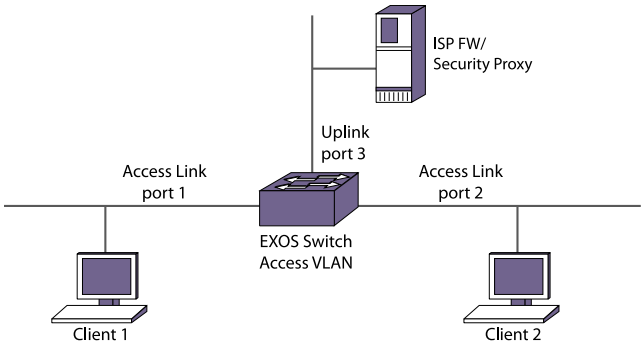



**Figure 89: Upstream Forwarding or Disabling Egress Flooding Example**

In this example, the three ports are in an ISP-access VLAN. Ports 1 and 2 are connected to clients 1 and 2, respectively, and port 3 is an uplink to the ISP network. Because clients 1 and 2 are in the same VLAN, client 1 could possibly learn about the other client's traffic by sniffing client 2's broadcast traffic; client 1 could then possibly launch an attack on client 2.

However, when you disable all egress flooding on ports 1 and 2, this sort of attack is impossible, for the following reasons:

- Broadcast and multicast traffic from the clients is forwarded only to the uplink port.
- Any packet with unlearned destination MAC addresses is forwarded only to the uplink port.
- One client cannot learn any information from the other client. Because egress flooding is disabled on the access ports, the only packets forwarded to each access port are those packets that are specifically targeted for one of the ports. There is no traffic leakage.

In this way, the communication between client 1 and client 2 is controlled. If client 1 needs to communicate with client 2 and has that IP address, client 1 sends out an ARP request to resolve the IP address for client 2.

### *Guidelines for Enabling or Disabling Egress Flooding*

The following guidelines apply to enabling and disabling egress flooding:

- Egress flooding can be disabled on ports that are in a load-sharing group. In a load-sharing group, the ports in the group take on the egress flooding state of the master port; each member port of the load-sharing group has the same state as the master port.
- *FDB* learning takes place on ingress ports and is independent of egress flooding; either can be enabled or disabled independently.
- Disabling unicast (or all) egress flooding to a port also prevents the flooding of packets with unknown MAC addresses to that port.

- Disabling broadcast (or all) egress flooding to a port also prevents the flooding of broadcast packets to that port.
- For BlackDiamond X-8 and 8800 series switches, SummitStack, and Summit family switches, the following guidelines apply:
  - You can enable or disable egress flooding for unicast, multicast, or broadcast MAC addresses, as well as for all packets on one or more ports.
  - Disabling multicasting egress flooding does not affect those packets within an *IGMP* membership group at all; those packets are still forwarded out.
  - If IGMP snooping is disabled, multicast packets with static FDB entries are forwarded according to the FDB entry.

*Configuring Egress Flooding*

To enable or disable egress flooding on BlackDiamond X8 and 8800 series switches, SummitStack, and the Summit family switches, use the following commands:

```
enable flooding [all_cast | broadcast | multicast | unicast] ports
[port_list | all]
```

```
disable flooding [all_cast | broadcast | multicast | unicast] ports
[port_list | all]
```

## Displaying Learning and Flooding Settings

To display the status of MAC learning and egress flooding, use the following commands:

```
show ports {mgmt | port_list | tag tag} information {detail}
```

```
show vlan {virtual-router vr-name}
```

```
show vman vman_list
```

The flags in the command display indicate the status.

## Creating Blackhole FDB Entries

A blackhole *FDB* entry discards all packets addressed to or received from the specified MAC address. A significant difference between the above *ACL* policy and the create fdb command **blackhole** option is the hardware used to implement the feature. Platforms with limited hardware ACL table sizes (for example, BlackDiamond 8800 series switches) are able to implement this feature using the FDB table instead of an ACL table.

- To create a blackhole FDB entry, use the command:

  ```
  create fdb mac_addr vlan vlan_name [ports port_list | blackhole]
  ```

  There is no software indication or notification when packets are discarded because they match blackhole entries.

The **blackhole** option is also supported through access lists.

Note
Blackhole is not supported on port-specific *VLAN* tags.

For example, the following ACL policy would also blackhole traffic destined to or sourced from a specific MAC address:

```
entry blackhole_dest {
if {
ethernet-destination-address 00:00:00:00:00:01;
} then {
deny;
}
}
entry blackhole_source {
if {
ethernet-source-address 00:00:00:00:00:01;
} then {
deny;
}
}
```

## Managing MAC Address Tracking

The MAC address tracking feature tracks *FDB* add, move, and delete events for specified MAC addresses and for specified ports.

When MAC address tracking is enabled for a port, this feature applies to all MAC addresses on the port.

When an event occurs for a specified address or port, the software generates an *EMS* message and can optionally send an *SNMP* trap. When MAC address tracking is enabled for a specific MAC address, this feature updates internal event counters for the address. You can use this feature with the Universal Port feature to configure the switch in response to MAC address change events (for an example, see Universal Port on page 309).

Note
When a MAC address is configured in the tracking table, but detected on a MAC tracking enabled port, the per MAC address statistical counters are not updated.

The MAC address tracking feature is always enabled; however, you must configure MAC addresses or ports before tracking begins. The default configuration contains no MAC addresses in the MAC address tracking table and disables this feature on all ports.

### Adding and Deleting MAC Addresses for Tracking

Use the following commands to add or delete MAC addresses in the MAC address tracking table:

```
create fdb mac-tracking entry mac_addr

delete fdb mac-tracking entry [mac_addr | all]
```

## Enabling and Disabling MAC Address Tracking on Ports

Use the following command to enable or disable MAC addresses tracking on specific ports:

```
configure fdb mac-tracking {[add|delete]} ports [port_list|all]
```

## Enabling and Disabling SNMP Traps for MAC Address Changes

The default switch configuration disables *SNMP* traps for MAC address changes.

Use the following commands to enable or disable SNMP traps for MAC address tracking events:

```
enable snmp traps fdb mac-tracking
```

```
disable snmp traps fdb mac-tracking
```

## Configuring Automatic Responses to MAC Tracking Events

The *EMS* messages produced by the MAC address tracking feature can be used to trigger Universal Port profiles. These are described in Event Management System Triggers on page 315.

The subcomponent name for MAC address tracking events is FDB.MACTracking.

## Displaying the Tracked MAC Addresses and Tracking Statistics

- To display the MAC address tracking feature configuration, including the list of tracked MAC addresses, use the command:

  ```
  show fdb mac-tracking configuration
  ```

- To display the counters for MAC address add, move, and delete events, use the command:

  ```
  show fdb mac-tracking statistics {mac_addr} {no-refresh}
  ```

## Clearing the Tracking Statistics Counters

There are several ways to clear the MAC tracking counters:

- Use the `clear counters` command.
- Use the **0** key while displaying the counters with the `show fdb mac-tracking statistics {mac_addr}` command.
- Enter the `clear counters fdb mac-tracking [mac_addr | all]` command.

# Data Center Solutions



This chapter provides information about Extreme Network's Data Center Solutions. It provides an overview of data centers and provides information about how to configure and manage data center features, including DCBX, *Extreme Network Virtualization (XNV)*, VM Tracking, Direct Attach to support VEPA, and FIP Snooping.

## Data Center Overview

Typical data centers support multiple Virtual Machines (VMs) on a single server. These VMs usually require network connectivity to provide their services to network users and to other VMs. The following sections introduce ExtremeXOS software features that support VM network connectivity.

Note

For additional information on using ExtremeXOS features to implement Data Center Bridging, see the application note titled *Enhanced Transmission Selection (ETS) Deployment and Configuration for ExtremeXOS* on the Extreme Networks website.

### Introduction to Data Center Bridging

*DCB (Data Center Bridging)* is a set of IEEE 802.1Q extensions to standard Ethernet, that provide an operational framework for unifying Local Area Networks (LAN), Storage Area Networks (SAN) and Inter-Process Communication (IPC) traffic between switches and endpoints onto a single transport layer.

#### *Data Center Bridging Exchange Protocol*

The Data Center Bridging eXchange (DCBX) protocol is used by *DCB* devices to exchange DCB configuration information with directly connected peers. In an ExtremeXOS enabled switch, the switch uses DCBX to advertise its DCB configuration to end stations. The end stations can then configure themselves to use the switch DCB services. If the peers do not support a common configuration for one or more features, the switch generates messages to alert network management. The switch does not accept configuration change requests from end stations.

The DCBX protocol advertises the following types of information:

- DCBX version information, so that the peers can negotiate a common version to use.
- Enhanced Transmission Selection (ETS) information for *QoS (Quality of Service)* parameters such as bandwidth allocation per traffic class (802.1p COS), priority for each traffic class, and the algorithm used for servicing traffic classes.
- Priority-based Flow Control (PFC) information for managing flow control between peers.
- Application priority information for prioritizing traffic for special applications that do not map directly to an established traffic class.

The ExtremeXOS software supports two versions of DCBX standards. The first version is a pre-standard version known as the baseline version, or more specifically as Baseline Version 1.01. The DCBX baseline version is specified in DCB Capability Exchange Protocol Base Specification Rev 1.01 and was developed outside of the IEEE and later submitted to the IEEE for standardization. The IEEE agreed to standardize DCBX as part of IEEE 802.1Qaz Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes. While IEEE 802.1Qaz has progressed through the standards process, many companies have released support for the baseline version. IEEE 802.1Qaz is nearing completion, and support is expected to start rolling out during 2011.

After you enable DCBX, the protocol collects most of the information to be advertised from other switch services such as QoS and PFC. The only DCBX feature that needs configuration is the application priority feature.

DCBX uses the *LLDP (Link Layer Discovery Protocol)* (IEEE 802.1AB) to exchange attributes between two link peers. DCBX attributes are packaged into organizationally specific TLVs, which are different for the Baseline and IEEE 802.1Qaz versions. Information on the TLV support differences is provided in the ExtremeXOS Command Reference under the command description for the command: `show lldp {port [all | port_list]} dcbx {ieee|baseline} {detailed}`

### *Custom Application Support*

The DCBX custom application support feature allows you to prioritize and manage traffic flow through the switch based on the application type. This feature allows you to configure DCBX handling of the following applications:

- Fiber Channel Over Ethernet (FCoE)
- FCoE Initiation Protocol (FIP)
- Internet Small Computer System Interface (iSCSI)
- Any application that can be defined by:
  - Ethertype value
  - Layer 4 port number
  - TCP port number
  - UDP port number

When you configure a custom application, you define a priority number that applies to traffic related to that application. DCBX advertises this priority to end stations in an application TLV. End stations that support this feature use the priority number for communications with the switch. The priority number maps to an 802.1p value, which determines which *QoS* profile in the switch manages the application traffic.

The software supports a maximum of eight application configurations.

*Enhanced Transmission Selection*

Enhanced Transmission Selection is defined in IEEE P802.1Qaz/D2.3, Virtual Bridged Local Area Networks-Amendment XX: Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes. This IEEE 802.1Qaz standard also defines one of the DCBX versions supported by the ExtremeXOS software.

ETS, and similar features in the Baseline DCBX standard, define methods for managing bandwidth allocation among traffic classes (called Priority Groups (PGs) in Baseline DCBX) and mapping 802.1p COS traffic to those traffic classes.

The rest of this section provides general guidelines for configuring the ExtremeXOS *QoS* feature to conform to the ETS requirements. After you configure QoS, DCBX advertises the ETS compatible configuration to DCBX peers on all DCBX enabled ports.

ETS configuration is affected by the following set of QoS objects:

- QoS scheduler
- QoS profile
- dot1p

By default, the scheduling is set to strict-priority.

The following command enables ETS compatible (weighted) scheduling:

```
configure qosscheduler [strict-priority | weighted-round-robin |
weighted-deficit-round-robin] {ports [ port_list | port_group | all}
```

Each QoS profile supports an IEEE ETS traffic class (TC) or a Baseline DCBX priority group (PG). To determine which QoS profile serves a TC or PG, add the number 1 to the TC or PG number. For example, TC 0 and PG 0 are served by QoS profile 1. ExtremeXOS switches support up to eight QoS profiles and can therefore support up to eight TCs or PGs. The following QoS configuration changes affect the ETS/PG configuration:

- QoS profile:
  - When you create or delete a QoS profile, you add or remove support for the corresponding TC or PG.
  - The weight configuration helps determine the bandwidth for a TC or PG.
  - The use-strict-priority configuration overrides ETS scheduling and selects strict priority scheduling for the corresponding TC or PG.
  - The dot1p configuration maps each 802.1p priority, and the associated TC and PG, to a QoS profile. If you change the 802.1p mapping, it will change which QoS profile services each TC or PG.
- Per port configuration parameters:
  - minbw: Sets a minimum guaranteed bandwidth in percent.
  - maxbw: Sets a maximum guaranteed bandwidth in percent.
  - committed_rate: Sets a minimum guaranteed bandwidth in Kbps or Mbps.
  - peak_rate: Sets a maximum guaranteed bandwidth in Kbps or Mbps.

For example, the following set of commands creates a QoS profile (qp5) in preparation to support iSCSI traffic, maps packets with 802.1p priority 4 to QoS profile 5, indicates that QoS profile 8 should use strict priority, and sets the weight for the ETS classes:

```
create qosprofile qp5
configure dot1p type 4 qosprofile qp5
configure qosprofile qp1 weight 1
configure qosprofile qp5 weight 2
configure qosprofile qp8 use-strict-priority
```

Note

All Extreme Networks *DCB*-capable switches are configured with qp1 and qp8 by default, and some platforms support additional QoS profiles by default. When stacking is used for Summit switches, qp7 is created by default for internal control communications, and is always set to strict priority.

DCBX only advertises the bandwidth for ETS classes, so in the example, the available bandwidth is divided only between qp1 and qp5. The total bandwidth for all ETS classes must add up to 100%, so if the weights don't divide evenly, one or more of the reported bandwidth numbers are rounded to satisfy this requirement. With this in mind, the above configuration results in reported bandwidth guarantees of 33% for TC/PG 0 (qp1) and 67% for TC/PG 4 (qp5).

Weighted round robin scheduling is packet based, so when packets are queued for both classes 0 and 4, the above configuration results in two TC/PG 4 packets being transmitted for each single TC/PG 0 packet. As such, the exact percentages are realized only when the average packet sizes for both classes are the same and the measurement is taken over a long enough period of time. Another consideration is that using the lowest weights possible to achieve the desired ratios results in a more even distribution of packets within a class (that is, less jitter). For example, using weights 1 and 2 are usually preferable to using weights 5 and 10—even though the resulting bandwidth percentages are the same.

Enhanced Transmission Selection allows you to configure QoS scheduling to be weighted-deficit-round-robin. In this approach, you can configure a weight in the range of 1–127 on the QoS profiles. The difference between weighted-round-robin (WRR) and weighted-deficit-round-robin (WDRR) is that, in the latter approach, the algorithm uses a “credit counter” mechanism.

The algorithm works in slightly different ways on different platforms:

**Platform:**

Summit X480, X460, X440 series switches; BlackDiamond 8800 series switches with 8900-G96T-c, 8900-10G24X-c, 8900-MSM128, 8900-G48T-xl, 8900-G48X-xl, and 8900-10G8X-xl modules; E4G-400, E4G-200 cell site routers.

**Methodology:**

- Credit counter—A token bucket that keeps track of bandwidth overuse relative to each queue’s specified weight.
- Weight—Relative bandwidth allocation to be serviced from a queue in each round compared with other queues. Range is between 1 and 127. A weight of 1 equals a unit of 128 bytes.
- MTU Quantum Value—2 Kbytes.

1. Set credit counter to quantum value for all queues.

2. Service queues in round robin order, according to the weight value. When a packet from a queue is sent, the size of the packet is subtracted from the credit counter. A queue is serviced until it is either empty or its credit counter is negative.
3. When all queues are either empty or their credit counter is less than 0, replenish credits by: MTU quantum value x weight of queue. No queue's credit can ever be more than quantum value x weight.

Repeat steps two and three until all queues are empty.

**Platform:**

Summit X670, X460-G2, X670-G2 and X770 series switches; BlackDiamond 8800 series switches with 8900-40G6X-xm module; BlackDiamond X8 series swithes with BDX-MM1, BDXA-FM960, BDXA-FM480, BDXA-40G24X, and BDXA-40G12X modules.

**Methodology:**

- Credit counter—A token bucket used to keep track of bandwidth overuse relative to each queue's specified weight.
- Weight—Relative bandwidth allocation to be serviced from a queue in each round compared with other queues. Range is between 1 and 127.
- K—Minimum value required to make all credit counters positive. This value is recalculated after each round.

1. Set credit counter for each queue to queue's weight value.
2. Service queues in round robin order, according to the weight value. When a packet from a queue is sent, the size of the packet is subtracted from the credit counter. A queue is serviced until it is either empty or its credit counter is negative.
3. When all queues are either empty or their credit counter is less than 0, replenish credits by: 2^K × weight of queue. K is calculated so that it is the minimum value required to make all credit counters positive. No queue's credit can ever be more than 2^K × weight of queue.

Repeat steps two and three until all queues are empty.

**Platform:**

BlackDiamond 8800 series switches with G48Te, G48Te2, G24Xc, G48Xc, G48Tc, 10G4Xc, 10G8Xc, MSM-48, S-G8Xc, S-10G1Xc, and S-10G2Xc modules.

**Methodology:**

These cards have a weight range of 1 to 15. Credit is replenished by 2^(weight – 1) × 10KB.

The number of bytes that can be transmitted in a single round is:

- Weight 0 = Strict Priority
- Weight 1 = 10 KB
- Weight 2 = 20 KB
- Weight 3 = 40 KB
- Weight 4 = 80 KB
- Weight 5 = 160 KB
- Weight 6 = 320 KB
- Weight 7 = 640 KB

- Weight 8 = 1,280 KB
- Weight 9 = 2,560 KB
- Weight 10 = 5,120 KB
- Weight 11 = 10 MB
- Weight 12 = 20 MB
- Weight 13 = 40 MB
- Weight 14 = 80 MB
- Weight 15 = 160 MB

When ETS scheduling is used without a minbw or committed_rate configured, packets from strict priority classes always preempt packets from ETS classes, so the reported percentages reflect the distribution of the bandwidth after strict priority classes use what they need.

Because of this, one might consider limiting the bandwidth for any strict priority classes using the *maxbw* parameter. For example, the following command limits TC/PG 7 to 20% of the interface bandwidth:

```
configure qosprofile qp8 maxbw 20 ports 1-24
```

The per-port bandwidth settings described above can also be used to either limit or guarantee bandwidth for an ETS class.

For example, the following command guarantees 40% of the bandwidth to TC/PG 0:

```
configure qosprofile qp1 minbw 40 ports 1-24
```

The DCBX protocol takes these minimum and maximum bandwidth guarantees into account when calculating the reported bandwidth. With the addition of this minimum bandwidth configuration, the reported bandwidth would change to 40% for class 0 (qp1) and 60% for class 4 (qp5).

The following are some important considerations when using minimum and maximum bandwidth guarantees:

- They change the scheduling dynamic such that a class with a minbw will have priority over other classes (including strict priority classes) until the minbw is met, which differs from the standard ETS scheduling behavior described in 802.1az
- If the minbw is set on multiple classes such that the total is 100%, these classes can starve other classes that do not have a configured minbw. So, for example, if the minbw for both class 0 and class 4 is set to 50% (100% total), traffic from these classes can starve class 7 traffic. This can lead to undesirable results since DCBX and other protocols are transmitted on class 7. In particular, DCBX may report the peer TLV as expired. This effect can be magnified when an egress port shaper is used to limit the egress bandwidth.
- If all ETS classes have a maxbw set, and the total is less than 100%, the total bandwidth reported by DCBX will be less than 100%. Extreme does not report an error in this case, but some DCBX peers may report an error.
- Packet size is a factor in the minimum and maximum bandwidth guarantees.

In light of these considerations, the following are a set of guidelines for using minimum and maximum bandwidth guarantees:

- If minbw guarantees are used for ETS classes, and strict priority classes exist:
  - Make sure that the total minbw reserved is less than 100%.
  - Configure minbw for the strict priority classes.
- If strict priority classes exist, you may want to configure a maxbw for the strict priority classes so they don't starve the ETS classes.
- If maxbw is configured on some ETS classes, ensure that either the total of the maxbw settings for all ETS classes is equal to 100%, or at least one ETS class does not have a maxbw configured.

For more information on the QoS features that support ETS, see QoS.

*Priority-based Flow Control*

Priority flow control (PFC) is defined in the IEEE 802.1Qbb standard as an extension of the IEEE 802.3x flow control standard. When buffer congestion is detected, IEEE 802.3x flow control allows the communicating device to pause all traffic on the port, whereas PFC allows the device to pause just a portion of the traffic and allow other traffic on the same port to continue.

The rest of this section provides general guidelines for configuring the ExtremeXOS PFC feature for *DCB* operation. After you configure PFC, DCBX advertises the PFC compatible configuration to DCBX peers on all DCBX enabled ports.

- PFC configuration is controlled per-port using the following command:

  `enable flow-control [tx-pause {priority priority} | rx-pause {qosprofile qosprofile}] ports [all | port_list]`

  The **rx-pause** option is configured on the *QoS* profile.

  The PFC priority to which a QoS profile responds is fixed and is determined by the QoS profile number such that qpN responds to a PFC frame for priority N-1.

  For example, the following command enables PFC priority 4 for qp5 on ports 1-24:

  ```
  enable flow-control rx-pause qosprofile qp5 ports 1-24
  ```

  After the above command is entered, if a PFC frame is received indicating that priority 4 should be paused, then qp5 will be paused. Note that qp5 is paused regardless of whether the packets mapped to qp5 have priority 4 or other priorities. For example, if we enter the command configure dot1p type 3 qosprofile qp5, priority 3 packets are queued in qp5, and a PFC pause frame for priority 4 pauses priority 3 frames, which might not be desired. For this reason, you should be careful about mapping multiple priorities to the same QoS profile when PFC is enabled for that profile.

The **tx-pause** option is configured on the priority itself. For example, the following command enables the transmittal of PFC Pause frames for priority 4 when frames with priority 4 are congested:

```
enable flow-control tx-pause priority 4 ports 1-24
```

The tx-pause configuration determines what is advertised in the DCBX PFC TLV. In order for PFC to work correctly, it is important to ensure that all switches in the DCB network are receiving and transmitting PFC consistently for each priority on all ports.

In summary, the following three commands ensure that PFC is enabled for priority 4 traffic on ports 1-24:

```
configure dot1p type 4 qosprofile qp5
enable flow-control rx-pause qosprofile qp5 ports 1-24
enable flow-control tx-pause priority 4 ports 1-24
```

For more information on PFC, see IEEE 802.1Qbb Priority Flow Control on page 186.

## Introduction to the XNV Feature

The Extreme Network Virtualization (*XNV*) feature, which is also known as Virtual Machine (VM) tracking, enables the ExtremeXOS software to support VM port movement, port configuration, and inventory on network switches. VM movement and operation on one or more VM servers is managed by a VM Manager (VMM) application. The XNV feature enables a network switch to respond to VM movement and report VM activity to network management software.

VM network access support enables a switch to support VMs as follows:

- Identify a VM by its MAC address and authenticate the VM connection to the network.
- Apply a custom port configuration in response to VM authentication.
- Remove a custom port configuration when a VM *FDB (forwarding database)* entry ages out.
- Detect a VM move between switch ports or switches and configure the old and new ports appropriately.

To support VM mobility, the XNV feature requires that each VM use unique, static MAC and IP addresses. Switch port operation for a VM can be configured with a policy file or an *ACL (Access Control List)*.

### *VM Port Configuration*

An important part of the *XNV* feature is the ability to configure a switch port to support a particular VM. A Virtual Port Profile (VPP) identifies a policy file or *ACL* rule to associate with a VM entry in the authentication database. You can define both ingress and egress policies in VPPs to configure a port separately for each direction. When the VPP is configured for a VM entry and the VM is detected on a port, any associated policy or rule is applied to the port in the specified direction.

The XNV feature supports two types of VPPs, Network VPPs (NVPPs) and Local VPPs (LVPPs).

NVPPs are stored on an FTP server called a repository server. The XNV feature supports file synchronization between XNV-enabled switches and the repository server. One of the advantages of the repository server is centralized storage for NVPPs.

LVPPs must be configured on each switch. LVPPs are a good choice for simple network topologies, but NVPPs offer easier network management for more complex network topologies.

### *VM Authentication Process*

The *XNV* feature supports three methods of authentication:

- NMS server authentication.
- Network authentication using a downloaded authentication database stored in the VMMAP file.
- Local authentication using a local database created with ExtremeXOS CLI commands.

The default VM authentication configuration uses all three methods in the following sequence: NMS server (first choice), network based VMMAP file (second choice), and finally, local database. If a service is not available, the switch tries the next authentication service in the sequence.

#### NMS Server Authentication

If NMS server authentication is enabled and a VM MAC address is detected on a VM-tracking enabled port, the software sends an Access-Request to the configured NMS server for authentication. When the switch receives a response, the switch does one of the following:

- When an Access-Accept packet is received with an NVPP specified, the policies are applied on VM enabled port.
- When an Access-Accept packet is received and no NVPP is specified, the port is authenticated and no policy is applied to the port.
- When an Access-Reject packet is received, the port is unauthenticated and no policy is applied.
- When an Access-Reject packet indicates that the NMS server timed-out or is not reachable, the switch tries to authenticate the VM MAC address based on the next authentication method configured, which can be either network authentication or local authentication.

The Access-Accept packet from the NMS server can include the following Vendor Specific Attributes (VSAs):

- VM name
- VM IP address
- VPP configured for the VM

An Access-Reject packet contains no VSA.

#### Network (VMMAP) Authentication

If network (VMMAP) authentication is enabled and a VM MAC address is detected on a VM-tracking enabled port, the switch uses the VMMAP file to authenticate the VM and applies the appropriate VPP.

#### Local Authentication

If local authentication is enabled and a VM MAC address is detected on a VM-tracking enabled port, the switch uses the local database to authenticate the VM and apply the appropriate VPP.

#### Authentication Failure

If all configured authentication methods fail, *EMS (Event Management System)* messages are logged and no VPP is applied.

Possible remedies include:

- Fix the authentication process that failed. Look for misconfiguration or down segments.
- Configure UPM to take action on the related EMS message.
- If one or two authentication methods are configured, configure additional authentication methods.

#### Duplicate VM MAC Detected

Each VM MAC must be unique. If duplicate MAC addresses are detected on the switch, whether on the same *VLAN (Virtual LAN)* or different VLANs, the switch supports only the last MAC detected.

*File Synchronization*

The *XNV* feature supports file synchronization between XNV-enabled switches and the repository server. The files stored on the repository server include the .map, .vpp, and .pol files. One of the advantages of the repository server is that multiple XNV-enabled switches can use the repository server to collect the network VM configuration files. The XNV feature provides for access to a secondary repository server if the primary repository server is unavailable.

Through file synchronization, the network files are periodically downloaded to the XNV-enabled switches, which allows these switches to continue to support VM authentication when the NMS server is unavailable.

*Network Management and Inventory*

The *XNV* feature is designed to support network management programs such as ExtremeManagement and Ridgeline. The ExtremeXOS software contains *SNMP (Simple Network Management Protocol)* MIBs, which allow network management programs to view VM network configuration data, discover the VM inventory, and make configuration changes. We recommend that you use ExtremeManagement to manage VM network connectivity.

For instructions on managing the XNV feature using the switch CLI, see Managing the XNV Feature, VM Tracking on page 585.

## Introduction to the Direct Attach Feature

The direct attach feature is a port configuration feature that supports VM-to-VM communication on a directly connected server that uses the Virtual Ethernet Port Aggregator (VEPA) feature on that server. Without VEPA and direct attach, a VM server must use a virtual Ethernet bridge or switch on the VM server to enable Ethernet communications between VMs. With VEPA, the VM server can rely on a directly connected switch to receive and reflect VM-to-VM messages between VMs on the same server.

The ExtremeXOS direct attach feature works with VEPA software on a VM server to intelligently forward unicast, flood, and broadcast traffic. Without direct attach, frames are never forwarded back out the same port on which they arrive. With direct attach, frames can be forwarded back out the ingress port, and VEPA software on the VM server ensures that the frames are forwarded appropriately.

For instructions on managing the Direct Attach feature, see Managing Direct Attach to Support VEPA on page 604.

# Managing the DCBX Feature

## Enabling DCBX on Switch Ports

DCBX uses *LLDP* to advertise *DCB* capabilities to DCB peers.

Use the following commands to enable LLDP and the DCBX feature on switch ports:

```
enable lldp ports [all | port_list] {receive-only | transmit-only}

configure lldp ports [all | port_list] [advertise | no-advertise]
vendor-specific dcbx {ieee|baseline}
```

## Configuring DCBX Application Priority Instances

Each DCBX application priority instance maps traffic from one of the supported application types to a TC or PG priority, which selects a specific *QoS* profile for traffic management. Supported application types include:

- Fiber Channel Over Ethernet (FCoE)
- FCoE Initiation Protocol (FIP)
- Internet Small Computer System Interface (iSCSI)

Use the following commands to add or delete DCBX application priority instances:

`configure lldp ports [`**`all`**` | `*`port_list`*`] `**`dcbx add application [name`** *`application_name`* **`| ethertype`** *`ethertype_value`* **`| L4-port`** *`port_number`* **`| tcp-port`** *`port_number`* **`| udp-port`** *`port_number`*`] `**`priority`** *`priority_value`*

`configure lldp ports [`**`all`**` | `*`port_list`*`] `**`dcbx delete application [all-applications | name`** *`application_name`* **`| ethertype`** *`ethertype_value`* **`| L4-port`** *`port_number`* **`| tcp-port`** *`port_number`* **`| udp-port`** *`port_number`*`]`

## Displaying DCBX Configuration and Statistics

Use the following commands to display DCBX feature configuration and statistics:

`show lldp {`**`port`**` [`**`all`**` | `*`port_list`*`]} {`**`detailed`**`}`

`show lldp {`**`port`**` [`**`all`**` | `*`port_list`*`]} `**`dcbx`**` {`**`ieee`**`|`**`baseline`**`} {`**`detailed`**`}`

## DCBX Configuration Example

The following is a sample DCBX configuration:

```
enable lldp ports 1
configure lldp port 1 advertise vendor-specific dcbx ieee
configure lldp port 1 advertise vendor-specific dcbx baseline

enable lldp ports 2
configure lldp port 2 advertise vendor-specific dcbx ieee
configure lldp port 2 advertise vendor-specific dcbx baseline

configure lldp ports 1 dcbx add application name iscsi priority 4
configure lldp ports 1 dcbx add application name fcoe priority 3
configure lldp ports 1 dcbx add application name fip priority 3
configure lldp ports 2 dcbx add application name iscsi priority 4
configure lldp ports 2 dcbx add application name fcoe priority 3
configure lldp ports 2 dcbx add application name fip priority 3
configure lldp ports 2 dcbx add application L4-port 25 priority 4
configure lldp ports 2 dcbx add application tcp-port 4500 priority 4
configure lldp ports 2 dcbx add application udp-port 45 priority 5
configure lldp ports 2 dcbx add application ethertype 2536 priority 4
```

# Managing the XNV Feature, VM Tracking

## Limitations

The following limitations apply to this release of the VM tracking feature:

- When VM tracking is configured on a port, all existing learned MAC addresses are flushed. MAC addresses will be relearned by the switch and the appropriate VPP (if any) for each VM will be applied.
- If a VM changes MAC addresses while moving between ports on a switch, the VM remains authenticated on the original port until the original MAC address ages out of the *FDB*.
- VM counters are cleared when a VM moves between ports on the same switch (because the *ACL*s are deleted and recreated).
- Each VPP entry supports a maximum of eight ingress and four egress ACL or policies.
- For Network VPP, only policy files can be mapped. For Local VPP, either ACL or policy files can be mapped. You cannot map a mixture of both ACL and policy files to a particular VPP.
- ExtremeXOS 15.6 does not support VM Tracking on the Summit X430.

## Managing VM Tracking on the Switch

Use the following steps to manage VM tracking on the switch:

- Issue the following command to enable the VM tracking feature on the switch:

  ```
  enable vm-tracking
  ```

- Issue the following command to disable the VM tracking feature on the switch:

  ```
  disable vm-tracking
  ```

  Note
  When the VM tracking feature is disabled, file synchronization with the repository server stops.

- Issue the following command to view the VM tracking feature configuration and the list of authenticated VMs:

  show vm-tracking

## Managing VM Tracking on Specific Ports

Before you enable the VM tracking feature on specific ports, you must enable VM tracking on the switch, configure the authentication method and sequence, and the VM authentication databases.

- When this configuration is complete, you can use the following command to enable VM tracking on one or more ports:

  ```
  enable vm-tracking ports port_list
  ```

- To disable the VM tracking feature on a group of ports, use the following command:

  ```
  disable vm-tracking ports port_list
  ```

- To view the VM tracking feature configuration on one or more ports, use the following command:

  ```
  show vm-tracking port port_list
  ```

## Configuring the Authentication Method and Sequence

You can configure VM authentication through the following services:

- NMS server
- Network based VMMAP file
- Local database

The default VM authentication configuration uses all three methods in the following sequence: NMS server (first choice), network based VMMAP file (second choice), and finally, local database. If a service is not available, the switch tries the next authentication service in the sequence.

To configure one or more authentication methods and a preferred sequence, use the following command:

```
configure vm-tracking authentication database-order [[nms] | [vm-map] |
[local] | [nms local] | [local nms] | [nms vm-map] | [vm-maplocal] |
[local vm-map] | [nms vm-map local] | [localnmsvm-map]]
```

## XNV and MLAG

Starting with EXOS 15.7 as part of ExtremeManagement NAC integration, as long as *MLAG (Multi-switch Link Aggregation Group)* peers have ISC connectivity, only one of the MLAG peers authenticates a VM that is learned on an MLAG port.

- When ISC connectivity between the MLAG peers is established, the peer with the highest IP address is chosen to be the authenticator. This peer will authenticate a VM based on the chosen authentication method.
- The result of the authentication is checkpointed by the authenticator to its peer so that the same VPP gets applied to the VM on both peers.
- When the MLAG peer that is the authenticator goes down, the other peer detects that the authenticator is down and re-authenticates the VM at the next authentication interval. Note that the peer that takes over as the authenticator does not re-authenticate the VMs immediately but waits for the re-authentication timer to expire.
- VMs learned on non-MLAG ports are authenticated by the detecting peer.
- All authentication-related configurations like *RADIUS (Remote Authentication Dial In User Service)* address, repository for VMMAP, local DB, etc. must be identical on both peers. This is an existing requirement and there is no change to this requirement.

## XNV Dynamic VLAN

Starting in release 15.3, when a virtual machine is detected, ExtremeXOS dynamically creates the *VLAN* that is required for the VM to send traffic. If a virtual machine shuts down or is moved, its VLAN is pruned to preserve bandwidth. This feature creates an adaptive infrastructure in which the network responds to changes dynamically in the virtual machine network.

### *Enabling/Disabling XNV Dynamic VLAN*

Enabling the *XNV* dynamic VLAN feature must be done on a per-port basis. XNV requires that the port on which dynamic VLANs is enabled is part of the "default or "base" VLAN as untagged. This "default", or "base", VLAN for the port is the VLAN on which untagged packets are classified to when no VLAN

configuration is available for the MAC. This default VLAN should be present, and you should manually add the port to this VLAN before you enable the feature. Enabling this feature on a port results in a failure if any of the following conditions are true:

- If XNV is not enabled, the command only results in a warning, and does not fail. XNV can be enabled later.
- The port is not an untagged member of any VLAN.

When a VLAN's MAC is detected on a port, XNV consults the configuration database to determine the VLAN configuration for the VM. For a case where the VM sends tagged traffic, the VLAN tag of the received frame is used to determine VLAN classification for the VM's traffic. If VLAN configuration exists for the VM and it conflicts with the actual tag present in received traffic, XNV reports an *EMS* message and does not trigger VLAN creation or port addition. However, if no configuration is present for the VM, XNV assumes that there are no restrictions for classifying traffic for the VM to the received VLAN.

For untagged traffic, XNV can determine the VLAN for the VM from any one of the three possible sources:

- VLAN configuration for the VM MAC entry.
- VLAN configuration for the VPP associated with the VM's MAC. The VPP can either be a network VPP or a local VPP.
- In case of untagged traffic from the VM, the "default" VLAN for the port that is specified as part of the dynamic VLAN enable configuration.

This list determines the order of precedence for VLAN classification for untagged traffic only. For tagged VLAN traffic, XNV validates the tag of the received traffic with then VLAN tag configuration for that VM.

In addition to the VLAN tag, you can specify the VR to which the dynamically created VLAN needs to be associated. The VR configuration is relevant only if a VLAN tag is configured for the VM.

**Table 68: Associating Dynamically Created VLANS to VRs**

| Configured VR on Port | Configured VR for VM (from VM Mapping Entry or VPP) | VLAN Already Exists on the Switch | Dynamic VLANs VR |
|---|---|---|---|
| None | None | No | *VR-Default* |
| None | None | Yes | VLAN's VR |
| None | VR-X | No | VR-X (Configured VR for VM) if VR-X is valid.) Otherwise an EMS error is displayed indicating the VR-X is invalid. |
| None | VR-X | Yes | VLAN's VR. An EMS error is displayed if the VLAN's VR is not VR-X. |
| VR-X | None | No | VR-X (Port's VR). |

**Table 68: Associating Dynamically Created VLANS to VRs (continued)**

| Configured VR on Port | Configured VR for VM (from VM Mapping Entry or VPP) | VLAN Already Exists on the Switch | Dynamic VLANs VR |
|---|---|---|---|
| VR-X | None | Yes | VR-X if VLAN's VR is VR-X. If it is not, an EMS error is displayed indicating the VR-X is invalid. |
| VR-X | VR-Y | No | Dynamic VLAN is not created when Port Level VR and VM-MAC VR are different, and *FDB* is learned on a system generated VMAN. An EMS warning is generated on the switch log, because a Dynamic VLAN cannot be created. |
| VR-X | VR-Y | Yes | VR-X if VLAN is part of VR-X. Otherwise, EMS error is displayed. |

When you disable dynamic VLAN on a port, XNV does the following:

- Triggers deletion of MAC-based entries on that port in the hardware.
- If the port has been added to any VLAN by XNV, XNV triggers a flush for those VLANs.
- If the port has been added to an VLAN by XNV, XNV requests VLAN manager to remove the port from the VLAN.

**Note**
It is up to the VLAN manager to decide if the port actually needs to be removed from the VLAN.

On deleting the ports from base/default VLAN the below warning message will be thrown and XNV Dynamic vlan gets disabled on that port:

```
 Warning: Removing the untagged VLAN from a port may disrupt network connectivity. IDM and
VMT may not be functional on the port without an untagged VLAN.
```

**Note**
This behavior is in effect from ExtremeXOS 16.1.

*Example*

```
create vlan v1
con v1 add ports 1untagged
enable vm-tracking
enable vm-tracking ports 1
enable vm-tracking dynamic-vlan ports 1
con vlan v1 delete ports 1
Warning: Removing the untagged VLAN from a port may disrupt network connectivity. IDM
```

```
and VMT may not be functional on the port without an untagged VLAN.

show vm-tracking
-------------------------------------------------------------
VM Tracking Global Configuration
-------------------------------------------------------------
VM Tracking                      : Enabled
VM  Tracking authentication order: nms vm-map local
VM Tracking nms reauth  period   : 0 (Re-authentication disabled)
VM Tracking blackhole policy       : none
-------------------------------------------------------------

Port                         : 1
VM Tracking                  : Enabled
VM Tracking Dynamic VLAN     : Disabled
```

When XNV is disabled on a port, the XNV dynamic VLAN feature is also disabled. The XNV dynamic VLAN configuration is not persistent, and needs to be re-enabled after XNV is re-enabled on that port.

### *Tracking XNV Per VM Statistics*

Beginning in release 15.3, each local and network VPP has the option to specify whether a counter needs to be installed to count traffic matching the virtual machine MAC which gets the VPP mapping. You can choose to install a counter to collect statistics for ingress traffic only, egress traffic only, or traffic in both directions.

Once the ingress counter installation option is selected for a specific local or network VPP and the virtual machine which has this VPP mapping is detected on the switch, the counter is installed with the name "xnv_ing_dyn_cnt_vmxxxxxxxxxxxx" for the port on which the VM MAC is detected. In this case, xxxxxxxxxxxx denotes the virtual machine MAC for which the counter is installed. In the same way, the egress counter is installed using the name "xnv_egr_dyn_cnt_vmxxxxxxxxxxxx" for that port.

You can view a list of packet/byte counts for this counter name using the command `show access dynamic-counter`. The counter is uninstalled only when the virtual machine MAC is deleted on the switch or the VPP is mapped to a virtual machine MAC which has the counter option set to none. If the VM MAC move happens then the counter installed on the previous port is uninstalled and the counter is installed on the new port. The counter values are not maintained during the MAC move.

## Managing the Repository Server

### *Selecting the Repository Server Directory*

All files for NMS and network authentication must be placed in the configured repository server directory. These files include the following:

- MANIFEST
- VMMAP
- NVPP policy files

By default, the *XNV* feature tries to access the FTP server with anonymous login and fetch the files from the pub directory within the FTP server root directory.

To configure a different directory for repository server files, use the following command:

```
configure vm-tracking repository
```

### *Creating the MANIFEST File*

The MANIFEST file identifies the VMMAP, NVPP, and policy files that are to be used for either NMS or network authentication. The MANIFEST file is downloaded to the switch at the specified refresh interval. Each time the MANIFEST file is downloaded, the switch scans the file and compares the file entries and timestamps to those files on the switch. If the switch detects newer files, it downloads those files to the switch.

You can create the MANIFEST file with a text editor. The MANIFEST file must be placed on the repository server as described in Selecting the Repository Server Directory on page 589.

The format of MANIFEST files is:

```
File1 yyyy-mm-dd hh:mm:ss
File2 yyyy-mm-dd hh:mm:ss
```

Because the definition for each file in the MANIFEST includes a date and time, you must update the MANIFEST file every time you update the VMMAP file or a policy file.

The following is a sample MANIFEST file:

```
a1.map 2010-07-07 18:57:00
a1.vpp 2010-07-07 18:57:00
a2.map 2010-07-07 18:57:00
a2.vpp 2010-07-07 18:57:00
policy1.pol 2010-07-07 18:57:00
epolicy1.pol 2010-07-07 18:57:00
```

The file extensions for the files in the MANIFEST file identify the supported file types:

- .map—VMMAP files
- .vpp—VPP files
- .pol—Policy files

### *Creating a VMMAP File*

Use a text editor to create a VMMAP file. VMMAP file entries must use the following XML format:

```
<VMLIST>
    <VM>
        <MAC>00:00:00:00:00:21</MAC>
        <NAME>network_vm1</NAME>
        <IPV4>10.10.10.10</IPV4>
        <VPP>nvpp1</VPP>
    </VM>
    <VM>
        <MAC>00:00:00:00:00:22</MAC>
        <NAME>network_vm2</NAME>
        <IPV4>20.20.20.20</IPV4>
        <VPP>nvpp2</VPP>
    </VM>
</VMLIST>
```

When creating VMMAP file entries, use the following guidelines:

- The VPP file supports up to 400 child nodes.
- The MAC address is required.
- If you do not want to specify a VM name, specify none.
- If you do not want to specify an IP address, specify 0.0.0.0.
- If you do not want to specify a VPP name, specify none.
- If a value such as the VM name contains any space characters, the entire value must be specified between double quotation marks ( " ).

For information on where to place the VMMAP file, see Selecting the Repository Server Directory on page 589.

*Creating VPP Files*

Use a text editor to create a VPP file. VPP file entries must use the following XML format:

```
<vppList>
  <vpp>
    <name>nvpp1</name>
      <last-updated>2002-05-30T09:00:00</last-updated>
    <policy>
      <name>policy1</name>
      <direction>ingress</direction>
      <order>1</order>
    </policy>
    <policy>
      <name>policy4</name>
      <direction>ingress</direction>
      <order>4</order>
    </policy>
    <policy>
      <name>epolicy1</name>
      <direction>egress</direction>
      <order>1</order>
    </policy>
    <policy>
      <name>epolicy4</name>
      <direction>egress</direction>
      <order>4</order>
    </policy>
  </vpp>
</vppList>
```

The VPP file supports up to 400 child nodes, and each VPP entry supports up to eight ingress and four egress *ACL* or policies. If multiple policies are defined within a VPP entry for either ingress or egress, the switch uses the policy with the lowest order number. If two ingress or egress policies have the same order number, the switch selects the policy based on which name is lexicographically lower.

To refresh all policies which are all associated and applied to each VPP, use the following command:

`refresh policy` *`policy-name`*

The NVPP policy files must be placed on the repository server as described in Selecting the Repository Server Directory on page 589.

### *Creating Policy Files*

For instructions on creating policy files, see Policy Manager on page 635.

To display the policy file or *ACL* associated with one or all VPPs, use the following command:

```
show vm-tracking vpp {vpp_name}
```

### *Managing Switch Access to the Repository Server*

- To enable and configure file synchronization between an *XNV*-enabled switch and a repository server, use the following command:

  ```
  configure vm-tracking repository [primary | secondary] server
  [ipaddress | hostname] {vr vr_name} {refresh-interval seconds} {path-
  name path_name} {user user_name {encrypted} password}
  ```

- To force file synchronization with the repository server, use the following command:

  ```
  run vm-tracking repository sync-now
  ```

- To remove the configuration for one or both repository servers, use the following command:

  ```
  unconfigure vm-tracking repository {primary | secondary}
  ```

- To display the repository server configuration and status, use the following command:

  ```
  show vm-tracking repository {primary | secondary}
  ```

## Manage NMS Server Authentication

NMS server authentication uses the *RADIUS* protocol to authenticate VM access to the network with the RADIUS server included with ExtremeManagement and Ridgeline. These products are designed to perform VM network management tasks, such as creating and associating NVPPs with VM authentication entries.

To use NMS authentication, you must do the following:

- Select NMS authentication as described in Configuring the Authentication Method and Sequence on page 586.
- Prepare the network repository server as described in Managing the Repository Server on page 589.
- Configure the NMS client software in the switch as described in Configure the NMS Client Software on page 593.
- Configure the NMS server as described in Configuring the NMS Server Software on page 592.

You can display NMS authenticated VMs as described in Displaying NMS Authenticated VMs on page 593.

### *Configuring the NMS Server Software*

The ExtremeManagement and Ridgeline include a *RADIUS* server that you can use for NMS server authentication. To configure this server, do the following:

1. Add the IP address of each *XNV*-enabled switch as a RADIUS client.
2. Add each VM MAC address as a username (in upper case and should not contain semicolon) and add the MAC address as the password.

3. Add a remote access policy with the Extreme Networks VSAs:
   - Vendor code: 1916
   - VSA ID: 213 (EXTREME_VM_NAME)

     Example: MyVM1
   - VSA ID: 214 (EXTREME_VM_VPP_NAME)

     Example: nvpp1
   - VSA ID: 215 (EXTREME_VM_IP_ADDR)

     Example: 11.1.1.254

For instructions on configuring the Ridgeline RADIUS server, refer to the Ridgeline documentation.

*Configure the NMS Client Software*

- The switch uses NMS client software to connect to an NMS server for VM authentication. Use the following commands to configure the NMS client software in the switch:

  `configure vm-tracking nms [primary | secondary] server [ipaddress | hostname] {udp_port} client-ip client_ip shared-secret {encrypted} secret {vr vr_name}`

  `configure vm-tracking nms timeout seconds`

  `configure vm-tracking timers reauth-period reauth_period`
- To remove the NMs client configuration for one or both NMS servers, use the following command:

  `unconfigure vm-tracking nms {server [primary | secondary]}`
- To display the NMS client configuration, use the following command:

  `show vm-tracking nms server {primary | secondary}`

*Displaying NMS Authenticated VMs*

To display the VMs and corresponding policies in the NMS authentication database, use the following command:

`show vm-tracking network-vm`

## Managing Network Authentication (Using the VMMAP File)

To use network authentication, you must do the following:

1. Select network authentication as described in Configuring the Authentication Method and Sequence on page 586.
2. Prepare the network repository server as described in Managing the Repository Server on page 589.

To display the VMs and corresponding policies in the network authentication database, use the following command:

`show vm-tracking network-vm`

## Manage Local Database Authentication

To use local database authentication, you must do the following:

1. Select local database authentication as described in Configuring the Authentication Method and Sequence on page 586.
2. Create and manage local VPPs (LVPPs) as described in Managing the Local VPP Database on page 594.
3. Create VM entries as described in Managing VM Entries in the Local Authentication Database on page 594.

### *Managing the Local VPP Database*

Only one dynamic *ACL* or policy can be added to a VPP. Ingress LVPPs apply to traffic flowing from the VM, into the switch port, and then to the client. Egress LVPPs apply to traffic flowing from the client, out the switch port, and to the VM.

For instructions on creating policy files, see Policy Manager on page 635. For instructions on creating dynamic ACLs, see ACLs on page 640.

- To create and configure entries in the LVPP database, use the following commands:

  `create vm-tracking vpp` *`vpp_name`*

  `configure vm-tracking vpp` *`vpp_name`* **`add`** `[`**`ingress`** `|` **`egress`**`]` `[`**`policy`** *`policy_name`* `|` **`dynamic-rule`** *`rule_name`*`]` `{`**`policy-order`** *`policy_order`*`}`
- To delete or unconfigure entries in the local VPP database, use the following commands:

  `delete vm-tracking vpp {`*`vpp_name`*`}`

  `unconfigure vm-tracking vpp` *`vpp_name`*
- To display the policy file or ACL associated with one or more VPPs , use the following command:

  `show vm-tracking vpp {`*`vpp_name`*`}`

### *Managing VM Entries in the Local Authentication Database*

- To create and configure entries in the local authentication database, use the following commands:

  `create vm-tracking local-vm` **`mac-address`** *`mac`* `{`**`name`** *`name`* `|` **`ipaddress`** *`ipaddress`* **`vpp`** *`vpp_name`* `}`

  `configure vm-tracking local-vm` **`mac-address`** *`mac`* `[`**`name`** *`name`* `|` **`ip-address`** *`ipaddress`* `|` **`vpp`** *`vpp_name`*`]`
- To remove a configuration parameter for a local authentication database entry, or to remove an entry, use the following commands:

  `unconfigure vm-tracking local-vm` **`mac-address`** *`mac`* `[`**`name`** `|` **`ip-address`** `|` **`vpp`**`]`

  `delete vm-tracking local-vm {`**`mac-address`** *`mac`*`}`
- To display the local VPP database entries, use the following command:

  `show vm-tracking local-vm {`**`mac-address`** *`mac`*`}`

## Example XNV Configuration

The following figure displays a sample *XNV* topology that will be used for the examples in the following sections:

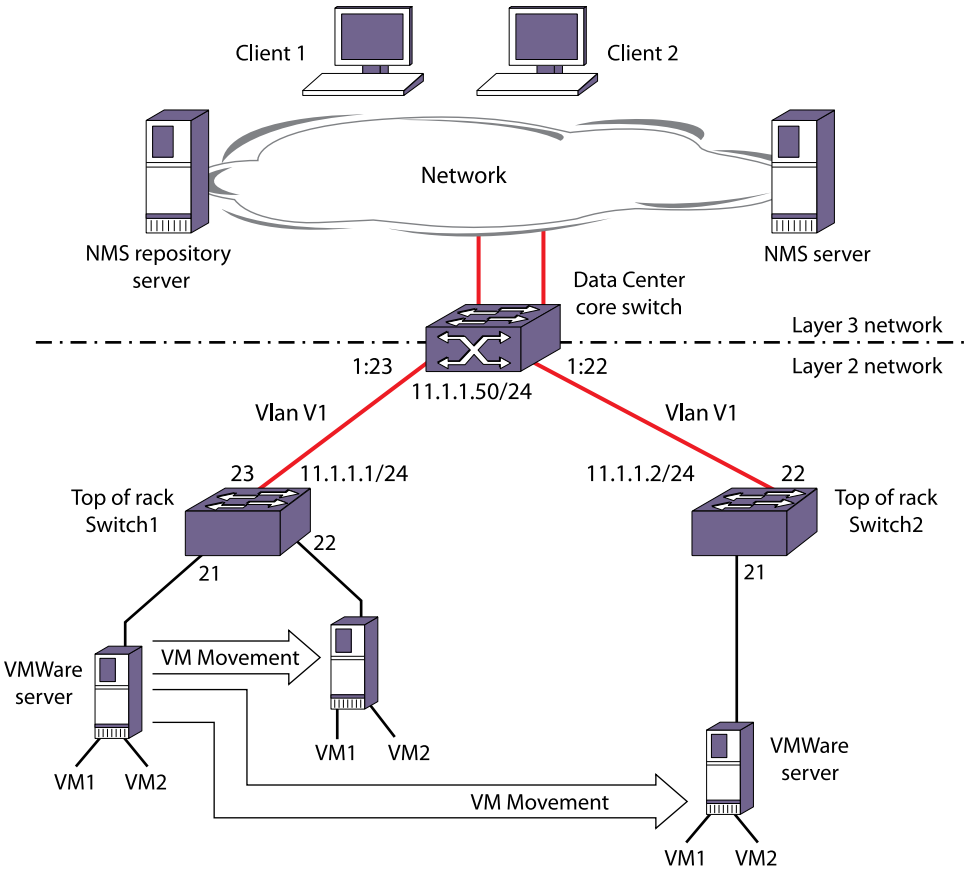



**Figure 90: Sample XNV Topology**

The example configuration supports the following:

- VM authentication using NMS server, network, or local authentication
- Ingress and egress port configuration for each VM
- VM movement from one switch port to another
- VM movement from one switch to another

**Note**

Ingress ACLs or policies apply to traffic flowing from the VM, into the switch port, and then to the client. Egress ACLs apply to traffic flowing from the client, out the switch port, and to the VM.

### *MAC and IP Addresses*

The following are the MAC and IP addresses for the example topology:

```
VM1 MAC address: 00:04:96:27:C8:23
VM2 MAC address: 00:04:96:27:C8:24
VM1 IP address: 11.1.1.101
VM2 IP address: 11.1.1.102
Client1 MAC address: 00:04:96:00:00:01
Client2 MAC address: 00:04:97:00:00:02
Repository server IP address: 10.127.8.1
NMS server IP address: 10.127.5.221
```

### *General VLAN Configuration*

The following is the core switch *VLAN* configuration:

```
create vlan v1
configure vlan v1 tag 100
configure vlan v1 add ports 1:22,1:23 tagged
configure vlan v1 ipaddress 11.1.1.50/24
```

The following is the Switch1 VLAN configuration:

```
create vlan v1
configure vlan v1 tag 100
configure vlan v1 add ports 21,22, 23 tagged
configure vlan v1 ipaddress 11.1.1.1/24
```

The following is the Switch2 VLAN configuration:

```
create vlan v1
configure vlan v1 tag 100
configure vlan v1 add ports 21,22 tagged
configure vlan v1 ipaddress 11.1.1.2/24
```

**Note**

For NMS server and network authentication, the NMS server and repository server must be accessible to all *XNV*-enabled switches through *VR-Mgmt*.

### *VMWare Server Setup*

The VMWare servers must be connected to Switch1 and Switch2 and should have dual quad-core processors. The VMWare servers require the following software:

- VMWare server: ESXi license
- Vsphere EXSI client
- V-Center client

Each physical VMWare server should be configured with two VMs. Use the V-Center client to trigger Vmotion.

### *Repository Server Setup*

The repository server setup for this topology is the same for NMS server authentication and network authentication. The following shows the FTP server setup:

```
FTP login: anonymous
Password: "" (no password)
Repository directory path: pub
[root@linux pub]# pwd
/var/ftp/pub
```

The following is an example MANIFEST file:

```
vm.map 2011-05-11 18:57:00
vpp.vpp 2011-05-11 18:57:00
nvpp1.pol 2011-05-11 18:57:00
nevpp1.pol 2011-05-11 18:57:00
nvpp2.pol 2011-05-11 18:57:00
nevpp2.pol 2011-05-11 18:57:00
```

The following is an example VMMAP file named vm.map:

```
<VMLIST>
  <VM>
    <MAC>00:04:96:27:C8:23</MAC>
    <NAME>vm_1</NAME>
    <IPV4>11.1.1.101</IPV4>
    <VPP>nvpp1</VPP>
    <CTag>1000</CTag>
    <VRName>Vr-Default</VRName>
  </VM>
  <VM>
    <MAC>00:04:96:27:C8:24</MAC>
    <NAME>vm_2</NAME>
    <IPV4>11.1.1.102</IPV4>
    <VPP>nvpp2</VPP>
  </VM>
</VMLIST>
```

The following is an example VPP file named vpp.vpp:

```
<vppList>
  <vpp>
    <name>nvpp1</name>
      <last-updated>2011-05-30T09:00:00</last-updated>
    <policy>
      <name>nvpp1.pol</name>
      <direction>ingress</direction>
      <order>1</order>
    </policy>
    <policy>
      <name>nevpp1.pol</name>
      <direction>egress</direction>
      <order>1</order>
      <CTag>1000</CTag>
      <VRName>Vr-Default</VRName>
    </policy>
  </vpp>
  <vpp>
    <name>nvpp2</name>
      <last-updated>2011-05-30T09:00:00</last-updated>
    <policy>
      <name>nvpp2.pol</name>
      <direction>ingress</direction>
      <order>1</order>
      </policy>
    <policy>
      <name>nevpp2.pol</name>
      <direction>egress</direction>
      <order>1</order>
    </policy>
  </vpp>
</vppList>
```

The following is the nvpp1.pol file:

```
entry nvpp1 {
if match all {
    ethernet-destination-address 00:04:96:00:00:00 / ff:ff:ff:00:00:00 ;
} then {
  deny ;
```

```
  count host1
} }
```

The following is the nvpp2.pol file:

```
entry nvpp2 {
if match all {
    ethernet-destination-address 00:04:97:00:00:00 / ff:ff:ff:00:00:00 ;
} then {
  deny ;
  count host2
} }
```

The following is the nevpp1.pol file:

```
entry nevpp1 {
if match all {
    ethernet-source-address 00:04:96:00:00:00 / ff:ff:ff:00:00:00 ;
} then {
  deny ;
  count h1
} }
```

The following is the nevpp2.pol file:

```
entry nevpp2 {
if match all {
  ethernet-source-address 00:04:97:00:00:00 / ff:ff:ff:00:00:00 ;
} then {
  deny ;
  count h2
} }
```

*Example ACL Rules*

The following are some example *ACL* rules:

```
entry etherType1 {
    if {
      ethernet-source-address 00:a1:f1:00:00:01;
      }
    then {
      permit;
      count etherType1;
      }
    }
entry denyall {
    if {
        source-address 10.21.1.1/32;
       }
    then {
          deny;
         }
    }
entry allowall {
    if {
        source-address 11.1.1.1/32;
        source-address 12.1.0.0/16;
       }
    then {
```

```
        allow;
       }
    }
entry destIp {
    if {
        destination-address 192.20.1.0/24;
        protocol UDP;
       }
    then {
        deny;
      count destIp;
       }
    }
entry denyAll {
    if {
       }
    then {
          deny;
          count denyAll;
         }
    }
```

*General Switch XNV Feature Configuration*

The following configuration enables the *XNV* feature on the switch and the specified ports:

```
enable vm-tracking
enable vm-tracking ports 21-22
```

*Local VM Authentication Configuration*

If you only want to use local authentication, configure the *XNV*-enabled switches as follows:

```
configure vm-tracking authentication database-order local
```

To enable dynamic *VLAN*, issue the following command:

```
enable vm-tracking dynamic-vlan ports 19
```

To add Uplinkports to Dynamic VLAN:

```
configure vlan dynamic-vlan uplink-ports add ports port_no
```

To delete the uplink port:

```
configure vlan dynamic-vlan uplink-ports delete ports port_no
```

The following is the policy1.pol file for Port 21 in the ingress direction:

```
entry nvpp1 {
if match all {
ethernet-destination-address 00:04:96:00:00:00 / ff:ff:ff:00:00:00 ;
} then {
deny ;
count host1
} }
```

The following is the policy2.pol file for Port 21 in the egress direction:

```
entry nevpp1 {
if match all {
ethernet-source-address 00:04:96:00:00:00 / ff:ff:ff:00:00:00 ;
```

```
} then {
deny ;
count h1
} }
```

The following commands configure VM authentication in the local database:

```
create vm-tracking local-vm mac-address 00:04:96:27:C8:23
configure vm-tracking local-vm mac-address 00:04:96:27:C8:23 ip-address 11.1.1.101
configure vm-tracking local-vm mac-address 00:04:96:27:C8:23 name myVm1
create vm-tracking vpp vpp1
configure vm-tracking vpp vpp1 add ingress policy policy1
configure vm-tracking vpp vpp1 add egress policy policy2
configure vm-tracking local-vm mac-address 00:04:96:27:C8:23 vpp vpp1
```

The following commands used to create VM-mac with vlan-tag, and Vr for Dynamic vlan creation:

```
create vm-tracking local-vm mac-address 00:00:00:00:00:01
configure vm-tracking local-vm mac-address 00:00:00:00:00:01 vpp lvpp1
configure vm-tracking local-vm mac-address 00:00:00:00:00:01 vlan-tag 1000 vr VR-Default
configure vm-tracking vpp lvpp1 vlan-tag 2000
```

The following commands display the switch XNV feature status after configuration:

```
* Switch.67 # show vm-tracking local-vm

MAC Address        IP Address        Type     Value
-------------------------------------------------------------------------------
00:00:00:00:00:01                    VM
                                     VPP      lvpp1
                                     VLAN Tag 1000
                                     VR Name  VR-Default
Number of Local VMs: 1
* Switch.69 # show vm-tracking vpp

VPP Name                          Type       Value
-------------------------------------------------------------------------------
lvpp1                             origin     local
                                  counters   none
                                  VLAN Tag   2000
                                  VR Name Vr-Default
ingress             policy1
egress              policy2
Number of Local VPPs : 1
Number of Network VPPs: 0
Switch.71 # show vm-tracking
-------------------------------------------------------------
   VM Tracking Global Configuration
-------------------------------------------------------------
VM Tracking                     : Enabled
VM Tracking authentication order: nms vm-map local
VM Tracking nms reauth period   : 0 (Re-authentication disabled)
VM Tracking blackhole policy    : none
-------------------------------------------------------------

Port                         : 19
VM Tracking                  : Enabled
VM Tracking Dynamic VLAN     : Enabled

                   Flags
MAC                APC   IP Address      Type     Value
-------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------

Flags :
  (A)uthenticated     : L - Local, N - NMS, V - VMMAP
  (P)olicy Applied    : B - All Ingress and Egress, E - All Egress, I - All Ingress
  (C)ounter Installed : B - Both Ingress and Egress, E - Egress Only, I - Ingress Only

Type :
  IEP - Ingress Error Policies
  EEP - Egress Error Policies

Number of Network VMs Authenticated: 0
Number of Local VMs Authenticated  : 0
Number of VMs Authenticated        : 0
Switch.73 # show policy
Policies at Policy Server:
PolicyName                   ClientUsage     Client          BindCount
--------------------------------------------------------------------------------
policy1                          1               acl             1
policy2                          1               acl             1
```

### *Network (VMMAP) Authentication Configuration*

If you only want to use network authentication, configure the *XNV*-enabled switches as follows:

```
configure vm-tracking authentication database-order vm-map
```

After the repository server is configured (see Repository Server Setup on page 596), the following commands can be used to display the switch XNV feature status:

```
* Switch.32 # show vm-tracking repository
---------------------------------
VMMAP FTP Server Information
---------------------------------
Primary VMMAP FTP Server  :
Server name:
IP Address      : 10.127.8.1
VR Name         : VR-Mgmt
Path Name       : /pub (default)
User Name       : anonymous (default)
Secondary VMMAP FTP Server : Unconfigured
Last sync           : 16:56:11          Last sync server : Primary
Last sync status    : Successful
* Switch.69 # show vm-tracking vpp
VPP Name                    Type               Name
--------------------------------------------------------------------------------
nvpp1                       origin             network
ingress             nvpp1
egress              nevpp1
nvpp2                       origin             network
ingress             nvpp2
egress              nevpp2
Number of Local VPPs : 0
Number of Network VPPs: 2
* Switch.15 # show vm-tracking
------------------------------------------------------------
VM Tracking Global Configuration
------------------------------------------------------------
VM Tracking                     : Enabled
VM Tracking authentication order: vm-map
VM Tracking nms reauth period   : 0 (Re-authentication disabled)
VM Tracking blackhole policy    : none
```

```
-------------------------------------------------------------
Port                : 21
VM TRACKING         : ENABLED
Flags
MAC                AP    IP Address      Type     Name
-------------------------------------------------------------------------------
00:04:96:27:c8:23  VB    11.1.1.101      VM       vm_1
VPP      nvpp1
00:04:96:27:c8:24  VB    11.1.1.102      VM       vm_2
VPP      nvpp2
-------------------------------------------------------------------------------
Flags :
(A)uthenticated  : L - Local, N - NMS, V - VMMAP
(P)olicy Applied : B - Both, E - Egress, I - Ingress
Number of Network VMs Authenticated   : 2
Number of Local VMs Authenticated     : 0
Number of VMs Authenticated           : 2
* Switch.16 # show vm-tracking network-vm
MAC Address        IP Address        Type     Name
-------------------------------------------------------------------------------
00:04:96:27:c8:23  11.1.1.101        VM       vm_1
VPP      nvpp1
00:04:96:27:c8:23  11.1.1.102        VM       vm_2
VPP      nvpp2
Number of Network VMs: 2
* Switch.16 # show policy
Policies at Policy Server:
PolicyName                    ClientUsage     Client          BindCount
-------------------------------------------------------------------------
vmt/nvpp1                          1          acl             1
vmt/nvpp2                          1          acl             1
vmt/nevpp1                         1          acl             1
vmt/nevpp2                         1          acl             1

show vm-tracking nms server
VM Tracking NMS : enabled
VM Tracking NMS server connect time out: 3 seconds

Primary VM Tracking NMS server:
  Server name :
  IP address : 10.127.6.202
  Server IP Port: 1812
  Client address: 10.127.11.101 (VR-Mgmt)
  Shared secret : qijxou
Access Requests : 0 Access Accepts : 0
Access Rejects : 0 Access Challenges : 0
Access Retransmits: 0 Client timeouts : 0
Bad authenticators: 0 Unknown types : 0
Round Trip Time : 0
```

*NMS Server Authentication Configuration*

- If you only want to use NMS server authentication, configure the *XNV*-enabled switches as follows:
  `configure vm-tracking authentication database-order nms`
- Configure the NMS server as follows:
  a. Add Switch1 and Switch2 as *RADIUS* clients.
  b. Add the MAC addresses for VM1 and VM2 as users, and configure the passwords to match the user names.

c. Add a remote access policy with the Extreme Networks VSAs:
   - Vendor code: 1916
   - VSA ID: 213 (EXTREME_VM_NAME)

     Example: MyVM1
   - VSA ID: 214 (EXTREME_VM_VPP_NAME)

     Example: nvpp1
   - VSA ID: 215 (EXTREME_VM_IP_ADDR)

     Example: 11.1.1.254

**Note**
For the Dynamic *VLAN* feature, the following VSAs are used:

EXTREME_VM_VLAN_ID with VSA ID as 216
EXTREME_VM_VR_NAME with VSA ID as 217

- The following command configures the switch as an NMS server client:

```
configure vm-tracking nms primary server 10.127.5.221 client-ip 10.127.8.12 shared-
secret secret
```

After the repository server is configured (see Repository Server Setup on page 596), the following commands can be used to display the switch XNV feature status:

```
 * Switch.33 # show vm-tracking nms server
VM Tracking NMS : enabled
VM Tracking NMS server connect time out: 3 seconds
Primary VM Tracking NMS server:
 Server name   :
 IP address    : 10.127.5.221
 Server IP Port: 1812
 Client address: 10.127.8.12 (VR-Mgmt)
 Shared secret : qijxou
 Access Requests   : 7              Access Accepts : 2
 Access Rejects    : 5              Access Challenges : 0
 Access Retransmits: 0              Client timeouts : 0
 Bad authenticators: 0              Unknown types : 0
 Round Trip Time   : 0
 * Switch.32 # show vm-tracking
-------------------------------------------------------------
VM Tracking Global Configuration
-------------------------------------------------------------
 VM Tracking                     : Enabled
 VM Tracking authentication order: nms
 VM Tracking nms reauth period   : 0 (Re-authentication disabled)
 VM Tracking blackhole policy    : none
-------------------------------------------------------------
 Port                        : 21
 VM TRACKING                 : ENABLED
 Flags
 MAC               AP    IP Address      Type     Name
-------------------------------------------------------------------------------
 00:04:96:27:c8:23  VB    11.1.1.101      VM       vm_1
 VPP      nvpp1
 00:04:96:27:c8:24  VB    11.1.1.102      VM       vm_2
 VPP      nvpp2
-------------------------------------------------------------------------------
```

```
 Flags :
 (A)uthenticated  : L - Local, N - NMS, V - VMMAP
 (P)olicy Applied : B - Both, E - Egress, I - Ingress
 Number of Network VMs Authenticated: 1
 Number of Local VMs Authenticated : 0
 Number of VMs Authenticated : 1

 * Switch.32 # show policy
Policies at Policy Server:
PolicyName ClientUsage Client BindCount
--------------------------------------------------------------------------
vmt/nvpp1 1 acl 1
vmt/nvpp2 1 acl 1
--------------------------------------------------------------------------
```

# Managing Direct Attach to Support VEPA

You should only enable the Direct Attach feature on ports that directly connect to a VM server running VEPA software.

- To enable or disable the direct attach feature on a port, enter the command:

  `configure port` *port* **`reflective-relay`** [**`on`** | **`off`**]

- To see if the direct attach feature (reflective-relay) is enabled on a switch port, enter the command:

  `show ports information detail`

Note

When the Direct Attach feature is configured on a port, the port number cannot be included in the port list for a static *FDB* entry. For example, the Direct-Attach enabled port can be the only port number specified in a static FDB entry, but it cannot be included in a port-list range for a static FDB entry.

# Managing the FIP Snooping Feature

## Introduction to FIP Snooping

Many data centers use Ethernet for TCP/IP networks and Fibre Channel for storage area networks (SANs).

Implementing Fibre Channel over Ethernet (FCoE) allows transmission over Ethernet networks, while preserving Fibre Channel's lossless, point-to-point transmission ability for reliable and efficient access of disk servers. FCoE is part of the International Committee for Information Technology Standards T11 FC-BB-5 standard.

FCoE Initialization Protocol (FIP) allows Ethernet nodes (Enode) to find, and set up virtual links with, FCoE forwarders (FCFs) that then connect to the fibre channel fabric.

FIP snooping monitors FCoE's virtual links and suppresses traffic not related to maintaining or establishing these virtual links to achieve a level of security comparable to native Fibre Channel.

### *FIP Snooping Requirements*

FIP snooping requires the following capabilities:

- Priority flow control (PFC) enabled
- Data center bridging capability exchange (DCBX) enabled
- FCoE application priority advertised by DCBX

## Extreme's Implementation of FIP Snooping

This section describes the Extreme Networks implementation of FIP snooping in more detail.

### *Supported Platforms*

FIP snooping is supported on the following Extreme platforms:

- BlackDiamond X8
- BlackDiamond 8800 series BD 8900-40G6X-xm
- Summit X670
- Summit X770

### *Limitations*

- *VLAN* discovery is not supported, only configured FIP VLANs.
- Virtual links between FCFs are not monitored.

## Example FIP Snooping Configuration

The following figure illustrates an example FIP snooping configuration.

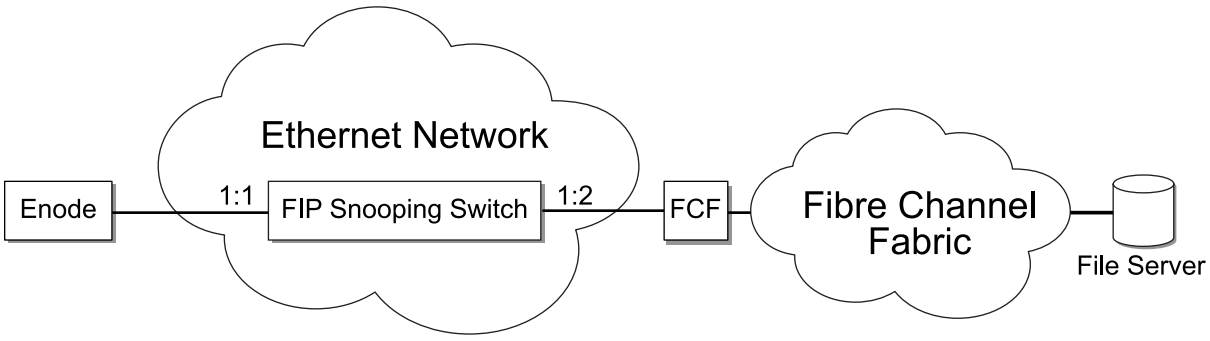
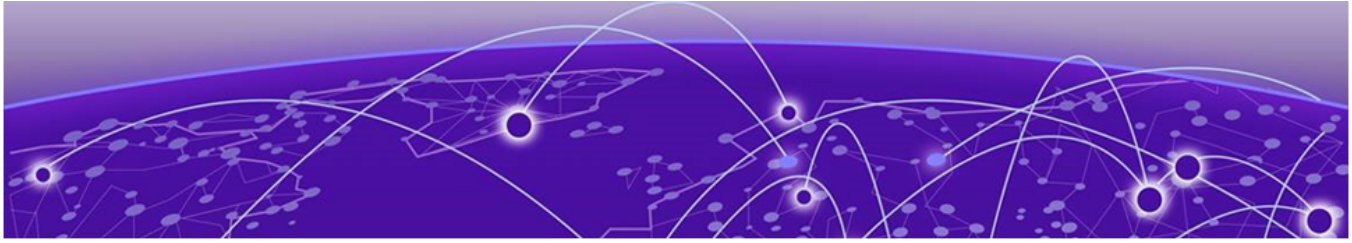



**Figure 91: Example FIP Snooping Configuration**

The following commands enable FIP snooping on *VLAN* "v1" with two ports (1:1 and 1:2) with PFC, jumbo frames, and DCBX enabled.

```
create vlan "v1"
configure vlan v1 tag 20
configure vlan v1 add ports 1:1-2 tagged
create qosprofile qp4
configure qosscheduler weighted-round-robin
configure qosprofile qp4 weight 1
enable jumbo-frame ports 1:1-2
enable flow-control rx-pause qosprofile qp4 ports 1:1-2
enable flow-control tx-pause priority 3 ports 1:1-2
```

```
enable lldp ports 1:1-2
configure lldp ports 1:1-2 advertise vendor-specific dcbx baseline
configure lldp ports 1:1-2 dcbx add application name fcoe priority 3
configure lldp ports 1:1-2 dcbx add application name fip priority 3
configure fip snooping add vlan v1
configure fip snooping vlan v1 port 1:1 location perimeter
configure fip snooping vlan v1 port 1:2 location fcf-to-enode
enable fip snooping vlan v1
```



# AVB



This chapter provides information about Audio Video Bridging support. It specifically discusses the AVB Feature Pack License, as well as how to configure and manage the AVB feature.

## Overview

Audio Video Bridging (AVB) supports the deployment of professional quality audio and/or video (AV) over standard Ethernet while coexisting with other "legacy" (or non-AV) Ethernet traffic. This supports "Network Convergence," or using one simple standard Ethernet network for all communication needs.

Note
AVB is not supported on SummitStack

To support AV applications, it is necessary for AVB systems to provide time synchronization and *QoS (Quality of Service)*.

Time Synchronization is needed so that multiple streams may be synchronized with respect to each other. For example:

- Voice and video
- Multiple audio streams for a multi-digital speaker deployment in a large venue
- Multiple Video streams in a security surveillance application

QoS is needed to ensure:

- Bandwidth guarantees sufficient for each application
- Worst Case Delay Bounds, particularly for interactive applications
- Traffic shaping to limit traffic burstiness and reduce buffering requirements

The time synchronization and QoS requirements for AVB systems are defined in the following set of IEEE Standards:

- IEEE 802.1AS: Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks (gPTP)
- IEEE 802.1Q
  - Clause 10: Multiple Registration Protocol (MRP) and Multiple MAC Registration Protocol (MMRP)
  - Clause 11: *VLAN (Virtual LAN)* Topology Management (MVRP)

  - Clause 34: Forwarding and Queuing for Time-Sensitive Streams (FQTSS)
  - Clause 35: Stream Reservation Protocol (SRP)
- IEEE 802.1BA: Audio Video Bridging (AVB) Systems

## AVB Feature Pack License

The AVB feature (including AVB, gPTP and MSRP commands) requires the AVB Feature Pack. After obtaining the AVB Feature Pack license, use the `enable license` command to install it. MRP and MVRP do not require the AVB Feature Pack. AVB is supported on the following platforms: Summit X430, X440, X450-G2, X460, X460-G2, X670, and X670-G2.

## Configuring and Managing AVB

AVB is not enabled in the default configuration, and must be enabled both globally on the switch and on the ports where you want to use it.

Note
The Summit X430 supports a maximum of eight ports for the AVB feature.

In the simplest case, when starting with a blank configuration, AVB may be enabled by executing the following two commands:

```
# enable avb

enable avb ports all
```

The status of AVB can be seen by using the following command:

```
# show avb
gPTP status          : Enabled
gPTP enabled ports   : *1s     *2m     *10m     *11m     *12m
                       *13m    *14m    *15m     *16m     *17m
                       *18m    *19m    *20m     *21m

MSRP status          : Enabled
MSRP enabled ports   : *1ab    *2ab    *10ab    *11ab    *12ab
                       *13ab   *14ab   *15ab    *16ab    *17ab
                       *18ab   *19ab   *20ab    *21ab

MVRP status          : Enabled
MVRP enabled ports   : *1      *2      *10      *11      *12
                       *13     *14     *15      *16      *17
                       *18     *19     *20      *21

Flags:         (*) Active,                  (!) Administratively disabled,
               (a) SR Class A allowed,      (b) SR Class B allowed,
               (d) Disabled gPTP port role, (m) Master gPTP port role,
               (p) Passive gPTP port role,  (s) Slave gPTP port role
```

The `show avb` command displays high level information about each of the three main protocols (gPTP, MSRP, and MVRP). Each protocol section indicates that all three protocols are enabled both globally, and on ports 1,2 and 11-21. The "*" indicates that we have link on each of the ports.

The gPTP status indicates that port 1 is a slave port, which means that the Grand Master Clock (GMC) is reachable through port 1. The gPTP status also indicates that the rest of the ports are master ports.

Furthermore, the fact that no ports are shown to be in the Disabled role means that gPTP is operational on all the ports.

ExtremeXOS provides static AVB configuration with the exception of of Best Clock Master Algorithm (BCMA), which runs as part of gPTP. The BCMA feature adds the ability to disable BCMA and to specify a slave port if desired. When BCMA functionality is on, BMCA is executed normally as described in IEEE 802.1AS. When BMCA is off, BMCA is not executed. In disabled mode, a port can be configured to be the slave-port. If no ports are configured to be slave-port, then the switch will become the Grandmaster Clock and all network-gptp enabled ports will be master ports. The `show network-gptp` command displays whether BCMA is on or off.

The "ab" on the MSRP status indicates that all ports are members of both the class A and class B domain domains. The MVRP status simply shows which ports are enabled and active.

The user interface for AVB includes the following five protocols:

- gPTP
- MRP
- MVRP
- MSRP
- FQTSS

The "avb" commands shown above are part of a set of AVB macro commands provided to simplify the process of enabling and disabling AVB. The AVB macro commands have the form:

```
[ enable | disable | unconfigure ] avb { ports [ all | port_list ] }
```

Using one of the macro commands is the same as executing the following three commands:

```
[ enable | disable | unconfigure ] network-clock gptp { ports [ all |
port_list ]}

[ enable | disable | unconfigure ] mvrp { ports [ all | port_list ]}

[ enable | disable | unconfigure ] msrp { ports [ all | port_list ]}
```

MRP does not need to be enabled or disabled, and the only MRP properties that may be configured are timer values. The defaults should be sufficient for most deployments, though it may be necessary to increase the leave-all and leave time values when supporting a large number of streams.

Multiple Registration Protocol/Multiple *VLAN* Registration Protocol is used for dynamically creating VLANs and/or dynamically adding ports to VLANs. As per IEEE Std 802.1Q-2011, some VLANs can be marked as forbidden VLANs on some ports so that when MVRP PDU is received on the port with the particular forbidden VLAN Id, the VLAN is not created and if the VLAN is already there, the port is not added to the VLAN. This functionality was added in 15.3.2.

Link aggregation allows an increased bandwidth and resilience by using a group of ports to carry traffic in parallel between switches. Multiple ports can be aggregated into one logical port. MVRP can be enabled on the logical port. The MVRP control packets will be transmitted on any available physical port of the *LAG (Link Aggregation Group)*. The peer on the other side will receive the packet and process as

if being received on the logical port. MVRP supports both dynamic (LACP) as well as static load sharing. Some restrictions that apply are:

- The individual ports of the LAG, including the master port, should not have MVRP configuration prior to grouping.
- MVRP can be enabled / disabled only on master port. The individual links cannot be configured.
- Once sharing is disabled, MVRP configuration of the master port will be lost (default will be disabled).
- The statistics and counters shown on the MVRP show commands will be a cumulative counter for all links added together. We do not maintain per link counters.
- The actual load sharing of the traffic is beyond MVRP's domain and should take place as per the configured LAG setting. MVRP just adds the LAG port to the VLAN(s).

MVRP data structure is based on port Instance. All dynamic VLANs created or propagated for a given port will be stored for each port Instance. For normal ports, the port Instance will correspond to the PIF port instance, and for LAG ports, the port Instance will correspond to the LIF port Instance. The port instance is not shown in any of the standard CLI show commands, though it is available as a part of the debug commands. Once MVRP is enabled on the master port, addition / deletion of individual links is supported. MVRP packets received on the newly added link will be accounted instantaneously.

MVRP LAG configuration examples:

- `enable sharing 13 grouping 13,14,15`
- `enable mvrp port 13`
- `enable sharing 13 grouping 13,14`
- `enable mvrp port 13`
- `configure sharing 13 add ports 15`

The VLAN registration is of three types:

- Forbidden—Port is forbidden to be added to the VLAN
- Normal—Port is allowed to be added to the VLAN
- Fixed—Port is statically added to the VLAN

The forbidden / normal setting is only for dynamic addition of ports to VLANs. Any static addition of ports to the VLANs, overrides this setting and marks the status as fixed. The forbidden setting can be used to control MSRP advertisements, in typical scaling scenarios. In addition to support for forbidden VLANs, support for periodic timer and extended-refresh timer has been added in 15.3.2.

The FQTSS settings are managed by MSRP, and may not be configured directly.

The disable commands disable the AVB protocols globally or per port without changing any other configured settings, while the unconfigure commands reset all AVB settings to the initial states, and release all switch resources that were allocated when the protocols were enabled.

More detailed configuration options are provided on a per-protocol basis using the corresponding configure commands:

```
configure network-clock gptp
configure mvrp
configure msrp
configure mrp
```

## MRP/MSRP LAG Support

*LAG* is supported for MRP/MSRP. A LAG may have one or multiple member ports, thus providing redundancy as well as additional bandwidth.

There are two modes of how a LAG runs with MRP/MSRP. The two MRP/MSRP LAG modes are single-port and cumulative. The concept of effective bandwidth is used below to simulate the available bandwidth of a physical port (port speed). As in the case of the physical port, a configurable percentage, deltaBandwidth, which by default is 75%, of the effective bandwidth is the final bandwidth available for MSRP streams.

In single-port mode, the LAG is simply a way to provide redundancy. Therefore the effective bandwidth is set to the minimum bandwidth of all member ports.

```
effective bw = min (bw of all active member ports)

MSRP reservable bw = deltalBandwidth * effective bw
```

All LAG member ports on ExtremeXOS must have the same speed to aggregate. Therefore, the minimum bandwidth is equivalent to the bandwidth of any member port.

In cumulative mode, the LAG trades redundancy for extra bandwidth. To calculate the available bandwidth, all bandwidth of one active member and partial bandwidth of other member ports are used to calculate the effective bandwidth.

```
effective bw = min (bw of all active member ports) + beta * sum (bw of
all other active member ports)

MSRP reservable bw = deltalBandwidth * effective bw
```

For example, if the LAG has three 1GB member ports, and deltaBandwidth = 75%, then in the single-port mode, the effective bandwidth is 1GB, of which 75%, that is 0.75GB is reservable for MSRP, exactly the same as in the case of a physical port.

For the same LAG in the cumulative mode, if beta = 50%, then 100% of one member port and 50% bandwidth of other two member ports contributes to the effective bandwidth, therefore:

```
effective bw = 1GB (one active member port) + 50% * 1GB (the next active
member port) + 50% * 1GB (the last active member port) = 2GB

MSRP reservable bw = 75% x effective bw = 1.5GB
```

When a LAG runs in the cumulative mode, the streams are roughly evenly distributed on all member links. Even though beta percent bandwidth of the other member ports contributes to the effective bandwidth, it does not mean that reservation on that member port is limited to beta percent of its bandwidth. That is just the way to calculate an estimate of how much bandwidth can be provided with a LAG, with a balanced redundancy requirement. All member ports are treated equally and programmed in the same way.

As in the case of physical ports, if the `(total requested BW) <= deltaBandwidth * (effective BW)`, then assume the LAG is able to handle all streams safely and provide a comfortable level of redundancy if the streams are reasonably evenly distributed. In extremely polarized cases, where many streams are hashed to a certain member link, packet drop is inevitable. When

`(total requested BW) > deltaBandwidth * (effective BW)`, then the LAG is not able to safely handle all streams, even though it may still be able to. At this time new reservation requests are declined.

The effective bandwidth is an aggregate of multiple physical ports and can exceed the bandwidth of any single physical port. The *QoS* profile is configured on a physical port, so if the reservation bandwidth request is less than the delta bandwidth of the physical port, the QoS profile will be configured to the requested bandwidth. To prevent the QoS profile configuration from exceeding the physical port bandwidth, the QoS profile configuration is limited to delta bandwidth of the physical port (which is typically 75% of the port bandwidth).

### gPTP LAG Support

Configuration of gPTP on a *LAG* port is supported. However, the protocol is still running on the member port, aka, physical port level. When gPTP is enabled on a LAG port, all member ports are enabled with the protocol, vice versa when it is disabled on the LAG port. All member ports share the same gPTP configurations. When adding or deleting LAG member port, gPTP is automatically enabled or disabled. Member ports go operational up/down when the physical port is administratively enabled/disabled.

It is recommended to perform configuration on the LAG level; however, the only keyword in the `enable/disable/configure network-gptp` CLIs allows per member port fine tuning. This per member port change is overwritten with the next LAG level configuration. Unconfigure gPTP on a member port is not allowed though; otherwise, the newly added member port may not have the reference of the master port for gptp configuration.

In the CLI display, all member ports with gPTP enabled, regardless port operational status, are listed, but noted as a LAG member port. The LAG itself is not displayed because it does not participate the protocol.

## Displaying AVB Information

The completae set of "show" commands are detailed in the *ExtremeXOS 16.2 Command Reference Guide*. Some of the more commonly used commands are outlined here.

### gPTP

Detailed information about gPTP can be displayed using the following set of commands:

```
show network-clock gptp ...
```

For example, the `show network-clock gptp ports` command can be used to view the gPTP properties of a given port, and is useful for debugging when the `summary avb` command shows that the port is not operational for gPTP.

```
# show network-clock gptp ports 1
Physical port number               : 1
gPTP port status                   : Enabled
Clock Identity                     : 00:04:96:ff:fe:51:ba:ea
gPTP Port Number                   : 1
IEEE 802.1AS Capable               : Yes
Port Role                          : 9 (Slave)
Announce Initial Interval          : 0 (1.000 secs)
Announce Current Interval          : 0 (1.000 secs)
```

```
Announce Receipt Timeout           : 3
Sync Initial Interval              : -3 (0.125 secs)
Sync Current Interval              : -3 (0.125 secs)
Sync Receipt Timeout               : 3
Sync Receipt Timeout Interval      : 375000000 ns
Measuring Propagation Delay        : Yes
Propagation Delay                  : 623 ns
Propagation Delay Threshold        : 3800 ns (auto)
Propagation Delay Asymmetry        : 0
Peer Delay Initial Interval        : 0 (1.000 secs)
Peer Delay Current Interval        : 0 (1.000 secs)
Peer Delay Allowed Lost Responses  : 3
Neighbor Rate Ratio                : 1.000020
PTP Version                        : 2
```

## MSRP

Detailed information about MSRP can be displayed using the following set of commands:

```
show msrp ...
```

Several that are commonly used are:

```
show msrp
show msrp streams
show msrp listeners
show msrp streams propagation
```

Examples of these commands are shown below.

The `show msrp` command displays the summary information included in the `show avb` command, but also displays the total number of streams and reservations on the switch.

```
# show msrp
MSRP Status                    : Enabled
MSRP Max Latency Frame Size    : 1522
MSRP Max Fan-in Ports          : No limit
MSRP Enabled Ports             : *1ab    *2ab    *10ab   *11ab    12
                                   13      14      15      16      17
                                   18      19      20      21
Total MSRP streams             : 2
Total MSRP reservations        : 6
Flags:  (*) Active,               (!) Administratively disabled,
        (a) SR Class A allowed,   (b) SR Class B allowed
```

The `show msrp streams` command displays all of the streams that the switch is aware of.

```
# show msrp streams
Stream Id              Destination     Port  Dec    Vid  Cls/Rn     BW
---------------------  --------------  ----  ----  ----  ------  ---------
00:50:c2:4e:db:02:00:00  91:e0:f0:00:ce:00     1  Adv      2  A/1      6.336 Mb
00:50:c2:4e:db:06:00:00  91:e0:f0:00:0e:82     2  Adv      2  A/1      6.336 Mb

Total Streams: 2
--------------------------------------------------------------------------------
BW       : Bandwidth,                Cls      : Traffic Class,
Dec      : Prop. Declaration Types,  Rn       : Rank

MSRP Declaration Types:
   Adv   : Talker Advertise,         AskFail  : Listener Asking Failed,
   Fail  : Talker Fail,              RdyFail  : Listener Ready Failed,
   Ready : Listener Ready
```

The `show msrp listeners` command displays all of the listeners the switch is aware of. If the declaration type is either Ready or RdyFail, a reservation has been made, and the Stream Age will show the length of time this reservation has been active.

```
# show msrp listeners
     Stream Id            Port  Dec       Dir      State      Stream Age

                                                  App  Reg  (days,hr:mm:ss)
----------------------  -----  ------- -------  ---  ---  ----------------
00:50:c2:4e:db:02:00:00      2  Ready   Ingress  VO   IN       0, 01:40:23

                            10  Ready   Ingress  VO   IN       0, 01:27:05
                            11  Ready   Ingress  VO   IN       0, 01:27:05
00:50:c2:4e:db:06:00:00      1  Ready   Ingress  VO   IN       0, 01:40:15
                            10  Ready   Ingress  VO   IN       0, 01:27:05
                            11  Ready   Ingress  VO   IN       0, 01:27:05
------------------------------------------------------------------------------
App      : Applicant State,                Dec     : MSRP Declaration Types
Dir      : Direction of MSRP attributes,   Reg     : Registrar State

MSRP Declaration Types:
AskFail  : Listener Asking Failed,         RdyFail : Listener Ready Failed,
Ready    : Listener Ready

Applicant States:
   AA    : Anxious active,                 AN      : Anxious new,
   AO    : Anxious observer,               AP      : Anxious passive,
   LA    : Leaving active,                 LO      : Leaving observer,
   QA    : Quiet active,                   QO      : Quiet observer,
   QP    : Quiet passive,                  VN      : Very anxious new,
   VO    : Very anxious observer,          VP      : Very anxious passive
Registrar States:
   IN    : In - Registered,                LV      : Leaving - Timing out
   MT    : Empty - Not Registered
```

The `show msrp streams propagation` command is useful for debugging the propagation of Talkers and Listners for each stream.

```
# show msrp streams propagation stream-id 00:50:c2:4e:db:02:00:00
     Stream Id             Destination      Port  Dec   Vid   Cls/Rn     BW
----------------------   -----------------  ----  ----  ----  ------  ---------
00:50:c2:4e:db:02:00:00   91:e0:f0:00:ce:00     1  Adv      2    A/1    6.336 Mb


    Talker Propagation:
         Ingress     Ingress     Propagated        Propagated     Egress
         DecType     Port        DecType             Ports        DecType
         -------     -------     ----------        ----------     -------
        Adv     -->  1  -->      Adv        -->          2   -->
Adv
                                                     10   -->  Adv

                                                     11   -->  Adv
    Listener Propagation:
        Egress      Egress        Propagated        Listener       Ingress
        DecType     Port          DecType             Ports        DecType
        -------     -------      ----------        ----------      -------
        Ready   <--       1   <--  Ready      <--          2  <--  Ready

                              <--  Ready      <--         10  <--  Ready
                              <--  Ready      <--         11  <--  Ready
Total Streams: 1
------------------------------------------------------------------------------
```

```
BW       : Bandwidth,                Cls      : Traffic Class,
Dec      : Prop. Declaration Types,  Rn       : Rank

MSRP Declaration Types:
   Adv   : Talker Advertise,         AskFail  : Listener Asking Failed,
   Fail  : Talker Fail,              RdyFail  : Listener Ready Failed,
  Ready  : Listener Ready
```

## MVRP

Other than the MVRP summary information displayed in the show avb command, information about dynamically created VLANs is shown using the "vlan" commands as follows.

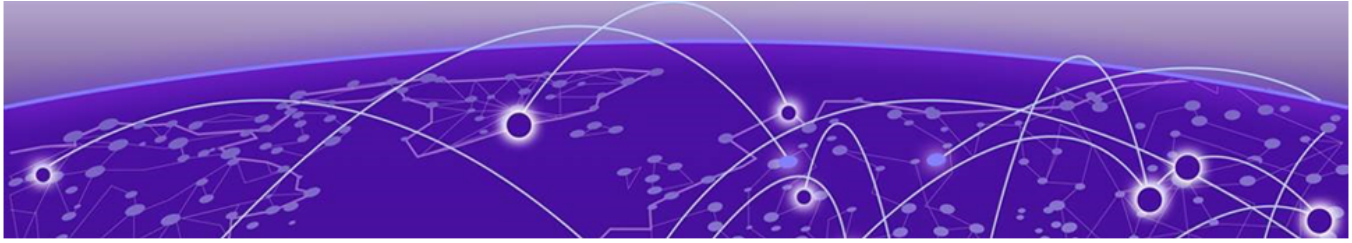
In the `show vlan` command, it can be seen that SYS_VLAN_0002 is a dynamically created *VLAN* due to the “d” flag.

```
# show vlan
-------------------------------------------------------------------------------------------
---
Name            VID  Protocol  Addr       Flags                       Proto  Ports
Virtual
                                                                             Active
router
                                                                             /Total
-------------------------------------------------------------------------------------------
---
Default         1    --------------------------------T--------------- ANY    4 /33  VR-
Default
Mgmt            4095 ------------------------------------------------ ANY    1 /1   VR-
Mgmt
SYS_VLAN_0002   2    --------------------------------T------d-------- ANY    4 /4   VR-
Default
-------------------------------------------------------------------------------------------
---
Flags : (B) BFD Enabled, (c) 802.1ad customer VLAN, (C) EAPS Control VLAN,
        (d) Dynamically created VLAN, (D) VLAN Admin Disabled,
        (e) CES Configured, (E) ESRP Enabled, (f) IP Forwarding Enabled,
        (F) Learning Disabled, (i) ISIS Enabled, (I) Inter-Switch Connection VLAN for
MLAG,
        (k) PTP Configured, (l) MPLS Enabled, (L) Loopback Enabled,
        (m) IPmc Forwarding Enabled, (M) Translation Member VLAN or Subscriber VLAN,

        (n) IP Multinetting Enabled, (N) Network Login VLAN, (o) OSPF Enabled,
        (O) Flooding Disabled, (p) PIM Enabled, (P) EAPS protected VLAN,
        (r) RIP Enabled, (R) Sub-VLAN IP Range Configured,
        (s) Sub-VLAN, (S) Super-VLAN, (t) Translation VLAN or Network VLAN,
        (T) Member of STP Domain, (v) VRRP Enabled, (V) VPLS Enabled, (W)VPWS Enabled,

        (Z) OpenFlow Enabled       Total number of VLAN(s) : 3
```

Details about SYS_VLAN_0002 can be displayed using the following command.

```
# show SYS_VLAN_0002
VLAN Interface with name SYS_VLAN_0002 created dynamically
    Admin State:      Enabled      Tagging:      802.1Q Tag 2
    Description:      None       Virtual router:   VR-Default
    IPv4 Forwarding:  Disabled
    IPv4 MC Forwarding:  Disabled
    IPv6 Forwarding:  Disabled
    IPv6 MC Forwarding:  Disabled
    IPv6:                None
    STPD:                   s0(Enabled)
```

```
    Protocol:            Match all unfiltered protocols
    Loopback:            Disabled
    NetLogin:            Disabled
    OpenFlow:            Disabled
    QosProfile:       None configured
    Flood Rate Limit QosProfile:       None configured
    Ports:   4.         (Number of active ports=4)
       Tag:       *1H,  *2H,   *10H,   *11H
         Flags:    (*) Active, (!) Disabled, (g) Load Sharing port
                   (b) Port blocked on the vlan, (m) Mac-Based port
                   (a) Egress traffic allowed for NetLogin
                   (u) Egress traffic unallowed for NetLogin
                   (t) Translate VLAN tag for Private-VLAN
                   (s) Private-VLAN System Port, (L) Loopback port
                   (e) Private-VLAN End Point Port
                   (x) VMAN Tag Translated port
                   (G) Multi-switch LAG Group port
                   (H) Dynamically added by MVRP
                   (U) Dynamically added uplink port
                   (V) Dynamically added by VM Tracking
```



# Layer 2 Tunneling and Filtering



This EXOS feature introduces the ability to tunnel and filter Layer 2 PDUs. Tunneling allows you to send Layer 2 PDUs across a service provider network, and be delivered to remote switches. It is useful when a network includes remote sites that are connected through a service provider network. Using tunneling, you can make the service provider network transparent to the customer network.

Filtering prevents Layer 2 PDUs from being received on a port.

## Layer 2 Protocol Tunneling

Layer 2 protocol tunneling (L2PT) is achieved by encapsulating the PDUs at the ingress PE device before transmitting them over the service provider network. The encapsulation prevents the PDUs from being processed by the switches in the SP network. At the egress PE device, the encapsulated packets are de-encapsulated, and transmitted to the CE device.

The encapsulation used for different types of networks is as follows:

- *VLAN (Virtual LAN)*/VMAN – The Destination Address (DA) MAC of the Layer 2 PDU is changed to the L2PT DA MAC. The switch shall also add any VLAN tags that may be required to the Layer 2 PDU before transmitting over the SP network.
- VPLS/VPWS – The DA MAC of the Layer 2 PDU is changed to L2PT DA MAC. The Layer 2 PDU is then treated like any other data packet by the *MPLS (Multiprotocol Label Switching)* stack. The MPLS stack shall add the labels and L2 headers as per its configuration to the Layer 2 PDU before transmitting over the SP network.

Tunneling is configured on a service by specifying a tunneling action for each interface of the service. The possible actions are:

- Tunnel – Configuring an interface of a service to **`tunnel`** for a protocol enables the interface to tunnel PDUs of the configured protocol that are received by the underlying port of the interface. Any PDUs that are received in its native format are tunneled instead of processing locally by the switch. Any PDUs of the protocol that are received in its encapsulated format are dropped by the switch (receiving an encapsulated packet on an interface configured to tunnel is considered proof of network misconfiguration, or loops).
- Encapsulate/Decapsulate – Configuring an interface of a service to encapsulate or de-encapsulate for a protocol enables the interface to transmit and receive PDUs of that protocol in its encapsulated

format. Native PDUs of the protocol may still be received by the underlying port of the interface, but they will not be tunneled and instead are processed locally by the switch.

- None – Configuring an interface of a service to **none** for protocol marks the interface as not participating in tunneling for that protocol. Native PDUs of the protocol that are received on the underlying port of the interface shall either be processed locally by the switch or be tunneled by another service which is configured to tunnel that protocol. Encapsulated PDUs that are received on the interface are treated like any other L2 packet.

An operator can specify a *CoS (Class of Service)* value for the tunneled PDUs. This can be useful since some L2 protocols may have a higher priority than others (for example, *STP (Spanning Tree Protocol)* may be considered higher priority than *LLDP (Link Layer Discovery Protocol)*). If a CoS value is specified for a protocol for which tunneling is enabled, the switch will transmit the encapsulated PDUs for that protocol with the operator specified CoS towards the network. The CoS value specified by the operator is transmitted on the SP network as follows:

- VLAN/VMAN – The CoS value is written to the PRI bits of the outermost VLAN tag if available.
- VPLS/VPWS – The CoS value is written to the EXP bits of the outermost MPLS label. The action taken by the switch for PDUs of a protocol is as described in the following table.

**Table 69: L2 PDU Actions**

| Ingress Action | Egress Action | Switch Action |
|---|---|---|
| None or Encap/Decap | NA | Process locally |
| Tunnel | None | Discard PDU at egress |
| Tunnel | Tunnel | Tx PDU natively |
| Tunnel | Encap/Decap | Tx PDU encapsulated |

The action taken by the switch for encapsulated PDUs for a protocol is as described in the following table.

**Table 70: L2 Encapsulated PDU Actions**

| Service has at least one I/F with tunnel action | Ingress Action | Egress Action | Switch Action |
|---|---|---|---|
| No | None or Encap/Decap | None or Encap/Decap | Forward |
| Yes | None or Tunnel | NA | Discard packet at ingress |
| Yes | Encap/Decap | None | Discard packet at egress |
| Yes | Encap/Decap | Tunnel | Tx PDU natively |
| Yes | Encap/Decap | Encap/Decap | Tx PDU encapsulated |

## Protocol Tunneling

To make L2PT configuration easier, in EXOS you can create L2PT profiles. An L2PT profile specifies the tunneling action and other parameters for protocols (specified using protocol filters) that should be tunneled. You can then apply the profile to the interfaces of the service that are participating in L2PT. And you can also change the profile when it is already bound to an interface.

The L2PT parameters that can be configured through a profile include the following:

- Tunneling Action
- Tunneling *CoS*

The following validity checks are performed when an entry for a protocol filter is created in an L2PT profile:

- Ensure that all protocols in the protocol filter define a destination MAC address.
- Ensure that all protocols in the protocol filter define a protocol identifier.
- Ensure that all protocols in the protocol filter are unique within the L2PT profile.
- If the action for the protocol filter is 'encapsulate:
  - Ensure that there are no entries with action as 'tunnel in the L2PT profile.
  - Ensure that the service interface is either a tagged *VLAN* port or a PW.
- If the action for the protocol filter is 'tunnel':
  - Ensure that there are no entries with action as 'encapsulate' in the L2PT profile.
  - For every service interface using the L2PT profile:
    - Ensure that none of the protocols in the protocol filter are filtered on the underlying port of the interface.
    - Ensure that none of the protocols in the protocol filter are tunneled on the underlying port of the interface.

The following validity checks are performed when a L2PT profile is bound to an interface of a service:

- If the profile specifies the action as 'tunnel' for protocol filter:
  - Ensure that the interface is not a PW.
  - Ensure that none of the protocols in the L2PT profile are filtered on the underlying port of the interface.
  - Ensure that none of the protocols in the L2PT profile are tunneled on the underlying port of the interface.

Typically, you will want to configure the tunneling action for all customer facing interfaces of the service that participate in L2PT as tunnel, and the tunneling action for all network facing interfaces as encapsulate/decapsulate. Once any interface of the service is configured to tunnel a protocol, the switch will configure all tagged ports and PWs of the service to encapsulate/decapsulate mode. You can override this implicit configuration by binding a profile to the service interface that specifies a different tunneling action.

For example, consider a VMAN service named c1 with customer facing ports 1, 2 and 3 and network facing ports 4, 5, 6. Ports 4, 5 and 6 are added as tagged to the VMAN and 1, 2 and 3 are added as untagged to the VMAN. The operator wants to tunnel LACP and EFM OAM on all customer facing ports at CoS 5. The configurations that he or she must make are as follows:

```
# Create a protocol filter
create protocol filter "my_slow_protocols_filter"

# Add LACP to the protocol filter
configure protocol filter "my_slow_protocols_filter"
add dest-mac 01:80:C2:00:00:02 etype 0x8809 field offset 14 value 01 mask FF

# Add EFM OAM to the protocol filter
configure protocol filter "my_slow_protocols_filter"
add dest-mac 01:80:C2:00:00:02 etype 0x8809 field offset 14 value 03 mask FF
```

```
# Create an L2PT profile for the customer facing ports named c1_l2pt_profile
create l2pt profile "c1_l2pt_profile"

# Enable CDP tunneling with CoS 5
configure l2pt profile "c1_l2pt_profile" add protocol filter
"my_slow_protocols_filter" action tunnel cos 5

# Bind c1_l2pt_profile to all customer facing ports
configure vman c1 ports 1,2,3 l2pt profile "c1_l2pt_profile"

# Please note that the network facing port 4, 5 and 6 don't have to be explicitly
# configured to encapsulate/decapsulate mode since the switch implicitly sets all
# tagged ports to encapsulate/decapsulate mode when an L2PT profile is bound to
# any port of the service.
```

The operator also has the option to configure the L2PT destination MAC address (i.e. the DA used by L2PT encapsulated PDUs). This is may be done using the following CLI command:

`configure l2pt encapsulation dest-mac mac_address`

The L2PT destination MAC address may only be changed when no L2PT profiles have been bound to any service interface. The default L2PT DA MAC is 01:00:0C:CD:CD:D0 (selected to be interoperable with Cisco and Juniper).

Use the following commands to view the status and statistics of L2PT:

`show [vlan | vman] vlan_name {ports port_list} l2pt {detail}`

`show {l2vpn} [vpls vpls_name | vpws vpws_name] {peer ipaddress} l2pt {detail}`

Use the following commands to clear L2PT stats:

`clear l2pt counters {[vlan | vman] vlan_name {ports port_list}}`

`clear l2pt counters {[vpls vpls_name {peer ipaddress} | vpws vpws_name]}`

## Implementing L2PT in EXOS

In EXOS, the L2PT data-plane is implemented almost entirely in software. When you attach a L2PT profile to a service interface, the following *ACL (Access Control List)* rules are configured:

- An ACL rule is added to copy and drop all packets with a destination address equal to the L2PT destination MAC address, and an outer *VLAN* ID equal to the VLAN tag of the service.
- For each protocol that is tunneled on the service interface, an ACL rule is added to copy and drop all packets with the same the destination address as the protocol. If the protocol defines an EtherType, then the rule is also qualified with the EtherType.
- If any protocol is tunneled on the service interface, an ACL rule is added to drop all packets received on the service interface with a destination address equal to the L2PT destination MAC address.

# Protocol Filtering

You can enable filtering of PDUs of a protocol on any port. If you enable filtering for a protocol on a port, the switch discards PDUs of that protocol on that port.

Use the following command to view protocol filter status and statistics:

```
show ports [port_list | all] protocol filter {detail}
```

Use the following command to clear protocol filtering stats:

```
clear counters ports {port_list} protocol filter
```

## Implementing Protocol Filtering in EXOS

In EXOS, the protocol filtering data-plane is implemented partially in hardware and partially in software. Filtering is performed only on the ingress. When a protocol filter is attached to a port, the following *ACL* rules are configured:

- For each protocol in the protocol filter: If the protocol does not define a user-defined field, and the protocol identifier is EtherType, or does not have a protocol identifier:
  - An ACL rule is added to drop all packets on the port that match the destination address of the packet. The rule is also qualified with the EtherType of the protocol if it defines one.

  Else:
  - An ACL rule is added to copy and drop all packets on the port that match the destination address of the packet. The rule is also qualified with the EtherType of the protocol if it defines one.

The protocol filtering data-plane inspects all packets received from ports that have protocol filters attached, and drops any packet that matches any of the protocols configured in the protocol filter.

## Protocol Filters

Both L2PT and protocol filtering allow you to tunnel or filter many protocols on an interface. For this purpose, EXOS supports creating protocol filters. A protocol filter contains a number of protocols to which you can apply some action (like tunneling and filtering). Each protocol in a protocol filter is defined using the following fields:

- The destination MAC address of PDUs of the protocol. This field is mandatory for all protocols that are to be tunneled or filtered.
- The protocol id (EtherType, LLC, SNAP). This field is mandatory for all protocols that are to be tunneled.
- User defined field. This is an arbitrary field in the PDU of the protocol that is specified using the offset of the field from the start of the PDU, the value of the field and a mask.

For example, use the following command to create a protocol filter that includes LACP and EFM OAM:

```
# Create a protocol filter
create protocol filter my_slow_protocols_filter

# Add LACP to the protocol filter
configure protocol filter my_slow_protocols_filteradd dest-mac
01:80:C2:00:00:02 etype 0x8809 field offset 14 value 01 mask FF

# Add EFM OAM to the protocol filter
configure protocol filter my_slow_protocols_filteradd dest-mac
01:80:C2:00:00:02 etype 0x8809 field offset 14 value 03 mask FF
```

The following validity checks are performed when a protocol is added to a protocol filter:

- Ensure that the protocol does not already exist in the protocol filter.
- If the protocol filter is used by any L2PT profile:
  - Ensure that the protocol defines a destination MAC address.
  - Ensure that the protocol defines a protocol identifier.
- For every L2PT profile that is using the protocol filter:
  - Ensure that the protocol is unique within the L2PT profile. If the action for the protocol filter is 'tunnel' in the L2PT profile:
    - For every service interface using the L2PT profile: ensure that the protocol is not filtered on the underlying port of the service interface.
    - It ensures that the protocol is not tunneled on the underlying port of the service interface.
- If the protocol filter is used by any port for the purpose of protocol filtering (`configure ports port# protocol filter filter-name`):
  - Ensure that the protocol defines a destination MAC address.
- For every port that has the protocol filter attached for the purpose of protocol filtering:
  - Ensure that the protocol is not tunneled by a service on that port.

Note

Protocol filters may be used with features other than L2PT and protocol filtering (for example, Protocol Based *VLAN*s). The validity tests listed above are only the ones relevant to L2PT and protocol filtering.

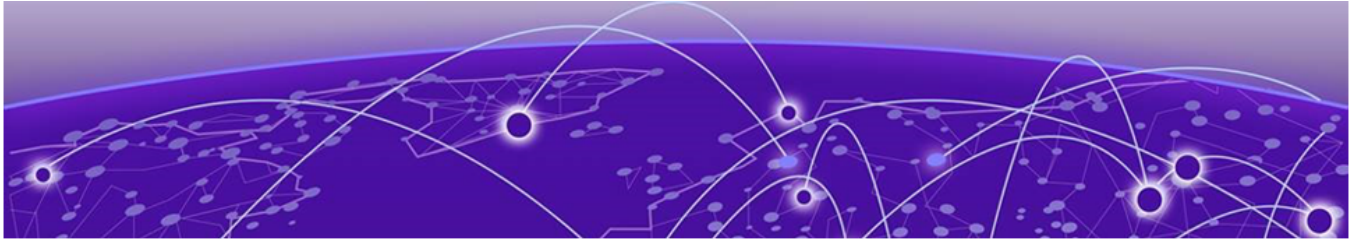
Protocol filters for the following protocols are created automatically by the switch when the switch is set to default configuration:

- Cisco Discovery Protocol (CDP)
- Unidirectional Link Detection (UDLD)
- VLAN Trunking Protocol (VTP)
- Port Aggregation Protocol (PAgP)
- Dynamic Trunking Protocol (DTP)
- Link Aggregation Control Protocol (LACP)
- *LLDP*
- *STP*
- *EDP (Extreme Discovery Protocol)*

## L2PT Limitations

- L2PT over VPLS/VPWS is not supported on Summit X480 and BlackDiamond 8K series switches.
- L2PT and protocol filtering is implemented in software, so the number of frames that can be filtered or tunneled is limited.
- Both L2PT and protocol filtering can be configured only through CLI. Configuration through *SNMP (Simple Network Management Protocol)*/XML is not supported for this release.
- If L2PT configurations are made on PWs, these configurations are lost on a restart of the *MPLS* process unless the L2PT process is also restarted.

- If L2PT configurations are made on a VPLS or VPWS service, dot1p tag inclusion must be enabled on the VPLS/VPWS.
- When tunneling protocols are point-to-point in nature, it is your responsibility to ensure that there are only two tunnel endpoints for the protocol.
- If a protocol that is configured to be tunneled on a service interface cannot be uniquely identified by its destination address and EtherType, then all packets with the same DA and EtherType of the protocol being tunneled (but that are not really PDUs of the protocol) will be slow path forwarded.
- Tagged protocol PDUs cannot be tunneled over *VLAN*s. Tagged protocol PDUs can only be tunneled over VMANs (the VMAN can be the service VMAN for a VPLS/VPWS service, or a standalone VMAN). Untagged protocol PDUs can be tunneled over both VLANs and VMANs (the VLAN/VMAN can be standalone, or be the service VMAN for a VPLS/VPWS service).
- Untagged protocol PDUs cannot be bypassed if the ingress port is an untagged VMAN port with a default CVID. Untagged protocol PDUs can be bypassed if the ingress port is an untagged VMAN port without a default CVID.
- In VPLS, only full-mesh configuration is supported for L2PT.
- L2PT is not supported on VLAN ports that have a port specific tag.



# Virtual Routers



This section provides information about Virtual Routers. It discusses how ExtremeXOS software supports Virtual Routers and VRFs, and provides specific information about how to configure and manage those virtual routers.

## Overview of Virtual Routers

The ExtremeXOS software supports virtual routers (VRs). This capability allows a single physical switch to be split into multiple *virtual router (VR)*s. This feature separates the traffic forwarded by a VR from the traffic on a different VR.

Each VR maintains a separate logical forwarding table, which allows the VRs to have overlapping IP addressing. Because each VR maintains its own separate routing information, packets arriving on one VR are never switched to another.

Note
VRs should not be connected together through a Layer 2 domain. Since there is a single MAC address per switch in the ExtremeXOS software, this same MAC address is used for all VRs. If two VRs on the same switch are connected through a Layer 2 domain, the intermediate Layer 2 switches learn the same MAC address of the switch on different ports, and may send traffic into the wrong VR.

Ports on the switch can either be used exclusively by one VR, or can be shared among two or more VRs. One reason to configure a port for the exclusive use of a single VR is to be sure that only packets from that VR egress from that port. One reason to configure a port to be shared by multiple VRs is to pass traffic from multiple VRs across a shared link.

Each *VLAN (Virtual LAN)* can belong to only one VR.

Because a single physical switch supports multiple VRs, some commands in the ExtremeXOS software require you to specify to which VR the command applies. For example, when you use the ping command, you must specify from which VR the ping packets are generated. Many commands that deal

with switch management use the management VR by default. See the *ExtremeXOS 16.2 Command Reference Guide* for the defaults for individual commands.

Note

The term VR is also used with the *VRRP (Virtual Router Redundancy Protocol)*. VRRP uses the term to refer to a single VR that spans more than one physical router, which allows multiple switches to provide redundant routing services to users. For more information about VRRP, see the VRRP Overview on page 1122.

## Types of Virtual Routers

### *System Virtual Routers*

The system VRs are the three VRs created at boot-up time. These system VRs cannot be deleted or renamed. They are named *VR-Mgmt*, *VR-Control*, and *VR-Default*. The following describes each system VR:

- VR-Mgmt

  VR-Mgmt enables remote management stations to access the switch through Telnet, SSH, and *SNMP (Simple Network Management Protocol)* sessions; it also owns the management port. No other ports can be added to VR-Mgmt, and the management port cannot be removed from it.

  The Mgmt *VLAN* is created in VR-Mgmt during ExtremeXOS system boot-up. No other VLAN can be created in this VR, and the Mgmt VLAN cannot be deleted from it.

  No routing protocol is running on or can be added to VR-Mgmt.

  VR-Mgmt is called VR-0 in ExtremeXOS releases before 11.0.

- VR-Control

  VR-Control is used for internal communications between all the modules and subsystems in the switch. It has no external visible ports, and you cannot assign any port to it.

  VR-Control has no VLAN interface, and no VLAN can be created for it.

  No routing protocol is running on or can be added to VR-Control.

  VR-Control is called VR-1 in ExtremeXOS releases before 11.0.

- VR-Default

  VR-Default is the default VR created by the ExtremeXOS system. By default, all data ports in the switch are assigned to VR-Default. Any data port can be added to and deleted from VR-Default.

  Users can create and delete VLANs in VR-Default. The Default VLAN is created in VR-Default during the ExtremeXOS system boot-up. The Default VLAN cannot be deleted from VR-Default.

  One instance of each routing protocol is spawned for VR-Default during the ExtremeXOS system boot-up, and these routing instances cannot be deleted.

  VR-Default is called VR-2 in ExtremeXOS releases before 11.0.

### *User Virtual Routers*

User VRs are the VRs created by users in addition to the system VRs, and each user VR supports Layer 3 routing and forwarding.

The routing tables for each VR are separate from the tables for other VRs, so user VRs can support overlapping address space.

**Note**

User VRs are supported only on the platforms listed for the VR feature in the following table in the *Feature License Requirements* document. When a modular switch or SummitStack contains modules or switches that do not support user VRs, the ports on those devices cannot be added to a user VR.

When a new user VR is created, by default, no ports are assigned, no *VLAN* interface is created, and no support for any routing protocols is added. User VRs support all switch routing protocols. When you add a protocol to a user VR, the user VR starts a process for that protocol. The ExtremeXOS software supports up to 63 user VRs, each of which supports protocols for that VR and all child Virtual Routers and Forwarding instances (VRFs).

**Note**

When using SNMPv2c for user created *VR*, "read community" in the *SNMP* tool should be set as "vr_name@community_name" where vr-name is user created virtual router name . Similarly for SNMPv3, "Context name" in SNMP tool should be set as "vr_name@community_name" where vr-name is user created virtual router name .

### *VRFs*

*VR* and Forwarding instances (VRFs) are similar to VRs.

VRFs are created as children of user VRs or *VR-Default*, and each VRF supports Layer 3 routing and forwarding. The routing tables for each VRF are separate from the tables for other VRs and VRFs, so VRFs can support overlapping address space. The primary differences between VRs and VRFs are:

- For each routing protocol added to a VRF, only one process is started in the user VR and VRF. The VRF protocol operates as one instance of the parent VR protocol, and additional child VRFs operate as additional instances of the same parent VR protocol process. VRFs allow a protocol process running in the parent VR to support many virtual router instances.
- ExtremeXOS supports up to 63 VRs and up to many more VRFs. (For the maximum number of supported VRFs, see the *ExtremeXOS Release Notes*.)

There are two types of VRFs:

**VPN VRFs**

Support *BGP (Border Gateway Protocol)*-based Layer 3 VPNs over *MPLS (Multiprotocol Label Switching)*. VPN VRF tables support entries for additional configuration parameters that enable Layer 3 VPN functionality over an BGP/MPLS backbone network.

**Non-VPN VRFs**

Support static routes and BGP. VRFs do not support dynamic routing protocols.

Use VRFs instead of VRs when your network plan calls for more than 63 virtual routers or when you want to create Layer 3 VPNs. Use VRs instead of VRFs when the routing protocol you want to use is not supported on a VRF.

When a new VRF is created, by default, no ports are assigned, no *VLAN* interface is created, and no support for any routing protocols is added. When you add a protocol to a VRF, an instance of the protocol is created in the protocol process running in the parent VR, if the protocol process exists. If no protocol process is running in the parent VR, a process is started and a protocol instance corresponding to this VRF is created within that process.

The rest of this chapter uses the following terms to identify the different types of VRs and VRFs to which features and commands apply:

- VR—All VRs and VRFs
- VRF—VPN and non-VPN VRFs
- VPN VRF—VPN VRFs only
- Non-VPN VRF—Non-VPN VRFs only

Note

VRFs are supported only on the platforms listed for the VRF feature in the *Feature License Requirements* document. When a modular switch or SummitStack contains modules or switches that do not support VRFs, the ports on those devices cannot be added to a VRF.

## VR Configuration Context

Each VR and VRF has its own configuration domain or context, in which you can enter commands to configure that VR. Some commands allow you to specify a VR to which the command applies.

For other commands, you must change context to that of the target VR or VRF before you execute the command. The current context is indicated in the command line interface (CLI) prompt by the name of the user VR or VRF. If no name appears in the prompt, the context is *VR-Default*.

For instructions on changing context, see Changing the VR Context on page 629.

Commands that apply to the current VR context include all the *BGP*, *OSPF (Open Shortest Path First)*, OSPFv3, PIM, IS-IS, *RIP (Routing Information Protocol)*, and *RIPng (Routing Information Protocol Next Generation)* commands, and the commands listed in the following table. Commands that apply to the current VRF context are limited to BGP commands and the commands listed in the following table.

**Table 71: Virtual Router Commands**

| |
|---|
| [enable \| disable] ipforwarding |
| clear iparp |
| clear counters iparp |
| configure iparp |
| configure iparp [**add** \| **delete**] |

[8] Other commands are available with these listed.

**Table 71: Virtual Router Commands (continued)**

| |
|---|
| [enable \| disable] iparp |
| show iparp |
| configure iproute [**add** \| **delete**] |
| show iproute |
| show ipstats |
| rtlookup |
| create [vlan \| vman] *vlan-name* |
| [enable \| disable] igmp |
| [enable \| disable] igmp snooping |
| [enable \| disable] ipmcforwarding |
| show igmp |
| show igmp snooping |
| show igmp group |
| show igmp snooping cache |

# Managing Virtual Routers

## Creating and Deleting User Virtual Routers

Before you delete a VR, you must delete all VLANs and child VRFs created in that VR. All of the ports assigned to this VR are deleted and made available to assign to other VRs and VRFs. Any routing protocol that is running on the VR is shut down and deleted gracefully.

- To create a user VR, use the following command and do not include the type or vr attributes:

  `create` **`virtual-router`** `name {`**`type`** `[`**`vrf`** `|` **`vpn-vrf`** `{`**`vr`** `parent_vr_name}]}`

  Note

  User VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document.

  A VR name cannot be the same as a *VLAN* name. You cannot name a user VR with the names *VR-Mgmt*, *VR-Control*, or *VR-Default* because these are the existing default system VRs. For backward compatibility, user VRs also cannot be named VR-0, VR-1 or VR-2 because these three names are the names for the system VRs in ExtremeXOS releases before 11.0.

  If you exceed the maximum number of VRs supported on your platform, a message similar to the following appears:

  ```
  Error: Maximum number of User VRs supported by the system is 63
  ```

- To display the virtual routers, use the following command:

  `show virtual-router {`*`name`*`}`

- To delete a user VR, use the following command:

  `delete virtual-router {`*`name`*`}`

## Creating and Deleting VRFs

Before you delete a VRF, you must delete all VLANs and stop all protocols that are assigned to that VRF. All of the ports assigned to a deleted VRF are deleted and made available to assign to other VRs and VRFs. Any routing protocol instance that is assigned to the VRF is deleted gracefully.

- To create a VRF, use the following command and include the type attribute:

  `create virtual-router` *name* {**type** [**vrf** | **vpn-vrf** {**vr** *parent_vr_name*}]}

  Note

  VRFs are supported only on the platforms listed for this feature in the *Feature License Requirements* document. To support a Layer 3 VPN, a VPN VRF must be created under the parent VR that will run the *MPLS* protocol.

  A VRF name cannot have the same name as a *VLAN* or VR.
- To display the VRFs, use the following command:

  `show virtual-router` {*name*}
- To delete a VRF, use the following command:

  `delete virtual-router` {*name*}

## Enabling and Disabling VRFs

VRFs are enabled when created. If you want to shut down a Layer 3 VPN, you can disable the corresponding VRF, which disables only the corresponding Layer 3 VPN and does not affect other routing services or Layer 3 VPNs.

To enable or disable a VRF, use the following commands:

`enable virtual-router` *vrf-name*

`disable virtual-router` *vrf-name*

## Configuring and Removing a VR Description

A VR description is a text message that can be used to label the VR. The text message is for viewing and *SNMP* MIB reports only; it has no affect on VR operation.

To configure or remove a description, use the following commands:

`configure vr` *name* **description** *string*

`unconfigure vr` *name* **description**

## Changing the VR Context

The VR context is introduced in VR Configuration Context on page 627.

To switch to a context for a different VR, use the following command:

`virtual-router` {*vr-name*}

The CLI prompt displays the VR context.

## Adding and Deleting Routing Protocols

When a user VR is created, no resources are allocated for routing protocols. You must add the routing protocols needed for your VR before you attempt to configure them. The maximum number of protocols supported is 64.

This provides for the following protocols:

- The basic seven protocols on *VR-Default* (*RIP*, *OSPF*, *BGP*, PIM, ISIS, *OSPFv3 (Open Shortest Path First version 3)*, and RIPNG)
- 1 *MPLS* protocol instance on any VR (only on platforms that support MPLS)
- 56 additional protocols for user VRs. Any combination of the 7 protocols supported on user VRs (RIP, OSPF, BGP, PIM, ISIS, OSPFv3, and RIPNG) can be assigned to user VRs, up to a maximum number of 56.

When you add a protocol to a user VR, the software starts a process to support the protocol, but it does not enable that protocol. After you add a protocol to a user VR, you must specifically enable and configure that protocol before it starts.

When you add a protocol to a VRF, a protocol process is started in the parent VR (if it is not already started) and a protocol instance is created inside that process for this VRF.

Note

You must add, configure, and enable a protocol for a VR before you start unicast or multicast forwarding on the VR and before you configure any features (such as VLANs) to use the VR.

- To add a protocol to a VR, use the command:

  `configure vr` *`vr_name`* `[`**`add`**` | `**`delete`**`] `**`protocol`**` [`**`ospf`**` | `**`ospf3`**` | `**`rip`**` | `**`ripng`**` | `**`bgp`**` | `**`isis`**` | `**`pim`**`]`

  If you add more than the maximum number of protocols, the following message appears:
  `Error: Maximum number of Protocols that can be started in the system is 64`
- To remove a protocol from a VR, use the command:

  `configure vr` *`vr-name`* **`delete protocol`** *`protocol-name`*

## Configuring Ports to Use One or More Virtual Routers

By default, all the user data ports belong to *VR-Default* and the default *VLAN*, Default. All these ports are used exclusively by VR-Default. To configure a port to use one or more virtual routers, you need to perform one or more of the tasks described in the following sections:

- Deleting Ports from a Virtual Router on page 630
- Adding Ports to a Single Virtual Router on page 631
- Adding Ports to Multiple Virtual Routers on page 631

### *Deleting Ports from a Virtual Router*

To configure a port for exclusive use by another VR, or for use by multiple VRs, it must first be deleted from *VR-Default*. You must delete the port from any *VLAN* it belongs to before deleting it from a VR.

To delete a port from a VR, use the command:

```
configure vr vr-name delete ports port_list
```

**Caution**

Do not create Layer 2 connections between ports assigned to different VRs in the same switch. Because each switch supports just one MAC address, every VR in the switch uses the same MAC address. A Layer 2 connection between two VRs can cause external devices to direct traffic to the wrong VR.

### *Adding Ports to a Single Virtual Router*

When you add a port to a VR, that port can only be used by that VR. To add a port to a single VR, use the command:

```
configure vr vr-name add ports port_list
```

The following example demonstrates how to remove all the ports on slot 3 from the Default *VLAN* in *VR-Default* and add them for the exclusive use of user VR helix:

```
configure vlan default delete ports 3:*
configure vr vr-default delete ports 3:*
configure vr helix add ports 3:*
```

### *Adding Ports to Multiple Virtual Routers*

To use a port in multiple VRs, do not add the port to a VR as described in the previous section.

To add the port to a *VLAN* in the desired VR, use the following command:

**Note**

See QoS for details about how multiple VRs per port can affect DiffServ and code replacement.

You should configure any protocols you want to use on a user VR before you add a VLAN to the user VR. When IP multicast forwarding will be supported on a user VR, add the PIM protocol before you enable IP multicast forwarding.

The following example demonstrates how to add port 3:5 to user VRs VR-green and VR-blue.

The tagged VLAN bldg_200 was previously configured in VR-green, and the tagged VLAN bldg_300 was previously configured in VR-blue.

```
configure vlan default delete ports 3:5
configure vr vr-default delete ports 3:5
configure vlan bldg_200 add ports 3:5 tagged
configure vlan bldg_300 add ports 3:5 tagged
```

## Displaying Ports and Protocols

To display the ports, protocols, and names of protocol processes for a VR, use the following command:

```
show virtual-router {name}
```

## Configuring the Routing Protocols and VLANs

After a user VR is created, the ports are added, and support for any required routing protocols is added, you can configure the VR.

- To create a *VLAN* in a VR, use the command:

  `create vlan vlan_name {description vlan-description} {vr name}`

  If you do not specify a VR in the create vlan command, the VLAN is created in the current VR context.

  VLAN names must conform to the guidelines specified in Object Names on page 16.

  **Note**
  All VLAN names and VLAN IDs on a switch must be unique, regardless of the VR in which they are created. You cannot have two VLANs with the same name, even if they are in different VRs.

- To display the VLANs in a specific VR, use the command:

  `show vlan virtual-router vr-name`, which is a specific form of this command:

  `show vlan {virtual-router vr-name}`

You can also configure routing protocols by using the standard ExtremeXOS software commands. The routing configurations of the different VRs are independent of each other.

## Configuration Tasks for Layer 3 VPNs

### *Layer 3 VPN Configuration Overview*

To configure VR and VRF support for a Layer 3 VPN, do the following:

1. Configure the PE to PE interfaces as follows:
   a. Select the VR that to which the *MPLS* protocol will be added.
   b. Assign PE facing VLANs and ports to the VR.
2. Configure the PE to CE interface as follows:
   a. Create a VPN VRF as described in Creating and Deleting VRFs on page 629.
   b. Assign a VPN ID to the VPN VRF as described in Configuring a VPN ID on page 632.
   c. Assign an RD to the VPN VRF as described in Configuring the Route Distinguisher on page 633.
   d. Configure RTs for the VPN VRF as described in Configuring Route Targets on page 633.
   e. Assign CE facing VLANs and ports to the VRF.
   f. Configure *BGP* as described in #unique_1314.
3. Configure BGP as described in #unique_1314.

### *Configuring a VPN ID*

A VPN ID is an optional configuration parameter that you can use to associate a Layer 3 VPN ID label with a VRF.

To configure a VPN ID, use the command:

`configure vr vrf_name vpn-id 3_byte_oui:4_vpn_index`

### *Configuring the Route Distinquisher*

The Route Distinquisher (RD) is added to the beginning of a VPN customer's IPv4 prefix to create a globally unique VPNv4 prefix for routing over the Layer 3 VPN. The IPv4 hosts that connect to a VRF must have unique IPv4 addresses. To eliminate duplicate IPv4 address issues between VRFs, configure a unique RD for each VRF.

To configure or unconfigure an RD, use the following commands:

```
configure vr vrf_name rd [2_byte_as_num:4_byte_number |
ip_address:2_byte_number | 4_byte_as_num:2_byte_number]

unconfigure vr vrf_name rd
```

### *Configuring Route Targets*

Route Targets (RTs) are used by *BGP* as extended communities that define which VRFs will export and import learned routes. A VRF import RT will receive Layer 3 VPN routes from any PE VRF that is configured to export to that RT. To enable BGP Layer 3 VPN communications, you must configure import and export RTs on every VRF that will participate in the Layer 3 VPN.

To add, delete, or configure an RT, use the command:

```
configure vr vrf_name route-target [import | export | both] [add |
delete] [route_target_extended_community]
```

### *Enabling and Disabling Layer 3 VPN SNMP Traps for a VR*

To enable or disable Layer 3 VPN SNMP traps for a VR, use the following commands:

```
enable snmp trap l3vpn {vr}
```

Note
You must enable this command in the parent VR of VPN-VRF.

```
disable snmp trap l3vpn {vr}
```

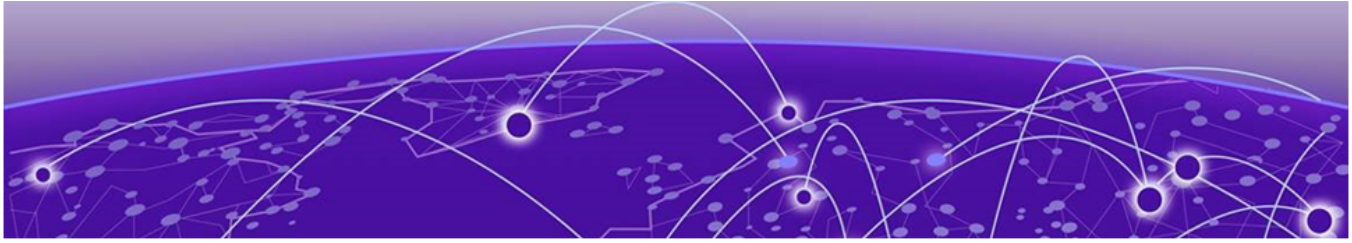
## Virtual Router Configuration Example

The following example demonstrates how to:

- Create a user VR named helix.
- Remove ports from the *VLAN* Default and *VR-Default*.
- Add ports to user VR helix.
- Add the *OSPF* protocol to user VR helix.
- Set the VR context to helix, so that subsequent VR commands affect VR helix.
- Create an incoming VLAN named helix-accounting-in.
- Create an outgoing VLAN named helix-accounting-out.
- Add ports that belong to user VR helix to the helix-accounting incoming and outgoing VLANs.

The CLI prompt is shown in this example to show how the VR context appears. At the end of the example, the VR is ready to be configured for OSPF, using ExtremeXOS software commands.

```
* BD10K.1 # create virtual-router helix
* BD10K.2 # configure vlan default delete ports 3:*
* BD10K.3 # configure vr vr-default delete ports 3:*
* BD10K.4 # configure vr helix add ports 3:*
* BD10K.5 # configure vr helix add protocol ospf
* BD10K.6 # virtual-router helix
* (vr helix) BD10K.8 # configure helix-accounting-in add ports 3:1
* (vr helix) BD10K.8 # configure helix-accounting-out add ports 3:2
* (vr helix) BD10K.9 #
```



# Policy Manager



This chapter provides information about how ExtremeXOS implements policy statements. It includes an overview of the Policy Manager, as well as specific information about how to create, edit, check, and apply policies.

## Policy Manager and Policies Overview

One of the processes that make up the ExtremeXOS system is the policy manager. The policy manager is responsible for maintaining a set of policy statements in a policy database and communicating these policy statements to the applications that request them.

Policies are used by the routing protocol applications to control the advertisement, reception, and use of routing information by the switch. Using policies, a set of routes can be selectively permitted (or denied) based on their attributes, for advertisements in the routing domain. The routing protocol application can also modify the attributes of the routing information, based on the policy statements.

Policies are also used by the *ACL (Access Control List)* application to perform packet filtering and forwarding decisions on packets. The ACL application will program these policies into the packet filtering hardware on the switch. Packets can be dropped, forwarded, moved to a different *QoS (Quality of Service)* profile, or counted, based on the policy statements provided by the policy manager.

## Creating and Editing Policies

A policy is created by writing a text file that contains a series of rule entries describing match conditions and actions to take.

Policies are created by writing a text file on a separate machine and then downloading it to the switch. Once on the switch, the file is then loaded into a policy database to be used by applications on the switch. Policy text files can also be created and edited directly on the switch.

Note

Although ExtremeXOS does not prohibit mixing *ACL* and routing type entries in a policy file, it is best to use separate policy files for ACL and routing policies instead of mixing the entries.

When you create a policy file, name the file with the policy name that you will use when applying the policy, and use “ .pol” as the filename extension. For example, the policy name “boundary” refers to the text file “boundary.pol”.

## Using the Edit Command

A VI-like editor is available on the switch to edit policies. There are many commands available with the editor. For information about the editor commands, use any tutorial or documentation about VI. The following is only a short introduction to the editor.

The editor operates in one of two modes: command and input.

1. To edit a policy file on the switch by launching the editor, enter the command:

   `edit policy` *`filename`*.pol

   When a file first opens, you are in the command mode.
2. To write in the file, use the keyboard arrow keys to position your cursor within the file, then press one of the following keys to enter input mode:
   a. **[i]**—Inserts text ahead of the initial cursor position.
   b. **[a]**—Appends text after the initial cursor position.
3. To escape the input mode and return to the command mode, press the **[Esc]** key.

Several commands can be used from the command mode. The following commands are the most commonly used:

- **[dd]**—Deletes the current line.
- **[yy]**—Copies the current line.
- **[p]**—Pastes the line copied.
- **[:w]**—Writes (saves) the file.
- **[:q]**—Quits the file if no changes were made.
- **[:q!]**—Forcefully quits the file without saving changes.
- **[:wq]**—Writes and quits the file.

## Using a Separate Machine to Edit Policies

You can also edit policies on a separate machine. Any common text editor can be used to create a policy file. The file is then transferred to the switch using TFTP and then applied.

To transfer policy files to the switch, enter the command:

`tftp [`*`host-name | ip-address`*`] {`**`-v`** *`vr_name`*`} [`**`-g | -p`**`] [{`**`-l [internal-memory`** *`local-file-internal`* **`| memorycard`** *`local-file-memcard | local-file`*`} {`**`-r`** *`remote-file`*`} | {`**`-r`** *`remote-file`*`} {`**`-l [internal-memory`** *`local-file-internal`* **`| memorycard`** *`local-file-memcard | local-file`*`]}]`

## Checking Policies

A policy file can be checked to see if it is syntactically correct. This command can only determine if the syntax of the policy file is correct and can be loaded into the policy manager database. Since a policy can be used by multiple applications, a particular application may have additional constraints on allowable policies.

To check the policy syntax, enter the command:

`check policy` *`policy_name`*

## Refreshing Policies

When a policy file is changed (such as adding, deleting an entry, adding/deleting/modifying a statement), the information in the policy database does not change until the policy is refreshed. The user must refresh the policy so that the latest copy of policy is used. When the policy is refreshed, the new policy file is read, processed, and stored in the server database.

Any clients that use the policy are updated.

- To refresh the policy, enter the command:

  `refresh policy policy_name`

  For *ACL* policies only, during the time that an ACL policy is refreshed, packets on the interface are blackholed, by default. This is to protect the switch during the short time that the policy is being applied to the hardware. It is conceivable that an unwanted packet could be forwarded by the switch as the new ACL is being set up in the hardware. You can disable this behavior.

  **Note**
  Performing a refresh on multiple ports requires the original and modified policy to coexist at the same time in the intermittent state. If this is not possible due to slice limitations, the refresh will fail with "ACL slice full" error.

- To control the behavior of the switch during an ACL refresh, enter the commands:

  `enable access-list refresh blackhole`

  `disable access-list refresh blackhole`

  In releases previous to ExtremeXOS 11.4, when ACLs were refreshed, all the ACL entries were removed, and new ACL entries were created to implement the newly applied policy.

  Beginning in release 11.4, the policy manager uses Smart Refresh to update the ACLs. When a change is detected, only the ACL changes needed to modify the ACLs are sent to the hardware, and the unchanged entries remain. This behavior avoids having to blackhole packets because the ACLs have been momentarily cleared. Smart Refresh works well up for up to 200 changes. If the number of changes exceeds 200, you will see this message: Policy file has more than 200 new rules. Smart refresh can not be carried out. Following this message, you will see a prompt based on the current blackhole configuration. If blackhole is disabled you will see the following prompt:

  ```
  Note, the current setting for Access-list Refresh Blackhole is Disabled. WARNING: If a
  full refresh is performed, it is possible packets that should be denied may be
  forwarded through the switch during the time the access list is being installed.
  Would you like to perform a full refresh?
  ```

  If blackhole is enabled, you will see the following prompt:

  ```
  Note, the current setting for Access-list Refresh Blackhole is Enabled.
  Would you like to perform a full refresh?
  ```

To take advantage of Smart Refresh, disable access-list refresh blackholing.

**Note**
Smart refresh is not performed for policies if the number of entries in the policy change during refresh.

# Applying Policies

*ACL* policies and routing policies are applied using different commands.

## Applying ACL Policies

A policy intended to be used as an *ACL* is applied to an interface, and the CLI command option is named *aclname*.

Supply the policy name in place of the *aclname* option.

- To apply an ACL policy, enter the command:

  `configure access-list` *aclname* [**any** | **ports** *port_list* ] {**ingress** | **egress**}

  When you use the **any** keyword, the ACL is applied to all the interfaces and is referred to as the wildcard ACL. This ACL is evaluated for any ports without specific ACLs, and it is also applied to any packets that do not match the specific ACLs applied to the interfaces.

  When an ACL is already configured on an interface, the command is rejected and an error message is displayed.
- To remove an ACL from an interface, enter the command:

  `unconfigure access-list` *policy_name* {**any** | **ports** *port_list* } {**ingress** | **egress**}
- To display the interfaces that have ACLs configured and the ACL that is configured on each. enter the command:

  `show access-list` {**any** | **ports** *port_list* } {**ingress** | **egress**}

## Applying Routing Policies

To apply a routing policy, use the command appropriate to the client. Different protocols support different ways to apply policies, but there are some generalities.

Commands that use the keyword **import-policy** are used to change the attributes of routes installed into the switch routing table by the protocol. These commands cannot be used to determine the routes to be added to the routing table.

To remove a routing policy, use the **none** option in the command.

The following are examples for the *BGP (Border Gateway Protocol)* and *RIP (Routing Information Protocol)* protocols:

`configure bgp import-policy [`*policy-name* |**none**]

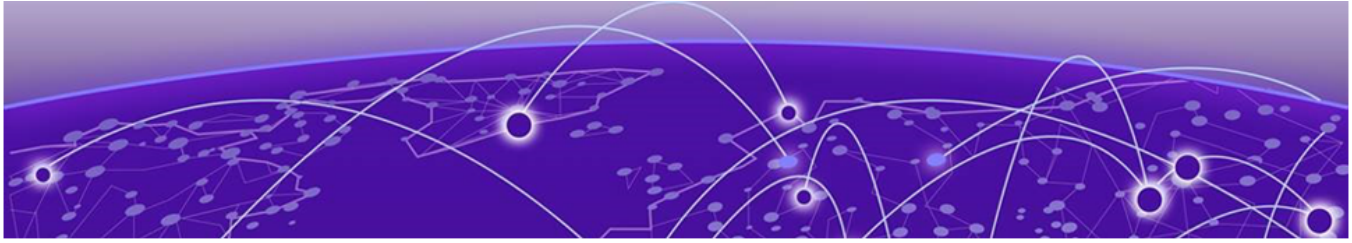
`configure rip import-policy [`*policy-name* |**none**]

Commands that use the keyword **route-policy** control the routes advertised or received by the protocol. Following are examples for BGP and RIP:

```
configure bgp neighbor [remoteaddr |all] {address-family [ipv4-unicast |
ipv4-multicast |ipv6-unicast |ipv6-multicast |vpnv4]} route-policy [in |
out] [none | policy]

configure bgp peer-group peer-group-name {address-family [ipv4-unicast |
ipv4-multicast | ipv6-unicast | ipv6-multicast |vpnv4]} route-policy [in
|out] [none | policy]

configure rip vlan [vlan_name | all] route-policy [in |out] [policy-name
| none]
```

Other examples of commands that use route policies include:

```
configure ospf area area-identifier external-filter [policy-map |none]

configure ospf area area-identifier interarea-filter [policy-map |none]

configure rip vlan [vlan_name |all] trusted-gateway [policy-name |none]
```



# ACLs



This chapter discusses *ACL (Access Control List)*s. It includes overview information, as well as sections on the following topics:

- ACL Two-stage policy
- ACL Rule syntax
- Layer-2 Protocol Tunneling ACLs
- ACL Byte Counters
- Dynamic ACLs
- ACL Evaluation Precedence
- Applying ACL Policy Files
- ACL Mechanisms
- Policy Based Routing
- ACL Troubleshooting

## ACLs Overview

*ACL*s are used to define packet filtering and forwarding rules for traffic traversing the switch. Each packet arriving on an ingress port and/or *VLAN (Virtual LAN)* is compared to the access list applied to that interface and is either permitted or denied. Packets egressing an interface can also be filtered on the platforms listed for this feature in the *Feature License Requirements* document.. However, only a subset of the filtering conditions available for ingress filtering are available for egress filtering.

In addition to forwarding or dropping packets that match an ACL, the switch can also perform additional operations such as incrementing counters, logging packet headers, mirroring traffic to a

monitor port, sending the packet to a *QoS (Quality of Service)* profile, and metering the packets matching the ACL to control bandwidth. (Metering is supported only on the platforms listed for this feature in the *Feature License Requirements* document.) Using ACLs has no impact on switch performance (with the minor exception of the mirror-cpu action modifier).

ACLs are typically applied to traffic that crosses Layer 3 router boundaries, but it is possible to use access lists within a Layer 2 virtual LAN (VLAN).

ACLs in ExtremeXOS apply to all traffic. This is somewhat different from the behavior in ExtremeWare. For example, if you deny all the traffic to a port, no traffic, including control packets, such as *OSPF (Open Shortest Path First)* or *RIP (Routing Information Protocol)*, will reach the switch and the adjacency will be dropped.

Note

Some locally CPU-generated packets are not subject to egress ACL processing.

You must explicitly allow those types of packets (if desired). In ExtremeWare, an ACL that denied “all” traffic would allow control packets (those bound for the CPU) to reach the switch.

ACLs are created in two different ways. One method is to create an ACL policy file and apply that ACL policy file to a list of ports, a VLAN, or to all interfaces. The second method to create an ACL is to use the CLI to specify a single rule, called a dynamic ACL; this is the default.

Note

ACLs applied to a VLAN are actually applied to all ports on the switch, without regard to VLAN membership. The result is that resources are consumed per chip on BlackDiamond 8000, BlackDiamond X8, and Summit family switches.

An ACL policy file is a text file that contains one or more ACL rule entries. This first method creates ACLs that are persistent across switch reboots, can contain a large number of rule entries, and are all applied at the same time. See ACL Rule Syntax on page 646 for information about creating ACL rule entries.

Policy files are also used to define routing policies. Routing policies are used to control the advertisement or recognition of routes communicated by routing protocols. ACL policy files and routing policy files are both handled by the policy manager, and the syntax for both types of files is checked by the policy manager.

Note

Although ExtremeXOS does not prohibit mixing ACL and routing type entries in a policy file, it is strongly recommended that you do not mix the entries and do use separate policy files for ACL and routing policies.

Dynamic ACLs persist across reboots; however, you can configure non-persistent dynamic ACLS that disappear when the switch reboots. Dynamic ACLs consist of only a single rule. Multiple dynamic ACLs can be applied to an interface. See Layer-2 Protocol Tunneling ACLs on page 664 for information about creating dynamic ACLs. The precedence of ACLs can be configured by defining zones and configuring the priority of both the zones and the ACLs within a zone. See Configuring ACL Priority on page 668 for more information.

# ACL Two-Stage Policy

The following diagram shows the three *ACL* processors available that can be used for filtering the packets at various stages of processing:

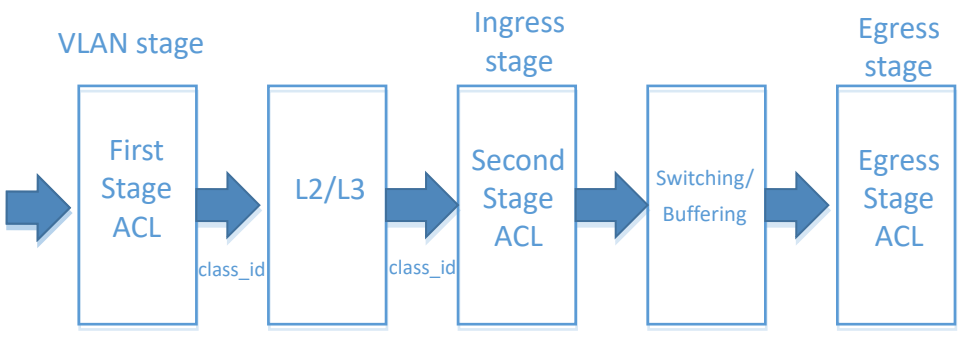



**Figure 92: ACL Stages**

First Stage ACL / *VLAN* Processor: is used to filter packets before ingress processing. It can be used to assign the VLAN, set a CLASS ID, or perform other more traditional ACL actions, such as drop or count. In general, this stage's scale, actions, and match criteria are more limited than the ingress stage. However, the high-level architecture of the first stage ACL is the same as the second stage ACL in that it is composed of a series of slices, or individual TCAM elements. First stage ACL rules are included in the policy file or dynamic ACLs in the same way as regular second stage ACL rules. To specify that a rule is to be added to the first stage ACL table, use the "class-id <class-id>" action.

Second Stage ACL / Ingress Filter Processor: is used to filter packets for ingress processing and is the primary hardware resource used for ingress user ACLs. While this stage follows the L2 and L3 lookups, the packet data presented to this stage is pre-modified, except in the case of tunneling. In general, this stage is the most capable and scalable of the 3 stages. Second stage ACL rules can additionally match the class-id specified as an action in a first stage ACL rule. This is done by listing "class-id <class-id>" in the match clause of the rule.

Egress ACL: is used to filter egressing packets after packet switching, queuing, and buffering operations have been performed. It can be used to deny or modify (e.g. 802.1p, DSCP) packets but cannot be used to redirect or change *QoS*. In general, this stage's scale, actions, and match criteria are more limited than the ingress stage.

The two-stage ingress classification is supported for static user ACL policies, dynamic CLI-driven user ACLs, and dynamic API-driven ACLs.

## Platform Support

This feature is supported on the following platforms:

- Summit X450-G2
- Summit X460
- Summit X460-G2

- Summit X480
- Summit X670
- Summit X670-G2
- Summit X770
- BlackDiamond 8K-G48Xc
- BlackDiamond 8K-G48Tc
- BlackDiamond 8K-G24Xc
- BlackDiamond 8K-10G4Xc
- BlackDiamond 8K-S-G8Xc
- BlackDiamond 8K-S-10G1Xc
- BlackDiamond 8K-S-10G2Xc
- BlackDiamond 8900-10G224X-c
- BlackDiamond xl-series
- BlackDiamond 8900-G96T-c
- Black Diamond 8900-40G6X-c
- Black Diamond X8
- E4G200 and E4G400

## Feature Description

Rules in the first classifier are set up with an action to set class_id. Rules in the second classifier are setup to use the class_id as the key to match on the identity specific policies. The class_id is the common attribute between the two classifiers/tables, uniquely identifies the role of the identity.

This feature introduces one new *ACL* action modifier for specifying the class-id from the first stage that will be input into the second stage. It also introduces one new ACL match criteria for matching the class-id within the second stage.

When a rule is installed in the first stage ACL table, it will be accounted for in the "Stage: LOOKUP" section of "show access-list usage acl-slice". When a rule is installed in the second stage ACL table, it will be accounted for in the "Stage: INGRESS" section of this command. For example:

```
X460G2-48x-10G4.9 # show access-list usage acl-slice port 1
Ports 1-54
Stage: INGRESS
Slices:          Used: 0  Available: 16
Virtual Slice  * (physical slice  0) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  1) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  2) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  3) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  4) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  5) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  6) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  7) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  8) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  9) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice 10) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice 11) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice 12) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice 13) Rules:   Used:      0  Available:    256
```

```
Virtual Slice  * (physical slice 14) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice 15) Rules:   Used:      0  Available:    256
Stage: EGRESS
Slices:          Used: 0  Available: 4
Virtual Slice  * (physical slice  0) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  1) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  2) Rules:   Used:      0  Available:    256
Virtual Slice  * (physical slice  3) Rules:   Used:      0  Available:    256
Stage: LOOKUP
Slices:          Used: 0  Available: 4
Virtual Slice  * (physical slice  0) Rules:   Used:      0  Available:    512
Virtual Slice  * (physical slice  1) Rules:   Used:      0  Available:    512
Virtual Slice  * (physical slice  2) Rules:   Used:      0  Available:    512
Virtual Slice  * (physical slice  3) Rules:   Used:      0  Available:    512
Stage: EXTERNAL

Virtual Slice :  (*) Physical slice not allocated to any virtual slice.
X460G2-48x-10G4.10 #
```

*Limitations*

- The second stage ACL will always override any qosprofile set in the First stage.
- A first stage ACL rule will not work for untagged traffic when "vlan-id" is used as a matching condition or when applied to a vlan .
- Matching "arp-sender-address" OR "arp-target-address" in the first stage ACL is not supported. However, matching both conditions is supported on select platforms.
- L4 port ranges are not supported in the first stage ACL.

**Table 72: First Stage ACL Support Actions**

| Platform Family | Platform | Permit | Deny | Count | Replace-dot1p-value | qosprofile | Replace-dot1p |
|---|---|---|---|---|---|---|---|
| Summit | X430 | N/A | N/A | N/A | N/A | N/A | N/A |
| | X440 | N/A | N/A | N/A | N/A | N/A | N/A |
| | X450-G2 | Y | Y | Y | Y | Y | Y |
| | X460 | Y | Y | Y | Y | Y | Y |
| | X460-G2 | Y | Y | Y | Y | Y | Y |
| | X480 | Y | Y | N | Y | Y | Y |
| | X670 | Y | Y | Y | Y | Y | Y |
| | X670-G2 | Y | Y | Y | Y | Y | Y |
| | X770 | Y | Y | Y | Y | Y | Y |

**Table 72: First Stage ACL Support Actions (continued)**

| Platform Family | Platform | Permit | Deny | Count | Replace-dot1p-value | qosprofile | Replace-dot1p |
|---|---|---|---|---|---|---|---|
| Black Diamond 8K | G48Xc | Y | Y | N | N | N | N |
| | G48Tc | Y | Y | N | N | N | N |
| | G48Te2 | Y | Y | N | N | N | N |
| | G24Xc | Y | Y | N | N | N | N |
| | 10G4Xc | Y | Y | N | N | N | N |
| | 10G8Xc | Y | Y | N | N | N | N |
| | S-G8Xc | Y | Y | N | N | N | N |
| | S-10G1Xc | Y | Y | N | N | N | N |
| | S-10G2Xc | Y | Y | N | N | N | N |
| | 8900-10G24X-c | Y | Y | N | Y | Y | Y |
| | xl-series | Y | Y | N | Y | Y | Y |
| | 8900-G96T-c | Y | Y | N | Y | Y | Y |
| | 8900-40G6X-c | Y | Y | Y | Y | Y | Y |
| | 8500-series | N/A | N/A | N/A | N/A | N/A | N/A |
| Black Diamond X8 | BDXA-series | Y | Y | Y | Y | Y | Y |
| | BDXB-series | Y | Y | Y | Y | Y | Y |
| | BDXC-series | Y | Y | Y | Y | Y | Y |

## Two-Stage Policy Example

The following example policy demonstrates how these new tokens can be used to create “user profiles” where each user is identified by source MAC address:

```
twostage_example1.pol:
 # First stage rules:

 entry firststage_1 {
if{
    ethernet-source-address 00:00:00:00:00:01;
} then {
    class-id 7;
}}
entry firststage_2 {
if {
    ethernet-source-address 00:00:00:00:00:02;
} then {
```

```
    class-id 8;
}}entry firststage_3 {
if {
    ethernet-source-address 00:00:00:00:00:03;
} then {
    class-id 7;
}}

 # Second stage rules:

entry secondstage_1 {
 if{
   class-id 7;
   destination-address 10.68.9.0/24;
} then {
    permit;
}}

entry secondstage_2 {
if {
    class-id 8;
    destination-address 10.68.0.0/16;
} then {
    permit;
}}entry secondstage_3 {
if {
} then {
    deny;
    }}
```

The above example policy would have the following resulting behavior:

1. MAC addresses 00:00:00:00:00:01 and 00:00:00:00:00:03 would be permitted to access subnet 10.68.9.0/24
2. MAC address 00:00:00:00:00:02 would be permitted to access subnet 10.68.0.0/16
3. All other traffic would be dropped.

## ACL Rule Syntax

An *ACL* rule entry consists of:

- A rule entry name, unique within the same ACL policy file or among Dynamic ACLs.
- Zero or more match conditions.
- Zero or one action (permit or deny). If no action is specified, the packet is permitted by default.
- Zero or more action modifiers.

Each rule entry uses the following syntax:

```
entry <ACLrulename>{
    if  {
        <match-conditions>;
    } then {
        <action>;
        <action-modifiers>;
    }
}
```

The following is an example of a rule entry:

```
entry  udpacl {
    if  {
        source-address 10.203.134.0/24;
        destination-address 140.158.18.16/32;
        protocol  udp;
        source-port 190;
        destination-port  1200 - 1250;
    } then {
        permit;
    }
}
```

An ACL rule is evaluated as follows:

- If the packet matches all the match conditions, the action and any action modifiers in the then statement are taken.
- For ingress ACLs, if a rule entry does not contain any match condition, the packet is considered to match and the action and any action modifiers in the rule entry's "then" statement are taken. For egress ACLs, if a rule entry does not contain any match condition, no packets will match. See Matching All Egress Packets on page 647 for more information.
- If the packet matches all the match conditions, and if there is no action specified in the then statement, the action permit is taken by default.
- If the packet does not match all the match conditions, the action in the then statement is ignored.

## Matching All Egress Packets

Unlike ingress *ACL*s, for egress ACLs you must specify either a source or destination address, instead of writing a rule with no match conditions.

For example, an ingress ACL deny all rule could be:

```
entry DenyAllIngress{
    if {
    } then {
        deny;
    }
}
```

The previous rule would not work as an egress ACL.

The following is an example of an egress ACL deny all rule:

```
entry DenyAllEgress{
    if {
        source-address 0.0.0.0/0;
    } then {
        deny;
    }
}
```

## Comments and Descriptions in ACL Policy Files

In *ACL* policy files, there are two types of textual additions that have no effect on the ACL actions: comments and descriptions. A comment is ignored by the policy manager and resides only in the policy file. Comments are not saved in the switch configuration and are not displayed by the show policy command. A description is saved in the policy manager and is displayed when the ACL is displayed.

- You can display the ACL using the following two commands:

  `show policy {`*`policy-name`* `|` **`detail`**`}`

  `show access-list {`**`any`** `|` **`ports`** *`port_list`* `|` **`vlan`** *`vlan_name`*`}` `{`**`ingress`** `|` **`egress`**`}`

  For example, the following policy, saved in the file denyping.pol, contains both a comment and a description:

```
# this line is a comment
@description "This line is a description for the denyping.pol"
entry ping_deny_echo-request {
    if {
        protocol icmp;
        icmp-type echo-request;
    } then {
        deny;
        count pingcount_deny;
    }
}
```

  Note that the description begins with the tag @description and is a text string enclosed in quotes.

- You can apply the policy to port 1 using the following command:

```
configure access-list denyping port 1
```

- and display the policy using the following command:

```
show policy denyping
```

  The output of this command is similar to the following:

```
Policies at Policy Server:
Policy: denyping
@description This line is a description for the denyping.pol
entry ping_deny_echo-request {
if match all {
protocol icmp ;
icmp-type echo-request ;
}
Then {
deny  ;
count pingcount_deny ;
}
}
Number of clients bound to policy: 1
Client: acl bound once
```

## Types of Rule Entries

In ExtremeXOS, each rule can be one of the following types:

- L2 rule—A rule containing only Layer 2 (L2) matching conditions, such as Ethernet MAC address and Ethernet type.
- L3 rule—A rule containing only Layer 3 (L3) matching conditions, such as source or destination IP address and protocol.
- L4 rule—A rule containing both Layer 3 (L3) and Layer 4 (L4) matching conditions, such as TCP/UDP port number.

## Match Conditions

You can specify multiple, single, or zero match conditions. If you do not specify a match condition, all packets match the rule entry. Commonly used match conditions are:

- `ethernet-source-address` *`mac-address mask`*—Ethernet source address
- `ethernet-destination-address` *`mac-address mask`*—Ethernet destination address and mask
- `ethernet-type` *`value`* {**`mask`** *`value`*}—Ethernet type, accepts an optional mask.
- `source-address` *`prefix`*—IP source address and mask
- `destination-address` *`prefix`*—IP destination address and mask
- `destination-port` *`value`* {**`mask`** *`value`*}—IP destination port, accepts optional mask
- `source-port` [*`value`* {**`mask`** *`value`*} | *`range`*]—TCP or UDP source port with optional mask or TCP or UDP source port range
- `destination-port` [*`port`* {**`mask`** *`value`*} | *`range`*]—TCP or UDP destination port with optional mask or TCP or UDP destination port range
- `ttl` *`value`* {**`mask`** *`value`*}—condition with optional mask that matches IPv4 Time-To-Live and IPv6 Hop Limit.
- `ip-tos` *`value`* {**`mask`** *`value`*}—this condition accepts optional masks
- `vlan-format`—matches packets based on their *VLAN* format. Can be one of the following values:
  - untagged—all untagged packets
  - single-tagged—all packets with only a single tag
  - double-tagged—all packets with a double tag
  - outer-tagged—all packets with at least one tag; for example, single tag or double tag
- `fragments`—matches any fragment of fragmented packet, including the first fragment
- `first-fragments`—matches only the first fragment of a fragmented packet.

Table 73 on page 654 describes all the possible match conditions.

## Actions

The actions are:

- `permit`—The packet is forwarded.
- `deny`—The packet is dropped.

The default action is permit, so if no action is specified in a rule entry, the packet is forwarded.

The following actions are supported on all platforms:

- `deny-cpu`—Prevents packets that are copied or switched to the CPU from reaching the CPU. The data-plane forwarding of these packets is unaffected. For example, use this action to match broadcast packets and prevent them from reaching the CPU, but still allow them to be flooded to other *VLAN* members. You can also use this action to match Spanning Tree Protocol packets and prevent them from reaching the CPU, and instead flood them to other VLAN members in certain configurations where Spanning Tree is enabled.
- `copy-cpu-off`—Prevents packets that are copied to the CPU from reaching the CPU. The data-plane forwarding of these packets is unaffected. For example, use this action to cancel the "mirror-cpu" action in another rule. This action does not prevent packets that are switched to the CPU (for example, broadcast, layer-3 unicast miss) from reaching the CPU.
- `copy-cpu-and-drop`—Overrides the above action to facilitate the above action in a "catch-all" rule. It sends matching packets only to the CPU.
- `add-vlan-id`—Adds a new outer VLAN id. If the packet is untagged it will add a vlan tag to the packet. If the packet is tagged, it will add additional VLAN tag. Only supported in VLAN Lookup stage (VFP).
- `replace-dscp-value`—Replaces the existing DSCP value of the packet
- `do-ipfix`—Records the matching packet. Can be used on both ingress and egress. Attempting to install a policy with this action on an unsupported chip will result in failure in HAL.
- `do-not-ipfix`—Cancels recording for the matching packet. Can be used to reduce demand on egress IPFIX capacity (and to reduce recording loss) during packet flooding situations. Attempting to install a policy with this action on an unsupported chip will result in failure in HAL.
- `redirect-port-copy-cpu-allowed`—Redirects a packet out of an output port, but does not enforce a requirement that Copy to CPU must be cancelled.
- `redirect-port-list-copy-cpu-allowed`—Redirects a packet out of an output port to a list of ports, but does not enforce a requirement that Copy to CPU must be cancelled.

## Action Modifiers

Additional actions can also be specified, independent of whether the packet is dropped or forwarded. These additional actions are called action modifiers. Not all action modifiers are available on all switches, and not all are available for both ingress and egress *ACL*s. The action modifiers are:

- `class-id` *`value 0-4095`*—Signifies that the rule will be installed in the LOOKUP stage access-list resource. Class-id range varies from platform to platform.
- `count` *`countername`*—Increments the counter named in the action modifier.
  - ingress—all platforms
  - egress—BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X450-G2, X460, X460-G2, X480, X670, X670-G2 and X770 series switches only. On egress, count does not work in combination with deny action.

**Note**
On egress, count does not work in combination with deny action in some platforms

- `add-vlan-id`—Adds a new outer *VLAN* id. If the packet is untagged it will add a vlan tag to the packet. If the packet is tagged, it will add additional VLAN tag. Only supported in VLAN Lookup stage (VFP).
- `byte-count` *`byte counter name`*—Increments the byte counter named in the action modifier (BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only).
- `packet-count` *`packet counter name`*—Increments the packet counter named in the action modifier (BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl- and xm-series modules, and Summit family switches only).
- `log`—Logs the packet header.
- `log-raw`—Logs the packet header in hex format.
- `meter` *`metername`*—Takes action depending on the traffic rate. (Ingress and egress meters are supported on the platforms listed for these features in the *Feature License Requirements* document.
- `mirror`—Rules that contain mirror as an action modifier will use a separate slice.
- `mirror-cpu`—Mirrors a copy of the packet to the CPU in order to log it. For Summit X460 and E4G400, it is supported in ingress/egress. In all other platforms, it is supported only in ingress.
- `qosprofile` *`qosprofilename`*—Forwards the packet to the specified *QoS* profile.
  - ingress—all platforms
  - egress—does not forward the packets to the specified qosprofile. If the action modifier “replace-dot1p” is present in the ACL rule, the dot1p field in the packet is replaced with the value from associated qosprofile. (BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules,E4G-200 and E4G-400 cell site routers, and Summit X460, X460-G2, X480, X670, X670-G2, and X770 series switches only).
- `redirect` *`ipv4 addr`*—Forwards the packet to the specified IPv4 address (BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only).
- `redirect-port` *`port`*—Overrides the forwarding decision and changes the egress port used. If the specified port is part of a load share group then this action will apply the load sharing algorithm. (BlackDiamond X8 series switches, BlackDiamond 8000a-, c-, e-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit family switches only.)
- `redirect-port-list` *`port_list`*—Supports multiple redirect ports as arguments. When used in an ACL, matching packets are now redirected to multiple ports as specified in the ACL while overriding the default forwarding decision. Maximum number of ports that can be mentioned in this list is 64. (Summit X440, X460, X480, X670, X770, E4G-200, E4G-400, BlackDiamond 8K - 8900-G96T-c, 8900-10G24X-c, 8900-G48T-xl, 8900-G48X-xl, 8900-10G8X-xl, 8900-40G6X-xm, BlackDiamond X8.)
- `redirect-port-no-sharing` *`port`*—Overrides the forwarding decision and changes the egress port used. If the specified port is part of a load share group then this action overrides the load sharing algorithm and directs matching packets to only this port. (BlackDiamond X8 and 8000 series switches, E4G-200 and E4G-400 cell site routers, and Summit family switches.)
- `redirect-name` *`name`*—Specifies the name of the flow-redirect that must be used to redirect matching traffic. (BlackDiamond X8 and 8000 series switches, E4G-200 and E4G-400 cell site routers, and Summit family switches except X430.)

- `replace-dscp`—Replaces the packet's DSCP field with the value from the associated QoS profile.
  - ingress—BlackDiamond X8, 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only.
  - egress—BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 series switches only.
- `replace-dot1p`—Replaces the packet's 802.1p field with the value from the associated QoS profile.
  - ingress—BlackDiamond X8, 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only.
  - egress—BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 series switches only.
- `replace-dot1p-value` *`value`*—Replaces the packet's 802.1p field with the value specified without affecting the QoS profile assignment.
  - ingress—BlackDiamond X8, 8000 c-, e-, xl-, and xm-series modules and the Summit family switches only.
  - egress—BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X450-G2, X460, X460-G2, X480, X670, X670-G2, and X770 series switches only.
- `replace-ethernet-destination-address` *`mac-address`*—Replaces the packet's destination MAC address; this is applicable only to layer-2 forwarded traffic. (BlackDiamond X8, 8000 c-, e-, xl-, and xm-series modules and the Summit family switches only.)

### *Counting Packets and Bytes*

When the *ACL* entry match conditions are met, the specified counter is incremented.

- The counter value can be displayed by the command:

  `show access-list counter {`*`countername`*`} {`**`any`** `|` **`ports`** *`port_list`* `|` **`vlan`** *`vlan_name`*`} {`**`ingress`** `|` **`egress`**`}`

  BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules and Summit family switches can use ACL byte counters as an alternative to ACL packet counters. See ACL Byte Counters on page 664 for more information.

  **Note**
  On BlackDiamond X8 series switches, BlackDiamond 8800 series switches and Summit family switches, the maximum number of packets that can be counted with token packet-count or count is 4,294,967,296. On the same switches, the maximum number of bytes that can be counted with byte-count is also 4,294,967,296 which is equivalent to 67,108,864 packets that are sized at 64 bytes.

  **Note**
  Each packet will increment only one counter in the egress direction. When there are multiple ACLs with action "count" applied in the port, only single counter based on the slice priority will work.

### *Logging Packets*

Packets are logged only when they go to the CPU, so packets in the fastpath are not automatically logged. You must use both the mirror-cpu action modifier and the log or log-raw action modifier if you want to log both slowpath and fastpath packets that match the *ACL* rule entry. Additionally, Kern.Info messages (or Kern.Card.Info on SummitStack) are not logged by default. You must configure an *EMS (Event Management System)* filter to log these messages, for example, `configure log filter DefaultFilter add event kern.info`. See the Status Monitoring and Statistics chapter for information about configuring EMS.

### *Metering Packets*

The `meter` *`metername`* action modifier applies a meter to the traffic defined by an *ACL*. For more information, see Meters on page 736.

### *Mirroring Packets*

You must enable port-mirroring on your switch. If you attempt to apply a policy that requires port-mirroring, the mirror action will be disabled until you enable the port-mirroring.

On the BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules and Summit X460, X480, X670, X450G2, X460G2, X670G2 and X770 series switches, mirroring can be configured on the same port as egress ACLs. Mirroring can send packets to port x and you can install your rule at egress port x, and the rule should match your mirrored traffic.

### *Redirecting Packets*

Packets are forwarded to the IPv4 address specified, without modifying the IP header (except the TTL is decremented and the IP checksum is updated). The IPv4 address must be in the IP ARP cache, otherwise the packet is forwarded normally. Only fast path traffic can be redirected. This capability can be used to implement Policy-Based Routing.

You may want to create a static ARP entry for the redirection IP address, so that there will always be a cache entry. See Policy-Based Routing on page 704 for more information.

### *Replacing DSCP or 802.1p Fields*

Specify a *QoS* profile for matching packets. The field values are replaced with the value associated with that profile. In the following example, DiffServ replacement is configured such that QP8 is mapped to code point 56. Matching packets are sent to QP8, and the DSCP value in the packet is set to 56.

```
entry voice_entry {
if {
source-address 2.2.2.2/32;
} then {
qosprofile qp8;
replace-dscp;
}
}
```

See Quality of Service on page 724 for more details about QoS profiles, and 802.1p and DSCP replacement.

## ACL Rule Syntax Details

The following table lists the match conditions that can be used with *ACL*s, and whether the condition can be used for ingress ACLs only, or with both ingress and egress. The conditions are not case-sensitive; for example, the match condition listed in the table as TCP-flags can also be written as tcp-flags. Within the following table are five different data types used in matching packets. The first table below lists general match conditions that apply to all traffic, unless otherwise noted. The second table lists the data types and details on using them.

**Table 73: ACL Match Conditions**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `ethernet-type` *`number`* | Ethernet packet type. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): ETHER-P-IP (0x0800), ETHER-P-8021Q (0x8100), ETHER-P-IPV6 (0x86DD). | Ethernet/ Ingress and Egress |
| `ethernet-source-address` *`mac-address`* | Ethernet source MAC address | Ethernet/ Ingress and Egress |
| `ethernet-source-address` *`mac-address`* `mask` *`mask`* or `ethernet-source-address` *`mac-address/mask`* | Ethernet source MAC address and mask. The mask is optional, and is in the same format as the MAC address, for example: `ethernet-source-address 00:01:02:03:01:01 mask ff:ff:ff:ff:00:00` or `ethernet-source-address 00:01:02:03:01:01 / ff:ff:ff:ff:00:00` Only those bits of the MAC address whose corresponding bit in the mask is set to 1 will be used as match criteria. So, the example above will match `00:01:02:03:xx:xx`. If the mask is not supplied then it will be assumed to be ff:ff:ff:ff:ff:ff. In other words, all bits of the MAC address will be used for matching. | Ethernet/ Ingress and Egress |
| `ethernet-destination-address` *`mac-address`* | Ethernet destination MAC address | Ethernet/ Ingress and Egress |

[9] However, packets using the Ethernet type for VMANs, 0x88a8 by default, are handled by VMAN ACLs.

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `ethernet-destination-address` *`mac-address`* `mask` *`mask`* or `ethernet-source-address` *`mac-address/mask`* | Ethernet destination MAC address and mask. The mask is optional, and is in the same format as the MAC address, for example: `ethernet-destination-address 00:01:02:03:01:01 mask ff:ff:ff:ff:00:00` or `ethernet-destination-address 00:01:02:03:01:01 / ff:ff:ff:ff:00:00` Only those bits of the MAC address whose corresponding bit in the mask is set to 1 will be used as match criteria. So, the example above will match `00:01:02:03:xx:xx`. If the mask is not supplied then it will be assumed to be ff:ff:ff:ff:ff:ff. In other words, all bits of the MAC address will be used for matching. | Ethernet/ Ingress and Egress |
| `source-address` *`prefix`* | IP source address and mask. Use either all IPv4 or all IPv6 addresses in an ACL. On BD8K, BDX8 and Summit series switches, using arbitrary mask arguments is supported. Masks are not restricted to be of a subnet type. Examples of arbitrary IPv4 and IPv6 masks include 10.22.3.4 and 1:0:0:ffff:2:4. The 1s in the mask indicate the corresponding bits of the source-address that should be used as part of the match criteria. | All IP/Ingress and Egress |
| `destination-address` *`prefix`* | IP destination address and mask. On BD8K, BDX8 and Summit series switches, using arbitrary mask arguments is supported. Masks are not restricted to be of a subnet type. Examples of arbitrary IPv4 and IPv6 masks include 10.22.3.4 and 1:0:0:ffff:2:4. The 1s in the mask indicate the corresponding bits of the destination-address that should be used as part of the match criteria. | All IP/Ingress and Egress |
| `source-port` {*`number`*\|*`range`*} | TCP or UDP source port. You must also specify the protocol match condition to determine which protocol is being used on the port, any time you use the this match condition. In place of the numeric value, you can specify one of the text synonyms listed under destination port. If no source-port is specified, the default source-port is "any." | TCP, UDP/ Ingress and Egress |
| `source-port number { mask` *`value`* `}` | TCP or UDP port and mask. The mask is optional, and it can be decimal value or a hexadecimal value. | TCP,UDP/ Ingress and Egress |

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `destination-port {number\|range}` | TCP or UDP destination port. You must also specify the protocol match condition to determine which protocol is being used on the port, any time you use the this match condition. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): afs(1483), bgp(179), biff(512), bootpc(68), bootps(67), cmd(514), cvspserver(2401), *DHCP (Dynamic Host Configuration Protocol)*(67), domain(53), eklogin(2105), ekshell(2106), exec(512), finger(79), ftp(21), ftp-data(20), http(80), https(443), ident(113), imap(143), kerberos-sec(88), klogin(543), kpasswd(761), krb-prop(754), krbupdate(760), kshell(544), ldap(389), login(513), mobileip-agent(434), mobileip-mn(435), msdp(639), netbios-dgm(138), netbios-ns(137), netbios-ssn(139), nfsd(2049), nntp(119), ntalk(518), ntp(123), pop3(110), pptp(1723), printer(515), radacct(1813), radius(1812), rip(520), rkinit(2108), smtp(25), snmp(161), snmptrap(162), snpp(444), socks(1080), ssh(22), sunrpc(111), syslog(514), tacacs-ds(65), talk(517), telnet(23), tftp(69), timed(525), who(513), xdmcp(177), zephyr-clt(2103), or zephyr-hm(2104). | TCP, UDP/ Ingress and Egress |
| `destination-port number {mask value}` | TCP or UDP port and mask. The mask is optional, and it can be decimal value or a hexadecimal value. Only those bits of the destination-port whose corresponding bit in the mask is set to 1 will be used as match criteria. | TCP,UDP/ Ingress and Egress |
| TCP-flags *bitfield* | TCP flags. Normally, you specify this match in conjunction with the protocol match statement. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): ACK(0x10), FIN(0x01), PUSH(0x08), RST(0x04), SYN(0x02), URG(0x20), SYN_ACK(0x12). | TCP/Ingress and Egress |
| `IGMP-msg-type number` | *IGMP (Internet Group Management Protocol)* message type. Possible values and text synonyms: v1-report(0x12), v2-report(0x16), v3-report(0x22), V2-leave (0x17), or query(0x11). | IGMP/Ingress only |

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `ICMP-Type` *number* | *ICMP (Internet Control Message Protocol)* type field. Normally, you specify this match in conjunction with the protocol match statement. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): echo-reply(0), echo-request(8), info-reply(16), info-request(15), mask-request(17), mask-reply(18), parameter-problem(12), redirect(5), router-advertisement(9), router-solicit(10), source-quench(4), time-exceeded(11), timestamp(13), timestamp-reply(14), or unreachable(3), v6-echo-request(128), v6-echo-reply(129), v6-mld-query(130), v6-mld-report(131), v6-mld-reduction(132), v6-router-soliciation(133), v6-router-advertisement(134), v6-neighbor-solicitation(135), v6-neighbor-advertisement(136), v6-redirect(137), v6-node-info-query(139), v6-node-info-reply(140), v6-unreachable(1), v6-packet-too-big(2), v6-time-exceeded(3), v6-parameter-problem(4), v6-echo-request(128), v6-echo-reply(129), v6-mld-query(130), v6-mld-report(131), v6-mld-reduction(132), v6-router-soliciation(133), v6-router-advertisement(134), v6-neighbor-solicitation(135), v6-neighbor-advertisement(136), v6-redirect(137), v6-node-info-query(139), v6-node-info-reply(140) v6-unreachable(1), v6-packet-too-big(2), v6-time-exceeded(3), v6-parameter-problem(4), v6-echo-request(128), v6-echo-reply(129), v6-mld-query(130), v6-mld-report(131), v6-mld-reduction(132), v6-router-soliciation(133), v6-router-advertisement(134), v6-neighbor-solicitation(135), v6-neighbor-advertisement(136), v6-redirect(137), v6-node-info-query(139), v6-node-info-reply(140). | ICMP/Ingress only |

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `ICMP-Code` *number* | ICMP code field. This value or keyword provides more specific information than the icmp-type. Because the value's meaning depends upon the associated icmp-type, you must specify the icmp-type along with the icmp-code (only available in IPv4). In place of the numeric value, you can specify one of the following text synonyms (the field values also listed); the keywords are grouped by the ICMP type with which they are associated:<br>Parameter-problem:<br>ip-header-bad(0), required-option-missing(1)<br>Redirect:<br>redirect-for-host (1), redirect-for-network (2), redirect-for-tos-and-host (3), redirect-for-tos-and-net (2)<br>Time-exceeded:<br>ttl-eq-zero-during-reassembly(1), ttl-eq-zero-during-transit(0)<br>Unreachable:<br>communication-prohibited-by-filtering(13), destination-host-prohibited(10), destination-host-unknown(7), destination-network-prohibited(9), destination-network-unknown(6), fragmentation-needed(4), host-precedence-violation(14), host-unreachable(1), host-unreachable-for-TOS(12), network-unreachable(0), network-unreachable-for-TOS(11), port-unreachable(3), precedence-cutoff-in-effect(15), protocol-unreachable(2), source-host-isolated(8), source-route-failed(5) | IPv4 only/ ICMP/Ingress only |
| `source-sap` | SSAP is a 1 byte field with possible values 0-255 decimal. The value can be specified in decimal or hexadecimal. The SSAP field can be found at byte offset 15 in 802.3 SNAP and LLC formatted packets. (Available on Summit family switches, SummitStack, and BlackDiamond 8000 c-, e-, xl-, and xm-series modules only.) | Ethernet/ Ingress Only |
| `destination-sap` | DSAP is a 1 byte field with possible values 0-255 decimal. The value can be specified in decimal or hexadecimal. The DSAP field can be found at byte offset 14 in 802.3 SNAP and LLC formatted packets. (Available on Summit family switches, SummitStack, and BlackDiamond 8000 c-, e-, xl-, and xm-series modules only.) | Ethernet/ Ingress Only |
| `snap-type` | SNAP type is a 2 byte field with possible values 0-65535 decimal. The value can be specified in decimal or hexadecimal. The SNAP type field can be found a byte offset 20 in 802.3 SNAP formatted packets. (Available on Summit family switches, SummitStack, and BlackDiamond 8000 c-, e-, xl-, and xm-series modules only.) | Ethernet/ Ingress Only |
| `ttl number {`**`mask`** *value*`}` | Time To Live with mask.The mask is optional, and it can be decimal value or a hexadecimal value.Only those bits of the ttl whose corresponding bit in the mask is set to 1 will be used as match criteria.This can be used to match IPv4 Time-To-Live and IPv6 Hop Limit. | All IP/Ingress and Egress. |

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `IP-TOS` *number* | IP TOS field. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): minimize-delay 16 (0x10), maximize-reliability 4(0x04), minimize-cost2 (0x02), and normal-service 0(0x00). | All IP/Ingress and Egress |
| `IP-TOS number {`*mask value*`}` | IP-TOS and mask.The mask is optional, and it can be decimal value or a hexadecimal value.Only those bits of the IP-TOS whose corresponding bit in the mask is set to 1 will be used as match criteria. | All IP/Ingress and Egress |
| `dscp` *value* | DSCP field. In place of the value, you can specify one of the DSCP numeric values (for example, 8, 16, or 24). | All IP/Ingress and Egress |
| `fragments` | BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only—IP fragmented packet including first fragment. FO = 0 (FO = Fragment Offset in IP header) | All IP, no L4 rules/Ingress only |
| `first-fragment` | Matches only first fragmented packet. FO==0. | All IP/Ingress only |
| `protocol` *number* | IP protocol field. For IPv6, this matches the Next Header field in the packet. In place of the numeric value, you can specify one of the following text synonyms (the field values are also listed): egp(8), gre(47), icmp(1), igmp(2), ipip(4), Ipv6 over ipv4(41), ospf(89), pim(103), rsvp(46), st(5), tcp(6), or udp(17). | All IP/Ingress and Egress |
| `vlan-format` *format* | *VLAN*-format matches packets based on its vlan format. Can be one of the 4 values: untagged - will match all untagged packets single-tagged - will match all packets with only single tag. double-tagged - will match all packets with double tag outer-tagged - will match all packets with at least one tag ex. single tag or double tag. | Ethernet/ Ingress and Egress |
| `vlan-id` *number* | Matches the VLAN tag number or the VLAN ID which is given to a VLAN when created. The ACL rule can only be applied to ports or any, and not VLANs.<br>The following restriction applies to all platforms:<br>The vlan-id match condition matches on the “outer” tag of a VMAN.The vlan-id ACL keyword can be used in egress ACL. | Ethernet/ Ingress and Egress |
| `vlan-id number {`**`mask`** *value*`}` | VLAN-id and mask.The mask is optional, and it can be decimal value or a hexadecimal value.Only those bits of the Vlan tag Number or vlan id whose corresponding bit in the mask is set to 1 will be used as match criteria. | Ethernet/ Ingress and Egress |
| `dot1p` *priority tag* | Creates an ACL with 802.1p match conditions, allowing the ACL to take action based on the VLAN tag priority. (Available on all platforms.) | All IP/Ingress |

[10] See the section Fragmented packet handlingfor details,
[11] See the section IPv6 Traffic with L4 Match Conditionsfor details about specifying a protocol/port match with IPv6.

**Table 73: ACL Match Conditions (continued)**

| Match Conditions | Description | Applicable IP Protocols/ Direction |
|---|---|---|
| `arp-sender-address` *`prefix`* and `arp-target-address` *`prefix`* | Matches the ARP sender protocol address and target protocol address respectively. *`prefix`* `=>` *`IPv4 address / mask length`*. They cannot be combined with an Ethernet-source-address or Ethernet-destination-address in the same rule. They can be used only when the ACL hardware database is configured to be internal for those platforms that support “external-table” ACL databases. (for example, Summit X480 switches and BlackDiamond 8900 and X8 xl-series modules). (Available on BlackDiamond X8 series switches, BlackDiamond 8800 switches and Summit family switches only.) | ARP packets/ Ingress |
| `cvid` | Use this match criteria in the following scenarios: Tagged VMAN ports: installing an ACL matching “cvid” on ingress or egress will match the inner vlan-id of a double tagged packet on a tagged VMAN port. Untagged VMAN ports: installing an ACL matching “cvid” on ingress or egress will match the single VLAN tag on an untagged VMAN port. CEP VMAN ports (with or without VPLS): installing an ACL matching “cvid” on ingress or egress will match the single VLAN tag on a CEP VMAN port (without translation). CEP VMAN ports with cvid translation (with or without translation): installing an ACL matching “cvid” on ingress will match the post-translation cvid. Installing an ACL matching “cvid” on egress will match the post-translation cvid. | Ethernet/ Ingress and Egress |
| `class-id` | This match condition can be specified on any rule within a policy file or within a list of dynamic access-lists. A rule cannot both match a class-id and specify a class-id as an action. When a “class-id” match criteria is specified, the associated rule will be programmed into the normal “INGRESS stage” access-list hardware resource. The range of valid class-id values varies per platform. | Ingress only. |
| unknown-l2-unicast | Matches the unknown L2 unicast packets | Ingress only |
| unknown-l2-multicast | Matches the unknown L2 multicast packets | Ingress only |
| unknown-l3-multicast | Matches the unknown L3 multicast packets | Ingress only |
| l2-da-hit | Matches the known L2 unicast packets | Ingress only |

Note

When you use a configured ACL that contains a match condition with any *mac-address*, IGMP snooping stops working and IGMP joins are flooded to all ports in a VLAN. When you unconfigure the ACL, IGMP joins stop flooding.

**Note**

An ACL that matches the *EAPS (Extreme Automatic Protection Switching)* `ethernet-destination-address (00:e0:2b:00:00:04)` or `ethernet-source-address (00:e0:2b:00:00:01)` match condition with the permit action should not be applied to an EAPS master node on EAPS ring ports. Doing so causes an EAPS PDU loop. For the EAPS master node, you should use the `copy-cpu-and-drop` action with either of these match conditions. For an EAPS transit node, use the `permit` action with either of these match conditions. This applies only to BlackDiamond 8000 series modules and Summit switches.

**Note**

Directed ARP response packets cannot be blocked with ACLs from reaching the CPU and being learned on BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules and the Summit family switches.

Along with the data types described in the following table, you can use the operators <, <=, >, and >= to specify match conditions. For example, the match condition, `source-port 190`, will match packets with a source port greater than 190. Be sure to use a space before and after an operator.

**Table 74: ACL Match Condition Data Types**

| Condition Data Type | Description |
|---|---|
| prefix | IP source and destination address prefixes. To specify the address prefix, use the notation prefix/prefix-length. For a host address, prefix-length should be set to 32. |
| number | Numeric value, such as TCP or UDP source and destination port number, IP protocol number. |
| range | A range of numeric values. To specify the numeric range, use the notation: number - number |
| bit-field | Used to match specific bits in an IP packet, such as TCP flags and the fragment flag. |
| mac-address | 6-byte hardware address. |

### *IPv6 Traffic with L4 Match Conditions*

If you apply an *ACL* policy using ACL that specifies L4 conditions, both IPv4 and IPv6 traffic will be matched. For example, the following ACL will match both IPv4 and IPv6 TCP packets that have their L4 destination port in the range of 120–150:

```
if {
protocol tcp;
destination-port 120 - 150;
}
Then {
permit;
count destIp;
}
}
```

*Fragmented Packet Handling*

Two keywords are used to support fragmentation in ACLs:

- fragments—FO field = 0 (FO means the fragment offset field in the IP header). BlackDiamond X8 series switches, BlackDiamond 8000 c-, e-, xl-, and xm-series modules, and Summit family switches only. This will match first fragment also (packets with FO = 0).
- first-fragments—FO == 0.

**Policy file syntax checker**

The **fragments** keyword cannot be used in a rule with L4 information. The syntax checker will reject such policy files.

The following rules are used to evaluate fragmented packets or rules that use the **fragments** or **first-fragments** keywords.

With no keyword specified, processing proceeds as follows:

- An L3-only rule that does not contain either the fragments or first-fragments keyword matches any IP packets.
- An L4 rule that does not contain either the fragments or first-fragments keyword matches non-fragmented or initial-fragment packets.

With the **fragments** keyword specified:

- An L3-only rule with the fragments keyword only matches fragmented packets.
- An L4 rule with the fragments keyword is not valid (see above).

With the **first-fragments** keyword specified:

- An L3-only rule with the first-fragments keyword matches initial fragment packets.
- An L4 rule with the first-fragments keyword matches initial fragment packets.

*Wide Key ACLs*

This feature allows the use of a 362-bit double-wide match key instead of a standard 181-bit single-wide key to be used with match conditions. A double-wide match key allows you to add more match conditions to an *ACL*. It also allows matching on a full destination-source IPv6 address.

The feature does not add any new match conditions but rather allows you to add additional condition combinations to any single-wide condition combination that is already supported. The existing supported condition combinations are described in the following table through the following table. The double wide condition combinations that can be appended under the set union operation to the single-wide condition combinations are as follows:

- OVID, DIP, SIP, IpInfo(First-Fragment,Fragments), IP-Proto, DSCP, TCP-Flag, L4SP, L4DP
- SIPv6, IP-Proto, DSCP, TCP-Flag, L4SP, L4DP

For example, your single-wide mode supports condition combination A, B, and C, and the double-wide mode adds condition combinations D1 and D2. Then in a single-wide mode, the conditions of your rule should be a subset of either {A}, or {B}, or {C} and in a double-wide mode, the conditions of your rule should be a subset of either {A U D1}, or {A U D2}, or {B U D1}, or {B U D2}, or {C U D1}, or {C U D2}.

The platforms that support this feature can operate either in double-wide mode or in the current single-wide mode. A individual switch or module cannot be configured to operate in a mixed double and single-wide mode. However, a BlackDiamond 8800 chassis or a SummitStack can have a mixture of modules and switches with some of them operating in a single-wide mode and some in a double-wide mode.

#### Limitations

Following are limitations associated with this feature:

- Double-wide mode provides richer condition combinations. However, when in a double-wide mode, you can install only one half as many rules into the internal ACL TCAM as you can when in a single-wide mode.
- Double-wide mode is supported only by internal TCAM hardware. External TCAM hardware does not support this feature and thus is not applicable to external TCAM ACLs.
- Only ingress ACLs support this feature. Egress and external ACLs do not support it.
- BlackDiamond 8000 10G24Xc2 and 10G24Xc module can operate in double-wide mode only in slices 8, 9, 10, and 11. Therefore, when you configure double-wide mode on these platforms, they operate in double mode on slices 8 through 11 and in single mode on slices 0 through 7.

#### Supported Platforms

Wide Key ACLs are available on BlackDiamond X8 Series Switches, BlackDiamond 8000 c- , xl-, and xm-series modules and Summit X460, X480, X670, X460-G2, X670-G2 and X450-G2 and X770 switches.

### Configuring Wide Key ACL Modes

To configure the TCAM width of a slot, switch in a SummitStack or stand-alone switch, use the following command:

```
configure access-list width [double | single] [slot slotNo | all]
```

You must reboot for the configuration to take effect.

If you attempt to configure a double wide mode on a slot or switch that does not support it, an error message is displayed.

When switching from single wide key mode to double wide key mode and rebooting, the following conditions apply:

- Configurations that have less that one-half the number of *ACL* rules that the single wide key mode supports, reboot successfully.
- Configurations that have more than one-half the number of ACL rules that the single wide key mode supports, fail after the reboot.

When switching from double wide key mode to single wide key mode and rebooting, the following conditions apply:

- Configurations that do not use the additional condition combinations that double wide key mode offers, reboot successfully.
- Configurations that use the additional condition combinations that the double wide key mode offers, fail after the reboot.

To display the wide key mode settings, use the following command:

```
show access-list width [slot slotNo | all]
```

## Layer-2 Protocol Tunneling ACLs

Three *ACL* match conditions and one ACL action interoperate with vendor-proprietary Layer-2 protocol tunneling on the platforms listed for this feature in the *Feature License Requirements* document.

The following fields within 802.3 Subnetwork Access Protocol (SNAP) and LLC formatted packets can be matched:

- Destination service access point (SAP)
- Source SAP

The following field can be matched within Subnetwork Access Protocol (SNAP) packets only:

- SNAP type

The following ACL action is added to the specified switches:

- Replacement of the Ethernet MAC destination address

This action replaces the destination MAC address of any matching Layer-2 forwarded packets on the supported platforms. This action can be used to effectively tunnel protocol packets, such as *STP (Spanning Tree Protocol)*, across a network by replacing the well-known protocol MAC address with a different proprietary or otherwise unique MAC address. After tunnel egress, the MAC destination address can be reverted back to the well-known MAC address.

Note

The "replace-ethernet-destination-address" action applies only to Layer-2 forwarded packets.

## ACL Byte Counters

An *ACL* byte counter associated with a particular rule, either dynamic or static, shows how many bytes of traffic have matched that ACL rule. You can use ACL byte counters as an alternative to packet counters on the platforms listed for this feature in the *Feature License Requirements* document.

A new ACL action token has been added to associate a byte counter with an ACL, and a new corresponding token for a packet counter.

Following are the two new tokens:

```
byte-count byte counter name
```

```
packet-count packet counter name
```

An ACL rule specifying both packet and byte counter is rejected.

Below is an example of an ACL rule that uses a byte counter:

```
entry CountBytes {
    if {
        ethernet-source-address 00:aa:00:00:00:10;
    } then {
        byte-count CountBytes;
        permit;
```

```
    }
}
```

Below are two examples of ACL rules that use packet counters. The "packet-count" token is a synonym of the existing "count" token.

```
entry CountPacket1 {
    if {
        ethernet-source-address 00:aa:00:00:00:10;
    } then {
        count CountPacket1;
        permit;
    }
}

entry CountPacket2 {
    if {
        ethernet-source-address 00:aa:00:00:00:10;
    } then {
        packet-count CountPacket2;
        permit;
    }
}
```

The output of the `show access-list counter` and `show access-list dynamic counter` commands has been changed to include a new "Byte Count" column in addition to the "Packet Count" column. When a rule utilizes a byte counter, the "Byte Count" field is incremented and the "Packet Count" field stays at zero. If a rule utilizes a packet counter, the "Packet Count" field is incremented and the "Byte Count" field stays at zero.

**Note**
Byte counters and packet counters cannot be used at the same time in the same rule.

## Dynamic ACLs

Dynamic *ACL*s are created using the CLI. They use a similar syntax and can accomplish the same actions as single rule entries used in ACL policy files. More than one dynamic ACL can be applied to an interface, and the precedence among the dynamic ACLs can be configured. By default, the priority among dynamic ACLs is established by the order in which they are configured.

**Note**
Dynamic ACLs have a higher precedence than ACLs applied using a policy file.

The steps involved in using a dynamic ACL on an interface are:

- Creating the Dynamic ACL Rule
- Configuring the ACL Rule on the Interface.
- Configuring ACL Priority
- Network-Zone Support in ACLs

## Creating the Dynamic ACL Rule

Creating a dynamic *ACL* rule is similar to creating an ACL policy file rule entry. You specify the name of the dynamic ACL rule, the match conditions, and the actions and action-modifiers. You can configure a dynamic ACL to be persistent or non-persistent across system reboots. User-created access-list names are not case sensitive. The match conditions, actions, and action-modifiers are the same as those that are available for ACL policy files (see ACL Rule Syntax on page 646). In contrast to the ACL policy file entries, dynamic ACLs are created directly in the CLI. Use the following command to create a dynamic ACL:

```
create access-list dynamic_rule conditions actions {non_permanent}
```

As an example of creating a dynamic ACL rule, compare an ACL policy file entry with the CLI command that creates the equivalent dynamic ACL rule.

The following ACL policy file entry will drop all *ICMP* echo-requests:

```
entry  icmp-echo {
    if  {
        protocol  icmp;
        icmp-type  echo-request;
    } then {
        deny;
    }
}
```

To create the equivalent dynamic ACL rule, use the following command:

```
create access-list icmp-echo "protocol icmp;icmp-type echo-request"
"deny"
```

Notice that the *conditions* parameter is a quoted string that corresponds to the match conditions in the if { ... } portion of the ACL policy file entry. The individual match conditions are concatenated into a single string. The *actions* parameter corresponds to the then { ... } portion of the ACL policy file entry.

From the command line you can get a list of match conditions and actions by using the following command:

```
check policy attribute {attr}
```

The ACL rule shown in the example will be saved when the save command is executed, because the optional keyword **non-permanent** was not configured. This allows the rule to persist across system reboots.

Note also that the sample ACL rule does not specify an application to which the rule belongs. The default application is CLI.

### Limitations

Dynamic ACL rule names must be unique, but can be the same as used in a policy file-based ACL. Any dynamic rule counter names must be unique. CLEAR-FLow rules can be specified only in policy files and therefore apply only to rules created in a policy file.

## Configuring the ACL Rule on the Interface

After a dynamic *ACL* rule has been created, it can be applied to a port, *VLAN*, or to the wildcard any interface. When the ACL is applied, you specify the precedence of the rule among the dynamic ACL rules. To configure the dynamic ACL rule on an interface, use the following command:

```
configure access-list add dynamic_rule [ [[first | last] {priority
p_number} {zone zone} ] | [[before | after] rule] | [ priority p_number
{zone zone} ]] [ any | vlan vlan_name | ports port_list ] {ingress |
egress}
```

To remove a dynamic ACL from an interface, use the following command:

```
configure access-list delete ruleName [ any | vlan vlan_name | ports
port_list | all] {ingress | egress}
```

An ACL can be created to be used when an edge port detects a loop. This ACL acts to block looped frames while allowing the port to remain in a forwarding state rather than shutting down. To configure a dynamic ACL for blocking looped *STP* BPDUs on port 6, for example, use the following:

```
create access-list bpdu1 "ethernet-destination-address    \
                          01:80:C2:00:00:00;" "deny; count bpdu1"
conf access-list add "bpdu1" first ports 6 ingress
```

To configure a dynamic ACL for blocking PVST frames on port 6, use the following:

```
create access-list bpdu2 "ethernet-destination-address \
                                 01:00:0c:cc:cc:cd;" "deny; count bpdu2"
 conf access-list add "bpdu2" first ports 6 ingress
```

To unconfigure the *STP* ACL, use the following:

```
conf access-list del "bpdu1" ports 6
del access-list "bpdu1"
```

### *Configuring ACLs on a Management Port*

Hardware *ACL* support is not possible on the management port. Untagged packets that are received on the management port are processed in software and can be filtered using ACLs. ACLs applied to the management port/vlan are installed only in software and not in the hardware.

For example, to block an *ICMP* echo-request on a management port, use the following:

```
create access-list echo "protocol icmp; icmp-type echo-request;" "deny; count echo"
conf access-list add "echo" first vlan "Mgmt" ingress
```

To unblock ICMP echo request on a management port, use the following:

```
conf access-list del "echo" vlan "Mgmt"
del access-list "echo"
```

To show ACL dropped packet counters, use the following command:

```
show access-list dynamic counter
```

## Configuring ACL Priority

Management of *ACLs* is flexible, with configurable priority for dynamic ACLs. This includes ACLs inserted by internal and external applications, as well as those inserted using the CLI. The priority is assigned by a system of zones, and within zones by numeric codes.

Zones are of two types:

- System Space—The System Space zones include the following:
  - SYSTEM_HIGH—This zone always has the highest priority.
  - SYSTEM_LOW—This zone always has the lowest priority.

  The priorities cannot be changed.

  No configuration is allowed by the user into System Space.

  Hal is the only application in a System Space zone.
- User Space—The User Space zones include the following:
  - DOS—This is the denial of service zone.
  - SYSTEM—This is the zone for applications that require a CPU-copy or mirror and for redirect ACLs.
  - SECURITY—This is the zone for ACLs installed by security appliances and internal security processes.

  User Space zones consist of default zones and created zones. Default zones group like functions and cannot be deleted.

  The administrator has the ability to create new zones and configure the priority of both default and created zones. See Configuring User Zones on page 669 for discussion of created zones and applications. Applications insert ACLs into zones.

To view both System Space and User Space zones, use the `show access-list zone` command.

The following table shows the priority of System Space zones and User Space zones together with the default assignments and priority of applications by zone.

**Table 75: Default Assignment and Priority of Applications, by Zone**

| Zone/Default Application | Default Priority | Platform |
|---|---|---|
| SYSTEM SPACE ZONES | | |
| hal | 1 | |
| USER SPACE ZONES | | |
| DOS | 2 | |
| hal | 1 | All platforms |
| Dos | 2 | All platforms |
| SYSTEM | 3 | |
| Cli | 1 | All platforms |

**Table 75: Default Assignment and Priority of Applications, by Zone (continued)**

| Zone/Default Application | Default Priority | Platform |
|---|---|---|
| IpSecurity | 2 | All platforms |
| *NetLogin* | 6 | All platforms |
| SECURITY | 4 | |
| GenericXml (Allows configuration of one additional external application) | 4 | All platforms |
| SYSTEM SPACE ZONES | | |
| hal | 1 | |

Note

The priority of static ACLs is determined by the order they are configured, with the first rule configured having the highest priority.

### *Configuring User Zones*

There is a configurable process for applications to insert an *ACL* into a zone according to the priority of the application within that zone. Applications can occupy multiple zones. For example, you can add the CLI application to the DOS zone, and assign it a higher priority than the DOS application. The DOS zone then has two applications, CLI and DOS application, and within the DOS zone, an ACL created by the CLI has a higher priority than an ACL inserted by the DOS application.

Another way to configure ACL priority is by creating new zones. For example, you might create a zone called MY_HIGH_ZONE, and assign that zone a priority below the DOS zone and above the System zone. You can add applications to that zone and assign their priority.

The example below shows the ACL zone priority that would result from adding the MacInMac and CLI applications to MY_HIGH_ZONE:

1. SYSTEM_HIGH_ZONE

   hal
2. DOS Zone

   hal

   DOS
3. MY_HIGH_ZONE

   MacInMac

   CLI
4. SYSTEM Zone

   Dot1Ag

   Dot1AgDefault

   MacInMac

   CLI

   *NetLogin*

5. SECURITY Zone

   FlowVSR

   FlowVSRTS

   Generic Xml
6. SYSTEM_LOW_ZONE

   hal

Applications can insert an ACL into any of the zones to which the application belongs.

If an application attempts to insert an ACL into a zone where the application is not configured, an error message appears, and the ACL is not installed. Therefore, you have full control of ACL priorities and you can configure the switch to install ACLs from an application at any priority level. In the example above, the application CLI can insert an ACL into either MY_HIGH_ZONE or the SYSTEM zone. The location of the ACL within the zone depends on the priority assigned by the application. An application can assign priority to an ACL using:

- priority attributes (first or last)
- relative priority
- priority numbers

The priority attributes first (highest priority) and last (lowest priority) can be applied to an ACL to establish its position within a zone.

Relative priority sets the ACL priority relative to another ACL already installed by the application in the same zone.

Priority numbers allow an application to specify the priority of an ACL within a zone. The priority numbers are unsigned integers from 0 to 7; a lower number represents a higher priority. This means that if an application adds an ACL at priority 5 and later adds another ACL at priority 3, the second ACL has higher priority.

If an application assigns the same priority number to two ACLs, the ACL added most recently has the higher priority. It is inserted in the priority map immediately ahead of the older ACL that has the same priority number. This effectively allows the application to create sub-zones within a zone. The attributes first and last can be used in combination with priority numbers to prioritize the ACLs within a sub-zone. For example, an ACL could be configured with the first attribute, along with the same priority number as other ACLs in the same zone, effectively assigning that ACL the highest priority within a sub-zone.

The `show configuration` command shows the current configuration of the entire switch in the form of CLI commands which can later be played back to configure the switch.

The `show configuration acl` command shows the current configuration of the ACL manager.

The new **`application`** keyword allows you to specify the application to which the ACL will be bound. Typically, applications create and insert ACLs on the switch; however the administrator can install ACLs "on behalf" of an application by specifying the **`application`** keyword. (This keyword is also used with the show config acl command to enable CLI playback). If no application is specified, the default application is CLI.

This means you have the ability to create, delete, and configure ACLs for any application.

- To create a zone, use the following command:

  `create access-list zone` *name* **zone-priority***number*

- To configure the priority of zones, use the following command:

  `configure access-list zone` *name* **zone-priority** *number*

- To add an application to a zone at a particular priority, or to change the priority of an application within a zone, use the following command:

  `configure access-list zone` *name* {**add**} **application***appl-name* **application_priority** *number*

  An application must occupy at least one zone.

- To move an application within a zone or to another zone, use the following command:

  `configure access-list zone` *name* **move-application***appl-name* `to-zone`*number* **application-priority***number*

  All applications can be configured to go into any and all zones.

  A change in the zone list results in a change in the order of dynamic ACLs that have been applied per interface. The changes in hardware are achieved by uninstalling and then reinstalling the dynamic ACLs in the new positions. There is a possibility, due to hardware constraints, that some ACLs will not be reinstalled. These occurrences are logged.

- To delete an application from a zone, use the following command:

  `configure access-list zone` *name* **delete application** *appl-name*

  When you delete an application from a zone, any ACLs that have been inserted into that zone for the deleted application are moved to the next higher zone in which the application appears.

- To delete a zone, use the following command:

  `delete access-list zone` *name*

  You must remove all applications from a zone before you can delete the zone. You cannot delete the default zones.

## Network-Zone Support in ACLs

ExtremeXOS Network-Zone support allows you to create a network-zone, add multiple IP addresses and/or MAC addresses to it, and use the network-zone in policy files.

This feature provides the ability to add a single attribute “source-zone,” or “destination-zone” to an entry of a policy file. This entry is then expanded into multiple entries depending upon the number of IP and/or MAC addresses configured in that particular zone. If the zone is added to the policy with the keyword “source-zone,” the attributes that are configured in that particular zone are added as either a “source-address” or an “ethernet-source-address” in the policy. Conversely, if the network-zone is added as a “destination-zone,” the attributes are added to the policies as a “destination-address,” or an “ethernet-destination-address.”

After you make changes in the zones and issue a refresh of a specific zone, or all the zones, the policies that have the corresponding zones configured as source-zone or destination-zone in their entries are expanded and refreshed in the hardware.

If you configure the following policy to a port or *VLAN*, or through applications like IdMgr or *Extreme Network Virtualization (XNV)*,

```
Policy: test
entry e1 {
if match all {
source-zone zone1 ;
}
Then {
permit  ;
}
}
```

and the network-zone “zone1” that is part of the policy is configured as below:

```
create access-list network-zone zone1
configure access-list network-zone zone1 add ipaddress 10.1.1.1 255.255.255.255
configure access-list network-zone zone1 add ipaddress 10.1.1.1 255.255.255.240
configure access-list network-zone zone1 add ipaddress 12.1.1.0 255.255.255.0
```

When you refresh the network-zone “zone1,” the policy is expanded as follows, and is applied in the hardware:

```
entry Cl0:0_10.1.1.1_E1 {
if match all {
source-address 10.1.1.1 / 32 ;
} then {
permit  ;
}
}
entry Cl0:0_10.1.1.1_E2 {
if match all {
source-address 10.1.1.1 / 28 ;
} then {
permit  ;
}
}

entry Cl0:0_12.1.1.0_E3 {
if match all {
source-address 12.1.1.0 / 24 ;
} then {
permit  ;
}
}
```

When the policy is configured with the network-zone, the zone may or may not have attributes configured in it. In cases where the network-zone does not have any attributes, the policy is created with entries that do not have the network-zone attribute alone.

So, if you create the following policy:

```
Policy: test2
entry e1 {
```

```
if match all {
source-zone zone2 ;
protocol udp ;
}
Then {
permit  ;
}
}
entry e2 {
if match all {
protocol tcp ;
}
Then {
permit  ;
}
}
```

And the network-zone “zone2” is just created, but is not configured with any attributes, the policy appears as follows and has only the second entry “e2,” and not “e1”.

```
entry e2 {
protocol tcp;
}
Then {
permit;
}
}
```

Once the network-zone “zone2” is configured with one or more attributes, and refreshed, the policy is updated accordingly. In this instance, the name of the entries that have a source-zone or a destination-zone are changed. This is because each entry in the original policy that has a source-zone/destination-zone is converted to a maximum of eight entries in the new policy.

A single policy can have one or more network-zones configured in it. It can also have the same network-zone in multiple entries with different combinations, as well as support for other attributes in the policy file. Similarly, the same network-zone can be configured to multiple policies. In cases where the policy has multiple network-zones, and only some of those network-zones are refreshed, the entries that correspond to those specific network-zones are alone refreshed, and not entries that correspond to the other network-zones.

After you refresh a network-zone, all the policies that have the specified network-zone are modified, and a refresh for each of those policies is sent to the hardware. The command succeeds only after receiving a successful return for all the policies that have this particular network-zone. If for some reason one of the policy’s refresh fails in the hardware, all the policies that are supposed to refresh are reverted back to their previous successful state, and the command is rejected.

Additionally, the configuration or refresh can fail if the attributes in the network-zone are not compatible with the attributes in the corresponding entries of the policy. For example, in platforms that do not support wide-key or UDF, a policy entry cannot have Layer 2 attributes and Layer4 attributes. In such cases, if the entry has “protocol tcp” and a network-zone that has an ethernet source address, the configuration fails in the hardware.

In cases where the refresh fails, the content of the policy and the content of the network-zone may go out of sync, because the policy reverts back to the last successful state, whereas the network-zone will contain the last configured values.

Here is an example:

```
create access-list network-zone zone1
configure access-list network-zone zone1 add ipaddress 10.1.1.1/32
configure access-list network-zone zone1 add ipaddress 10.1.1.1/28
```

Once this configuration is refreshed and is successfully installed in the hardware, the policy will look like the following:

```
entry Cl0:0_10.1.1.1_E1 {
if match all {
source-address 10.1.1.1 / 32 ;
} then {
permit  ;
}
}
entry Cl0:0_10.1.1.1_E2 {
if match all {
source-address 10.1.1.1 / 28 ;
} then {
permit  ;
}
}
```

If you remove 10.1.1.1/28, and adds 10.1.1.1/24 to the network-zone and perform a refresh,

```
configure access-list network-zone zone1 delete ipaddress 10.1.1.1/28
configure access-list network-zone zone1 add ipaddress 12.1.1.0/24
```

and if for some reason the policy refresh fails, the policy and the network-zone will look like this:

```
entry Cl0:0_10.1.1.1_E1 {
if match all {
source-address 10.1.1.1 / 32 ;
} then {
permit  ;
}
}
entry Cl0:0_10.1.1.1_E2 {
if match all {
source-address 10.1.1.1 / 28 ;
} then {
permit  ;
}
}

create access-list network-zone zone1
configure access-list network-zone zone1 add ipaddress 10.1.1.1 255.255.255.255
configure access-list network-zone zone1 add ipaddress 12.1.1.0 255.255.255.0
```

### *Configuring Network-Zone Support in ACLs*

Use the following command to configure the network-zone support in ACLs.:

#### Creating a Network-Zone

To create a network-zone with the specified name, enter the command below. You can then associate this network-zone with the policy file using either the *source-zone* or *destination-zone* attribute.

`create access-list network-zone` *zone_name*

Here is an example:

```
Switch# create access-list network-zone zone1
```

If you try to create a network-zone that is already created, the following error message is displayed on the console, and the command is rejected:

```
Switch# create access-list network-zone zone1
Error:  Network Zone "zone1" already exists.
```

#### Deleting a Network-Zone

To delete a network-zone, and all the configurations that belong to the zone, enter the following command:

`delete access-list network-zone` *zone_name*

Here is an example:

```
Switch# delete access-list network-zone zone1
```

If you try to delete a network-zone that is bound with one or more policy files, the following error message is displayed, and the command is rejected:

```
Switch# delete access-list network-zone zone1
Error:  Network Zone "zone1" - Unable to delete zone. Zone has one
or more policies.
```

#### Adding or Removing Network-Zone Attributes

To add or remove IP/MAC addresses to or from the network-zone, enter the following command:

`configure access-list network-zone` *zone_name* [**add** | **delete**] [**mac-address** *macaddress* {*macmask*} | **ipaddress** [*ipaddress* {*netmask*} | *ipNetmask* | *ipv6_address_mask*]]

Here is an example:

```
Switch# configure access-list network-zone zone1 add ipaddress 11.1.1.1/24
```

If you try to add the same IP/MAC with the same or narrow mask, the configuration is rejected, with the following error message:

```
Switch# configure access-list network-zone "zone1" add ipaddress 11.1.1.1/32
Error:  Network Zone "zone1" - Zone already has the same entity value with same or wider
mask.
```

If the you try to add more than eight attributes to a network-zone, the following error message is issued:

```
Switch# configure access-list network-zone "zone1" add ipaddress 11.1.1.1/24
Error:  Network Zone "zone1" - Reached maximum number of attributes. Unable to add more.
```

#### Refreshing Network-Zones

To refresh a specific network-zone, or all the network-zones, enter the following command:

`refresh access-list` **`network-zone`** `[`*`zone_name`* `|` **`all`**`]`

Here is an example:

```
Switch# refresh access-list network-zone zone1
```

Note

When you issue the command to refresh a network-zone, or all network-zones, it can take a long time to clear the CLI because each individual policy must be converted before it is refreshed. The command succeeds, or fails, only after it receives a success response for all policy refresh, or when a first refresh failure is received from the hardware.

If the refresh fails for a specific zone, the following error message is printed on the console.

```
Switch# refresh access-list network-zone zone1
Error: Refresh failed for network-zone "zone1".
```

### *Monitoring Network-Zone Support in ACLs*

#### show access-list network-zone

To monitor various network statistics, use the `show access-list network-zone` command. This command's output displays the network-zones configured, the number of attributes configured, and the number of policy files that has the specific zones in it.

Example

```
Switch# sh access-list network-zone
============================================================
Network Zone                        No. of    No. of Policies
Entities       Bound
============================================================
zone1                                 5              2
zone2                                 3              1
zone3                                 0              0
============================================================
Total Network Zones : 3
```

The following example displays detailed information about a specific network zone, the attributes configured in the zone, and the policies bound to the zone:

```
Switch# show access-list network-zone zone1
Network-zone     : zone1
Total Attributes : 3
Attributes       : 10.1.1.1 / 32
10.1.1.1 / 30
10.1.1.0 / 24
No. of Policies  : 1
Policies         : test
Switch # sh access-list network-zone zone2
Network-zone     : zone2
No. of Entities  : 3
Entities         : 00:00:00:00:00:22 / ff:ff:ff:ff:ff:ff
00:00:00:00:00:23 / ff:ff:ff:ff:00:00
00:00:00:00:00:24 / ff:ff:ff:ff:ff:00
No. of Policies : 0
```

## CVID ACL Match Criteria

This feature adds support for the EXOS *ACL* match criteria "cvid." It provides the ability to specify access-lists that filter on the inner-*VLAN*-id field of a double tagged packet, the customer VLAN id field of a single tagged packet entering a VMAN UNI/CEP port, or the port-cvid inserted into an untagged packet entering a VMAN UNI port. You can use this feature to perform service-level, or customer-level (cvid) rate-limiting and accounting.

You can utilize this match criteria in the following scenarios:

- Tagged VMAN ports: installing an ACL matching "cvid" on ingress or egress will match the inner vlan-id of a double tagged packet on a tagged VMAN port.
- Untagged VMAN ports: installing an ACL matching "cvid" on ingress or egress will match the single VLAN tag on an untagged VMAN port.
- CEP VMAN ports (with or without VPLS): installing an ACL matching "cvid" on ingress or egress will match the single VLAN tag on a CEP VMAN port (without translation).
- CEP VMAN ports with cvid translation (with or without translation): installing an ACL matching "cvid" on ingress will match the post-translation cvid. Installing an ACL matching "cvid" on egress will match the post-translation cvid.

  As an example of CEP VMAN ports, consider the following configuration:

```
create vman vm1 tag 100
config vman vm1 add port 1 cep cvid 7 translate 8
config vman vm1 add port 2 tag
```

  Now consider the following ACL policy applied to "access" port 1:

```
     test.pol:
     entry one {
       if {
          cvid 7;
        } then {
           count count7;
        }
     }
     entry two {
        if {
           cvid 8;
        } then {
           count count8;
        }
     }
     config access-list test port 1
     config access-list test port 1 egress
```

  This results in "count8" incrementing for ingress, and "count7" incrementing on egress.

Here is another example policy:

```
entry one {
           if{
             cvid 7;
             vlan-id 100;   #SVID
           } then {
              count foo;
           }
         }
```

And here's an example that allow you to perform service-level, or customer-level (cvid) rate-limiting and accounting:

```
doubletag.pol:
               entry customer1 {
                 if{
                   cvid 8;
                 } then{
                    count cust1;
                 }
               }
create vman vm1 tag 100
config vman vm1 add port 21
config vman vm1 add port 22 tag
config access-list doubletag port 21
config access-list doubletag port 21 egress
```

### Limitations

The CVID ACL match criteria support has the following limitations:

- Any platform that does not support egress ACLs will not support this match criteria on egress.
- Using "cvid" with an egress ACL will not match egress packets matching the port-cvid (since the cvid will be stripped).
- Using "cvid" does not provide symmetrical results when you apply them to VMAN CEP ports that also enable cvid translation. Ingress ACLs match the CVID after ingress translation, while egress ACLs also match the CVID after egress translation.

### Supported Platforms

CVID ACl match criteria is supported on the following platforms:

- All Summit platforms
- BlackDiamond 8K platform
- BlackDiamond X8 platform

## ACL Evaluation Precedence

The section Precedence on page 678 describes *ACL* evaluation precedence for different platforms.

### Precedence

This section describes the precedence for evaluation among *ACL* rules for BlackDiamond X8 series switches, BlackDiamond 8800 series switches, SummitStack, and Summit family switches. In many cases there will be more than one ACL rule entry for an interface. This section describes how multiple rule entries are evaluated.

Multiple rule entries do consume hardware resources. If you find that your situation runs up against those limits, there are steps you can take to conserve resources by modifying the order of the ACL entries that you create. For details, see ACL Mechanisms on page 684.

### *Rule Evaluation*

When there are multiple rule entries applied to an interface, evaluation proceeds as follows:

- A packet is compared to all the rule entry match conditions at the same time.
- For each rule where the packet matches all the match conditions, the action and any action modifiers in the then statement are taken. If there are any actions or action modifiers that conflict (deny vs. permit, etc), only the one with higher precedence is taken.
- If a packet matches no rule entries in the *ACL*, it is permitted.

Often there will be a lowest-precedence rule entry that matches all packets. This entry will match any packets not otherwise processed, so that the user can specify an action to overwrite the default permit action. This lowest-precedence rule entry is usually the last entry in the ACL policy file applied to the interface.

### *Precedence of Dynamic ACLs*

Dynamic ACLs have a higher precedence than any ACLs applied using policy files. The precedence among any dynamic ACLs is determined as they are configured. The precedence of ACLs applied using policy files is determined by the rule's relative order in the policy file.

### *Precedence of L2/L3/L4 ACL Entries*

Rule precedence is solely determined by the rule's relative order. L2, L3, and L4 rules are evaluated in the order found in the file or by dynamic *ACL* configuration.

### *Precedence Among Interface Types*

As an example of precedence among interface types, suppose a physical port 1:2 is a member port of the *VLAN* yellow. *ACL*s could be configured on the port, either singly or as part of a port list, on the VLAN yellow, and on all ports in the switch (the wildcard ACL). For all packets crossing this port, the port-based ACL has highest precedence, followed by the VLAN-based ACL and then the wildcard ACL.

### *Precedence with Egress ACLs*

Egress *ACL*s are supported on the BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X460-G2, X670-G2 and X450-G2, X460, X480, X670, and X770 series switches. For these, egress ACL lookup happens at egress, and diffserv, dot1p and other non-ACL feature examination happen at ingress. Therefore, egress ACL happens at the last moment and has precedence.

### *Redundant Rules*

For BlackDiamond X8 series switches, BlackDiamond 8800 series switches, E4G-200 and E4G-400 cell site routers, and Summit family switches, eliminate redundant rules (any with the EXACT same match criteria) in the policy file. If two rules have identical match conditions, but different actions, the second rule is rejected by the hardware.

For example, the two following *ACL* entries are not allowed:

```
entry DenyNMR {
    if {
        protocol 17;
        destination-port 161;
    } then {
        deny;
```

```
        count denyNMR;
    }
}
entry DenyNIC {
    if {
        protocol 17;
        destination-port 161;
    } then {
        deny;
        count denyNIC;
    }
}
```

## Applying ACL Policy Files

A policy file intended to be used as an *ACL* is applied to a port, *VLAN*, or to all interfaces (the any keyword). Use the name of the policy file for the *aclname* parameter in the CLI command. To apply an ACL policy, use the following command:

`configure access-list` *aclname* [**any** | **ports** *portlist* | **vlan** *vlanname*] {**ingress** | **egress**}

If you use the **any** keyword, the ACL is applied to all the interfaces and is referred to as the wildcard ACL. This ACL is evaluated for any ports without specific ACLs, and it is also applied to any packets that do not match the specific ACLs applied to the interfaces.

If an ACL is already configured on an interface, the command will be rejected and an error message displayed.

To remove an ACL from an interface, use the following command:

`unconfigure access-list` *aclname* [**any** | **ports** *portlist* | **vlan** *vlanname*] {**ingress** | **egress**}

To display which interfaces have ACLs configured, and which ACL is on which interface, use the following command:

`show access-list` *aclname* [**any** | **ports** *portlist* | **vlan** *vlanname*] {**ingress** | **egress**}

Note

If an ACL needs to be installed for traffic that is L3 routed, and the ingress/egress ports are on different packet-processing units or different slots, and any of the following features are enabled, we recommend that you install the policy on a per-port basis rather than applying it as a wildcard, or VLAN-based ACL.

- *MLAG (Multi-switch Link Aggregation Group)*
- PVLAN
- Multiport-*FDB (forwarding database)*

### Displaying and Clearing ACL Counters

To display the *ACL* counters, use the following command:

```
show access-list counter {countername} {any | ports port_list | vlan
vlan_name} {ingress | egress}
```

To clear the access list counters, use the following command:

```
clear access-list {dynamic} counter {countername} {any | ports port_list
| vlan vlan_name} {ingress | egress}
```

## Example ACL Rule Entries

The following entry accepts all the UDP packets from the 10.203.134.0/24 subnet that are destined for the host 140.158.18.16, with source port 190 and a destination port in the range of 1200 to 1250:

```
entry  udpacl {
    if  {
        source-address 10.203.134.0/24;
        destination-address 140.158.18.16/32;
        protocol  udp;
        source-port 190;
        destination-port  1200 - 1250;
    } then {
        permit;
    }
}
```

The following rule entry accepts TCP packets from the 10.203.134.0/24 subnet with a source port larger than 190 and ACK & SYN bits set and also increments the counter tcpcnt. The packets will be forwarded using *QoS* profile QP3. This example works only with BlackDiamond 8000 c-, e, xl-, and xm-series modules, and Summit family switches, since the match condition source-port > 190 alone will create more than 118 rules in the hardware:

```
entry  tcpacl {
    if  {
        source-address 10.203.134.0/24;
        protocol  TCP;
        source-port > 190;
        tcp-flags  syn_ack;
    } then {
        permit;
        count tcpcnt ;
        qosprofile qp3;
    }
}
```

The following example denies *ICMP* echo request (ping) packets originating from the 10.203.134.0/24 subnet, and increments the counter icmpcnt:

```
entry  icmp {
    if  {
        source-address 10.203.134.0/24;
        protocol  icmp;
        icmp-type  echo-request;
    } then {
        deny;
        count icmpcnt;
    }
}
```

The following example prevents TCP connections from being established from the 10.10.20.0/24 subnet, but allows established connections to continue, and allows TCP connections to be established to that subnet. A TCP connection is established by sending a TCP packet with the SYN flag set, so this example blocks TCP SYN packets. This example emulates the behavior of the ExtremeWare permit-established *ACL* command:

```
entry  permit-established {
    if  {
        source-address 10.10.20.0/24;
        protocol  TCP;
        tcp-flags  syn;
    } then {
        deny;
    }
}
```

The following entry denies every packet and increments the counter default:

```
entry  default {
    if  {
    } then {
        deny;
        count default;
    }
}
```

The following entry permits only those packets with destination MAC addresses whose first 32 bits match 00:01:02:03:

```
entry rule1 {
    if {
        ethernet-destination-address 00:01:02:03:01:01 ff:ff:ff:ff:00:00 ;
    } then {
        permit  ;
    }
}
```

The following entry denies IPv6 packets from source addresses in the 2001:db8:c0a8::/48 subnets and to destination addresses in the 2001:db8:c0a0:1234::/64 subnets:

```
entry ipv6entry {
    if {
        source-address 2001:DB8:C0A8:: / 48;
        destination-address 2001:DB8:C0A0:1234:: / 64;
    } then {
        deny;
    }
}
```

Access lists have entries to match an Ethernet type, so be careful when configuring access lists to deny all traffic. For example, the following rule entries permit traffic only to destination 10.200.250.2 and block any other packet.

```
entry test_policy_4 {
    if {
        source-address 0.0.0.0/0;
```

```
        destination-address 10.200.250.2/32;
    } then {
        permit;
        count test_policy_permit;
    }
}
# deny everyone else
entry test_policy_99 {
    if {
    } then {
        deny;
        count test_policy_deny;
    }
}
```

Since the deny section does not specify an Ethernet type, all traffic other than IP packets destined to 10.200.250.2/32 are blocked, including the ARP packets. To allow ARP packets, add an entry for the Ethernet type, 1x0806, as shown below.

```
entry test_policy_5 {
    if {
        ethernet-type 0x0806;
    } then {
        permit;
        count test_policy_permit;
    }
}
```

The following entries use vlan-ids to set up meters based on individual VLANs.

```
myServices.pol
    entry voiceService {
        if {
            vlan-id 100;
        } then {
            meter voiceServiceMeter;
        }
    }
    entry videoService {
        if {
            vlan-id 101;
        } then {
            meter videoServiceMeter;
        }
    }
…and so on.
```

To bind this ACL to a port with vlan-id match criteria use the following command:

```
config access-list myServices port <N>
```

The following entry shows how to take action based on *VLAN* tag priority information. In this example, the **dot1p** match keyword is used to allow and count every tagged packet with a VLAN priority tag of 3.

```
entry count_specific_packets {
    if {
        dot1p 3;
    } then {
        count allowed_pkts;
```

```
        permit;
    }
}
```

# ACL Mechanisms

For many applications of *ACL*s, it is not necessary to know the details of how ACLs work. However, if you find that your application of ACLs is constrained by hardware limitations, you can often rearrange the ACLs to use the hardware more efficiently. The following sections go into some detail about how the ACLs use hardware resources, and provide some suggestions about how to reorder or rewrite ACLs to use fewer resources.

- ACL Slices and Rules.
- ACL Counters—Shared and Dedicated.

## ACL Slices and Rules

The Summit family switches (whether or not included in a SummitStack), BlackDiamond X8 series switches and BlackDiamond 8000 c-, e, xl-, and xm-series modules use a mechanism different from the earlier Summit series and BlackDiamond 8800 series switches to implement *ACL*s. The same architecture and guidelines apply to both platforms.

Note
This feature applies only to BlackDiamond X8 series switches, BlackDiamond 8000 series modules and Summit family switches.

Instead of the per port masks used in earlier switches, these platforms use slices that can apply to any of the supported ports. An ACL applied to a port may be supported by any of the slices.

For Summit family switches and BlackDiamond 8800 a- and e- series modules, the slice support is as follows:

- BlackDiamond 8800 a-series modules—Each group of 24 ports has 16 slices with each slice having enough memory for 128 ingress rules and actions.
- BlackDiamond 8800 e-series modules—Each group of 24 ports has 8 slices with each slice having enough memory for 128 ingress rules and actions.
- Summit X430 switches—Each group of 48 ports has 4 slices; with each slice having enough memory for 256 ingress rules each, which adds up to 1024 ingress rules.
- Summit X440 series switches—each group of 24 ports has 4 slices with each slice having enough memory for 256 ingress rules.
- Summit X450G2 switches—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 512 egress rules, which adds up to 2048 rules
  - Each group of 48 ports has 16 slices with each slice having enough memory for 256 ingress rules , which adds up to 4096 ingress rules.
- Summit X460 series switches and E4G400 routers —
  - Each group of 24 ports has 4 slices with each slice having enough memory for 128 egress rules.
  - Each group of 24 ports has 16 slices with each slice having enough memory for 256 ingress rules.

- Summit X460G2 switches—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 512 egress rules, which adds up to 2048 rules
  - Each group of 48 ports has 16 slices with each slice having enough memory for 256 ingress rules , which adds up to 4096 ingress rules.
- Summit X480 series switches—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 48 ports has 16 internal slices with each slice having enough memory for 512 ingress rules plus the external slice.
- Summit X670 switches and BlackDiamond X8 series switches—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 48 ports has 10 slices; the first 4 (0-3) slices hold 128 ingress rules each, and the last 6 (4-9) slices hold 256 ingress rules each, which adds up to 2048 ingress rules.
- Summit X670G2 switches—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 256 egress rules, which adds up to 1024 rules
  - Each group of 48 ports has 12 slices; the first 4 (0-3) slices hold 512 ingress rules each, and the last 8 (4-11) slices hold 256 ingress rules each, which adds up to 4096 ingress rules.
- Summit X770 switches—
  - Each group of 104 ports has 4 slices with each slice having enough memory for 256 egress rules .
  - Each group of 104 ports has 12 slices; the first 4 (0-3) slices hold 512 ingress rules each, and the last 8 (4-11) slices hold 256 ingress rules each, which adds up to 4096 ingress rules.
- E4G200 switches—
  - Each group of 12 ports has 4 slices with each slice having enough memory for 128 egress rules.
  - Each group of 12 ports has 8 internal slices with each slice having enough memory for 256 ingress rules.
- BlackDiamond X8 series switches—

  10G48X-

  - Each group of 24 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 24 ports has 10 slices; the first 4 (0-3) slices hold 128 ingress rules each, and the last 6 (4-9) slices hold 256 ingress rules each, which adds up to 2048 ingress rules.

  10G48T-

  - Each group of 24 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 24 ports has 10 slices; the first 4 (0-3) slices hold 128 ingress rules each, and the last 6 (4-9) slices hold 256 ingress rules each, which adds up to 2048 ingress rules.

  40G12X-

  - Each group of 6 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 6 ports has 10 slices; the first 4 (0-3) slices hold 128 ingress rules each, and the last 6 (4-9) slices hold 256 ingress rules each, which adds up to 2048 ingress rules.

  40G24X-

  - Each group of 6 ports has 4 slices with each slice having enough memory for 256 egress rules.

- Each group of 6 ports has 10 slices; the first 4 (0-3) slices hold 128 ingress rules each, and the last 6 (4-9) slices hold 256 ingress rules each, which adds up to 2048 ingress rules.

**Note**

Egress ACLs are supported on BlackDiamond X8 series switches, BlackDiamond 8000 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X460, X480, X670, X770, X460-G2, X670-G2 and X450-G2 series switches only.

The following figure shows the 16 slices and associated rule memory for BlackDiamond 8800 a-series module.

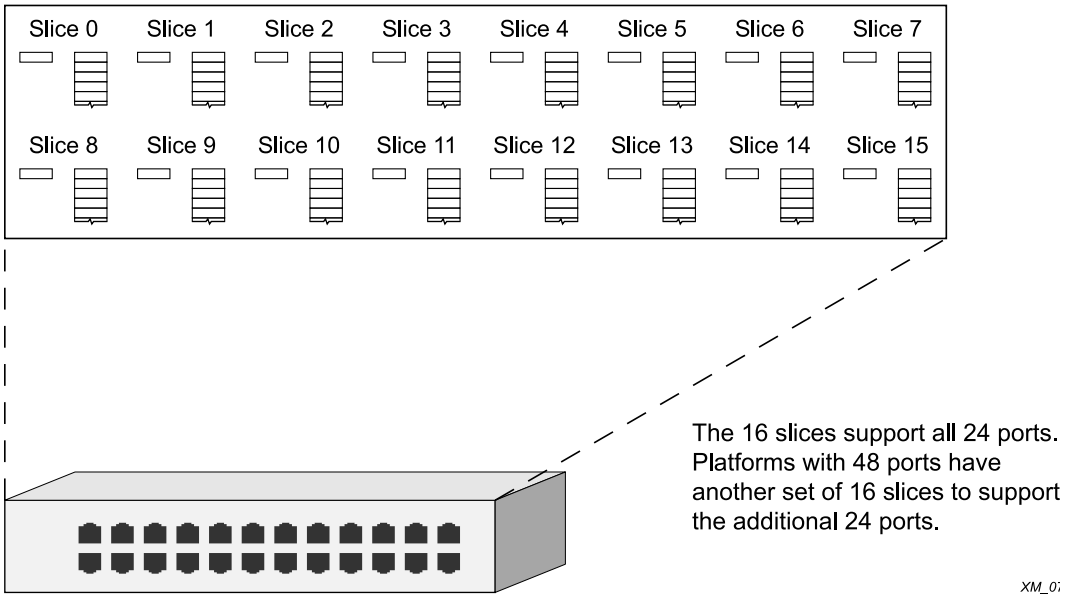



**Figure 93: Slice Support for BlackDiamond 8800 a-Series Modules**

The following figure shows the 8 slices and associated rule memory for a BlackDiamond 8000 e-series module.

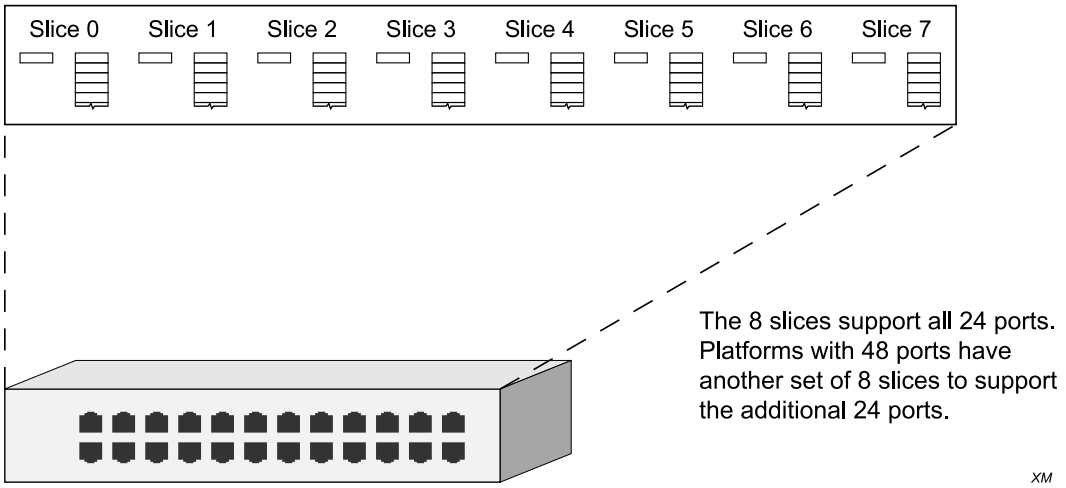



**Figure 94: Slice Support for BlackDiamond 8000 e-Series Modules**

For BlackDiamond 8000 c-, xl-, and xm-series modules, the slice support for the cards is as follows:

- 10G1Xc—
  - Its single port has 4 slices with each slice having enough memory for 128 egress rules.
  - Its single port has 16 slices with each slice having enough memory for 256 ingress rules.
- G8Xc—
  - Its 8 ports have 4 slices with each slice having enough memory for 128 egress rules.
  - Its 8 ports have 16 slices with each slice having enough memory for 256 ingress rules.
- 10G4Xc/10G8Xc—
  - Each group of 2 ports has 4 slices with each slice having enough memory for 128 egress rules.
  - Each group of 2 ports has 16 slices with each slice having enough memory for 256 ingress rules.
- 10G24X-c—
  - Each group of 12 ports has 4 slices with each slice having enough memory for 128 egress rules.
  - Each group of 12 ports has 12 slices with each of the first 8 slices having enough memory for 128 ingress rules and each of the last 4 slices having enough memory for 256 ingress rules.
- G96T-c—
  - Each group of 48 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 48 ports has 16 slices with each slice having enough memory for 512 ingress rules.
- G48Tc/G48Xc/G24Xc—
  - Each group of 24 ports has 4 slices with each slice having enough memory for 128 egress rules.
  - Each group of 24 ports has 16 slices with each slice having enough memory for 256 ingress rules.
- G48X-xl/G48T-xl—
  - Its 48 ports have 4 slices with each slice having enough memory for 256 egress rules.
  - Its 48 ports have 16 internal slices with each slice having enough memory for 512 ingress rules.
- 10G8X-xl—
  - Each group of 4 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 4 ports has 16 internal slices with each slice having enough memory for 512 ingress rules.
- 40G6X-xm and BlackDiamond X8 series switches—
  - Each group of 24 ports has 4 slices with each slice having enough memory for 256 egress rules.
  - Each group of 24 ports has 10 slices with each slice having enough memory for 256 ingress rules.

This architecture also allows a single slice to implement ACLs that are applied to more than one port. When an ACL entry is applied, if its match conditions do not conflict with an already existing ACL, the entry is added to the rule memory of an already populated slice. Because the slices are much more flexible than masks, a much wider variety of rule entries can use the same slice.

When ACLs are applied, the system programs each slice to select parts of the packet information to be loaded into it. For example, one possible way a slice can be programmed allows it to hold the information about a packet's ingress port, source and destination IP address, IP protocol, source and destination Layer 4 ports, DSCP value, TCP flag, and if it is a first fragment. Any rule entry that consists of match conditions drawn from that list is compatible with that slice. This list of conditions is just one example. A complete description of possible ways to program a slice is discussed in Compatible and Conflicting Rules on page 690.

In the following example, the two rule entries are compatible and require only one slice in hardware even though they are applied to different ports. The following entry is applied to port 1:

```
entry ex_A {
    if {
        source-address 10.10.10.0/24 ;
        destination-port 23 ;
        protocol tcp ;
    } then {
        deny ;
    }
}
```

and the following entry is applied to port 2:

```
entry ex_B {
    if {
        destination-address 192.168.0.0/16 ;
        source-port 1000 ;
        protocol tcp ;
    } then {
        deny ;
    }
}
```

Both of these ACLs could be supported on the same slice, since the match conditions are taken from the example list discussed earlier. This example is shown in the following figure. In the example, we refer to slice A, even though the slices are numbered. Slice A just means that one slice is used, but does not specify a particular slice. Some rules require more than one slice, so we use letters to show that different slices are used, but not which specific slices.

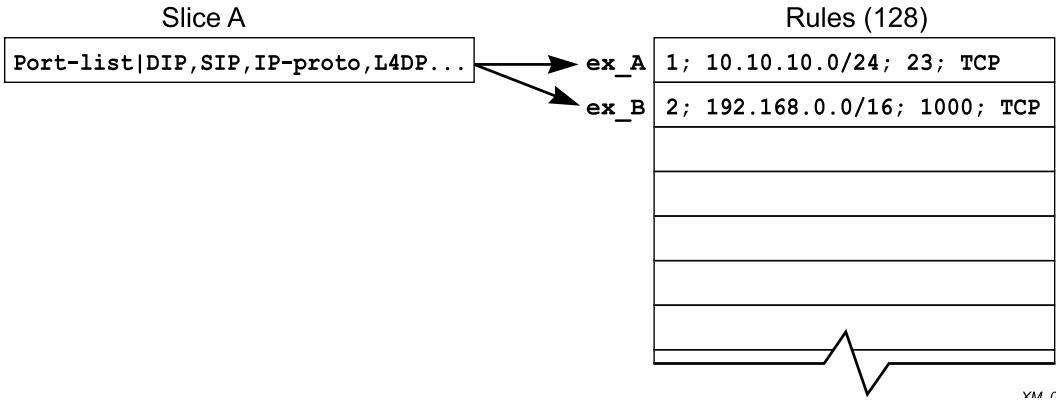



**Figure 95: ACL Entry ex_A and ex_B**

There are cases where compatible ACLs require using a different slice. If the memory associated with a slice is filled with rule entries, then another slice will be used to process any other compatible entries.

For example, consider the following 129 rule entries applied to ports 3-7:

```
entry one {
    if {
        source-address 10.66.10.0/24 ;
        destination-port 23 ;
        protocol tcp ;
    } then {
        deny ;
```

```
    }
}
entry two {
    if {
        destination-address 192.168.0.0/16 ;
        source-port 1000 ;
        protocol tcp ;
    } then {
        deny ;
    }
}
entry three {
    if {
        source-address 10.5.2.246/32 ;
        destination-address 10.0.1.16/32 ;
        protocol udp ;
        source-port 100 ;
        destination-port 200 ;
    } then {
        deny ;
    }
}
....
[The 125 intervening entries are not displayed in this example]
....
entry onehundred_twentynine {
    if {
        protocol udp ;
        destination-port 1714 ;
    } then {
        deny ;
    }
}
```

The following figure shows the result of applying the 129 entries; 128 of the entries are applied to one slice, and the final entry is applied to a different slice. If another compatible entry is applied from another port, for example, it will use Slice B.

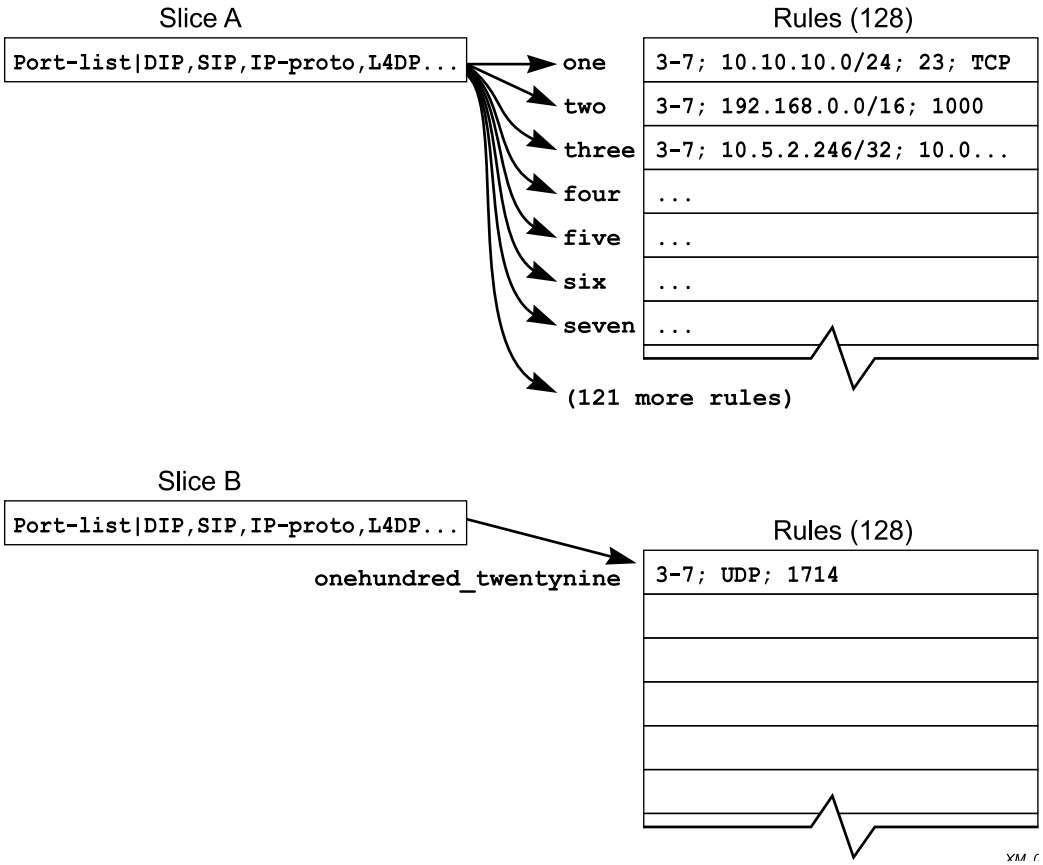



**Figure 96: ACL Entry one Through onehundred_twentynine**

As entries are configured on the switch, the slices are programmed to implement the rules, and the rule memory is filled with the matching values for the rules. If a compatible slice is available, each entry is added to that slice.

### *Compatible and Conflicting Rules*

The slices can support a variety of different *ACL* match conditions, but there are some limitations on how you combine the match conditions in a single slice. A slice is divided up into fields, and each field uses a single selector. A selector is a combination of match conditions or packet conditions that are used together. To show all the possible combinations, the conditions in the following table are abbreviated.

**Table 76: Abbreviations Used in Field Selector Table**

| Abbreviation | Condition |
|---|---|
| Ingress | |
| DIP | destination address <prefix> (IPv4 addresses only) |
| DIPv6/128 | destination address <prefix> (IPv6 address with a prefix length longer than 64) |
| DIPv6/64 | destination address <prefix> (IPv6 address with a prefix length up to 64) |
| DSCP | dscp <number> |

**Table 76: Abbreviations Used in Field Selector Table (continued)**

| Abbreviation | Condition |
|---|---|
| Etype | ethernet-type <number> |
| First Fragment | first ip fragment |
| FL | IPv6 Flow Label |
| Fragments | fragments |
| IP-Proto | protocol <number> |
| L4DP | destination-port <number> (a single port) |
| L4-Range | A Layer 4 port range. For example, if you specify "protocol UDP" and "port 200 - 1200" in an entry, you have used a Layer 4 range. There are a total of sixteen Layer 4 port ranges. Also, you can have a source port range, or a destination port range, but not both kinds of ranges together in the same entry. |
| L4SP | source-port <number> (a single port) |
| MACDA | ethernet-destination-address <mac-address> <mask> |
| MACSA | ethernet-source-address <mac-address> |
| NH | IPv6 Next Header field. Use protocol <number> to match. See IP-Proto |
| OVID | This is not a match condition used in ACLs, but is used when an ACL is applied to *VLAN*s. An ACL applied to a port uses a different field selector than an ACL applied to a VLAN. VLAN IDs are outer VLAN IDs unless specified as inner VLAN IDs. |
| packet-type | This selector is used internally and not accessible by users through explicit ACLs. |
| Port-list | This is not a match condition used in ACLs, but is used when an ACL is applied to ports, or to all ports (the wildcard ACL). An ACL applied to a port uses a different field selector than an ACL applied to a VLAN. |
| SIP | source address <prefix> (IPv4 addresses only) |
| SIPv6/128 | source address <prefix> (IPv6 address with a prefix length longer than 64) |
| SIPv6/64 | source address <prefix> (IPv6 address with a prefix length up to 64) |
| TC | IPv6 Traffic Class field. Use dscp <number> |
| TCP-Flags | TCP-flags <bitfield> |
| TPID | 802.1Q Tag Protocol Identifier |
| TTL | Time-to-live |
| UDF | User-defined field. This selector is used internally and not accessible by users through explicit ACLs. |
| VID-inner | Inner VLAN ID |
| VRF | *virtual router (VR)* and forwarding instance |
| Egress | |
| DestIPv6 | destination-address <ipv6> |
| DIP | destination-address |

**Table 76: Abbreviations Used in Field Selector Table (continued)**

| Abbreviation | Condition |
|---|---|
| Etype | ethernet-type |
| IP-Proto | protocol |
| L4DP | destination-port. Support only single L4 ports and not port ranges. |
| L4SP | source-port. Support only single L4 ports and not port ranges. |
| MACDA | ethernet-destination-address |
| MACSA | ethernet-source-address |
| NH | IPv6 Next Header field. |
| SIP | source-address |
| SIPv6 | source-address <ipv6> |
| TC | IPv6 Traffic Class field. |
| Tcp-Flags | tcp-flags |
| TOS | ip-tos or diffserv-codepoint |
| VlanId | vlan-id |

The following ingress conditions are not supported on egress:

- fragments
- first-fragment
- *IGMP*-msg-type
- *ICMP*-type
- ICMP-code

The tables that follow list all the combinations of match conditions that are available. The possible choices for different collections of switches and modules are listed in the tables as follows:

- BlackDiamond 8800 a-series and G48Te2 Modules
- BlackDiamond 8000 e-Series Modules (Continued)
- BlackDiamond 8800 c-Series Modules
- BlackDiamond 8900 10G24X-c Module
- BlackDiamond 8900 xl-Series and G96Tc Modules and Summit X480 Series Switches
- BlackDiamond 8900 40G6X-xm Module, BlackDiamond X8 series switches and Summit X460, X460-G2, X670, X450-G2, X670-G2, and X770 Switches

Note

It is not possible for the BlackDiamond X8 and Summit X670 series switches to have ICMP/IGMP code and type fields on egress. ICMP/IGMP type requires UDF (user defined fields). Ingress Pipeline has UDF but Egress pipeline hardware does not have UDF. So it cannot match ICMP/IGMP types on egress pipeline.

Any number of match conditions in a single row for a particular field may be matched. For example if Field 1 has row 1 (Port-list) selected, Field 2 has row 8 (MACDA, MACSA, Etype, OVID) selected, and

Field 3 has row 7 (Dst-Port) selected, any combination of Port-list, MACDA, MACSA, Etype, OVID, and Dst-Port may be used as match conditions.

If an ACL requires the use of field selectors from two different rows, it must be implemented on two different slices.

**Table 77: Field Selectors, G48Te2 Series Modules**

| Field 1 | Field 2 | Field 3 |
|---|---|---|
| Port-list | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IP-Fl | IpInfo(First-Fragment, Fragments) |
| L4DP, L4SP | DIP, SIP, IP-Proto, L4DP, L4-range, DSCP, TCP-Flag, IP-flag | Port |
| OVID, VID-inner | DIP, SIP, IP-Proto, L4-range, L4SP, DSCP, TCP-Flag, IP-flag | DSCP, TCP-Flag |
| Etype, OVID | DIPv6/128 | OVID |
| IpInfo(First-Fragment, Fragments), OVID | SIPv6/128 | IP-Proto, DSCP |
| Port, Dst-Port | DIPv6/64,SIPv6/64 | L4-Range |
| Etype, IP-Proto | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | Dst-Port |
| | MACDA, MACSA, Etype, OVID | |
| | MACDA, DIP, Etype, OVID | |
| | MACSA, SIP, Etype, OVID | |
| | "User Defined Field" 1 | |
| | "User Defined Field" 2 | |

**Table 78: Field Selectors, BlackDiamond 8800 G48Te and G48Pe Modules**

| Field 1 | Field 2 | Field 3 |
|---|---|---|
| Port-list | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IP-Flag | IpInfo(First-Fragment, Fragments) |
| L4DP, L4SP | DIP, SIP, IP-Proto, L4DP, L4-range, DSCP, TCP-Flag, IP-flag | Port |
| OVID, VID-inner | DIP, SIP, IP-Proto, L4-range, L4SP, DSCP, TCP-Flag, IP-flag | DSCP, TCP-Flag |
| Etype, OVID | DIPv6/128 | OVID |
| Port, Dst-Port | SIPv6/128 | Dst-Port |
| Etype, IP-Proto | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | |
| | MACDA, MACSA, Etype, OVID | |
| | MACDA, DIP, Etype, OVID | |
| | MACSA, SIP, Etype, OVID | |

**Table 78: Field Selectors, BlackDiamond 8800 G48Te and G48Pe Modules (continued)**

| Field 1 | Field 2 | Field 3 |
|---|---|---|
| | "User Defined Field" 1 | |
| | "User Defined Field" 2 | |

**Table 79: Field Selectors, BlackDiamond 8800 c-Series Modules**

| Field 1 | Field 2 | Field 3 |
|---|---|---|
| Port-list | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IP-Flag | IpInfo(First-Fragment, Fragments) |
| L4DP, L4SP | DIP, SIP, IP-Proto, L4DP, L4-range, DSCP, TCP-Flag, IP-flag | Port |
| OVID, VID-inner | DIP, SIP, IP-Proto, L4-range, L4SP, DSCP, TCP-Flag, IP-flag | DSCP, TCP-Flag |
| Etype, OVID | DIPv6/128 | OVID |
| IpInfo(First-Fragment, Fragments), OVID | SIPv6/128 | IP-Proto, DSCP |
| Port, Dst-Port | DIPv6/64, SIPv6/64 | L4-Range |
| Etype, IP-Proto | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | Dst-Port |
| VRF, OVID | MACDA, MACSA, Etype, OVID | |
| DSCP, VRF, IP-Proto | MACDA, DIP, Etype, OVID | |
| | MACSA, SIP, Etype, OVID | |
| | "User Defined Field" 1 | |
| | "User Defined Field" 2 | |
| | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IpInfo(First-Fragment, Fragments) | |
| | DIP, SIP, IP-Proto, L4DP, L4-range, DSCP, TCP-Flag, IpInfo(First-Fragment, Fragments) | |
| | DIP, SIP, IP-Proto, L4-range, L4SP, DSCP, TCP-Flag, IpInfo(First-Fragment, Fragments) | |

**Table 80: Field Selectors, BlackDiamond 8900 10G24X-c Module**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| Port-list | L4DP, L4SP | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IP-Flag | IpInfo(First-Fragment, Fragments) |
| | OVID, VID-inner | DIPv6/128 | Port |
| | Etype, OVID | SIPv6/128 | DSCP, TCP-Flag |
| | IpInfo(First-Fragment, Fragments), OVID | DIPv6/64, SIPv6/64 | OVID |
| | Port, Dst-Port | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | IP-Proto, DSCP |

**Table 80: Field Selectors, BlackDiamond 8900 10G24X-c Module (continued)**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| | Etype, IP-Proto | MACDA, MACSA, Etype, OVID | L4-Range |
| | VRF, OVID | MACDA, DIP, Etype, OVID | Dst-Port |
| | DSCP, VRF, IP-Proto | MACSA, SIP, Etype, OVID | |
| | | "User Defined Field" 1 | |
| | | "User Defined Field" 2 | |
| | | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IpInfo(First-Fragment, Fragments) | |

**Table 81: Field Selectors, BlackDiamond 8900 xl-series and G96Tc Modules and Summit X480 Series Switches**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| Port-list | DstPort | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, TCP-Flag, IP-Flag | OVID(12bit) |
| | TPID, OVID, VID-inner | DIP, SIP, IP-Proto, L4SP, L4DP, DSCP, IpInfo(First-Fragment, Fragments), TCP-Flag | DstPort |
| | Etype, OVID | SIPv6/128 | OVID |
| | InnerTPID, VID-inner | DIPv6/128 | OVID, VID-inner |
| | OVID | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | Etype, OVID |
| | DSCP, IP-Proto | MACDA, MACSA, Etype, OVID | VID-inner |
| | | MACSA, SIP, Etype, OVID | InnerTPID, OuterTPID |
| | | MACDA, DIP, Etype, OVID | |
| | | “User Defined Field” | |
| | | SIPV6/64, DIPV6/64 | |
| | | DIPV6/64 | |

**Table 82: Field Selectors, BlackDiamond 8900 40G6X-xm Module, BlackDiamond X8 Series Switches and Summit X460, X670, and X770 Series Switches**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| Port-list | OVID, VID-inner | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, IPFlag, TCP-Flag | OVID |
| | Etype, OVID | DIP, SIP, IP-Proto, L4DP, L4SP, DSCP, IpInfo(First-Fragment, Fragments) TCP-Flag | OVID, IpInfo(First-Fragment, Fragments) |
| | VID-inner | DIPv6/128 | OVID, VID-inner |
| | IpInfo(First-Fragment, Fragments), OVID | SIPv6/128 | OVID, Etype |

**Table 82: Field Selectors, BlackDiamond 8900 40G6X-xm Module, BlackDiamond X8 Series Switches and Summit X460, X670, and X770 Series Switches (continued)**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| | OVID | DIPv6/64, IP-Proto, DSCP, FL, TCP-Flag | VID-Inner |
| | IP-Proto, DSCP | MACDA, MACSA, OVID, Etype | L4-Range |
| | "User Defined Field" 1 | MACSA, OVID, Etype, SIP | FL |
| | | MACDA, OVID, Etype, DIP, IP-Proto | UDF1[95..64] |
| | | "User Defined Field" 1 | |
| | | "User Defined Field" 2 | |
| | | DIPv6/64, SIPv6/64 | |

**Table 83: Field Selectors, Summit X440**

| Fixed Field | Field 1 | Field 2 | Field 3 |
|---|---|---|---|
| Ingress Port List | Vlan, EtherType | TTL, TcpControl, IpFlags, TOS, l4DstPort, L4SrcPort, IpProtocol, DstIp, SrcIp | Vlan, EtherType |
| | DstPort, DstMod, DstTrunk, SrcPort, SrcMod, SrcTrunk | TTL, TcpControl, IpFrag, TOS, l4DstPort, L4SrcPort, IpProtocol, DstIp, SrcIp | RangeCheck(l4 ports or vlans) |
| | IpProtocol, TOS, VlanId | SrcIp6 | |
| | 4-byte UDF | DstIp6 | |
| | | TTL, TcpControl, IP6FlowLabel, TOS, IpProtocol, Ip6High | |
| | | Vlan, EtherType, SrcMac, DstMac | |
| | | Vlan, EtherType, SrcMac, SrcIp | |
| | | Vlan, EtherType, TTL, IpProtocol, DstIp, DstMac | |
| | | 16-byte UDF | |
| | | SrcIp6High, DstIp6High | |

### Egress ACLs

Each of the 4 egress slices can be configured to one of the 3 combinations below. The rules that can be installed into a particular slice should be a subset of the combination to which that slice is configured.

Following is the table of the available combinations:

- Combination 1:

  ```
  <vlan-id, ethernet-source-address, ethernet-destination-address,
  ethernet-type>
  ```

- Combination 2:

```
<vlan-id, diffserv-codepoint/ip-tos, destination-address, source-
address, protocol, destination-port, source-port, tcp-flags>
```

- Combination 3:

```
<vlan-id, ip-tos, destination-address<ipv6>, source-address<ipv6>,
protocol>
```

Use the following table through the following table to determine which *ACL* entries are compatible. If the entries are compatible, they can be on the same slice.

For example, the earlier example entries are applied to ports:

```
entry ex_A {
    if {
        source-address 10.10.10.0/24 ;
        destination-port 23 ;
        protocol tcp ;
    } then {
        deny ;
    }
}
entry ex_B {
    if {
        destination-address 192.168.0.0/16 ;
        source-port 1000 ;
    } then {
        deny ;
    }
}
```

Entry ex_A consists of the following conditions (using the abbreviations from the following table), SIP, L4DP, and IP-Proto. Entry ex_B is DIP, L4SP. Since they are applied to ports, the selector for Field 1 is Port-list (the first item). The selector for Field 2 would be the first item, and Field 3 could be any item.

Our other example entries are also compatible with the entries ex_A and ex_B:

```
entry one {
    if {
        source-address 10.66.10.0/24 ;
        destination-port 23 ;
        protocol tcp ;
    } then {
        deny ;
    }
}
entry two {
    if {
        destination-address 192.168.0.0/16 ;
        source-port 1000 ;
    } then {
        deny ;
    }
}
entry three {
    if {
        source-address 10.5.2.246/32 ;
        destination-address 10.0.1.16/32 ;
        protocol udp ;
        source-port 100 ;
```

```
        destination-port 200 ;
    } then {
        deny ;
    }
}
```

Entry one is SIP, L4DP, and IP-Proto; entry two is DIP, and L4SP; entry three is SIP, DIP, IP-Proto, L4SP, and L4DP. All of these examples can use the first item in Field 2 in the tables.

However, if we add the following entry:

```
entry alpha {
    if {
        ethernet-destination-address 00:e0:2b:11:22:33 ;
    } then {
        deny ;
    }
}
```

this will not be compatible with the earlier one. Entry alpha is MACDA, and there is no MACDA in the first item for Field 2. Any entry with MACDA will have to use selector 7 or 8 from the following table (or 6 or 7 from the following table, depending on the platform). If an entry requires choosing a different selector from the table, it is not compatible and must go into a different slice.

### *Single Virtual Group for User ACLs*

Prior to ExtremeXOS 16.1, when two user rules in two separate slices are matched by a packet, the non-conflicting actions from both of the rules are executed. This feature allows you to put all user rules into a single virtual group. When rules are in a single virtual group, even when two rules in two separate virtual slices are matched, only the actions of the highest precedence rule are executed. In effect, in this mode, multiple slices behave as a big single virtual slice.

Normally *ACL* hardware works in the following way: on arrival of a packet, all N slices are searched in parallel to find a possible match in each of these N slices. In each slice, upon finding the first match, the search within that slice stops. In other slices the search continues until a match is found or the end of the slice is reached. Thus, a single slice can produce only a single match, but all N slices combined together can produce up to N matches for a given packet.

In the case of multiple matches in multiple slices, all the actions of the rule in the highest priority virtual slice are executed. In addition, all the actions from the lower priority rules from the lower priority virtual slices are executed if those additional actions do not conflict with the actions of the highest priority rule. An example of non-conflicting with each other actions would be "permit" and "count". An example of conflicting with each other actions would be "permit" and "deny".

However, in more recent chipsets a new mode of operation was introduced where you can combine a few virtual slices into one big virtual group. In this mode of operation, even if a packet gets multiple matches from multiple virtual slices within the same virtual group, only the actions of the highest priority rule are executed, whereas the actions from the lower priority rules are not executed at all, even if those actions do not conflict with the actions of the highest priority rule.

This feature allows choosing between the old way of operation where every virtual slice is in its own virtual group and multiple matches are possible, and the new way, where all user ACL's virtual slices are in the same virtual group and multiple matches are not possible.

Note

Some platforms that do not support virtual groups. On those chipsets even if single virtual group feature is enabled, ACL will operate in the old way and multiple matches would be still possible.

### *Rule Evaluation and Actions*

When a packet ingresses the switch, its header is loaded into all the slices, and the header value is compared with the rule values. If the values match, the rule action is taken. Conflicting actions are resolved by the precedence of the entries. However, if rule entries are on different slices, then *ACL* counters can be incremented on each slice that contains a counter-incrementing rule.

### *Slice and Rule Use by Feature*

A number of slices and rules are used by features present on the switch. You consume these resources when the feature is enabled.

- dot1p examination - enabled by default - 1 slice, 8 rules per chip
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
- *IGMP* snooping - enabled by default - 2 slice, 2 rules
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
  - Slice B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=IP-Proto, TOS)
- *VLAN* without IP configured - 2 rules - 2 slices
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
  - Slice C (F1=Port-list, F2=SIP, DIP, IP-proto, L4SP, L4DP, DSCP, F3=packet-type)
- IP interface - disabled by default - 2 slices, 3 rules (plus IGMP snooping rules above)
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
  - Slice C (F1=Port-list, F2=SIP, DIP, IP-proto, L4SP, L4DP, DSCP, F3=packet-type)
- VLAN *QoS* - disabled by default - 1 slice, n rules (n VLANs)
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- port QoS - disabled by default - 1 slice, 1 rule
  - Slice D (F1=anything, F2=anything, F3=anything)
- *VRRP (Virtual Router Redundancy Protocol)* - 2 slices, 2 rules
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- *EAPS* - 1 slice, 1 rule (master), n rules (transit - n domains)
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- *ESRP (Extreme Standby Router Protocol)* - 2 slices, 2 rules
  - Slice A (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=packet-type)
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- IPv6 - 2 slices, 3 rules
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
  - Slice (F1=Port-list, F2=DIPv6, IPv6 Next Header Field, TC, F3=anything)

- Netlogin - 1 slice, 1 rule
  - Slice A or B (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- VLAN Mirroring - 1 slice, n rules (n VLANs)
  - Slice E (F1=Port-list, F2=MACDA, MACSA, Etype, VID, F3=anything)
- Unicast Multiport *FDB*
  - 1 slice, 1+n rules in 24 port Summit series switches
  - 1 slice, 2+ n rules in 48 port Summit series and G48Ta, G48Pe cards
- VLAN Aggregation
  - 1 slice, 4 rules for the first subvlan configured and 1 slice, 2 rules for subsequent subvlan configuration
- Private VLAN
  - 2 slices, 3 rules when adding an non-isolated VLAN with loop-back port a to private VLAN
  - 1 slice, 3 rules when adding an isolated subscriber VLAN (without loopback port) to a private VLAN. 3 additional rules when a loopback port is configured in the above isolated subscriber VLAN
- ESRP Aware - 1 slice, 1 rule
  - Field 1: {Drop, OuterVlan, EtherType, PacketFormat, HiGig, Stage, StageIngress, Ip4, Ip6}
  - Field 2: {SrcIp, DstIp, L4SrcPort, L4DstPort, IpProtocol, DSCP, Ttl, Ip6HopLimit, TcpControl, IpFlags}
  - Field 3: {RangeCheck}
- *ACL* rule with mirror action is installed in a separate slice, and this slice cannot be shared by other rules without a mirror action.

Note

The user ACLs may not be compatible with the slice used by this ESRP rule. This may result in the reduction the number of rules available to the user by 127.

Note

Additional rule is created for every active IPv6 interface and for routes with prefix greater than 64 in following cards for Black Diamond. These rules occupy a different slice. G48Ta,10G1xc,G48Te, G48Pe, G48Ta, G48Xa, 10G4Xa, 10G4Ca, G48Te2, G24Xc, G48Xc, G48Tc, 10G4Xc, 10G8Xc, S-G8Xc, S-10G1Xc.

To display the number of slices used by the ACLs on the slices that support a particular port, use the following command:

```
show access-list usage acl-slice port port
```

To display the number of rules used by the ACLs on the slices that support a particular port, use the following command:

```
show access-list usage acl-rule port port
```

To display the number of Layer 4 ranges used by the ACLs on the slices that support a particular port, use the following command:

```
show access-list usage acl-range port port
```

### *System Configuration Example*

The following example shows incremental configurations and their corresponding *ACL* resource consumption taken on a BlackDiamond 8800 switch with an a-series card.

- Default configuration including: dot1p examination and *IGMP* snooping:
  - 2 slices, 10 rules
- Add an IP interface to the configuration:
  - 2 slices, 13 rules
- Add port-based *QoS* to the configuration:
  - 2 slices, 14 rules
- Add *VLAN*-based QoS to the configuration:
  - 2 slices, 15 rules
- Add *VRRP* to the configuration:
  - 2 slices, 17 rules
- Add *EAPS* (Master mode) to the configuration:
  - 2 slices, 18 rules
- Add *ESRP* to the configuration:
  - 2 slices, 21 rules
- Add IPv6 routing (slowpath) to the configuration:
  - 4 slices, 24 rules
- Add Netlogin to the configuration:
  - 5 slices, 25 rules

**Note**

The slice and rule usage numbers given in this section are for the ExtremeXOS 12.4.1 release. They may vary slightly depending on the ExtremeXOS release.

### *ACL Error Messages*

Errors may happen when installing an *ACL* policy on a port, *VLAN*, or all interfaces (wildcard). Following is a list of the most common error conditions and their resulting CLI error message:

- `Error: ACL install operation failed - slice hardware full for port 3:1`

  Slice resource exceeded: This happens when all slices are allocated for a given chip and an additional incompatible rule (see Egress ACLs on page 696) is installed which requires allocation of another slice.

- `Error: ACL install operation failed - rule hardware full for port 3:1`

  Rule resource exceeded: This happens when all slices are allocated for a given chip and there is an attempt to install a compatible rule to the lowest precedence slice which already has 128 rules. This condition can be triggered with less than the full capacity number of rules installed. For example, if 15 of the slices each have less than 128 rules and there is an attempt to install 129 compatible rules, this error message will be displayed.

- `Error: ACL install operation failed - layer-4 port range hardware full for port 3:1`

Layer-4 port range exceeded: This happens when more than 32 Layer 4 port ranges are installed on a single chip.

- `Error: ACL install operation failed - conditions specified in rule "r1" cannot be satisfied by hardware on port 3:1`

  Incompatible fields selected: This happens when the selected conditions can not be satisfied by the available single-slice field selections described in Compatible and Conflicting Rules on page 690.

- `Error: ACL install operation failed - user-defined-field (UDF) hardware full for port 3:1`

  UDF exceeded: This happens in the rare case that the two available user-defined fields are exceeded on a given chip. UDF fields are used to qualify conditions that are not natively supported by the hardware. Some ACL rules that use UDF are: Source MAC address + Destination IP address combination, Destination MAC address + Source IP address combination, *ICMP* Type, and ICMP Code.

## ACL Counters-Shared and Dedicated

You can configure rule compression in *ACL*s to be either shared or dedicated.

Note
This feature only applies to BlackDiamond X8, BlackDiamond 8800, and Summit family switches.

In the dedicated mode, ACL rules that have counters are assigned a separate rule space and the counter accurately shows the count of matching events. If the ACL with counter is applied to ports 1 and 2, and 10 packets ingress via port 1 and 20 packets ingress via port 2, the ACL counter value for ports 1 and 2 is 10 and 20 packets respectively. More space is used and the process is slower than shared. Dedicated is the default setting.

In the shared mode, ACL space is reused even with counters. ACL counters count packets ingressing via all ports in the whole unit. If the ACL with the counter is applied to ports 1 and 2, and 10 packets ingress via port 1, and 20 packets ingress via port 2, the ACL counter value is 30 each of ports 1 and 2 instead of 10 and 20. The process is faster—as fast as applying an ACL without the counters—and saves space.

Note
Port-Counter shared mode will not work when ports are on across slots and units.

The shared/dedicated setting is global to the switch; that is, the option does not support setting some ACL rules with shared counters and some with dedicated counters.

Use the following command to configure the shared or dedicated mode:

`configure access-list rule-compression port-counters [shared | dedicated]`

Use the following command to view the configuration:

`show access-list configuration`

The shared or dedicated mode does not affect any ACLs that have already been configured. Only ACLs entered after the command is entered are affected.

**Note**
To configure all ACLs in the shared mode, enter the command before any ACLs are configured or have been saved in the configuration when a switch is booted.

## External TCAM ACLs

In addition to internal *ACL* tables, BlackDiamond 8900 and X8 xl-series modules and Summit X480 series switches can install ACL rules into a ternary content addressable memory (TCAM). External TCAMs can hold a much greater number of ACL rules than internal ACL memories. External TCAMs are used for user ACLs when the switch runs in either acl-only mode or 12-and-13-and-acl mode. If the switch is not running in one of these two modes, internal ACL memory is used instead.

**Note**
This feature applies only to BlackDiamond 8900 and X8 xl-series modules and Summit X480 series switches.

To set the system to acl-only mode, issue the following command, save, and reboot:

```
configure forwarding external-tables acl-only
```

To set the system to l2-and-l3-and-acl mode, issue the following command, save, and reboot:

```
configure forwarding external-tables l2-and-l3-and-acl
```

In acl-only mode, the following condition sets and the following number of rules are supported:

```
Ipv4 Rules: (The maximum is 61440 such rules.)
{
  <ethernet-source-address>, <ethernet-destination-address>,
  <vlan or vlan-id>, <source-address ipv4 addr>,
  <destination-address ipv4 addr>, <protocol>,
  <source-port l4 port or port-range>,
  <destination-port l4 port or port-range>,
  (Note, only one l4 port range per rule is supported)
  <tcp-flags>
}
Ipv6 Rules: (The maximum is 2048 such rules.)
{
  <ethernet-source-address>, <ethernet-destination-address>,
  <vlan or vlan-id>, <source-address ipv6 addr>,
  <destination-address ipv6 addr>, <diffserv-codepoint>, <protocol>,
  <source-port l4 port or port-range>,
  <destination-port l4 port or port-range>,
  (Note, only one l4 port range per rule is supported)
  <tcp-flags>
}
```

In `l2-and-l3-and-acl mode`, the following condition sets and the following number of rules are supported:

```
Ipv4 Rules: (The maximum is 57344 such rules.)
{
  <vlan or vlan-id>, <source-address ipv4 addr>,
```

```
  <destination-address ipv4 addr>, <protocol>,
  <source-port l4 port or port-range>,
  <destination-port l4 port or port-range>,
  (Note, only one l4 port range per rule is supported)
  <tcp-flags>
}
```

**Note**

In either of the two available external TCAM ACL modes, configuring more that 55000 rules is not recommended, because when the number of rules is greater than 55000, the system runs low on memory and can experience unexpected crashes.

# Policy-Based Routing

**Note**

Policy-Based Routing is available only on the platforms listed for this feature see the *Feature License Requirements* document. Refer to Load Sharing Rules and Restrictions for All Switches on page 257 for information on applying ACLs to *LAG (Link Aggregation Group)* ports.

## Layer 3 Policy-Based Redirect

Policy-Based Routing allows you to bypass standard Layer 3 forwarding decisions for certain flows. Typically, in a Layer 3 environment, when an IP packet hits an Ethernet switch or router, the Layer 3 processing determines the next-hop and outgoing interface for the packet (based only on the packet's destination address). The Layer 3 processing does so by looking up the IP Forwarding Table; this forwarding table itself is populated either by static routes or by routes learned dynamically from routing protocols such as *OSPF* and *RIP*.

With Policy-Based Routing, you can configure policies to use a different next-hop than what the routing lookup would have chosen. The switch first compares packets to the *ACL* rule entries. If there is a match, the packet is forwarded to the destination identified by the redirect action modifier. If there is no match, the packet is forwarded based on normal routing, in other words, by looking up a route in the IP Forwarding Table.

When there is a match with a redirect ACL rule, the matched traffic is redirected to the next-hop specified in the action.

**Note**

The IP packet itself is not modified, but only redirected to the port where the next-hop entry resides. The original IP destination address and source address are retained in the packet. The TTL is decremented and the IP checksum is recalculated.

The applications for Policy-Based Routing are quite diverse, since the functionality can be used to set policies on how flows identified by any Layer 2 to Layer 7 field (bounded by the switch's ACL syntax) are forwarded.

Deployment scenarios include:

- Forwarding flows for certain applications, for example, all HTTP traffic to designated server(s).
- Redirecting flows from certain source IP addresses for security and other applications.

Policy-Based Routing is implemented using ACLs, so it inherits the capabilities and limitations of ACLs. All the matching conditions used for ACLs can be used for Policy-Based Routing. The destination IP address must be an IPv4 unicast address. For IPv6 scenarios refer the section on Policy-Based Redirection Redundancy on page 708.

When a switch finds a matching ACL rule, it forwards the packet to the redirect IP address as specified in the rule without modifying the packet (except as noted above).

The traffic flow is redirected only after applying the ACL to the port and only when the redirect IP address's adjacency is resolved. When the ARP or NDP table does not have the information to reach the redirect IP address, the packet is routed based on the Layer 3 routing table. When the switch does not know how to reach the redirect IP address in the rule, the rule is installed with a warning, and traffic is not redirected until the address is resolved in the ARP or NDP table. After the address is resolved, the traffic is redirected.

To configure Policy-Based Routing, you configure an ACL on your switch. You can apply an ACL policy file, or use a dynamic ACL.

The following is an example ACL rule entry that redirects any TCP traffic with a destination port of 81 to the device at IP address 3.3.3.2:

```
entry redirect_port_81 {
    if {
        protocol tcp;
        destination-port 81;
    } then {
        redirect 3.3.3.2;
    }
}
```

Use the following procedure:

1. Issue the following command to prevent the redirect IP address from clearing from the ARP or NDP table due to a timeout: `enable iparp refresh`
2. Configure the ACL, either by applying an ACL policy file similar to the example, or a dynamic ACL.
3. Ping or send traffic so that the redirect IP adjacency is resolved.

You may want to create a static ARP or NDP entry for the redirect IP address, so that there will always be a cache entry.

**Note**

An ACL can be rejected on modules and switches that support Policy-Based Routing, because these have different amounts of hardware resources and one module or switch has exhausted its resources.

## Layer 2 Policy-Based Redirect

This feature allows matching packets to override the normal forwarding decision and be Layer 2 switched to the specified physical port. This is accomplished using an additional packet *ACL* lookup. While similar to the Layer 3 Policy-Based Redirect feature, it differs in that the packet is not modified for Layer 3 routing based on a new IP redirect next-hop. Instead, the packet uses the packet format based on the forwarding decision. When the packet is Layer 2-switched, the packet egresses the redirect port

unmodified. When the packet is Layer 3-switched, the packet egresses with the Layer 3 packet modifications of the next-hop found by the normal Layer 3 forwarding lookups.

The following ACL actions are added in support of this feature:

`redirect-port` *`port`*`; redirect-port-list` *`port-list`*

Note

The `redirect-port` or `redirect-port-list` commands will not work for L3 switched packets matching ACL, if distributed IP ARP feature is turned ON.

You must specify the *`port`* argument in the correct format for the switch platform. On supporting switches and modules, this argument must be in the format *`slot`*:*`port`* and on Summit family switches, this argument must be in the format *`port`*.

The *`port-list`* argument is simply a comma-separated list of *`port`* arguments. White space between *`port`* arguments is not allowed.

Here is an example of valid *`port-list`* syntax:

```
redirect-port-list 2:1,2:5,5:3; and redirect-port-list 3,24,5;
```

Here is an example of invalid *`port-list`* syntax :

```
redirect-port-list 2:1 2:5 5:3;
redirect-port-list 2, 4, 5;
```

The policy shown below redirects any TCP traffic with source Layer 4 port 81 to physical port 3:2.

```
entry one {
    if {
        protocol tcp;
        source-port 81;
        destination-port 200 ;
    } then {
        count num_pkts_redirected;
        redirect-port 3:2;
    }
```

The policy shown below redirects any in-profile traffic as defined by the meter configuration to physical port 14. The out-of-profile traffic is subject to the action specified in the meter “out-action” configuration.

```
entry one {
    if {
    } then {
        meter redirected_traffic;
        count num_pkts_redirected;
        redirect-port 14;
    }
```

The policy shown below redirects all traffic with source IP matching 192.168.1.1/24; to physical ports 2:10 and 4:7.

```
entry one {
if {
```

```
source-address 192.168.1.1/24;
} then {
count num_pkts_redirected;
redirect-port-list 2:10,4:7;
}
```

If an incorrect port format is used or if the port number specified is out of range, the following error message is displayed:

```
*BD-8810.68 # check policy l2pbr
Error:  Policy l2pbr has syntax errors
Line 7 : 12:3 is not a valid port.
BD-8810.70 # check policy l2pbr
Error:  Policy l2pbr has syntax errors
Line 7 : 77 is not a valid port.
```

When this feature is used on BlackDiamond 8000 series modules, the traffic egressing the redirect-port can be either tagged or untagged depending on the redirect-port *VLAN* configuration. The following table provides the details.

**Table 84: VLAN Format of Traffic Egressing Redirect-Port**

| ACL Hardware Type | Redirect-Port Not in Egress VLAN | Redirect-Port Tagged in Egress VLAN | Redirect-Port Untagged in Egress VLAN |
|---|---|---|---|
| BlackDiamond 8000 c-, e-, xl-, and xm-series modules | Dropped | VLAN Tagged | Untagged |

Be aware of the following important implementation notes:

- Using the “redirect-port” action with a disabled port causes traffic to be dropped.
- Using the “redirect-port” action overrides Layer 2 echo kill; the result is that a packet can be made to egress the ingress port at Layer 2.
- For systems with a- and e- series hardware that has the larger table size, packets with IP options do not match ACLs using the “redirect-port” action. Systems with hardware that has the smaller table size do not have this capability. On these systems, packets with IP options will match ACLs that use the “redirect-port” action, and will be dropped.
- The redirect-port-list action modifier is targeted towards L2 scenarios. This action is not supported in slow path ACLs. The following list summarizes the behavior of the redirect-port-list action modifier under certain situations.

  The following list summarizes the behavior of redirect-port-list action modifier under certain situations.

  - When a Unicast packet matches the applied ACL, the packet is redirected to all ports specified in the redirect port-list as long as these ports are part of the true egress VLAN.
  - When a Broadcast/Multicast packet matches the applied ACL, the packet is redirected only to ports specified in the redirect port-list that are part of the ingress VLAN. Matched multicast packets will get L2 switched.
  - When a *LAG* port is part of redirect-port-list, then packets matching applied ACL will be load shared between LAG member ports based on Layer 2 source and destination MAC addresses.

## LAG Port Selection

This feature allows you to apply an *ACL* that causes matching packets to egress a specific port in a link aggregation (or load-sharing) group.

**Note**
This feature applies only to BlackDiamond 8000 series modules and Summit family switches.

The following ACL action is added in support of this feature:

```
redirect-port-no-sharing port
```

The ACL overrides any load-sharing algorithm hash that is generated based on the lookup results.

Limitations include the following:

- If the selected port in a load-sharing group is down, the packets will be dropped.
- Like the redirect-port action, the specified port must be a member of the egress *VLAN*.

Following is an example of a configuration and ACL policy that directs traffic matching 10.66.4.10 to *LAG* port 3:

```
enable sharing 2 group 2,3
radiomgmt.pol:
entry one {
if {
destination-address 10.66.4.10/32;
} then {
redirect-port-no-sharing 3;
}
}
config access-list radiomgmt any
```

This example would direct inband management traffic to specific radios connected to specific ports within a load-sharing group.

## Policy-Based Redirection Redundancy

*Multiple Next-hop Support*

As discussed above, Layer 3 and Layer 2 policy-based redirect support only one next-hop for one policy-based entry. Multiple next-hops with different priorities can be configured. A higher priority is denoted with a higher number; for example, "priority 5" has a higher precedence than "priority 1." When a high priority next-hop becomes unreachable, another preconfigured next-hop, based on priority, replaces the first. This is done by first creating a flow-redirect name that is used to hold next-hop information. User-created flow-redirect names are not case-sensitive.

**Note**
As of ExtremeXOS 16.1, there is no limitation in creating the flow-redirects. Number of Next hops has been increased to 4096 next hops. If more than 4096 next hops are attempted to be created, an error message is displayed.

Use the following command:

```
create flow-redirect flow_redirect_name
```

To delete the flow-redirect name, use:

```
delete flow-redirect flow_redirect_name
```

Then information for each next-hop, including a defined priority, is added one by one to the new flow-redirect name. Use the following command:

```
configure flow-redirect flow_redirect_name add nexthop ipaddress
priority number
```

Note

You can add IPv4 or IPv6 next-hops to a flow-redirect policy, but both types are not supported in the same policy.

To delete a next-hop, use the following command:

```
configure flow-redirect flow_redirect_name delete nexthop {ipaddress |
all }
```

Because an *ACL* does not recognize the virtual routing concept, one policy-based routing is used for multiple virtual routing entries when a *VLAN*-based *VR* is used for one port. Configuring a virtual router into a flow-redirect allows policy-based routing to work for only one specific virtual router. Use the following command:

```
configure flow-redirect flow_redirect_name vr vr_name
```

Note

Configuring the virtual router parameter is not supported on BlackDiamond 8800 series switches and Summit family switches. Flow-redirect does not work on user-created virtual routers.

Finally, a new action modifier, redirect-name, is used to specify the flow-redirect name in an ACL rule entry.

```
entry redirect_redundancy {
    if match all {
        source-address 1.1.1.100/24 ;
    } then          {
        permit ;
        redirect-name <name>
    }
}
```

### *Health Checking for ARP, NDP, and Ping*

Policy-based redirection redundancy requires the determination of the reachability or unreachability of the active next hop and the other configured next hops. This feature can use ARP, NDP or Ping checking to make the determination.

Note

IPv6 Policy-Based Routing is not supported for traffic with Hop-by-Hop extension headers. Traffic will continue to be hardware forwarded, and will not be processed in slow path.

To configure health checking for a specific flow-redirect-name, use the following command:

```
configure flow-redirect flow_redirect_name health-check [ping | arp |
neighbor-discovery]
```

To configure the ping interval and miss count for a next-hop, use the following command:

```
configure flow-redirect flow_redirect_name nexthop ip_address ping
health-check interval seconds miss number
```

*Packet Forward/Drop*

The default behavior for policy-based routing when all next-hops are unreachable is to route packets based on the routing table. Policy-based routing redundancy adds an option to drop the packets when all next-hops for the policy-based routing become unreachable.

To configure this option, use the following command:

```
configure flow-redirect flow_redirect_name no-active [drop|forward]
```

**Configuring Packet Forward Drop**

Traffic from the Source IP = 211.10.15.0/24, 211.10.16.0/24 network blocks should be redirected into two routers: 192.168.2.2 and 192.168.2.3. The 192.168.2.2 router is preferred to 192.168.2.3. If router 192.168.2.2 is not reachable, 192.168.2.3 should be used. If both routers are not reachable, the default route is used.

1. Create a flow-redirect to keep next-hop IP address and health check information.

   ```
   create flow-redirect premium_subscriber
   config flow-redirect premium_subscriber add next-hop 192.168.2.2
   priority 200
   config flow-redirect premium_subscriber add next-hop 192.168.2.3
   priority 100
   ```

2. Add an *ACL* entry with a flow-redirect name action to the existing ACL policy (For example: premium_user.pol).

   ```
   entry premium_15 {
   if match {
   source-address 211.10.15.0/24;
   } then {
   permit;
   redirect-name premium_subscriber;
   }
   }
   entry premium_16 {
   if match {
   source-address 211.10.16.0/24;
   } then {
   permit;
   redirect-name premium_subscriber;
   }
   }
   ```

3. Apply the modified ACL policy file or dynamic ACL into a port, *VLAN*, or VLAN and Port. (For example: user1 VLAN: 192.168.1.0/30, user2 VLAN: 192.168.1.4/30.)

   ```
   config access-list premium_user vlan user1 ingress
   config access-list premium_user vlan user2 ingress
   ```

4. Finally, check the current flow-redirect status.

   ```
   BD-8810.47 # show flow-redirect "premium_subscriber"
   Name         : premium_subscriber           VR Name         : VR-Default
   ```

```
NO-ACTIVE NH : FORWARD                        HC TYPE         : PING
NH COUNT     : 2                              ACTIVE IP       : 192.168.2.3
Index     STATE     Pri         IP ADDRESS             STATUS INTERVAL MISS
================================================================================
0         ENABLED   200         192.168.2.2             DOWN
2
           2
1         ENABLED   100         192.168.2.3             UP       2      2

BD-8810.48 # show flow-redirect
Flow-Redirect Name     NH_CNT     ACTIVE IP    VR Name         D/F  HC
=========================================================================
premium_subscriber     2          192.168.2.3  VR-Default       F   PING
```

## ACL Troubleshooting

On BlackDiamond 8800 series switches, SummitStack, and Summit family switches, the following commands are designed to help troubleshoot and resolve *ACL* configuration issues:

- `show access-list usage acl-mask port` *`port`*
- `show access-list usage acl-range port` *`port`*
- `show access-list usage acl-rule port` *`port`*
- `show access-list usage acl-slice port` *`port`*

The **`acl-mask`** keyword is not relevant for the a-series or e-series models.

If you enter this command and specify an a-series or e-series port, the following error message appears:

```
This command is not applicable to the specified port.
```

Use the **`acl-rule`** keyword to display the total number of ACL rules that are available and consumed for the specified port.

If this keyword is specified on an a-series or e-series port, the first part of the command output details the port list using this resource because the ACL hardware rules are shared by all ports on a given ASIC (24x1G ports). If you enter the same command and specify any of the listed ports, the command output is identical.

```
*switch# show access-list usage acl-rule port 4:1 Ports 4:1-4:12, 4:25-4:36
Total Rules:     Used: 46  Available: 2002
```

The **`acl-slice`** keyword is used to display ACL resource consumption for each of the independent TCAMs, or slices, that make up the hardware ACLs.

Each slice is a 128-entry TCAM. The command output displays the number of consumed and available TCAM rules for each slice as follows.

```
*switch# show access-list usage acl-slice port 4:1
Ports 4:1-4:12, 4:25-4:36
Slices:          Used: 8  Available: 8
Slice 0 Rules:   Used: 1  Available: 127
Slice 1 Rules:   Used: 1  Available: 127
Slice 2 Rules:   Used: 1  Available: 127
Slice 3 Rules:   Used: 8  Available: 120
Slice 4 Rules:   Used: 8  Available: 120
Slice 5 Rules:   Used: 2  Available: 126
Slice 6 Rules:   Used: 1  Available: 127
Slice 7 Rules:   Used: 24 Available: 104
```

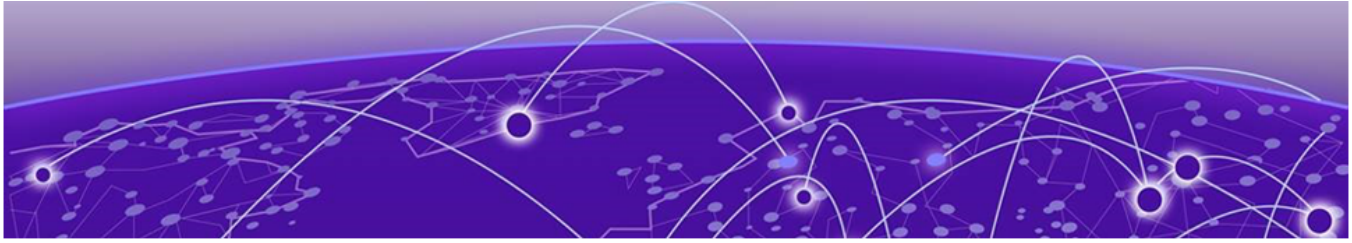
Use the **acl-range** keyword to view the Layer-4 port range hardware resource on an a-series or e-series model switch.

Each a-series and e-series ASIC has 16 Layer-4 port range checkers that are shared among the 24 1G ports. The first part of the command output lists the ports that utilizes this resource. The second part of the command output lists the number of range checkers that are consumed and the number available for use.

```
switch # show access-list usage acl-range port 4:1
Ports 4:1-4:12, 4:25-4:36
L4 Port Ranges:  Used: 0  Available: 16
```

If the **acl-slice** or **acl-range** keyword is specified with an e-series port, the following error message will appear:

```
This command is not applicable to the specified port.
```



# Routing Policies



This chapter provides information about Routing Policies. It includes an overview, specific information about Routing Policy File Syntax, how to apply Routing Policies, and offers some Policy examples.

## Routing Policies Overview

Routing policies are used to control the advertisement or recognition of routes communicated by routing protocols, such as Routing Information Protocol (*RIP (Routing Information Protocol)*), *OSPF (Open Shortest Path First)*, Intermediate System-Intermediate System (IS-IS) and *BGP (Border Gateway Protocol)*.

Routing policies can be used to "hide" entire networks or to trust only specific sources for routes or ranges of routes. The capabilities of routing policies are specific to the type of routing protocol involved, but these policies are sometimes more efficient and easier to implement than access lists.

Routing policies can also modify and filter routing information received and advertised by a switch.

A similar type of policy is an *ACL (Access Control List)* policy, used to control, at the hardware level, the packets accessing the switch. ACL policy files and routing policy files are both handled by the policy manager and the syntax for both types of files is checked by the policy manager.

Note

Although ExtremeXOS does not prohibit mixing ACL and routing type entries in a policy file, it is strongly recommended that you do not mix the entries, and you use separate policy files for ACL and routing policies.

## Routing Policy File Syntax

A routing policy file contains one or more policy rule entries. Each routing policy entry consists of:

- A policy entry rule name, unique within the same policy.
- Zero or one match type. If no type is specified, the match type is all, so all match conditions must be satisfied.
- Zero or more match conditions. If no match condition is specified, then every routing entity matches.
- Zero or more actions. If no action is specified, the packet is permitted by default.

Each policy entry in the file uses the following syntax:

```
entry <routingrulename>{
    if <match-type> {
        <match-conditions>;
    } then {
        <action>;
    }
}
```

The following is an example of a policy entry:

```
entry ip_entry {
    if match any {
        nlri  10.203.134.0/24;
        nlri  10.204.134.0/24;
    } then {
        next-hop   192.168.174.92;
        origin   egp;
    }
}
```

Policy entries are evaluated in order, from the beginning of the file to the end, as follows:

- If a match occurs, the action in the then statement is taken:
  - if the action contains an explicit permit or deny, the evaluation process terminates.
  - if the action does not contain an explicit permit or deny, the action is an implicit permit, and the evaluation process terminates.
- If a match does not occur, the next policy entry is evaluated.
- If no match has occurred after evaluating all policy entries, the default action is deny.

Often a policy has a rule entry at the end of the policy with no match conditions. This entry matches anything not otherwise processed, so that the user can specify an action to override the default deny action.

Policy match type, match conditions and action statements are discussed in the following sections:

- Policy Match Type on page 714
- Policy Match Conditions on page 715
- Autonomous System Expressions on page 716
- Policy Action Statements on page 718

## Policy Match Type

The two possible choices for the match type are:

- match all—All the match conditions must be true for a match to occur. This is the default.
- match any—If any match condition is true, then a match occurs.

## Policy Match Conditions

The following table lists the possible policy entry match conditions.

**Table 85: Policy Match Conditions**

| Match Condition | Description |
|---|---|
| as-path [<as-number> \| <as-path-regular-expression>]; | Where <as-number> is a valid 2-byte AS number in the range of 1 to 65535 or a 4-byte AS number in the range of 65536 to 4294967294.<br>Where <as-path-regular-expression> is a multi-character regular expression (with 2-byte unsigned Integer being an Atom). Regular expression will consist of the AS-Numbers and various regular expression symbols. Regular expressions must be enclosed in double quotes (""). |
| community [no-advertise \| no-export \| no-export-subconfed \| number <community_num> \| <community_regular_expression> \| <as_num> : <num>]; | Where no-advertise, no-export and no-export-subconfed are the standard communities defined by RFC. <community_num> is a four-byte unsigned integer, <as_num> is a two-byte or four-byte AS-Number and <num> is the 2-bytes community number.<br>Community regular expression is a multi-character regular expression (with four byte unsigned integer being an Atom). Regular expression is enclosed in double quotes (""). |
| med <number>; | Where <number> is a 4-byte unsigned integer. |
| next-hop [<ipaddress> \| <ipaddress-regular-expression>]; | Where <ipaddress> is a valid IP address in dotted decimal format. |
| nlri [ *ipaddress* \| any ] *mask-length* {exact}; nlri [*ipaddress* \| any] mask *mask* {exact}; nlri [ *ipv6address* \| any-ipv6 ]/ *mask-length* {exact}; | Where *ipaddress* and *mask* are IPv4 addresses and masks, *mask-length* is an integer with maximum value of 32 for IPv4 addresses. The keyword **any** matches any IPv4 address with a given (or larger) mask/mask-length.<br>Similarly *ipv6address* is an IPv6 address and *masklength* is an integer with a maximum value of 128 for IPv6 addresses. The keyword **any-ipv6** matches any IPv6 address with a given (or larger) mask-length. |
| origin [igp \| egp \| incomplete]; | Where igp, egp and incomplete are the *BGP* route origin values. |

**Table 85: Policy Match Conditions (continued)**

| Match Condition | Description |
|---|---|
| tag <number>; | Where <number> is a 4-byte unsigned number. |
| route-origin [direct \| static \| icmp \| egp \| ggp \| hello \| rip \| isis \| esis \| cisco-igrp \| ospf \| bgp \| idrp \| dvmrp \| mospf \| pim-dm \| pim-sm \| ospf-intra \| ospf-inter \| ospf-extern1 \| ospf-extern2 \| bootp \| e-bgp \| i-bgp \| mbgp \| i-mbgp \| e-mbgp \| isis-level-1 \| isis-level-2 \| isis-level-1-external \| isis-level-2-external]; | Matches the origin (different from BGP route origin) of a route.<br>A match statement "route-origin bgp" will match routes whose origin are "I-bgp" or "e-bgp" or "I-mbgp" or "e-mbgp". Similarly, the match statement "route-origin ospf" will match routes whose origin is "ospf-inta" or "ospf-inter" or "ospf-as-external" or "ospf-extern-1" or "ospf-extern-2" |

Note

When entering an AS number in a policy file, you must enter a unique 2-byte or 4-byte AS number. The transition AS number, AS 23456, is not supported in policy files.

### *Autonomous System Expressions*

The **`AS-path`** keyword uses a regular expression string to match against the autonomous system (AS) path. The following table lists the regular expressions that can be used in the match conditions for *BGP* AS path and community. It also shows examples of regular expressions and the AS paths they match.

**Table 86: AS Regular Expression Notation**

| Character | Definition |
|---|---|
| N | As number |
| N1 - N2 | Range of AS numbers, where N1 and N2 are AS numbers and N1 < N2 |
| [Nx ... Ny] | Group of AS numbers, where Nx and Ny are AS numbers or a range of AS numbers |
| [^Nx ... Ny] | Any AS numbers other than the ones in the group |
| . | Matches any number |
| ^ | Matches the beginning of the AS path |
| $ | Matches the end of the AS path |
| _ | Matches the beginning or end, or a space |
| - | Separates the beginning and end of a range of numbers |
| * | Matches 0 or more instances |
| + | Matches 1 or more instances |
| ? | Matches 0 or 1 instance |
| { | Start of AS SET segment in the AS path |
| } | End of AS SET segment in the AS path |

**Table 86: AS Regular Expression Notation (continued)**

| Character | Definition |
|---|---|
| ( | Start of a confederation segment in the AS path |
| ) | End of a confederation segment in the AS path |

**Table 87: Policy Regular Expression Examples**

| Attribute | Regular Expression | Example Matches |
|---|---|---|
| AS path is 64496 | “64496” | 64496 |
| Zero or more occurrences of AS number 1234 | “64496*” | 64496<br>64496 64496 |
| Start of AS path set | “64496 64500 { 64505” | 64496 64500 64505 { 64511<br>64509 64496 64500 { 64505 64507 |
| End of AS path set | “64500 } 64505” | 64500 } 64505 56 |
| Path that starts with 99 followed by 34 | “^64511 64505 “ | 64511 64505 45 |
| Path that ends with 99 | “64511 $” | 45 66 64511 |
| Path of any length that begins with AS numbers 64496 64497 64498 | “64496 64497 64498 .*” | 64496 64497 64498 64499 64500 64501 64502 64503 64504 |
| Path of any length that ends with AS numbers 64496 64497 64498 | “.* 64496 64497 64498 $” | 64496 64497 64498 64502 64503 64504 64496 64497 64498 |

Following are additional examples of using regular expressions in the AS-Path statement.

The following AS-Path statement matches AS paths that contain only (begin and end with) AS number 64511:

```
as-path "^64511$"
```

The following AS-Path statement matches AS paths beginning with AS number 64500, ending with AS number 64511, and containing no other AS paths:

```
as-path "^64500 64511$"
```

The following AS-Path statement matches AS paths beginning with AS number 64496, followed by any AS number from 65500 - 65505, and ending with either AS number 65507, 65509, or 65511:

```
as-path "^64496 65500-65505 [65507 65509 65511]$"
```

The following AS-Path statement matches AS paths beginning with AS number 65511 and ending with any AS number from 65500 - 65505:

```
as-path "65511 [65500-65505]$"
```

The following AS-Path statement matches AS paths beginning with AS number 65511 and ending with any additional AS number, or beginning and ending with AS number 65511:

```
as-path "65511 .?"
```

## Policy Action Statements

The following table lists policy action statements. These are the actions taken when the policy match conditions are met in a policy entry.

**Table 88: Policy Actions**

| Action | Description |
|---|---|
| as-path "<as_num> {<as_num1> <as_num2> <as_num3> .... <as_numN>}"; | Prepends the entire list of as-numbers to the as-path of the route. |
| community set [no-advertise \| no-export \| noexport- subconfed \| <community_num> \| <as_num> : <community_num>]; | Replaces the existing community attribute of a route by the community specified by the action statement. Community must be enclosed in double quotes (""). |
| community [add \| delete] [no-advertise \| no-export \| no-export-subconfed \| <community_num> {<community_num1> <community_num2> .... <community_numN>} \| <as_num> : <community_num> {<as_num1> <community_num1> <as_num2> <community_num2> ....}]; | Adds/deletes communities to/from a route's community attribute. Communities must be enclosed in double quotes (""). |
| community remove; | Strips off the entire community attribute from a route. Communities must be enclosed in double quotes (""). |
| cost <cost(0-4261412864)>; | Sets the cost/metric for a route. |
| cost-type {ase-type-1 \| ase-type-2 \| external \| internal}; | Sets the cost type for a route. |
| dampening half-life <minutes (1-45)> reuse-limit <number (1-20000)> suppress-limit <number (1-20000)> max-suppress <minutes (1-255)>; | Sets the *BGP* route flap dampening parameters. |
| deny; | Denies the route. |
| local-preference <number>; | Sets the BGP local preference for a route. |
| med {add \| delete} <number>; | Performs MED arithmetic. Add means the value of the MED in the route will be incremented by <number>, and delete means the value of the MED in the route will be decremented by <number>. |
| med {internal \| remove}; | Internal means that the Interior Gateway Protocol (IGP) distance to the next hop will be taken as the MED for a route. Remove means take out the MED attribute from the route. |
| med set <number>; | Sets the MED attribute for a route. |
| next-hop <ipv4 address \| ipv6 address>; | Sets the next hop attribute for a route. |

**Table 88: Policy Actions (continued)**

| Action | Description |
|---|---|
| nlri [<ipaddress> \| any]/<mask-length> {exact};nlri [<ipaddress> \| any] mask <mask> {exact}; | These set statements are used for building a list of IP addresses. This is used by PIM to set up the RP list. |
| origin {igp \| egp \| incomplete}; | Sets the BGP route origin values. |
| permit; | Permits the route. |
| tag <number>; | Sets the tag number for a route. |
| weight <number> | Sets the weight for a BGP route. |

Note

Multiple communities couldn't be used in "community set" attribute in the BGP policy file. The way to set multiple communities in BGP policy file could be accomplished by two set of attributes as given in below example:

```
entry permit-anything-else {
   if {
   } then {
     community set "2342:6788";
     community add "2342:6789 2342:6790";
 permit;
   }
}
```

## Applying Routing Policies

To apply a routing policy, use the command appropriate to the client. Different protocols support different ways to apply policies, but there are some generalities.

Commands that use the keyword **import-policy** are used to change the attributes of routes installed into the switch routing table by the protocol. These commands cannot be used to determine the routes to be added to the routing table. The following are examples for the *BGP* and *RIP* protocols:

`configure bgp import-policy [`*policy-name* `|` **none**`]`

`configure rip import-policy [`*policy-name* `|` **none**`]`

Commands that use the keyword **route-policy** control the routes advertised or received by the protocol. For BGP and RIP, here are some examples:

`configure bgp neighbor [`*remoteaddr* `|` **all**`] {`**address-family** `[`**ipv4-unicast** `|` **ipv4-multicast** `|`**ipv6-unicast** `|` **ipv6-multicast** `|`**vpnv4**`]}` **route-policy** `[`**in** `|` **out**`] [`**none** `|` *policy*`]`

`configure bgp peer-group` *peer-group-name* `{`**address-family** `[`**ipv4-unicast** `|` **ipv4-multicast** `|`**ipv6-unicast** `|` **ipv6-multicast** `|`**vpnv4**`]}` **route-policy** `[`**in** `|`**out**`] [`**none** `|` *policy*`]`

```
configure rip vlan [vlan_name | all] route-policy [in | out] [policy-
name | none]
```

Other examples of commands that use routing policies include:

```
configure ospf area area-identifier external-filter [policy-map |none]
```

```
configure ospf add vlan [vlan-name | all] area area-identifier {passive}
{vr vrf_name}
```

```
configure rip vlan [vlan_name | all] trusted-gateway [policy-name |
none]
```

To remove a routing policy, use the **none** option in the command.

## Policy Examples

### Translating an Access Profile to a Policy

You may be more familiar with using access profiles on other Extreme Networks switches. This example shows the policy equivalent to an ExtremeWare access profile.

ExtremeWare Access-Profile:

```
Seq_No Action     IP Address       IP Mask            Exact
5      permit     22.16.0.0        255.252.0.0        No
10     permit     192.168.0.0      255.255.192.0      Yes
15     deny       any              255.0.0.0          No
20     permit     10.10.0.0        255.255.192.0      No
25     deny       22.44.66.0       255.255.254.0      Yes
```

Equivalent ExtremeXOS policy map definition:

```
entry  entry-5  {
    If  {
        nlri    22.16.0.0/14;
    }
    then  {
        permit;
    }
}
entry  entry-10  {
    if  {
        nlri   192.168.0.0/18 exact;
    }
    then  {
        permit;
    }
}
entry  entry-15  {
    if  {
        nlri   any/8;
    }
    then  {
        deny;
    }
```

```
}
entry  entry-20  {
    if  {
        nlri   10.10.0.0/18;
}
Then  {
    permit;
}
}
entry  entry-25  {
    if  {
        nlri   22.44.66.0/23  exact;
    }
    then  {
        deny;
    }
}
```

The policy above can be optimized by combining some of the if statements into a single expression.

The compact form of the policy looks like this:

```
entry  permit_entry  {
    If match any {
        nlri    22.16.0.0/14;
        nlri   192.168.0.0/18 exact ;
        nlri   10.10.0.0/18;
    }
    then  {
        permit;
    }
}
entry  deny_entry  {
    if match any {
        nlri   any/8;
        nlri   22.44.66.0/23  exact;
    }
    then  {
        deny;
    }
}
```

## Translating a Route Map to a Policy

You may be more familiar with using route maps on other Extreme Networks switches. This example shows the policy equivalent to an ExtremeWare route map.

ExtremeWare route map:

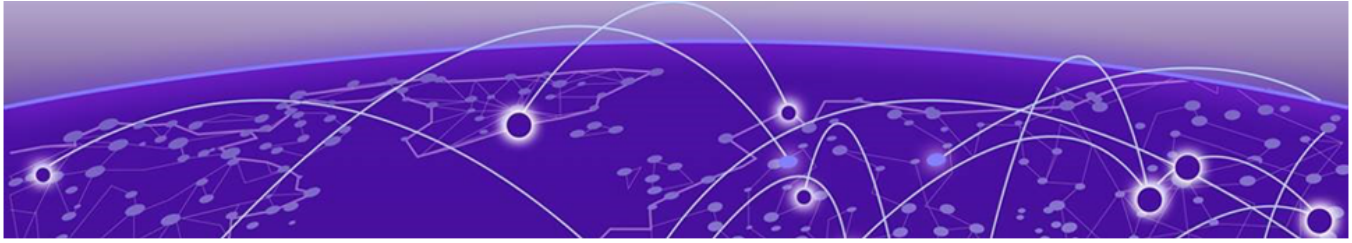
```
Route Map : rt
Entry : 10     Action : permit
match origin incomplete
Entry : 20     Action : deny
match community  6553800
Entry : 30     Action : permit
match med  30
set next-hop 10.201.23.10
set as-path 64502
set as-path 64503
set as-path 64504
```

```
set as-path 64504
Entry : 40     Action : permit
set local-preference 120
set weight 2
Entry : 50     Action : permit
match origin incomplete
match community  19661200
set dampening half-life 20 reuse-limit 1000 suppress-limit 3000 max-suppress 40
Entry : 60     Action : permit
match next-hop 192.168.1.5
set community add 949616660
```

Equivalent policy:

```
entry  entry-10  {
    If  {
        origin    incomplete;
    }
    then  {
        permit;
    }
}
entry  entry-20  {
    if  {
        community   6553800;
    }
    then  {
        deny;
    }
}
entry  entry-30  {
    if  {
        med   30;
    }
    then  {
        next-hop     10.201.23.10;
        as-path 64502;
        as-path 64503;
        as-path     64504;
        as-path     64504;
        permit;
    }
}
entry  entry-40  {
    if  {
    }
    then  {
        local-preference 120;
        weight         2;
        permit;
}
}
entry  entry-50 match any {
    if  {
        origin         incomplete;
        community     19661200;
    }
    then  {
        dampening half-life 20 reuse-limit 1000 suppress-limit 3000 max-suppress 40
        permit;
    }
}
entry  entry-60 {
```

```
    if  {
        next-hop 192.168.1.5;
    }
    then  {
        community  add  949616660;
        permit;
    }
}
entry  deny_rest {
    if  {
    }
    then  {
        deny;
    }
}
```



# Quality of Service



This chapter discusses the *QoS (Quality of Service)* feature, and allows you to configure a switch to provide different levels of service to different groups of traffic. In this section you will find both overview information, as well as specific information on how to configure and monitor the QoS feature.

## QoS Overview

Quality of Service (QoS) is a feature that allows you to configure a switch to provide different levels of service to different groups of traffic. For example, QoS allows you to do the following:

- Give some traffic groups higher priority access to network resources.
- Reserve bandwidth for special traffic groups.
- Restrict some traffic groups to bandwidth or data rates defined in a Service Level Agreement (SLA).
- Count frames and packets that exceed specified limits and optionally discard them (rate limiting).
- Queue or buffer frames and packets that exceed specified limits and forward them later (rate shaping).
- Modify QoS related fields in forwarded frames and packets (remarking).

The following figure shows the QoS components that provide these features on Extreme Networks switches.

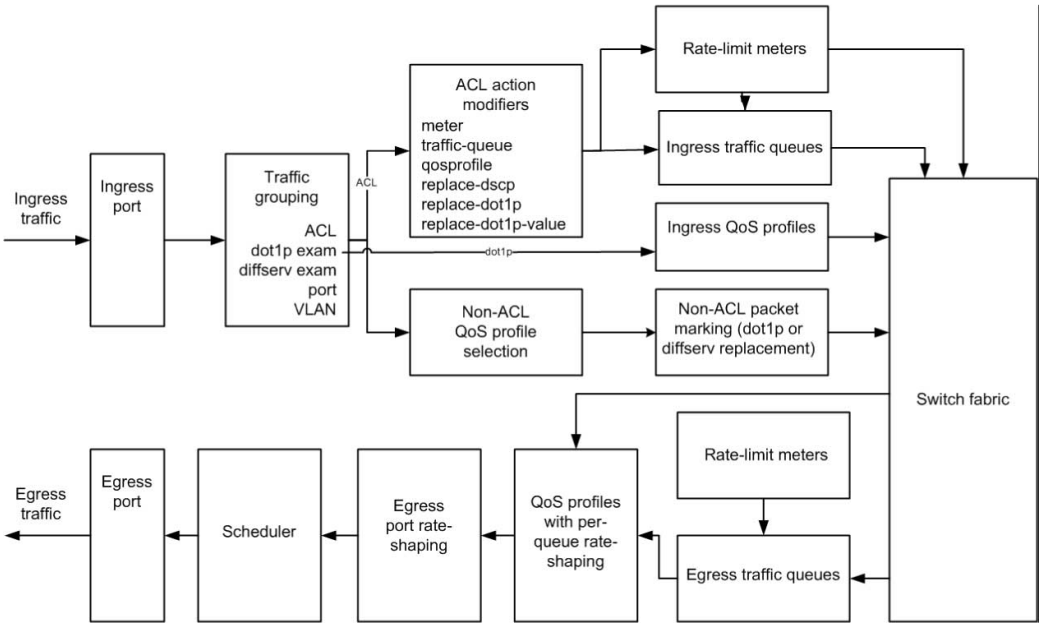



**Figure 97: QoS on Extreme Networks Switches**

In the figure above, data enters the ingress port and is sorted into traffic groups, which can be classified as either *ACL (Access Control List)*-based or non-ACL-based.

The ACL-based traffic groups provide the most control of QoS features and can be used to apply ingress and egress rate limiting and rate shaping as follows:

- Subject ingress traffic to rate limit meters
- Specify ingress hardware queues (QoS profiles) for rate limiting and rate shaping
- Specify ingress software traffic queues for rate limiting and rate shaping (these can be associated with egress traffic queues for additional QoS control)
- Specify egress software traffic queues for rate limiting and rate shaping
- Specify egress QoS profiles for rate limiting and rate shaping
- Change the dot1p or Differential Services (DiffServ) values in egress frames or packets

Non-ACL-based traffic groups specify an ingress or egress QoS profile for rate limiting and rate shaping. These groups cannot use ingress or egress software traffic queues. However, non-ACL-based traffic groups can use the packet marking feature to change the dot1p or DiffServ values in egress frames or packets.

The ingress rate-limiting and rate-shaping features allow you to apply QoS to incoming traffic before it reaches the switch fabric. If some out-of-profile traffic needs to be dropped, it is better to drop it before it consumes resources in the switch fabric.

All ingress traffic is linked to an egress traffic queue or QoS profile before it reaches the switch fabric. This information is forwarded with the traffic to the egress interface, where it selects the appropriate egress traffic queue or QoS profile. Egress traffic from all traffic queues and QoS profiles is forwarded to the egress port rate-shaping feature, which applies QoS to the entire port. When multiple QoS profiles are contending for egress bandwidth, the scheduler determines which queues are serviced.

The following sections provide more information on QoS:



## Applications and Types of QoS

Different applications have different *QoS* requirements. The following table summarizes the QoS guidelines for different types of network traffic.

**Table 89: Traffic Type and QoS Guidelines**

| Traffic Type | Key QoS Parameters |
|---|---|
| Voice | Minimum bandwidth, priority |
| Video | Medium bandwidth, priority, buffering (varies) |
| Database | Minimum bandwidth |
| Web browsing | Minimum bandwidth for critical applications, maximum bandwidth for noncritical applications |
| File server | Minimum bandwidth |

Consider the parameters in the table above as general guidelines and not as strict recommendations. After QoS parameters have been set, you can monitor the performance of the application to determine if the actual behavior of the applications matches your expectations. It is very important to understand the needs and behavior of the particular applications you want to protect or limit. Behavioral aspects to consider include bandwidth needs, sensitivity to latency and jitter, and sensitivity and impact of packet loss.

Note

Full-duplex links should be used when deploying policy-based QoS. Half-duplex operation on links can make delivery of guaranteed minimum bandwidth impossible.

### Voice Applications

Voice applications, or voice over IP (VoIP), typically demand small amounts of bandwidth. However, the bandwidth must be constant and predictable because voice applications are typically sensitive to latency (inter-packet delay) and jitter (variation in inter-packet delay). The most important *QoS* parameter to establish for voice applications is minimum bandwidth, followed by priority.

### Video Applications

Video applications are similar in needs to voice applications, with the exception that bandwidth requirements are somewhat larger, depending on the encoding. It is important to understand the behavior of the video application being used. For example, in the playback of stored video streams, some applications can transmit large amounts of data for multiple streams in one spike, with the expectation that the end stations will buffer significant amounts of video-stream data. This can present a problem to the network infrastructure, because the network must be capable of buffering the transmitted spikes where there are speed differences (for example, going from gigabit Ethernet to Fast Ethernet). Key *QoS* parameters for video applications include minimum bandwidth and priority, and possibly buffering (depending upon the behavior of the application).

### Critical Database Applications

Database applications, such as those associated with Enterprise Resource Planning (ERP), typically do not demand significant bandwidth and are tolerant of delay. You can establish a minimum bandwidth using a priority less than that of delay-sensitive applications.

### web Browsing Applications

*QoS* needs for web browsing applications cannot be generalized into a single category. For example, ERP applications that use a browser front-end might be more important than retrieving daily news information. Traffic groupings can typically be distinguished from each other by their server source and destinations. Most browser-based applications are distinguished by the dataflow being asymmetric (small dataflows from the browser client, large dataflows from the server to the browser client).

An exception to this might be created by some Java™-based applications. In addition, web-based applications are generally tolerant of latency, jitter, and some packet loss; however, small packet loss might have a large impact on perceived performance because of the nature of TCP. The relevant parameter for protecting browser applications is minimum bandwidth. The relevant parameter for preventing non-critical browser applications from overwhelming the network is maximum bandwidth.

### File Server Applications

With some dependencies on the network operating system, file serving typically poses the greatest demand on bandwidth, although file server applications are very tolerant of latency, jitter, and some packet loss, depending on the network operating system and the use of TCP or UDP.

## Traffic Groups

A traffic group defines the ingress traffic to which you want to apply some level of *QoS*. You can use the ExtremeXOS software to define traffic groups based on the following:

- Frame or packet header information such as IP address or MAC address
- *CoS (Class of Service)* 802.1p bits in the frame header
- DiffServ information in a packet header
- Ingress port number
- *VLAN (Virtual LAN)* ID

Traffic groups that are defined based on frame or packet information are usually defined in Access Control Lists (ACLs). The exception to this rule is the CoS and DiffServ information, which you can use to define traffic groups without ACLs.

The function of the CoS and DiffServ traffic groups is sometimes referred to as explicit packet marking, and it uses information contained within a frame or packet to explicitly determine a class of service. An advantage of explicit packet marking is that the class of service information can be carried throughout the network infrastructure, without repeating what can be complex traffic group policies at each switch location. Another advantage is that end stations can perform their own packet marking on an application-specific basis. Extreme Networks switch products have the capability of observing and manipulating packet marking information with no performance penalty.

The CoS and DiffServ capabilities (on supported platforms) are not impacted by the switching or routing configuration of the switch. For example, 802.1p information can be preserved across a routed switch boundary and DiffServ code points can be observed or overwritten across a Layer 2 switch boundary.

During QoS configuration, you configure the QoS level first by configuring QoS profiles, traffic queues, and meters, and then you define a traffic group and assign the traffic group to the QoS configuration.

## ACL-Based Traffic Groups

An *ACL*-based traffic group allows you to use ACL rules in an ACL policy file to define the traffic to which you want to apply *QoS*. An ACL-based traffic group requires more effort to create, but the ACL rules give you more control over which traffic is selected for the traffic group. For example, you can use an ACL to add traffic to a traffic group based on the following frame or packet components:

- MAC source or destination address
- Ethertype
- IP source or destination address
- IP protocol
- TCP flag
- TCP, UDP, or other Layer 4 protocol
- TCP or UDP port information
- IP fragmentation

Depending on the platform you are using, traffic in an ACL traffic group can be processed as follows:

- Assigned to an ingress meter for rate limiting
- Marked for an egress QoS profile for rate shaping
- Marked for an egress traffic queue for rate shaping
- Marked for DSCP replacement on egress
- Marked for 802.1p priority replacement on egress
- Assigned to an egress meter for rate limiting

When you are deciding whether to use an ACL-based traffic group or another type of traffic group, consider what QoS features you want to apply to the traffic group. Some QoS features can only apply to ACL-based traffic groups.

Note
ACLs are discussed in detail in the ACLs chapter.

## CoS 802.1p-Based Traffic Groups

CoS 802.1p-based traffic groups forward traffic to *QoS* features based on the three 802.1p priority bits in an Ethernet frame. The 802.1p priority bits are located between the 802.1Q type field and the 802.1Q *VLAN* ID as shown in the following figure.

Typically *CoS* will be used in conjunction with policy, however it may also work independently. In a policy-enabled system, policy may use any CoS index setting (0-255) to configure ACLs to assign a qosprofile, meter, dot1p replacement and TOS-rewrite value. By default, 802.1p priorities 0-7 are mapped to CoS indices 0-7.

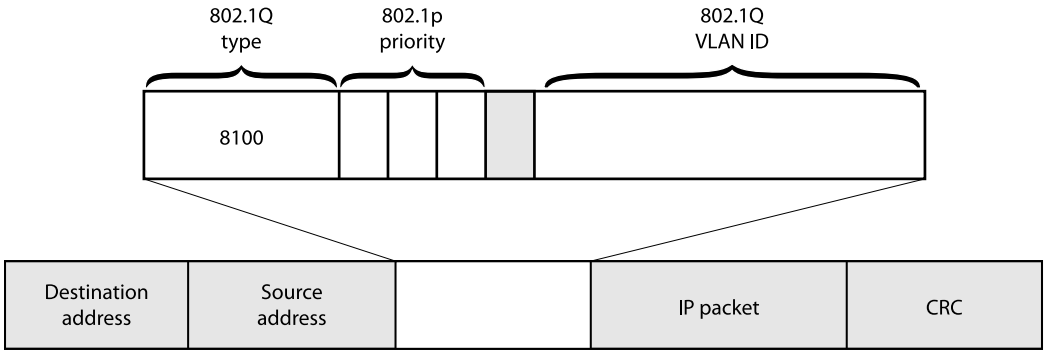



**Figure 98: 802.1p Priority Bits**

The three 802.1p priority bits define up to eight traffic groups that are predefined in the ExtremeXOS software.

On BlackDiamond X8, 8800, SummitStack, and Summit family switches, the traffic groups direct traffic to egress QoS profiles for egress rate shaping (see the following table).

Note
See #unique_1425 for information regarding VMANs using 802.1p information to direct frames to appropriate egress QoS queues.

You do not need to define 802.1p-based traffic groups. You can enable or disable the use of these traffic groups by enabling or disabling the 802.1p examination feature. You can also configure which 802.1p values map to which QoS profiles.

A related feature is the 802.1p replacement feature, which allows you to configure the software to replace the 802.1p bits in an ingress frame with a different value in the egress frame. For more information on 802.1p replacement, see Configuring 802.1p or DSCP Replacement on page 744.

## DiffServ-Based Traffic Groups

DiffServ-based traffic groups forward traffic to egress QoS profiles based on the type-of-service (TOS) or traffic class (TC) information in an IP packet. In many systems, this TOS or TC information is replaced with a DiffServ field that uses six of the eight bits for a DiffServ code point (DSCP) as shown in the following figure. (The other two bits are not used.)

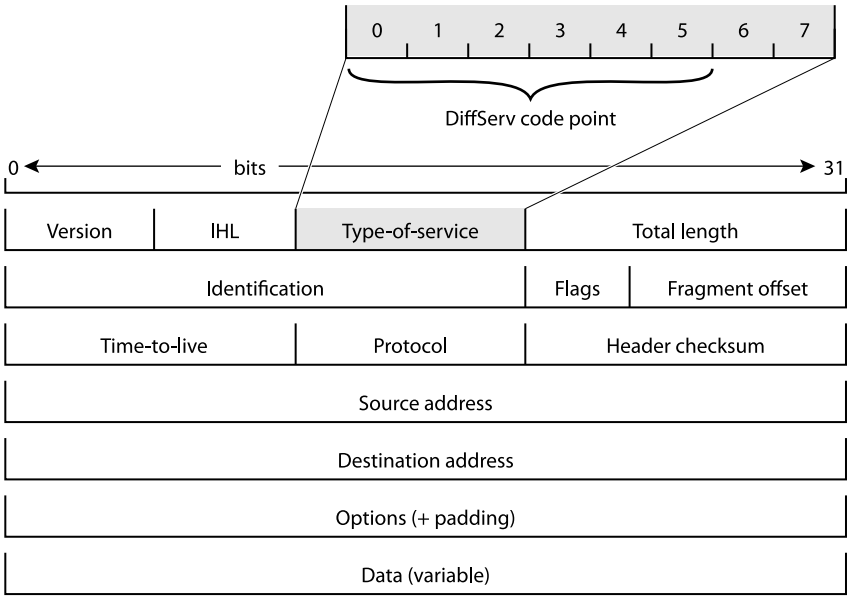



**Figure 99: DiffServe Code Point**

Because the DSCP uses six bits, it has 64 possible values (2^6 = 64). By default, the values are grouped and assigned to the default QoS profiles as listed in the following table.

**Table 90: Default DSCP-to-QoS-Profile Mapping**

| Traffic Group Code Point | BlackDiamond 8800 and X8 Series Switches, SummitStack, and Summit Family Switches QoS Profile |
|---|---|
| 0-7 | QP1 |
| 8-15 | QP1 |
| 16-23 | QP1 |
| 24-31 | QP1 |
| 32-39 | QP1 |
| 40-47 | QP1 |

**Table 90: Default DSCP-to-QoS-Profile Mapping (continued)**

| Traffic Group Code Point | BlackDiamond 8800 and X8 Series Switches, SummitStack, and Summit Family Switches QoS Profile |
|---|---|
| 48-55 | QP1 |
| 56-63 | QP8 |

**Note**
The default DiffServ examination mappings apply on ports in more than one VR. If you attempt to configure DiffServe examination or replacement on a port that is in more than one *virtual router (VR)*, the system returns the following message:

```
Warning: Port belongs to more than one VR. Port properties related to diff serv
and code replacement will not take effect.
```

You do not need to define these traffic groups. You can enable or disable the use of these traffic groups by enabling or disabling the DiffServ examination feature as described in Configuring a DiffServ-Based Traffic Group on page 750. You can also configure which DSCP values map to which queues.

**Note**
When DiffServ examination is enabled on 1 Gigabit Ethernet ports for BlackDiamond 8800 series switches, SummitStack, and Summit family switches, 802.1p replacement is enabled and cannot be disabled. The ingress 802.1p value is replaced with the 802.1p value assigned to the egress QoS profile.

A related feature is the DiffServ replacement feature, which allows you to configure the software to replace the DSCP in an ingress frame with a different value in the egress frame. For more information on DiffServ replacement, see Configuring 802.1p or DSCP Replacement on page 744.

## Port-Based Traffic Groups

Port-based traffic groups forward traffic to egress *QoS* profiles based on the incoming port number. There are no default port-based traffic groups; you must configure each port-based traffic group.

**Note**
On BlackDiamond X8 series switches, BlackDiamond 8800 series switches, SummitStack, and Summit family switches, port-based traffic groups apply to all packets.

## VLAN-Based Traffic Groups

*VLAN*-based traffic groups forward traffic to egress *QoS* profiles based on the VLAN membership of the ingress port. There are no default VLAN-based traffic groups; you must configure each VLAN-based traffic group.

**Note**
On BlackDiamond X8 series switches, BlackDiamond 8800 series switches, SummitStack, and Summit family switches, VLAN-based traffic groups apply to all packets.

## Precedence of Traffic Groups

The ExtremeXOS software allows you to define multiple traffic groups, and you can configure traffic groups in such a way that multiple traffic groups apply to an ingress frame or packet. When an ingress frame or packet matches two or more traffic groups, the software chooses one traffic group based on the precedence defined for the switch platform. In general, the more specific traffic group definition takes precedence. The following table shows the traffic group precedence for the supported switch platforms (number 1 is the highest precedence).

Note

Beginning in ExtremeXOS 16.1, the port, *VLAN* and DiffServ-based traffic groups will only work if 802.1p examination is disabled.

**Table 91: Traffic Group Precedence**

| BlackDiamond X8 series switches, BlackDiamond 8800 Series Switches, SummitStack, and Summit Family Switches |
|---|
| 1. *ACL*-based traffic groups for IP packets (specifies IP address information)<br>2. ACL-based traffic groups for Ethernet frames (specifies MAC address information)<br>3. *CoS* 802.1p-based traffic groups<br>4. Port-based traffic groups<br>5. VLAN-based traffic groups<br>6. DiffServ-based traffic groups |

# Introduction to Rate Limiting, Rate Shaping, and Scheduling

The terms *rate limiting* and *rate shaping* are used throughout this chapter to describe *QoS* features. Some QoS features perform both rate limiting and rate shaping. Rate limiting is the process of restricting traffic to a peak rate (PR). Rate shaping is the process of reshaping traffic throughput to give preference to higher priority traffic or to buffer traffic until forwarding resources become available.

Both rate limiting and rate shaping allow you to take action on traffic that exceeds the configured limits. These actions include forwarding traffic, dropping traffic, and marking the excess traffic for possible drops later in the communication path. Software counters allow you to record traffic statistics such as total packets forwarded and total packets dropped.

## Single-Rate QoS

Single-rate *QoS* defines a single rate for traffic that is subject to QoS. Single-rate QoS is the most basic form of rate limiting and is well suited for constant rate traffic flows such as video or where more complex dual-rate QoS is not needed. The traffic that meets the rate requirement is considered in-profile. Traffic that does not meet the specified rate is considered out-of-profile. A two-color system is often used to describe or mark the single-rate QoS result. In-profile traffic is marked green, and out-of-profile traffic is marked red.

Single-rate rate-limiters pass traffic that is in-profile or marked green. Out-of-profile traffic (marked red) is subject to whatever action is configured for out-of-profile traffic. Out of profile traffic can be forwarded if bandwidth is available, dropped, or marked for a possible drop later in the communication path.

All traffic that arrives at or below the PR is considered in-profile and marked green. All traffic that arrives above the PR is considered out-of-profile and marked red. When the traffic exceeds the capacity of the rate-limiting component, all traffic that is marked red is dropped.

Another type of single-rate QoS is used on BlackDiamond 8800 switches, SummitStack, and Summit family switches. A committed information rate (CIR) establishes a reserved traffic rate, and a peak burst size (PBS) establishes a maximum size for a traffic stream. If a traffic stream is at or below the CIR and the PBS, it is considered to be within profile and marked green. If a traffic stream exceeds either the CIR or the PBS, it is considered out-of-profile and marked red. On these switches, you can configure the single-rate rate-limiting components to drop traffic marked red or set a drop precedence for that traffic. You can also specify a DSCP value to mark the out-of-profile traffic.

## Dual-rate QoS

Dual-rate *QoS* defines two rates for traffic that is subject to QoS. The lower of the two rates is the CIR, which establishes a reserved traffic rate. The higher of the two rates is the peak rate, which establishes an upper limit for traffic. Dual-rate QoS is well suited to bursty traffic patterns which are common with typical data traffic. Dual-rate QoS is widely used in legacy Frame Relay and ATM leased lines.

Note

You must configure the peak rate higher than the committed rate.

A three-color system is used with dual-rate QoS. As with single-rate QoS, traffic at or below the CIR is considered in-profile, marked green, and forwarded. The traffic that is above the CIR and below the PIR is out-of-profile and marked yellow. Traffic above the PIR is also out-of-profile and marked red. When incoming traffic is already marked yellow and is out of profile, it is marked red. Different switch platforms take different actions on the traffic marked yellow or red.

## Rate Specification Options

The ExtremeXOS software allows you to specify the CIR and PR in gigabits per second (Gbps), megabits per second (Mbps), or kilobits per second (Kbps). Most commands also allow you to specify the CIR and PR as a percentage of the maximum port bandwidth using the **minbw** (CIR) and **maxbw** (PR) options. The default value on all minimum bandwidth parameters is 0%, and the default value on all maximum bandwidth parameters is 100%.

*QoS* can be applied at different locations in the traffic path using the following rate-limiting and rate-shaping components:

- Ingress meters
- Egress traffic queues
- Egress meters
- Egress QoS profiles
- Egress ports

The CIR or minimum bandwidth configuration for a rate-limiting or rate-shaping component is a bandwidth guarantee for that component at a particular location in the traffic path. The guarantees for all components at a specified location should add up to less than 100% and should account for the traffic needs of the other components. For example, if you configure 25% minimum bandwidth for four

out of eight queues at a particular location, there will be no available bandwidth for the remaining four queues when traffic exceeds the port capacity. Bandwidth unused by a queue can be used by other queues.

The rate-shaping configuration is configured at the location to which it applies on most platforms.

## Disabling Rate Limiting and Rate Shaping

All switch platforms provide multiple *QoS* components in the traffic path that provide rate limiting or rate shaping. These components give you control over where and how the rate shaping is applied. However, your application might not require rate shaping at every component. The default configuration for most components provides no rate shaping. When rate shaping is disabled on a component, the CIR is set to 0 (minbw=0%) and the PR is set to the maximum bandwidth (maxbw=100%). This setting reserves no bandwidth for the component and allows the component to use 100% of the port bandwidth. If you need to remove rate shaping from a QoS component, configure these settings on that component.

## Scheduling

Scheduling is the process that determines which traffic is forwarded when multiple *QoS* components are competing for egress bandwidth. The ExtremeXOS software supports the following scheduling methods:

- Strict priority queuing: All higher priority queues are serviced before lower priority queues. This ensures that high priority traffic receives access to available network bandwidth immediately, but can result in lower priority traffic being starved. As long as a queued packet remains in a higher-priority queue, any lower-priority queues are not serviced.
- Weighted fair queuing: All queues are given access to a relative amount of bandwidth based on the weight assigned to the queue. When you configure a QoS profile with a weight of 4, that queue is serviced four times as frequently as a queue with a weight of 1. The hardware services higher-weighted queues more frequently, but lower-weighted queues continue to be serviced at all times. Initially, the weight for all queues is set to 1, which gives them equal weight. If all queues are set to 4, for example, all queues still have equal weight, but each queue is serviced for a longer period.
- Round-robin priority: All queues are given access based on the configured priority level and a round-robin algorithm.

Scheduling takes place on the egress interface and includes consideration for the color-marking of egress frames and packets. Green-marked traffic has the highest priority and is forwarded based on the scheduling method. When multiple queues are competing for bandwidth, yellow-marked traffic might be dropped or remarked red. Red-marked traffic is dropped when no bandwidth is available. If yellow-marked traffic is forwarded to the egress port rate-shaping component, it can be dropped there if the egress port is congested.

*Limitation*

- Summit X670-G2 and X770 only: In hybrid scheduling (WRR/WDRR,SP), strict priority (use-strict-priority) is allowed only for contiguous queues on the port. Switch displays the folllowing error message when SP is configured on non contiguous queues: `Error: strict-priority queues must be contiguous on this platform`

## Introduction to WRED

The weighted random early detection (WRED) feature is supported on some platforms to avoid congestion in traffic queues or *QoS* profiles. WRED improves upon the TCP congestion control mechanism.

The TCP congestion control mechanism on hosts detects congestion when packets are lost and lowers the packet transmission rate in response. At the switch, packets are dropped when a queue is full. When multiple hosts forward packets that are dropped, multiple hosts reduce the transmission rate. This creates a global synchronization problem as multiple hosts overwhelm a queue and then simultaneously lower their transmission rate and under utilize the queue.

WRED is an extension to random early detection (RED), which calculates an average queue size and randomly discards packets in proportion to the queue usage. At low usage levels, no packets are discarded. As the average queue size exceeds configured thresholds, packets are discarded in proportion to the queue usage. Discarding packets early causes some (but not all) hosts to reduce their transmission rate, which reduces queue congestion. The random nature of the discard process reduces the global synchronization problem.

WRED extends RED by applying different discard rules for different types of traffic. Typically, WRED is used on core routers and takes action based on the packet contents established at edge routers. Edge routers can use the IP precedence or DSCP value to mark packets as committed (green), conforming (yellow), or exceeded (red). The marking process and these colors are described in Introduction to Rate Limiting, Rate Shaping, and Scheduling on page 732.

The ExtremeXOS WRED implementation varies per platform and allows you to configure the following:

- Minimum threshold for dropped packets.
- Maximum threshold for dropped packets.
- Maximum drop rate.
- An average weight control that determines how WRED calculates the average queue size.

WRED does not drop packets at calculated average queue sizes below the minimum threshold (although packets would be dropped if the queue fills).

When the calculated average queue size rises above the maximum threshold, WRED drops packets at the maximum drop rate. When the calculated average falls between the minimum and maximum thresholds, packets are randomly dropped at a proportional rate between zero and the maximum drop rate. As the queue fills, more packets are dropped.

The average weight parameter provides some control over how the average queue size is calculated and the probability of packet drop. Increasing the avg_weight value reduces the probability that traffic is dropped. Conversely, decreasing the avg_weight value increases the probability that traffic is dropped.

On BlackDiamond 8900 c- and xl-series modules and Summit X460 and X480 switches, you can configure up to three WRED profiles or configurations per QoS profile, enabling you to create custom WRED configurations for up to 24 traffic flows (3 WRED profiles x 8 QoS profiles).

Each QoS profile supports WRED profiles for the following colors of traffic:

- TCP green

- TCP red
- Non-TCP any

On BlackDiamond X8 series switches, BlackDiamond 8900 xm-series modules, and Summit X460-G2, X670, X670-G2, and X770 switches, you can configure up to four WRED profiles or configurations per QoS profile, enabling you to create custom WRED configurations for up to 32 traffic flows (4 WRED profiles x 8 QoS profiles).

Each QoS profile supports WRED profiles for the following colors of traffic:

- TCP green
- Non-TCP green
- TCP red
- Non-TCP red

Without support for non-TCP traffic management, non-TCP traffic could monopolize a QoS profile in which TCP traffic is regulated, effectively giving non-TCP traffic priority over TCP traffic. With support for both TCP and non-TCP traffic, WRED allows you to regulate different types of traffic independently, giving you greater control over which type of traffic is dropped first and most frequently.

The typical WRED configuration establishes the lowest probability for packet drop for green traffic, which conforms to established limits along the transmission path. A typical WRED configuration establishes a higher probability for packet drop for red colored traffic, because it has already exceeded established limits earlier in the transmission path. All traffic (green and red) is dropped when the queue is full, so the goal is to configure the WRED settings for each color in such a way as to prevent the queue from filling frequently.

## Meters

Meters are used to define ingress rate-limiting and rate-shaping on BlackDiamond X8 and 8800 series, SummitStack, and Summit family switches. Some platforms also support meters for egress traffic. The following sections provide information on meters for specific platforms.

### Meters on BlackDiamond X8 and 8800 Series Switches, SummitStack, and Summit Family Switches

The BlackDiamond X8 and 8800 series switches, SummitStack, and Summit family switches use a single-rate meter to determine if ingress traffic is in-profile or out-of-profile.

On BlackDiamond X8, c-, xl-, and xm-series modules and Summit X450-G2, X460-G2, X480, X670, X670-G2, and X770 switches, you can also use single-rate meters to determine if egress traffic is in-profile or out-of-profile.

When *ACL* meters are applied to a *VLAN* or to any, the rate limit is applied to each port group. To determine which ports are contained within a port group, use any of the following commands:

```
show access-list usage acl-range port port

show access-list usage acl-rule port port

show access-list usage acl-slice port port
```

The out-of-profile actions are drop, set the drop precedence, or mark the DSCP with a configured value. Additionally, each meter has an associated out-of-profile counter that counts the number of packets that were above the committed-rate (and subject to the out-of-profile-action).

On BlackDiamond X8 and 8000 series modules and Summit family switches, the meters are a per-chip, per-slice resource (see ACLs for complete information.)

# QoS Profiles

*QoS* profiles are queues that provide ingress or egress rate limiting and rate shaping. The following sections provide more information on QoS profiles.

## Egress Qos Profiles

Egress *QoS* profiles are supported on all ExtremeXOS switches and allow you to provide dual-rate egress rate-shaping for all traffic groups on all egress ports. Any configuration you apply to an egress QoS profile is applied to the same egress QoS profile on all other ports, unless a QoS profile parameter has been overridden for a port.

When you are configuring *ACL*-based traffic groups, you can use the qosprofile action modifier to select an egress QoS profile. For DiffServ-, port-, and *VLAN*-based traffic groups, the traffic group configuration selects the egress QoS profile. For *CoS* dot1p traffic groups on all platforms, the dot1p value selects the egress QoS profile.

Egress QoS profile operation depends on the switch type and is described in the following sections.

### *Egress QoS Profiles on BlackDiamond X8 series switches, BlackDiamond 8800 and Summit Family Switches*

BlackDiamond X8 series switches, BlackDiamond 8800 series switches, SummitStack, and Summit family switches have two default egress QoS profiles named QP1 and QP8. You can configure up to six additional QoS profiles (QP2 through QP7) on the switch. However, on a SummitStack, you cannot create QoS profile QP7, as this profile is reserved for system control traffic. The default settings for egress QoS profiles are summarized in the following table.

**Table 92: Default QoS Profile Parameters on all Platforms**

| Ingress 802.1p Priority Value | Egress QoS Profile Name | Queue Service Priority Value | Buffer | Weight | Notes |
|---|---|---|---|---|---|
| 0-6 | QP1 | 1 (Low) | 100% | 1 | This QoS profile is part of the default configuration and cannot be deleted. |
| | QP2 | 2 (LowHi) | 100% | 1 | You must create this QoS profile before using it. |
| | QP3 | 3 (Normal) | 100% | 1 | You must create this QoS profile before using it. |

12 The QoS Profile Name cannot be changed.
13 The Queue Service Priority value cannot be changed.

**Table 92: Default QoS Profile Parameters on all Platforms (continued)**

| Ingress 802.1p Priority Value | Egress QoS Profile Name | Queue Service Priority Value | Buffer | Weight | Notes |
|---|---|---|---|---|---|
| | QP4 | 4 (NormalHi) | 100% | 1 | You must create this QoS profile before using it. |
| | QP5 | 5 (Medium) | 100% | 1 | You must create this QoS profile before using it. |
| | QP6 | 6 (MediumHi) | 100% | 1 | You must create this QoS profile before using it. |
| | QP7 | 7 (High) | 100% | 1 | You must create this QoS profile before using it. You cannot create this QoS profile on SummitStack. |
| 7 | QP8 | 8 (HighHi) | 100% | 1 | This QoS profile is part of the default configuration and cannot be deleted. |

For CoS 802.1p traffic groups, the ingress 802.1p priority value selects a specific QoS profile as shown in the table above. This mapping can be changed as described in Changing the 802.1p Priority to QoS Profile Mapping on page 750. For traffic groups other than 802.1p-based groups, the traffic group configuration selects a specific egress QoS profile by name.

The default dual-rate QoS configuration is 0% for minimum bandwidth and 100% for maximum bandwidth.

The QoS profile for each port receives a default buffer reservation. All unreserved buffer space is part of a buffer pool, which can be used by QoS profiles when reserved space runs out, provided that the configuration for that QoS profile and port allows it.

You can increase the size of the shared buffer pool by reducing the global buffer reservation for a QoS profile on all switch ports. You can restrict buffer usage for a QoS profile in amounts ranging from 1 to 100%, in whole integers.

You can also override the global buffer reservation to increase or decrease the buffer space allotment for a specific QoS profile on one or more ports. Using the buffer override feature, you can override the global setting to use from 1–10,000% of the configured global allotment. The system does not drop any packets as long as reserved packet buffer memory for the port and QoS profile or shared packet memory for the port (`configure port` *`port_list`* **`shared-packet-buffer`** command) remains available.

Note

In a SummitStack, the scheduling algorithm is automatically programmed by ExtremeXOS for the stacking links only, and might be different from the algorithm you select.

Use of all eight queues on all ports can result in insufficient buffering to sustain zero packet loss throughput during full-mesh connectivity with large packets.

12 The QoS Profile Name cannot be changed.
13 The Queue Service Priority value cannot be changed.

When multiple QoS profiles are contending for port bandwidth and the egress traffic in each profile is within profile, the scheduler determines how the QoS profiles are serviced as described in Scheduling on page 734. In strict-priority mode, the queues are serviced based on the queue service priority value. In weighted fair-queuing mode, the queues are serviced based on the configured weight.

When configured to do so, the priority of a QoS profile can determine the 802.1p bits used in the priority field of a forwarded frame (see Introduction to Rate Limiting, Rate Shaping, and Scheduling on page 732). The priority of a QoS profile can determine the DiffServ code point value used in an IP packet when the packet is forwarded (see Replacement of DSCP on Egress on page 745).

A QoS profile change does not alter the behavior of the switch until it is assigned to a traffic group.

*Egress QoS Profiles on BlackDiamond X8 Series Switches, BlackDiamond 8900 xm-Series Modules, and Summit X670 Series Switches*

The egress QoS profiles on BlackDiamond X8 series switches, BlackDiamond 8900 xm-series modules, and Summit X670 series switches operate very similar to those for other BlackDiamond 8000 series modules and Summit family switches. This section describes the behaviors that are unique to the BlackDiamond 8900 xm-series modules, and the Summit X670 series switches.

The unicast and multicast queues in hardware on BlackDiamond 8900 xm-series modules, and Summit X670 series switches, are organized differently from other BlackDiamond 8000 series modules and Summit family switches.

For optimum use of the QoS profiles on these platforms, we recommend the following:

- Be aware that hardware on these platforms may occasionally reorder packets within a traffic flow.
- Multicast, broadcast, and flooded traffic flows for QP5-8 share a single multicast queue and are prioritized equally with the other traffic flows from QP1-8. We recommend that you do not direct multicast flows to QP5-8.

## Multicast Traffic Queues

On BlackDiamond X8, BlackDiamond 8900 xm-series modules, and Summit X670 series switches, multicast, broadcast, and flooded traffic flows for QP5-8 share a single multicast queue and are prioritized equally with the other traffic flows from QP1-8. We recommend that you do not direct multicast flows to QP5-8.

## Egress Port Rate Limiting and Rate Shaping

Egress port rate limiting and rate shaping allow you to define limits for all egress traffic coming from the egress *QoS* profiles and traffic queues.

On the BlackDiamond X8 and 8800 series, SummitStack, and Summit family switches, you can apply single-rate rate-limiting and control shared packet buffer space (`configure port port_list` **`shared-packet-buffer`** command) for each port.

You can also configure ports to pass an unlimited flow as described in Disabling Rate Limiting and Rate Shaping on page 734.

# Class of Service (CoS)

*CoS* provides a means to configure 802.1p-based traffic grouping, *QoS* rate-shaping and scheduling, ingress and egress metering and flood rate-limiting via *SNMP (Simple Network Management Protocol)*. The CoS MIB is fully integrated with ONEPolicy to provide a solution to management entities such as ExtremeManagement for QoS on ExtremeXOS

## Class of Service (CoS) Settings

Each of the 256 *CoS* entries, indexed 0-255, are mapped to a 802.1p priority (0-7), ToS value (0-255), virtual txq-reference (0-15) and virtual irl-reference (0-31). A virtual reference is used to abstract the hardware capabilities from the CoS row. This allows for the management entity such as ExtremeManagement to create user CoS profiles based on the user and not multiple user profiles for each hardware platform. The virtual references are then mapped to actual hardware resources. On ExtremeXOS, the txq-reference mappings 0-7 are fixed to *QoS* profiles 1-8 and txq-reference 8-15 map to QoS profile 8. The irl-reference mappings are also fixed to the per-port ingress meters. This mapping is dependent on the number of per-port ingress meters supported on the platform.CoS entries 0-7 are equivalent to the dot1p traffic groupings for priorities 0-7.

## Port Groups

Port groups allow the user to define ports by similar functionality or role. The user can create port groups and use them to configure *QoS* profiles, ingress meters and flood rate-limiters for ports that share the same role. In addition, ExtremeManagement will create and manipulate port groups through the *CoS* MIB via *SNMP*. These groups will be used to configure the ExtremeManagement-specified CoS resources that are mapped to QoS profiles, ingress meters and flood rate-limiters. Since the ExtremeManagement CoS feature is port group- oriented, extremeXOS QoS entities (e.g. QoS profiles, ingress meter, and flood rate-limiting) that are configured on a per-port basis are not exposed through the CoS MIB via SNMP.

## CoS Port Resource

The TXQ, IRL and Flood Control port resource tables in *CoS* are configured using the *QoS* profile, ingress meter and rate-limit flood resource commands in the CLI. These port resources are mapped to CoS when the resources in the CLI are configured using a port group.

## CoS Global Enable Action

An *SNMP* set to enable *CoS* will automatically enable per-port dot1p examination and diffserv replacement controls as needed for the functionality of CoS. Diffserv replacement is enabled for a *QoS* profile on all ports if the corresponding tosValue for CoS index 0-7 is configured. Otherwise, diffserv replacement is disabled for the corresponding QoS profile on all ports. Dot1p replacement is disabled for all QoS profiles on all ports. In addition, all QoS profiles are created and the dot1p examination to QoS profile mapping are fixed to match the TXQ reference mapping for CoS indexes 0-7. The internal CoS state is used to warn the user in the CLI in the event that the controls setup by CoS are modified. A subsequent SNMP set to disable CoS disables the internal CoS state. This turns off the warning messages in the CLI; however, it does not restore the configuration to its original state before the global enable action occurred.

When the internal CoS state is enabled, the global enable action is performed on any new blade that is inserted into the system (stack or chassis).

## Meter and Flood Actions

ExtremeXOS 16.1 has been expanded to implement the syslog, trap, and disable-port out-actions for meters and flood rate-limiters. This requires additional counters and a polling task to periodically check the thresholds for the configured events.

## Meter and Flood Limitations

*Port-Meter Out-of-profile Limitations*

The *CoS* MIB supports an out-of-profile status and counter for each per-port meter on each port. On all 16.1 supported platforms, the hardware only supports a global counter for each per-port meter. The out-of-profile status is supported for each per-port meter on each port; however, the status is only valid while the meter is out-of-profile. When the per-port meter rate is no longer exceeded, the out-of-profile status is automatically cleared in the hardware. The ability to detect an out-of-profile per-port meter on a given port is dependent on the out-of-profile status of the per-port meter at the time of the software polling interval.

Due to this limitation, the out-of-profile actions can only accurately be performed if the exceeded per-port meter rate is sustained for the duration of the software polling interval. Any change in the global out-of-profile counter for the per-port meter will be evenly distributed between the ports that indicate an out-of-profile status at the time of the software polling interval. This should give a rough approximation of the out-of-profile count of the per-port meter on each port.

*Flood Out-of-profile Limitations*

The CoS MIB supports an out-of-profile status and counter on each port for each flood type (unknown unicast, multicast, and broadcast). On all ExtremeXOS 16.1 supported platforms, the hardware only supports a single counter and status per port for all flood types. Any out-of-profile actions configured for one flood type on a given port will be performed if any flood type is out-of-profile on that port. The out-of-profile counter will be the aggregate count of all flood types on a given port.

# Configuring QoS

## Configuring QoS on BlackDiamond X8 and 8800, SummitStack, and Summit Family Switches

The following figure shows the *QoS* configuration components for BlackDiamond X8 and 8800 series switches, SummitStack, and Summit family switches.

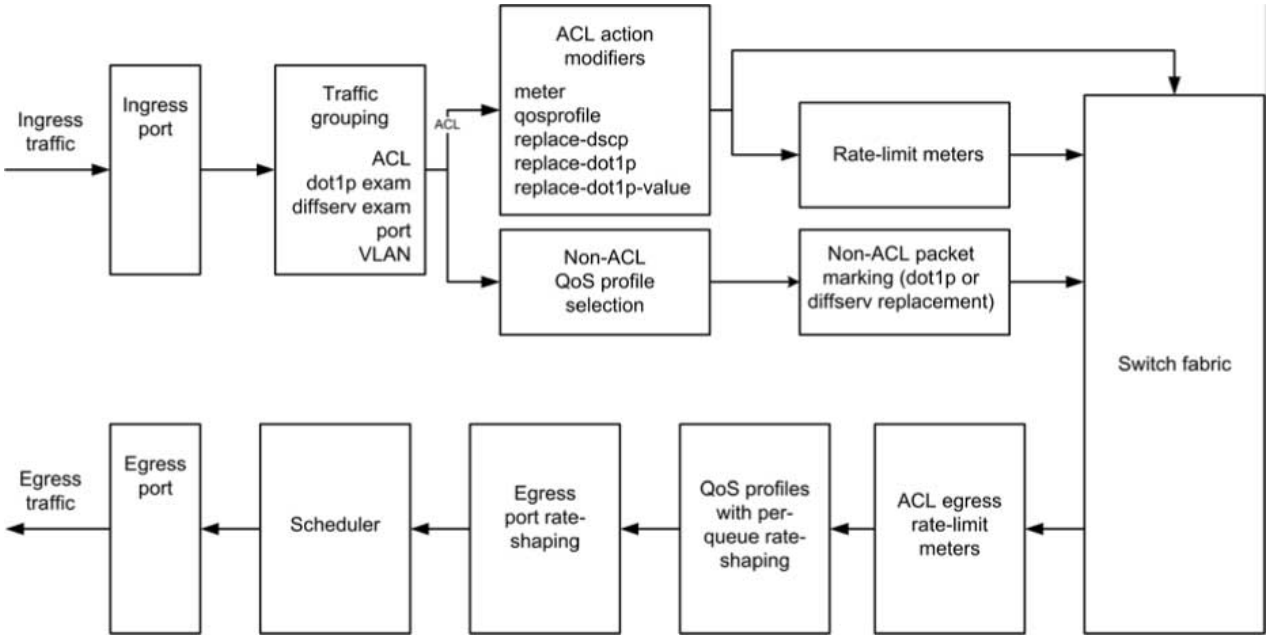



**Figure 100: QoS on BlackDiamond X8 and 8800 Series Switches, SummitStack, and Summit Family Switches**

### *QoS Configuration Guidelines for BlackDiamond X8 and 8800, SummitStack, and Summit Family Switches*

The considerations below apply only to QoS on the BlackDiamond X8 and 8800 series switches, SummitStack, and Summit family switches:

- The following QoS features share resources:
  - *ACL*s
  - dot1p
  - *VLAN*-based QoS
  - Port-based QoS
- You might receive an error message when configuring a QoS feature in the above list; it is possible that the shared resource is depleted. In this case, unconfigure one of the other QoS features and reconfigure the one you are working on.
- On a SummitStack, you cannot create QoS profile QP7. This QoS profile is reserved for system control traffic.
- These switches allow dynamic creation and deletion of QoS queues, with QP1 and QP8 always available.
- ACL egress rate-limit meters are supported only on BlackDiamond X8, BlackDiamond c-, xl-, and xm-series modules, and Summit x460, X480, X670, and X770 switches.

### *Configuration Summary for BlackDiamond X8 and 8800, SummitStack, and Summit Family Switches*

Use the following procedure to configure *QoS* on BlackDiamond X8 and 8800 series switches, SummitStack, and Summit family switches:

1. Configure basic Layer 2 connectivity (prerequisite).
2. Configure QoS scheduling, if needed, as described in Selecting the QoS Scheduling Method on page 743.

3. Configure ingress and egress rate-limiting as needed:
   a. Create a meter as described in Creating Meters on page 752.
   b. Configure the meter as described in Configuring a Meter on page 752.
   c. Apply the meter to ingress traffic as described in Applying a Meter to Ingress or Egress Traffic on page 753.
4. Configure non-*ACL*-based egress QoS profile selection as described in the following sections:
   - Configuring a CoS 802.1p-Based Traffic Group on page 749
   - Configuring a DiffServ-Based Traffic Group on page 750
   - Configuring a Port-Based Traffic Group on page 751
   - Configuring a VLAN-Based Traffic Group on page 752
5. Configure 802.1p or DiffServ packet marking as described in Configuring 802.1p or DSCP Replacement on page 744.
6. Configure egress QoS profile rate shaping as needed:
   a. Create egress QoS profiles as described in Creating or Deleting an Egress QoS Profile on page 747.
   b. Configure egress QoS profile rate shaping parameters as described in Configuring an Egress QoS Profile on page 747.
7. Configure egress port rate shaping as described in Configuring Egress Port Rate Limits on page 748.
8. Finalize ACL traffic-based group configuration as described in Configuring an ACL-Based Traffic Group on page 749.
9. Verify the configuration using the commands described in Displaying QoS Configuration and Performance on page 754.

## Selecting the QoS Scheduling Method

*QoS* scheduling determines the order of QoS profile service and varies between platforms. The BlackDiamond X8, BlackDiamond 8800 series switches, SummitStack, and Summit family switches support three scheduling methods: strict-priority, weighted-round-robin, and weighted-deficit-round-robin. These scheduling methods are described in Scheduling on page 734. Scheduling can be applied globally or on a per-port or per-PortGroup basis. If no port_list or port_group is specified in the command, the global scheduling algorithm is set and applied to ports that are not configured via a port_list or port_group. Specifying a port_list or port_group in the command configures the scheduling algorithm for specific ports.

QoS profiles can also be overriden on a global, per-port or per-port_group basis. If no port_list or port_group is specified, the global qosprofile configuration is changed and applied to all ports that are not configured on a per-port or per-port_group basis.

1. Select the QoS scheduling method for a switch.

   ```
   configure qosscheduler [strict-priority | weighted-round-robin |
   weighted-deficit-round-robin ] {ports [port_list | port_group | all ]}
   ```

   Note
   In a SummitStack, the scheduling algorithm is automatically programmed by ExtremeXOS for the stacking links only, and will likely be different than the algorithm you select.

2. Override the weighted-round-robin switch configuration on a specific QoS profile.

   `configure qosprofile qosprofile use-strict-priority`

## Configuring 802.1p or DSCP Replacement

*Replacement of 802.1p Priority Information on Egress*

By default, 802.1p priority information is not replaced or manipulated, and the information observed on ingress is preserved when forwarding the frame. This behavior is not affected by the switching or routing configuration of the switch. However, the switch is capable of inserting and/or overwriting 802.1p priority information when it transmits an 802.1Q tagged frame as described below.

**Replacement in ACL-Based Traffic Groups**

If you are using *ACL*-based traffic groups, you can use the `replace-dot1p` action modifier to replace the ingress 802.1p priority value with the 802.1p priority value of the egress *QoS* profile as listed in the following table. To specify a specific 802.1p priority value on egress, use the `replace-dot1p-value` action modifier.

Note

If you are using ACL-based traffic groups, you must use ACL action modifiers to replace the 802.1p priority. Traffic that meets any ACL match conditions is not subject to non-ACL-based 802.1p priority replacement.

**Replacement in Non-ACL-Based Traffic Groups**

For non-ACL-based traffic groups, you can enable or disable 802.1p priority replacement on specific ingress ports. When 802.1p priority replacement is enabled, the default egress 802.1p priority value is set to the priority value of the egress QoS profile as listed in the following table.

**Table 93: Default Queue-to-802.1p Priority Replacement Value**

| Egress QoS Profile | 802.1p Priority Replacement Value |
|---|---|
| QP1 | 0 |
| QP2 | 1 |
| QP3 | 2 |
| QP4 | 3 |
| QP5 | 4 |
| QP6 | 5 |
| QP7 | 6 |
| QP8 | 7 |

To enable 802.1p priority replacement on egress, use the following command:

[14] You cannot direct traffic to this QoS profile in a SummitStack.

```
enable dot1p replacement ports [port_list | all]
```

Note
The port in this command is the ingress port.

To disable this feature, use the following command:

```
disable dot1p replacement ports [port_list | all]
```

Note
If only DiffServ traffic groups are enabled, then 802.1p priority enforcement for 802.1q tagged packets continues for non-IP packets using the default 802.1p map shown in the following table.

On the BlackDiamond X8 series switches, BlackDiamond 8800 series switches, SummitStack, and Summit family switches, only QP1 and QP8 exist by default; you must create QP2 to QP7 (QP2 to QP5 in a SummitStack). If you have not created these QPs, the replacement feature will not take effect.

When DiffServ examination is enabled on 1 Gigabit Ethernet ports for BlackDiamond 8800 series switches, SummitStack, and Summit family switches, 802.1p replacement is enabled and cannot be disabled. The ingress 802.1p value is replaced with the 802.1p value assigned to the egress QoS profile.

### *Replacement of DSCP on Egress*

The switch can be configured to change the DSCP in a packet before forwarding the packet. This is done with no impact on switch performance and can be configured as described in the following sections:

#### Replacement in ACL-Based Traffic Groups

If you are using *ACL*-based traffic groups, you can use the replace-dscp action modifier to replace the ingress DSCP value with the DSCP value of the egress *QoS* profile as listed in the following table. This action modifier functions for both IPv4 and IPv6 traffic.

Note
If you are using ACL-based traffic groups, you must use ACL action modifiers to replace the DSCP. Traffic that meets any ACL match conditions is not subject to non-ACL-based DSCP priority replacement. For all platforms, we recommend that you use ACL-based traffic groups when configuring DSCP replacement.

#### Replacement in Non-ACL-Based Traffic Groups

For non-ACL-based traffic groups, you can enable or disable DSCP replacement on specific ingress ports. When DSCP replacement is enabled, the DSCP value used on egress is determined by either the

QoS profile or the 802.1p priority value. The following table shows the default mappings of QoS profiles and 802.1p priority values to DSCPs.

**Table 94: Default QoS Profile and 802.1p Priority Value Mapping to DiffServ Code Points**

| BlackDiamond X8 Series Switches, BlackDiamond 8800 Series Switches, SummitStack, and Summit Family Switches QoS Profile | 802.1p Priority Value | DSCP |
|---|---|---|
| QP1 | 0 | 0 |
| | 1 | 8 |
| | 2 | 16 |
| | 3 | 24 |
| | 4 | 32 |
| | 5 | 40 |
| | 6 | 48 |
| QP8 | 7 | 56 |

**Replacing a DSCP on Egress**

- To replace DSCPs by enabling DiffServ replacement, use the command:
  `enable diffserv replacement ports [`*`port_list`* `|` **`all`**`]`

  Note
  The port in this command is the ingress port.

- To disable this feature, use the command:
  `disable diffserv replacement ports [`*`port_list`* `|` **`all`**`]`
- To view the current DiffServ replacement configuration, use the command:
  `show diffserv replacement`
- To change the DSCP mapping, use the commands:
  `configure diffserv replacement [{`**`qosprofile`**`}` *`qosprofile`* `|` **`priority`** *`priority`*`]` **`code-point`** *`code_point`*

  `unconfigure diffserv replacement`

*DiffServ Example*

In this example, we use DiffServ to signal a class of service throughput and assign any traffic coming from network 10.1.2.x with a specific DSCP. This allows all other network switches to send and observe the DSCP instead of repeating the same *QoS* configuration on every network switch.

To configure the switch:

1. Using ACLs, assign a traffic grouping for traffic from network 10.1.2.x to QP3:

   ```
   configure access-list qp3sub any
   ```

   The following is a sample policy file example:

   ```
   #filename: qp3sub.pol
   ```

```
entry QP3-subnet {
    if {
        source-address 10.1.2.0/24
     } then {
        Qosprofile qp3;
        replace-dscp;
        }
```

2. Configure the switch so that other switches can signal calls of service that this switch should observe.

```
enable diffserv examination ports all
```

Note

The switch only observes the DSCPs if the traffic does not match the configured access list. Otherwise, the *ACL* QoS setting overrides the QoS DiffServ configuration.

## Configuring Egress QoS Profile Rate Shaping

### *Creating or Deleting an Egress QoS Profile*

The default configuration for most platforms defines eight egress *QoS* profiles. On BlackDiamond 8800 switches, SummitStack, and Summit family switches, the default configuration defines two egress QoS profiles.

- Use the following command to create an additional egress QoS profile:
  `create qosprofile [QP2| QP3 | QP4 | QP5 | QP6 | QP7]`
- Use the following command to delete an egress QoS profile:
  `delete qosprofile [QP2| QP3 | QP4 | QP5 | QP6 | QP7]`

### *Configuring an Egress QoS Profile*

- Egress *QoS* profile rate shaping is disabled by default on all ports. On all platforms, use the following commands to configure egress QoS profile rate shaping on one or more ports:

  `configure qosprofile egress qosprofile [{minbw minbw_number} {maxbw maxbw_number} | {peak_rate peak_bps [K | M]}] [ports [port_list | all]]`

  `configure qosprofile qosprofile [{minbw minbw_number} {maxbw maxbw_number} | {{committed_rate committed_bps [K | M]} {peak_rate peak_bps [K | M]} | [ports [port_list | all]]`

  `configure {qosprofile} qosprofile [[{maxbuffer buffer_percentage} {use-strict-priority}] | [maxbuffer buffer_percentage ports [port_list | all]]]`

  Note

  You must use these commands on all platforms if you want to configure the buffer size or weight value. Otherwise, you can use the command in the following description.

  You cannot configure the priority for the QoS profile on BlackDiamond X8, 8800 series switches, SummitStack, or Summit family switches.

- To remove the limit on egress bandwidth per QoS profile per port, re-issue this command using the default values.

- To display the current configuration for the QoS profile, use the following command.

  `show qosprofile [ all | port_list | port_group]`

*Configuring WRED on an Egress QoS Profile*

- To configure or unconfigure WRED on an egress QoS profile, use the following commands:

  `configure {qosprofile} {egress} qosprofile [wred [{color [tcp [green | red] | non-tcp [any|red]] [{min-threshold min_thresh} {max-threshold } {max-drop-rate max_drop_rate}]} | avg-weight avg_weight]] ports [port_list | all]`

  `unconfigure qosprofile wred {ports [port_list | all]}`
- To display the WRED configuration settings, use the following command:

  `show wredprofile {ports [port_list | all]}`

## Configuring Egress Port Rate Limits

*Configuring Egress Traffic*

- The default behavior is to have no limit on the egress traffic per port on BlackDiamond X8 Series Switches, BlackDiamond 8800 Series Switches, SummitStack, and Summit Family Switches. To configure egress rate limiting, use the following command:

  `configure ports [port_list | port_group]rate-limit egress [no-limit | cir-rate [Kbps | Mbps | Gbps] {max-burst-size burst-size [Kb | Mb]}]`
- To view the configured egress port rate-limiting behavior, use the following command.

  `show port {mgmt | port_list | tag tag} information {detail}`

  You must use the *detail* parameter to display the egress port rate configuration and, if configured, the maximum burst size. Refer to Displaying Port Information on page 303 for more information on the `show ports information` command.

  You can also display this information using the following command:

  `show configuration vlan`

  The following is sample output from the show configuration vlan command for configured egress rate limiting:

```
# Module vlan configuration.
#
configure vlan Default tag 1
config port 3:1 rate-limit egress 128 Kbps max-burst-size 200 Kb
config port 3:2 rate-limit egress 128 Kbps
config port 3:10 rate-limit egress 73 Kbps max-burst-size 128 Kb
configure vlan Default add ports 3:1-48 untagged
```

**Note**

Refer to FDB on page 561 for more information on limiting broadcast, multicast, or unknown MAC traffic ingressing the port.

## Configuring Traffic Groups

### *Configuring an ACL-Based Traffic Group*

*ACL*-based traffic groups are introduced in ACL-Based Traffic Groups on page 728. An ACL can implement multiple *QoS* features, so it is usually best to finalize the ACL after all other features have been configured.

To configure an ACL-based traffic group, do the following:

1. Create an ACL policy file and add rules to the file using the following guidelines:
   a. Use ACL match conditions to identify the traffic for the traffic group.
   b. Use ACL action modifiers to apply QoS features such as ingress meter or traffic queue selection, egress QoS profile or traffic queue selection, and 802.1p priority replacement to the traffic group.
2. Apply the ACL policy file to the ports where you want to define the traffic groups. You can apply the file to specific ports, all ports, or all ports in a *VLAN*.

**Note**
ACLs are described in detail in the ACLs chapter.

### *Configuring a CoS 802.1p-Based Traffic Group*

As described in CoS 802.1p-Based Traffic Groups on page 729, the default switch configuration defines CoS 802.1p-based traffic groups. The configuration options for these groups are described in the following sections:

- Enabling and Disabling 802.1p Examination on page 749
- Changing the 802.1p Priority to QoS Profile Mapping on page 750

**Note**
If you are using *ACL*-based traffic groups, use the `qosprofile` or `traffic-queue` action modifier to select a forwarding queue. Traffic that meets any ACL match conditions is not evaluated by other traffic groups.

#### Enabling and Disabling 802.1p Examination

*CoS* 802.1p examination is supported on all platforms and enabled by default. However, you can only disable and enable this feature on BlackDiamond 8800 series switches, SummitStack, and Summit

family switches. To free *ACL* resources, disable this feature whenever another *QoS* traffic grouping is configured. (See ACLs for information on available ACL resources.)

Note
If you disable this feature when no other QoS traffic grouping is in effect, 802.1p priority enforcement of 802.1q tagged packets continues. If only DiffServ traffic groups are enabled, then 802.1p priority enforcement for 802.1q tagged packets continues for non-IP packets using the default 802.1p map shown in the following table.

- To disable the 802.1p examination feature on BlackDiamond X8 and 8800 switches, SummitStack, and Summit family switches, use the following vommand.

  `disable dot1p examination ports [port_list | all]`

  Note
  802.1p examination cannot be disabled for 802.1p priority level 6 in a SummitStack. When 802.1p examination is disabled on a SummitStack, the precedence for the remaining examination becomes lower than that of all other traffic groupings.

- To re-enable the 802.1p examination feature on BlackDiamond 8800 switches, SummitStack, and Summit family switches, use the following command:

  `enable dot1p examination ports [port_list | all]`

- To display whether the 802.1p examination feature is enabled or disabled, use the following command:

  `show ports {mgmt | port_list | tag tag} information {detail}`

#### Changing the 802.1p Priority to QoS Profile Mapping

To view the current 802.1p priority to *QoS* profile mapping on a switch, enter the following command:

```
show dot1p
```

To change the mapping on BlackDiamond 8800 series switches, SummitStack, and Summit family switches, enter the following command:

`configure dot1p type dot1p_priority {qosprofile} qosprofile`

### *Configuring a DiffServ-Based Traffic Group*

As described in DiffServ-Based Traffic Groups on page 730, the default switch configuration defines DiffServ-based traffic groups. The configuration options for these groups are described in the following sections:

- Enabling and Disabling Diffserv Examination on page 751
- Changing the DSCP to QoS Profile Mapping on page 751

Note
If you are using *ACL*-based traffic groups, use the `qosprofile` or `traffic-queue` action modifier to select a forwarding queue. Traffic that meets any ACL match conditions is not evaluated by other traffic groups.

#### Enabling and Disabling Diffserv Examination

When a packet arrives at the switch on an ingress port and Diffserv examination is enabled, the switch uses the DSCP value to select the egress *QoS* profile that forwards the packet. The QoS profile configuration defines the forwarding characteristics for all traffic assigned to the QoS profile.

- The DiffServ examination feature is disabled by default. To enable DiffServ examination, use the following command:

  `enable diffserv examination ports [port_list | all]`

  Note

  When DiffServ examination is enabled on 1 Gigabit Ethernet ports for BlackDiamond 8800 series switches, SummitStack, and Summit family switches, 802.1p replacement is enabled and cannot be disabled. The ingress 802.1p value is replaced with the 802.1p value assigned to the egress QoS profile.

  Note

  When DiffServ examination is enabled on a Summit X670, or X770, the following warning message does not apply:

  ```
  "Warning: Enabling diffserv examination will cause dot1p replacement of
  802.1q tagged packets."
  ```

- To disable DiffServ examination, use the following command:

  `disable diffserv examination ports [port_list | all]`

#### Changing the DSCP to QoS Profile Mapping

- You can change the egress *QoS* profile assignment for each of the 64 code points. To view the current DSCP to QoS profile mapping, use the following command:

  `show diffserv examination`

- On BlackDiamond X8, 8800, SummitStack, and Summit family switches, use the following commands to change the DSCP to QoS profile mapping:

  `configure diffserv examination code-point code_point {qosprofile} qosprofile`

  `unconfigure diffserv examination`

  After a QoS profile is assigned, the rest of the switches in the network prioritize the packet using the characteristics specified by the QoS profile.

### *Configuring a Port-Based Traffic Group*

A port-based traffic group links a physical ingress port to an egress *QoS* profile for traffic forwarding.

To configure a port-based traffic group, use the following command:

```
configure ports port_list {qosprofile} qosprofile
```

Note

If you are using *ACL*-based traffic groups, use the `qosprofile` or `traffic-queue` action modifier to select a forwarding queue. Traffic that meets any ACL match conditions is not evaluated by other traffic groups.

On the BlackDiamond X8 switch, port-based traffic groups apply only to untagged packets. On the BlackDiamond 8800 series switches, SummitStack, and Summit family switches, port-based traffic groups apply to all packets.

### *Configuring a VLAN-Based Traffic Group*

A *VLAN*-based traffic group links all ports in a VLAN to an egress *QoS* profile for traffic forwarding. All intra-VLAN switched traffic and all routed traffic sourced from the named VLAN uses the specified QoS profile.

To configure a VLAN-based traffic group, use the following command:

```
configure vlan vlan_name {qosprofile} qosprofile
```

Note

If you are using *ACL*-based traffic groups, use the `qosprofile` or `traffic-queue` action modifier to select a forwarding queue. Traffic that meets any ACL match conditions is not evaluated by other traffic groups.

On the BlackDiamond X8 VLAN-based traffic groups apply only to untagged packets. On the BlackDiamond 8800 series switches, SummitStack, and Summit family switches, VLAN-based traffic groups apply to all packets.

## Creating and Managing Meters

You can configure meters to define bandwidth requirements on BlackDiamond 8800, SummitStack, and Summit family switches.

### *Creating Meters*

- To create a meter, use the following command:

  ```
  create meter meter-name
  ```

- To display the meters already configured on the switch, use the command:

  ```
  show meter meter_name
  ```

ExtremeXOS now creates 8 to 16 default ingress meters that are used for per-port metering. These are named "ingmeter<n>" where n is 0-15.

### *Configuring a Meter*

After you create a meter, you can configure the meter using the command for the platform you are using. To configure a *QoS* meter on all platforms, use the following command:

```
configure meter metername {committed-rate cir [Gbps | Mbps | Kbps |
Pps]} {max-burst-size burst-size [Kb | Mb | packets]} {out-actions
```

```
[{disable-port} {drop | set-drop-precedence {dscp [none | dscp-value]}}
{log} {trap}]} {ports [port_group | port_list]}
```

Note

The disable-port, log, and trap options are only allowed for per-port meters (i.e. ingmeter0, ingmeter1,...).

*Applying a Meter to Ingress or Egress Traffic*

You can apply a meter to ingress traffic using an *ACL* on BlackDiamond 8800 series switches, SummitStack, and Summit family switches. You can apply a meter to egress traffic using an ACL on BlackDiamond c-, xl-, and xm-series modules, Summit X480, X460, X670, and X770 series switches, and the E4G400 router.

Use rules within the ACL to identify the ingress traffic to which you want to apply the meter. Apply the meter by specifying the meter name with the `meter metername` action modifier. For information on completing the ACL configuration, see Configuring an ACL-Based Traffic Group on page 749.

*Deleting a Meter*

To delete a meter, use the following command:

```
delete meter meter-name
```

Note

The associated meters are not deleted when you delete any type of traffic queue. Those meters remain and can be associated with other traffic queues. To display the configured meters, use the `show meter meter_name` command.

## Adjusting the Byte Count Used to Calculate Traffic Rates

You can configure a per-packet byte adjustment that the system uses to calculate the ingress traffic rate, traffic utilization, and traffic statistics. You configure either the number of bytes you want subtracted from each packet ingressing the specified ports or the number of bytes you want added to the packet ingressing the specified ports.

By default, all bytes are counted for the ingressing traffic rate. After you issue this command, the default number of bytes removed is 0; you can add or subtract from one to four bytes from each ingressing packet on the specified ports for calculating the ingressing traffic rate.

- To display the number of bytes added to or subtracted from the packet to calculate the ingressing traffic rate, traffic utilization, and traffic statistics, enter the command:

  ```
  show ports port_list information detail
  ```

  Note

  You must use the **detail** keyword to display this information.

- To unconfigure this setting, re-issue the command and enter the value `0`.

## Controlling Flooding, Multicast, and Broadcast Traffic on Ingress Ports

On BlackDiamond 8800 and X8 series switches, SummitStack, and Summit family switches, you can control ingress flooding of broadcast and multicast traffic and traffic for unknown destination MAC addresses.

To control ingress flooding of broadcast and multicast traffic and traffic for unknown destination MAC addresses, enter the command:

```
configure ports port_list | port_grouprate-limit flood [broadcast |
multicast | unknown-destmac] [no-limit | pps]
```

# Displaying QoS Configuration and Performance

## Displaying Traffic Group Configuration Data

### *Displaying 802.1p Priority to QoS Profile Mappings*

To display the 802.1p priority to egress QoS profile mappings, enter the command:

```
show dot1p
```

### *Displaying DiffServ DSCP to QoS Profile Mappings*

To display the DiffServ DSCP to egress QoS profile mappings, enter the command:

```
show diffserv examination
```

### *Displaying Port and VLAN QoS Settings*

To display QoS settings assigned to ports or VLANs, enter the command:

```
show port {mgmt | port_list | tag tag} information {detail}
```

Note
To ensure that you display the QoS information, you must use the **detail** keyword.

On the BlackDiamond 8800 series, SummitStack, and Summit family switches, this command displays which QoS profile, if any, is configured.

## Displaying Performance Statistics

### *Displaying QoS Profile Traffic Statistics*

After you have created QoS policies that manage the traffic through the switch, you can use the QoS monitor to determine whether the application performance meets your expectations.

The QoS monitor allows you to display the traffic packet counts in a real-time or a snapshot display for the specified ports.

- View QoS profile traffic statistics on BlackDiamond X8 series switches, BlackDiamond 8000 series modules, SummitStack, and the Summit X440, X460, X480, X670, and X770 series switches by entering the command:.

  `show ports port_list qosmonitor {congestion} {no-refresh}`

  Note

  On a Summit X440 stack master slot, the QoS monitor displays the traffic packet count only for data traffic that is switched or routed. It does not capture the CPU/System-generated packet count.

  On BlackDiamond 8800 a-, c-, and e-series modules, only one port per slot or module can be monitored at any one time. This restriction does not apply to BlackDiamond 8900 series modules.

  On BlackDiamond X8 series switches, BlackDiamond 8900 xm-series modules and Summit X670 series switches, QP1-4 support one unicast and one multicast queue for each QoS profile. The QoS monitor counters for QP1-4 tally the unicast and multicast traffic for these QoS profiles. QoS monitor counters for QP5-8 tally only the unicast traffic for these QoS profiles.

- View or clear the WRED statistics on BlackDiamond X8 series switches, BlackDiamond 8900 c-, xl-, and xm-series modules, E4G-200 and E4G-400 cell site routers, and Summit X460, X480, X670, and X770 switches by entering the command:

  `show ports port_list wred {no-refresh}`

  `clear counters wred`

- View QoS profile traffic statistics on BlackDiamond 8800 by entering the command:

  `show ports port_list qosmonitor {ingress | egress} {bytes | packets} {no-refresh}`

*Displaying Congestion Statistics*

- To display a count of the packets dropped due to congestion on a port, enter the command:

  `show ports port_list congestion {no-refresh}`

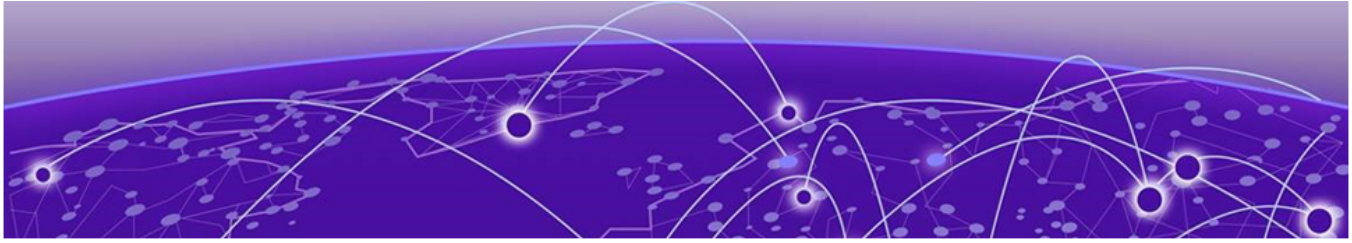
- To display a count of the packets dropped due to congestion in the *QoS* profiles for a port, enter the command:

  `show ports port_list qosmonitor {congestion} {no-refresh}`

  Note

  On BlackDiamond 8800 c-, and e-series modules, and Summit X440, only one port per slot or module can be monitored at any one time. This restriction does not apply to BlackDiamond 8900 series modules.

  On BlackDiamond 8900 xm-series modules and Summit X670 series switches, QP1-4 support one unicast and one multicast queue for each QoS profile. The congestion counters for QP1-4 tally the unicast and multicast traffic for these QoS profiles. Congestion counters for QP5-8 tally only the unicast traffic for these QoS profiles.



# Network Login



This chapter offers information about Network Login procedures. This section provides an overview, specific configuration information, and specific information about Local Database Authentication, 802.1X Authentication, Web-based Authentication, and MAC-based Authentication.

## Network Login Overview

Network login controls the admission of user packets into a network by allowing MAC addresses from users that are properly authenticated. Network login is controlled on a per-port basis.

When network login is enabled on a port, that port does not forward any packets until authentication takes place.

Unknown broadcast/unicast/multicast (BUM) traffic is not allowed on the egress side of network login-enabled ports until authentication is successful. To allow BUM traffic to egress on network login-enabled, unauthenticated ports, use `configure netlogin ports [port_list | all] allow egress-traffic [none | unicast | broadcast | all_cast]` .

Network login is capable of three types of authentication: web-based, MAC-based, and 802.1X. In addition, network login has two different modes of operation: campus mode and ISP mode. The authentication types and modes of operation can be used in any combination.

When web-based network login is enabled on a switch port, that port is placed into a non-forwarding state until authentication takes place. To authenticate, a user must open a web browser and provide the appropriate credentials. These credentials are either approved, in which case the port is placed in forwarding mode, or not approved, in which case the port remains blocked. You can initiate user logout by submitting a logout request or closing the logout window.

The following capabilities are included with network login:

- Web-based login using HTTP available on each port.
- Web-based login using HTTPS.

- Multiple supplicants for web-based, MAC-based, and 802.1X authentication on each port.

ExtremeXOS *NetLogin* provides the AAA (Authentication, Authorization, and Accounting) functionality, which is an important block of the network infrastructure and security and provides a model or framework to determine who is requesting network access, network resources that can be accessed by the requesting party, and when the resources are used. NetLogin supports all popular methods of authentication: MAC-based, Web-based, and IEEE 802.1X. NetLogin can help network administrators to control access into the network; it also provides flexibility to configure specific backend resources to which user access is allowed.

Together with IP Security, administrators can enhance security in the network by controlling access to upstream network and resources by the hosts or clients. IP Security is a collection of powerful features that allow network administrators to design security in combination with standard authentication and authorization techniques. IP Security features as ExtremeXOS 15.2 include:

- *DHCP (Dynamic Host Configuration Protocol)* Snooping and concept of trusted DHCP Servers
- Source IP Lockdown
- ARP Learning and Validation
- Gratuitous ARP Protection

When NetLogin and IP Security features are enabled on a port, NetLogin performs the first or the basic function of authenticating and authorizing the user. Further course of action is determined by IP

Security in case a violation is detected. The violation action will then determine further access to the network.

| Scenario | Notes | Expected Behavior |
|---|---|---|
| NetLogin + DHCP Snooping and trusted DHCP Servers/Ports.<br>Violation: DHCP Server Packets seen on netlogin enabled ports (i.e. a host is masquerading post authentication). | We recommend that you enable NetLogin on the client-facing ports. Enabling DHCP Snooping on all ports, including client/host facing ports and ports connected to the upstream network automation and provisioning infrastructure, ensures that all DHCP messages are processed by the switch and a DHCP binding database is maintained.<br>For the combination of NetLogin and DHCP Security to work correctly, we recommend that you configure at least one uplink (or server facing port) as a trusted port. This ensures that all other ports (normally client-facing ports) automatically become untrusted, and can be monitored for any violations that might occur. You can configure more than one uplink port as "trusted" as this allows flexibility in network design.<br>In addition to controlling which DHCP Servers can communicate with the downstream clients, the trusted DHCP Server configuration can be used. | Action: None<br>NetLogin authenticates the client, and IP Security flags a violation. No action is taken because of configuration.<br>Action: Drop-packet<br>The packet is dropped, and an *EMS (Event Management System)* event is logged.<br>Action: block-mac<br>NetLogin initially authenticates the client, and subsequently when the violation occurs, IP Security reports violation, which causes the MAC address to be blocked either permanently or for a configured duration on the switch. The *FDB (forwarding database)* will be flushed after FDB entry ages out and the netlogin entry is unauthenticated and removed from the switch.<br>Action: block-port<br>NetLogin initially authenticates the client, and subsequently when the violation occurs, IP Security reports the violation, which causes the port to be administratively disabled. As a result, all authenticated clients on the ports are immediately unauthenticated and removed from the switch. This can occur either for a certain configured duration or permanently.<br>We do not recommend that you use this configuration if there are multiple supplicants on the port (for example, conference rooms, groups of users, etc. accessing the network through an intermediate port extender, or hub).<br>In addition, for network troubleshooting and debugging purposes, an *SNMP (Simple Network Management Protocol)* trap can be sent to a central network manager such as ExtremeManagement or Ridgeline. |
| NetLogin + Source IP Lockdown<br>Violation: After a client/host successfully authenticates to the | The Source IP lockdown feature can determine if a client/host should be allowed access to the network based on inspection of the source IP address used in the packets. If a client is not | It is not mandatory to configure specific violation actions in DHCP Snooping (which is a prerequisite to this feature). |

| Scenario | Notes | Expected Behavior |
|---|---|---|
| network, it performs a source IP address violation. | using an IP address present in DHCP binding database, a violation is flagged for the client and further action is determined by configuration. This helps prevent clients from using source addresses not assigned by a centralized network automation and provisioning infrastructure.<br>By default the switch denies all IP traffic from clients when source IP lockdown is enabled. In order for the clients to get a valid IP address, DHCP packets are allowed to be forwarded through the switch. NetLogin authentication (all three forms) will still proceed, and clients presenting the valid credentials (per authentication scheme) are authenticated. Post authentication (and authorization - for membership to a *VLAN (Virtual LAN)*), once a valid DHCP binding is found for the client, access control lists are automatically (dynamically) applied to permit traffic from the client. | If configured, DHCP Snooping filters and violation actions take precedence over source IP lockdown. This is to ensure that successfully authenticated clients (with NetLogin) do not masquerade as rogue DHCP servers post authentication. In this case, DHCP violation is detected and actions are determined per configuration for vilation-action. |
| NetLogin + DHCP Secured ARP | DHCP Secured ARP allows administrators to control how the ARP table is populated. By default, the switch learns IP ARP bindings and builds the ARP table by tracking the ARP requests and replies. When DHCP Secured ARP is used for design, IP ARP learning method can be disabled. This is recommended for security purposes.<br>When combined with NetLogin, this feature ensures that a client (even after success authentication) cannot override the ARP entry on the switch, thereby preventing duplicate addresses and ensuring proper network operation.<br>ARP entries populated from DHCP are known as Secure ARP entries and are flushed/removed only when the address is released. | Same behavior as expected in the case of "NetLogin + DHCP Snooping and Trusted Ports/Server". Apart from securing the IP ARP table, in this case, violations detected as part of DHCP Snooping and validation is flagged and actions are determined by configuration. |

| Scenario | Notes | Expected Behavior |
|---|---|---|
| NetLogin + ARP Validation | ARP Validation helps check different fields in the ARP packet. Source MAC, Destination MAC, and Source IP addresses can be checked for validity. For a complete reference to the different checks performed, please refer to ARP Validation Options. All checks are performed post authentication. | Violation actions are same as the options presented in the DHCP Snooping and Validation Cases. Same expected behavior and functionality as in the case of "NetLogin + DHCP Snooping and Trusted Ports/Server". |
| NetLogin + Gratuitous ARP Protection | Gratuitous ARP is a method by which a client/host can resolve its own IP address, and is useful in scenarios where duplicate addresses need to be detected, or a host can announce that it has either used an IP address on a different NIC card, or even if a client has moved from one location to another. While ExtremeXOS supports Gratuitous ARP, protection can also be enabled to mitigate any risk or threats that can arise because of any m-i-m attacks. The switch will automatically send ARP packets to not only protect its own IP address but to also safeguard addresses of any NetLogin authenticated clients on the switch. For this, it is recommended that both DHCP Secured ARP and Gratuitous ARP protection be enabled on the switch. It is not mandatory that DHCP Snooping be enabled for this feature, but becomes a prerequisite if DHCP Secured ARP is also configured. | ARP Packets are sent out when a violation is detected. For all other violations detected by DHCP Snooping and Trust, the corresponding violation actions are determined by configuration, and expected behavior is the same as the case of "NetLogin + DHCP Snooping and Trusted Servers/Ports". |

## Web-Based, MAC-Based, and 802.1X Authentication

Authentication is handled as a web-based process, MAC-based process, or as described in the IEEE 802.1X specification.

Web-based network login does not require any specific client software and can work with any HTTP-compliant web browser. By contrast, 802.1X authentication may require additional software installed on the client workstation, making it less suitable for a user walk-up situation, such as a cybercafé or coffee shop. A workstation running Windows 7 or Windows 8 supports 802.1X natively, and does not require additional authentication software. Extreme Networks supports a smooth transition from web-based to 802.1X authentication.

Note

When both HTTP and HTTPS are enabled on the switch and sending HTTP requests from the Netlogin client, HTTPS takes preference and the switch responds with a HTTPS response.

MAC-based authentication is used for supplicants that do not support a network login mode, or supplicants that are not aware of the existence of such security measures, for example an IP phone.

If a MAC address is detected on a MAC-based enabled network login port, an authentication request is sent once to the AAA application. AAA tries to authenticate the MAC address against the configured *RADIUS (Remote Authentication Dial In User Service)* server and its configured parameters (timeout, retries, and so on) or the configured local database.

The credentials used for this are the supplicant's MAC address in ASCII representation and a locally configured password on the switch. If no password is configured, the MAC address is also used as the password. You can also group MAC addresses together using a mask.

*DHCP* is required for web-based network login because the underlying protocol used to carry authentication request-response is HTTP. The client requires an IP address to send and receive HTTP packets. Before the client is authenticated, however, the only connection that exists is to the authenticator. As a result, the authenticator must be furnished with a temporary DHCP server to distribute the IP address.

The switch responds to DHCP requests for unauthenticated clients when DHCP parameters such as `dhcp-address-range` and `dhcp-options` are configured on the network login *VLAN*. The switch can also answer DHCP requests following authentication if DHCP is enabled on the specified VLAN. If network login clients are required to obtain DHCP leases from an external DHCP server elsewhere on the network, DHCP should not be enabled on the VLAN.

The DHCP allocation for network login has a short time duration of 10 seconds and is intended to perform web-based network login only. As soon as the client is authenticated, it is deprived of this address. The client must obtain an operational address from another DHCP server in the network. DHCP is not required for 802.1X, because 802.1X uses only Layer 2 frames (EAPOL) or MAC-based network login.

URL redirection (applicable to web-based mode only) is a mechanism to redirect any HTTP request to the base URL of the authenticator when the port is in unauthenticated mode. In other words, when the user tries to log in to the network using the browser, the user is first redirected to the network login page. Only after a successful login is the user connected to the network. URL redirection requires that the switch is configured with a DNS client.

Web-based, MAC-based, and 802.1X authentication each have advantages and disadvantages, as summarized in Advantages of Web-Based Authentication.

*Advantages of Web-Based Authentication:*

- Works with any operating system that is capable of obtaining an IP address using DHCP. There is no need for special client side software; only a web browser is needed.

*Disadvantages of Web-Based Authentication:*

- The login process involves manipulation of IP addresses and must be done outside the scope of a normal computer login process. It is not tied to a Windows login. The client must bring up a login page and initiate a login.
- Supplicants cannot be re-authenticated transparently. They cannot be re-authenticated from the authenticator side.
- This method is not as effective in maintaining privacy protection.

*Advantages of MAC-Based Authentication:*

- Works with any operating system or network enabled device.
- Works silently; the user, client, or device does not know that it gets authenticated.
- Ease of management - set of devices can easily be grouped by the vendor part of the MAC address.

*Disadvantages of MAC-Based Authentication:*

- There is no re-authentication mechanism. The *FDB* aging timer determines the logout.
- Security is based on the MAC address of the client, so the network is more vulnerable to spoofing attacks.

*Advantages of 802.1X Authentication:*

- In cases where the 802.1X is natively supported, login and authentication happens transparently.
- Authentication happens at Layer 2. It does not involve getting a temporary IP address and subsequent release of the address to obtain a permanent IP address.
- Allows for periodic, transparent re-authentication of supplicants.

*Disadvantages of 802.1X Authentication:*

- 802.1X native support is available only on newer operating systems, such as Windows 7 or Windows 8.
- 802.1X requires an EAP-capable RADIUS Server. Most current RADIUS servers support EAP, so this is not a major disadvantage.
- Transport Layer Security (TLS) and Tunneled TLS (TTLS) authentication methods involve Public Key Infrastructure (PKI), which adds to the administrative requirements.

## Multiple Supplicant Support

An important enhancement over the IEEE 802.1X standard is that ExtremeXOS supports multiple supplicants (clients) to be individually authenticated on the same port.

This feature makes it possible for two or more client stations to be connected to the same port, with some being authenticated while others are not. A port's authentication state is the logical “OR” of the individual MAC's authentication states. In other words, a port is authenticated if any of its connected clients is authenticated. Multiple clients can be connected to a single port of authentication server through a hub or Layer 2 switch.

Multiple supplicants are supported in ISP mode for web-based, 802.1X, and MAC-based authentication. In addition, multiple supplicants are supported in Campus mode if you configure and enable network login MAC-based *VLAN*s. For more information, see Configuring Network Login MAC-Based VLANs on page 795.

The choice of web-based versus 802.1X authentication is again on a per-MAC basis. Among multiple clients on the same port, it is possible that some clients use web-based mode to authenticate, and some others use 802.1X, but the restriction is that they must be in the same untagged VLAN. This restriction is

not applicable if you configure network login MAC-based VLANs. For more information, see Configuring Network Login MAC-Based VLANs.

Note

With multiple supplicant support, after the first MAC is authenticated, the port is transitioned to the authenticated state and other unauthenticated MACs can listen to all data destined for the first MAC. Be aware of this as unauthenticated MACs can listen to all broadcast and multicast traffic directed to a network login-authenticated port.

## Network Login Multiple Authentication Support

The client or supplicant connected to the netlogin-enabled port(s) are authenticated by only one authentication protocol. If enabled globally and at the port, MAC-based authentication takes precedence if enabled globally and at the port. Dot1x takes precedence over MAC based authentication if Dot1x is supported by the supplicant. In this case the MAC-based authentication information is cleared as the client gets authenticated via Dot1x. Web based authentication happens only when the port belongs to the netlogin *VLAN*. The final authentication method used with its associated actions will be applied while any previous authenticated protocol information will be cleared.

This feature supports multiple authentication protocols on a netlogin-enabled port. The user must specify the authentication protocol priority or order per port which dictates the action for the client or supplicant that is getting authenticated on this port. Use the CLI to configure the authentication protocol order. By default the protocol precedence order for a netlogin enabled port is

1. Dot1x
2. Web-based
3. MAC

For example, if the following is the authentication protocol order configured on a netlogin enabled port in which all three authentication protocols are enabled:

1. Dot1x
2. MAC
3. Web-based

When user “john” tries to authenticate with his login credentials through Dot1x enabled supplicant or client, it sends the EAPOL packet to the ExtremeXOS switch or authenticator. Upon receipt of the EAPOL packet, the ExtremeXOS kernel *FDB* Module informs the user interface FDBMgr about the new MAC detection. The FDBMgr in turn informs the netlogin process about the new MAC or client. The netlogin process tries to authenticate the client/MAC through *RADIUS*. On receiving the authentication result from AAA process, the netlogin process checks for the protocol precedence configured by the user for that port and also finds if this client is being authenticated by any other authentication protocol. In this case no other authentication protocol has authenticated this MAC yet and the netlogin process will apply the action (VLAN movement, UPM security profile, etc.;) corresponding to MAC based authentication.

The ExtremeXOS switch or authenticator then sends the credentials of user “john” to the authentication server (RADIUS) a second time for Dot1x protocol authentication, Once the authentication result is received the netlogin process again checks the protocol precedence to find that the user “john”’s host/MAC is already authenticated via MAC based authentication. Since Dot1x is configured as the highest precedence protocol for this port the netlogin process will remove MAC based authentication

actions for this client and apply the Dot1x protocol action for “john” on this port. The MAC based authenticated client will continue to exist and will do the periodic reauthentication for the configured time. The “show netlogin” output will show the client’s highest precedence protocol or action applied authentication protocol details only.

When another user “sam” tries to authenticate from the same host or MAC through web based authentication method (provided the netlogin enabled port is still present in netlogin VLAN) the user “sam” will get authenticated but the web based authentication protocol action will not be applied since user “john” is already authenticated from this MAC with user configured highest precedence Dot1x protocol in this port.

Note

On changing the protocol precedence the action for the current highest precedence protocol (if client is authenticated by this protocol) will take effect immediately.

Note

On disabling the highest precedence protocol on this port the next precedence protocol (if client is authenticated by this protocol) action will take effect immediately.

*Support for Attaching and Detaching the UPM profile*

When the user or device gets authenticated the netlogin process will check for the protocol precedence configured by the user in this port and apply or remove the action accordingly. From the previous example when the user “john” tries to authenticate using Dot1x the ExtremeXOS switch or authenticator will authenticate the MAC using MAC based authentication and apply the action corresponding to the MAC based authentication protocol which includes applying UPM profile, creating *VLAN* (if netlogin dynamic VLAN is enabled on this port), VLAN movement etc.; when the user “john” then gets authenticated through Dot1x the EXOS switch or authenticator determines that Dot1x is the highest precedence protocol configured by the user in this port and removes the actions of MAC based authentication protocol and applies the Dot1x authentication protocol action that includes applying UPM profile, creating VLAN (if netlogin dynamic VLAN is enabled on this port), VLAN movement etc.; the MAC based authenticated client details still remains and continues to get reauthenticated for the configured time.

The netlogin process does the following when the user or MAC is being unauthenticated:

1. Sends accounting stop message to *RADIUS* through AAA.
2. Logs unauthentication *EMS* message for this client for this authentication protocol.
3. Sends “extremeNetloginUserLogout” *SNMP* trap message for this authentication protocol.
4. Informs IDM about unauthentication of this client for this authentication protocol.
5. Informs UPM about the client’s unauthentication for this protocol.

After performing the above said actions the netlogin process applies the highest precedence authentication protocol action configured for this port.

## Campus and ISP Modes

Network login supports two modes of operation, Campus and ISP.

Campus mode is intended for mobile users who tend to move from one port to another and connect at various locations in the network. ISP mode is meant for users who connect through the same port and *VLAN* each time (the switch functions as an ISP).

In Campus mode, the clients are placed into a permanent VLAN following authentication with access to network resources. For wired ports, the port is moved from the temporary to the permanent VLAN.

In ISP mode, the port and VLAN remain constant. Before the supplicant is authenticated, the port is in an unauthenticated state. After authentication, the port forwards packets.

You do not explicitly configure the mode of operation; rather, the presence of any Extreme Networks Vendor Specific Attribute (VSA) that has a VLAN name or VLAN ID (any VLAN attribute) in the *RADIUS* server determines the mode of operation. If a VLAN attribute is present, it is assumed to be Campus mode. If a VLAN attribute is not present, it is assumed to be ISP mode.

Note

When a client is authenticated in multiple VLANs using campus mode: 1) If any of the authenticated VLANs are deleted manually from a port or globally, the client is unauthenticated from all VLANs; and 2) If traffic is not seen on a particular VLAN then the *FDB* entry ages out and is deleted; the client itself remains authenticated and the FDB entry is reinstalled either when traffic is detected on that VLAN or when the client reauthenticates. For additional information on Campus and ISP mode operation on ports that support network login and *STP (Spanning Tree Protocol)*, see Exclusions and Limitations.

## Network Login and Hitless Failover

When you install two management modules (nodes) in a BlackDiamond chassis or when redundancy is available in a SummitStack, one node assumes the role of primary and the another node assumes the role of backup node.

Note

This section applies only to modular switches.

The primary node executes the switch's management functions, and the backup node acts in a standby role. Hitless failover transfers switch management control from the primary node to the backup node.

Note

Not all platforms support hitless failover in the same software release. To verify if the software version you are running supports hitless failover, see the following table in Managing the Switch. For more information about protocol, platform, and MSM support for hitless failover, see Understanding Hitless Failover Support.

Network login supports hitless failover by relaying current client authentication information from the master node to the backup node. For example, if a client moves to the authenticated state, or moves

from an authenticated state to an unauthenticated state, the primary node conveys this information to the backup node. If failover occurs, your authenticated client continues to operate as before the failover.

Note

If you use 802.1X network login, authenticated clients remain authenticated during failover; however, shortly after failover, all authenticated clients automatically re-authenticate themselves. Re-authentication occurs without user intervention.

If failover occurs during the authentication or re-authentication of a client, the client must repeat the authentication process.

Note

Before initiating failover, review the section Synchronizing Nodes—Modular Switches and SummitStack Only to confirm that your switch (or SummitStack) and both (or all) nodes are running software that supports the synchronize command.

*Initial Hitless Failover*

To initiate hitless failover on a network that uses network login:

1. Confirm that the nodes are synchronized and have identical software and switch configurations using the show switch {detail} command.

   If the primary and backup nodes, are not synchronized and both nodes are running a version of ExtremeXOS that supports synchronization, proceed to step Step 2.

   If the primary and backup nodes, are synchronized, proceed to step Step 3.

   The output displays the status of the primary and backup nodes, with the primary node showing MASTER and the backup node showing BACKUP (InSync).
2. If the primary and backup nodes, are not synchronized, use the synchronize command to replicate all saved images and configurations from the primary to the backup.

   After you confirm that the nodes are synchronized, proceed to step Step 3.
3. If the nodes are synchronized, use the show ports tdm information command to initiate failover.

For more detailed information about verifying the status of the nodes and system redundancy, see Understanding System Redundancy. For more information about hitless failover, see Understanding Hitless Failover Support.

## Configuring Network Login

This section provides a general overview of the commands used for:

- Enabling or Disabling Network Login on the Switch on page 767
- Enabling or Disabling Network Login on a Specific Port on page 767
- Configuring the Move Fail Action on page 767
- Displaying Network Login Settings on page 768
- Exclusions and Limitations on page 768

This section also describes information about the Exclusions and Limitations on page 768 of network login.

For more detailed information about a specific mode of network login, including configuration examples, refer to the following sections:

- 802.1X Authentication on page 773
- Web-Based Authentication on page 783
- MAC-Based Authentication on page 791

Note

When *STP* with Edge-safeguard and the Network login feature are enabled on same port, the port moves to a disabled state when it detects a loop in the network.

## Enabling or Disabling Network Login on the Switch

By default, network login is disabled.

To enable or disable network login and specify the authentication method, use the following commands:

- `netlogin vlan vlan_name`
- `enable netlogin [{dot1x} {mac} {web-based}]`
- `disable netlogin [{dot1x} {mac} {web-based}]`

## Enabling or Disabling Network Login on a Specific Port

Network login must be disabled on a port before you can delete a *VLAN* that contains that port.By default, all methods of network login are disabled on all ports.

- To enable network login on a port to specify the ports and the authentication method, use the following command:

  `enable netlogin ports ports [{dot1x} {mac} {web-based}]`

  Note

  When network login and *STP* are enabled on the same port, network login operates only when the STP port is in the forwarding state.

- To disable network login, use the following command:

  `disable netlogin ports ports [{dot1x} {mac} {web-based}]`

## Configuring the Move Fail Action

If network login fails to perform Campus mode login, you can configure the switch to authenticate the client in the original *VLAN* or deny authentication even if the user name and password are correct. For example, this may occur if a destination VLAN does not exist.

To configure the behavior of network login if a VLAN move fails, use the following command:

`configure netlogin move-fail-action [authenticate | deny]`

By default, the setting is deny.

The following describes the parameters of this command if two clients want to move to a different untagged VLAN on the same port:

- authenticate—Network login authenticates the first client that requests a move and moves that client to the requested VLAN. Network login authenticates the second client but does not move that client to the requested VLAN. The second client moves to the first client's authenticated VLAN.
- deny—Network login authenticates the first client that requests a move and moves that client. Network login does not authenticate the second client.

The dot1x client is not informed of the VLAN move-fail because it always receives EAP-Success or EAP-Fail directly based on the authentication result, not based on both authentication and the VLAN move result.

## Displaying Network Login Settings

To display the network login settings and parameters, use the following command:

```
show netlogin {port portlist vlan vlan_name} {dot1x {detail}}
```

## Exclusions and Limitations

The following are limitations and exclusions for network login:

- When using *NetLogin* MAC-based *VLAN* mode, moving a port as untagged from a pre-authentication VLAN to a post-authentication VLAN is not supported when both VLANs are configured with Protocol Filter IP.
- All unauthenticated MACs will be seeing broadcasts and multicasts sent to the port if even a single MAC is authenticated on that port.
- Network login must be disabled on a port before that port can be deleted from a VLAN.
- In Campus mode on all switches with untagged VLANs and the network login ports' mode configured as port-based-VLAN, after the port moves to the destination VLAN, the original VLAN for that port is not displayed.
- A network login VLAN port should not be a part of following protocols:
  - Ethernet Automatic Protection Switching (EAPS)
  - *ESRP (Extreme Standby Router Protocol)*
  - Link Aggregation
- Network login and ELRP are not supported on the same port.
- Network login and *STP* operate on the same port as follows:
  - At least one VLAN on the intended port should be configured both for network login and STP.
  - Network login and STP operate together only in network login ISP mode.
  - When STP blocks a port, network login does not process authentication requests and BPDUs are the only traffic in and out of the port. All user data forwarding stops.
  - When STP places a port in forwarding state, network login operates and BPDUs and user data flow in and out of the port. The forwarding state is the only STP state that allows network login and user data forwarding.

- If a network login client is authenticated in ISP mode and STP blocks one of the authenticated VLANS on a given port, the client is unauthenticated only from the port or VLAN which is blocked.
- All clients that are going through authentication and are learned on a blocked port or VLAN are cleared.

Note
When STP with edge-safeguard and network login feature is enabled on the same port, the port goes into the disabled state after detecting a loop in the network.

## Authenticating Users

Network login uses two types of databases to authenticate users trying to access the network:

- *RADIUS* servers
- Local database

All three network login protocols, web-based, MAC-based, and 802.1X, support RADIUS authentication. Only web-based and MAC-based network login support local database authentication.

Note
Beginning in ExtremeXOS 16.1, when RADIUS authentication is listed first and sends an access-reject it is honored and the local user database is not consulted - the bullet points listed just below indicate when the local database is used. .

The network login authenticated entry is cleared when there is an *FDB* timeout. This applies to web-based, MAC-Based, and 802.1X authentication.

## Local Database Authentication

You can configure the switch to use its local database for web-based and MAC-based network login authentication.

802.1X network login does not support local database authentication. Local authentication essentially mimics the functionality of the remote *RADIUS* server locally. This method of authentication is useful in the following situations:

- If both the primary and secondary (if configured) RADIUS servers timeout or are unable to respond to authentication requests.
- If no RADIUS servers are configured.
- If the RADIUS server used for network login authentication is disabled.

If any of the above conditions are met, the switch checks for a local user account and attempts to authenticate against that local account.

For local authentication to occur, you must configure the switch's local database with a user name and password for network login. We recommend a maximum of 64 local accounts. If you need more than 64 local accounts, we recommend using RADIUS for authentication. You can also specify the destination *VLAN* to enter upon a successful authentication.

You can also use local database authentication in conjunction with network login MAC-based VLANs. For more detailed information about network login MAC-based VLANs, see Configuring Network Login MAC-Based VLANs on page 795.

## Creating a Local Network Login Account--User Name and Password Only

We recommend creating a maximum of 64 local accounts. If you need more than 64 local accounts, we recommend using *RADIUS* for authentication. For information about RADIUS authentication, see Configuring the RADIUS Client on page 913.

User names are not case-sensitive. Passwords are case-sensitive. User names must have a minimum of one character and a maximum of 32 characters. Passwords must have a minimum of zero characters and a maximum of 32 characters. If you use RADIUS for authentication, we recommend that you use the same user name and password for both local authentication and RADIUS authentication.

- To create a local network login user name and password, use the following command and specify the parameter: *user-name*

  `create netlogin local-user user-name {encrypted} {password} {vlan-vsa [[{tagged | untagged} [vlan_name] | vlan_tag]]} {security-profile security_profile}`

  If you attempt to create a user name with more than 32 characters, the switch displays the following messages:

  ```
  %% Invalid name detected at '^' marker.
  %% Name cannot exceed 32 characters.
  ```

  If you attempt to create a password with more than 32 characters, the switch displays the following message after you re-enter the password:

  ```
  Password cannot exceed 32 characters
  ```

  The encrypted option is used by the switch to encrypt the password. Do not use this option through the command line interface (CLI).
- After you enter a local network login user name, press **[Return]**. The switch prompts you twice to enter the password.

The following example:

- Creates a new local network login user name.
- Creates a password associated with the local network login user name.

```
# create netlogin local-user megtest
password: <Enter the password. The switch does not display the password.>
Reenter password: <Re-enter the password. The switch does not display the password.>
```

For information about specifying the destination *VLAN*, see the next section Specifying a Destination VLAN on page 771.

Note

If you do not specify a password or the keyword **encrypted**, you are prompted for one.

## Specifying a Destination VLAN

If you configure a local network login account with a destination *VLAN*, upon successful authentication, the client transitions to the permanent, destination VLAN.

You can specify the destination VLAN when you initially create the local network login account or at a later time.

### *Adding VLANs when Creating a Local Network Login Account*

- To specify the destination *VLAN* when creating the local network login account, use the following command and specify the **vlan-vsa** option with the associated parameters:

  `create netlogin local-user user-name {encrypted} {password} {vlan-vsa [[{tagged | untagged} [vlan_name] | vlan_tag]]} {security-profile security_profile}`

  Where the following is true:

  - **tagged**—Specifies that the client be added as tagged.
  - **untagged**—Specifies that the client be added as untagged.
  - **vlan_name**—Specifies the name of the destination VLAN.
  - **vlan_tag**—Specifies the VLAN ID, tag, of the destination VLAN.

The following example:

- Creates a new local network login user name.
- Creates a password associated with the local network login user name.
- Adds the VLAN test1 as the destination VLAN.

The following is a sample display from this command:

```
create netlogin local-user megtest vlan-vsa "test1"
password: <Enter the password. The switch does not display the password.>
Reenter password: <Re-enter the password. The switch does not display the password.>
```

### *Adding VLANs at a Later Time*

To specify the destination *VLAN* after you created the local network login account, use the following command:

`configure netlogin local-user user-name {vlan-vsa [[{tagged | untagged} [vlan_name | vlan_tag]] | none]}`

Where the following is true:

- **tagged**—Specifies that the client be added as tagged
- **untagged**—Specifies that the client be added as untagged
- **vlan_name**—Specifies the name of the destination VLAN
- **vlan_tag**—Specifies the VLAN ID, tag, of the destination VLAN
- **none**—Specifies that the VSA 211 wildcard (*) is applied, only if you do not specify tagged or untagged

The following example:

- Modifies a previously created local network login account.
- Specifies that clients are added as tagged to the VLAN.
- Adds the VLAN blue as the destination VLAN.

```
configure netlogin local-user megtest vlan-vsa tagged "blue"
```

## Modifying an Existing Local Network Login Account

After you create a local network login user name and password, you can update the following attributes of that account:

- Password of the local network login account.
- Destination *VLAN* attributes including: adding clients tagged or untagged, the name of the VLAN, and the VLAN ID.

If you try modifying a local network login account that is not present on the switch, or you incorrectly enter the name of the account, output similar to the following appears:

```
* Switch # configure netlogin local-user purplenet
^
%% Invalid input detected at '^' marker.
```

To confirm the names of the local network login accounts on your switch, use the following command:

`show netlogin local-users`

### *Updating the Local Network Login Password*

Passwords are case-sensitive. Passwords must have a minimum of zero characters and a maximum of 32 characters.

1. Update the password of an existing local network login account with the following command:
   `configure netlogin local-user` *`user_name`*

   where *`user_name`* specifies the name of the existing local network login account.
2. After you enter the local network login user name, press [Enter].
   The switch prompts you to enter a password.
3. At the prompt enter the new password and press [Enter].
   The switch then prompts you to reenter the password.

   If you attempt to create a password with more than 32 characters, the switch displays the following message after you re-enter the password:

   ```
   Password cannot exceed 32 characters
   ```

   The following example modifies the password for the existing local network login account megtest.The following is a sample display from this command:

   ```
   configure netlogin local-user megtest
   password: <Enter the new password. The switch does not display the password.>
   Reenter password: <Re-enter the new password. The switch does not display the
   password.>
   ```

After you complete these steps, the password has been updated.

*Updating VLAN Attributes*

You can add a destination *VLAN*, change the destination VLAN, or remove the destination VLAN from an existing local network login account.

To make any of these VLAN updates, use the following command:

```
configure netlogin local-user user-name {vlan-vsa [[{tagged | untagged}
[vlan_name | vlan_tag]] | none]}
```

Where the following are true:

- *user_name*—Specifies the name of the existing local network login account
- tagged—Specifies that the client be added as tagged
- untagged—Specifies that the client be added as untagged
- *vlan_name*—Specifies the name of the destination VLAN
- *vlan_tag*—Specifies the VLAN ID, tag, of the destination VLAN
- none—Specifies that the VSA 211 wildcard (*) is applied, only if you do not specify tagged or untagged

## Displaying Local Network Login Accounts

To display a list of local network login accounts on the switch, including *VLAN* information, use the following command:

```
show netlogin local-users
```

## Deleting a Local Network Login Account

To delete a local network login user name and password, use the following command:

```
delete netlogin local-user user-name
```

# 802.1X Authentication

802.1X authentication methods govern interactions between the supplicant (client) and the authentication server.

The most commonly used methods are Transport Layer Security (TLS); Tunneled TLS (TTLS), which is a Funk/Certicom standards proposal; and PEAP.

TLS is the most secure of the currently available protocols, although TTLS is advertised to be as strong as TLS. Both TLS and TTLS are certificate-based and require a Public Key Infrastructure (PKI) that can issue, renew, and revoke certificates. TTLS is easier to deploy, as it requires only server certificates, by contrast with TLS, which requires client and server certificates. With TTLS, the client can use the RSA Data Security, Inc. *MD5 (Message-Digest algorithm 5)* Message-Digest Algorithm mode of user name/password authentication.

If you plan to use 802.1X authentication, refer to the documentation for your particular *RADIUS* server and 802.1X client on how to set up a PKI configuration.

## Interoperability Requirements

For network login to operate, the user (supplicant) software and the authentication server must support common authentication methods.Not all combinations provide the appropriate functionality.

### *Supplicant Side*

The supported 802.1X supplicants (clients) are Windows 7 and Windows 8 native clients, and Meetinghouse AEGIS.

A Windows 7 or Windows 8 802.1X supplicant can be authenticated as a computer or as a user. Computer authentication requires a certificate installed in the computer certificate store, and user authentication requires a certificate installed in the individual user's certificate store.

By default, the Windows 7 or Windows 8 machine performs computer authentication as soon as the computer is powered on, or at link-up when no user is logged into the machine. User authentication is performed at link-up when the user is logged in.

Windows 7 or Windows 8 also supports guest authentication, but this is disabled by default. Refer to relevant Microsoft documentation for further information. You can configure a Windows 7 or Windows 8 machine to perform computer authentication at link-up even if a user is logged in.

### *Authentication Server Side*

The *RADIUS* server used for authentication must be EAP-capable. Consider the following when choosing a RADIUS server:

- Types of authentication methods supported on RADIUS, as mentioned previously.
- Need to support VSAs. Parameters such as *`Extreme-Netlogin-Vlan-Name`* (destination vlan for port movement after authentication) and *`Extreme-NetLogin-Only`* (authorization for network login only) are brought back as VSAs.
- Need to support both EAP and traditional user name-password authentication. These are used by network login and switch console login respectively.

**Note**

For information on how to use and configure your RADIUS server, refer to Configuring the RADIUS Client on page 913 and to the documentation that came with your RADIUS server.

## Enabling and Disabling 802.1X Network Login

Network Login must be disabled on a port before you can delete a *VLAN* that contains that port. You can set a reauthentication maximum counter value to indicate the number of reauthentication trials after which the supplicant is denied access or given limited access.

- To enable 802.1X network login on the switch, use the following command:

  ```
  enable netlogin dot1x
  ```

Any combination of types of authentication can be enabled on the same switch. At least one of the authentication types must be specified on the CLI.

- To disable 802.1X network login on the switch, use the following command:

  `disable netlogin dot1x`
- To enable 802.1X network login on one or more ports, use the following command:

  `enable netlogin ports portlist dot1x`
- To disable 802.1X network login on one or more ports, use the following command:

  `disable netlogin ports portlist dot1x`
- To configure the reauthentication counter values, use the following command:

  `configure netlogin dot1x timers`
- To unconfigure the reauthentication counter values, use the following command:

  `unconfigure netlogin dot1x guest-vlan`

## 802.1X Network Login Configuration Example

The following configuration example shows the Extreme Networks switch configuration needed to support the 802.1X network login example.

**Note**

In the following sample configuration, any lines marked (Default) represent default settings and do not need to be explicitly configured.

```
create vlan “temp”
create vlan “corp”
configure vlan “default” delete ports 4:1-4:4
# Configuration Information for VLAN corp
# No VLAN-ID is associated with VLAN corp.
configure vlan “corp” protocol “ANY” (Default)
configure vlan “corp” ipaddress 10.203.0.224 255.255.255.0
# Configuration Information for VLAN Mgmt
configure vlan “Mgmt” ipaddress 10.10.20.30 255.255.255.0
# Network Login Configuration
configure netlogin vlan “temp”
enable netlogin dot1x
enable netlogin ports 1:10-1:14, 4:1-4:4 dot1x
# RADIUS Configuration
configure radius netlogin primary server 10.0.1.2 1812 client-ip 10.10.20.30 vr “VR-Mgmt”
configure radius netlogin primary shared-secret purple
enable radius
```

The following example is for the FreeRADIUS server; the configuration might be different for your *RADIUS* server:

```
#RADIUS Server Setting, in this example the user name is eaptest
eaptest Auth-Type := EAP, User-Password == "eaptest"
Session-Timeout = 120,
Termination-Action =1
```

For information about how to use and configure your RADIUS server, refer to Configuring the RADIUS Client on page 913 and the documentation that came with your RADIUS server.

## Configuring Guest VLANs

Ordinarily, a client that does not respond to 802.1X authentication remains disabled and cannot access the network.

802.1X authentication supports the concept of "guest *VLAN*s" that allow such a supplicant (client) limited or restricted network access. If a supplicant connected to a port does not respond to the 802.1X authentication requests from the switch, the port moves to the configured guest VLAN. A port always moves untagged into the guest VLAN.

Note

The supplicant does not move to a guest VLAN if it fails authentication after an 802.1X exchange; the supplicant moves to the guest VLAN only if it does not respond to an 802.1X authentication request.

When the authentication server sends an 802.1X request to the supplicant, there is a specified time interval for the supplicant to respond. By default, the switch uses the supplicant response timer to authenticate the supplicant every 30 seconds for a maximum of three tries. If the supplicant does not respond within the specified time, the authentication server sends another request. After the third 802.1X request without a supplicant response, the port is placed in the guest VLAN, if the guest VLAN feature has been configured for the port. The number of authentication attempts is not a user-configured parameter.

If a supplicant on a port in the guest VLAN becomes 802.1X-capable, the switch starts processing the 802.1X responses from the supplicant. If the supplicant is successfully authenticated, the port moves from the guest VLAN to the destination VLAN specified by the *RADIUS* server. If the RADIUS server does not specify a destination VLAN, the port moves to the VLAN it belonged to before it was placed in the guest VLAN. After a port has been authenticated and moved to a destination VLAN, it is periodically re-authenticated. If the port fails authentication, it moves to the VLAN it belonged to originally.

Note

A guest VLAN is not a normal network login VLAN. A guest VLAN performs authentication only if authentication is initiated by the supplicant.

### *Using Guest VLANs*

Suppose you have a meeting that includes company employees and visitors from outside the company.

In this scenario, your employees have 802.1X enabled supplicants but your visitors do not. By configuring a guest *VLAN*, when your employees log into the network, they are granted network access (based on their user credentials and 802.1X enabled supplicants). However, when the visitors attempt to log into the network, they are granted limited network access because they do not have 802.1X enabled supplicant. The visitors might be able to reach the Internet, but they are unable to access the corporate network.

For example, in the following figure Host A has 802.1x capability and Host B does not. When Host A is authenticated, it is given full access to the network. Host B does not have 802.1X capability and therefore does not respond to 802.1X requests from the switch. If port B is configured with the guest VLAN, port B is moved to the guest VLAN. Then Host B will be able to access the Internet but not the corporate network. After Host B is equipped with 802.1X capability, it can be authenticated and allowed to be part of the corporate network.

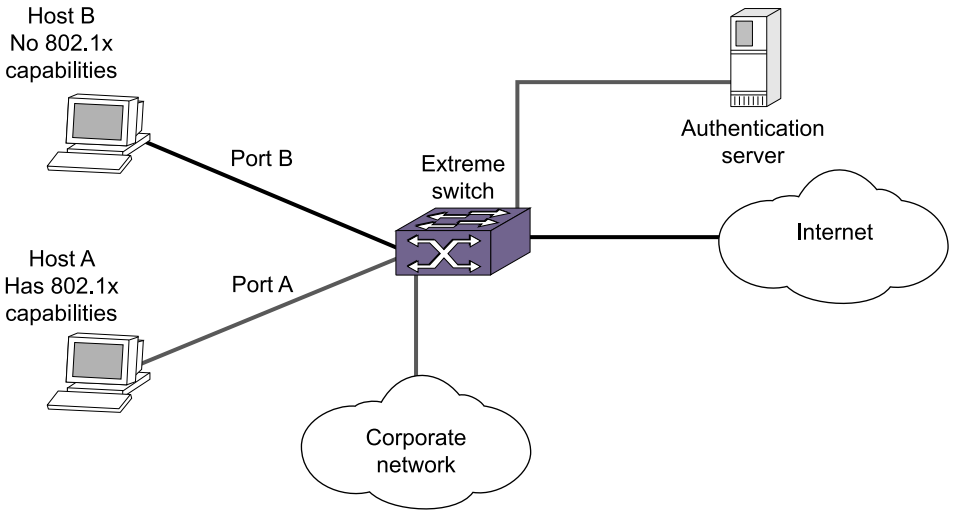



**Figure 101: Guest VLAN for Network Login**

### *Guidelines for Configuring Guest VLANs*

Keep in mind the following guidelines when configuring guest *VLAN*s:

- You must create a VLAN and configure it as a guest VLAN before enabling the guest VLAN feature.
- Configure guest VLANs only on network login ports with 802.1x enabled .
- Movement to guest VLANs is not supported on network login ports with MAC-based or web-based authentication.
- 802.1x must be the only authentication method enabled on the port for movement to guest VLAN.
- No supplicant on the port has 802.1x capability.

### *Creating Guest VLANs*

If you configure a guest *VLAN*, and a supplicant has 802.1X disabled and does not respond to 802.1X authentication requests from the switch, the supplicant moves to the guest VLAN. Upon entering the guest VLAN, the supplicant gains limited network access.

**Note**

You can configure guest VLANs on a per port basis, which allows you to configure more than one guest VLAN per VR. In ExtremeXOS 11.5 and earlier, you can only configure guest VLANs on a per VR basis, which allows you to configure only one guest VLAN per VR.

To create a guest VLAN, use the following command:

```
configure netlogin dot1x guest-vlan vlan_name {ports port_list}
```

### *Enabling Guest VLANs*

To enable the guest *VLAN*, use the following command:

```
enable netlogin dot1x guest-vlan ports [all | ports]
```

### *Modifying the Supplicant Response Timer*

The default supplicant response timeout is 30 seconds, and the range is 1-120 seconds. The number of authentication attempts is not a user-configured parameter.

To modify the supplicant response timer, use the following command and specify the *supp-resp-timeout* parameter:

```
configure netlogin dot1x timers [{server-timeout server_timeout} {quiet-period quiet_period} {reauth-period reauth_period {reauth-max max_num_reauths}} {supp-resp-timeout supp_resp_timeout}]
```

### *Disabling Guest VLANs*

To disable the guest *VLAN*, use the following command:

```
disable netlogin dot1x guest-vlan ports [all | ports]
```

### *Unconfiguring Guest VLANs*

To unconfigure the guest *VLAN*, use the following command:

```
unconfigure netlogin dot1x guest-vlan {ports port_list | vlan_name}
```

### *Display Guest VLAN Settings*

To display the guest *VLAN* settings, use the following command:

```
show netlogin guest-vlan {vlan_name}
```

If you specify the vlan_name, the switch displays information for only that guest VLAN.

The output displays the following information in a tabular format:

- Port—Specifies the 802.1X enabled port configured for the guest VLAN.
- Guest-vlan—Displays guest VLAN name and status: enable/disable.
- Vlan—Specifies the name of the guest VLAN.

## Post-authentication VLAN Movement

After the supplicant has been successfully authenticated and the port has been moved to a *VLAN*, the supplicant can move to a VLAN other than the one it was authenticated on.

This occurs when the switch receives an Access-Accept message from the *RADIUS* server with a VSA that defines a new VLAN. The supplicant remains authenticated during this transition. This occurs on both untagged and tagged VLANs. For example, suppose a supplicant submits the required credentials for network access; however, it is not running the current, approved anti-virus software or it does not have the appropriate software updates installed. If this occurs, the supplicant is authenticated but has limited network access until the problem is resolved. After you update the supplicant's anti-virus software, or install the software updates, the RADIUS server re-authenticates the supplicant by sending ACCESS-ACCEPT messages with the accompanying VLAN attributes, thereby allowing the supplicant to enter its permanent VLAN with full network access.

This is normal and expected behavior; no configuration is necessary.

## 802.1X Authentication and Network Access Protection

802.1X authentication in combination with Microsoft's Network Access Protection (NAP) provide additional integrity checks for end users and supplicants that attempt to access the network.

NAP allows network administrators to create system health policies to ensure supplicants that access or communicate with the network meet administrator-defined system health requirements. For example, if a supplicant has the appropriate software updates or anti-virus software installed, the supplicant is deemed healthy and granted network access. On the other hand, if a supplicant does not have the appropriate software updates or anti-virus software installed, the supplicant is deemed unhealthy and is placed in a quarantine *VLAN* until the appropriate update or anti-virus software is installed. After the supplicant is healthy, it is granted network access. For more information about NAP, please refer to the documentation that came with your Microsoft Windows or Microsoft Server software.

To configure your network for NAP, the minimum required components are:

- Extreme Networks switches running ExtremeXOS 11.6 or later.
- *RADIUS* server that supports NAP (Microsoft Windows Vista operating system refers to this as a network policy server (NPS), formerly known as the internet authentication server (IAS)).
- Remediation servers that receive unhealthy supplicants. The remediation servers contain the appropriate software updates, anti-virus software, and so on to make a supplicant healthy.

In addition to the required hardware and software, you must configure NAP-specific VSAs on your RADIUS server. By configuring these VSAs, you ensure supplicant authentication and authorization to the network and the switch creates dynamic Access Control Lists (ACLs) to move unhealthy supplicants to the quarantine VLAN for remediation. For more information see, Using NAP-Specific VSAs to Authenticate 802.1X Supplicants.

The following figure displays a sample network that uses NAP to protect the network.

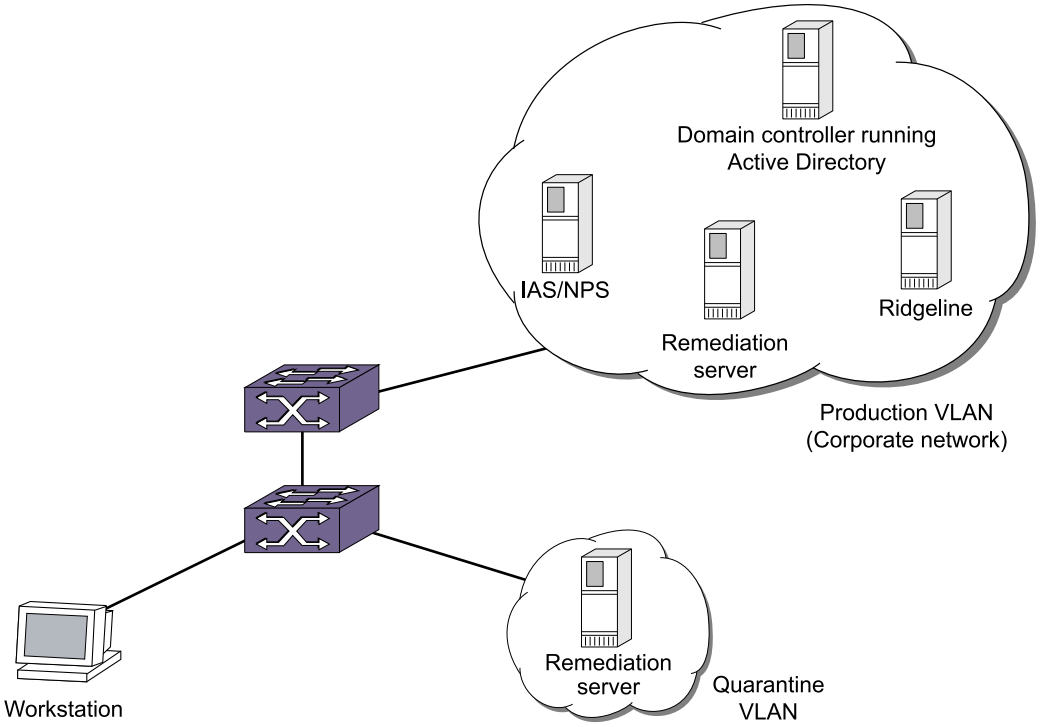



**Figure 102: Sample Network Using NAP to Provide Enhanced Security**

*Example Scenarios Using NAP*

Using the following figure, the following two scenarios describe some sample actions taken when an 802.1X-enabled supplicant initiates a connection to the network.

The scenarios assume the following:

- Scenario 1 has a healthy 802.1X-enabled supplicant.
- Scenario 2 has an unhealthy 802.1X-enabled supplicant.
- 802.1X network login has been configured and enabled on the switch.
- The *RADIUS* server has been configured using the NAP-specific VSAs for authenticating supplicants.
- The remediation servers have been configured with the appropriate software updates, anti-virus software, and so on.
- The ExtremeManagement and Ridgeline servers have been configured to receive traps from the switch. The traps sent from the switch inform Ridgeline of the state of the supplicant. In these scenarios, you configure the NMS as the syslog target.
- *VLAN*s Production and Quarantine have already been created and configured.

Note

You can dynamically create the quarantine VLAN if you configure dynamic VLAN creation on the switch. For more information see, Configuring Dynamic VLANs for Network Login on page 798.

**Scenario 1--Healthy Supplicant**

The steps to authenticate a healthy supplicant are:

1. The 802.1X supplicant initiates a connection to the 802.1X network access server (NAS), which in this scenario is the Extreme Networks switch.
2. The supplicant passes its authentication credentials to the switch using PEAP and an inner authentication method such as MS-CHAPv2.
3. The *RADIUS* server requests a statement of health (SoH) from the supplicant.

   Only NAP-capable supplicants create an SoH, which contains information about whether or not the supplicant is compliant with the system health requirements defined by the network administrator.
4. If the SoH indicates that the supplicant is healthy, the RADIUS server sends an Access-Accept message with a RADIUS VSA indicating which *VLAN* the healthy supplicant is moved to (in this example, the Production VLAN).
5. The switch authenticates the supplicant and moves it into the Production VLAN.
6. The switch sends a trap to the NMS indicating that the supplicant has been successfully authenticated and the VLAN into which it has been moved.

**Scenario 2--Unhealthy Supplicant**

The steps to authenticate an unhealthy supplicant are:

1. The 802.1X supplicant initiates a connection to the 802.1X network access server (NAS), which in this scenario is the Extreme Networks switch.
2. The supplicant passes its authentication credentials to the switch using PEAP and an inner authentication method such as MS-CHAPv2.
3. The *RADIUS* server requests a statement of health (SoH) from the supplicant.

   Only NAP-capable supplicants create an SoH, which contains information about whether or not the supplicant is compliant with the system health requirements defined by the network administrator.
4. If the SoH indicates that the supplicant is unhealthy, the RADIUS server sends an Access-Accept message with RADIUS VSAs indicating which:
   - *VLAN* the unhealthy supplicant is moved to (in this example, the Quarantine VLAN).
   - the remediation server(s) from which the supplicant can get software updates, anti-virus software and so on to remediate itself.
5. When the switch receives the VLAN and remediation server information from the RADIUS server, the switch:
   - Moves the supplicant into the Quarantine VLAN.
   - Applies ACLs to ensure the supplicant in the Quarantine VLAN can access only the remediation servers
   - Drops all other traffic not originating/destined from/to the remediation servers
   - sends a trap to Ridgeline indicating that the supplicant has been authenticated but has restricted access in the Quarantine VLAN for remediation.
6. The supplicant connects to the remediation server to get software updates, anti-virus software, and so on to get healthy.
7. After the supplicant is healthy, it restarts the authentication process and is moved to the Production VLAN, as a healthy supplicant with full network access.

### *Using NAP-Specific VSAs to Authenticate 802.1X Supplicants*

The following table contains the VSA definitions for 802.1X network login in conjunction with devices and servers that support NAP.

The Microsoft Vendor ID is 311.

**Note**

For more information about NAP and the VSAs supported by NAP, please refer to the documentation that came with your Microsoft operating system or server.

**Table 95: NAP-Specific VSA Definitions for 802.1X Network Login**

| VSA | Vendor Type | Type | Sent-in | Description |
|---|---|---|---|---|
| MS-Quarantine-State | 45 | Integer | Access-Accept | Indicates the network access level that the *RADIUS* server authorizes the user. The network access server (the switch) also enforces the network access level. A value of "0" gives the user full network access. A value of "1" gives the user limited network access. A value of "2" gives the user full network access within a specified time period. |
| MS-IPv4-Remediation-Servers | 52 | Integer | Access-Accept | Indicates the IP address(es) of the remediation server(s) that an unhealthy supplicant moves to in order to get healthy. |

### *ACLS for Remediation Servers*

The NAP VSA, MS-IPv4-Remediation-Servers, contains a list of IP addresses that an unhealthy and therefore quarantined supplicant should be allowed access to so that it can remediate itself and become healthy.

The way a quarantine is implemented on the switch is simply by moving the client/port to a user-designated 'quarantine' *VLAN* whose VLANID/Name is sent in the Access-Accept message. It is up to the user to ensure that the quarantine VLAN does indeed have limited access to the rest of the network. Typically, this can be done by disabling IP forwarding on that VLAN so no routed traffic can get out of that VLAN. Also, with dynamic VLAN creation, the quarantine VLAN being supplied by *RADIUS* could be dynamically created on the switch, once dynamic VLAN creation is enabled on it. The remediation server(s) would need to be accessible via the uplink port, regardless of whether the quarantine VLAN is pre-configured or dynamically created, since IP forwarding is not enabled on it.

To get around this restriction, network login has been enhanced so when a MS-Quarantine-State attribute is present in the Access-Accept message with extremeSessionStatus being either 'Quarantined' or 'On Probation,' then a 'deny all traffic' dynamic *ACL (Access Control List)* will be applied on the VLAN. If such an ACL is already present on that VLAN, then no new ACL will be applied.

When the last authenticated client has been removed from the quarantine VLAN, then the above ACL will be removed.

Additionally, if the MS-IPv4-Remediation-Servers VSA is present in the Access-Accept message, for each IP address present in the VSA a 'permit all traffic to/from this IP address' ACL will be applied on the quarantine VLAN. This will allow traffic to/from the remediation servers to pass unhindered in the Quarantine VLAN while all other traffic will be dropped.

## Web-Based Authentication

This section describes web-based network login.

For web-based authentication, you need to configure the switch DNS name, default redirect page, session refresh, and logout-privilege. URL redirection requires the switch to be assigned a DNS name. The default name is network-access.com. Any DNS query coming to the switch to resolve switch DNS name in unauthenticated mode is resolved by the DNS server on the switch in terms of the interface (to which the network login port is connected to).

Note

When both HTTP and HTTPS are enabled on the switch and sending HTTP requests from the Netlogin client, HTTPS takes preference and the switch responds with a HTTPS response.

### Enabling and Disabling Web-Based Network Login

- To enable web-based network login on the switch, use the following command:

  `enable netlogin web-based`

  Any combination of types of authentication can be enabled on the same switch. At least one of the authentication types must be specified on the CLI.
- To disable web-based network login on the switch, use the following command:

  `disable netlogin web-based`
- To enable web-based network login on one or more ports, use the following command:

  enable netlogin ports *`portlist`* web-based

  Network Login must be disabled on a port before you can delete a *VLAN* that contains that port.
- To disable web-based network login on one or more ports, use the following command:

  disable netlogin ports *`portlist`* web-based

### Configuring the Base URL

- To configure the network login base URL, use the following command:

  `configure netlogin base-url url`

  Where *`url`* is the DNS name of the switch.

  For example, configure netlogin base-url network-access.com makes the switch send DNS responses back to the network login clients when a DNS query is made for network-access.com.

### Configuring the Redirect Page

To configure the network login redirect page, use the following command:

```
configure netlogin redirect-page url
```

Where *url* defines the redirection information for the users after they have logged in. You must configure a complete URL starting with http:// or https://. By default, the redirect URL value is "http://www.extremenetworks.com" and default re-direction will take maximum of 20 seconds i.e default netlogin-lease-timer + 10 seconds. Re-direct time can be changed by tuning netlogin-lease-timer.

You can also configure the redirect value to a specific port number, such as 8080. For example, you can configure the network login redirect page to the URL value "http://www.extremenetworks.com:8080". The default port value is 80.

This redirection information is used only in case the redirection info is missing from *RADIUS* server. For example, configure netlogin base-url http://www.extremenetworks.com redirects all users to this URL after they get logged in.

For more information about SSH2, see Network Login. For information about installing the SSH module, see Software Upgrade and Boot Options.

## Configuring Proxy Ports

To configure the ports to be hijacked and redirected, use the following command:

```
configure netlogin add proxy-port tcp_port {http | https}
```

```
configure netlogin delete proxy-port
```

For each hijacked or proxy port, you must specify whether the port is to be used for HTTP or HTTPS traffic. No more than five hijack or proxy ports are supported for HTTP in addition to port 80 (for HTTP) and port 443 (for HTTPS), both of which cannot be deleted.

## Configuring Session Refresh

To enable or disable the network login session refresh, use the following commands:

```
enable netlogin session-refresh {refresh_minutes}
```

```
disable netlogin session-refresh
```

Where *minutes* ranges from 1 - 255. The default setting is 3 minutes. The commands enable netlogin session-refresh and configure netlogin session-refresh make the logout window refresh itself at every configured time interval. Session refresh is enabled by default. When you configure the network login session refresh for the logout window, ensure that the *FDB* aging timer is greater than the network login session refresh timer.

Note

If an attempt is made to authenticate the client in a non-existent *VLAN*, and the move fail action setting is authenticate, then the client is successfully authenticated in the port's original VLAN, but subsequent session refreshes fail and cause the client to become unauthenticated.

When web-based Network login is configured with proxy ports and session-refresh are also enabled, you must configure the web browser to bypass the web proxy server for the IP address of the VLAN into which the client moves after authentication.

## Configuring Logout Privilege

- To enable or disable network login logout privilege, use the following commands:

  ```
  enable netlogin logout-privilege
  ```

  ```
  disable netlogin logout-privilege
  ```

  These commands turn the privilege for network login users to logout by popping up (or not popping up) the logout window. Logout-privilege is enabled by default.
- You can configure the number of times a refresh failure is ignored before it results in the client being unauthenticated by using the following commands:

  ```
  configure netlogin allowed-refresh-failures
  ```

  ```
  unconfigure netlogin allowed-refresh-failures
  ```

  You can set the number of failures to be from between 0 and 5. The default number of logout failures is 0.

## Configuring the Login Page

You can fully customize the HTML login page and also add custom embedded graphical images to it. This page and the associated graphics must be uploaded to the switch so that they can be served up as the initial login page at the base URL. Both HTTP and HTTPS are supported as a means of authenticating the user via the custom page.

In general, the steps for setting up a custom login page and graphical images (if any) are as follows:

1. Write the custom webpage.
2. TFTP the page and any embedded JPEG or GIF graphical images that it references onto the switch.
3. Enable and configure web-based Network Login on the switch. When the custom page is present on the switch, it will override the configured banner.

### *Configuring a Network Login Banner*

- To configure a banner on a network login screen, use the following command:

  ```
  configure netlogin banner banner
  ```
- To display configured banners from the network login screen, use the following command:

  ```
  show netlogin banner
  ```
- To clear configured network login banners, use the following command:

  ```
  unconfigure netlogin banner
  ```

### *Login Page Contents*

The customized web-page must have the file name `netlogin_login_page.html`.

While the contents of the page are left up to the customer, they must contain the following elements:

- An HTML submit form with action="/hello" and method="post" that is used to send the Network Login username and password to the switch. The form must contain the following:
  - A username input field with name="extremenetloginuser"
  - A password input field with name="extremenetloginpassword"
- Optionally, one or more graphical images embedded using the tags

  `<img src="netlogin_<xxx>.jpg">` or

  `<img src="netlogin_<xxx>.jpeg">` or

  `<img src="<netlogin_<xxx>.gif">`

  where `<xxx>` is user-configurable.

The following is a sample custom page, where the embedded graphical image is named netlogin_welcome.jpg:

```
<html lang="en">

<head>
<title>Network Login Page</title>
</head>
<body>
<form action="/hello" method="post">
<img src="netlogin_welcome.jpg">
<br/>
Please log in:

<br/>
User:

<input type="text" name="extremenetloginuser" />
<br/>
Password:

<input type="password" name="extremenetloginpassword" />
<br/>
<input type="submit" value="Submit" />
</form>
</body>
</html>
```

*Uploading the Login File to the Switch*

To upload the page and the JPEG/GIF files, the switch TFTP command must be used.

For example, assuming the page resides on a TFTP server with IP address 10.255.49.19, the command used would be:

```
BD-8810.1 # tftp get 10.255.49.19 netlogin_login_page.html
```

*General Guidelines*

The following general guidelines are applicable to the login page:

- When the custom web page is not present on the switch, the system falls back to the using the default banner. The web page may be added (or removed) from the switch at any time, at which point the switch will stop (or start) using the banner.
- The graphical image file names referenced in the web page must not have any path information prepended.
- Both uppercase and lowercase names (or a mixture) for the graphical image filenames are supported, but the user and password tag names should be either all uppercase or all lowercase, not a mixture of the two.
- More than one form may exist on the page. This can be useful when, for example, in addition to the main username and password that is typed in by the user, an additional special username and password needs to be auto-filled and sent. This could be used when end users without a valid username or password need to be given restricted access to the network.

*Limitations*

The following limitations apply to the login page:

- When the client is in the unauthenticated state, any embedded URLs in the custom page are inaccessible to it.
- Only JPEG and GIF graphical images are supported.
- It is the web page writer's responsibility to write the HTML page correctly and without errors.
- Only TFTP is supported as a method to upload the web-page and graphical images to the switch.

## Customizable Authentication Failure Response

In the event of web-based network login authentication failure, you can use a custom authentication failure page to recover.

When a customized login page is in effect, by default, any authentication failure results in the following failure response being delivered to the browser:

```
Login Incorrect. Click here to try again.
```

Clicking on the indicated link will bring the user back to the initial custom login page.

You may choose to override the above default response with a custom one. This custom failure response page must be uploaded to the switch using TFTP with the name `netlogin_login_fail_page.html`. When authentication fails, the switch responds with this page. If the page is deleted from the switch, the response reverts back to the default.

The same graphical images that are uploaded to the switch for the custom login page can also be embedded in the custom authentication failure page.

Note

The custom authentication failure page can be used only when authentication is being done via the custom login page.

## Customizable Graphical Image in Logout Popup Window

You can embed a graphical image in the logout popup window.

This image appears in the window in addition to the text that is displayed. The image must be TFTPed to the switch in the same manner as the custom login image, and must have the filename `netlogin_logout_image.jpg` or `netlogin_logout_image.gif` (depending on whether the image is JPEG or GIF). If no such image is present on the switch, then the logout popup contains only text.

## Web-Based Network Login Configuration Example

The following configuration example shows both the Extreme Networks switch configuration and the *RADIUS* server entries needed to support the example.

*VLAN* corp is assumed to be a corporate subnet which has connections to DNS, WINS servers, network routers, and so on. VLAN temp is a temporary VLAN and is created to provide connections to unauthenticated network login clients. Unauthenticated ports belong to the VLAN temp. This kind of configuration provides better security as unauthenticated clients do not connect to the corporate subnet and will not be able to send or receive any data. They have to get authenticated in order to have access to the network.

- ISP Mode—Network login clients connected to ports 1:10–1:14, VLAN corp, will be logged into the network in ISP mode. This is controlled by the fact that the VLAN in which they reside in unauthenticated mode and the RADIUS server Vendor Specific Attributes (VSA), Extreme-Netlogin-Vlan, are the same, corp. So there will be no port movement. Also if this VSA is missing from RADIUS server, it is assumed to be ISP Mode.
- Campus Mode—On the other hand, clients connected to ports 4:1–4:4, VLAN temp, will be logged into the network in Campus mode since the port will move to the VLAN corp after getting authenticated. A port moves back and forth from one VLAN to the other as its authentication state changes.

Both ISP and Campus mode are not tied to ports but to a user profile. In other words, if the VSA Extreme:Extreme-Netlogin-Vlan represents a VLAN different from the one in which the user currently resides, then VLAN movement will occur after login and after logout. In following example, it is assumed that campus users are connected to ports 4:1–4:4, while ISP users are logged in through ports 1:10–1:14.

**Note**

In the following sample configuration, any lines marked (Default) represent default settings and do not need to be explicitly configured.

```
create vlan “temp”
create vlan “corp”
configure vlan “default” delete ports 4:1-4:4
enable ipforwarding
# Configuration Information for VLAN temp
# No VLAN-ID is associated with VLAN temp.
configure vlan “temp” ipaddress 198.162.32.10 255.255.255.0
# Configuration Information for VLAN corp
# No VLAN-ID is associated with VLAN corp.
configure vlan “corp” ipaddress 10.203.0.224 255.255.255.0
configure vlan “corp” add port 1:10 untagged
configure vlan “corp” add port 1:11 untagged
```

```
configure vlan “corp” add port 1:12 untagged
configure vlan “corp” add port 1:13 untagged
configure vlan “corp” add port 1:14 untagged
# Network Login Configuration
configure vlan “temp” dhcp-address-range 198.162.32.20 - 198.162.32.80
configure vlan “temp” dhcp-options default-gateway 198.162.32.1
configure vlan “temp” dhcp-options dns-server 10.0.1.1
configure vlan “temp” dhcp-options wins-server 10.0.1.85
configure netlogin vlan “temp”
enable netlogin web-based
enable netlogin ports 1:10-1:14,4:1-4:4 web-based
configure netlogin base-url "network-access.com" (Default)
configure netlogin redirect-page http://www.extremenetworks.com (Default)
enable netlogin logout-privilege (Default)
disable netlogin session-refresh  3 (Default)
# DNS Client Configuration
configure dns-client add name-server 10.0.1.1
configure dns-client add name-server 10.0.1.85
#RADIUS Client Configuration
configure radius netlogin primary server 10.0.1.2 1812 client-ip 10.10.20.30 vr “Vr-Mgmt”
configure radius netlogin primary shared-secret purple
enable radius
```

For this example, the following lines (for a FreeRADIUS server) should be added to the RADIUS server users file for each user:

```
Extreme:Extreme-Netlogin-Only = Enabled (if no CLI authorization)
Extreme:Extreme-Netlogin-Vlan = "corp" (destination vlan for CAMPUS mode network login)
```

**Note**

For information about how to use and configure your RADIUS server, refer to Configuring the RADIUS Client on page 913 and the documentation that came with your RADIUS server.

## Web-Based Authentication User Login

To use web-based authentication:

1. Set up the Windows IP configuration for *DHCP*.
2. Plug into the port that has web-based network login enabled.
3. Log in to Windows.

4. Release any old IP settings and renew the DHCP lease.

   This is done differently depending on the version of Windows the user is running:

   Windows 9x—Use the winipcfg tool. Choose the Ethernet adapter that is connected to the port on which network login is enabled. Use the buttons to release the IP configuration and renew the DHCP lease.

   Windows 7 or Windows 8—Use the ipconfig command line utility. Use the command ipconfig/release to release the IP configuration and ipconfig/renew to get the temporary IP address from the switch. If you have more than one Ethernet adapter, specify the adapter by using a number for the adapter following the `ipconfig` command. You can find the adapter number using the command `ipconfig/all`.

   **Note**

   The idea of explicit release/renew is required to bring the network login client machine in the same subnet as the connected *VLAN*. When using web-based authentication, this requirement is mandatory after every logout and before login again as the port moves back and forth between the temporary and permanent VLANs.

   At this point, the client will have its temporary IP address. In this example, the client should have obtained an IP address in the range 198.162.32.20–198.162.32.80.

5. Bring up the browser and enter any URL as `http://www.123.net or http://1.2.3.4` or switch IP address as `http://<IP address>/login` (where IP address could be either temporary or Permanent VLAN Interface for Campus mode).

   URL redirection redirects any URL and IP address to the network login page. This is significant where security matters most, as no knowledge of VLAN interfaces is required to be provided to network login users, because they can login using a URL or IP address.

   **Note**

   URL redirection requires that the switch be configured with a DNS client.

   A page opens with a link for Network Login.

6. Click the Network Login link.

   A dialog box opens requesting a user name and password.

7. Enter the user name and password configured on the *RADIUS* server. After the user has successfully logged in, the user will be redirected to the URL configured on the RADIUS server. During the user login process, the following takes place:

   a. Authentication is done through the RADIUS server.
   b. After successful authentication, the connection information configured on the RADIUS server is returned to the switch:
      - The permanent VLAN
      - The URL to be redirected to (optional)
      - The URL description (optional)
   c. The port is moved to the permanent VLAN.
   d. You can verify this using the `show vlan` command. For more information on the `show vlan` command, see Displaying VLAN Information on page 515.

After a successful login has been achieved, there are several ways that a port can return to a non-authenticated, non-forwarding state:

- The user successfully logs out using the logout web browser window.
- The link from the user to the switch's port is lost.
- There is no activity on the port for 20 minutes.
- An administrator changes the port state.

Note

Because network login is sensitive to state changes during the authentication process, we recommend that you do not log out until the login process is complete. The login process is complete when you receive a permanent address.

## MAC-Based Authentication

MAC-based authentication is used for supplicants that do not support a network login mode, or supplicants that are not aware of the existence of such security measure (for example, an IP phone).

If a MAC address is detected on a MAC-based enabled network login port, an authentication request is sent once to the AAA application. AAA tries to authenticate the MAC address against the configured *RADIUS* server and its configured parameters (timeout, retries, and so on) or the local database.

In a MAC-based authentication environment the authentication verification is done only once at MAC address detection. However, forced reauthentication is allowed through the Session-Timeout VSA supplied by RADIUS. When this VSA is present the switch re-authenticates the client based on the value supplied by the VSA. If no VSA is present, there is no re-authentication.

The credentials used for this are the supplicants MAC address in ASCII representation, and a locally configured password on the switch. If no password is configured, the MAC address is used as the password. You can also group MAC addresses together using a mask.

You can configure a MAC list or a table of MAC entries to filter and authenticate clients based on their MAC addresses. If a match is found in the table of MAC entries, authentication occurs. If no match is found in the table of MAC entries, and a default entry exists, the default will be used to authenticate the client. All entries in the list are automatically sorted in longest prefix order. All passwords are stored and showed encrypted.

You can associate a MAC address with one or more ports. By learning a MAC address, the port confirms the supplicant before sending an authorization request to the RADIUS server. This additional step protects your network against unauthorized supplicants because the port accepts only authorization requests from the MAC address learned on that port. The port blocks all other requests that do not have a matching entry.

### Enabling and Disable MAC-Based Network Login

Network Login must be disabled on a port before you can delete a *VLAN* that contains that port.

- To enable MAC-based network login on the switch, use the following commands:

  ```
  configure netlogin vlan vlan_name

  enable netlogin mac
  ```

Any combination of types of authentication can be enabled on the same switch. At least one of the authentication types must be specified on the CLI.

- To disable MAC-based network login on the switch, use the following command:

  ```
  disable netlogin mac
  ```

- To enable MAC-based network login on one or more ports, use the following command:

  ```
  enable netlogin ports portlist mac
  ```

- To disable MAC-based network login on one or more ports, use the following command:

  ```
  disable netlogin ports portlist mac
  ```

## Associating a MAC Address to a Specific Port

You can configure the switch to accept and authenticate a client with a specific MAC address. Only MAC addresses that have a match for the specific ports are sent for authentication. For example, if you associate a MAC address with one or more ports, only authentication requests for that MAC address received on the port(s) are sent to the configured *RADIUS* server or local database. The port(s) block all other authentication requests that do not have a matching entry. This is also known as secure MAC.

- To associate a MAC address with one or more ports, specify the **ports** option when using the following command:

  ```
  configure netlogin add mac-list [mac {mask} | default] {encrypted}
  {password} {ports port_list}
  ```

  You must enable MAC-based network login on the switch and the specified ports.

  If MAC-based network login is not enabled on the specified port(s), the switch displays a warning message similar to the following:

  ```
  WARNING: Not all specified ports have MAC-Based NetLogin enabled.
  ```

For a sample configuration, see Securing MAC Configuration Example on page 793.

## Adding and Deleting MAC Addresses

- To add a MAC address to the table, use the following command:

  ```
  configure netlogin add mac-list [mac {mask} | default] {encrypted}
  {password} {ports port_list}
  ```

- To remove a MAC address from the table, use the following command:

  ```
  configure netlogin delete mac-list [mac {mask} | default]
  ```

## Displaying the MAC Address List

To display the MAC address table, use the following command:

```
show netlogin mac-list
```

When a client needs authentication the best match will be used to authenticate to the server.

MAC-based authentication is VR aware, so there is one MAC list per VR.

Assume we have a supplicant with MAC address 00:04:96:05:40:00, and the switch has the following table:

```
MAC Address/Mask        Password (encrypted)      Port(s)
--------------------  ----------------------  --------------
00:00:00:00:00:10/48    <not configured>          1:1-1:5
00:00:00:00:00:11/48    <not configured>          1:6-1:10
00:00:00:00:00:12/48    <not configured>          any
00:01:30:70:0C:00/48    yaqu                      any
00:01:30:32:7D:00/48    ravdqsr                   any
00:04:96:00:00:00/24    <not configured>          any
```

The user name used to authenticate against the *RADIUS* server would be "000496000000," as this is the supplicant's MAC address with the configured mask applied. Although this is the default, ExtremeXOS 16.1 allows for a hyphenated mac address to be sent - `configure netlogin mac username format hyphenated`.

Note that the commands are VR aware, and therefore one MAC list table exists per VR.

## Configuring Reauthentication Period

This timer is applicable only in the case where the client is authenticated in authentication failure vlan or authentication service unavailable vlan and the *RADIUS* server provides no session-timeout attribute during authentication. If the switch does receive the session-timeout attribute during authentication, the switch uses that value to set the reauthentication period.

To configure the reauthentication period for network login MAC-based authentication, use the following command:

```
configure netlogin mac timers reauth-period
```

For more information on RADIUS server attributes, see Configuring the RADIUS Client on page 913.

## Securing MAC Configuration Example

The following configuration example shows how to configure secure MAC on your Extreme Networks switch. To configure secure MAC:

1. Create a *VLAN* used for network login.
2. Configure the VLAN for network login.
3. Enable MAC-based network login on the switch.
4. Enable MAC-based network login on the ports used for authentication.
5. Specify one or more ports to accept authentication requests from a specific MAC address.

   In the following example, authentication requests from MAC address:

   - 00:00:00:00:00:10 are only accepted on ports 1:1 through 1:5
   - 00:00:00:00:00:11 are only accepted on ports 1:6 through 1:10
   - 00:00:00:00:00:12 are accepted on all other ports

```
create vlan nlvlan
```

```
configure netlogin vlan nlvlan
enable netlogin mac
enable netlogin ports 1:1-1:10 mac
configure netlogin add mac-list 00:00:00:00:00:10 ports 1:1-1:5
configure netlogin add mac-list 00:00:00:00:00:11 ports 1:6-1:10
configure netlogin add mac-list 00:00:00:00:00:12
```

To view your network login configuration, use the following commands:

```
show netlogin {port portlist vlan vlan_name} {dot1x {detail}} {mac}
{web-based}

show netlogin mac-list
```

## MAC-Based Network Login Configuration Example

The following configuration example shows the Extreme Networks switch configuration needed to support the MAC-based network login example.

```
create vlan "temp"
create vlan "corp"
configure vlan "default" delete ports 4:1-4:4
# Configuration Information for VLAN corp
# No VLAN-ID is associated with VLAN corp.
configure vlan "corp" ipaddress 10.203.0.224 255.255.255.0
# Network Login Configuration
configure netlogin vlan "temp"
enable netlogin mac
enable netlogin ports 4:1-4:4 mac
configure netlogin add mac-list default <password>
# RADIUS Client Configuration
configure radius netlogin primary server 10.0.1.2 1812 client-ip 10.10.20.30 vr "VR-Mgmt"
configure radius netlogin primary shared-secret purple
enable radius
```

The following example is a user's file entry for a specific MAC address on a FreeRADIUS server:

```
00E018A8C540 Auth-Type := Local,  User-Password == "00E018A8C540"
```

Note

For information about how to use and configure your *RADIUS* server, refer to Configuring the RADIUS Client on page 913 and the documentation that came with your RADIUS server.

## MAC-Based Authentication Delay

Prior to ExtremeXOS 16.1, the default behavior was to authenticate the client with all enabled authentication methods on that port for backward compatibility. To delay MAC authentication the user must configure the MAC authentication delay period using the CLI. The MAC authentication delay period's default value is 0 seconds for backward compatibility. The MAC authentication delay period configurable range is 0 to 120 seconds.

The following example explains both the pre-ExtremeXOS 16.1 behavior and the added MAC Authentication Delay feature:

Assume MAC, dot1X and Web-based authentication methods are enabled on a port. When the client is connected to the port the first packet from the client triggers ExtremeXOS to do MAC authentication, authenticates the client using *RADIUS*, and applies the action. When the user "Adam" tries to do the dot1X authentication, ExtremeXOS triggers the dot1X authentication, authenticates "Adam" using RADIUS, and applies the high preferred authentication method's action. If dot1x authentication is configured as preferred over MAC authentication, then the MAC authentication action is unapplied and the dot1X authentication action is applied. In this case the switch authenticates the client using both MAC and dot1x authentication method. This is the existing behavior in which the MAC authentication delay interval is 0 second.

If the customer requirement is to delay/bypass the MAC authentication then the the MAC authentication delay period must be configured on a per port basis. In this case, the moment ExtremeXOS detects the first packet from the client connected port it will wait for the MAC authentication delay period for other authentication methods to be triggered to authenticate the client. In this case the user "Adam" will do dot1X authentication to authenticate himself. The time ExtemeXOS waits for the dot1X authentication to trigger is termed as MAC authentication delay period and it is user configurable.

# Additional Network Login Configuration Details

This section describes additional, optional network login configurations.These configurations are not required to run network login; however, depending on your network settings and environment, you can use the commands described in this section to enhance your network login settings.

Review the earlier sections of this chapter for general information about network login and information about MAC-based, web-based, and 802.1X authentication methods.

## Configuring Network Login MAC-Based VLANs

Currently, network login allows only a single, untagged *VLAN* to exist on a port. This limits the flexibility for untagged supplicants because they must be in the same VLAN.

BlackDiamond 8800, BlackDiamond X8, and Summit family switches support network login MAC-based VLANs. Network login MAC-based VLANs allow a port assigned to a VLAN to operate in a MAC-based fashion. This means that each individual untagged supplicant, identified by its MAC address, can be in different VLANs.

Network login MAC-based VLAN utilizes VSA information from both the network login local database and the *RADIUS* server. After successfully performing the Campus mode operation, the supplicant is added untagged to the destination VLAN.

To support this feature, you must configure the network login port's mode of operation.

### *Network Login MAC-Based VLANs Rules and Restrictions*

This section summarizes the rules and restrictions for configuring network login MAC-based *VLAN*s:

- You must configure and enable network login on the switch and before you configure network login MAC-based VLANs.

  If you attempt to configure the port's mode of operation before enabling network login, the switch displays an error message similar to the following:

```
ERROR: The following ports do not have NetLogin enabled; 1
```

- On ExtremeXOS versions prior to 12.0 on switches other than the Summit family, 10 Gigabit Ethernet ports such as those on the 10G4X I/O module and the uplink ports on the Summit family of switches do not support network login MAC-based VLANs.

  If you attempt to configure network login MAC-based VLANs on 10 Gigabit Ethernet ports, the switch displays an error message similar to the following:

  ```
  ERROR: The following ports do not support the MAC-Based VLAN mode; 1,
  2, 10
  ```

  In ExtremeXOS version 12.0 and later, on the SummitStack and Summit family switches, and on the BlackDiamond 8800 and X8 switches, you can configure mac-based-VLANs on 10 Gigabit Ethernet ports.
- You can have a maximum of 1,024 MAC addresses per I/O module or per switch.

### *Configuring the Port Mode*

To support network login MAC-based *VLAN*s on a network login port, you must configure that port's mode of operation.

Specify MAC-based operation using the following command and specifying mac-based-vlans:

```
configure netlogin ports [all | port_list] mode [mac-based-vlans | port-
based-vlans]
```

By default, the network login port's mode of operation is port-based-vlans. If you modify the mode of operation to mac-based-vlans and later disable all network login protocols on that port, the mode of operation automatically returns to port-based-vlans.

When you change the network login port's mode of operation, the switch deletes all currently known supplicants from the port and restores all VLANs associated with that port to their original state. In addition, by selecting mac-based-vlans, you are unable to manually add or delete untagged VLANs from this port. Network login now controls these VLANs.

With network login MAC-based operation, every authenticated client has an additional *FDB* flag that indicates a translation MAC address. If the supplicant's requested VLAN does not exist on the port, the switch adds the requested VLAN.

### *Displaying Network Login MAC-Based VLAN Information*

The following commands display important information for network login MAC-based *VLAN*s.

#### FDB Information

To view *FDB* entries, use the following command:

```
show fdb netlogin [all | mac-based-vlans]
```

By specifying netlogin, you see only FDB entries related to network login or network login MAC-based *VLAN*s.

The flags associated with network login include:

- v—Indicates the FDB entry was added because the port is part of a MAC-Based virtual port/VLAN combination
- n—Indicates the FDB entry was added by network login

**VLAN and Port Information**

- To view the *VLAN*s that network login adds temporarily in MAC-based mode, use the following command:

  `show ports port_list information detail`

  By specifying **information** and **detail**, the output displays the temporarily added VLANs in network login MAC-based mode.
- To confirm this, review the following output of this command:

  VLAN cfg—The term "MAC-based" appears next to the tag number.

  Netlogin port mode—This output was added to display the port mode of operation. "Mac-based" appears and the network login port mode of operation.
- To view information about the ports that are temporarily added in MAC-based mode for network login, due to discovered MAC addresses, use the following command:

  `show vlan detail`

  By specifying **detail**, the output displays detailed information including the ports associated with the VLAN.

  The flags associated with network login include:
  - a—Indicates that egress traffic is allowed for network login
  - u—Indicates that egress traffic is not allowed for network login.

  m—Indicates that the network login port operates in MAC-based mode.

  Note

  If network login is enabled together with *STP*, the 'a' and 'u' flags are controlled by network login only when the STP port state is 'Forwarding.'

### *Network Login MAC-Based VLAN Example*

The following example configures the network login MAC-based *VLAN* feature:

```
create vlan users12
create vlan nlvlan
configure netlogin vlan nlvlan
enable netlogin mac
enable netlogin ports 1:1-1:10 mac
configure netlogin ports 1:1-1:10 mode mac-based-vlans
configure netlogin add mac-list default MySecretPassword
```

Expanding upon the previous example, you can also utilize the local database for authentication rather than the *RADIUS* server:

```
create netlogin local-user 000000000012 vlan-vsa untagged default
create netlogin local-user 000000000010 vlan-vsa untagged users12
```

For more information about local database authentication, see Local Database Authentication on page 769.

## Configuring Dynamic VLANs for Network Login

During an authentication request, network login receives a destination *VLAN* (if configured on the *RADIUS* server) to put the authenticated user in.

The VLAN must exist on the switch for network login to authenticate the client on that VLAN.

You can configure the switch to dynamically create a VLAN after receiving an authentication response from a RADIUS server. A dynamically created VLAN is only a Layer 2 bridging mechanism; this VLAN does not work with routing protocols to forward traffic. If configured for dynamic VLAN creation, the switch automatically creates a supplicant VLAN that contains both the supplicant's physical port and one or more uplink ports. After the switch unauthenticates all of the supplicants from the dynamically created VLAN, the switch deletes that VLAN.

Note

Dynamically created VLANs do not support the session refresh feature of web-based network login because dynamically created VLANs do not have an IP address.

By dynamically creating and deleting VLANs, you minimize the number of active VLANs configured on your edge switches. In addition, the dynamic VLAN name can be stored on the RADIUS server and supplied to the switch during authentication, simplifying switch management. A key difference between dynamically created VLANs and other VLANs is that the switch does not save dynamically created VLANs. Even if you use the save command, the switch does not save a dynamically created VLAN.

After you configure network login on the switch, the two steps to configure dynamic VLANs are:

- Specifying the tagged uplink port(s) to be added to each dynamically created VLAN.
- Enabling the switch to create dynamic VLANs.

### *Specifying the Uplink Ports*

The uplink ports send traffic to and from the supplicants from the core of the network. Uplink ports should not be configured for network login (network login is disabled on uplink ports).

To specify one or more ports as tagged uplink ports that are added to the dynamically created *VLAN*, use the following command:

```
configure netlogin dynamic-vlan uplink-ports [port_list | none]
```

By default, the setting is none.

If you specify an uplink port with network login enabled, the configuration fails and the switch displays an error message similar to the following:

```
ERROR: The following ports have NetLogin enabled: 1, 2
```

If this occurs, select a port with network login disabled.

- 
- 

*Enabling Dynamic VLANs for Network Login*

By default, the setting is disabled.

To enable the switch to create dynamic *VLAN*s, use the following command:

```
configure netlogin dynamic-vlan [disable | enable]
```

When enabled, the switch dynamically creates VLANs. Remember, dynamically created VLANs are not permanent nor are user-created VLANs. The switch uses the VLAN ID supplied by the *RADIUS* attributes (as described below) to create the VLAN. The switch only creates a dynamic VLAN if the requested VLAN, indicated by the VLAN ID, does not currently exist on the switch.

The RADIUS server uses VSAs to forward VLAN information.The forwarded information can include only a VLAN ID (no VLAN name). The following list specifies the supported VSAs for configuring dynamic VLANs:

- Extreme: Netlogin-VLAN-ID (VSA 209)
- Extreme: Netlogin-Extended-VLAN (VSA 211)
- IETF: Tunnel-Private-Group-ID (VSA 81)

Note

If the ASCII string contains only numbers, it is interpreted as the VLAN ID. Dynamic VLANS support only numerical VLAN IDs; VLAN names are not supported.

The switch automatically generates the VLAN name in the following format: SYS_VLAN_TAG where *TAG* specifies the VLAN ID. For example, a dynamic VLAN with an ID of 10 has the name `SYS_VLAN_0010`.

Note

Like all VLAN names, dynamic VLAN names are unique. If you create a VLAN and use the name of an existing dynamic VLAN, the switch now sees the dynamic VLAN as a user-created VLAN and will save this VLAN to the switch configuration. If this occurs, the switch does not delete the VLAN after the supplicants are authenticated and moved to the permanent VLAN.

For more information on Extreme Networks VSAs, see Extreme Networks VSAs on page 919.

*Dynamic VLAN Example with Web-Based Network Login*

After you finish the web-based network login configuration as described in Web-Based Network Login Configuration Example on page 788, complete the dynamic *VLAN* configuration by:

- Assigning one or more non-network-login ports as uplink ports

  **Note**

  Do not enable network login on uplink ports. If you specify an uplink port with network login enabled, the configuration fails and the switch displays an error message.

- Enabling the switch to dynamically create VLANs

Whether you have MAC-based, web-based, or 802.1X authentication, you use the same two commands to configure dynamic VLANs on the switch.

The following example configures dynamic VLANs on the switch:

```
configure netlogin dynamic-vlan uplink ports 2:1-2:2

configure netlogin dynamic-vlan enable
```

*Displaying Dynamic VLAN Information*

- To display summary information about all of the *VLAN*s on the switch, including any dynamically VLANs currently operating on the switch, use the following command:

  `show vlan`

  If the switch has dynamically created VLANs, the VLAN name begins with `SYS_NLD_`.
- To display the status of dynamic VLAN configuration on the switch, use the following command:

  `show netlogin`

  The switch displays the current state of dynamic VLAN creation (enabled or disabled) and the uplink port(s) associated with the dynamic VLAN.

## Configuring Network Login Port Restart

You can configure network login (*NetLogin*) to restart specific network-login-enabled ports when the last authenticated supplicant unauthenticates, regardless of the configured authentication methods on the port.

This feature, known as *network login port restart*, is available with all network login authentication methods although is most practical with web-based network login. This section describes how this feature behaves with web-based network login; MAC-based and 802.1X network login do not experience any differences in behavior if you enable network login port restart.

Currently with web-based network login, if you have an authenticated supplicant and log out of the network, you must manually release the IP address allocated to you by the *DHCP* server. The DHCP server dynamically manages and allocates IP addresses to supplicants. When a supplicant accesses the network, the DHCP server provides an IP address to that supplicant. DHCP cannot renegotiate their leases, which is why you must manually release the IP address.

For example, if the idle timer expires on the switch, the switch disconnects your network session. If this occurs, it may be unclear why you are unable to access the network. After you manually renew the IP address, you are redirected to the network login login page and can log back into the network. To solve this situation in a single supplicant per port environment, port restart triggers the DHCP client on the PC to restart the DHCP address assignment process.

*Guidelines for Using Network Login Port Restart*

Configure network login port restart on ports with directly attached supplicants.

If you use a hub to connect multiple supplicants, only the last unauthenticated supplicant causes the port to restart. Although the hub does not inflict harm to your network, in this situation, the previously unauthenticated supplicants do not get the benefit of the port restart configuration.

*Enabling Network Login Port Restart*

To enable network login port restart, use the following command:

```
configure netlogin ports [all | port_list] restart
```

*Disabling Network Login Port Restart*

To disable network login port restart, use the following command:

```
configure netlogin [ports port_list | all] allow egress-traffic
{all_cast | broadcast | none | unicast}
```

```
configure netlogin ports [all | port_list] no-restart
```

*Displaying the Port Restart Configuration*

To display the network login settings on the port, including the configuration for port restart, use the following command:

```
show netlogin port port_list
```

Output from this command includes the enable/disable state for network login port restart.

## Authentication Failure and Services Unavailable Handling

ExtremeXOS provides the following features for handling network login authentication failures, and for handling instances of services unavailable:

- Configure Authentication Failure VLAN
- Configure Authentication Services Unavailable VLAN
- Configure Reauthentication Period (for more information see configure netlogin mac timers reauth-period)

You can use these features to set and control the response to network login authentication failure and instances of services unavailable.

*Configuring Authentication Failure VLAN*

When a network login client fails authentication, it is moved to authentication failure *VLAN* and given restricted access.

To configure the authentication failure VLAN, use the following commands:

```
configure netlogin authentication failure vlan
```

```
unconfigure netlogin authentication failure vlan
```

```
enable netlogin authentication failure vlan ports
```

```
disable netlogin authentication failure vlan ports
```

Use the command `enable netlogin authentication failure vlan` to configure authentication failure VLAN on network-login-enabled ports. When a supplicant fails authentication, it is moved to the authentication failure VLAN and is given limited access until it passes the authentication.

Through either a *RADIUS* or local server, the other database is used to authenticate the client depending on the authentication database order for that particular network login method (mac, web, or dot1x). If the final result is authentication failure and if the authentication failure VLAN is configured and enabled on that port, then the client is moved there.

For example, if the network login MAC authentication database order is local, radius and the authentication of a MAC client fails through local database, then the RADIUS server is used to authenticate. If the RADIUS server also fails authentication, the client is moved to the authentication failure VLAN. This applies for all authentication database orders (radius,local; local,radius; radius; local).

In the above example if authentication through local fails but passes through the RADIUS server, the client is moved to appropriate destination VLAN. If the local server authentication fails and the RADIUS server is not available, the client is not moved to authentication failure VLAN.

**Dependency on authentication database order**

There are four different authentication orders which can be configured per authentication method. These four orders are the following:

- *RADIUS*
- Local
- RADIUS, Local
- Local, RADIUS

For each authentication order, the end result is considered in deciding whether to authenticate the client through the authentication failure *VLAN* or the authentication service unavailable VLAN (if configured).

For example, if the authentication order is radius, local, with the RADIUS server unavailable, and local authentication failed, the client is authenticated in the authentication failure VLAN (if one is configured on the port).

For local authentication, if the user is not created in the local database, it is considered as service unavailable. If the user is configured but the password does not match, it is considered as an authentication failure.

For RADIUS server authentication, if for some reason the user cannot be authenticated due to problems with the RADIUS configuration, the RADIUS server not running, or some other problem then it is considered as an authentication service unavailable. If the actual authentication fails then it is considered as an authentication failure.

### *Configuring Authentication Services Unavailable VLAN*

When the authentication service is not available for authentication, the supplicant is moved to authentication service unavailable *VLAN* and given restricted access.

To configure the authentication services unavailable VLAN, use the following commands:

```
configure netlogin authentication service-unavailable vlan
```

```
unconfigure netlogin authentication service-unavailable vlan
```

```
enable netlogin authentication service-unavailable vlan ports
```

```
disable netlogin authentication service-unavailable vlan ports
```

If a network login port has web enabled, authentication failure VLAN and authentication service unavailable VLAN configuration are not applicable to MAC and dot1x clients connected to that port. For example, if port 1:2 has network login MAC and web authentication enabled and authentication failure VLAN is configured and enabled on it, and if a MAC client connected to that port fails authentication, it is not moved to authentication failure VLAN.

# ONEPolicy



ONEPolicy provides for the configuration of role-based profiles for securing and provisioning network resources based upon the role the user or device plays within the enterprise. By first defining the user or device role, network resources can be granularly tailored to a specific user, system, service, or port-based context by configuring and assigning rules to the policy role. A policy role can be configured for any combination of Class of Service, *VLAN (Virtual LAN)* assignment, or default behavior based upon L2, L3, and L4 packet fields. Hybrid authentication allows either policy or dynamic VLAN assignment, or both, to be applied through *RADIUS (Remote Authentication Dial In User Service)* authorization.

Note

For ExtremeXOS 16.1, the software only allows policy to be enabled if all the devices in the stack support policy. At the time of configuration the device will provision the lowest common denominator of functionality. If a device attempts to join the stack after policy is enabled, it must be able to support the existing level of functionality or it will not be allowed to participate in policy.

## ONEPolicy Overview

The three primary benefits of using policy in your network are provisioning and control of network resources, security, and centralized operational efficiency. Policy provides for the provisioning and control of network resources by creating policy roles that allow you to determine network provisioning and control at the appropriate network layer, for a given user or device. With a role defined, rules can be created based upon up to 15 traffic classification types for traffic drop or forwarding. A *CoS (Class of Service)* can be associated with each role for purposes of setting priority, forwarding queue, rate limiting, and rate shaping.

Security can be enhanced by allowing only intended users and devices access to network protocols and capabilities. Some examples are:

- Ensuring that only approved stations can use *SNMP (Simple Network Management Protocol)*, preventing unauthorized stations from viewing, reading, and writing network management information
- Preventing edge clients from attaching network services that are appropriately restricted to data centers and managed by the enterprise IT organization such as *DHCP (Dynamic Host Configuration Protocol)* and DNS services
- Identifying and restricting routing to legitimate routing IP addresses to prevent DoS, spoofing, data integrity and other routing related security issues
- Ensuring that FTP/TFTP file transfers and firmware upgrades only originate from authorized file and configuration management servers
- Preventing clients from using legacy protocols

ExtremeManagement Policy Manager provides a centralized point and click configuration, and one click pushing of defined policy out to all network elements. Use ExtremeManagement Policy Manager for ease of initial configuration and response to security and provisioning issues that may come up during real-time network operation.

Note

When ONEPolicy is enabled certain *MPLS (Multiprotocol Label Switching)*, PSTag, VXLAN, and OpenFlow configurations may not operate.

## Implementing ONEPolicy

To implement ONEPolicy:

- Identify the roles of users and devices in your organization that access the network.
- Create a policy role for each identified user role.
- Associate classification rules and administrative profiles with each policy role.
- Optionally, configure a class of service and associate it directly with the policy role or through a classification rule.
- Optionally, enable hybrid authentication, which allows *RADIUS* filter-ID and tunnel attributes to be used to dynamically assign policy roles and VLANs to authenticating users.
- Optionally, set device response to invalid policy.

## ExtremeManagement Policy Manager

ExtremeManagement Policy Manager is a management GUI that automates the definition and enforcement of network-wide policy rules. It eliminates the need to configure policies on a device-by-device basis using complex CLI commands. The Policy Manager's GUI provides ease of classification rule and policy role creation, because you only define policies once using an easy to understand point and click GUI— and regardless of the number of moves, adds or changes to the policy role, Policy Manager automatically enforces roles on Extreme security-enabled infrastructure devices.

This section presents policy configuration from the perspective of the CLI. Though it is possible to configure policy from the CLI, CLI policy configuration in even a small network can be complex from an

operational point of view. It is highly recommended that policy configuration be performed using the ExtremeManagement Policy Manager.

The ExtremeManagement Policy Manager provides:

- Ease of rule and policy role creation.
- The ability to store and retrieve roles and policies.
- The ability, with a single click, to enforce policy across multiple devices.

## Roles in a Secure Network

The capacity to define roles is directly derived from the ability of supported devices to isolate packet flows by inspecting Layer 2, Layer 3, and Layer 4 packet fields while maintaining line rate. This capability allows for the granular application of a policy to a:

- Specific user (MAC, IP address or interface)
- Group of users (masked MAC or IP address)
- System (IP address)
- Service (such as TCP or UDP)
- Port (physical or application)

Because users, devices, and applications are all identifiable within a flow, a network administrator has the capacity to define and control network access and usage by the actual role the user or device plays in the network. The nature of the security challenge, application access, or amount of network resource required by a given attached user or device, is very much dependent upon the “role” that user or device plays in the enterprise. Defining and applying each role assures that network access and resource usage align with the security requirements, network capabilities, and legitimate user needs as defined by the network administrator.

## The Policy Role

A role, such as sales, admin, or engineering, is first identified and defined in the abstract as the basis for configuring a policy role. Once a role is defined, a policy role is configured and applied to the appropriate context using a set of rules that can control and prioritize various types of network traffic. The rules that make up a policy role contain both classification definitions and actions to be enforced when a classification is matched. Classifications include Layer 2, Layer 3, and Layer 4 packet fields. Policy actions that can be enforced include *VLAN* assignment, filtering, inbound rate limiting, outbound rate shaping, priority class mapping.

## ONEPolicy Roles

### Defining a Policy Role

The policy role is a container that holds all aspects of policy configuration for a specific role. Policy roles are identified by a numeric profile-index value between 1 and the maximum number of roles supported on the platform. Policy roles are configured using the `configure policy profile` command. Policy configuration is either directly specified with the `configure policy profile` command or is associated with the role by specifying the *`profile-index`* value within the command syntax where the given policy option is configured. For example, when configuring a policy maptable entry using the

`configure policy maptable` command (see VLAN to Policy Mapping on page 808), the command syntax requires that you identify the policy role the maptable entry will be associated with, by specifying the profile-index value.

When modifying an existing policy role the default behavior is to replace the existing role with the new policy role configuration. Use the append option to limit the change to the existing policy role to the options specified in the entered command.

A policy role can also be identified by a text name of between 1 and 64 characters. This name value is used by the *RADIUS* filter-ID attribute to identify the policy role to be applied by the switch with a successful authentication.

## Setting Default VLAN for Policy Role

A default *VLAN* can be configured for a policy role. A default VLAN will only be used when either a VLAN is not specifically assigned by a classification rule or all policy role classification rules are missed. To configure a default VLAN, enable pvid-status and specify the port VLAN to be used. pvid-status is disabled by default.

Note

ExtremeXOS supports the assignment of port VLAN-IDs 1 - 4094. Use of VLAN ID 4094 is supported by stackable and standalone devices running v6.51.xx and higher. VLAN-IDs 0 and 4095 can not be assigned as port VLAN-IDs, but do have special meanings within a policy context and can be assigned to the pvid parameter. Within a policy context:

- 0 - Specifies an explicit deny all VLANs
- 4095 - Specifies an explicit permit all VLANs

## Assigning a Class of Service to Policy Role

How a packet is treated as it transits the link can be configured in the *CoS*. It is through a CoS that *QoS (Quality of Service)* is implemented. A CoS can be configured for the following values:

- 802.1p priority
- IP Type of Service (ToS) rewrite value
- Priority Transmit Queue (TxQ) along with a forwarding behavior
- Inbound rate limiter per transmit queue
- Outbound rate shaper per transmit queue

CoS configurations are identified by a numeric value between 0 - 255. 0 - 7 are fixed 802.1p CoS configurations. CoS configurations 8 - 255 are user configurable. Policy uses the cos option followed by the CoS configuration ID value to associate a CoS with a policy role.

## Adding Tagged and Untagged Ports to the VLAN Egress Lists

The *VLAN* egress list contains a list of ports that a frame for this VLAN can exit. Specified ports are automatically assigned to the VLAN egress list for this policy role as tagged or untagged. Ports are

added to the VLAN egress list using the egress-vlans and untagged-vlans options of the `configure policy profile` command.

Note
Egress policy is not supported.

### Overwriting VLAN Tags Priority and Classification Settings

TCI overwrite supports the application of rules to a policy role that overwrite the current user priority and other classification information in the *VLAN* tag's TCI field. TCI overwrite must be enabled for both the policy role and the port the role is applied to.

Use the `configure policy profile` *`profile_index`* `tci-overwrite` command to enable TCI overwrite on a policy role.

## VLAN to Policy Mapping

*VLAN*-to-Policy mapping provides for the manual configuration of a VLAN-to-Policy association that creates a policy maptable entry between the specified VLAN and the specified policy role. A policy maptable holds the VLAN-to-Policy mappings. When an incoming tagged VLAN packet is seen by the switch, a lookup of the policy maptable determines whether a VLAN-to-policy mapping exists. This feature can be used at the distribution layer in environments where non-policy capable edge switches are deployed and there is no possibility of applying Extreme policy at the edge. Tagged frames received at the distribution layer interface for a VLAN with an entry in the policy maptable will have the associated policy applied to the frame.

Use the `configure policy maptable` command specifying a single VLAN ID or range of IDs and the policy profile-index to create a policy maptable entry.

## Applying ONEPolicy Using the RADIUS Response Attributes

If an authentication method that requires communication with an authentication server is configured for a user, the *RADIUS* filter-ID attribute can be used to dynamically assign a policy role to the authenticating user. Supported RADIUS attributes are sent to the switch in the RADIUS access-accept message. The RADIUS filter-ID can also be applied in hybrid authentication mode. Hybrid authentication mode determines how the RADIUS filter-ID and the three RFC 3580 *VLAN* tunnel attributes (VLAN Authorization), when either or all are included in the RADIUS access-accept message, will be handled by the switch. The three VLAN tunnel attributes define the base VLAN-ID to be applied to the user. In either case, conflict resolution between RADIUS attributes is provided by the maptable response feature.

Note
The maptable response feature is only applicable if VLAN Authorization is enabled (`configure policy vlanauthorization enable`).

Note

VLAN-to-policy mapping to maptable response configuration behavior is as follows:

- If the RADIUS response is set to policy, any VLAN-to-policy maptable configuration is ignored for all platforms.
- If the RADIUS response is set to tunnel, VLAN-to-policy mapping can occur on a modular switch platform.
- If the RADIUS response is set to both and both the filter-ID and tunnel attributes are present, VLAN-to-policy mapping configuration is ignored. See the "When Policy Maptable Response is Both" section of the Configuring User Authentication feature guide for exceptions to this behavior.

Use the policy option of the `configure policy maptable response` command to configure the switch to dynamically assign a policy using the RADIUS filter-ID in the RADIUS response message.

## NetLogin Authentication

*NetLogin* provides the dynamic authentication of users as a frontend to policy. Supported authentication type for ExtremeXOS 16.1 includes 8021.1X, MAC Authentication, and Web Authentication.

Unknown unicast/multicast/broadcast traffic is allowed to egress NetLogin-enabled ports, even if the porst are not authenticated. Configuring the port authentication mode as optional/required does not affect egress traffic (`configure netlogin ports [all | port_list] [allowed-users allowed_users | authentication mode [optional | required] | trap [all-traps | no-traps | [{success} {failed} {terminated} {max-reached}]]]` ).

## Applying Policy Using Hybrid Authentication Mode

Hybrid authentication is an authentication capability that allows the switch to use both the filter-ID and tunnel attributes in the *RADIUS* response message to determine how to treat the authenticating user. Hybrid authentication is configured by specifying the **both** option in the `configure policy maptable response` command. The both option:

- Applies the *VLAN* tunnel attributes if they exist and the filter-ID attribute does not
- Applies the filter-ID attribute if it exists and the VLAN tunnel attributes do not
- Applies both the filter-ID and the VLAN tunnel attributes if all attributes exist

  If all attributes exist, the following rules apply:
  - The policy role will be enforced, with the exception that any port PVID specified in the role will be replaced with the VLAN tunnel attributes
  - The policy map is ignored because the policy role is explicitly assigned
  - VLAN classification rules are assigned as defined by the policy role

  vlanauthorization must be enabled or the VLAN tunnel attributes are ignored and the default VLAN is used. Please see the *Configuring User Authentication* feature guide located at www.extremenetworks.com/documentation/ for a complete VLAN Authorization discussion.

Hybrid Mode support eliminates the dependency of VLAN assignment based on roles. As a result, VLANs can be assigned via the tunnel-private-group-ID, as defined per RFC3580, while assigning roles via the filter-ID. This separation gives administrators more flexibility to segment their networks for efficiency beyond the role limits.

## Device Response to Invalid Policy

The action that the device should take when asked to apply an invalid or unknown policy can be specified. The available actions are:

- Ignore the result and search for the next policy assignment rule. If all rules are missed, the default policy is applied.
- Block traffic.
- Forward traffic as if no policy has been assigned using 802.1D/Q rules.

Use the `configure policy invalid action` command to specify a default action to take when asked to apply an invalid or unknown policy.

# Classification Rules

Classification rules associate specific traffic classifications or policy behaviors with the policy role. There are two aspects of classification rule configuration:

- The association of a traffic classification with a policy role by assigning the traffic classification to an administrative profile.
- The assignment of policy rules that define desired policy behaviors for the specified traffic classification type.

Both the administrative profile and policy rules are associated with the policy role by specifying the **`admin-pid`** option, in the case of an administrative profile, or a **`profile-index`** value, in the case of the policy rule. Administrative profiles and policy rules are configured using the `configure policy rule` command.

The administrative profile assigns a traffic classification to a policy role by using the **`admin-profile`** option of the `configure policy rule command`.

Note

Standard policy supports the *VLAN* tag traffic classification for administrative profiles. All other traffic classifications are enhanced policy in an administrative profile context.

Policy rules are based on traffic classifications. Table 96 provides the supported policy rule traffic classification command options and definitions. All other traffic classifications are supported by standard policy.

A detailed discussion of supported traffic classifications is available in the “Traffic Classification Rules” section of the ExtremeManagement Policy Manager online help.

**Table 96: Administrative Policy and Policy Rule Traffic Classifications**

| Traffic Classification | Description | Attribute ID | Enhanced Rule |
|---|---|---|---|
| macsource | Classifies based on MAC source address. | 1 | |
| macdest | Classifies based on MAC destination address. | 2 | |
| ipsourcesocket | Classifies based on source IP address. | 12 | |
| ipdestsocket | Classifies based on destination IP address. | 13 | |
| ip6dest | Classifies based on destination IPv6 address. | 10 | |
| ipttl | Classifies based on TTL. | 20 | |
| ip frag | Classifies based on IP fragmentation value. | 14 | |
| udpsourceportip | Classifies based on UDP source port and optional post-fix IP address. | 15 | |
| udpdestportip | Classifies based on UDP destination port and optional post-fix IP address. | 16 | |
| tcpsourceportip | Classifies based on TCP source port and optional post-fix IP address. | 17 | |
| tcpdestportip | Classifies based on TCP destination port and optional post-fix IP address. | 18 | |
| iptos | Classifies based on Type of Service field in IP packet. | 21 | |
| ipproto | Classifies based on protocol field in IP packet. | 22 | |
| ether | Classifies based on type field in Ethernet II packet. | 25 | |
| port | Classifies based on port-string. | 31 | |

A data value is associated with most traffic classifications to identify the specific network element for that classification. For data value and associated mask details, see the “Valid Values for Policy Classification Rules” table in the `configure policy rule` command discussion of the command reference guide for your platform.

## Specifying Storage Type

ONEPolicy provides for specifying the storage type for this rule entry. Storage types are volatile and non-volatile. Volatile storage does not persist after a reset of the device. Non-volatile storage does persist after a reset of the device. Use the **`storage-type`** option to specify the desired storage type for this policy rule entry in an enhanced policy context.

## Forward and Drop

Packets for this entry can be either forwarded or dropped for this traffic classification using the **`forward`** and **`drop`** policy rule options.

## Allowed Traffic Rule-Type on a Port

Use the `show policy allowed-type` command to display a table of the current allowed and disallowed traffic rule-types for the specified port(s).

See Table 96 on page 811 for a listing of supported allowed traffic classification rule-types.

## Quality of Service in a Policy Rules Context

*QoS* can be specified directly in a policy role as stated in Assigning a Class of Service to Policy Role on page 807. A *CoS* can also be applied to a policy rule. The CoS specified at the policy role level is the default and is only used if no rule is triggered. Therefore, if a CoS is applied to both the policy role and a policy rule, the CoS specified in the policy rule takes precedence over the CoS in the policy role for the traffic classification context specified in the policy rule. As stated in the policy role discussion, CoS configuration details are beyond the scope of this document. See the *QoS Configuration* feature guide located at www.extremenetworks.com/documentation/ for a complete discussion of QoS configuration.

## Blocking Non-Edge Protocols at the Edge Network Layer

Edge clients should be prevented from acting as servers for a number of IP services. If non-edge IP services accidently or maliciously attach to the edge of the network, they are capable of disrupting network operation. IP services should only be allowed where and when your network design requires. This section identifies ten IP Services you should consider blocking at the edge unless allowing them is part of your network architecture.

**Table 97: Non-Edge Protocols**

| Protocol | Policy Effect |
|---|---|
| *DHCP* Server Protocol | Every network needs DHCP. Automatically mitigate the accidental or malicious connection of a DHCP server to the edge of your network to prevent DoS or data integrity issues, by blocking DHCP on the source port for this device. |
| DNS Server Protocol | DNS is critical to network operations. Automatically protect your name servers from malicious attack or unauthorized spoofing and redirection, by blocking DNS on the source port for this device. |
| Routing Topology Protocols | *RIP (Routing Information Protocol)*, *OSPF (Open Shortest Path First)*, and *BGP (Border Gateway Protocol)* topology protocols should only originate from authorized router connection points to ensure reliable network operations. |
| Router Source MAC and Router Source IP Address | Routers and default gateways should not be moving around your network without approved change processes being authorized. Prevent DoS, spoofing, data integrity and other router security issues by blocking router source MAC and router source IP addresses at the edge. |
| SMTP/POP Server Protocols | Prevent data theft and worm propagation by blocking SMTP at the edge. |

**Table 97: Non-Edge Protocols (continued)**

| Protocol | Policy Effect |
|---|---|
| *SNMP* Protocol | Only approved management stations or management data collection points need to be speaking SNMP. Prevent unauthorized users from using SNMP to view, read, or write management information. |
| FTP and TFTP Server Protocols | Ensure file transfers and firmware upgrades are only originating from authorized file and configuration management servers. |
| Web Server Protocol | Stop malicious proxies and application-layer attacks by ensuring only the right Web servers can connect from the right location at the right time, by blocking HTTP on the source port for this device. |
| Legacy Protocols | If IPX, AppleTalk, DECnet or other protocols should no longer be running on your network, prevent clients from using them. Some organizations even take the approach that unless a protocol is specifically allowed, all others are denied. |

## Standard and Enhanced Policy Considerations

This section itemizes additional policy considerations for the stackable and standalone platforms, and provides a table cross-referencing standard and enhanced policy capability and policy capability to traffic classification rules.

Not all stackable fixed switch platforms support policy. On some stackable and standalone fixed switch platforms policy support requires a purchased license. See the software release notes that come with your device for policy support.

Table 98 provides a listing of policy capabilities by standard and enhanced support level. Standard policy capabilities are further granulated based upon traffic classification support. See Table 99 on page 814 for a cross-reference of traffic classification to policy capability support.

**Table 98: Standard and Enhanced Policy Capability Cross-Reference**

| Policy Support Level | Policy Capability |
|---|---|
| Standard | • Dynamic PID Assign Rule – The ability to dynamically assign a policy based upon a traffic classification (macsource and port-string). See **Dynamic** in the following table.<br>• Admin PID Assign Rule – The ability to administratively assign a policy based upon a traffic classification (macsource and port-string). See **Admin** in the following table.<br>• *VLAN* Forwarding – The ability to assign a forwarding VLAN rule through the default profile/role PVID only.<br>• Deny – The ability to assign a drop traffic rule. See **Drop** in the following table.<br>• Permit – The ability to assign a forward traffic rule. See **Forward** in the following table.<br>• *CoS* Assign Rule – The ability to assign a CoS rule. See **CoS** in the following table.<br>• Priority – The ability to assign traffic priority using a CoS assignment. See **CoS** in the following table.<br>• Longest Prefix Rules – The ability to always look at the highest bit mask for an exact traffic classification match. |
| Enhanced | • TCI Overwrite – The ability to overwrite user priority and other VLAN tag TCI field classification information.<br>• Invalid Policy Action – The ability to set a drop, forward, or default-policy behavior based upon an invalid action. |

The following table provides a cross-reference of standard ( ) and enhanced (X) policy capability to traffic classification rule.

**Table 99: Policy Capability to Traffic Classification Rule Cross-Reference**

| Traffic Classification Rule | Dynamic | Admin | VLAN | CoS | Drop | Forward | Syslog | Trap | Disable |
|---|---|---|---|---|---|---|---|---|---|
| MAC Source Address | X | X | | X | X | X | | | |
| MAC Destination Address | | | | X | X | X | | | |
| IPX Source Address | | | | X | X | X | | | |
| IPv6 Destination Address | | | | X | X | X | | | |
| IPX Destination Address | | | | | | | | | |

**Table 99: Policy Capability to Traffic Classification Rule Cross-Reference (continued)**

| Traffic Classification Rule | Dynamic | Admin | VLAN | CoS | Drop | Forward | Syslog | Trap | Disable |
|---|---|---|---|---|---|---|---|---|---|
| IPX Source Socket | | | | | | | | | |
| IPX Destination Socket | | | | | | | | | |
| IPX Transmission Control | | | | | | | | | |
| IPX Type Field | | | | | | | | | |
| IP Source Address | | | | X | X | X | | | |
| IP Destination Address | | | | X | X | X | | | |
| IP Fragmentation | | | | X | X | X | | | |
| UPD Port Source | | | | X | X | X | | | |
| UDP Port Destination | | | | X | X | X | | | |
| TCP Port Source | | | | X | X | X | | | |
| TCP Port Destination | | | | X | X | X | | | |
| *ICMP (Internet Control Message Protocol)* Packet Type | | | | | | | | | |
| Time-To-Live (TTL) | | | | X | X | X | | | |
| IP Type of Service | | | | X | X | X | | | |
| IP Protocol | | | | X | X | X | | | |
| Ether II Packet Type | | | | X | X | X | | | |
| LLC DSAP/ SSAP/CTRL | | | | | | | | | |
| VLAN Tag | | | | | | | | | |
| TCI-Overwrite | | | | | | | | | |
| Port String | X | X | | X | X | X | | | |

## Configuring ONEPolicy

This section presents configuration procedures and tables including command description and syntax in the following policy areas: profile, classification, and display.

Procedure 1 describes how to configure policy roles and related functionality.

**Table 100: Procedure 1**

| Step | Task | Command(s) |
|---|---|---|
| 1 | Create a policy role.<br>• name – (Optional) Specifies a name for this policy profile; used by the filter-ID attribute. This is a string from 1 to 64 characters.<br>• pvid-status – (Optional) Enables or disables PVID override for this policy profile. If all the classification rules associated with this profile are missed, then this parameter, if specified, determines the default *VLAN* for this profile.<br>• pvid – (Optional) Specifies the PVID to assign to packets, if PVID override is enabled and invoked as the default behavior.<br>• cos-status – (Optional) Enables or disables Class of Service override for this policy profile. If all the classification rules associated with this profile are missed, then this parameter, if specified, determines the default *CoS* assignment.<br>• cos – (Optional) Specifies a CoS value to assign to packets, if CoS override is enabled and invoked as the default behavior. Valid values are 0 to 255.<br>• egress-vlans – (Optional) Specifies the port to which this policy profile is applied should be added to the egress list of the VLANs defined by egress-vlans. Packets will be formatted as tagged.<br><br>Note: Egress policy is not supported in ExtremeXOS 16.1.<br>• untagged-vlans – (Optional) Specifies the port to which this policy profile is applied should be added to the egress list of the VLANs defined by untagged-vlans. Packets will be formatted as untagged.<br>• append – (Optional) Appends any egress, forbidden, or untagged specified VLANs to the existing list. If append is not specified, all previous settings for this VLAN list are replaced<br>• clear – (Optional) Clears any egress, forbidden or untagged VLANs specified from the existing list.<br>• tci-overwrite – (Optional) Enhanced policy that enables or disables TCI (Tag Control | `configure policy profile` *`profile-index`* [**`name`** *`name`*] [**`pvid-status`** {**`enable`** \| **`disable`**}] [**`pvid`** *`pvid`*] [**`cos-status`** {**`enable`** \| **`disable`**}] [**`cos`** *`cos`*] [**`egress-vlans`** *`egress-vlans`*] [**`untagged-vlans`** *`untagged-vlans`*] [**`append`**] [**`clear`**] [**`tci-overwrite`** {**`enable`** \| **`disable`**}] |

**Table 100: Procedure 1 (continued)**

| Step | Task | Command(s) |
|---|---|---|
| | Information) overwrite for this profile. When enabled, rules configured for this profile are allowed to overwrite user priority and other classification information in the VLAN tag's TCI field. | |
| 2 | Optionally, for enhanced policy capable devices, assign the action the device will apply to an invalid or unknown policy.<br>• default-policy – Instructs the device to ignore this result and search for the next policy assignment rule.<br>• drop – Instructs the device to block traffic.<br>• forward – Instructs the device to forward traffic. | `configure policy invalid action {`**`default-policy`** `\|` **`drop`** `\|` **`forward`**`}` |
| 3 | Optionally, for enhanced policy capable devices, set a policy maptable entry that associates a VLAN with a policy profile. | `configure policy maptable {`*`vlan-list profile-index`*`}` |
| 4 | Optionally, set a policy maptable response.<br>• tunnel – Applies the VLAN tunnel attribute.<br>• policy – Applies the policy specified in the filter-ID.<br>• both – An enhanced policy option that applies either or all the filter-ID and VLAN tunnel attributes or the policy depending upon whether one or both are present. | `configure policy maptable response {`**`tunnel`** `\|` **`policy`** `\|` **`both`**`}` |

Procedure 2 describes how to configure classification rules as an administrative profile or to assign policy rules to a policy role.

**Table 101: Procedure 2**

| Step | Task | Command(s) |
|---|---|---|
| 1 | Optionally set an administrative profile to assign traffic classifications to a policy role. See Table 96 on page 811 for traffic classification-type descriptions and enhanced policy information. See the set policy rule command discussion in the command reference guide that comes with your device for traffic classification data and mask information.<br>• port-string – Applies this administratively-assigned rule to a specific ingress port. The `configure policy port` command is also supported as an alternative way to administratively assign a profile rule to a port.<br>• storage-type – (Optional) Adds or removes this entry from non-volatile storage.<br>• admin-pid – Associates this administrative profile with a policy profile index ID. Valid values are 1 - 1023. | `configure policy rule admin-profile {macsource \| port} [data] [mask mask] port-string port-string [storage-type {non-volatile \| volatile}] [admin-pid admin-pid]` |
| 2 | Optionally configure policy rules to associate with a policy role. See Table 96 on page 811 for traffic classification-type descriptions and enhanced policy information. See the `configure policy rule` command discussion in the command reference guide that comes with your device for traffic classification data and mask information.<br>• port-string – (Optional) Applies this policy rule to a specific ingress port. The `set policy port` command is also supported as an alternative way to assign a profile rule to a port.<br>• storage-type – (Optional) Adds or removes this entry from non-volatile storage.<br>• drop \| forward – (Optional) Specifies that packets within this classification will be dropped or forwarded.<br>• cos – (Optional) Specifies that this rule will classify to a Class-of-Service ID. Valid values are 0 - 255. A value of -1 indicates that no CoS forwarding behavior modification is desired. | `configure policy rule profile-index classification-type [data] [mask mask] [port-string port-string] [storage-type {non-volatile \| volatile}] \| [drop \| forward] [admin-pid admin-pid] [cos cos]` |
| 3 | Optionally, for enhanced policy capable devices, assign a policy role to a port. | `configure policy port <ports> admin-id admin_id` |

The following table describes how to display policy information and statistics.

**Table 102: Displaying Policy Configuration and Statistics**

| Task | Command(s) |
|---|---|
| Display policy role information. | `show policy profile {all \| profile-index [-detail]}` |
| Display the action the device should take if asked to apply an invalid or unknown policy, or the number of times the device has detected an invalid/unknown policy, or both action and count information. | `show policy invalid {action \| count \| all}` |
| Display VLAN-ID to policy role mappings table. | `show policy maptable [vlan-list]` |
| Display policy classification and admin rule information. | `show policy rule [classification-type] [data] [mask mask] [port-string port-string] [storage-type {non-volatile \| volatile}] \| [drop \| forward] [dynamic-pid dynamic-pid] [cos cos] [admin-pid admin-pid] [-verbose] [-wide]` |
| Display all policy classification capabilities for this device. | `show policy capability` |
| Display a list of currently supported traffic rules applied to the administrative profile for one or more ports. | `show policy allowed-type ports [detail]` |
| Display status of dynamically assigned roles. | `show policy dynamic override` |

## ONEPolicy Configuration Example

This section presents a college-based ONEPolicy configuration example. The following figure displays an overview of the policy configuration. This overview display is followed by a complete discussion of the configuration example.

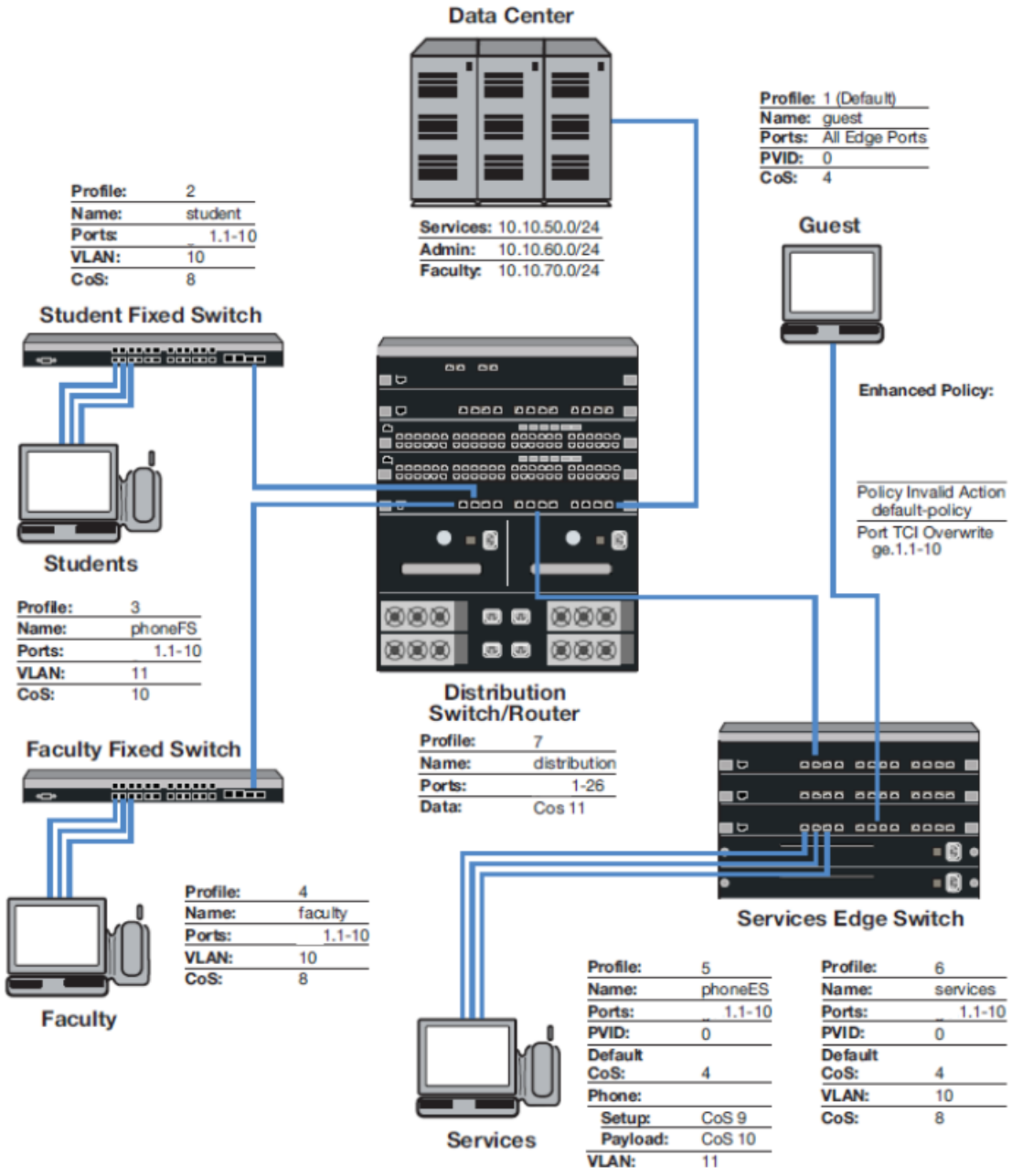



**Figure 103: College-Based Policy Configuration**

## Roles

The example defines the following roles:

- guest – Used as the default policy for all unauthenticated ports. Connects a PC to the network providing internet only access to the network. Provides guest access to a limited number of the edge switch ports to be used specifically for internet only access. Policy is applied using the port level

default configuration, or by authentication, in the case of the Services Edge Switch port internet only access PCs.

- student – Connects a dorm room PC to the network through a “Student” Fixed Switch port. A configured *CoS* rate limits the PC. Configured rules deny access to administrative and faculty servers. The PC authenticates using *RADIUS*. Hybrid authentication is enabled. The student policy role is applied using the filter-ID attribute. The base *VLAN* is applied using the tunnel attributes returned in the RADIUS response message. If all rules are missed, the settings configured in the student policy profile are applied.
- phoneFS – Connects a dorm room or faculty office VoIP phone to the network using a stackable fixed switch port. A configured CoS rate limits the phone and applies a high priority. The phone authenticates using RADIUS. Hybrid authentication is enabled. Policy is applied using the filter-ID returned in the RADIUS response message. The base VLAN is applied using the tunnel attributes returned in the RADIUS response message. If all rules are missed, the settings configured in the phoneFS policy profile are applied.
- faculty – Connects a faculty office PC to the network through a “Faculty” Fixed Switch port. A configured CoS rate limits the PC. A configured rule denies access to the administrative servers. The PC authenticates using RADIUS. Hybrid authentication is enabled. The faculty policy role is applied using the filter-ID attribute. The base VLAN is applied using the tunnel attributes returned in the RADIUS response message for the authenticating user. If all rules are missed, the settings configured in the faculty policy profile are applied.
- phoneES – Connects a services VoIP phone to the network using a Services Edge Switch port. A configured CoS rate limits the phone for both setup and payload, and applies a high priority. The phone authenticates using RADIUS. Tunnel authentication is enabled. The base VLAN is applied using the tunnel attributes returned in the RADIUS response message. Policy is applied using a maptable configuration. If all rules are missed, the settings configured in the phoneES policy profile are applied.
- services – Connects a services PC to the network through the Services Edge Switch port. A configured CoS rate limits the PC. Services are denied access to both the student and faculty servers. The PC authenticates using RADIUS. The base VLAN is applied using the tunnel attributes returned in the RADIUS response message for the authenticating user. The services policy role is applied using a policy maptable setting. The policy invalid action and TCI overwrite are enabled for this role. If all rules are missed, the settings configured in the services policy profile are applied.
- distribution – The Distribution policy role is applied at the Distribution Switch providing rate limiting.

## Policy Domains

It is useful to break up policy implementation into logical domains for ease of understanding and configuration. For this example, it is useful to consider four domains: basic edge, standard edge on the Fixed Switch, premium edge on the Services Edge Switch, and premium distribution on the Distribution Switch.

### *Basic Edge*

Protocols not appropriate to the edge should be blocked. For this example we will block *DHCP*, DNS, *SNMP*, SSH, Telnet and FTP at the edge on the data *VLAN*. We will forward destination port DHCP and DNS and source port for IP address request to facilitate auto configuration and IP address assignment. See Blocking Non-Edge Protocols at the Edge Network Layer on page 812 for a listing of protocols you should consider blocking at the edge.

*Standard Edge*

Edge Switch platforms will be rate-limited using a configured *CoS* that will be applied to the student and faculty, and phoneFS policy roles. Fixed Switch support for hybrid authentication depends upon the platform and firmware release. The Fixed Switch in this example supports the hybrid authentication capability. Hybrid authentication will be enabled.

*Premium Edge*

The Edge Switch will be rate-limited using a configured CoS that is applied to the services and phoneES policy role. This premium edge platform will be enabled for the following capabilities:

- Invalid policy action set to drop
- TCI overwrite enabled

*Premium Distribution*

The Distribution Switch Router will be rate-limited using a configured CoS. Premium distribution will be enabled for the following policy capabilities:

- Invalid policy action set to drop
- TCI overwrite enabled

## Platform Configuration

This section provides the CLI based policy configuration on the following platforms:

- Student Fixed Switch
- Faculty Fixed Switch
- Services Edge Switch
- Distribution Switch

In CLI mode, configuration takes place on each platform. When using the ExtremeManagement Policy Manager, configuration takes place at a central location and is pushed out to the appropriate network devices. For this configuration example, *CoS* related configuration will be specified as a final CoS. See the *QoS* Configuration feature guide located at http://documentation.extremenetworks.com for a complete discussion of QoS configuration.

Note

CLI command prompts used in this configuration example have the following meanings:

- `System->`—Input on all platforms used in this example.
- `Fixed Switch->`—Input on all Fixed Switches.
- `StudentFS->`—Input on the student Fixed Switch.
- `FacultyFS->`—Input on the faculty Fixed Switch.
- `Services->`—Input on the services S-Series device.
- `Distribution ->`—Input on the distribution S-Series device.

## Configuring Guest Policy on Edge Platforms

All edge ports will be set with a default guest policy using the `configure policy port` command. This guest policy provides for an internet only access to the network. Users on all ports will attempt to authenticate. If the authentication succeeds, the policy returned by authentication or, in the case of the Services Edge Switch configuration, the maptable setting, overrides the default port policy setting. If authentication fails, the guest policy is used. On the Services Edge Switch , five ports are used by PCs at locations throughout the campus, such as the library, to provide access to the internet. The PCs attached to these five ports will authenticate with the guest policy role. Public facing services would also be configured for guest status in a school or enterprise scenario. Public facing services are not part of this example.

### *Configuring the Policy Role*

The guest role is configured with:

- A profile-index value of 1
- A name of guest
- A PVID set to 0
- A *CoS* set to 4

Create the guest policy profile on all platforms:

```
configure policy profile 1 name guest pvid-status enable pvid 0 cos-status enable cos 4
```

### *Assigning Traffic Classification Rules*

For cases where discovery must take place to assign an IP address, DNS and *DHCP* traffic must be allowed. Forwarding of traffic is allowed on UDP source port 68 (IP address request) and UDP destination ports 53 (DNS) and 67 (DHCP).

```
configure policy rule 1 udpsourceport 68 mask 16 forward
configure policy rule 1 udpdestportIP 53 mask 16 forward configure policy rule 1
udpdestportIP 67 mask 16 forward
```

Guest policy allows internet traffic. TCP destination Ports 80, 8080, and 443 will be allowed traffic forwarding.

```
configure policy rule 1 tcpdestportIP 80 mask 16 forward
configure policy rule 1 tcpdestportIP 443 mask 16 forward
configure policy rule 1 tcpdestport 8080 mask 16 forward
```

ARP forwarding is required on ether port 0x806.

```
configure policy rule 1 ether 0x806 mask 16 forward
```

### *Assigning the Guest Policy Profile to All Edge Ports*

Assign the guest policy profile to all Fixed Switch and Services Edge Switch ports.

```
configure policy port 1-47 1
```

## Configuring Policy for the Edge Student Fixed Switch

### *Configuring the Policy Role*

The student role is configured with:

- A profile-index value of 2
- A name of student
- A port *VLAN* of 10
- A *CoS* of 8

Create a policy role that applies a CoS 8 to data VLAN 10 and configures it to rate-limit traffic to 1M with a moderate priority of 5.

```
StudentFS->configure policy profile 2 name student pvid-status
enable pvid 10 cos-status enable cos 8
```

### *Assigning Hybrid Authentication*

Configure the *RADIUS* server user accounts with the appropriate tunnel information using VLAN authorization and policy filter-ID for student role members and devices. Enable hybrid authentication, allowing the switch to use both the filter-ID and tunnel attributes in the RADIUS response message. Set a VLAN-to-policy mapping as backup incase the response does not include the RADIUS filter-ID attribute. This mapping is ignored in case RADIUS filter-ID attribute is present in the RADIUS response message.

```
StudentFS->configure policy maptable response both
StudentFS->configure policy maptable 10 2
```

### *Assigning Traffic Classification Rules*

Forward traffic on UDP source port for IP address request (68), and UDP destination ports for protocols *DHCP* (67) and DNS (53). Drop traffic on UDP source ports for protocols DHCP (67) and DNS (53). Drop traffic for protocols *SNMP* (161), SSH (22), Telnet (23) and FTP (20 and 21) on both the data and phone VLANs.

```
StudentFS->configure policy rule 2 udpsourceport 68 mask 16 forward
StudentFS->configure policy rule 2 udpdestport 67 mask 16 forward
StudentFS->configure policy rule 2 udpdestport 53 mask 16 forward
StudentFS->configure policy rule 2 udpsourceportIP 67 mask 16 drop
StudentFS->configure policy rule 2 udpsourceportIP 53 mask 16 drop
StudentFS->configure policy rule 2 udpdestportIP 16 mask 16 drop
StudentFS->configure policy rule 2 tcpdestportIP 22 mask 16 drop
StudentFS->configure policy rule 2 tcpdestportIP 23 mask 16 drop
StudentFS->configure policy rule 2 tcpdestportIP 20 mask 16 drop
StudentFS->configure policy rule 2 tcpdestportIP 21 mask 16 drop
```

Students should only be allowed access to the services server (subnet 10.10.50.0/24) and should be denied access to both the administrative (subnet 10.10.60.0/24) and faculty servers (subnet 10.10.70.0/24).

```
StudentFS->configure policy rule 2 ipdest 10.10.60.0 mask 24 drop
StudentFS->configure policy rule 2 ipdest 10.10.70.0 mask 24 drop
```

## Configuring PhoneFS Policy for the Edge Fixed Switch

### *Configuring the Policy Role*

The phoneFS role is configured on both the dorm room and faculty office Fixed Switches with:

- A profile-index of 3
- A name of phoneFS
- A port *VLAN* of 11
- A *CoS* of 10

Because we can not apply separate rate limits to the phone setup and payload ports on the Fixed Switch using policy rules, apply CoS 10 with the higher payload appropriate rate limit of 100k bps and a high priority of 6 to the phoneFS role.

```
Fixed Switch->configure policy profile 3 name phoneFS pvid-status enable pvid 11
cos-status enable cos 10
```

### *Assigning Traffic Classification Rules*

Drop traffic for protocols *SNMP* (161), SSH (22), Telnet (23) and FTP (20 and 21) on the phone VLAN. Forward traffic on UDP source port for IP address request (68) and forward traffic on UDP destination ports for protocols *DHCP* (67) and DNS (53) on the phone VLAN, to facilitate phone auto configuration and IP address assignment.

```
Fixed Switch->configure policy rule 3 udpdestportIP 161 mask 16 drop
Fixed Switch->configure policy rule 3 tcpdestportIP 22 mask 16 drop
Fixed Switch->configure policy rule 3 tcpdestportIP 23 mask 16 drop
Fixed Switch->configure policy rule 3 tcpdestportIP 20 mask 16 drop
Fixed Switch->configure policy rule 3 tcpdestportIP 21 mask 16 drop
Fixed Switch->configure policy rule 3 udpsourceport 68 mask 16 forward
Fixed Switch->configure policy rule 3 udpdestportIP 67 mask 16 forward
Fixed Switch->configure policy rule 3 udpdestportIP 53 mask 16 forward
```

### *Assigning Hybrid Authentication*

Configure the *RADIUS* server user accounts with the appropriate tunnel information using VLAN authorization and policy filter-ID for phoneFS role members and devices. Enable hybrid authentication, allowing the switch to use both the filter-ID and tunnel attributes in the RADIUS response message. Set a VLAN-to-policy mapping as backup incase the response does not include the RADIUS filter-ID attribute. This mapping is ignored if RADIUS filter-ID attribute is present in the RADIUS response message.

```
Fixed Switch->configure policy maptable response both
Fixed Switch->configure policy maptable 11 3
```

## Configuring Policy for the Edge Faculty Fixed Switch

### *Configuring the Policy Role*

The faculty role is configured with:

- A profile-index value of 4
- A name of faculty
- A port *VLAN* of 10
- A *CoS* of 8

Create a policy role that applies a CoS 8 to data VLAN 10 and configures it to rate-limit traffic to 1M with a moderate priority of 5.

```
FacultyFS->configure policy profile 4 name faculty pvid-status enable pvid 10
cos-status enable cos 8
```

### *Assigning Hybrid Authentication*

CConfigure the *RADIUS* server user accounts with the appropriate tunnel information using VLAN authorization and policy filter-ID for faculty role members and devices. Enable hybrid authentication. Set a VLAN-to-policy mapping. This mapping is ignored if the RADIUS filter-ID attribute is present in the RADIUS response message.

```
FacultyFS->configure policy maptable response both
FacultyFS->configure policy maptable 10 4
```

### *Assigning Traffic Classification Rules*

Forward traffic on UDP source port for IP address request (68), and UDP destination ports for protocols *DHCP* (67) and DNS (53). Drop traffic on UDP source ports for protocols DHCP (67) and DNS (53). Drop traffic for protocols *SNMP* (161), SSH (22), Telnet (23) and FTP (20 and 21) on both the data and phone VLANs

```
FacultyFS->configure policy rule 4 udpsourceport 68 mask 16 forward
FacultyFS->configure policy rule 4 udpdestport 67 mask 16 forward
FacultyFS->configure policy rule 4 udpdestport 53 mask 16 forward
FacultyFS->configure policy rule 4 udpsourceportIP 67 mask 16 drop
FacultyFS->configure policy rule 4 udpsourceportIP 53 mask 16 drop
FacultyFS->configure policy rule 4 udpdestportIP 16 mask 16 drop
FacultyFS->configure policy rule 4 tcpdestportIP 22 mask 16 drop
FacultyFS->configure policy rule 4 tcpdestportIP 23 mask 16 drop
FacultyFS->configure policy rule 4 tcpdestportIP 20 mask 16 drop
FacultyFS->configure policy rule 4 tcpdestportIP 21 mask 16 drop
```

Faculty should only be allowed access to the services (subnet 10.10.50.0/24) and the faculty servers (subnet 10.10.70.0/24) and should be denied access to the administrative server (subnet 10.10.60.0/24).

```
FacultyFS->configure policy rule 4 ipdest 10.10.60.0 mask 24 drop
```

## Configuring PhoneES Policy for the Services Edge Switch

### *Configuring the Policy Role*

The phoneES role is configured on the Services Edge Switch with:

- A profile-index of 5 Policy
- A name of phoneES
- A default port *VLAN* of 0
- A default *CoS* of 4

Because VLANs can be applied to Services Edge Switch ports using the appropriate traffic classification, the explicit deny all PVID 0 will be applied at policy creation. Separate rate limits can be applied to the phone setup and payload ports on the Services Edge Switch using policy rules. A default CoS of 4 will be applied at policy role creation.

```
ServicesES->configure policy profile 5 name phoneES pvid-status enable pvid 0
cos-status enable cos 4
```

### *Assigning Traffic Classification Rules*

Forward traffic on UDP source port for IP address request (68) and and forward traffic on UDP destination ports for protocols *DHCP* (67) and DNS (53) on the phone VLAN, to facilitate phone auto configuration and IP address assignment. Drop traffic for protocols *SNMP* (161), SSH (22), Telnet (23) and FTP (20 and 21) on the phone VLAN.

```
ServicesES->configure policy rule 5 udpsourceport 68 mask 16 forward
ServicesES->configure policy rule 5 udpdestportIP 67 mask 16 forward
ServicesES->configure policy rule 5 udpdestportIP 53 mask 16 forward
ServicesES->configure policy rule 5 udpdestportIP 161 mask 16 drop
ServicesES->configure policy rule 5 tcpdestportIP 22 mask 16 drop
ServicesES->configure policy rule 5 tcpdestportIP 23 mask 16 drop
ServicesES->configure policy rule 5 tcpdestportIP 20 mask 16 drop
ServicesES->configure policy rule 5 tcpdestportIP 21 mask 16 drop
```

Apply a CoS 9 to phone setup data on VLAN 11, rate limiting the data to 5 pps with a high priority of 7 on port 2427.

Apply a CoS 10 to phone payload data on VLAN 11, rate limiting the data to 100k bps with a high priority of 7 for both source and destination on port 5004.

```
ServicesES->configure policy rule 5 upddestIP 2427 mask 16 vlan 11 cos 9
ServicesES->configure policy rule 5 updsourceIP 5004 mask 16 vlan 11 cos 10
ServicesES->configure policy rule 5 upddestIP 5004 mask 16 vlan 11 cos 10
```

### *Assigning the VLAN-to-Policy Association*

The nature of services related devices that might connect to a switch port is not as static as with the student or faculty roles. Services related network needs can run the gamut from temporary multimedia events to standard office users. There may be multiple VLAN and policy role associations that take care of services related needs, depending upon the connected user. This may include the requirement for multiple services related roles.

For services, the network administrator desires greater resource usage flexibility in assigning the policy to VLAN association. Authentication in this case will return only the tunnel attributes in the response message based upon the requirements of the authenticating user. Setting the VLAN-to-policy association will be handled by the maptable configuration, allowing for ease in changing the policy associated with a VLAN on the fly using Policy Manager. Specify that the tunnel attributes returned in the *RADIUS* response message will be used by the authenticating user. Associate VLAN 11 with policy role 5 using the `configure policy maptable` command.

```
ServicesES->configure policy maptable response tunnel
ServicesES->configure policy maptable 11 5
```

## Configuring Policy for the Services Edge Switch

### *Configuring the Policy Role*

The services role is configured with:

- A profile-index value of 6
- A name of services
- A default port *VLAN* of 0

- A default *CoS* when no rule overrides CoS
- TCI overwrite enabled

```
ServicesES->set policy profile 6 name services pvid-status enable pvid 0
cos-status enable cos 4 tci-overwrite enable
```

### *Assigning the VLAN-to-Policy Association*

Setting the VLAN-to-policy association will be handled by the policy maptable setting, allowing for ease in changing the policy associated with a VLAN on the fly using Policy Manager. Specify that the tunnel attributes returned in the *RADIUS* response message will be used by the authenticating user. Associate VLAN 10 with policy role 6 using the `set policy maptable` command.

```
ServicesES->set policy maptable response tunnel
ServicesES->set policy maptable 10 6
```

### *Assigning Traffic Classification Rules*

Forward traffic on UDP source port for IP address request (68) and forward traffic on UDP destination ports for protocols *DHCP* (67) and DNS (53) on the data VLAN, to facilitate PC auto configuration and IP address assignment. Drop traffic for protocols *SNMP* (161), SSH (22), Telnet (23) and FTP (20 and 21) on the phone VLAN.

```
ServicesES->configure policy rule 6 udpsourceportIP 68 mask 16 vlan 10 forward
ServicesES->configure policy rule 6 udpdestportIP 67 mask 16 vlan 10 forward
ServicesES->configure policy rule 6 udpdestportIP 53 mask 16 vlan 10 forward
ServicesES->configure policy rule 6 udpdestportIP 67 mask 16 vlan 10 drop
ServicesES->configure policy rule 6 udpdestportIP 53 mask 16 vlan 10 drop
ServicesES->configure policy rule 6 udpdestportIP 161 mask 16 drop
ServicesES->configure policy rule 6 tcpdestportIP 22 mask 16 drop
ServicesES->configure policy rule 6 tcpdestportIP 23 mask 16 drop
ServicesES->configure policy rule 6 tcpdestportIP 20 mask 16 drop
ServicesES->configure policy rule 6 tcpdestportIP 21 mask 16 drop
```

Apply a CoS 8 to data VLAN 10 and configure it to rate-limit traffic to 1M and moderate priority of 5 for services IP subnet 10.10.30.0 mask 28.

```
ServicesES->configure policy rule 6 ipsource 10.10.30.0 mask 28 vlan 10 cos 8
```

Services should only be allowed access to the services server (subnet 10.10.50.0/24) and should be denied access to the faculty servers (subnet 10.10.70.0/24) and administrative servers (subnet 10.10.60.0/24).

```
ServicesES->configure policy rule 6 ipdest 10.10.60.0 mask 24 drop
ServicesES->configure policy rule 6 ipdest 10.10.70.0 mask 24 drop
```

### *Enable Enhanced Edge Switch Capabilities on the Services Edge Switch Platform*

The Services Edge Switch platform supports invalid action set to default policy should an invalid policy occur.

```
ServicesES->configure policy invalid action default-policy
```

## Configuring the Distribution Layer Role

### *Configuring the Policy Role*

The distribution role is configured with:

- A profile-index value of 7
- A name of distribution
- A default *CoS* when no rule overrides CoS
- TCI overwrite enabled

```
Distribution(rw)->configure policy profile 7 name distribution cos-status enable cos 4
tci-overwrite enable
```

### *Assigning Traffic Classification Rules*

Assign a CoS to distribution up and down stream link ports, rate-limiting the traffic to 25M.

```
Distribution(rw)->configure policy rule 7 port 1-26 cos 11
Distribution(rw)->configure policy rule 7 port 1-26 cos 11
```

### *Enable Enhanced Policy Capability*

The following enhanced policy capability is enabled: invalid action set to default policy should an invalid policy occur.

```
ServicesES(rw)->configure policy invalid action default-policy
```

This completes the policy configuration for this school example.

# Terms and Definitions

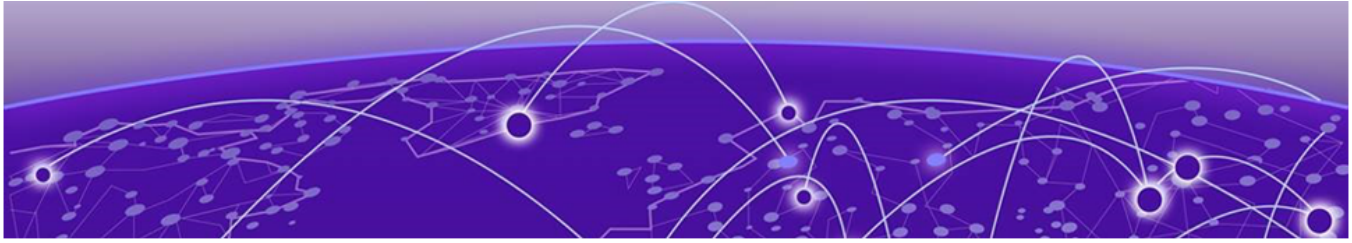
The following table lists terms and definitions used in this policy configuration discussion.

**Table 103: Policy Configuration Terms and Definitions**

| Term | Definition |
|---|---|
| Administrative Profile | A logical container that assigns a traffic classification to a policy role. |
| *CoS* | A logical container for packet priority, queue, and forwarding treatment that determines how the firmware treats a packet as it transits the link. |
| Filter-ID | A string that is formatted in the *RADIUS* access-accept packet sent back from the authentication server to the switch during the authentication process. In the Extreme policy context, the string contains the name of the policy role to be applied to the authenticating user or device. |
| Hybrid Authentication | An authentication feature that allows the switch to use both the filter-ID and tunnel attributes in the RADIUS response message to determine how to treat the authenticating user. |
| Policy | A component of Secure Networks that provides for the configuration of a role based profile for the securing and provisioning of network resources based upon the function the user or device plays within the enterprise network. |

**Table 103: Policy Configuration Terms and Definitions (continued)**

| Term | Definition |
|---|---|
| Policy Maptable | A logical entity that can be configured to provide *VLAN* to policy role mappings. |
| Policy Profile/Role | A logical container for the rules that define a particular policy role. |
| Policy Rule | A logical container providing for the specification of policy behaviors associated with a policy role. |
| Role | The grouping of individual users or devices into a logical behavioral profile for the purpose of applying policy. |
| Rule Precedence | A numeric traffic classification value, associated with the policy role, the ordering of which on a precedence list determines the sequence in which classification rules are applied to a packet.<br><br>Note: Rule precedence is fixed (i.e. not configurable) in ExtremeXOS 16.1. |
| TCI Overwrite | A policy feature, when enabled in a policy role-based tci-overwrite only, allows for the overwrite of the current user priority and other classification information in the VLAN tag's TCI field. |
| Traffic Classification | A network element such as MAC or IP address, packet type, port, or VLAN used as the basis for identifying the traffic to which the policy will be applied. |
| Untagged and Tagged VLAN | Untagged VLAN frames are classified to the VLAN associated with the port it enters. Tagged VLAN frames are classified to the VLAN specified in the VLAN tag; the PVID is ignored. |
| VLAN Authorization | An aspect of RFC3580 that provides for the inclusion of the VLAN tunnel attribute in the RADIUS Access-Accept packet defining the base VLAN-ID to be applied to the authenticating user or device. |
| VLAN Egress List | A configured list of ports that a frame for this VLAN can exit. |



# Identity Management



This chapter offers detailed information about the ExtremeXOS Identity Management feature. It provides an overview, as well as specific information on how to configure, manage and monitor this feature.

## Identity Management Overview

The identity management feature allows you to learn more about the users and devices (such as phones and routers) that connect to a switch. In this chapter, users and devices are collectively called *identities*. The Identity Management feature:

- Captures identity information when users and devices connect to and disconnect from the switch.
- Stores captured identity information and identity event data in a local database.
- Generates *EMS (Event Management System)* messages for user and device events.
- Makes collected identity information available for viewing by admin-level users and to management applications such as ExtremeManagement or Ridgeline through XML APIs.
- Uses locally collected identity information to query an LDAP server and collect additional information about connected identities.
- Supports custom configurations called *roles*, which are selected based on identity information collected locally and from an LDAP server.
- Uses roles to enable traffic filtering, counting, and metering on ports (using ACLs and policies) in response to identity events (connections, disconnections, and time-outs).
- Supports the configuration of blacklist to deny all access to an identity and whitelists to permit all access to an identity.
- Supports the configuration of greylist to enable the network administrator to choose usernames whose identity is not required to be maintained. When these usernames are added to greylist, the Identity Management module does not create an identity when these users log on.

- Integrates with UPM to modify the switch configuration in response to discovered identities.
- Services users under different domains by allowing different domains to be configured and then associating different LDAP servers for those different domains.

Note

This chapter discusses identity management features that are managed using the switch CLI. Related features are described in other chapters and in the ExtremeManagement and Ridgeline product documentation. For a description of identity management that ties all the related components together, see the application note titled *Deploying an Identity Aware Network*, which is available from the www.extremenetworks.com.

## Identity Information Capture

The identity management feature collects user and device data whenever users or devices connect to or disconnect from the switch. The table below lists the identity management attributes that the identity manager process collects from the listed switch software components.

**Table 104: Identity (User/Device) Attributes and Source Software Components**

| Attribute | NetLogin | *LLDP (Link Layer Discovery Protocol)* | *FDB (forwarding database)* | IP-Security | Kerberos Snooping |
|---|---|---|---|---|---|
| User's MAC address | X | X | X | X | X |
| Authentication and unauthentication time stamp | X | X | X | X | X |
| User's port | X | X | X | X | X |
| User's VLANs | X | | X | X | X |
| User's identity | X | X | | | X |
| IPv4 to MAC binding | | | X | X | X |
| *NetLogin* authentication protocol | X | | | | |
| Authentication failures | X | | | | |
| Device capabilities[a] | | X | | | |
| Device model name[a] | | X | | | |
| Device manufacturer name[a] | | X | | | |

a. Identity manager receives these attributes only from LLDP enabled ports when the remote device is configured to send the corresponding TLV.

The software components in the table above trigger identity attribute collection when a user or device connects to the switch. All components provide the MAC address, authentication and unauthentication time stamps, and the port to which the identity connected. When multiple components are triggered by a user or device connection, the triggers usually happen at different times. Identity manager responds to all identity event triggers, adding additional information to the identity database each time it becomes available.

To capture all the available attribute information listed in the following table, enable the following features:

- Network Login
- LLDP
- IP Security

By default, the identity management feature collects information from all devices connected to identity management enabled ports which does Kerberos authentication using Kerberos snooping. Kerberos authentication, or ticketing, is used by Microsoft's Active Directory. The Kerberos snooping feature collects identity attributes from Kerberos Version 5 traffic. This feature does not capture information from earlier versions of Kerberos.

Note

We recommend that you enable CPU DoS protect in combination with Kerberos snooping to make sure the CPU is not flooded with mirrored Kerberos packets in the event of a DoS attack on Kerberos TCP/UDP ports. If the rate limiting capability is leveraged on capable platforms, it is applied on CPU mirrored packets.

Because an identity entry in the identity manager database can contain information from various software components (listed in the table above), when a component other than a network login triggers an identity removal, only the attributes supplied by that component are removed from the identity. When network login triggers an identity removal, all attributes for that identity are removed from the identity manager database.

## Identity Names

After identity attributes are captured, they can be viewed with show commands on the switch. The identity ID Name assigned to each identity depends on the identity attributes collected. For example, if a MAC address detected by *FDB* is not correlated by at least one other software component, the identity is considered an unknown identity, and identity manager creates an identity entry with the name unknown_<MAC-Address>, where MAC-Address is replaced with the actual MAC address.

When an FDB detected MAC address is correlated by another software component, the identity is considered a known identity, and the identity manager names the identity based on the identity attributes.

For example, if a user name is collected, the user name becomes the ID name. If a username is not discovered, identity manager creates a name based on the MAC address.

Identity manager can change the ID name when additional attributes are learned, or when the identity status changes between known and unknown. For example, if *LLDP* sends an identity removal trigger to the identity manager for an LLDP-based identity, and if a valid FDB entry exists for the removed identity, the identity manager reestablishes the identity as an unknown identity (unknown_<MAC-Address>).

Note

If FDB triggers the removal of the MAC address for an unknown identity, the identity manager deletes the corresponding unknown identity after a period of time.

## Application of ACLs and Policies for Identities

Each time the identity manager detects a new identity or an identity change, it evaluates the identity attributes to determine which role to apply to the identity. A role is a switch configuration entity that identifies ACLs to apply to a port in response to an identity presence.

### *How Roles Affect Ports*

A role is a configuration entity to which you can add multiple policy files or dynamic *ACL (Access Control List)* rules. When an identity is matched to a role, any policies or rules attached to that role are applied to the port to which the identity connected. These rules or policies permit or deny traffic, increment traffic counters, or implement traffic meters. When identity manager detects a removal trigger for an identity, all rules or policies associated with the identity are removed from the port on which the identity was detected.

### *Authenticated and Unauthenticated Roles*

The identity management feature supports two default roles—authenticated and unauthenticated. No default rules or policies are configured for these roles, but you can add rules or policies to these roles.

Authenticated identities are known identities that meet the following requirements:

- Are not included in the blacklist or whitelist.
- Do not meet the match criteria for any user-defined roles.
- Cannot meet the match criteria for any user-defined role with LDAP attributes because no LDAP server is available or because LDAP queries are disabled.
- Are detected either through network login (using any of the network login methods) or through Kerberos snooping.

The unauthenticated role applies to all identities that do not match any other default or user-defined role.

For example, the following identities are placed in the unauthenticated role:

- A device detected by *LLDP* that has not authenticated through network login and does not match any other default or user-defined role.
- A user who does not successfully log in using Kerberos login and does not match any other default or user-defined role.
- A device discovered through IP ARP or *DHCP (Dynamic Host Configuration Protocol)* snooping that does not match any other default or user-defined role.
- Any identity classified as an unknown identity.

**Note**

The unauthenticated role is not applied to network login and Kerberos users because those users are either authenticated or denied by network login.

One option for configuring the unauthenticated role policy/rule is to allow DNS, DHCP, and Kerberos traffic, and deny all other traffic. This configuration allows identities to attempt log in, and denies access to identities that do not successfully log in.

*Blacklist and Whitelist Roles*

Blacklist and whitelist roles are special roles that are evaluated before all the other role types. If an identity is listed in a blacklist, that identity is denied all access to the network without consideration of any other roles to which it might belong. Similarly, if a discovered identity is found in the whitelist, that identity is granted complete network access, and no further role processing occurs for that identity.

You can configure identities in a blacklist or whitelist using any one the following identity attributes:

- MAC address
- IPv4 address
- Username (with or without a domain name)

The type of identity attribute specified in a blacklist or whitelist impacts the locations from which an identity can access a switch. For example, if a MAC address or an IP address is specified in a blacklist, no access is permitted from any user at devices with the specified address. If a username is specified in a whitelist, that user is permitted access from all locations.

When an identity accesses the switch and that identity is in a blacklist or whitelist, the switch installs a specific deny or allow *ACL* on the port through which the identity attempts access. The installed ACL is an active ACL that explicitly denies or allows traffic from that identity. There is no passive action that takes place if the identity is not listed in the ACL. When the identity is not listed in a blacklist or whitelist, the switch checks for matches to other roles as described in Role Precedence and Priority on page 840.

*Greylist Roles*

Greylist feature enables the network administrator to choose usernames whose identity is not required to be maintained. When these usernames are added to greylist, the Identity Management module does not create an identity when these users log on.

This will be useful in a scenario wherein multiple users log in from same device at the same time. For example, actual user has logged into computer after Kerberos authentication. Later, Anti-Virus Agent (AVAgent) software starts within the same computer and does Kerberos authentication.

This will result in losing actual user identity and creating identity for AVAgent. Configuring AVAgent's username in greylist will prevent the above situation and actual user identity along with policies will be retained when AVAgent user logs in.

**List Precedence Configuration**

Greylist entries have higher precedence over blacklist and whitelist entries by default. This means that IDM consults with greylist first, upon detection of user, and then decides if the identity needs to be created. If there is no matching greylist entry, IDM proceeds with role identification for the user. However, greylist precedence is configurable. The following are three possibilities for greylist precedence configuration:

- greylist, blacklist, whitelist
- blacklist, greylist, whitelist
- blacklist, whitelist, greylist

At this time, blacklist always has precendence over whitelist. To change list precedence, disable IDM first. Disabling IDM is required since reverting roles and revoking policies due to greylist entries may increase processing load. When precedence configuration is changed, each entry present in the list with

lower precedence (new precedence) is checked with each entry present in all the lists with higher precedence. If any existing entry becomes ineffective, details of those entries are displayed at the CLI prompt.

*User-Defined Roles*

User-defined roles allow you to create custom roles that can restrict, count, and meter traffic for identities you want to control. CLI commands allow you to do the following:

- Create a user defined role.
- Configure identity match criteria that determine which identities use a role.
- Add dynamic *ACL* rules or policies to a role so that those policies are applied to ports to which a matching identity connects.
- Assign a priority level to each role to determine which role applies when multiple roles are matched to an identity.
- Establish hierarchical roles that can be used to support topologies built around a company organization structure or a geographical layout.

When specifying match criteria for a role, you can specify identity attributes collected by identity manager (see Identity Information Capture on page 832) and those collected from an LDAP server. When configured for an LDAP server, identity manager can send a query to the server with locally collected attributes and retrieve additional attributes for the identity, such as an employee department or title. The use of an LDAP server allows you to design roles that serve departments or localities.

**Identity Attributes on an LDAP Server**

When identity manager is configured to connect to an LDAP server, identity manager can query the server for the identity attributes listed in the following table.

**Table 105: LDAP Attributes for Role Selection**

| Attribute | Active Directory LDAP Attribute | Attributes Allowed in Identity Manager Match Criteria |
|---|---|---|
| City | l | l or location |
| State | st | st or state |
| Country | co | co or country |
| Employee ID | employeeID | employeeID |
| Title | title | title |
| Department | department | department |
| Company | company | company |
| Email Address | mail | mail or email |

An LDAP query contains one or more of the identity attributes listed in Table 104 on page 832.

If an LDAP server fails to respond, the next configured LDAP server is contacted. When a server query succeeds, all further LDAP queries are sent to that LDAP server. All LDAP servers should be configured to synchronize the user information available in each of them.

Note

Identity manager supports a maximum of eight LDAP servers.

#### Match Criteria for Selecting User-Defined Roles

When you create a user-defined role, you must define the match criteria that determines which identities will use the role. The match criteria is a group of identity attributes and the attribute values that must match identities to which this role is assigned. For example, the following command creates a role named US-Engr that matches employees whose title is Engineer and who work in United States:

```
* Switch.23 # create identity-management role US-Engr match-criteria "title contains
Engineer; AND country == US;" priority 100
```

The match criteria are a series of attributes, operators, and attribute values, all of which are described in the *ExtremeXOS 16.2 Command Reference Guide*. Each role can define up to 16 attribute requirements, and there are additional operators such as not equal. Beginning in ExtremeXOS 15.3, the match criteria attributes are combined using the AND or OR operators, not a combination of both. When multiple roles are matched to an identity, the role with the highest priority (lowest numerical value) applies.

In the preceding example, identity manager must be configured to query an LDAP server because the identity attributes listed in the match criteria are not discovered locally.

The match criteria for a role establish the role as on of two types:

- Local user-defined role
- LDAP user-defined role

A local user-defined role uses only the following locally discovered attributes (which are listed in the following table) in the match criteria:

- User's MAC address
- MAC OUI
- User's port
- User's identity
- IPv4-to-MAC binding
- Device capabilities
- Device model name
- Device manufacturer name

Because a local user-defined role does not require LDAP provided attributes, the role can be matched to identities when an LDAP server is unavailable, or when LDAP processing is disabled for network login authenticated identities. A local user-defined role can serve as a backup role to an LDAP user-defined role.

An LDAP user-defined role uses one or more of the LDAP attributes listed in Identity Attributes on an LDAP Server on page 836 in the match criteria, and it can also use the attributes listed in Identity Information Capture on page 832. An LDAP user-defined role gives you more flexibility in selecting

attributes for the match criteria. However, if no LDAP server is available, and the identity attributes do not match a local user-defined role, one of the two default roles is applied to the identity.

Role Policy Order

Previously, the policy or dynamic rule associated to the role occurred in the order of configuration. There was no way for you to change the order of the policy or dynamic rule associated with the role. ExtremeXOS 15.2 supported the ability to change the order of the policy or dynamic rule associated with the role. You can also change the order of the policy or dynamic rule during the run time. Even if the role is assigned to some identities, the policy or the dynamic rule associated to the role can be changed.

Role Hierarchy

To make it easier to manage users, the role management feature supports hierarchical roles. Hierarchical roles can be defined to reflect different organizational and functional structures. To create a role hierarchy, one or more roles are defined as child roles of what becomes a parent role. Each role can have a maximum of eight child roles and only one parent role. This feature can support up to five levels of parent and child roles. With respect to role hierarchy and match criteria, there is no restriction across roles. Beginning in 15.3 release, a user can have the parent role with AND, and the child role with OR, or vice versa. The inheritance of match criteria to the child role from the parent role uses AND as in previous releases.

Role Inheritance

Child roles inherit the policies of the parent role and all roles above the parent in the hierarchy. When an identity is assigned to a role, the policies and rules defined by that role and all higher roles in the hierarchy are applied.

When the parent role is deleted or when the parent-child relationship is deleted, the child role no longer inherits the parent role's policies and the policies are immediately removed from all identities mapped to the child role.

Because the maximum role hierarchy depth allowed is five levels, the maximum number of policies and dynamic ACLs that can be applied to a role is 40 (five role levels x eight policies/rules per role). The figure below shows an example of hierarchical role management.

Match Criteria Inheritance

Beginning in release 15.2, the child role can inherit the match criteria of the parent role.

This means that the match criteria does not need to be duplicated in all levels of hierarchy.

For example, if you have roles called Employee, India employee, and India engineer in a hierarchy, previously the match criteria of the three roles would have been:

```
“company == Extreme”
“company == Extreme; AND country == India”
“company == Extreme; AND country == India; AND department = Engineering”
```

This can be simplified into the following since the child role automatically inherits the parent role’s match criteria:

```
“company == Extreme”
“country == India”
“department = Engineering”
```

Once this support is enabled, user identity must satisfy not only the role's match criteria, but its parent and ancestors also. This support can be enabled/disabled using CLI or XML. You no longer have to repeat the match criteria configured in the parent role in the child roles also.

Note

- Role match criteria inheritance can only be enabled if all of the existing roles have higher priority than their descendants. If this condition is not satisfied, match criteria inheritance will fail.
- Once this feature is enabled, you cannot configure a child role with lesser priority (higher priority number) than its parent.
- Enabling this feature changes the order of the roles according to the parent-child relationship.
- Incoming identities are matched against the child role and then against the parent irrespective of the order of creation.

For example, Role A and Role B have match criteria MC-A and MC-B, respectively. Role B is a child role of Role A. When match criteria inheritance is disabled, an identity matches Role B criteria, and then it is placed under Role B with no further check.

When match criteria inheritance is enabled, the same identity, after satisfying Role B's match criteria, is then checked against Role A's match criteria. Once the identity satisfies child and parent match criteria, it is placed under Role B.

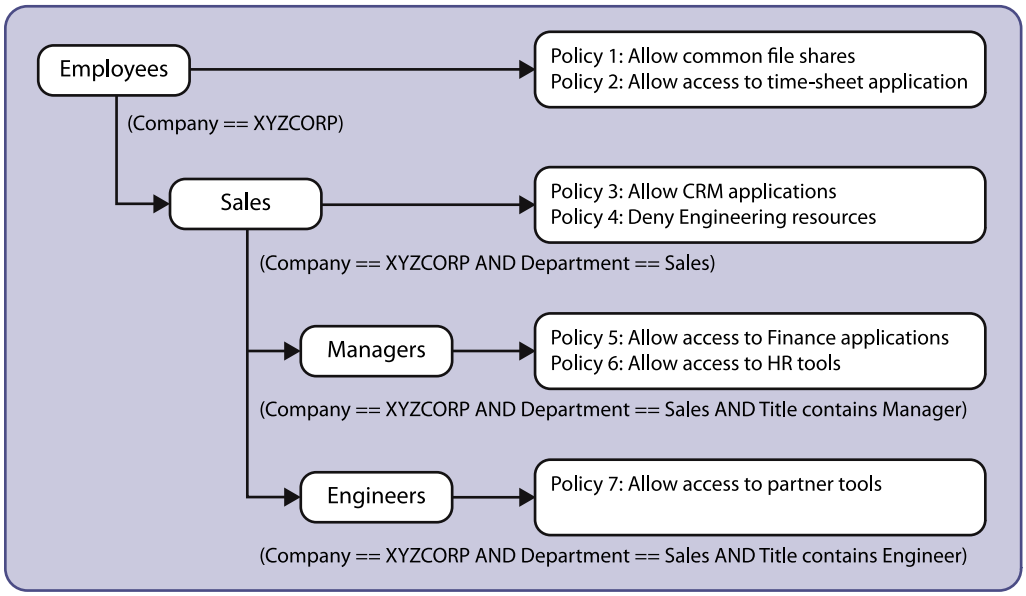



**Figure 104: Hierarchical Role Management Example**

Support for OR in Match Criteria

Prior to release 15.3, only AND was supported in the match criteria of user roles. From 15.3, OR is now supported also, so the user can have either AND or OR in the match criteria, but not both. For a particular role, the user can have all match criteria with AND, or have all the match criteria with OR. There is no restriction across roles with respect to role hierarchy and match criteria inheritance. The user

can have the parent role with AND, and the child role with OR, or vice versa. The inheritance of match criteria to the child role from the parent uses AND as in previous releases.

```
Examples:
create identity-management role "EniBuildservers" match-criteria match-criteria "userName
contains
enibuild; OR ip-address == 10.120.89.0/24;
```

This example creates a role for enibuild servers, whose name contain enibuild or whose ip-addresses are in the range 10.120.89.1 - 10.120.89.255.

```
If the parent role has the match criteria as
"company == Extreme" AND "Title == Manager;"
And the child role has the match critera as
"country == India;" OR "country == USA"
And the grandchild has the match criteria as
"department == Engineering"
```

All the managers who belong to Extreme Engineering, from both India and the USA, will be placed in the grandchild role.

Context Based Roles

Context based roles apply additional rules or policies to a port based on context related attributes for an identity. For example, consider a campus environment where a student logs into the network through a PC and also through a smart phone. Suppose that a role named Student already exists and applies basic policies for a student. Also suppose that the administrator wants to apply additional policies for students accessing the network through smart phones.

To apply the additional policies, the administrator can create a role called Student_smartPhone as a child role of Student. The match criteria could include "title == Student; AND mac-oui == 00:1b:63:00:00:00/FF:FF:FF:00:00:00;", where the MAC address is the address of the smart phone. The additional policies can be added to the new child role. When the student logs in from the PC, the rules applicable to the Student role apply. When the student logs in from the smart phone, the policies for the Student_smartPhone role apply, in addition to those inherited from the Student role.

Note

A student logging on through a smart phone is placed under the Student_smartPhone role only if that role has a higher priority (lower numerical value) than the Student role.

*Role Precedence and Priority*

Roles are evaluated for identities in the following sequence:

1. The blacklist role is searched for the identity. If the identity is in the blacklist, the identity is denied access and role evaluation stops.
2. The whitelist role is searched for the identity. If the identity is in the whitelist, the identity is allowed access and role evaluation stops.
3. A local user-defined role is searched for the identity. If the identity is mapped to a local user-defined role, the identity is allowed access and role evaluation stops for all unknown/*LLDP* users. For Kerberos and network login users (except those authenticated through the local network login database), a query is sent to an LDAP server with the user attributes. If the Kerberos and network

login users (except those authenticated through the local network login database) do not map to any local user-defined role , the identity is placed in authenticated role.

Note
The LDAP query can be disabled for specific types of network login users, and the LDAP query is disabled for locally authenticated network login identities.

4. When the switch receives LDAP attributes for an identity, the software evaluates the user-defined roles. If one or more user-defined roles match the identity attributes, and if those roles have a higher priority (lower numerical value) than the current role applied to the identity, the policies for the current role are removed and the policies for the user-defined role with the highest priority are applied.

Note
To support a change from the one role to another, the role priority for the new role must be higher than the current role.

5. Authenticated identities that cannot be placed in a user-defined role remain assigned to the authenticated role.
6. The unauthenticated role is applied to all identities that do not match any other roles.

### *Application of Rules or Policies*

When the software makes the final determination of which default or user-configured role applies to the identity, the policies and rules configured for that role are applied to the port to which the identity connected. This feature supports up to eight policies and dynamic *ACL* rules per role (eight total).

When a dynamic ACL or policy is added to a role, it is immediately installed for all identities mapped to that role. Effective configuration of the dynamic ACLs and policies will ensure that intruders are avoided at the port of entry on the edge switch, thereby reducing noise in the network.

Note
The identity management feature supports wide key ACLs, which allow you to create many more match conditions for each ACL. For more information, see Wide Key ACLs on page 662.

The dynamic rules or policies that are installed for an identity, as determined by its role, are customized for that identity by inserting the MAC or IP address of the identity as the source address in the ACL rules. In ExtremeXOS release 12.5, identity manager inserted the IP address of the identity in all the ACL rules to be installed for that identity. Beginning with release 12.6, identity manager can insert either the MAC address or the IP address of the identity in all the ACL rules to be installed for that identity. By default, the MAC address of the identity is used to install the ACLs. Every network entity has a MAC address, but not all network devices have an IP address, so we recommend that you use the default configuration to install ACLs for network entities based on the source MAC address.

For additional information on creating ACLs, see ACLs on page 640. For additional information on creating policies, see Policy Manager on page 635.

### *Role-Based Policy Enforcement*

After user information is retrieved from the directory server, it is matched against a configured set of criteria and the user is then assigned to a specific role.

Pre-defined roles contain details of attributes with corresponding values to be used as match-criteria and the policies that need to be applied for that role. The administrator will be provided with a set of CLI commands to map association between role, match-criteria, and policies.

The following is a list of LDAP attributes that can be looked up in the LDAP server:

- Employee/User ID
- Title
- Department
- Company
- City
- State
- Country
- Email ID

#### Association Between Role and Attribute

Using CLI, various roles can be created with corresponding match criteria specified in attributes and values.

#### Association Between Role and Policy

When a policy is added to a role, the newly added policy will be applied to both existing users mapped to that role as well as new users who get mapped to this role in the future.

#### Match Criteria Inheritance

Beginning in release 15.2, a child role can inherit the match criteria of the parent role. The match criteria now does not need to be duplicated in all levels of the hierarchy.

### *Role-Based VLAN*

Available in EXOS 15.4 and beyond, when an identity is detected and the role is determined, EXOS dynamically creates the *VLAN (Virtual LAN)* that is required for the identity to send traffic. If the identity is deleted, aged out, or moved, its VLAN is removed to preserve bandwidth. MVRP is leveraged by this feature to add uplink ports to the dynamically created VLAN.

#### Enable and Disable Identity Management Role-Based VLAN

Enabling this feature in EXOS must be done on a per-port basis. Identity management (IDM) requires that the port on which role-based *VLAN* is enabled be part of a “default” or “base” (not necessarily the “Default” VLAN) VLAN as untagged. This “default” or “base” VLAN for the port is the VLAN on which untagged packets are classified to when no VLAN configuration is available for the MAC. This default VLAN should be present before enabling the feature and the port should have already been added to this VLAN by the user manually before enabling the feature.

Enabling this feature on a port results in a failure if any of the following conditions are true:

- IDM is not enabled globally.
- IDM is not enabled on the port.
- The port is not an untagged member of any VLAN.

When an identity's MAC address is detected on a port, identity management consults its configuration database to determine the VLAN configuration for the role to which this identity is placed under. When

the identity is sending tagged traffic it will work as in previous releases. Role based VLAN for tagged traffic is not supported in this release. If no configuration is present for the identity's role, IDM assumes that there are no restrictions for traffic classification and the traffic is classified to the default/base VLAN (received VLAN). In addition to the VLAN tag, you can specify the VR to which the dynamically created VLAN needs to be associated. The VR configuration is relevant only if a VLAN tag is configured for the role.

The following table specifies the VR configuration:

**Table 106: Identity Management Role-Based VLAN**

| Configured VR on Port | Configured VR for Role | VLAN already exists on the switch | Role-based Dynamic VLAN's VR |
|---|---|---|---|
| None | None | No | *VR-Default* |
| None | None | Yes | VLAN's VR if it is Default Else *EMS* error |
| None | VR-X | No | VR-X |
| None | VR-X | Yes | VLAN's VR if it is VR-X (Role's VR) Else EMS error |
| VR-X | None | No | EMS error |
| VR-X | None | Yes | EMS error |
| VR-X | VR-Y | No | EMS error |
| VR-X | VR-Y | Yes | EMS error |

When you disable role based VLAN on a port, identity management does the following:

1. Triggers deletion of MAC-based entries in that port in the hardware.
2. If the port has been added to any VLAN by identity management, identity management triggers deletion of MAC-based entries on that port in the hardware..
3. If the port has been added to any VLAN by IDM, IDM requests VLAN manager to remove the port from the VLAN. (Note: It is up to VLAN Manager to decide if the port actually needs to be removed from the VLAN).

When IDM is disabled on a port, the IDM based VLAN feature is also operationally disabled. However IDM role based VLAN configuration is persistent and will come into effect once IDM is re-enabled on that port.

#### MAC Learning and Provisioning of VLAN

The first step in determining *VLAN* configuration for an identity is to learn the identity's MAC. For untagged traffic the port is added as untagged to a "catcher/learning" VLAN that is used to learn MACs. Identity Management (IDM) role based VLAN is not supported for tagged traffic.

Upon receiving the first packet from the identity, the following actions are completed:

1. *FDB* Manager learns the identity's MAC and informs IDM.
2. IDM creates an identity for the newly learned MAC and determines the role for the identity.
3. IDM checks the role's configuration to see if the identities in this role need to be associated with a VLAN.

4. If the identity in this role is associated with a VLAN tag, IDM checks to see if a VLAN with the configured tag is already present.
5. If not, IDM creates VLAN “SYS_VLAN_<Configured-Role-VLAN-Tag>” and adds the port (on which the identity is detected) to VLAN “SYS_VLAN_< Configured-Role-VLAN-Tag>” as untagged. If a VLAN with configured tag already exists, IDM simply adds the port to the VLAN as untagged.
6. In addition, IDM adds a MAC entry for identity’s MAC in the hardware to classify all untagged traffic from this identity to be associated with VLAN “SYS_VLAN_<Configured-Role-VLAN-Tag>”.
7. IDM does not explicitly add uplink ports to VLAN “SYS_VLAN_<Configured-Role-VLAN-Tag>”. It is assumed that user would have enabled MVRP on the uplink ports or the uplink ports are configured statically. Creation of the VLAN is sufficient for MVRP to advertise membership for VLAN “SYS_VLAN_<Configured-Role-VLAN-Tag>” over those ports.
8. If no VLAN configuration exists for Role, IDM adds a MAC entry to associate identity’s MAC with the default/base VLAN configured for the port.

Note

All of the IDM enabled ports should be part of a default/base VLAN to enable IDM role based VLAN on the port.

Tagged Traffic from Identity

Note

This section assumes that the IDM enabled port and the uplink ports are already added to the VLAN as tagged.

1. FDB Manager learns the identity’s MAC and informs IDM.
2. IDM creates an identity for the newly learned MAC and determines the role for the identity.
3. IDM checks the role’s configuration to see if the identities in this role need to be associated with a VLAN.
4. If the identity in this role is associated with a VLAN tag, IDM checks to see if a VLAN with configured tag is already present.
5. IDM also checks if the role configured tag matches the incoming VLAN tag of the identity. If not, an *EMS* error is generated.

Untagged Traffic from Identity

The following figure shows a topology of untagged traffic from an identity:

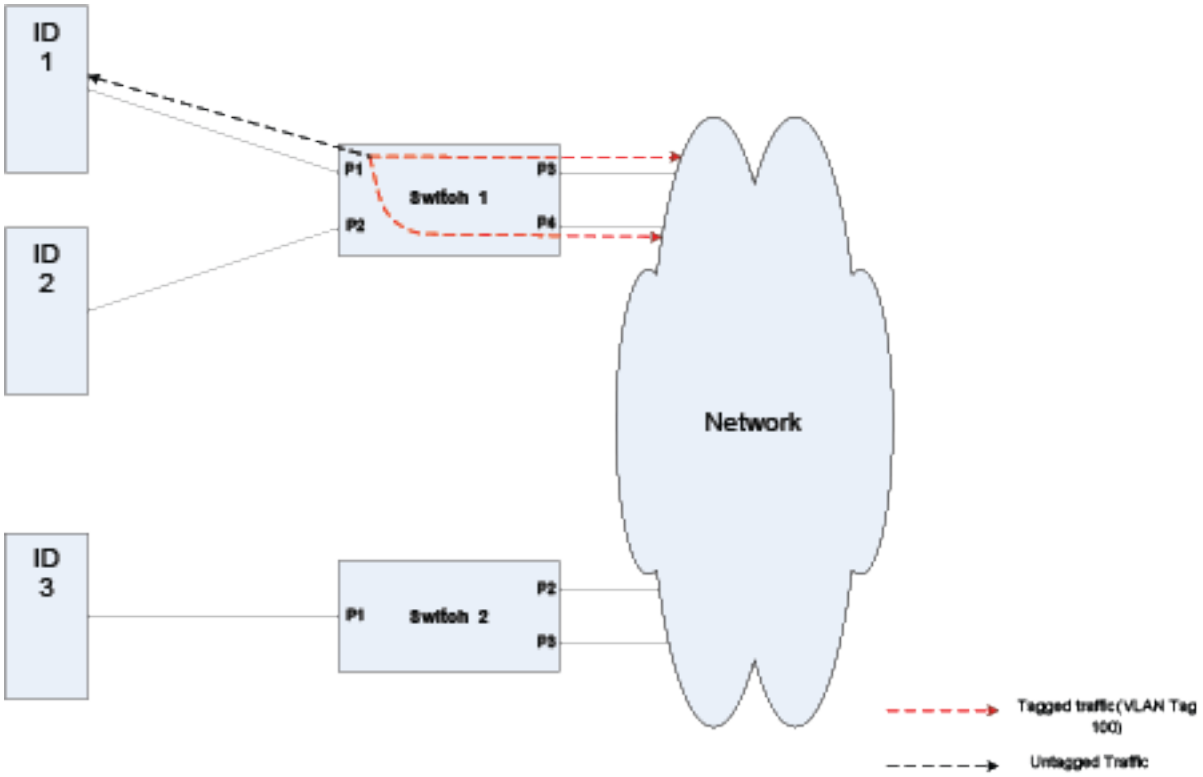



**Figure 105: Untagged Traffic Topology**

- FDB Manager learns the identity's MAC on Switch1's port P1 and informs IDM.
- IDM creates an identity for this MAC and determine the role for this new identity. IDM checks Role configuration to see if the identities in this role is associated with a VLAN.
- If the identity in this role is associated with a VLAN tag (say VLAN ID 100), IDM checks to see if a VLAN with tag 100 is already present. [If VLAN is already present the assumption is the user has already added the uplink port to the VLAN].
- If not IDM will create VLAN "SYS_VLAN_100" on Switch 1 and adds port P1 to VLAN "SYS_VLAN_100" as untagged. If a VLAN with tag 100 already exists, IDM simply adds the port to the VLAN as untagged.
- In addition IDM will add a MAC entry for identity's MAC in H/W to classify all untagged traffic from this identity to be associated with VLAN "SYS_VLAN_100".
- IDM does not explicitly add uplink ports (ports P3 & P4) in this case to VLAN "SYS_VLAN_100". It is assumed that user would have enabled MVRP on the uplink ports or the uplink ports are configured statically. Creation of the VLAN is sufficient for MVRP to advertise membership for VLAN "SYS_VLAN_100" over those ports.
- If no VLAN configuration exists for Role, IDM adds a MAC entry to associate identity's MAC with the default/base VLAN configured for the port.

### *Group Attributes Support*

Network users can be mapped to a role based on group membership (distribution list) information. When a user is detected by identity manager, it retrieves the groups in which the detected user is member of from the LDAP server. Identity manager places the user under the appropriate role, based on group information and existing eight LDAP attributes.

You can specify the group name in the role's match criteria while creating the role. For example, the role creation command will appear as follows:

```
1  Create identity-management role Role1 match-criteria "memberOf==EXOSCLI-Review"
2  Create identity-management role Role2 match-criteria "title==Engineer; AND
memberOf==PI_SW"
```

A role's match criteria accepts all of the following operators: ==, !=, contains, AND, and OR.

Note

A combination of AND and OR is not supported in the match criteria definition of the role.

You can specify groups of the following types in match-criteria:

- direct-membership: the user is a direct member of the group specified in role match-criteria.
- hierarchical-membership: the user is not a direct member of the group specified, but comes under a specified group, per the hierarchy of the Active Directory (i.e., nested groups). Hierarchical groups are supported in Windows Server 2003 and later. Some LDAP servers may require special OID to perform a hierarchical search.

When a user is detected by identity manager, the following things occur:

- Identity manager retrieves eight LDAP attributes as supported before the 15.3 release, and also the Distinguished Name of the user.
- If any role's match criteria contains group attribute, a second LDAP query is created using the Distinguished Name of the user to retrieve all of the groups that the user is a member of. If an OID is configured for the hierarchical search, it will be used to form this LDAP query.
- Role determination takes place based on the group membership information and other LDAP attribute values.

The following optimizations are completed with respect to the LDAP query for Group Attributes:

- All of the group names used in every role configuration are collected and stored in a global database. When the LDAP query returns a list of the user's groups, the group names are cached against the user and used for role determination. The optimization is that only the group names used for role configuration are cached. The rest of the group names are discarded.
- The second LDAP query is not sent if the group attribute (i.e., memberOf) is not used in any role.

### *Network Zone Support for Policy Files*

Network zone support helps users create a network zone and add multiple IP address and/or MAC addresses, which can then be used in the policy files.

With this new support, you can add a single attribute “source-zone” or “destination-zone” to an entry of a policy file. This entry is then expanded into multiple entries depending upon the number of IP and/or MAC addresses configured in that particular zone. If the zone is added to the policy with the keyword “source-zone,” the attributes that are configured in those particular zones, will be added as either source-address or ethernet-source-address in the policy, whereas, if the network-zone is added as destination-zone, the attributes will be added to the policies as destination-address or ethernet-destination-address.

Once you complete the changes in the zones, and issue a refresh of a specific zone, or all the zones, the policies that have corresponding zones configured as source-zone or destination-zone in their entries will be exapnded, and then refreshed in the hardware.

If you configure a policy such as the following to a port or *VLAN* through applications such as IdMgr, *Extreme Network Virtualization (XNV)* or CLI:

```
Policy: test
entry e1 {
if match all {
    source-zone zone1 ;
}
Then {
    permit  ;
}
}
```

Upon refreshing the network-zone zone1, the policy will be expanded as below:

```
entry Cl0:0_10.1.1.1_E1 {
  if match all {
    source-address 10.1.1.1 / 32 ;
  } then {
    permit  ;
  }
}
entry Cl0:0_10.1.1.1_E2 {
  if match all {
    source-address 10.1.1.1 / 28 ;
  } then {
    permit  ;
  }
}

entry Cl0:0_12.1.1.0_E3 {
  if match all {
    source-address 12.1.1.0 / 24 ;
  } then {
    permit  ;
  }
}
```

This converted policy will be the one to be applied in the hardware.

Note

When the policy is configured in the network-zone, the zone may or may not have attributes configured with it. In cases where the network-zone does not have any attributes, the policy will be created with entries that do not have the network-zone attribute alone. For example, if you have a policy similar to the following:

```
Policy: test2
entry e1 {
if match all {
    source-zone zone2 ;
    protocol udp ;
}
then {
    permit  ;
}
}
entry e2 {
if match all {
```

```
    protocol tcp ;
}
then {
    permit  ;
}
}
```

and the network-zone “zone02” is created but not configured with any attributes, the policy would be as follows:

```
entry e2 {
protocol tcp;
}
then {
    permit;
}
}
```

Once the network-zone “zone2” is configured with one or more attributes and refreshed, the policy will be updated accordingly. Here the name of the entries that have source-zone or destination-zone will be changed. This is because each entry in the original policy that has a source-zone/destination-zone will be converted to a maximum of eight entries in the new policy.

A single policy can have one or more network-zones configured in it, and can also have the same network-zone in multiple entries with different combinations as other attributes are supported in the policy file. Similarly, the same network-zone can be configured to multiple policies. In cases where the policy has multiple network-zones, and only some of those network-zones are refreshed, the entries that correspond to those network-zones will be refreshed alone, and not the entries that correspond to the other network-zones.

Once a refresh of a network zone is issued, all the policies that have the specified network-zone will be modified, and a refresh for each of those policies will be sent to the hardware. The command will succeed only after getting a success return for all the policies that have this particular network-zone. If one of the policy’s refresh fails in the hardware, all of the policies that are supposed to refresh will revert to their previous successful state and the command will be rejected.

The configuration or refresh may fail if the attributes in the network zone are not compatible with the attributes in the corresponding entries of the policy. For example, in platforms that do not support wide-key or UDF, a policy entry cannot have Layer 2 attributes and Layer 4 attributes. In this case, if the entry has “protocol tcp,” and a network_zone that has an ethernet source address, the configuration will fail in the hardware.

Note

In the refresh failed case, the content of the policy and the content of the network-zone may go out of sync, as the policy will be reverted back to the last successful state, whereas the network zone will contain the last configured values.

For example, if we have the the network-zone configuration as follows:

```
create access-list network-zone zone1
configure access-list network-zone zone1 add ipaddress 10.1.1.1/32
configure access-list network-zone zone1 add ipaddress 10.1.1.1/28
```

and this is refreshed, and has been successfully installed in the hardware, the policy will look like this:

```
entry Cl0:0_10.1.1.1_E1 {
  if match all {
    source-address 10.1.1.1 / 32 ;
  } then {
    permit  ;
  }
}

entry Cl0:0_10.1.1.1_E2 {
  if match all {
    source-address 10.1.1.1 / 28 ;
  } then {
    permit  ;
  }
}
```

Now, if the user removes 10.1.1.1/28, and adds 10.1.1.1/24 to the network zone as below:

```
configure access-list network-zone zone1 delete ipaddress 10.1.1.1/28
configure access-list network-zone zone1 add ipaddress 12.1.1.0/24
```

and then does a refresh network-zone, and for some reason, the policy refresh fails, the policy and the network-zone will look as below:

```
entry Cl0:0_10.1.1.1_E1 {
  if match all {
    source-address 10.1.1.1 / 32 ;
  } then {
    permit  ;
  }
}

entry Cl0:0_10.1.1.1_E2 {
  if match all {
    source-address 10.1.1.1 / 28 ;
  } then {
    permit  ;
  }
}
create access-list network-zone zone1
configure access-list network-zone zone1 add ipaddress 10.1.1.1 255.255.255.255
configure access-list network-zone zone1 add ipaddress 12.1.1.0 255.255.255.0
```

*Role Refresh*

Role refresh allows you to enter a CLI command that triggers a reevaluation of role selection for one or all users. A role refresh can also trigger reevaluation of role selection for all users using a specific role.

After role evaluation completes for an identity, the role remains the same as long as the identity is present at the original location and no new high priority role matching this identity's attributes is created. Consider a situation where a Kerberos user is always present at a particular location. The switch detects traffic to and from the user periodically, so the user identity is never aged out. The user's role at this location remains the same as the role determined by identity manager when the user was detected at this location for the first time.

A network administrator might want to refresh a role for the following reasons:

- The user's LDAP attributes have changed. For example, the user's job title is changed from Engineer to Manager or his department is changed from Engineering to Marketing.
- The administrator has created a new role, which is more applicable to the user than his previous role. For example, the user was initially placed under the Engineer role because his department was Engineering, and now a new role called Test Engineer is a better match that considers both the user's department and title.

For both of the above situations, a role refresh triggers a role evaluation that would not otherwise occur as long as the user remains active at the current location. If the role refresh finds an LDAP user-defined role that matches the identity being refreshed, the identity manager queries the LDAP server to update the attributes provided by the LDAP server.

### Switch Configuration Changes in Response to Identity Management Events

You can configure automatic switch configuration changes in response to identity management events. To do this, configure UPM profiles to respond to identity management events as described in Event Management System Triggers on page 315.

Identity management events generate corresponding UPM events, including:

- IDENTITY-DETECT
- IDENTITY-UNDETECT
- IDENTITY-ROLE-ASSOCIATE
- IDENTITY-ROLE-DISSOCIATE

For instructions on displaying a complete identity management event list, see Event Management System Triggers on page 315. The component name for identity management events is IdMgr.

## Identity Management Feature Limitations

In the current release, the identity management feature has the following limitations:

- IPv4 support only. IPv6 to MAC bindings are not captured.
- For Kerberos snooping, clients must have a direct Layer 2 connection to the switch; that is, the connection must not cross a Layer 3 boundary. If the connection does cross a Layer 3 boundary, the gateway's MAC address gets associated with the identity.
- Kerberos snooping does not work on fragmented IPv4 packets.
- Kerberos identities are not detected when both server and client ports are added to identity management.
- Kerberos does not have a logout mechanism, so mapped identities are valid for the time period defined by the Kerberos aging timer or the Force aging timer.
- Kerberos snooping applied ACLs can conflict with other ACLs in the system. The identity management feature registers itself in the user space SYSTEM zone; for details, see .

# Configuring Identity Management

Note

When a switch is managed by NMS, it is possible to modify/delete configurations that are created by NMS. After editing via CLI, these changes are not reflected to NMS. Therefore, CLI configurations will be lost when NMS modifies the configuration further, because NMS configuration overrides CLI configuration. The following warning is printed so that users are aware that the CLI change is not permanent:

```
WARNING: An object that was created by Network Management System(NMS) has been
modified. The modification will not be reflected to the NMS and will not be
preserved if the object is subsequently modified by the NMS.
```

Note

The Identity Manager role-based *VLAN* feature will not be enabled on Netlogin enabled ports.

## Basic Identity Management Feature Configuration

The following sections describe basic identity management configuration tasks.

### *Configuring the Maximum Database Size*

- To configure the maximum size for the identity management database, use the following command:
  `configure identity-management database memory-size Kbytes`
- To set the maximum database size to the default value, use the following command:
  `unconfigure identity-management database memory-size`

### *Selecting the Access-List Source-Address Type*

The identity management feature can install *ACL*s for identities based on the source MAC or source IP address. By default the MAC address of the identity is used to install the ACLs. Every network entity has a MAC address, but not all network devices have an IP address, so we recommend that you use the default mac selection to install ACLs for network entities based on the source MAC address.

- To change the configuration for the access-list source-address type, use the following command:
  `configure identity-management access-list source-address [mac | ip]`

  Note

  You must disable identity management to change the current access-list source-address type configuration.

  By default, the identity's MAC address is used for applying the dynamic ACLs and policies. The dynamic ACLs or policies that are associated to roles should not have any source MAC address specified because the identity management feature will dynamically insert the identity's MAC address as the source MAC address. Similarly, if the ACL source address type is configured as ip, the dynamic ACLs or policies that are associated to roles should not have any source IP address specified.

### *Enabling and Disabling Identity Management*

- To enable or disable the identity management feature, use the following admin-level commands.
  `enable identity-management`

```
disable identity-management
```

Identity manager does not detect and create identities for *FDB* blackhole and static entries.

Note

When the identity management feature is first enabled, the FDB entries previously learned on identity-management enabled ports are flushed.

*Enabling and Disabling Identity Management on Ports*

- To add or delete identity management on specific ports, use the following command:

  ```
  configure identity-management {add | delete} ports [port_list | all]
  ```

- To return to the default value, which removes all ports from the port list, use the following command:

  ```
  unconfigure identity-management ports
  ```

## Enabling and Disabling SNMP Traps

- To enable the transmission of *SNMP (Simple Network Management Protocol)* traps for identity management low memory conditions, use the following command:

  ```
  enable snmp traps identity-management
  ```

- To disable SNMP traps for identity management, use the following command:

  ```
  disable snmp traps identity-management
  ```

*Adjusting the Aging Time for Stale Entries*

The stale-entry aging time defines when event entries in the identity management database become stale. To preserve memory, the software periodically uses a cleanup process to remove the stale entries.

- To adjust the period at which stale database entries are deleted (regardless of the database usage level), use the following command:

  ```
  configure identity-management stale-entry aging-time seconds
  ```

  Note

  For additional information on the stale-entry aging time and how it can be automatically adjusted by the software, see the command description for the above command.

- To set the stale-entry aging time to the default value, use the following command:

  ```
  unconfigure identity-management stale-entry aging-time
  ```

*Resetting the Identity Management Configuration to the Default Values*

To reset the identity management configuration to the default values, use the following command:

```
unconfigure identity-management
```

## Adding and Deleting Entries in the Blacklist and Whitelist

To add or delete entries in the blacklist or whitelist, use the following commands:

```
configure identity-management blacklist add [mac mac_address {macmask} |
ip ip_address {netmask} | ipNetmask] | user user_name]

configure identity-management whitelist add [mac mac_address {macmask} |
ip ip_address {netmask} | ipNetmask] | user user_name]

configure identity-management blacklist delete [all | mac mac_address
{macmask} | ip ip_address {netmask} | ipNetmask] | user user_name]

configure identity-management whitelist delete [all | mac mac_address
{macmask} | ip ip_address {netmask} | ipNetmask] | user user_name]
```

*Configuring Entries in Greylist*

- To add or delete entries in greylist, use the following commands:

  ```
  configure identity-management greylist add user username

  configure identity-management greylist delete [all | user username]
  ```

- To display the entries in greylist, use the following command:

  ```
  show identity-management greylist
  ```

*Configuring List-Precedence*

- To configure or reset list-precedence, use the following commands:

  ```
  configure identity-management list-precedence listname1 listname2
  listname3

  unconfigure identity-management list-precedencee
  ```

- To display the list-precedence configuration, use the following command:

  ```
  show identity-management list-precedence
  ```

- To display the entries in the blacklist or whitelist, use the following commands:

  ```
  show identity-management blacklist

  show identity-management whitelist
  ```

## Configuring Kerberos Snooping

Kerberos authentication or ticketing is used by Microsoft's Active Directory and by various Unix systems (including Linux and MAC OSX). The Kerberos snooping feature in the ExtremeXOS software collects identity information from Kerberos Version 5 traffic. This feature does not capture information from earlier versions of Kerberos.

Note

We recommend that you enable CPU DoS protect in combination with this feature to make sure the CPU is not flooded with mirrored Kerberos packets in the event of a DoS attack on Kerberos TCP/UDP ports. If the rate limiting capability is leveraged on capable platforms, it is applied on CPU mirrored packets.

Kerberos snooping is enabled when you enable identity management.

Note

Kerberos identities are not detected when both server and client ports are added to identity management.

### *Configuring a Kerberos Server List*

By default, the identity management feature collects information from all Kerberos servers. However, this can subject the switch to DoS attacks targeted at Kerberos servers. To reduce the opportunities for DoS attacks, you can configure a Kerberos server list for identity management. When a Kerberos server list exists, identity management collects information only from the servers in the list.

- To add a server to the Kerberos server list, use the following command:

  `configure identity-management kerberos snooping add server ip_address`

- To delete a server from the Kerberos server list, use the following comma nd:

  `configure identity-management kerberos snooping delete server [ip_address|all]`

Note

Identity management supports configuration of up to 20 Kerberos servers.

### *Adjusting the Kerberos Snooping Aging Time*

Kerberos does not provide any service for un-authentication or logout. Kerberos does provide a ticket lifetime, but that value is encrypted and cannot be detected during snooping. To enable the aging and removal of snooped Kerberos entries, this timer defines the maximum age for a snooped entry. When a MAC address with a corresponding Kerberos entry in identity manager is aged out, the Kerberos snooping aging timer starts. If the MAC address becomes active before the Kerberos snooping aging timer expires, the timer is reset and the Kerberos entry remains active. If the MAC address is inactive when the Kerberos snooping aging timer expires, the Kerberos entry is removed.

- To configure the Kerberos snooping aging time, use the following command:

  `configure identity-management kerberos snooping aging time minutes`

- To reset the Kerberos snooping aging time to the default value, use the following command:

  `unconfigure identity-management kerberos snooping {aging time}`

Note

The default value for this command is `none`, which means that an identity discovered through Kerberos snooping is removed immediately on the aging out of the identity MAC address by the *FDB* manager.

### *Forced Kerberos Logout*

To force the removal of all identities discovered through Kerberos snooping, use the following command:

`configure identity-management kerberos snooping force-aging time [none | minutes]`

## Configure Default and User-Defined Roles

*Creating and Deleting User-Defined Roles*

- To create or delete a role, use the following commands:

  ```
  create identity-management role role_name match-criteria
  match_criteria {priority pri_value}

  delete identity-management role {role-name | all}
  ```

- To create or delete a child role, use the following commands:

  ```
  configure identity-management role role_name add child-role child_role

  configure identity-management role role_name delete child-role
  [child_role | all]
  ```

*Configuring Rules or Policies for Default and User-Defined Roles*

The default authenticated and unauthenticated roles contain no rules or policies. When you first create a user-define role, it also contains no rules or policies.

To add or delete a rule or policy from a role, use the following commands:

```
configure identity-management role role_name [add dynamic-rule rule_name
{ first | last | { [before | after] ref_rule_name}}]

configure identity-management role role_name add policy policy-name
{first | last {[before | after] ref_policy_name}}

configure identity-management role role_name delete dynamic-rule
[rule_name | all]

configure identity-management role role_name delete policy [policy-name
| all]
```

*Configuring LDAP Server Access*

- To add or remove LDAP server connections for retrieving identity attributes, use the following commands:

  ```
  configure ldap {domain domain_name} add server [host_ipaddr |
  host_name] {server_port} {client-ip client_ipaddr} {vr vr_name}
  {encrypted sasl digest-md5}
  ```

- To create a new domain, use the following command:

  ```
  create ldap domain domain_name {default}
  ```

- To configure credentials for accessing an LDAP server, use the following command:

  ```
  configure ldap {domain [domain_name|all]} bind-user [user_name
  {encrypted} password | anonymous]
  ```

- To specify a base domain name to be added to usernames in LDAP queries, use the following command:

  ```
  configure ldap {domain [domain_name|all]} base-dn [base_dn | none |
  default]
  ```

- To specify a domain as default, use the following command:

  ```
  configure ldap domain domain_name [default | non-default]
  ```

- To enable or disable LDAP queries for specific network login types, use the following command:

  ```
  configure ldap { domain [ domain_name | all ] } [enable|disable]
  netlogin [dot1x | mac | web-based]
  ```

- To configure bind-user for LDAP queries, use the following command:

  ```
  configure ldap {domain [domain_name|all]} bind-user [user_name
  {encrypted} password | anonymous] {domain [domain_name|all]}
  ```

- To delete an LDAP server, use the following commands:

  ```
  configure ldap {domain [domain_name|all]} delete server [host_ipaddr |
  host_name] {server_port} {vr vr_name}
  ```

- To delete a domain, use the following command:

  ```
  delete ldap domain [domain_name | all]
  ```

*Support for Multiple Windows Domains*

Some organizations are large enough to use multiple Windows domains (sub-domains) in their networks. Each Windows domain can have its own LDAP server.

In previous releases, identity manager supported up to eight LDAP servers which are assumed to be replicas on the same domain (default base-dn). From the 15.2 release, identity manager supports multiple Windows domains.

**LDAP Servers in Different Domains**

In 15.2, identity manager can service users under different domains. You can configure different domains and add different LDAP servers for these different domains. When adding an LDAP server to identity manager, you can specify the domain under which the server is to be added.

- You can configure a base-dn and a bind user for each domain.
- Base-dn is assumed to be the same as the domain name unless explicitly configured otherwise. (Base-dn is the LDAP directory under which the users are to be searched.)
- For users upgrading from older configurations, the base-dn configured on an older version now becomes the default domain name. This can be changed later if required.
- For users upgrading from older configurations, the LDAP servers configured on older versions are now servers under the default domain.
- You can now add up to eight LDAP servers to each of the user-configured domains.

**LDAP Connections**

Identity manager tries to maintain LDAP connections with one of the servers in each of the configured domains. LDAP queries for users logging on to those domains will be sent to the respective servers or to a server on the default domain if the user does not fall under any configured domain. The LDAP server might choose to close the connection after a timeout.

**LDAP Process**

Identity manager tries to bind to one of the configured LDAP servers in each of the user-configured domains.

When a new user is detected, the user's domain is used to determine the LDAP server to be contacted for the user's details.

If there is a match, the LDAP server corresponding to that domain is chosen and the LDAP search request for the user attributes is sent to that LDAP server.

If the domain does not match any of the configured domains, LDAP query is sent to a server in the default domain.

*Changing the Role Priority*

The role priority is defined when a role is created. To change the priority for a role, use the following commands:

```
configure identity-management role role_name priority pri_value
```

# Managing the Identity Management Feature

## Clearing the Identity Management Counters

To clear the statistics counters for the identity management feature, use the following commands:

```
clear counters
```

```
clear counters identity-management
```

## Refreshing the Role Selection for Users

To refresh role evaluation for a specified user, for all users, or for all users currently using the specified role, use the following commands:

```
refresh identity-management role user [user_name {domain domain_name} |
all {role role_name}]
```

## Enabling/Disabling Snooping Identities

The identity management feature makes the edge of the network more intelligent by providing access to the devices/users in the network. The identity manager detects the identities through the following protocols:

- *FDB*
- IPARP
- IPSecurity *DHCP* Snooping
- *LLDP*
- Netlogin
- Kerberos

By default, identity management detects identities through all of the above mentioned protocols. There is no way for the administrator to disable the detection of the identities tht are triggered through the above protocols.

This feature now provides the administrator an option to enable/disable the detection of the identities that are triggered through any of the above protocols. The administrator can now control the identity detection through any of the protocol triggers at port level. This configuration can be applied to identity management-enabled ports only. An error is received if this configuration is applied to identity management-disabled ports.

As part of this feature, the limitation of FDB entries getting cleared on enabling identity management on a port is removed. The identity mangement module will retrieve the FDB entries learned on the identity management-enabled ports and create the identity accordingly.

Note

All types of Netlogin identity will not be detected if the netlogin detection is disabled.

Enabling Kerberos identity detection does not create identities for previously authenticated clients.

## Displaying Identity Management Information

### Displaying Database Entries

Log in as an admin-level user.

To display the entries in the identity management database, enter the following command:

```
show identity-management entries {user id_name} {domain domain} {ports port_list} {mac mac_address} {vlan vlan_name} {ipaddress ip_address} {detail}
```

### Displaying Configuration Information

Use the following command to display the current configuration of the identity management feature:

```
show identity-management
```

```
show identity-management role {role-name} {detail}
```

### Displaying Statistics

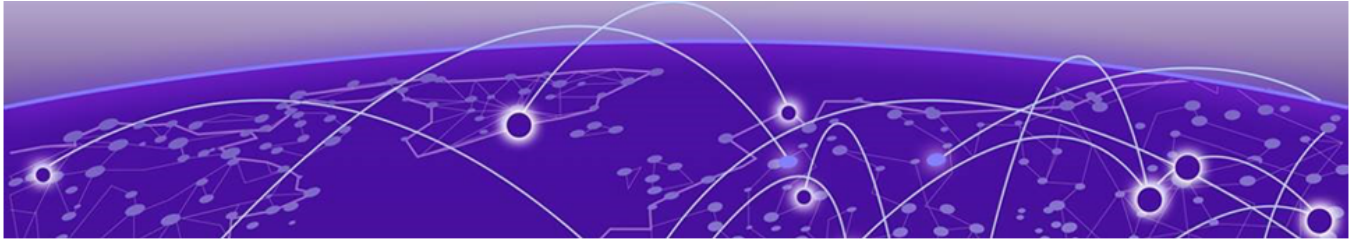
- To display operating statistics for the identity management feature, use the following command:

  ```
  show identity-management statistics
  ```

- To clear the statistics counters for the identity management feature, enter either of the following commands:

  ```
  clear counters
  ```

  ```
  clear counters identity-management
  ```



# Security



## Security Features Overview

Security is a term that covers several different aspects of network use and operation.

One general type of security is control of the devices or users that can access the network. Ways of doing this include authenticating the user at the point of logging in, controlling access by defining limits on certain types of traffic, or protecting the operation of the switch itself. Security measures in this last category include routing policies that can limit the visibility of parts of the network or denial of service protection that prevents the CPU from being overloaded. Finally, management functions for the switch can be protected from unauthorized use. This type of protection uses various types of user authentication.

ExtremeXOS has enhanced security features designed to protect, rapidly detect, and correct anomalies in your network. Extreme Networks products incorporate a number of features designed to enhance the security of your network while resolving issues with minimal network disruption. No one feature can

ensure security, but by using a number of features in concert, you can substantially improve the security of your network.

The following list provides a brief overview of some of the available security features:

- *ACL (Access Control List)*s are policy files used by the ACL application to perform packet filtering and forwarding decisions on incoming traffic and packets. Each packet arriving on an ingress port is compared to the ACL applied to that port and is either permitted or denied.

  For more information about using ACLs to control and limit network access, see ACLs Overview on page 640.
- CLEAR-Flow—A security rules engine available only on BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules, and Summit X440, X460, X480, X670, and X770 series switches in a non-stack configuration. CLEAR-Flow inspects Layer 2 and Layer 3 packets, isolates suspicious traffic, and enforces policy-based mitigation actions. Policy-based mitigation actions include the switch taking an immediate, predetermined action or sending a copy of the traffic off-switch for analysis. Working together, CLEAR-Flow provides a rapid response to network threats.

  For more information about CLEAR-Flow, see CLEAR-Flow on page 948.
- Denial of Service Protection—DoS protection is a dynamic response mechanism used by the switch to prevent critical network or computing resources from being overwhelmed and rendered inoperative. In essence, DoS protection protects the switch, CPU, and memory from attacks and attempts to characterize the attack (or problem) and filter out the offending traffic so that other functions can continue. If the switch determines it is under attack, the switch reviews the packets in the input buffer and assembles ACLs that automatically stop the offending packets from reaching the CPU. For increased security, you can enable DoS protection and establish CLEAR-Flow rules at the same time.

  For more information about DoS attacks and DoS protection, see Denial of Service Protection.
- Network Login—Controls the admission of user packets and access rights thereby preventing unauthorized access to the network. Network login is controlled on a per port basis. When network login is enabled on a port in a *VLAN (Virtual LAN)*, that port does not forward any packets until authentication takes place. Network login is capable of three types of authentication: web-based, MAC-based, and 802.1X.

  For more information about network login, see Network Login Overview on page 756.
- Policy Files—Text files that contain a series of rule entries describing match conditions and actions to take. Policy files are used by both routing protocol applications (routing policies) and the ACL application (ACLs).

  For more information about policy files, see Creating and Editing Policies on page 635.
- Routing Policies—Policy files used by routing protocol applications to control the advertisement, reception, and use of routing information by the switch. By using policies, a set of routes can be selectively permitted or denied based on their attributes for advertisements in the routing domain. Routing policies can be used to “hide” entire networks or to trust only specific sources for routes or ranges of routes.

For more information about using routing policies to control and limit network access, see Routing Policies Overview on page 713.

- sFlow—A technology designed to monitor network traffic by using a statistical sampling of packets received on each port. sFlow also uses IP headers to gather information about the network. By gathering statistics about the network, sFlow becomes an early warning system, notifying you when there is a spike in traffic activity. Upon analysis, common response mechanisms include applying an ACL, changing *QoS (Quality of Service)* parameters, or modifying VLAN settings.

  For more information, see Using sFlow on page 486.

## Safe Defaults Mode

When you set up your switch for the first time, you must connect to the console port to access the switch.

After logging in to the switch, you enter safe defaults mode. Although *SNMP (Simple Network Management Protocol)*, Telnet, and switch ports are enabled by default, the script prompts you to confirm those settings. By answering `N (No)` to each question, you keep the default settings.

```
Would you like to disable Telnet? [y/N]: No
Would you like to disable SNMP [y/N]: No
Would you like unconfigured ports to be turned off by default [y/N]: No
Would you like to change the failsafe account username and password now? [y/N]: No
Would you like to permit failsafe account access via the management port? [y/N]: No
```

In addition, if you keep the default settings for SNMP and Telnet, the switch returns the following interactive script:

```
Since you have chosen less secure management methods, please remember to increase the
security of your network by taking the following actions:
* change your admin password
* change your failsafe account username and password
* change your SNMP public and private strings
* consider using SNMPv3 to secure network management traffic
```

For more detailed information about safe defaults mode, see Using Safe Defaults Mode on page 24.

## MAC Security

The switch maintains a database of all media access control (MAC) addresses received on all of its ports.

The switch uses the information in this database to decide whether a frame should be forwarded or filtered. MAC security (formerly known as MAC address security) allows you to control the way the *FDB (forwarding database)* is learned and populated. For more information, see FDB on page 561.

MAC security includes several types of control. You can:

- Limit the number of dynamically learned MAC addresses allowed per virtual port. For more information, see Limiting Dynamic MAC Addresses on page 871.

- "Lock" the FDB entries for a virtual port, so that the current entries will not change, and no additional addresses can be learned on the port. For information, see MAC Address Lockdown on page 874.

  Note
  You can either limit dynamic MAC FDB entries or lockdown the current MAC FDB entries, but not both.

- Set a timer on the learned addresses that limits the length of time the learned addresses will be maintained if the devices are disconnected or become inactive. For more information, see MAC Address Lockdown with Timeout on page 874.

  Note
  When limit-learning is configured in the port which is also associated with some other vlan where learning is disabled, then few packets with new MAC address beyond learning limit will get flooded. This flooding will take place for fraction of second until new black-hole entry is created in hardware.

- Use ACLS to prioritize or stop packet flows based on the source MAC address of the ingress virtual LAN (*VLAN*) or the destination MAC address of the egress VLAN. For more information about *ACL* policies, see Security on page 859.
- Enhance security, depending on your network configuration, by disabling Layer 2 flooding. For more information about enabling and disabling Layer 2 flooding, see Managing Egress Flooding on page 569.

## MAC Locking

MAC locking helps prevent unauthorized access to the network by limiting access based on a device's MAC address. MAC locking enables the binding of specific MAC addresses to specific ports on a switch. MAC locking locks a port to one or more MAC addresses, preventing connection of unauthorized devices via a port. With MAC locking enabled, the only frames forwarded on a MAC locked port are those with the configured or dynamically selected MAC addresses for that port.

Frames received on a port with a Source MAC address not bound to the port are discarded or optionally allowed to dynamically bind to the port, up to a user-controlled maximum number of MAC addresses per port.

There are two different types of MAC locking:

• Static MAC locking - Locking one or more specified MAC addresses to a port.

• First Arrival MAC locking - Locking one or more MAC addresses to a port based on first arrival of received frames after First Arrival MAC locking is enabled. The configuration specifies the maximum number of end users that will be allowed. As each new end user is identified, it is MAC locked up to the maximum number of users. Once the maximum number of users has been MAC locked, all other users will be denied access to the port until a MAC locked address is either aged, if aging is configured, or the session for that user ends.

The MAC locking feature is disabled in the device, by default. MAC locking must be enabled both globally and on port level. Once enabled, ports can be configured for static and First Arrival MAC locking.

Existing limit learning and lock learning features are supported on a port-*VLAN* combination. The MAC locking feature implemented in ExtremeXOS 15.7 supports MAC locking functionality on a port basis.

### *MAC Locking Limitations*

The following limitations exist for the MAC locking feature:

- MAC locking cannot be enabled along with existing limit learning and lock learning features.
- MAC locking is be supported on ports in the VPLS service *VLAN*.

### *MAC Locking Functionality*

This feature provides a global enable/disable control and controls per port. Additional controls are provided to create, destroy, enable and disable static MAC to Interface bindings. The MAC and port, together are required to be unique (the same MAC may be bound to multiple ports).

The configuration of a maximum number of static and dynamic entries (each individually) is provided per port. First Arrival bindings (also known as first arrival MAC locking stations) are not persistent through a device reset. Static bindings (also known as static MAC locking stations) are maintained through a device reset. First Arrival bindings are removed in the event of link loss, or after the *FDB* entry with the MAC locked MAC ages out. A port, may have both the static and dynamic entries, at any instance.

Dynamic locking may be disabled by setting the maximum number of First Arrival MAC addresses to zero.

Controls are provided per port to convert all current first arrival entries to static entries. This converts only the first arrival MAC bindings to static bindings. This is not to be confused with MAC locking where all dynamic FDB entries are converted to static permanent locked FDB entries and further learning is disabled. Controls are provided to clear current (both static and dynamic) bindings. Each port is configurable to support the aging of dynamically locked bindings.

The device keeps a record of the number of MAC locked stations and when the configured threshold is reached, a threshold *SNMP* trap/notification and/or a log message is issued based on the per-port configuration. The Source MAC Address of the frame causing the last invalid attempt is also recorded. In the event that the device is so configured, a violation SNMP Trap/ Notification and/or a violation log message is issued – these controls will be exercised per port.

### *Configuring MAC Locking*

The following content describes the MAC locking commands:

#### Enabling and Disabling MAC Locking Globally

To enable or disable the MAC locking feature, use the following commands:

- `enable mac-locking`
- `disable mac-locking`

#### Enabling and Disabling MAC Locking on Ports

To enable MAC locking on specific ports, use the following command:

- `enable mac-locking ports [`**`all`**`| port_list]`

To Disable MAC locking on specific ports, use the following command:

- `disable mac-locking ports [all| port_list]`

#### Configuring First-arrival/Dynamic MAC Locking Limit

To configure first arrival MAC locking on a port, use the following command:

- `configure mac-locking ports port_list first-arrival limit-learning learn_limit`

When the configured limit is reached, no further entries are learnt. However, if the learnt entries are aged out, new MAC addresses can be learned.

By default, Aging is disabled for first arrival MAC locking entries on a configured port. When the *FDB* entries are aged out, they are removed from the FDB, but they are retained in the MAC locking table. So when the first arrival limit is reached, only those entries in the MAC locking table can be learned again if these devices start sending out traffic. Any new MAC addresses cannot be learned.

The maximum number of dynamic MAC addresses that can be locked on a port is 600.

Note

There is no command to unconfigure first arrival or static MAC locking limit. The value can be reset giving the default learn limit specified in the help text. When First arrival MAC locking is configured to a value that is lower than the number of MACs that are locked, all the MAC locking bindings on the port are cleared.

#### Configuring Static MAC Locking Limit

To configure static MAC locking on a port, use the following command:

- `configure mac-locking ports port_list static limit-learning learn_limit`

When the configured limit is reached, no further entries are learned (either black holed or not further learned depending on the configured action). However, if the learned entries are cleared or deleted, new MAC addresses can be created and learned. The maximum number of MAC addresses that will be locked on a port configured for static MAC locking is 64. Aging is not applicable for the static MAC locking entries.

Note

There is no command to unconfigure first arrival or static MAC locking limit. The value can be reset giving the default learn limit specified in the help text. CLI doesn't allow changing the static MAC locking limit value to a value lower than the number of MACs locked in the MAC lock station table. Some or all of the created MAC locking stations should be removed to change the limit to a lower value.

**Note**

Assume that port 2:22 is enabled for MAC Locking. The maximum static entry limit value is configured to 5 on port 2:22. If the user wants to configure the maximum static entry limit value to 3,

1. The device will display an error, as maximum number of static stations already locked on the port and the value cannot be reduced.
2. In the same example, the port 2:22 has only 1 static station locked. If the maximum static entry limit value is reduced to 3, the device will allow to reduce the value.

Scenario A:

```
* Slot-1 DUT1.94 # show mac-locking ports 2:22

MAC locking is globally enabled.
Port   MAC  Trap      Log       FA    Limit     Link    Max Max   Last Violating
       Lock Thr|Viol Thr|Viol Aging   Action    Down    Stc FA    MAC Address
       Stat                         Cfg|Stat   Action
-----  ---- -------- -------- ----- --------    ------ --- ---   -----------------
2:22   ena  off|on    off|on   dis   ena|ena   clear    5   45   00:11:11:11:11:04

Legend:

Stat              - Status                    Thr|Viol - Threshold | Violation

Max Stc           - Max Static Count          Max FA   - Max First-Arrival Count

dis               - Disabled                  ena       - Enabled

retain            - Retain MACs               clear     - Clear MACs

Limit Action Cfg  - If port should be disabled when learnt limit is exceeded

             dis   - Port to be disabled when learn limit is exceeded

             ena   - Port to remain enabled when learn limit is exceeded

Limit Action Stat - Port status on exceeding learn limit

*  Slot-1 DUT1.95 #

*  Slot-1 DUT1.95 # show mac-locking stations ports 2:22

Port   MAC Address         Status     State           Aging

-----  -----------------   --------   -------------   -----

2:22   00:11:11:11:11:00   active     static          false

2:22   00:11:11:11:11:01   active     static          false

2:22   00:11:11:11:11:02   active     static          false

2:22   00:11:11:11:11:03   active     static          false

2:22   00:11:11:11:11:04   active     static          false

*  Slot-1 DUT1.96 # configure mac-locking ports 2:22 static limit-learning 3

Error: Static limit-learning value cannot be reduced to 3 for port 2:22 as 5 static
```

```
stations are already created.


Configuration failed on backup Node, command execution aborted!

* Slot-1 DUT1.97 #
```

Scenario B:

```
* Slot-1 DUT1.109 # show mac-locking stations ports 2:22

Port   MAC Address         Status     State           Aging

-----  -----------------   --------   -------------   -----

2:22   00:11:11:11:11:00   active     static          false

2:22   00:11:11:11:11:01   active     first-arrival   false

2:22   00:11:11:11:11:02   active     first-arrival   false

2:22   00:11:11:11:11:03   active     first-arrival   false

2:22   00:11:11:11:11:04   active     first-arrival   false

* Slot-1 DUT1.109 # show mac-locking ports 2:22


MAC locking is globally enabled.

Port   MAC  Trap      Log       FA    Limit     Link    Max Max   Last Violating

       Lock Thr|Viol Thr|Viol Aging   Action    Down    Stc FA    MAC Address

       Stat                          Cfg|Stat   Action

-----  ---- --------  --------  ----- --------  ------  --- ---   -----------------

2:22   ena  off|on    off|on    dis   ena|ena   clear     5   45   00:11:11:11:11:04

Legend:

Stat              - Status                    Thr|Viol - Threshold | Violation

Max Stc           - Max Static Count          Max FA   - Max First-Arrival Count

dis               - Disabled                  ena       - Enabled

retain            - Retain MACs               clear     - Clear MACs

Limit Action Cfg  - If port should be disabled when learnt limit is exceeded

             dis   - Port to be disabled when learn limit is exceeded
             ena   - Port to remain enabled when learn limit is exceeded

Limit Action Stat - Port status on exceeding learn limit

* Slot-1 DUT1.110 # configure mac-locking ports 2:22 static limit-learning 3

* Slot-1 DUT1.111 # show mac-locking ports 2:22

MAC locking is globally enabled.
```

```
Port   MAC  Trap      Log       FA    Limit     Link    Max Max   Last Violating

       Lock Thr|Viol Thr|Viol Aging   Action    Down    Stc FA    MAC Address

       Stat                          Cfg|Stat Action

-----  ---- --------  --------  -----  --------  ------  --- ---   -----------------

2:22   ena  off|on    off|on    dis    ena|ena   clear     3   45   00:11:11:11:11:04

Legend:

Stat              - Status                    Thr|Viol - Threshold | Violation

Max Stc           - Max Static Count          Max FA   - Max First-Arrival Count

dis               - Disabled                  ena       - Enabled

retain            - Retain MACs               clear     - Clear MACs

Limit Action Cfg  - If port should be disabled when learned limit is exceeded

             dis   - Port to be disabled when learn limit is exceeded
             ena   - Port to remain enabled when learn limit is exceeded

Limit Action Stat - Port status on exceeding learn limit
```

#### Create/Enable/Disable static MAC Locking Entries

To create a static MAC locking entry (also known as MAC locking station) and enable or disable MAC locking for the specific MAC address and port, use the following command:

`configure mac-locking ports` *`port_list`* `static [`**`add`**` | `**`enable`**` | `**`disable`**`] station` *`station_mac_address`*

Note
A static MAC locking station is enabled by default.

To disable the static MAC locking station, use the following command.

`configure mac-locking ports` *`port_list`* `static disable station` *`station_mac_address`*

When created and enabled, a static MAC lock configuration allows only the end station designated by the MAC address to participate in forwarding of traffic.

The disabled entries are also counted when calculating the total number of locked stations. Static MAC locking stations that are disabled are only shown in “show mac-locking stations static” command. When “static” keyword is not given in “show mac-locking stations”, the disabled entries are not shown.

#### Enable/Disable Aging of First-arrival MAC Addresses

To enable or disable first arrival MAC address aging, use the following command.

`configure mac-locking ports` *`port_list`* `first-arrival aging [`**`enable`**` | `**`disable`**`]`

Dynamic MAC locking mode MAC address aging is disabled by default.

This is applicable only to MAC addresses locked by first-arrival locking and not to MAC addresses locked by static locking.

When First arrival aging is disabled, MAC locking stations are retained even when the corresponding FDB entry ages out.

When First arrival aging is enabled, MAC locking station starts aging when all the FDB entries corresponding to the station MAC are removed. MAC lock stations do not start aging when FDB entries are present.

When “firstarrival aging” is configured to be enabled in first-arrival locking, when an FDB entry ages out, the entry is no more locked and so new MAC addresses can be learned till the configured first-arrival limit is reached.

Note
First arrival Aging – Age out time for First Arrival MAC locking station is same as FDB aging time that is configured using `configure fdb agingtime`.

#### Move First-arrival MACs to Static Entries

To move all current first-arrival MACs to static entries on a port, use the following command:

`configure mac-locking ports` *`port_list`* `first-arrival move-to-static`

This command converts dynamic MAC locked stations to static MAC locked stations. There is no change to FDB entries.

The static MAC locked station entries are saved in configuration and so are preserved across reboots.

Note
Ensure the static limit can accommodate the entries before moving them from to static. Otherwise, the device may throw the following error: `Error: Some dynamic stations could not be converted to static stations for port <port_list>`.

Note
An FDB entry created from the CLI will not be removed when a static MAC lock station is created and disabled for the corresponding MAC address. It is necessary to delete the FDB entry from the CLI. MAC-Locking does not remove user created FDB entries.

## *Managing MAC Locking*

#### Enable/Disable Clearing First-arrival MACs on Link Change

To manage the behavior of first arrival MAC locking on link state change, use the following command.

`configure mac-locking ports` *`port_list`* `first-arrival link-down-action [`**`clear-macs`**` | `**`retain-macs`**`]`

Clear MAC on link change is enabled by default.

When the link goes down, by default, all the first arrival MAC locking addresses will be removed. When link-down-action is configured to “retain-macs”, the first arrival MAC locking addresses will be retained even when the link goes down.

#### Disable/Enable port when MAC threshold is reached

This command is used to configure the disabling of ports when the configured MAC threshold is met. This is used for both “first arrival” and “static” MAC locking methods.

`configure mac-locking ports` *`port_list`* `learn-limit-action [`**`disable-port`** | **`remain-enabled`**

The port is disabled when the configured MAC threshold is met. All the *FDB* entries learned on this port are flushed as the port is disabled. This configuration can be reset using the `clear mac-locking disabled-state ports` *`port_list`* command. When MAC locking is disabled on the port, the port comes back up.

#### Clearing the Disabled-state of a Port

This command is used to return the behavior of first arrival MAC locking with link state change to its default value of enabled.

`clear mac-locking disabled-state ports` *`port_list`*

#### Delete Static MAC Locking Entries

To delete MAC locking for all static MAC address or the specified static MAC address on the given port, use the following command:

`configure mac-locking ports` *`port_list`* `static delete station [`*`station_mac_address`* | **`all`**`]`

#### Clearing MAC Locking entries

The following command is used to clear MAC locking station entries for the given parameters:

`clear mac-locking station [`**`all`** | {**`mac`** *`station_mac_address`*} {**`first-arrival`** | **`static`**} {**`ports`** *`port_list`*}]

This command clears MAC locking configuration by port/mac/first arrival/static etc.

Note
Clearing static MAC locking stations will remove them from the configuration. The cleared static MAC locking stations will not be saved across reboots.

#### Displaying MAC Locking Information

This command is used to display the status of MAC locking on one or more ports.

`show mac-locking {`**`ports`** *`port_list`*}

If port is not specified, MAC locking status will be displayed for all ports.

Sample output:

```
Slot-1 Stack.2 # show mac-locking

MAC locking is globally enabled.

Port   MAC  Trap      Log       FA    Limit     Link    Max Max   Last Violating
```

```
    Lock Thr|Viol    Thr|Viol   Aging Action    Down    Stc FA    MAC Address
   Stat                         Cfg|Stat Action
-----  ---- -------- -------- ----- --------    ------ --- ---    -----------------
1:1    dis  off|off   off|off   dis  ena|ena    clear    64 600    00:00:00:00:00:00
1:2    dis  off|off   off|off   dis  ena|ena    clear    64 600    00:00:00:00:00:00
1:3    dis  off|off   off|off   dis  ena|ena    clear    64 600    00:00:00:00:00:00
1:4    dis  off|off   off|off   dis  ena|ena    clear    64 600    00:00:00:00:00:00
1:5    dis  off|off   off|off   dis  ena|ena    clear    64 600    00:00:00:00:00:00

Legend:
Stat              - Status                    Thr|Viol - Threshold | Violation
Max Stc           - Max Static Count          Max FA   - Max First-Arrival Count
dis               - Disabled                  ena       - Enabled
retain            - Retain MACs               clear     - Clear MACs
Limit Action Cfg  - If port should be disabled when learnt limit is exceeded
             dis   - Port to be disabled when learn limit is exceeded
             ena   - Port to remain enabled when learn limit is exceeded
Limit Action Stat - Port status on exceeding learn limit
```

The following command displays MAC locking stations for different parameters:

`show mac-locking stations {`**`first-arrival`** | **`static`**`} {`**`ports`** *`port_list`*`}`

*MAC Locking Configuration Example*

The following command enables MAC locking both globally for the device (stack) and at the port level for ports 2:1 through 2: 5:

```
Slot-1 Stack.1 # enable mac-locking

Slot-1 Stack.2 # enable mac-locking ports 2:1-2:5
```

The following command lines enable port 2:1 for a maximum of 3 static MAC address entries. This is followed by four static MAC address creation entries. The fourth entry fails because the maximum allowed has been set to 3.

```
* Slot-1 Stack.3 # configure mac-locking ports 2:1 static limit-learning 3
* Slot-1 Stack.4 # configure mac-locking ports 2:1 static add station 00:22:33:44:55:66
* Slot-1 Stack.5 # configure mac-locking ports 2:1 static add station 00:22:33:44:55:77
* Slot-1 Stack.6 # configure mac-locking ports 2:1 static add station 00:22:33:44:55:88
* Slot-1 Stack.7 # configure mac-locking ports 2:1 static add station 00:22:33:44:55:99
Error: Station 00:22:33:44:55:99 cannot be added as maximum static limit of 3 is already
reached on port 2:1.


* Slot-1 Stack.10 # show mac-locking stations static ports 2:1
Port   MAC Address         Status     State           Aging
-----  ------------------  --------   -------------   -----
2:1    00:22:33:44:55:66   active     static          false
2:1    00:22:33:44:55:77   active     static          false
2:1    00:22:33:44:55:88   active     static          false

Total for specified ports: 3 Static: 3 First-Arrival: 0
* Slot-1 Stack.11 #
```

The following commands configure ports 2:2 through 2:5 for dynamic MAC locking with a maximum of 15 users on each port. This is followed by a line enabling MAC locking trap messaging on ports 2:1 through 5:

```
* Slot-1 Stack.12 # configure mac-locking ports 2:2-2:5 first-arrival limit-learning 15
```

```
* Slot-1 Stack.13 # configure mac-locking ports 2:2-2:5 trap on
* Slot-1 Stack.14 # show mac-locking ports 2:1-2:5

MAC locking is globally enabled.

Port   MAC  Trap      Log       FA    Limit     Link    Max Max   Last Violating
       Lock Thr|Viol Thr|Viol Aging   Action    Down    Stc FA    MAC Address
       Stat                         Cfg|Stat   Action
-----  ---- -------- -------- ----- --------    ------ --- ---    -----------------
2:1    ena  off|off   off|off   dis  ena|ena   clear     3 600    00:00:00:00:00:00
2:2    ena  off|on    off|off   dis  ena|ena   clear    64   15   00:00:00:00:00:00
2:3    ena  off|on    off|off   dis  ena|ena   clear    64   15   00:00:00:00:00:00
2:4    ena  off|on    off|off   dis  ena|ena   clear    64   15   00:00:00:00:00:00
2:5    ena  off|on    off|off   dis  ena|ena   clear    64   15   00:00:00:00:00:00

Legend:
Stat              - Status                    Thr|Viol - Threshold | Violation
Max Stc           - Max Static Count          Max FA   - Max First-Arrival Count
dis               - Disabled                  ena      - Enabled
retain            - Retain MACs               clear     - Clear MACs
Limit Action Cfg  - If port should be disabled when learnt limit is exceeded
             dis   - Port to be disabled when learn limit is exceeded
             ena   - Port to remain enabled when learn limit is exceeded
Limit Action Stat - Port status on exceeding learn limit
* Slot-1 Stack.15 #
```

## Limiting Dynamic MAC Addresses

You can set a predefined limit on the number of dynamic MAC addresses that can participate in the network.

After the *FDB* reaches the MAC limit, all new source MAC addresses are blackholed at both the ingress and egress points. These dynamic blackhole entries prevent the MAC addresses from learning and responding to *ICMP (Internet Control Message Protocol)* and address resolution protocol (ARP) packets.

**Note**

Blackhole FDB entries added due to MAC security violations on BlackDiamond 8800 series switches, SummitStack, and Summit family switches are removed after each FDB aging period regardless of whether the MAC addresses in question are still sending traffic. If the MAC addresses are still sending traffic, the blackhole entries will be re-added after they have been deleted.

### *Configuring Limit Learning*

- To limit the number of dynamic MAC addresses that can participate in the network, use the **limit-learning** option in following command:

  `configure ports` *port_list* {**tagged** *tag*} **vlan** *vlan_name* | *vlan_list* [**limit-learning** *number* {**action** [**blackhole** | **stop-learning**]} | **lock-learning** | **unlimited-learning** | **unlock-learning**]

  This command specifies the number of dynamically learned MAC entries allowed for these ports in this *VLAN*. The range is 0 to 500,000 addresses.

  When the learned limit is reached, all new source MAC addresses are blackholed at the ingress and egress points. This prevents these MAC addresses from learning and responding to *ICMP* and ARP packets.

Dynamically learned entries still get aged and can be cleared. If entries are cleared or aged out after the learning limit has been reached, new entries will then be able to be learned until the limit is reached again.

Permanent static and permanent dynamic entries can still be added and deleted using the `create fdb` and `disable flooding ports` commands. These override any dynamically learned entries.

For ports that have a learning limit in place, the following traffic still flows to the port:

- Packets destined for permanent MAC addresses and other non-blackholed MAC addresses
- Broadcast traffic
- *EDP (Extreme Discovery Protocol)* traffic

Traffic from the permanent MAC and any other non-blackholed MAC addresses still flows from the virtual port.

- To remove the learning limit, use the **unlimited-learning** option.

  `configure ports` *port_list* {**tagged** *tag*} **vlan** *vlan_name* | *vlan_list* [**limit-learning** *number* {**action** [**blackhole** | **stop-learning**]} | **lock-learning** | **unlimited-learning** | **unlock-learning**]

  The MAC limit-learning feature includes a stop-learning argument that protects the switch from exhausting *FDB* resources with blackhole entries. When limit-learning is configured with stop-learning, the switch is protected from exhausting FDB resources by not creating blackhole entries. Any additional learning and forwarding is prevented, but packet forwarding is not impacted for existing FDB entries.

  On the BlackDiamond 8800 series switches and the Summit X440, X460, X480, X670, and X770 series switches, the VLANs in a port are impacted when the configured learning limit is reached.

### *Display Limit Learning Information*

- To verify the configuration, enter the commands:

  `show vlan` *vlan name* **security**

  This command displays the MAC security information for the specified *VLAN*.

  `show ports` {**mgmt** | *portlist*} **info** {**detail**}

  This command displays detailed information, including MAC security information, for the specified port.

### *Example of Limit Learning*

In the following figure, three devices are connected through a hub to a single port on the Extreme Networks device.

If a learning limit of 3 is set for that port, and you connect a fourth device to the same port, the switch does not learn the MAC address of the new device; rather, the switch blackholes the address.

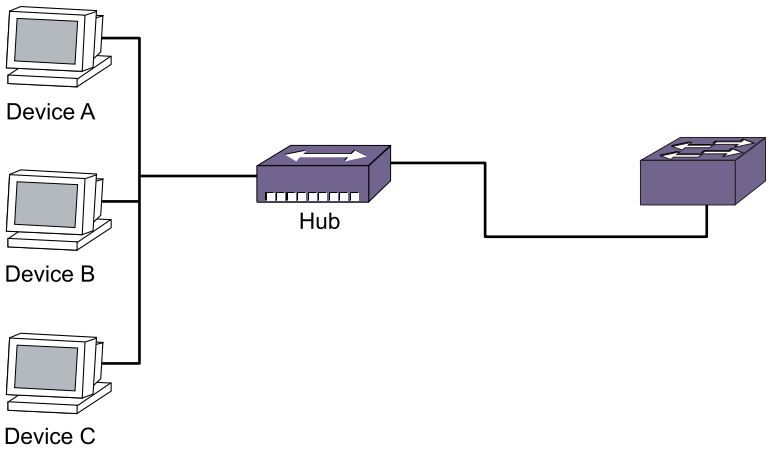



**Figure 106: Switch Configured for Limit Learning**

*Limiting MAC Addresses with ESRP Enabled*

If you configure a MAC address limit on *VLAN*s that participate in an *ESRP (Extreme Standby Router Protocol)* domain, you should add an additional back-to-back link (that has no MAC address limit on these ports) between the ESRP-enabled switches.

Doing so prevents ESRP protocol data units (PDUs) from being dropped due to MAC address limit settings.

The following figure is an example of configuring a MAC address limit on a VLAN participating in an ESRP domain.

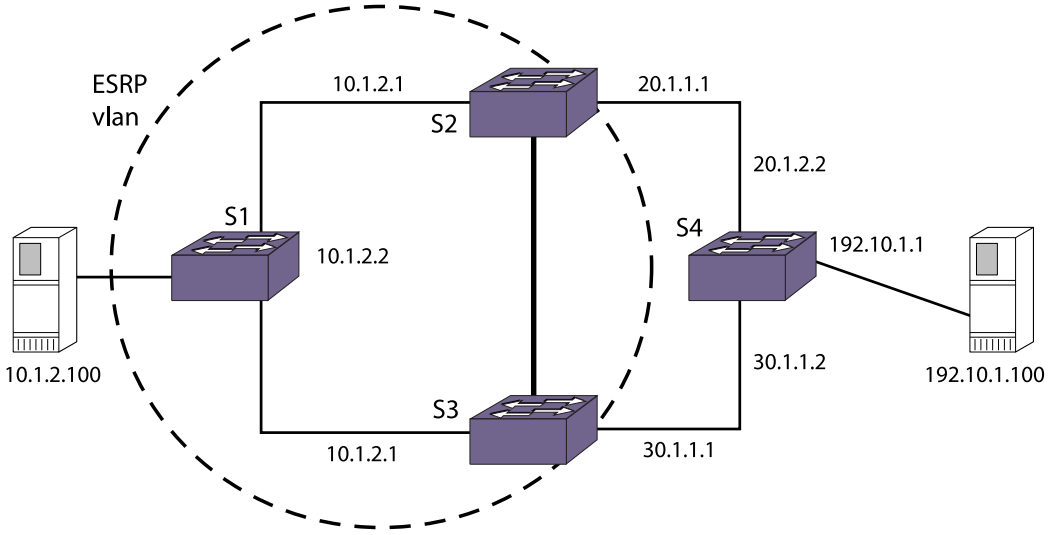



**Figure 107: MAC Address Limits and VLANs Participating in ESRP**

In the preceding figure, S2 and S3 are ESRP-enabled switches, while S1 is an ESRP-aware (regular Layer 2) switch. Configuring a MAC address limit on all S1 ports might prevent ESRP communication between S2 and S3. To resolve this, you should add a back-to-back link between S2 and S3. This link is not needed if MAC address limiting is configured only on S2 and S3, but not on S1.

## MAC Address Lockdown

- In contrast to limiting learning on virtual ports, you can lockdown the existing dynamic *FDB* entries and prevent any additional learning using the **lock-learning** option from the following command:

  `configure ports` *port_list* {**tagged** *tag*} **vlan** *vlan_name* | *vlan_list* [**limit-learning** *number* {**action** [**blackhole** | **stop-learning**]} | **lock-learning** | **unlimited-learning** | **unlock-learning**]

  This command causes all dynamic FDB entries associated with the specified *VLAN* and ports to be converted to locked static entries. It also sets the learning limit to 0, so that no new entries can be learned. All new source MAC addresses are blackholed.

  Note

  Blackhole FDB entries added due to MAC security violations on BlackDiamond 8800 series switches, SummitStack, and Summit family switches are removed after each FDB aging period regardless of whether the MAC addresses in question are still sending traffic. If the MAC addresses are still sending traffic, the blackhole entries will be re-added after they have been deleted.

  Locked entries do not get aged, but can be deleted like a regular permanent entry.

  For ports that have lock-down in effect, the following traffic still flows to the port:
  - Packets destined for the permanent MAC and other non-blackholed MAC addresses
  - Broadcast traffic
  - *EDP* traffic
  - Traffic from the permanent MAC still flows from the virtual port.
- Remove MAC address lockdown, use the **unlock-learning** option.

  `configure ports` *port_list* {**tagged** *tag*} **vlan** *vlan_name* | *vlan_list* [**limit-learning** *number* {**action** [**blackhole** | **stop-learning**]} | **lock-learning** | **unlimited-learning** | **unlock-learning**]

  When you remove the lockdown using the unlock-learning option, the learning-limit is reset to unlimited, and all associated entries in the FDB are flushed.
- Display the locked entries on the switch.

  `show fdb`

  Locked MAC address entries have the “l” flag.

## MAC Address Lockdown with Timeout

The MAC address lockdown with timeout feature provides a timer for aging out MAC addresses on a per port basis and overrides the *FDB* aging time. That is, when this feature is enabled on a port, MAC addresses learned on that port age out based on the MAC lockdown timeout corresponding to the port, not based on the FDB aging time. By default, the MAC address lockdown timer is disabled.

When this feature is enabled on a port, MAC addresses learned on that port remain locked for the MAC lockdown timeout duration corresponding to the port, even when the port goes down. As a result, when a device is directly connected to the switch and then disconnected, the MAC address corresponding to the device will be locked up for the MAC lockdown timeout duration corresponding to that port. If the

same device reconnects to the port before the MAC lockdown timer expires and sends traffic, the stored MAC address becomes active and the MAC lockdown timer is restarted. If the device is not reconnected for the MAC lockdown timeout duration, the MAC entry is removed.

MAC lockdown timeout entries are dynamically learned by the switch, which means these entries are not saved or restored during a switch reboot. If the switch reboots, the local MAC entry table is empty, and the switch needs to relearn the MAC addresses.

MAC address lockdown with timeout is configured by individual ports. The lockdown timer and address learning limits are configured separately for a port.

Note

You cannot enable the lockdown timeout feature on a port that already has MAC address lockdown enabled. For more information about MAC address lockdown, see MAC Address Lockdown on page 874.

MAC address learning limits and the lockdown timer work together in the following ways:

- When the learning limit has been reached on a port, a new device attempting to connect to the port has its MAC address blackholed.
- As long as the timer is still running for a MAC entry, a new device cannot connect in place of the device that entry represents. That is, if a device has disconnected from a port, a new device cannot replace it until the lockdown timer for the first device has expired. This condition is true if the limit on the port is set to 1 or if the limit (greater than 1) on the port has been reached.
- If a learning limit is already configured on a port when you enable the lockdown timeout feature, the configured limit will continue to apply. Existing blackholed entries are therefore not affected. If you enable this feature on a port with no configured learning limit, the default maximum learning limit (unlimited learning) is used.

### *Understanding the Lockdown Timer*

The lockdown timer works in the following ways:

- When you enable this feature on a port, existing MAC entries for the port begin aging out based on the configured MAC lockdown timer value.
- If you move a device from one port to another, its MAC address entry is updated with the new port information, including the lockdown timer value configured for that port.
- If this feature is enabled on a port and you decrease the lockdown timer value for that port, all of the MAC *FDB* entries for that port will time out and be removed at the next polling interval.
- When you disable the lockdown timer on a port, existing MAC address entries for the port will time out based on the FDB aging period.

### *Examples of Active and Inactive Devices*

The following figure shows three devices (A, B, and C) connected through a hub to an Extreme Networks device with MAC lockdown timeout configured on the ports.

When each device starts sending traffic, the source MAC address of the device is learned and *FDB* entries are created. The MAC lockdown timer is set at 100 seconds.

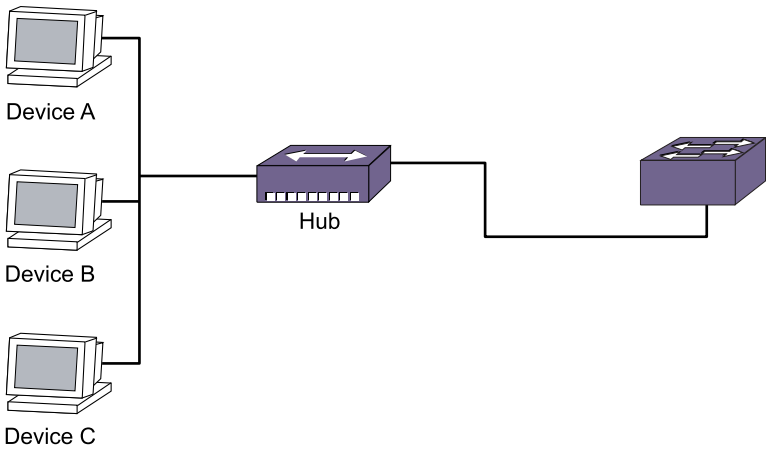



**Figure 108: Devices Using MAC Address Lockdown**

#### Device Inactivity for Less than the MAC Lockdown Timer

As long as a device continues to send traffic, the MAC entry for that device is refreshed, and the MAC lockdown timer corresponding to the MAC entry is refreshed.

Therefore, as long as the device is active, the timer does not expire. The traffic can be continuous or can occur in bursts within the MAC lockdown timeout duration for the port.

In this example, Device A starts sending traffic. When the MAC address of Device A is learned and added to the FDB, the MAC lockdown timer is started for this entry.

Device A stops sending traffic and resumes sending traffic after 50 seconds have elapsed. At this point the MAC entry for Device A is refreshed and the MAC lockdown timer is restarted.

#### Device Inactivity for Longer than the MAC Lockdown Timer

When a device stops sending traffic and does not resume within the MAC lockdown timer interval for the port, the MAC lockdown timer expires, and the MAC entry is removed from the FDB.

In this example, Devices A, B, and C start sending traffic. As each MAC address is learned, the MAC lockdown timer is started for each entry.

Device A stops sending traffic; Devices B and C continue sending traffic. After 100 seconds, the MAC lockdown timer for the Device A entry is removed from the FDB. Because Devices B and C have continued to send traffic, their MAC entries continue to be refreshed and their MAC lockdown timers continue to be restarted.

### *Examples of Disconnecting and Reconnecting Devices*

The following figure shows Device A connected to an Extreme Networks device with MAC lockdown timeout configured for the ports.

sWhen Device A starts sending traffic, the source MAC address is learned on the port, the *FDB* entry is created, and the MAC lockdown timer is started for the entry. The MAC lockdown timer is set at 3,000 seconds.

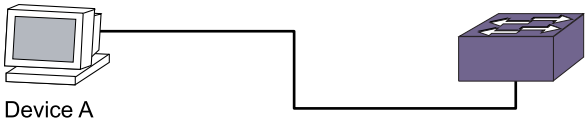



**Figure 109: Single Device with MAC Lockdown Timeout**

#### Disconnecting a Device

In this example, Device A is disconnected from the port, triggering a port-down action.

The MAC entry for Device A is removed from the hardware FDB; however, the MAC entry for the device is maintained in the software. The MAC lockdown timer for this entry starts when the port goes down.

After 3,000 seconds, the MAC entry for Device A is removed from the software.

#### Disconnecting and Reconnecting a Device

When Device A is disconnected from the port, the resulting port-down action causes the MAC entry for Device A to be removed from the hardware FDB.

The MAC entry in software is maintained, and the MAC lockdown timer is started for the port.

After only 1,000 seconds have elapsed, Device A is reconnected to the same port and starts sending traffic. A MAC entry is created in the hardware FDB, and the MAC lockdown timer is restarted for the MAC entry in the software.

If Device A is reconnected but does not send any traffic for 3,000 seconds, no MAC entry is created in the hardware FDB, and the MAC lockdown timer will expire after reaching 3,000 seconds.

#### Disconnecting and Reconnecting Devices with MAC Limit Learning

In this example, a MAC learning limit of 1 has also been configured on the ports in addition to the MAC lockdown timer of 3000 seconds.

When Device A is disconnected, the resulting port-down action removes the MAC entry for Device A from the hardware FDB. The MAC entry for Device A is maintained in the software, and the MAC lockdown timer for this entry is restarted when the port goes down.

After 1000 seconds, a different device is connected to the same port and starts sending traffic. Because the MAC learning limit is set to 1 and the MAC lockdown timer is still running, the MAC address of the new device is not learned. Instead, the new MAC address is blackholed in the hardware.

When the MAC lockdown timer for Device A expires, its MAC entry is removed from the software. If the new device is still connected to the same port and sends traffic, the MAC address for the new device is learned and added to the FDB. The MAC lockdown timer for the new device is started, and the blackhole entry that was created for this device is deleted.

### *Example of Port Movement*

The following figure shows Device A connected to port X.

Port X has a MAC lockdown timer setting of 100 seconds, and port Y has a MAC lockdown timer setting of 200 seconds.

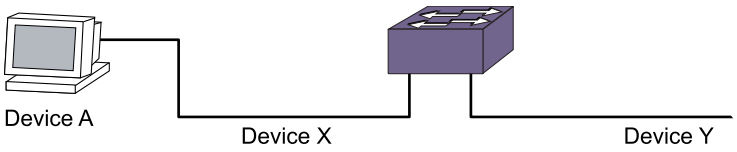



**Figure 110: Port Movement with MAC Lockdown Timeout**

Device A starts sending traffic on port X. The MAC address for Device A is learned and added to the *FDB*, and the MAC lockdown timer (100 seconds) is started for this entry.

After 50 seconds, Device A is disconnected from port X and connected to port Y where it begins sending traffic. When Device A starts sending traffic on port Y, the existing MAC entry for Device A is refreshed, and port X in the entry is replaced with port Y. At the same time, the MAC lockdown timer for the entry is restarted for a duration of 200 seconds (the configured MAC lockdown timer setting for port Y).

*Configuring MAC Address Lockdown with Timeout*

- Configure the MAC lockdown timeout value on one or more specified ports, or on all ports using the following command:

  `configure mac-lockdown-timeout ports [`**`all`**` | `*`port_list`*`] `**`aging-time`** *`seconds`*

*Enabling and Disabling MAC Address Lockdown with Timeout*

- Enable the MAC lockdown timeout feature on one or more specified ports, or on all ports. using the command:

  `enable mac-lockdown-timeout ports [`**`all`**` | `*`port_list`*`]`
- Disable the MAC lockdown timeout feature on one or more specified ports, or on all ports using the command:

  `disable mac-lockdown-timeout ports [`**`all`**` | `*`port_list`*`]`

*Displaying MAC Address Lockdown Information*

- Display configuration information about the MAC lockdown timeout feature using the command:

  `show mac-lockdown-timeout ports [`**`all`**` | `*`port_list`*`]`

  Output from this command includes the configured timeout value and whether the feature is enabled or disabled.
- Display the MAC entries learned on one or more ports, or on all ports using the command:

  `show mac-lockdown-timeout fdb ports [`**`all`**` | `*`port_list`*`]`

  Output from this command also lists the aging time of the port.

## DHCP Server

ExtremeXOS has *DHCP (Dynamic Host Configuration Protocol)* support.

In simple terms, a DHCP server dynamically manages and allocates IP addresses to clients. When a client accesses the network, the DHCP server provides an IP address to that client. The client is not required to receive the same IP address each time it accesses the network. A DHCP server with limited configuration capabilities is included in the switch to provide IP addresses to clients.

## Enabling and Disabling DHCP

*DHCP* is enabled on a per port, per *VLAN* basis.

- Enable or disable DHCP on a port in a VLAN using the commands:

  `enable dhcp ports` *port_list* **vlan** *vlan_name*

  `disable dhcp ports` *port_list* **vlan** *vlan_name*

## Configuring the DHCP Server

The following commands allow you to configure the *DHCP* server included in the switch. The parameters available to configure include the IP address range, IP address lease, and multiple DHCP options. Until EXOS 15.1, the DHCP server had a limited set of known DHCP options it could send out on request, i.e., Default Gateway, DNS, and WINS server(s). General option support has been added in EXOS 15.2. This allows you to add support for any option needed, with no EXOS code changes. The three options mentioned above can also be overwritten to support a larger number of servers, if needed. This feature allows the switch administrator to add an option based on DHCP option code value, and support various was of setting the value.

- Configure the range of IP addresses assigned by the DHCP server using the command:

  `configure vlan` *vlan_name* **dhcp-address-range** *ipaddress1* **-** *ipaddress2*

- Remove the address range information using the command:

  `unconfigure vlan` *vlan_name* **dhcp-address-range**

- Set how long the IP address lease assigned by the server exists using the command:

  `configure vlan` *vlan_name* **dhcp-lease-timer** *lease-timer*

  Note

  The ExtremeXOS DHCP server allows the configuration of a DHCP lease timer value greater than two seconds only. The timer value range is 3–4294967295. If the DHCP lease timer is not configured, the ExtremeXOS DHCP server offers an IP address with the default lease time of 7200 seconds.

- To set the generic DHCP option code, default gateway, Domain Name Servers (DNS) addresses, or Windows Internet Naming Service (WINS) server, use the following command:

  `configure {`**vlan**`}` *vlan_name* **dhcp-options** `[`**code** *option_number* `[`**16-bit** *value1* `{`*value2* `{`*value3* `{`*value4*`}}}` | **32-bit** *value1* `{`*value2* `{`*value3* `{`*value4*`}}}` | **flag** `[`**on** | **off**`]` | **hex** *string_value* | **ipaddress** *ipaddress1* `{`*ipaddress2* `{`*ipaddress3* `{`*ipaddress4*`}}}` | **string** *string_value*`]` | **default-gateway** | **dns-server** `{`**primary** | **secondary**`}` | **wins-server**`]` *ipaddress*

- To remove the generic DHCP option code, default gateway, DNS server addresses, and WINS server information for a particular *VLAN*, use the following command:

  `unconfigure {`**vlan**`}` *vlan_name* **dhcp-options** `{[` **default-gateway** | **dns-server** `{`**primary** | **secondary**`}` | **wins-server**`]}`

- Remove all the DHCP information for a particular VLAN using the command:

  `unconfigure vlan` *vlan_name* `dhcp`

You can clear the DHCP address allocation table selected entries, or all entries.

- Clear entries using the command:

  `clear vlan vlan_name dhcp-address-allocation [[all {offered | assigned | declined | expired}] | ipaddress]`

  You would use this command to troubleshoot IP address allocation on the VLAN.

## Displaying DHCP Information

The last two commands were retained for compatibility with earlier versions of ExtremeWare.

- Display the *DHCP* configuration, including the DHCP range, DHCP lease timer, network login lease timer, DHCP-enabled ports, IP address, MAC address, and time assigned to each end device using the command:

  `show dhcp-server {vlan vlan_name}`

- View only the address allocation of the DHCP server on a *VLAN* using the command:

  `show {vlan} vlan_name dhcp-address-allocation`

- View only the configuration of the DHCP server on a VLAN using the command:

  `show {vlan} vlan_name dhcp-config`

# IP Security

This section describes a collection of IP security features implemented in ExtremeXOS software.

If you configure any of the features described in this section, you can enhance your network security by controlling which hosts are granted or not granted access to your network.

The following figure displays the dependencies of IP security. Any feature that appears directly above another feature depends on it. For example, to configure ARP validation, you must configure *DHCP* snooping and trusted DHCP server.

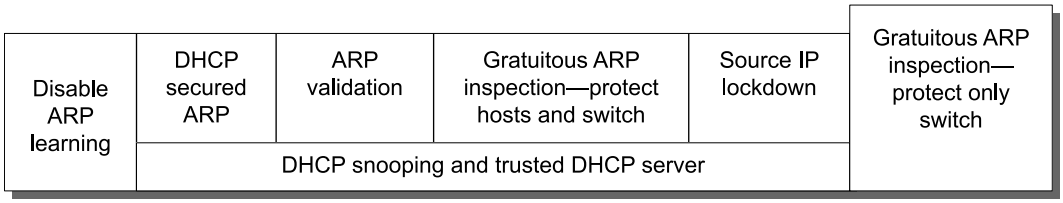



**Figure 111: IP Security Dependencies**

Note

IP security features are supported on link aggregation ports with the exception of DHCP snooping with the **block-mac** option and source IP lockdown. You can enable IP security on pre-existing trunks, but you cannot make IP security-enabled ports into trunks without first disabling IP security.

Enabling IP security and Network Login on the same port is not supported.

## DHCP Snooping and Trusted DHCP Server

A fundamental requirement for most of the IP security features described in this section is to configure *DHCP* snooping and trusted DHCP server.

DHCP snooping enhances security by filtering untrusted DHCP messages and by building and maintaining a DHCP bindings database. Trusted DHCP server also enhances security by forwarding DHCP packets from only configured trusted servers within your network.

The DHCP bindings database contains the IP address, MAC Address, *VLAN* ID, and port number of the untrusted interface or client. If the switch receives a DHCP ACK message and the IP address does not exist in the DHCP bindings database, the switch creates an entry in the DHCP bindings database. If the switch receives a DHCP RELEASE, NAK or DECLINE message and the IP address exists in the DHCP bindings database, the switch removes the entry.

You can enable DHCP snooping on a per port, per VLAN basis and trusted DHCP server on a per-vlan basis. If configured for DHCP snooping, the switch snoops DHCP packets on the indicated ports and builds a DHCP bindings database of IP address and MAC address bindings from the received packets. If configured for trusted DHCP server, the switch forwards only DHCP packets from the trusted servers. The switch drops DHCP packets from other DHCP snooping-enabled ports.

In addition, to prevent rogue DHCP servers from farming out IP addresses, you can optionally configure a specific port or set of ports as trusted ports. Trusted ports do not block traffic; rather, the switch forwards any DHCP server packets that appear on trusted ports. When configured to do so, the switch drops packets from DHCP snooping-enabled ports and causes one of the following user-configurable actions: disables the port temporarily, disables the port permanently, blocks the violating MAC address temporarily, blocks the violating MAC address permanently, and so on.

### *Configuring DHCP Snooping*

*DHCP* snooping is disabled on the switch by default.

- To enable DHCP snooping on the switch, use the command:
  `enable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports] violation-action [drop-packet {[block-mac | block-port] [duration duration_in_seconds | permanently] | none}] {snmp-trap}`

  Note
  Snooping IP fragmented DHCP packets is not supported.

  The violation action setting determines what action(s) the switch takes when a rogue DHCP server packet is seen on an untrusted port or the IP address of the originating server is not among those of the configured trusted DHCP servers.

  The DHCP server packets are DHCP OFFER, ACK and NAK. The following list describes the violation actions:

  block-mac

  The switch automatically generates an *ACL* to block the MAC address on that port. The switch does not blackhole that MAC address in the *FDB*. The switch can either temporarily or permanently block the MAC address.

  block-port

  The switch blocks all traffic on that port by disabling the port either temporarily or permanently.

none

The switch takes no action to drop the rogue DHCP packet or block the port, and so on. In this case, DHCP snooping continues to build and manage the DHCP bindings database and DHCP forwarding will continue in hardware as before. This option can be used when the intent is only to monitor the IP addresses being assigned by the DHCP server.

Note

You must enable DHCP snooping on both the DHCP server port as well as on the client port. The latter ensures that DHCP client packets (DHCP Request, DHCP Release etc.) are processed appropriately.

Note

DHCP snooping does not work when the client and server are in different VRs and server reachability is established by inter-VR leaked routes on client VR.

Any violation that occurs causes the switch to generate an *EMS (Event Management System)* log message. You can configure to suppress the log messages by configuring EMS log filters. For more information about EMS, see Using the Event Management System/Logging on page 470.

- To disable DHCP snooping on the switch, use the command:

  `disable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports]`

*Configuring Trusted DHCP Server*

- Configure a trusted *DHCP* server on the switch using the command:

  `configure trusted-servers {vlan} vlan_name add server ip_address trust-for dhcp-server`

  If you configure one or more trusted ports, the switch assumes that all DHCP server packets on the trusted port are valid. You can configure a maximum of eight trusted DHCP servers on the switch.

  For more information about configuring trusted ports, see the next section Configure Trusted DHCP Ports.

- Delete a trusted DHCP server using the command:

  `configure trusted-servers vlan vlan_name delete server ip_address trust-for dhcp-server`

*Configuring Trusted DHCP Ports*

Trusted ports do not block traffic; rather, the switch forwards any *DHCP* server packets that appear on trusted ports. Depending on your DHCP snooping configuration, the switch drops packets and can disable the port temporarily, disable the port permanently, block the MAC address temporarily, block the MAC address permanently, and so on.

- Enable trusted ports on the switch using the command:

  `configure trusted-ports [ports|all] trust-for dhcp-server`

- Disable trusted ports on the switch using the command:

  `unconfigure trusted-ports [ports |all] trust-for dhcp-server`

### *Displaying DHCP Snooping and Trusted Server Information*

- Display the *DHCP* snooping configuration settings using the command:

  `show ip-security dhcp-snooping {vlan} vlan_name`

  Below is sample output from this command:

```
DHCP Snooping enabled on ports: 1:2, 1:3, 1:4, 1:7, 1:9
Trusted Ports: 1:7
Trusted DHCP Servers: None
-----------------------------------------
Port            Violation-action
-----------------------------------------
1:2             none
1:3             drop-packet
1:4             drop-packet, block-mac permanently
1:7             none
1:9             drop-packet, snmp-trap
```

- Display the DHCP bindings database using the command:

  `show ip-security dhcp-snooping entries {vlan} vlan_name`

  Below is sample output from this command:

```
-----------------------------------------
Vlan: dhcpVlan
-----------------------------------------
Server    Client
IP Addr         MAC Addr            Port      Port
-------         --------            ------    ------
172.16.100.9    00:90:27:c6:b7:65   1:1       1:2
```

### *Clearing DHCP Snooping Entries*

- To clear Existing *DHCP* snooping entries, use the command:

  `clear ip-security dhcp-snooping entries {vlan} vlan_name`

> **Note**
> This will also clear out any associated source IP lockdown and DHCP secured ARP entries.

### *Configuring the DHCP Relay Agent Option (Option 82) at Layer 2*

This section describes how to configure the *DHCP* Relay agent option for Layer 2 forwarded DHCP packets.

The DHCP relay agent option feature inserts a piece of information, called option 82, into any DHCP request packet that is to be relayed by the switch. Similarly, if a DHCP reply received by the switch contains a valid relay agent option, the option will be stripped from the packet before it is relayed to the client. This is a Layer 2 option that functions only when the switch is not configured as a Layer 3 BOOTP relay.

The Agent remote ID sub-option always contains the Ethernet MAC address of the relaying switch. You can display the Ethernet MAC address of the switch by issuing the show switch command.

The contents of the inserted option 82 sub-options is as follows:

**Table 107: Contents of the Inserted Option 82 Sub-options**

| Code (1 byte) | Length (1 byte) | Sub-Option (1 byte) | Length (1 byte) | Value (1-32 bytes) | Sub-Option (1 byte) | Length (1 byte) | Switch MAC address (6 bytes) |
|---|---|---|---|---|---|---|---|
| 82 | | 1 (Circuit ID) | 1-32 | vlan_info-port_info | 2 (Remote ID) | 6 | |

To enable the DHCP relay agent option at Layer 2, use the following command:

```
configure ip-security dhcp-snooping information option
```

Note

When DHCP relay is configured in a DHCP snooping environment, the relay agent IP address should be configured as the trusted server.

When DHCP option 82 is enabled, two types of packets need to be handled:

- DHCP Request: When the switch (relay agent) receives a DHCP request, option 82 is added at the end of the packet. If the option has already been enabled, then the action taken depends on the configured policy (drop packet, keep existing option 82 value, or replace the existing option). Unless configured otherwise using the `configure ip-security dhcp-snooping information circuit-id vlan-information` *`vlan_info`* {**`vlan`**} [*`vlan_name`* | **`all`**] , the vlan_info portion of the circuit ID added will be the *VLAN* ID of the ingress VLAN.
- DHCP Reply: When the option 82 information check is enabled, the packets received from the DHCP server are checked for option 82 information. If the remote ID sub-option is the switch's MAC address, the packet is sent to the client; if not, the packet is dropped. If the check is not enabled. The packets are forwarded as-is.

To disable the DHCP relay agent option, use the following command:

```
unconfigure ip-security dhcp-snooping information option
```

In some instances, a DHCP server may not properly handle a DHCP request packet containing a relay agent option.

To prevent DHCP reply packets with invalid or missing relay agent options from being forwarded to the client, use the following command:

```
configure ip-security dhcp-snooping information check
```

To disable checking of DHCP replies, use this command:

```
unconfigure ip-security dhcp-snooping information check
```

A DHCP relay agent may receive a client DHCP packet that has been forwarded from another relay agent.

If this relayed packet already contains a relay agent option, then the switch will handle this packet according to the configured DHCP relay agent option policy. The possible actions are to replace the option information, to keep the information, or to drop packets containing option 82 information. To configure this policy, use the following command:

```
configure ip-security dhcp-snooping information policy [drop | keep |
replace]
```

The default relay policy is replace.

To configure the policy to the default, use this command:

```
unconfigure ip-security dhcp-snooping information policy
```

The Layer 2 relay agent option allows you to configure the circuit ID on a VLAN or port basis., the Circuit-ID can contain a variable length (up to 32 bytes long) ASCII string with the following format:

```
<VLAN Info>-<Port Info>
```

If the configuration of either VLAN Info or Port Info causes the total string length of <VLAN Info>-<Port Info> to exceed 32 bytes, then it is truncated to 32 bytes. The string is not NULL terminated, since the total circuit ID length is being specified.

For a DHCP client packet ingressing on a VLAN with the VLAN ID equal to 200 and the ingress port at 3:5, the following are true:

- When neither VLAN Info or Port Info is specified, circuit ID value is = 200-3005
- When VLAN Info is configured to SomeInfo and Port Info is not specified, the circuit ID value is SomeInfo-3005
- When VLAN Info is not specified and Port Info is configured to User1, the circuit ID value is 200-User1
- When VLAN Info is configured to SomeInfo and Port Info to User1, the circuit ID value is SomeInfo-User1

VLAN Info is configurable per VLAN.

When not explicitly configured for a VLAN, VLAN Info defaults to the ASCII string representation of the ingress VLAN ID. To configure the circuit ID on a VLAN, use the following command:

```
configure ip-security dhcp-snooping information circuit-id vlan-
information vlan_info {vlan} [vlan_name | all]
```

To unconfigure the circuit ID on a VLAN, use the following command:

```
unconfigure ip-security dhcp-snooping information circuit-id vlan-
information {vlan} [vlan_name|all]
```

Port Info is configurable.

When not explicitly configured for a port, port info defaults to the ASCII representation of the ingress port's *SNMP* ifIndex. To configure the port information portion of the circuit-ID, use the following command:

```
configure ip-security dhcp-snooping information circuit-id port-
information port_info port port
```

To unconfigure the port information portion of the circuit-ID, use the following command:

```
unconfigure ip-security dhcp-snooping information circuit-id port-
information ports [port_list | all]
```

**Note**

When this feature is enabled, all DHCP traffic must be forwarded in slowpath only, which means that this feature functions only in the context of IP Security and only on interfaces where DHCP snooping is enabled in enforcement (violation-action of 'drop') mode. In other words, with DHCP snooping not configured with a violation-action of 'none' (which is pure monitoring mode).

### *Configuring DHCP Binding*

The *DHCP* bindings database contains the IP address, MAC address, *VLAN* ID, and port number of the client. You can add or delete the static IP to the MAC DHCP binding entries using the following commands:

```
configure ip-security dhcp-bindings add
```

```
configure ip-security dhcp-bindings delete
```

You can specify the storage details of the DHCP binding database. Use the following commands to specify the DHCP binding database location, filename, write-interval, and write threshold limits:

```
configure ip-security dhcp-bindings storage filename
```

```
configure ip-security dhcp-bindings storage location
```

```
configure ip-security dhcp-bindings storage
```

You can upload the DHCP binding database periodically by enabling the DHCP binding restoration. Binding write intervals occur in minutes, 5 to 1440 (24 hours). The default is 30 minutes.

Upload the latest DHCP binding database using the upload command:

```
enable ip-security dhcp-bindings restoration
```

You can also upload the DHCP binding database by the number of DHCP entries (the write-threshold is 25 to 200).

The periodic backup of the DHCP binding database can be disabled using the following command:

```
disable ip-security dhcp-bindings restoration
```

For information about configuring option 82 at Layer 3, see Configuring the DHCP Relay Agent Option (Option 82) at Layer 3 on page 1283.

#### Example of Option 82 Configuration

The following example describes Option 82 configuration for circuit ID fields.

```
create vlan v1
conf v1 add ports 21
enable ip-security dhcp-snooping v1 ports all violation-action drop-packet
configure trusted-ports 21 trust-for dhcp-server
conf ip-security dhcp-snooping information option
conf ip-security dhcp-snooping information check
```

```
conf ip-security dhcp-snooping information circuit-id vlan-information ServiceProvider-1
v1
conf ip-security dhcp-snooping information circuit-id port-information cutomer-1 port 1
conf ip-security dhcp-snooping information circuit-id port-information cutomer-2 port 2
CLI display output
==================

* switch # sh ip-security dhcp-snooping v1
DHCP Snooping enabled on ports: 21
Trusted Ports: 21
Trusted DHCP Servers: None
Bindings Restoration     : Disabled
Bindings Filename        :
Bindings File Location   :
Primary Server  : None
Secondary Server: None
Bindings Write Interval  : 30 minutes
Bindings last uploaded at:
-----------------------------------
Port            Violation-action
-----------------------------------
21              drop-packet

* switch # show ip-security dhcp-snooping information-option
Information option insertion: Enabled
Information option checking : Enabled
Information option policy   : Replace
* switch #

* switch # sh ip-security dhcp-snooping information-option circuit-id vlan-information
Vlan            Circuit-ID vlan information string
----            ----------------------------------
Default         1 (Default i.e. vlan-id)
Mgmt            4095 (Default i.e. vlan-id)
v1              ServiceProvider-1
Note: The full Circuit ID string has the form '<Vlan Info>-<Port Info>'
* switch


* switch # sh ip-security dhcp-snooping information-option circuit-id port-information
ports all
Port            Circuit-ID Port information string
----            ----------------------------------
1               cutomer-1
2               cutomer-2
3               1003
4               1004
5               1005
6               1006
7               1007
8               1008
9               1009
10              1010
11              1011
12              1012
13              1013
14              1014
15              1015
16              1016
17              1017
18              1018
19              1019
20              1020
21              1021
22              1022
23              1023
24              1024
```

```
25              1025
26              1026
Note: The full Circuit ID string has the form '<Vlan Info>-<Port Info>'
* switch #
```

## Source IP Lockdown

Another type of IP security prevents IP address spoofing by automatically placing source IP address filters on specified ports. This feature, called *source IP lockdown*, allows only traffic from a valid *DHCP*-assigned address obtained by a DHCP snooping-enabled port to enter the network. In this way, the network is protected from attacks that use random source addresses for their traffic. With source IP lockdown enabled, end systems that have a DHCP address assigned by a trusted DHCP server can access the network, but traffic from others, including those with static IP addresses, is dropped at the switch.

Source IP lockdown is linked to the “DHCP snooping” feature. The same DHCP bindings database created when you enable DHCP snooping is also used by source IP lockdown to create ACLs that permit traffic from DHCP clients. All other traffic is dropped. In addition, the DHCP snooping violation action setting determines what action(s) the switch takes when a rogue DHCP server packet is seen on an untrusted port.

When source IP lockdown is enabled on a port, a default *ACL* is created to deny all IP traffic on that port. Then an ACL is created to permit DHCP traffic on specified ports. Each time source IP lockdown is enabled on another port, the switch creates ACLs to allow DHCP packets and to deny all IP traffic for that particular port.

Source IP lockdown is enabled on a per-port basis; it is not available at the *VLAN* level. If source IP lockdown is enabled on a port, the feature is active on the port for all VLANs to which the port belongs.

**Note**
The source IP lockdown feature works only when hosts are assigned IP address using DHCP; source IP lockdown does not function for statically configured IP addresses.

The source IP lockdown ACLs listed in table are applied per port (in order of precedence from highest to lowest).

**Table 108: Source IP Lockdowns Applied Per-port**

| ACL Name | Match Condition | Action | When Applied | Comments |
|---|---|---|---|---|
| esSrcIpLockdown_<portIfIndex>_<source IP in hex> | Source IP | Permit | Runtime | Multiple ACLs of this type can be applied, one for each permitted client. |
| esSrcIpLockdown_<portIfIndex>_1 | Proto UDP, Dest Port 67 | Permit | Configuration time | |
| esSrcIpLockdown_<portIfIndex>_2 | Proto UDP, Dest Port 68 | Permit | Configuration time | |

**Table 108: Source IP Lockdowns Applied Per-port (continued)**

| ACL Name | Match Condition | Action | When Applied | Comments |
|---|---|---|---|---|
| esSrcIpLockdown_<portIfIndex>_3 | Ethertype ARP | Permit | Configuration time | |
| esSrcIpLockdown_<portIfIndex>_4 | All | Deny + count | Configuration time | |

The counter has the same name as that of the rule of the catch-all ACL, so the counter is also named esSrcIpLockdown_<portIfIndex>_4.

### *Configuring Source IP Lockdown*

To configure source IP lockdown, you must enable *DHCP* snooping on the ports connected to the DHCP server and DHCP client before you enable source IP lockdown. You must enable source IP lockdown on the ports connected to the DHCP client, not on the ports connected to the DHCP server.

- Enable DHCP snooping using the command:
  `enable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports] violation-action [drop-packet {[block-mac | block-port] [duration duration_in_seconds | permanently] | none]}] {snmp-trap}`
  For more information about DHCP snooping see, Configuring DHCP Snooping on page 881.

  Source IP lockdown is disabled on the switch by default.
- To enable source IP lockdown, use the command:
  `enable ip-security source-ip-lockdown ports [all | ports]`
- To disable source IP lockdown, use the command
  `disable ip-security source-ip-lockdown ports [all | ports]`

### *Displaying Source IP Lockdown Information*

- Display the source IP lockdown configuration on the switch using the ocmmand:.
  `show ip-security source-ip-lockdown`
  Below is sample output from this command:

```
Ports           Locked IP Address
23              10.0.0.101
```

### *Clear Source IP Lockdown Information*

- Remove existing source IP lockdown entries on the switch using the command:
  `clear ip-security source-ip-lockdown entries ports [ports | all]`

## ARP Learning

The address resolution protocol (ARP) is part of the TCP/IP suite used to dynamically associate a device's physical address (MAC address) with its logical address (IP address).

The switch broadcasts an ARP request that contains the IP address, and the device with that IP address sends back its MAC address so that traffic can be transmitted across the network. The switch maintains

an ARP table (also known as an ARP cache) that displays each MAC address and its corresponding IP address.

By default, the switch builds its ARP table by tracking ARP requests and replies, which is known as ARP learning. You can disable ARP learning so that the only entries in the ARP table are either manually added or those created by *DHCP* secured ARP; the switch does not add entries by tracking ARP requests and replies. By disabling ARP learning and adding a permanent entry or configuring DHCP secured ARP, you can centrally manage and allocate client IP addresses and prevent duplicate IP addresses from interrupting network operation.

### *Configuring ARP Learning*

As previously described, ARP learning is enabled by default. The switch builds its ARP table by tracking ARP requests and replies.

- Disable ARP learning on one or more ports in a *VLAN* using the command:

  `disable ip-security arp learning learn-from-arp {vlan} vlan_name ports [all | ports]`

- Re-enable ARP learning on one or more ports in a VLAN using the command:

  `enable ip-security arp learning learn-from-arp {vlan} vlan_name ports [all | ports]`

### *Adding a Permanent Entry to the ARP Table*

If you disable ARP learning, you must either manually add a permanent entry to the ARP table or configure *DHCP* secured ARP to populate the ARP table.

- Manually add a permanent entry to the ARP table using the command:

  `configure iparp add ip_addr {vr vr_name} mac`

For more detailed information about this command and IP routing, see IPv4 Unicast Routing on page 1243.

### *Configuring DHCP Secured ARP*

Before you configure *DHCP* secured ARP, you must enable DHCP snooping on the switch.

Another method available to populate the ARP table is DHCP secured ARP. DHCP secured ARP requires that ARP entries be added to or deleted from the ARP table only when the DHCP server assigns or re-assigns an IP address. These entries are known as a secure ARP entry. If configured, the switch adds the MAC address and its corresponding IP address to the ARP table as a permanent ARP entry. Regardless of other ARP requests and replies seen by the switch, the switch does not update secure ARP entries. DHCP secured ARP is linked to the “DHCP snooping” feature. The same DHCP bindings database created when you enabled DHCP snooping is also used by DHCP secured ARP to create secure ARP

entries. The switch only removes secure ARP entries when the corresponding DHCP entry is removed from the trusted DHCP bindings database.

Note
If you enable DHCP secured ARP on the switch without disabling ARP learning, ARP learning continues which allows insecure entries to be added to the ARP table.

- Enable DHCP snooping using the command:

  `enable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports] violation-action [drop-packet {[block-mac | block-port] [duration duration_in_seconds | permanently] | none]}] {snmp-trap}`

  For more information about DHCP snooping see, Configuring DHCP Snooping on page 881.

  DHCP secured ARP learning is disabled by default.
- Enable DHCP secured ARP using the command:

  `enable ip-security arp learning learn-from-dhcp {vlan} vlan_name ports [all | ports]`

  DHCP Secured ARP must be enabled on the DHCP server port as well as the DHCP client ports.
- Disable DHCP secured ARP using the command:

  `disable ip-security arp learning learn-from-dhcp {vlan} vlan_name ports [all | ports]`

Note
You must enable DHCP secured ARP on the DHCP server as well as on the client ports. DHCP snooping, as always, must also be enabled on both the server and client ports.

*Displaying ARP Information*

- Display how the switch builds an ARP table and learns MAC addresses for devices on a specific *VLAN* and associated member ports using the command:.

  `show ip-security arp learning {vlan} vlan_name | vlan_list`

  Below is sample output from this command:

```
Port                Learn-from
---------------------------------
2:1                 ARP
2:2                 DHCP
2:3                 ARP
2:4                 None
2:5                 ARP
2:6                 ARP
2:7                 ARP
2:8                 ARP
```

- View the ARP table, including permanent and DHCP secured ARP entries using the command:

  `show iparp {ip_address | mac | vlan {vlan_name | vlan_list} | permanent} {vr vr_name}`

Note
DHCP secured ARP entries are stored as static entries in the ARP table.

## Gratuitous ARP Protection

When a host sends an ARP request to resolve its own IP address it is called *gratuitous ARP*. A gratuitous ARP request is sent with the following parameters:

- Destination MAC address—FF:FF:FF:FF:FF:FF (broadcast)
- Source MAC address—Host's MAC address
- Source IP address = Destination IP address—IP address to be resolved

In a network, gratuitous ARP is used to:

- Detect duplicate IP address.

  In a properly configured network, there is no ARP reply for a gratuitous ARP request. However, if another host in the network is configured with the same IP address as the source host, then the source host receives an ARP reply.
- Announce that an IP address has moved or bonded to a new network interface card (NIC).

  If you change a system NIC, the MAC address to its IP address mapping also changes. When you reboot the host, it sends an ARP request packet for its own IP address. All of the hosts in the network receive and process this packet. Each host updates their old mapping in the ARP table with this new mapping
- Notify a Layer 2 switch that a host has moved from one port to another port.

However, hosts can launch man-in-the-middle attacks by sending out gratuitous ARP requests for the router's IP address. This results in hosts sending their router traffic to the attacker, and the attacker forwarding that data to the router. This allows passwords, keys, and other information to be intercepted.

To protect against this type of attack, the router sends out its own gratuitous ARP request to override the attacker whenever a gratuitous ARP request broadcast packet with the router's IP address as the source is received on the network.

If you enable both *DHCP* secured ARP and gratuitous ARP protection, the switch protects its own IP address and those of the hosts that appear as secure entries in the ARP table.

### *Configuring Gratuitous ARP*

You enable the gratuitous ARP feature on a per *VLAN* basis, not on a per port basis. The validation is done for all gratuitous ARP packets received on a VLAN in which this feature is enabled irrespective of the port in which the packet is received.

When enabled, the switch generates gratuitous ARP packets when it receives a gratuitous ARP request where either of the following is true:

- The sender IP is the same as the switch VLAN IP address and the sender MAC address is not the switch MAC address.
- The sender IP is the same as the IP of a static entry in the ARP table and the sender MAC address is not the static entry's MAC address.

When the switch generates an ARP packet, the switch generates logs and traps.

- Enable gratuitous ARP protection using the command:

  `enable ip-security arp gratuitous-protection {vlan} [all | vlan_name]`

- In addition, to protect the IP addresses of the hosts that appear as secure entries in the ARP table, use the following commands to enable *DHCP* snooping, DHCP secured ARP, and gratuitous ARP on the switch:

  `enable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports] violation-action [drop-packet {[block-mac | block-port] [duration duration_in_seconds | permanently] | none}] {snmp-trap}`

  `enable ip-security arp learning learn-from-dhcp {vlan} vlan_name ports [all | ports]`

  `enable ip-security arp gratuitous-protection {vlan} [all | vlan_name]`

- Disable gratuitous ARP protection using the command:

  `disable ip-security arp gratuitous-protection {vlan} [all | vlan_name]`

- In ExtremeXOS 11.5 and earlier, you enable gratuitous ARP protection using the following command:

  `enable iparp gratuitous protect vlan vlan-name`

- In ExtremeXOS11.5 and earlier, you disable gratuitous ARP protection with the following command:

  `disable iparp gratuitous protect vlan vlan-name`

*Displaying Gratuitous ARP Information*

- Display information about gratuitous ARP using the command:

  `show ip-security arp gratuitous-protection`

  Below is sample output from this command:

```
Gratuitous ARP Protection enabled on following VLANs:
Default, test
```

## ARP Validation

ARP validation is also linked to the "*DHCP* snooping" feature. The same DHCP bindings database created when you enabled DHCP snooping is also used to validate ARP entries arriving on the specified ports.

| Validation Option | ARP Request Packet Type | ARP Response Packet Type |
|---|---|---|
| DHCP | | Source IP is not present in the DHCP snooping database OR is present but Source Hardware Address doesn't match the MAC in the DHCP bindings entry. |
| IP | Source IP == Mcast OR<br>Target IP == Mcast OR<br>Source IP is not present in the DHCP snooping database OR<br>Source IP exists in the DHCP bindings database but Source Hardware Address doesn't match the MAC in the DHCP bindings entry. | Source IP == Mcast OR<br>Target IP == Mcast |
| Source-MAC | Ethernet source MAC does not match the Source Hardware Address. | Ethernet source MAC does not match the Source Hardware Address. |
| Destination-MAC | | Ethernet destination MAC does not match the Target Hardware Address. |

Depending on the options specified when enabling ARP validation, the following validations are done. Note that the 'DHCP' option does not have to be specified explicitly, it is always implied when ARP validation is enabled.

### *Configuring ARP Validation*

Before you configure ARP validation, you must enable *DHCP* snooping on the switch.

- Enable DHCP snooping using the command:.

  `enable ip-security dhcp-snooping {vlan} vlan_name ports [all | ports] violation-action [drop-packet {[block-mac | block-port] [duration duration_in_seconds | permanently] | none]}] {snmp-trap}`

  For more information about DHCP snooping see, Configuring DHCP Snooping on page 881.

  ARP validation is disabled by default.

- Enable and configure ARP validation using the command:

  `enable ip-security arp validation {destination-mac} {source-mac} {ip} {vlan} vlan_name [all | ports] violation-action [drop-packet {[block-port] [duration duration_in_seconds | permanently]}] {snmp-trap}`

  The violation action setting determines what action(s) the switch takes when an invalid ARP is received.

Any violation that occurs causes the switch to generate an *EMS* log message. You can configure to suppress the log messages by configuring EMS log filters. For more information about EMS, see the section Using the Event Management System/Logging on page 470.

- Disable ARP validation using the command:

  `disable ip-security arp validation {vlan} vlan_name [all | ports]`

*Displaying ARP Validation Information*

- Display information about ARP validation using the command:

  `show ip-security arp validation {vlan} vlan_name`

  Below is sample output from this command:

```
-------------------------------------------------------------
Port    Validation              Violation-action
-------------------------------------------------------------
7       DHCP                    drop-packet, block-port for 120 seconds, snmp-trap
23      DHCP                    drop-packet, block-port for 120 seconds, snmp-trap
```

## Denial of Service Protection

A Denial-of-Service (DoS) attack occurs when a critical network or computing resource is overwhelmed and rendered inoperative in a way that legitimate requests for service cannot succeed.

In its simplest form, a DoS attack is indistinguishable from normal heavy traffic. There are some operations in any switch or router that are more costly than others, and although normal traffic is not a problem, exception traffic must be handled by the switch's CPU in software.

Some packets that the switch processes in the CPU software include:

- Traffic resulting from new MAC learning (Only the BlackDiamond 8800 and the Summit family switches learn in hardware.)

Note

When certain features such as Network Login are enabled, hardware learning is disabled to let software control new MAC learning.

- Routing and control protocols including *ICMP*, *BGP (Border Gateway Protocol)*, *OSPF (Open Shortest Path First)*, *STP (Spanning Tree Protocol)*, EAPS, *ESRP*, etc.
- Switch management traffic (switch access by Telnet, SSH, HTTP, *SNMP*, etc.)
- Other packets directed to the switch that must be discarded by the CPU

If any one of these functions is overwhelmed, the CPU may be too busy to service other functions and switch performance will suffer. Even with very fast CPUs, there will always be ways to overwhelm the CPU with packets that require costly processing.

DoS Protection is designed to help prevent this degraded performance by attempting to characterize the problem and filter out the offending traffic so that other functions can continue. When a flood of CPU bound packets reach the switch, DoS Protection will count these packets. When the packet count nears the alert threshold, packets headers will be saved. If the threshold is reached, then these headers are analyzed, and a hardware *ACL* is created to limit the flow of these packets to the CPU. This ACL will remain in place to provide relief to the CPU. Periodically, the ACL will expire, and if the attack is still

occurring, it will be re-enabled. With the ACL in place, the CPU will have the cycles to process legitimate traffic and continue other services.

Note
User-created ACLs take precedence over the automatically applied DoS protect ACLs.

DoS Protection will send a notification when the notify threshold is reached.

You can also specify some ports as trusted ports, so that DoS protection will not be applied to those ports.

## Configuring Simulated Denial of Service Protection

The conservative way to deploy DoS protection is to use the simulated mode first. In simulated mode, DoS protection is enabled, but no *ACL*s are generated.

- Enable the simulated mode using the command:
  `enable dos-protect simulated`

  This mode is useful to gather information about normal traffic levels on the switch. This will assist in configuring denial of service protection so that legitimate traffic is not blocked.

  The following topics describe how to configure DoS protection, including alert thresholds, notify thresholds, ACL expiration time, and so on.

## Configuring Denial of Service Protection

- Enable or disable DoS protection using the command:
  `enable dos-protect`
  `disable dos-protect`

  After enabling DoS protection, the switch will count the packets handled by the CPU and periodically evaluate whether to send a notification and/or create an *ACL* to block offending traffic.

  You can configure a number of the values used by DoS protection if the default values are not appropriate for your situation.

  The values that you can configure are:
  - interval—How often, in seconds, the switch evaluates the DoS counter (default: 1 second)
  - alert threshold—The number of packets received in an interval that will generate an ACL (default: 4000 packets)
  - notify threshold—The number of packets received in an interval that will generate a notice (default: 3500 packets)
  - ACL expiration time—The amount of time, in seconds, that the ACL will remain in place (default: 5 seconds)
- Configure the interval at which the switch checks for DoS attacks using the command:
  `configure dos-protect interval` *`seconds`*
- Configure the alert threshold using the command:
  `configure dos-protect type l3-protect alert-threshold` *`packets`*

- Configure the notification threshold using the command:

  `configure dos-protect type l3-protect notify-threshold` *packets*
- Configure the ACL expiration time using the command:

  `configure dos-protect acl-expire` *seconds*

*Configuring Trusted Ports*

Traffic from trusted ports will be ignored when DoS protect counts the packets to the CPU. If we know that a machine connected to a certain port on the switch is a safe "trusted" machine, and we know that we will not get a DoS attack from that machine, the port where this machine is connected to can be configured as a trusted port, even though a large amount of traffic is going through this port.

- Configure the trusted ports list using the command:

  `configure dos-protect` **`trusted-ports`** `[`**`ports`** `[`*`ports`* `|` **`all`**`] |` **`add-ports`** `[`*`ports-to-add`* `|` **`all`**`] |` **`delete-ports`** `[`*`ports-to-delete`* `|` **`all`**`]]`

*Displaying DoS Protection Settings*

- Display the DoS protection settings using the command:

  `show dos-protect {`**`detail`**`}`

## Protocol Anomaly Protection

The Extreme chipsets contain built-in hardware protocol checkers that support port security features for security applications, such as stateless DoS protection.

The protocol checkers allow users to drop the packets based on the following conditions, which are checked for ingress packets prior to the L2/L3 entry table:

- SIP = DIP for IPv4/IPv6 packets.
- TCP_SYN Flag = 0 for IPv4/IPv6 packets
- TCP Packets with control flags = 0 and sequence number = 0 for IPv4/IPv6 packets
- TCP Packets with FIN, URG & PSH bits set & seq. number = 0 for IPv4/IPv6 packets
- TCP Packets with SYN & FIN bits are set for IPv4/IPv6 packets
- TCP Source Port number = TCP Destination Port number for IPv4/IPv6 packets
- First TCP fragment does not have the full TCP header (less than 20 bytes) for IPv4/IPv6 packets
- TCP header has fragment offset value as 1 for IPv4/IPv6 packets
- UDP Source Port number = UDP Destination Port number for IPv4/IPv6 packets
- CMP ping packets payload is larger than programmed value of *ICMP* max size for IPv4/IPv6 packets
- Fragmented ICMP packets for IPv4/IPv6 packets

The protocol anomaly detection security functionality is supported by a set of anomaly-protection enable, disable, configure, clear, and show CLI commands. For further details, see the *ExtremeXOS 16.2 Command Reference Guide*.

## Flood Rate Limitation

Flood rate limitation, or storm control, is used to minimize the network impact of ingress flooding traffic. You can configure ports to accept a specified rate of packets per second. When that rate is exceeded,

the port blocks traffic and drops subsequent packets until the traffic again drops below the configured rate.

- Configure the rate limit.

  `configure ports port_list rate-limit flood [broadcast | multicast | unknown-destmac] [no-limit | pps]`

- Display rate limiting statistics.

  `show ports {port_list} rate-limit flood {no-refresh}`

Note

Summit X440, X460, X480, X670, and X770 ; 8900-MSM128, 8900-Series I/O modules, BDX-MM1, and BDX-Series I/O modules, implement rate limiting granularity at millisecond intervals. The traffic bursts are monitored at millisecond intervals and the actions are performed within sub-seconds (when applicable). When the switch evaluates the traffic pattern for bursts against the configured value in pps, the value is calibrated on a per-millisecond interval. For example, using the `configure port 1 rate-limit flood broadcast1000` command would be equivalent to 1 packet per millisecond.

## Authenticating Management Sessions Through a TACACS+ Server

You can use a Terminal Access Controller Access Control System Plus (TACACS+) server to authenticate management sessions for multiple switches.

A TACACS+ server allows you to centralize the authentication database, so that you do not have to maintain a separate local database on each switch. TACACS+ servers provide the following services:

- Username and password authentication
- Command authorization (the TACACS+ server validates whether the user is authorized to execute each command within the subset of commands, based on login privilege level)
- Accounting service (tracks authentication and authorization events)

Note

You can use a local database on each switch as a backup authentication service if the TACACS + service is unavailable. When the TACACS+ service is operating, privileges defined on the TACACS+ server take precedence over privileges configured in the local database.

To use TACACS+ server features, you need the following components:

- TACACS+ client software, which is included in the ExtremeXOS software.
- A TACACS+ server, which is a third-party product.

Note

TACACS+ provides many of the same features provided by *RADIUS (Remote Authentication Dial In User Service)*. You cannot use RADIUS and TACACS+ at the same time.

TACACS+ is a communications protocol that is used between client and server to implement the TACACS+ service. The TACACS+ client component of the ExtremeXOS software should be compatible with any TACACS+ compliant server product.

Note
The switch allows local authentication when the client IP is excluded in TACACS+ server by default. To disallow local authentication when the client IP is excluded in TACACS+ server the local authentication disallow option should be used.

For information on installing, configuring, and managing a TACACS+ server, see the product documentation for that server.

The following describes how to configure the ExtremeXOS TACACS+ client component in the ExtremeXOS software: Configuring the TACACS+ Client for Authentication and Authorization on page 899.

## Configuring the TACACS+ Client for Authentication and Authorization

### *Changing the TACACS+ Server*

Use the following steps to change TACACS+ server configuration to avoid service interruption with respect to authentication and authorization.

1. Unconfigure the existing primary TACACS+ server.

Note
After this step, TACACS+ will failover to secondary server.

2. Configure the new primary TACACS+ server.
3. Configure the shared-secret password for primary TACACS+ server.

Note
Only after configuring shared-secret password for primary server, TACACS+ will fallback to primary server from secondary.

4. Unconfigure existing secondary TACACS+ server.
5. Configure new secondary TACACS+ server.
6. Configure shared-secret password for secondary TACACS+ server

To unconfigure the existing TACACS+ server, use the following command:

```
unconfigure tacacs server [primary | secondary]
```

To configure a TACACS+ server, use the following command:

```
configure tacacs [primary | secondary] server [ipaddress | hostname]
{tcp_port} client-ip ipaddress {vr vr_name}
```

To configure shared-secret password for TACACS+ server, use the following command.

```
configure tacacs [primary | secondary] shared-secret {encrypted} string
```

When only a single TACACS+ server is configured, it is essential to disable tacacs-authorization(if enabled) before reconfiguring TACACS+ server.

Note

Command `disable tacacs` is not required while changing TACACS+ servers. And it is recommended to `disable tacacs-authorization` (if enabled), before disabling TACACS+.

### *Specifying TACACS+ Server Addresses*

Before the TACACS+ client software can communicate with a TACACS+ server, you must specify the server address in the client software. You can specify up to two TACACS+ servers, and you can use either an IP address or a host name to identify each server.

- To configure the TACACS+ servers in the client software, use the following command:

  `configure tacacs [`**`primary`** `|` **`secondary`**`] server [`*`ipaddress`* `|` *`hostname`*`] {`*`tcp_port`*`} client-ip` *`ipaddress`* `{`**`vr`** *`vr_name`*`}`

  To configure the primary TACACS+ server, specify **`primary`**. To configure the secondary TACACS+ server, specify **`secondary`**.

### *Configuring the TACACS+ Client Timeout Value*

- To configure the timeout if a server fails to respond, use the following command:

  `configure tacacs timeout` *`seconds`*

  To detect and recover from a TACACS+ server failure when the timeout has expired, the switch makes one authentication attempt before trying the next designated TACACS+ server or reverting to the local database for authentication. In the event that the switch still has IP connectivity to the TACACS+ server, but a TCP session cannot be established, (such as a failed TACACS+ daemon on the server), fail over happens immediately regardless of the configured timeout value.

  For example, if the timeout value is set for three seconds (the default value), it will take three seconds to fail over from the primary TACACS+ server to the secondary TACACS+ server. If both the primary and the secondary servers fail or are unavailable, it takes approximately six seconds to revert to the local database for authentication.

### *Configuring the Shared Secret Password for TACACS+ Communications*

The shared secret is a password that is configured on each network device and TACACS+ server. The shared secret is used to verify communication between network devices and the server.

- To configure the shared secret for client communications with TACACS+ servers, use the following command:

  `configure tacacs [`**`primary`** `|` **`secondary`**`] shared-secret {`**`encrypted`**`}` *`string`*

  To configure the shared secret for a primary TACACS+ server, specify **`primary`**. To configure the shared secret for a secondary TACACS+ server, specify **`secondary`**.

  Do not use the **`encrypted`** keyword to set the shared secret. The **`encrypted`** keyword prevents the display of the shared secret in the `show configuration` command output.

### *Enabling and Disabling the TACACS+ Client Service*

The TACACS+ client service can be enabled or disabled without affecting the client configuration. When the client service is disabled, the client does not communicate with the TACACS+ server, so authentication must take place through the another service such as the local database or a *RADIUS* server.

**Note**
You cannot use RADIUS and TACACS+ at the same time.

- To enable the TACACS+ client service, use the following command: `enable tacacs`
- To disable the TACACS+ client service, use the following command: `disable tacacs`

### *TACACS+ Configuration Example*

This section provides a sample TACACS+ client configuration that:

- Specifies the primary TACACS+ server.
- Specifies the shared secret for the primary TACACS+ server.
- Specifies the secondary TACACS+ server.
- Specifies the shared secret for the secondary TACACS+ server.
- Enables the TACACS+ client on the switch.

All other client configuration parameters use the default settings as described earlier in this section or elsewhere in the *ExtremeXOS 16.2 User Guide*.

```
configure tacacs primary server 10.201.31.238 client-ip 10.201.31.85 vr "VR-Default"
configure tacacs primary shared-secret purple
configure tacacs secondary server 10.201.31.235 client-ip 10.201.31.85 vr "VR-Default"
configure tacacs secondary shared-secret purple
enable tacacs
```

To display the TACACS+ client configuration, use the `show tacacs` command. Below is sample output from this command:

```
TACACS+: enabled
TACACS+ Authorization: disabled
TACACS+ Accounting : disabled
TACACS+ Server Connect Timeout sec: 3
Primary TACACS+ Server:
    Server name   :
    IP address    :  10.201.31.238
    Server IP Port:  49
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
Secondary TACACS+ Server:
    Server name   :
    IP address    :  10.201.31.235
    Server IP Port:  49
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
TACACS+ Acct Server Connect Timeout sec: 3
Primary TACACS+ Accounting Server:Not configured
Secondary TACACS+ Accounting Server:Not configured
```

## Configuring the TACACS+ Client for Accounting

### *Specifying the Accounting Server Addresses*

Before the TACACS+ client software can communicate with a TACACS+ accounting server, you must specify the server address in the client software. You can specify up to two accounting servers, and you can use either an IP address or a host name to identify each server.

- To specify TACACS+ accounting servers, use the following command:

  `configure tacacs-accounting [primary | secondary] server [ipaddress | hostname] {udp_port} client-ip ipaddress {vr vr_name}`

  To configure the primary TACACS+ accounting server, specify **primary**. To configure the secondary TACACS+ accounting server, specify **secondary**.

### *Configuring the TACACS+ Client Accounting Timeout Value*

- To configure the timeout if a server fails to respond, use the following command:

  `configure tacacs-accounting timeout seconds`

  To detect and recover from a TACACS+ accounting server failure when the timeout has expired, the switch makes one authentication attempt before trying the next designated TACACS+ accounting server or reverting to the local database for authentication. In the event that the switch still has IP connectivity to the TACACS+ accounting server, but a TCP session cannot be established, (such as a failed TACACS+ daemon on the accounting server), failover happens immediately regardless of the configured timeout value. If the user does not have a local account or the user is disabled locally, the user's login will fail.

  For example, if the timeout value is set for three seconds (the default value), it takes three seconds to failover from the primary TACACS+ accounting server to the secondary TACACS+ accounting server. If both the primary and the secondary servers fail or are unavailable, it takes approximately six seconds to revert to the local database for authentication.

### *Configuring the Shared Secret Password for TACACS+ Accounting Servers*

The shared secret is a password that is configured on each network device and TACACS+ accounting server. The shared secret is used to verify communication between network devices and the server.

- To configure the shared secret for client communications with TACACS+ accounting servers, use the following command:

  `configure tacacs-accounting [primary | secondary] shared-secret {encrypted} string`

  To configure the primary TACACS+ accounting server, specify **primary**. To configure the secondary TACACS+ accounting server, specify **secondary**.

  Do not use the **encrypted** keyword to set the shared secret. The **encrypted** keyword prevents the display of the shared secret in the `show configuration` command output.

### *Enabling and Disabling TACACS+ Accounting*

After you configure the TACACS+ client with the TACACS+ accounting server information, you must enable accounting in the TACACS+ client before the switch begins transmitting the information. You

must enable TACACS+ authentication in the client for accounting information to be generated. You can enable and disable accounting without affecting the current state of TACACS+ authentication.

- To enable TACACS+ accounting, use the following command:

  `enable tacacs-accounting`

- To disable TACACS+ accounting, use the following command:

  `disable tacacs-accounting`

### *TACACS+ Accounting Configuration Example*

This section provides a sample TACACS+ client configuration for TACACS+ accounting that:

- Specifies the primary TACACS+ accounting server.
- Specifies the shared secret for the primary TACACS+ accounting server.
- Specifies the secondary TACACS+ accounting server.
- Specifies the shared secret for the secondary TACACS+ accounting server.
- Enables TACACS+ accounting on the switch.

All other client configuration features use the default settings as described earlier in this section or in the *ExtremeXOS 16.2 Command Reference Guide*.

```
configure tacacs-accounting primary server 10.201.31.238 client-ip 10.201.31.85 vr "VR-
Default"
configure tacacs-accounting primary shared-secret purple
configure tacacs-accounting secondary server 10.201.31.235 client-ip 10.201.31.85 vr "VR-
Default"
config tacacs-accounting secondary shared-secret purple
enable tacacs-accounting
```

To display the TACACS+ client accounting configuration, use the `show tacacs` or the `show tacacs-accounting` command. The following is sample output from the `show tacacs` command:

```
TACACS+: enabled
TACACS+ Authorization: enabled
TACACS+ Accounting : enabled
TACACS+ Server Connect Timeout sec: 3
Primary TACACS+ Server:
    Server name   :
    IP address    :  10.201.31.238
    Server IP Port:  49
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
Secondary TACACS+ Server:
    Server name   :
    IP address    :  10.201.31.235
    Server IP Port:  49
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
TACACS+ Acct Server Connect Timeout sec: 3
Primary TACACS+ Accounting Server:
    Server name   :
    IP address    :  10.201.31.238
    Server IP Port:  49
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
Secondary TACACS+ Accounting Server:
    Server name   :
    IP address    :  10.201.31.235
    Server IP Port:  49
```

```
    Client address:  10.201.31.85 (VR-Default)
    Shared secret :  purple
```

# Authenticating Management Sessions Through a RADIUS Server

You can use a Remote Authentication Dial In User Service (*RADIUS*) server to authenticate management sessions for multiple switches. A RADIUS server allows you to centralize the authentication database, so that you do not have to maintain a separate local database on each switch. RADIUS servers provide the following services for management sessions:

- Username and password authentication
- Command authorization (the RADIUS server validates whether the user is authorized to execute each command)
- Accounting service (tracks authentication and authorization events)

Note
You can use a local database on each switch as a backup authentication service if the RADIUS service is unavailable. When the RADIUS service is operating, privileges defined on the RADIUS server take precedence over privileges configured in the local database.

To use RADIUS server features, you need the following components:

- RADIUS client software, which is included in the ExtremeXOS software.
- A RADIUS server, which is a third-party product.

Note
RADIUS provides many of the same features provided by TACACS+. You cannot use RADIUS and TACACS+ at the same time.

RADIUS is a communications protocol (RFC 2865) that is used between client and server to implement the RADIUS service.

The RADIUS client component of the ExtremeXOS software should be compatible with any RADIUS compliant server product.

Note
The switch allows local authentication when the client IP is excluded in RADIUS server.

The following sections provide more information on management session authentication:

- How Extreme Switches Work with RADIUS Servers on page 904
- Configuration Overview for Authenticating Management Sessions on page 906

## How Extreme Switches Work with RADIUS Servers

When configured for use with a *RADIUS* server, an ExtremeXOS switch operates as a RADIUS client. In RADIUS server configuration, the client component is configured as a client or as a Network Access Server (NAS). Typically, an ExtremeXOS NAS provides network access to supplicants such as PCs or phones.

When a supplicant requests authentication from a switch that is configured for RADIUS server authentication, the following events occur:

1. The switch sends an authentication request in the form of a RADIUS Access-Request message.
2. The RADIUS server looks up the user in the users file.
3. The RADIUS server accepts or rejects the authentication and returns a RADIUS Access-Accept or Access-Reject message.

Note
A user rejected by the Radius/TACACS server can not be authenticated via local database.

4. If authentication is accepted, the Access-Accept message can contain standard RADIUS attributes and Vendor Specific Attributes (VSAs) that can be used to configure the switch.
5. If authentication is accepted, the Access-Accept message can enable command authorization for that user on the switch. Command authorization uses the RADIUS server to approve or deny the execution of each command the user enters.

The ExtremeXOS switch initiates all communications with the RADIUS server. For basic authentication, the switch sends the Access-Request message, and communications with the RADIUS server is complete when the switch receives the Access-Accept or Access-Reject message. For command authorization, communications starts each time a user configured for command authorization enters a switch command. RADIUS server communications ends when command use is allowed or denied.

A key component of RADIUS server management is the attributes and VSAs that the RADIUS server can be configured to send in Access-Accept messages. VSAs are custom attributes for a specific vendor, such as Extreme Networks. These attributes store information about a particular user and the configuration options available to the user. The RADIUS client in ExtremeXOS accepts these attributes and uses them to configure the switch in response to authentication events. The RADIUS server does not process attributes; it simply sends them when authentication is accepted. It is the switch that processes attributes.

User authentication and attributes are managed on a RADIUS server by editing text files. On the FreeRADIUS server, the user ID, password, attributes, and VSAs are stored in the users file, and VSAs are defined in the dictionary file. The dictionary file associates numbers with each attribute. When you edit the users file, you specify the text version of each attribute you define. When the RADIUS server sends attributes to the switch, it sends the attribute type numbers to reduce the network load. Some attribute values are sent as numbers too.

Command authorization is also managed on a RADIUS server by editing text files. On a FreeRADIUS server, the profiles file is divided into sections called *profiles*. Each profile lists command access definitions. In the users file, you can use the Profile-Name attribute to select the command profile that applies to each user managed by command authorization.

The ExtremeXOS software supports backup authentication and authorization by eight total servers for redundancy.

RADIUS servers can be optionally configured to work with directory services such as LDAP or Microsoft Active Directory. Because ExtremeXOS switches operate with RADIUS servers, they can benefit from the pairing of the RADIUS server and a directory service. Some guidelines for configuring FreeRADIUS with LDAP are provided later in this chapter. Since the use of the directory service requires

configuration of the RADIUS server and directory service, the appropriate documentation to follow is the documentation for those products.

## Configuration Overview for Authenticating Management Sessions

To configure the switch *RADIUS* client and the RADIUS server to authenticate management sessions, do the following:

1. Configure the switch RADIUS client for authentication as described in Configuring the RADIUS Client for Authentication and Authorization on page 913.
2. If you want to use RADIUS accounting, configure the switch RADIUS accounting client as described in Configuring the RADIUS Client for Accounting on page 915.
3. Configure the RADIUS server for authentication as described in Configuring User Authentication (Users File) on page 916.
4. If you want to use RADIUS accounting, configure a RADIUS accounting server as described in the documentation for your RADIUS product.

# Authenticating Network Login Users Through a RADIUS Server

You can use a *RADIUS* server to authenticate network login users and supply configuration data that the switch can use to make dynamic configuration changes to accommodate network login users.

A RADIUS server allows you to centralize the authentication database, so that you do not have to maintain a separate local database on each switch. RADIUS servers provide the following services for network login sessions:

- Username and password authentication
- Standard RADIUS attributes and Extreme Networks VSAs that the switch can use for dynamic configuration
- Accounting service (tracks authentication and authorization events)

To use RADIUS server features, you need the following components:

- RADIUS client software, which is included in the ExtremeXOS software.
- A RADIUS server, which is a third-party product.

Note
RADIUS provides many of the same features provided by TACACS+, but the network login feature does not work with TACACS+.

The following sections provide more information on network login session authentication:

- Differences Between Network Login Authentication and Management Session Authentication on page 906
- Configuration Overview for Authenticating Network Login Users on page 907

## Differences Between Network Login Authentication and Management Session Authentication

Network login authentication is very similar to management session authentication.

The differences are:

- Network login authentication grants network access to devices connected to a switch port, and management session authentication grants management access to the switch for configuration and management.
- The user name for network login authentication can be a MAC address.
- Standard *RADIUS* attributes and Extreme Networks VSAs can be used with the network login and universal port features to configure switch ports and general switch configuration parameters.
- Command authorization is not applicable because network login controls network access, not management session access.

Except for the above differences, network login authentication is the same as described in How Extreme Switches Work with RADIUS Servers on page 904.

## Configuration Overview for Authenticating Network Login Users

To configure the switch *RADIUS* client and the RADIUS server to authenticate network login users, do the following:

1. Configure the switch RADIUS client for authentication as described in Configuring the RADIUS Client for Authentication and Authorization on page 913.
2. If you want to use RADIUS accounting, configure the switch RADIUS accounting client as described in Configuring the RADIUS Client for Accounting on page 915.
3. Configure network login on the switch as described in Security.
4. Configure the RADIUS server for authentication and Extreme Networks VSAs as described in Configuring User Authentication (Users File) on page 916.
5. If you want to use the universal port feature to run configuration scripts at authentication, configure the switch universal port feature as described in Universal Port on page 309.
6. If you want to use RADIUS accounting, configure a RADIUS accounting server as described in the documentation for your RADIUS product.

# Authentication

The *RADIUS* client software sends authentication requests using standard mechanisms for PAP, CHAP (RFC 2865 (13)) and EAP (RFC 3579 (12)).

## Authentication Retransmission Algorithm

Two retransmission algorithms are used in combination: Back-off Round Robin, and simple Round Robin. The focus of this retransmission algorithm is to provide for server redundancy.

Note

The reason for using a combination of back-off and round-robin rather than the standard back-off algorithm where all configured transmissions occur to server 1 before transmitting to server 2 is to allow for more than one server to be tried prior to 802.1x timeout when EAP authentication is occurring.

The standard algorithm is as described in the following figure and is a combination of back-off and round-robin.

This algorithm always uses the highest priority server first regardless of past transaction history. If the highest priority server can handle the entire load all, transactions will go to that server. Consider three *RADIUS* servers – 1, 2 and 3 with the configurable number of retries set to 2:

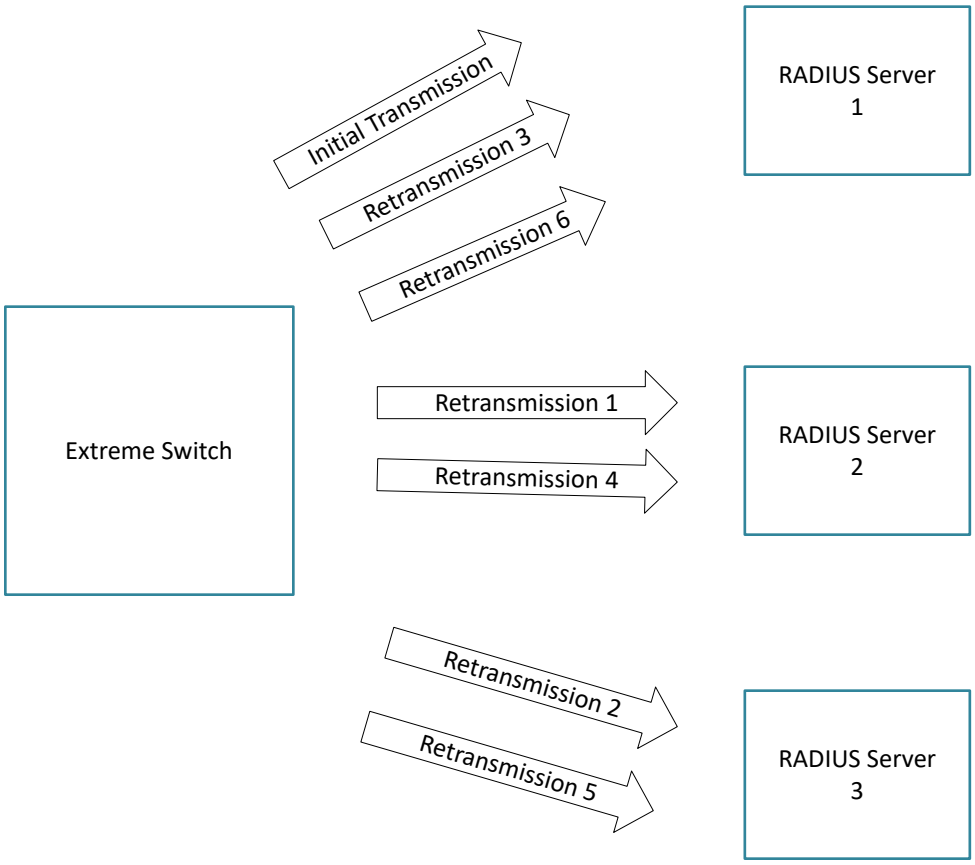



**Figure 112: Authentication Retransmission Algorithm for a Single RADIUS Transaction (No Servers Responding 1)**

This figure shows the entire retransmission algorithm for a single RADIUS transaction if none of the servers were to respond. No more transmissions will occur for this transaction if a response is received by the RADIUS client software within the configurable timeout period.

The round-robin retransmission algorithm is depicted in the following figure and is simply round-robin.

The configurable round-robin retransmission algorithm for RADIUS authentication aims to spread the load among all the configured servers. In large-scale deployments with high rates of authentication this algorithm will provide for better performance than the default algorithm. The initial transmission for each potential authentication will go to the next server in the list. If 999 sessions were to be authenticated across three servers and no timeouts were to occur, then 333 responses would be sent to each server.

Consider three RADIUS servers – 1, 2 and 3 with the configurable number of retries set to 2 and where the prior session sent its initial request to server 1:

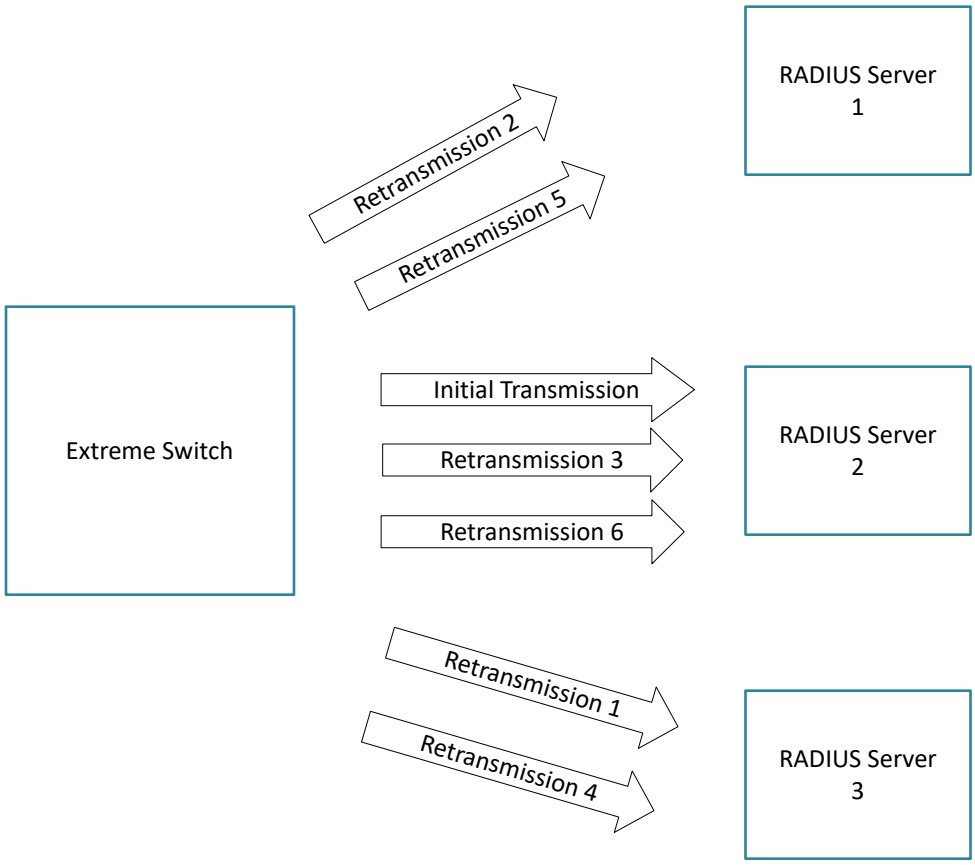



**Figure 113: Authentication Retransmission Algorithm for a Single RADIUS Transaction (No Servers Responding 2)**

This figure shows the entire retransmission algorithm for a single RADIUS transaction if none of the servers were to respond. No more transmissions will occur for this transaction if a response is received by the RADIUS client software within the configurable timeout period. All servers are considered the same priority when using this transmission algorithm with each server taking its turn receiving the initial transmission.

# Accounting

Accounting Start and Stop requests are sent for user sessions in accordance with RFC 2866 (7).

## Accounting Retransmission Algorithm

The *RADIUS* client accounting software uses the standard back-off retransmission algorithm. Since accounting transactions generally do not require timeliness this algorithm focuses on redundancy rather than expediency. The highest priority server is always tried first regardless of past transaction history. Consider three accounting servers – 1, 2 and 3 each with the configurable number of retries set to 2:

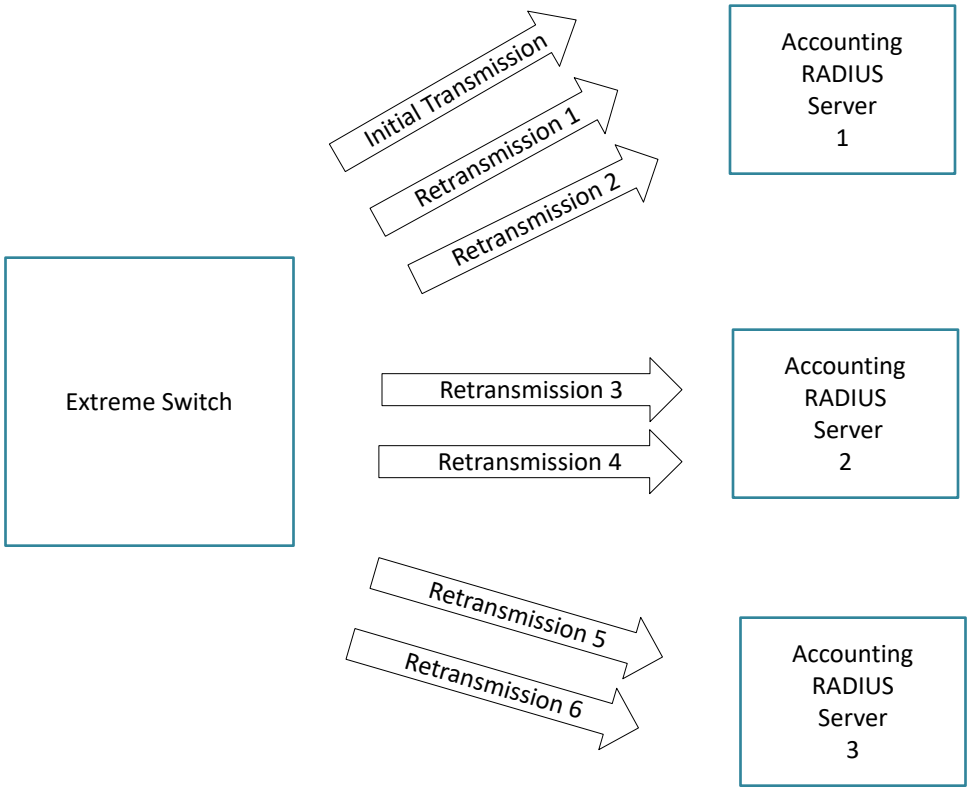



**Figure 114: Accounting Retransmission Algorithm (No Servers Responding)**

This figure shows the entire retransmission algorithm for a single RADIUS accounting transaction if none of the servers were to respond. All of the transactions to each of the servers are attempted before another server is attempted. If the initial server can handle the entire accounting transmission load, then all transactions will go to that server.

## Authentication NMS Realm

The NMS realm type is supported in addition to the management and network realms. These servers are used for the Extreme Network Virtualization or Virtual Machine Tracking feature. This realm only supports the existing subset of two servers (primary and secondary).

## Per Realm Authentication Enable/Disable

*RADIUS* authentication can be enabled on a per realm basis for the management and network realms.

## Supported RADIUS Attributes

The following is a lit of *RADIUS* attributes.

Note

Although this is the list of attributes that is supported by the RADIUS software, usage of these attributes is dependent on the features provided by the system as a whole.

Table Legend:

- PRQ—PAP Authentication Request
- CRQ—CHAP Authentication Request
- EARQ—EAP Authentication Request
- AC—Access Challenge
- AA—Access Accept
- AR—Access Reject

**Table 109: Authentication Attributes**

| Attribute Name | RFC | Attr # | PRQ | CRQ | EARQ | AC | AA | AR |
|---|---|---|---|---|---|---|---|---|
| User-Name | 2865 | 1 | X | X | X | | X | X |
| User-Password | 2865 | 2 | X | | | | | |
| CHAP-Password | 2865 | 3 | | X | | | | |
| NAS-IP-Address | 2865 | 4 | X | X | X | | | |
| NAS-Port | 2865 | 5 | X | X | X | | | |
| Service-Type | 2865 | 6 | X | X | X | | X | |
| Framed-Protocol | 2865 | 7 | | X | X | | | |
| Filter-ID | 2865 | 11 | | | | | X | |
| Framed-MTU | 2865 | 12 | X | X | X | | | |
| State | 2865 | 24 | | | X | X | | |
| Class | 2865 | 25 | | | | | X | |
| Session-Timeout | 2865 | 27 | | | | | X | |
| Idle-Timeout | 2865 | 28 | | | | | X | |
| Termination-Action | 2865 | 29 | | | | | X | |
| Called-Station-ID | 2865 | 30 | X | X | X | | | |
| Calling-Station-ID | 2865 | 31 | X | X | X | | | |
| NAS-Identifier | 2865 | 32 | X | X | X | | | |
| CHAP-Challenge | 2865 | 60 | | X | | | | |
| NAS-Port-Type | 2865 | 61 | X | X | X | | | |
| Tunnel-Type | 2868 | 64 | | | | | X | |
| Tunnel-Medium | 2868 | 65 | | | | | X | |
| Message-Authenticator | 2869 | 80 | | | X | X | X | X |
| Tunnel-Private-Group-ID | 2868 | 81 | | | | | X | |
| Acct-Interim-Interval | 2868 | 85 | | | | | X | |
| NAS-Port-ID | 2869 | 87 | X | X | X | | | |
| NAS-IPV6-Address | 3162 | 95 | X | X | X | | | |
| MS-MPPE-Send-Key | 2548 | 16 | | | | | X | |

**Table 109: Authentication Attributes (continued)**

| Attribute Name | RFC | Attr # | PRQ | CRQ | EARQ | AC | AA | AR |
|---|---|---|---|---|---|---|---|---|
| MS-MPPE-Recv-Key | 2548 | 17 | | | | | X | |
| MS-Quarantine-State | VSA | 45 | | | | | X | X |
| MS-IPv4-Remediation-servers | VSA | 52 | | | | | X | X |
| Extreme-CLI-Authorization | VSA | 201 | | | | | X | |
| Extreme-Shell-Command | VSA | 202 | X | | | | | |
| Extreme-Netlogin-*VLAN* | VSA | 203 | | | | | X | |
| Extreme-Netlogin-URL | VSA | 204 | | | | | X | |
| Extreme-Netlogin-URL-Desc | VSA | 205 | | | | | X | |
| Extreme-Netlogin-Only | VSA | 206 | | | | | X | |
| Extreme-Netlogin-VLAN-Tag | VSA | 209 | | | | | X | |
| Extreme-Netlogin-Extended-VLAN | VSA | 211 | | | | | X | |
| Extreme-Security-Profile | VSA | 212 | | | | | X | |
| Extreme-VM-Name | VSA | 213 | | | | | X | |
| Extreme-VM-VPP-Name | VSA | 214 | | | | | X | |
| Extreme-VM-IP-Addr | VSA | 215 | | | | | X | |
| Extreme-VM-VLAN-Tag | VSA | 216 | | | | | X | |
| Extreme-VM-VR-Name | VSA | 217 | | | | | X | |

Table Legend:

- ASTA—Accounting Start Request
- ASTO—Accounting Stop Request
- AR—Accounting Response

**Table 110: Accounting Attributes**

| Attribute Name | RFC | Attr # | ASTA | ASTO | AR |
|---|---|---|---|---|---|
| User-Name | 2865 | 1 | X | X | |
| NAS-IP-Address | 2865 | 4 | X | X | |
| NAS-Port | 2865 | 4 | X | X | |
| Class | 2865 | 25 | X | X | |
| Calling-Station-ID | 2865 | 31 | X | X | |
| Acct-Status-Type | 2866 | 40 | X | X | |

**Table 110: Accounting Attributes (continued)**

| Attribute Name | RFC | Attr # | ASTA | ASTO | AR |
|---|---|---|---|---|---|
| Acct-Delay-Time | 2866 | 41 | X | X | |
| Acct-Session-ID | 2866 | 44 | X | X | |
| Acct-Authentic | 2866 | 45 | X | X | |
| Acct-Session-Time | 2866 | 46 | | X | |
| Acct-Terminate-Cause | 2866 | 49 | | X | |
| NAS-IPV6-Address | 3162 | 95 | X | X | |
| ETS-Auth-Client-Type | VSA | 1 | X | X | |
| Acct-Input-Octets | 2866 | 42 | | X | |
| Acct-Output-Octets | 2866 | 43 | | X | |
| Acct-Input-Packets | 2866 | 47 | | X | |
| Acct-Output-Packets | 2866 | 48 | | X | |
| Login-Service | 2865 | 15 | | X | |

## Configuring the RADIUS Client

For information on installing, configuring, and managing a *RADIUS* server, see the product documentation for that server and the guidelines in RADIUS Server Configuration Guidelines on page 916.

### Configuring the RADIUS Client for Authentication and Authorization

*Specifying the RADIUS Server Addresses*

Before the *RADIUS* client software can communicate with a RADIUS server, you must specify the server address in the client software. You can specify up to eight RADIUS servers, and you can use either an IP address or a host name to identify each server.

- To configure the RADIUS servers in the client software, use the following command:

  `configure radius {mgmt-access | netlogin} {primary | secondary | index} server {host_ipaddr | host_ipV6addr | hostname} {udp_port} client-ip {client_ipaddr | client_ipV6addr} {vr vr_name} {shared-secret {encrypted} secret}`

  **Note**
  It is recommended to enable loopback mode on the *VLAN* associated with RADIUS if the radius connectivity is established via a front panel port on a summit stack.

  The default port value for authentication is 1812. The client IP address is the IP address used by the RADIUS server for communicating back to the switch.

  To configure the primary RADIUS server, specify **primary**. To configure the secondary RADIUS server, specify **secondary**.

By default, switch management and network login use the same primary and secondary RADIUS servers for authentication. To specify one pair of RADIUS servers for switch management and another pair for network login, use the **mgmt-access** and **netlogin** keywords.

### *Configuring the RADIUS Client Timeout Value*

- To configure the timeout if a server fails to respond, use the following command:
  `configure radius {mgmt-access {primary | secondary} | netlogin {primary | secondary} | index } timeout sec`

  If the timeout expires, another authentication attempt is made. After three failed attempts to authenticate, the alternate server is used. After six failed attempts, local user authentication is used. If the user does not have a local account or the user is disabled locally, the user's login will fail.

  If you do not specify the **mgmt-access** or **netlogin** keyword, the timeout interval applies to both switch management and netlogin *RADIUS* servers.

### *Configuring the Shared Secret Password for RADIUS Communications*

The shared secret is a password that is configured on each network device (*RADIUS* client) and RADIUS server. The shared secret is used to verify communication between network devices and the server.

- To configure the shared secret for client communications with RADIUS servers, use the following command:
  `configure radius {mgmt-access | netlogin} {primary | secondary} shared-secret {encrypted} string`

  To configure the shared secret for a primary RADIUS server, specify **primary**. To configure the shared secret for a secondary RADIUS server, specify **secondary**.

  If you do not specify the **mgmt-access** or **netlogin** keyword, the secret applies to both the primary and secondary switch management and network login RADIUS servers.

  Do not use the **encrypted** keyword to set the shared secret. The **encrypted** keyword prevents the display of the shared secret in the `show configuration` command output.

### *Enabling and Disabling the RADIUS Client Service*

The *RADIUS* client service can be enabled or disabled without affecting the client configuration. When the client service is disabled, the client does not communicate with the RADIUS server, so authentication must take place through the another service such as the local database or a TACACS+ server.

Note
You cannot use RADIUS and TACACS+ at the same time.

- To enable the RADIUS client service, use the following command:
  `enable radius {mgmt-access | netlogin}`
- To disable the RADIUS client service, use the following command:
  `disable radius {mgmt-access | netlogin}`

  If you do not specify the **mgmt-access** or **netlogin** keywords, RADIUS authentication is enabled or disabled on the switch for both management and network login.

## Configuring the RADIUS Client for Accounting

### *Specifying the RADIUS Accounting Server Addresses*

Before the *RADIUS* client software can communicate with a RADIUS accounting server, you must specify the server address in the client software. You can specify up to two accounting servers, and you can use either an IP address or a host name to identify each server.

- To specify RADIUS accounting servers, use the following command:

  `configure radius-accounting { mgmt-access | netlogin } [ primary | secondary | index ] server [ host_ipaddr | host_ipV6addr | hostname] {udp_port} client-ip [ client_ipaddr | client_ipV6addr] {vr vr_name} {shared-secret {encrypted} secret}`

  The default port value for accounting is 1813. The client IP address is the IP address used by the RADIUS server for communicating back to the switch.

  To configure the primary RADIUS accounting server, specify **primary**. To configure the secondary RADIUS accounting server, specify **secondary**.

  By default, switch management and network login use the same primary and secondary RADIUS servers for accounting. To specify one pair of RADIUS accounting servers for switch management and another pair for network login, use the **mgmt-access** and **netlogin** keywords.

### *Configuring the RADIUS Client Accounting Timeout Value*

- To configure the timeout if a server fails to respond, use the following command:

  `configure radius-accounting {mgmt-access {primary | secondary} | netlogin {primary | secondary} | index } timeout sec`

  If the timeout expires, another authentication attempt is made. After three failed attempts to authenticate, the alternate server is used.

### *Configure the Shared Secret Password for RADIUS Accounting Servers*

The shared secret is a password that is configured on each network device (*RADIUS* client) and RADIUS accounting server. The shared secret is used to verify communication between network devices and the server.

- To configure the shared secret for client communications with RADIUS accounting servers, use the following command:

  `configure radius-accounting {mgmt-access | netlogin} [primary | secondary] shared-secret {encrypted} string`

  To configure the primary RADIUS accounting server, specify **primary**. To configure the secondary RADIUS accounting server, specify **secondary**.

  If you do not specify the **mgmt-access** or **netlogin** keywords, the secret applies to both the primary and secondary switch management and network login RADIUS accounting servers.

  Do not use the **encrypted** keyword to set the shared secret. The **encrypted** keyword prevents the display of the shared secret in the `show configuration` command output.

*Enabling and Disabling RADIUS Accounting*

After you configure the *RADIUS* client with the RADIUS accounting server information, you must enable accounting in the RADIUS client before the switch begins transmitting the information. You must enable RADIUS authentication in the client for accounting information to be generated. You can enable and disable accounting without affecting the current state of RADIUS authentication.

- To enable RADIUS accounting, use the following command:

  `enable radius-accounting {mgmt-access | netlogin}`

- To disable RADIUS accounting, use the following command:

  `disable radius-accounting {mgmt-access | netlogin}`

  If you do not specify a keyword, RADIUS accounting is enabled or disabled on the switch for both management and network login.

## RADIUS Server Configuration Guidelines

Note

For information on how to use and configure your *RADIUS* server, refer to the documentation that came with your RADIUS server.

### Configuring User Authentication (Users File)

User authentication is configured in the users file on a FreeRADIUS server. Other *RADIUS* servers might use a different name and a different syntax for configuration, but the basic components of the users file and user authentication are the same.

For Extreme Networks switches, there are three types of users file entries:

- Session management entries
- Network login user entries
- Network login MAC address entries

Note

The “users” file is case-sensitive, and punctuation is very important for FreeRADIUS.

*Session Management Entries*

The following is an example of a session management entry for read-write access:

```
eric
Cleartext-Password := "eric"
Service-Type = Administrative-User,
Extreme-CLI-Authorization = Enabled
```

The key components of the example above are the user name, password, service-type, and Extreme-CLI-Authorization VSA. For simple authentication, you only need to enter the user name ("eric" in this example) and a password as described in the *RADIUS* server documentation.

Enter the attributes for each user and separate them from the others with commas as described in the RADIUS server documentation.

For more information on the Extreme-CLI-Authorization VSA, see Extreme Networks VSAs on page 919.

### *Network Login User Entries*

The following is an example of a network login user entry:

```
Jim             Auth-Type := EAP, User-Password == "12345"
                Session-Timeout = 60,
                Termination-Action = 1,
                Extreme-Security-Profile = "user-auth LOGOFF-PROFILE=avaya-
                  remove;qos=\"QP1\";",
                Extreme-Netlogin-Vlan = voice-avaya
```

The key components of the example above are the user name, password, attributes, and Extreme Networks VSAs. For simple authentication, you only need to enter the user name (Jim in this example) and a password as described in the *RADIUS* server documentation.

Enter the attributes for each user and separate them from the others with commas as described in the RADIUS server documentation.

In the example above, the Session-Timeout and Termination-Action attributes are examples of standard RADIUS attributes, and these are described in Standard RADIUS Attributes Used by Extreme Switches on page 918.

The Extreme-Security-Profile and Extreme-Netlogin-Vlan attributes are examples of Extreme Networks VSAs and are described in Extreme Networks VSAs on page 919.

### *Network Login MAC Address Entries*

The following is an example of a network login MAC address entry:

```
00040D9D12AF
Auth-Type := Local,
User-Password == "00040D9D12AF"
Session-Timeout = 60,
Termination-Action = 1,
Extreme-Security-Profile = "user-auth LOGOFF-PROFILE=avaya remove;qos=\"QP1\";",
Extreme-Netlogin-Vlan = voice-avaya
```

The key components of the example above are the MAC address, password (which is set to the MAC address), attributes, and Extreme Networks VSAs. For simple authentication, you only need to enter the MAC address (00040D9D12AF in this example) and a password as described in the *RADIUS* server documentation.

Enter the attributes for each user and separate them from the others with commas as described in the RADIUS server documentation.

In the example above, the Session-Timeout and Termination-Action attributes are examples of standard RADIUS attributes, and these are described in Standard RADIUS Attributes Used by Extreme Switches on page 918.

The Extreme-Security-Profile and Extreme-Netlogin-*VLAN* attributes are examples of Extreme Networks VSAs and are described in Extreme Networks VSAs on page 919.

### *Standard RADIUS Attributes Used by Extreme Switches*

The ExtremeXOS software uses standard *RADIUS* attributes to send information in an Access-Request message to a RADIUS server.

The software also accepts some standard RADIUS attributes in the Access-Accept message that the RADIUS server sends to the switch after successful authentication. The switch ignores attributes that it is not programmed to use.

The following table lists the standard RADIUS attributes used by the ExtremeXOS software.

**Table 111: Standard RADIUS Attributes Used by Network Login**

| Attribute | RFC | Attribute Type | Format | Sent-in | Description |
|---|---|---|---|---|---|
| User-Name | RFC 2138 | 1 | String | Access-Request | Specifies a user name for authentication. |
| Calling-Station-ID | RFC 2865 | 31 | String | Access-Request | Identifies the phone number for the supplicant requesting authentication. |
| EAP-Message | RFC 3579 | 79 | String | Access-Request, Access-Challenge, Access-Accept, and Access Reject | Encapsulates EAP packets. |
| Login-IP-Host | RFC 2138 | 14 | Address | Access-Request and Access-Accept | Specifies a host to log into after successful authentication. |
| Message-Authenticator | RFC 3579 | 80 | String | Access-Request, Access-Challenge, Access-Accept, and Access Reject | Contains a hash of the entire message that is used to authenticate the message. |
| NAS-Port-Type | RFC 2865 | 61 | Integer | Access-Request | Identifies the port type for the port through which authentication is requested. |
| Service-Type | RFC 2138 | 6 | String | Access-Accept | Specifies the granted service type in an Access-Accept message. See Attribute 6: Service Type below. |
| Session-Timeout | RFC 2865 | 27 | Integer | Access-Accept, Access-Challenge | Specifies how long the user session can last before authentication is required. |
| State | RFC 2865 | 24 | String | Access-Challenge, Access-Request | Site specific. |
| Termination-Action | RFC 2865 | 29 | Integer | Access-Accept | Specifies how the switch should respond to service termination. |

**Table 111: Standard RADIUS Attributes Used by Network Login (continued)**

| Attribute | RFC | Attribute Type | Format | Sent-in | Description |
|---|---|---|---|---|---|
| Tunnel-Medium-Type | RFC 2868 | 65 | Integer | Access-Accept | Specifies the transport medium used when creating a tunnel for protocols (for example, VLANs) that can operate over multiple transports. |
| Tunnel-Private-Group-ID | RFC 2868 | 81 | String | Access-Accept | Specifies the *VLAN* ID of the destination VLAN after successful authentication; used to derive the VLAN name. |
| Tunnel-Type | RFC 2868 | 64 | Integer | Access-Accept | Specifies the tunneling protocol that is used. |
| User-Password | RFC 2138 | 2 | String | Access-Request | Specifies a password for authentication. |

**Attribute 6: Service Type**

Extreme Networks switches have two levels of user privilege:

- read-only
- read-write

Because no command line interface (CLI) commands are available to modify the privilege level, access rights are determined when you log in. For a RADIUS server to identify the administrative privileges of a user, Extreme Networks switches expect a RADIUS server to transmit the Service-Type attribute in the Access-Accept packet, after successfully authenticating the user.

Extreme Networks switches grant a RADIUS-authenticated user read-write privilege if a Service-Type value of 6 is transmitted as part of the Access-Accept message from the RADIUS server. Other Service-Type values or no value, result in the switch granting read-only access to the user. Different implementations of RADIUS handle attribute transmission differently. You should consult the documentation for your specific implementation of RADIUS when you configure users for read-write access.

### *Extreme Networks VSAs*

The following table contains the Vendor Specific Attribute (VSA) definitions that a *RADIUS* server can send to an Extreme switch after successful authentication.

These attributes must be configured on the RADIUS server along with the Extreme Networks Vendor ID, which is 1916.

**Table 112: VSA Definitions for Web-Based, MAC-Based, and 802.1X Network Login**

| VSA | Attribute Type | Format | Sent-in | Description |
|---|---|---|---|---|
| Extreme-CLI-Authorization | 201 | Integer | Access-Accept | Specifies whether command authorization is to be enabled or disabled for the user on the ExtremeXOS switch. |
| Extreme-Netlogin-VLAN-Name | 203 | String | Access-Accept | Name of destination *VLAN* after successful authentication (must already exist on switch). |
| Extreme-Netlogin-URL | 204 | String | Access-Accept | Destination web page after successful authentication. |
| Extreme-Netlogin-URL-Desc | 205 | String | Access-Accept | Text description of network login URL attribute. |
| Extreme-Netlogin-Only | 206 | Integer | Access-Accept | Indication of whether the user can authenticate using other means, such as telnet, console, SSH, or Vista. A value of "1" (enabled) indicates that the user can only authenticate via network login. A value of "0" (disabled) indicates that the user can also authenticate via other methods. |
| Extreme-User-Location | 208 | String | | |
| Extreme-Netlogin-VLAN-ID | 209 | Integer | Access-Accept | ID of destination VLAN after successful authentication (must already exist on switch). |
| Extreme-Netlogin-Extended-VLAN | 211 | String | Access-Accept | Name or ID of the destination VLAN after successful authentication (must already exist on switch).<br><br>**Note:** When using this attribute, specify whether the port should be moved tagged or untagged to the VLAN. See the guidelines listed in the section VSA 211: Extreme-Netlogin-Extended-Vlan on page 924 below for more information. |
| Extreme-Security-Profile | 212 | String | Access-Accept | Specifies a universal port profile to execute on the switch. For more information, see Universal Port on page 309. |
| EXTREME_VM_NAME | 213 | String | Access-Accept | Specifies the name of the VM that is being authenticated . Example: MyVM1 |
| EXTREME_VM_VPP_NAME | 214 | String | Access-Accept | Specifies the VPP to which the VM is to be mapped. Example: nvpp1 |
| EXTREME_VM_IP_ADDR | 215 | String | Access-Accept | Specifies the IP address of the VM . Example: 11.1.1.254 |

**Table 112: VSA Definitions for Web-Based, MAC-Based, and 802.1X Network Login (continued)**

| VSA | Attribute Type | Format | Sent-in | Description |
|---|---|---|---|---|
| EXTREME_VM_CTag | 216 | Integer | Access-Accept | Specifies the ID or tag of the destination VLAN for the VM . Example: 101 |
| EXTREME_VM_VR_Name | 217 | String | Access-Accept | Specifies the VR in which the destination VLAN is to be placed. Example : UserVR1 |

The examples in the following sections are formatted for use in the FreeRADIUS users file. If you use another RADIUS server, the format might be different.

Note

For information on how to use and configure your RADIUS server, refer to the documentation that came with your RADIUS server.

For untagged VLAN movement with 802.1X netlogin, you can use all current Extreme Networks VLAN VSAs: VSA 203, VSA 209, and VSA 211.

#### VSA 201: Extreme-CLI-Authorization

This attribute specifies whether command authorization is to be enabled or disabled for the user on the ExtremeXOS switch.

If command authorization is disabled, the user has full access to all CLI commands. If command authorization is enabled, each command the user enters is accepted or rejected based on the contents of the profiles file on the RADIUS server.

When added to the RADIUS users file, the following example enables command authorization for the associated user:

**Extreme: Extreme-CLI-Authorization = enabled**

When added to the RADIUS users file, the following example disables command authorization for the associated user:

```
Extreme: Extreme-CLI-Authorization = disabled
```

#### VSA 203: Extreme-Netlogin-VLAN-Name

This attribute specifies a destination VLAN name that the RADIUS server sends to the switch after successful authentication.

The VLAN must already exist on the switch. When the switch receives the VSA, it adds the authenticated user to the VLAN.

The following describes the guidelines for VSA 203:

- For untagged VLAN movement with 802.1X netlogin, you can use all current Extreme Networks VLAN VSAs: VSA 203, VSA 209, and VSA 211.
- To specify the VLAN name, use an ASCII string.
- When using this VSA, do not specify whether the VLAN is tagged or untagged.

Because the RADIUS server can identify a target VLAN with multiple attributes, the switch selects the appropriate VLAN or VLANs using the order:

- Extreme-Netlogin-Extended-VLAN (VSA 211)
- Extreme-Netlogin-VLAN-Name (VSA 203)
- Extreme-Netlogin-VLAN-ID (VSA 209)
- Tunnel-Private-Group-ID, but only if Tunnel-Type == VLAN(13) and Tunnel-Medium-Type == 802 (6) (see Standard RADIUS Attributes Used by Extreme Switches on page 918)

If none of the previously described attributes are present ISP mode is assumed, and the client remains in the configured VLAN.

When added to the RADIUS users file, the following example specifies the destination VLAN name, purple, for the associated user:

```
Extreme: Extreme-Netlogin-VLAN-Name = purple
```

#### VSA 204: Extreme-Netlogin-URL

The Extreme-*NetLogin*-Url attribute specifies a web page URL that the RADIUS server sends to the switch after successful authentication. When the switch receives the attribute in response to a web-based network login, the switch redirects the web client to display the specified web page. If a login method other than web-based is used, the switch ignores this attribute.

The following describes the guidelines for VSA 204:

- To specify the URL to display after authentication, use an ASCII string.
- If you do not specify a URL, the network login infrastructure uses the default redirect page URL, , or the URL that you configured using the `configure netlogin redirect-page` command.
- VSA 204 applies only to the web-based authentication mode of Network Login.

The following example specifies the redirection URL to use after successful authentication.

To configure the redirect URL as http://www.myhomepage.com, add the following line:

```
Extreme: Netlogin-URL = http://www.myhomepage.com
```

#### VSA 205: Extreme-Netlogin-URL-Desc

The Extreme-NetLogin-Url-Desc attribute provides a redirection description that the RADIUS server sends to the switch after successful authentication. When the switch receives this attribute in response to a web-based network login, the switch temporarily displays the redirect message while the web client is redirected to the web page specified by attribute 204. If a login method other than web-based is used, the switch ignores this attribute.

The following describes the guidelines for VSA 205:

- To let the user know where they will be redirected to after authentication (specified by VSA 204), use an ASCII string to provide a brief description of the URL.
- VSA 205 applies only to the web-based authentication mode of Network Login.

The following example specifies a redirect description to send to the switch after successful authentication:

```
Extreme: Netlogin-URL-Desc = "Authentication successful. Stand by for the home page."
```

#### VSA 206: Extreme-Netlogin-Only

The Extreme-Netlogin-Only attribute can be used to allow normal authentication or restrict authentication to only the network login method.

When this attribute is assigned to a user and authentication is successful, the RADIUS server sends the configured value back to the switch. The configured value is either disabled or enabled.

The Extreme switch uses the value received from the RADIUS server to determine if the authentication is valid. If the configured value is disabled, all normal authentication processes are supported (Telnet and SSH, for example), so the switch accepts the authentication. If the configured value is enabled, the switch verifies whether network login was used for authentication. If network login was used for authentication, the switch accepts the authentication. If an authentication method other than network login was used, the switch rejects the authentication.

Add the following line to the RADIUS server users file for users who are not restricted to network login authentication:

```
Extreme:Extreme-Netlogin-Only = Disabled
```

Add the following line to the RADIUS server users file for users who are restricted to network login authentication:

```
Extreme:Extreme-Netlogin-Only = Enabled
```

To reduce the quantity of information sent to the switch, the RADIUS server sends either a 1 for the enabled configuration or a 0 for the disabled configuration.

These values must be configured in the RADIUS dictionary file as shown in Configuring the Dictionary File on page 926.

#### VSA 209: Extreme-Netlogin-VLAN-ID

This attribute specifies a destination VLAN ID (or VLAN tag) that the RADIUS server sends to the switch after successful authentication.

The VLAN must already exist on the switch. When the switch receives the VSA, it adds the authenticated user to the VLAN.

The following describes the guidelines for VSA 209:

- For untagged VLAN movement with 802.1X netlogin, you can use all current Extreme Networks VLAN VSAs: VSA 203, VSA 209, and VSA 211.
- To specify the VLAN ID, use an ASCII string.
- When using this VSA, do not specify whether the VLAN is tagged or untagged.

Because the RADIUS server can identify a target VLAN with multiple attributes, the switch selects the appropriate VLAN or VLANs using the order:

- Extreme-Netlogin-Extended-VLAN (VSA 211)
- Extreme-Netlogin-VLAN-Name (VSA 203)
- Extreme-Netlogin-VLAN-ID (VSA 209)
- Tunnel-Private-Group-ID, but only if Tunnel-Type == VLAN(13) and Tunnel-Medium-Type == 802 (6) (see Standard RADIUS Attributes Used by Extreme Switches on page 918)

If none of the previously described attributes are present ISP mode is assumed, and the client remains in the configured VLAN.

When added to the RADIUS users file, the following example specifies the destination VLAN ID, 234, for the associated user:

```
Extreme:Extreme-Netlogin-VLAN-ID = 234
```

#### VSA 211: Extreme-Netlogin-Extended-Vlan

This attribute specifies one or more destination VLANs that the RADIUS server sends to the switch after successful authentication.

You can specify VLANS by VLAN name or ID (tag). The VLANs may either already exist on the switch or, if you have enabled dynamic VLANs and a non-existent VLAN tag is given, the VLAN is created.

When the switch receives the VSA, it does the following:

- Unauthenticates the user on all VLANs where it is currently authenticated during reauthentication.
- Authenticates the user on all VLANs in the VSA, or none of them.

In cases where the client is already authenticated, if a single VLAN move fails from a list of VLANs in the VSA and the move-fail-action is **authenticate**, then it is left as-is. If the client is not already authenticated (first time authentication), then it is authenticated on learnedOnVlan if possible. If move-fail-action is **deny** then the client is unauthenticated from all the VLANs where it is currently authenticated. There is no partial success.

Note

If there is one or more invalid VLAN in the VSA, the supplicant is not authenticated on any one of them.

For example, if the VSA is Uvoice;Tdata and the VLAN **data** does not have a tag or the VLAN does not exist, then the port movement fails. Even if a single VLAN in the list is invalid the entire list is discarded and the action taken is based on move-fail-action configuration.

The following describes the guidelines for VSA 211:

- For tagged VLAN movement with 802.1X netlogin, you must use VSA 211.
- To specify the VLAN name or the VLAN ID, use an ASCII string; however, you cannot specify both the VLAN name and the VLAN ID at the same time. If the string only contains numbers, it is interpreted as the VLAN ID.
- A maximum of 10 VLANs are allowed per VSA.
- For tagged VLANs, specify **T** for tagged before the VLAN name or VLAN ID.
- For untagged VLANs, specify **U** for untagged before the VLAN name or VLAN ID.
- For movement based on the incoming port's traffic, specify the wildcard * before the VLAN name or VLAN ID. The behavior can be either tagged or untagged, based on the incoming port's traffic, and mimics the behavior of VSA 203 and VSA 209, respectively.
- Multiple VLAN names or VLAN IDs are separated by semicolons. When multiple vlans are defined in single VSA 211, the wildcard * is not allowed.
- There cannot be more than one untagged VLAN in a single VSA.
- The same VLAN cannot be both untagged and tagged in a single VSA.

- A client or supplicant can be authenticated in a only one untagged VLAN.
- The ports configured for an untagged VLAN different from the netlogin VLAN can never be added tagged to the same VLAN.
- A port can be in more than one untagged VLAN when MAC-based VLANs are enabled.

When added to the RADIUS users file, the following examples specify VLANs for the switch to assign after authentication:

```
Extreme-Netlogin-Extended-VLAN = Tvoice (Tagged VLAN named voice)
Extreme-Netlogin-Extended-VLAN = Udata (Untagged VLAN named data)
Extreme-Netlogin-Extended-VLAN = *orange (VLAN named orange, tagging dependent on
incoming traffic)
Extreme-Netlogin-Extended-VLAN = T229 (Tagged VLAN with ID 229)
Extreme-Netlogin-Extended-VLAN = U4091 (Untagged VLAN with ID 4091)
Extreme-Netlogin-Extended-VLAN = *145 (VLAN with ID 145, tagging dependent on incoming
traffic)
in FreeRADIUS, a tagged VLAN voice and a tagged VLAN mktg would be configured as the
following:
Extreme-Netlogin-Extended-VLAN = "Tvoice;Tmktg;"
```

An untagged VLAN **data** and a tagged VLAN **mktg** is configured as the following:

```
Extreme-Netlogin-Extended-VLAN = "Udata;Tmktg;"
```

A tagged VLAN with VLAN ID 229 and a tagged VLAN with VLAN ID 227 is configured in FreeRADIUS as the following:

```
Extreme-Netlogin-Extended-VLAN = "T229;T227;"
```

An untagged VLAN with VLAN ID 4091 and a tagged VLAN with VLAN ID 2001 is configured as the following:

```
Extreme-Netlogin-Extended-VLAN = "U4091;T2001;"
```

#### VSA 212: Extreme-Security-Profile

This attribute specifies a profile name that the RADIUS server sends to the switch after successful authentication. The switch uses this profile name to run a special type of script called a *profile*. The profile is stored on the switch and can be used to modify the switch configuration in response to authentication. Profiles are created using the Universal Port feature, which is described in Universal Port on page 309.

The following describes the guidelines for VSA 212:

- This VSA must contain a profile name.
- This VSA can include optional variables for use in profile execution.
- The variable entry format is: `<var1>=<value1>;<var2>=<value2>;...`
- Each profile variable must be separated from the others by a semicolon.

When added to the RADIUS users file, the following example provides to the switch the profile name p1, variable QOS=QP8, and variable LOGOFF-PROFILE=P2:

```
EXTREME-SECURITY-PROFILE= "p1 QOS=\"QP8\";LOGOFF-PROFILE=P2;"
```

#### VSA 213: EXTREME_VM_NAME

This VSA is used in context with the *Extreme Network Virtualization (XNV)* feature, especially with the NMS authentication of VMs. Use this VSA to specify the name of the VM that is being authenticated. An example would be: MyVM1

#### VSA 214: EXTREME_VM_VPP_NAME

This VSA is used in context with the XNV feature, especially with the NMS authentication of VMs. Use this VSA to specify the VPP to which the VM is to be mapped. An example would be: nvpp1

#### VSA 215: EXTREME_VM_IP_ADDR

This VSA is used in context with the XNV feature, especially with the NMS authentication of VMs. Use this VSA to specify the IP address of the VM. An example would be: 11.1.1.254

#### VSA 216: EXTREME_VM_VLAN_ID

This VSA corresponds to XNV with Dynamic VLANs. Use this VSA to specify the ID or tag of the destination VLAN for the VM. An example would be: 101

#### VSA 217: EXTREME_VM_VR_NAME

This VSA corresponds to XNV with Dynamic VLANs. Use this VSA to specify the VR in which the destination VLAN is to be placed. An example would be: UserVR1

## Configuring the Dictionary File

Before you can use Extreme Networks VSAs on a *RADIUS* server, you must define the VSAs.

On the FreeRADIUS server, you define the VSAs in the dictionary file in the /etc/raddb directory. You must define the vendor ID for Extreme Networks, each of the VSAs you plan to use, and the values to send for the VSAs. The following example shows the entries to add to a FreeRADIUS server dictionary file for Extreme Networks VSAs:

```
VENDOR          Extreme         1916
ATTRIBUTE    Extreme-CLI-Authorization      201  integer     Extreme
ATTRIBUTE    Extreme-Shell-Command          202   string     Extreme
ATTRIBUTE    Extreme-Netlogin-Vlan          203   string     Extreme
ATTRIBUTE    Extreme-Netlogin-Url           204   string     Extreme
ATTRIBUTE    Extreme-Netlogin-Url-Desc      205   string     Extreme
ATTRIBUTE    Extreme-Netlogin-Only          206  integer     Extreme
ATTRIBUTE    Extreme-User-Location          208   string     Extreme
ATTRIBUTE    Extreme-Netlogin-Vlan-Tag      209  integer     Extreme
ATTRIBUTE    Extreme-Netlogin-Extended-Vlan 211   string     Extreme
ATTRIBUTE    Extreme-Security-Profile       212   string     Extreme
VALUE        Extreme-CLI-Authorization   Disabled       0
VALUE        Extreme-CLI-Authorization   Enabled        1
VALUE        Extreme-Netlogin-Only       Disabled       0
VALUE        Extreme-Netlogin-Only       Enabled        1
# End of Dictionary
```

The lines that begin with VALUE provide the integers that the RADIUS server sends to the switch when the corresponding text is configured in the RADIUS users file. For example, if the Extreme-CLI-Authorization attribute is set to Enabled for a particular user, the RADIUS server sends the value 1 to the switch (which reduces total bytes transferred). The ExtremeXOS software is designed to interpret the integer values as shown above, so be sure to use these values.

## Additional RADIUS Configuration Examples

### *Installing and Testing the FreeRADIUS Server*

*RADIUS* is a client/server protocol based on UDP.

The example presented in this section describes a RADIUS server that is a daemon process running on a Linux server.

The following example shows how to install and test a FreeRADIUS server:

```
tar -zxvf freeradius-1.0.2.tar.gz        (extract with gunzip and tar)
./configure
make
make install                             (run this command as root)
radiusd                                  (start RADIUS server, or...)
radiusd -X                               (start RADIUS server in debug mode)
radtest test test localhost 0 testing123 (test RADIUS server)
```

If radtest receives a response, the FreeRADIUS server is up and running.

**Note**

RADIUS server software can be obtained from several sources. This solution uses the FreeRADIUS software available on the following URLs: www.freeradius.org and www.redhat.com. Another free tool, NTRadPing, can be used to test authentication and authorization requests from Windows clients. NTRadPing displays detailed responses such as attribute values sent back from the RADIUS server.

### *Configuring the FreeRADIUS Server*

Configuring the *RADIUS* server involves configuring the RADIUS server and the RADIUS client (for authentication and authorization). FreeRADIUS configuration files are usually stored in the /etc/raddb folder. The following example demonstrates how to configure the FreeRADIUS server for authentication and LDAP support:

1. Modify the radiusd.conf file global settings:

```
log_auth = yes              (log authentication requests to the log file)
log_auth_badpass = no       (don't log passwords if request rejected)
log_auth_goodpass = no      (don't log passwords if request accepted)
```

2. Modify LDAP Settings:

```
modules {
            ldap {
                server = "ldaptest.extremenetworks.com"
                basedn = "o=ldaptestdemo,dc=extremenetworks,dc=com"
                filter = "(cn=%{Stripped-User-Name:-%{User-Name}})"
                base_filter = "(objectclass=radiusprofile)"
                start_tls = no
                dictionary_mapping = ${raddbdir}/ldap.attrmap
                authtype = ldap
                ldap_connections_number = 5
                timeout = 4
        timelimit = 3
        net_timeout = 1
                }
}
```

3. Uncomment LDAP from the authorize section:

```
            authorize {
          preprocess
          chap
          mschap
          suffix
          ldap
          eap
          files
              }
```

4. Uncomment LDAP from the authenticate section:

```
      authenticate {
      Auth-Type PAP {
      pap
          }
      Auth-Type CHAP {
      chap
          }
      Auth-Type MS-CHAP {
      mschap
          }
          unix
      ldap
          eap
```

An Extreme Networks edge switch serves as a network access server (NAS) for workstations and as a RADIUS client for the RADIUS server.

RADIUS clients are configured in /etc/raddb/clients.conf. There are two ways to configure RADIUS clients. Either group the NAS by IP subnet or list the NAS by host name or IP address.

5. Configure the RADIUS client using the second method.

```
client 192.168.1.1 {
          secret = extreme1
          shortname = ldap-demo
}
```

*Configuring the RADIUS-to-LDAP Attribute Mappings*

Attributes are configured in /etc/freeradius/ladp.attrmap. This file maps *RADIUS* attributes to LDAP attributes. Samba has NT/LM password hashes. Hence, the default mapping for LM-Password and NT-Password must be changed.

- Configure attribute mappings.

```
checkItem User-Password userPassword.
checkItem LMPassword sambaLMPassword
checkItem NTPassword sambaNTPassword
replyItem Tunnel-Type radiusTunnelType
replyItem Tunnel-Medium-Type radiusTunnelMediumType
replyItem Tunnel-Private-Group-Id radiusTunnelPrivateGroupId
```

*Configuring Additional Attributes Mappings*

Attributes are configured in /etc/freeradius/ladp.attrmap:

```
## Attributes for Extreme Networks Vendor-Specific RADIUS
replyItem   Extreme-Security-Profile radiusExtremeSecurityProfile
```

```
replyItem   Extreme-Netlogin-Vlan-Tag radiusExtremeNetloginVlanTag
replyItem   Extreme-Netlogin-Extended-Vlan radiusExtremeNetloginExtendedVlan
```

### *Modifying the RADIUS Schema*

Additional attributes for *RADIUS* must be configured to extend the RADIUS-LDAP-V3.schema under the /etc/openldap directory.

- Use the following commands to modify the RADIUS schema:

```
attributetype
( 1.3.6.1.4.1.3317.4.3.1.61
NAME 'radiusExtremeSecurityProfile'
DESC ''
EQUALITY caseIgnoreIA5Match
SYNTAX 1.3.6.1.4.1.1466.115.121.1.26
)
attributetype
( 1.3.6.1.4.1.3317.4.3.1.62
NAME 'radiusExtremeNetloginVlanTag'
DESC ''
EQUALITY caseIgnoreIA5Match
SYNTAX 1.3.6.1.4.1.1466.115.121.1.26
)
attributetype
( 1.3.6.1.4.1.3317.4.3.1.63
NAME 'radiusExtremeNetloginExtendedVlan'
DESC ''
EQUALITY caseIgnoreIA5Match
SYNTAX 1.3.6.1.4.1.1466.115.121.1.26
)
```

### *Configuring the Authentication Method for Supplicants*

The authentication method is configured in /etc/raddb/eap.conf. The authentication method used by Free*RADIUS* is the PEAP (Protected EAP) method. To activate PEAP, a TLS tunnel is required to encrypt communication between supplicant and RADIUS server. This means that server certificates are required.

- Configure the authentication method.

```
peap {
default_eap_type = mschapv2
}
tls {
private_key_password = whatever
private_key_file = ${raddbdir}/certs/cert-srv.pem
certificate_file = ${raddbdir}/certs/cert-srv.pem
CA_file = ${raddbdir}/certs/demoCA/cacert.pem
dh_file = ${raddbdir}/certs/dh
random_file = ${raddbdir}/certs/random
fragment_size = 1024
include_length = yes
}
```

### *Starting the FreeRADIUS Server*

- Start *RADIUS* in the foreground with debugging enabled.

  ```
  radiusd -X -f
  ```

## Implementation Notes for Specific RADIUS Servers

### *Cistron RADIUS*

Cistron *RADIUS* is a popular server, distributed under GPL. Cistron Radius can be found at:

When you configure the Cistron server for use with Extreme switches, you must pay close attention to the users file setup. The Cistron Radius dictionary associates the word Administrative-User with Service-Type value 6, and expects the Service-Type entry to appear alone on one line with a leading tab character.

The following is a user file example for read-write access:

```
adminuser   Auth-Type = System
Service-Type = Administrative-User,
Filter-Id = "unlim"
```

### *RSA Ace*

For users of Cistron's RSA SecureID® product, RSA offers *RADIUS* capability as part of their RSA/Ace Server® server software. It is mandatory to configure a matching shared-secret key on the switch and RSA Ace server for successful authentication.

### *Steel-Belted Radius*

For users who have the Steel-Belted Radius (SBR) server from Juniper Networks, it is possible to limit the number of concurrent login sessions using the same user account. This feature allows the use of shared user accounts, but limits the number of simultaneous logins to a defined value. Using this feature requires Steel-Belted Radius for *RADIUS* authentication and accounting. To limit the maximum concurrent login sessions under the same user account:

1. Configure RADIUS and RADIUS-Accounting on the switch.

   The RADIUS and RADIUS-Accounting servers used for this feature must reside on the same physical RADIUS server. Standard RADIUS and RADIUS-Accounting configuration is required as described earlier in this chapter.
2. Modify the SBR vendor.ini file and user accounts.
   a. To configure the SBR server, the file vendor.ini must be modified to change the Extreme Networks configuration value of ignore-ports to yes as shown in the example below:

      ```
      vendor-product       = Extreme Networks
      dictionary           = Extreme
      ignore-ports         = yes
      port-number-usage    = per-port-type
      help-id              = 2000
      ```

   b. After modifying the vendor.ini file, the desired user accounts must be configured for the Max-Concurrent connections. Using the SBR Administrator application, enable the check box for Max-Concurrent connections and fill in the desired number of maximum sessions.

### *Microsoft IAS*

To use Extreme Networks VSAs with the Internet Authentication Service (IAS) in Microsoft® Windows Server™ 2003, you must first create a Remote Access Policy and apply it so that user authentication occurs using a specific authentication type such as EAP-TLS, PEAP, or PAP. The following procedure

assumes that the Remote Access Policy has already been created and configured and describes how to define Extreme Networks VSAs in Microsoft IAS:

1. Open the IAS administration GUI application.
2. In the left window pane, select the **Remote Access Policies** section of the tree.
3. In the right window pane, double-click the desired Remote-Access policy name so you can edit it.
4. Click the **Edit-Profile** button in the lower-left corner, and then select the **Advanced** tab.
5. If any attributes already appear in the **Parameters** window, remove them by selecting the attribute and clicking the **Remove** button.
6. When the **Parameters** window is empty, proceed to the next step.
7. Click the **Add** button, which brings up the **Add Attributes** dialog window.
8. Scroll down the displayed list of *RADIUS* attributes and select the attribute named **Vendor-Specific**.
9. Double-click the **Vendor-Specific** attribute or click the **Add** button.

   The **Multivalued Attribute Information** dialog box should appear.
10. Click the **Add** button, which brings up the **Vendor-Specific Attribute Information** dialog window.
    a. Select the first radio button for **Enter Vendor Code** and enter the Extreme Networks vendor code value of `1916` in the text-box.
    b. Select the second radio button for **Yes, It conforms**.
    c. Verify both settings, and click the **Configure Attribute** button to proceed.

       The **Configure VSA (RFC compliant)** dialog window should now appear.

       The settings for this dialog window varies, depending on which product and attribute you wish to use in your network.
    d. In the first text-box enter the Extreme Networks VSA number for the attribute you want to configure (see Extreme Networks VSAs on page 919).
    e. Use the pull-down menu to select the **Attribute** format, which is the same as the attribute Type listed in Extreme Networks VSAs on page 919.

       Note
       For values of format integer you will have to select the type **Decimal** from the pull-down menu.

    f. Configure the desired value for the attribute.
    g. Once the desired values have been entered, click **OK**.
11. Click **OK** two more times to return to the **Add Attributes** dialog window.
12. Select **Close**, and then click **OK** twice to complete the editing of the Remote Access Policy profile.
13. To apply the configuration changes, stop and restart the Microsoft IAS service.

    After restarting the IAS service, new authentications should correctly return the Extreme Networks VSA after successful authentication. Users who were previously authenticated have to re-authenticate to before the new VSAs apply to them.
14. If you experience problems with the newly configured VSAs, use the following troubleshooting guidelines:
    a. If you have multiple IAS Remote Access Policies, verify that the user is being authenticated with the correct policy.
    b. Check the IAS System Log events within Microsoft Event Viewer to verify the user is authenticated through the policy where VSA settings are configured.

c. Check whether the VSA configuration performed above is correct.

   A mismatch in any of the VSA settings could cause authentication or VSA failure.

d. Verify that attributes such as "*VLAN* tag" or "VLAN name" correctly match the configuration of your ExtremeXOS switch and overall network topology.

   Invalid, or incorrect values returned in the VSA could prevent authenticated users from accessing network resources.

## Setting Up Open LDAP

To integrate an ExtremeXOS switch in an LDAP environment, a *RADIUS* server must be configured to communicate with the LDAP database. The following components are required to install the access control solution:

- Linux server with Linux Red Hat 4.0
- FREERADIUS 1.1.x
- OpenLDAP 2.3.x
- Extreme Networks switches
- Windows 7/Windows 8 clients

To configure Universal Port for use in an LDAP environment, use the following procedure:

1. Install and configure a RADIUS server on an existing Linux server as described in Installing and Testing the FreeRADIUS Server on page 927.
2. Install and configure OpenLDAP as described later in this section.
3. Add vendor specific attributes to the RADIUS and LDAP servers as described in Installing and Testing the FreeRADIUS Server on page 927 and later in this section.
4. Configure the edge switches as described in this guide.
5. Configure each supplicant as described in Configuring a Windows 7/Windows 8 Supplicant for 802.1X Authentication on page 936.

For complete instructions on setting up an LDAP server, see the product documentation for the LDAP server.

### *Installing OpenLDAP*

OpenLDAP software is an open source implementation of Lightweight Directory Access Protocol. This can be obtained from the site: www.openldap.org. To install OpenLDAP packages:

1. Verify the Red Hat Linux installed releases.

   The release number is stored in the /etc/redhat-release file.

2. Verify the version of OpenLDAP currently installed by entering the command `rpm -qa | grep openldap` at the Linux prompt.

```
# rpm -qa |grep openldap
openldap-2.3.xx-x
openldap-clients-2.3.xx-x
openldap-servers-2.3.xx-x
```

3. If you have a default Red Hat Linux installation, there is at least one OpenLDAP Red Hat Package Manager (RPM) installed.

   The LDAP RPMs can be found on the Red Hat CD or downloaded from one of the following RPM download sources:

   www.rpmfind.net and search for **openldap** and select the RPM based on the distribution

   www.redhat.com and select **Download**, and then search for **openldap**.
4. After downloading the RPMs to the Linux server, change to the download directory and start the installation using the rpm command:

   ```
   # rpm -ivh openldap*
   ```
5. Verify that the OpenLDAP RPMs have been installed with the `rpm -qa | grep openldap` command at the Linux prompt.

   ```
   # rpm -qa | grep openldap
   openldap-2.3.xx-x
   openldap-clients-2.3.xx-x
   openldap-servers-2.3.xx-x
   ```

### *Configuring OpenLDAP*

Once the build is complete, the slapd and slurpd daemons are located in /usr/local/libexec.

The config files are in /etc/openldap and ready to start the main server daemon, slapd.

### *Configuring slapd for Startup*

Before you start slapd, edit /etc/openldap/slapd.conf to include the location to store the data and details on who is allowed to access the data.

The following configuration changes need to be made:

- Change the suffix.
- Change the rootDN.
- Use slappasswd to generate rootpw.
- Add rootpw entry.

Use the following commands to configure slapd for startup:

```
database (use default)
suffix "dc=xxxxxx,dc=org"
rootdn "cn=xxxx,dc=xxxxxx,dc=org"
rootpw {SSHA}c5PemO1KWqz0254r4rnFVmxKA/evs4Hu
directory /var/lib/ldap
allow bind_v2
pidfile /var/run/slapd.pid
```

### *Adding New Schemas*

The *RADIUS* schema and Samba schema for PEAP authentication must be included into the slapd.conf file. After modifying the file, the LDAP server must be restarted to load the new schemas.

- Use the following commands to add new schemas:

  ```
  cp /usr/share/doc/freeradius-1.0.1/RADIUS-LDAPv3.schema /etc/openldap/schema/
  ```

  ```
  cp /usr/share/doc/samba-3.0.10/LDAP/samba.schema /etc/openldap/schema
  ```

- Use the following commands to modify slapd.conf:

```
include/etc/openldap/schema/RADIUS-LDAPv3.schema
include/etc/openldap/schema/samba.schema
```

*Populating the LDAP Database with Organization and User Entries*

- Use the following commands to make the user entry in the LDAP directory (slapd.conf):

```
dn: uid=newperson3,o=ldaptestdemo,dc=extremenetworks,dc=com
objectClass: top
objectClass: person
objectClass: radiusprofile << Defined in the RADIUS-LDAPv3 schema
objectClass: sambaSamAccount
sn: ldaptestdemo
uid: newperson3 <<< This username given in the Odyssey client
cn: newperson3
radiusTunnelMediumType: IEEE-802
radiusTunnelType: VLAN
radiusTunnelPrivateGroupId: 2 <<< Value of the VLAN tag
sambaNTPassword: A3A685F89364D4A5182B028FBE79AC38
sambaLMPassword: C23413A8A1E7665FC2265B23734E0DAC
userPassword:: e1NIQX00MXZzNXNYbTRPaHNwUjBFUU9raWdxbldySW89
sambaSID: S-1-0-0-28976
```

The Samba-related attributes can be populated in the LDAP server already if there is an LDAP-enabled Samba infrastructure in place.

**Note**

If the Samba related entries are not present, then the values for sambaNTPassword and sambaNMPPassword can be created by running the `mkntpwd` command.

```
cd /usr/share/doc/samba-3.0.10/LDAP/smbldap-tools/mkntpwd
make
./mkntpwd -L <password>   (provides value for sambaLMPassword attribute)
./mkntpwd -N <password>   (provides value for sambaNTPassword attribute)
```

*Restarting the LDAP Server*

- Use the following syntax to stop and start LDAP services:

  `service ldap restart`

  For phone authentication (which uses EAP based md5 authentication), the password is stored in clear text in the UserPassword field for the phone entries in LDAP.

*LDAP Configuration Example*

This configuration example is for Summit switches, but can also be used for other Extreme switches that support 802.1X.

Use the following commands to activate the switch for 802.1X port-based authentication:

```
create vlan voice
create vlan data
create vlan ldap
configure voice tag 10
configure data tag 20
configure ldap ipaddress 192.168.1.1/24
cnable ipforwarding
create vlan nvlan
en netlogin dot1x
en netlogin port 13-24 dot1x
```

```
configure radius netlogin primary server 192.168.1.2 1812 client-ip 192.168.1.1 vr VR-
Default
configure radius netlogin primary shared-secret extreme1
enable radius netlogin
enable netlogin dot1x
```

Configure the ports to run a script when a user is authenticated through *RADIUS* and LDAP:

```
configure upm event user-authenticate profile a-avaya ports 1-23
LDAP UID entries:
```

In the LDAP phone UID entry in the users file, use the following attribute to specify a profile to run on the switch:

```
Extreme-Security-Profile
```

To add the port as tagged in the voice *VLAN*, use the following attribute in the users file:

```
Extreme-Netlogin-Extended-Vlan = TVoice (use UData for a PC)
```

**Note**

It depends on the end-station to determine the fields required for authentication; XP uses EAP-PEAP and must have encrypted fields for the UID password. Avaya phones authenticate with MD-5 and must have an unencrypted field in LDAP.

#### Scripts

The following a-avaya script tells the phone to configure itself in the voice VLAN, and to send tagged frames.

The script also informs the phone of the file server and call server:

```
create upm profile a-avaya
create log message Starting_UPM_Script_AUTH-AVAYA
set var callServer 10.147.12.12
set var fileServer 10.147.10.3
set var voiceVlan voice
set var CleanupProfile CleanPort
set var sendTraps false
#
create log message Starting_UPM_AUTH-AVAYA_Port_$EVENT.USER_PORT
#*********************************************************
# adds the detected port to the device "unauthenticated" profile port list
#*********************************************************
create log message Updating_Unauthenticated_Port_List_Port_$EVENT.USER_PORT
#configure upm event user-unauthenticated profile CleanupProfile ports $EVENT.USER_PORT
#*********************************************************
# Configure the LLDP options that the phone needs
#*********************************************************
configure lldp port $EVENT.USER_PORT advertise vendor-specific dot1 vlan-name vlan
$voiceVlan
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme call-server
$callServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme file-server
$fileServer
configure lldp port $EVENT.USER_PORT advertise vendor-specific avaya-extreme dot1q-
framing tagged
configure lldp port $EVENT.USER_PORT advertise vendor-specific med capabilities
#configure lldp port $EVENT.USER_PORT advertise vendor-specific med policy application
voice vlan $voiceVlan dscp 46
# If port is PoE capable, uncomment the following lines
#*************************************************************
```

```
# Configure the POE limits for the port based on the phone requirement
#*****************************************************************
configure lldp port $EVENT.USER_PORT advertise vendor-specific med power-via-mdi
#configure inline-power operator-limit $EVENT.DEVICE_POWER ports $EVENT.USER_PORT
create log message UPM_Script_A-AVAYA_Finished_Port_$EVENT.USER_PORT
```

Note

Parts of the scripts make use of the QP8 profile. This is NOT recommended because the QP8 profile is used by EAPS. For voice, use QP7 for QOS.

This script uses tagging for the phone and the ports for the voice VLAN. This is NOT necessary; use multiple supplicant and untagged for the phones.

## Configuring a Windows 7/Windows 8 Supplicant for 802.1X Authentication

This section provides an overview procedure for configuring a Windows 7/Windows 8 supplicant. For complete instructions on setting up a Windows 7/Windows 8 supplicant, see the product documentation for Microsoft Windows XP.

Note

For enhanced security, install the FreeRADIUS server CA certificate (the CA that signed the certificate installed in eap.conf).

1. Open the network configuration panel, select the network card, enter the properties.
2. Click the **Authentication** tab.

   The Authentication dialog appears.
3. Enable 802.1X and disable authenticate as computer.
4. Choose EAP type of Protected EAP, then click **Properties**.
5. Deselect the Validate server certificate and select eap-mschapv2 as the authentication method.
6. Click **Configure**.
7. Select or clear the check box depending on whether you want to use the logon name and password, then click **OK**.

## Hypertext Transfer Protocol

The Hypertext Transfer Protocol (HTTP) is a set of rules for transferring and exchanging information (data, voice, images, and so on) on the World Wide Web. HTTP is based on a request-response model. An HTTP client initiates requests by establishing a TCP connection to a port on a remote host (port 80 by default). An HTTP server listening on that port waits for and then responds to the request; in many instances, the client is requesting a specific URL or IP address. Upon receiving a request, the destination server sends back the associated file or files and then closes the connection.

The web server in ExtremeXOS allows HTTP clients to access the switch on port 80 (by default) as well as the network login page without additional encryption or security measures. For information about secure HTTP transmission, including Secure Socket Layer (SSL), see Secure Socket Layer on page 944.

By default, HTTP is enabled on the switch.

- If you disabled HTTP access, you can re-enable HTTP access on the default port (80) using the following command:

  ```
  enable web http
  ```

- To disable HTTP, use the following command:

  ```
  disable web http
  ```

# Secure Shell 2

Secure Shell 2 (SSH2) is a feature of the ExtremeXOS software that allows you to encrypt session data between a network administrator using SSH2 client software and the switch or to send encrypted data from the switch to an SSH2 client on a remote system.

Configuration, image, public key, and policy files can be transferred to the switch using the Secure Copy Protocol 2 (SCP2).

The ExtremeXOS SSH2 switch application works with the following clients: Putty, SSH2 (version 2.x or later) from SSH Communication Security, and OpenSSH (version 2.5 or later).

## Enabling SSH2 for Inbound Switch Access

You must enable SSH2 on the switch before you can connect to the switch using an external SSH2 client.

Enabling SSH2 involves two steps:

1. Generating or specifying an authentication key for the SSH2 sessions.
2. Enabling SSH2 access by specifying a TCP port to be used for communication and specifying on which *virtual router (VR)* SSH2 is enabled.

After SSH2 has enabled, it TCP port 22 and is available on all virtual routers by default.

### *Standard Key Authentication*

An authentication key must be generated before the switch can accept incoming SSH2 sessions. This can be done automatically by the switch, or you can enter a previously generated key.

- To have the key generated by the switch, use the following command:

  ```
  configure ssh2 key
  ```

  The key generation process can take up to ten minutes. After the key has been generated, you should save your configuration to preserve the key.

- To use a key that has been previously created, use the following command:

  ```
  configure ssh2 key {pregenerated}
  ```

  The switch prompts you to enter the pregenerated key.

  Note

  The pregenerated key must be one that was generated by the switch. To get such a key, you can use the command `show ssh2 private-key` to display the key on the console. Copy the key to a text editor and remove the carriage return/line feeds. Finally, copy and paste the key into the command line. The key must be entered as one line.

The key generation process generates the SSH2 private host key. The SSH2 public host key is derived from the private host key and is automatically transmitted to the SSH2 client at the beginning of an SSH2 session.

### *User Key Based Authentication*

Public key authentication is an alternative method to password authentication that SSH uses to verify identity. You can generate a key pair consisting of a private key and a public key. The public key is used by the ExtremeXOS SSH server to authenticate the user.

In ExtremeXOS, user public keys are stored in the switch's configuration file; these keys are then associated (or bound) to a user.

The keys are configured on the switch in one of two ways:

- By copying the key to the switch using scp2/sftp2 with the switch acting as the server.
- By configuring the key using the CLI.

RSA and DSA encryption keys are both supported.

The public key can be loaded onto the switch using SCP or SFTP, where the switch is the server. The administrator can do this by using the SCP2 or SFTP2 client software to connect to and copy the key file to the switch. The public key file must have the extension ssh; for example, id_dsa_2048.ssh. When the .ssh file is copied to the switch, the key is loaded into the memory. The loaded public keys are saved to the configuration file (*.cfg) when the save command is issued via the CLI.

The key name is derived from the file name. For example, the key name for the file id_dsa_2048.ssh will be id_dsa_2048.

The key is associated with a user either implicitly, by pre-pending the user name to the file or explicitly using the CLI.

In order for a key to be bound or associated to a user, the user must be known. In other words, that user must have an entry in the local database on the switch. Once the user is authenticated, the user's rights (read-only or read/write) are obtained from the database.

The key can be associated with a user by pre-pending the user name to the file name. For example, admin.id_dsa_2048.ssh.

If the user specified in the filename does not exist on the switch, the key is still accepted, but will not be associated to any user. Once the user is added, the key can be associated with the user via the CLI. If the user name is not pre-pended to the filename, the key is accepted by the switch but is not associated with any user. The key can be then be associated with the user via the CLI.

You can also enter or paste the key using the CLI. There cannot be any carriage returns or new lines in the key. See the appropriate reference page in the *ExtremeXOS 16.2 Command Reference Guide* for additional details.

The host and user public keys can be written to a file in the config directory using the `create sshd2 key-file` command. This enables the administrator to copy the public key to an outside server.

### *Enabling SSH2*

- To enable SSH2, use the following command:

  `enable ssh2 {`**`access-profile`** `[`*`access_profile`* `|` **`none`**`]} {`**`port`** *`tcp_port_number`*`} {`**`vr`** `[`*`vr_name`* `|` **`all`** `|` **`default`**`]}`

You can also specify a TCP port number to be used for SSH2 communication. The TCP port number is 22 by default. The switch accepts IPv6 connections.

Before you initiate a session from an SSH2 client, ensure that the client is configured for any non-default access list or TCP port information that you have configured on the switch. After these tasks are accomplished, you may establish an SSH2-encrypted session with the switch. Clients must have a valid user name and password on the switch in order to log in to the switch after the SSH2 session has been established.

Up to eight active SSH2 sessions can run on the switch concurrently. If you enable the idle timer using the `enable idletimeout` command, the SSH2 connection times out after 20 minutes of inactivityby default. If you disable the idle timer using the `disable idletimeout` command, the SSH2 connection times out after 60 minutes of inactivity by default. This timeout value can be modified using the `configure ssh2 idletimeout` *`minutes`* command wherein *`minutes`* can be from 1 to 240 ". For more information please refer to the command help for "configure ssh2".

General technical information is also available at http://www.openssh.com/specs.html.

## Viewing SSH2 Information

- To view the status of SSH2 sessions on the switch, use the following command:

  `show management`

  The `show management` command displays information about the switch including the enable/disable state for SSH2 sessions and whether a valid key is present.

## Using ACLs to Control SSH2 Access

You can restrict SSH2 access by creating and implementing an *ACL* policy.

You configure an ACL policy to permit or deny a specific list of IP addresses and subnet masks for the SSH2 port.

The two methods to load ACL policies to the switch are:

- Use the `edit policy` command to launch a VI-like editor on the switch. You can create the policy directly on the switch.
- Use the `tftp` command to transfer a policy that you created using a text editor on another system to the switch.

For more information about creating and implementing ACLs and policies, see Security and ACLs on page 640.

### *Sample SSH2 Policies*

The following are sample policies that you can apply to restrict SSH2 access.

In the following example, named MyAccessProfile.pol, the switch permits connections from the subnet 10.203.133.0/24 and denies connections from all other addresses:

```
MyAccessProfile.pol
Entry  AllowTheseSubnets {
if  {
source-address 10.203.133.0 /24;
```

```
}
Then
{
permit;
}
}
```

In the following example, named MyAccessProfile.pol, the switch permits connections from the subnets 10.203.133.0/24 or 10.203.135.0/24 and denies connections from all other addresses:

```
MyAccessProfile.pol
Entry  AllowTheseSubnets {
if match any {
source-address 10.203.133.0 /24;
source-address 10.203.135.0 /24;
}
Then
{
permit;
}
}
```

In the following example, named MyAccessProfile_2.pol, the switch does not permit connections from the subnet 10.203.133.0/24 but accepts connections from all other addresses:

```
MyAccessProfile_2.pol
Entry dontAllowTheseSubnets {
if {
source-address 10.203.133.0 /24;
}
Then
{
deny;
}
}
Entry  AllowTheRest {
If  {
; #none specified
}
Then
{
permit;
}
}
```

In the following example, named MyAccessProfile_2.pol, the switch does not permit connections from the subnets 10.203.133.0/24 or 10.203.135.0/24 but accepts connections from all other addresses:

```
MyAccessProfile_2.pol
Entry dontAllowTheseSubnets {
if match any {
source-address 10.203.133.0 /24;
source-address 10.203.135.0 /24
}
Then
{
deny;
}
}
Entry  AllowTheRest {
If  {
; #none specified
}
```

```
Then
{
permit;
}
}
```

*Configuring SSH2 to Use ACL Policies*

This section assumes that you have already loaded the policy on the switch. For more information about creating and implementing *ACL*s and policies, see Security and ACLs.

- To configure SSH2 to use an ACL policy to restrict access, use the following command:

  `enable ssh2 {`**`access-profile`** `[`*`access_profile`* `|` **`none`**`]} {`**`port`** *`tcp_port_number`*`} {`**`vr`** `[`*`vr_name`* `|` **`all`** `|` **`default`**`]}`

  Use the **none** option to remove a previously configured ACL.

  In the ACL policy file for SSH2, the source-address field is the only supported match condition. Any other match conditions are ignored.

## Using SCP2 from an External SSH2 Client

In ExtremeXOS, the SCP2 protocol is supported for transferring configuration, image and public key policy files to the switch from the SCP2 client. The user must have administrator-level access to the switch. The switch can be specified by its switch name or IP address. ExtremeXOS only allows SCP2 to transfer to the switch files named as follows:

- *.cfg—ExtremeXOS configuration files
- *.pol—ExtremeXOS policy files
- *.xos—ExtremeXOS core image files
- *.xmod—ExtremeXOS modular package files
- *.ssh—Public key files

In the following examples, you are using a Linux system to move files to and from the switch at 192.168.0.120, using the switch administrator account admin. You are logged into your Linux system as user.

- To transfer the primary configuration file from the switch to your current Linux directory using SCP2, use the following command:

  ```
  [user@linux-server]# scp2 admin@192.168.0.120:primary.cfg primary.cfg
  ```

- To copy the policy filename test.pol from your Linux system to the switch, use the following command:

  ```
  [user@linux-server]# scp2 test.pol admin@192.168.0.120:test.pol
  ```

- To copy the image file test.xos from your Linux system to the switch, use the following command:

  ```
  [user@linux-server]# scp2 test.xos admin@192.168.0.120:test.xos
  ```

  Now you can use the command install image test.xos to install the image in the switch.

- To copy the SSH image file test.xmod from your Linux system to the switch, use the following command:

  ```
  [user@linux-server]# scp2 test.xmod admin@192.168.0.120:test.xmod
  ```

Now you can use the command install image test.xmod to install the image in the switch.

- To load the public key id_rsa.pub from your Linux system to the switch, use the following command:

```
[user@linux-server]# scp2 id_rsa.pub admin@192.168.0.120:test.ssh
```

This command loads the key into memory, which can be viewed with the command `show sshd2 user-key`.

## Understanding the SSH2 Client Functions on the Switch

An Extreme Networks switch can function as an SSH2 client. This means you can connect from the switch to a remote device running an SSH2 server and send commands to that device. You can also use SCP2 to transfer files to and from the remote device.

Note
ExtremeXOS 15.7.1 upgraded from openssh-3.9p1 to openssh-6.5p1.

Note
ExtremeXOS 16.2 adds the openssl-fips-ecp-2.0.9 open source library.

You do not need to enable SSH2 or generate an authentication key to use the SSH2 and SCP2 commands from the ExtremeXOS CLI.

Note
User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document.

- To send commands to a remote system using SSH2, use the following command:

  ssh2 {**cipher** [*cipher*} {**mac** *mac*} {**compression** [**on** | **off**]} {**port** *port*} {**user** *username*} {**vr** *vr_name*} *user@host* {*remote_command*}

  The remote commands can be any command acceptable by the remote system. You can specify the login user name as a separate argument or as part of the user@host specification. If the login user name for the remote system is the same as your user name on the switch, you can omit the username parameter entirely.

  For example, to obtain a directory listing from a remote Linux system with IP address 10.10.0.2 using SSH2, enter the following command:`ssh2 admin@10.10.0.2 ls`

- To initiate a file copy from a remote system to the switch using SCP2, use the following command:

  scp2 {**cipher** *cipher*} {**mac** *mac*} {**compression** [**on** | **off**]} {**port** *port*} {**vr** *vr_name*} [ *user@host:file local-file* | *local-file user@host:file* ]

  For example, to copy the configuration file test.cfg on host system1 to the switch, enter the following command:

```
scp2 admin@system1:test.cfg localtest.cfg
```

- To initiate a file copy to a remote system from the switch using SCP2, use the following command:

  scp2 {**cipher** *cipher*} {**mac** *mac*} {**compression** [**on** | **off**]} {**port** *port*} {**vr** *vr_name*} [ *user@host:file local-file* | *local-file user@host:file* ]

For example, to copy the configuration file engineering.cfg from the switch to host system1, enter the following command:

```
scp2 engineering.cfg admin@system1:engineering.cfg
```

*SSH/SCP Client Upgrade Limitations*

- Only password-based authentication is supported for SSH/SCP client.
- SCP client will not support upload of Image or BootROM files (i.e, *.xos, *.xmod, and *.xtr).
- SCP client will not support download of Image or BootRom files (like *.xos, *.xmod, and *.xtr) from EXOS SCP server.
- Only supports transfer of files such as *.cfg, *.xsf, *.py, *.pol and *.ssh files to/from the switch.
- Current version of openssl is not FIPS compliant.

## Using SFTP from an External SSH2 Client

The SFTP protocol is supported for transferring configuration, and policy files to the switch from the SFTP client. You must have administrator-level access to the switch. The switch can be specified by its switch name or IP address.

ExtremeXOS requires that SFTP transfer to the switch files named as follows:

- *.cfg—ExtremeXOS configuration files
- *.pol—ExtremeXOS policy files
- *.xos—ExtremeXOS core image file
- *.xmod—ExtremeXOS modular package file
- *.ssh—Public key files

In the following examples, you are using a Linux system to move files to and from the switch at 192.168.0.120, using the switch administrator account admin. You are logged into your Linux system as account user.

- To transfer the primary configuration file from the switch to your current Linux directory using SCP2, use the following command:

```
user@linux-server]# sftp admin@192.168.0.120
password: <Enter password>
sftp> put primary.cfg
```

- To copy the policy filename test.pol from your Linux system to the switch, use the following command:

```
user@linux-server]# sftp admin@192.168.0.120
password: <Enter password>
sftp> put test.pol
```

- To copy the image file test.xos from your Linux system to the switch, use the following command:

```
user@linux-server]# sftp admin@192.168.0.120
password: <Enter password>
sftp> put test.xos
```

- To load the public keyed_rsa.pub from your Linux system to the switch, use the following command:

```
user@linux-server]# sftp admin@192.168.0.120
password: <Enter password>
sftp> put id_rsa.pub id_rsa.ssh
```

For image file transfers, only one image file at a time can be available for installation. In other words, if test.xos needs to be installed, you must follow these steps:

a. Transfer test.xos into the switch using scp/sftp.
b. Install the test.xos image using the "install image" command.

- For image file transfers using SFTP or SCP (with the switch acting as the server), once the image is copied to the switch, validation of image is done by the switch, as indicated by the following log message:

  ```
  <Info:AAA.LogSsh> Validating Image file, this could take approximately
  30 seconds.. test.xos
  ```

  In stacking switches, you must receive a log message from all slots before proceeding with the installation.

  For example, in a four-switch stack, the installation can be proceed only after the following log messages are received:

  ```
  04/19/2007 17:41:09.71 <Info:AAA.LogSsh> Slot-1: Sent file "test.xos" info to backup
  04/19/2007 17:41:09.71 <Info:AAA.LogSsh> Slot-1: Sent file "test.xos" info to standby
  slot 3
  04/19/2007 17:41:09.71 <Info:AAA.LogSsh> Slot-1: Sent file "test-12.0.0.13.xos" info
  to standby slot 4
  ```

## Secure Socket Layer

Secure Socket Layer (SSLv3) is a feature of ExtremeXOS that allows you to authenticate and encrypt data over an SSL connection to provide secure communication.

The existing web server in ExtremeXOS allows HTTP clients to access the network login page. By using HTTPS on the web server, clients securely access the network login page using an HTTPS enabled web browser. Since SSL encrypts the data exchanged between the server and the client, you protect your data, including network login credentials, from unwanted exposure.

HTTPS access is provided through SSL and the Transport Layer Security (TLS1.0). These protocols enable clients to verify the authenticity of the server to which they are connecting, thereby ensuring that users are not compromised by intruders.

You must upload or generate a certificate for SSL server use. Before you can upload a certificate, you must purchase and obtain an SSL certificate from an Internet security vendor. The following security algorithms are supported:

- RSA for public key cryptography (generation of certificate and public-private key pair, certificate signing). RSA key size between 1024 and 4096 bits.
- Symmetric ciphers (for data encryption): RC4 and 3DES.
- Message Authentication Code (MAC) algorithms: RSA Data Security, Inc. *MD5 (Message-Digest algorithm 5)* Message-Digest Algorithm and SHA.

## Enabling and Disabling SSL

This section describes how to enable and disable SSL on your switch.

Note

Before ExtremeXOS 11.2, the Extreme Networks SSH module did not include SSL. To use SSL for secure HTTPS web-based login, you must upgrade your core software image to ExtremeXOS 11.2 or later, install the SSH module that works in concert with that core software image, and reboot the switch.

Keep in mind the following guidelines when using SSL:

- To use SSL with web-based login (secure HTTP access, HTTPS) you must specify the HTTPS protocol when configuring the redirect URL.
- If you are downloading the SSH module for the first time and want to immediately use SSL for secure HTTPS web-based login, restart the thttpd process after installing the SSH module. For more detailed information about activating the SSH module, see #unique_1918.
- To enable SSL and allow secure HTTP (HTTPS) access on the default port (443), use the following command:

  `enable web https`
- To disable SSL and HTTPS, use the following command:

  `disable web https`

## Creating Certificates and Private Keys

When you generate a certificate, the certificate is stored in the configuration file, and the private key is stored in the EEPROM. The certificate generated is in PEM format. By default ExtremeXOS uses the SHA-512 hashing algorithm to create the certificate. The certificate hashing algorithm can be configured using command `configure ssl certificate hash-algorithm` *`hash-algorithm`*. ExtremeXOS supports *MD5*, SHA-224, SHA-256, SHA-384 and SHA-512. The configured algorithm are used to create certificates from next time onwards. Use the `show ssl` command to check the currently configured Signature hashing algorithm.

- To create a self-signed certificate and private key that can be saved in the EEPROM, use the following command:

  `configure ssl certificate privkeylen` *`length`* **`country`** *`code`* **`organization`** *`org_name`* **`common-name`** *`name`*

  Make sure to specify the following:

  - Country code (maximum size of 2 characters)
  - Organization name (maximum size of 64 characters)
  - Common name (maximum size of 64)

  Any existing certificate and private key is overwritten.

  The size of the certificate depends on the RSA key length (privkeylen) and the length of the other parameters (country, organization name, and so forth) supplied by the user. If the RSA key length is 1024, then the certificate is approximately 1 kb. For an RSA key length of 4096, the certificate length is approximately 2 kb, and the private key length is approximately 3 kb.

*Downloading a Certificate Key from a TFTP Server*

You can download a certificate key from files stored in a TFTP server. If the operation is successful, any existing certificate is overwritten. After a successful download, the software attempts to match the public key in the certificate against the private key stored. If the private and public keys do not match, the switch displays a warning message similar to the following: `Warning: The Private Key does not match with the Public Key in the certificate.` This warning acts as a reminder to also download the private key.

Downloaded certificates and keys are not saved across switch reboots unless you save your current switch configuration. After you use the `save` command, the downloaded certificate is stored in the configuration file and the private key is stored in the EEPROM.

- To download a certificate key from files stored in a TFTP server, use the following command:

  `download ssl` *`ipaddress`* **`certificate`** *`cert_file`*

  Note
  For security measures, you can only download a certificate key in the *VR-Mgmt* virtual router.

- To see whether the private key matches with the public key stored in the certificate, use the following command:

  `show ssl {`**`detail`**`}`

  This command also displays:

  - HTTPS port configured. This is the port on which the clients will connect.
  - Length of the RSA key (the number of bits used to generate the private key).
  - Basic information about the stored certificate.

*Downloading a Private Key from a TFTP Server*

For security reasons, when downloading private keys, we recommend obtaining a pre-generated key rather than downloading a private key from a TFTP server. See Configuring Pregenerated Certificates and Keys on page 947 for more information.

- To download a private key from files stored in a TFTP server, use the following command:

  `download ssl` *`ipaddress`* **`privkey`** *`key_file`*

  If the operation is successful, the existing private key is overwritten. After the download is successful, a check is performed to find out whether the private key downloaded matches the public key stored in the certificate. If the private and public keys do not match, the switch displays a warning message similar to the following: `Warning: The Private Key does not match with the Public Key in the certificate.` This warning acts as a reminder to also download the corresponding certificate.

Downloaded certificates and keys are not saved across switch reboots unless you save your current switch configuration. After you use the `save` command, the downloaded certificate is stored in the configuration file and the private key is stored in the EEPROM.

*Configuring Pregenerated Certificates and Keys*

- To get the pregenerated certificate from the user, use the following command:

  ```
  configure ssl certificate pregenerated
  ```

  You can copy and paste the certificate into the command line followed by a blank line to end the command.

  This command is also used when downloading or uploading the configuration. Do not modify the certificate stored in the uploaded configuration file because the certificate is signed using the issuer's private key.

  The certificate and private key file should be in PEM format and generated using RSA as the cryptography algorithm.
- To get the pregenerated private key from the user, use the following command:

  ```
  configure ssl privkey pregenerated
  ```

  You can copy and paste the key into the command line followed by a blank line to end the command.

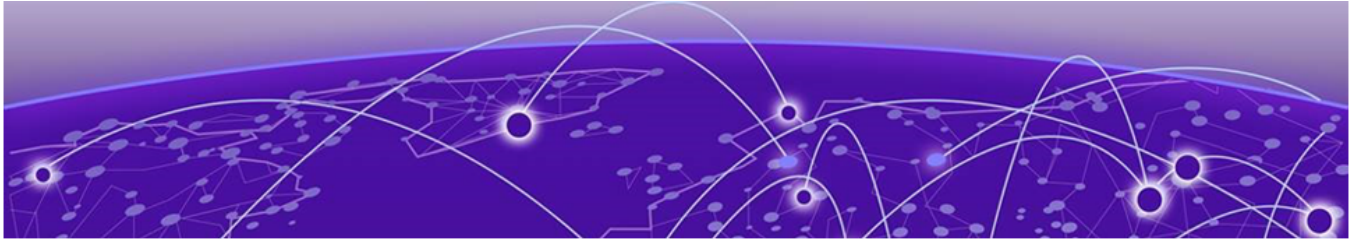
  This command is also used when downloading or uploading the configuration. The private key is stored in the EEPROM.

  The certificate and private key file should be in PEM format and generated using RSA as the cryptography algorithm.

## Displaying SSL Information

- To display whether the switch has a valid private and public key pair and the state of HTTPS access, use the following command:

  ```
  show ssl
  ```



# CLEAR-Flow



This chapter offers detailed information about the ExtremeXOS' implementation of CLEAR-Flow. This section provides an overview, as well as specific information on how to configure CLEAR-Flow, add CLEAR-Flow rules, and provides examples.

## CLEAR-Flow Overview

CLEAR-Flow is a broad framework for implementing security, monitoring, and anomaly detection in ExtremeXOS software. Instead of simply looking at the source and destination of traffic, CLEAR-Flow allows you to specify certain types of traffic that require more attention. After certain criteria for this traffic are met, the switch can either take an immediate, predetermined action, or send a copy of the traffic off-switch for analysis.

CLEAR-Flow is an extension to Access Control Lists (ACLs). You create *ACL (Access Control List)* policy rules to count packets of interest. CLEAR-Flow rules are added to the policy to monitor these ACL counter statistics. The CLEAR-Flow agent monitors the counters for the situations of interest to you and your network. You can monitor the cumulative value of a counter, the change to a counter over a sampling interval, the ratio of two counters, or even the ratio of the changes of two counters over an interval. For example, you can monitor the ratio between TCP SYN and TCP packets. An abnormally large ratio may indicate a SYN attack.

The counters used in CLEAR-Flow are either defined by you in an ACL entry, or can be a predefined counter. See Predefined CLEAR-Flow Counters on page 959 for a list and description of these counters.

If the rule conditions are met, the CLEAR-Flow actions configured in the rule are executed. The switch can respond by modifying an ACL that will block, prioritize, or mirror the traffic, executing a set of CLI

commands, or sending a report using a *SNMP (Simple Network Management Protocol)* trap or *EMS (Event Management System)* log message.

Note

CLEAR-Flow is available on platforms with an Edge, Advanced Edge, or Core license. These include BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules, BlackDiamond X8 series switches and Summit X440, X460, X480, X670, and X770 series switches. For more license information, see the *Feature License Requirements* document.

CLEAR-Flow is supported only on ingress. Any limitations on a given platform for a regular ACL also hold true for CLEAR-Flow.

## Configuring CLEAR-Flow

CLEAR-Flow is an extension to *ACL*s, so you must be familiar with configuring ACLs before you add CLEAR-Flow rules to your ACL policies. Creating ACLs is described in detail in the ACLs chapter. The chapter describes how to create ACL policies, the syntax of an ACL policy file, and how to apply ACL policies to the switch.

In this chapter, you will find information about the CLEAR-Flow rules that you add to ACL policies, including the CLEAR-Flow rules' syntax and behavior.

After creating the ACLs that contain CLEAR-Flow rules, and after applying the ACLs to the appropriate interface, you enable CLEAR-Flow on the switch. When CLEAR-Flow is enabled, the agent on the switch evaluates the rules, and when any rules are triggered, the CLEAR-Flow actions are executed.

- To enable CLEAR-Flow, enter the command:

  ```
  enable clear-flow
  ```

- To disable CLEAR-Flow, enter the command:

  ```
  disable clear-flow
  ```

  When you disable the CLEAR-Flow agent on the switch, CLEAR-Flow sampling stops and all rules are left in the current state.

  Note

  Any actions triggered while CLEAR-Flow is enabled will continue when CLEAR-Flow is disabled, unless explicitly stopped.

## Displaying CLEAR-Flow Configuration and Activity

- Dsplay the state of the CLEAR-Flow agent, any CLEAR-Flow policies on each interface, and the number of CLEAR-Flow rules by entering the command:

  ```
  show clear-flow
  ```

- Display the CLEAR-Flow rules and configuration by entering the command:.

  ```
  show clear-flow rule
  ```

- Display all the rules by entering the command:

  ```
  show clear-flow rule-all
  ```

  When CLEAR-Flow is enabled, any rules that satisfy the threshold will trigger and take action.

- Display the CLEAR-Flow rules that have been triggered by entering the command:.

  ```
  show clear-flow rule-triggered
  ```

- Display which ACLs have been modified by CLEAR-Flow rules.

  ```
  show clear-flow acl-modified
  ```

## Adding CLEAR-Flow Rules to ACLs

As described in the *ACLs* chapter, each ACL policy file consists of a number of named entries. Each entry consists of match conditions and actions to take if the entry is matched. CLEAR-Flow builds on the ACL concept to include rules that are periodically checked, and actions to take if a rule is triggered. The CLEAR-Flow entries are similar to the ACL entries.

The syntax of a CLEAR-Flow rule entry is:

```
entry <CLFrulename> {
if <match-type> { <match-conditions>;
}
Then {
<actions>;
}
}
```

Or you can specify an optional else clause:

```
entry <CLFrulename> {
if <match-type> { <match-conditions>;
}
Then {
<actions>;
} else {
<actions>;
}
}
```

In the CLEAR-Flow rule syntax, the *CLFrulename* is the name of the rule (maximum of 31 characters). The *match-type* specifies whether the rule is triggered when any of the expressions that make up the conditions are true (logical OR), or only when all of the expressions are true (logical AND). The *match-type* is an optional element. The *match-conditions* specifies the conditions that will trigger the rule, and how often to evaluate the rule. The *actions* in the then clause is the list of actions to take when the rule is triggered, and the optional else clause *actions* is the list of actions to take after the rule is triggered, and when the *match-conditions* later become false.

Note

When you create an ACL policy file that contains CLEAR-Flow rules, the CLEAR-Flow rules do not have any precedence, unlike the ACL entries. Each CLEAR-Flow rule specifies how often it should be evaluated. The order of evaluation depends on the sampling time and when the CLEAR-Flow agent receives the counter statistics. The order of the CLEAR-Flow rules in the policy file does not have any significance.

## CLEAR-Flow Rule Match Type

Match types are optional; the possible choices are:

- match all—All the match expressions must be true for a match to occur. This is the default.
- match any—If any match expression is true, then a match occurs.

## CLEAR-Flow Rule Match Conditions

In a CLEAR-Flow rule, the *match-conditions* portion consists of one to four expressions, an optional global-rule statement, and an optional period statement:

```
entry <CLFrulename> {
if <match-type> { <expression>;
<expression>;
<expression>;
<expression>;
global-rule;
period <interval>;
}
Then {
<actions>;
} else {
<actions>;
}
}
```

In the following example, the CLEAR-Flow rule (named cflow_count_rule_example) will be evaluated every ten seconds. The actions statements will be triggered if the value of counter1 (defined earlier in the *ACL* policy file) is greater than 1,000,000:

```
entry cflow_count_rule_example  {
if { count counter1 > 1000000 ;
period 10 ;
}
Then {
<actions>;
}
}
```

The global-rule statement is optional and affects how the counters are treated. An ACL that defines counters can be applied to more than one interface. You can specify the global-rule statement so that counters are evaluated for all the applied interfaces. For example, if a policy that defines a counter is applied to port 1:1 and 2:1, a CLEAR-Flow rule that used the global-rule statement would sum up the counts from both ports. Without the global-rule statement, the CLEAR-Flow rule would look at only the counts received on one port at a time.

The period *interval* statement is optional and sets the sampling interval, in seconds. This statement specifies how often the rule is evaluated by the CLEAR-Flow agent. If not specified, the default value is 5 seconds.

The five CLEAR-Flow rule expressions are: count; delta; ratio; delta-ratio; and rule. All of these expressions check the values of counters to evaluate if an action should be taken. The counters are either defined in the ACL entries that are defined on the switch, or are the predefined CLEAR-Flow

counters. When you use a counter statement in an ACL, you are defining the counter used by CLEAR-Flow to monitor your system.

*Count Expression*

A CLEAR-Flow count expression compares a counter with the threshold value.

The following is the syntax for a CLEAR-Flow count expression:

```
count <counterName> REL_OPER <countThreshold> ;
hysteresis <hysteresis> ;
```

The value of *countThreshold* and *hysteresis* can be specified as floating point numbers. The count statement specifies how to compare a counter with its threshold. The *counterName* is the name of an *ACL* counter referred to by an ACL rule entry and the *countThreshold* is the value compared with the counter. The REL_OPER is selected from the relational operators for greater than, greater than or equal to, less than, or less than or equal to (>, >=, <, <=).

The hysteresis *hysteresis* statement is optional and sets a hysteresis value for the threshold. After the count statement is true, the value of the threshold is adjusted so that a change smaller than the hysteresis value will not cause the statement to become false. For statements using the REL_OPER > or >=, the hysteresis value is subtracted from the threshold; for < or <=, the hysteresis value is added to the threshold.

Following is an example of a count expression used in a CLEAR-Flow rule:

```
entry cflow_count_rule_example  {
if { count counter1 > 1000000 ;
period 10 ;
}
Then {
<actions>;
}
}
```

Actions are discussed in CLEAR-Flow Rule Actions on page 957.

The following table is an example of evaluating the CLEAR-Flow count expression above multiple times. Notice that the rule is not triggered until evaluation 3, when the value of the counter is greater than 1,000,000.

**Table 113: Count Expression Evaluation Example**

| Evaluation | counter1 value | Rule triggered? |
|---|---|---|
| 1 | 384625 | No |
| 2 | 769250 | No |
| 3 | 1153875 | Yes |
| 4 | 1538500 | Yes |

See Count Expression Example on page 963 for a full example of an ACL and a CLEAR-Flow rule using a count expression.

*Delta Expression*

A CLEAR-Flow delta expression computes the difference from one sample to the next of a counter value.

This difference is compared with the threshold value. The following is the syntax for a CLEAR-Flow delta expression:

```
delta <counterName> REL_OPER <countThreshold> ;
hysteresis <hysteresis> ;
```

The values of *countThreshold* and *hysteresis* can be specified as floating point numbers. The delta expression specifies how to compare the difference in a counter value from one sample to the next with its threshold. The *counterName* is the name of an *ACL* counter referred to by an ACL rule entry and the *countThreshold* is the value compared with the difference in the counter from one sample to the next. The REL_OPER is selected from the relational operators for greater than, greater than or equal to, less than, or less than or equal to (>, >=, <, <=).

The hysteresis *hysteresis* statement is optional and sets a hysteresis value for the threshold. After the delta statement is true, the value of the threshold is adjusted so that a change smaller than the hysteresis value will not cause the statement to become false. For statements using the REL_OPER > or >=, the hysteresis value is subtracted from the threshold; for < or <=, the hysteresis value is added to the threshold.

For example, the following delta expression:

```
delta counter1 >= 100 ;
hysteresis 10 ;
```

will only be true after the delta of the counter reaches at least 100. At the time it becomes true, the hysteresis value is subtracted from the threshold (setting the threshold to 90). With the threshold now at 90, the condition would stay true until the delta of the counter becomes less than 90.

If the expression becomes false, the threshold is reset to its original value. You would use the hysteresis value to prevent the expression from vacillating between the true and false states if the difference between the counter values is near the threshold. If the hysteresis value is greater than the threshold value, the hysteresis value will be set to 0.

The following table is an example of evaluating the CLEAR-Flow delta expression above multiple times. Notice that the rule is not triggered until evaluation 4, when the delta value (the change in the counter value from one evaluation to the next) is greater than or equal to 100. After the rule is triggered, it remains triggered until the delta value is less than 90 (the original threshold minus the hysteresis), at evaluation 7. At evaluation 9, the rule is again triggered when the delta reaches 100. The rule will remain triggered until the delta drops below 90.

**Table 114: Delta Expression Evaluation Example**

| Evaluation | counter1 value | Delta value | Rule Triggered? |
|---|---|---|---|
| 1 | 397 | N/A | No |
| 2 | 467 | 70 | No |
| 3 | 547 | 80 | No |
| 4 | 657 | 110 | Yes |

**Table 114: Delta Expression Evaluation Example (continued)**

| Evaluation | counter1 value | Delta value | Rule Triggered? |
|---|---|---|---|
| 5 | 757 | 100 | Yes |
| 6 | 852 | 95 | Yes |
| 7 | 932 | 80 | No |
| 8 | 1031 | 99 | No |
| 9 | 1131 | 100 | Yes |
| 10 | 1230 | 99 | Yes |

See Delta Expression Example on page 963 for a full example of an ACL and a CLEAR-Flow rule using a delta expression.

### *Ratio Expression*

A CLEAR-Flow ratio expression compares the ratio of two counter values with the threshold value.

The following is the syntax for a CLEAR-Flow ratio expression:

```
ratio <counterNameA> <counterNameB> REL_OPER <countThreshold> ;
min-value <min-value> ;
hysteresis <hysteresis> ;
```

The values of *countThreshold* and *hysteresis* can be specified as floating point numbers, and the ratio is computed as a floating point number. The ratio statement specifies how to compare the ratio of two counters with its threshold. To compute the ratio, the value of *counterNameA* is divided by the value of *counterNameB*. That ratio is then compared with the *countThreshold*. The REL_OPER is selected from the relational operators for greater than, greater than or equal to, less than, or less than or equal to (>, >=, <, <=).

The min-value statement is optional, and sets a minimum value for the counters. If either counter is less than the minimum value, the expression evaluates to false. If not specified, the minimum value is 1.

The hysteresis *hysteresis* statement is optional and sets a hysteresis value for the threshold. After the ratio statement is true, the value of the threshold is adjusted so that a change smaller than the hysteresis value will not cause the statement to become false. For statements using the REL_OPER > or >=, the hysteresis value is subtracted from the threshold; for < or <=, the hysteresis value is added to the threshold.

For example, the following ratio expression:

```
ratio counter1 counter2 >= 5 ;
min-value 100;
hysteresis 1 ;
```

is true only after the ratio of the counters reaches at least 5 and the counter values are at least 100. At the time it became true, the hysteresis value would be subtracted from the threshold (setting the threshold to 4). With the threshold now at 4, the condition would stay true until the ratio of the counters became less than 4.

If the statement becomes false, the threshold is reset to its original value. You can use the hysteresis value to prevent the rule from vacillating between the true and false states if the ratio between the counter values is near the threshold. If the hysteresis value is greater than the threshold value, the hysteresis value will be set to 0.

The following table is an example of evaluating the CLEAR-Flow ratio expression above multiple times. Notice that the rule is not triggered at the first evaluation because both counters have not yet reached the min-value of 100. The rule first triggers at evaluation 3, when ratio of the two counters exceeds 5. After the rule is triggered, it remains triggered until the ratio value is less than 4 (the original threshold minus the hysteresis), at evaluation 5. At evaluation 7, the rule is again triggered when the ratio reaches 5. The rule will remain triggered until the ratio drops below 4.

**Table 115: Ratio Expression Evaluation Example**

| Evaluation | counter1 value | counter2 value | Ratio | Rule Triggered? |
|---|---|---|---|---|
| 1 | 427 | 70 | 6 | No |
| 2 | 941 | 235 | 4 | No |
| 3 | 2475 | 412 | 6 | Yes |
| 4 | 2308 | 570 | 4 | Yes |
| 5 | 2313 | 771 | 3 | No |
| 6 | 3597 | 899 | 4 | No |
| 7 | 5340 | 1065 | 5 | Yes |

See Ratio Expression Example on page 964 for a full example of an *ACL* and a CLEAR-Flow rule using a ratio expression.

### *Delta-Ratio Expression*

A CLEAR-Flow delta-ratio expression is a combination of the delta and ratio expressions.

The CLEAR-Flow agent computes the difference from one sample to the next for each of the two counters. The ratio of the differences is then compared to the threshold value. The following is the syntax for a CLEAR-Flow delta-ratio expression (note the similarity to the delta expression):

```
delta-ratio <counterNameA> <counterNameB> REL_OPER <countThreshold> ;
min-value <min-value> ;
hysteresis <hysteresis> ;
```

The values of *countThreshold* and *hysteresis* can be specified as floating point numbers, and the delta-ratio is computed as a floating point number. The delta-ratio statement specifies how to compare the ratio of the counter differences with its threshold. The difference of the sample values of *counterNameA* is divided by the difference of the sample values of *counterNameB*, to compute the ratio that is compared with the *countThreshold*. The REL_OPER is selected from the relational operators for greater than, greater than or equal to, less than, or less than or equal to (>, >=, <, <=).

The min-value statement is optional and sets a minimum value for the counters. If either counter is less than the minimum value, the expression evaluates to false. If not specified, the minimum value is 1.

The hysteresis *hysteresis* statement is optional, and sets a hysteresis value for the threshold. After the ratio statement is true, the value of the threshold is adjusted so that a change smaller than the

hysteresis value will not cause the statement to become false. For statements using the REL_OPER > or >=, the hysteresis value is subtracted from the threshold; for < or <=, the hysteresis value is added to the threshold.

For example, the following delta-ratio expression:

```
delta-ratio counter1 counter2 >= 5 ;
min-value 100 ;
hysteresis 1 ;
```

will only be true after the ratio of the deltas of the counters reached at least 5. At the time it became true, the hysteresis value would be subtracted from the threshold (setting the threshold to 4). With the threshold now at 4, the condition would stay true until the ratio of the deltas of the counters became less than 4.

If the statement becomes false, the threshold is reset to its original value. You can use the hysteresis value to prevent the rule from vacillating between the true and false states if the ratio of the deltas of the counters is near the threshold. If the hysteresis value is greater than the threshold value, the hysteresis value will be set to 0.

The following table is an example of evaluating the CLEAR-Flow delta-ratio expression above multiple times. Notice that the rule is not triggered at the second evaluation because both counters have not yet reached the min-value of 100. The rule first triggers at evaluation 4, when ratio of the two counters exceeds 5. After the rule is triggered, it remains triggered until the ratio value is less than 4 (the original threshold minus the hysteresis), at evaluation 6. At evaluation 8, the rule is again triggered when the ratio reaches 5. The rule will remain triggered until the ratio drops below 4.

**Table 116: Delta-Ratio Expression Evaluation Example**

| Evaluation | counter1 value | counter1 delta | counter2 value | counter2 delta | Ratio | Rule Triggered? |
|---|---|---|---|---|---|---|
| 1 | 110 | N/A | 20 | N/A | N/A | No |
| 2 | 537 | 427 | 90 | 70 | 6 | No |
| 3 | 1478 | 941 | 325 | 235 | 4 | No |
| 4 | 3953 | 2475 | 737 | 412 | 6 | Yes |
| 5 | 6261 | 2308 | 1307 | 570 | 4 | Yes |
| 6 | 8574 | 2313 | 2078 | 771 | 3 | No |
| 7 | 12171 | 3597 | 2977 | 899 | 4 | No |
| 8 | 17511 | 5340 | 4042 | 1065 | 5 | Yes |

See Delta-Ratio Expression Example on page 965 for a full example of an *ACL* and a CLEAR-Flow rule using a delta-ratio expression.

### *Rule-True-Count Expression*

A CLEAR-Flow rule-true-count expression compares how many times a CLEAR-Flow rule is true with a threshold value.

One use is to combine multiple rules together into a complex rule. The following is the syntax for a CLEAR-Flow rule-true-count expression:

```
rule-true-count <ruleName> REL_OPER <countThreshold> ;
```

The rule-true-count statement specifies how to compare how many times a CLEAR-Flow rule is true with the expression threshold. The *ruleName* is the name of the CLEAR-Flow rule to monitor and the *countThreshold* is the value compared with the number of times the rule is true. The REL_OPER is selected from the relational operators for greater than, greater than or equal to, less than, or less than or equal to (>, >=, <, <=).

For example, the following delta-ratio expression:

```
rule-true-count cflow_count_rule_example >= 5 ;
```

will only be true after the CLEAR-Flow rule cflow_count_rule_example has been true at least five times. If the rule cflow_count_rule_example becomes true and remains true, and the period for cflow_count_rule_example is the default five seconds, the rule would have to be true for at least 20 seconds before the rule-true-count expression will become true. If the period of the rule cflow_count_rule_example is 10 seconds, it will need to be true for at least 40 seconds before the rule-true_count expression becomes true.

## CLEAR-Flow Rule Actions

CLEAR-Flow rules specify an action to take when the rule is triggered and can optionally specify an action to take when the expression is false.

Because more than one action can be taken in a single rule, the collection of actions is referred to as an action list. The following sections describe the different rule actions:

- Permit/Deny on page 957
- QoS Profile on page 958
- Mirror on page 958
- SNMP Trap on page 958
- Syslog on page 958
- CLI on page 959

Additionally, the *SNMP* trap, syslog, and CLI rule actions can use keyword substitution to make the rule actions more flexible. The keyword substitutions are described at the end of the rule action descriptions. See Keyword Substitution on page 959 for more information.

### *Permit/Deny*

This action modifies an existing *ACL* rule to permit or block traffic that matches that rule.

- To change an ACL to permit, use the following syntax:

  `permit <ACLRuleName>`
- To change an ACL to deny, use the following syntax:

  `deny <ACLRuleName>`

*QoS Profile*

This action modifies an existing *ACL* rule to set the *QoS (Quality of Service)* profile for traffic that matches that rule.

- To change the ACL to forward to QoS profile <QPx>, use the following syntax:

  `qosprofile <ACLRuleName> <QPx>`

  `For example:`

  ```
  qosprofile acl_rule_1 QP3
  ```

*Mirror*

This action modifies an existing *ACL* rule to mirror traffic that matches that rule, or to stop mirroring that traffic. The mirroring port must be enabled when mirroring on an ACL rule is turned on. This could be configured earlier, or use the CLI action to execute CLI commands to configure mirroring at the same time.

- To change the ACL to mirror traffic, use the following syntax:

  `mirror [add|delete] <ACLRuleName>`

  For example (enabling mirroring from within CLEAR-Flow rule):

  ```
  enable mirror to port 7:4 tagged
  mirror add acl_rule_1
  ```

*SNMP Trap*

This action sends an *SNMP* trap message to the trap server, with a configurable ID and message string, when the rule is triggered. The message is sent periodically with interval *period* seconds. If *period* is 0, or if this optional parameter is not present, the message is sent only once when the rule is triggered. The interval must be a multiple of the rule sampling/evaluation interval, or the value will be rounded down to a multiple of the rule sampling/evaluation interval.

- To send an SNMP trap, use the following syntax:

  `snmptrap <id> <message> <period>`

*Syslog*

This action sends log messages to the ExtremeXOS *EMS* sever. The possible values for message level are: DEBU, INFO, NOTI, WARN, ERRO, and CRIT.

The message is sent periodically with interval *period* seconds. If *period* is 0, or if this optional parameter is not present, the message is sent only once when the rule is triggered. The interval must be a multiple of the rule sampling/evaluation interval, or the value will be rounded down to a multiple of the rule sampling/evaluation interval.

The messages are logged on both management modules (MSMs/MMs), so if the backup log is sent to the primary MSM/MM, then the primary MSM/MM will have duplicate log messages.

- To send a log message, use the following syntax:

  `syslog <message> <level> <period>`

*CLI*

This action executes a CLI command. There is no authentication or checking the validity of each command. If a command fails, the CLI will log a message in the *EMS* log.

- To execute a CLI command, use the following syntax:

  `cli <cliCommand>`

  where `<cliCommand>` is a quoted string.

*Keyword Substitution*

To make the *SNMP* trap, syslog, and CLI actions more flexible, keyword substitutions are supported in the syslog and SNMP trap message strings, as well as in the CLI command strings.

The following table lists the keywords and their substitutions.

If a keyword is not supported, or a counter name is not found, a string of "unknownKeyword[$keyword]" will be substituted.

For the $vlanName and $port keyword, the keyword **`all`** will be substituted for those rules in the wildcard *ACL* Some CLI commands do not support the **`all`** keyword, so caution must be used with CLI commands that use this feature.

A maximum of ten different counter substitutions can be used per rule, including counters used in expressions. For example, if a rule uses four counters in its expressions, then we can use six more different counters in keyword substitutions, for a total of ten.

**Table 117: Keyword Substitutions**

| Keyword | Substitution |
|---|---|
| $policyName | Replace with the policy name. |
| $ruleName | Replace with the CLEAR-Flow rule name. |
| $<counterName> | Replace with counter value for the indicated counter name. |
| $ruleValue | Replace with the current expression value. |
| $ruleThreshold | Replace with the expression threshold value. |
| $ruleInterval | Replace with the rule sampling/evaluation interval. |
| $vlanName | Replace with the interface *VLAN (Virtual LAN)* name. |
| $port | Replace with the interface port number. |

*Predefined CLEAR-Flow Counters*

A number of packet statistics are gathered by the ExtremeXOS kernel.

To allow you to use these statistics in CLEAR-Flow expressions, these kernel counters are now available for use with CLEAR-Flow. Most of the counter names are based directly on well known names from

common kernel structures and MIBs. The names are modified from their familiar form by pre-pending the characters sys_ to the counter names.

**Table 118: Predefined CLEAR-Flow Counters**

| Counter Name | Description |
|---|---|
| sys_IpInReceives | The total number of input IP packets received from interfaces, including those received in error. |
| sys_IpInHdrErrors | The number of input IP packets discarded due to errors in their IP headers, including bad checksums, version number mismatch, other format errors, time-to-live exceeded, errors discovered in processing their IP options, etc. |
| sys_IpInAddrErrors | The number of input IP packets discarded because the IP address in their IP header's destination field was not a valid address to be received at this entity. This count includes invalid addresses (for example, 0.0.0.0) and addresses of unsupported Classes (for example, Class E). |
| sys_IpForwDatagrams | The number of input IP packets for which this entity was not their final IP destination, as a result of which an attempt was made to find a route to forward them to that final destination. |
| sys_IpInUnknownProtos | The number of locally-addressed IP packets received successfully but discarded because of an unknown or unsupported protocol. |
| sys_IpInDiscards | The number of input IP packets for which no problems were encountered to prevent their continued processing, but which were discarded (for example, for lack of buffer space). Note that this counter does not include any IP packets discarded while awaiting re-assembly. |
| sys_IpInDelivers | The total number of input IP packets successfully delivered to IP user-protocols (including *ICMP (Internet Control Message Protocol)*). |
| sys_IpOutRequests | The total number of IP packets which local IP user-protocols (including ICMP) supplied to IP in requests for transmission. Note that this counter does not include any IP packets counted in ipForwDatagrams. |
| sys_IpOutDiscards | The number of output IP packets for which no problem was encountered to prevent their transmission to their destination, but which were discarded (for example, for lack of buffer space). Note that this counter would include IP packets counted in ipForwDatagrams if any such packets met this (discretionary) discard criterion. |
| sys_IpOutNoRoutes | The number of IP packets discarded because no route could be found to transmit them to their destination. Note that this counter includes any packets counted in ipForwDatagrams which meet this `no-route' criterion. |
| sys_IpReasmTimeout | The maximum number of seconds which received fragments are held while they are awaiting reassembly at this entity. |
| sys_IpReasmReqds | The number of IP fragments received which needed to be reassembled at this entity. |
| sys_IpReasmOKs | The number of IP packets successfully re-assembled. |

[15] Most of these descriptions can be found in RFC 2011, SNMPv2 Management Information Base for the Internet Protocol using SMIv2.

**Table 118: Predefined CLEAR-Flow Counters (continued)**

| Counter Name | Description |
|---|---|
| sys_IpReasmFails | The number of failures detected by the IP re-assembly algorithm (for whatever reason: timed out, errors, etc.). Note that this is not necessarily a count of discarded IP fragments since some algorithms (notably the algorithm in RFC 815) can lose track of the number of fragments by combining them as they are received. |
| sys_IpFragOKs | The number of IP packets that have been successfully fragmented at this entity. |
| sys_IpFragFails | The number of IP packets that have been discarded because they needed to be fragmented at this entity but could not be, for example, because their Don't Fragment flag was set. |
| sys_IpFragCreates | The number of IP packet fragments that have been generated as a result of fragmentation at this entity. |
| sys_IcmpInMsgs | The total number of ICMP messages which the entity received. Note that this counter includes all those counted by icmpInErrors. |
| sys_IcmpInErrors | The number of ICMP messages which the entity received but determined as having ICMP-specific errors (bad ICMP checksums, bad length, etc.). |
| sys_IcmpInDestUnreachs | The number of ICMP Destination Unreachable messages received. |
| sys_IcmpInTimeExcds | The number of ICMP Time Exceeded messages received. |
| sys_IcmpInParmProbs | The number of ICMP Parameter Problem messages received. |
| sys_IcmpInSrcQuenchs | The number of ICMP Source Quench messages received. |
| sys_IcmpInRedirects | The number of ICMP Redirect messages received. |
| sys_IcmpInEchos | The number of ICMP Echo (request) messages received. |
| sys_IcmpInEchoReps | The number of ICMP Echo Reply messages received. |
| sys_IcmpInTimestamps | The number of ICMP Timestamp (request) messages received. |
| sys_IcmpInTimestampReps | The number of ICMP Timestamp Reply messages received. |
| sys_IcmpInAddrMasks | The number of ICMP Address Mask Request messages received. |
| sys_IcmpInAddrMaskReps | The number of ICMP Address Mask Reply messages received. |
| sys_IcmpOutMsgs | The total number of ICMP messages which this entity attempted to send. Note that this counter includes all those counted by IcmpOutErrors. |
| sys_IcmpOutErrors | The number of ICMP messages which this entity did not send due to problems discovered within ICMP such as a lack of buffers. This value should not include errors discovered outside the ICMP layer such as the inability of IP to route the resultant datagram. In some implementations there may be no types of error which contribute to this counter's value. |
| sys_IcmpOutDestUnreachs | The number of ICMP Destination Unreachable messages sent. |
| sys_IcmpOutTimeExcds | The number of ICMP Time Exceeded messages sent. |
| sys_IcmpOutParmProbs | The number of ICMP Parameter Problem messages sent. |
| sys_IcmpOutSrcQuenchs | The number of ICMP Source Quench messages sent. |
| sys_IcmpOutRedirects | The number of ICMP Redirect messages sent. |
| sys_IcmpOutEchos | The number of ICMP Echo (request) messages sent. |

**Table 118: Predefined CLEAR-Flow Counters (continued)**

| Counter Name | Description |
|---|---|
| sys_IcmpOutEchoReps | The number of ICMP Echo Reply messages sent. |
| sys_IcmpOutTimestamps | The number of ICMP Timestamp (request) messages sent. |
| sys_IcmpOutTimestampReps | The number of ICMP Timestamp Reply messages sent. |
| sys_IcmpOutAddrMasks | The number of ICMP Address Mask Request messages sent. |
| sys_IcmpOutAddrMaskReps | The number of ICMP Address Mask Reply messages sent. |
| sys_IcmpInProtoUnreachs | The number of incoming ICMP packets addressed to a not-in-use / unreachable / invalid protocol. This message is in the general category of ICMP destination unreachable error messages. |
| sys_IcmpInBadLen | The number of incoming bad ICMP length packets. |
| sys_IcmpInBadCode | The number of incoming ICMP packets with a bad code field value. |
| sys_IcmpInTooShort | The number of incoming short ICMP packets. |
| sys_IcmpInBadChksum | The number of incoming ICMP packets with bad checksums. |
| sys_IcmpInRouterAdv | The number of incoming ICMP router advertisements. Router advertisements are used by IP hosts to discover addresses of neighboring routers. |
| sys_IcmpOutProtoUnreachs | The number of outgoing ICMP packets addressed to a not-in-use / unreachable / invalid protocol. This message is in the general category of ICMP destination unreachable error messages. |
| sys_IcmpOutRouterAdv | The number of outgoing ICMP router advertisements. Router advertisements are used by IP hosts to discover addresses of neighboring routers. |
| sys_IgmpInQueries | The number of Host Membership Query messages that have been received on this interface. |
| sys_IgmpInReports | The number of Host Membership Report messages that have been received on this interface for this group address. |
| sys_IgmpInLeaves | The number of incoming *IGMP (Internet Group Management Protocol)* leave requests. |
| sys_IgmpInErrors | The number of incoming IGMP errors. |
| sys_IgmpOutQueries | The number of Host Membership Query messages that have been sent on this interface. |
| sys_IgmpOutReports | The number of Host Membership Report messages that have been sent on this interface for this group address. |
| sys_IgmpOutLeaves | The number of outgoing IGMP leave requests. |

[16] The length of an ICMP packet depends on the type and code field.

# CLEAR-Flow Rule Examples

In the examples that follow, one to two *ACL* rule entries are followed by a CLEAR-Flow rule entry. The examples illustrate the four CLEAR-Flow rule expressions: count, delta, ratio, and delta-ratio.

## Count Expression Example

In the following example, every ten seconds the CLEAR-Flow agent will request the counter1 statistics from the hardware.

After it receives the counter value, it will evaluate the CLEAR-Flow rule. If the value of counter1 is greater than 1,000,000 packets, the CLEAR-Flow agent will send a trap message to the *SNMP* master, and change the *ACL* acl_rule1 to block traffic (acl_rule1 is modified to a deny rule).

Since there is no period configured for the snmptrap statement, the message is sent only once.

```
entry acl_rule1 {
if {
destination-address 192.168.16.0/24;
destination-port 2049;
protocol tcp;
} then {
count counter1;
}
}
entry cflow_count_rule_example  {
if { count counter1 > 1000000 ;
period 10 ;
}
Then {
snmptrap 123 "Traffic on acl_rule1 exceeds threshold";
deny acl_rule1;
}
}
```

## Delta Expression Example

In this example, every ten seconds the CLEAR-Flow agent will request the counter1 statistics from the hardware.

After it receives the counter value, it will then evaluate the rule. If the delta (change) of the counter1 value from the last sampled value ten seconds ago is greater than or equal to 1,000 packets, the CLEAR-Flow agent will send a trap message to the *SNMP* master and change the *ACL* acl_rule1 to move the traffic to QP3. In addition, reduce the peak rate to 5 Kbps on QP3. As long as the delta continues to be greater than or equal to 1000 packets, the CLEAR-Flow agent will repeatedly send a trap message every 120 seconds. When the delta falls below the threshold, the agent will execute the two actions in the else portion; it will send a single SNMP trap message, return the traffic to QP1, and reset QP3 to its original bandwidth.

```
entry acl_rule1 {
if {
destination-address 192.168.16.0/24;
destination-port 2049;
protocol tcp;
} then {
count counter1;
```

```
}
}
entry cflow_delta_rule_example  {
if  { delta counter1 >= 100000 ;
period 10 ;
} then {
snmptrap 123 "Traffic to 192.168.16.0/24 exceed rate limit" 120;
qosprofile acl_rule1 QP3;
cli "configure qosprofile qp3 peak_rate 5 K ports all" ;
} else {
snmptrap 123 "Traffic to 192.168.16.0/24 falls below rate limit";
qosprofile acl_rule1 QP1;
cli "configure qosprofile qp3 maxbw 100 ports all" ;
}
}
```

## Ratio Expression Example

In this example, every two seconds the CLEAR-Flow agent will request the counter1 and counter2 statistics from the hardware.

After it receives the two counter values, it will then check each counter value against its minimum valid threshold, which is 1,000. If both of the counter values is greater than 1,000, it then calculates the ratio of counter1 and counter2. If the ratio is greater than 5, the agent will execute the actions in the then clause, which consists of logging a message to the syslog server. Before logging the syslog string, the agent will replace the $ruleName keyword with the string cflow_ratio_rule_example, the $ruleValue keyword with the calculated ratio value, and the $ruleThreshold keyword with a value of 5. If either of the counter values is below the minimum value of 1,000, or the ratio is below the threshold of 5, the expression is false and no action is taken.

```
entry acl_rule1 {
if {
protocol udp;
} then {
count counter1;
}
}
entry acl_rule2 {
if {
protocol tcp;
} then {
count counter2;
}
}
entry cflow_ratio_rule_example  {
if  { ratio counter1 counter2 > 5 ;
period 2;
min-value 1000;
}
Then {
syslog "Rule $ruleName threshold ratio $ruleValue exceeds limit $ruleThreshold";
}
}
```

## Delta-Ratio Expression Example

In this example, every two seconds, the CLEAR-Flow agent will request the tcpSynCounter and tcpCounter values from the hardware.

After it receives the two counter values, it will first calculate the delta for each of the counters and then check each counter's delta value for its minimum value, which is 100. If both of the counters' delta values are greater than 100, it then calculates the ratio of the delta of two counters. If the ratio is greater than 10, then the agent will log a warning message and deny all SYN traffic on the interface. No period value for the syslog message is given, so the message will be logged once when the expression first becomes true. When the expression transitions from true to false, a different message will be logged and the SYN traffic on the interface will be permitted again. The delta-ratio value has to fall below a threshold of 8 for the expression to be evaluated to be false.

```
entry acl_syn {
if {
protocol tcp_flags SYN;
} then {
count tcpSynCounter;
}
}
entry acl_tcp {
if {
protocol tcp;
} then {
count tcpCounter;
}
}
entry cflow_delta_ratio_rule_example  {
if  { delta-ratio tcpSynCounter tcpCounter > 10 ;
period 2;
min-value 100;
              threshold 8;
            } then {
syslog "Syn attack on port $port is detected" WARN;
deny acl_syn;
} else {
syslog "Syn attack on port $port is no longer detected" WARN;
permit acl_syn;
}
}
```

# EAPS



This chapter provides an overview and discusses various topologies of Extreme's Automatic Protection Switching (EAPS) feature. The chapter offers configuration and monitoring details, and also provides configuration examples.

## EAPS Protocol Overview

The EAPS protocol provides fast protection switching to Layer 2 switches interconnected in an Ethernet ring topology, such as a Metropolitan Area Network (MAN) or large campus (see the following figure).

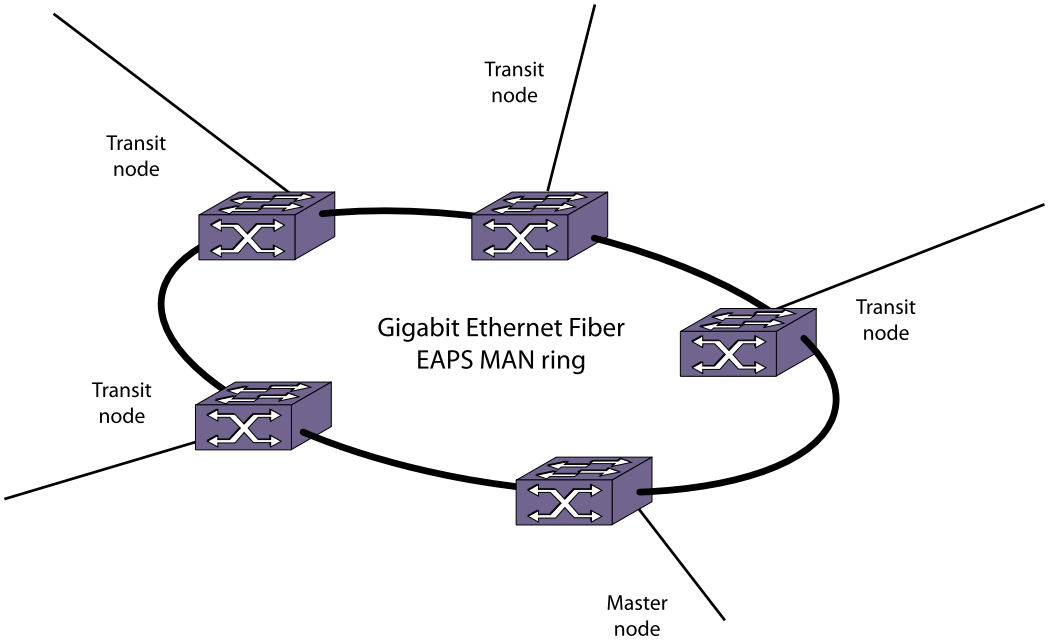



**Figure 115: Gigabit Ethernet Fiber EAPS MAN Ring**

## EAPS Benefits

EAPS offers the following benefits:

- Fast Recovery time for link or node failures—When a link failure or switch failure occurs, EAPS provides fast recovery times. EAPS provides resiliency for voice, video and data services.
- Scalable network segmentation and fault isolation—EAPS domains can protect groups of multiple VLANs, allowing scalable growth and broadcast loop protection. EAPS domains provide logical and physical segmentation, which means the failures in one EAPS ring do not impact network service for other rings and VLANs.
- Resilient foundation for non-stop IP routing services—EAPS provides a resilient foundation for upper level routing protocols such as *OSPF (Open Shortest Path First)* and *BGP (Border Gateway Protocol)*, minimizing route-flapping and dropped neighbors within the routed IP network.
- Predictable convergence regardless of failure location—EAPS provides consistent and predictable recovery behavior regardless of where link failures occur. The simple blocking architecture and predictable performance of EAPS allows for enforceable Service Level Agreements (SLAs). This allows easier network troubleshooting and failure scenario analysis without lengthy testing or debugging on live production networks.

EAPS protection switching is similar to what can be achieved with the *STP (Spanning Tree Protocol)*, but EAPS offers the advantage of converging in less than one second when a link in the ring breaks.

An Ethernet ring built using EAPS can have resilience comparable to that provided by SONET rings, at a lower cost and with fewer restraints (such as ring size). The EAPS technology developed by Extreme Networks to increase the availability and robustness of Ethernet rings is described in RFC 3619: Extreme Networks' Ethernet Automatic Protection Switching (EAPS) Version 1.

## EAPS Single Ring Topology

The simplest EAPS configuration operates on a single ring.

This section describes how this type of EAPS configuration operates. Later sections describe more complex configurations.

An EAPS domain consists of one master node and one or more transit nodes (see the following figure), and includes one control *VLAN (Virtual LAN)* and one or more protected VLANs.

A domain is a single instance of the EAPS protocol that defines the scope of protocol operation. A single logical EAPS domain typically exists on a given physical ring topology (fiber or copper).

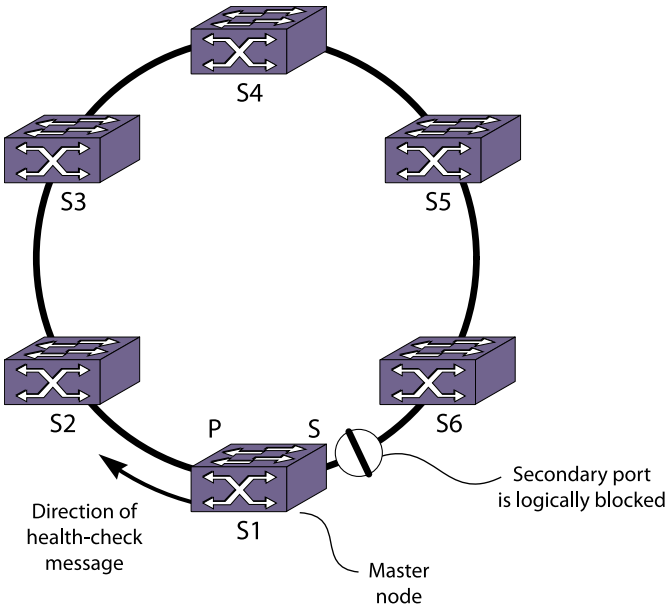



**Figure 116: EAPS Operation**

A protected VLAN is a user data VLAN that uses the ring for a protected connection between all edge ports. The protected VLAN uses 802.1q trunking on the ring ports and supports tagged and untagged edge ports.

One ring port of the master node is designated the master node's primary port (P), and another port is designated as the master node's secondary port (S) to the ring. In normal operation, the master node blocks the secondary port for all protected VLAN traffic, thereby preventing a loop in the ring. (The spanning tree protocol, *STP*, provides the same type of protection.) Traditional Ethernet bridge learning and forwarding database mechanisms direct user data around the ring within the protected VLANs.

Note

Although primary and secondary ports are configured on transit nodes, both port types operate identically as long as the transit node remains a transit node. If the transit node is reconfigured as a master node, the configured states of the primary and secondary ports apply.

The control VLAN is a dedicated 802.1q tagged VLAN that is used to transmit and receive EAPS control frames on the ring. The control VLAN can contain only two EAPS ring ports on each node. Each EAPS domain has a unique control VLAN, and control traffic is not blocked by the master node at any time. The control VLAN carries the following EAPS control messages around the ring:

- Health-check messages, which are sent from the master node primary port. Transit nodes forward health-check messages toward the master node secondary port on the control VLAN. When the master node receives a health check message on the secondary port, the EAPS ring is considered intact.
- Link-down alert messages, which are sent from a transit node to the master node when the transit node detects a local link failure.
- Flush-*FDB (forwarding database)* messages, which are sent by the master node to all transit nodes when ring topology changes occur. Upon receiving this control frame, the transit node clears its MAC address forwarding table (FDB) and relearns the ring topology.

When the master node detects a failure, due to an absence of health-check messages or a received link-down alert, it transitions the EAPS domain to the Failed state and unblocks its secondary port to allow data connectivity in the protected VLANs.

## EAPS Multiple Ring Topology

EAPS works with multiple ring networks to support more complex topologies for interconnecting multiple EAPS domains. This allows larger EAPS end-to-end networks to be built from edge to core.

**Note**
Minimal EAPS support is provided at all license levels. EAPS multiple ring topologies and common link topologies are supported at higher license levels as described in the *Feature License Requirements* document.

The simplest multiple ring topology uses a single switch to join two EAPS rings.

The common link feature uses two switches, which share a common link, to provide redundancy and link multiple EAPS rings.

### *Two Rings Connected by One Switch*

The following figure shows how a data *VLAN* can span two rings interconnected by a common switch—a figure eight topology.

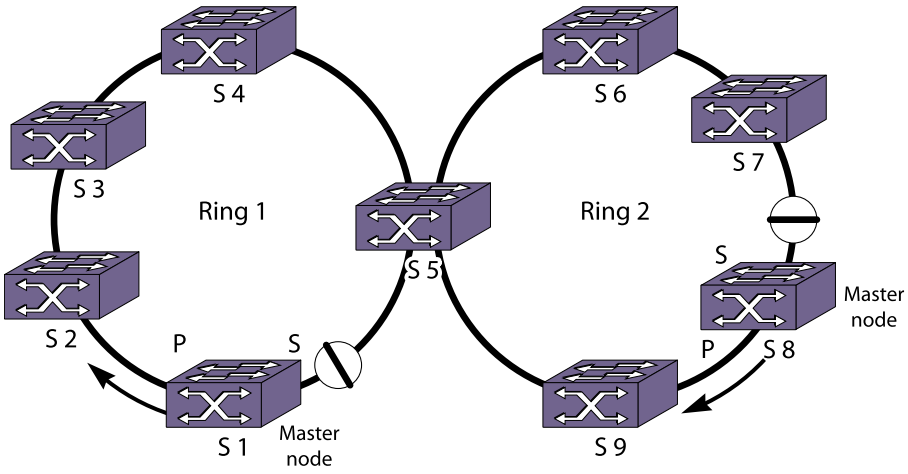



**Figure 117: Two Rings Interconnected by One Switch**

A data VLAN that spans multiple physical rings or EAPS domains and is protected by EAPS is called an overlapping VLAN. An overlapping VLAN requires loop protection for each EAPS domain to which it belongs.

In the following figure, there is an EAPS domain with its own control VLAN running on ring 1 and another EAPS domain with its own control VLAN running on ring 2. A data VLAN that spans both rings is added as a protected VLAN to both EAPS domains to create an overlapping VLAN. Switch S5 has two instances of EAPS domains running on it, one for each ring.

*Multiple Rings Sharing an EAPS Common Link*

**EAPS Common Link Operation**

The following figure shows an example of a multiple ring topology that uses the *EAPS (Extreme Automatic Protection Switching)* common link feature to provide redundancy for the switches that connect the rings.

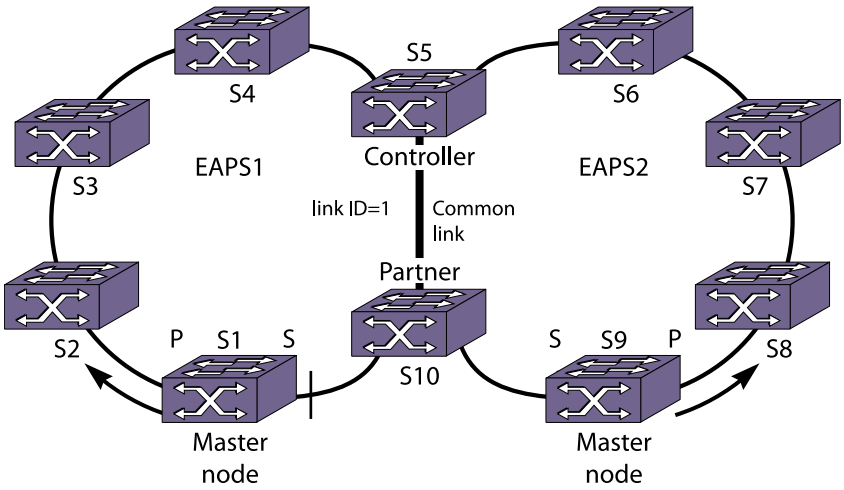



**Figure 118: Multiple Rings Sharing a Common Link**

An EAPS common link is a physical link that carries overlapping *VLAN*s that are protected by more than one EAPS domain.

In the example shown earlier in the preceding figure, switch S5 could be a single point of failure. If switch S5 were to go down, users on Ring 1 would not be able to communicate with users on Ring 2. To make the network more resilient, you can add another switch. A second switch, S10, connects to both rings and to S5 through a common link, which is common to both rings.

The EAPS common link in the following figure requires special configuration to prevent a loop that spans both rings. The software entity that requires configuration is the eaps shared-port, so the common link feature is sometimes called the *shared port* feature.

Note
If the shared port is not configured and the common link goes down, a superloop between the multiple EAPS domains occurs.

The correct EAPS common link configuration requires an EAPS shared port at each end of the common link. The role of the shared port (and switch) at each end of the common link must be configured as either controller or partner. Each common link requires one controller and one partner for each EAPS domain. Typically the controller and partner nodes are distribution or core switches. A controller or partner can also perform the role of master or transit node within its EAPS domain.

During normal operation, the master node on each ring protects the ring as described in EAPS Single Ring Topology on page 967. The controller and partner nodes work together to protect the overlapping VLANs from problems caused by a common link failure or a failed controller (see the following figure).

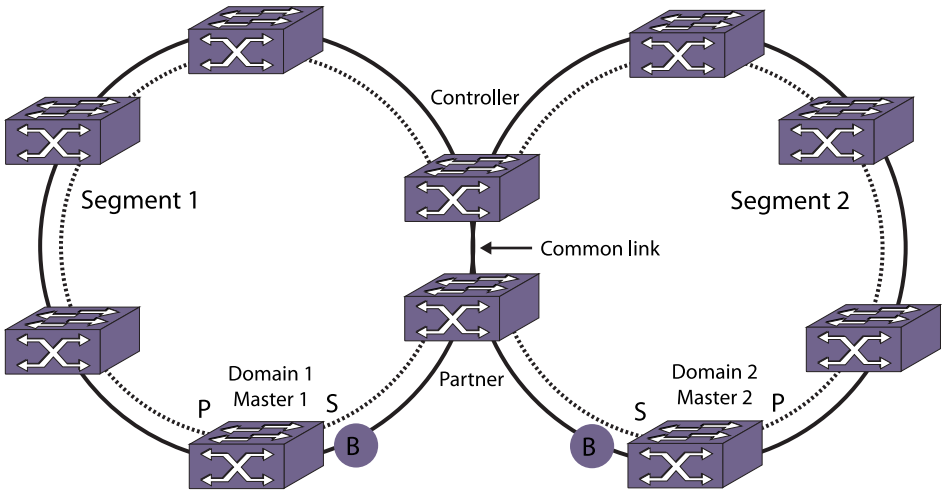



**Figure 119: Master Node Operation in a Multiple Ring Topology**

If a link failure occurs in one of the outer rings, only a single EAPS domain is affected. The EAPS master detects the failure in its domain, and converges around the failure. In this case, the controller does not take any blocking action, and EAPS domains on other rings are not affected. Likewise, when the link is restored, only the local EAPS domain is affected. The controller and any EAPS domains on other rings are not affected, and continue forwarding traffic normally.

To detect common-link faults, the controller and partner nodes send segment health check messages at one-second intervals to each other through each segment. A *segment* is the ring communication path between the controller and partner. The common link completes the ring, but it is a separate entity from the segment. To discover segments and their up or down status, segment health-check messages are sent from controller to partner, and also from partner to controller (see the following figure).

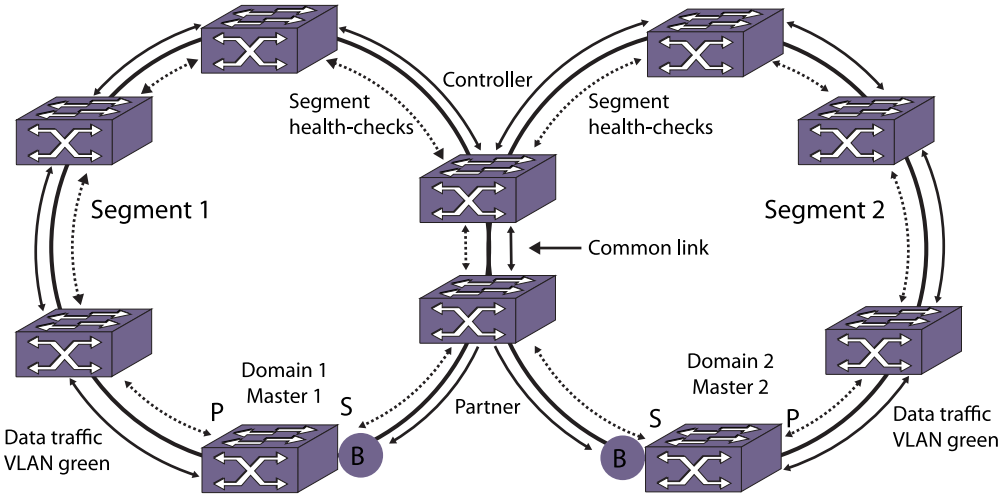



**Figure 120: Segment Health-Check Messages**

#### Common Link Fault Detection and Response

With one exception, when a common link fails, each master node detects the failure and unblocks its secondary port, as shown in the following figure.

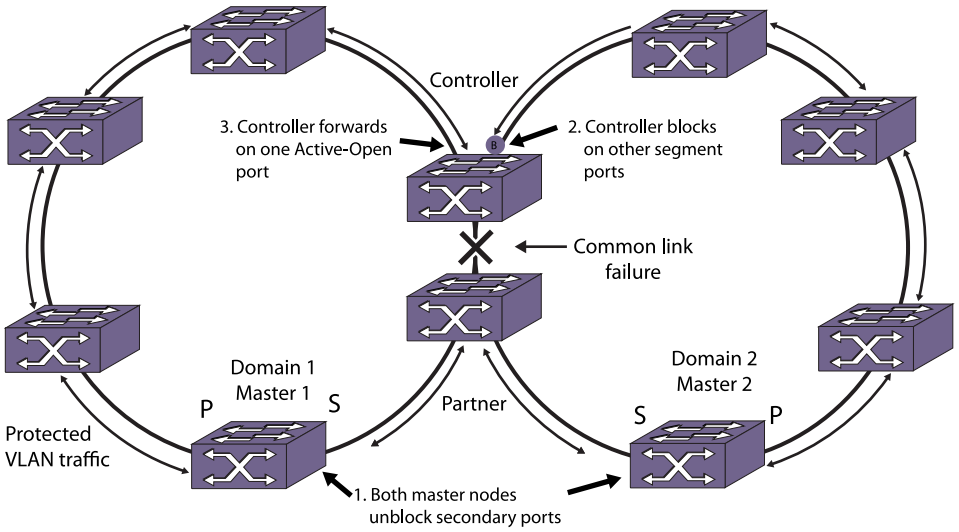



**Figure 121: Common Link Failure**

Because the secondary port of each master node is now unblocked, the new topology introduces a broadcast loop spanning the outer rings.

The controller and partner nodes immediately detect the loop, and the controller does the following:

- Selects an active-open port for protected *VLAN* communications.
- Blocks protected VLAN communications on all segment ports except the active-open port.

Note

When a controller goes into or out of the blocking state, the controller sends a flush-fdb message to flush the *FDB* in each of the switches in its segments. In a network with multiple EAPS ports in the blocking state, the flush-fdb message gets propagated across the boundaries of the EAPS domains.

The exception mentioned above occurs when the partner node is also a master node, and the shared port that fails is configured as a primary port. In this situation, the master node waits for a link-down PDU from the controller node before opening the secondary port. This delay prevents a loop that might otherwise develop if the master/partner node detects the link failure before the controller node.

Note

If the common link and a ring link fail, and if the common link restores before the ring link, traffic down time can be as long as three seconds. This extended delay is required to prevent loops during the recovery of multiple failed links.

#### Common Link Recovery

When a common link recovers, each master node detects that the ring is complete and immediately blocks their secondary ports. The controller also detects the recovery and puts its shared port to the common link into a temporary blocking state called pre-forwarding as shown in the following figure.

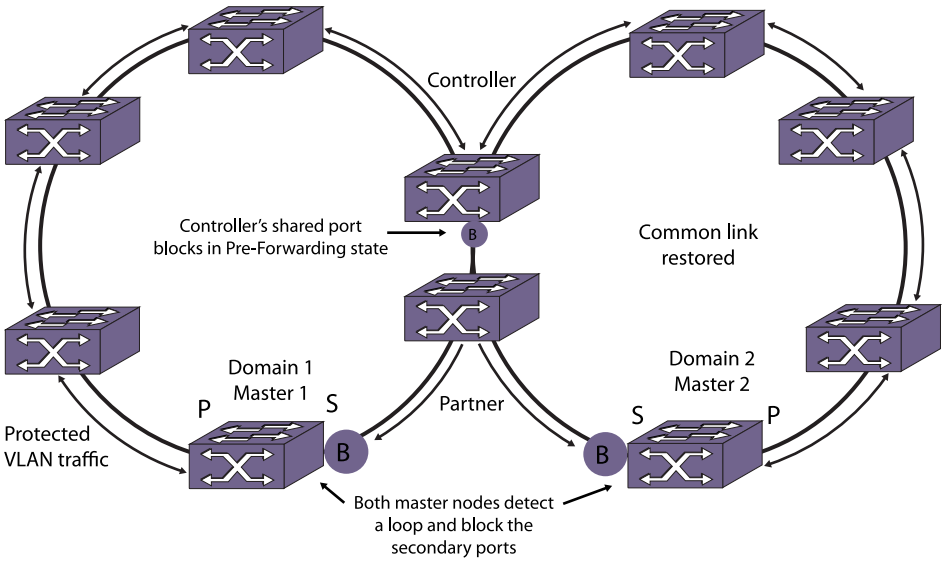



**Figure 122: Common Link in Pre-Forwarding State**

Because the topology has changed, the EAPS nodes must learn the new traffic paths. Each master node notifies all switches in their domain to clear their FDB tables, and traditional Ethernet bridge learning and forwarding mechanisms establish the new traffic paths. Once the controller receives flush-fdb messages for all of its connected EAPS domains, the controller shared-port state for the common link changes to forwarding, the controller state changes to Ready, and traffic flows normally as shown in the following figure.

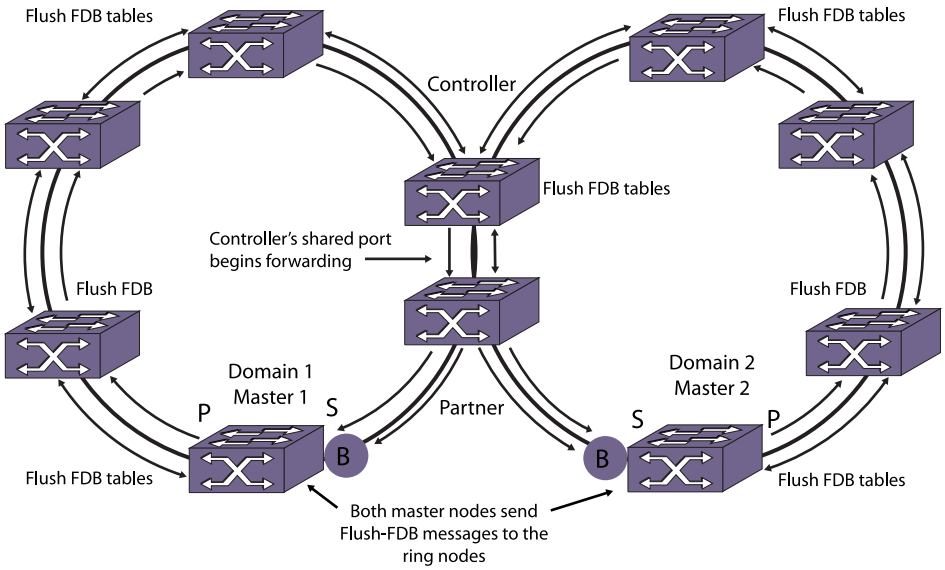



**Figure 123: Common-Link Restored**

#### Controller and Partner Node States

EAPS controller and partner nodes can be in the following states:

- Ready—Indicates that the EAPS domains are running, the common-link neighbor can be reached through segment health-checks, and the common link is up.
- Blocking—Indicates that the EAPS domains are running, the common-link neighbor can be reached through segment health-checks, but the common-link is down. Only the controller node (and not the partner) performs blocking.
- Preforwarding—Indicates the EAPS domain was in a blocking state, and the common link was restored. The controller port is temporarily blocked to prevent a loop during state transition from Blocking to Ready.
- Idle—Indicates the EAPS common-link neighbor cannot be reached through segment health-check messages.

### *Spatial Reuse with an EAPS Common Link*

The common-link topology supports multiple EAPS domains (spatial reuse) on each ring as shown in the following figure.

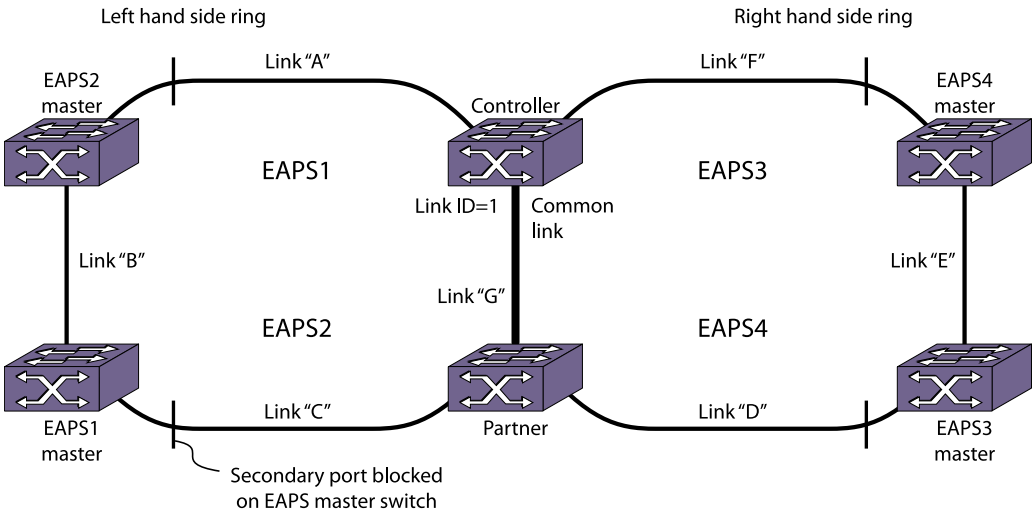



**Figure 124: EAPS Common Link Topology with Spatial Reuse**

Note

If you are using the older method of enabling *standard mode* instead of EAPSv2 to block the super loop in a shared-port environment, you can continue to do so. In all other scenarios, we recommendsthat you do not use both *STP* and EAPS on the same port.

### *Additional Common Link Topology Examples*

#### Basic Core Topology

The following figure shows a core topology with two access rings. In this topology, there are two EAPS common links.

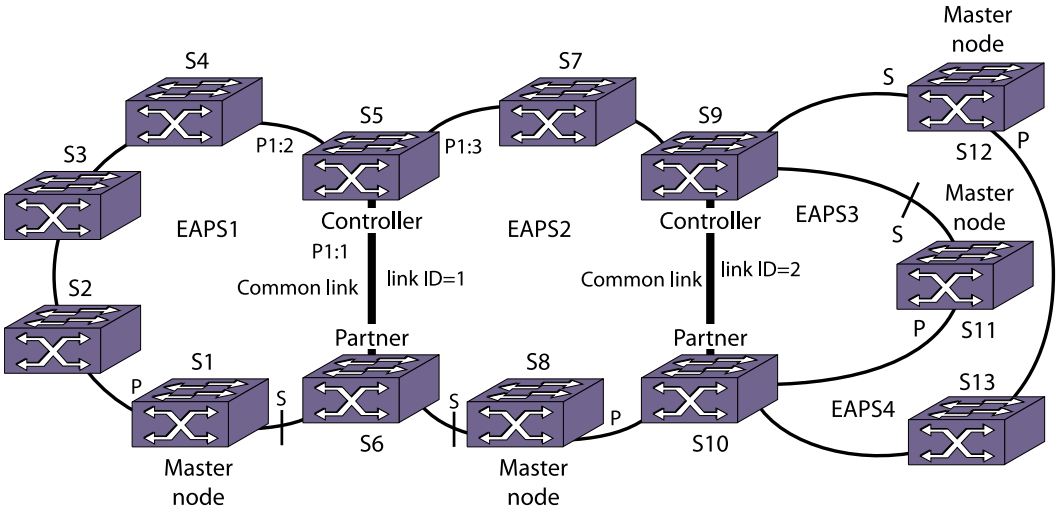



**Figure 125: Basic Core Topology**

### Right-Angle Topology

In the right-angle topology, there are still two EAPS common links, but the common links are adjacent to each other.

To configure a right-angle topology, there must be two common links configured on one of the switches. The following figure shows a right-angle topology.

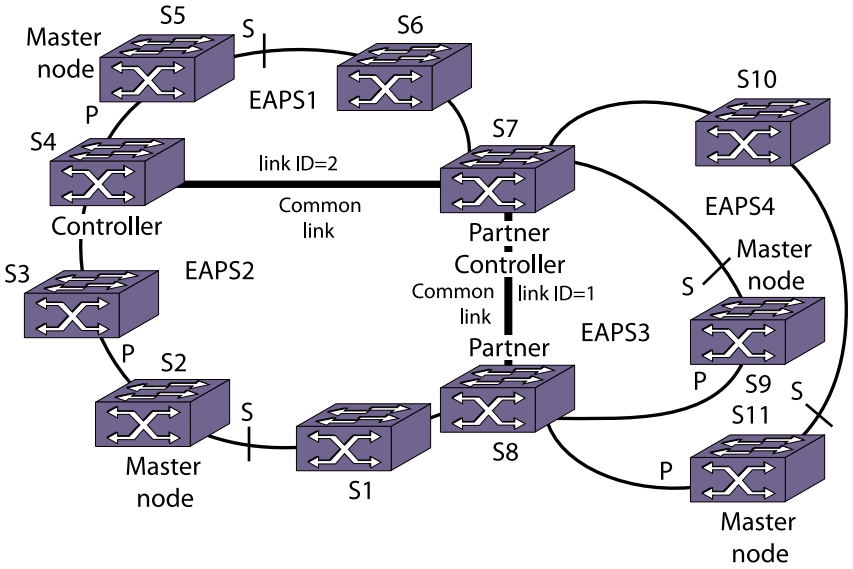



**Figure 126: Right-Angle Topology**

### Combined Basic Core and Right-Angle Topology

The following figure shows a combination basic core and right-angle topology.

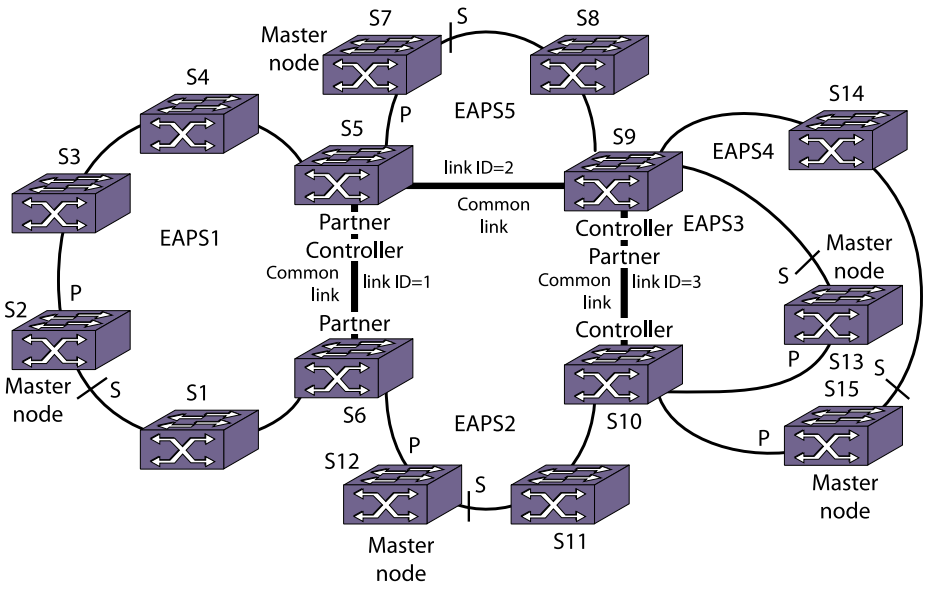



**Figure 127: Basic Core and Right Angle Topology**

The following figure shows an extension of the basic core and right angle configuration.

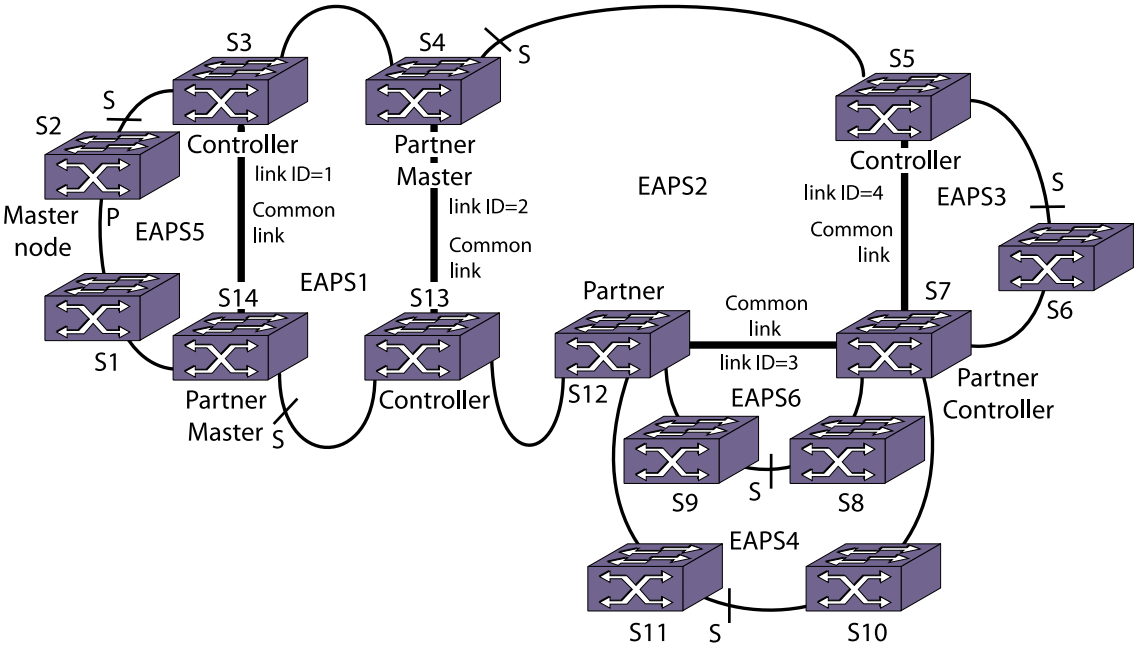



**Figure 128: Advanced Basic Core and Right Angle Topology**

#### Large Core and Access Ring Topology

The following figure shows a single large core ring with multiple access rings hanging off of it.

This is an extension of a basic core configuration.

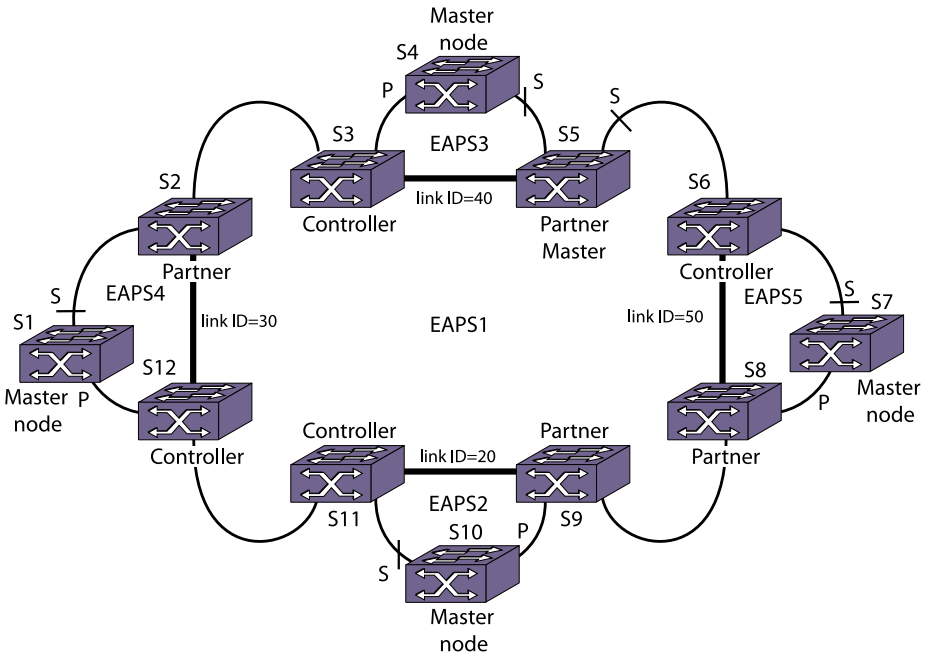



**Figure 129: Large Core and Access Ring Topology**

## Fast Convergence

The fast convergence mode allows EAPS to converge more rapidly. In EAPS fast convergence mode, the link filters on EAPS ring ports are turned off. In this case, an instant notification is sent to the EAPS process if a port's state transitions from up to down or vice-versa.

You must configure fast convergence for the entire switch, not by EAPS domain.

For optimum performance and convergence, it is recommended to use fiber cables.

## EAPS and Hitless Failover--Modular Switches and SummitStack Only

When you install two Management Switch Fabric Modules (MSMs) or Management Modules (MMs) in a BlackDiamond chassis or use redundancy in a SummitStack, one MSM/MM (node) assumes the role of primary and another node assumes the role of backup.

The primary node executes the switch's management functions, and the backup node acts in a standby role. Hitless failover transfers switch management control from the primary to the backup and maintains the state of EAPS. EAPS supports hitless failover. You do not explicitly configure hitless failover support; rather, if you have two MSMs/MMs installed in a chassis or you are operating with redundancy in a SummitStack, hitless failover is available.

Note

Not all platforms support hitless failover in the same software release. To verify if the software version you are running supports hitless failover, see the following table in Managing the Switch on page 39. For more information about protocol, platform, and MSM/MM support for hitless failover, see Understanding Hitless Failover Support on page 59.

To support hitless failover, the primary node replicates all EAPS PDUs to the backup, which allows the backup to be aware of the EAPS domain state. Since both nodes receive EAPS PDUs, each node maintains equivalent EAPS states.

By knowing the state of the EAPS domain, the EAPS process running on the backup node can quickly recover after a primary node failover. Although both nodes receive EAPS PDUs, only the primary transmits EAPS PDUs to neighboring switches and actively participates in EAPS.

Note

For instructions on how to manually initiate hitless failover, see Relinquishing Primary Status on page 55.

## EAPS Licensing

Different EAPS features are offered at different license levels.

For complete information about software licensing, including how to obtain and upgrade your license and what licenses are appropriate for these features, see the *Feature License Requirements* document.

# Configuring EAPS

## Single Ring Configuration Tasks

To configure and enable an EAPS protected ring, do the following on each ring node:

1. Create an EAPS domain and assign a name to the domain as described in Creating and Deleting an EAPS Domain on page 979.
2. Create and add the control *VLAN* to the domain as described in Adding the EAPS Control VLAN on page 979.
3. Create and add the protected VLAN(s) to the domain as described in Adding Protected VLANs on page 979.
4. Configure the EAPS mode (master or transit) for the switch in the domain as described in Defining the Switch Mode (Master or Transit) on page 980.
5. Configure the EAPS ring ports, including the master primary and secondary ring ports, as described in Configuring the Ring Ports on page 980.
6. If desired, configure the polling timers and timeout action as described in Configuring the Polling Timers and Timeout Action on page 981.*
7. Enable EAPS for the entire switch as described in Enabling and Disabling EAPS on the Switch on page 982.

8. If desired, enable Fast Convergence as described in Enabling and Disabling Fast Convergence on page 983.*
9. Enable EAPS for the specified domain as described in Enabling and Disabling an EAPS Domain on page 983.

Note

If you configure a VMAN on a switch running EAPS, make sure you configure the VMAN attributes on all of the switches that participate in the EAPS domain. For more information about VMANs, see VMAN (PBN) and PBBN.

### *Creating and Deleting an EAPS Domain*

Each EAPS domain is identified by a unique domain name.

- To create an EAPS domain, use the following command:

  `create eaps` *`name`*

- To delete an EAPS domain, use the following command:

  `delete eaps` *`name`*

### *Adding the EAPS Control VLAN*

You must create and configure one control *VLAN* for each EAPS domain. For instructions on creating a VLAN, see VLANs on page 502.

- To configure EAPS to use a VLAN as the EAPS control VLAN for a domain, use the following command:

  `configure eaps` *`name`* **`add control`** {**`vlan`**} *`vlan_name`*

Note

A control VLAN cannot belong to more than one EAPS domain. If the domain is active, you cannot delete the domain or modify the configuration of the control VLAN.

The control VLAN must NOT be configured with an IP address. In addition, only ring ports may be added to this control VLAN. No other ports can be members of this VLAN. Failure to observe these restrictions can result in a loop in the network.

The ring ports of the control VLAN must be tagged.

By default, EAPS PDUs are automatically assigned to *QoS (Quality of Service)* profile QP8. This ensures that the control VLAN messages reach their intended destinations. You do not need to configure a QoS profile for the control VLAN.

### *Adding Protected VLANs*

You must add one or more protected *VLAN*s to each EAPS domain. The protected VLANs are the data-carrying VLANs.

Note

When you configure a protected VLAN, the ring ports of the protected VLAN must be tagged (except in the case of the default VLAN).

For instructions on creating a VLAN, see VLANs on page 502.

- To configure a VLAN as an EAPS protected VLAN, use the following command:

  `configure eaps` *name* **`add protected`** {**`vlan`**} *vlan_name*

### *Configuring the EAPS Domain Priority*

The EAPS domain priority feature allows you to select the EAPS domains that are serviced first when a break occurs in an EAPS ring. For example, you might set up a network topology with two or more domains on the same physical ring, such as in spatial reuse. In this topology, you could configure one domain as high priority and the others as normal priority. You would then add a small subset of the total protected VLANs to the high priority domain, and add the rest of the protected vlans to the normal priority domain. The secondary port of the normal and high priority domains can be the same, or as is typically the case of spatial reuse, opposite. If a ring fault occurs in this topology, the protected VLANs in the high priority domain are the first to recover.

- To configure the EAPS domain priority, use the following command:

  `configure eaps` *name* **`priority`** {**`high`** | **`normal`**}

### *Defining the Switch Mode (Master or Transit)*

We recommend keeping the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary.

1. Configure the EAPS switch mode for a domain using the following command:

   `configure eaps` *name* **`mode`** [**`master`** | **`transit`**]

   One switch on the ring must be configured as the master node for the specified domain; all other switches on the same ring and domain are configured as transit nodes.

   If you configure a switch to be a transit node for an EAPS domain, the default switch configuration displays the following message and prompts you to confirm the command:

   ```
   WARNING: Make sure this specific EAPS domain has a Master node in the ring.  If you
   change this node from EAPS master to EAPS transit, you could cause a loop in the
   network. Are you sure you want to change mode to transit? (y/n)
   ```

2. When prompted, do one of the following:
   - Enter `y` to identify the switch as a transit node.
   - Enter `n` or press **[Return]** to cancel the command.

For more information see, Disabling EAPS Loop Protection Warning Messages on page 984.

### *Configuring the Ring Ports*

Each node on the ring connects to the ring through two ring ports. The ports that you choose on each switch should be tagged and added to the control *VLAN* and all protected VLANs. For information on adding tagged ports to a VLAN, see VLANs on page 502.

On the master node, one ring port must be configured as the primary port, and the other must be configured as the secondary port.

We recommend that you keep the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary.

1. To configure a node port as primary or secondary, use the following command:

   ```
   configure eaps name [primary | secondary] port ports
   ```

   If you attempt to add an EAPS ring port to a VLAN that is not protected by EAPS, the default switch configuration prompts you to confirm the command with the following message:

   ```
   Make sure <vlan_name> is protected by EAPS. Adding EAPS ring ports to a VLAN could
   cause a loop in the network. Do you really want to add these ports (y/n)
   ```

2. When prompted, do one of the following:
   - Enter `y` to identify the switch as a transit node.
   - Enter `n` or press **[Return]** to cancel the command.

For information on configuring a VLAN for EAPS, see the following sections:

- Adding the EAPS Control VLAN on page 979
- Adding Protected VLANs on page 979

For more information see, Disabling EAPS Loop Protection Warning Messages on page 984.

### *Configuring the Polling Timers and Timeout Action*

The polling timers provide an alternate way to detect ring breaks. In a ring that uses only Extreme Networks switches, the master switch learns about a ring break by receiving a link-down PDU. When the ring uses only Extreme networks switches, the polling timers are not needed and can remain configured for the default values.

In a ring that contains switches made by other companies, the polling timers provide an alternate way to detect ring breaks. The master periodically sends hello PDUs at intervals determined by the hello PDU timer and waits for a reply. If a hello PDU reply is not received before the failtime timer expires, the switch detects a failure and responds by either sending an alert or opening the secondary port. The response action is defined by a configuration command.

- Set the polling timer values the master node uses for detecting ring failures.

  ```
  configure eaps name hellotime seconds milliseconds
  configure eaps name failtime seconds milliseconds
  ```

  Note

  These commands apply only to the master node. If you configure the polling timers for a transit node, they are ignored. If you later reconfigure that transit node as the master node, the polling timer values are used as the current values.

  Use the **hellotime** keyword and its associated parameters to specify the amount of time the master node waits between transmissions of health check messages on the control *VLAN*. The combined value for seconds and milliseconds must be greater than 0. The default value is 1 second.

Use the **failtime** keyword and its associated parameters to specify the amount of time the master node waits before the failtimer expires. The combined value for seconds and milliseconds must be greater than the configured value for **hellotime**. The default value is 3 seconds.

Note

Increasing the failtime value increases the time it takes to detect a ring break using the polling timers, but it can also reduce the possibility of incorrectly declaring a failure when the network is congested.

- Configure the action taken when a ring break is detected.

  `configure eaps` *`name`* **`failtime expiry-action`** [**`open-secondary-port`** | **`send-alert`**]

  Use the *`send-alert`* parameter to send an alert when the failtimer expires. Instead of going into a failed state, the master node remains in a Complete or Init state, maintains the secondary port blocking, and writes a critical error message to syslog warning the user that there is a fault in the ring. An *SNMP (Simple Network Management Protocol)* trap is also sent.

  Use the *`open-secondary-port`* parameter to open the secondary port when the failtimer expires.

### *Enabling and Disabling EAPS on the Switch*

We recommend that you keep the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary.

- To enable the EAPS function for the entire switch, use the following command:

  ```
  enable eaps
  ```

- To disable the EAPS function for the entire switch, use the following command:

  ```
  disable eaps
  ```

  If you enter the command to disable EAPS, the default switch configuration displays the following warning message and prompts you to confirm the command:

  ```
  WARNING: Disabling EAPS on the switch could cause a loop in the network! Are you sure you want to disable EAPS? (y/n)
  ```

- When prompted, do one of the following:

  a. Enter `y` to disable EAPS for the entire switch.

  b. Enter `n` or press **[Return]** to cancel the command.

For more information see, Disabling EAPS Loop Protection Warning Messages on page 984.

### *Enabling and Disabling Fast Convergence*

You can enable or disable fast convergence for the entire switch to improve EAPS convergence times.

Note

Possible factors affecting EAPS fast convergence time:

- The medium type of the link being flapped (Fiber link-down events are detected faster than copper, causing better convergence).
- Number of VLANs protected by the EAPS domain (convergence time increases with the number of protected VLANs).
- Number of *FDB* entries present during the switch over (convergence time increases with the number of FDBs learned).
- Topology change event (link down or link up) causes the master node to send an FDB flush to all transits. In the event ofa shared port failure, FDB is flushed twice, causing an increase in convergence time.
- Number of hops between the switch where the link flap happens and the master node (convergence increases with the number of hops).

- To enable or disable fast convergence on the switch, use the following command:

  `configure eaps` **`fast-convergence`**`[`**`off`** `|` **`on`**`]`

### *Enabling and Disabling an EAPS Domain*

We recommend that you keep the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary.

- To enable a specific EAPS domain, use the following command:

  `enable eaps {`*`name`*`}`

- To disable a specific EAPS domain, use the following command:

  `disable eaps {`*`name`*`}`

  If you enter the `disable eaps` command, the default switch configuration displays the following warning message and prompts you to confirm the command:

  ```
  WARNING: Disabling specific EAPS domain could cause a loop in the network! Are you sure
  you want to disable this specific EAPS domain? (y/n)
  ```

- When prompted, do one of the following:

  a. Enter `y` to disable EAPS for the specified domain.

  b. Enter `n` or press **[Return]** to cancel the command.

For more information see, Disabling EAPS Loop Protection Warning Messages on page 984.

### *Configuring EAPS Support for Multicast Traffic*

The ExtremeXOS software provides several commands for configuring how EAPS supports multicast traffic after an EAPS topology change.

**Note**

EAPS multicast flooding must be enabled before the add-ring-ports feature will operate. For information on enabling EAPS multicast flooding, see the command:

```
configure eaps multicast temporary-flooding [on | off]
```

### *Unconfiguring an EAPS Ring Port*

Unconfiguring an EAPS port sets its internal configuration state to INVALID, which causes the port to appear in the Idle state with a port status of Unknown. This occurs when you use the `show eaps {eapsDomain} {detail}` command to display the status information about the port.

We recommend that you keep the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary.

1. To unconfigure an EAPS primary or secondary ring port for an EAPS domain, use the following command:

   ```
   unconfigure eaps eapsDomain [primary | secondary] port
   ```

   To prevent loops in the network, the switch displays by default a warning message and prompts you to unconfigure the specified EAPS primary or secondary ring port.
2. When prompted, do one of the following:
   a. Enter `y` to unconfigure the specified port.
   b. Enter `n` or press **[Return]** to cancel this action.

   The following command example unconfigures this node's EAPS primary ring port on the domain "eaps_1":

   ```
   unconfigure eaps eaps_1 primary port
   WARNING: Unconfiguring the Primary port from the EAPS domain could cause a loop in
   The network! Are you sure you want to unconfigure the Primary EAPS Port? (y/n)
   ```
3. Enter `y` to continue and unconfigure the EAPS primary ring port. Enter `n` to cancel this action.

   The switch displays a similar warning message if you unconfigure the secondary EAPS port.

For more information see, Disabling EAPS Loop Protection Warning Messages on page 984.

### *Disabling EAPS Loop Protection Warning Messages*

The switch displays by default loop protection messages when configuring the following EAPS parameters:

- Adding EAPS primary or secondary ring ports to a *VLAN*
- Deleting a protected VLAN
- Disabling the global EAPS setting on the switch
- Disabling an EAPS domain
- Configuring an EAPS domain as a transit node
- Unconfiguring EAPS primary or secondary ring ports from an EAPS domain

We recommend keeping the loop protection warning messages enabled. If you have considerable knowledge and experience with EAPS, you might find the EAPS loop protection warning messages unnecessary. For example, if you use a script to configure your EAPS settings, disabling the warning messages allows you to configure EAPS without replying to each interactive yes/no question.

- To disable loop protection messages, use the following command:

  ```
  configure eaps config-warnings off
  ```

- To re-enable loop protection messages, use the following command:

  ```
  configure eaps config-warnings on
  ```

## Common Link Topology Configuration Tasks

To create a common link topology, you must configure the shared ports at each end of the common link.

### *EAPS Shared Port Configuration Rules*

The following rules apply to EAPS shared port configurations:

- Each common link in the EAPS network must have a unique link ID, which is configured at the shared port at each end of the link.
- The shared port mode configured on each side of a common link must be different from the other; one must be a controller and one must be a partner.
- The controller and partner shared ports on either side of a common link must have the same
- link ID. The common link is established only when the shared ports at each end of the common link have the same link ID.
- There can be up to two shared ports per switch.
- There cannot be more than one controller on a switch.

  Valid combinations on any one switch are:

  - 1 controller
  - 1 partner
  - 1 controller and 1 partner
  - 2 partners

- A shared port cannot be configured on an EAPS master's secondary port.

**Note**

When a common link fails, one of the segment ports becomes the active-open port, and all other segment ports are blocked to prevent a loop for the protected VLANs. For some topologies, you can improve network performance during a common link failure by selecting the port numbers to which segments connect. For information on how the active-open port is selected, see Common Link Fault Detection and Response.

### *Common Link Configuration Overview*

To configure and enable a common link to serve multiple rings, do the following on the controller and partner nodes:

1. Create a shared port for the common link as described in Creating and Deleting a Shared Port on page 986.

2. Configure the shared port as either a controller or a partner as described in Defining the Mode of the Shared Port on page 986.
3. Configure the link ID on the shared port as described in Configuring the Link ID of the Shared Port on page 986.
4. If desired, configure the polling timers and timeout action as described in Configuring the Shared Port Timers and Timeout Action on page 987.

   This step can be configured at any time, even after the EAPS domains are running.
5. Configure EAPS on each ring as described in Single Ring Configuration Tasks on page 978.

### *Creating and Deleting a Shared Port*

To configure a common link, you must create a shared port on each switch belonging to the common link.

- To create a shared port, use the following command:

  `create eaps shared-port` *`ports`*

  Where *`ports`* is the common link port.

  Note
  A switch can have a maximum of two shared ports.

- To delete a shared port on the switch, use the following command:

  `delete eaps shared-port` *`ports`*

### *Defining the Mode of the Shared Port*

The shared port on one end of the common link must be configured to be the controller. This is the end responsible for blocking ports when the common link fails, thereby preventing the superloop.

The shared port on the other end of the common link must be configured to be the partner. This end does not participate in any form of blocking. It is responsible for only sending and receiving health-check messages.

- To configure the mode of the shared port, use the following command:

  `configure eaps shared-port` *`ports`* **`mode`** *`controller | partner`*

### *Configuring the Link ID of the Shared Port*

Each common link in the EAPS network must have a unique link ID. The controller and partner shared ports that belong to the same common link must have matching link IDs. No other instance in the network should have that link ID.

If you have multiple adjacent common links, we recommend that you configure the link IDs in ascending order of adjacency. For example, if you have an EAPS configuration with three adjacent common links, moving from left to right of the topology, configure the link IDs from the lowest to the highest value.

- To configure the link ID of the shared port, use the following command:

  `configure eaps shared-port` *`ports`* **`link-id`** *`id`*

  The link ID range is 1–65534.

*Configuring the Shared Port Timers and Timeout Action*

- To configure the shared port timers, use the following commands:

  `configure eaps shared-port` *`port`* **`common-path-timers`** `{[`**`health-interval`** `|` **`timeout`**`]` *`seconds`*`}`

  `configure eaps shared-port` *`port`* **`segment-timers health-interval`** *`seconds`*

  `configure eaps shared-port` *`port`* **`segment-timers timeout`** *`seconds`*
- To configure the time out action for segment timers, use the following command:

  `configure eaps shared-port` *`port`* **`segment-timers expiry-action`** `[`**`segment-down`** `|` **`send-alert`**`]`

*Unconfiguring an EAPS Shared Port*

- To unconfigure a link ID on a shared port, use the following command:

  `unconfigure eaps shared-port` *`ports`* **`link-id`**
- To unconfigure the mode on a shared port, use the following command:

  `unconfigure eaps shared-port` *`ports`* **`mode`**
- To delete a shared port, use the following command:

  `delete eaps shared-port` *`ports`*

## Clearing the EAPS Counters

The EAPS counters continue to increment until you explicitly clear the information. By clearing the counters, you can see fresh statistics for the time period you are monitoring.

- To clear the counters used by EAPS, use the following commands:

  `clear counters`

  `clear eaps counters`

# Displaying EAPS Information

## Displaying Single Ring Status and Configuration Information

- To display EAPS status and configuration information, use the following command:

  `show eaps {`*`eapsDomain`*`} {`**`detail`**`}`

  Note

  You might see a slightly different display, depending on whether you enter the command on the master node or the transit node.

  If you specify a domain with the optional *`eapsDomain`* parameter, the command displays status information for a specific EAPS domain.

  The display from the `show eaps detail` command shows all the information shown in the `show eaps` *`eapsDomain`* command for all configured EAPS domains.

## Displaying Domain Counter Information

- To display EAPS counter information for one or all domains, use the following command:

  `show eaps counters [`*`eapsDomain`* `|` **`global`**`]`

  If you specify the name of an EAPS domain, the switch displays counter information related to only that domain.

  If you specify the **global** keyword, the switch displays a list of the counter totals for all domains. To see the counters for a specific domain, you must specify the domain name.

  Note
  If a PDU is received, processed, and consumed, only the Rx counter increments. If a PDU is forwarded in slow path, both the Rx counter and Fw counter increment.

## Displaying Common Link Status and Configuration Information

Each controller and partner node can display status and configuration information for the shared port or ports on the corresponding side of the common link.

- To display EAPS common link information, use the following command:

  `show eaps shared-port {`*`port`*`} {`**`detail`**`}`

  If you enter the `show eaps shared-port` command without an argument or keyword, the command displays a summary of status information for all configured EAPS shared ports on the switch.

  If you specify a shared port, the command displays information about that specific port.

  You can use the **detail** keyword to display more detailed status information about the segments and VLANs associated with each shared port.

## Displaying Common Link Counter Information

Each controller and partner node can display counter information for the shared port or ports through which the switch connects to a common link.

- To display EAPS shared port counter information, use the following command:

  `show eaps counters shared-port [`**`global`** `|` *`port`* `{`**`segment-port`** *`segport`* `{`*`eapsDomain`*`}}]`

  If you specify the **global** keyword, the switch displays a list of counters that show the totals for all shared ports together. To view the counters for a single shared port, enter the command with the port number.

  If you specify a particular EAPS segment port, the switch displays counter information related to only that segment port for the specified EAPS domain.

# Configuration Examples

## Migrating from STP to EAPS

This section explains how to migrate or reconfigure an existing *STP* network to an EAPS network.

**Note**
Actual implementation steps on a production network may differ based on the physical topology, switch models, and software versions deployed.

The sample STP network is a simple two-switch topology connected with two Gigabit Ethernet trunk links, which form a broadcast loop. Both Extreme Networks switches are configured for 802.1D mode STP running on a single data *VLAN* named Data. The sample STP network for migration to EAPS is shown in the following figure.

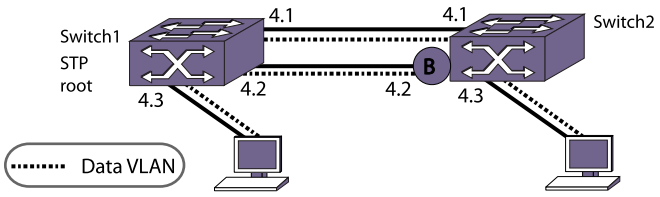



**Figure 130: Sample STP Network for Migration to EAPS**

### *Creating and Configuring the EAPS Domain*

- The first step in the migration process is to create an EAPS Domain and configure the EAPS mode, then define the primary and secondary ports for the domain. Follow this step for both switches. Switch2 is configured as EAPS Master to ensure the same port blocking state is maintained as in the original *STP* topology.

  Switch 1 EAPS domain configuration:

```
* SWITCH#1.1 # create eaps new-eaps
* SWITCH#1.2 # configure new-eaps mode transit
* SWITCH#1.3 # configure new-eaps primary port 4:1
* SWITCH#1.4 # configure new-eaps secondary port 4:2
```

  Switch 2 EAPS domain configuration:

```
* SWITCH#2.1 # create eaps new-eaps
* SWITCH#2.2 # configure new-eaps mode master
* SWITCH#2.3 # configure new-eaps primary port 4:1
* SWITCH#2.4 # configure new-eaps secondary port 4:2
```

### *Creating and Configuring the EAPS Control VLAN*

1. You must create the EAPS control *VLAN* and configure the 802.1q tag and ring ports.
2. Configure the control VLANs as part of the EAPS domain. Do this for both switches.

   Switch 1 control VLAN configuration:

```
* SWITCH#1.5 # create vlan control-1
* SWITCH#1.6 # configure vlan control-1 tag 4001
* SWITCH#1.8 # configure vlan control-1 add port 4:1,4:2 tagged
* SWITCH#1.9 # configure eaps new-eaps add control vlan control-1
```

Switch 2 control VLAN configuration:

```
* SWITCH#2.5 # create vlan control-1
* SWITCH#2.6 # configure vlan control-1 tag 4001
* SWITCH#2.8 # configure vlan control-1 add port 4:1,4:2 tagged
* SWITCH#2.9 # configure eaps new-eaps add control vlan control-1
```

### *Enabling EAPS and Verify EAPS Status*

1. Enable the EAPS protocol and the EAPS domain.
2. Confirm that the master node is in Complete state and its secondary port is blocking.

   Switch 1 commands to enable EAPS and the domain:

```
* SWITCH#1.10 # enable eaps
* SWITCH#1.11 # enable eaps new-eaps
```

   Switch 2 commands to enable EAPS and verify status:

```
* SWITCH#2.10 # enable eaps
* SWITCH#2.11 # enable eaps new-eaps
* SWITCH#2.12 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: Off
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain         State           Mo  En  Pri   Sec   Control-Vlan VID   Count
--------------------------------------------------------------------------------
new-eaps       Complete        M   Y   4:1   4:2   control-1    (4001) 0
--------------------------------------------------------------------------------
```

### *Configuring the STP Protected VLAN as an EAPS Protected VLAN*

Configure the data *VLAN* (currently protected by *standard mode* as an untagged VLAN) as an EAPS protected VLAN.

1. Assign an 802.1q tag to the data VLAN, as this might not be required with the previous *STP* configuration.
2. Next, the data VLAN is added to the EAPS domain as a protected VLAN.
3. Configure the VLAN port changes at the end to prevent any broadcast loop from forming during the transition from STP to EAPS protection.

   A warning message is displayed on the CLI, but this can be ignored, as it is just a reminder that the ring ports have not been added to the protected VLAN yet.
4. Change the port membership for the data VLAN from untagged to 802.1q tagged trunk ports.

   Switch#2 commands to add EAPS protected VLAN and tagged ports:

```
* SWITCH#2.13 # configure vlan data tag 1000
* SWITCH#2.14 # configure new-eaps add protect vlan data
WARNING: Primary port [4:1] is not tagged on vlan "data", EAPS="new-eaps"
WARNING: Secondary port [4:2] is not tagged on vlan "data", EAPS="new-eaps"
* SWITCH#2.15 # configure data add port 4:1,4:2 tagged
```

Switch#1 commands to add EAPS protected VLAN and tagged ports:

```
* SWITCH#1.13 # configure vlan data tag 1000
* SWITCH#1.14 # configure new-eaps add protect vlan data
WARNING: Primary port [4:1] is not tagged on vlan "data", EAPS="new-eaps"
WARNING: Secondary port [4:2] is not tagged on vlan "data", EAPS="new-eaps"
* SWITCH#1.15 # configure data add port 4:1,4:2 tagged
```

### *Verifying the EAPS Blocking State for the Protected VLAN*

- To ensure there is no potential for a broadcast storm, confirm that EAPS is successfully blocking the protected *VLAN*, as shown in the following example:

```
* SWITCH#2.16 # show new-eaps
Name: new-eaps
State: Complete                                Running: Yes
Enabled: Yes Mode: Master
Primary port:   4:1           Port status: Up         Tag status: Tagged
Secondary port: 4:2           Port status: Blocked    Tag status: Tagged
Hello timer interval: 1  sec  0  millisec
Fail timer interval:  3  sec
Fail Timer expiry action: Send alert
Last valid EAPS update: From Master Id 00:04:96:10:51:50, at Fri Sep 10 13:38:39 2004
EAPS Domain's Controller Vlan: control-1 4001
EAPS Domain's Protected Vlan(s): data 1000
Number of Protected Vlans: 1
```

After you verify that EAPS is protecting the data VLAN, you can safely remove the *STP* configuration.

### *Verifying the STP Status and Disabling STP*

Once you have successfully verified that EAPS has taken over loop prevention for the data *VLAN*, you no longer need the *STP* configuration.

Now, verify whether the data VLAN is removed from the STP domain, and then disable the STP protocol.

Switch 2 commands to verify STP status and disable STP:

```
* SWITCH#2.17 # show stp s0
Stpd: s0                Stp: ENABLED            Number of Ports: 0
Rapid Root Failover: Disabled
Operational Mode: 802.1D                        Default Binding Mode: 802.1D
802.1Q Tag: (none)
Ports: (none)
Participating Vlans: (none)
Auto-bind Vlans: Default
Bridge Priority: 32768
BridgeID:               80:00:00:04:96:10:51:50
Designated root:        80:00:00:04:96:10:51:50
RootPathCost: 0         Root Port: ----
MaxAge: 20s             HelloTime: 2s           ForwardDelay: 15s
CfgBrMaxAge: 20s        CfgBrHelloTime: 2s      CfgBrForwardDelay: 15s
Topology Change Time: 35s                       Hold time: 1s
Topology Change Detected: FALSE                 Topology Change: FALSE
Number of Topology Changes: 4
Time Since Last Topology Change: 1435s
* SWITCH#2.18 # show s0 port
Port Mode State Cost Flags Priority Port ID Designated Bridge
* SWITCH#2.19 # disable stp
```

Switch 1 commands to verify STP status and disable STP:

```
* SWITCH#1.16 # show stp s0
Stpd: s0                Stp: ENABLED            Number of Ports: 0
Rapid Root Failover: Disabled
Operational Mode: 802.1D                        Default Binding Mode: 802.1D
802.1Q Tag: (none)
Ports: (none)
Participating Vlans: (none)
Auto-bind Vlans: Default
Bridge Priority: 1
BridgeID:               00:01:00:04:96:10:30:10
Designated root:        00:01:00:04:96:10:30:10
RootPathCost: 0         Root Port: ----
MaxAge: 20s             HelloTime: 2s           ForwardDelay: 15s
CfgBrMaxAge: 20s        CfgBrHelloTime: 2s      CfgBrForwardDelay: 15s
Topology Change Time: 35s                       Hold time: 1s
Topology Change Detected: FALSE                 Topology Change: FALSE
Number of Topology Changes: 2
Time Since Last Topology Change: 11267s
* SWITCH#1.17 # show stp s0 po
Port Mode State Cost Flags Priority Port ID Designated Bridge
* SWITCH#1.18 # disable stp s0
* SWITCH#1.19 # disable stp
```

The network should now be successfully migrated from STP to EAPS.

## Designing and Implementing a Highly Resilient Enterprise Network Using EAPS

Network managers can design and employ a highly resilient end-to-end enterprise network using the Extreme Networks switching platform and the EAPS protocol as shown in the following figure.

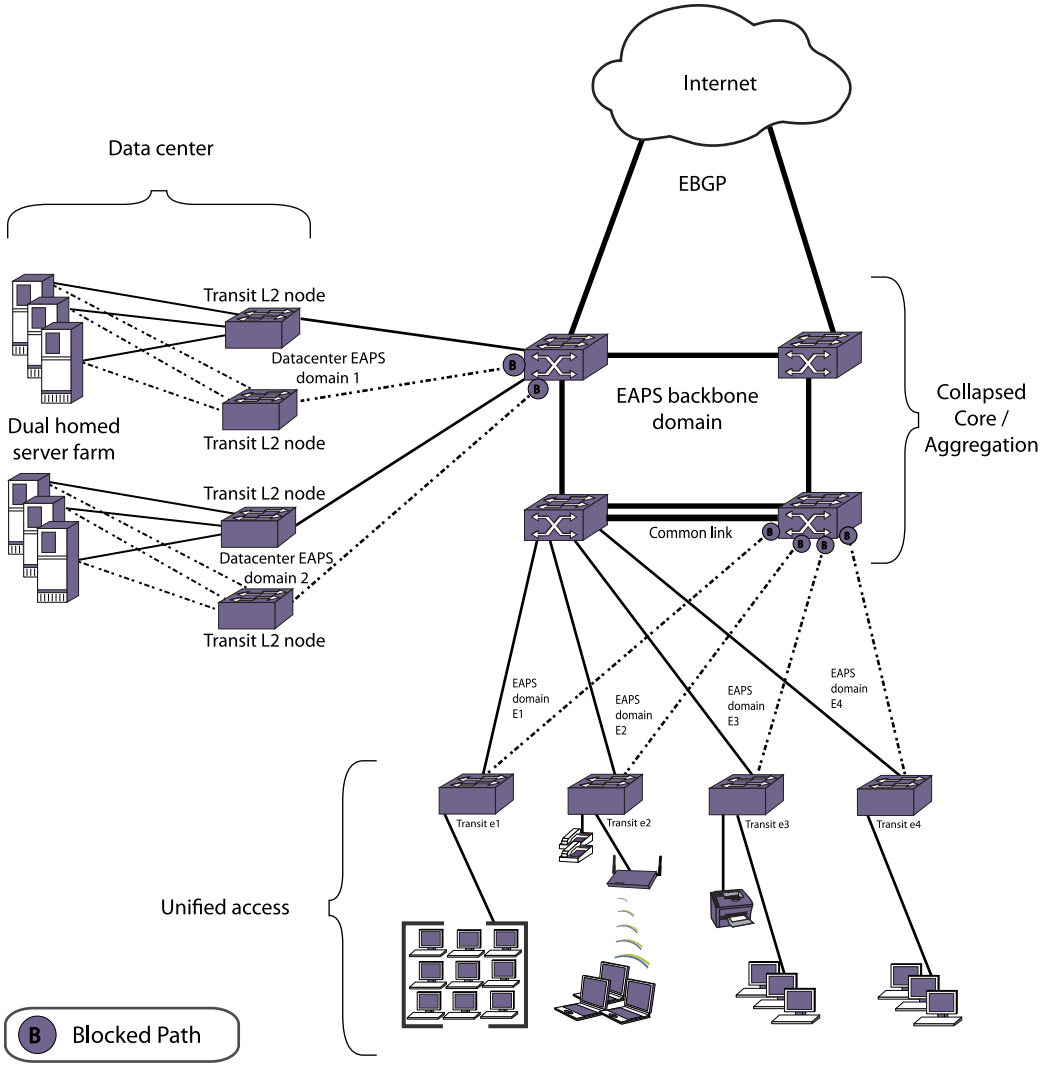



**Figure 131: Extreme Networks EAPS Everywhere**

EAPS can be used in the network edge to provide link resiliency for Ethernet and IP services in a partial-meshed design. In the aggregation layer, EAPS interconnects multiple edge and core domains. When combined with *VRRP (Virtual Router Redundancy Protocol)* and *OSPF* in the aggregation layer, EAPS provides the foundation for highly resilient IP routing by protecting against link and switch failures.

In the network core, EAPS is used with OSPF to provide a high-performance IP routing backbone with zero downtime or route flaps. Using EAPS and dual-homed server farms in the data center provides high availability for mission-critical server resources.

The collapsed core/aggregation layer and data center also make use of EAPS resilient ring topology to ensure network availability to all critical sources.

### *Designing and Configuring the Unified Access Layer*

The unified access network layer makes use of EAPS in a partial-meshed ring topology for maximum resiliency. The edge of the network is the first point of entry for client devices such as PCs, servers, VoIP phones, wireless devices, and printers.

Utilizing EAPS and redundant uplink ports on edge switches increases network resiliency and availability. Edge switches connect their primary and secondary uplink trunk ports to one or more switches in the aggregation network layer (as shown in the following figure). If the primary uplink port fails, traffic can use the alternate secondary uplink.

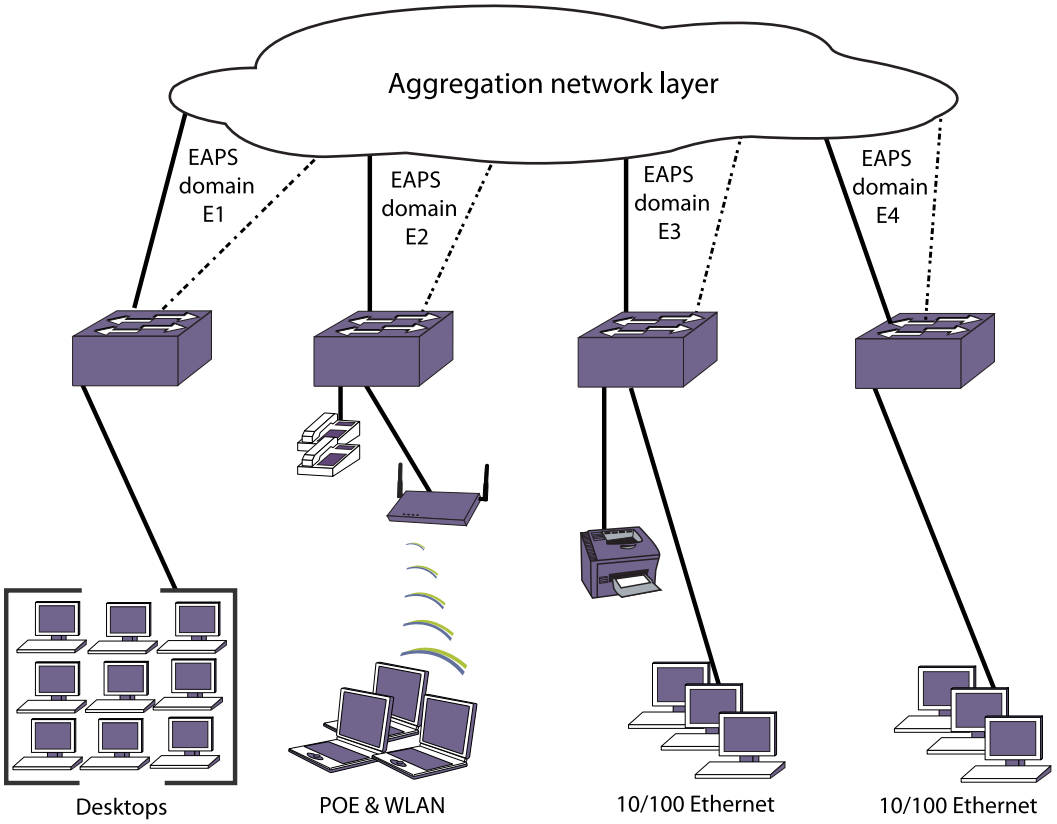



**Figure 132: Converged Network Edge (Unified Access Layer)**

In this sample network, each edge switch is configured with a unique EAPS domain and control *VLAN*. Protected VLANs can overlap across multiple EAPS domains, or remain local to their own domain.

By putting each edge switch and VLAN into a separate EAPS domain, you gain resiliency and management benefits. First, any link or switch failures in one ring do not affect the other edge switches. Also, this type of modular design allows you to add edge switches easily without impacting other parts of the network. Troubleshooting becomes easier as the scope of failures can be quickly isolated to a specific EAPS ring or switch.

This section describes how to design the access edge network switches as EAPS transit nodes to provide Ethernet L2 connectivity services. In this example, upstream aggregation switches perform Layer 3 (L3) inter-VLAN routing functions. Although not discussed in the scope of this section, the edge switches could also be configured with additional routing, *QoS*, *WLAN (Wireless Local Area Network)*, or security features.

### Creating and Configuring the EAPS Domain

- Create the EAPS domain, configure the switch as a transit node, and define the EAPS primary and secondary ports as shown in the following example:

```
* Edge-Switch#1:1 # create eaps e1-domain
* Edge-Switch#1:2 # configure eaps e1-domain mode transit
```

```
* Edge-Switch#1:3 # configure eaps e1-domain primary port 49
* Edge-Switch#1:4 # configure eaps e1-domain secondary port 50
```

#### Creating and Configuring the EAPS Control VLAN

1. Create the EAPS control *VLAN* and configure its 802.1q tag and ring ports.
2. Configure the control VLAN as part of the EAPS domain. The control VLAN only contains the EAPS primary and secondary ports configured earlier. The following commands accomplish these tasks:

```
* Edge-Switch#1:5 # create vlan control-1
* Edge-Switch#1:6 # configure vlan control-1 tag 4000
* Edge-Switch#1:8 # configure vlan control-1 add port 49,50 tagged
* Edge-Switch#1:9 # configure eaps e1-domain add control vlan control-1
```

#### Creating and Configuring EAPS Protected VLANs

1. Create at least one EAPS protected *VLAN*, and configure its 802.1q tag and ports.
2. Configure the protected VLAN as part of the EAPS domain.

   The Protect VLAN contains the EAPS primary and secondary ports as tagged VLAN ports. Additional VLAN ports connected to client devices such as a PC could be untagged or tagged. The following commands accomplish these tasks and should be repeated for all additional protected VLANs:

```
* Edge-Switch#1:10 # create vlan purple-1
* Edge-Switch#1:11 # configure purple-1 tag 1
* Edge-Switch#1:12 # configure purple-1 add port 49,50 tagged
* Edge-Switch#1:13 # configure purple-1 add port 1 untagged
* Edge-Switch#1:14 # configure eaps e1-domain add protect vlan purple-1
```

#### Enabling the EAPS Protocol and EAPS Domain

- Enable EAPS to run on the domain as shown in the following example:

```
* Edge-Switch#1:15 # enable eaps
* Edge-Switch#1:16 # enable eaps e1-domain
```

#### Verifying the EAPS Configuration and Status

- The command in the following example allows you to verify that the EAPS configuration is correct and that the EAPS state is Links-Up.

  Both ring ports must be plugged in to see the Links-Up state.

```
* Edge-Switch#1:17 # show eaps e1-domain detail
Name: "e1-domain" (instance=0) Priority: High
State: Links-Up                Running: Yes
Enabled: Yes                   Mode: Transit
Primary port:   49            Port status: Up         Tag status: Tagged
Secondary port: 50            Port status: Up         Tag status: Tagged
Hello Timer interval: 1 sec  0  millisec
Fail Timer interval: 3 sec
Preforwarding Timer interval: 0 sec
Last valid EAPS update: From Master Id 00:04:96:10:51:50, at Sun Sep 5 23:20:10 2004
EAPS Domain has following Controller Vlan:
Vlan Name                     VID
"control-1"                   4000
EAPS Domain has following Protected Vlan(s):
Vlan Name                     VID
"purple-1"                    0001
Number of Protected Vlans: 1
```

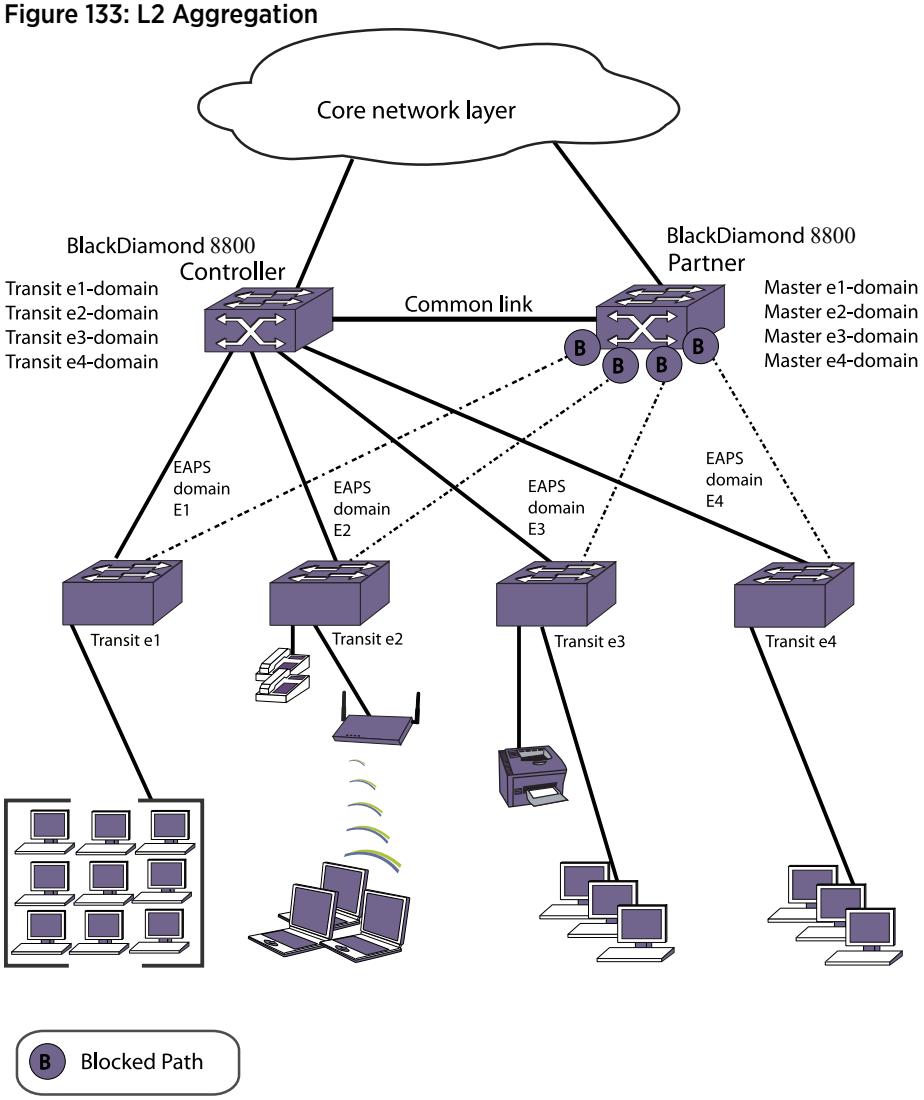
*Designing and Configuring the Aggregation Layer*

The network switches in the aggregation layer provide additional resiliency benefits.

In the following example, aggregation switches are typically deployed in pairs that protect against single switch failures. Each aggregation switch is physically connected to all edge switches and participates in multiple EAPS domains. The aggregation switches can serve a different role within each EAPS domain, with one switch acting as a transit node and the other as a master node.

In this example, we have a common link with overlapping domains (and protected VLANs), which includes an EAPS controller and partner configuration. The result is a partial-mesh network design of EAPS from the access edge to the aggregation layer (see the following figure).

**Figure 133: L2 Aggregation**




The aggregation switches are configured to act as multi-function EAPS nodes to provide L2 connectivity services. After EAPS and L2 connectivity is configured, additional L3 routing configuration can be added.

Using redundant aggregation switches helps protect against a single point of failure at the switch level, while EAPS domains provide fault isolation and minimize the impact that failures have on the network.

With shared port configurations, the partial-mesh physical design is maintained without broadcast loops, regardless of where a failure might occur.

To configure the L2 aggregate switches, complete the tasks described in the following sections on all aggregate switches:

1. Create and configure the EAPS domains.
2. Create and configure the EAPS control VLANs.
3. Create and configure the EAPS shared ports.
4. Enable the EAPS protocol and EAPS domain.
5. Create and configure the EAPS protected VLANs.
6. Verify the EAPS configuration and operating state.

**Creating and Configuring the EAPS Domains**

- Create the EAPS domains for each ring (one domain for one edge switch) and configure the EAPS mode.

  Define the primary and secondary ports for each domain. In this example, however, the primary port is the same as the common link. One aggregation switch has EAPS mode configured as master and partner, while the other aggregation switch is configured as transit and controller.

  EAPS master node configuration:

```
* AGG-SWITCH#2.1 # create eaps e1-domain
* AGG-SWITCH#2.2 # create eaps e2-domain
* AGG-SWITCH#2.3 # create eaps e3-domain
* AGG-SWITCH#2.4 # create eaps e4-domain
* AGG-SWITCH#2.5 # configure eaps e1-domain mode master
* AGG-SWITCH#2.6 # configure eaps e2-domain mode master
* AGG-SWITCH#2.7 # configure eaps e3-domain mode master
* AGG-SWITCH#2.8 # configure eaps e4-domain mode master
* AGG-SWITCH#2.9 # configure eaps e1-domain primary port 2:1
* AGG-SWITCH#2.10 # configure eaps e1-domain secondary port 1:1
* AGG-SWITCH#2.11 # configure eaps e2-domain primary port 2:1
* AGG-SWITCH#2.12 # configure eaps e2-domain secondary port 1:4
* AGG-SWITCH#2.13 # configure eaps e3-domain primary port 2:1
* AGG-SWITCH#2.14 # configure eaps e3-domain secondary port 3:1
* AGG-SWITCH#2.15 # configure eaps e4-domain primary port 2:1
* AGG-SWITCH#2.16 # configure eaps e4-domain secondary port 3:2
```

  EAPS transit node configuration:

```
* AGG-SWITCH#1.1 # create eaps e1-domain
* AGG-SWITCH#1.2 # create eaps e2-domain
* AGG-SWITCH#1.3 # create eaps e3-domain
* AGG-SWITCH#1.4 # create eaps e4-domain
* AGG-SWITCH#1.5 # configure eaps e1-domain mode transit
* AGG-SWITCH#1.6 # configure eaps e2-domain mode transit
* AGG-SWITCH#1.7 # configure eaps e3-domain mode transit
* AGG-SWITCH#1.8 # configure eaps e4-domain mode transit
* AGG-SWITCH#1.9 # configure eaps e1-domain primary port 2:1
* AGG-SWITCH#1.10 # configure eaps e1-domain secondary port 1:1
* AGG-SWITCH#1.11 # configure eaps e2-domain primary port 2:1
* AGG-SWITCH#1.12 # configure eaps e2-domain secondary port 1:4
* AGG-SWITCH#1.13 # configure eaps e3-domain primary port 2:1
* AGG-SWITCH#1.14 # configure eaps e3-domain secondary port 3:1
```

```
* AGG-SWITCH#1.15 # configure eaps e4-domain primary port 2:1
* AGG-SWITCH#1.16 # configure eaps e4-domain secondary port 3:2
```

#### Creating and Configuring the EAPS Control VLANs

1. Create the EAPS control *VLAN*s (one for each domain) and configure the 802.1q tag and ring ports for each.
2. Configure the control VLANs as part of their respective EAPS domain.

   The control VLAN only contains the EAPS primary and secondary ports configured earlier. The following commands are entered on both aggregate switches:

```
* AGG-SWITCH.17 # create vlan control-1
* AGG-SWITCH.18 # create vlan control-2
* AGG-SWITCH.19 # create vlan control-3
* AGG-SWITCH.20 # create vlan control-4
* AGG-SWITCH.21 # configure vlan control-1 tag 4001
* AGG-SWITCH.22 # configure vlan control-2 tag 4002
* AGG-SWITCH.23 # configure vlan control-3 tag 4003
* AGG-SWITCH.24 # configure vlan control-4 tag 4004
* AGG-SWITCH.29 # configure vlan control-1 add port 2:1,1:1 tagged
* AGG-SWITCH.30 # configure vlan control-2 add port 2:1,1:4 tagged
* AGG-SWITCH.31 # configure vlan control-3 add port 2:1,3:1 tagged
* AGG-SWITCH.32 # configure vlan control-4 add port 2:1,3:2 tagged
* AGG-SWITCH.33 # configure eaps e1-domain add control vlan control-1
* AGG-SWITCH.34 # configure eaps e2-domain add control vlan control-2
* AGG-SWITCH.35 # configure eaps e3-domain add control vlan control-3
* AGG-SWITCH.36 # configure eaps e4-domain add control vlan control-4
```

#### Creating and Configuring the EAPS Shared Ports

- Create the EAPS shared ports, which are used to connect a common-link between the aggregate switches.

  On the first switch, define the shared port mode as partner, and define the link ID. Repeat this step on the other aggregate switch, but configure the shared port mode as controller. The link ID matches the value configured for the partner.

  The following shows an example configuration for the partner:

```
* AGG-SWITCH#2.37 # create eaps shared-port 2:1
* AGG-SWITCH#2.38 # configure eaps shared-port 2:1 mode partner
* AGG-SWITCH#2.39 # configure eaps shared-port 2:1 link-id 21
```

#### Enabling the EAPS Protocol and EAPS Domain

- Enable the EAPS protocol on the switch, and enable EAPS to run on each domain created.

  The following commands are entered on both aggregate switches.

```
* AGG-SWITCH.40 # enable eaps
* AGG-SWITCH.41 # enable eaps e1-domain
* AGG-SWITCH.42 # enable eaps e2-domain
* AGG-SWITCH.43 # enable eaps e3-domain
* AGG-SWITCH.44 # enable eaps e4-domain
```

#### Creating and Configuring the EAPS Protected VLANs

1. Create the EAPS protected *VLAN*s for each domain.
2. Configure an 802.1q tag and the ports for each protected VLAN.
3. Configure each protected VLAN as part of the EAPS domain.

   Depending on the scope of the VLAN, it could be added to multiple EAPS domains. This type of VLAN is referred to as an *overlapping protected VLAN*, and requires shared port configurations.

In this example, there is one overlapping protected VLAN, purple-1, while all other VLANs are isolated to a single EAPS domain (VLANs green-1, orange-1, and red-1). Protected VLAN configuration, such as 802.1q tagging, must match on the edge switch. The commands in the following example are entered on both aggregate switches.

This procedure can also be repeated for additional protected VLANs as needed:

```
* AGG-SWITCH.44 # create vlan purple-1
* AGG-SWITCH.45 # create vlan green-1
* AGG-SWITCH.46 # create vlan orange-1
* AGG-SWITCH.47 # create vlan red-1
* AGG-SWITCH.48 # configure purple-1 tag 1
* AGG-SWITCH.49 # configure green-1 tag 2
* AGG-SWITCH.50 # configure orange-1 tag 3
* AGG-SWITCH.51 # configure red-1 tag 4
* AGG-SWITCH.52 # configure eaps e1-domain add protect vlan purple-1
* AGG-SWITCH.53 # configure eaps e2-domain add protect vlan purple-1
* AGG-SWITCH.54 # configure eaps e3-domain add protect vlan purple-1
* AGG-SWITCH.55 # configure eaps e4-domain add protect vlan purple-1
* AGG-SWITCH.56 # configure eaps e2-domain add protect vlan green-1
* AGG-SWITCH.57 # configure eaps e3-domain add protect vlan orange-1
* AGG-SWITCH.58 # configure eaps e4-domain add protect vlan red-1
* AGG-SWITCH.59 # configure vlan purple-1 add port 2:1,1:1,1:4,3:1,3:2 tagged
* AGG-SWITCH.60 # configure vlan green-1 add port 2:1,1:4 tagged
* AGG-SWITCH.61 # configure vlan orange-1 add port 2:1,3:1 tagged
* AGG-SWITCH.62 # configure vlan red-1 add port 2:1,3:2 tagged
```

#### Verifying the EAPS Configuration and Operating State

1. When the configuration is complete, confirm that the EAPS domain and shared port configuration is correct.
2. Verify whether the EAPS state is Complete and the shared port status is Ready.

   Both ring ports must be plugged in to see the Links-Up state. This verification is performed on both aggregate switches.

   EAPS master and partner node status verification example:

```
* AGG-SWITCH#2.63 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: Off
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 4
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e1-domain Complete M Y 2:1 1:1 control-1 (4001) 1
e2-domain Complete M Y 2:1 1:4 control-2 (4002) 2
e3-domain Complete M Y 2:1 3:1 control-3 (4003) 2
e4-domain Complete M Y 2:1 3:2 control-4 (4004) 2
--------------------------------------------------------------------------------
* AGG-SWITCH#2.64 # show eaps shared-port
EAPS shared-port count: 1
--------------------------------------------------------------------------------
Link Domain Vlan RB RB
Shared-port Mode Id Up State count count Nbr State Id
--------------------------------------------------------------------------------
```

```
2:1 Partner 21 Y Ready 4 4 Yes None None
--------------------------------------------------------------------------------
```

EAPS transit and controller node status verification example:

```
* AGG-SWITCH#1.63 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: Off
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 4
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e1-domain Links-Up M Y 2:1 1:1 control-1 (4001) 1
e2-domain Links-Up M Y 2:1 1:4 control-2 (4002) 2
e3-domain Links-Up M Y 2:1 3:1 control-3 (4003) 2
e4-domain Links-Up M Y 2:1 3:2 control-4 (4004) 2
--------------------------------------------------------------------------------
* AGG-SWITCH#1.64 # show eaps shared-port
EAPS shared-port count: 1
--------------------------------------------------------------------------------
Link Domain Vlan RB RB
Shared-port Mode Id Up State count count Nbr State Id
--------------------------------------------------------------------------------
2:1 Controller 21 Y Ready 4 4 Yes None None
--------------------------------------------------------------------------------
```

### *Designing and Configuring L3 Services on top of EAPS*

This section explains how to run L3 routing services on top of EAPS as a foundation.

In this example, *OSPF* is used as the dynamic IP routing protocol to communicate between different *VLAN*s. To provide redundancy at the router level, *VRRP* is used to protect against an aggregation switch failure. VRRP allows one aggregation switch to route IP traffic, and if it fails the other aggregation switch takes over the IP routing role. Each EAPS protected VLAN provides L3 connectivity to the clients by configuring IP addressing, OSPF routing, and VRRP on the aggregation switches.

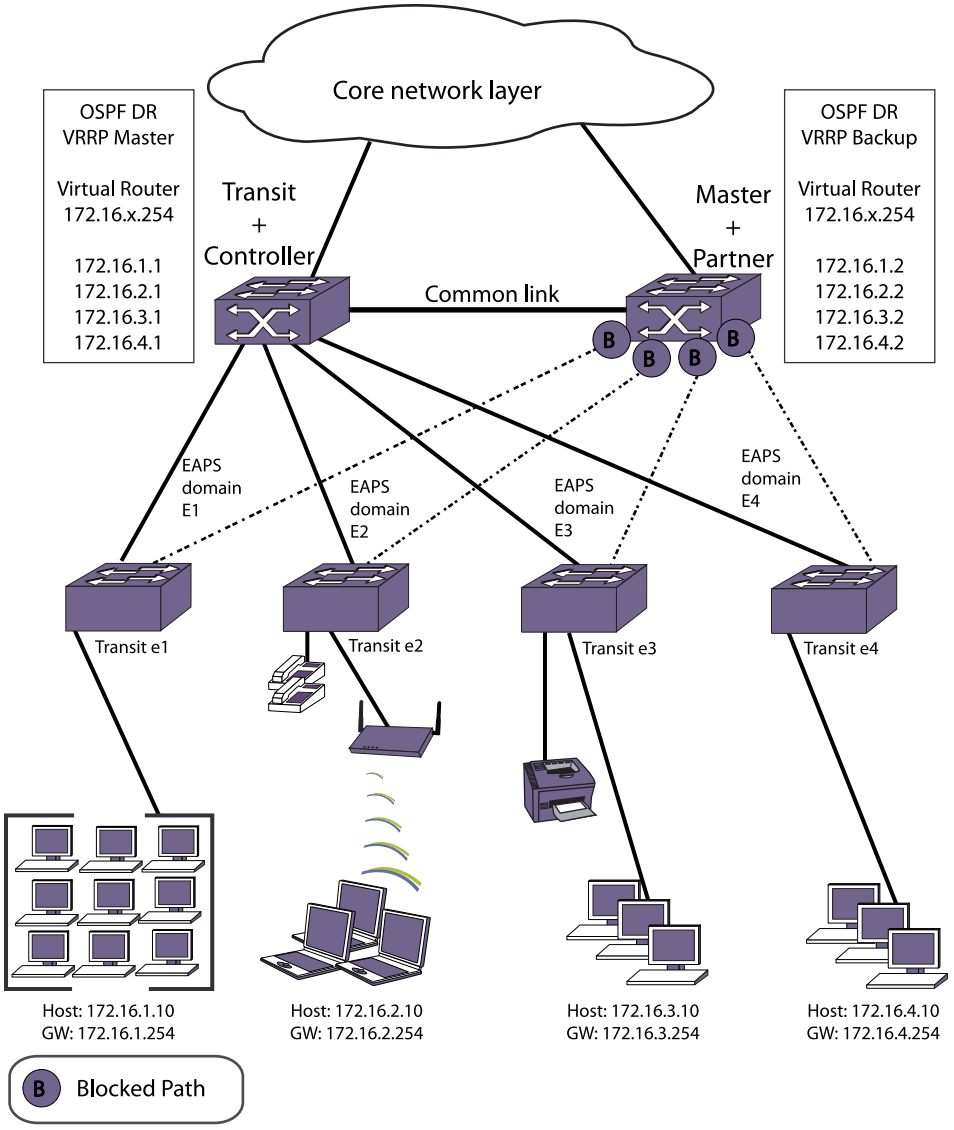



**Figure 134: L2 and L3 Aggregation Network Layer**

IP routing is added to the design on the access network switches by configuring each EAPS protected VLAN as an OSPF interface. Because these are broadcast OSPF interfaces, we need to specify a Designated Router (DR) and Backup Designated Router (BDR). While the EAPS transit and controller node is not blocking any ports, it is configured as the OSPF DR.

The EAPS master and partner node is then configured as the BDR. Similarly, the EAPS transit and controller node is also configured as the VRRP master, which provides L3 routing to the hosts. The EAPS master and partner node is configured as the VRRP backup router for redundancy.

Using redundant aggregation switches with VRRP protects against a single point of failure at the switch level. Client devices receive non-stop IP routing services in the event of link or aggregation switch failure without any reconfiguration. OSPF provides fast convergence from any routing failures. EAPS provides the resilient L2 foundation and minimizes the occurrence of routing interface flaps or dropped OSPF neighbor adjacencies.

To configure L3 on the aggregation switches, completed the tasks described in the following sections:

1. Configure OSPF on the EAPS protected VLANs.
2. Configure OSPF on the EAPS protected VLANs.
3. Configure VRRP on the EAPS protected VLANs.
4. Verify OSPF and VRRP configuration status.

#### Configuring IP Addresses on the EAPS Protected VLANs

Client host stations need the IP address configuration to match their protected VLANs. The edge switches do not require IP addresses, but this could optionally be done for management or troubleshooting purposes.

The following example shows IP address configuration:

```
* AGG-SWITCH#1.1 # configure vlan green-1 ipaddress 172.16.1.1/24
* AGG-SWITCH#1.2 # configure vlan purple-1 ipaddress 172.16.2.1/24
* AGG-SWITCH#1.3 # configure vlan orange-1 ipaddress 172.16.3.1/24
* AGG-SWITCH#1.4 # configure vlan red-1 ipaddress 172.16.4.1/24
* AGG-SWITCH#2.1 # configure vlan green-1 ipaddress 172.16.1.2/24
* AGG-SWITCH#2.2 # configure vlan purple-1 ipaddress 172.16.2.2/24
* AGG-SWITCH#2.3 # configure vlan orange-1 ipaddress 172.16.3.2/24
* AGG-SWITCH#2.4 # configure vlan red-1 ipaddress 172.16.4.2/24
```

#### Configuring OSPF on the EAPS Protected VLANs

Because *OSPF* broadcast networks are being used, configure the DR and BDR for each *VLAN*. Configure the EAPS transit and controller as the DR by using a higher OSPF priority value since it is not performing L2 blocking. The EAPS master and partner switch is configured as the BDR. In this example, all edge EAPS protected VLANs are placed in the OSPF backbone area, but another OSPF area could be created if desired.

Example OSPF DR configuration:

```
* AGG-SWITCH#1.5 # enable ipforwarding vlan green-1
* AGG-SWITCH#1.6 # enable ipforwarding vlan purple-1
* AGG-SWITCH#1.7 # enable ipforwarding vlan orange-1
* AGG-SWITCH#1.8 # enable ipforwarding vlan red-1
* AGG-SWITCH#1.9 # configure ospf routerid 172.16.1.1
* AGG-SWITCH#1.10 # configure ospf add vlan green-1 area 0.0.0.0
* AGG-SWITCH#1.11 # configure ospf add vlan purple-1 area 0.0.0.0
* AGG-SWITCH#1.12 # configure ospf add vlan orange-1 area 0.0.0.0
* AGG-SWITCH#1.13 # configure ospf add vlan red-1 area 0.0.0.0
* AGG-SWITCH#1.14 # configure ospf vlan green-1 priority 110
* AGG-SWITCH#1.15 # configure ospf vlan purple-1 priority 110
* AGG-SWITCH#1.16 # configure ospf vlan orange-1 priority 110
* AGG-SWITCH#1.17 # configure ospf vlan red-1 priority 110
* AGG-SWITCH#1.18 # enable ospf
```

Example OSPF BDR configuration:

```
* AGG-SWITCH#2.5 # enable ipforwarding vlan green-1
* AGG-SWITCH#2.6 # enable ipforwarding vlan purple-1
* AGG-SWITCH#2.7 # enable ipforwarding vlan orange-1
* AGG-SWITCH#2.8 # enable ipforwarding vlan red-1
* AGG-SWITCH#2.9 # configure ospf routerid 172.16.1.2
* AGG-SWITCH#2.10 # configure ospf add vlan green-1 area 0.0.0.0
* AGG-SWITCH#2.11 # configure ospf add vlan purple-1 area 0.0.0.0
* AGG-SWITCH#2.12 # configure ospf add vlan orange-1 area 0.0.0.0
* AGG-SWITCH#2.13 # configure ospf add vlan red-1 area 0.0.0.0
```

```
* AGG-SWITCH#2.14 # configure ospf vlan green-1 priority 100
* AGG-SWITCH#2.15 # configure ospf vlan purple-1 priority 100
* AGG-SWITCH#2.16 # configure ospf vlan orange-1 priority 100
* AGG-SWITCH#2.17 # configure ospf vlan red-1 priority 100
* AGG-SWITCH#2.18 # enable ospf
```

#### Configuring VRRP on the EAPS Protected VLANs

The *VRRP virtual router (VR)* is configured with the virtual IP address of 172.16.x.254 for each *VLAN* (example VLAN green-1 = 172.16.1.254). The VRRP virtual router IP address is configured as the default gateway of each client machine. Since it is not performing L2 blocking, configure the EAPS transit and controller as VRRP master router by using a higher priority value. The EAPS master and partner switch is configured as the VRRP backup router.

Example VRRP master router configuration:

```
* AGG-SWITCH#1.19 # create vrrp vlan green-1 vrid 1
* AGG-SWITCH#1.20 # configure vrrp vlan green-1 vrid 1 priority 110
* AGG-SWITCH#1.21 # configure vrrp vlan green-1 vrid 1 add 172.16.1.254
* AGG-SWITCH#1.22 # enable vrrp vlan green-1 vrid 1
* AGG-SWITCH#1.23 # create vrrp vlan purple-1 vrid 1
* AGG-SWITCH#1.24 # configure vrrp vlan purple-1 vrid 1 priority 110
* AGG-SWITCH#1.25 # configure vrrp vlan purple-1 vrid 1 add 172.16.2.254
* AGG-SWITCH#1.26 # enable vrrp vlan purple-1 vrid 1
* AGG-SWITCH#1.27 # create vrrp vlan orange-1 vrid 1
* AGG-SWITCH#1.28 # configure vrrp vlan orange-1 vrid 1 priority 110
* AGG-SWITCH#1.29 # configure vrrp vlan orange-1 vrid 1 add 172.16.3.254
* AGG-SWITCH#1.30 # enable vrrp vlan orange-1 vrid 1
* AGG-SWITCH#1.31 # create vrrp vlan red-1 vrid 1
* AGG-SWITCH#1.32 # configure vrrp vlan red-1 vrid 1 priority 110
* AGG-SWITCH#1.33 # configure vrrp vlan red-1 vrid 1 add 172.16.4.254
* AGG-SWITCH#1.34 # enable vrrp vlan red-1 vrid 1
```

Example VRRP backup router configuration:

```
* AGG-SWITCH#2.19 # create vrrp vlan green-1 vrid 1
* AGG-SWITCH#2.20 # configure vrrp vlan green-1 vrid 1 priority 100
* AGG-SWITCH#2.21 # configure vrrp vlan green-1 vrid 1 add 172.16.1.254
* AGG-SWITCH#2.22 # enable vrrp vlan green-1 vrid 1
* AGG-SWITCH#2.23 # create vrrp vlan purple-1 vrid 1
* AGG-SWITCH#2.24 # configure vrrp vlan purple-1 vrid 1 priority 100
* AGG-SWITCH#2.25 # configure vrrp vlan purple-1 vrid 1 add 172.16.2.254
* AGG-SWITCH#2.26 # enable vrrp vlan purple-1 vrid 1
* AGG-SWITCH#2.27 # create vrrp vlan orange-1 vrid 1
* AGG-SWITCH#2.28 # configure vrrp vlan orange-1 vrid 1 priority 100
* AGG-SWITCH#2.29 # configure vrrp vlan orange-1 vrid 1 add 172.16.3.254
* AGG-SWITCH#2.30 # enable vrrp vlan orange-1 vrid 1
* AGG-SWITCH#2.31 # create vrrp vlan red-1 vrid 1
* AGG-SWITCH#2.32 # configure vrrp vlan red-1 vrid 1 priority 100
* AGG-SWITCH#2.33 # configure vrrp vlan red-1 vrid 1 add 172.16.4.254
* AGG-SWITCH#2.34 # enable vrrp vlan red-1 vrid 1
```

#### Verifying OSPF and VRRP Configuration Status

1. Verify the *OSPF* neighbor adjacencies are established and that the DR and BDR status is correct.
2. Verify that the *VRRP* virtual router is running and the VRRP master/backup status is correct.

   OSPF and VRRP verification example:

   ```
   * AGG-SWITCH#1.35 # show ospf neighbor
   Neighbor ID Pri State Up/Dead Time Address Interface
   172.16.1.2 100 FULL /BDR 00:18:01:08/00:00:00:03 172.16.3.2 orange-1
   ```

```
172.16.1.2 100 FULL /BDR 00:18:01:08/00:00:00:03 172.16.4.2 red-1
172.16.1.2 100 FULL /BDR 00:17:54:17/00:00:00:03 172.16.1.2 green-1
172.16.1.2 100 FULL /BDR 00:17:54:07/00:00:00:03 172.16.2.2 purple-1
* AGG-SWITCH#1.36 # show vrrp
VLAN Name VRID Pri Virtual IP Addr State Master Mac Address TP/TR/TV/P/T
green-1(En) 0001 110 172.16.1.254 MSTR 00:00:5e:00:01:01 0 0 0 Y 1
purple-(En) 0001 110 172.16.2.254 MSTR 00:00:5e:00:01:01 0 0 0 Y 1
orange-(En) 0001 110 172.16.3.254 MSTR 00:00:5e:00:01:01 0 0 0 Y 1
red-1(En) 0001 110 172.16.4.254 MSTR 00:00:5e:00:01:01 0 0 0 Y 1
En-Enabled, Ds-Disabled, Pri-Priority, T-Advert Timer, P-Preempt
TP-Tracked Pings, TR-Tracked Routes, TV-Tracked VLANs
* AGG-SWITCH#2.35 # show ospf neighbor
Neighbor ID Pri State Up/Dead Time Address Interface
172.16.1.1 110 FULL /DR 00:18:01:08/00:00:00:03 172.16.3.1 orange-1
172.16.1.1 110 FULL /DR 00:18:01:08/00:00:00:03 172.16.4.1 red-1
172.16.1.1 110 FULL /DR 00:17:54:17/00:00:00:03 172.16.1.1 green-1
172.16.1.1 110 FULL /DR 00:17:54:07/00:00:00:03 172.16.2.1 purple-1
* AGG-SWITCH#2.36 # show vrrp
VLAN Name VRID Pri Virtual IP Addr State Master Mac Address TP/TR/TV/P/T
green-1(En) 0001 100 172.16.1.254 BKUP 00:00:5e:00:01:01 0 0 0 Y 1
purple-(En) 0001 100 172.16.2.254 BKUP 00:00:5e:00:01:01 0 0 0 Y 1
orange-(En) 0001 100 172.16.3.254 BKUP 00:00:5e:00:01:01 0 0 0 Y 1
red-1(En) 0001 100 172.16.4.254 BKUP 00:00:5e:00:01:01 0 0 0 Y 1
En-Enabled, Ds-Disabled, Pri-Priority, T-Advert Timer, P-Preempt
TP-Tracked Pings, TR-Tracked Routes, TV-Tracked VLANs
```

### *Designing and Configuring the Core Layer with EAPS*

The core switches provide high performance backbone routing between the edge, aggregation, data center, and external Internet networks.

An additional high availability backbone ring is built that combines EAPS and *OSPF*. Using EAPS and OSPF together increases the stability of IP routing tables. Since EAPS provides 50-millisecond convergence for link failures, OSPF adjacencies do not flap. In this example, the backbone ring is formed by adding two core L2/L3 switches and connecting them to the two existing aggregation switches. The core switches also provide routing to the Internet using BGP (see the following figure).

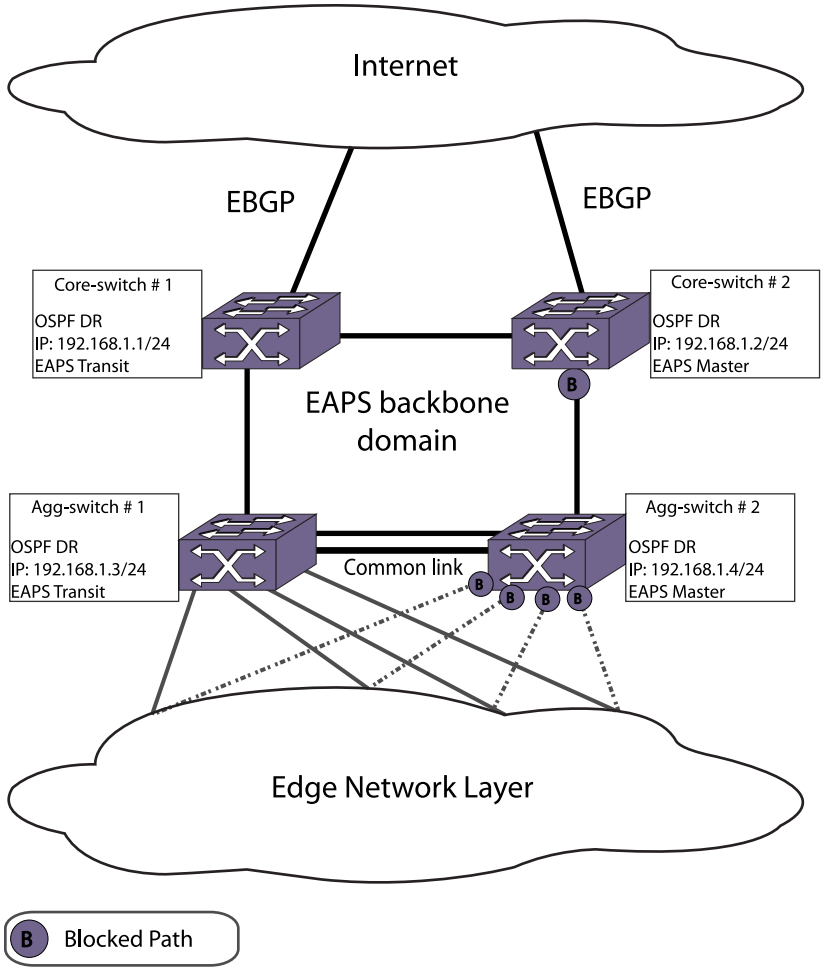



**Figure 135: Core EAPS and OSPF Network Layer**

Using redundant core switches protects against a single point of failure at the switch level. OSPF provides fast convergence from any routing failures. EAPS provides the resilient L2 foundation and minimizes the occurrence of routing interface flaps or dropped OSPF neighbor adjacencies. Combining EAPS and OSPF provides the highest level of network resiliency and routing stability.

Configuring the core switches requires a new EAPS domain with a single EAPS protected *VLAN* with OSPF forming the backbone IP network. Additional configuration is needed on the aggregation switches to connect them to the backbone EAPS and OSPF ring. Since the steps are similar to previous configuration examples, the L2 (EAPS) and L3 (OSPF) configurations are combined. Since the BGP configuration is independent of EAPS configuration, BGP configuration is not discussed here.

To configure backbone connectivity on the core and aggregation switches, complete the tasks described in the following sections:

1. Create and configure the backbone EAPS domain.
2. Create and configure the backbone EAPS protected VLANs.
3. Configure an IP address and OSPF on the backbone VLAN.
4. Verify EAPS and OSPF configuration status.

#### Creating and Configuring the Backbone EAPS Domain

1. Create the backbone EAPS domains and configure the EAPS mode.
2. Define the primary and secondary ports for each domain.
   Configure on both core and aggregation switches.

   Core-Switch 1 EAPS configuration:

```
* CORE-SWITCH#1.1 # create eaps e5-domain
* CORE-SWITCH#1.2 # configure eaps e5-domain mode transit
* CORE-SWITCH#1.3 # configure eaps e5-domain primary port 2:1
* CORE-SWITCH#1.4 # configure eaps e5-domain secondary port 2:4
```

   Core-Switch 2 EAPS configuration:

```
* CORE-SWITCH#2.1 # create eaps e5-domain
* CORE-SWITCH#2.2 # configure eaps e5-domain mode master
* CORE-SWITCH#2.3 # configure eaps e5-domain primary port 2:1
* CORE-SWITCH#2.4 # configure eaps e5-domain secondary port 2:4
```

   Agg-Switch 1 EAPS configuration:

```
* AGG-SWITCH#1.1 # create eaps e5-domain
* AGG-SWITCH#1.2 # configure eaps e5-domain mode transit
* AGG-SWITCH#1.3 # configure eaps e5-domain primary port 2:1
* AGG-SWITCH#1.4 # configure eaps e5-domain secondary port 2:4
```

   Agg-Switch 2 EAPS configuration:

```
* AGG-SWITCH#2.1 # create eaps e5-domain
* AGG-SWITCH#2.2 # configure eaps e5-domain mode transit
* AGG-SWITCH#2.3 # configure eaps e5-domain primary port 2:1
* AGG-SWITCH#2.4 # configure eaps e5-domain secondary port 2:4
```

#### Creating and Configuring the Backbone EAPS Control VLAN

1. Create the EAPS control *VLAN* and configure its 802.1q tag, and ring ports.
2. Configure the control VLANs as part of the backbone EAPS domain. Enable EAPS and the backbone EAPS domain. Configure on both core and aggregation switches (EAPS is already enabled on aggregation switches).
   Core-Switch#1 control VLAN configuration:

```
* CORE-SWITCH#1.1 # create vlan control-5
* CORE-SWITCH#1.2 # configure vlan control-5 tag 4005
* CORE-SWITCH#1.4 # configure vlan control-5 add port 2:1,2:4 tagged
* CORE-SWITCH#1.5 # configure eaps e5-domain add control vlan control-5
* CORE-SWITCH#1.6 # enable eaps
* CORE-SWITCH#1.7 # enable eaps e5-domain
```

   Core-Switch#2 control VLAN configuration:

```
* CORE-SWITCH#2.1 # create vlan control-5
* CORE-SWITCH#2.2 # configure vlan control-5 tag 4005
* CORE-SWITCH#2.4 # configure vlan control-5 add port 2:1,2:4 tagged
* CORE-SWITCH#2.5 # configure eaps e5-domain add control vlan control-5
* CORE-SWITCH#2.6 # enable eaps
* CORE-SWITCH#2.7 # enable eaps e5-domain
```

   Agg-Switch#1 control VLAN configuration:

```
* AGG-SWITCH#1.1 # create vlan control-5
* AGG-SWITCH#1.2 # configure vlan control-5 tag 4005
* AGG-SWITCH#1.4 # configure vlan control-5 add port 2:4,2:6 tagged
* AGG-SWITCH#1.5 # configure eaps e5-domain add control vlan control-5
* AGG-SWITCH#1.6 # enable eaps e5-domain
```

Agg-Switch#2 control VLAN configuration:

```
* AGG-SWITCH#2.1 # create vlan control-5
* AGG-SWITCH#2.2 # configure vlan control-5 tag 4005
* AGG-SWITCH#2.4 # configure vlan control-5 add port 2:4,2:6 tagged
* AGG-SWITCH#2.5 # configure eaps e5-domain add control vlan control-5
* AGG-SWITCH#1.6 # enable eaps e5-domain
```

#### Creating and Configuring the Backbone EAPS Protected VLANs

1. Create the EAPS protected *VLAN* for the backbone domain.
2. Configure the 802.1q tag and ports for the protected VLANs.

   Because this VLAN is only used for transit routing, there are no other ports besides the ring ports.
3. Configure the protected VLAN as part of the EAPS domain. Do this configuration on both the core and aggregate switches.

   Core-Switch#1 protected VLAN configuration:

```
* CORE-SWITCH#1.8 # create vlan backbone
* CORE-SWITCH#1.9 # configure vlan backbone tag 3000
* CORE-SWITCH#1.10 # configure vlan backbone add port 2:1,2:4 tagged
* CORE-SWITCH#1.11 # configure eaps e5-domain add protect vlan backbone
```

   Core-Switch#2 protected VLAN configuration:

```
* CORE-SWITCH#2.8 # create vlan backbone
* CORE-SWITCH#2.9 # configure vlan backbone tag 3000
* CORE-SWITCH#2.10 # configure vlan backbone add port 2:1,2:4 tagged
* CORE-SWITCH#2.11 # configure eaps e5-domain add protect vlan backbone
```

   Agg-Switch#1 protected VLAN configuration:

```
* AGG-SWITCH#1.7 # create vlan backbone
* AGG-SWITCH#1.8 # configure vlan backbone tag 3000
* AGG-SWITCH#1.9 # configure vlan backbone add port 2:4,2:6 tagged
* AGG-SWITCH#1.10 # configure eaps e5-domain add protect vlan backbone
```

   Agg-Switch#2 protected VLAN configuration:

```
* AGG-SWITCH#2.7 # create vlan backbone
* AGG-SWITCH#2.8 # configure vlan backbone tag 3000
* AGG-SWITCH#2.9 # configure vlan backbone add port 2:4,2:6 tagged
* AGG-SWITCH#2.10 # configure eaps e5-domain add protect vlan backbone
```

#### Configuring an IP Address and OSPF on the Backbone VLAN

1. Configure an IP address and enable IP forwarding (routing) on the backbone protected *VLAN*.
2. *OSPF* is configured and because an OSPF broadcast network is used, configure the designated router and backup designated router for each VLAN.

   Since it is not performing L2 blocking, configure the EAPS transit core switch as the DR by using a higher OSPF priority value. The EAPS master core switch is configured as the BDR. The aggregation transit switches need not perform DR/BDR duties for the backbone VLAN, so their OSPF priority is configured at 0 to force ODR behavior.

   Core-Switch#1 OSPF configuration:

```
* CORE-SWITCH#1.12 # configure vlan backbone ipaddress 192.168.1.1/24
* CORE-SWITCH#1.13 # enable ipforwarding vlan backbone
* CORE-SWITCH#1.14 # configure ospf routerid 192.168.1.1
* CORE-SWITCH#1.15 # configure ospf add vlan backbone area 0.0.0.0
```

```
* CORE-SWITCH#1.16 # configure ospf vlan backbone priority 110
* CORE-SWITCH#1.17 # enable ospf
```

Core-Switch#2 OSPF configuration:

```
* CORE-SWITCH#2.12 # configure vlan backbone ipaddress 192.168.1.2/24
* CORE-SWITCH#2.13 # enable ipforwarding vlan backbone
* CORE-SWITCH#2.14 # configure ospf routerid 192.168.1.2
* CORE-SWITCH#2.15 # configure ospf add vlan backbone area 0.0.0.0
* CORE-SWITCH#2.16 # configure ospf vlan backbone priority 100
* CORE-SWITCH#2.17 # enable ospf
```

Agg-Switch#1 OSPF configuration:

```
* AGG-SWITCH#1.11 # configure vlan backbone ipaddress 192.168.1.3/24
* AGG-SWITCH#1.12 # enable ipforwarding vlan backbone
* AGG-SWITCH#1.13 # configure ospf add vlan backbone area 0.0.0.0
* AGG-SWITCH#1.14 # configure ospf vlan backbone priority 0
```

Agg-Switch#2 OSPF configuration:

```
* AGG-SWITCH#2.11 # configure vlan backbone ipaddress 192.168.1.4/24
* AGG-SWITCH#2.12 # enable ipforwarding vlan backbone
* AGG-SWITCH#2.13 # configure ospf add vlan backbone area 0.0.0.0
* AGG-SWITCH#2.14 # configure ospf vlan backbone priority 0
```

#### Verifying EAPS and OSPF Configuration Status

1. Verify that the backbone EAPS domain and *OSPF* configuration is correct.
2. Confirm that the OSPF neighbor adjacencies and DR/BDR/ODR status are correct. Verify this status on both aggregate switches.

   Core-Switch#1 EAPS and OSPF status example:

```
* CORE-SWITCH#1.18 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e5-domain Links-Up T Y 2:1 2:4 control-5 (4005) 1
--------------------------------------------------------------------------------
* CORE-SWITCH#1.19 # show ospf neighbor
Neighbor ID Pri State Up/Dead Time Address Interface
192.168.1.3 0 2WAY /DROTHER00:05:23:17/00:00:00:07 192.168.1.3 backbone
192.168.1.4 0 2WAY /DROTHER00:05:23:17/00:00:00:07 192.168.1.4 backbone
192.168.1.2 100 FULL /BDR 00:05:23:17/00:00:00:09 192.168.1.2 backbone
```

   Core-Switch#2 EAPS and OSPF status example:

```
* CORE-SWITCH#2.18 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
```

```
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e5-domain Complete T Y 2:1 2:4 control-5 (4005) 1
--------------------------------------------------------------------------------
* CORE-SWITCH#2.19 # show ospf neighbor
Neighbor ID Pri State Up/Dead Time Address Interface
192.168.1.3 0 2WAY /DROTHER00:05:23:17/00:00:00:07 192.168.1.3 backbone
192.168.1.4 0 2WAY /DROTHER00:05:23:17/00:00:00:07 192.168.1.4 backbone
192.168.1.1 110 FULL /DR 00:05:23:17/00:00:00:09 192.168.1.1 backbone
```

Agg-Switch#1 EAPS and OSPF status example:

```
* AGG-SWITCH#1.15 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 5
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e1-domain Links-Up T Y 1:1 2:1 control-1 (4001) 2
e2-domain Links-Up T Y 1:4 2:1 control-2 (4002) 2
e3-domain Links-Up T Y 3:1 2:1 control-3 (4003) 2
e4-domain Links-Up T Y 3:2 2:1 control-4 (4004) 2
e5-domain Links-Up T Y 2:4 2:6 control-5 (4005) 1
--------------------------------------------------------------------------------
* AGG-SWITCH#1.16 # show ospf neighbor
Neighbor ID Pri State Up/Dead Time Address Interface
192.168.1.1 110 FULL /DR 00:00:28:51/00:00:00:01 192.168.1.1 backbone
192.168.1.2 100 FULL /BDR 00:00:28:51/00:00:00:01 192.168.1.2 backbone
192.168.1.4 0 2WAY /DROTHER00:05:45:40/00:00:00:03 192.168.1.4 backbone
172.16.1.2 100 FULL /BDR 00:18:01:08/00:00:00:03 172.16.3.2 orange-1
172.16.1.2 100 FULL /BDR 00:18:01:08/00:00:00:03 172.16.4.2 red-1
172.16.1.2 100 FULL /BDR 00:17:54:17/00:00:00:03 172.16.1.2 green-1
172.16.1.2 100 FULL /BDR 00:17:54:07/00:00:00:03 172.16.2.2 purple-1
```

Agg-Switch#2 EAPS and OSPF status example:

```
* AGG-SWITCH#2.15 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 5
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain State Mo En Pri Sec Control-Vlan VID Count
--------------------------------------------------------------------------------
e1-domain Complete M Y 2:1 1:1 control-1 (4001) 2
e2-domain Complete M Y 2:1 1:4 control-2 (4002) 2
e3-domain Complete M Y 2:1 3:1 control-3 (4003) 2
e4-domain Complete M Y 2:1 3:2 control-4 (4004) 2
```

```
e5-domain Links-Up T Y 2:4 2:6 control-5 (4005) 1
--------------------------------------------------------------------------------
* AGG-SWITCH#2.16 # show ospf neighbor
Interface
192.168.1.1 110 FULL /DR 00:00:28:51/00:00:00:01 192.168.1.1 backbone
192.168.1.2 100 FULL /BDR 00:00:28:51/00:00:00:01 192.168.1.2 backbone
192.168.1.3 0 2WAY /DROTHER00:05:45:40/00:00:00:03 192.168.1.3 backbone
172.16.1.1 110 FULL /DR 00:18:01:08/00:00:00:03 172.16.3.1 orange-1
172.16.1.1 110 FULL /DR 00:18:01:08/00:00:00:03 172.16.4.1 red-1
172.16.1.1 110 FULL /DR 00:17:54:17/00:00:00:03 172.16.1.1 green-1
172.16.1.1 110 FULL /DR 00:17:54:07/00:00:00:03 172.16.2.1 purple-1
```

### *Designing and Configuring the Data Center Switches with EAPS*

Building from the network core, you can expand the network with additional EAPS rings to provide resiliency to mission-critical server farms.

The core switches provide high performance backbone routing between the data center and the rest of the network, which includes both internal and external (Internet) destinations. The core switch acts as the EAPS master node for each ring, while the data center switches act as EAPS transit nodes to complete the ring. The core switch also acts as the *OSPF* routing node to provide gateway routing functionality to the server-farms. For an additional level of resiliency, each server is dual-homed (dual attached) to both EAPS transit L2 switches. Even if a switch or link fails, the servers are available.

The network design and configuration is similar to the edge and aggregation EAPS and OSPF layers. The modular approach is simple and scalable, and allows additional data center rings to be added to provide room for growth. In our example, server-farms are isolated into separate categories such as external and internal service groups, which yield additional security and resiliency benefits.

To configure the data center switches, you need a new EAPS domain with a single EAPS protected *VLAN* to form the server-farm network. In this example, two data center switches are configured as EAPS transit nodes (L2 switch only) and attach to the existing core switch acting as the EAPS master. Each server in the server-farm is dual-homed to both EAPS transit switches in the data center for additional physical resiliency. IP routing functionality is performed by the core switch via OSPF, which provides L3 connectivity to the rest of the network.

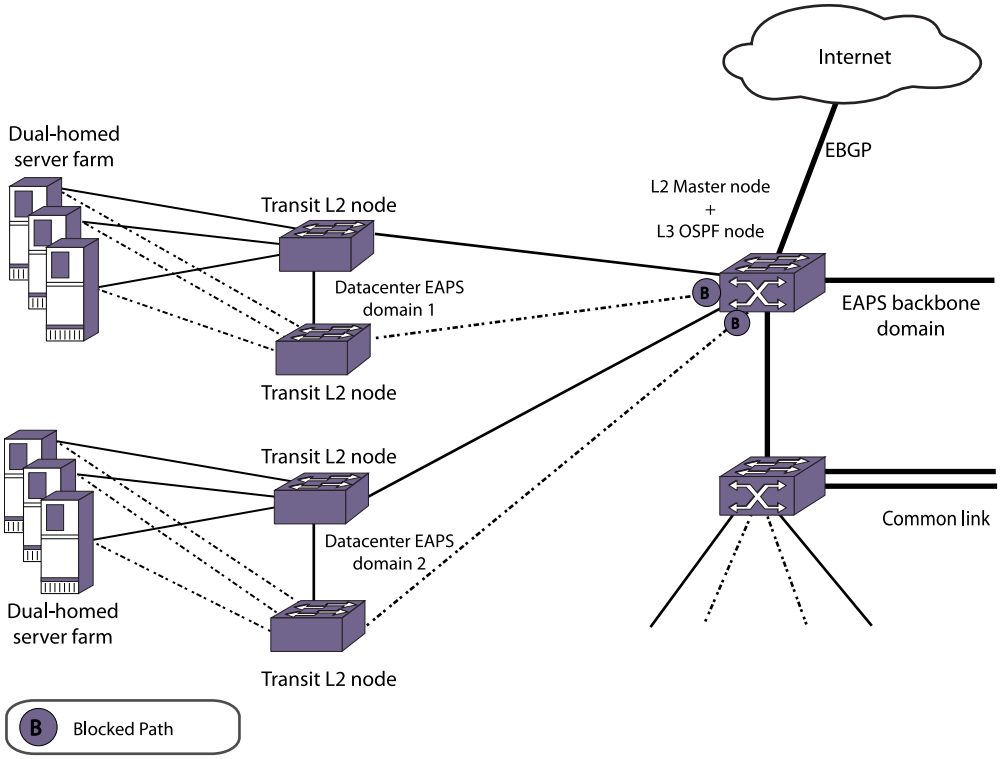



**Figure 136: Data Center EAPS and OSPF Network Layer**

To configure data center connectivity, complete the tasks described in the following sections:

1. Create and configure the data center EAPS domain.
2. Create and configure the data center EAPS Control VLAN.
3. Create and configure the data center EAPS protected VLANs.
4. Configure an IP address and OSPF on the backbone VLAN.
5. Verify EAPS and OSPF configuration status.

### Creating and Configuring the Data Center EAPS Domain

Create the backbone EAPS domains, configure the EAPS mode, and define the primary and secondary ports for each domain. Do this configuration on both core and aggregation switches.

Core-Switch#1 EAPS configuration:

```
* CORE-SWITCH#1.1 # create eaps e6-domain
* CORE-SWITCH#1.2 # configure eaps e6-domain mode master
* CORE-SWITCH#1.3 # configure eaps e6-domain primary port 4:1
* CORE-SWITCH#1.4 # configure eaps e6-domain secondary port 4:2
```

Data center-Switch#1 EAPS configuration:

```
* DC-SWITCH#1.1 # create eaps e6-domain
* DC-SWITCH#1.2 # configure eaps e6-domain mode transit
* DC-SWITCH#1.3 # configure eaps e6-domain primary port 49
* DC-SWITCH#1.4 # configure eaps e6-domain secondary port 50
```

Datacenter -Switch#2 EAPS configuration:

```
* DC-SWITCH#2.1 # create eaps e6-domain
* DC-SWITCH#2.2 # configure eaps e6-domain mode transit
```

```
* DC-SWITCH#2.3 # configure eaps e6-domain primary port 49
* DC-SWITCH#2.4 # configure eaps e6-domain secondary port 50
```

#### Creating and Configuring the Data Center EAPS Control VLAN

1. Create the EAPS control *VLAN* and configure its 802.1q tag, and ring ports.
2. Configure the control VLANs as part of the data center EAPS domain. Enable EAPS and the data center EAPS domain. You need to do this configuration on the core and data center L2 switches.
   Core-Switch#1 control VLAN configuration:

   ```
   * CORE-SWITCH#1.1 # create vlan control-6
   * CORE-SWITCH#1.2 # configure vlan control-6 tag 4006
   * CORE-SWITCH#1.4 # configure vlan control-6 add port 4:1,4:2 tagged
   * CORE-SWITCH#1.5 # configure eaps e5-domain add control vlan control-6
   * CORE-SWITCH#1.6 # enable eaps e6-domain
   ```

   Data center-Switch#1 control VLAN configuration:

   ```
   * DC-SWITCH#1.1 # create vlan control-6
   * DC-SWITCH#1.2 # configure vlan control-6 tag 4006
   * DC-SWITCH#1.4 # configure vlan control-6 add port 49,50 tagged
   * DC-SWITCH#1.5 # configure eaps e6-domain add control vlan control-6
   * DC-SWITCH#1.6 # enable eaps
   * DC-SWITCH#1.7 # enable eaps e6-domain
   ```

   Dat acenter-Switch#2 control VLAN configuration:

   ```
   * DC-SWITCH#2.1 # create vlan control-6
   * DC-SWITCH#2.2 # configure vlan control-6 tag 4006
   * DC-SWITCH#2.4 # configure vlan control-6 add port 49,50 tagged
   * DC-SWITCH#2.5 # configure eaps e6-domain add control vlan control-6
   * DC-SWITCH#2.6 # enable eaps
   * DC-SWITCH#2.7 # enable eaps e6-domain
   ```

#### Create and Configure the Data Center EAPS Protected VLANs

1. Create the EAPS protected *VLAN* for the data center domain.
2. Configure the 802.1q tag and ports for the protected VLANs.
   Because each server is dual-homed to each data center switch, add a VLAN port on each switch for each server.
3. Configure the protected VLAN as part of the EAPS domain. Do this configuration on the core and data center switches.
   Core-Switch#1 protected VLAN configuration:

   ```
   * CORE-SWITCH#1.7 # create vlan srvfarm-1
   * CORE-SWITCH#1.8 # configure vlan srvfarm-1 tag 1000
   * CORE-SWITCH#1.9 # configure vlan srvfarm-1 add port 4:1,4:2 tagged
   * CORE-SWITCH#1.10 # configure eaps e6-domain add protect vlan srvfarm-1
   ```

   Data center-Switch#1 protected VLAN configuration:

   ```
   * DC-SWITCH#1.8 # create vlan srvfarm-1
   * DC-SWITCH#1.9 # configure vlan srvfarm-1 tag 1000
   * DC-SWITCH#1.10 # configure vlan srvfarm-1 add port 49,50 tagged
   * DC-SWITCH#1.11 # configure vlan srvfarm-1 add port 1 untagged
   * DC-SWITCH#1.12 # configure eaps e5-domain add protect vlan srvfarm-1
   ```

   Data center-Switch#2 protected VLAN configuration:

   ```
   * DC-SWITCH#2.8 # create vlan srvfarm-1
   * DC-SWITCH#2.9 # configure vlan srvfarm-1 tag 1000
   ```

```
* DC-SWITCH#2.10 # configure vlan srvfarm-1 add port 49,50 tagged
* DC-SWITCH#2.11 # configure vlan srvfarm-1 add port 1 untagged
* DC-SWITCH#2.12 # configure eaps e5-domain add protect vlan srvfarm-1
```

#### Configuring an IP Address and OSPF on the Backbone VLAN

Configure an IP address and enable IP forwarding (routing) on the data center protected *VLAN*.

This step is only performed on the core switch. Servers are configured accordingly with the core switch IP address as their default gateway. Since there are no additional routers on this VLAN, configure it as an *OSPF* passive interface. In this example, the data center VLAN is placed on the backbone OSPF area, but additional OSPF areas can be configured if needed.

Core-Switch#1 OSPF configuration:

```
* CORE-SWITCH#1.11 # configure vlan srvfarm-1 ipaddress 10.10.10.10/24
* CORE-SWITCH#1.12 # enable ipforwarding vlan srvfarm-1
* CORE-SWITCH#1.13 # configure ospf add vlan srvfarm-1 area 0.0.0.0 passive
```

#### Verifying EAPS and OSPF Configuration Status

1. Verify that the data center EAPS domain and *OSPF* configuration is correct.
2. Verify whether the data center subnet is advertised to other routers through OSPF.

   Core-Switch#2 route verification example:

```
* CORE-SWITCH#2.1 # show iproute 10.10.10.0/24
Ori  Destination        Gateway         Mtr  Flags        VLAN       Duration
#oa  10.10.10.0/24      192.168.1.1     6    UG-D---um--f backbone   0d:0h:25m:5s
Origin(Ori): (b) BlackHole, (be) EBGP, (bg) BGP, (bi) IBGP, (bo) BOOTP
(ct) CBT, (d) Direct, (df) DownIF, (dv) DVMRP, (e1) ISISL1Ext
(e2) ISISL2Ext, (h) Hardcoded, (i) ICMP, (i1) ISISL1 (i2) ISISL2
(is) ISIS, (mb) MBGP, (mbe) MBGPExt, (mbi) MBGPInter, (mp) MPLS Lsp
(mo) MOSPF (o) OSPF, (o1) OSPFExt1, (o2) OSPFExt2
(oa) OSPFIntra, (oe) OSPFAsExt, (or) OSPFInter, (pd) PIM-DM, (ps) PIM-SM
(r) RIP, (ra) RtAdvrt, (s) Static, (sv) SLB_VIP, (un) UnKnown
(*) Preferred unicast route (@) Preferred multicast route
(#) Preferred unicast and multicast route
Flags: (B) BlackHole, (D) Dynamic, (G) Gateway, (H) Host Route
(L) Matching LDP LSP, (l) Calculated LDP LSP, (m) Multicast
(P) LPM-routing, (R) Modified, (S) Static, (s) Static LSP
(T) Matching RSVP-TE LSP, (t) Calculated RSVP-TE LSP, (u) Unicast, (U) Up
(f) Provided to FIB (c) Compressed Route
Mask distribution:
1 routes at length 16 1 routes at length 24
Route Origin distribution:
1 routes from OSPFIntra 1 routes from OSPFExt1
Total number of routes = 2
Total number of compressed routes = 0
```

Core-Switch#1 EAPS status:

```
* CORE-SWITCH#1.14 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 2
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain         State           Mo  En  Pri   Sec   Control-Vlan VID   Count
```

```
--------------------------------------------------------------------------------
e5-domain      Links-Up        T   Y   2:1   2:4   control-5    (4005) 1
e6-domain      Complete        T   Y   4:1   4:2   control-6    (4006) 1
--------------------------------------------------------------------------------
```

Data center-Switch#1 EAPS status:

```
* DC-SWITCH#1.15 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain         State           Mo  En  Pri   Sec   Control-Vlan VID   Count
--------------------------------------------------------------------------------
e6-domain      Links-Up        T   Y   49    50    control-6    (4006) 1
--------------------------------------------------------------------------------
```

Data center-Switch#2 EAPS status:

```
* DC-SWITCH#2.15 # show eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: On
EAPS Display Config Warnings: On
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
--------------------------------------------------------------------------------
Domain         State           Mo  En  Pri   Sec   Control-Vlan VID   Count
--------------------------------------------------------------------------------
e6-domain      Links-Up        M   Y   49    50    control-6    (4006) 1
--------------------------------------------------------------------------------
```

## CFM Support in EAPS

ExtremeXOS provides Connectivity Fault Management (CFM) support within EAPS protocol.

CFM reports fault connectivity failures to EAPS, and EAPS communicates with the CFM process to set up point-to-point DOWN MEPs (Management Endpoints) to monitor link connectivity. The CFM module notifies EAPS of any link-connectivity issues, and triggers EAPS to take necessary action.

802.1ag CFM supports link monitoring. It does this by sending out PDUs at designated transmit intervals. If the CFM fails to receive PDUs, it assumes the link is out of service, and notifies its clients. In this instance, EAPS acts as a CFM client.

First, you will create a down MEP within the CFM CLI. Configure the CLI to create a MEP group that associates this down MEP with a remote MEP (RMEP). There is a 1:1 relationship between a port and the down MEP, and as such, each MEP group is tied to a single port. Using the EAPS CLI, you can add the MEP groups you wish to monitor. For each MEP group added to EAPS, EAPS will receive UP/DOWN notifications from CFM when CFM detects a MEP state change for that group. Each MEP group

corresponds to an EAPS ring port. Notifications from those MEP groups that are inadvertently added, that do not correspond to an EAPS ring port, are ignored in EAPS.

The CFM configuration is independent of EAPS, and MEPs and MEP groups may use different *VLAN*s other than the EAPS control VLAN to monitor links.

When EAPS receives a CFM notification that the link failed, EAPS blocks that port on all of the EAPS control VLANs. This prevents EAPS control PDUs from being hardware forwarded on the link, in case the link is still up. Any EAPS PDUs that are received on a CFM failed port are dropped in EAPS.

*Configuring EAPS for CFM Support*

- Use the following command to configure EAPS for CFM support:

For additional configuration details for CFM support, refer to Configuring CFM on page 392.

**Binding to a MEP Group**

- To bind to a MEP Group, use the following command:

  `configure eaps cfm [add | delete] group group_name`

  This command notifies CFM that EAPs is interested in notifications for this MEP and RMEP pair. This MEP should already be bound to a physical port, so when notification is received, EAPS associates that notification with a ring-port failure.

*Create MPs and the CCM Transmission Interval*

Within an MA, you configure the following MPs:

- Maintenance end points (MEPs), which are one of the following types:
- ◦ UP MEPs—transmit CCMs and maintain CCM database
  ◦ DOWN MEPs—transmit CCMs and maintain CCM database
- Maintenance intermediate points (MIPs)—pass CCMs through

Each MEP must have an ID that is unique for that MEP throughout the MA.

- To configure UP and DOWN MEPs and its unique MEP ID, use the following command:

  `configure cfm domain domain_name association association_name [ports port_list add [[end-point [up|down] mepid {group group_name}] | [intermediate-point]]`

- To change the MEP ID on an existing MEP, use the following command:

  `configure cfm domain domain-name association association_name ports port_list end-point [up | down] mepid mepid`

- To delete UP and DOWN MEPs, use the following command:

  `configure cfm domain domain-name association association_name ports port_list delete end-point [up | down] intermediate-point`

- To configure a MIP, use the following command:

  `configure cfm domain domain_name association association_name [ports port_list add [[end-point [up|down] mepid {group group_name}] | [intermediate-point]]`

- To delete a MIP, use the following command:

  `configure cfm domain` *domain_name* **association** *association_name* [**ports** *port_list* **delete** [[**end-point** [**up**|**down**] *mepid* {**group** *group_name*}] | [**intermediate-point**]]

- To configure the transmission interval for the MEP to send CCMs, use the following command:

  `configure cfm domain` *domain_name* **association** *association_name* {**ports** *port_list* **end-point** [**up** | **down**]} **transmit-interval** [**3**|**10**|**100**|**1000**|**10000**|**60000**|**600000**]

- To unconfigure the transmission interval for the MEP to send CCMs and return it to the default, use the following command:

  `unconfigure cfm domain` *domain_name* **association** *association_name* {**ports** *port_list* **end-point** [**up** | **down**]} **transmit-interval**

- To enable of disable a MEP, use the following command:

  `configure cfm domain` *domain_name* **association** *association_name* **ports** *port_list* **end-point** [**up** | **down**] [**enable** | **disable**]

#### Displaying EAPS MEP Group Bindings

- Display EAPS MEP group bindings with the command: `show eaps cfm groups`

```
X480-48t.2 # sh eaps cfm groups
-------------------------------------------------------------------------
MEP Group Name                  Status Port   MEP ID
-------------------------------------------------------------------------
eapsCfmGrp1                     Up     41     11
eapsCfmGrp2                     Up     31     12
```

#### Displaying EAPS Output Change

- Display EAPS output changes using the command `show eaps`

  The existing output places a ! next to a CFM monitored ring port if the CFM indicates the MEP group for that port is down.

```
X480-48t.1 # sh eaps
EAPS Enabled: Yes
EAPS Fast-Convergence: Off
EAPS Display Config Warnings: Off
EAPS Multicast Add Ring Ports: Off
EAPS Multicast Send IGMP Query: On
EAPS Multicast Temporary Flooding: Off
EAPS Multicast Temporary Flooding Duration: 15 sec
Number of EAPS instances: 1
# EAPS domain configuration :
-------------------------------------------------------------------------------
Domain      State        Mo  En  Pri    Sec    Control-Vlan VID   Count Prio
-------------------------------------------------------------------------------
d2          Failed       M   Y   !41     31    v2           (101 ) 1     N
-------------------------------------------------------------------------------
Flags : (!) CFM Down
```

### *Configuration Example*

Below is a sample configuration of CFM support in EAPS:

```
switch 1 # sh configuration cfm
#
# Module dot1ag configuration.
```

```
#
create cfm domain string "MD1" md-level 6
configure cfm domain "MD1" add association string "MD1v1" vlan "v1"
configure cfm domain "MD1" add association string "MD1v2" vlan "v2"
configure cfm domain "MD1" association "MD1v1" ports 17 add end-point down 6
configure cfm domain "MD1" association "MD1v1" ports 23 add end-point down 5
configure cfm domain "MD1" association "MD1v2" ports 31 add end-point down 13
configure cfm domain "MD1" association "MD1v1" ports 17 end-point down add group
"eapsCfmGrp1"
configure cfm domain "MD1" association "MD1v1" ports 23 end-point down add group
"eapsCfmGrp2"
configure cfm domain "MD1" association "MD1v2" ports 31 end-point down add group
"eapsCfmGrp3"
configure cfm group "eapsCfmGrp1" add rmep 2
configure cfm group "eapsCfmGrp2" add rmep 4
configure cfm group "eapsCfmGrp3" add rmep 12
switch 2 # sh configuration "eaps"s
#
# Module eaps configuration.
#
enable eaps
create eaps d1
configure eaps d1 mode transit
configure eaps d1 primary port 17
configure eaps d1 secondary port 23
enable eaps d1
create eaps d2
configure eaps d2 mode transit
configure eaps d2 primary port 31
configure eaps d2 secondary port 23
enable eaps d2
configure eaps d1 add control vlan v1
configure eaps d1 add protected vlan pv1
configure eaps d2 add control vlan v2
configure eaps d2 add protected vlan pv2
create eaps shared-port 23
configure eaps shared-port 23 mode partner
configure eaps shared-port 23 link-id 100
configure eaps cfm add group eapsCfmGrp1
configure eaps cfm add group eapsCfmGrp2
configure eaps cfm add group eapsCfmGrp3
```

*Limitations*

CFM PDU transmit intervals are limited by the supported limits of CFM module. Platforms that do not support CFM in hardware are limited to a minimum interval of 100 ms.

The maximum number of down MEPs is limited by the CFM module. This is as low as 32 MEPs in some platforms.

*Platforms Supported*

All ExtremeXOS platforms support this feature; however, not all platforms support hardware-based CFM.

Platforms with no hardware-based CFM support are limited to software-based CFM transmit intervals of 100 ms or higher. Hardware-based intervals can go as low as 3.3 ms.

Currently, only the x460 and E4G platforms support hardware-based CFM.

# ERPS



This chapter provides an overview to *ERPS (Ethernet Ring Protection Switching)*, and discusses various ERPS features. The chapter also offers configuration details, provides configuration examples, and shows you how to debug ERPS.

## ERPS Overview

The basic concept of G.8032/*ERPS* is that traffic may flow on all links of a ring network except on one link called the Ring Protection Link (RPL).

The RPL owner is the node that blocks the RPL, and the other node of the RPL is called the *RPL neighbor node*. All other nodes are called *non-RPL nodes*. When a link fails, the RPL owner unblocks the RPL to allow connectivity to the nodes in the ring. The G.8032/ERPS rings utilize a channel (dedicated path) for carrying their control traffic which is the R-APS messages (Ring Automatic Protection Switching).

The ring protection architecture relies on the existence of an APS protocol to coordinate ring protection actions around an Ethernet ring, as shown in the following figure.

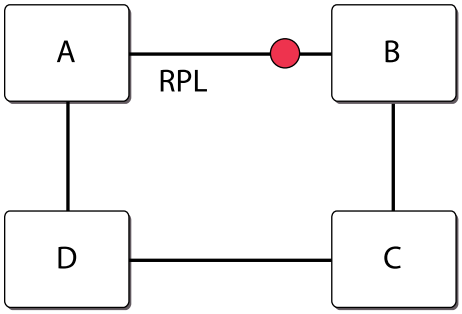



**Figure 137: Simple Ring with RPL, RPL Owner, RPL Neighbor, and Non-RPL Nodes**

More complex topologies include ladder ring networks which are called *sub-rings* in G.8032 terminology. In these networks, there could exist one or more rings and sub-rings which complete their

connectivity through the interconnected nodes of the ring(s). Multiple ladder networks are supported only if the following conditions are met:

- R-APS channels are not shared across Ethernet ring interconnections.
- On each ring port, each traffic channel and each R-APS channel are controlled by the Ethernet Ring Protection (ERP) Control process of only one Ethernet ring.
- Each major ring or sub-ring must have its own RPL.

Note

One important aspect of sub-rings is that they complete their channel through the virtual channel (when using the virtual channel mode), which can span the network and cross the sub-ring boundaries. This entails that the virtual channel is provisioned on all the nodes it spans across.

In the following figure, the ring comprises nodes A, B, C, and D with links A–B, B–C, C–D, and D–A while the control channel for this ring has its own dedicated *VLAN (Virtual LAN)*. The sub-ring consists of nodes D, F, E, and C with links D–F, F–E, and E–C. D and C are interconnected nodes. The channel for the sub-ring spans the links C–E, E–F, and F–D and their nodes while the virtual channel comprises the links D-A, A-B, B-C and D–C and their nodes. This means that the virtual channel for the sub-ring needs to not only exist on the interconnected nodes, but also on the nodes A and B.

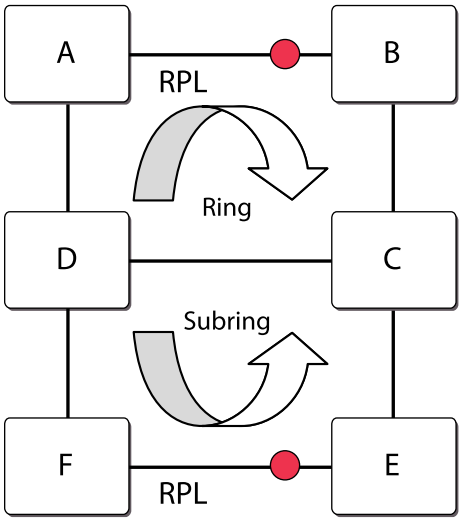



**Figure 138: Ring and Sub-ring Network**

When using G.8032 in networks, take care to design the virtual channel paths, since the VLAN provisioning has to exist on all the nodes through which the virtual channel can pass and which is solely dedicated to the sub-ring in question.

Sub-ring topology changes may impact flow forwarding over the domain of the other (interconnected) network, as such topology change events are signaled to the domain of the other network using the Topology Change signal.

## Supported ERPS Features

The following are the *ERPS* features supported in the current release:

- G.8032 version 1 support.
- G.8032 version 2 support with a restricted VC option.
- Revertive mode support for version 1 and 2.
- Basic interoperability with EAPS with G.8032 acting as an access ring. Flush notifications will be sent Link monitoring using CFM or native local link monitoring methods.
- Support for hardware accelerated CFM in specific platforms that have this capability.
- G.8032 version 2 with no Virtual Channel support.
- Support for attaching to a CFM DOWN-MEP configured external to ERPS.

## G.8032 Version 2

The concept of sub-rings is introduced to add multiple rings to the main ring. A sub-ring is an incomplete ring that completes its path through the main ring or other sub-rings. The control path for the sub-ring completes either through the implementation of a virtual channel, or by changing the flow of control packets in the sub-rings. Virtual channels are supported through the use of the sub-rings control channel being configured as a data *VLAN* in the main ring.

You can configure the sub-ring in “no virtual channel” mode, where the control path for the sub-ring is through all the nodes of the sub-ring (including the RPL owner and neighbor). You must be careful, however, to avoid using the sub-ring’s control channel across the main ring because that will cause a loop. ExtremeXOS supports the use of CFM, in conjunction with Manual Switch (MS), to protect the sub-rings against multiple failures in the main ring.

### CFM Link Monitoring

To enable CFM to report link events, the link must first be registered with CFM. *ERPS* acts as a client of CFM and creates the required Management Entity Points (MEPs). For G.8032 v1/v2 implementation:

- Creating a DOWN-MEP is by creating the DOWN-MEP with the CFM commands, and then assigning a group name to it. This group can then be associated to the ERPS ring.

Here is an example:

```
switch # sh  cfm
Domain: "erps_6",  MD Level: 6
   Association: "erps_MA_100", Destination MAC Type:  Multicast, VLAN "v2" with 2 cfm
ports
   Transmit Interval: 1000 ms
     port 27; Down End Point, mepid: 11, transmit-interval: 10000 ms (configured),

        MEP State: Enabled, CCM Message: Enabled,  Send SenderId TLV: Disabled
     port 37; Down End Point, mepid: 21,
        transmit-interval: 10000 ms (configured),
        MEP State: Enabled, CCM Message: Enabled, Send SenderId TLV: Disabled
   Association: "erps_MA_100", Destination MAC Type: Multicast, VLAN "v2" with 2 cfm
ports
   Transmit Interval: 1000 ms
Total Number of Domain             : 1
Total Number of Association        : 2
Total Number of Up MEP             : 0
```

```
Total Number of Down MEP           : 2
Total Number of MIP                : 0
Total Number of Number of CFM port : 4
Total Number of VPLS MIP(Static/Up): 0 / 0

switch # show cfm detail
Domain/       Port     MP  Remote End-Point  Remote End-Point   MEP   Life       Flags
Association                  MAC Address       IP Address       ID    time   Age
================================================================================
erps_6
  erps_MA_100  27      DE     00:04:96:34:e3:43  0.0.0.0        10    35000  4430   DM

               37      DE     00:04:96:27:fb:7b 0.0.0.0         20    35000  2790   DM
================================================================================
Maintenance Point: (UE) Up End-Point, (DE) Down End-Point
Flags: S - Static  Entry D - Dynamic Entry
           CCM Destination MAC: (U) Unicast (M)  Multicast
NOTE: The Domain and Association names are truncated to 13 characters, Lifetime and Age
are in milliseconds.
================================================================================

Total Number of Dynamic Up RMEP      : 0
Total Number of Dynamic Down RMEP    : 2
Total Number of Active Static RMEP   : 0
Total Number of Inactive Static RMEP : 0
```

**Note**

You must configure a remote MEP-ID for the local MEPs so that a specific association can be maintained between the two ends.

## Revertive and Non-revertive Mode

In the revertive mode, you can revert back to the RPL being blocked once the Signal Fault has cleared. In non-revertive mode, the SF remains blocked even after the fault clears. Reversion is handled in the following way:

- The reception of an R-APS No Request (NR) message causes the RPL owner node to start the wait-to-restore (WTR) timer.
- The WTR timer is cancelled if, during the WTR period, a request with a higher priority than NR is accepted by the RPL owner node, or is declared locally at the RPL owner node.
- When the WTR timer expires, without the presence of any other higher priority request, the RPL owner node initiates reversion by blocking its traffic channel over the RPL, transmitting an R-APS (NR, RB) message over both ring ports, informing the Ethernet ring that the RPL is blocked, and performing a flush FDB action. The *ERPS* Ring will be in the idle state.
- The acceptance of the R-APS (NR, RB) message causes all Ethernet ring nodes to unblock any blocked non-RPL link that does not have an SF condition. If it is an R-APS (NR, RB) message without a DNF indication, all Ethernet ring nodes perform a necessary flush *FDB (forwarding database)* action.

In non-revertive operation, the Ethernet ring does not automatically revert when all ring links and Ethernet ring nodes have recovered and no external requests are active. Non-revertive operation is handled in the following way:

- The RPL owner node does not generate a response on reception of an R-APS (NR) messages.
- When other healthy Ethernet ring nodes receive the NR (node ID) message, no action is taken in response to the message.

- When the operator issues a clear command for non-revertive mode at the RPL owner node, the non-revertive operation is cleared, the RPL owner node transmits an R-APS (NR, RB) message in both directions, repeatedly. The ERPS Ring will be in pending state.
- Upon receiving an R-APS (NR, RB) message, any blocking Ethernet ring node should unblock its non-failed ring port. If it is an R-APS (NR, RB) message without a DNF indication, all Ethernet ring nodes perform a necessary flush FDB action.

## Force Switch/Clearing

In the absence of any failure in the ring network, an operator-initiated Force Switch (FS) results in the RPL getting unblocked, and the node on which the FS has been issued is blocked. This condition is indicated by the transmission of R-APS FS messages, which are continuous until this condition is unconfigured. Two or more Forced Switches are allowed in the Ethernet ring, but this may cause the segmentation of an Ethernet ring. It is the responsibility of the operator to prevent this effect if it is undesirable.

You can remove a Forced Switch condition by issuing a clear command to the same Ethernet ring node where the Forced Switch is presented. The clear command removes existing local operator commands and triggers reversion in case the Ethernet ring is in revertive behavior. The Ethernet ring node where the Forced Switch was cleared continuously transmits the R-APS (NR) message on both ring ports, informing that no request is present at the Ethernet ring node.

## Manual Switch

Manual Switch is similar to the Force Switch except that only one Manual Switch is allowed for an Ethernet ring. The processing of which node retains the Manual Switch is based on the priority table and the node state. However only one Manual Switch is retained at the end for the ring.

Clearing the Manual Switch is done similar to the Force Switch.

### *Virtual Channel for Sub-rings*

While the standard describes how the sub-rings can function with a virtual channel, in this implementation sub-rings will function only with the presence of virtual channels.

### *Channel Blocking*

The R-APS control channel is blocked, as is traffic on the blocked ports for the control traffic entering on one ring port and getting forwarded to the other ring port. However, locally generated or delivered control traffic on the blocked port is supported.

### *Traffic Blocking*

Traffic is always blocked for the protected VLANs on the blocked ports of the ring/sub-ring in a G.8032 network.

*Signal Failure and Recovery*

In the absence of a higher priority request in the node, the following Signal Failure (SF) actions are taken.

- An Ethernet ring node detecting an SF condition on one of its ring ports blocks the traffic channel and R-APS channel on the failed ring port.
- The Ethernet ring node detecting an SF condition transmits an R-APS message indicating SF on both ring ports. The R-APS (SF) message informs other Ethernet ring nodes of the SF condition. R-APS (SF) messages are continuously transmitted by the Ethernet ring node detecting the SF condition while this condition persists. (The Periodic timer determines the interval of sending the SF after the first three.) For sub-ring interconnection nodes, the R-APS (SF) message is transmitted on the R-APS channel of the Sub-Ring port.
- Assuming the Ethernet ring node was in an idle state before the SF condition occurred, upon detection of this SF condition the Ethernet ring node triggers a local *FDB* flush.
- An Ethernet ring node accepting an R-APS (SF) message unblocks any blocked ring port that does not have an SF condition. This action unblocks the traffic channel on the RPL.
- An Ethernet ring node accepting an R-APS (SF) message stops transmission of other R-APS messages.
- An Ethernet ring node accepting an R-APS (SF) message without a DNF indication performs a flush FDB.

An Ethernet ring node that has one or more ring ports in an SF condition (upon detection of clearance of the SF condition) keeps at least one of these ring ports blocked for the traffic channel and for the R-APS channel, until the RPL is blocked as a result of Ethernet ring protection reversion, or until there is another higher priority request (for example, an SF condition) in the Ethernet ring. An Ethernet ring node that has one ring port in an SF condition, and detects clearing of this SF condition, continuously transmits the R-APS (NR) message with its own Node ID as the priority information over both ring ports, informing that no request is present at the Ethernet ring node and initiates a guard timer as described in sub-clause 10.1.5. Another recovered Ethernet ring node (or Nodes) holding the link block receives the message and compares the Node ID information with its own Node ID. If the received R-APS (NR) message has the higher priority, the Ethernet ring node unblocks its ring ports. Otherwise, the block remains unchanged. There is only one link with one-end block. The Ethernet ring nodes stop transmitting R-APS (NR) messages when they accept an R-APS (NR, RB), or when another higher priority request is received

## Timers

This section discusses the various timers associated with *ERPS*.

*Guard Timer*

The guard timer is used to prevent Ethernet ring nodes from acting upon outdated R-APS messages, and to prevent the possibility of forming a closed loop. The guard timer is activated whenever an Ethernet ring node receives an indication that a local switching request has cleared (i.e., local clear SF, clear). The guard timer can be configured in 10 ms steps, between 10 ms and two seconds, with a default value of 500 ms. This timer period should be greater than the maximum expected forwarding delay in which an R-APS message traverses the entire ring. The longer the period on the guard timer, the longer an Ethernet ring node is unaware of new or existing relevant requests transmitted from other Ethernet ring nodes, and is unable to react to them.

A guard timer is used in every Ethernet ring node. Once a guard timer is started, it expires by itself. While the guard timer is running, any received R-APS Request/State and Status information is blocked and not forwarded to the Priority Logic. When the guard timer is not running, the R-APS Request/State and Status information is forwarded unchanged.

*Hold-off Timer*

W hen a new defect, or more severe defect occurs (new SF), this event is not be reported immediately to protection switching if the provisioned hold-off timer is a non-zero value. Instead, the hold-off timer is started. When the hold-off timer expires, the trail that started the timer is checked to see if a defect still exists. If one does exist, that defect is reported to protection switching. The suggested range of the hold-off timer is 0 to 10 seconds in steps of 100 ms with an accuracy of ±5 ms. The default value for a hold-off timer is 0 seconds.

*Delay Timers*

In revertive mode, the wait-to-restore (WTR) timer is used to prevent frequent operation of the protection switching caused by intermittent signal failure defects. The wait-to-block (WTB) timer is used when clearing Forced Switch and Manual Switch commands. As multiple Forced Switch commands are allowed to coexist in an Ethernet ring, the WTB timer ensures that clearing of a single Forced Switch command does not trigger the re-blocking of the RPL. When clearing a Manual Switch command, the WTB timer prevents the formation of a closed loop due to a possible timing anomaly where the RPL owner node receives an outdated remote MS request during the recovery process.

## Sample Configuration

Here is a sample configuration of the *ERPS* feature:

```
create vlan cv1
config vlan cv1 tag 10
config vlan cv1 add port 5,6 tagged

create vlan pv1
config vlan pv1 tag 1000
config vlan pv1 add port 1
config vlan pv1 add port 5,6 tagged

create erps ring1
configure erps ring1 add ring-ports east 5
configure erps ring1 add ring-ports west 6
configure erps ring1 add control “cv1”
configure erps ring1 add protected vlan “pv1”
configure erps ring1 add protection-port 5
configure erps ring1 revert enabled wait-to-restore 500
enable erps r1
enable erps
```

*CFM DOWN-MEP Configuration to Provide Link Monitoring/Notifications*

```
create cfm domain string "MD3" md-level 3
configure cfm domain "MD3" add association string "MD3vsub1" vlan "vsub1"
configure cfm domain "MD3" association "MD3vsub1" ports 20 add end-point down 14
configure cfm domain "MD3" association "MD3vsub1" ports 24 add end-point down 13
configure cfm domain "MD3" association "MD3vsub1" ports 20 end-point down add group
"erpsDn1"
configure cfm domain "MD3" association "MD3vsub1" ports 24 end-point down add group
"erpsDn2"
configure cfm group "erpsDn1" add rmep 15
```

```
configure cfm group "erpsDn2" add rmep 12
configure erps subring1 cfm port east add group erpsDn2
configure erps subring1 cfm port west add group erpsDn1
```

*Sub-ring Configuration*

First, configure a main ring on the Interconnected node:

```
create erps main-ring1
configure erps main-ring1 add ring-ports east 5
configure erps main-ring1 add ring-ports west 6
configure erps ring1 add control “cv1”
```

Next, configure a sub-ring on the interconnected node:

```
create erps sub-ring1
configure erps sub-ring1 add ring-ports east 10
configure erps sub-ring1 add control “subv1”
configure erps main-ring1 add sub-ring sub-ring1
enable erps main-ring1
enable erps sub-ring1
```

*Virtual Channel for Sub-ring*

```
configure vlan subv1 add port 5 6 tagged
configure main-ring1 add protected vlan subv1
```

*No Virtual Channel for Sub-ring*

```
configure erps subring1 subring-mode no-virtualChannel
```

*Sub-ring Protection using UP MEP*

```
create cfm domain string "ERPS-UP" md-level 4
configure cfm domain "ERPS-UP" add association string "ERPS-UP-cfmVlan" vlan "cfmVlan"
configure cfm domain "ERPS-UP" association "ERPS-UP-cfmVlan" ports 24 add end-point up
21
configure cfm domain "ERPS-UP" association "ERPS-UP-cfmVlan" ports 24 end-point up add
group "erpsUp1"
configure cfm group "erpsUp1" add rmep 22
```

# Configuring ERPS

## ERPS Version 1 Commands

- To create or delete an *ERPS* ring, use the following commands:

  `create erps` *`ring-name`*

  `delete erps` *`ring-name`*

- To add or delete a control *VLAN* on the ERPS ring, use the following commands:

  `configure erps` *`ring-name`* **`add control`** {**`vlan`**} *`vlan_name`*

  `configure erps` *`ring-name`* **`delete control`** {**`vlan`**} *`vlan_name`*

- To add or delete a protected VLAN on the ERPS ring, use the following commands:

  `configure erps` *`ring-name`* **`add protected`** {**`vlan`**} *`vlan_name`*

```
configure erps ring-name delete protected {vlan} vlan_name
```

- To add ring ports on the ERPS ring, use the following command:

  ```
  configure erps ring-name ring-ports [east | west] port
  ```

- To delete ring ports on the ERPS ring, use the following command:

  ```
  unconfigure erps ring-name ring-ports west
  ```

- To add or delete RPL (ring protection link) owner configuration for the ERPS ring, use the following commands:

  ```
  configure erps ring-name protection-port port
  ```

  ```
  unconfigure erps ring-name protection-port
  ```

- To add or delete RPL (ring protection link) neighbor configuration for the ERPS ring, use the following commands:

  ```
  configure erps ring-name neighbor-port port
  ```

  ```
  unconfigure erps ring-name neighbor-port
  ```

- To add or delete ERPS revert operation along with the wait-to-restore time interval, use the following commands:

  ```
  configure {erps} ring-name revert [ enable | disable ]
  ```

- To configure the periodic timer, use the following command:

  ```
  configure {erps} ring-name timer periodic [ default | milliseconds ]
  ```

- To configure the guard timer, use the following command:

  ```
  configure {erps} ring-name timer guard [ default | milliseconds ]
  ```

- To configure the hold-off timer, use the following command:

  ```
  configure {erps} ring-name timer hold-off [ default | milliseconds ]
  ```

- To configure the wait-to-restore timer, use the following command:

  ```
  configure {erps} ring-name timer wait-to-restore [ default |
  milliseconds ]
  ```

- To associate and disassociate fault monitoring entities on the ERPS ring ports, use the following commands:

  ```
  configure erps ring-name cfm md-level level
  ```

  ```
  unconfigure {erps} ring-name cfm
  ```

- To rename the ERPS ring/sub-ring, use the following command:

  ```
  configure erps old-ring-name name new-ring-name
  ```

- To enable or disable ERPS, use the following commands:

  ```
  enable erps
  ```

  ```
  disable erps
  ```

- To enable or disable an existing ERPS ring/sub-ring, , use the following command:

  ```
  enable erps ring-name
  ```

  ```
  disable erps ring-name
  ```

- Run or clear force and manual switch triggers to the ERPS ring/sub-ring.

  `configure erps` *`ring-name`* **`dynamic-state`** [**`force-switch`** | **`manual-switch`** | **`clear`**] **`port`** *`slot:port`*
- To display global information for ERPS, use the following command:

  `show erps`
- To display specific details about an ERPS ring, use the following command:

  `show erps` *`ring-name`*
- To display ERPS statistics, use the following command:

  `show erps` *`ring-name`* **`statistics`**
- To clear statistics on an ERPS ring, use the following command:

  `clear counters erps` *`ring-name`*
- To debug ERPS, use the following commands:

  `debug erps` [**`options`**]

  `debug erps show` *`ring-name`*

## ERPS Version 2 Commands

- To set the rings to which to propagate topology change events, use the following command:

  `configure erps` *`ring-name`* [**`add`** | **`delete`**] **`topology-change`** *`ring-list`*
- To enable or disable the ability of *ERPS* to allow the topology-change bit to be set (to send out Flush events), , use the following commands:

  `enable erps` *`ring-name`* **`topology-change`**

  `disable erps` *`ring-name`* **`topology-change`**
- To add or delete a sub-ring to the main ring, use the following command:

  `configure` {**`erps`**} *`ring-name`* [**`add`** | **`delete`**] **`sub-ring-name`** *`sub_ring`*
- To configure the wait-to-block timer, use the following command:

  `configure` {**`erps`**} *`ring-name`* **`timer wait-to-block`** [ **`default`** | *`milliseconds`*]
- To add or delete an ERPS sub-ring to the EAPS domain, use the following commands:

  `configure` {**`erps`**} *`ring-name`* **`notify-topology-change`** {**`eaps`**} *`domain_name`*

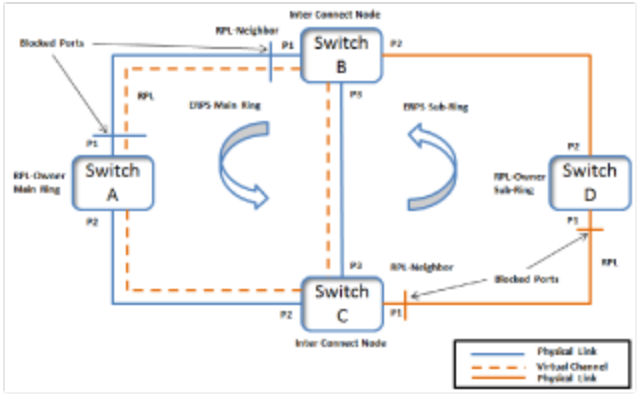
  `unconfigure` {**`erps`**} *`ring-name`* **`notify-topology-change`** {**`eaps`**} *`domain_name`*
- To configure a wait-to-block timer, use the following command:

  `configure` {**`erps`**} *`ring-name`* **`timer wait-to-block`** [ **`default`** | *`milliseconds`*]
- To configure sub-ring mode, use the following command:

  `configure erps` *`ring_name`* **`subring-mode`** [**`no-virtualChannel`** | **`virtualChannel`**]

# Sample Configuration

The following is a sample *ERPS* configuration:




## Configurations of Switch A

```
#VLAN Configuration
create vlan c_vlan tag 10
create vlan c_vlan_sub tag 20
create vlan p_vlan tag 100
configure c_vlan add ports 1,2 tagged
configure c_vlan_sub add ports 1,2 tagged
configure p_vlan add ports 1,2 tagged
#CFM Down MEP configuration in ERPS main ring RPL-owner
create cfm domain string MD6 md-level 6
configure cfm domain MD6 add association string MDlevel6 vlan c_vlan
configure cfm domain MD6 association MDlevel6 ports 1 add end-point down 601
configure cfm domain MD6 association MDlevel6 ports 2 add end-point down 602
configure cfm domain MD6 association MDlevel6 ports 1 end-point down add group AB
configure cfm domain MD6 association MDlevel6 ports 2 end-point down add group AC
configure cfm group AB add rmep 603
configure cfm group AC add rmep 604
# ERPS Configuration
create erps main_ring
configure erps main_ring add control vlan c_vlan
configure erps main_ring ring-port east 1
configure erps main_ring ring-port west 2
configure erps main_ring protection-port 1
configure erps main_ring cfm port east add group AB
configure erps main_ring cfm port west add group AC
configure erps main_ring add protected vlan p_vlan
configure erps main_ring add protected vlan c_vlan_sub
enable erps
enable erps main_ring
```

## Configurations of Switch B

```
# VLAN Configuration
create vlan c_vlan tag 10
create vlan c_vlan_sub tag 20
create vlan p_vlan tag 100
configure c_vlan add ports 1,3 tagged
configure c_vlan_sub add ports 1,3,2 tagged
configure p_vlan add ports 1,3,2 tagged
# CFM Down MEP configuration for ERPS main ring
create cfm domain string MD6 md-level 6
configure cfm domain MD6 add association string MDlevel6 vlan c_vlan
configure cfm domain MD6 association MDlevel6 ports 1 add end-point down 603
```

```
configure cfm domain MD6 association MDlevel6 ports 3 add end-point down 402
configure cfm domain MD6 association MDlevel6 ports 1 end-point down add group BA
configure cfm domain MD6 association MDlevel6 ports 3 end-point down add group BC
configure cfm group BA add rmep 601
configure cfm group BC add rmep 404
# CFM Down MEP Configuration for ERPS sub-ring
create cfm domain string MD3 md-level 3
configure cfm domain MD3 add association string MDlevel3 vlan c_vlan_sub
configure cfm domain MD3 association MDlevel3 ports 2 add end-point down 303
configure cfm domain MD3 association MDlevel3 ports 2 end-point down add group BD
configure cfm group BD add rmep 301
# ERPS Configuration for main ring
create erps main_ring
configure erps main_ring add control vlan c_vlan
configure erps main_ring ring-port east 1
configure erps main_ring ring-port west 3
configure erps main_ring neighbor-port 1
configure erps main_ring cfm port east add group BA
configure erps main_ring cfm port west add group BC
configure erps main_ring add protected vlan p_vlan
configure erps main_ring add protected vlan c_vlan_sub
enable erps
enable erps main_ring
# ERPS Configuration for sub-ring
create erps sub_ring
configure erps sub_ring add control vlan c_vlan_sub
configure erps sub_ring ring-port east 2
configure erps sub_ring cfm port east add group BD
configure erps sub_ring add protected vlan p_vlan
configure erps main_ring add sub-ring sub_ring
enable erps sub_ring
```

## Configurations of Switch C

```
# VLAN Configuration
create vlan c_vlan tag 10
create vlan c_vlan_sub tag 20
create vlan p_vlan tag 100
configure c_vlan add ports 2,3 tagged
configure c_vlan_sub add ports 2,3,1 tagged
configure p_vlan add ports 2,3,1 tagged
# CFM Down MEP configuration for ERPS main ring
create cfm domain string MD6 md-level 6
configure cfm domain MD6 add association string MDlevel6 vlan c_vlan
configure cfm domain MD6 association MDlevel6 ports 2 add end-point down 604
configure cfm domain MD6 association MDlevel6 ports 3 add end-point down 404
configure cfm domain MD6 association MDlevel6 ports 2 end-point down add group CA
configure cfm domain MD6 association MDlevel6 ports 3 end-point down add group CB
configure cfm group CA add rmep 602
configure cfm group CB add rmep 402
# CFM Down MEP configurations for ERPS sub-ring
create cfm domain string MD3 md-level 3
configure cfm domain MD3 add association string MDlevel3 vlan c_vlan_sub
configure cfm domain MD3 association MDlevel3 ports 1 add end-point down 304
configure cfm domain MD3 association MDlevel3 ports 1 end-point down add group CD
configure cfm group CD add rmep 302
# ERPS Configuration for main ring
create erps main_ring
configure erps main_ring add control vlan c_vlan
configure erps main_ring ring-port east 2
configure erps main_ring ring-port west 3
configure erps main_ring cfm port east add group CA
configure erps main_ring cfm port west add group CB
```

```
configure erps main_ring add protected vlan p_vlan
configure erps main_ring add protected vlan c_vlan_sub
enable erps
enable erps main_ring
# ERPS Configuration for sub-ring
create erps sub_ring
configure erps sub_ring add control vlan c_vlan_sub
configure erps sub_ring ring-port east 1
configure erps sub_ring neighbor-port 1
configure erps sub_ring cfm port east add group CD
configure erps sub_ring add protected vlan p_vlan
configure erps main_ring add sub-ring sub_ring
enable erps sub_ring
```

## Configurations of Switch D

```
# VLAN Configuratio
ncreate vlan c_vlan_sub tag 20
create vlan p_vlan tag 100
configure c_vlan_sub add ports 2,1 tagged
configure p_vlan add ports 2,1 tagged
# CFM Down MEP configurations for ERPS sub-ring
create cfm domain string MD3 md-level 3
configure cfm domain MD3 add association string MDlevel3 vlan c_vlan_sub
configure cfm domain MD3 association MDlevel3 ports 2 add end-point down 301
configure cfm domain MD3 association MDlevel3 ports 1 add end-point down 302
configure cfm domain MD3 association MDlevel3 ports 2 end-point down add group d5d2
configure cfm domain MD3 association MDlevel3 ports 1 end-point down add group d5d4
configure cfm group d5d2 add rmep 303
configure cfm group d5d4 add rmep 304
# ERPS Configuration for sub-ring
create erps sub_ring
configure erps sub_ring add control vlan c_vlan_sub
configure erps sub_ring ring-port east 2
configure erps sub_ring ring-port west 1
configure erps sub_ring protection-port 1
configure erps sub_ring cfm port east add group d5d2
configure erps sub_ring cfm port west add group d5d4
configure erps sub_ring add protected vlan p_vlan
enable erps
enable erps sub_ring
```

**Note**

ERPS Virtual channel is enabled in the above configuration by creating a control vlan of the sub-ring [c_vlan_sub] and added as a protected vlan in switch A, B and C's main ring.

# Debugging ERPS

1. Check the output of `show erps` *`ring`* **`statistics`** to see if any error/dropped counters are incrementing.
   a. If they are, check the state of the ring ports and trace these links to the neighbor node to see the state of the links.

      The output of `show log` after turning on the filters for *ERPS* should provide more information on what is happening on the switch.
2. Check the output of `show erps` and `show erps` *`ring`* to see if the node state is as expected.

   In steady state, the node should be in "Idle" and the failed state ring should be in "Protected" state.

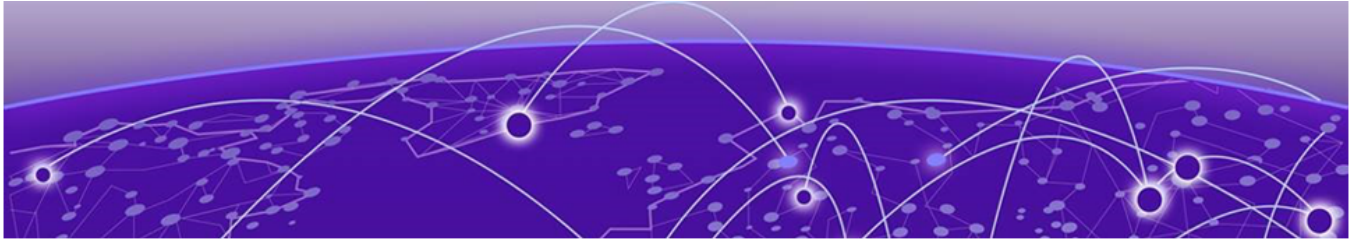
## ERPS Feature Limitations

The following are *ERPS* feature limitations:

- Backup MSM Failover and checkpointing for both v1 and v2 are not available in the current release.
- In platforms that do not have hardware OAM (operations and management), the optimum CFM interval recommended is one second for link monitoring, which will give rise to approximately three-second overhead in convergence times.

Note
For optimum performance and convergence, it is recommended to use fiber cables.

- Other than the basic EAPS interoperability stated above, all other EAPS related interoperability is not supported.
- There is no interoperability with *STP (Spanning Tree Protocol)* in the current release.
- *SNMP (Simple Network Management Protocol)* is not supported in the current release.



# STP



Using the Spanning Tree Protocol (STP) functionality of the switch makes your network more fault tolerant. This chapter explains more about STP and the STP features supported by ExtremeXOS.

Note
STP is a part of the 802.1D bridge specification defined by the IEEE Computer Society. To explain STP in terms used by the IEEE 802.1D specification, the switch will be referred to as a *bridge*.

ExtremeXOS version 12.0 and later supports the new edition of the IEEE 802.1D standard (known as IEEE 802.1D-2004 ) for STP, which incorporates enhancements from the IEEE 802.1t-2001, IEEE 802.1W, and IEEE 802.1y standards. The IEEE 802.1D-2004 standard is backward compatible with the IEEE 802.1D-1998 standard. For more information, see Compatibility Between IEEE 802.1D-1998 and IEEE 802.1D-2004 STP Bridges on page 1033.

## Spanning Tree Protocol Overview

*STP (Spanning Tree Protocol)* is a bridge-based mechanism for providing fault tolerance on networks.

STP allows you to implement parallel paths for network traffic and to ensure that redundant paths are:

- Disabled when the main paths are operational.
- Enabled if the main path fails.

Note
STP and *ESRP (Extreme Standby Router Protocol)* cannot be configured on the same *VLAN (Virtual LAN)* simultaneously.

**Note**
If you are transitioning from EOS to ExtremeXOS, please note that ExtremeXOS blocks on a more granular (VLAN) level, instead of at the port level as EOS does.

## Compatibility Between IEEE 802.1D-1998 and IEEE 802.1D-2004 STP Bridges

The IEEE 802.1D-2004 compliant bridges interoperate with the IEEE 802.1D-1998 compliant bridges.

To ensure seamless operation of your *STP* network, read this section before you configure STP on any Extreme Networks device running ExtremeXOS 11.6 or later.

Differences in behavior between the two standards include the:

- Default port path cost
- Bridge priority
- Port priority
- Edge port behavior

This section describes the bridge behavior differences in more detail.

### *Default Port Path Cost*

The 802.1D-2004 standard modified the default port path cost value to allow for higher link speeds.

A higher link speed can create a situation whereby an 802.1D-1998 compliant bridge could become the more favorable transit path.

For example, in the following figure, bridge A is the root bridge running the new 802.1D-2004 standard, bridges B and C are running the old 802.1D-1998 standard, and bridges D, E, and F are running the new 802.1D-2004 standard. In addition, all ports are 100 Mbps links. The ports on bridges B and C have a default path cost of 19, and the ports on bridge A, D, E, and F have a default path cost of 200,000.

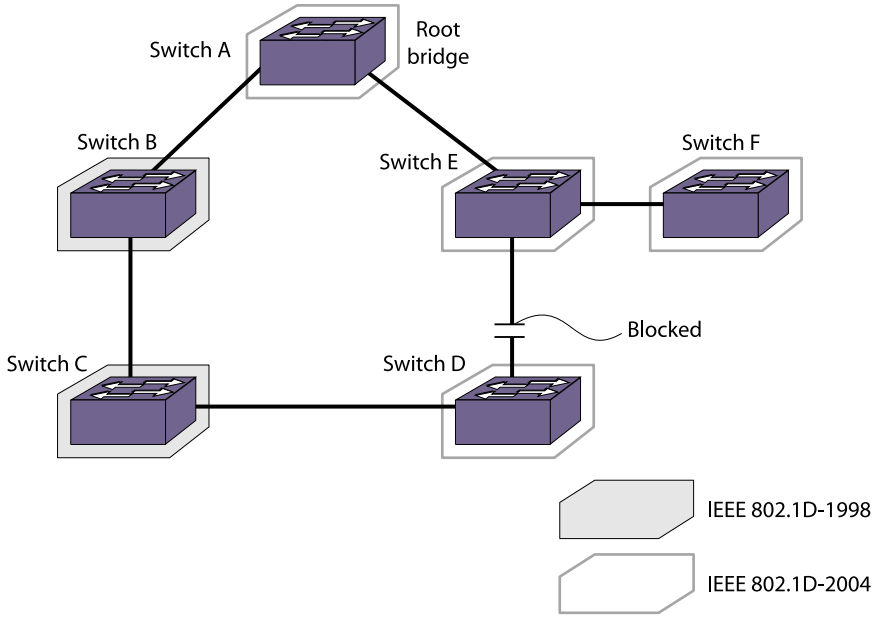



**Figure 139: 802.1D-1998 and 802.1D-2004 Mixed Bridge Topology**

If you use the default port path costs, bridge D blocks its port to bridge E, and all traffic between bridges D and E must traverse all of bridges in the network. Bridge D blocks its port to bridge E because the path cost to the root bridge is less by going across bridges B and C (with a combined root cost of 38) compared with going across bridge E (with a root cost of 200,000). In fact, if there were 100 bridges between bridges B, C, and D running the old 802.1D-1998 standard with the default port path costs, bridge D would still use that path because the path cost is still higher going across bridge E.

As a workaround and to prevent this situation, configure the port path cost to make links with the same speed use the same path host value. In the example described above, configure the port path cost for the 802.1D-2004 compliant bridges (bridges A, D, E, and F) to 19.

Note
You cannot configure the port path cost on bridges B and C to 200,000 because the path cost range setting for 802.1D-1998 compliant bridges is 1 to 65,535.

To configure the port path cost, use the following command:

```
configure stpd stpd_name ports cost [auto | cost] port_list
```

*Bridge Priority*

By configuring the *STPD (Spanning Tree Domain)* bridge priority, you make the bridge more or less likely to become the root bridge.

Unlike the 802.1D-1998 standard, the 802.1D-2004 standard restricts the bridge priority to a 16-bit number that must be a multiple of 4,096. The new priority range is 0 to 61,440 and is subject to the multiple of 4,096 restriction. The old priority range was 0 to 65,535 and was not subject to the multiple

of 4,096 restriction (except for *MSTP (Multiple Spanning Tree Protocol)* configurations). The default bridge priority remains the same at 32,768.

If you have an ExtremeXOS 11.5 or earlier configuration that contains an *STP* or RSTP bridge priority that is not a multiple of 4,096, the switch rejects the entry and the bridge priority returns to the default value while loading the structure. The MSTP implementation in ExtremeXOS already uses multiples of 4,096 to determine the bridge priority.

To configure the bridge priority, use the following command:

`configure stpd` *`stpd_name`* **`priority`** *`priority`*

For example, to lower the numerical value of the priority (which gives the priority a higher precedence), you subtract 4,096 from the default priority: 32,768 - 4,096 = 28,672. If you modify the priority by a value other than 4,096, the switch automatically changes the priority to the lower priority value. For example, if you configure a priority of 31,000, the switch automatically changes the priority to 28,672.

### *Port Priority*

The port priority value is always paired with the port number to make up the 16-bit port identifier, which is used in various *STP* operations and the STP state machines.

Unlike the 802.1D-1998 standard, the 802.1D-2004 standard uses only the four most significant bits for the port priority and it must be a multiple of 16. The new priority range available is 0 to 240 and is subject to the multiple of 16 restriction. The 802.1D-1998 standard uses the eight most significant bits for the port priority. The old priority range was 0 to 31 and was not subject to the multiple of 16 restriction.

To preserve backward compatibility and to use ExtremeXOS 11.5 or earlier configurations, the existing `configure stpd ports priority` command is available. If you have an ExtremeXOS 11.5 or earlier configuration, the switch interprets the port priority based on the 802.1D-1998 standard. If the switch reads a value that is not supported in ExtremeXOS 11.6 or later, the switch rejects the entry.

When you save the port priority value, the switch saves it as the command `configure stpd ports port-priority` with the corresponding change in value.

For example, if the switch reads the `configure stpd ports priority 16` command from an ExtremeXOS 11.5 or earlier configuration, (which is equivalent to the command `configure stpd ports priority 8` entered through CLI), the switch saves the value as `configure stpd ports port-priority 128`.

### *Edge Port Behavior*

In ExtremeXOS 11.5 or earlier, Extreme Networks had two edge port implementations: edge port and edge port with safeguard.

The 802.1D-2004 standard has a bridge detection state machine, which introduced a third implementation of edge port behavior. The following list describes the behaviors of the different edge port implementations:

- Edge port (ExtremeXOS 11.5 and earlier):
  - The port does not send bridge protocol data units (BPDUs).
  - The port does not run a state machine.

- If BPDUs are received, the port discards the BPDU and enters the blocking state.
- If subsequent BPDUs are not received, the port remains in the forwarding state.

- Edge port with safeguard configured (ExtremeXOS 11.5 and 11.4 only):
  - The port sends BPDUs.
  - When configured for *MSTP*, the port runs a partial state machine.
  - If BPDUs are received, the port enters the blocking state.
  - If subsequent BPDUs are not received, the port attempts to enter the forwarding state.
- Edge port running 802.1D-2004 with safeguard enabled:
  - The port sends BPDUs.
  - The port runs a state machine.
  - If BPDUs are received, the port behaves as a normal RSTP port by entering the forwarding state and participating in RSTP.
  - If subsequent BPDUs are not received, the port attempts to become the edge port again.

Edge port with safeguard prevents accidental or deliberate misconfigurations (loops) by having edge ports enter the blocking state upon receiving a BPDU. The 802.1D-2004 standard implements a bridge detection mechanism that causes an edge port to transition to a non-edge port upon receiving a BPDU; however, if the former edge port does not receive any subsequent BPDUs during a pre-determined interval, the port attempts to become an edge port.

If an 802.1D-2004 compliant safeguard port (edge port) connects to an 802.1D-1998 compliant edge port with safeguard configured, the old safeguard port enters the blocking state. Although the new safeguard port becomes a designated port, the link is not complete (and thus no loop is formed) because one side of the link is blocked.

*Restricted Role*

In a large metro environment, to prevent external bridges from influencing the spanning tree active topology, the following commands have been introduced for Rapid Spanning Tree Protocol (RSTP) and *MSTP*.

- `configure stpd` *`stpd_name`* **`ports restricted-role enable`** *`port_list`*
  - This command enables restricted role on a specified port in the core network to prevent external bridges from influencing the spanning tree active topology.
  - Restricted role should not be enabled with edge mode.
  - stpd_name—Specifies an *STPD* name on the switch.
  - port_list—Specifies one or more ports or slots and ports.
  - Enabling restricted role causes a port to not be selected as a root port, even if it has the best spanning tree priority vector. Such a port is selected as an alternate port after the root port is selected. The restricted role is disabled by default. If set, it can cause a lack of spanning tree connectivity.
  - A network administrator enables restricted role to prevent external bridges from influencing the spanning tree active topology.
- `configure stpd` *`stpd_name`* **`ports restricted-role disable`** *`port_list`*
  - This command disables restricted role on a specified port in the core network.
  - stpd_name—Specifies an STPD name on the switch.

- port_list—Specifies one or more ports or slots and ports.
- Restricted role is disabled by default. If set, it can cause a lack of spanning tree connectivity. A network administrator enables restricted role to prevent external bridges from influencing the spanning tree active topology.

## Loop Protect

*STP* depends on continuous reception of Type 2 BPDUs (RSTP/*MSTP*) based on the port role. The designated port transmits BPDUs and the non-designated port receives BPDUs. If one of the ports in a physical redundant topology no longer receives BPDUs, then STP assumes that the topology is loop free. This leads the blocking port from the alternate or backup port becomes designated and moves to a forwarding State causing a loop.

Loop Protect protects the network from loops. The Loop Protect feature is achieved by ports receiving BPDUs (RSTP/MSTP only) on point-to-point ISLs before their states are allowed to become forwarding. Further, if a BPDU timeout occurs on a port, its state becomes listening until a new BPDU is received. In this way, both upstream and downstream facing ports are protected. When a root or alternate port loses its path to the root bridge, due to message age expiration, it takes on the role of designated port and will not forward traffic until a BPDU is received. When a port is intended to be the designated port in an ISL, it constantly proposes and will not forward until a BPDU is received. It will revert to listening if it stops getting a response. Loop Protect also overrides the port admin setting. This protects against misconfiguration (such as disabling STP on a port) and protocol failure by the connected bridge.

Loop Protect has the capability to

- Control port forwarding state based on reception of agreement BPDUs
- Control port forwarding state based on reception of disputed BPDUs
- Disable a port based on frequency of failure events
- Communicate port non-forwarding status through traps.

**Note**
Loop Protect Traps are not supported in this ExtremeXOS 15.7.

By default, Loop Protect is disabled on all ports.

### *Loop Protect Port Modes*

Ports work in two Loop Protect operational modes.

- If the port has the partner loop protect as capable then it works in full functional mode.
- If the port has the partner loop protect as incapable then it works limited functional mode.

In full mode, when RSTP/*MSTP* BPDUs is received in point-to-point link and the port is designated, a Loop Protect timer is set to 3 times hello time, when this timer expires then port will be moved to blocking state. Limited mode adds a further requirement that the flags field in the BPDU indicates a root role.

Message age expiration and the expiration of the Loop Protect timer are both events for which Loop Protect generates traps and a debug message. In addition, user can configure Loop Protect to forcefully disable port when one or more events occur. When the configured number of events happens within a given window of time, the port will be forced into disable and held there until you manually unlock it.

The following example shows the loop due to the misconfiguration in *STP*:

The following figure shows that Switch 1 is elected as Root. Switch 2 and Switch 3 elect the root port. Switch3's port connected to Switch2 is elected as Alternate port and it is port state is in blocking state.

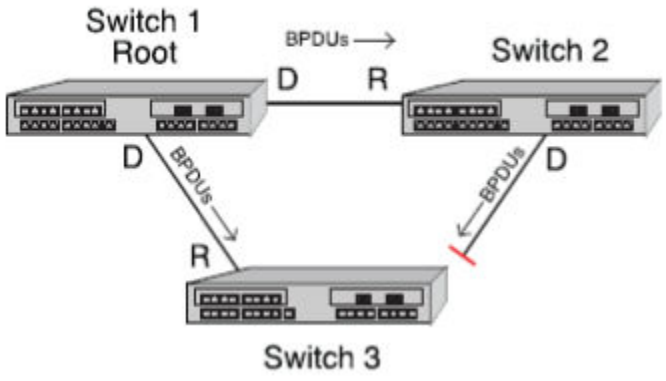



**Figure 140: Switch 1 Elected as Root**

The next figure shows that if the user accidentally disables the STP on Switch 2 port connected to Switch 3, Switch 2 will stop sending the BDPU to Switch 3 since STP is disabled. Switch 3 assumes that neighbor is down and it changes the port to forwarding state which will eventually create a loop.

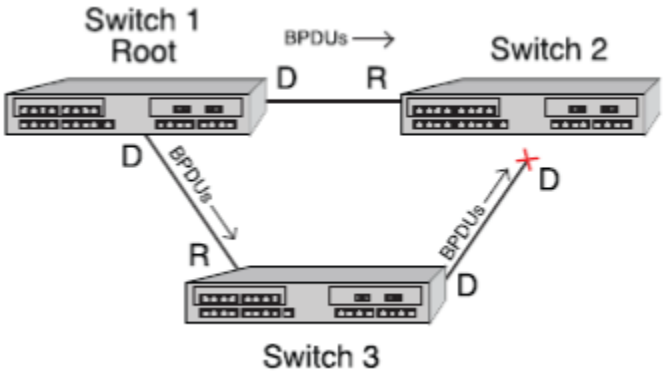



**Figure 141: STP Disabled on Switch 2**

The last figure shows that, with loop protect enabled switch 3 will not go to forwarding state until it receives a BPDU from switch 2 and the port state will be in discarding state.

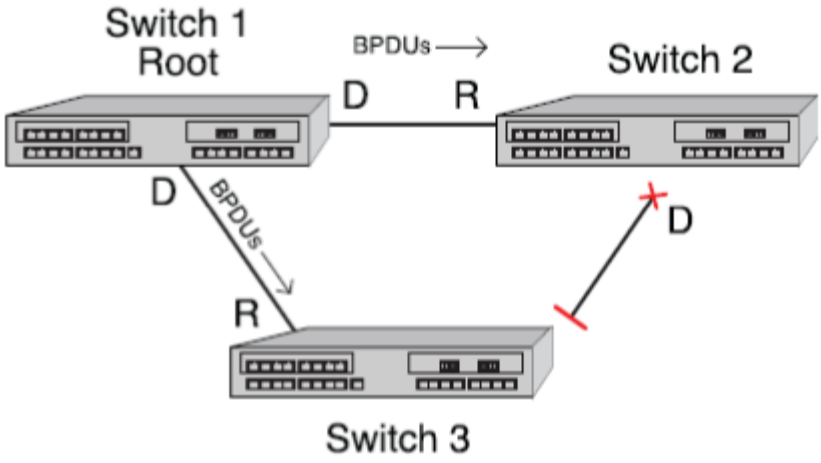



**Figure 142: Loop Protect Enabled**

When the Loop protect feature is enabled:

- On a Point-to-point Link, BPDU must be received before going to Forwarding state.
- If a BPDU timeout occurs on a port, its state becomes DISCARDING until a BPDU is received.
- When a root or alternate port loses its path to the root bridge due to a message age expiration it takes on the role of designated port. It will not forward traffic until a BPDU is received.
- When a port is intended to be the designated port in a point-to-point link it constantly proposes and will not forward until a BPDU is received, and will revert to discarding if it fails to get a response.
- If the partner is not Loop Protect Capable (Alternate Agreement not supported), designated port will not be allowed to forward unless receiving agreements from a port with root role.
- Legacy Spanning Tree (802.1d) or shared media devices should be connected in a non-redundant fashion to avoid the possibility of looping.

You can enable the port by giving the command `enable port` *`port-list`*.

### *STP Filter Configuration*

This new ExtremeXOS 15.7.1 CLI feature includes processing and forwarding of *STP* BPDU's either by system wide or port based filter installation.

In previous releases this processing and forwarding of STP BPDUs was based only on ports, now this has been changed to system wide (default). Prior to ExtremeXOS 15.7, blocked ports transmission of the BPDU's was restricted, but for the Loop Protect feature when any proposal agreement BPDU is received in the blocked port, then reply for the proposal agreement is to be transmitted. Due to this MAC movement is reported and the packet was not processed. This feature prevents this issue.

You can roll back to the previous implementation (port-based) if any issues are encountered .

## Backup Root

When the root bridge of spanning tree instance is lost, its information may be retained in the network until the aging mechanism causes it to be removed. This leads to a delay in convergence on for the new

root. This backup root feature is used to get faster convergence when the root bridge connectivity is lost.

Backup root feature enabled bridge port should be connected to Root with point to point link. When backup root bridge lost contact with the root bridge, automatically backup root bridge lowers its bridge priority (an increment below the priority of the lost root). This will cause the backup root bridge to become the new root with information superior to any information held by any bridge in the network. While save and reboot automatically changed priority will not be saved and restored. At the time of reboots it will have its priority restored to the original configured value by administrator.

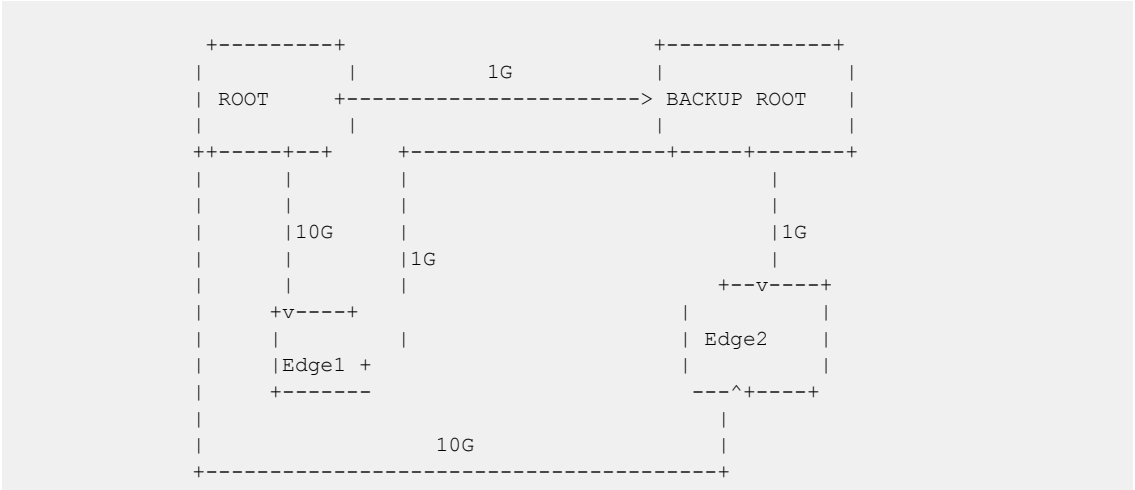
If the priority of the root bridge is Zero and backup root has lost the connectivity of the root bridge then automatic assignment of the priority value for the backup root will be the initial configured value.

This backup root feature is activated only when there is loss of connectivity with the root bridge. Raising the priority on the root will not cause the Backup root feature to be activated.

*Configuration Recommendations:*

- Enable the Backup Root feature on both the root and backup root.
- Configure all bridges except the root and backup root with the maximum bridge priority value (61440 with 802.1t).
- Configure the root and backup root to have the next lowest priority (57344 with 802.1t).
- We also recommend having multiple links between the root and backup root. This helps prevent the Backup Root feature from activating due to a simple link failure rather than a bridge failure.
- Care must be taken when this feature is used as it may result in sub-optimal traffic forwarding paths.

  Consider the following topology: 10G link connected from the edge switches to the aggregation Root while the edge switches to the Backup Root switch with 1G links to reduce the cost and acts as a backup link. In this case, when the link between aggregation Root and Backup Root fails, the traffic will now use the 1G links providing reduced performance of the network. The aggregation old root is still in the network which can be used to forward the traffic. This feature has moved the root of the stp domain.



## Multisource Detection

Multisource Detection is a feature that prevents network disruption due to excessive topology changes caused by a full duplex port transmitting multiple BPDUs with different source MAC addresses, and different BPDU information. When a port is point-to-point, the received priority information comes from the most recently received BPDU. When a port is non-point-to-point, the received information reflects the best priority information out of all the received BPDUs. Typical scenarios for multisource detection are when a switch is connected to a device which has been improperly configured to forward received BPDUs out on other ports or has been configured to not run the Spanning Tree protocol and treats BPDUs as multicast packets by transmitting them out all other forwarding ports. In these situations, the connected port is acting as a shared media device. The way to detect shared media is the duplex setting. Since the port is full duplex it treats the connection as point-to-point.

When Loop Protect is configured for the port, if multisource detection is triggered, the port will go to the listening state and no longer be part of the active topology. Loop protect does not operate on shared media ports.

## Restrict Topology Change Notification

Restricted Topology Change Notification (TCN) is a Spanning Tree Protocol feature that allows/disallows TCN propagation on specified ports. When Restricted TCN is disabled, TCN propagation is allowed. The port propagates received TCNs and topology changes to other ports.

Restricted TCN is disabled by default. When Restricted TCN is enabled, the port does not propagate received TCNs and topology changes to other ports. Enable Restricted TCN to prevent unnecessary address flushing in the core region of the network caused by activation of bridges external to the core network.

A possible reason for not allowing TCN propagation is when bridges are not under the full control of the administrator or because MAC operational state for the attached or downstream LANs transitions frequently, causing disruption throughout the network. Rapid Spanning Tree responds to TCNs by selectively flushing the filter database. Persistent TCNs are disruptive, causing persistent address flushing, which in turn causes increased flooding in the network.

Restricted TCN is a useful tool when it is not possible to remove the source of the TCNs.

## BPDU Dispute Threshold

The BPDU received from the peer is inferior with designated role and state as learning or forwarding. Since this condition could be caused by a unidirectional link failure, the interface is assigned to blocking sate and marked as disputed.

The disputed BPDU threshold is an integer variable that represents the number of disputed.

BPDUs must be received on a given port/*STP* domain until a disputed BPDU trap is sent. For example, if the threshold is 10, then a trap is issued when 10, 20, 30, etc. disputed BPDUs have been received.

If the threshold is configured as "none," a trap is not sent. The trap indicates port, STP domain and total Disputed BPDU count.

The default is none.

## BPDU Restrict on Edge Safeguard

BPDU restrict causes a port on which this feature is configured to be disabled as soon as an *STP* BPDU is received on that port, thus allowing you to enforce the STP domain borders and keep the active topology predictable.

The following figure shows a BPDU restrict example.

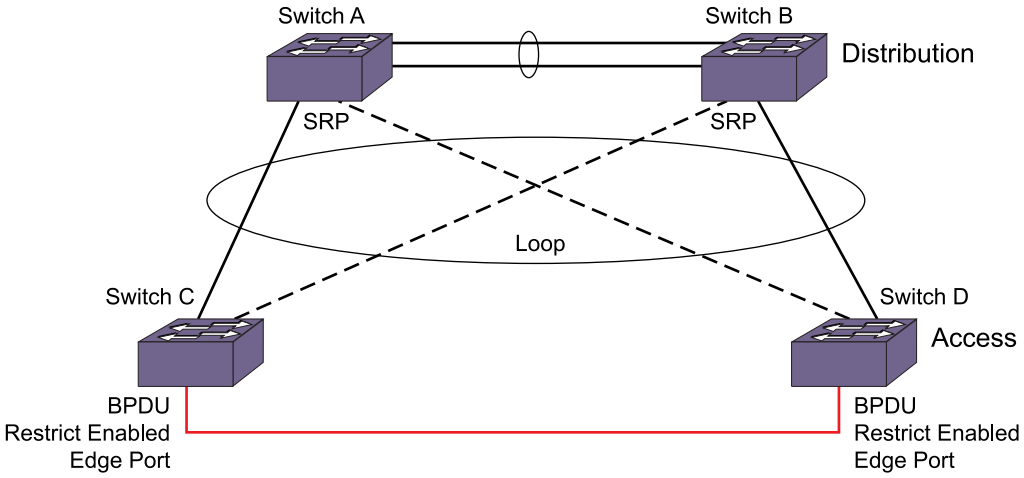



**Figure 143: BPDU Restrict**

In this figure, loops on the LAN access switches are not prevented since the ports towards the distribution switches are not running STP but Software Redundant Ports (SRP). Currently, ExtremeXOS software cannot run STP on ports that are configured for SRP. STP on the access switch is unaware of the alternate path and therefore cannot prevent the loop that exists across the switches. Configuring a port as an edge mode port alone cannot prevent the loop between the switches because edge ports never send BPDUs. The edge safeguard feature is not able to prevent the loops because STP does not have the information about the alternate path.

To prevent the loops across the switches, the edge safeguard feature can be configured with the BPDU restrict function. When running in BPDU restrict mode, edge safeguard ports send STP BPDUs at a rate of one very two seconds. The port is disabled as soon as an STP BPDU is received on the BPDU restrict port, thereby preventing the loop. Flexibility is provided with an option to re-enable the port after a user specified time period. If a user enables a port while STP has disabled it, the port is operationally enabled; STP is notified and then stops any recovery timeout that has started.

When an *STPD* is disabled for a BPDU restrict configured port, an STP port in 802.1D operation mode begins forwarding immediately, but in the RSTP or *MSTP* operation modes, the port remains in the disabled state.

BPDU restrict is available on all of the three operational modes of STP: 802.1D, RSTP, and MSTP.

Although edge safeguard is not available in 802.1D operation mode, when you configure BPDU restrict you do so in a similar way, that is, as an extension of edge safeguard; then only BPDU restrict is available on the port and not edge safeguard.

To configure BPDU restrict, use the command:

- `configure {stpd} stpd_name ports edge-safeguard enable port_list {bpdu-restrict} {recovery-timeout {seconds}}`

- BPDU restrict can also be configured by using the following commands:
  - `configure {stpd} stpd_name ports bpdu-restrict [enable | disable] port_list {recovery-timeout {seconds}}`
  - `configure stpd stpd_name ports link-type [[auto | broadcast | point-to-point] port_list | edge port_list {edge-safeguard [enable | disable] {bpdu-restrict} {recovery-timeout seconds}}]`

To include BPDU restrict functionality when configuring link types or edge safeguard, see Configuring Link Types on page 1061 and Configuring Edge Safeguard on page 1061.

The example below shows a BPDU restrict configuration:

```
* switch # configure s1 ports edge-safeguard enable 9 bpdu-restrict recovery-timeout 400.
```

The following is sample output from the `show s1 ports` command resulting from the configuration:

```
switch # show s1 ports
Port   Mode   State      Cost  Flags     Priority Port ID Designated Bridge
9      EMISTP FORWARDING 20000 eDee-w-G-- 128      8009    80:00:00:04:96:26:5f:4e
Total Ports: 1
------------------------ Flags: ----------------------------
1:                e=Enable, d=Disable
2: (Port role)    R=Root, D=Designated, A=Alternate, B=Backup, M=Master
3: (Config type)  b=broadcast, p=point-to-point, e=edge, a=auto
4: (Oper. type)   b=broadcast, p=point-to-point, e=edge
5:                p=proposing, a=agree
6: (partner mode) d = 802.1d, w = 802.1w, m = mstp
7:                i = edgeport inconsistency
8:                S = edgeport safe guard active
s = edgeport safe guard configured but inactive
8:                G = edgeport safe guard bpdu restrict active in 802.1w and mstp
g = edgeport safe guard bpdu restrict active in 802.1d
9:                B = Boundary, I = Internal
10:               r = Restricted Role
switch # show configuration stp
#
# Module stp configuration.
#
configure mstp region 000496265f4e
configure stpd s0 delete vlan default ports all
disable stpd s0 auto-bind vlan default
create stpd s1
configure stpd s1 mode dot1w
enable stpd s0 auto-bind vlan Default
configure stpd s1 add vlan v1 ports 9 emistp
configure stpd s1 ports mode emistp 9
configure stpd s1 ports cost auto 9
configure stpd s1 ports port-priority 128 9
configure stpd s1 ports link-type edge 9
configure stpd s1 ports edge-safeguard enable 9 recovery-timeout 400
configure stpd s1 ports bpdu-restrict enable 9 recovery-timeout 400
enable stpd s1 ports 9
configure stpd s1 tag 10
enable stpd s1
```

The following is sample output for STP operation mode dot1d from the `show configuration stp` command:

```
switch # show configuration stp
#
# Module stp configuration.
#
configure mstp region region2
configure stpd s0 delete vlan default ports all
disable stpd s0 auto-bind vlan default
create stpd s1
enable stpd s0 auto-bind vlan Default
configure stpd s1 add vlan v1 ports 9 emistp
configure stpd s1 ports mode emistp 9
configure stpd s1 ports cost auto 9
configure stpd s1 ports priority 16 9
configure stpd s1 ports link-type edge 9
configure stpd s1 ports edge-safeguard enable 9 recovery-timeout 400
configure stpd s1 ports bpdu-restrict enable 9 recovery-timeout 400
enable stpd s1 ports 9
configure stpd s1 tag 10
enable stpd s1
```

## Span Tree Domains

The switch can be partitioned into multiple virtual bridges. Each virtual bridge can run an independent Spanning Tree instance. Each Spanning Tree instance is called a *STPD*. Each STPD has its own root bridge and active path. After an STPD is created, one or more *VLAN*s can be assigned to it.

A physical port can belong to multiple STPDs. In addition, a VLAN can span multiple STPDs.

The key points to remember when configuring VLANs and *STP* are:

- Each VLAN forms an independent broadcast domain.
- STP blocks paths to create a loop-free environment.
- Within any given STPD, all VLANs belonging to it use the same spanning tree.
- To create an STPD, use the command:

  `create stpd stpd_name {description stpd-description}`
- To delete an STPD, use the command:

  `delete stpd stpd_name`

  User-created STPD names are not case-sensitive.

For detailed information about configuring STP and various STP parameters on the switch, see Configure STP on the Switch on page 1084.

### Member VLANs

When you add a *VLAN* to an *STPD*, that VLAN becomes a member of the STPD. The two types of member VLANs in an STPD are:

- Carrier
- Protected

### *Carrier VLAN*

A carrier VLAN defines the scope of the *STPD*, which includes the physical and logical ports that belong to the STPD and if configured, the 802.1Q tag used to transport Extreme Multiple Instance Spanning Tree Protocol (EMISTP) or Per VLAN Spanning Tree (PVST+) encapsulated bridge protocol data units (BPDUs).

See Encapsulation Modes on page 1047 for more information about encapsulating *STP* BPDUs.

Only one carrier VLAN can exist in a given STPD, although some of its ports can be outside the control of any STPD at the same time.

If you configure EMISTP or PVST+, the STPD ID must be identical to the VLAN ID of the carrier VLAN in that STPD. See Specifying the Carrier VLAN on page 1045 for an example.

If you have an 802.1D configuration, we recommend that you configure the StpdID to be identical to the VLAN ID of the carrier VLAN in that STPD. See Basic 802.1D Configuration Example on page 1087 for an example.

If you configure Multiple Spanning Tree (*MSTP*—IEEE 802.1Q-2003, formerly IEEE 802.1s), you do not need carrier VLANs for MSTP operation. With MSTP, you configure a Common and Internal Spanning Tree (CIST) that controls the connectivity of interconnecting MSTP regions and sends BPDUs across the regions to communicate the status of MSTP regions. All VLANs participating in the MSTP region have the same privileges. For more information about MSTP, see Multiple Spanning Tree Protocol on page 1070.

### *Protected VLAN*

Protected *VLAN*s are all other VLANs that are members of the *STPD*.

These VLANs “piggyback” on the carrier VLAN. Protected VLANs do not transmit or receive *STP* BPDUs, but they are affected by STP state changes and inherit the state of the carrier VLAN. Protected VLANs can participate in multiple STPDs, but any particular port in the VLAN can belong to only one STPD. Also known as non-carrier VLANs.

If you configure *MSTP*, all member VLANs in an MSTP region are protected VLANs. These VLANs do not transmit or receive STP BPDUs, but they are affected by STP state changes communicated by the CIST to the MSTP regions. Multiple spanning tree instances (MSTIs) cannot share the same protected VLAN; however, any port in a protected VLAN can belong to multiple MSTIs. For more information about MSTP, see Multiple Spanning Tree Protocol on page 1070.

### *Specifying the Carrier VLAN*

The following example:

- Creates and enables an *STPD* named s8.
- Creates a carrier *VLAN* named v5.
- Assigns VLAN v5 to STPD s8.
- Creates the same tag ID for the VLAN and the STPD (the carrier VLAN’s ID must be identical to the STPD’s ID).

```
create vlan v5
configure vlan v5 tag 100
```

```
configure vlan v5 add ports 1:1-1:20 tagged
create stpd s8
configure stpd s8 add vlan v5 ports all emistp
configure stpd s8 tag 100
enable stpd s8
```

Notice how the tag number for the VLAN v5 (100) is identical to the tag for STPD s8. By using identical tags, you have selected the carrier VLAN. The carrier VLAN's ID is now identical to the STPD's ID.

## STPD Modes

An *STPD* has three modes of operation:

- 802.1D mode

  Use this mode for backward compatibility with previous *STP* versions and for compatibility with third-party switches using IEEE standard 802.1D. When configured in this mode, all rapid configuration mechanisms are disabled.

- 802.1w mode

  Use this mode for compatibility with Rapid Spanning Tree (RSTP). When configured in this mode, all rapid configuration mechanisms are enabled. The benefit of this mode is available on point-to-point links only and when the peer is likewise configured in 802.1w mode. If you do not select point-to-point links and the peer is not configured for 802.1w mode, the STPD fails back to 802.1D mode.

  You can enable or disable RSTP on a per STPD basis only; you cannot enable RSTP on a per port basis.

  For more information about RSTP and RSTP features, see Rapid Spanning Tree Protocol on page 1059.

- *MSTP* mode

  Use this mode for compatibility with MSTP. MSTP is an extension of RSTP and offers the benefit of better scaling with fast convergence. When configured in this mode, all rapid configuration mechanisms are enabled. The benefit of MSTP is available only on point-to-point links and when you configure the peer in MSTP or 802.1w mode. If you do not select point-to-point links and the peer is not configured in 802.1w mode, the STPD fails back to 802.1D mode.

  You must first configure a CIST before configuring any MSTIs in the region. You cannot delete or disable a CIST if any of the MSTIs are active in the system.

  You can create only one MSTP region on the switch, and all switches that participate in the region must have the same regional configurations. You can enable or disable an MSTP on a per STPD basis only; you cannot enable MSTP on a per port basis.

  If configured in MSTP mode, an STPD uses the 802.1D BPDU encapsulation mode by default. To ensure correct operation of your MSTP STPDs, do not configure EMISTP or PVST+ encapsulation mode for MSTP STPDs.

  For more information about MSTP and MSTP features, see Multiple Spanning Tree Protocol on page 1070.

By default:

- The STPD operates in 802.1D mode.
- The default device configuration contains a single STPD called s0.
- The default *VLAN* is a member of STPD s0 with autobind enabled.

To configure the mode of operation of an STPD, use the following command:

```
configure stpd stpd_name mode [dot1d | dot1w | mstp [cist | msti instance]]
```

All STP parameters default to the IEEE 802.1D values, as appropriate.

## Encapsulation Modes

You can configure ports within an *STPD* to accept specific BPDU encapsulations.

This *STP* port encapsulation is separate from the STP mode of operation. For example, you can configure a port to accept the PVST+ BPDU encapsulation while running in 802.1D mode.

An STP port has three possible encapsulation modes:

- 802.1D mode

  Use this mode for backward compatibility with previous STP versions and for compatibility with third-party switches using IEEE standard 802.1D. BPDUs are sent untagged in 802.1D mode. Because of this, any given physical interface can have only one STPD running in 802.1D mode.

  This encapsulation mode supports the following STPD modes of operation: 802.1D, 802.1w, and *MSTP*.
- Extreme Multiple Instance Spanning Tree Protocol (EMISTP) mode

  EMISTP mode is proprietary to Extreme Networks and is an extension of STP that allows a physical port to belong to multiple STPDs by assigning the port to multiple *VLAN*s. EMISTP adds significant flexibility to STP network design. BPDUs are sent with an 802.1Q tag having an STPD instance Identifier (STPD ID) in the VLAN ID field.

  This encapsulation mode supports the following STPD modes of operation: 802.1D and 802.1w.
- Per VLAN Spanning Tree (PVST+) mode

  This mode implements PVST+ in compatibility with third-party switches running this version of STP. The STPDs running in this mode have a one-to-one relationship with VLANs and send and process packets in PVST+ format.

  This encapsulation mode supports the following STPD modes of operation: 802.1D and 802.1w.

These encapsulation modes are for STP ports, not for physical ports. When a physical port belongs to multiple STPDs, it is associated with multiple STP ports. It is possible for the physical port to run in different modes for different domains to which it belongs.

If configured in MSTP mode, an STPD uses the 802.1D BPDU encapsulation mode by default. To ensure correct operation of your MSTP STPDs, do not configure EMISTP or PVST+ encapsulation mode for MSTP STPDs.

- To configure the BPDU encapsulation mode for one or more STP ports, use the command:
  - `configure stpd stpd_name ports mode [dot1d | emistp | pvst-plus] port_list`
- To configure the default BPDU encapsulation mode on a per STPD basis, use the command:
  - `configure stpd stpd_name default-encapsulation [dot1d | emistp | pvst-plus]`

Instead of accepting the default encapsulation modes of dot1d for the default STPD s0 and emistp for all other STPDs, this command allows you to specify the type of BPDU encapsulation to use for all ports added to the STPD (if not otherwise specified).

*STPD Identifier*

An StpdID is used to identify each *STP* domain.

When assigning the StpdID when configuring the domain, ensure that the carrier *VLAN* of that *STPD* does not belong to another STPD. Unless all ports are running in 802.1D mode, an STPD with ports running in either EMISTP mode or PVST+ mode must be configured with an StpdID.

An StpdID must be identical to the VLAN ID of the carrier VLAN in that STP domain. For an 802.1D STPD, the VLAN ID can be either a user-defined ID or one automatically assigned by the switch.

Note
If an STPD contains at least one port not in 802.1D mode, you must configure the STPD with an StpdID.

*MSTP* uses two different methods to identify the STPDs that are part of the MSTP network. An instance ID of 0 identifies the CIST. The switch assigns this ID automatically when you configure the CIST STPD. An *MSTI (Multiple Spanning Tree Instances)* identifier (MSTI ID) identifies each STP domain that is part of an MSTP region. You assign the MSTI ID when configuring the STPD that participates in the MSTP region. In an MSTP region, MSTI IDs only have local significance. You can reuse MSTI IDs across MSTP regions. For more information about MSTP and MSTP features, see Multiple Spanning Tree Protocol on page 1070.

## STP States

Each port that belongs to a member *VLAN* participating in *STP* exists in one of the following states:

### Blocking

A port in the blocking state does not accept ingress traffic, perform traffic forwarding, or learn MAC source addresses. The port receives STP BPDUs. During STP initialization, the switch always enters the blocking state.

### Listening

A port in the listening state does not accept ingress traffic, perform traffic forwarding, or learn MAC source addresses. The port receives STP BPDUs. This is the first transitional state a port enters after

being in the blocking state. The bridge listens for BPDUs from neighboring bridge(s) to determine whether the port should or should not be blocked.

Learning

A port in the learning state does not accept ingress traffic or perform traffic forwarding, but it begins to learn MAC source addresses. The port also receives and processes STP BPDUs. This is the second transitional state after listening. From learning, the port will change to either blocking or forwarding.

Forwarding

A port in the forwarding state accepts ingress traffic, learns new MAC source addresses, forwards traffic, and receives and processes STP BPDUs.

Disabled

A port in the disabled state does not participate in STP; however, it will forward traffic and learn new MAC source addresses.

## Binding Ports

There are two ways to bind (add) ports to an *STPD*: manually and automatically. By default, ports are manually added to an STPD.

Note
The default *VLAN* and STPD S0 are already on the switch.

### *Manually Binding Ports*

- To manually bind ports, use the commands:

  `configure stpd` *stpd_name* **`addvlan`** *vlan_name* **`ports`** `[`**`all`** `|` *port_list*`]` `{[`**`dot1d`** `|`**`emistp`** `|`**`pvst-plus`**`]}`

  `configure vlan` *vlan_name* **`addports`** `[`**`all`** `|` *port_list*`]` `{`**`tagged`** `{`*tag*`}` `|` **`untagged`**`}`**`stpd`** *stpd_name* `{[`**`dot1d`** `|`**`emistp`** `|`**`pvst-plus`**`]}`

  The first command adds all ports or a list of ports within the specified *VLAN* to an *STPD*. For EMISTP and PVST+, the carrier VLAN must already exist on the same set of ports. The second command adds all ports or a list of ports to the specified VLAN and STPD at the same time. If the ports are added to the VLAN but not to the STPD, the ports remain in the VLAN.

  For EMISTP and PVST+, if the specified VLAN is not the carrier VLAN and the specified ports are not bound to the carrier VLAN, the system displays an error message. If you configure *MSTP* on your switch, MSTP does not need carrier VLANs.

  Note
  The carrier VLAN's ID must be identical to the ID of the STP domain.

  If you add a protected VLAN or port, that addition inherits the carrier VLAN's encapsulation mode, unless you specify the encapsulation mode when you execute the `configure stpd add vlan` or `configure vlan add ports stpd` commands. If you specify an encapsulation mode (dot1d, emistp, or pvst-plus), the *STP* port mode is changed to match; otherwise, the STP port inherits either the carrier VLANs encapsulation mode on that port or the STPD's default encapsulation mode.

For MSTP, you do not need carrier a VLAN. A CIST controls the connectivity of interconnecting MSTP regions and sends BPDUs across the regions to communicate region status. You must use the dot1d encapsulation mode in an MSTP environment. For more information about MSTP, see the section Multiple Spanning Tree Protocol on page 1070.

- To remove ports, use the command:

  `configure stpd stpd_name deletevlan vlan_name ports [all | port_list]`

  If you manually delete a protected VLAN or port, only that VLAN or port is removed. If you manually delete a carrier VLAN or port, all VLANs on that port (both carrier and protected) are deleted from that STPD.

To learn more about member VLANs, see Member VLANs on page 1044. For more detailed information about these command line interface (CLI) commands, see the *ExtremeXOS 16.2 Command Reference Guide*.

### *Automatically Binding Ports*

- To automatically bind ports to an *STPD* when the ports are added to a *VLAN*, use the command:

  `enable stpd stpd_name auto-bind vlan vlan_name`

  The autobind feature is disabled on user-created STPDs. The autobind feature is enabled on the default VLAN that participates in the default STPD S0.

  For EMISTP or PVST+, when you issue this command, any port or list of ports that you add to the carrier VLAN are automatically added to the STPD with autobind enabled. In addition, any port or list of ports that you remove from a carrier VLAN are automatically removed from the STPD. This feature allows the STPD to increase or decrease its span as ports are added to or removed from a carrier VLAN.

  **Note**
  The carrier VLAN's ID must be identical to the ID of the *STP* domain.

  Enabling autobind on a protected VLAN does not expand the boundary of the STPD.

  If the same set of ports are members of the protected VLAN and the carrier VLAN, protected VLANs are aware of STP state changes. For example, assume you have the following scenario:

  - Carrier VLAN named v1
  - v1 contains ports 3:1-3:2
  - Protected VLAN named v2
  - v2 contains ports 3:1-3:4

  Since v1 contains ports 3:1-3:2, v2 is aware only of the STP changes for ports 3:1 and 3:2, respectively. Ports 3:3 and 3:4 are not part of the STPD, which is why v2 is not aware of any STP changes for those ports.

  In addition, enabling autobind on a protected VLAN causes ports to be automatically added or removed as the carrier VLAN changes.

  For *MSTP*, when you issue this command, any port or list of ports that gets automatically added to an *MSTI* are automatically inherited by the CIST. In addition, any port or list of ports that you remove from an MSTI protected VLAN are automatically removed from the CIST. For more information, see Automatically Inheriting Ports--MSTP Only on page 1051.

- To remove ports, enter the command:

  `configure stpd stpd_name` **`deletevlan`** `vlan_name` **`ports`** `[` **`all`** `|` `port_list]`

  If you manually delete a port from the STPD on a VLAN that has been added by autobind, ExtremeXOS records the deletion so that the port does not get automatically added to the STPD after a system restart.

To learn more about the member VLANs, see Member VLANs on page 1044. For more detailed information about these CLI commands, see the *ExtremeXOS 16.2 Command Reference Guide*.

*Automatically Inheriting Ports--MSTP Only*

In an MSTP environment, whether you manually or automatically bind a port to an *MSTI* in an *MSTP* region, the switch automatically binds that port to the CIST.

The CIST handles BPDU processing for itself and all of the MSTIs; therefore, the CIST must inherit ports from the MSTIs in order to transmit and receive BPDUs. You can only delete ports from the CIST if it is no longer a member of an MSTI.

For more information about MSTP, see Multiple Spanning Tree Protocol on page 1070.

## Rapid Root Failover

ExtremeXOS supports rapid root failover for faster *STP* failover recovery times in STP 802.1D mode. If the active root port link goes down, ExtremeXOS recalculates STP and elects a new root port. The rapid root failover feature allows the new root port to immediately begin forwarding, skipping the standard listening and learning phases. Rapid root failover occurs only when the link goes down and not when there is any other root port failure, such as missing BPDUs.

The default setting for this feature is disabled.

- To enable rapid root failover, enter the command:

  `enable stpd stpd_name` **`rapid-root-failover`**

- To display the configuration, enter the command:

  `show stpd {stpd_name |` **`detail`**`}`

## STP and Hitless Failover--Modular Switches Only

When you install two management modules (MSM/MM) in a BlackDiamond chassis or you are using redundancy in a SummitStack, one node assumes the role of primary and the other node assumes the role of backup. The primary executes the switch's management functions, and the backup acts in a standby role. Hitless failover transfers switch management control from the primary to the backup and maintains the state of *STP*. STP supports hitless failover. You do not explicitly configure hitless failover support; rather, if you have two nodes installed, hitless failover is available.

**Note**

Not all platforms support hitless failover in the same software release. To verify if the software version you are running supports hitless failover, see the following table in Managing the Switch on page 39. For more information about protocol, platform, and MSM/MM support for hitless failover, see Understanding Hitless Failover Support on page 59.

To support hitless failover, the primary node replicates STP BPDUs to the backup, which allows the nodes to run STP in parallel. Although both primary and backup node receive STP BPDUs, only the primary transmits STP BPDUs to neighboring switches and participates in STP.

Note

Before initiating failover, review the section Synchronizing Nodes--Modular Switches and SummitStack Only on page 1550 to confirm that both primary and backup nodes are running software that supports the `synchronize` command.

To initiate hitless failover on a network that uses STP:

1. Confirm that the nodes are synchronized and have identical software and switch configurations using the command:

   `show switch {detail}`

   The output displays the status of the primary and backup nodes, with the primary node showing MASTER and the backup node showing BACKUP (InSync).

   If the primary and backup nodes are not synchronized and both nodes are running a version of ExtremeXOS that supports synchronization, proceed to 2.

   If the primary and backup nodes are synchronized, proceed to 3 on page 1052.
2. If the primary and backup nodes are not synchronized, use the `synchronize` command to replicate all saved images and configurations from the primary to the backup.

   After you confirm the nodes are synchronized, proceed to 3.
3. If the nodes are synchronized, use the `run failover` (formerly `run msm-failover`) command to initiate failover.

For more detailed information about verifying the status of the primary and backup nodes, and system redundancy, see Understanding System Redundancy on page 54. For more information about hitless failover, see Understanding Hitless Failover Support on page 59.

## STP Configurations

When you assign *VLAN*s to an *STPD*, pay careful attention to the *STP* configuration and its effect on the forwarding of VLAN traffic.

This section describes three types of STP configurations:

- Basic STP
- Multiple STPDs on a single port (which uses EMISTP)
- A VLAN that spans multiple STPDs

### Basic STP Configuration

This section describes a basic, 802.1D *STP* configuration. The following figure illustrates a network that uses *VLAN* tagging for trunk connections.

The following four VLANs have been defined:

- Sales is defined on switch A, switch B, and switch M.
- Personnel is defined on switch A, switch B, and switch M.
- Manufacturing is defined on switch Y, switch Z, and switch M.
- Engineering is defined on switch Y, switch Z, and switch M.
- Marketing is defined on all switches (switch A, switch B, switch Y, switch Z, and switch M).

Two STPDs are defined:

- STPD1 contains VLANs Sales and Personnel.
- STPD2 contains VLANs Manufacturing and Engineering.

The carrier and protected VLANs are also defined:

- Sales is the carrier VLAN on STPD1.
- Personnel is a protected VLAN on STPD1.
- Manufacturing is a protected VLAN on STPD2.
- Engineering is the carrier VLAN on STPD2.
- Marketing is a member of both STPD1 and STPD2 and is a protected VLAN.

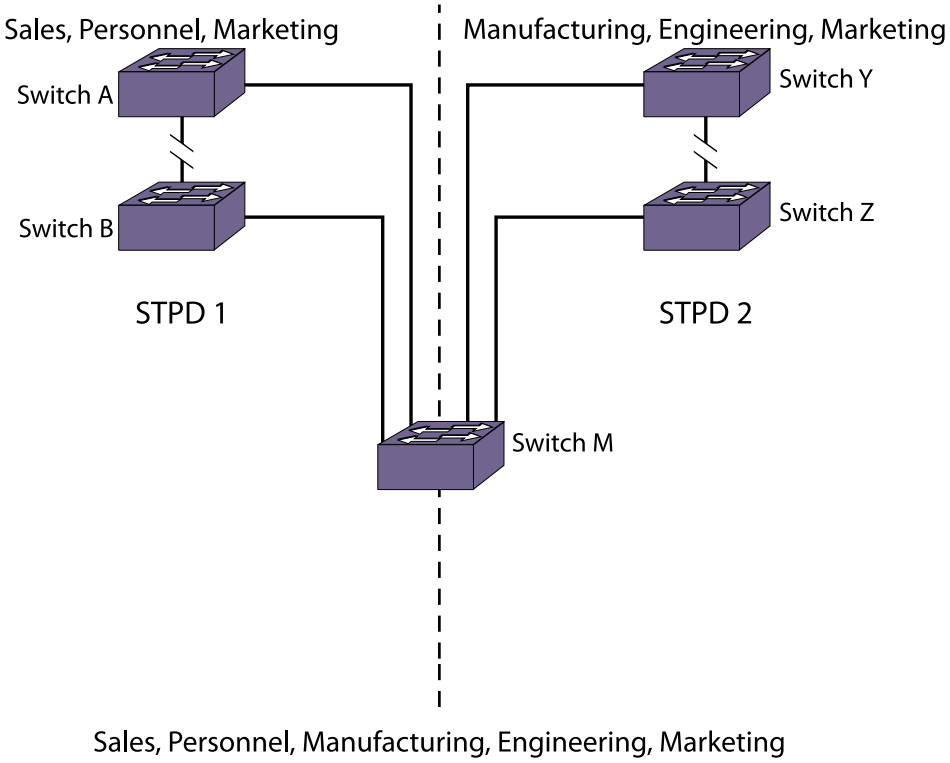



**Figure 144: Multiple STPDs**

When the switches in this configuration boot-up, STP configures each *STPD* such that the topology contains no active loops. STP could configure the topology in a number of ways to make it loop-free.

In the following figure, the connection between switch A and switch B is put into blocking state, and the connection between switch Y and switch Z is put into blocking state. After STP converges, all the VLANs can communicate, and all bridging loops are prevented.

The protected VLAN Marketing, which has been assigned to both STPD1 and STPD2, communicates using all five switches. The topology has no loops, because STP has already blocked the port connection between switch A and switch B and between switch Y and switch Z.

Within a single STPD, you must be extra careful when configuring your VLANs. The following figure illustrates a network that has been incorrectly set up using a single STPD so that the STP configuration disables the ability of the switches to forward VLAN traffic.

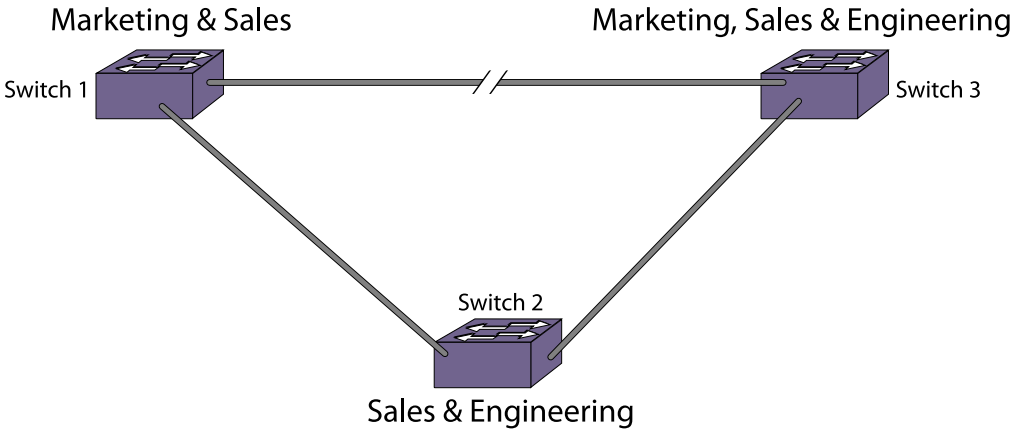



**Figure 145: Incorrect Tag-Based STPD Configuration**

The tag-based network in the following figure has the following configuration:

- Switch 1 contains VLAN Marketing and VLAN Sales.
- Switch 2 contains VLAN Engineering and VLAN Sales.
- Switch 3 contains VLAN Marketing, VLAN Engineering, and VLAN Sales.
- The tagged trunk connections for three switches form a triangular loop that is not permitted in an STP topology.
- All VLANs in each switch are members of the same STPD.

STP can block traffic between switch 1 and switch 3 by disabling the trunk ports for that connection on each switch.

Switch 2 has no ports assigned to VLAN Marketing. Therefore, if the trunk for VLAN Marketing on switches 1 and 3 is blocked, the traffic for VLAN Marketing will not be able to traverse the switches.

Note
If an STPD contains multiple VLANs, all VLANs should be configured on all ports in that domain, except for ports that connect to hosts (edge ports).

## Multiple STPDs on a Port

Traditional 802.1D *STP* has some inherent limitations when addressing networks that have multiple *VLAN*s and multiple *STPD*s.

For example, consider the sample depicted in the following figure.

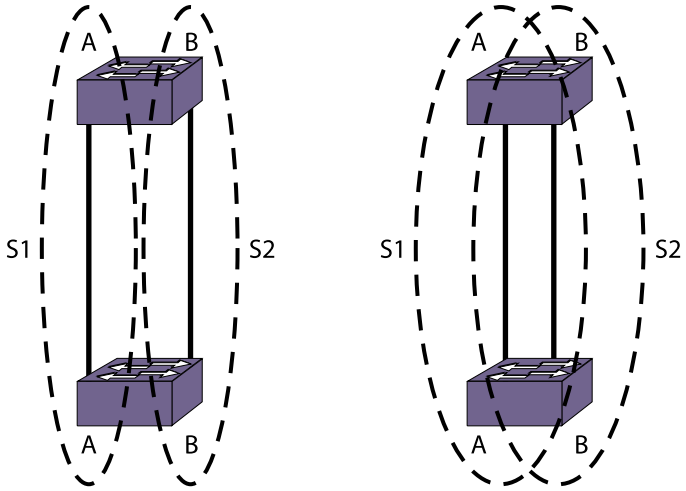



**Figure 146: Limitations of Traditional STPD**

The two switches are connected by a pair of parallel links. Both switches run two VLANs, A and B. To achieve load-balancing between the two links using the traditional approach, you would have to associate A and B with two different STPDs, called S1 and S2, respectively, and make the left link carry VLAN A traffic while the right link carries VLAN B traffic (or vice versa). If the right link fails, S2 is broken and VLAN B traffic is disrupted.

To optimize the solution, you can use the Extreme Multiple Instance Spanning (EMISTP) mode, which allows a port to belong to multiple STPDs. EMISTP adds significant flexibility to STP network design. Referring to the figure above, using EMISTP, you can configure all four ports to belong to both VLANs.

Assuming that S1 and S2 still correspond to VLANs A and B respectively, you can fine-tune STP parameters to make the left link active in S1 and blocking in S2, while the right link is active in S2 and blocking in S1. Again, if the right link fails, the left link is elected active by the STP algorithm for S2, without affecting normal switching of data traffic.

Using EMISTP, an STPD becomes more of an abstract concept. The STPD does not necessarily correspond to a physical domain; it is better regarded as a vehicle to carry VLANs that have STP instances. Because VLANs can overlap, so do STPDs. However, even if the different STPDs share the entire topology or part of the redundant topology, the STPDs react to topology change events in an independent fashion.

## VLANs Spanning Multiple STPDs

Traditionally, the mapping from *VLAN*s to *STP* instances have been one-to-one or many-to-one.

In both cases, a VLAN is wholly contained in a single instance. In practical deployment there are cases in which a one-to-many mapping is desirable. In a typical large enterprise network, for example, VLANs span multiple sites and/or buildings. Each site represents a redundant looped area. However, between any two sites the topology is usually very simple.

Alternatively, the same VLAN may span multiple large geographical areas (because they belong to the same enterprise) and may traverse a great many nodes.

In this case, it is desirable to have multiple STP domains operating in a single VLAN, one for each looped area.

The justifications include the following:

- The complexity of the STP algorithm increases, and performance drops, with the size and complexity of the network. The 802.1D standard specifies a maximum network diameter of seven hops. By segregating a big VLAN into multiple STPDs, you reduce complexity and enhance performance.
- Local to each site, there may be other smaller VLANs that share the same redundant looped area with the large VLAN. Some STPDs must be created to protect those VLANs. The ability to partition VLANs allows the large VLAN to be "piggybacked" in those STPDs in a site-specific fashion.

The following figure has five domains. VLANs green, blue, brown, and yellow are local to each domain. VLAN red spans all of the four domains. Using a VLAN that spans multiple STPDS, you do not have to create a separate domain for VLAN red. Instead, VLAN red is “piggybacked” onto those domains local to other VLANs.

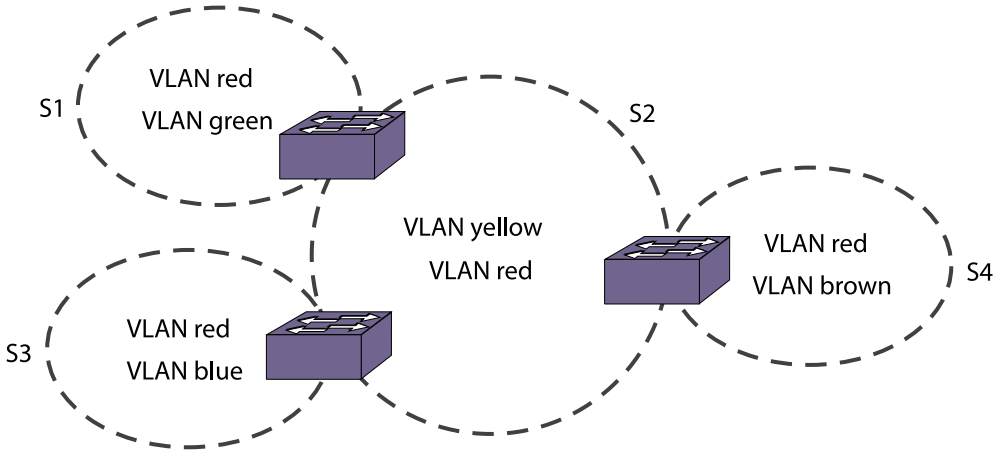



**Figure 147: VLANs Spanning Multiple STPDs**

In addition, the configuration in the figure has these features:

- Each site can be administered by a different organization or department within the enterprise. Having a site-specific STP implementation makes the administration more flexible and convenient.
- Between the sites the connections usually traverse distribution switches in ways that are known beforehand to be “safe” with STP. In other words, the looped areas are already well defined.

## EMISTP Deployment Constraints

Although EMISTP greatly enhances *STP* capability, these features must deployed with care.

This section describes configuration issues that, if not followed, could lead to an improper deployment of EMISTP. This section also provides the following restrictive principles to abide by in network design:

- Although a physical port can belong to multiple *STPD*s, any *VLAN* on that port can be in only one domain. Put another way, a VLAN cannot belong to two STPDs on the same physical port.
- Although a VLAN can span multiple domains, any LAN segment in that VLAN must be in the same STPD. VLANs traverse STPDs only inside switches, not across links. On a single switch, however,

bridge ports for the same VLAN can be assigned to different STPDs. This scenario is illustrated in the following figure.

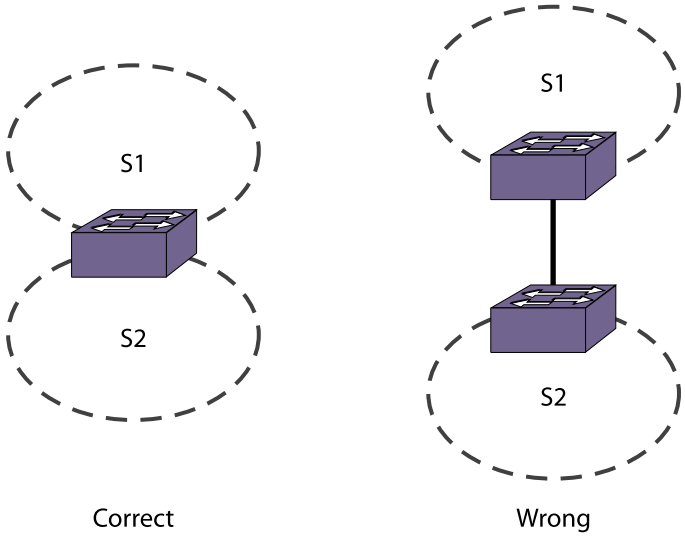



**Figure 148: VLANs Traverse Domains Inside Switches**

- The VLAN partition feature is deployed under the premise that the overall inter-domain topology for that VLAN is loop-free. Consider the case in the following figure, VLAN red (the only VLAN in the figure) spans STPDs 1, 2, and 3. Inside each domain, STP produces a loop-free topology. However, VLAN red is still looped, because the three domains form a ring among themselves.

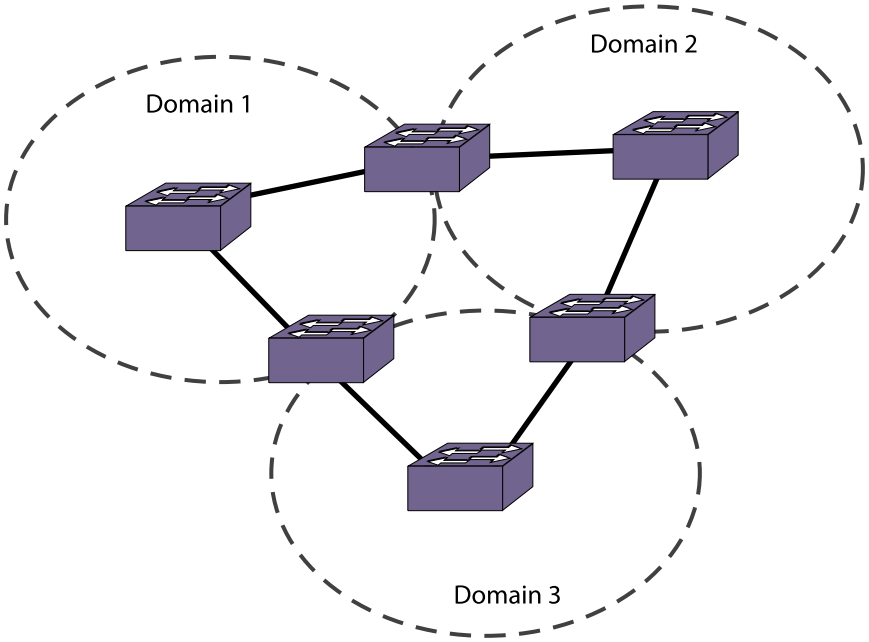



**Figure 149: Looped VLAN Topology**

- A necessary (but not sufficient) condition for a loop-free inter-domain topology is that every two domains only meet at a single crossing point.

Note
You can use MSTP to overcome the EMISTP constraints described in this section.

## Per VLAN Spanning Tree

Switching products that implement Per *VLAN* Spanning Tree (PVST) have been in existence for many years and are widely deployed.

To support *STP* configurations that use PVST, ExtremeXOS has an operational mode called PVST+.

Note
In this document, PVST and PVST+ are used interchangeably. PVST+ is an enhanced version of PVST that is interoperable with 802.1Q STP. The following discussions are in regard to PVST+, if not specifically mentioned.

### STPD VLAN Mapping

Each *VLAN* participating in PVST+ must be in a separate *STPD*, and the VLAN number (VLAN ID) must be the same as the STPD identifier (STPD ID).

As a result, PVST+ protected VLANs cannot be partitioned.

This fact does not exclude other non-PVST+ protected VLANs from being grouped into the same STPD. A protected PVST+ VLAN can be joined by multiple non-PVST+ protected VLANs to be in the same STPD.

Note

When PVST+ is used to interoperate with other networking devices, each VLAN participating in PVST+ must be in a separate *STP* domain.

### Native VLAN

In PVST+, the native *VLAN* must be peered with the default VLAN on Extreme Networks devices, as both are the only VLANs allowed to send and receive untagged packets on the physical port.

Third-party PVST+ devices send VLAN 1 packets in a special manner. ExtremeXOS does not support PVST+ for VLAN 1. Therefore, when the switch receives a packet for VLAN 1, the packet is dropped.

When a PVST+ instance is disabled, the fact that PVST+ uses a different packet format raises an issue. If the *STPD* also contains ports not in PVST+ mode, the flooded packet has an incompatible format with those ports. The packet is not recognized by the devices connected to those ports.

## Rapid Spanning Tree Protocol

The Rapid Spanning Tree Protocol (RSTP), originally in the IEEE 802.1w standard and now part of the IEEE 802.1D-2004 standard, provides an enhanced spanning tree algorithm that improves the convergence speed of bridged networks.

RSTP takes advantage of point-to-point links in the network and actively confirms that a port can safely transition to the forwarding state without relying on any timer configurations. If a network topology change or failure occurs, RSTP rapidly recovers network connectivity by confirming the change locally before propagating that change to other devices across the network. For broadcast links, there is no difference in convergence time between *STP* and RSTP.

RSTP supersedes legacy STP protocols, supports the existing STP parameters and configurations, and allows for seamless interoperability with legacy STP.

### RSTP Concepts

#### *Port Roles*

RSTP uses information from BPDUs to assign port roles for each LAN segment. Port roles are not user-configurable. Port role assignments are determined based on the following criteria:

- A unique bridge identifier (MAC address) associated with each bridge
- The path cost associated with each bridge port
- A port identifier associated with each bridge port

RSTP assigns one of the following port roles to bridge ports in the network, as described in the following table.

**Table 119: RSTP Port Roles**

| Port Role | Description |
|---|---|
| Root | Provides the shortest (lowest) path cost to the root bridge. Each bridge has only one root port; the root bridge does not have a root port. If a bridge has two or more ports with the same path cost, the port with the best port identifier (lowest MAC address) becomes the root port. |
| Designated | Provides the shortest path connection to the root bridge for the attached LAN segment. To prevent loops in the network, there is only one designated port on each LAN segment. To select the designated port, all bridges that are connected to a particular segment listen to each other's BPDUs and agree on the bridge sending the best BPDU. The corresponding port on that bridge becomes the designated port. If there are two or more ports connected to the LAN, the port with the best port identifier becomes the designated port. |
| Alternate | Provides an alternate path to the root bridge and the root port. |
| Backup | Supports the designated port on the same attached LAN segment. Backup ports exist only when the bridge is connected as a self-loop or to a shared-media segment. |
| Disabled | A port in the disabled state does not participate in RSTP; however, it will forward traffic and learn new MAC source addresses. |

When RSTP stabilizes:

- All root ports and designated ports are in the forwarding state.
- All alternate ports and backup ports are in the blocking state.

RSTP makes the distinction between the alternate and backup port roles to describe the rapid transition of the alternate port to the forwarding state if the root port fails.

To prevent a port from becoming an alternate or backup port, use the command:

`configure stpd` *`stpd_name`* **`ports active-role enable`** *`port`* .

To revert to the default that allows a port to be elected to any *STP* port role, use the command:

`configure stpd` *`stpd_name`* **`ports active-role disable`** *`port`*

To view the active-role status, use the command: `show stpd ports`.

*Link Types*

With RSTP, you can configure the link type of a port in an *STPD*.

RSTP tries to rapidly move designated point-to-point links into the forwarding state when a network topology change or failure occurs. For rapid convergence to occur, the port must be configured as a point-to-point link.

The following table describes the link types.

**Table 120: RSTP Link Types**

| Port Link Type | Description |
|---|---|
| Auto | Specifies the switch to automatically determine the port link type. An auto link behaves like a point-to-point link if the link is in full-duplex mode or if link aggregation is enabled on the port. Otherwise, the link behaves like a broadcast link used for 802.1w configurations. |
| Edge | Specifies a port that does not have a bridge attached. An edge port is held in the *STP* forwarding state unless a BPDU is received by the port. In that case, the port behaves as a normal RSTP port. The port is no longer considered an edge port. If the port does not receive subsequent BPDUs during a pre-determined time, the port attempts to become an edge port.<br>ExtremeXOS 11.5 or earlier—An edge port is placed and held in the STP forwarding state unless a BPDU is received by the port. In that case, an edge port enters and remains in the blocking state until it stops receiving BPDUs and the message age timer expires. |
| Broadcast | Specifies a port attached to a LAN segment with more than two bridges. A port with a broadcast link type cannot participate in rapid reconfiguration using RSTP or *MSTP*. By default, all ports are broadcast links. |
| Point-to-point | Specifies a port attached to a LAN segment with only two bridges. A port with point-to-point link type can participate in rapid reconfiguration. Used for 802.1w and MSTP configurations. |

### Configuring Link Types

By default, all ports are broadcast links.

- To configure the ports in an *STPD*, enter the command:.

  `configure stpd` *stpd_name* **`ports link-type`** `[[`**`auto`** `|` **`broadcast`** `|` **`point-to-point`**`]` *port_list* `|` **`edge`** *port_list* `{`**`edge-safeguard`** `[`**`enable`** `|` **`disable`**`]` `{`**`bpdu-restrict`**`}` `{`**`recovery-timeout`** *seconds*`}}]`

  Where the following is true:

  - auto—Configures the ports as auto links. If the link is in full-duplex mode or if link aggregation is enabled on the port, an auto link behaves like a point-to-point link.
  - broadcast—Configures the ports as broadcast ports. By default, all ports are broadcast links.
  - point-to-point—Configures the ports for rapid reconfiguration in an RSTP or *MSTP* environment.
  - edge—Configures the ports as edge ports. For information about edge safeguard, see Configuring Edge Safeguard on page 1061.

- To change the existing configuration of a port in an STPD, and return the port to factory defaults, enter the command:

  `unconfigure stpd` *stpd_name* **`ports link-type`** *port_list*

- To display detailed information about the ports in an STPD, enter the command:

  `show {`**`stpd`**`}` *stpd_name* **`ports`** `{[`**`detail`** `|` *port_list* `{`**`detail`**`}]}`

### Configuring Edge Safeguard

Loop prevention and detection on an edge port configured for RSTP is called *edge safeguard*. You can configure edge safeguard on RSTP edge ports to prevent accidental or deliberate misconfigurations (loops) resulting from connecting two edge ports together or by connecting a hub or other non-*STP*

switch to an edge port. Edge safeguard also limits the impact of broadcast storms that might occur on edge ports. This advanced loop prevention mechanism improves network resiliency but does not interfere with the rapid convergence of edge ports.

An edge port configured with edge safeguard immediately enters the forwarding state and transmits BPDUs. If a loop is detected, STP blocks the port. By default, an edge port without edge safeguard configured immediately enters the forwarding state but does not transmit BPDUs unless a BPDU is received by that edge port.

You can also configure edge safeguard for loop prevention and detection on an *MSTP* edge port.

- To configure an edge port and enable edge safeguard on that port, use the command:

  configure stpd *stpd_name* **ports link-type** [[**auto** | **broadcast** | **point-to-point**] *port_list* | **edge** *port_list* {**edge-safeguard** [**enable** | **disable**] {**bpdu-restrict**} {**recovery-timeout** *seconds*}}]

- If you have already configured a port as an edge port and you want to enable edge safeguard on the port, use the following command:

  configure {**stpd**} *stpd_name* **ports edge-safeguard enable** *port_list* {**bpdu-restrict**} {**recovery-timeout** {*seconds*}}

- To disable edge safeguard on an edge port, enter the command:

  configure {**stpd**} *stpd_name* **ports edge-safeguard disable** *port_list* {**bpdu-restrict**} {**recovery-timeout** {*seconds*}}

  configure stpd *stpd_name* **ports link-type** [[**auto** | **broadcast** | **point-to-point**] *port_list* | **edge** *port_list* {**edge-safeguard** [**enable** | **disable**] {**bpdu-restrict**} {**recovery-timeout** *seconds*}}]

In ExtremeXOS 11.5 and earlier, ports that connect to non-STP devices are edge ports. Edge ports do not participate in RSTP, and their role is not confirmed. Edge ports immediately enter the forwarding state unless the port receives a BPDU. In that case, edge ports enter the blocking state. The edge port remains in the blocking state until it stops receiving BPDUs and the message age timer expires.

ExtremeXOS 11.6 and later support an enhanced bridge detection method, which is part of the 802.1D-2004 standard. Ports that connect to non-STP devices are still considered edge ports. However, if you have an 802.1D-2004 compliant edge port, the bridge detection mechanism causes the edge port to transition to a non-edge port upon receiving a BPDU. If the former edge port does not receive a subsequent BPDU during a pre-determined interval, the port attempts to become an edge port.

In ExtremeXOS 12.0.3 and 12.1.4 onwards, STP edge safeguard disables a port when a remote loop is detected. ExtremeXOS versions prior to 12.0.3 and 12.1.4 place the port in blocking mode. The change was made because BPDUs are still processed when a port is in a blocking state. A remote loop causes BPDUs to be exponentially duplicated which caused high CPU utilization on the switch even though the port was transitioned to a blocked state.

#### Configuring Auto Edge

This feature helps to automatically determine if a port is an edge port or non-edge port. If no BPDU is received for a period of time on auto edge enabled ports, the switch marks those ports as edge ports. If BPDU is received, the switch marks those ports as non-edge ports. By default auto edge is enabled on all ports.

### *RSTP Timers*

For RSTP to rapidly recover network connectivity, RSTP requires timer expiration. RSTP derives many of the timer values from the existing configured *STP* timers to meet its rapid recovery requirements rather than relying on additional timer configurations.

Table 121 on page 1063 describes the user-configurable timers, and the Table 122 on page 1063 describes the timers that are derived from other timers and are not user configurable.

**Table 121: User-Configurable Timers**

| Timer | Description |
|---|---|
| Hello | The root bridge uses the hello timer to send out configuration BPDUs through all of its forwarding ports at a predetermined, regular time interval. The default value is 2 seconds. The range is 1 to 10 seconds. |
| Forward delay | A port moving from the blocking state to the forwarding state uses the forward delay timer to transition through the listening and learning states. In RSTP, this timer complements the rapid configuration behavior. If none of the rapid rules are in effect, the port uses legacy STP rules to move to the forwarding state. The default is 15 seconds. The range is 4 to 30 seconds. |

**Table 122: Derived Timers**

| Timer | Description |
|---|---|
| TCN | The root port uses the topology change notification (TCN) timer when it detects a change in the network topology. The TCN timer stops when the topology change timer expires or upon receipt of a topology change acknowledgement. The default value is the same as the value for the bridge hello timer. |
| Topology change | The topology change timer determines the total time it takes the forwarding ports to send configuration BPDUs. The default value for the topology change timer depends upon the mode of the port:<br>802.1D mode—The sum of the forward delay timer value (default value is 15 seconds; range of 4 to 30 seconds) and the maximum age timer value (default value is 20 seconds; range of 6 to 40 seconds).<br>802.1w mode—Double the hello timer value (default value is 4 seconds). |
| Message age | A port uses the message age timer to time out receiving BPDUs. When a port receives a superior or equal BPDU, the timer restarts. When the timer expires, the port becomes a designated port and a configuration update occurs. If the bridge operates in 1w mode and receives an inferior BPDU, the timer expires early. The default value is the same as the *STPD* bridge max age parameter. |
| Hold | A port uses the hold timer to restrict the rate that successive BPDUs can be sent. The default value is the same as the value for the bridge hello timer. |
| Recent backup | The timer starts when a port leaves the backup role. When this timer is running, the port cannot become a root port. The default value is double the hello time (4 seconds). |
| Recent root | The timer starts when a port leaves the root port role. When this timer is running, another port cannot become a root port unless the associated port is put into the blocking state. The default value is the same as the forward delay time. |

The protocol migration timer is neither user-configurable nor derived; it has a set value of 3 seconds. The timer starts when a port transitions from STP (802.1D) mode to RSTP (802.1w) mode and vice-versa. This timer must expire before further mode transitions can occur.

## RSTP Operation

In an RSTP environment, a point-to-point link LAN segment has two bridges.

A switch that considers itself the unique, designated bridge for the attached LAN segment sends a "propose" message to the other bridge to request a confirmation of its role. The other bridge on that LAN segment replies with an "agree" message if it agrees with the proposal. The receiving bridge immediately moves its designated port into the forwarding state.

Before a bridge replies with an "agree" message, it reverts all of its designated ports into the blocking state. This introduces a temporary partition into the network. The bridge then sends another "propose" message on all of its designated ports for further confirmation. Because all of the connections are blocked, the bridge immediately sends an "agree" message to unblock the proposing port without having to wait for further confirmations to come back or without the worry of temporary loops.

Beginning with the root bridge, each bridge in the network engages in the exchange of "propose" and "agree" messages until they reach the edge ports. Edge ports connect to non-*STP* devices and do not participate in RSTP. Their role does not need to be confirmed. If you have an 802.1D-2004 compliant edge port, the bridge detection mechanism causes the edge port to transition to a non-edge port upon receiving a BPDU. If the former edge port does not receive a subsequent BPDU during a pre-determined interval, the port attempts to become an edge port.

RSTP attempts to transition root ports and designated ports to the forwarding state and alternate ports and backup ports to the blocking state as rapidly as possible.

A port transitions to the forwarding state if any of the port:

- Has been in either a root or designated port role long enough that the spanning tree information supporting this role assignment has reached all of the bridges in the network;

  Note
  RSTP is backward-compatible with STP, so if a port does not move to the forwarding state with any of the RSTP rapid transition rules, a forward delay timer starts and STP behavior takes over.

- Is now a root port and no other ports have a recent role assignment that contradicts with its root port role;
- Is a designated port and attaches to another bridge by a point-to-point link and receives an "agree" message from the other bridge port; or
- Is an edge port. An edge port is a port connected to a non-STP device and is in the forwarding state.

The following sections provide more information about RSTP behavior.

### *Root Port Rapid Behavior*

In the following figure, the diagram on the left displays the initial network topology with a single bridge having the following:

- Two ports are connected to a shared LAN segment.

- One port is the designated port.
- One port is the backup port.

The diagram on the right displays a new bridge that:

- Is connected to the LAN segment.
- Has a superior *STP* bridge priority.
- Becomes the root bridge and sends a BPDU to the LAN that is received by both ports on the old bridge.

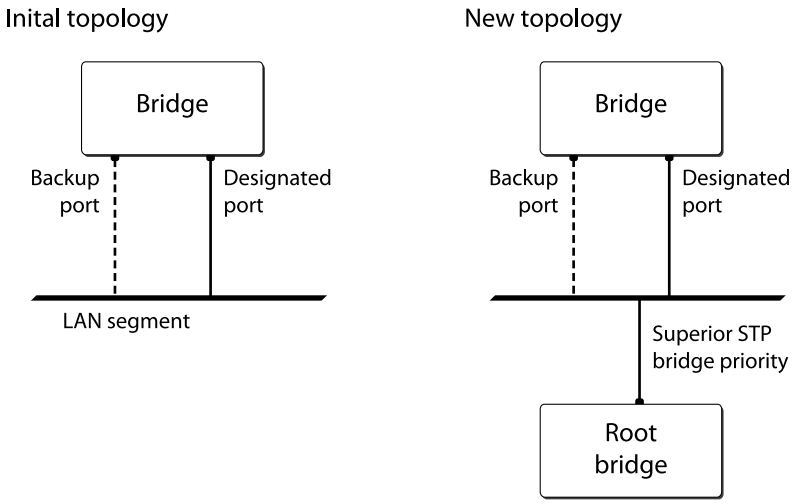



**Figure 150: Example of Root Port Rapid Behavior**

If the backup port receives the BPDU first, STP processes this packet and temporarily elects this port as the new root port while the designated port's role remains unchanged. If the new root port is immediately put into the forwarding state, there is a loop between these two ports.

To prevent this type of loop from occurring, the recent backup timer starts. The root port transition rule does not allow a new root port to be in the forwarding state until the recent backup timer expires.

Another situation may arise if you have more than one bridge and you lower the port cost for the alternate port, which makes it the new root port. The previous root port is now an alternate port. Depending on your STP implementation, STP may set the new root port to the forwarding state before setting the alternate port to the blocking state. This may cause a loop.

To prevent this type of loop from occurring, the recent root timer starts when the port leaves the root port role. The timer stops if the port enters the blocking state. RSTP requires that the recent root timer stop on the previous root port before the new root port can enter the forwarding state.

### *Designated Port Rapid Behavior*

When a port becomes a new designated port, or the *STP* priority changes on an existing designated port, the port becomes an unsynced designated port.

For an unsynced designated port to rapidly move into the forwarding state, the port must propose a confirmation of its role on the attached LAN segment (unless the port is an edge port). Upon receiving an "agree" message, the port immediately enters the forwarding state.

If the receiving bridge does not agree and it has a superior STP priority, the receiving bridge replies with its own BPDU. Otherwise, the receiving bridge keeps silent, and the proposing port enters the forwarding state and starts the forward delay timer.

The link between the new designated port and the LAN segment must be a point-to-point link. If there is a multi-access link, the “propose” message is sent to multiple recipients. If only one of the recipients agrees with the proposal, the port can erroneously enter the forwarding state after receiving a single “agree” message.

### *Receiving Bridge Behavior*

The receiving bridge must decide whether or not to accept a proposal from a port.

Upon receiving a proposal for a root port, the receiving bridge:

- Processes the BPDU and computes the new *STP* topology.
- Synchronizes all of the designated ports if the receiving port is the root port of the new topology.
- Puts all unsynced, designated ports into the blocking state.
- Sends down further “propose” messages.
- Sends back an “agree” message through the root port.

If the receiving bridge receives a proposal for a designated port, the bridge replies with its own BPDU. If the proposal is for an alternate or backup port, the bridge keeps silent.

### *Propagating Topology Change Information*

When a change occurs in the topology of the network, such events are communicated through the network.

In an RSTP environment, only non-edge ports entering the forwarding state cause a topology change. A loss of network connectivity is not considered a topology change; however, a gain in network connectivity must be communicated. When an RSTP bridge detects a topology change, that bridge starts the topology change timer, sets the topology change flag on its BPDUs, floods all of the forwarding ports in the network (including the root ports), and flushes the learned MAC address entries.

### *Rapid Reconvergence*

This section describes the RSTP rapid behavior following a topology change.

In this example, the bridge priorities are assigned based on the order of their alphabetical letters; bridge A has a higher priority than bridge F.

Suppose you have a network, as shown in the following figure, with six bridges (bridge A through bridge F) where the following is true:

- Bridge A is the root bridge.
- Bridge D contains an alternate port in the blocking state.
- All other ports in the network are in the forwarding state.

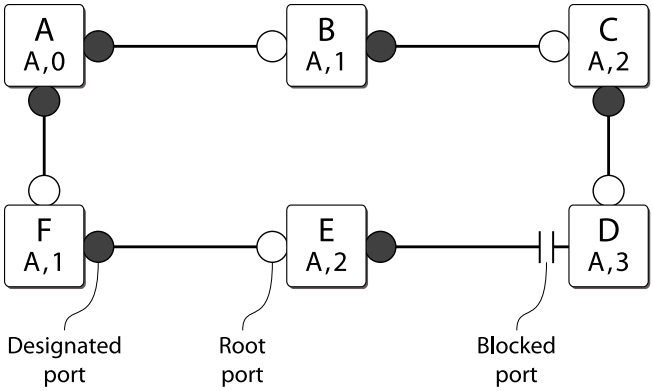



**Figure 151: Initial Network Configuration**

The network reconverges in the following way:

If the link between bridge A and bridge F goes down, bridge F detects the root port is down. At this point, bridge F:

- Immediately disables that port from the *STP*.
- Performs a configuration update.

As shown in the following figure, after the configuration update, bridge F:

- Considers itself the new root bridge.
- Sends a BPDU message on its designated port to bridge E.

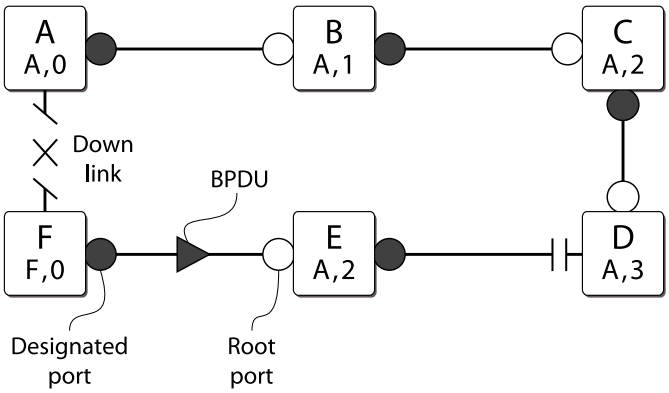



**Figure 152: Down Link Detected**

- Bridge E believes that bridge A is the root bridge. When bridge E receives the BPDU on its root port from bridge F, bridge E:
  - Determines that it received an inferior BPDU.
  - Immediately begins the max age timer on its root port.
  - Performs a configuration update.

As shown in the following figure, after the configuration update, bridge E:

- Regards itself as the new root bridge.
- Sends BPDU messages on both of its designated ports to bridges F and D, respectively.

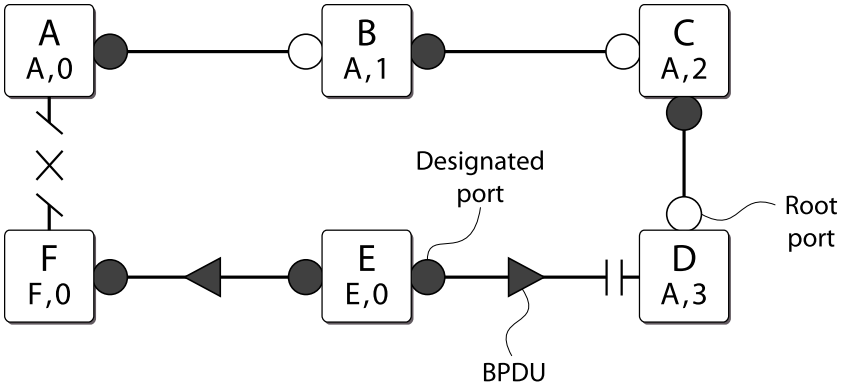



**Figure 153: New Root Bridge Selected**

As shown in the following figure, when bridge F receives the superior BPDU and configuration update from bridge E, bridge F:

- Decides that the receiving port is the root port.
- Determines that bridge E is the root bridge.

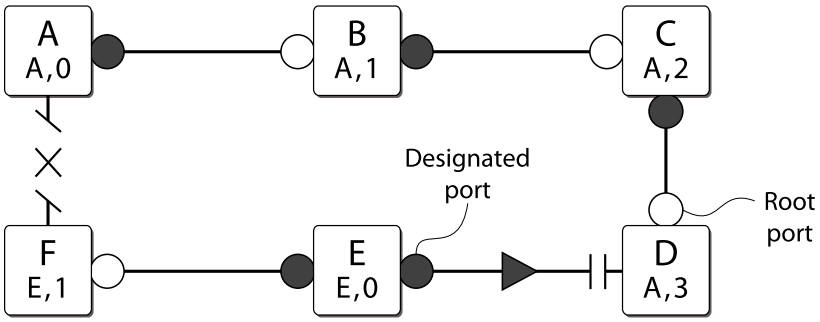



**Figure 154: Communicating New Root Bridge Status to Neighbors**

Bridge D believes that bridge A is the root bridge. When bridge D receives the BPDU from bridge E on its alternate port, bridge D:

- Immediately begins the max age timer on its alternate port.
- Performs a configuration update.

As shown in the following figure, after the configuration update, bridge D:

- Moves the alternate port to a designated port.
- Sends a “propose” message to bridge E to solicit confirmation of its designated role and to rapidly move the port into the designated state.

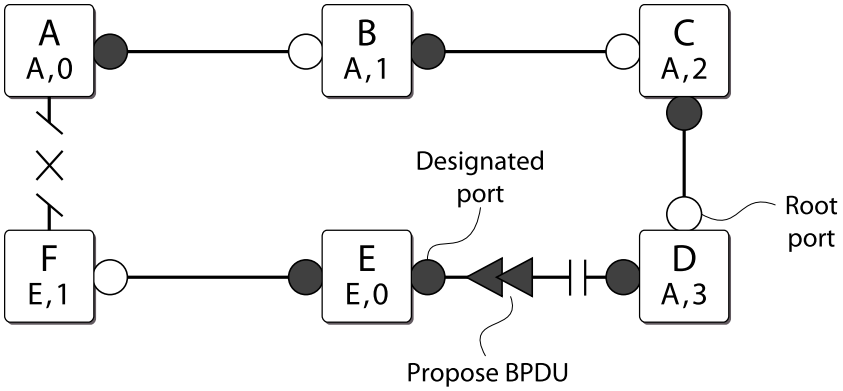



**Figure 155: Sending a Propose Message to Confirm a Port Role**

Upon receiving the proposal, bridge E (as shown in the following figure):

- Performs a configuration update.
- Changes its receiving port to a root port.

The existing designated port enters the blocking state.

Bridge E then sends:

- A “propose” message to bridge F.
- An “agree” message from its root port to bridge D.

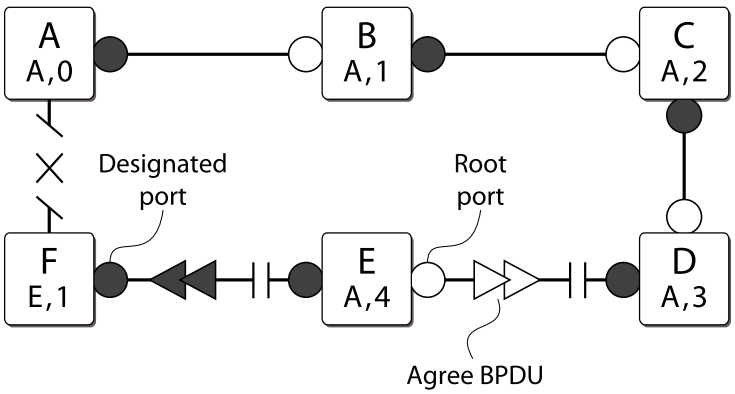



**Figure 156: Communicating Port Status to Neighbors**

To complete the topology change (as shown in the following figure):

- Bridge D moves the port that received the “agree” message into the forwarding state.
- Bridge F confirms that its receiving port (the port that received the “propose” message) is the root port, and immediately replies with an “agree” message to bridge E to unblock the proposing port.

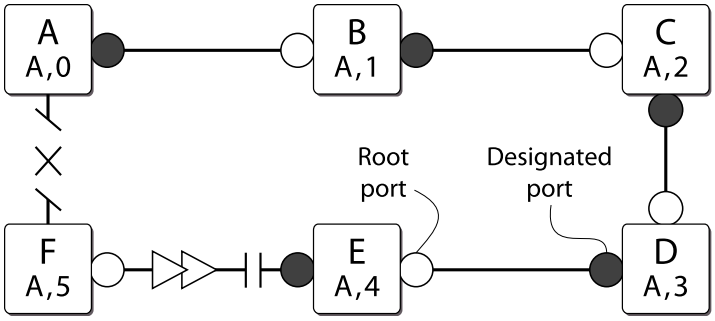



**Figure 157: Completing the Topology Change**

The following figure displays the new topology.

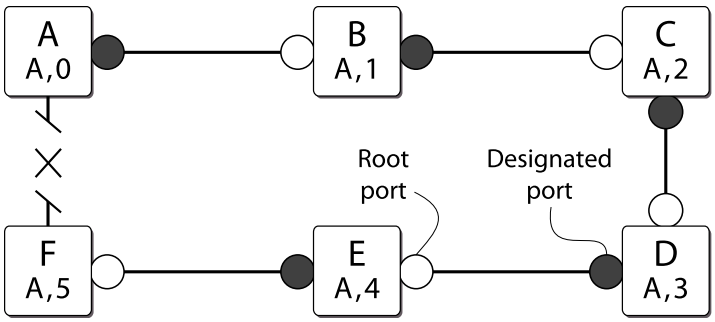



**Figure 158: Final Network Configuration**

*Compatibility With STP (802.1D)*

RSTP interoperates with legacy *STP* protocols; however, the rapid convergence benefits are lost when interacting with legacy STP bridges.

Each RSTP bridge contains a port protocol migration state machine to ensure that the ports in the *STPD* operate in the correct, configured mode. The state machine is a protocol entity within each bridge configured to run in 802.1w mode. For example, a compatibility issue occurs if you configure 802.1w mode and the bridge receives an 802.1D BPDU on a port. The receiving port starts the protocol migration timer and remains in 802.1D mode until the bridge stops receiving 802.1D BPDUs. Each time the bridge receives an 802.1D BPDU, the timer restarts. When the port migration timer expires, no more 802.1D BPDUs have been received, and the bridge returns to its configured setting, which is 802.1w mode.

## Multiple Spanning Tree Protocol

The *MSTP*, based on IEEE 802.1Q-2003 (formerly known as IEEE 802.1s), allows the bundling of multiple *VLAN*s into one spanning tree topology.

This concept is not new to Extreme Networks. Like MSTP, Extreme Networks proprietary EMISTP implementation can achieve the same capabilities of sharing a virtual network topology among multiple VLANs; however, MSTP overcomes some of the challenges facing EMISTP, including enhanced loop protection mechanisms and new capabilities to achieve better scaling.

MSTP logically divides a Layer 2 network into regions. Each region has a unique identifier and contains multiple spanning tree instances (MSTIs). An *MSTI* is a spanning tree domain that operates within and is bounded by a region. MSTIs control the topology inside the regions. The Common and Internal

Spanning Tree (CIST) is a single spanning tree domain that interconnects MSTP regions. The CIST is responsible for creating a loop-free topology by exchanging and propagating BPDUs across regions to form a Common Spanning Tree (CST).

MSTP uses RSTP as its converging algorithm and is interoperable with the legacy *STP* protocols: STP (802.1D) and RSTP (802.1w).

MSTP has three major advantages over 802.1D, 802.1w, and other proprietary implementations:

- To save control path bandwidth and provide improved scalability, MSTP uses regions to localize BPDU traffic. BPDUs containing information about MSTIs contained within an MSTP region do not cross that region's boundary.
- A single BPDU transmitted from a port can contain information for up to 64 STPDs. MSTP BPDU processing utilizes less resources compared to 802.1D or 802.1w where one BPDU corresponds to one *STPD*.
- In a typical network, a group of VLANs usually share the same physical topology. Dedicating a spanning tree per VLAN like PVST+ is CPU intensive and does not scale very well. MSTP makes it possible for a single STPD to handle multiple VLANs.

## MSTP Concepts

### *MSTP Regions*

An *MSTP* network consists of either individual MSTP regions connected to the rest of the network with 802.1D and 802.1w bridges or as individual MSTP regions connected to each other.

An MSTP region defines the logical boundary of the network. With MSTP, you can divide a large network into smaller areas similar to an *OSPF (Open Shortest Path First)* area or a *BGP (Border Gateway Protocol)* Autonomous System, which contain a group of switches under a single administration. Each MSTP region has a unique identifier and is bound together by one CIST that spans the entire network. A bridge participates in only one MSTP region at a time.

An MSTP region can hide its internal STPDs and present itself as a virtual 802.1w bridge to other interconnected regions or 802.1w bridges because the port roles are encoded in 802.1w and MSTP BPDUs.

By default, the switch uses the MAC address of the switch to generate an MSTP region. Since each MAC address is unique, every switch is in its own region by default. For multiple switches to be part of an MSTP region, you must configure each switch in the region with the same MSTP region identifiers. See Configuring MSTP Region Identifiers on page 1072 for information.

In the following figure, all bridges inside MSTP regions 1 and 2 are MSTP bridges; bridges outside of the regions are either 802.1D or 802.1w bridges.

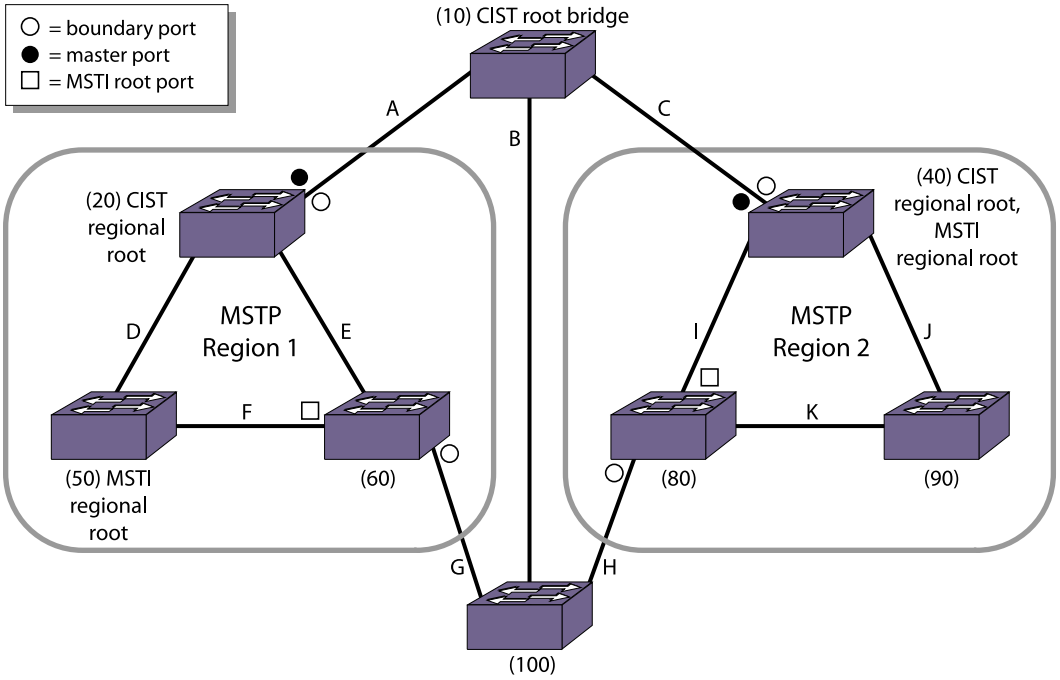



**Figure 159: Sample MSTP Topology with Two MSTP Regions**

### Configuring MSTP Region Identifiers

For multiple switches to be part of an *MSTP* region, you must configure each switch in the region with the same MSTP configuration attributes, also known as MSTP region identifiers. The following list describes the MSTP region identifiers:

- Region Name—This indicates the name of the MSTP region. In the Extreme Networks implementation, the maximum length of the name is 32 characters and can be a combination of alphanumeric characters and underscores ( _ ).
- Format Selector—This indicates a number to identify the format of MSTP BPDUs. The default is 0.
- Revision Level—An unsigned integer encoded within a fixed field of 2 octets that identifies the revision of the current MST configuration. The revision number is not incremented automatically each time that the MST configuration is committed.

The switches inside a region exchange BPDUs that contain information for MSTIs.

The switches connected outside of the region exchange CIST information. By having devices look at the region identifiers, MSTP discovers the logical boundary of a region:

- To configure the MSTP region name, use the command:

  `configure mstp region` *`regionName`*

  The maximum length of the region name is 32 characters and can be a combination of alphanumeric characters and underscores ( _ ). You can configure only one MSTP region on the switch at any given time.

  If you have an active MSTP region, we recommend that you disable all active STPDs in the region before renaming the region on all of the participating switches.
- To configure the number used to identify MSTP BPDUs, use the command:

  `configure mstp format` *`format_identifier`*

By default, the value used to identify the MSTP BPDUs is 0. The range is 0 to 255.

If you have an active MSTP region, we recommend that you disable all active STPDs in the region before modifying the value used to identify MSTP BPDUs on all participating switches.

- To configure the MSTP revision level, use the command:

  `configure mstp revision` *`revision`*

  Although this command is available on the CLI, this command is reserved for future use.

#### Unconfiguring an MSTP Region

Before you unconfigure an *MSTP* region, we recommend that you disable all active STPDs in the region.

To unconfigure the MSTP region on the switch, use the command:

```
unconfigure mstp region
```

After you issue this command, all of the MSTP settings return to their default values. See Configuring MSTP Region Identifiers on page 1072 for information about the default settings.

### *Common and Internal Spanning Tree*

*MSTP* logically divides a Layer 2 network into regions. The Common and Internal Spanning Tree (CIST) is a single spanning tree domain that interconnects MSTP regions. The CIST is responsible for creating a loop-free topology by exchanging and propagating BPDUs across regions to form a Common Spanning Tree (CST).

In essence, the CIST is similar to having a large spanning tree across the entire network. The CIST has its own root bridge that is common to all MSTP regions, and each MSTP region elects a CIST regional root that connects that region to the CIST, thereby forming a CST.

The switch assigns the CIST an instance ID of 0, which allows the CIST to send BPDUs for itself in addition to all of the MSTIs within an MSTP region. Inside a region, the BPDUs contain CIST records and piggybacked M-records. The CIST records contain information about the CIST, and the M-records contain information about the MSTIs. Boundary ports exchange only CIST record BPDUs.

All MSTP configurations require a CIST domain. You must first configure the CIST domain before configuring any MSTIs. By default, all *MSTI* ports in the region are inherited by the CIST. You cannot delete or disable a CIST if any of the MSTIs are active in the system.

#### Configuring the CIST

- Configure an *STPD* as the CIST, specifying the **`mstp cist`** keywords in the following command:

  `configure stpd` *`stpd_name`* **`mode`** `[`**`dot1d`** | **`dot1w`** | **`mstp`** `[`**`cist`** | **`msti`** *`instance`*`]]`

  You can enable *MSTP* on a per STPD basis only. By specifying the **`mstp cist`** keywords, you can configure the mode of operation for the STPD as MSTP and identify the STPD to be the CIST.

#### CIST Root Bridge

In a Layer 2 network, the bridge with the lowest bridge ID becomes the CIST root bridge. The parameters (vectors) that define the root bridge include the following:

- User-defined bridge priority (by default, the bridge priority is 32,768)

- MAC address

The CIST root bridge can be either inside or outside an *MSTP* region. The CIST root bridge is unique for all regions and non-MSTP bridges, regardless of its location.

For more information about configuring the bridge ID, see the `configure stpd priority` command.

### CIST Regional Root Bridge

Within an *MSTP* region, the bridge with the lowest path cost to the CIST root bridge is the CIST regional root bridge.

The path cost, also known as the CIST external path cost, is a function of the link speed and number of hops. If there is more than one bridge with the same path cost, the bridge with the lowest bridge ID becomes the CIST regional root. If the CIST root is inside an MSTP region, the same bridge is the CIST regional root for that region because it has the lowest path cost to the CIST root. If the CIST root is outside an MSTP region, all regions connect to the CIST root via their CIST regional roots.

The total path cost to the CIST root bridge from any bridge in an MSTP region consists of the CIST internal path cost (the path cost of the bridge to the CIST regional root bridge) and the CIST external path cost. To build a loop-free topology within a region, the CIST uses the external and internal path costs, and the *MSTI* uses only the internal path cost.

Looking at MSTP region 1 in the following figure, the total path cost for the bridge with ID 60 consists of an external path cost of A and an internal path cost of E.

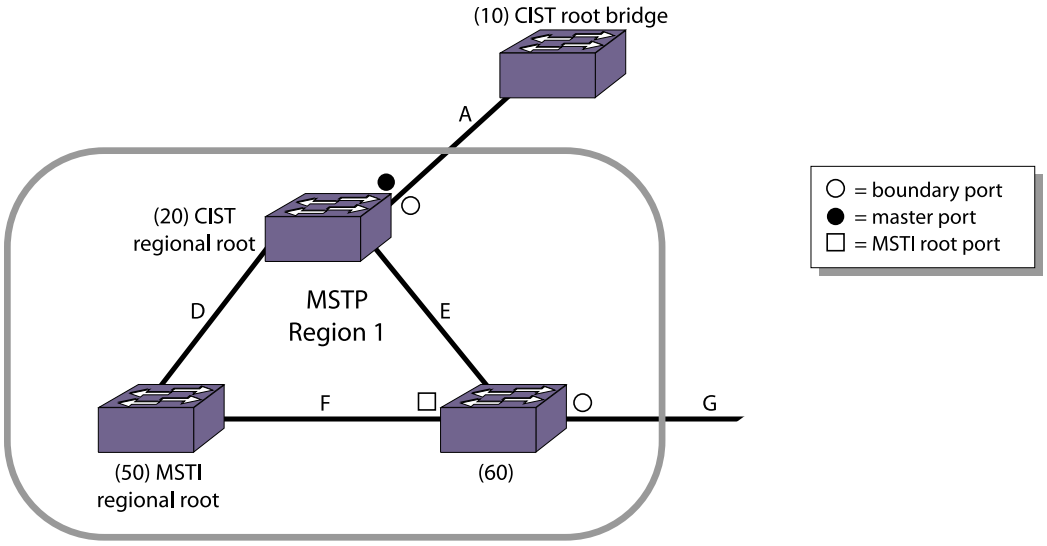



**Figure 160: Closeup of MSTP Region 1**

### CIST Root Port

The port on the *CIST regional root bridge* that connects to the CIST root bridge is the *CIST root port* (also known as the master port for *MSTI*s).

The CIST root port is the master port for all MSTIs in that region, and it is the only port that connects the entire region to the CIST root.

If a bridge is both the CIST root bridge and the CIST regional root bridge, there is no CIST root port on that bridge.

#### Enabling the CIST

To enable the CIST, use the following command and specify the CIST domain as the *stpd_name*:

```
enable stpd {stpd_name}
```

### *Multiple Spanning Tree Instances*

*MSTI*s control the topology inside an *MSTP* region. An MSTI is a spanning tree domain that operates within and is bounded by a region; an MSTI does not exchange BPDUs with or send notifications to other regions. You must identify an MSTI on a per region basis. The MSTI ID does not have any significance outside of its region so you can reuse IDs across regions. An MSTI consists of a group of VLANs, which can share the same network topology. Each MSTI has its own root bridge and a tree spanning its bridges and LAN segments.

You must first configure a CIST before configuring any MSTIs in the region. You cannot delete or disable a CIST if any of the MSTIs are active in the system.

You can map multiple *VLAN*s to an MSTI; however, multiple MSTIs cannot share the same VLAN.

#### Configuring the MSTI and the MSTI ID

*MSTP* uses the *MSTI* ID, not an Stpd ID, to identify the spanning tree contained within the region. As previously described, the MSTI ID only has significance within its local region, so you can re-use IDs across regions.

To configure the MSTI that is inside an MSTP region and its associated MSTI ID, use the following command and specify the **mstp** [**msti** *instance*] parameters:

```
configure stpd stpd_name mode [dot1d | dot1w | mstp [cist | msti
instance]]
```

The range of the MSTI instance ID is 1–4094.

MSTP STPDs use 802.1D BPDU encapsulation mode by default. To ensure correct operation of your MSTP STPDs, do not configure EMISTP or PVST+ encapsulation mode for MSTP STPDs. For more information, see Encapsulation Modes on page 1047.

#### MSTI Regional Root Bridge

Each *MSTI* independently chooses its own root bridge. For example, if two MSTIs are bounded to a region, there is a maximum of two MSTI regional roots and one CIST regional root.

The bridge with the lowest bridge ID becomes the MSTI regional root bridge. The parameters that define the root bridge include the following:

- User-defined bridge priority (by default, the bridge priority is 32,768)
- MAC address

Within an *MSTP* region, the cost from a bridge to the MSTI regional root bridge is known as the MSTI internal path cost. Looking at MSTP region 1 in Figure 160 on page 1074, the bridge with ID 60 has a path cost of F to the MSTI regional root bridge.

The MSTI regional root bridge can be the same as or different from the CIST regional root bridge of that region. You achieve this by assigning different priorities to the *STP* instances configured as the MSTIs and the CIST. For more information about configuring the bridge ID, see the `configure stpd priority` command in the *ExtremeXOS 16.2 Command Reference Guide*.

#### MSTI Root Port

The port on the bridge that has the lowest path cost to the *MSTI* regional root bridge is the MSTI root port.

If a bridge has two or more ports with the same path cost, the port with the best port identifier becomes the root port.

#### Enabling the MSTI

To enable the *MSTI*, use the following command and specify the MSTI domain as the *stpd_name*:

```
enable stpd {stpd_name}
```

Note

If two switches are configured for the same CIST and MSTI region, in order for them to understand that they are in the same region, both must also belong to the same *VLAN* which is added to the *STP* domain. If they belong to different VLANs, each switch believes that each belongs to a different region. When an *MSTP* BPDU is sent, it carries a VID digest created by VLAN memberships in the CIST domain and the MSTI domain.

### *Boundary Ports*

Boundary ports are bridge ports that are only connected to other *MSTP* regions or 802.1D or 802.1w bridges.

The ports that are not at a region boundary are called internal ports. The boundary ports exchange only CIST BPDUs. A CIST BPDU originated from the CIST root enters a region through the CIST root port and egresses through boundary ports. This behavior simulates a region similar to an 802.1w bridge, which receives BPDUs on its root ports and forwards updated BPDUs on designated ports.

The following figure shows an MSTP network that consists of two MSTP regions. Each region has its own CIST regional root and is connected to the CIST root through master ports. The CIST regional roots in each region are the MSTP bridges having the lowest CIST external root path cost. The CIST root is the bridge with the lowest bridge ID and is an 802.1w bridge outside of either MSTP region.

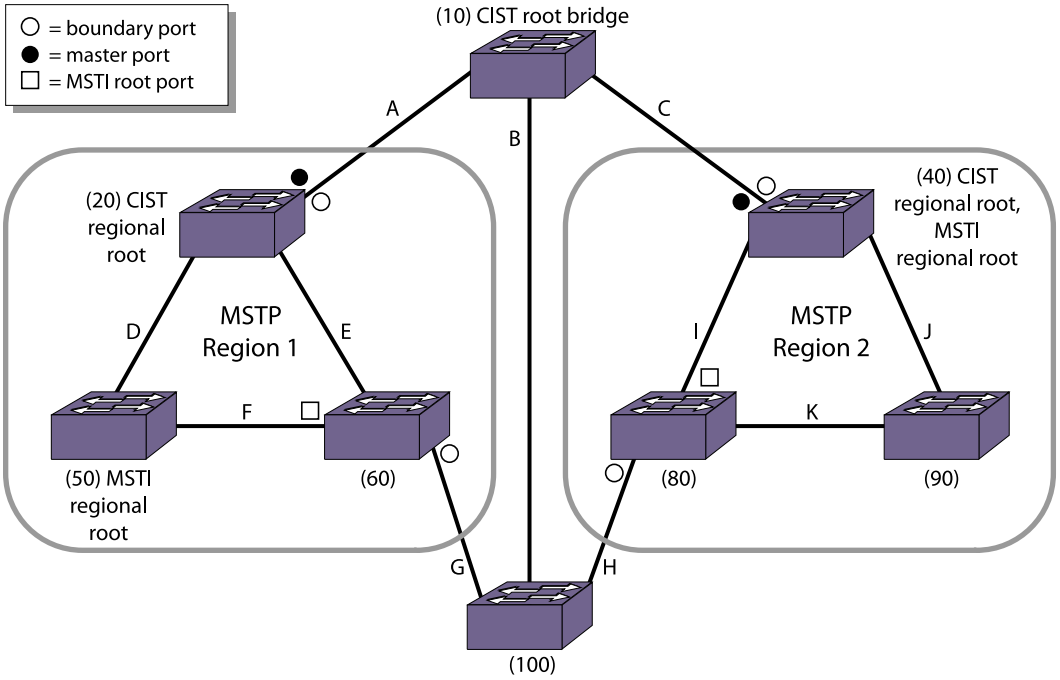



**Figure 161: Sample MSTP Topology with Two MSTP Regions**

MSTP Region 1 and MSTP Region 2 are connected to the CIST root through directly connected ports, identified as master ports. The bridge with ID 100 connects to the CIST root through Region 1, Region 2, or segment B. For this bridge, either Region 1 or Region 2 can be the designated region or segment B can be the designated segment. The CIST BPDUs egressing from the boundary ports carry the CIST regional root as the designated bridge. This positions the entire MSTP region as one virtual bridge.

The CIST controls the port roles and the state of the boundary ports. A master port is always forwarding for all CIST and *MSTI* VLANs. If the CIST sets a boundary port to the discarding state, the CIST blocks traffic for all VLANs mapped to it and the MSTIs within that region. Each MSTI blocks traffic for their member VLANs and puts their internal ports into the forwarding or blocking state depending on the MSTI port roles.

### *MSTP Port Roles*

*MSTP* uses the same port roles as RSTP (Root, Designated, Alternate, and Backup).

In addition to these port roles, MSTP introduces a new port role: Master. A Master port is the port that connects an *MSTI* to the CIST root.

### *MSTP Port States*

*MSTP* uses the same port states as RSTP (Listening, Learning, Forwarding, and Blocking).

In the Extreme Networks MSTP implementation, the listening state is not truly implemented as *FDB (forwarding database)* learning cannot be done when the port is not in the forwarding state. Ports in the blocking state listen but do not accept ingress traffic, perform traffic forwarding, or learn MAC source address; however, the port receives and processes BPDUs.

For more information about all of the *STP* port states, see STP States on page 1048.

### *MSTP Link Types*

*MSTP* uses the same link types as *STP* and RSTP, respectively.

In an MSTP environment, configure the same link types for the CIST and all MSTIs.

For more information about the link types, see Link Types on page 1060.

### *MSTP Edge Safeguard*

You can configure edge safeguard for loop prevention and detection on an *MSTP* edge port. For more information, see Configuring Edge Safeguard on page 1061.

**Note**

In MSTP, configuring edge safeguard at CIST will be inherited in all *MSTI*s.

In MSTP, an edge port needs to be added to a CIST before adding it to an MSTI.

### *MSTP Timers*

*MSTP* uses the same timers as *STP* and RSTP. For more information, see RSTP Timers on page 1063.

### *MSTP Hop Counts*

In an *MSTP* environment, the hop count has the same purpose as the maxage timer for 802.1D and 802.1w environments. The CIST hop count is used within and outside a region. The *MSTI* hop count is used only inside of the region. In addition, if the other end is an 802.1D or 802.1w bridge, the maxage timer is used for interoperability between the protocols.

The BPDUs use hop counts to age out information and to notify neighbors of a topology change.

To configure the hop count.

```
configure stpd stpd_name max-hop-count hopcount
```

By default, the hop count of a BPDU is 20 hops. The range is 6 to 40 hops.

### *Configuring MSTP on the Switch*

To configure and enable *MSTP*:

1. Create the MSTP region using the following command:

   ```
   configure mstp region regionName
   ```

2. Create and configure the CIST, which forms the CST, using the following commands:

   ```
   create stpd stpd_name {description stpd-description}
   configure stpd stpd_name mode mstp cist
   ```

**Note**

You can configure the default *STPD*, S0 as the CIST.

No *VLAN* can be bound to the CIST and no ports can be added to the CIST. Therefore, the VLAN should be bound to the *MSTI* and the “show MSTI port” command will show the VLAN ports. The ports added to the MSTI are bound automatically to the CIST even though they are not added to it.

3. Enable the CIST using hte command:
   `enable stpd {stpd_name}`
4. Create and configure MSTIs using the commands:
   `create stpd stpd_name {description stpd-description}`
   `configure stpd stpd_name mode mstp cist instance`
5. Add VLANs to the MSTIs using one of the following commands:
   a. Manually binding ports
      `configure stpd stpd_name add vlan vlan_name ports [all | port_list] {[dot1d | emistp | pvst-plus]}`
      `configure vlan vlan_name add ports [all | port_list] {tagged {tag} | untagged} stpd stpd_name {[dot1d | emistp | pvst-plus]}`
   b. Automatically binding ports to an STPD when ports are added to a member VLAN
      `enable stpd stpd_name auto-bind vlan vlan_name`
6. Enable the MSTIs using the command:.
   `enable stpd {stpd_name}`

For a more detailed configuration example, see MSTP Configuration Example on page 1089.

## MSTP Operation

To further illustrate how *MSTP* operates and converges, the following figure displays a network with two MSTP regions. Each region contains three MSTP bridges and one *MSTI*. The overall network topology also contains one CIST root bridge (Switch A, which has the lowest bridge ID), one interconnecting 802.1w bridge (Switch D), and 10 full duplex, point-to-point segments. *VLAN* Default spans all of the bridges and segments in the network, VLAN engineering is local to its respective region, and *STPD* S0 is configured as the CIST on all bridges.

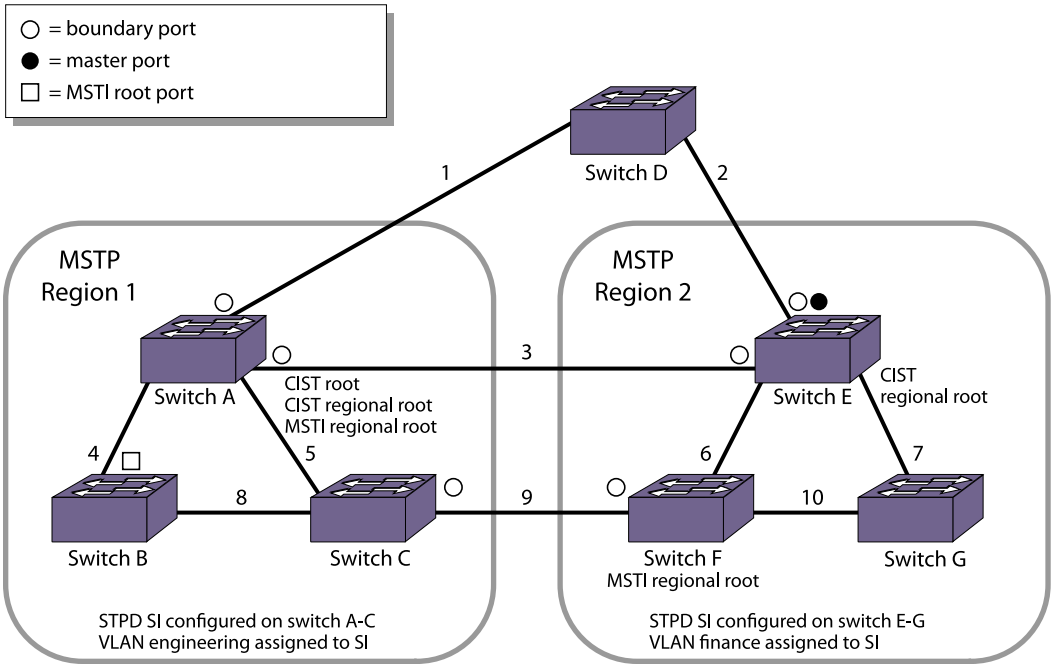



**Figure 162: MSTP Topology with the CIST Root Bridge Contained within a Region**

MSTP Region 1 consists of the following:

- Three bridges named Switch A, Switch B, and Switch C
- One MSTI STPD named S1 with an MSTI ID of 1
- VLAN Engineering mapped to the MSTI STPD, S1
- Switch A as the CIST root bridge (this is the CIST root bridge for all regions)
- Switch A as the CIST regional root bridge
- Switch A as the MSTI regional root bridge
- Three boundary ports that connect to MSTP Region 2 and other 802.1D or 802.1w bridges

MSTP Region 2 consists of the following:

- Three bridges named Switch E, Switch F, and Switch G
- One MSTI STPD named S1 with an MSTI ID of 1

Note

The MSTI ID does not have any significance outside of its region so you can reuse IDs across regions.

- VLAN finance mapped to the MSTI STPD, S1
- Switch E as the CIST regional root bridge
- Switch F as the MSTI regional root bridge
- One master port that connects to the CIST
- Three boundary ports that connect to MSTP Region 1 and other 802.1D or 802.1w bridges

The following sequence describes how the MSTP topology convergences:

1. Determining the CIST root bridge, MSTP regions, and region boundaries.

   Each bridge believes that it is the root bridge, so each bridge initially sends root bridge BPDUs throughout the network. As bridges receive BPDUs and compare vectors, the bridge with the lowest Bridge ID is elected the CIST root bridge. In our example, Switch A has the lowest Bridge ID and is the CIST root bridge.

   The bridges in the MSTP regions (Switches A, B, C, E, F, and G) advertise their region information along with their bridge vectors.

   Segments 1, 3, and 9 receive BPDUs from other regions and are identified as boundary ports for Region 1. Similarly, segments 2, 3, and 9 are identified as boundary ports for Region 2.
2. Controlling boundary ports.

   The CIST regional root is advertised as the Bridge ID in the BPDUs exiting the region. By sending CIST BPDUs across regional boundaries, the CIST views the MSTP regions as virtual 802.1w bridges. The CIST takes control of the boundary ports and only CIST BPDUs enter or exit a region boundary.

   Each MSTP region has a CIST regional root bridge that communicates to the CIST root bridge. The bridge with the lowest path cost becomes the CIST regional root bridge. The port on the CIST regional root bridge that connects to the CIST root bridge is the CIST root port.

For Region 1, Switch A has the lowest cost (0 in this example) and becomes the CIST regional root. Since the bridge is both the CIST root bridge and the CIST regional root bridge, there is no CIST root port on the bridge.

For Region 2, Switch E is the CIST regional root bridge and so a port on that bridge becomes the CIST root port.

3. Identifying MSTI regional roots.

   Each MSTI in a region has an MSTI regional root bridge. MSTI regional roots are selected independently of the CIST root and CIST regional root. The MSTP BPDUs have M-records for each MSTI. Bridges belonging to an MSTI compare vectors in their M-records to elect the MSTI regional root.

4. Converging the CIST.

   The CIST views every region as a virtual bridge and calculates the topology using the 802.1w algorithm. The CIST calculates the topology both inside and outside of a region.

5. Converging MSTIs.

   After the CIST identifies the boundary ports, each MSTI in a domain converge their own trees using 802.1w.

   At this point, all CIST and MSTIs have assigned port roles (Root, Designated, Alternate, and Backup) to their respective spanning trees. All root and designated ports transition to the forwarding state while the remaining ports remain in the discarding state.

   Propagating topology change information is similar to that described for RSTP.

For more information see, Propagating Topology Change Information on page 1066.

For a configuration example, see MSTP Configuration Example on page 1089.

## STP and Network Login

*STP* and network login can be enabled on the same port. This feature can be used to prevent loops while providing redundancy and security on aggregated as well as end switches.

Note

You should be aware that an STP topology change will affect the network login clients. See STP Rules and Restrictions on page 1083 for further information.

The following figure shows STP and network login enabled on ports 2 and 3 of Switch 2 and Switch 3 for a typical aggregation scenario.

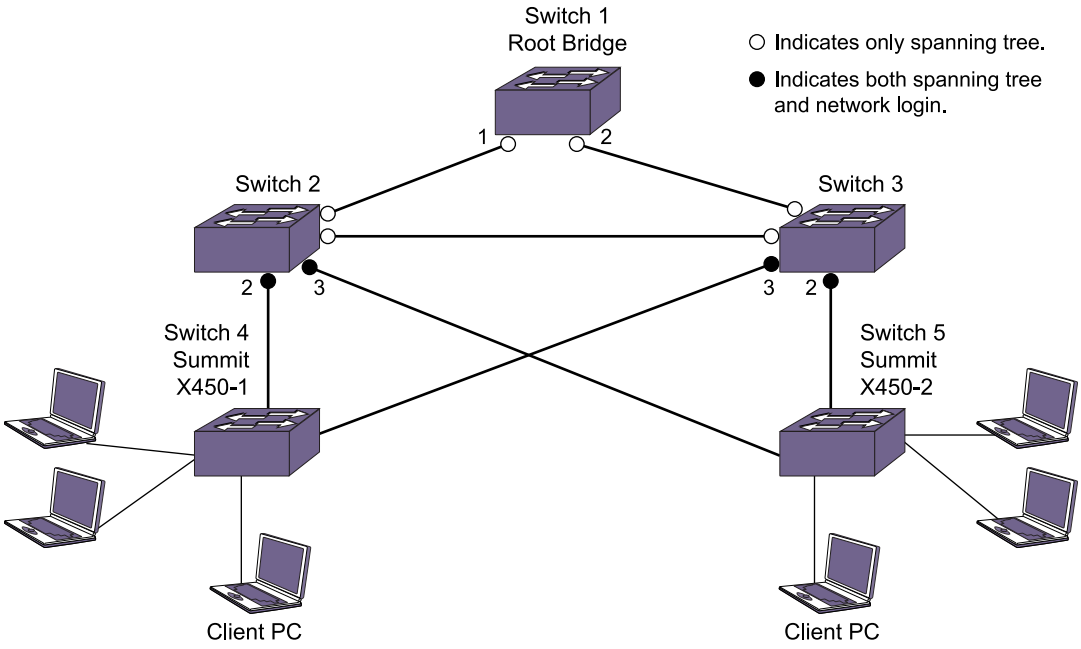



**Figure 163: STP and Network Login Enabled**

This relieves the administrator from having to configure network login on all the edge ports. All the traffic can be monitored and resiliency is provided at the aggregation side.

The following figure shows a typical scenario for protecting loops and monitoring traffic on the edge side.

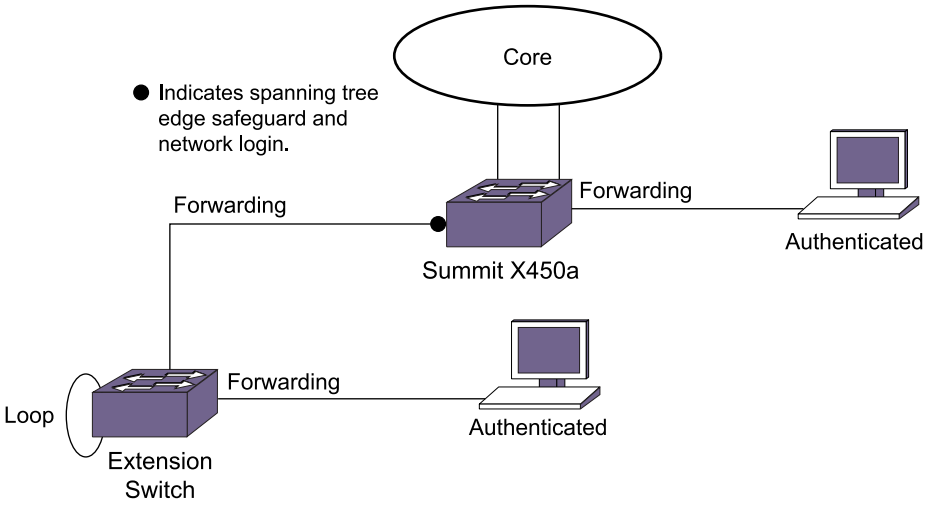



**Figure 164: Traffic Monitoring on the Edge Side**

In huge networks, it is not easy to control or prevent end users from connecting devices other than workstations to the edge ports. This feature helps prevent the network loops that occur when end users connect a switch or hub to the existing edge port in order to increase the number of end user ports.

## STP Rules and Restrictions

This section summarizes the rules and restrictions for configuring *STP* are:

- The carrier *VLAN* must span all ports of the *STPD*. (This is not applicable to *MSTP*.)
- The StpdID must be the VLAN ID of the carrier VLAN; the carrier VLAN cannot be partitioned. (This is not applicable to MSTP.)
- A default VLAN cannot be partitioned. If a VLAN traverses multiple STPDs, the VLAN must be tagged.
- An STPD can carry, at most, one VLAN running in PVST+ mode, and its STPD ID must be identical with that VLAN ID. In addition, the PVST+ VLAN cannot be partitioned.
- The default VLAN of a PVST+ port must be identical to the native VLAN on the PVST+ device connected to that port.
- If an STPD contains both PVST+ and non-PVST+ ports, that STPD must be enabled. If that STPD is disabled, the BPDUs are flooded in the format of the incoming STP port, which may be incompatible with those of the connected devices.
- The 802.1D ports must be untagged and the EMISTP/PVST+ ports must be tagged in the carrier VLAN.
- An STPD with multiple VLANs must contain only VLANs that belong to the same *virtual router (VR)* instance.
- STP and network login operate on the same port as follows:
  - STP (802.1D), RSTP (802.1w), and MSTP (802.1s) support both network login and STP on the same port.
  - At least one VLAN on the intended port should be configured both for STP and network login.
  - STP and network login operate together only in network login ISP mode.
  - When STP blocks a port, network login does not process authentication requests. All network traffic, except STP BPDUs, is blocked.
  - When STP places a port in forwarding state, all network traffic is allowed and network login starts processing authentication requests.
- STP cannot be configured on the following ports:
  - A mirroring target port.
  - A software-controlled redundant port.
- When you are using the older method of enabling STP instead of using EAPSv2 to block the super loop in a shared-port environment, you can continue to do so. In all other scenarios, it is not recommended to use both STP and EAPS on the same port.
- MSTP and 802.1D STPDs cannot share a physical port.
- Only one MSTP region can be configured on a switch.
- In an MSTP environment, a VLAN can belong to one of the MSTIs.
- A VLAN can belong to only one MSTP domain.
- MSTP is not interoperable with PVST+.
- No VLAN can be bound to the CIST.

## Configure STP on the Switch

Note
If you are transitioning from EOS to ExtremeXOS, please note that ExtremeXOS blocks on a more granular (*VLAN*) level, instead of at the port level as EOS does.

To configure basic STP:

1. Create one or more STPDs using the command:

   `create stpd stpd_name {description stpd-description}`
2. Add one or more VLANs to the *STPD* using the command:

   `configure stpd stpd_name add vlan vlan_name | vlan_list ports [all | port_list] {[dot1d | emistp | pvst-plus]}`
3. Define the carrier VLAN using the command:.

   `configure stpd stpd_name tag stpd_tag`

   Note
   The carrier VLAN's ID must be identical to the StpdID.

4. Enable *STP* for one or more STPDs using the command:

   `enable stpd {stpd_name}`
5. After you have created the STPD, you can optionally configure STP parameters for the STPD.

   Note
   You should not configure any STP parameters unless you have considerable knowledge and experience with STP. The default STP parameters are adequate for most networks.

   The following parameters can be configured on each STPD:
   - Hello time (In an *MSTP* environment, configure this only on the CIST.)
   - Forward delay
   - Max age (In an MSTP environment, configure this only on the CIST.)
   - Max hop count (MSTP only)
   - Bridge priority
   - Domain description
   - StpdID (STP, RSTP, EMISTP, and PVST+ only)
   - *MSTI* ID (MSTP only)

   The following parameters can be configured on each port:
   - Path cost
   - Port priority

- Port mode

**Note**

The device supports the RFC 1493 Bridge MIB, RSTP-03, and Extreme Networks STP MIB. Parameters of the s0 default STPD support RFC 1493 and RSTP-03. Parameters of any other STPD support the Extreme Networks STP MIB.

If an STPD contains at least one port not in 802.1D (dot1D) mode, the STPD must be configured with an StpdID.

The following section provides more detailed STP configuration examples, including 802.1D, EMISTP, RSTP, and MSTP.

## STP FDB Flush Criteria

When there are more than 1000 VLANs and more than 70 ports participating in *STP*, the number of messages exchanged between STP/*FDB*/HAL modules can consume a lot of system memory when trying to flush the FDB during a STP topology change. To help avoid this high consumption, you can set the flush type from the default of vlan-and-port to port-based.

To set the flush type, enter the command:

```
configure stpd flush-method [vlan-and-port | port-only]
```

# Display STP Settings

- To display *STPD* settings, use the command:

  ```
  show stpd {stpd_name | detail}
  ```

  To display more detailed information for one or more STPDs, specify the **detail** option.

  This command displays the following information:
  - STPD name
  - STPD state
  - STPD mode of operation
  - Domain description
  - Rapid Root Failover
  - Tag
  - Ports
  - Active VLANs
  - Bridge priority
  - Bridge ID
  - Designated root
  - STPD configuration information

  If you have *MSTP* configured on the switch, this command displays additional information:
  - MSTP Region
  - Format Identifier
  - Revision Level

- Common and Internal Spanning Tree (CIST)
- Total number of *MSTI*

- To display the state of a port that participates in *STP*, use the command:

  `show {stpd} stpd_name ports {[detail | port_list {detail}]}`

  To display more detailed information for one or more ports in the specified STPD, including participating *VLAN*s, specify the **detail** option.

  This command displays the following information:
  - STPD port configuration
  - STPD port mode of operation
  - STPD path cost
  - STPD priority
  - STPD state (root bridge, etc.)
  - Port role (root designated, alternate, etc.)
  - STPD port state (forwarding, blocking, etc.)
  - Configured port link type
  - Operational port link type
  - Edge port settings (inconsistent behavior, edge safeguard setting)
  - MSTP port role (internal or boundary)

  If you have MSTP configured and specify the **detail** option, this command displays additional information:
  - MSTP internal path cost
  - MSTP timers
  - STPD VLAN Settings
- If you have a VLAN that spans multiple STPDs, use the `show {vlan} vlan_name stpd` command to display the STP configuration of the ports assigned to that specific VLAN.

  The command displays the following:
  - STPD port configuration
  - STPD port mode of operation
  - STPD path cost
  - STPD priority
  - STPD state (root bridge, etc.)
  - Port role (root designated, alternate, etc.)
  - STPD port state (forwarding, blocking, etc.)
  - Configured port link type
  - Operational port link type

## STP Configuration Examples

Note

If you are transitioning from EOS to ExtremeXOS, please note that ExtremeXOS blocks on a more granular (*VLAN*) level, instead of at the port level as EOS does.

## Basic 802.1D Configuration Example

The following example:

- Removes ports from the *VLAN* Default that will be added to VLAN Engineering.
- Creates the VLAN Engineering.
- Assigns a VLAN ID to the VLAN Engineering.

> **Note**
> If you do not explicitly configure the VLAN ID in your 802.1D deployment, use the `show vlan` command to see the internal VLAN ID automatically assigned by the switch.

- Adds ports to the VLAN Engineering.
- Creates an *STPD* named Backbone_st.
- Configures the default encapsulation mode of dot1d for all ports added to STPD Backbone_st.
- Enables autobind to automatically add or remove ports from the STPD.
- Assigns the Engineering VLAN to the STPD.
- Assigns the carrier VLAN.
- Enables *STP*.

> **Note**
> To assign the carrier VLAN, the StpdID must be identical to the VLAN ID of the carrier VLAN.

```
configure vlan default delete ports 2:5-2:10
create vlan engineering
configure vlan engineering tag 150
configure vlan engineering add ports 2:5-2:10 untagged
create stpd s1
configure stpd s1 default-encapsulation dot1d
enable stpd s1 auto-bind vlan engineering
configure stpd s1 tag 150
enable stpd s1
```

By default, the port encapsulation mode for user-defined STPDs is emistp. In this example, you set it to dot1d.

## EMISTP Configuration Example

The following figure is an example of EMISTP.

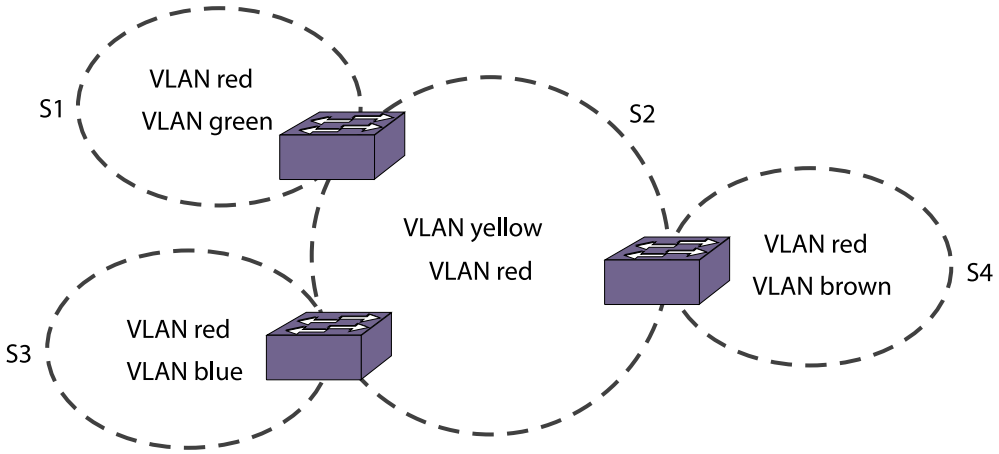



**Figure 165: EMISTP Configuration Example**

Note
By default, all ports added to a user-defined *STPD* are in emistp mode, unless otherwise specified.

The following commands configure the switch located between S1 and S2:

```
create vlan red
configure red tag 100
configure red add ports 1:1-1:4 tagged
create vlan green
configure green tag 200
configure green add ports 1:1-1:2 tagged
create vlan yellow
configure yellow tag 300
configure yellow add ports 1:3-1:4 tagged
create stpd s1
configure stpd s1 add green ports all
configure stpd s1 tag 200
configure stpd s1 add red ports 1:1-1:2 emistp
enable stpd s1
create stpd s2
configure stpd s2 add yellow ports all
configure stpd s2 tag 300
configure stpd s2 add red ports 1:3-1:4 emistp
enable stpd s2
```

## RSTP 802.1w Configuration Example

The following figure is an example of a network with multiple STPDs that can benefit from RSTP.

For RSTP to work:

1. Create an *STPD*.
2. Configure the mode of operation for the STPD.
3. Create the *VLAN*s and assign the VLAN ID and the VLAN ports.
4. Assign the carrier VLAN.
5. Add the protected VLANs to the STPD.

6. Configure the port link types.
7. Enable *STP*.

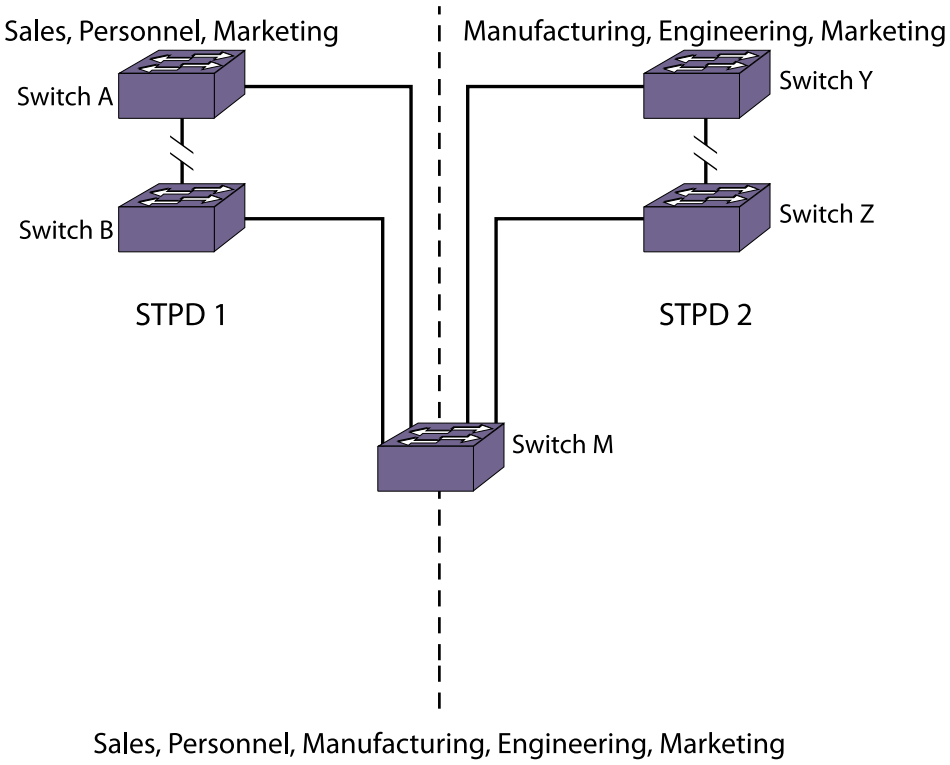



**Figure 166: RSTP Example**

In this example, the commands configure Switch A in STPD1 for rapid reconvergence.

Use the same commands to configure each switch and STPD in the network.

```
create stpd stpd1
configure stpd stpd1 mode dot1w
create vlan sales
create vlan personnel
create vlan marketing
configure vlan sales tag 100
configure vlan personnel tag 200
configure vlan marketing tag 300
configure vlan sales add ports 1:1,2:1 tagged
configure vlan personnel add ports 1:1,2:1 tagged
configure vlan marketing add ports 1:1,2:1 tagged
configure stpd stpd1 add vlan sales ports all
configure stpd stpd1 add vlan personnel ports all
configure stpd stpd1 add vlan marketing ports all
configure stpd stpd1 ports link-type point-to-point 1:1,2:1
configure stpd stpd1 tag 100
enable stpd stpd1
```

## MSTP Configuration Example

You must first configure a CIST before configuring any MSTIs in the region. You cannot delete or disable a CIST if any of the MSTIs are active in the system.

The following figure is an example with multiple STPDs that can benefit from *MSTP*. In this example, we have two MSTP regions that connect to each other and one external 802.1w bridge.

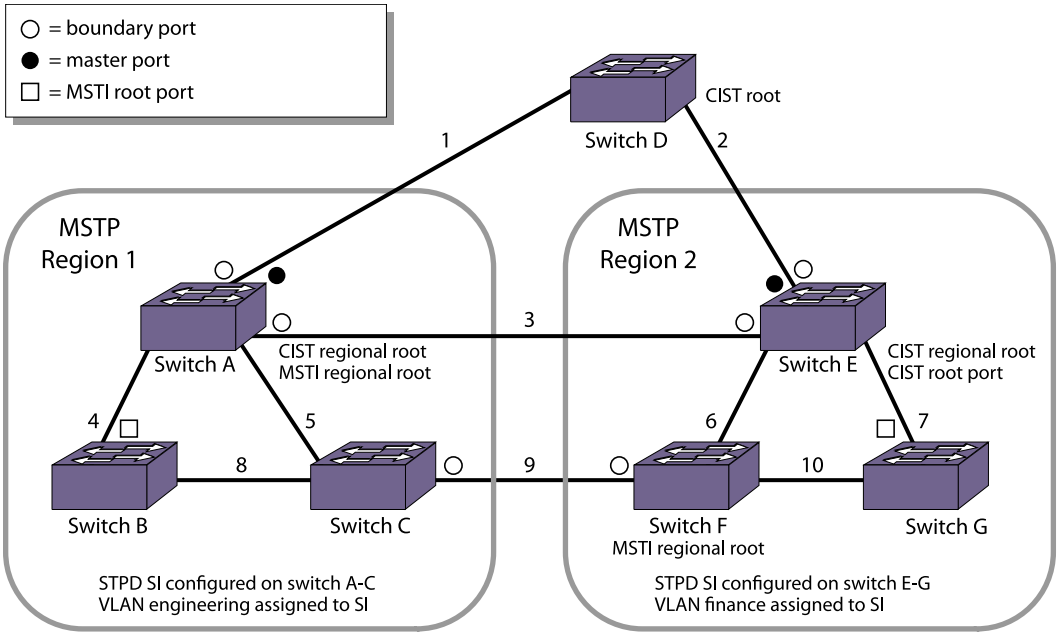



**Figure 167: MSTP Configuration Example**

For MSTP to work, complete the following steps on all switches in Region 1 and Region 2:

- Remove ports from the *VLAN* Default that will be added to VLAN Engineering.
- Create the VLAN Engineering.
- Assign a VLAN ID to the VLAN Engineering.

> **Note**
> If you do not explicitly configure the VLAN ID in your MSTP deployment, use the `show vlan` command to see the internal VLAN ID automatically assigned by the switch.

- Add ports to the VLAN Engineering.
- Create the MSTP region.

> **Note**
> You can configure only one MSTP region on the switch at any given time.

- Create the *STPD* to be used as the CIST, and configure the mode of operation for the STPD.
- Specify the priority for the CIST.
- Enable the CIST.
- Create the STPD to be used as an *MSTI* and configure the mode of operation for the STPD.
- Specify the priority for the MSTI.
- Assign the VLAN Engineering to the MSTI.
- Configure the port link type.
- Enable the MSTI.

On the external switch (the switch that is not in a region):

- Create an STPD that has the same name as the CIST, and configure the mode of operation for the STPD.

- Specify the priority of the STPD.
- Enable the STPD.

Note

In the following sample configurations, any lines marked (Default) represent default settings and do not need to be explicitly configured. STPD s0 already exists on the switch.

In the following example, the commands configure Switch A in Region 1 for MSTP. Use the same commands to configure each switch in Region 1:

```
create vlan engineering
configure vlan engineering tag 2
configure vlan engineering add port 2-3 tagged
configure mstp region region1
create stpd s0 (Default)
disable stpd s0 auto-bind vlan Default
configure stpd s0 mode mstp cist
configure stpd s0 priority 32768 (Default)
enable stpd s0
create stpd s1
configure stpd s1 mode mstp msti 1
configure stpd s1 priority 32768 (Default)
enable stpd s1 auto-bind vlan engineering
configure stpd s0 ports link-type point-to-point 2-3
enable stpd s1
```

In the following example, the commands configure Switch E in Region 2 for MSTP. Use the same commands to configure each switch in Region 2:

```
create vlan finance
configure vlan finance tag 2
configure vlan finance add port 2-3 tagged
configure mstp region region2
create stpd s0 (Default)
configure stpd s0 mode mstp cist
configure stpd s0 priority 32768 (Default)
disable stpd s0 auto-bind vlan Default
enable stpd s0
create stpd s1
configure stpd s1 mode mstp msti 1
configure stpd s1 priority 32768 (Default)
enable stpd s1 auto-bind vlan finance
configure stpd s0 ports link-type point-to-point 2-3
```

In the following example, the commands configure switch D, the external switch. Switch D becomes the CIST root bridge:

```
create stpd s0 (Default)
configure stpd s0 mode dot1w
configure stpd s0 priority 28672
enable stpd s0 auto-bind vlan Default
configure stpd s0 ports link-type point-to-point 4-5
enable stpd s0
```

# ESRP



*ESRP (Extreme Standby Router Protocol)*™ allows multiple switches to provide redundant routing services to users. This chapter discusses how to configure, operate and display the ESRP feature. It also provides detailed configuration examples.

## ESRP Overview

The *ESRP*™, like the Virtual Router Redundancy Protocol (VRRP), allows multiple switches to provide redundant routing services to users.

ESRP is used to eliminate the single point of failure associated with manually configuring a default gateway address on each host in a network. Without using ESRP, if the configured default gateway fails, you must reconfigure each host on the network to use a different router as the default gateway. ESRP provides a redundant path for the hosts. Using ESRP, if the default gateway fails, the backup router assumes forwarding responsibilities.

Note
Support for ESRP operation over IPv6 networks was added in ExtremeXOS release 12.6.

In addition to providing Layer 3 routing redundancy for IP and IPX, ESRP also provides Layer 2 redundancy features for fast failure recovery and to provide for dual-homed system design. In some instances, depending on network system design, ESRP can provide better resiliency than using Spanning Tree Protocol (STP) or Virtual Router Redundancy Protocol (*VRRP (Virtual Router Redundancy Protocol)*). You can use Layer 3 and Layer 2 redundancy features in combination or independently. ESRP is available only on Extreme Networks switches. An example ESRP topology is shown in the following figure.

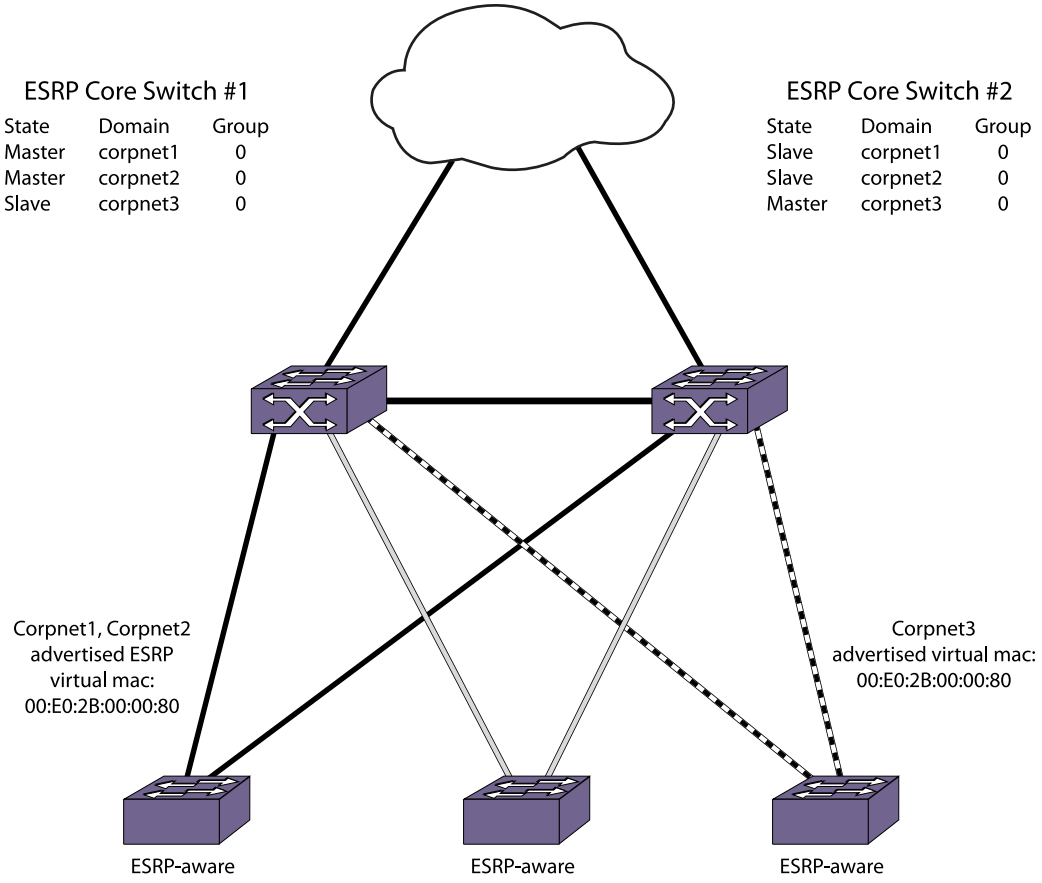



**Figure 168: Example of a Basic ESRP Topology**

In the figure above, ESRP Core Switch #1 and ESRP Core Switch #2 are both configured with three ESRP domains. Each domain represents a separate ESRP instance and supports a unique set of VLANs. Each domain is configured to use one master *VLAN (Virtual LAN)* and can support additional member VLANs. The switches exchange keep-alive packets for each VLAN independently.

Unless groups are configured, each ESRP domain supports two routers, one operating in the master state, and one operating in the slave state. Within an ESRP domain, any ESRP router can become the master, but only one ESRP router can be master at a time. Only the master can actively provide Layer 3 routing and/or Layer 2 switching for each VLAN. The master handles the forwarding, ARP requests, NDP messages, and routing for a particular VLAN. The slave router stands by, ready to take over if the master is no longer available.

Each switch in an ESRP topology has its own unique IP address (or IPX NetID) and a MAC address, which are required for basic IP connectivity. For each ESRP domain, there is a shared virtual IP address or IPX NetID and a MAC address, which are used for network client communications. The virtual IP address or IPX NetID is configured on all ESRP routers in a domain, and it is configured as the default gateway address on network clients in that domain. If the master ESRP router becomes unavailable, the backup ESRP router takes over using the same virtual IP address or IPX NetID.

The topology in the figure above shows that one switch serves as the master for the Corpnet1 and Corpnet2 domains, and the other switch serves as master for the Corpnet3 domain. This topology demonstrates the load sharing capability of ESRP. If one switch served as master for all ESRP domains, all traffic would be routed through that master, and the slave switch would be idle. Dividing the ESRP

domain mastership between routers allows domain clients access to more bandwidth and reduces the likelihood of exceeding the capacity of a single master router.

You can use ESRP to achieve edge-level or aggregation-level redundancy.

Deploying ESRP in this area of the network allows you to simplify your network design, which is important in designing a stable network. ESRP also works well in meshed networks where Layer 2 loop protection and Layer 3 redundancy are simultaneously required.

Note
For complete information about platform support for ESRP, see the *Feature License Requirements* document.

## ESRP Master Election

The system determines the *ESRP* master switch (providing Layer 3 routing and/or Layer 2 switching services for a *VLAN*) using the following default factors:

- Stickiness—The switch with the higher sticky value has higher priority. When an ESRP domain claims master, its sticky value is set to 1 (available only in extended mode).
- Active ports—The switch that has the greatest number of active ports takes highest precedence.
- Tracking information—Various types of tracking are used to determine if the switch performing the master ESRP function has connectivity to the outside world. ExtremeXOS software supports the following types of tracking:
  - VLAN—Tracks any active port connectivity to one designated VLAN. An ESRP domain can track one VLAN, and the tracked VLAN should not be a member of any other ESRP domain in the system.
  - IP unicast route table entry—Tracks specific learned routes from the IP route table.
  - Ping—Tracks *ICMP (Internet Control Message Protocol)* ping connectivity to specified devices.
  - Environment (health checks)—Tracks the environment of the switch, including power supply and chassis temperature.

  If any of the configured tracking mechanisms fail, the master ESRP switch relinquishes status as master, and remains in slave mode for as long as the tracking mechanism continues to fail.
- ESRP priority—This is a user-defined field. The range of the priority value is 0 to 255; a higher number has higher priority, except for 255. The default priority setting is 0. A priority setting of 255 makes an ESRP switch a standby switch that remains in slave mode until you change the priority setting. We recommend this setting for system maintenance. A switch with a priority setting of 255 will never become the master.
- System MAC address—The switch with the higher MAC address has higher priority.
- Active port weight—The switch that has the highest port weight takes precedence. The bandwidth of the port automatically determines the port weight (available only in extended mode).

You can configure the precedence order of the factors used by the system to determine the master ESRP switch. For more information about configuring the ESRP election metrics, see Configuring ESRP Election Algorithms on page 1096.

### *Master Switch Behavior*

If a switch is master, it actively provides Layer 3 routing services to other VLANs, and Layer 2 switching between all the ports of that *VLAN*.

Additionally, the switch exchanges *ESRP* packets with other switches that are in slave mode.

### *Pre-Master Switch Behavior*

A pre-master switch is ready to transition to master, but is going through possible loop detection before changing to the master state.

Upon entering the pre-master state, the switch sends *ESRP* packets to other switches on that same *VLAN*. If the switch finds itself superior to its neighbor, and successfully executes loop detection techniques, the switch transitions to master. This temporary state avoids the possibility of having simultaneous masters.

### *Slave Switch Behavior*

If a switch is in slave mode, it exchanges *ESRP* packets with other switches on that same *VLAN*.

When a switch is in slave mode, it does not perform Layer 3 routing or Layer 2 switching services for the VLAN. From a Layer 3 routing protocol perspective (for example, *RIP (Routing Information Protocol)* or *OSPF (Open Shortest Path First)*), when in slave mode for the VLAN, the switch marks the router interface associated with that VLAN as down. From a Layer 2 switching perspective, no forwarding occurs between the member ports of the VLAN; this prevents loops and maintains redundancy.

If you configure the switch to use the optional ESRP Host Attach configuration, the switch continues Layer 2 forwarding to the master. For more information, see ESRP Host Attach on page 1114.

### *Neutral Switch Behavior*

The neutral state is the initial state entered into by the switch.

In a neutral state, the switch waits for *ESRP* to initialize and run. A neutral switch does not participate in ESRP elections. If the switch leaves the neutral state, it enters the slave state.

### *Electing the Master Switch*

A new master can be elected. This is done in one of the following ways:

- A communicated parameter change
- Loss of communication between master and slave(s)

If a parameter determines the master changes (for example, link loss or priority change), the election of the new master typically occurs within one second. A parameter change triggers a handshake between the routers. As long as both routers agree upon the state transition, new master election is immediate.

If a switch in slave mode loses its connection with the master, a new election occurs (using the same precedence order indicated ESRP Master Election on page 1094 or using a configured precedence order described in Configuring ESRP Election Algorithms on page 1096). The new election typically takes place in three times the defined timer cycle (8 seconds by default).

Before the switch transitions to the master state, it enters a temporary pre-master state. While in the pre-master state, the switch sends *ESRP* PDUs until the pre-master state timeout expires. Depending

upon the election algorithm, the switch may then enter the master or slave state. Traffic is unaffected by the pre-master state because the master continues to operate normally. The pre-master state avoids the possibility of having simultaneous masters.

To configure the pre-master state timeout, use the following command:

```
configure esrp esrpDomain timer premaster seconds
```

**Caution**

Configure the pre-master state timeout only with guidance from Extreme Networks support. Misconfiguration can severely degrade the performance of ESRP and your switch.

### *ESRP Failover Time*

*ESRP* Failover time is largely determined by the following factors:

- ESRP hello timer setting.
- ESRP neighbor timer setting.
- The routing protocol being used for interrouter connectivity if Layer 3 redundancy is used; *OSPF* failover time is faster than *RIP* failover time.

The failover time associated with the ESRP protocol depends on the timer setting and the nature of the failure. The default hello timer setting is two seconds; the range is 2-1024 seconds. The default neighbor timer setting is eight seconds; the range is 3*hello to 1024 seconds. The failover time depends on the type of event that caused ESRP to failover. In most cases, a non-hardware failover is less than one second, and a hardware failover is eight seconds.

If routing is configured, the failover of the particular routing protocol (such as RIP V1, RIP V2, or OSPF) is added to the failover time associated with ESRP.

If you use OSPF, make your OSPF configuration passive. A passive configuration acts as a stub area and helps decrease the time it takes for recalculating the network. A passive configuration also maintains a stable OSPF core.

For more information about the ESRP timers and configuring the ESRP timers, see the *ExtremeXOS 16.2 Command Reference Guide*.

### *Configuring ESRP Election Algorithms*

You configure the switch to use one of 15 different election algorithms to select the *ESRP* master. ESRP uses the default election policy for extended mode. If you have an ESRP domain operating in standard mode, the domain ignores the sticky and weight algorithms.

To change the election algorithm, you must first disable the ESRP domain and then configure the new election algorithm. If you attempt to change the election algorithm without disabling the domain first, an error message appears.

- To disable the ESRP domain, use the following command:

  ```
  disable esrp {esrpDomain}
  ```

- To modify the election algorithm, use the following command:

  ```
  configure esrp esrpDomain add elrp-poll ports [ports | all]
  ```

If you attempt to use an election algorithm not supported by the switch, an error message similar to the following appears:

```
ERROR: Specified election-policy is not supported! Supported Policies:
1. sticky > ports > weight > track > priority > mac 2. ports > track >
priority 3. sticky > ports > track > priority 4. ports > track >
priority > mac 5. sticky > ports > track > priority > mac 6. priority
> mac 7. sticky > priority > mac 8. priority > ports > track > mac 9.
sticky > priority > ports > track > mac 10. priority > track > ports >
mac 11. sticky > priority > track > ports > mac 12. track > ports >
priority 13. sticky > track > ports > priority 14. track > ports >
priority > mac 15. sticky > track > ports > priority > mac
```

### ESRP Election Algorithms

The following table describes the *ESRP* election algorithms. Each algorithm considers the election factors in a different order of precedence. The election algorithms that use sticky and weight are only available in extended mode.

**Table 123: ESRP Election Algorithms**

| Election Algorithm | Description |
|---|---|
| ports > track > priority | Specifies that this ESRP domain should consider election factors in the following order: active ports, tracking information, ESRP priority. |
| ports > track > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: active ports, tracking information, ESRP priority, MAC address.<br><br>**Note:** This is the default election algorithm for standard mode. |
| priority > mac | Specifies that this ESRP domain should consider election factors in the following order: ESRP priority, MAC address. |
| priority > ports > track > mac | Specifies that this ESRP domain should consider election factors in the following order: ESRP priority, active ports, tracking information, MAC address. |
| priority > track > ports > mac | Specifies that this ESRP domain should consider election factors in the following order: ESRP priority, tracking information, active ports, MAC address. |
| sticky > ports > track > priority | Specifies that this ESRP domain should consider election factors in the following order: stickiness, active ports, tracking information, ESRP priority. |
| sticky > ports > track > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, active ports, tracking information, ESRP priority, MAC address. |
| sticky > ports > weight > track > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, active ports, port weight, tracking information, ESRP priority, MAC address. This is the default election algorithm for extended mode. |

**Table 123: ESRP Election Algorithms (continued)**

| Election Algorithm | Description |
|---|---|
| sticky > priority > ports > track > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, ESRP priority, active ports, tracking information, MAC address. |
| sticky > priority > track > ports > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, ESRP priority, tracking information, active ports, MAC address. |
| sticky > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, ESRP priority, MAC address. |
| sticky > track > ports > priority | Specifies that this ESRP domain should consider election factors in the following order: stickiness, tracking information, active ports, ESRP priority. |
| sticky > track > ports > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: stickiness, tracking information, active ports, ESRP priority, MAC address. |
| track > ports > priority | Specifies that this ESRP domain should consider election factors in the following order: tracking information, active ports, ESRP priority. |
| track > ports > priority > mac | Specifies that this ESRP domain should consider election factors in the following order: tracking information, active ports, ESRP priority, MAC address. |

Caution

All switches in the ESRP network must use the same election algorithm; otherwise, loss of connectivity, broadcast storms, or other unpredictable behavior may occur.

Note

If you have a network that contains a combination of switches running ExtremeXOS software and ExtremeWare, only the ports > track > priority >mac election algorithm is compatible with ExtremeWare releases before version 6.0.

## ESRP Domains

*ESRP* domains allow you to configure multiple *VLAN*s under the control of a single instance of the ESRP protocol. By grouping multiple VLANs under one ESRP domain, the ESRP protocol can scale to provide protection to large numbers of VLANs. All VLANs within an ESRP domain simultaneously share the same active and standby router and failover router, as long as one port of each member VLAN belongs to the domain master.

Depending on the election policy used, when a port in a member VLAN belongs to the domain master, the member VLAN ports are considered when determining the ESRP master. You can configure a maximum of 64 ESRP domains in a network.

If you disable an ESRP domain, the switch notifies its neighbor that the ESRP domain is going down, and the neighbor clears its neighbor table. If the master switch receives this information, it enters the

neutral state to prevent a network loop. If the slave switch receives this information, it also enters the neutral state.

ESRP packets do not identify themselves to which domain they belong; you either configure a domain ID or the ESRP domain uses the 802.1Q tag (VLANid) of the master VLAN.

A domain ID in the packet clearly classifies the packet, associates a received ESRP PDU to a specific ESRP domain, and tells the receiving port where the packet came from.

**Note**
Active Ports Count (ports): Total number of active physical ports of master VLANs of the ESRP domain.

## ESRP Groups

ExtremeXOS software supports running multiple instances of *ESRP* within the same *VLAN* or broadcast domain. This functionality is called an *ESRP group*. Although other uses exist, the most typical application for multiple ESRP groups is when two or more sets of ESRP switches are providing fast-failover protection within a subnet.

A maximum of seven distinct ESRP groups can be supported on a single ESRP switch, and a maximum of seven ESRP groups can be defined within the same network broadcast domain. You can configure a maximum of 32 ESRP groups in a network.

For example, two ESRP switches provide Layer 2/Layer 3 connectivity and redundancy for the subnet, while another two ESRP switches provide Layer 2 connectivity and redundancy for a portion of the same subnet. Figure 169 shows ESRP groups.

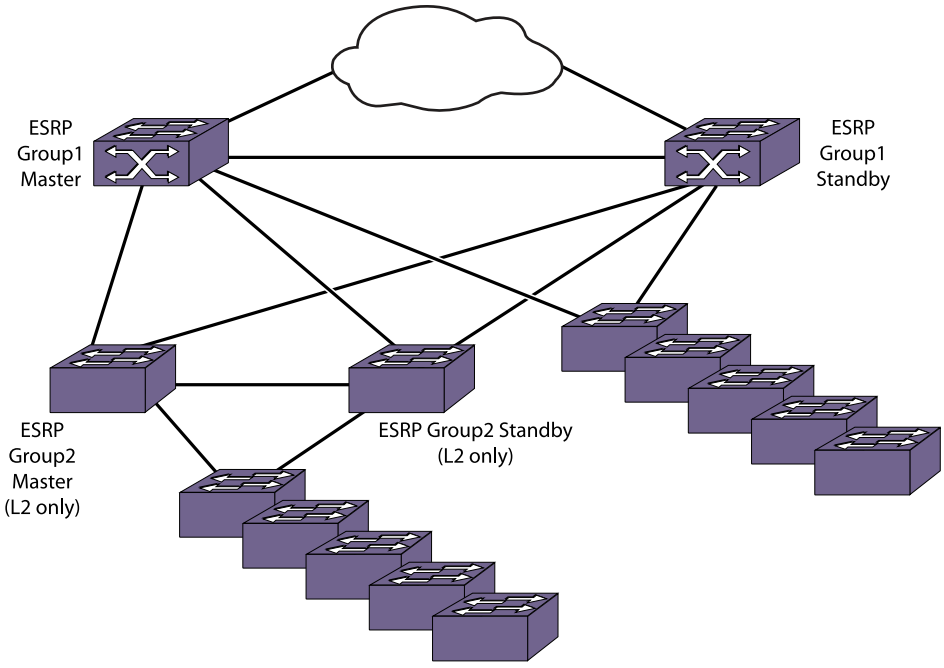



**Figure 169: ESRP Groups**

An additional use for ESRP groups is ESRP HA, as described ESRP Host Attach on page 1114.

## ESRP Extended Mode Features

The *ESRP* extended mode is enabled by default and provides the maximum ESRP feature set.

You can use ESRP extended mode only when all switches that participate in ESRP are running ExtremeXOS software.

Note

If you want the ESRP feature on a switch running ExtremeXOS software to interoperate with a switch running ExtremeWare software, you must configure ESRP in the ExtremeXOS software for ESRP standard mode as described in Configuring Interoperability with ExtremeWare on page 1106.

The following list describes the ESRP extended mode features that are not available in standard mode:

- Handshaking

  In standard mode, events such as link flapping cause the ESRP master switch to generate a large number of packets and to increase processing time.

  To prevent this, extended mode supports handshaking, which occurs when a switch requests a state change, forces its neighbor to acknowledge the change, and the neighbor sends an acknowledgement to the requesting switch. For example, if a slave switch wants to become the master, it enters the pre-master state, notifies the neighbor switch, and forces the neighbor to acknowledge the change. The neighbor then sends an acknowledgement back to the slave switch. While the requesting switch waits for the acknowledgements, future updates are suppressed to make sure the neighbor does not act on incorrect data.
- Stickiness

  In standard mode, if an event causes the ESRP master switch to fail over to the slave, it becomes the new master. If another event occurs, the new master switch returns to the slave and you have experienced two network interruptions.

  To prevent this, extended mode supports the sticky election metric. The default election algorithm uses the sticky metric. For example, if an event causes the ESRP master switch to fail over to the slave, it becomes the new master and has a higher sticky value. If another event occurs, for example adding active ports to the slave, the new master does not fail back to the original master even if the slave has more active ports. After sticky is set on the master, regardless of changes to its neighbor's election algorithm, the new master retains its position. Sticky algorithms provide for fewer network interruptions than non-sticky algorithms. Sticky is set only on the master switch.
- Port weight

  In standard mode, the port count calculation does not take into account the available bandwidth of the ports. For example, a switch with a one Gigabit Ethernet uplink may be unable to become master because another switch has a load-shared group of four fast Ethernet links. The active port count calculation considers only the number of active ports, not the bandwidth of those ports.

  In extended mode, the active port count calculation considers the number of active ports and the port weight configuration considers the bandwidth of those ports. You enable port weight only on the load-shared master port.
- Domain ID

In standard mode, ESRP packets do not contain domain information; therefore, the only information about the packet comes from the receiving port.

The concept of domain ID is applicable only to extended mode. A domain ID in the packet clearly classifies the packet, associates a received ESRP PDU to a specific ESRP domain, and tells the receiving port where the packet came from. In extended mode, you must have a domain ID for each ESRP domain. Each switch participating in ESRP for a particular domain must have the same domain ID configured.

The ESRP domain ID is determined from one of the following user-configured parameters:

- ESRP domain number created with the command:

  #unique_2176
- 802.1Q tag (VLANid) of the tagged master *VLAN*
- Hello messages

  In standard mode, both the master switch and slave switch send periodic ESRP hello messages. This causes an increase in packet processing by both the master and slave.

  In extended mode, the master switch sends periodic ESRP hello messages. This reduces the amount of packet processing, increases the amount of available link bandwidth, and does not impact communicating state changes between switches.

  Note

  If a switch running ExtremeXOS software detects a neighbor switch that is running ExtremeWare, the ExtremeXOS switch toggles to standard mode, and the configured mode of operation remains as extended. For more information, see the following bullet about the ESRP Automatic Toggle Feature.
- ESRP Automatic Toggle Feature

  ESRP includes an automatic toggle feature, which toggles to the same mode of operation as an ESRP neighbor. For example, if an ExtremeXOS switch is operating in ESRP extended mode and detects a neighbor switch that is running ExtremeWare, the ExtremeXOS switch automatically changes to standard mode for that domain. This action causes the switch to enter the neutral state and re-elect the ESRP master.

  Note

  The automatic toggle feature toggles the ESRP operational mode, not the configured mode. If the switch is configured for ESRP extended mode and the switch toggles to standard mode, the switch enters extended mode the next time the switch boots.

## Linking ESRP Switches

When considering system design using *ESRP*, we recommend using a direct link.

Direct links between ESRP switches are useful under the following conditions:

- A direct link can provide a more direct routed path, if the ESRP switches are routing and supporting multiple *VLAN*s where the master/slave configuration is split such that one switch is master for some VLANs and a second switch is master for other VLANs. The direct link can contain a unique router-to-router VLAN/subnet, so that the most direct routed path between two VLANs with different

master switches uses a direct link, instead of forwarding traffic through another set of connected routers.

- A direct link can be used as a highly reliable method to exchange ESRP hello messages, so that the possibility of having multiple masters for the same VLAN is lessened if all downstream Layer 2 switches fail.
- A direct link is necessary for the ESRP host attach option. The direct link is used to provide Layer 2 forwarding services through an ESRP slave switch.

Direct links may contain a router-to-router VLAN, along with other VLANs participating in an ESRP domain. If multiple VLANs are used on the direct links, use 802.1Q tagging. The direct links may be aggregated into a load-shared group, if desired. If multiple ESRP domains share a host port, each VLAN must be in a different ESRP group.

## ESRP-Aware Switches

Extreme Networks switches that are not actively participating in *ESRP* but are connected on a network that has other Extreme Networks switches running ESRP are ESRP-aware. When ESRP-aware switches are attached to ESRP-enabled switches, the ESRP-aware switches reliably perform failover and failback scenarios in the prescribed recovery times.

If Extreme Networks switches running ESRP are connected to Layer 2 switches that are manufactured by third-party vendors, the failover times for traffic local to that segment may appear longer, depending on the application involved and the *FDB (forwarding database)* timer used by the other vendor's Layer 2 switch. ESRP can be used with Layer 2 switches from other vendors, but the recovery times vary.

The *VLAN*s associated with the ports connecting an ESRP-aware switch to an ESRP-enabled switch must be configured using an 802.1Q tag on the connecting port; or, if only a single VLAN is involved, as untagged using the protocol filter 'any.' ESRP does not function correctly if the ESRP-aware switch interconnection port is configured for a protocol-sensitive VLAN using untagged traffic. You can also use port restart in this scenario. For more information, see ESRP Port Restart on page 1113.

The following sections provide information on managing ESRP-aware switches:

- Configuring ESRP-Aware Switches on page 1105
- ESRP Configuration Examples

## ExtremeWare Compatibility

The ExtremeXOS software has two modes of *ESRP* operation: standard and extended.

Select standard ESRP if your network contains some switches running ExtremeWare, others running ExtremeXOS software, and a combination of those switches participating in ESRP. Standard ESRP is backward compatible with and supports the ESRP functionality of ExtremeWare.

Select extended ESRP if your network contains switches running only ExtremeXOS software. Extended mode ESRP supports and is compatible with switches running ExtremeXOS software. By default, the ExtremeXOS software operates in extended mode.

In addition to the modes of operation, ESRP has an auto-toggle feature. Depending on the mode of operation configured on the neighbor switch, the mode of operation at this end will toggle to the same mode of operation as the neighbor.

For more detailed information about the ESRP modes of operation, see Configuring Interoperability with ExtremeWare on page 1106.

# Configuring ESRP

## Guidelines

To participate in *ESRP*, the following must be true:

- A *VLAN* can belong to only one ESRP domain.
- The IP address for the VLANs participating in an ESRP domain must be identical.
- For operation over IPv6, both an IPv6 and an IPv4 address must be present on the master VLAN for every participating router.
- All switches in the ESRP network must use the same election algorithm, otherwise loss of connectivity, broadcast storms, or other unpredictable behavior may occur.
- If you have an untagged master VLAN, you must specify an ESRP domain ID. The domain ID must be identical on all switches participating in ESRP for that particular domain.
- If you have a tagged master VLAN, ESRP uses the 802.1Q tag (VLANid) of the master VLAN for the ESRP domain ID. If you do not use the VLANid as the domain ID, you must specify a different domain ID. As previously described, the domain ID must be identical on all switches participating in ESRP for that particular domain.

Note

If you configure the *OSPF* routing protocol and ESRP, you must manually configure an OSPF router identifier. Be sure that you configure a unique OSPF router ID on each switch running ESRP. For more information, see OSPF

We recommend that all switches participating in ESRP run the same version of ExtremeXOS software.

## Configuration Overview

The following procedure can be used to configure a simple ESRP topology:

1. Create and configure a *VLAN* that will become the master VLAN. (See VLANs on page 502.)
2. As needed, create and configure the VLANs that will become the member VLANs. (See VLANs on page 502.)
3. Create the *ESRP* domain as described in Creating and Deleting an ESRP Domain on page 1104.
4. If your configuration requires an ESRP domain ID, configure it as described in Configuring the ESRP Domain ID on page 1104.
5. Add the master VLAN to the ESRP domain as described in Adding and Deleting a Master VLAN on page 1105.
6. If your configuration requires member VLANs, add the member VLANs to the ESRP domain as described in Adding and Deleting a Member VLAN on page 1105.

7. Enable ESRP for the specified ESRP domain as described in Enabling and Disabling an ESRP Domain on page 1105.

You can also configure other ESRP domain parameters, including ESRP:

- Mode of operation as described in Configuring Interoperability with ExtremeWare on page 1106.
- Timers as described in the *ExtremeXOS 16.2 Command Reference Guide*.
- Election algorithms as described in Configuring ESRP Election Algorithms on page 1096.
- Tracking as described in ESRP Tracking on page 1110.
- Port restart as described in ESRP Port Restart on page 1113.
- Host attach as described in ESRP Host Attach on page 1114.
- Groups as described in ESRP Groups on page 1099.

For more detailed information about all of the commands used to create, configure, enable, and disable an ESRP domain, refer to the *ExtremeXOS 16.2 Command Reference Guide*.

## Creating and Deleting an ESRP Domain

You can specify a unique *ESRP* domain name to identify each ESRP domain in your network.

- To create an ESRP domain, use the following command:

  ```
  create esrp esrp_domain {type [vpls-redundancy | standard]}
  ```

  The *esrpDomain* parameter is a character string of up to 32 characters that identifies the ESRP domain to be created.

  **Note**

  If you use the same name across categories (for example, *STPD (Spanning Tree Domain)* and ESRP names) we recommend that you specify the appropriate keyword as well as the actual name. If you do not specify the keyword, the switch may display an error message.

- To delete an ESRP domain, use the following command:

  ```
  delete esrp esrpDomain
  ```

## Configuring the ESRP Domain ID

If you choose not use the 802.1Q tag (VLANid) of the master *VLAN*, or you have an untagged master VLAN, you must create a domain ID before you can enable the *ESRP* domain. For more information about ESRP domains and the ESRP domain ID, see ESRP Domains on page 1098.

- To configure an ESRP domain ID, use the following command:

  ```
  configure esrp esrpDomain domain-id number
  ```

  The *number* parameter specifies the number of the domain ID. The user-configured ID range is 4096 through 65,535.

## Adding and Deleting a Master VLAN

The master *VLAN* is the VLAN on the *ESRP* domain that exchanges ESRP PDUs and data between a pair of ESRP-enabled devices. You must configure one master VLAN for each ESRP domain, and a master VLAN can belong to only one ESRP domain.

- To add a master VLAN to an ESRP domain, use the following command:

  `configure esrp esrpDomain add master vlan_name`

  The *esrpDomain* parameter specifies the name of the ESRP domain, and the *vlan_name* parameter specifies the name of the master VLAN.
- To delete a master VLAN, you must first disable the ESRP domain before removing the master VLAN using the `disable esrp {esrpDomain}` command.
- To delete a master VLAN from an ESRP domain, use the following command:.

  `configure esrp esrpDomain delete master vlan_name`

## Adding and Deleting a Member VLAN

The member *VLAN* can belong to only one *ESRP* domain, and you configure zero or more member VLANs for each ESRP domain.The state of the ESRP device determines whether the member VLAN is in the forwarding or blocking state.

- To add a member VLAN to an ESRP domain. use the following command:

  `configure esrp esrpDomain add member vlan_name`

  The *esrpDomain* parameter specifies the name of the ESRP domain, and the *vlan_name* parameter specifies the name of the member VLAN.
- To delete a member VLAN from an ESRP domain, use the following command:

  `configure esrp esrpDomain delete member vlan_name`

## Enabling and Disabling an ESRP Domain

- To enable a specific *ESRP* domain, use the following command:

  `enable esrp esrpDomain`
- To disable a specific ESRP domain, use the following command:

  `disable esrp {esrpDomain}`

## Configuring ESRP-Aware Switches

For an Extreme Networks switch to be *ESRP*-aware, you must create an ESRP domain on the aware switch, add a master *VLAN* to that ESRP domain, add a member VLAN to that ESRP domain if configured, and configure a domain ID if necessary.

To participate as an ESRP-aware switch, the following must be true:

- The ESRP domain name must identical on all switches (ESRP-enabled and ESRP-aware) participating in ESRP for that particular domain.
- The master VLAN name and IP address must be identical on all switches (ESRP-enabled and ESRP-aware) participating in ESRP for that particular domain.

- If configured, the member VLAN name and IP address must be identical on all switches (ESRP-enabled and ESRP-aware) participating in ESRP for that particular domain.
- The domain ID must be identical on all switches (ESRP-enabled or ESRP-aware) participating in ESRP for that particular domain.
- If you have an untagged master VLAN, you must specify an ESRP domain ID.
- If you have a tagged master VLAN, ESRP uses the 802.1Q tag (VLANid) of the master VLAN for the ESRP domain ID. If you do not use the VLANid as the domain ID, you must specify a different domain ID.

Note

Before you begin, make a note of the ESRP domain parameters on the ESRP-enabled switch. That way you can easily refer to your notes while creating the ESRP domain on the ESRP-aware switch.

To configure an ESRP-aware switch:

1. Create an ESRP domain using the command:

   ```
   create esrp esrp_domain {type [vpls-redundancy | standard]}
   ```

   For complete information about software licensing for this feature, see the *Feature License Requirements* document.
2. Add a master VLAN to your ESRP domain using the command:

   ```
   configure esrp esrpDomain add master vlan_name
   ```
3. If configured, add the member VLANs to your ESRP domain using the command:

   ```
   configure esrp esrpDomain add member vlan_name
   ```
4. If necessary, configure a domain ID for the ESRP domain using the command:

   ```
   configure esrp esrpDomain domain-id number
   ```

## Configuring Interoperability with ExtremeWare

The *ESRP* feature in ExtremeXOS software supports interoperability with switches that are running ExtremeWare software. If you want the ESRP feature on an ExtremeXOS switch to interoperate with an ExtremeWare switch, you must configure ESRP in the ExtremeXOS software for ESRP standard mode using the following command:

```
configure esrp mode [extended | standard]
```

Note

By default, the ESRP feature operates in extended mode. ESRP extended mode provides additional features that are not available in standard mode. Use ESRP extended mode only if all switches that participate in ESRP are running ExtremeXOS software. For more information on additional features supported in extended mode, see ESRP Extended Mode Features on page 1100.

ExtremeWare switches forward only those ESRP hello messages that apply to the ESRP group to which the switch belongs. ExtremeWare switches do not forward ESRP hello messages for other ESRP groups in the same *VLAN*. This limitation does not apply to ExtremeXOS switches operating in standard mode.

# Operation with Other ExtremeXOS Features

## ESRP and IP Multinetting

When configuring *ESRP* and IP multinetting on the same switch, the same set of IP addresses must be configured for all involved VLANs.

## ESRP and STP

A switch running *ESRP* should not simultaneously participate in *STP (Spanning Tree Protocol)* for the same *VLAN*(s). Other switches in the VLAN being protected by ESRP may run STP; the switch running ESRP forwards, but does not filter, STP BPDUs. Therefore, you can combine ESRP and STP on a network and a VLAN, but you must do so on separate devices. You should be careful to maintain ESRP connectivity between ESRP master and slave switches when you design a network that uses ESRP and STP.

## ESRP and VRRP

Do not configure *ESRP* and VRRP on the same *VLAN* or port. This configuration is not allowed or supported.

## ESRP Groups and Host Attach

*ESRP* domains that share ESRP HA ports must be members of different ESRP groups.

## Port Configurations and ESRP

The following ports cannot be part of a *VLAN* that participates in an *ESRP* domain:

- A mirroring target port
- A software-controlled redundant port
- A Netlogin port

In addition, the following ESRP ports cannot be a mirroring, software-controlled redundant port, or Netlogin port:

- Host Attach port
- Don't-count port (This port has a port weight of 0.)
- Restart port

## Using ELRP with ESRP

Extreme Loop Recovery Protocol (ELRP) is a feature of ExtremeXOS software that allows you to prevent, detect, and recover from Layer 2 loops in the network.

You can use ELRP with other protocols, including *ESRP*.

Note

The ExtremeXOS software does not support ELRP and Network Login on the same port. When used on a VPLS service *VLAN*, ELRP does not detect loops involving the VPLS pseudowires.

For more information about standalone ELRP, see Using ELRP to Perform Loop Tests on page 1570.

With ELRP, each switch, except for the sender, treats the ELRP protocol data unit (PDU) as a Layer 2 multicast packet. The sender uses the source and destination MAC addresses to identify the packet it sends and receives. When the sender receives its original packet back, that triggers loop detection and prevention. After a loop is detected, the loop recovery agent is notified of the event and takes the necessary actions to recover from the loop. ELRP operates only on the sending switch; therefore, ELRP operates transparently across the network.

How a loop recovers is dependent upon the protocol that uses the loop detection services provided by ELRP.

If you are using ELRP in an ESRP environment, ESRP may recover by transitioning the ESRP domain from master to slave. The following sections describe how ESRP uses ELRP to recover from a loop and the switch behavior:

- Using ELRP with ESRP to Recover Loops on page 1108
- Configuring ELRP on page 1109
- Displaying ELRP Information on page 1110

## *Using ELRP with ESRP to Recover Loops*

ELRP sends loop-detect packets to notify *ESRP* about loops in the network.

In an ESRP environment, when the current master goes down, one of the slaves becomes the master and continues to forward Layer 2 and Layer 3 traffic for the ESRP domain. If a situation occurs when a slave incorrectly concludes that the master is down, the slave incorrectly assumes the role of master. This introduces more than one master on the ESRP domain which causes temporary loops and disruption in the network.

### ELRP on an ESRP Pre-Master Switch

A pre-master switch is an ESRP switch that is ready to transition to master but is going through possible loop detection.

A pre-master periodically sends out ELRP loop-detect packets (ELRP PDUs) for a specified number of times and waits to make sure that none of the sent ELRP PDUs are received. Transition to master occurs only after this additional check is completed. If any of the ELRP PDUs are received, the switch transitions from pre-master to slave state. You configure pre-master ELRP loop detection on a per ESRP domain basis.

### ELRP on an ESRP Master Switch

A master switch is an ESRP switch that sends ELRP PDUs on its ESRP domain ports.

If the master switch receives an ELRP PDU that it sent, the master transitions to the slave. While in the slave state, the switch transitions to the pre-master rate and periodically checks for loops before transitioning to the master. The pre-master process is described in ELRP on an ESRP Pre-Master Switch on page 1108. You configure the master ELRP loop detection on a per ESRP domain basis.

### *Configuring ELRP*

By default, ELRP is disabled. The following sections describe the commands used to configure ELRP for use with *ESRP*:

- Configuring Pre-Master Polling on page 1109
- Configuring Master Polling on page 1109
- Configuring Ports on page 1109

Note

When used on a virtual private LAN service (VPLS) *VLAN*, ELRP does not detect loops involving the VPLS pseudowires.

#### Configuring Pre-Master Polling

If you enable the use of ELRP by *ESRP* in the pre-master state, ESRP requests ELRP packets sent to ensure that there is no loop in the network before changing to the master state. If no packets are received, there is no loop in the network. By default, the use of ELRP by ESRP in the pre-master state is disabled.

- To enable the use of ELRP by ESRP in the pre-master state on a per-ESRP domain basis, and to configure how often and how many ELRP PDUs are sent in the pre-master state, use the following command:

  ```
  configure esrp esrpDomain elrp-premaster-poll enable {count count | interval interval}
  ```

- To disable the use of ELRP by ESRP in the pre-master state, use the following command:

  ```
  configure esrp esrpDomain elrp-premaster-poll disable
  ```

#### Configuring Master Polling

If you enable the use of ELRP by *ESRP* in the master state, ESRP requests that ELRP packets are periodically sent to ensure that there is no loop in the network while ESRP is in the master state. By default, the use of ELRP by ESRP in the master state is disabled.

- To enable the use of ELRP by ESRP in the master state on a per-ESRP domain basis, and to configure how often the master checks for loops in the network, use the following command:

  ```
  configure esrp esrpDomain elrp-master-poll enable {interval interval}
  ```

- To disable the use of ELRP by ESRP in the master state, use the following command:

  ```
  configure esrp esrpDomain elrp-master-poll disable
  ```

#### Configuring Ports

You can configure one or more ports of an *ESRP* domain where ELRP packet transmission is requested by ESRP. This allows the ports in your network that might experience loops, such as ports that connect

to the master, slave, or ESRP-aware switches, to receive ELRP packets. You do not need to send ELRP packets to host ports.

Note

The ExtremeXOS software does not support ELRP and Network Login on the same port.

By default, all ports of the ESRP domain have ELRP transmission enabled on the ports.

If you change your network configuration, and a port no longer connects to a master, slave, or ESRP-aware switch, you can disable ELRP transmission on that port.

- To disable ELRP transmission, use the following command:

  `configure esrp esrpDomain delete elrp-poll ports [ports | all]`
- To enable ELRP transmission on a port, use the following command:

  `configure esrp esrpDomain add elrp-poll ports [ports | all]`

*Displaying ELRP Information*

To display summary ELRP information, use the following command:

```
show elrp
```

In addition to displaying the enabled/disabled state of ELRP, the command displays the total number of:

- Clients registered with ELRP
- ELRP packets transmitted
- ELRP packets received

For more information about the output associated with the `show elrp` command, see the *ExtremeXOS 16.2 Command Reference Guide*.

# Advanced ESRP Features

## ESRP Tracking

Tracking information is used to track various forms of connectivity from the *ESRP* switch to the outside world.

*ESRP Environment Tracking*

You can configure *ESRP* to track hardware status. If a power supply fails, if the chassis is overheating, or if a non-fully loaded power supply is detected, the priority for the ESRP domain will change to the failover settings.

Note

ExtremeXOS software determines the maximum available power required for the switch by calculating the number of power supplies and the power required by the installed modules. Enabling environmental tracking on the switch without enough power budget causes tracking to fail. In this case, the tracking failure occurs by design.

To configure the failover priority for an ESRP domain:

1. Set the failover priority using the following command:

   `configure esrp esrpDomain add track-environment failover priority`

2. Assign the priority flag precedence over the active ports count using the following command:

   `configure esrp esrpDomain election-policy [ports > track > priority | ports > track > priority > mac | priority > mac | priority > ports > track > mac | priority > track > ports > mac | sticky > ports > track > priority | sticky > ports > track > priority > mac | sticky > ports > weight > track > priority > mac | sticky > priority > mac | sticky > priority > ports > track > mac | sticky > priority > track > ports > mac | sticky > track > ports > priority | sticky > track > ports > priority > mac | track > ports > priority | track > ports > priority > mac]`

Because the priority of both ESRP domains are set to the same value, ESRP will use the active ports count to determine the master ESRP domain.

### *ESRP VLAN Tracking*

You can configure an *ESRP* domain to track port connectivity to a specified *VLAN* as criteria for ESRP failover. The number of VLAN active ports are tracked. If the switch is no longer connected to the specified VLAN, the switch automatically relinquishes master status and remains in slave mode. You can track a maximum of one VLAN.

- To add a tracked VLAN, use the following command:

  `configure esrp esrpDomain add track-vlan vlan_name`

- To delete a tracked VLAN, use the following command:

  `configure esrp esrpDomain delete track-vlan vlan_name`

### *ESRP Unicast Route Table Tracking*

You can configure *ESRP* to track specified IPv4 routes in the route table as criteria for ESRP failover. If all of the configured routes are not available within the route table, the switch automatically relinquishes master status and remains in slave mode. You can track a maximum of eight routes per route table.

Note

ESRP route tracking is not supported for IPv6 destinations..

- To add a tracked route, use the following command:

  `configure esrp esrpDomain add track-iproute ipaddress/masklength`

- To delete a tracked route, use the following command:

  `configure esrp esrpDomain delete track-iproute ipaddress/masklength`

### *ESRP Ping Tracking*

You can configure *ESRP* to track connectivity using a simple ping to any IPv4 device. This may represent the default route of the switch, or any device meaningful to network connectivity of the master ESRP

switch. The switch automatically relinquishes master status and remains in slave mode if a ping keepalive fails. You can configure a maximum of eight ping tracks.

Note

ESRP ping tracking is not supported for IPv6 destinations. The ESRP ping tracking option cannot be configured to ping an IP address within an ESRP *VLAN* subnet. It should be configured on some other normal VLAN across the router boundary.

- To configure ping tracking, use the following command:

  `configure esrp esrpDomain add track-ping ipaddress frequency seconds miss misses`

  The *seconds* parameter specifies the number of seconds between ping requests. The range is one to 600 seconds.

  The *misses* parameter specifies the number of consecutive ping failures that will initiate failover to an ESRP slave. The range is one to 256 pings.
- To disable ping tracking, use the following command:

  `configure esrp esrpDomain delete track-ping ipaddress`

### *Displaying ESRP Tracking Information*

You can view the status of *ESRP* tracking on a per domain basis. The information displayed includes the type of tracking used by the ESRP domain and how you configured the tracking option.

To view the status of tracked devices, use the following command:

`show esrp name`

### *ESRP Tracking Example*

The following figure is an example of *ESRP* tracking.

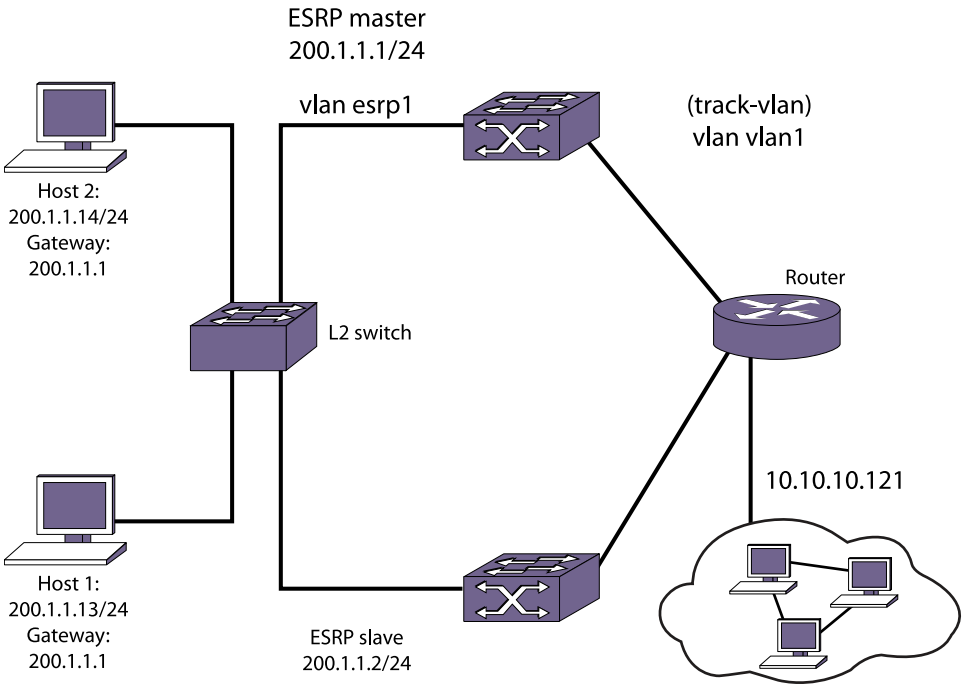



**Figure 170: ESRP Tracking**

- To configure *VLAN* tracking, use the following command:
  ```
  configure esrp esrp1 add track-vlan vlan1
  ```
  Using the tracking mechanism, if VLAN1 fails, the ESRP master realizes that there is no path to the upstream router via the master switch and implements an ESRP failover to the slave switch.
- To configure route table tracking, use the following command:
  ```
  configure esrp esrp1 add track-iproute 10.10.10.0/24
  ```
  The IPv4 route specified in this command must exist in the IP routing table. When the route is no longer available, the switch implements an ESRP failover to the slave switch.
- To configure ping tracking, use the following command:
  ```
  configure esrp esrp1 add track-ping 10.10.10.121 frequency 2 miss 2
  ```
  The specified IPv4 address is tracked. If the fail rate is exceeded, the switch implements an ESRP failover to the slave switch.

## ESRP Port Restart

You can configure *ESRP* to restart ports in the ESRP master domain when the downstream switch is from a third-party vendor. This action takes down and restarts the port link to clear and refresh the downstream ARP table.

- To configure port restart, use the following command:
  ```
  configure esrp ports ports restart
  ```
- To disable port restart, use the following command:
  ```
  configure esrp ports ports no-restart
  ```

If a switch becomes a slave, ESRP takes down (disconnects) the physical links of member ports that have port restart enabled. The disconnection of these ports causes downstream devices to remove the ports from their *FDB* tables. This feature allows you to use ESRP in networks that include equipment from other vendors. After 2 seconds, the ports re-establish connection with the ESRP switch.

- To remove a port from the restart configuration, delete the port from the *VLAN* and re-add it.

## ESRP Host Attach

*ESRP* host attach (HA) is an optional ESRP configuration that allows you to connect active hosts directly to an ESRP master or slave switch.

Normally, the Layer 2 redundancy and loop prevention capabilities of ESRP do not allow packet forwarding from the slave ESRP switch. ESRP HA allows configured ports that do not represent loops to the network to continue Layer 2 operation independent of their ESRP status.

ESRP HA is designed for redundancy for dual-homed server connections. HA allows the network to continue Layer 2 forwarding regardless of the ESRP status. Do not use ESRP HA to interconnect devices on the slave ESRP switch instead of connecting directly to the ESRP master switch.

The ESRP HA option is useful if you are using dual-homed network interface cards (NICs) for server farms, as shown in the following figure. The ESRP HA option is also useful where an unblocked Layer 2 environment is necessary to allow high-availability security.

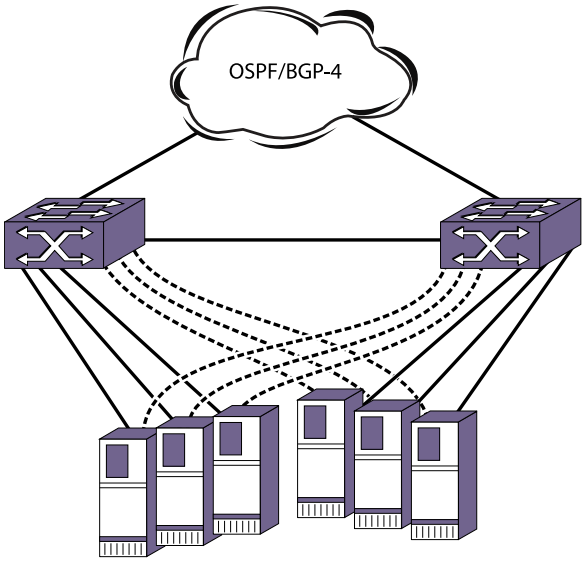



**Figure 171: ESRP Host Attach**

ESRP VLANs that share ESRP HA ports must be members of different ESRP groups. Each port can have a maximum of seven VLANs.

If you use load sharing with the ESRP HA feature, configure the load-sharing group first and then enable HA on the group.

Other applications allow lower-cost redundant routing configurations because hosts can be directly attached to the switch involved with ESRP. HA also requires at least one link between the master and the slave ESRP switch for carrying traffic and to exchange ESRP hello packets.

Note

Do not use the ESRP HA feature with the following protocols: *STP* or *VRRP*. A broadcast storm may occur.

ESRP domains that share ESRP HA ports must be members of different ESRP groups.

To configure a port to be a host port, use the following command: `configure esrp ports ports mode [host | normal]`

## ESRP Port Weight and Don't Count

In an *ESRP* domain, the switch automatically calculates the port weight based on the bandwidth of the port. ESRP uses the port weight to determine the master ESRP switch.

For load-shared ports, configure the master port in the load-share group with the port weight. A load-shared port has an aggregate weight of all of its member ports. If you add or delete a load-shared port (or trunk), the master load-shared port weight is updated.

If you do not want to count host ports and normal ports as active, configure the ESRP port weight on those ports. Their weight becomes 0 and that allows the port to be part of the *VLAN*, but if a link failure occurs, it will not trigger a reconvergence. With this configuration, ESRP experiences fewer state changes due to frequent client activities like rebooting and unplugging laptops. This port is known as a don't-count port.

- To configure the port weight on either a host attach port or a normal port, use the following command:
  `configure esrp ports ports weight [auto | port-weight]`

## Selective Forwarding

An *ESRP*-aware switch floods ESRP PDUs from all ports in an ESRP-aware *VLAN*. This flooding creates unnecessary network traffic because some ports forward ESRP PDUs to switches that are not running the same ESRP groups. You can select the ports that are appropriate for forwarding ESRP PDUs by configuring selective forwarding on an ESRP-aware VLAN and thus reduce this excess traffic. Configuring selective forwarding creates a port list of only those ports that forward to the ESRP groups that are associated with an ESRP-aware VLAN. This ESRP-aware port list is then used for forwarding ESRP PDUs.

Note

We recommend keeping the default settings unless you have considerable knowledge and experience with ESRP.

- To configure selective forwarding, use the following command:
  `configure esrp domain aware add selective-forward-ports portlist {group group number}`

When an ESRP-aware switch receives an ESRP PDU on a domain, the software looks up the group to which the PDU belongs. If the group is found, the ESRP-aware switch processes the PDU then and forwards it according to the group's specified aware selective forwarding port list. If no selective forwarding port list is configured, the switch forwards the PDU from all of the ports of the domain's master VLAN. If the group is not found, the PDU is forwarded on all ports.

When a user adds one or more ports to the ESRP-aware port list (for example, 5:1 and 6:2) that are not part of the master VLAN, the following message appears:

```
Warning: Port 5:1, 6:2 not currently a member of master vlan
```

The ports will still be added to the ESRP-aware port list; however, PDUs will not be forwarded out of those ports until they are added to the master VLAN.

- To disable selective forwarding, use the following command:

  `configure esrp domain aware delete selective-forward-ports all| portlist {group group number}`

*Display Selective Forwarding Information*

To display all selective forwarding information for a given domain, use the following command:

`show esrp domain aware {selective-forward-ports | statistics}`

## Display ESRP Information

- To view *ESRP* information, use the command `show esrp`

  Output from this command includes:

  - The operational state of an ESRP domain and the state of its neighbor.
  - ESRP port configurations.
- To view more detailed information about an ESRP domain on an ESRP enabled switch or an ESRP aware switch, use the following command and specify the domain name:

  `show esrp { {name} | {type [vpls-redundancy | standard]} }`
- To view ESRP counter information for a specific domain, use the following command:

  `show esrp {name} counters`
- To view ESRP-aware information for a specific domain (including the group number, MAC address for the master, and the age of information), use the following command:

  `show esrp domain aware {selective-forward-ports | statistics}`

## ESRP Configuration Examples

### Single Domain Using Layer 2 and Layer 3 Redundancy

The example shown in the following figure uses four Extreme Networks devices as edge switches that perform Layer 2 switching for *VLAN* Sales.

The edge switches are dual-homed two switches that are configured to support *ESRP* domain esrp1. The ESRP switches perform Layer 2 switching and Layer 3 routing between the edge switches and the outside world. Each ESRP switch has the VLAN Sales configured using the identical IP address. The

ESRP switches then connect to the routed enterprise normally, using the desired routing protocol (for example, *RIP* or *OSPF*).

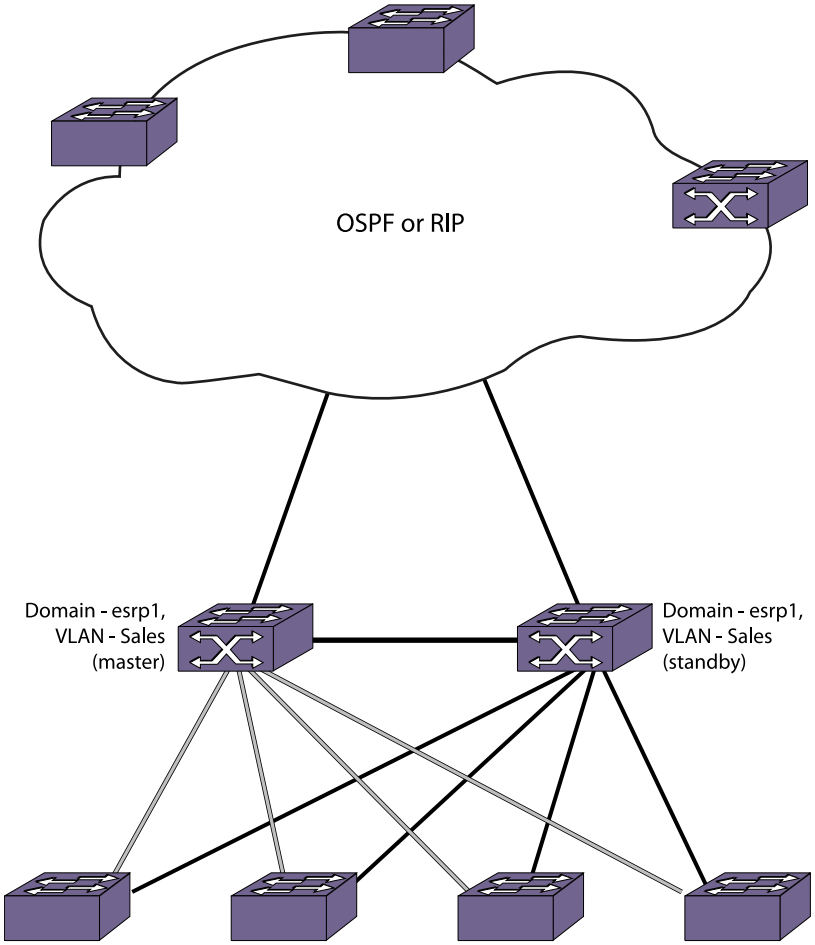



**Figure 172: Single ESRP Domain Using Layer 2 and Layer 3 Redundancy**

In the figure above, the ESRP master performs both Layer 2 switching and Layer 3 routing services for VLAN Sales. To prevent bridging loops in the VLAN, the ESRP slave performs no switching or routing for VLAN Sales while the ESRP master is operating.

There are four paths between each ESRP switch and the edge switches for VLAN Sales. All the paths are used to send ESRP packets, allowing for four redundant paths for communication. The edge switches, being ESRP-aware, allow traffic within the VLAN to failover quickly because these edge switches sense when a master/slave transition occurs and flush *FDB* entries associated with the uplinks to the ESRP-enabled switches.

This example assumes the following:

- ESRP election algorithm used is the default for standard mode (ports > track > priority > mac).
- The inter-router backbone is running OSPF, with other routed VLANs already properly configured. Similar commands would be used to configure a switch on a network running RIP.
- Ports added to the VLAN have already been removed from VLAN default.

- The same IP address is specified for all VLANs participating in ESRP.
- The master is determined by the programmed MAC address of the switch because the number of active links for the VLAN and the priority are identical for both switches.

**Note**

If your network has switches running ExtremeWare and ExtremeXOS software participating in ESRP, we recommend that the ExtremeXOS switches operate in ESRP standard mode. To change the mode of operation, use the command:

`configure esrp mode [extended | standard]`.

The following commands are used to configure both ESRP switches:

```
create vlan sales
configure vlan sales add ports 1:1-1:4
configure vlan sales ipaddress 10.1.2.3/24
enable ipforwarding
create esrp esrp1
configure esrp esrp1 domain-id 4096
configure esrp esrp1 add master sales
enable esrp esrp1
configure ospf add vlan sales area 0.0.0.0 passive
configure ospf routerid 5.5.5.5
enable ospf
```

## Multiple Domains Using Layer 2 and Layer 3 Redundancy

The example shown in the following figure illustrates an *ESRP* configuration that has multiple domains using Layer 2 and Layer 3 redundancy.

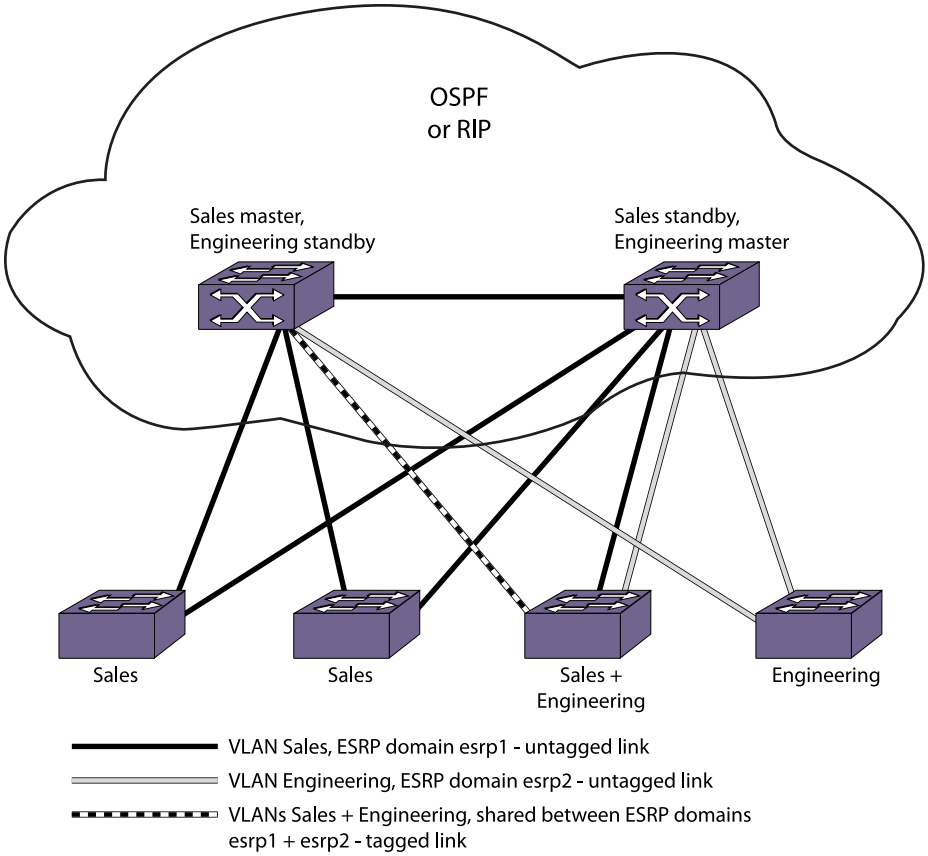



**Figure 173: Multiple ESRP Domains Using Layer 2 and Layer 3 Redundancy**

This example builds on the previous example. It has the following features:

- An additional *VLAN*, Engineering, is added that uses Layer 2 redundancy.
- VLAN Sales uses three active links to each ESRP switch.
- VLAN Engineering has two active links to each ESRP switch.
- One of the edge devices carries traffic for both VLANs.
- The link between the third edge device and the first ESRP switch uses 802.1Q tagging to carry traffic from both VLANs on one link. The ESRP switch counts the link active for each VLAN.
- The second ESRP switch has a separate physical port for each VLAN connected to the third edge switch.

In this example, the ESRP switches are configured such that VLAN Sales normally uses the first ESRP switch and VLAN Engineering normally uses the second ESRP switch. This is accomplished by manipulating the ESRP priority setting for each VLAN for the particular ESRP switch.

Configuration commands for the first ESRP switch are as follows:

```
create vlan sales
configure vlan sales tag 10
configure vlan sales add ports 1:1-1:2
configure vlan sales add ports 1:3, 1:5 tagged
configure vlan sales ipaddress 10.1.2.3/24
create vlan engineering
```

```
configure vlan engineering tag 20
configure vlan engineering add ports 1:4
configure vlan engineering add ports 1:3, 1:5 tagged
configure vlan engineering ipaddress 10.4.5.6/24
create esrp esrp1
configure esrp esrp1 domain-id 4096
configure esrp esrp1 add master sales
configure esrp esrp1 priority 5
enable esrp esrp1
create esrp esrp2
configure esrp esrp2 domain-id 4097
configure esrp esrp2 add master engineering
enable esrp esrp2
```

Configuration commands for the second ESRP switch are as follows:

```
create vlan sales
configure vlan sales tag 10
configure vlan sales add ports 1:1-1:3
configure vlan sales ipaddress 10.1.2.3/24
configure vlan sales add ports 1:5 tagged
create vlan engineering
configure vlan engineering tag 20
configure vlan engineering add ports 1:4, 2:1
configure vlan engineering ipaddress 10.4.5.6/24
configure vlan engineering add ports 1:5 tagged
create esrp esrp1
configure esrp esrp1 domain-id 4096
configure esrp 1 add master sales
enable esrp esrp1
create esrp esrp2
configure esrp esrp2 domain-id 4097
configure esrp esrp2 add master engineering
configure esrp esrp2 priority 5
enable esrp esrp2
```

## ESRP Over IPv6 Configuration Example

The example shown in the following figure illustrates an *ESRP* configuration that can operate over IPv6.

**Note**

To support operation over IPv6, the master VLANs configured in this example require both an IPv4 and an IPv6 address. ESRP route tracking and ESRP ping tracking are not supported for IPv6 addresses.

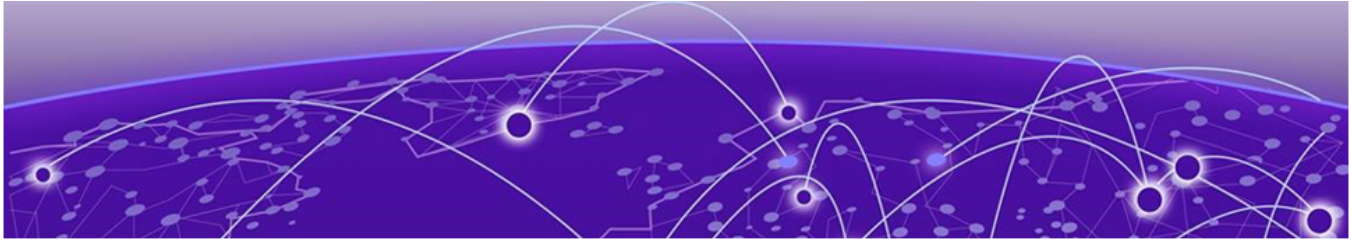
Configuration commands for the first ESRP switch are as follows:

```
create vlan sales
configure vlan sales tag 10
configure vlan sales add ports 1:1-1:2
configure vlan sales add ports 1:3, 1:5 tagged
configure vlan sales ipaddress 10.1.2.3/24
configure vlan sales ipaddress 2001:db8:36::1/48
create vlan engineering
configure vlan engineering tag 20
configure vlan engineering add ports 1:4
configure vlan engineering add ports 1:3, 1:5 tagged
configure vlan engineering ipaddress 10.4.5.6/24
```

```
configure vlan engineering ipaddress 2001:db8:36::2/48
create esrp esrp1
configure esrp esrp1 domain-id 4096
configure esrp esrp1 add master sales
configure esrp esrp1 priority 5
enable esrp esrp1
create esrp esrp2
configure esrp esrp2 domain-id 4097
configure esrp esrp2 add master engineering
enable esrp esrp2
```

Configuration commands for the second ESRP switch are as follows:

```
create vlan sales
configure vlan sales tag 10
configure vlan sales add ports 1:1-1:3
configure vlan sales ipaddress 10.1.2.3/24
configure vlan sales ipaddress 2001:db8:36::1/48
configure vlan sales add ports 1:5 tagged
create vlan engineering
configure vlan engineering tag 20
configure vlan engineering add ports 1:4, 2:1
configure vlan engineering ipaddress 10.4.5.6/24
configure vlan engineering ipaddress 2001:db8:36::2/48
configure vlan engineering add ports 1:5 tagged
create esrp esrp1
configure esrp esrp1 domain-id 4096
configure esrp 1 add master sales
enable esrp esrp1
create esrp esrp2
configure esrp esrp2 domain-id 4097
configure esrp esrp2 add master engineering
configure esrp esrp2 priority 5
enable esrp esrp2
```



# VRRP



This chapter assumes that you are already familiar with the Virtual Router Redundancy Protocol (VRRP). If not, refer to the following publications for additional information:

- RFC 2338—Virtual Router Redundancy Protocol (VRRP)
- RFC 2787—Definitions of Managed Objects for the Virtual Router Redundancy Protocol
- RFC 3768—Virtual Router Redundancy Protocol (VRRP) Version 2 for IPv4 www.ietf.org/rfc/rfc3768.txt
- RFC 5798—Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6 http://tools.ietf.org/html/rfc5798
- Draft IETF VRRP Specification v2.06 http://tools.ietf.org/html/draft-ietf-vrrp-spec-v2-06

## VRRP Overview

*VRRP (Virtual Router Redundancy Protocol)*, like the *ESRP (Extreme Standby Router Protocol)*, allows multiple switches to provide redundant routing services to users.

VRRP is used to eliminate the single point of failure associated with manually configuring a default gateway address on each host in a network. Without using VRRP, if the configured default gateway fails, you must reconfigure each host on the network to use a different router as the default gateway. VRRP provides a redundant path for the hosts. Using VRRP, if the default gateway fails, the backup router assumes forwarding responsibilities. An example VRRP topology is shown in Figure 174.

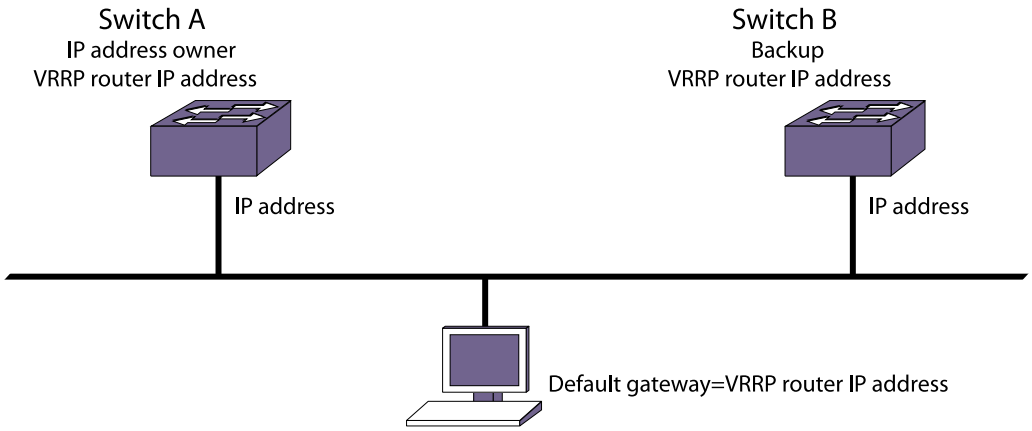



**Figure 174: Simple VRRP Network**

Switches A and B in the figure above are both configured with the same VRRP router ID on the same *VLAN (Virtual LAN)*, which establishes a VRRP relationship between the two routers. Because a single switch can support multiple VRRP relationships, each relationship is referred to as a *VRRP router instance*. Within a VRRP router instance, any VRRP router can become the master, but only one VRRP router can be master at a time. The master processes all client communications, and the other VRRP routers in the VRRP routing instance stand by, ready to take over if the master is no longer available.

Each switch in a VRRP topology has its own unique IP and MAC addresses, which are required for basic IP connectivity. For each VRRP router instance, there are shared VRRP IP and MAC addresses, which are used for network client communications. The VRRP router IP address is configured on all VRRP routers in a VRRP routing instance, and it is configured as the default gateway address on network clients. If the master VRRP router becomes unavailable, the backup VRRP router takes over using the same VRRP router IP address.

If the VRRP router IP address matches the actual VLAN IP address of the IP address owner has the highest priority value (255) and will always become the master when VRRP is enabled and operating correctly. If the switch or the VRRP process on the switch stops responding, a backup switch (Switch B in the following figure) takes over the master role and serves as the default gateway for network clients.

VRRP supports multiple backup routers. If the master VRRP router stops working, one of the backup routers takes over as described in ESRP Master Election on page 1094.

VRRP also supports multiple VRRP router instances, which can be used to enable load sharing. The following figure shows a VRRP load-sharing configuration.

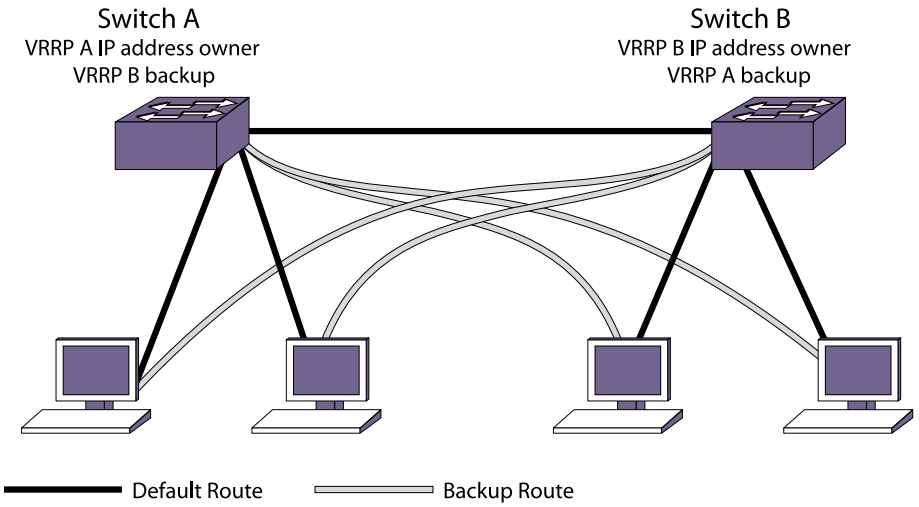



**Figure 175: VRRP Load-sharing Configuration**

Switches A and B in the above figure are each configured with two VRRP router instances. Switch A is the IP address owner and default master for VRRP instance 1, and Switch B is the IP address owner and default master for VRRP instance 2. Half the network clients are configured to use VRRP instance 1 as the primary gateway and VRRP instance 2 as the backup gateway. The other half of the network clients are configured to use VRRP instance 2 as the primary gateway and VRRP instance 1 as the backup gateway. When both switches are operating with VRRP, each switch supports half the clients in a load-sharing topology. If either switch fails, or if VRRP is disabled on a switch, the remaining switch supports all network clients.

Note
We recommend that you do not enable VRRP on aggregated *VLAN*s, which are also known as super VLANs.

## VRRP Master Election

When a *VRRP* configured network starts, VRRP uses an election algorithm to dynamically assign master responsibility to one of the VRRP routers on the network.

Note
On BlackDiamond 8800 series switches, when a port belongs to two different VRRP instances with the same VRID, and one of the instances is a master VRID and the other a standby VRID, broadcast packets belonging to the standby VRRP *VLAN* generated by the master VRRP in that VLAN are not forwarded.

The VRRP master is determined by the following factors:

- VRRP priority—Possible values are 0 through 255. Value 255 is reserved for the VRRP router IP address owner, and value 0 is reserved for the master router, to indicate it is releasing master responsibility. Values 1–254 can be configured on backup routers to influence which backup router becomes the master when the master VRRP router is no longer available. The higher number has higher priority. The default value for backup routers is 100.
- Higher IP address—If multiple backup routers have the same configured priority, the router with the highest IP address becomes the master.

If the master router becomes unavailable, the election process begins and the backup router that wins the election assumes the role of master.

Note
In VRRP IPv6, master election happens based on the interface link local address when the priorities are the same. The highest link local address switch is selected as the VRRP master.

A new master is elected when one of the following things happen:

- VRRP is disabled on the master router.
- Loss of communication occurs between master and backup router(s).

If VRRP is disabled on the master interface, the master router sends an advertisement with the priority set to 0 to all backup routers. This signals the backup routers that they do not need to wait for the master down interval to expire, and the master election process can begin immediately.

The master down interval is set using the following formula: 3 x advertisement interval + skew time.

The advertisement interval is a user-configurable option, and the skew time is (256-priority/256).

Note
The formula for VRRPv2; ((256 - priority) / 256). The VRRPv3 Skew time calculation will be different. Please refer to RFC 5798 (http://tools.ietf.org/html/rfc5798), Skew time. (((256 - priority) x Master_Adver_Interval) / 256)

Note
An extremely busy CPU can create a short dual master situation. To avoid this, increase the advertisement interval.

## VRRP Master Preemption

*VRRP* master preemption is a feature that allows a VRRP backup router with a higher VRRP priority to take control from a lower priority backup that is acting as the master.

VRRP election occurs as described in ESRP Master Election on page 1094. VRRP preemption occurs when a VRRP backup router is added to the network or recovers, and that backup router has a higher priority than the current backup VRRP that is operating as the master.

Note
The VRRP router IP address owner always preempts, independent of the VRRP preemption setting.

When a VRRP backup router preempts the master, it does so in one of the following ways:

- If the preempt delay timer is configured for between 1 and 3600 seconds and the lower-priority master is still operating, the router preempts the master when the timer expires.
- If the preempt delay timer is configured for 0, the router preempts the master after three times length of the hello interval.
- If the higher priority router stops receiving advertisements from the current master for three times the length of the hello interval, it takes over mastership immediately.

The preempt delay timer provides time for a recovering router to complete start up before preempting a lower-priority router. If the preempt delay timer is configured too low, traffic is lost between the time the preempting router takes control and the time when it has completed startup.

## VRRP Tracking

Tracking information is used to track various forms of connectivity from the *VRRP* router to the outside world.

### *VRRP Tracking Mode*

When a *VRRP* tracked entity fails, the VRRP router behavior is controlled by the tracking mode. The mode can be all, or any. The default mode is all.

When the mode is all, the master role is relinquished when one of the following events occur:

- All of the tracked VLANs fail.
- All of the tracked routes fail.
- All of the tracked pings fail.

**Note**
Mastership is relinquished and the switch goes to INIT state.

When the mode is any, the master role is relinquished when any of the tracked VLANs, routes, or pings fail.

### *VRRP VLAN Tracking*

You can configure *VRRP* to track active *VLAN*s (active ports in a VLAN or Loopback) of up to eight specified VLANs as criteria for failover.

If no active ports remain on the specified VLANs, the router automatically relinquishes master status based on the tracking mode.

When a tracking condition is in a failed state, VRRP behaves as though it is locally disabled; so it is neither master nor backup (which are both active states).

### *VRRP Route Table Tracking*

You can configure *VRRP* to track specified routes in the route table as criteria for VRRP failover.

If any of the configured routes are not available within the route table, the router automatically relinquishes master status based on the tracking mode.

**Note**
*MPLS (Multiprotocol Label Switching)* LSPs are not considered to be part of VRRP route tracking.

### *VRRP Ping Tracking*

You can configure *VRRP* to track connectivity using a simple ping to any outside responder.

The responder may represent the default route of the router, or any device meaningful to network connectivity of the master VRRP router. If pinging the responder consecutively fails the specified number of times, the router automatically relinquishes master status based on the tracking mode.

## VRRP Address Support for IPv4

For IPv4 traffic, a primary IPv4 address is selected from the set of real interface addresses. *VRRP* advertisements are always sent using the primary IPv4 address as the source of the IPv4 packet.

The VRRP MAC address for an IPv4 VRRP router instance is an IEEE 802 MAC address in the following hexadecimal format (in Internet-standard bit-order):

```
00-00-5E-00-01-<vrid>
```

The first three octets are derived from the IANA Organizational Unique Identifier (OUI). The next two octets (00-01) indicate the address block assigned to the VRRP router for the IPv4 protocol, and VRID is the VRRP instance identifier. This mapping provides for up to 255 IPv4 VRRP routers on a network.

When a VRRP router instance becomes active, the master router issues a gratuitous ARP response that contains the VRRP router MAC address for each VRRP router IP address.

The master also always responds to ARP requests for VRRP router IP addresses with an ARP response containing the VRRP MAC address. Hosts on the network use the VRRP router MAC address when they send traffic to the default gateway.

## VRRP Address Support for IPv6

IPv6 *VRRP* router advertisements are always sent using the VRRP virtual link-local address as the source address.

Hosts on the LAN can use this link-local address as their default route gateway. If no VRRP link-local address is configured, a default value is derived as follows:

```
FE80::5E00:02{VRID}
```

VRID is the VRRP instance identifier.

Note

The host portion of this address corresponds to the virtual MAC address associated with the VRRP router.

When a backup VRRP router assumes the role of master, it must use the same link-local address in router advertisements as the previous master. Therefore, when configuring a backup VRRP router, you must either configure a virtual link local address that matches the link local address on the IP address owner or allow the virtual link local address to default to the derived value in the same manner as on the master.

Router advertisement prefixes are configured based on the VRRP IP addresses. The mask used for the prefix will be the smallest mask used by all VRRP IP addresses on a *VLAN* interface. The ExtremeXOS software supports multiple IPv6 addresses on an interface that can overlap. For example, you can add

both 200d::1/48 and 200d::2/96 to a VLAN. IPv6 VRRP routers advertise the smallest mask that applies to all VRRP IP addresses. In this example, if VRRP IP address 200d::100 is added, the mask is 48.

The VRRP MAC address for an IPv6 VRRP router instance is an IEEE 802 MAC address in the following hexadecimal format (in Internet-standard bit-order):

```
00-00-5E-00-02-<vrid>
```

The first three octets are derived from the IANA OUI. The next two octets (00-02) indicate the address block assigned to the VRRP router for IPv6 protocol, and VRID is the VRRP instance identifier. This mapping provides for up to 255 IPv6 VRRP routers on a network.

When a VRRP router assumes the role of master, it issues an unsolicited neighbor discovery (ND) neighbor advertisement message for each of the VRRP router IP addresses.

The master also always responds to ND neighbor solicitations with ND neighbor advertisements using the VRRP MAC address.

## NTP VRRP Virtual IP support

This feature allows switches to configure the *VRRP* virtual IP as an NTP server address. The NTP server, when configured on the VRRP master, listens on the actual IP and virtual IP address for NTP clients.

On the VRRP backup node, only one socket is opened and bound to the physical IP address alone. Once a node transitions to the VRRP master, the ExtremeXOS software re-triggers a listen on the interface to ntpd for it to open a socket and bind to the VRRP VIP.

A flag configuration is added for the IPv4 cases, and these are propagated to *VLAN* Manager clients. NTP uses this to trigger a listen on the interface. For the master node to process non-ping packets destined to the VIP, the software already has a configuration command in VRRP (accept-mode on/off).

Note
If you want to configure VRRP VIP as the server address on NTP clients, enable accept mode.

### *Limitations*

The following limitations exist for NTP VRRP Virtual IP support:

- Summit switches configured as NTP clients need to have the following bootrom version:
  - X480, X460, X440, X670, and X770 – 2.0.1.7
  - NWI – 1.0.5.7
- We do not recommend FHRP Virtual IPs for NTP configuration because they can cause undesirable behavior when the NTP servers are not in sync, or if the delay is asymmetric. Ensure that both servers derive their clock information from the same source.

  This problem can be more acute if a node connected to VRRP peers using *MLAG (Multi-switch Link Aggregation Group)* and VRRP is in active-active mode. In this case, it is possible that every other packet could be sent to a different switch due to *LAG (Link Aggregation Group)* hashing at the remote node.

## VRRPv3 Interoperation with VRRPv2

RFC 5798 states that VRRPv2 and VRRPv3 interoperation is optional, and that mixing these two versions should only be done when transitioning from VRRPv2 to VRRPv3.

VRRPv2 and VRRPv3 interoperation should not be implemented as a permanent solution.

This release supports configuration for the following modes of operation: VRRPv2, VRRPv3, and VRRPv2 and VRRPv3 interoperation.

## VRRP Active-Active

*VRRP* Active-Active mode allows you to have two active VRRP masters in conjunction with *MLAG* by applying an *ACL (Access Control List)* on the IST links in order to block VRRP updates.

When you configure VRRP with MLAG, you have the option to make VRRP operate in active-active mode. For MLAG peers to operate in VRRP active-active mode, configure the following ACL on both ends of the ISC port.

```
entry vrrp-act {
if match all {
    destination-address 224.0.0.18/32 ;
} then {
    deny ;
       }
}
```

There are two caveats that you need to be aware of that are illustrated in the following figure:

- An ARP request from 10.0.0.4 results in duplicate ARP replies (one from each MLAG switch).
- For this to work correctly, you have to configure the virtual IP address to be a different address from either of the MLAG peer interface addresses. When an MLAG switch generates an ARP request it uses the vMAC instead of its own switch MAC, and the response (if the reverse path hashing chooses the other MLAG switch) is consumed by the peer MLAG switch.

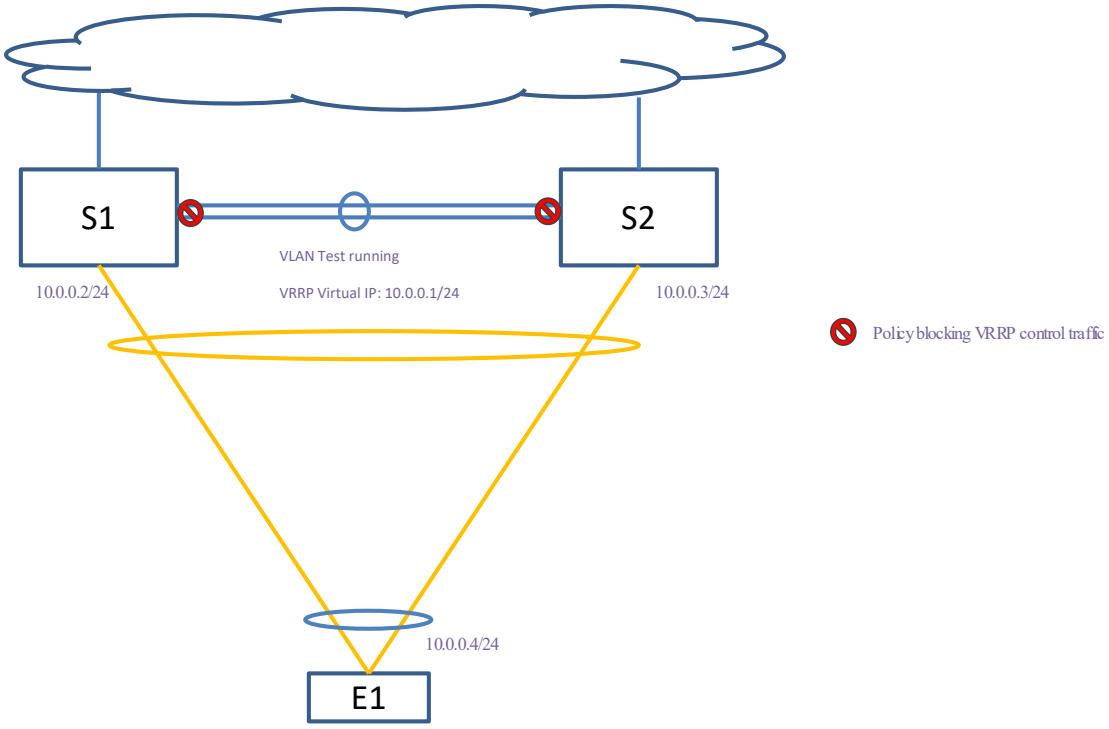



**Figure 176: VRRP Active-Active**

## VRRP Fabric Routing

Prior to ExtremeXOS 16.2, *VRRP* had one master router which would do L3 routing and one or more backup routers which would do L2 forwarding of packets to the master router. This caused loss of bandwidth in the links that connect the master and backup routers. This issue is present in any topology wherein host traffic is flowing via backup router. In case of multiple backup routers, traffic from hosts attached to some backup routers would have to traverse multiple links to reach the master router. This caused loss of bandwidth in multiple links towards the master.

The VRRP fabric routing feature introduced in ExtremeXOS 16.2 allows the backup router to take part in L3 routing for the packets it receives with DA=VMAC. A backup router enabled with this feature is called a Fabric Routing Enabled Backup (FREB) router. This feature allows load sharing of traffic between VRRP routers and saves the bandwidth of links connecting the master and backup. This bandwidth saved can be used to provide better connectivity between the segments of the network wherein each segment is connected to only one VRRP router directly. This solution is applicable for all topologies such as *MLAG*, EAPS, or *STP (Spanning Tree Protocol)*.

The VRRP FREB router receives and processes VRRP advertisements from the master, but does not generate advertisements. Broadcast ARP requests and multicast neighbor solicitations to resolve the virtual IP are responded to by the master router. The Fabric Routing Enabled Backup router behaves as mentioned in RFC 5798 with the following exceptions:

- The backup router performs forwarding functionality, similar to the master.
- The backup router responds to unicast ARP/Neighbor Solicitation requests it receives.

### *Configuring VRRP Fabric Routing*

To configure the fabric routing feature, enter the following command:

```
configure vrrp vlan vlan_name vrid vridval fabric-routing [on | off]
```

This configuration can be present on all VRRP routers, regardless of VRRP state of the router. Fabric routing will be enabled only when the VRRP router is in backup state.

## *Fabric Routing Functionality*

The *VRRP* backup router enabled with fabric routing mode is called a VRRP FREB router (Fabric Routing Enabled Backup router). The VRRP FREB router is responsible for the following:

- Performing routing for the packets destined for subnets other than the received interface's subnet. These packets should also have a destination MAC matching Virtual MAC. These packets will be routed by hardware.
- Forwarding the packets having a Destination IP matching Virtual IP, towards VRRP Master. These packets will have Destination MAC matching Virtual MAC. These packets will be forwarded by hardware.
- Responding for unicast ARP requests with target IP matching Virtual IP.
- Responding for unicast Neighbor Solicitation requests with target IP matching Virtual IP.
- Does not respond to broadcast ARP / multicast Neighbor solicitations targeted for Virtual IP.
- Does not respond for *ICMP (Internet Control Message Protocol)* requests destined for Virtual IP.
- Does not generate gratuitous ARP/ Neighbor Advertisements.
- Does not advertise Router Advertisement Prefixes in this state.

#### VRRP Router Accepts Packets Destined to Virtual IP

Only the master serves as protocol servers like NTP, telnet, SSH, etc and accepts connections destined to the virtual IP. By doing so, hosts always connect to the same *virtual router (VR)* by using the virtual IP, at any given point of time. This ensures that the host is getting a consistent response from the protocol server. This arrangement allows the host to reach the NTP server using the same IP i.e. virtual IP, even if VRRP mastership moves to a different router. A network monitoring tool is another example, which can use virtual IP to collect data about VRRP domain, by connecting to current VRRP Master. It is not recommended to change the configurations of the switch, when a management session is connected using virtual IP. When VRRP FREB router sits in between the host and VRRP Master router, FREB router does hardware forwarding of these packets from host towards VRRP Master, at Layer 3.

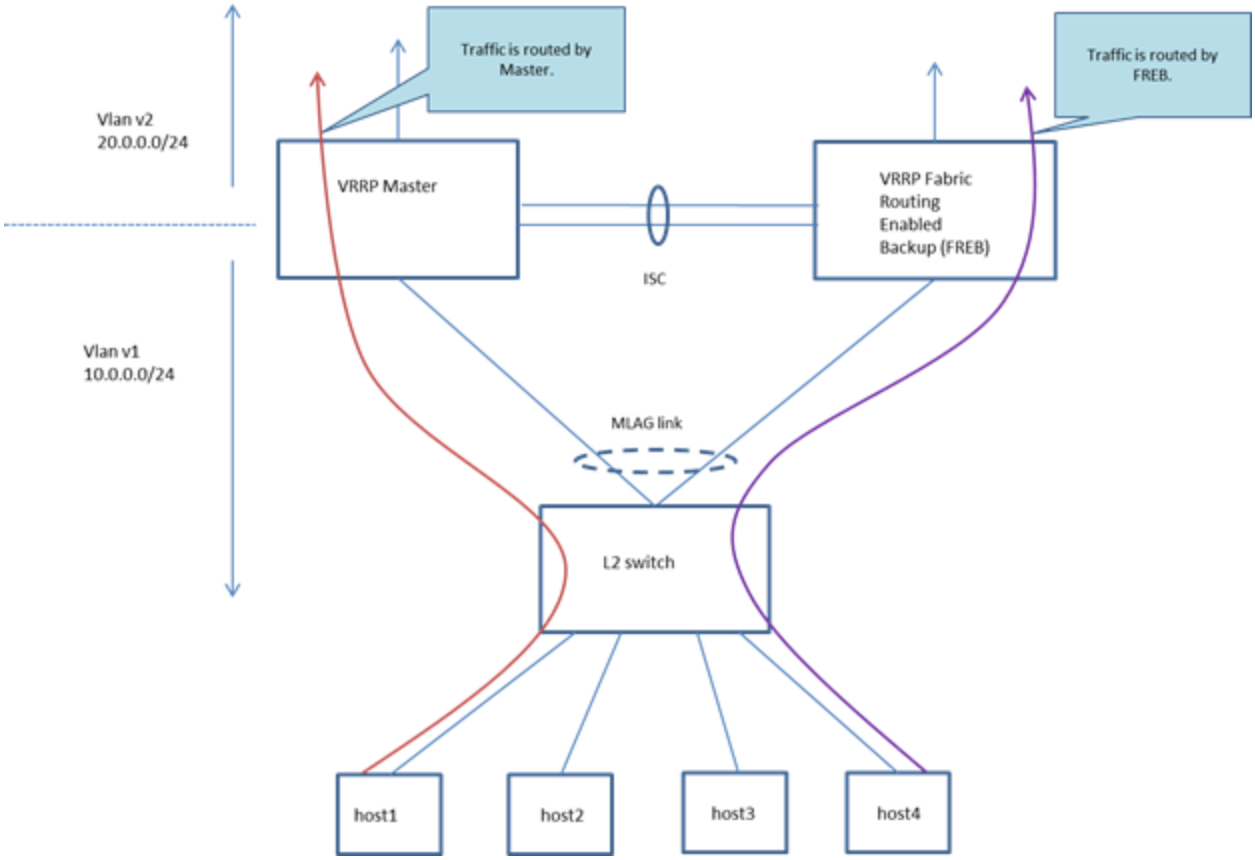



**Figure 177: VRRP Fabric Routing**

Note

A caveat is that TTL/hop count is decremented for the packets destined for virtual IP, when forwarded by FREB. This may be a problem to run any protocol that expects TTL not to be decremented, between host and Master.

Hosts can generate unicast ARP to validate a ARP cache entry. Similarly, unicast Neighbor Solicitation is generated to perform Neighbor Unreachability Detection for a neighbor. These requests are periodic. The unicast ARP/NS requests will be responded by FREB, if it receives the request. A downside of allowing VRRP Master to respond these requests is that it may take considerable CPU cycles when large numbers of hosts are present in VRRP domain.

### *Enabling Fabric Routing Mode*

A *VRRP* router configured with fabric routing mode will be in backup state. There is some difference in behavior based on enabling sequence which is described below.

When enabling VRRP VR after configuring fabric routing mode, assume that fabric routing mode has already been configured for VRRP VR and VR is administratively enabled now. Initially the VRRP state machine will start from INIT state. It will move to backup state and wait for Master_Down_Interval duration to receive an advertisement from current master. Based on VRRP Master election (as described earlier in the sections), VRRP router will either become master or will assume forwarding responsibility

while in backup state (i.e. FREB). Waiting in backup state avoids a burst of messages exchanged in software during the transition. This is useful in scaled setups.

Note

For the duration it waits in backup state, traffic will be forwarded to the VRRP master where it gets routed. This behavior is same as existing behavior without fabric routing.

When configuring fabric route mode when the VR is already in backup state, if the VR has received advertisements within the last advertisement interval, the router will take forwarding responsibility immediately.

The hold off timer will be applied before FREB starts IP forwarding. In case of reboot or process restart, the VR will wait for 60 seconds before taking forwarding responsibility. This will allow the routing protocols in the restarting system to converge.

To enable fabric routing, use the following command:

- `configure vrrp {vlan vlan_name vr vr_id | all} fabric-routing [on | off]`

To configure a specific VRRP instance as FREB:

```
configure vrrp vlan v1 vrid 1 fabric-routing on
```

To configure all the VRRP instances as FREB:

```
configure vrrp all fabric-routing on
```

Configured fabric routing can be seen in the show command output of the `show vrrp detail` or `show vrrp vlan <vlan>` CLI command:

```
# sh vrrp detail
VLAN:  v1       VRID:  1        VRRP:  Disabled State:  Backup
Virtual Router:  VR-Default
Priority:  100(backup)  Advertisement Interval:  1 sec
Version: v2     Preempt:  Yes   Preempt Delay: 0 sec
 Virtual IP Addresses:
 Accept mode:  Off
 Host-Mobility:  Off
 Host-Mobility Exclude-Ports:
 Tracking mode:  ALL
 Tracked Pings:  -
 Tracked IP Routes:  -
 Tracked VLANs:  -
 Fabric Routing: On

  * indicates a tracking condition has failed
```

### *Fabric Routing Limitations*

The following are limitation of *VRRP* fabric routing:

- VRRP Fabric routing is not supported in following modules: G48Te2, G24Xc, G48Xc, G48Tc, 10G4Xc, 10G8Xc, MSM-48c, S-G8Xc, S-10G1Xc and S-10G2Xc modules. Fabric routing feature does not function correctly when VRRP instances are configured on these modules (when VRRP *VLAN* has ports in these modules).

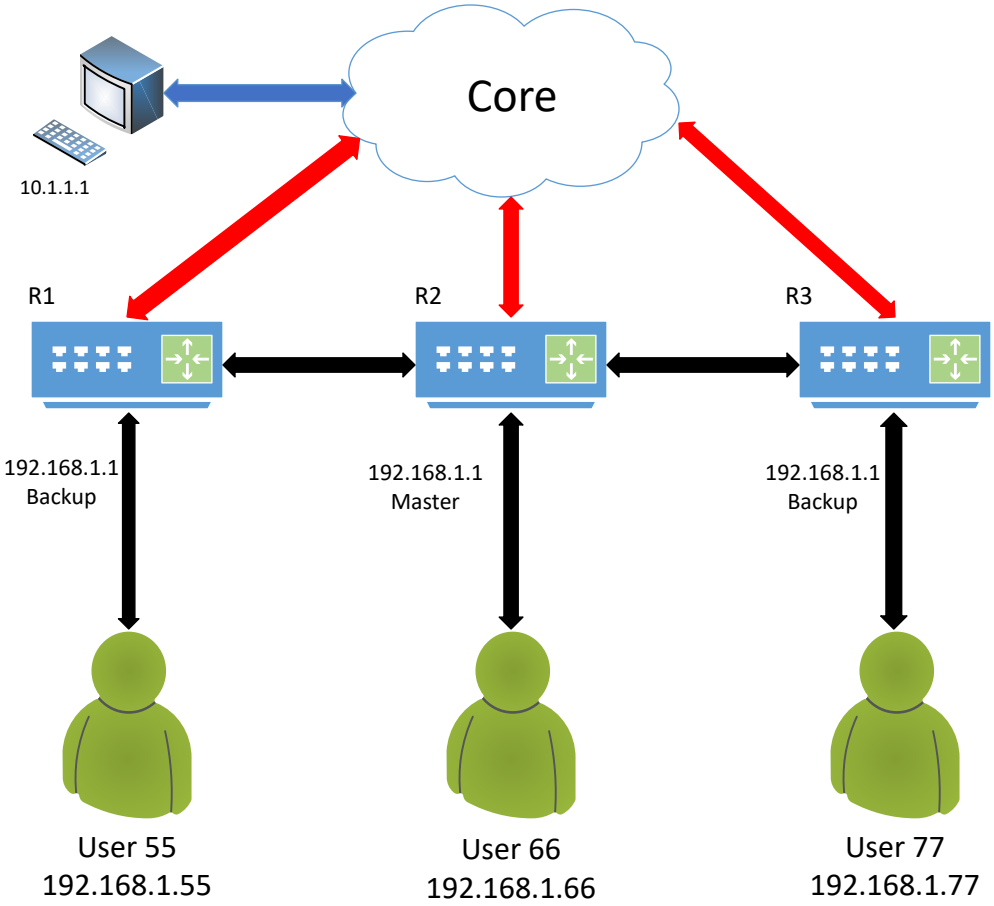
- VRRP fabric routing is not supported for the configuration of virtual IP address being same as interface IP address (i.e. IP owner case).
- VRRP fabric routing is not supported for PVLAN and VLAN aggregation topologies.

## VRRP Host-Mobility

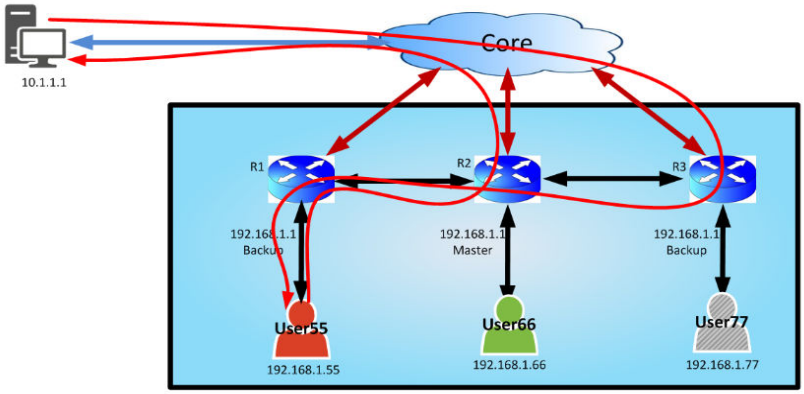
*VRRP* host-mobility solves the asymmetric routing problem associated with VRRP where the path to return to an end host may be different and longer than necessary. This feature uses host routes to indicate where in the network an end host resides. Using other routing protocols such as *OSPF (Open Shortest Path First)*, other routers will then pick the shortest path back to the end host when multiple paths are available via *ECMP (Equal Cost Multi Paths)* route entries.

### *Asymmetric Routing and VRRP*

The optimum flow path for frames sourced by clients is that they are forwarded by the first encountered router in the path. The optimum flow path for return frames for these flows is through the same router that forwarded the initial frames of the flow. This is referred to as a "symmetric routing path." Unfortunately, standard VRRP elects one of a group of routers to forward frames from a client, but the return path may be through any of the routers in the VRRP group. This condition is known as "asymmetric routing". In the following figure there are three devices that are routers in a VRRP group servicing the 192.168.1.0/24 network with the VRRP virtual gateway 192.168.1.1. Router R2 has been elected VRRP master and this is the only router that will route frames from the clients to the core, even though the clients are not locally connected to the master. The other devices bridge frames from the clients to the master; therefore, the frames from User 55 and User 77 must traverse two devices when routing the frame to the core.




R1, R2, and R3 all advertise 192.168.1.0/24 to the core. This means that return traffic to the users will be sent from the core to any of these routers. A non-optimal path for the flows in either direction may result. Flows from User 55 to Core would traverse R1, R2. The return path could traverse R3, R2, and R1.



Fabric Routing solves part of this path problem by allowing the backup routers to forward frames for the clients without having to send the frame to the master router for the group. Now frames from User 55 will be routed by the first router that receives the frame (R1). Unfortunately, the return path to User 55 from the core is still via Equal Cost Multipath (ECMP) to any of the routers. Again, the frame may have to still traverse R3, R2, and finally R1 to reach User 55 from the core.

### *Host-Mobility Solving Asymmetric Routing*

The host-mobility feature uses ARP/ND to learn each user connected to the IP Network on the router that is closest to the end host. It then uses routing protocols such as OSPF to advertise the hosts IP/IPv6 address. These “Host Routes” will be advertised to other routers, making them the preferred route to the end host’s IP address. As the host route propagates through the network, the most efficient path to the end host will be revealed as every other path will have an additional cost. With host-mobility, User 55’s complete route will be advertised to R2, R3, and the core. The core will then receive a route to User 55’s address from all three routers. The most direct route will have fewer hops and a lower cost than the routes reported by R2 and R3. When traffic destined to User 55 enters the core the most direct path (R1) is selected.

### *VRRP Host-Mobility Feature Detail*

The ExtremeXOS Host-Mobility feature is implemented using the user space application VRRP, which listens for ARP/ND updates from the kernel using the *FDB (forwarding database)*. When ARP/ND entries are inserted into the kernel, the FDB notifies the host-mobility application. If the new ARP/ND entry exists on an interface that has VRRP, host-mobility is configured on the VRRP & VRID pair, and the port is not being excluded, the host-mobility application will create a host route in the route table.

The host routes in the route table will be propagated via existing routing protocols such as OSPF.

As ARP/ND entries are removed from the kernel, the host-mobility portion of the VRRP application is notified. Normally if the interface is still up when entries are removed, the host-mobility application sends an ARP/ND packet to the IP address of the host associated with the ARP/ND entry to resolve the address again. The FDB attempts to resolve ARP addresses when near the age period. Due to the FDB keeping ARPs up to date, host-mobility does not need to try to resolve aging entries when aging. When the entry is removed by a clear CLI command, FDB does not send an ARP. When the entry is removed, Host-Mobility requests an ARP. After five seconds, if no route has been added, the host routes are removed from the route table.

Ports are designated as host or router ports; by default the ports are host ports. ARP/ND entries learned on router ports do not generate host route entries. Only ARP/ND entries learned on host ports generate route entries. Host route entries are removed if the ARP/ND entry is learned or changes to a router port.

When a host moves from one switch to another on the same Layer 2 network, a new host route is created from the new location. In this case there may be two host routes throughout the network advertising for the same host. While not optimal, this situation does not cause any network delivery problems. Eventually the original host route is removed when the ARP/ND entry is removed from the switch or moves from a host port to a router port.

A redistribution command is required for OSPF or other routing protocols to distribute the new host routes. The command to redistribute is `enable ospf export host-mobility cost 0 type ase-type-2`. A new routing “Origin” type is added to route table for host-mobility routes.

Since OSPF will redistribute host-mobility routes, it is possible that a route could be learned for a host that the local device can reach at a lesser cost. Host-mobility monitors routes added by OSPF. If OSPF adds a route that is part of a subnet that the device is a member of, then host-mobility triggers ARP/ND to be performed. By performing ARP/ND, the lower cost route to host is added as an ARP/ND entry in the proper table and is used while routing traffic. If no response is received, then the OSPF route will continue to be used instead.

## VRRP Guidelines

The following guidelines apply to using *VRRP*:

- VRRP packets are encapsulated IP packets.
- The VRRP IPv4 multicast address is 224.0.0.18.
- The VRRP IPv6 multicast address is ff02::12.
- The maximum number of supported VRIDs per interface is 31.
- An interconnect link between VRRP routers should not be used, except when VRRP routers have hosts directly attached.
- VRRP instance scale is dependent on platform and license. Please refer to the ExtremeXOS release notes for details.
- VRRP and other L2 redundancy protocols can be simultaneously enabled on the same switch.
- When VRRP and BOOTP/*DHCP (Dynamic Host Configuration Protocol)* relay are both enabled on the switch, the relayed BOOTP agent IP address is the actual switch IP address, not the virtual IP address.
- We do not recommend simultaneously enabling VRRP and *ESRP* on the same switch.
- VRRP and ESRP cannot be configured on the same *VLAN* or port. This configuration is not allowed.
- RFC 5798 describes a situation where a master VRRP router takes on a duplicate IP address due to interaction with the duplicate address detection (DAD) feature. To prevent such duplicate addresses, the DAD feature is disabled whenever a VRRP router is configured for IPv6 or IPv4.
- A VRRP router instance can be configured with multiple IP addresses on the same subnet or on different subnets, provided that all virtual IP addresses match the subnet address of a VLAN on the switch. For example, if a host switch has VLAN IP addresses in the 1.1.1.x and 2.2.2.x subnets, then that VRRP router instance can contain virtual IP addresses in both those subnets as well.
- If a VRRP router instance is assigned priority 255, then the host router must own all the IP addresses assigned to the VRRP router instance. That is, each virtual IP address must match an IP address configured for a VLAN on the router.
- When a VRRPv2 instance spans routers using ExtremeXOS version 12.6 and earlier and routers using ExtremeXOS version 12.7 and later, routers using ExtremeXOS version 12.6 and earlier log packet-size warning messages.
- VRRP scaling numbers differs based on the license and hardware used; please refer the release notes for individual scaling limits.
- The maximum number of VIPs supported for a single VRRP instance is 255.

Note

A maximum of 511 VRRP instances are allowed without One-To-Many Mirroring. If using OTM Mirroring, 507 VRRP instances are allowed.

## VRRP and Hitless Failover

When you install two Management Switch Fabric Module (MSM) or Management Modules (MMs) in a BlackDiamond chassis, one MSM/MM (node) assumes the role of primary and the other node assumes the role of backup.

Note
This section applies to Modular Switches and SummitStack only.

The primary node executes the switch's management functions, and the backup acts in a standby role. Hitless failover transfers switch management control from the primary to the backup and maintains the state of *VRRP*. While VRRP supports hitless failover, you do not explicitly configure hitless failover support; rather, if you have two nodes installed, hitless failover is available.

To support hitless failover, the primary node replicates VRRP protocol data units (PDUs) to the backup, which allows the nodes to run VRRP in parallel. Although both nodes receive VRRP PDUs, only the primary transmits VRRP PDUs to neighboring switches and participates in VRRP.

To initiate hitless failover on a network that uses VRRP:

1. Confirm that the primary and backup nodes are synchronized and have identical software and switch configurations using the `show switch {detail}` command.

   The output displays the status of the nodes, with the primary node showing MASTER and the backup node showing BACKUP (InSync).

   a. If the primary and backup nodes are not synchronized and both nodes are running a version of ExtremeXOS that supports synchronization, proceed to step 2 in STP and Hitless Failover--Modular Switches Only on page 1051.
   b. If the primary and backup nodes are synchronized, proceed to step 3 in STP and Hitless Failover--Modular Switches Only on page 1051.
2. If the primary and backup nodes are not synchronized, use the `synchronize` command to replicate all saved images and configurations from the primary to the backup.

   After you confirm the primary and backup nodes are synchronized, proceed to step 3 in STP and Hitless Failover--Modular Switches Only on page 1051.
3. If the primary and backup nodes are synchronized, use the `run failover` (formerly `run msm-failover`) command to initiate failover.

For more detailed information about verifying the status of the nodes and system redundancy, see Understanding System Redundancy on page 54. For more information about hitless failover, see Understanding Hitless Failover Support on page 59.

Note
For complete information about software licensing, including how to obtain and upgrade your license and what licenses are appropriate for these features, see the *Feature License Requirements* document.

# Configuring VRRP

The following procedure can be used to configure a simple *VRRP* topology:

1. Configure the VRRP IP address owner router as follows:
   a. Create a *VLAN* to serve as the VRRP router VLAN.
   b. Add an IP address to the VRRP VLAN.
   c. Add the VRRP version for the VRRP instance (see Configuring VRRP Version Support on page 1141).
   d. Create the VRRP router instance for the intended VRRP master (see Creating and Deleting VRRP Router Instances on page 1139).
   e. Add the IP address defined in Step 1b as a VRRP router IP address (see Adding and Deleting VRRP Router IP Addresses on page 1139).
   f. Enable VRRP on the switch (see Enabling and Disabling VRRP and VRRP Router Instances on page 1143).
2. Configure each backup VRRP router as follows:
   a. Create a VLAN to serve as the VRRP router VLAN. This name must match the name used for the appropriate VRRP IP address owner.
   b. Add an IP address to the VRRP VLAN. This address must be different from the IP address assigned to the intended VRRP master, but it must use the same subnet.
   c. Create the VRRP router instance that will serve as the backup instance (see Creating and Deleting VRRP Router Instances on page 1139).
   d. Configure the priority for the backup VRRP router to a value in the range of 1–254 (see Configuring VRRP Router Priority on page 1140).
   e. Add the same VRRP router IP address that was added to the intended VRRP master instance (see Adding and Deleting VRRP Router IP Addresses on page 1139).
   f. Enable VRRP on the switch (See Enabling and Disabling VRRP and VRRP Router Instances on page 1143).
3. Configure network workstations to use the VRRP router IP address as the default gateway address.

## Creating and Deleting VRRP Router Instances

- To create a *VRRP* router instance, use the following command:

  `create vrrp vlan` *`vlan_name`* **`vrid`** *`vridval`*
- To delete a VRRP router instance, use the following command:

  `delete vrrp vlan` *`vlan_name`* **`vrid`** *`vridval`*

## Adding and Deleting VRRP Router IP Addresses

- To add a *VRRP* router IP address to a switch, use the following command:

  `configure vrrp vlan` *`vlan_name`* **`vrid`** *`vridval`* **`add`** *`ipaddress`*

Note
A VRRP routing instance can support IPv4 or IPv6 addresses, but it cannot support both.

- To delete a VRRP router IP address from a switch, use the following command:

  `configure vrrp vlan vlan_name vrid vridval delete ipaddress`

## Adding an IPv6 Link Local Address to a VRRP Router

To add an IPv6 link local address to a *VRRP* router, enter the following command:

`configure vrrp vlan vlan_name vrid vridval add virtual-link-local vll_addr`

## Configuring the VRRP Router Advertisement Interval

To configure the *VRRP* router advertisement interval, use the following command:

`configure vrrp vlan vlan_name vrid vridval advertisement-interval interval [{seconds} | centiseconds]`

Note

We recommend that you configure the same router advertisement interval in all VRRP routers. VRRPv3 supports a 40 second maximum advertisement interval.

Note

If the advertisement interval is different on the master and backup switches, the master switch's advertisement interval is used by the backup as well.

## Configuring VRRP Router Authentication

To configure *VRRP* router authentication, use the following command:

`configure vrrp vlan vlan_name vrid vridval authentication [none | simplepassword password]`

Note

VRRP router authentication is obsolete in VRRPv3. For backward compatibility, this feature is still supported in VRRPv2.

## Configuring Master Preemption

- To enable *VRRP* master preemption and configure the preempt delay timer, use the following command:

  `configure vrrp vlan vlan_name vrid vridval preempt {delay seconds}`
- To disable VRRP master preemption, use the following command:

  `configure vrrp vlan vlan_name vrid vridval dont-preempt`

## Configuring VRRP Router Priority

To configure the priority for a *VRRP* router, use the following command:

```
configure vrrp vlan vlan_name vrid vridval priority priorityval
```

Note

If VRRPv3 and VRRPv2 routers participate in VRRP instance, the VRRPv3 routers should be configured with a higher priority to ensure that they win master elections over VRRPv2 routers.

## Configuring the Accept Mode

To configure the accept mode, enter the following command:

```
configure vrrp vlan vlan_name vrid vridval accept-mode [on | off]
```

## Configuring NTP VRRP Virtual IP support

To configure the *VRRP* virtual IP as NTP server address, enter the following command:

```
configure vrrp vlan vlan_name vrid vridval accept mode [on|off]
```

Enabling the **accept** mode allows the switch to process non-ping packets that have a destination IP set to the virtual IP address.

## Configuring VRRP Version Support

To configure *VRRP* version support, use the following command:

```
configure vrrp vlan vlan_name vrid vridval version [v3-v2 | v3 | v2]
```

## Configuring VRRP Tracking

### *Configuring the Tracking Mode*

To configure the tracking mode, enter the following command:

```
configure vrrp vlan vlan_name vrid vridval track-mode [all | any]
```

### *Adding and Deleting Tracked Routes*

- To add a tracked route, enter the following command:

  ```
  configure vrrp vlan vlan_name vrid vridval add track-iproute
  ipaddress/masklength
  ```

- To delete a tracked route, enter the following command:

  ```
  configure vrrp vlan vlan_name vrid vridval delete track-iproute
  ipaddress/masklength
  ```

### *Adding and Deleting Tracked VLANs*

- To add a tracked *VLAN*, enter the following command:

  ```
  configure vrrp vlan vlan_name vrid vridval add track-vlan
  target_vlan_name
  ```

- To delete a tracked VLAN, enter the following command:
  `configure vrrp vlan vlan_name vrid vridval delete track-vlan target_vlan_name`

*Adding and Deleting Tracked Pings*

- To add a tracked ping, enter the following command:
  `configure vrrp vlan vlan_name vrid vridval add track-ping ipaddress frequency seconds miss misses`
- To delete a tracked ping, enter the following command:
  `configure vrrp vlan vlan_name vrid vridval delete track-ping ipaddress`

## Configuring VRRP Fabric Routing

To configure the fabric routing feature, enter the following command:

`configure vrrp {vlan vlan_name vr vr_id | all} fabric-routing [on | off]`

## Configuring Include and Exclude Host Ports

To configure include and exclude host ports:

```
configure vrrp vlan vlan1 vrid 1 host-mobility on excluded-ports add 1,10
configure vrrp vlan vlan1 vrid 1 host-mobility off excluded-ports delete 10
```

The `show vrrp details` output shows:

```
VLAN: v1 VRID: 1 VRRP: Disabled State: Backup
Virtual Router: VR-Default
Priority: 100(backup) Advertisement Interval: 1 sec
Version: v3 Preempt: Yes Preempt Delay: 0 sec
Virtual IP Addresses:
Accept mode: Off
Host-Mobility: On
Host-Mobility Exclude-Ports: 1
Tracking mode: ALL
Tracked Pings: -
Tracked IP Routes: -
Tracked VLANs: -
Fabric Routing: On
* indicates a tracking condition has failed
```

## Redistributing Host Routes Using OSPF

To redistribute the host route using *OSPF*:

```
enable ospf export host-mobility cost 0 type ase-type-2
```

## Configuring Host-Mobility

- To enable host-mobility, use the following command:

  `configure vrrp {vlan} vlan_name vrid vridval host-mobility [{on | off} {exclude-ports [add | delete] port_list}]`

- To enable redistribution of host-mobility routes to *OSPF*, use the following command:

  ```
  enable ospf export [bgp | direct | e-bgp | host-mobility |i-bgp | rip
  | static | isis | isis-level-1 | isis-level-1-external | isis-level-2
  | isis-level-2-external] [cost cost type [ase-type-1 | ase-type-2]
  {tag number} | policy-map]
  ```

- To define that host-mobility is the preferred route, use the command:

  ```
  configure iproute {ipv4} priority [blackhole | bootp | ebgp | host-
  mobility | ibgp | icmp | isis | isis-level-1 | isis-level-1-external |
  isis-level-2 | isis-level-2-external | mpls | ospf-as-external | ospf-
  extern1 | ospf-extern2 | ospf-inter | ospf-intra | rip | static host-
  mobility] priority {vr vrname}
  ```

# Managing VRRP

## Enabling and Disabling VRRP and VRRP Router Instances

- To enable *VRRP* or a VRRP router instance, enter the following command:

  ```
  enable vrrp {vlan vlan_name vrid vridval}
  ```

- To disable VRRP or a VRRP router instance, enter the following command:

  ```
  disable vrrp {vlan vlan_name vrid vridval}
  ```

## Clearing VRRP Counters

To clear the *VRRP* counters, enter the following command:

```
clear counters vrrp {{vlan vlan_name} {vrid vridval}}
```

# Displaying VRRP Information

## Displaying VRRP Router Information

To display *VRRP* router information for one or all VRs, enter the following command:

```
show vrrp {virtual-router {vr_name}} {detail}
```

## Displaying VRRP Router Information and Statistics for VLANs

To display *VRRP* information or statistics for a *VLAN*, enter the following command:

- Display VRRP information or statistics for a VLAN.

  ```
  show vrrp vlan vlan_name | vlan_list {stats}
  ```

## Displaying VRRP Tracking Information

To view the status of tracked devices, enter the following command:

```
show vrrp {virtual-router {vr_name}} {detail}
```

### Displaying Host-Mobility Information

- To display the host-mobility configuration, use the following command:

  ```
  show vrrp {virtual-router {vr_name}} {detail}
  ```

- To display the *OSPF* host-mobility export configuration that is used to redistribute host-mobility routes, use the following command:

  ```
  show ospf
  ```

- To display the *OSPFv3 (Open Shortest Path First version 3)* host-mobility export configuration that is used to redistribute host-mobility routes, use the following command:

  ```
  show ospfv3
  ```

- To display the *VRRP* host-mobility routes that are created, use the following command:

  ```
  show iproute {ipv4} {priority | vlan vlan_name | permanent |
  ip_address netmask | summary} {multicast | unicast} {vr vrname}}
  ```

- To display the VRRP host-mobility route priority configuration, use the following command:

## VRRP Configuration Examples

### Simple VRRP Network Examples

The topology for a simple *VRRP* network example is shown in Figure 174 on page 1123.

Switch A is the IP address owner, and Switch B is configured as the backup.

The configuration commands for switch A are as follows:

```
configure vlan vlan1 ipaddress 192.168.1.3/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 priority 255
configure vrrp vlan vlan1 vrid 1 add 192.168.1.3
enable vrrp
```

The configuration commands for switch B are as follows:

```
configure vlan vlan1 ipaddress 192.168.1.5/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 add 192.168.1.3
enable vrrp
```

### VRRP Load Sharing Example

You can use two or more *VRRP*-enabled switches to provide a fully redundant VRRP configuration that supports load sharing on your network.

The topology for a load sharing example is shown in Figure 175 on page 1124. Switch A is the IP address owner for VRRP router instance 1 and the backup for VRRP instance 2. Switch B is the IP address owner for VRRP router instance 2 and the backup for VRRP instance 1.

The configuration commands for switch A are as follows:

```
configure vlan vlan1 ipaddress 192.168.1.3/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 priority 255
configure vrrp vlan vlan1 vrid 1 add 192.168.1.3
create vrrp vlan vlan1 vrid 2
configure vrrp vlan vlan1 vrid 2 add 192.168.1.5
enable vrrp
```

The configuration commands for switch B are as follows:

```
configure vlan vlan1 ipaddress 192.168.1.5/24
create vrrp vlan vlan1 vrid 2
configure vrrp vlan vlan1 vrid 2 priority 255
configure vrrp vlan vlan1 vrid 2 add 192.168.1.5
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 add 192.168.1.3
enable vrrp
```

## VRRP Tracking

The following figure is an example of *VRRP* tracking.

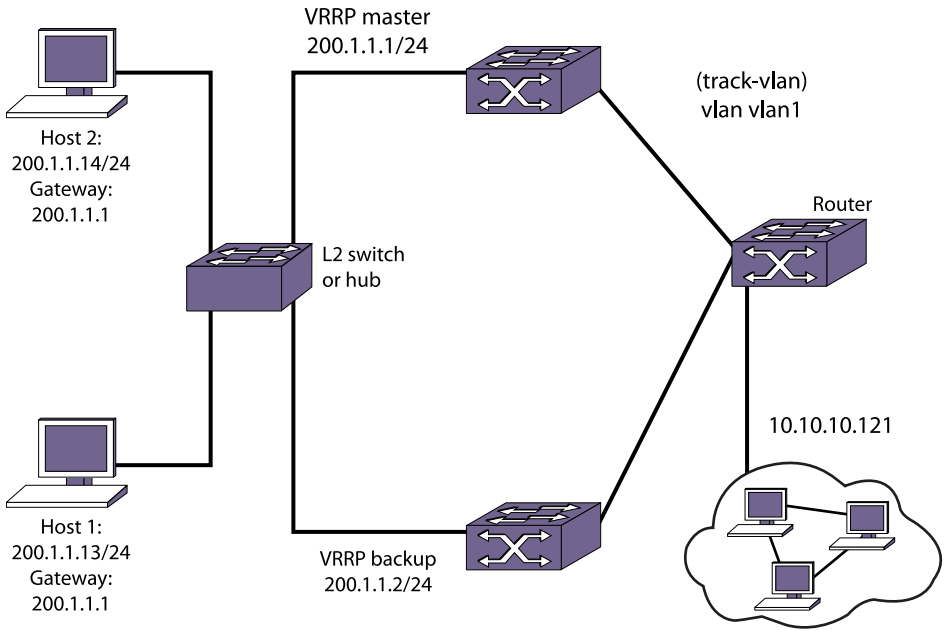



**Figure 178: VRRP Tracking**

- Configure *VLAN* tracking, as shown in Adding and Deleting Tracked VLANs on page 1141 .

  ```
  configure vrrp vlan vrrp1 vrid 2 add track-vlan vlan1
  ```

  Using the tracking mechanism, if VLAN1 fails, the VRRP master realizes that there is no path to the upstream router through the master switch and implements a VRRP failover to the backup.

- Configure route table tracking, as shown in Adding and Deleting Tracked Routes on page 1141.

  ```
  configure vrrp vlan vrrp1 vrid 2 add track-iproute 10.10.10.0/24
  ```

  The route specified in this command must exist in the IP routing table. When the route is no longer available, the switch implements a VRRP failover to the backup.
- Configure ping tracking, as shown in Adding and Deleting Tracked Pings on page 1142.

  ```
  configure vrrp vlan vrrp1 vrid 2 add track-ping 10.10.10.121 frequency
  2 miss 2
  ```

  The specified IP address is tracked. If the fail rate is exceeded, the switch implements a VRRP failover to the backup. A VRRP node with a priority of 255 may not recover from a ping-tracking failure if there is a Layer 2 switch between it and another VRRP node. In cases where a Layer 2 switch is used to connect VRRP nodes, we recommend that those nodes have priorities of less than 255.

## VRRP Host-Mobility Configuration Example

Switch A

```
create vrrp vlan vlan1 tag 10
configure vrrp vlan vlan1 ad port 2,7 tag
configure vlan vlan1 ipaddress 1.1.1.1/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 add 1.1.1.5
enable vrrp
configure vrrp vlan vlan1 vrid 1 host-mobility on excluded-ports add 2
configure ospf add vlan vlan1 area 0 passive
create vlan uplink tag 100
configure vlan uplink add port 5 tag
configure vlan uplink ipaddress 10.1.1.1/24
configure ospf add vlan uplink area 0
configure ospf add vlan vlan1 area 0 passive
enable ospf export host-mobility cost 0 type ase-type-2
enable ospf
```

Switch B

```
create vrrp vlan vlan1 tag 10
configure vrrp vlan vlan1 ad port 2,3,7 tag
configure vlan vlan1 ipaddress 1.1.1.2/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 add 1.1.1.5
enable vrrp
configure vrrp vlan vlan1 vrid 1 host-mobility on excluded-ports add 2,3
create vlan uplink tag 100
configure vlan uplink add port 10 tag
configure vlan uplink ipaddress 10.1.1.2/24
configure ospf add vlan uplink area 0
configure ospf add vlan vlan1 area 0 passive
enable ospf export host-mobility cost 0 type ase-type-2
enable ospf
```

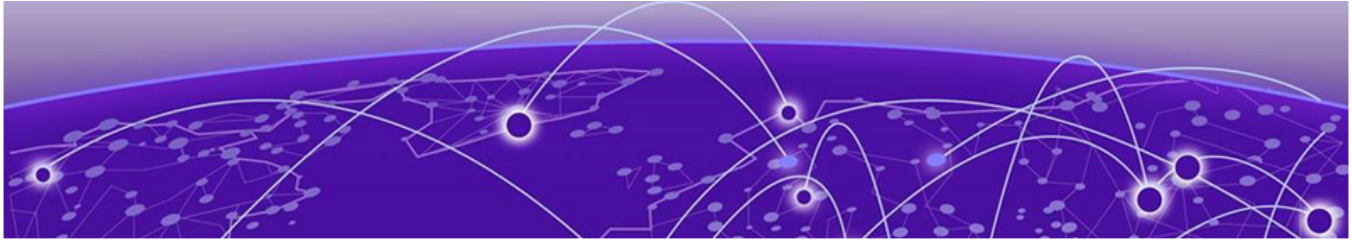
Switch C

```
create vrrp vlan vlan1 tag 10
configure vrrp vlan vlan1 ad port 3,7 tag
configure vlan vlan1 ipaddress 1.1.1.1/24
create vrrp vlan vlan1 vrid 1
configure vrrp vlan vlan1 vrid 1 add 1.1.1.5
enable vrrp
```

```
configure vrrp vlan vlan1 vrid 1 host-mobility on excluded-ports add 3
create vlan uplink tag 100
configure vlan uplink add port 15 tag
configure vlan uplink ipaddress 10.1.1.3/24
configure ospf add vlan uplink area 0
configure ospf add vlan vlan1 area 0 passive
enable ospf export host-mobility cost 0 type ase-type-2
enable ospf
```

Switch D

```
create vlan uplink tag 100
configure vlan uplink add port 5,10,15 tag
configure vlan uplink ipaddress 10.1.1.4/24
configure ospf add vlan all area 0
enable ospf
```



# MPLS



*MPLS (Multiprotocol Label Switching)* is a connection-oriented technology that allows routers to make protocol-independent forwarding decisions based on fixed-length labels. This chapter provides an overview and discusses how to configure, monitor and troubleshoot the MPLS feature. The chapter provides specific configuration examples.

## MPLS Overview

*MPLS* provides advanced IP services for switches that contain advanced ASICs that support MPLS.

To configure MPLS on your switch, you need the MPLS enabled feature pack license .

Note
MPLS and MPLS subfeatures are supported on the platforms listed for specific features in the *Feature License Requirements* document.

Note
ExtremeXOS MPLS does not support Graceful Restart. Any restart of the MPLS process or failover to a backup node requires re-signaling of all MPLS-related connections and entities.

### How MPLS Works

*MPLS* is a connection-oriented technology that allows routers to make protocol-independent forwarding decisions based on fixed-length labels.

The use of MPLS labels enables routers to avoid the processing overhead of delving deeply into each packet and performing complex route lookup operations based upon destination IP addresses.

MPLS protocols are designed primarily for routed IP networks and work together to establish multiple, unidirectional Label Switched Path (LSP) connections through an MPLS network. Once established, an LSP can be used to carry IP traffic or to tunnel other types of traffic, such as bridged MAC frames. The tunnel aspects of LSPs, which are important in supporting virtual private networks (VPNs), result from the fact that forwarding is based solely on labels and not on any other information carried within the packet.

The MPLS protocols operate on Label Switch Routers (LSRs). The router where an LSP originates is called the ingress LSR, while the router where an LSP terminates is called the egress LSR. Ingress and egress LSRs are also referred to as Label Edge Routers (LERs). For any particular LSP, a router is either an ingress LER, an intermediate LSR, or an egress LER. However, a router may function as an LER for one LSP, while simultaneously functioning as an intermediate LSR for another LSP.

The following figure illustrates an MPLS network.

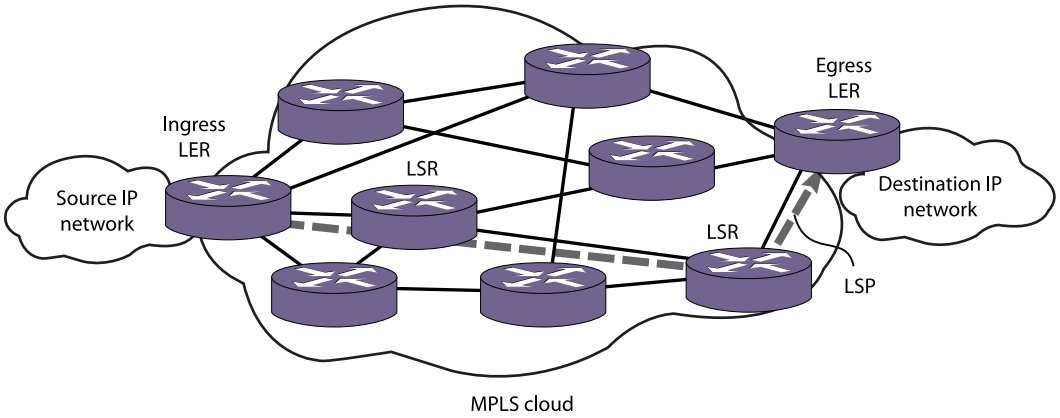



**Figure 179: MPLS Network**

In an MPLS environment, incoming packets are initially assigned labels by the ingress LER. The labels allow more efficient packet handling by MPLS-capable routers at each point along the forwarding path.

An MPLS label essentially consists of a short fixed-length value carried within each packet header that identifies a Forwarding Equivalence Class (FEC). The FEC tells the router how to handle the packet. An FEC is defined to be a group of packets that are forwarded in the same manner. Examples of FECs include an IP prefix, a host address, or a *VLAN (Virtual LAN)* ID.

Note

The label concept in MPLS is analogous to other connection identifiers, such as an ATM VPI/VCI or a Frame Relay DLCI.

By mapping to a specific FEC, the MPLS label efficiently provides the router with all of the local link information needed for immediate forwarding to the next hop. MPLS creates an LSP along which each LSR can make forwarding decisions based solely upon the content of the labels. At each hop, the LSR simply strips off the Incoming label and applies a new Outgoing label that tells the next LSR how to forward the packet. This allows packets to be tunneled through an IP network.

## MPLS Protocol Preference

When LSPs from different protocols exist to the same destination/FEC, LSPs from only one protocol are used at a time for that destination/FEC. The preference, or precedence, for each LSP user is defined as follows.

For both IP/L3VPN and L2VPN the order of preference is:

- RSVP-TE
- LDP
- static

## MPLS Terms and Acronyms

The following table defines common *MPLS* terms and acronyms.

**Table 124: MPLS Terms and Acronyms**

| Term or Acronym | Description |
|---|---|
| CSPF | Constrained Shortest Path First. Route selection determined by an algorithm based on available link bandwidth and path cost. |
| DoD | Downstream-on-Demand. Distribution of labels as a result of explicit upstream label requests. |
| DU | Downstream Unsolicited. Distribution of labels downstream without an explicit label request. |
| EXP bits | A three-bit experimental field in an MPLS shim header. |
| FEC | Forward Equivalence Class. A group of packets that are forwarded in the same manner (for example, over the same Label Switched Path). |
| Label | A short, fixed-length identifier used to forward packets from a given link. |
| Label stack | A set of one or more MPLS labels used by MPLS to forward packets to the appropriate destination. |
| Label swapping | Lookup and replacement of an incoming label with the appropriate outgoing label. |
| LDP | Label Distribution Protocol. A protocol defined by the IETF used to establish an MPLS Label Switched Path (LSP). |
| LER | Label Edge Router. A Label Switch Router that is at the beginning (ingress) or end (egress) of an LSP. |
| LSP | Label Switched Path. The unidirectional MPLS connection between two routers over which packets are sent. LSPs are established using LDP or RSVP-TE. |
| LSR | Label Switch Router. A router that receives and transmits packets on an MPLS network. |
| MPLS | MultiProtocol Label Switching. A set of protocols defined by the IETF used to transmit information based on a label-switching forwarding algorithm. |
| NHLFE | Next Hop Label Forwarding Entry. The NHLFE represents the MPLS router next hop along the LSP. |
| PHP | Penultimate Hop Popping. A label stack optimization used for conserving the number of allocated labels. |

**Table 124: MPLS Terms and Acronyms (continued)**

| Term or Acronym | Description |
|---|---|
| PW | Pseudowire. A logical point-to-point connection. |
| RSVP | Resource ReSerVation Protocol (RSVP). A resource setup protocol designed for an integrated services network. |
| RSVP-TE | Resource ReSerVation Protocol-Traffic Engineering (RSVP-TE). The combination of RSVP and MPLS label signaling to provide traffic engineered LSPs as specified in RFC 3209, RSVP-TE: Extensions to RSVP for LSP Tunnels. |
| Shim header | MPLS-specific header information that is inserted between layer-2 and layer-3 information in the data packet. |
| SP | Service Provider. An entity that provides network services for individuals or organizations. |
| TE | Traffic Engineering. The provisioning of an autonomous flow along a specified network path. |
| Transport LSP | Any active LSP used to forward traffic through an MPLS network. |
| VPLS | Virtual Private LAN Service (VPLS). A multipoint Layer 2 VPN service that has the property that all PW tunnels within a VPN are signaled with the same vcid, where the vcid represents the VPN identifier. |
| VPN | Virtual Private Network (VPN). A logical private network domain that spans a public or service provider network infrastructure. |
| VPWS | Virtual Private Wire Service (VPWS). A point-to-point Layer 2 VPN service that operates over MPLS. |

## LDP Support

The Label Distribution Protocol (LDP) is a protocol defined by the IETF for the purpose of establishing *MPLS* LSPs. Using LDP, peer LSRs exchange label binding information to create LSPs. The LDP features supported in this release include:

- Downstream unsolicited label advertisement.
- Liberal label retention.
- Ordered control mode.
- Advertisement of labels for direct interfaces, *RIP (Routing Information Protocol)* routes, and static routes.
- Ability to use multiple IGP routing protocols (for example IS-IS, *OSPF (Open Shortest Path First)*, and *BGP (Border Gateway Protocol)*).
- LDP loop detection.
- Configurable LDP timers.

Note

To use BGP as an IGP routing protocol, issue the `enable mpls ldp bgp-routes` command.

### *LDP Neighbor Discovery*

LDP includes a neighbor discovery protocol that runs over UDP.

Using the basic discovery mechanism, each LSR periodically multicasts a hello message to a well-known UDP port to which all LSRs listen. These hello messages are transmitted to the all routers on this subnet multicast group. When a neighbor is discovered, a hello-adjacency is formed and the LSR with the numerically greater IP address is denoted as the active LSR.

Hello messages must continue to be received periodically for the hello-adjacency to be maintained. The hold time that specifies the duration for which a hello message remains valid can be negotiated by the peer LSRs as part of the HELLO exchange. During the HELLO exchange, each LSR proposes a value and the lower of the two is used as the hold time.

Targeted LDP hello-adjacencies between potentially non-directly connected LSRs are supported using an extended discovery mechanism. In this case, targeted hello messages are periodically sent to a specific IP address.

After the hello-adjacency is formed, the active LSR initiates establishment of a TCP connection to the peer LSR. At this point, an LDP session is initiated over the TCP connection. The LDP session consists of an exchange of LDP messages that are used to set up, maintain, and release the session.

### *Advertising Labels*

You can control whether label mappings are advertised for:

- Direct routes
- *RIP* routes
- Static routes

In these cases, the switch is acting as the egress LER for these LSPs.

### *Propagating Labels*

LDP propagates label mappings for FECs that exactly match a routing table entry.

In the case of label mappings received from an LDP peer, LDP checks for an exactly matching entry with a next hop IP address that is associated with the LDP peer from which a label mapping was received.

### *Label Advertisement Modes*

LDP provides two modes for advertising labels:

- Downstream-on-demand (DoD)
- Downstream unsolicited (DU)

Using DoD mode, label bindings are only distributed in response to explicit requests. A typical LSP establishment flow begins when the ingress LER originates a label request message to request a label binding for a particular FEC (for a particular IP address prefix or IP host address). The label request message follows the normal routed path to the FEC. The egress LER responds with a label mapping message that includes a label binding for the FEC. The label mapping message then follows the routed path back to the ingress LSR, and a label binding is provided by each LSR along the path. LSP establishment is complete when the ingress LER receives the label mapping message.

Conversely, using DU mode, an LSR may distribute label bindings to LSRs that have not specifically requested them. These bindings are distributed using the label mapping message, as in downstream-on-demand mode. From an LDP message perspective, the primary difference using DU mode is the lack of a preceding label request message.

Architecturally, the difference is more significant, because the DU mode is often associated with a topology-driven strategy, where labels are routinely assigned to entries as they are inserted into the routing database. In either case, an LSR only uses a label binding to switch traffic if the binding was received from the current next hop for the associated FEC.

Both label advertisement modes can be concurrently deployed in the same network. However, for a given adjacency, the two LSRs must agree on the discipline. Negotiation procedures specify that DU mode be used when a conflict exists when using Ethernet links. Label request messages can still be used when *MPLS* is operating in unsolicited mode.

The Extreme LDP implementation supports DU mode only.

### *Label Retention Modes*

LDP provides two modes for label retention:

- Conservative
- Liberal

Using conservative label retention mode, an LSR retains only the label-to-FEC mappings that it currently needs (mappings received from the current next hop for the FEC).

Using liberal retention mode, LSRs keep all the mappings that have been advertised to them. The trade-off is memory resources saved by conservative mode versus the potential of quicker response to routing changes made possible by liberal retention (for example, when the label binding for a new next hop is already resident in memory).

The Extreme *MPLS* implementation supports liberal label retention only.

### *LSP Control Modes*

LDP provides two LSP control modes:

- Independent
- Ordered

Using independent LSP control, each LSR makes independent decisions to bind labels to FECs. By contrast, using ordered LSP control, the initial label for an LSP is always assigned by the egress LSR for the associated FEC (either in response to a label request message or by virtue of sending an unsolicited label mapping message).

More specifically, using ordered LSP control, an LSR only binds a label to a particular FEC if it is the egress LSR for the FEC, or if it has already received a label binding for the FEC from its next hop for the FEC. True to its name, the mode provides a more controlled environment that yields benefits such as preventing loops and ensuring use of consistent FECs throughout the network.

The Extreme *MPLS* implementation supports ordered LSP control only.

## MPLS Routing

This section describes how *MPLS* and IP routing work together to forward information on your network.

MPLS provides a great deal of flexibility for routing packets. Received IP unicast frames can be routed normally or tunneled through LSPs. If a matching FEC exists for a received packet, the packet may be transmitted using an LSP that is associated with the FEC. The packet is encapsulated using an MPLS shim header before being transmitted.

Received MPLS packets can be label switched or routed normally toward the destination. Packets that are in the middle of an LSP are label switched. The incoming label is swapped for a new outgoing label and the packet is transmitted to the next LSR. For packets that have arrived at the end of an LSP (the egress end of the LSP), the label is popped. If this label is the bottom of the stack, the shim header is stripped and the packets are routed to the destination as normal IP packets.

Note
Multicast routing is not supported.

An MPLS domain is generally defined to be an *OSPF* or IS-IS autonomous system (AS). You can use MPLS to reach destinations inside one of these AS types. You can also use MPLS to tunnel through all or part of an AS in order to reach destinations outside of the AS.

### *MPLS Layer Details*

*MPLS* can be thought of as a shim-layer between Layer 2 and Layer 3 of the protocol stack. MPLS provides connection services to Layer 3 functions while making use of link-layer services from Layer 2. To achieve this, MPLS defines a shim header that is inserted between the link layer header and the network layer header of transmitted frames. The format of a 32-bit MPLS shim header is illustrated in the following figure.

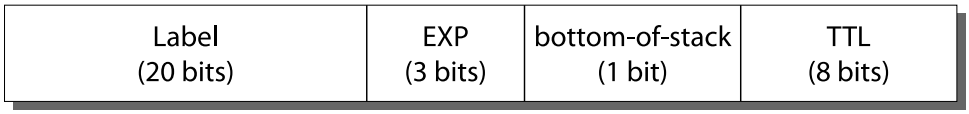



**Figure 180: MPLS Shim Header**

#### MPLS Shim Header

The *MPLS* shim header contains the following fields:

- 20-bit label
- 3-bit experimental (EXP) field

  The EXP field can be used to identify different traffic classes to support the DiffServ *QoS (Quality of Service)* model.
- 1-bit bottom-of-stack flag

  The bottom-of-stack bit is set to 1 to indicate the last stack entry.
- 8-bit Time-To-Live (TTL) field.

  The TTL field is used for loop mitigation, similar to the TTL field carried in IP headers.

#### MPLS Label Stack

The format of an *MPLS* label stack containing two MPLS shim header entries is shown in the following figure.

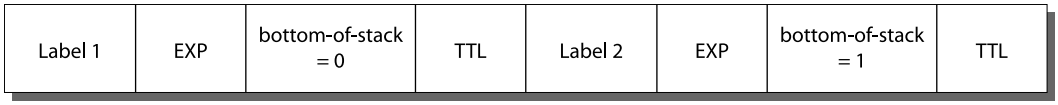



**Figure 181: MPLS Label Stack**

The following figure illustrates the format of a unicast MPLS frame on an Ethernet link. The MAC addresses are those of the adjacent MPLS router interfaces. The x8847 Ethertype value indicates that the frame contains an MPLS unicast packet. A different Ethertype value (x8848) is used to identify MPLS multicast packets.

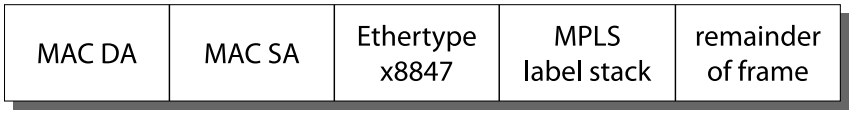



**Figure 182: MPLS Unicast Frame on Ethernet**

The following figure shows the format of a unicast MPLS frame that contains an 802.1Q *VLAN* tag. In both cases, the Ethertype values no longer identify the network layer protocol type. This implies that, generally, the protocol type must be inferable from the MPLS label value(s). For example, when only one type of protocol is carried on a given LSP.


**Figure 183: MPLS Unicast Frame on Tagged Ethernet VLAN**

Note
For more detailed information on MPLS encapsulations, see RFC 3032, MPLS Label Stack Encoding.

### Penultimate Hop Popping

Penultimate hop popping (PHP) is an LSR label stack processing optimization feature. When enabled, the LSR can pop (or discard) the remaining label stack and forward the packet to the last router along the LSP as a normal Ethernet packet.

By popping the label stack one hop prior to the LSP egress router, the egress router is spared having to do two lookups. After the label stack has been popped by the penultimate hop LSR, the LSP egress router must only perform an address lookup to forward the packet to the destination.

PHP label advertisements using implicit NULL labels can be optionally enabled. Support for receiving implicit NULL label advertisements by neighbor LSRs is always enabled. For example, if an LSR advertises implicit NULL labels for IP prefixes, the neighbor LSRs must support PHP.

Note
PHP should not be enabled on egress LER when there is a X670G2 or X770 in a network acting as a PHP LSR.

### Label Binding

Label binding is the process of, and the rules used to, associate labels with FECs.

LSRs construct label mappings and forwarding tables that comprise two types of labels: labels that are locally assigned and labels that are remotely assigned.

Locally assigned labels are labels that are chosen and assigned locally by the LSR. For example, when the LSR assigns a label for an advertised direct interface. This binding information is communicated to neighboring LSRs. Neighbor LSRs view this binding information as remotely assigned.

Remotely assigned labels are labels that are assigned based on binding information received from another LSR.

### Label Space Partitioning

The Extreme implementation partitions its label space as described in the following illustration.

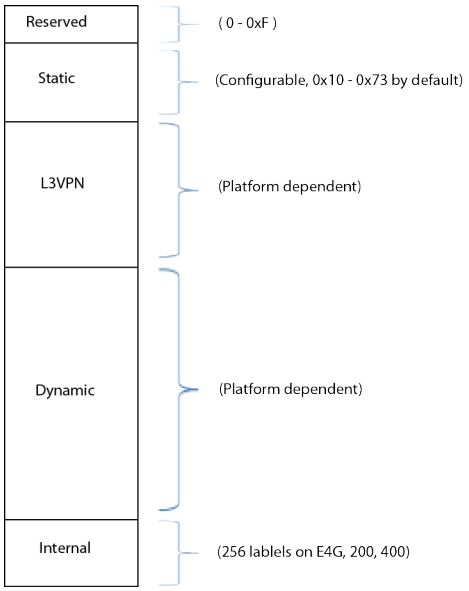



**Figure 184: MPLS Label Space Partitions**

Platforms that support *MPLS* divide the incoming (MPLS egress) label space into static, L3VPN, and dynamic label ranges. The static label range starts immediately after the reserved labels at 16 (0x10), and spans up through the configured maximum number of static labels. In ExtremeXOS 15.4, the `configure mpls label max-static` command is added to allow the static partition range to be changed. Any changes made to the static range affect subsequent ranges.

VRFs used for L3VPNs are allocated from a label range that immediately follows the static range. The remaining label space is called the dynamic label space, and is used by LDP and RSVP-TE. On E4G platforms, an additional block of labels at the end of the dynamic range is used for the internal data path within the hardware. The `show mpls label usage` command displays the label ranges on a given system. The outgoing (MPLS ingress) label space spans the full 20-bit label range, and is used for all outgoing labels for both LSPs and PWs, static and signaled.

No hard limits are imposed on the maximum size of the label stack, other than the constraint of not exceeding the maximum frame size supported by the physical links comprising the LSP.

Jumbo frames should be enabled on all ports, and the jumbo frame size should be set to accommodate the addition of a maximally-sized label stack. For example, a jumbo frame size of at least 1530 bytes is

needed to support a two-level label stack on a tagged Ethernet port, and a jumbo frame size of at least 1548 bytes is needed to support a VPLS encapsulated MPLS frame.

### *Routing Using Matching and Calculated LSP Next Hops*

Normally, a route table prefix is associated with a gateway or next hop IP address.

Using *MPLS*, a route table prefix can also be associated with an LSP that can be used as the next hop. There are two types of LSP next hops that can be used to route a packet to its destination:

- Matching LSP next hop

  An LSP is considered matching with respect to an FEC if it has been associated with that FEC via LDP or RSVP-TE. An example of this is an IPv4 prefix for which a matching label mapping has been received by LDP. Matching LSPs are supported for all route origin types.

- Calculated LSP next hop

  An LSP is considered calculated with respect to an FEC if it has been associated with that FEC through a routing protocol. Both *OSPF* and *BGP* can perform the calculations necessary to associate a route table prefix with an LSP next hop.

The following figure illustrates the concept of matching and calculated LSPs.

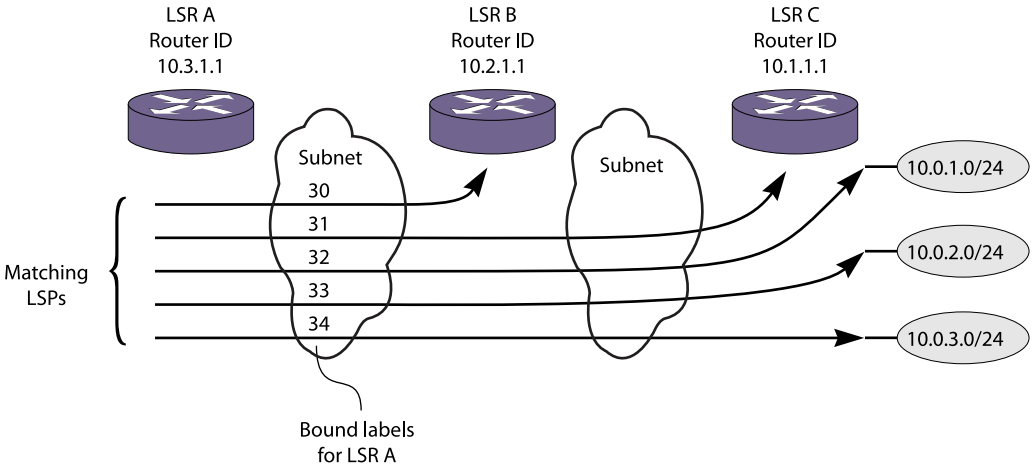



**Figure 185: Matching and Calculated LSP Next Hops**

The following table describes the label bindings in the MPLS forwarding table for LSR A that are maintained for FECs reachable via LSR A to LSR C, shown in the figure above.

**Table 125: Label Bindings for LSR A**

| Destination | Next Hop | Matching LSP Next Hop Label | Calculated LSP Next Hop Label |
|---|---|---|---|
| 10.1.1.1/32 | 10.2.1.1 | 31 | 30 |
| 10.0.1.0/24 | 10.2.1.1 | 32 | 31 |
| 10.0.2.0/24 | 10.2.1.1 | 33 | 31 |
| 10.0.3.0/24 | 10.2.1.1 | 34 | 31 |

### *Matching LSP Next Hops*

A matching LSP next hop is always preferred over a calculated LSP next hop.

Matching LSP next hop entries are added to the route table when an LSP becomes operational. They are deleted when an LSP is no longer operational.

### *OSPF Calculated LSP Next Hops*

Managing calculated LSP next hop entries is more involved.

The *OSPF* Shortest Path First (SPF) algorithm checks the availability of LSPs to remote OSPF routers during a calculation. The intra-area SPF algorithm begins with the calculating router as the root of a graph. The graph is expanded by examining the networks connected to the root and then examining the routers connected to those networks. Continuing in this manner, the graph is built as a series of parent and child nodes. A check is made for a matching LSP next hop as each entry is added. A check is also made for an LSP next hop that can be inherited from the parent node. These inherited LSP next hops are referred to as calculated LSP next hops. Thus, for each route table entry, the modified SPF algorithm determines whether a matching LSP next hop is available and whether a calculated LSP next hop is available for use whenever a matching LSP next hop is not present.

The modification to the SPF algorithm described above is important, because it enables the capabilities provided by LDP or RVSP-TE LSPs to be fully utilized, while minimizing the resources devoted to label management.

For example, in a network where all the LERs/LSRs implement this feature (such as an all-Extreme *MPLS* network), labels only need to be advertised for the router IDs of the LERs/LSRs. Yet, LSPs can still be used to route traffic destined for any OSPF route.

More specifically, LSPs can be used for all routes advertised by OSPF, with the possible exception of LDP LSPs to routes summarized by OSPF area border routers (ABRs). The problem with using routes summarized by OSPF ABRs is that route summarization can prevent label mappings from being propagated for the links internal to the area being summarized, since an LSR only propagates LDP labels for FECs that exactly match a routing table entry.

### *BGP Calculated LSP Next Hops*

*BGP* can also calculate how to use LSPs to reach BGP next hops.

For example, an IBGP session is established across the *OSPF*/*MPLS* backbone, and the communicating routers run both OSPF and IBGP. When an IBGP route is installed, BGP determines whether a matching LSP next hop exists to the destination. If not, it checks for an LSP to the BGP next hop. If an LSP exists to the BGP next hop, that LSP is used as an LSP next hop for the IBGP route.

The recalculation requirements for BGP are similar to those for OSPF; when an LSP to a BGP next hop router changes state, the BGP routing table entries must be checked to ensure their LSP next hop information is still valid.

### *LSP Precedence and Interaction*

A longest prefix match (LPM) is determined for all packets.

If an LSP next hop is available, routed IP traffic may be forwarded over an LSP using the LSP next hop. With respect to a given prefix, LSP next hops can be either matching or calculated, and can be based on

LDP, RSVP-TE, or static LSPs. Matching LSP next hops are preferred over calculated LSP next hops. RSVP-TE LSPs are preferred over LDP LSPs, and LDP LSPs are preferred over static LSPs. Also, RSVP-TE LSPs and static LSPs can be individually configured to enable or disable their use as LSP next hops.

Therefore, if a more preferred LSP is established, routed IP traffic may begin to use a new LSP next hop. Likewise, if a preferred LSP is torn down, routed traffic may begin to use the next best LSP next hop. These changes can take place when there is an *OSPF* routing topology change, an LDP label advertisement event, or a RSVP-TE signaling action.

### *Multivendor Support for Calculated LSPs*

Unfortunately, some *MPLS* implementations do not support the ability to forward packets received on an egress LSP to their *OSPF* router ID and/or *BGP* next hop address.

If your MPLS network includes equipment that does not support this type of IP forwarding, you can use configuration commands to explicitly control the use of calculated LSP next hops.

The following commands enable and disable all use of LSP next hops. No IP traffic is routed over an LSP when mpls-next-hop IP routing capability is disabled.

- `enable iproute mpls-next-hop`
- `disable iproute mpls-next-hop`

**Note**

You can enable the use of LSP next hops, or you can enable *DHCP (Dynamic Host Configuration Protocol)*/BOOTP relay. The software does not support both features at the same time.

These commands enable and disable the calculation of LSP next hops by OSPF:

- `enable ospf mpls-next-hop {vr vrf_name}`
- `disable ospf mpls-next-hop {vr vrf_name}`

These commands enable and disable the calculation of LSP next hops by BGP:

- `enable bgp mpls-next-hop`
- `disable bgp mpls-next-hop`

## Layer 2 VPN over MPLS Overview (VPLS and VPWS)

Layer 2 virtual private networking (VPN) services over *MPLS* include Virtual Private LAN Services (VPLS) and Virtual Private Wire Services (VPWS).

These services enable Layer 2 VPN service offerings in a simple manner that is easy to deploy and operate. Layer-2 VPN services, based on a combination of Ethernet and MPLS/IP technologies, are designed to enable service providers to offer Business Ethernet private line services. These services use a simple Layer 2 interface at the customer edge and benefit from the resilience and scalability of an MPLS/IP core.

VPLS provides a virtual LAN between multiple locations. VPWS provides a virtual dedicated line between only two locations.

### *Layer 2 VPN Support*

The LDP Layer 2 VPN implementation includes support for:

- LDP signaling support for pseudowire ID (PWid) FEC for pseudowire (PW) establishment.
- Use of LDP, RSVP-TE, or Static to establish transport LSPs.
- Tunnel endpoints, identified via configured IP addresses.
- Different *VLAN* IDs at each end of a PW, with the VLAN ID set by the egress switch to match that of the locally configured VLAN.
- Operations as VPLS, H-VPLS, or VPWS node.
- VLAN, VMAN, and port edge services (no port-qualified VLAN service).
- Flooding of Layer 2 packets to multiple PWs when operating as a VPLS or H-VPLS node.

Note

The implementation does not include support for pseudowire participation in running the Spanning Tree Protocol.

### *Layer 2 VPN Service Deliminators*

Service deliminators are used to define how the customer is identified to the Layer 2 VPN service.

There are multiple types of service deliminators. The ExtremeXOS software currently supports three types. The first is a *VLAN* service. This service transparently interconnects two or more VLAN segments together over an *MPLS* network. The configured VLAN IDs for the customer switch interfaces are not required to match, as long as the egress LSR overwrites the VLAN tag with the locally defined VLAN ID.

The second service is a VMAN service, which interconnects two or more VMAN segments. The service operates like the VLAN service but uses the VMAN tag instead of the VLAN tag to identify the service. As with the VLAN service, the interconnected VMAN segments do not need to have matching VMAN IDs.

The third service is a port service, which transparently interconnects two or more ports together over an MPLS network. Traffic is transported unmodified between ports.

The VLAN and VMAN services are configured by adding the service VLAN or a VMAN to a VPLS or VPWS. The port service is not explicitly configured but is emulated using a combination of Layer 2 VPN capabilities. First a VMAN must be configured and the port added untagged to the VMAN. The service VMAN is then added to the VPLS or VPWS. At this point all traffic received on the port is VMAN encapsulated for transmission across the Layer 2 VPN. To transmit the traffic across the Layer 2 VPN as it was received on the port, the VPLS or VPWS is configured to exclude the service tag. By excluding the service tag, the VMAN tag is stripped prior to being transmitted from the switch. This configuration provides port mode service and allows one or multiple ports to be associated with a Layer 2 VPN.

### *MPLS Pseudowires*

*MPLS* pseudowire (PW) tunnels are logical connections between two LERs over an LSP.

LDP Pseudowires are signaled based on the configured PW identifier (pwid). The signaled PW label is used to create a two-label-stack shim header on PW encapsulated packets. The outer label is the transport LSP label obtained from LDP or RSVP-TE and the inner label is the signaled PW label. LERs

also signal the PW type when attempting to establish a PW. The ExtremeXOS software supports only the PWid type FEC. The Generalized ID FEC type is currently not supported.

Note
MPLS PWs can also be configured with statically assigned labels.

#### Transporting 802.1Q Tagged Frames

When an 802.1Q Ethernet frame is encapsulated for transport over a VC tunnel, the entire frame is included, except for the preamble and FCS.

There is a configuration option that determines whether the 4-byte *VLAN* tag field is included in the transmitted packet. By default, the tag field is not included. If the tag field is not included, the egress LER may add one. If it is included, the tag service identifier may be overwritten by the egress LER. The ability to add a tag field or to overwrite the service identifier at the egress node allows two (possibly independently administered) VLAN segments with different VLAN IDs to be treated as a single VLAN.

The following command can be used to include the VLAN tag field:

`configure l2vpn [` **`vpls`** *`vpls_name`* `|` **`vpws`** *`vpws_name`*`] {` **`dot1q`** `[` **`ethertype`** *`hex_number`* `|` **`tag`** `[` **`include`** `|` **`exclude`**`]]} {` **`mtu`** *`number`*`}`

This command can also be used to control the overwriting of the 802.1Q ethertype when the VLAN tag field is included. In this case, the ingress node prior to transmitting the encapsulated packet overwrites the ethertype. This allows devices with a configurable VLAN tag ethertype to interoperate.

#### Establishing LDP LSPs to PW Endpoints

Establishing a PW requires both an LSP and a targetted LDP session between the two endpoints.

The local PW endpoint is the *MPLS* LSR ID. The remote PW endpoint is identified using an IP address configuration parameter.

When using LDP to establish the LSPs, each endpoint needs to advertise a label mapping for an LSP to its local endpoint address. To ensure that its LDP peers use the label mapping, a corresponding IGP route should also be advertised for the address. The IGP route can come from any of the supported routing protocols, such as *OSPF*, or IS-IS. For example, when using OSPF, an OSPF route with prefix length 32 should be advertised for the configured IP address.

We recommend that you configure a loopback *VLAN* using the IP address of the local endpoint (the MPLS LSR ID). Use prefix length 32 for the IP address configured for the loopback VLAN. When you configure a loopback VLAN, the IP address used to identify the endpoint remains active, even when one or more of the LSR VLAN interfaces go down. Should a remote peer normally use one of the down interfaces, the normal IGP and LDP recovery procedures allow the PW to use one of the remaining up interfaces to minimize the network outage.

You should also configure the loopback VLAN for MPLS using the `configure mpls add vlan` *`vlan_name`* command. The addition of the loopback VLAN to MPLS causes LDP to include the IP address in LDP address messages. Some implementations (including the ExtremeXOS software) require

this information to determine the correct LDP session over which to advertise label mappings for VC FECs (see Using LDP to Signal PW Label Mappings on page 1162).

Note
Neither MPLS nor LDP have to be enabled on the loopback VLAN.

There are two options to initiate the LDP advertisement of an LSP to the local MPLS LSR ID when a loopback VLAN has been configured for that IP address:

- Configure MPLS LDP to advertise a direct interface whose IP address matches the LSR ID and has prefix length 32. Use the `configure mpls ldp advertise direct lsr-id` command to do this.
- Configure MPLS LDP to advertise direct interfaces using the `configure mpls ldp advertise direct all` command.

Note
This causes LDP to advertise label mappings for all VLANs that have an IP address configured and have IP forwarding enabled.

While both of the above methods initiate the advertisement of a label mapping for an LSP to the local endpoint, the first method is the preferred method.

### Using LDP to Signal PW Label Mappings

Just as LDP advertises label mappings for LSPs, it can also advertise label mappings for Layer 2 VPNs.

In this case, the signaled FEC information describes a particular Layer 2 VPN. This FEC is often called a Virtual Circuit FEC, or VC FEC. The VC FEC information includes a PWid that is a 32-bit numeric field. Unlike LSP label advertisements that are usually sent to all possible upstream peers, the VC FEC information is sent only to the configured remote endpoint.

When the first Layer 2 VPN is configured to a remote peer, *MPLS* automatically creates a targeted hello adjacency entity for establishing an LDP session. Once the session is established, LDP passes the VC FEC label mapping associated with the Layer 2 VPN. Once VC FECs for the same PW ID have been exchanged in each direction, MPLS is ready to associate the PW with an LSP to the remote endpoint as described in Message Types on page 1187.

To determine the correct LDP session over which to send a VC FEC, MPLS checks the IP addresses learned from its LDP peers via LDP address messages. The ExtremeXOS software MPLS expects to find the IP address of a remote PW peer among the addresses received over the LDP session to that peer.

To ensure that the local endpoint IP address is included in LDP address messages, it is highly recommended to configure MPLS on a loopback *VLAN* as described in Establishing LDP LSPs to PW Endpoints on page 1161.

Use the command `configure mpls add vlan vlan_name` to configure MPLS on the loopback VLAN. It is not required that LDP or MPLS be enabled on the VLAN for the associated IP address to be advertised in LDP address messages.

### Statically Configured Pseudo-Wires

Static *MPLS* PWs are configurable point-to-point emulated circuits that have statically configured MPLS PW labels. Static PWs do not use targeted Label Distribution Protocol (LDP) to negotiate setup and exchange peer status. They can use any type of MPLS tunnel Label Switch Path (LSP). When used in conjunction with static routes and static LSPs, no routing protocol (such as *OSPF* or ISIS), and no label distribution protocol (such as LDP or RSVP-TE) are needed to provision and manage static PWs. Managing this kind of network can provide a disruptive architectural solution for building large backhaul networks that are easy to provision, operate, and incrementally expand. Because protocols are no longer required to set up emulated circuits over MPLS, you now have the capability to proactively, or on-demand, verify end-to-end PW connectivity, to provide remote endpoint status, and offer options to configure redundant PWs to maintain network high availability.

Statically configured PWs provide greater administrative and management control over the network. It also allows MPLS PWs to be configured across a network when no label distribution protocol is running. This can simplify operational management and reduce equipment interoperability issues that can arise when deploying routing packet networks.

You can use the `configure mpls label max-static` command to configure the max number of static labels (labels reserved for static configuration). The maximum number of static labels depends on the underlying hardware platform, and at least 100 labels are reserved for dynamic or signaled labels, such as those used by LDP and RSVP-TE. Use the `show mpls label usage` command to display the current label ranges and usage.

ExtremeXOS supports TDM PWs for CES and Ethernet PWs, for L2VPN VPLS/VPWS, and both types of PWs can be statically configured. The PWs in a VPLS can be a mix of signaled and statically configured, but the corresponding peer PW must be of the same type.

Static PWs are created by adding a peer with configured labels. If the configured labels are not in the allowable range, or are already in use by some other statically configured entity such as static LSPs, then the command is rejected. Once a static PW is created, the labels for that PW can be changed without deleting and re-adding the peer. The CES or L2VPN can remain operational during the change; however, the PW will go down and come back up. If the configured PW labels are accepted, but have not yet replaced the “in-use” labels in hardware, the `show l2vpn detail` and `show ces detail` commands will output an additional line showing the “pending” rx and tx labels. This line will only be shown if necessary, and generally, would not be shown since this is only a transient condition with a small window for its occurrence.

Since static PWs are not signaled, the remote parameters, such as remote Virtual Circuit (VC) status and remote I/F MTU, are not "none" for Ethernet PWs. The local VC status is still calculated and displayed, but is not sent to the peer. Additionally, since the “standby” VC status bit cannot be signaled, PW redundancy cannot be configured for L2VPNs that have static PWs, and Hierarchical Virtual Private LAN Services (H-VPLS) is not supported. TDM PWs already use an associated channel through the “control word”, so the remote VC status is available for those types of PWs. For TDM PWs, the normal TDM signaled parameters, such as payload size and bit-rate, are not sent to the peer and are displayed as a “N/A” in the `show ces detail` command output. There is no OAM support for static PWs, and VCCV is not supported.

Use the `show ces` and `show l2vpn` commands to display PWs that are statically configured. Since static PWs are not signaled, a static PW in a state similar to an LDP PW in signal state, will display in a down state.

Configuring Static Pseudowires

- To configure TDM Circuit Emulation Service over *MPLS* Static PW, use the following commands:

  `configure ces ces_name add peer ipaddress ipaddress fec-id-type pseudo-wire pw_id {static-pw transmit-label outgoing_pw_label receive-label incoming_pw_label}{lsp lsp_name}`

  Use this command to statically configure a new MPLS TDM PW for the specified CES. Both the outgoing (MPLS ingress) and incoming (MPLS egress) PW labels must be specified. The peer must be similarly configured with a static PW that has the reverse PW label mappings. Locally, the *incoming_pw_label* must be unique and is allocated out of the static label space. The *outgoing_pw_label* must match the peer's configured incoming PW label.

  Optionally, you can configure the PW to use any type of tunnel LSP: LDP, RSVP-TE, or Static. In the case of RSVP-TE and LDP, those protocols must be configured and enabled, and an LSP must be established before traffic can be transmitted over the static PW.

  For Static LSPs, only the MPLS ingress LSP (or outgoing LSP) is specified. Unlike signaled PWs, there is no end-to-end PW communication that is used to verify that the PW endpoint is operational, and in the case of static LSPs, that the data path to the PW endpoint is viable.

  In the event of a network fault, if a secondary RSVP-TE LSP is configured or the routing topology changes such that there is an alternate LDP LSP, the static PW will automatically switch LSPs in order to maintain connectivity with the PW endpoint. Static LSPs can be protected proactively by configuring BFD to verify the static LSPs IP next hop connectivity.

  Optionally, the underlying LSP for the PW can be explicitly specified using a named LSP. When a named LSP is explicitly specified, only the specified named LSP is used to carry the PW. In the event that a specified named LSP is withdrawn, the CES remains operationally down until the named LSP is restored.

- To configure L2PN VPLS/VPWS Service over MPLS Static PW, use the following commands:

  `configure {l2vpn} vpls vpls_name add peer ipaddress ipaddress {core} {full-mesh} {static-pw transmit-label outgoing_pw_label receive-label incoming_pw_label}`

  `configure {l2vpn} vpws vpws_name add peer ipaddress ipaddress {static-pw transmit-label outgoing_pw_label receive-label incoming_pw_label}`

  Use these commands to statically configure a new MPLS Ethernet PW for the specified VPLS or VPWS. Both the outgoing (MPLS ingress) and incoming (MPLS egress) PW labels must be specified. You must similarly configure the peer with a static PW that has the reverse PW label mappings. Locally, the *incoming_pw_label* must be unique and is allocated out of the static label space. The *outgoing_pw_label* must match the peer's configured incoming PW label.

  Just like a signaled PW, a static PW can optionally be configured to use any type of tunnel LSP: LDP, RSVP-TE, or Static. In the case of RSVP-TE and LDP, those protocols must be configured and enabled and an LSP must be established before traffic can be transmitted over the static PW. For Static LSPs, only the MPLS ingress LSP (or outgoing LSP) is specified. Unlike signaled PWs, there is no end-to-end PW communication that is used to verify that the PW endpoint is operational, and in the case of static LSPs, that the data path to the PW endpoint is viable.

In the event of a network fault, if a secondary RSVP-TE LSP is configured or the routing topology changes such that there is an alternate LDP LSP, the static PW will automatically switch LSPs in order to maintain connectivity with the PW endpoint. Static LSPs can be protected proactively by configuring BFD to verify the static LSPs IP next hop connectivity. Optionally, the underlying LSP for the PW can be explicitly specified using a named LSP. When a named LSP is explicitly specified, only the specified named LSP is used to carry the PW.

In the event that a specified named LSP is withdrawn, the VPLS/VPWS remains operationally down until the named LSP is restored.

Since VC Status signaling is not supported, the VC Status "standby" bit cannot be used to allow support for PW redundancy and H-VPLS. Consequently, only "core full-mesh" PWs are allowed to have statically configured labels.

- To modify the current configuration of a PW for the specified *ces_mpls_name*, use the following command:

  `configure ces` *ces_name* **peer ipaddress** *ipaddress* **static-pw** [{**transmit-label** *outgoing_pw_label*} {**receive-label** *incoming_pw_label*}]

  Network administrators can use this command to modify the current configuration of a PW for the specified *ces_mpls_name*. The *incoming_pw_label* must be locally unique and is allocated out of the static label space. The *outgoing_pw_label* can be any value and must match the peer's configured incoming PW label.

- To change labels for L2PN VPLS/VPWS service over MPLS Static PW, use the following commands:

  `configure {`**l2vpn**`} vpls` *vpls_name* **peer** *ipaddress* **static-pw** [{**transmit-label** *outgoing_pw_label*} {**receive-label** *incoming_pw_label*}]

  `configure l2vpn vpws` *vpws_name* **peer static-pw** [{**transmit-label** *outgoing_pw_label*} {**receive-label** *incoming_pw_label*}]

  Use these commands to change the labels of a statically configured Ethernet PW for a VPLS or VPWS that already exists. You can specify either, or both the outgoing (MPLS ingress) and incoming (MPLS egress) PW labels. The peer must be similarly configured with a static PW that has the reverse PW label mappings. Locally, the *incoming_pw_label* must be unique and is allocated out of the static label space. The *outgoing_pw_label* must match the peer's configured incoming PW label. The CES or L2VPN can remain operational during the change; however, the PW will go down and come back up.

- To display TDM Circuit Emulation Service, use the following command:

  `show ces {`*ces_name*`} {`**detail**`}`

  For CES services created for use with MPLS, the "Type: Static/Signaled" line in the CES section of the output will show "N/A" until a PW is configured, since this the PW type is not known until the peer is added to the CES. The PW section of the output includes a "PW Signaling" line that will display "LDP" or "None (Static)", depending on the PW configuration.

  Since the configured labels can be changed while the current labels are in-use, there is a small window where the configured labels and in-use labels are different. If you issue the `show ces detail` command during this window, an extra line is output to indicate the extra information.

- To display L2PN VPLS/VPWS service, use the following command:

  `show [ {`**l2vpn**`} vpls {`*vpls_name*`} |` **l2vpn vpws** `{`*vpws_name*`} |` **l2vpn** `] {{`**peer** *ipaddress*`} {`**detail**`} |` **summary**`}`

The non-detail version of this command includes a peer flag that indicates the signaling protocol, if any, for a PW/peer. An “L” indicates LDP is used to signal the PW. A “T” indicates that no signaling is done, and therefore, this is a static PW.

The detail version of this command now displays a “PW Signaling” line that displays “LDP” or “None (Static)”, depending on the PW configuration. The “Local PW Status” shows “--” instead of “Not Signaled”, since the PW status is not currently signaled. For informational purposes, any local faults are still shown.

The “Remote PW Status” and “Remote I/F MTU” always show “--”.Since the configured labels can be changed while the current labels are in-use, there is a small window where the configured labels and in-use labels are different. If you issue the `show l2vpn detail` command during this window, an extra line is output to indicate this extra information.

- To display the MPLS label ranges and usage statistics, use the following command:

  ```
  show mpls label usage
  ```

  This command displays the label ranges on the current running system, including configurable and non-configurable ranges. The output also includes hardware resource usage to provide a better picture about the MPLS hardware utilization and capacity.

**LSP Selection**

A PW can be configured to use any available LSP to the peer endpoint IP address, or the PW can be configured to use one or more specific named LSPs.

In either case, the LSP has to egress (terminate) at the remote endpoint. In the case of an LDP LSP, the LSP's FEC has to be a /32 prefix length to the endpoint IP address. In the case of an RSVP-TE LSP or static LSP, the destination address has to be that of the remote endpoint. When configured to use any available LSP, *MPLS* gives preference to RSVP-TE LSPs, then to LDP LSPs, and lastly, to static LSPs. As a single LSP is chosen to carry the PW traffic, if multiple LSPs of the chosen type exist, the decision of which LSP of this type to use is non-deterministic.

The `configure l2vpn [vpls vpls_name | vpws vpws_name] peer ipaddress [add | delete] mpls lsp lsp_name` command forces the PW to use the specified named LSP. If multiple named LSPs are configured, only one is used to carry the PW. The decision of which of the multiple configured LSPs to use is non-deterministic.

RSVP-TE can be configured to allow specific types of traffic on an LSP. By default, LSPs are used to transport all traffic. Optionally, named LSPs can be configured to allow only IP traffic or only VPN traffic. This can be used to control the LSP selection for specific types of packets. For example, if both LDP and RSVP-TE LSPs exist and the RSVP-TE LSPs are configured to transport only VPN traffic, all IP traffic is forwarded using LDP LSPs. Since RSVP-TE LSPs are preferred over LDP LSPs, VPN traffic flows over the RSVP-TE LSPs. The following command configures this behavior for the specified RSVP-TE LSP:

```
configure mpls rsvp-te lsp lasp_name transport ip-traffic deny
```

For more information see Pseudowire Label Switch Path Load Sharing on page 1166.

*Pseudowire Label Switch Path Load Sharing*

Pseudowire (PW) Label Switch Path (LSP) Load Sharing provides the ability for L2VPN PWs to use upto 16 or 64 Transport LSPs, depending on the platform, for carrying tunneled data across the *MPLS*

network. Previously, Ethernet PWscould only use one tunnel LSP. This feature is related to traffic distribution over *LAG (Link Aggregation Group)*s.

The current HW hashing algorithm on MPLS Transit Nodes uses MPLS labels to derive a hash value. In previous implementations where an L2VPN PW used only one Transport LSP, there were only two MPLS labels to hash on (the outer transport label, and inner VC label) for a given PW. In an eight-port LAG that L2VPN tunneled traffic would only be distributed to one or two links in the LAG. When a PW uses multiple Tunnel LSPs, there are more labels for the HW to use for hashing, and this allows improved distribution over LAG in MPLS Transit Nodes

**Limitations**

This feature has the following limitations:

- Support only for RSVP TE and Static LSPs (Named LSPs).
- Named LSPs must be configured for each PW.
- LSP-Sharing must be enabled in order to be able to configure more than one Transport LSP for a given PW.
- Multiple PWs can use the same LSP, or set of LSPs (same as today).
- A maximum of 16 or 64 LSPs per PW can be configured, depending on the platform.
- If LSP-sharing is disabled when more than 1 LSP is programmed into ExtremeXOS, all LSPs used by the PW, except one, will be removed from ExtremeXOS (HAL). The configuration will not be modified.
- A single default load sharing hashing algorithm is used. When multiple Transport LSPs are configured for a PW, the PW will be UP as long as one Transport LSP is available
- HW counters are only supported for PW packet counts (VC LSP), not Transport LSP (outer label).
- No support for VPWS.
- *ECMP (Equal Cost Multi Paths)* is not supported in slow-path forwarding.
- ECMP is not supported for flood traffic (unknown unicast, multicast, broadcast). These packets will only go over one LSP.
- When multiple LSPs are configured for use by a PW, the HW counts packets for the PW only, and not individual LSPs associated with the PW. So, if there are six LSPs in use for a given PW, you will not be able to display packet counts for each individual Transport LSP.

**Platforms Supported**

This feature is supported on the following platforms:

- Summit X460-G2 (16 LSPs)
- Summit X670-G2 (16 LSPs)
- Summit X670 (64 LSPs per PW)
- Summit X770 (16 LSPs per PW)
- BlackDiamondX8 (64 LSPs per PW)

Note

On a Summit Stack and X770 platforms 16 named LSPs are supported. On X670 and BlackDiamond X8 64 Named LSPs are supported

**Configuring Pseudowire LSP Sharing**

- To enable or disable LSP sharing for L2VPNs, use the following commands:

  ```
  [enable|disable] l2vpn sharing
  ```

  When LSP sharing is disabled, only one named LSP is used for a PW. When LSP sharing is enabled,up to 16 or 64 Named LSPs, depending on the platform, are used for a PW.

  If LSP Sharing is disabled, and more than one Transport LSP is programmed into HW, all but one Transport LSP are removed from HW, and the configuration is preserved.

  If LSP Sharing is enabled, and more than one Transport LSP was previously configured, the remaining LSPs are programmed into HW as they become available for use.
- To configure LSP sharing parameters for PW LSP sharing, use the following command:

  ```
  configure sharing address-based custom [ipv4 [L3-and-L4 | source-only
  | destination-only | source-and-destination] | hash-algorithm [xor |
  crc-16] crc-32 [lower | upper]]
  ```
- To display if PW LSP sharing is enabled, use the following command:

  ```
  show l2vpn sharing
  ```
- To display the status of the L2VPN Sharing configuration, use the following command:

  ```
  show vpls detail
  ```

  If L2VPN Sharing is enabled, and more than one Transport LSP is configured, the output will display the status of each Transport LSP.
- To display an informational message when multiple transport LSPs are configured for a VPLS PW, and when LSP sharing is not enabled, use the following command:

  ```
  configure vpls vpls1 peer 20.20.20.83 add mpls lsp lsp2
  NOTE: To share LSPs in HW, use the command: enable l2vpn sharing.
  ```

  Note

  This message will only be displayed once per switch boot.

*Layer 2 VPN Domains*

Layer 2 VPN domains are created by adding PWs to each peer LSR to build a fully-meshed interconnected VPLS.

For each peer added, a PW is signaled that is used to carry traffic from the local LSR to the remote peer LSR. Flood traffic from the local service (broadcast, multicast, and unknown unicast packets) is replicated and forwarded across all PWs in the VPLS. Each peer receives one copy of the packet for delivery to its locally attached service. As MAC learning occurs on PWs, unicast packets to a known destination MAC address are forwarded to the peer over the PW from which the MAC address was learned.

*MAC Learning*

Learned MAC addresses are associated with the PWs from which the packets are received.

The learned MAC address is always inserted into the *FDB (forwarding database)* as though it was learned on the local service *VLAN* (and not the VLAN identified in the dot1q tag in the received PW

packet). MAC addresses learned from PWs use a different FDB aging timer than those MAC addresses learned on Ethernet interfaces. Different FDB aging timers are maintained for Ethernet and pseudowire Layer 2 VPN FDB entries. By default, both aging timers are set to 300 seconds. However, the aging timers for each type of FDB entry can be configured to different values. Note that PW FDB entries are not refreshed; they age out based on the configured aging timer setting, unless you have disabled the aging timer. Ethernet FDB entries automatically refresh with use, and do not age out unless they are not used for the length of time configured for the aging timer. Any MAC address associated with a PW is automatically cleared from the FDB when the PW label is withdrawn.

Note
MAC learning is disabled for VPWS.

### *Spanning Tree Protocols*

There is some debate as to the benefit of supporting *STP (Spanning Tree Protocol)* within a Layer 2 VPN.

The idea is that STP protocols can be used to provide redundant VPN data paths that can be unblocked if the STP detects a spanning tree topology failure. In general, it is believed that introducing STP to VPLS increases network complexity with very little real benefit.

*MPLS* already provides a significant level of redundancy for the LSP over which a PW is carried. For example, if a PW is using an LDP established LSP, provided there are parallel routed paths to the PW endpoint, the PW automatically shifts from a withdrawn or failed LSP to the next best available LSP. For transport LSPs established using RSVP-TE, secondary LSPs can be configured that can be hot-swapped in the event of a primary LSP failure. Fast-reroute detour LSPs can also be used to protect RSVP-TE LSPs. Thus, even though the underlying transport LSP might have changed, the Layer 2 VPN data plane remains unaffected.

For these reasons, VPLS and STP are not normally enabled on the same *VLAN*. The exception is for local customer network redundancy such as shown in VPLS STP Redundancy Overview on page 1179.

When STP is not enabled on a VPLS VLAN, the BPDU functional address is not inserted into the *FDB* for this VLAN and all received BPDU packets are flooded across the Layer 2 VPN. In this scenario, a single large spanning tree topology spanning all interconnected Layer 2 VPN service sites is constructed. Note that this is not a recommended configuration for a Layer 2 VPN service. Depending on the packet latency within the backbone network, STP timers might need to be tuned to build and maintain reliable topologies.

Currently, most ExtremeXOS software Layer 2 protocols cannot be configured on MPLS Layer 2 VPN domains. Likewise, the following protocols cannot be enabled on a Layer 2 VPN service VLAN:

- *VRRP (Virtual Router Redundancy Protocol)*
- *ESRP (Extreme Standby Router Protocol)*
- EAPS control VLAN

### *IP Protocol Considerations*

The ExtremeXOS software allows an IP address to be configured for a Layer 2 VPN service *VLAN*. This is permitted to allow the switch to use the IP ping and traceroute functions to and from other nodes on the VLAN. It is envisioned that any such IP address is configured temporarily to assist in network verification.

As such, the ExtremeXOS software does not allow IP forwarding to be enabled on a Layer 2 VPN service VLAN for either IPv4 or IPv6. Therefore, any higher-level protocol that requires IP forwarding to be enabled cannot be enabled on a Layer 2 VPN service VLAN. For example, *OSPF* cannot be enabled on a VPLS service VLAN at the PE router. However, OSPF routers can be attached to the service vlan transparently at the customer interfaces. In addition, *IGMP (Internet Group Management Protocol)* snooping cannot be enabled on a VPLS service VLAN.

*MPLS Layer 2 VPN Characteristics*

Characteristics of a Layer 2 VPN include:

- Use of LDP or RSVP-TE to establish the underlying pseudowire transport LSPs.
- Pseudowire endpoints are identified through configured VPLS/VPWS peer IP addresses.
- Configuration of a Layer 2 VPN pseudowire ID doubles as VPN ID.
- Customer packet *VLAN* tags may be overwritten on egress from the pseudowire allowing local service VLANs or VMANs interconnected over the same Layer 2 VPN to have different IDs.
- Customer packet VLAN tags can be included or excluded from packets transmitted on the pseudowire.
- Customer packet VLAN tag ethertype value can be modified before packet is transmitted on the pseudowire, allowing local service VLANs or VMANs interconnected over the same VPLS to have different 802.1Q tag ethertype values.
- Support for full-mesh VPN architectures.
- Support for VLAN, VMAN, and port Layer 2 VPN services.
- Support for enabling and disabling VPLS and/or the VPLS service.
- Support for enabling and disabling VPWS and/or the VPWS service.

## Layer 3 VPN over MPLS Overview

Layer 3 VPN services over *MPLS* are supported using *BGP*. For more information, see #unique_1314.

## H-VPLS Overview

VPLS requires a full mesh of pseudowires between all Provider Edge (PE) peers.

As *MPLS* is pushed to the edge of the network, this requirement presents a number of problems. One problem is the increased number of pseudowires required to service a large set of VPLS peers. In a full-mesh VPLS, pseudowires must be established between all VPLS peers across the core. Full-mesh networks do not scale well due to the number pseudowires that are required, which is p(p-1), where p is the number of peer devices in the network. Hierarchical VPLS (H-VPLS) networks can dramatically increase network scalability by eliminating the $p^2$ scaling problem.

In a hierarchical VPLS network, a spoke node (often a Multi-Tenant Unit—MTU) is only required to establish a pseudowire to a single core PE. Thus the number of pseudowires required in the provider's network is c(c-1) + s, where c is the number of core PE nodes and s is the number of spoke MTU edge devices. This is a significant reduction in the number of pseudowires that need to be established and maintained. For example, a 10-core PE network with 50 MTU devices per core PE requires almost 260,000 pseudowires using a full-mesh VPLS design. A hierarchical VPLS design requires only 590 pseudowires.

An example H-VPLS network is shown in the following figure.

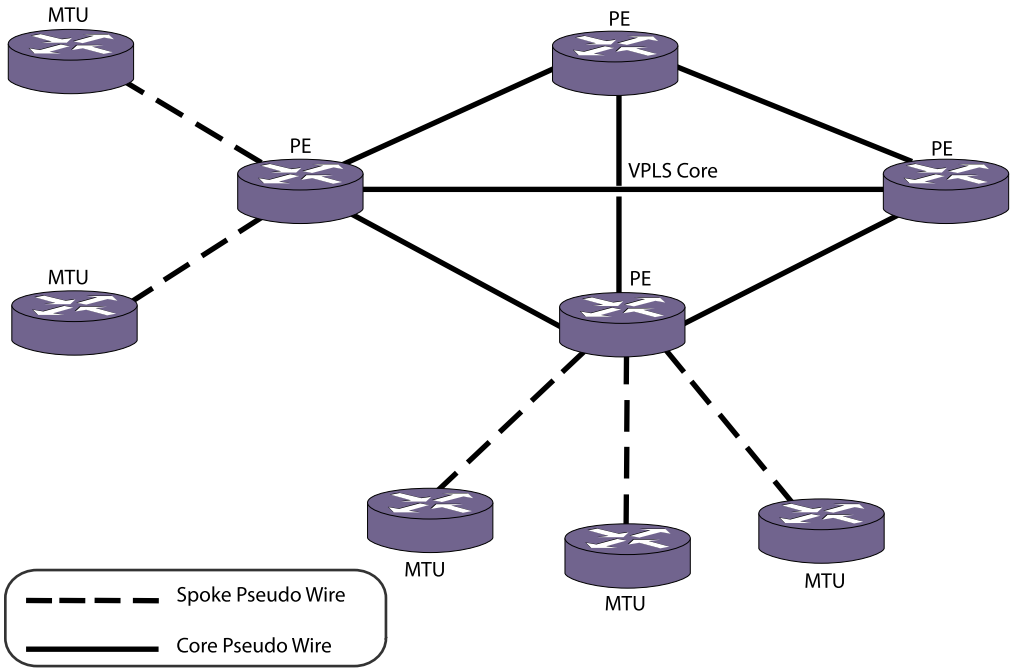



**Figure 186: Example H-VPLS Network**

H-VPLS spokes allow VPLS domains to be constructed hierarchically in a partial-mesh or hub-and-spoke configuration. This is useful for increasing the scaling of VPLS domains that can be supported. Within the context of H-VPLS, a spoke is a VPLS connection between two VPLS peers. Typically, one spoke node provides connectivity to the customer *VLAN* or customer service while its peer, a core node, provides repeater connectivity to other VPLS peers.

The pseudowire hierarchy must be known because the forwarding rules for spoke and core pseudowires are different. Flood traffic received on a core pseudowire from another full-mesh core PE must not be transmitted over other core pseudowires to other PEs. However, flood traffic received on a core pseudowire is transmitted on all spoke pseudowires in the VPLS. Unlike core pseudowires in a full-mesh VPLS, flood traffic received on a spoke pseudowire must be transmitted on all other pseudowires in the VPLS, including pseudowires to other core PEs.

H-VPLS introduces the definition of a pseudowire type. In previous ExtremeXOS releases, only core peers were supported in an interconnected full-mesh configuration. Therefore, all pseudowires were considered to be of the type core. A new spoke pseudowire type is introduced and is highlighted in Figure 187 on page 1173. A VPLS core node that has multiple spoke pseudowires but no configured core pseudowires is informally referred to as a hub.

### *Eliminating Packet Replication by the MTU*

A scaling problem inherent in a full-mesh VPLS network is packet replication.

In a full-mesh configuration, until a node learns over which pseudowire a MAC address is reachable, unknown unicast frames must be flooded on all pseudowires within the VPLS. Packet replication is always true for broadcast and multicast traffic. As the number of VPLS peers increase, the packet replication burden on a node increases. MTU devices attached to a full-mesh core most likely cannot maintain wire-speed forwarding as the number of VPLS peers increase. Hierarchical VPLS eliminates

this MTU burden by requiring only a single pseudowire connection between a spoke and its core PE peer. Packet replication is pushed to the PEs, where it is more suitably handled.

*Simplifying Customer Service Provisioning*

Bandwidth provisioning between an MTU and a PE is extremely difficult with a full-mesh VPLS design.

Since each VPLS instance can require multiple tunnel LSPs, the bandwidth requirements for each tunnel LSP must be separately accepted and individually enforced by every PE a tunnel LSP traverses. Because the provider requirement is to manage the provisioned bandwidth for the VPLS and not each tunnel LSP, the MTU has the added responsibility of rate limiting the aggregate egress traffic across multiple tunnel LSPs on the uplink. Due to packet replication issues described previously, this is not practical.

Hierarchical VPLS designs simplify bandwidth provisioning and management. Because tunnel LSPs from the MTU are terminated at the PE, tunnel LSP resources are easily shared and managed between customers. Thus, traffic for multiple VPLS instances can be transported across a single tunnel LSP. In some cases only a single best-effort tunnel LSP is required between the MTU and the PE. Traffic for each customer is carried over a different pseudowire on the same tunnel LSP. This allows the tunnel LSP to be signaled once, with the desired bandwidth and priority parameters sufficient for providing best-effort service for customers connected to the spoke peer. If a customer upgrades their service or a new customer is connected that requires guaranteed bandwidth, a second tunnel LSP could be signaled with the SLA bandwidth parameters. Once established, the second tunnel LSP can carry traffic for a single customer as a premium service.

*Redundant Spoke Pseudowire Connections*

Redundant spoke pseudowires to PE peers can be configured from an MTU to provide backup connectivity into a VPLS core.

The addition of a redundant spoke pseudowire is optional. By default, when the MPLS Feature Pack license has been applied to the switch, the spoke pseudowire to the primary peer is used to forward packets. In the event of a network failure over the primary pseudowire, the spoke pseudowire to the secondary peer is used to provide redundant VPLS connectivity. An example network is shown in the following figure.

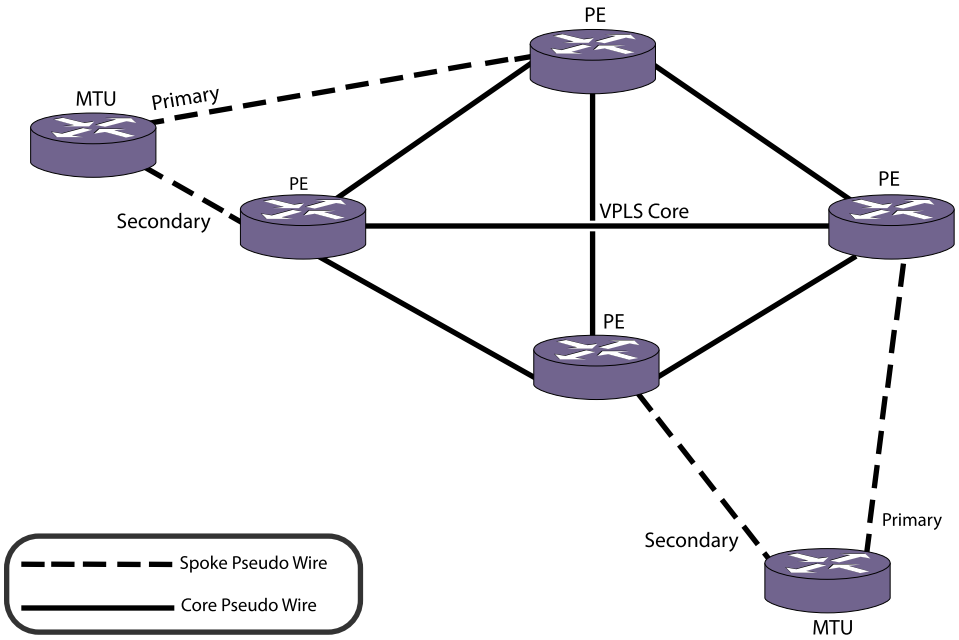



**Figure 187: Example H-VPLS Network with Redundant Spokes**

When both the primary and secondary pseudowires are established, the MTU is responsible for blocking the secondary pseudowire. Any packets received on the secondary pseudowire while the primary pseudowire is active are discarded. This behavior prevents packet-forwarding loops within the L2 VPN.

Since the MTU is responsible for choosing which pseudowire to the VPLS is active, the MTU is uniquely responsible for preventing network loops. The MTU uses only one spoke pseudowire per VPLS and only the label stack associated with the active pseudowire is programmed into the hardware. If the active pseudowire fails, then the label stack for the active pseudowire is removed from hardware. The secondary pseudowire label stack is then installed in the hardware in order to use the redundant VPLS link from the MTU into the VPLS core. Customer connectivity through the MTU should experience minimal disruption.

*IETF RFC 6870* defines the "Preferential Forwarding" status bit to designate which pseudowire is active and can be used to forward user packets. The MTU sets this bit to indicate that a pseudowire is standby (not active) and should not be used to forward user packets.

When a failover occurs from a primary pseudowire to a secondary pseudowire, the MTU clears its *FDB* database of MAC addresses learned over the primary pseudowire. It then begins learning MAC addresses over the new active pseudowire. To inform other nodes to clear their learned MAC database, the MTU can send a MAC address-withdraw message (if this feature is enabled) to the peer PE node of the secondary pseudowire. This PE node can subsequently send its own MAC address-withdraw message to the other VPLS full-mesh core nodes. Upon receipt of the MAC address withdrawal message, each core node clears its database. In this manner, other core nodes re-learn MAC addresses from the correct pseudowire or port.

Packets can be received out-of-order by the VPLS destination device during certain pseudowire failover events. In the redundant VPLS spoke configuration, when the primary pseudowire fails, traffic is immediately switched to the secondary pseudowire. For a very short period of time, there may be packets that are in route via both pseudowires. No attempt to prevent mis-ordered packets from being received is made.

The command to configure the VPLS peer from an MTU to a PE and from a PE to an MTU is fundamentally the same. However, the optional primary and secondary pseudowire keywords are only applicable on the MTU since the MTU is responsible for preventing loops within the VPLS. A switch cannot be configured with a primary and a secondary pseudowire to the same peer within a VPLS. This is an invalid configuration since it provides no redundant protection for a failed PE node.

*MAC Address Withdrawal TLV Support*

MAC address withdrawal is a feature that is used to inform other nodes that certain FDB MAC address entries should be immediately unlearned, rather than waiting for them to age out. Traffic destined to these unlearned MAC addresses is then flooded until the MAC addresses are learned again. MAC address withdrawal is enabled by default when the MPLS Feature Pack License has been applied to the switch, but can be disabled. When this feature is disabled, traffic destined to MAC addresses previously learned over a now unusable pseudowire (PW) is not flooded until the FDB entries eventually age out. However, it can take the VPLS network longer to adjust than when MAC address withdrawal is enabled. This section describes how this feature operates when it is enabled.

After certain network recovery events, MAC addresses should be unlearned. For example, if the MTU's last usable transport LSP for the primary spoke pseudowire goes down, the MTU decides to activate and switch to the secondary pseudowire and send the primary and secondary core PE nodes an LDP notification message containing the appropriate PW status codes. The core PE node of the primary pseudowire flushes its FDB of any MAC addresses learned from the MTU, but does not send any MAC address-withdraw messages. It is the responsibility of the MTU to initiate the sending of MAC address-withdraw messages. If MAC address withdrawal is enabled, the MTU sends a MAC address-withdraw message to the core PE node of the secondary pseudowire. This node, in turn, sends its own MAC address-withdraw message to all the other core PE nodes in the VPLS causing each of these other core PE nodes to flush their FDB of any matching learned MAC addresses specified in the address-withdraw message. By withdrawing MAC addresses immediately, the other core PE nodes are forced to flood traffic destined to the MAC addresses specified in the address-withdraw message. If an alternate VPLS path exists, the new path can be quickly learned without having to wait for the FDB MAC entry to age out.

When a node needs to withdraw a MAC address, it can signal the MAC withdrawal for a VPLS using an address-withdraw message in one of two ways: the MAC address is explicitly specified in a MAC TLV; or an empty MAC TLV is sent indicating that all MAC addresses not learned from that node should be unlearned. Because this information may be propagated to multiple VPLS nodes, a control plane processing trade-off exists. To reduce the processing load, ExtremeXOS sends an empty MAC TLV. Additionally, ExtremeXOS supports the processing of multiple withdraw messages per VPLS, since other vendors may choose not to send an empty MAC TLV.

*Event Log Messages*

The H-VPLS feature has full *EMS (Event Management System)* logging capabilities to capture error and debug information. Messages are logged with the *MPLS* component identifier.

*SNMP Support*

No *SNMP (Simple Network Management Protocol)* support is provided for H-VPLS.

## Protected VPLS and H-VPLS with ESRP Redundancy Overview

Protected VPLS Access enables redundant nodes at the entry to a VPLS or H-VPLS network.

This feature provides fault tolerant connectivity from the customer *VLAN* to the backbone VPLS. This could be implemented by running Layer 2 protocols across the VPLS to block switch ports, but this can lead to sub optimal spanning tree topologies across the VPLS backbone and relatively long outages while the *STP* converges. Instead, the ExtremeXOS software has been enhanced to provide the ability to configure redundant VPLS switches using a dual-homed design that provides fast failover for protected access points.

The first figure below shows fault-tolerant access in a full mesh core VPLS network while the second figure shows fault-tolerant access in a hierarchical VPLS network.

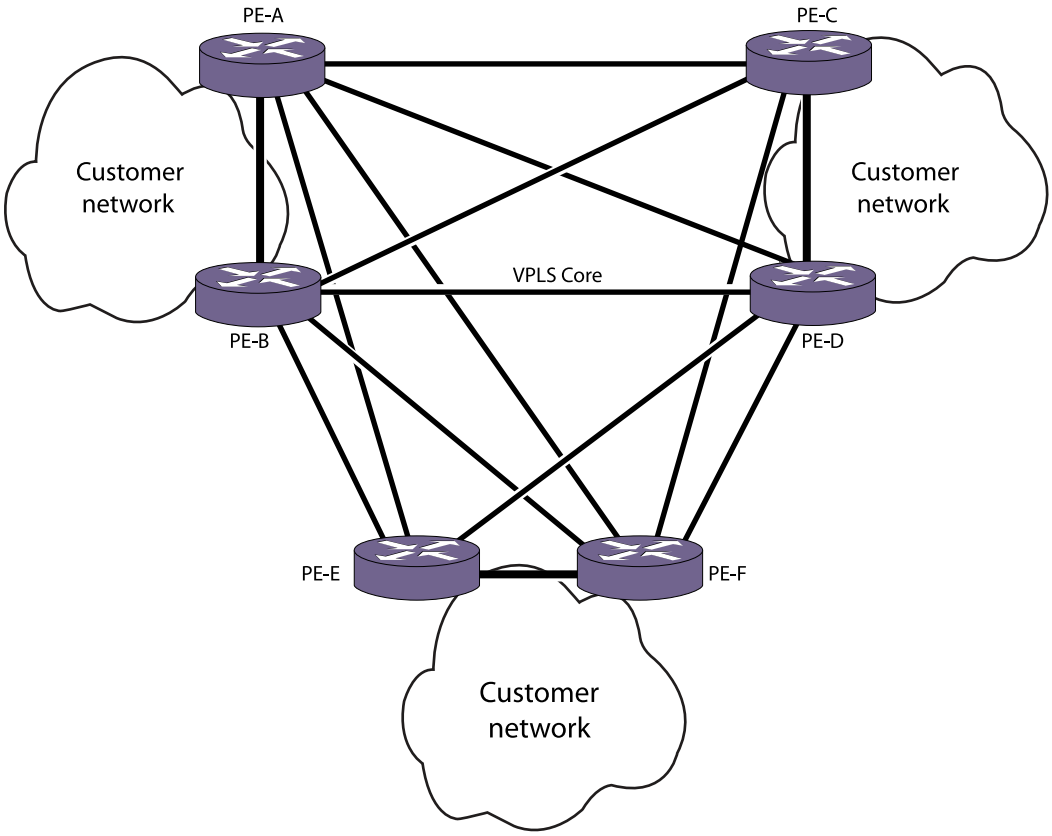



**Figure 188: Example Protected Access VPLS Network**

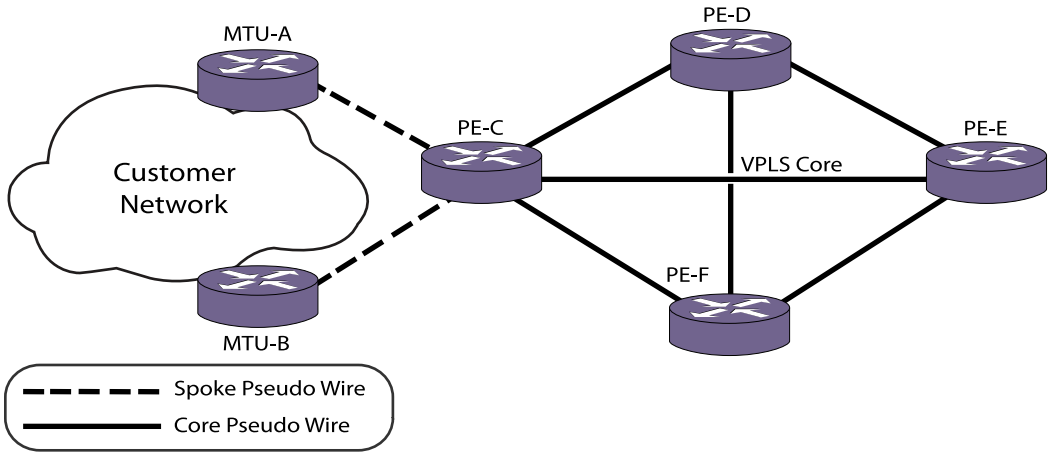



**Figure 189: Example Protected Access H-VPLS Network**

In the above figure, fault tolerance is provided at the customer site by MTU-A and MTU-B. A failure of either MTU-A or MTU-B does not result in any loss of customer connectivity beyond the failover time from one MTU to the other.

*ESRP* is employed to ensure that only one VPLS switch is active at any instant in time for a protected customer access point. Only the ESRP master switch forwards packets between the access network and the backbone VPLS network. This active primary switch retains this status based on a set of predefined tracking criteria. If the configured criteria can be better satisfied by the inactive secondary VPLS switch, the primary VPLS switch relinquishes control and the secondary switch assumes the active role. The secondary switch can also autonomously assume the active role if it detects that the primary switch has failed. This use of ESRP helps to prevent duplicate packet delivery and to prevent broadcast loops when the customer network is a loop topology.

### *Fault Tolerant Access Points Assumptions and Limitations*

The following assumptions and limitations are associated with a fault tolerant access point network:

- This feature does not interoperate with the VPLS redundancy feature in ExtremeWare.
- This feature supports a maximum of two redundant VPLS switches per protected *VLAN* access point.
- This feature operates only with *ESRP* extended mode.
- For ESRP to communicate between neighbor switches, you must configure a separate control VLAN with the same network layout as the set of protected customer VLANs. For example, consider two customer VLANs, VLAN-X and VLAN-Y. Both require protected VPLS access. If both VLAN-X and VLAN-Y have the same network layout (for example, both are part of a single EAPS domain), you must create an ESRP control VLAN that has the same layout as VLAN-X and VLAN-Y. Conversely, if VLAN-X and VLAN-Y do not have the same layout, you must create two separate ESRP domains with each control VLAN following the layout of the associated service VLAN(s).
- All VPLS switches in the control VLAN need to have the same ESRP domain configured. VPLS switches that provide protected access to the VPLS network need to have ESRP enabled while other nodes in the control VLAN need to be ESRP aware.
- The software does not validate the configuration between switches to determine if all VPLS switches for a protected VLAN are configured to be part of the same ESRP.

*H-VPLS Redundant Edge Network*

In the following figure, PE-C still represents a single point of failure. To remove this exposure, the fault tolerant access points can be combined with the redundant spoke pseudowires to produce a redundant edge configuration, as shown below.

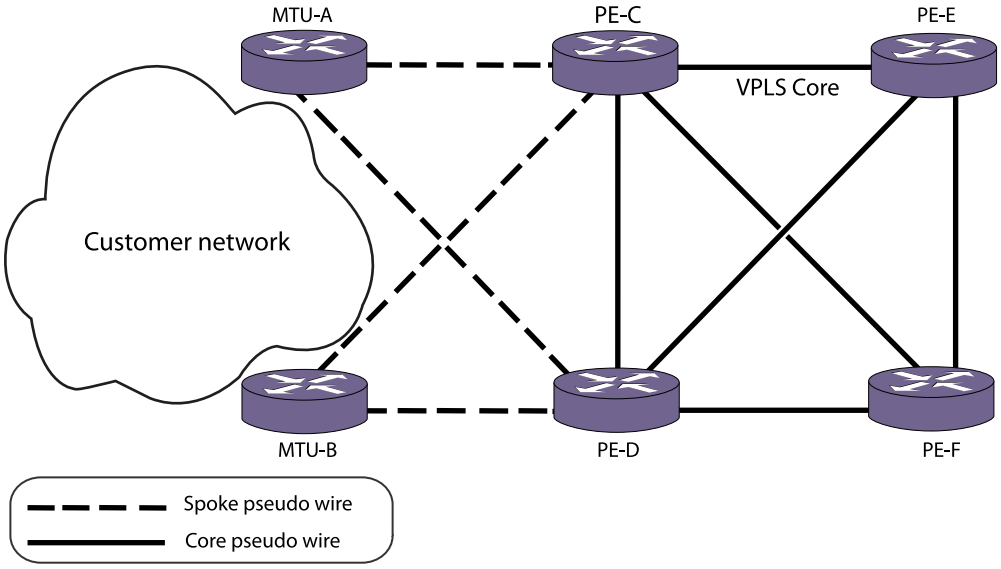



**Figure 190: Example Redundant Edge H-VPLS Network**

In the network shown in the above figure, only one of the pseudowires between the MTUs and their attaching PEs is active at any instant in time. This network provides fault tolerance for a failure at either MTU or at either of the attaching PEs, as well as for the active pseudowire in use between the MTUs and PEs.

*Fault Tolerant VPLS Operation*

To provide the fault tolerance shown in the following figure through the following figure, the redundant network nodes communicate with each other to determine the active primary and inactive secondary status and elect an active master node.

For a VPLS domain type, *ESRP* considers election factors in the following order: standby, active ports, tracking information, stickiness, ESRP priority, and MAC. For more information on the ESRP election priority, see ESRP on page 1092.

For fault tolerant VPLS to function correctly, the ExtremeXOS software imposes restrictions on the configuration options on the VPLS redundancy type ESRP domain. The following table lists the configuration restrictions on the control *VLAN*.

**Table 126: ESRP Configuration Restrictions for VPLS Type Domains**

| No. | Parameter | Restrictions | Remarks |
|---|---|---|---|
| 1 | `mode` | Not configurable | Only extended mode is supported. |
| 2 | `elrp poll` | Always enabled on control VLAN ports | Enabled because control VLANs can have loops. That is, a control VLAN is protecting an S-VLAN in an EAPS ring. |

**Table 126: ESRP Configuration Restrictions for VPLS Type Domains (continued)**

| No. | Parameter | Restrictions | Remarks |
|---|---|---|---|
| 3 | *master VLAN* | No restrictions | The ESRP control VLAN is configured as the master VLAN on the ESRP master and slave. |
| 4 | *member VLAN* | Not configurable | Member VLANs are not configured for the domain since we would need the slave node to perform L2 switching and L3 forwarding. |
| 5 | *track-environment* | Not configurable | Tracking is always done on pseudowires. |
| 6 | *track-VLAN* | Not configurable | Tracking is always done on pseudowires. |
| 7 | *track-IProute* | Not configurable | Tracking is always done on pseudowires. |
| 8 | *track-Ping* | Not configurable | Tracking is always done on pseudowires. |
| 9 | *domain-id* | None | |
| 10 | *elrp-master-poll* | No restrictions | Not configurable, disabled by default. |
| 11 | *elrp-premaster-poll* | No restrictions | Not configurable, disabled by default. |
| 12 | *election policy* | Not configurable | Always set to: standby > ports > track > sticky > priority > mac. |
| 13 | *name* | No restrictions | |
| 14 | *priority* | No restrictions | |
| 15 | *timer* | No restrictions | When the service VLAN is part of an EAPS ring, it is strongly recommended that the hello timer is always greater than the EAPS master health checkup timer because an ESRP switch before EAPS could cause traffic loops. |

The MTU nodes in the following figure signal the PE nodes about active/inactive state using the Status TLV defined in RFC 4447, Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP). This operation is described in Redundant Spoke Pseudowire Connections on page 1172.

### *Performance of Fault Tolerant VPLS Access Points*

Switching times for a VPLS type *ESRP* domain are the same as that for regular domains. Actual time depends on the configured hello timer values and is expected to be in the order of seconds.

### *Deployment and Application Considerations*

The following configuration guidelines should be observed when deploying fault tolerant access points:

- When this feature is deployed with EAPS access rings, it is strongly recommended that the EAPS health check time is configured to a value that is less than or equal to the *ESRP* hello time. Current defaults are one second and two seconds for EAPS health check and ESRP hello, respectively.
- All nodes in the control *VLAN* other than the two VPLS enabled nodes are configured as ESRP-Aware.

*Event Log Messages*

*MPLS* has full *EMS* logging capabilities to capture error as well as debug information. Messages are logged with the MPLS component identifier.

*SNMP Support*

No *SNMP* support is provided for protected VPLS access.

## VPLS STP Redundancy Overview

The following figure shows an example network that uses *STP* to support redundant links to an H-VPLS network.

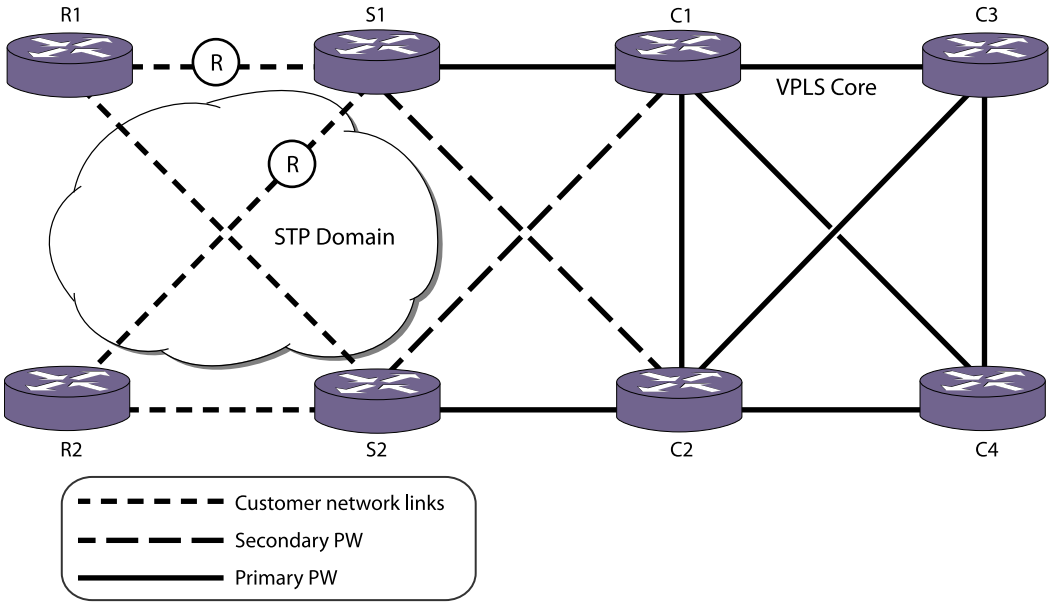



**Figure 191: Redundant Edge H-VPLS Network with STP Example**

The topology in the preceding figure uses redundant VPLS spoke nodes (S1 and S2) and an STP customer network to protect customer access. The redundant VPLS nodes provide protection from the loss of a VPLS node, and STP provides protection from the loss of a node or link in the customer access network. Within the VPLS nodes, VPLS and STP work together to react to topology changes in the customer access *VLAN*.

This topology uses the restricted role feature on access switch ports to control path redundancy. In the following figure, the VPLS nodes S1 and S2 are the lowest priority STP bridges (STP prefers lower priority for root bridge election and shortest path calculation). The S1 ports connected to the R links are configured for STP restricted role mode. To prevent network loops, the restricted role mode in S1 blocks an STP enabled port when STP BPDUs with better information are received from the access network. As shown, the customer traffic uses S2 to access the VPLS network. Should one of the two restricted ports on S1 become unblocked due to a topology change, customer traffic could use both S1 and S2 to access the VPLS network.

The selection of primary and secondary PWs for this configuration is arbitrary. Therefore data paths traversing the spoke nodes S1 or S2 could use either or both core nodes C1 and C2.

In the network shown in the preceding figure, traffic destined to R1 from the VPLS core traverses C2, S2, and R1.

*Failure Recovery Scenario without VPLS STP Redundancy*

If VPLS *STP* redundancy is not configured on the network shown in the following figure and the S2-R1 link fails, S1 unblocks S1-R1 and S2 flushes its *FDB* on the S2-R1 port.

The VPLS core nodes C3 and C4 are unaware of this change and continue to forward any traffic destined for R1 to C2. C2 continues to forward traffic towards S2. This results in data loss because S2 is not able to reach R1 over the customer network.

This data loss continues until one of the following events occurs:

- The FDB entry for R1 ages out in the VPLS core nodes
- R1 sends traffic into the VPLS core allowing its MAC address to be relearned by the VPLS core nodes

Depending on the type of data traffic from R1, the latter scenario might not occur quickly.

*Failure Recovery Scenario with VPLS STP Redundancy*

If VPLS *STP* redundancy is configured on switches S1 and S2 in the following figure and the S2-R1 link fails, S1 unblocks S1-R1 and S2 flushes its *FDB* on the S2-R1 port.

When S2 flushes its FDB, it also sends a flush message to its core peer C2. Upon receipt of this flush message, C2 flushes its FDB entries learned over the PW to S2. C2 also sends flush messages to its core peers C1, C3, and C4. These core nodes then flush their FDB entries learned over their PWs to C2. Any traffic from the VPLS core destined for R1 is flooded until such time as traffic from R1 is forwarded into the VPLS core.

Note

In the example shown in Figure 191 on page 1179, the core nodes are unaware that STP redundancy is configured on S1 and S2. There is no STP redundancy configuration on C1 and C2.

The following figure shows an H-VPLS configuration where the STP network directly connects to redundant spoke nodes. A similar configuration is possible where the STP network directly connects to redundant VPLS core nodes. In this case, the core nodes participating in STP are configured for STP redundancy and originate flush messages to their core peers whenever STP causes a flush on an STP port.

*Requirements and Limitations*

The configuration in the following figure has the following requirements and limitations:

- The VPLS nodes with *STP* redundancy (Switches S1 and S2) must always be the lowest priority STP bridges to ensure that STP port blocking is done by one of the VPLS nodes. For example, the following priorities for the nodes in the following figure will work correctly: S1, priority 8192; S2, priority 4096; R1, priority 32768; and R2, priority 32768.
- For VPLS STP redundancy to work properly, the VPLS nodes must be directly connected to the STP nodes. For example, if a node R3 is added between S2 and R2, node S2 cannot directly detect a failure of the link between R3 and R2.

*Enabling and Disabling VPLS STP Redundancy*

- To enable or disable VPLS *STP* redundancy, use the following commands:

  `configure {l2vpn} vpls vpls_name redundancy [esrp esrpDomain | eaps | stp]`

  `unconfigure {l2vpn} vpls vpls_name redundancy [eaps | esrp | stp]`

## VPLS EAPS Redundancy Overview

To protect your customer access network from link and node failures, you can use the VPLS with Redundant EAPS configuration shown in the following figure.

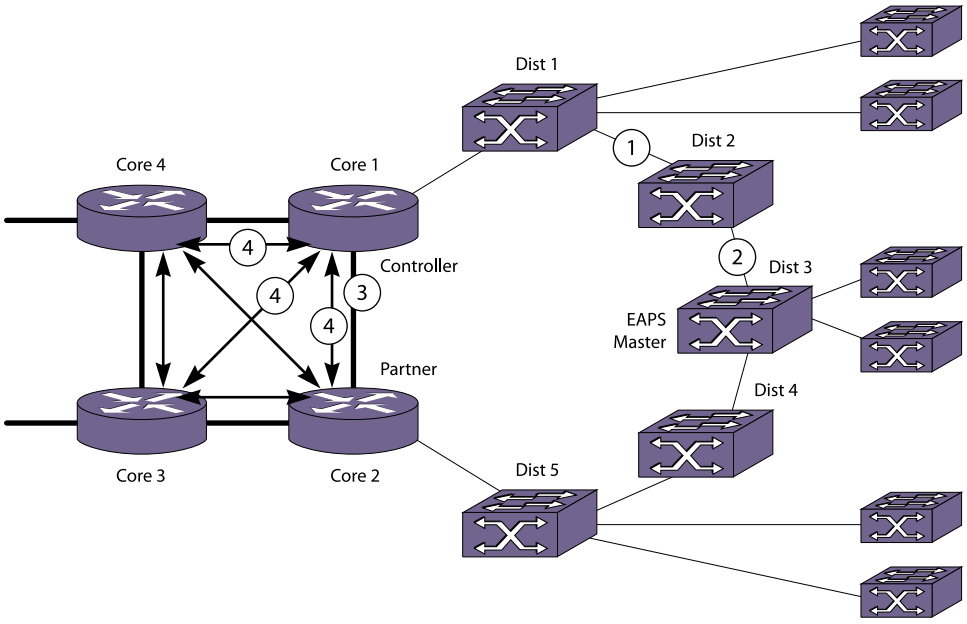



**Figure 192: VPLS with Redundant EAPS Configuration Example**

The topology in the above figure uses redundant VPLS core nodes and an EAPS ring to protect customer access. The redundant VPLS nodes provide protection from the loss of a VPLS node, and the EAPS ring provides protection from the loss of a node or link on the EAPS ring. Within the VPLS nodes, VPLS and EAPS work together to control the use of PWs so that no loops are created.

During normal operation, the Dist2 node is the EAPS ring master, and it blocks the port leading to the link at point 2 in the following figure. The protected *VLAN*s on the EAPS ring do not include the EAPS common link at point 3 in the above figure. The protected VLANS on the EAPS ring use VPLS PWs between the redundant VPLS nodes to connect the two ring segments. This difference in normal EAPS operation is established during configuration by configuring an EAPS-protected VLAN on the core node with only one port on the ring.

To see the commands used to create the configuration shown above, see VPLS with Redundant EAPS Configuration Example on page 1229.

### *Requirements and Limitations*

The solution shown in VPLS EAPS Redundancy Overview on page 1181 has the following requirements and limitations:

- The redundant VPLS nodes must be core nodes.
- An EAPS common link is required between the redundant VPLS nodes.
- The redundant VPLS nodes (Core 1 and Core 2) use PWs to connect to each other and the other core nodes (Core 3 and Core 4).
- The redundant VPLS nodes use PWs to support the EAPS-protected VLANs, not the EAPS common link.
- Works only with EAPS customer attachment ring.
- The EAPS master should *not* be on a VPLS node.
- EAPS state used by VPLS to control state of PWs (Active/Ready).
- EAPS monitors common link state and ring state.
- VPLS on controller node uses EAPS state to set Active/Ready PWs.
- EAPS blocks customer-facing ports as normal.

### *Failure Recovery Scenario 1*

Suppose that a failure occurs at point 1 in the following figure.

The EAPS master detects the topology change (either through a failure notification from a node on the ring or through a hello timeout), and it unblocks the port on the protected *VLAN* at point 2. The Dist 2 node now connects to the VPLS through Core 2 instead of through Core 1.

When the topology changes either on an access ring or the shared port link, the path used to reach customer devices can change. For example, in the following figure, the path that Dist 2 takes to reach other parts of the VPLS network changes following the failure on the access ring at point 1. Prior to the failure, Dist 2 used Core 1 to reach the VPLS network. Following the failure, Dist 2 accesses the VPLS network using Core 2.

When the EAPS master detects a topology change, it sends a flush *FDB* message to its transit nodes. The transit nodes re-learn all the MAC addresses on the ring. However, this flush FDB message is not propagated over the VPLS network. As a result, Core 3 in the following figure still expects to find Dist 2 through the PW between Core 3 and Core 1. Any traffic destined for Dist 2 that is sent to Core 1 will not reach its destination.

To correct this problem, EAPS informs VPLS about any received EAPS flush FDB messages on both the controller and the partner nodes, and VPLS performs a local flush of any MAC addresses learned from the originating nodes. In this example, the EAPS processes in both Core 1 and Core 2 notify VPLS because neither node knows where the access ring is broken. The VPLS services in Core 1 and Core 2 send flush messages to the other VPLS nodes.

### *Failure Recovery Scenario 2*

The recovery scenario is more complicated when the common link fails between Core 1 and Core 2.

The common link is configured to be an EAPS common link, and this configuration requires that one node be designated as the EAPS controller node and the other node be designated as the EAPS partner

node. As described below, the selection of which node is assigned which role must consider the overall customer topology.

If the shared port link fails, the EAPS master node unblocks its port, and VPLS on the EAPS controller node takes the additional action to remove the PWs (point 4 in the following figure) associated with the VPLS from hardware.

In this recovery mode, all traffic to and from the access ring and the rest of the VPLS instance passes through Core 2. When the EAPS controller node detects that the common link is repaired, it enters the preforwarding state as normal. When the controller node exits the preforwarding state, EAPS informs VPLS so that VPLS can reestablish the PWs.

*Failure Recovery Scenario 3*

Now suppose there is a failure on the shared port link (point 3) and on the access ring at point 1.

The recovery actions for this double-failure need to be somewhat different. In this case, even though the core link has failed, both core nodes do not receive a copy of the ring traffic. For example, the only path to the VPLS network for Dist 1 is through the controller core node. In this case, the controller node does not take down its PWs.

It is possible that the customer access network could have parallel EAPS rings that attach to Core 1 and Core 2 as shown in the following figure. In this example, the network connections are broken at each point X and as long as any of the parallel EAPS rings are complete, there is a path to both core VPLS nodes. Thus, the controller node must take down its PWs as long as any of the parallel EAPS rings is complete.

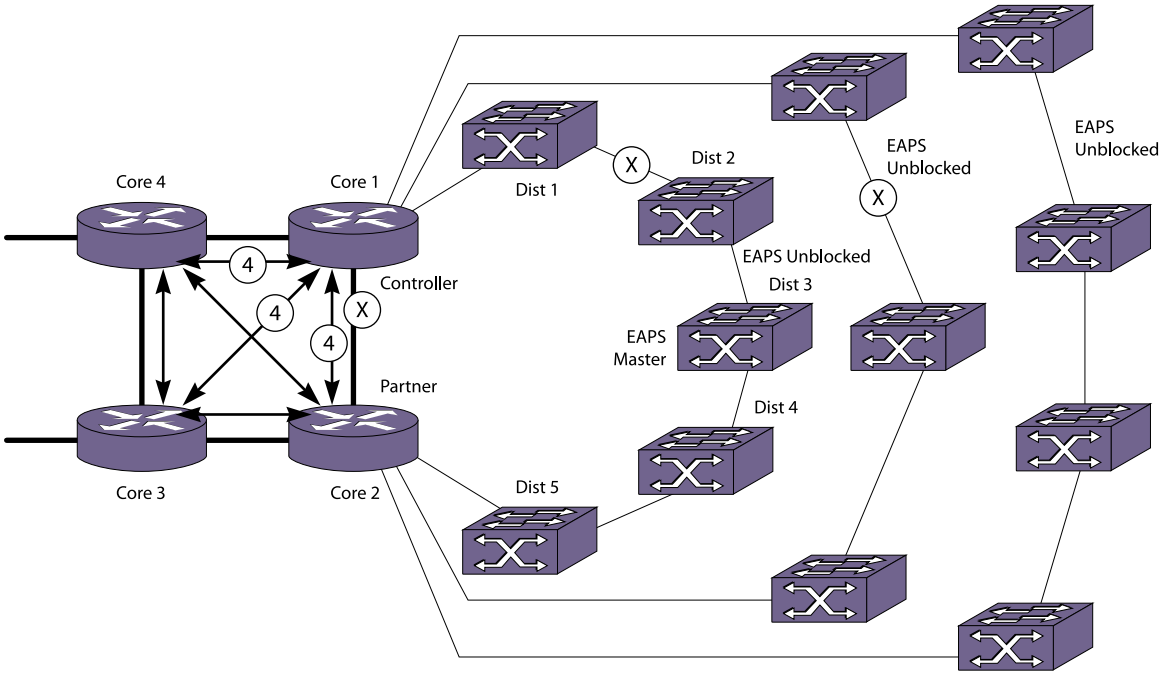



**Figure 193: VPLS with Parallel Redundant EAPS Rings Configuration Example**

**PW Management for Network Failures**

The following table shows how the controller node responds to multiple failures in a network with parallel EAPS rings.

| Ring State | Common Link Up | Common Link Down |
|---|---|---|
| Any parallel ring Up | PWs Active | PWs Inactive |
| All parallel rings Down | PWs Active | PWs Active |

## Selective VLAN Mapping to VPLS

ExtremeXOS currently supports filtering a set of CVIDs received on a VMAN port by classifying the port as a CEP. Since a VPLS can already be configured to carry VMAN traffic, ExtremeXOS 15.4 release extends support for mapping a CEP to a VPLS.

VMAN is the *VLAN* stacking feature in EXOS. It has two types of ports: access and network. The access port can be both aware or unaware of VLAN. The customer edge port is the VLAN aware port of VMAN. The CEP allows several configurable options:

- An Ethernet port can be associated with multiple VMANs based on the CVIDs.
- Multiple CVIDs on multiple Ethernet ports can be associated with a VMAN.
- A range of CVIDs can be specified instead of individually configured.
- CVID translations and egress filtering.

VMAN with CEP ports can work as a way to achieve some of the requirements. However, you cannot currently assign it to a VPLS.

When implemented, a VMAN with CNP, or CEP, or both can be assigned as a service to a VPLS. This feature does not add any other capability beyond associating a VMAN with CEP port to a VPLS. Existing VMAN features such as CVID translation and egress filtering do not change. The existing VPLS feature to include or exclude SVLAN also does not change.

### *Supported Platforms*

VLAN mapping to VPLS is supported on the following platforms:

- Summit X460-G2 (supported from ExtremeXOS 15.6)
- Summit X670-G2 (supported from ExtremeXOS 15.6)
- Summit X670
- Summit X770
- BlackDiamond X8

### *Limitations*

The Selective VLAN Mapping to VPLS feature has the following limitations:

- You cannot assign multiple VMANs to a VPLS.
- You cannot use *SNMP* and XML to assign a VMAN with CEP port to VPLS.

*Configuring Selective VLAN Mapping to VPLS*

The CEP port is configured as normal. The existing command to associate VMAN to VPLS also does not change. The following example illustrates the three-step process:

1. Create VMAN with CEP ports.
2. Create VPLS.
3. Associate the VMAN to the VPLS.

```
# create vman vm1
configure vman vm1 add ports 3 cep cvid 2 – 3

create vpls vsi1 fec-id-type pseudo-wire 35
configure vpls vsi1 add peer 192.168.0.2 core

configure vpls vsi1 add service vman vm1
```

Additionally, the display does not change. To show the mapping, you need to configure the following two-step process:

1. Show the VPLS.
2. Show the VMAN.

Here is an example:

```
# show vpls vsi1
L2VPN Name    VPN ID Flags   Services Name  Peer IP         State Flags
------------  ------ ------- -------------- --------------- ----- ------
VSI1          35     EAX--L- vm1            192.168.0.2     Up    C---V-

VPN Flags: (E) Admin Enabled, (A) Oper Enabled, (I) Include Tag,
           (X) Exclude Tag, (T) Ethertype  Configured,
           (V) VCCV HC Enabled, (W) VPN Type VPWS, (L) VPN Type
           VPLS, (M) CFM MIP Configured
Peer Flags:(C) Core Peer, (S) Spoke Peer, (A) Active Core,
           (p) Configured Primary Core, (s) Configured Secondary Core,
           (N) Named LSP Configured, (V) VCCV HC Capabilities Negotiated,
           (F) VCCV HC Failed

# show vpls vsi1 detail
L2VPN Name: VSI1
    VPN ID                 : 35              Admin State    : Enabled
    Source Address         : 192.168.0.1     Oper State     : Enabled
    VCCV Status            : Disabled        MTU            : 1500
    VCCV Interval Time     : 5 sec.          Ethertype      : 0x88a8
    VCCV Fault Multiplier  : 4               .1q tag        : exclude
    L2VPN Type             : VPLS            Redundancy     : None
    Service Interface      : vm1

    Peer IP: 192.168.0.2
        PW State          : Up
        PW Uptime         : 0d:0h:4m:20s
        PW Installed      : True
        Local PW Status   : No Faults
        Remote PW Status  : No Faults
        Remote I/F MTU    : 1500
        PW Mode           : Core-to-Core
        Transport LSP     : LDP LSP (Not Configured)
        Next Hop I/F      : vlan2
        Next Hop Addr     : 11.0.2.2       Tx Label      : 0x00173
        PW Rx Label       : 0x00174        PW Tx Label   : 0x00174
        PW Rx Pkts        : 185920064      PW Tx Pkts    : 186031288
        PW Rx Bytes       : 16732807110    PW Tx Bytes   : 16742817810
```

```
        MAC Limit         : No Limit
        VCCV HC Status    : Not Sending (VCCV Not Enabled For This L2VPN)
            CC Type       : Rtr Alert                Total Pkts Sent : 0
            CV Type       : LSP Ping                 Total Pkts Rcvd : 0
            Send Next Pkt : --
            Total Failures: 0               Pkts During Last Failure : 0
            Last Failure Tm: --
1.26 # show vman vm1
VMAN Interface with name vm1 created by user
    Admin State:          Enabled    Tagging:Untagged (Internal tag 4091)
    Description:          None
    Virtual router:       VR-Default
    IPv4 Forwarding:      Disabled
    IPv4 MC Forwarding:   Disabled
    IPv6 Forwarding:      Disabled
    IPv6 MC Forwarding:   Disabled
    IPv6:                 None
    STPD:                 None
    Protocol:             Match all unfiltered protocols
    Loopback:             Disabled
    NetLogin:             Disabled
    OpenFlow:             Disabled
    QosProfile:           None configured
    Egress Rate Limit Designated Port: None configured
    Flood Rate Limit QosProfile:       None configured
    Ports:   1.            (Number of active ports=1)
        CEP:         *3: CVID 2-3
         Flags:     (*) Active, (!) Disabled, (g) Load Sharing port
          (b) Port blocked on the vlan, (m) Mac-Based port
          (a) Egress traffic allowed for NetLogin
          (u) Egress traffic unallowed for NetLogin
          (t) Translate VLAN tag for Private-VLAN
          (s) Private-VLAN System Port, (L) Loopback port
          (e) Private-VLAN End Point Port
          (x) VMAN Tag Translated port
          (G) Multi-switch LAG Group port
          (H) Dynamically added by MVRP
          (U) Dynamically added uplink port
          (V) Dynamically added by VM Tracking
```

## RSVP-TE Overview

RSVP is a protocol that defines procedures for signaling *QoS* requirements and reserving the necessary resources for a router to provide a requested service to all nodes along a data path.

RSVP is not a routing protocol. It works in conjunction with unicast and multicast routing protocols. An RSVP process consults a local routing database to obtain routing information. Routing protocols determine where packets get forwarded; RSVP is concerned with the QoS of those packets that are forwarded in accordance with the routing protocol.

Reservation requests for a flow follow the same path through the network as the data comprising the flow. RSVP reservations are unidirectional in nature, and the source initiates the reservation procedure by transmitting a path message containing a traffic specification (Tspec) object. The Tspec describes the source traffic characteristics in terms of peak data rate, average data rate, burst size, and minimum/maximum packet sizes.

RSVP-TE is a set of traffic engineering extensions to RSVP. RSVP-TE extensions enable RSVP use for traffic engineering in *MPLS* environments. The primary extensions add support for assigning MPLS

labels and specifying explicit paths as a sequence of loose and strict routes. These extensions are supported by including label request and explicit route objects in the path message. A destination responds to a label request by including a label object in its reserve message. Labels are then subsequently assigned at each node the reserve message traverses. Thus, RSVP-TE operates in downstream-on-demand label advertisement mode with ordered LSP control.

The ExtremeXOS software implementation of RSVP-TE complies with RFC 3209 and includes support for:

- Configuration on a per *VLAN* interface.
- Operation as either edge or core MPLS router.
- Support for specifying explicitly routed paths.
- Support for both loose and strict route objects.
- Recording the route of an established path.
- Bandwidth reservation and policy per LSP.
- Signaling QoS along the RSVP path using the Tspec and Adspec objects.
- Fixed Filter (FF) and Shared Explicit (SE) reservation styles.
- Specifying RSVP-TE session attributes.
- Scaling enhancements using Refresh Overhead Reduction extensions.
- Improved link failure detection using the RSVP-TE Hello Message.
- Ability to reroute traffic over pre-configured backup LSPs.

## *RSVP Elements*

### Message Types

RSVP messages are passed between RSVP-capable routers to establish, remove, and confirm resource reservations along specified paths.

RSVP messages are sent as raw IP datagrams with protocol number 46. Each LSR along the path must process RSVP control messages so that it can maintain RSVP session state information. Therefore, most RSVP messages are transmitted with the IP Router Alert Option. Including the IP Router Alert provides a convenient mechanism allowing the IP routing hardware to intercept IP packets destined to a different IP address and deliver them to the RSVP control plane for processing. This is needed to set up and refresh RSVP-TE LSPs that follow an explicitly specified network path and thus may not use the normal routed next hop IP address. RSVP has two basic message types, path message and reserve message, as shown in the following figure.

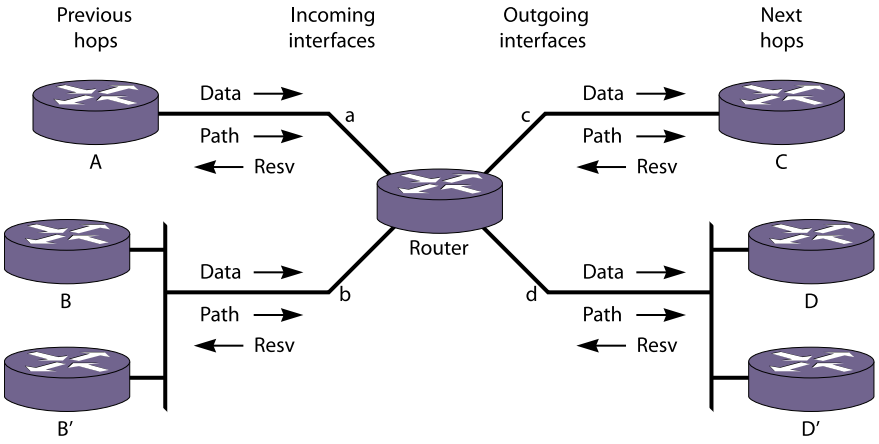



**Figure 194: RSVP Messages**

RSVP has the following message types:

- Path message
- Reserve message
- Path tear message
- Reserve tear message
- Path error message
- Reserve error message
- Reserve confirm message

Path Message: The RSVP path message is used to store state information about each node in the path. Each RSVP sender transmits path messages downstream along routed paths to set up and maintain RSVP sessions. Path messages follow the exact same path as the data flow, creating path states in each LSR along the path. The IP source address of the path message must be an address of the sender it describes and the IP destination address must be the endpoint address for the session. The path message is transmitted with the IP Router Alert option since each router along the path must process the path message. Each LSR is responsible for refreshing its path status by periodically transmitting a path message to the downstream LSR.

In addition to the previous hop address, the path message contains the sender Tspec and Adspec. The reservation message carries the flowspec.

Reserve Message: Each receiver host transmits an RSVP reservation request to its upstream neighbor. Reserve messages carry reservation requests hop-by-hop along the reverse path. The IP destination address of a reserve message is the unicast address of the previous-hop LSR, obtained from the session's path state. The IP source address is the address of the node that originated the message. The reserve message creates and maintains a reserve state in each node on the path. Each LSR is responsible for refreshing its reserve status by periodically transmitting a reserve message to the upstream LSR.

Reserve messages are eventually delivered to the sender, so that the sender can configure appropriate traffic control parameters for the first hop node.

[17] The routed path may be the best routed path or an explicitly specified routed path using EROs.
[18] IP Router Alert option is described in RFC 2113.

Path Tear Message: Path tear messages delete path state information reserved along the path. The message is initiated by the path sender or by any LSR in which a path state time-out occurs or an LSP is preempted (due to bandwidth reservations), and is sent downstream to the session's path endpoint. Path tear messages are transmitted with the IP Router Alert option and are routed exactly the same as path messages. The IP destination address must be the path endpoint and the source IP address must be the sender address obtained from the session's path state for the path that is being torn down.

When a path state is deleted as the result of the path tear message, the related reservation state must also be adjusted to maintain consistency in the node. The adjustment depends on the reservation style.

Reserve Tear Message: Reserve tear messages delete reservation state information. The message is initiated by the path endpoint or any node along the path in which a reservation state has timed out or an LSP is preempted (due to bandwidth reservations), and is sent upstream to the session's path sender. Reserve tear messages are routed exactly the same as reserve messages. The IP destination address of a reserve message is the unicast address of the previous-hop node, obtained from the session's reservation state. The IP source address is the address of the node that originated the message.

If no reservation state matches the reserve tear message, the message is discarded. The reserve tear message can delete any subset of the filter specification in FF-style or SE-style reservation state. Reservation styles are described in the following table.

Path Error Message: Path error messages are used to report processing errors for path messages. These messages are sent upstream to the sender that issued the path message. The message is routed hop-by-hop using the path state information maintained in each node. Path error messages are informational and do not modify the path state within any node.

Reserve Error Message: Reserve error messages are used to report processing errors for reserve messages. In addition, reserve error messages are used to report the spontaneous disruption of a reservation. Reserve error messages travel downstream to the endpoint of the session. The message is forwarded hop-by-hop using the reservation state information maintained in each node. Reserve error messages are informational and do not modify the reservation state within any node.

Reserve Confirm Message: Reserve confirm messages are optionally transmitted to acknowledge a reservation request. These messages are transmitted from the sender to the endpoint. The destination IP address is the IP address of the endpoint and the source IP address is the address of the sender. Since none of the intermediate path nodes need to process a reserve confirm message, the message is transmitted without the IP Router Alert option.

**Reservation Styles**

A reservation style is a set of options that is included in the reservation request.

One reservation style concerns how reservations requested by different senders within the same session are handled. This type of reservation style is handled in one of two ways: either create a distinct reservation for each sender in the session, or use a single reservation that is shared among all packets of the selected senders.

Another reservation style concerns how senders are selected. Again, there are two choices: an explicit list of all selected senders, or a wildcard that implies all senders in the session.

The following table describes the relationship between reservation attributes and styles.

**Table 127: Reservation Attributes and Styles**

| Sender Selection | Distinct Reservation Style | Shared Reservation Style |
|---|---|---|
| Explicit | Fixed filter (FF) | Shared explicit (SE) |
| Wildcard | Not defined | Wildcard filter (WF) |

Fixed Filter: The fixed filter (FF) reservation style uses a distinct reservation and an explicit sender selection. This means that each resource reservation is for a specific sender. The session resources are not shared with other senders' packets. Because each reservation is identified with a single sender, a unique label is assigned by the endpoint to each sender (i.e., point-to-point LSP reservation).

Shared Explicit: The shared explicit (SE) reservation style uses a shared reservation and an explicit sender selection. This means that a single resource reservation is created that is shared by multiple senders. The endpoint may specify which senders are to be included for the reservation. Because different senders are explicitly listed in the RESV message, different labels may be assigned to each sender. Thus, multiple shared-resource LSPs to the same endpoint can be created (i.e., multipoint-to-point LSP reservation). The Extreme *MPLS* implementation requests SE reservation style when signaling RSVP-TE LSPs.

Wildcard: The wildcard (WF) reservation style uses the shared reservation and wildcard sender options. A wildcard reservation creates a single reservation that is shared by data flows from all upstream senders.

The Extreme MPLS implementation does not support WF reservation style.

*RSVP Traffic Engineering*

*MPLS* Traffic Engineering (TE) extends RSVP to support several unique capabilities.

By coupling RSVP and MPLS, LSPs can be signaled along explicit paths with specific resource reservations. Additional RSVP objects have been defined to provide TE extensions. These objects include the Label Request, Label, Explicit Route, Record Route, and Session Attribute. Extreme's RSVP-TE implementation supports all of these TE objects.

**RSVP Tunneling**

An RSVP tunnel sends traffic from an ingress node through an LSP. The traffic that flows through the LSP is opaque (or tunneled) to the intermediate nodes along the path. Traffic flowing through the tunnel to an intermediate node along the path is identified by the previous hop and is forwarded, based on the label value(s), to the downstream node.

RSVP-TE can:

- Establish tunnels with or without *QoS* requirements.
- Dynamically reroute an established tunnel.
- Observe the actual route traversed by a tunnel.
- Identify and diagnose tunnels.
- Use administrative policy control to preempt an established tunnel.
- Perform downstream-on-demand label allocation, distribution, and binding.

Some LSRs require their neighboring LSRs to include their Router ID in the Extended Tunnel ID field when sending RSVP-TE messages. The Extended Tunnel ID is a globally unique identifier present in the RSVP common header Session object (see RFC 3209). To provide maximum compatibility with other vendors' implementations, the ExtremeXOS *MPLS* implementation accepts RSVP-TE messages regardless of the Extended Tunnel ID value and always inserts the local Router ID into the Extended Tunnel ID field prior to transmission of an RSVP-TE message.

### RSVP Objects

This section describes the RSVP objects that are used to establish RSVP-TE LSPs:

- Label
- Label request
- Explicit
- Record route
- Session attribute

Label: The label object is carried in the reserve message and is used to communicate a next hop label for the requested tunnel endpoint IP address upstream towards the sender.

Label Request: To create an RSVP-TE LSP, the sender on the *MPLS* path creates an RSVP path message and inserts the label request object into the path message.

A label request object specifies that a label binding for the tunneled path is requested. It also provides information about the network layer protocol that is carried by the tunnel. The network layer protocol sent through a tunnel is not assumed to be IP and cannot be deduced from the Layer 2 protocol header, which simply identifies the higher layer protocol as MPLS. Therefore, the Layer 3 Protocol ID (PID) value must be set in the Label Request Object, so that the egress node can properly handle the tunneled data.

Note

The ExtremeXOS RSVP-TE implementation supports only Label Request objects with no Label Range. Label Ranges are used to signal ATM VPI/VCI or Frame Relay DLCI information for the LSP. These types of Label Requests are not supported. In the ExtremeXOS RSVP-TE implementation, the L3 PID value, which identifies the Layer 3 protocol of the encapsulated traffic, is always set to 0x0800 (IP).

Explicit Route: The explicit route object specifies the route of the traffic as a sequence of nodes. Nodes may be loosely or strictly specified.

The explicit route object is used by the MPLS sender if the sender knows about a route that:

- Has a high likelihood of meeting the *QoS* requirements of the tunnel.
- Uses the network resources efficiently.
- Satisfies policy criteria.

If any of the above criteria are met, the sender can decide to use the explicit route for some or all of its sessions. To do this, the sender node adds an explicit route object to the path message.

After the session has been established, the sender node can dynamically reroute the session (if, for example, if discovers a better route) by changing the explicit route object.

Record Route: The record route object is used by the sender to receive information about the actual route traversed by the RSVP-TE LSP. It is also used by the sender to request notification if there are changes to the routing path. Intermediate or transit nodes can optionally use the RRO to provide loop detection.

To use the object, the sender adds the record route object to the path message.

Session Attribute: The session attribute object can also be added to the path message. It is used for identifying and diagnosing the session. The session attribute includes the following information:

- Setup and hold priorities
- Resource affinities
- Local protection

**ERO Exclude Option**

In order to allow more flexibility when traffic engineering redundant paths for RSVP-TE LSPs, this feature adds the “exclude” option to the “configure mpls rsvp-te path <path> add ero” command. An “include” option is also available, but is optional. If neither option is used, “include” will be used for backward compatibility.

A hop that is excluded is avoided when CSPF does path calculations. By adding an “exclude” hop to a path, LSPs can be set up to avoid certain links. An example of this is to configure a secondary RSVP-TE LSP to avoid the hops that a primary LSP has been configured to traverse. This can allow the secondary LSP more freedom to route through other parts of the network. Often, without the “exclude” option, the secondary LSP must be configured more tightly than desired to ensure that its path never overlaps with the primary LSP’s path.

*Establishing RSVP-TE LSPs*

Establishing LSPs requires every LSR along the path to support RSVP and the TE extensions defined in RFC 3209.

The LSP endpoints attempt to detect non-RSVP capable LSRs by comparing the time-to-live (TTL) value maintained in the RSVP common header with that of the IP TTL. If these values are different, it is assumed that a non-RSVP capable LSR exists along the path. By including the Label Request object in the path message, RSVP capable routers that do not support the TE extensions can be detected. RSVP routers that do not support TE extensions reply with the Unknown object class error.

RSVP-TE LSPs are referred to as named LSPs. These LSPs have configurable names that are used to identify the LSP within the CLI. The command `create mpls rsvp-te lsp` *`lsp_name`* `destination` *`ipaddress`* allocates the internal resources for the LSP. The newly created LSP is not signaled until the LSP has been configured. The LSP can be configured to take a specific path through the network or the administrator can let the switch choose the best path by specifying the path any. Up to three paths may be configured for an LSP to provide redundancy. The command `configure mpls rsvp-te lsp` *`lsp_name`* `add path` configures an LSP. Optionally, RSVP-TE profiles may be applied to an LSP to change its properties. An RSVP-TE profile is a specific CLI container used to hold configuration parameters associated with timers, bandwidth reservation, limits, and other miscellaneous properties.

Once the RSVP-TE LSP is configured, the LSP is immediately signaled. If signaled successfully, the LSP becomes active. The commands `disable mpls rsvp-te lsp` *`lsp_name`* and `enable mpls`

`rsvp-te lsp` *`lsp_name`* are used to tear down and re-signal the LSP. Disabling the LSP causes the LER to send a path tear message to the destination, forcing the LSP down and all resources along the path to be freed. Enabling the LSP instructs the LER to send a path message to the destination re-establishing the LSP. The configuration of the LSP is not modified by the enable or disable LSP commands.

*RSVP-TE Implementation*

**Explicit Route Path LSPs**

An explicit route is a specified path through a routed network topology.

The path can be strictly or loosely specified. If strictly specified, each node or group of nodes along the path must be configured. Thus, no deviation from the specified path is allowed.

Loosely specified paths allow for local flexibility in fulfilling the requested path to the destination. This feature allows for significant leeway by the LSR in choosing the next hop when incomplete information about the details of the path is generated by the LER. Each node along the path may use other metrics to pick the next hop along the path, such as bandwidth available, class of service, or link cost. The command configure mpls rsvp-te path *`path_name`* add ero is used to add an Explicit Route Object to a path container.

An explicit routed path is encoded using the explicit route object (ERO) and is transmitted in the path message. The ERO consists of a list of subobjects, each of which describes an abstract node. By definition, an abstract node can be an IP prefix or an autonomous system (AS) number. The ExtremeXOS RSVP-TE implementation supports only IPv4 abstract nodes. The ExtremeXOS RSVP-TE implementation supports both strict and loose IPv4 abstract nodes. Received path messages with EROs that contain any other subobject type result in the transmittal of an Unknown object class error message. All LSRs along the specified path must support the inclusion of the ERO in the path message for an explicitly routed path to be successfully set up.

An LSR receiving a path message containing an ERO must determine the next hop for this path.

The steps for selection of the next hop are as follows:

1. The receiving LSR evaluates the first subobject. If the subobject type is not supported or there is no subobject, a Bad ERO error is returned. The abstract node is evaluated to ensure that this LSR was the valid next hop for the path message. If the subobject is a strict abstract node, the abstract node definition must match the local interface address. If it does, then this LSR is considered to be a member of the abstract node. Additionally, if the /32 address matches a local interface address, the path message must have been received on the direct interface corresponding to the /32 address. If the abstract node is an IP prefix, the subnet configured for the interface from which the path message was received must match the abstract node definition. In the event that this LSR is not part of the strict abstract node definition, a Bad initial subobject error is returned. If the subobject is a loose abstract node, the LSR determines if the abstract node definition corresponds to this LSR. If it doesn't, the path message is transmitted along the best-routed or constrained optimized path to the endpoint and the ERO is not modified. If it is, then processing of the ERO continues.
2. If there is no second subobject, the ERO is removed from the path message. If this LSR is not the end of the path, the next hop is determined by the constrained optimized path (through Constrained Shortest Path First—CSPF) to the path message endpoint.

3. If there is a second subobject, a check is made to determine if this LSR is a member of the abstract node. If it is, the first subobject is deleted and the second subobject becomes the first subobject. This process is repeated until either there is only one subobject or this LSR is not a member of the abstract node as defined by the second subobject. Processing of the ERO is then repeated with step 2. By repeating steps 2 and 3, any redundant subobjects that are part of this LSRs abstract node can be removed from the ERO. If this operation were not performed, the next hop LSR might reject the path message.
4. The LSR uses its CSPF to determine the next hop to the second subobject. If the first object is a /32 address, the first subobject is removed, since it would not be part of the next hop's abstract node. The path message is then sent along the explicit path to the path message endpoint. No determination is made to verify that the abstract node defined in the subobject is topologically adjacent to this LSR. The next hop should verify this as part of its processing as defined in step 1.

If CSPF determines that a specific path needs to be taken through the network, additional EROs are inserted into the path message.

**Route Recording**

The route a path takes can be recorded.

Recording the path allows the ingress LER to know, on a hop-by-hop basis, which LSRs the path traverses. Knowing the actual path of an LSP can be especially useful for diagnosing various network issues.

Network path recording is configurable per LSP. This feature is configured by enabling route recording for a specific RSVP-TE profile using the command `configure mpls rsvp-te lsp profile` *`lsp_profile_name`* `record enabled` and associating the profile to an LSP. The ExtremeXOS software sets the label recording desired flag in the path message if route recording has been enabled for the LSP.

If route recording is enabled, the record route object (RRO) is inserted into the path message using a single RRO subobject, representing the ingress LER. When a path message that contains an RRO is received by an Extreme LSR, an RRO IPv4 subobject representing the /32 address of the outgoing interface of the path message is pushed onto the top of the first RRO. The updated RRO is returned in the reserve message.

The label recording flag is supported by the ExtremeXOS software and is set automatically when route recording is enabled. The **`route-only`** option can be used when enabling route recording in the profile to prevent the label recording flag from being set. If an Extreme LSR receives a path message with the label recording flag set in the RRO, the LSR encodes the LSP label into a label subobject and pushes it onto the RRO.

If a path message is received that contains an RRO, the Extreme LSR uses the RRO to perform loop detection. The RRO is scanned to verify that the path message has not already traversed this LSR. If the RRO contains an IPv4 subobject that represents a local LSR interface, the path message is dropped and a Routing Problem error message is sent to the originating LER with an error value of Loop detected.

[19] RRO is organized as a LIFO stack.

### LSP Session Attributes

Session attributes are signaled for configured RSVP-TE LSPs using the session attribute object without resource affinities (that is, LSP_TUNNEL Type).

The ExtremeXOS software uses the setup and hold priority values to preempt established LSPs in order to satisfy bandwidth requests. Lower hold priority LSPs are preempted in order to satisfy the bandwidth request in a path message with a higher setup priority. LSP attributes are configured by setting the priorities for a specific RSVP-TE profile using the command `configure mpls rsvp-te lsp profile` *`lsp_profile_name`* `setup-priority` *`priority`* `hold-priority` *`priority`* and associating the profile to the configured LSP.

### Bandwidth Reservation

As mentioned previously, RSVP reservations are unidirectional in nature.

The source initiates the reservation procedure by transmitting a path message containing a sender Tspec object. The Tspec describes the source traffic characteristics in terms of peak data rate, average data rate, burst size, and minimum/maximum packet sizes. The path message can also contain an optional AdSpec object that is updated by network elements along the path to indicate information such as the availability of particular *QoS* services, the maximum bandwidth available along the path, the minimum path latency, and the path maximum transmission unit (MTU).

The ExtremeXOS software supports LSR bandwidth reservation requests per LSP. Only the Int-Serv Controlled-Load service request is supported. Bandwidth is always reserved on the physical ports that the LSP traverses. Depending on the platform, the bandwidth reservation may also be policed. The network administrator can verify that the requested bandwidth was actually reserved. In those cases when the bandwidth reserved is less than the requested bandwidth, the LSP can be manually torn down, re-signaled using a different path, or accepted. The LSR automatically attempts to find a path that best satisfies the bandwidth request. Constrained path selections are supported using *OSPF*-TE. Best effort LSPs are provisioned by specifying a reserved bandwidth as best-effort. The reserved LSP bandwidth is configured by setting the bps rate for a specific RSVP-TE profile, using the `configure mpls rsvp-te lsp profile` *`lsp_profile_name`* `bandwidth` command and associating the profile to an LSP.

Accounting of bandwidth reserved through an Extreme LSR RSVP-TE enabled *VLAN* is supported. The maximum available bandwidth per physical port or trunk group is enforced. Thus, the available bandwidth specified in the Adspec object is not modified as the path message is forwarded to the LSP endpoint. As reserve messages are processed, the reserved bandwidth specified in the Flowspec is added to the total reserved bandwidth allocated for the physical ports.

Because LSP bandwidth is dynamically allocated, a configuration command is provided to reserve port bandwidth for use by *MPLS*. The command `configure mpls rsvp-te bandwidth committed-rate` pre-reserves bandwidth from the specified MPLS enabled VLAN for RSVP-TE traffic only. This pre-allocation of bandwidth is useful since other applications may compete with MPLS for available bandwidth. By pre-reserving a portion of the MPLS interface's bandwidth capacity, MPLS is guaranteed to have that amount of the MPLS interface's bandwidth to meet RSVP-TE LSP reservation requests.

CIR bandwidth for the receive direction is not tracked by TE IGPs, such as OSPF-TE, and configuring it is not required. Configuring CIR bandwidth for the receive direction does not prevent an LSP from going operational due to lack of receive bandwidth; however, it can be useful for tracking and informational

purposes. An Info level log (MPLS.RSVPTE.IfRxBwdthExcd) is generated if the setup of a TE LSP requires receive bandwidth greater than that which is currently available for the receive direction on a particular interface. This generally happens only when TE LSPs with different previous hops ingress the switch on the same interface (for example, from a multi-access link) and egress the switch on different interfaces.

### Bandwidth Management for RSVP-TE LSPs

If an RSVP-TE LSP is signaled through a switch with bandwidth parameters, the LSP bandwidth request is granted or rejected based on the availability of bandwidth resources on the physical ports that the LSP traverses.

Data traffic through these switches is not policed and there are no guarantees that the packets using the LSP are not dropped.

Note
Per LSP rate limiting is not supported in this release.

The available bandwidth for each *OSPF* interface is continually updated within the OSPF area. As RSVP-TE LSPs are established and torn down, the reserved bandwidth associated with these LSPs is used to update the total bandwidth available through each OSPF interface. RSVP-TE and CSPF can use the bandwidth information to determine the appropriate path that each LSP should take through the network based on the LSP's profile parameters. LSP parameters that can affect the CSPF TE path calculation include the LSP setup priority and bandwidth configuration.

Available bandwidth is calculated for eight *CoS (Class of Service)* levels. Each CoS uniquely maps to an LSP hold priority. Thus, when an LSP is set up through the switch, the reserved bandwidth consumed is associated with a CoS based on the signaled LSP hold priority. The available bandwidth is recalculated and is advertised to its OSPF neighbors. Advertised bandwidth is calculated using graduated bandwidth reporting methodology. Using this scheme, higher CoS levels advertise available bandwidth that includes allocated bandwidth for lower CoS levels. The reasoning for doing this is that higher priority LSPs can preempt lower priority LSP. Thus, even though the bandwidth has been allocated to a lower priority LSP, it is still available for use by higher priority LSPs.

In the following example, an interface is configured to reserve 250 Mbps for *MPLS* traffic.

The following LSPs are established through this interface. Remember, hold priority value of 0 is the highest priority and 7 is the lowest.

- LSP A, hold priority = 7, reserved = 50 Mbps
- LSP B, hold priority = 5, reserved = 100 Mbps
- LSP C, hold priority = 2, reserved = 25 Mbps
- LSP D, hold priority = 1, reserved = 25 Mbps

OSPF advertises the following available bandwidth for each CoS.

CoS 0 is the highest and CoS 7 is the lowest:

- CoS 0 (hold = 0): 250 Mbps (No LSPs; all bandwidth available)
- CoS 1 (hold = 1): 225 Mbps (LSP D)
- CoS 2 (hold = 2): 200 Mbps (LSP C & D)
- CoS 3 (hold = 3): 200 Mbps (LSP C & D)

- CoS 4 (hold = 4): 200 Mbps (LSP C & D)
- CoS 5 (hold = 5): 100 Mbps (LSP B, C & D)
- CoS 6 (hold = 6): 100 Mbps (LSP B, C & D)
- CoS 7 (hold = 7): 50 Mbps (LSP A, B, C & D)

CSPF calculations only use the available bandwidth for the desired CoS, as specified by the LSP hold priority. Thus in this example, if LSP E, with a configured setup priority of 6, requires 150 Mbps, CSPF calculates a path to the destination that does not go through the above interface, since only 100 Mbps worth of bandwidth is available.

**Redundant LSPs**

There are three methods for provisioning redundant RSVP-TE LSPs at the ingress LER, also referred to as head-end LSP protection:

- Configured secondary (or backup) LSPs
- Fast reroute (detour) LSPs
- Multipath LSPs

Secondary RSVP-TE LSPs can be configured to provide backup LSPs in the event that the primary LSP fails. You can create up to two secondary LSPs for each primary LSP. The secondary LSPs are fully provisioned, pre-established RSVP-TE LSPs that are maintained as inactive until needed. If the primary LSP is torn down, the associated LSP next hop is removed from the route table, and a new LSP next hop representing one of the secondary LSPs is installed as the preferred LSP. If there are multiple secondary LSPs available, the secondary LSP is randomly selected. If the primary LSP is re-established, the primary LSP next hop information is re-installed and the secondary LSP returns to inactive state.

If both the primary and secondary paths for an LSP fail, and there are no other RSVP-TE LSPs active to the destination, an LDP LSP can be used if available.

Operation with L2 VPNs is similar. If a primary path fails, and a secondary LSP is available, VPLS uses the secondary LSP. When the primary LSP is re-established, VPLS again uses the primary LSP.

Specifying redundant LSPs is accomplished by assigning secondary paths to an LSP. The `configure mpls rsvp-te lsp` *`lsp_name`* `add path` *`path_name`* `secondary` command can configure the specified path as a backup LSP. A path different from the primary path must be specified. It is recommended that defined paths be configured using EROs to specify different paths through the network. Relying on the routing topology, by configuring the path to any, can create two LSPs that take the same path. It is important to understand that the configured LSP signals multiple LSPs, up to three (one primary and two secondary), but only one LSP can be used to forward traffic at any one time.

Fast Reroute LSPs are based on the on IETF RFC 4090, Fast Reroute Extensions to RSVP-TE for LSP Tunnels, which defines RSVP-TE extensions to establish backup LSP tunnels for local repair of LSP tunnels. To respond to failures, these mechanisms enable the re-direction of traffic onto backup LSP tunnels in tens of milliseconds, and this meets the needs of real-time applications such as voice over IP (VoIP). This timing requirement is satisfied by computing and signaling backup LSP tunnels in advance of a failure and by re-directing traffic as close to the failure point as possible. In this way the time for redirection includes no path computation and no signaling delays, which include delays to propagate failure notification between label-switched routers (LSRs). Speed of repair is the primary advantage of using fast-reroute backup methods.

There are two backup methods; the detour LSP method (which is also called the one-to-one backup method) and the facility backup method (which is also called the by-pass tunnel method). The software supports only the detour LSP method.

Based on the RFC-4090 there are two different methods to uniquely identify a backup path:

1. Path-specific method
2. Sender template-specific method

The software supports only the path-specific method, which uses a new object, the DETOUR object, to distinguish between PATH messages for a backup path and the protected LSP.

The following figure illustrates the terminology used to describe fast-reroute configuration and operation.

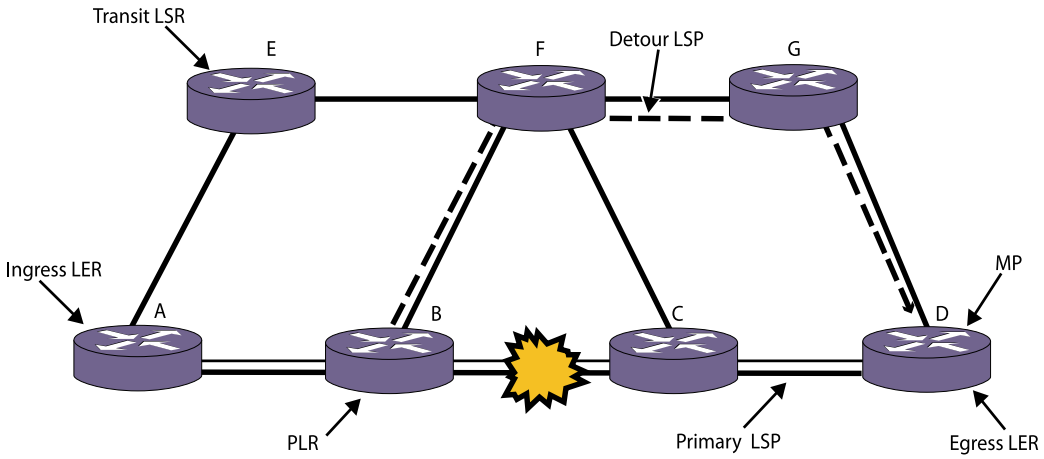



**Figure 195: Fast-Reroute Terminology**

The primary LSP in the following figure is established between *MPLS* routers A and D. Router A is the ingress LER, and Router D is the egress LER. When used with fast-reroute protection, the primary LSP is also called the protected LSP, as it is protected by the detour LSP created by the fast-reroute feature. The detour LSP provides a route around a protected component. In the following figure, the link between Router B and Router C has failed. The detour LSP, which is indicated by the dashed line, provides a path around the failure.

Routers B and C are transit LSRs for the primary LSP. With respect to a specific LSP, any router that is not the ingress or egress LER is a transit LSR. Routers F and G are transit LSRs for the detour LSP.

The origin of the detour LSP is called the Point of Local Repair (PLR), and the termination of the detour LSP is called the Merge Point. A protected LSP is an explicitly-routed LSP that is provided with protection. A detour LSP is also an explicitly-routed LSP. If you configure a series of one or more hops (EROs), then based on the currently set DYNAMIC_FULL option in the Constrained-based Shorted Path First (CSPF) routing component, the CSPF will calculate and try to fill in the gaps to build a complete list of EROs.

You can configure up to two secondary LSPs for each standard TE (non-FRR) LSP or for each protected FRR LSP. If a standard TE LSP fails, then one of the secondary LSPs becomes active. If that secondary LSP fails, the other secondary LSP becomes active. If a protected FRR LSP fails, its detour LSP becomes active. If the detour LSP fails, then one of the secondary LSPs becomes active, and if that secondary LSP fails, the other secondary LSP becomes active. If all configured backup and secondary paths for an

LSP fail, a different active RSVP-TE LSP to the destination can be used. Otherwise, an LDP LSP can be used if available.

The primary advantage of detour LSPs is the repair speed. The cost of detour LSPs is resources. Each backup LSP reserves resources that cannot be used by other LSPs. Another cost is that currently there is no automatic way to redirect traffic from a detour LSP back to a primary LSP when the protected LSP recovers. Redirecting traffic from the detour LSP to the primary LSP requires a series of CLI commands.

Fast reroute protection is configured primarily on the ingress LER, however, it must be enabled on all transit LSRs and the egress LER also. After configuration is complete and fast-reroute protection is enabled on the primary LSP, the primary and detour LSPs are signalled. Provided that the resources are available, detour LSPs are set up at each transit LSP along the primary LSP.

Multiple RSVP-TE LSPs can exist or be configured to the same destination. The paths do not need to be equal cost; all that is required is that all the LSPs to the same destination must have IP transport enabled. In this scenario, LSP next hop information is communicated to the route table for up to eight different named RSVP-TE LSPs. Locally originated traffic is distributed across each LSP based on standard IP address hash algorithms. If one of the LSPs fails, the traffic is redistributed across the remaining active named LSPs. Unlike the backup LSP mechanism, all of the redundant multipath LSPs are unique named LSPs and in general have primary configured paths.

#### Improving LSP Scaling

RSVP maintains path and reserve state by periodically sending refresh messages.

Refresh messages allow each LSR along the path to properly maintain reservation state information and to recover from network failures. Because refresh messages are periodically sent for each path reservation, scaling the number of RSVP-TE LSPs is an issue. Additionally, network requirements for faster failure detection and improved LSP recovery times further exacerbate the scaling issue.

Several techniques are described in RFC 2961 RSVP Refresh Overhead Reduction to improve the scalability of RSVP. These techniques include the bundle message, message ID extension, and summary refresh extension. Support for these extensions is signaled between RSVP peers via the refresh-reduction-capable bit in the flags field of the common RSVP header. Additionally, the hello extension, described in RFC 3209, provides a fourth scaling mechanism for RSVP. The hello extension is designed so that either peer can use the mechanism regardless of how the other peer is configured. Therefore, support for the hello extension is not signaled between RSVP peers. The ExtremeXOS software supports and is compliant with the RSVP-TE scaling features described in RFC 2961.

These features include the following:

- Bundle message
- Summary refresh extension
- Message ID extension
- Hello extension

Bundle Message: RSVP bundle messages aggregate multiple RSVP messages within a single PDU. The messages are addressed directly to peer LSRs. Therefore, bundle messages are not sent with the IP Router Alert option. Bundling multiple RSVP messages into a single PDU reduces the per packet overhead associated with local delivery and local origination. Each bundle message must contain at least one RSVP message. Transmission of RSVP messages may be delayed up to the number of seconds

configured for bundle time. The size of the bundle message is limited to the RSVP-TE interface MTU size. Bundle messaging is enabled using the enable mpls rsvp-te bundle-message command.

Summary Refresh Extension: A summary refresh message is used to refresh RSVP states along an LSP without having to explicitly send path and reserve refresh messages. This can substantially reduce the RSVP control bandwidth overhead. Summary refresh messages contain a list of message_ID objects. Each message_ID object identifies a path and reserve state to be refreshed. When summary refresh support is enabled, path and reserve refresh messages are suppressed. If the message identifier value indicates that the RSVP state has changed, the receiving LSR notifies the sender by transmitting a message_ID_NACK message. The summary refresh rate is enabled using the enable mpls rsvp-te summary-refresh command.

Message ID Extension: The message ID extension provides reliable delivery of RSVP messages. It also provides a simple mechanism for identifying refresh messages, which can greatly reduce refresh message processing on the receiving LSR. The message ID extension defines three new objects: message_ID, message_ID_ACK, and message_ID_NACK. The message_ID object contains a unique message identifier based on the sender's IP address. Only one message_ID object is inserted into an RSVP message. The receiving LSR can use the message_ID object to quickly refresh path and reserve states. If the message identifier value in the message_ID object is greater than the locally saved message identifier value, then the RSVP message represents a new or modified state. The receiving LSR must acknowledge an RSVP message using the message_ID_ACK object if the sender set the ACK_desired flag in the message_ID object, otherwise the message_ID acknowledgement is optional. The message_ID_ACK object may be included in any unrelated RSVP message or in an RSVP ACK message. Message ID extension is required for both bundle message and summary refresh, so this capability is automatically enabled if either of the other capabilities is enabled.

Hello Extension: The RSVP hello message provides a quick and simple mechanism for detecting the loss of a peer RSVP-TE LSR. The hello protocol is implemented using the RSVP soft-state model. RSVP hello messages may be enabled independently of each LSR peer. The hello protocol consists of two new objects: hello_request and hello_ACK. If configured, an LSR sends a hello_request every hello interval. If a hello_ACK is not received within a specified amount of time, the sending LSR assumes that its peer LSR is no longer active. Once a peer LSR is deemed inactive, all reservation states associated with LSPs established to or through the peer LSR must be freed and the LSPs torn down. The hello interval is configurable using the command configure mpls rsvp-te timers session hello-time.

You can improve LSP scaling by configuring the following RSVP-TE parameters:

- Refresh time
- Summary refresh time
- Bundle time

Refresh Time: The refresh time specifies the interval for sending refresh path messages. RSVP refresh messages provide soft state link-level keep-alive information for previously established paths and enable the switch to detect when an LSP is no longer active. RSVP sessions are torn down if an RSVP refresh message is not received from a neighbor within [(keep-multiplier + 0.5) * 1.5 * refresh-time] seconds. The valid refresh time may be set to any value between 1 and 36000 seconds. The default setting is 30 seconds. Configuring a longer refresh time reduces both switch and network overhead.

Summary Refresh Time: The summary refresh time, specified in tenths of a second, indicates the time interval for sending summary refresh RSVP messages. The summary refresh time must be less than the configured refresh time. The default summary refresh time is zero, indicating that no summary refresh

RSVP messages are sent. The summary refresh time value may be set to any value between zero to 100 (or 10 seconds). If configured, the bundled and summary refresh RSVP messages are only sent to RSVP-TE peers supporting RSVP refresh reduction.

Bundle Time: The bundle time, specified in tenths of a second, indicates the maximum amount of time a transmit buffer is held so that multiple RSVP messages can be bundled into a single PDU. The default bundle time is zero, indicating that RSVP message bundling is not enabled. The bundle time value can be set to any value between zero and 30 (or 3 seconds).

## Supporting Quality of Service Features

*QoS* LSP support is an important attribute of *MPLS*.

MPLS supports the Differentiated Services (DiffServ) model of QoS. The DiffServ QoS model is supported by mapping different traffic classes to different LSPs, or by using the EXP bits in the MPLS shim header to identify traffic classes with particular forwarding requirements.

## Propagation of IP TTL

There are two modes of operation for routed IP packets: pipe TTL mode and uniform TTL mode.

Currently, switches that run Extreme OS support only the pipe TTL mode. In pipe TTL mode, the LSP is viewed as a point-to-point link between the ingress LSR and the egress LSR; intermediate LSRs in the *MPLS* network are not viewed as router hops from an IP TTL perspective. Thus, the IP TTL is decremented once by the ingress LSR, and once by the egress LSR.

In this mode, the MPLS TTL is independent of the IP TTL. The MPLS TTL is set to 255 on all packets originated by the switch and on all packets that enter a pseudowire.

## IXP MPLS Enhancements

ExtremeXOS Release 15.4 adds the following IXP *MPLS* feature enhancements:

- *EMS* logs reflect the operation status changes of LSPs, PWs, and Interfaces.
- Show command displays details for transit and egress RSVP-TE LSPs.
- Metaswitch fixpack enhances logging of PATH failures.
- RSVP-TE description field now displays LSP Name/Path.
- MPLS commands are added to “show tech-support” functionality.
- Add support for the Metaswitch dmptrace facility.

### *EMS Logging of MPLS Protocols, LSPs, PWs, and Interfaces*

The following EMS logs allow you to track MPLS protocols, some LSPs, and PW and MPLS I/F operational states. These logs are added at the “Info” level, and are not generated in the default log output (the default level for MPLS is “Warning”).

#### LSPs

- MPLS.ChgStaticIngrLSPState
- MPLS.ChgStaticTrnstLSPState

- MPLS.ChgStaticEgrLSPState
- MPLS.RSVPTE.ChgIngrLSPState

PWs

- MPLS.L2VPN.ChgPWState

MPLS Protocol Interfaces

- MPLS.ChgIfState
- MPLS.ChgProtoState
- MPLS.LDP.ChgProtoState
- MPLS.RSVPTE.ChgProtoState

Note
CES.TDM.PWUp and CES.TDM.PWDown EMS logs already exist.

*Enhanced Show Details for Transit and Egress RSVP-TE LSPs*

The following fields are added to the `show mpls rsvp-te lsp` command output on transit and egress nodes:

- Administrative & Operational Statuses
- Setup & Hold Priorities

*Enhance Usability and Ability to Debug MPLS*

- Apply fixpack to provide better debugging for RSVP-TE Path message handling
- Use RSVP-TE "Tunnel Description" field in MPLS TE MIB to indicate LSP and Path

  Because primary and secondary RSVP-TE LSPs use the same LSP name, and the idea of Redundant LSPs is an EXOS feature, determining which path is associated with an LSP could only be done through the CLI. To eliminate this problem, the RSVP-TE "Tunnel Description" MIB field (RFC 3812: mplsTunnelDescr) now returns the "LSPName (PathName) on the Ingress node.

  Note
  The tunnel description is not signaled as part of RSVP messaging, so this information is not available at transit and egress nodes.

- Add MPLS (including RSVP-TE) debug information to `show tech-support mpls` command.
- Add support for the Metaswitch dmptrace facility

## Configuring MPLS

*MPLS* has the following configuration constraints:

- IP flow redirection—IP flow redirection commands and MPLS are mutually exclusive functions. Both functions cannot be enabled simultaneously.
- IGMP snooping—*OSPF* and LDP session establishment require the MSM/MM to receive and process IP multicast frames. Therefore, *IGMP* snooping must be enabled to support MPLS.
- VPLS—VPLS requires that IGMP snooping be disabled on customer-facing VLANs.

## Configuration Overview

Note
BlackDiamond 8800 series switches require specific software and hardware to support *MPLS*. For more information, see the *Feature License Requirements* document.

1. If MPLS will be used on a SummitStack or BlackDiamond 8800 series switch, select the enhanced protocol as described in Selecting the Enhanced Protocol on page 1203.
2. If you want to use MPLS in a different VR (the default is *VR-Default*), move MPLS to the appropriate VR as described in Moving MPLS From VR to VR on page 1203.
3. Create and configure the VLANs that will use MPLS.
4. Configure an MPLS LSR ID for each switch that serves as an LSR in the network. (See Configuring the MPLS LSR ID on page 1204.)
5. Add MPLS support to the appropriate VLANs. (See Adding MPLS Support to VLANs on page 1204.)
6. Enable MPLS on each switch that serves as an LSR. (See Enabling and Disabling MPLS on an LSR on page 1205.)
7. Enable MPLS on the appropriate VLANs. (See Enabling and Disabling MPLS on a VLAN on page 1205.)
8. Enable LDP on each switch that serves as an LSR. (See Enabling LDP on the Switch on page 1205.)
9. Enable LDP on the appropriate VLANs. (See Enabling and Disabling LDP on a VLAN on page 1206.)
10. Configure an IGP, such as *OSPF* or IS-IS, for the appropriate switches and VLANs.

## Selecting the Enhanced Protocol

The enhanced protocol is required to support *MPLS* on the stacking links between Summit X460, X480, and X670 series switches, and on the switch fabric of BlackDiamond 8800 series switches. On Summit X460, X480, and X670 series switches, the standard protocol previously used on these switches does not support MPLS.

Note
The Summit X670-G2 and X770 only support the enhanced protocol.

For information on selecting the enhanced stacking port protocol, see #unique_2403.

- On BlackDiamond 8800 series switches, use the following command to select the enhanced protocol for the switch fabric:

  ```
  configure forwarding switch-fabric protocol [standard | enhanced]
  ```

## Moving MPLS From VR to VR

By default, *MPLS* is enabled on the *VR-Default virtual router (VR)*. You can operate MPLS in VR-Default or in any user VR, but MPLS only operates in one VR. If you want to use MPLS in a user VR, you must delete it from VR-Default first.

- To delete MPLS from a VR, do the following:
  a. Disable MPLS on the LSR as described in Enabling and Disabling MPLS on an LSR on page 1205.
  b. Disable MPLS on all VLANs and reset the MPLS configuration as described in Resetting MPLS Configuration Parameter Values on page 1212.

c. Remove MPLS from the VR by entering the following command:

```
configure vr vr-name delete protocol protocol-name
```

Note
When you enter the command to delete MPLS from a VR, the software displays a prompt to remind you that the MPLS configuration is lost when MPLS is deleted.

- To add MPLS to VR, do the following:
  a. Use the following command:

  ```
  configure vr name add protocol mpls
  ```

  b. Add any other protocols that you want to use with MPLS.

After you add MPLS to a VR, you must configure and enable MPLS. MPLS commands operate only in the VR to which MPLS is added. An error message appears if you enter an MPLS command in a VR that does not support MPLS.

## Configuring the MPLS LSR ID

The MPLS LSR ID must be configured before *MPLS* can be used. The address chosen must be a routable IP address on the switch. It is suggested that this be set to the same IP address as a routing protocol ID (for example, the *OSPF* Router ID).

- To configure the MPLS LSR ID, use the following command:

  ```
  configure mpls lsr-id ipaddress
  ```

Note
The MPLS LSR ID must be configured before MPLS can be enabled. The LSR ID should be configured on a loopback *VLAN*.

## Adding MPLS Support to VLANs

To use *MPLS* on a *VLAN*, MPLS must first be configured on that VLAN.

- To configure a specific VLAN or all VLANs, use the following command:

  ```
  configure mpls add {vlan} vlan_name
  ```

  MPLS must be configured on a VLAN before it can be used to transmit or receive MPLS-encapsulated frames. By default, MPLS is not configured on a newly created VLAN.

  If you have enabled MPLS on an *OSPF* interface that is used to reach a particular destination, make sure that you enable MPLS on all additional OSPF interfaces that can reach that same destination (for example, enable MPLS on all VLANs that are connected to the backbone network).

## Enabling and Disabling MPLS on an LSR

- To enable *MPLS* on an LSR, use the following command:

  ```
  enable mpls
  ```

  **Note**
  Refer to the ExtremeXOS Command Reference description for the `enable mpls` command for special requirements for BlackDiamond 8800 series switches and SummitStack.

  By default, MPLS is disabled on the switch. This command starts the MPLS process, allowing MPLS protocols to run and MPLS packets to be forwarded.

- To disable MPLS on an LSR, use the following command:

  ```
  disable mpls
  ```

  When you disable MPLS, LDP and RSVP-TE are effectively disabled. The MPLS configuration remains intact.

## Enabling and Disabling MPLS on a VLAN

After *MPLS* is enabled globally, MPLS must be enabled on the MPLS configured *VLAN*s. Configuring a VLAN for MPLS does not enable MPLS on the VLAN. The VLAN must be specifically MPLS enabled in order to send and receive MPLS packets over an interface.

- To enable MPLS on specific VLANs, use the following command:

  `enable mpls [{vlan}vlan_name|vlan all]`

- To disable MPLS on specific VLANs, use the following command:

  `disable mpls [{vlan} vlan_name | vlan all]`

  MPLS must be enabled on all VLANs that transmit or receive MPLS-encapsulated frames. By default, MPLS and MPLS label distribution protocols are disabled when MPLS is first configured on a VLAN.

## Enabling LDP on the Switch

- To enable LDP on the switch, use the following command:

  ```
  enable mpls protocol ldp
  ```

  By default, LDP is disabled on the switch. This command enables *MPLS* to process LDP control packets and to advertise and receive LDP labels.

  **Note**
  Globally enabling the LDP protocol does not enable LDP on MPLS configured *VLAN*s. See the next section for instructions on enabling LDP on a VLAN.

## Enabling and Disabling LDP on a VLAN

Each *VLAN* must be specifically LDP-enabled in order to set up LDP adjacencies over the interface. Before LDP can be enabled on a VLAN, the VLAN must be configured to support *MPLS* as described in Adding MPLS Support to VLANs on page 1204.

- To enable LDP on a VLAN, use the following command:

  `enable mpls ldp [{vlan} vlan_name | vlan all]`

  This command enables LDP on one or all VLANs for which MPLS has been configured.

  **Note**
  If you have enabled LDP and MPLS on an IGP interface that is used to reach a particular destination, make sure that you enable LDP and MPLS on all additional IGP interfaces that can reach that same destination (for example, enable LDP and MPLS on all *OSPF* VLANs that are connected to the backbone network).

- To disable LDP on a VLAN, use the following command:

  `disable mpls ldp [{vlan} vlan_name | vlan all]`

  This command disables LDP on one or all VLANs for which MPLS has been configured. This command terminates all LDP hello adjacencies and all established LDP LSPs that use the specified interface(s).

## Creating Static LSPs

Static LSPs are label switched paths that are manually configured at each LSR in the prospective path. Static LSPs are too labor intensive for building complex network topologies, but they can be useful for defining simple one-hop LSPs to MTUs that might not have the CPU power necessary to support routing and label distribution protocols.

The following figure shows two static LSPs configured between *VLAN* A and VLAN B. The dashed line shows the static LSP for unidirectional communications from VLAN A to VLAN B. The solid line shows the static LSP for communications in the reverse direction.

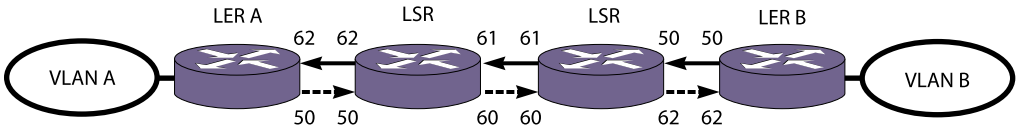



**Figure 196: Static LSP Example**

The path that an LSP takes and the labels the LSP uses at every hop do not change when the network topology changes due to links and nodes going up or down. Once enabled, a static LSP remains administratively up until it is manually disabled.

Static LSP configuration is different for ingress, transit, and egress LSRs. At the ingress LER, only an egress label is defined. At transit LSRs, both ingress and egress labels must be defined, and at the egress LER, only the ingress label is defined. During configuration, you must ensure that the egress label number for each LSR matches the ingress label number for the downstream LSR.

When creating static LSPs, consider the following guidelines:

- It is your responsibility to ensure that the egress label of an LSR matches the ingress label of the downstream LSR. The software does not detect or report label mismatches. Mismatches result in dropped or mis-routed packets.
- The operational state of the LSP is set to down at the head-end on local failures. However, there is no mechanism to detect the LSP going down when a failure occurs on a downstream node. When a failure occurs at a downstream node, traffic may be black-holed for the duration of the failure.
- The traffic profile for static LSPs is not configurable in this release. All static LSPs are given best effort treatment.
- The maximum number of static LSPs configurable on any given node is 1024. The maximum number of ingress static LSPs that are used to forward traffic to a single destination is 16 (as limited by *ECMP*).
- When multiple LSPs exist for the same destination, unless forced otherwise, signaled LSPs are preferred to static LSPs. When choosing an LSP for a FEC, the software prefers RSVP-TE LSPs first, LDP next, and finally static LSPs.
- Since the software has no knowledge of the cost or hop-count associated with each static LSP, all static LSPs to the same destination are equally preferred by IP routing.
- We recommend that the same LSP name be used on every LSR along the path of the static LSP. The software does not check for naming consistency across LSRs. However the switch does report an error when the configured name is not unique to the LSP on that LSR.

To configure a static LSP, use the following procedure at each node on the path:

1. Create a namespace for the LSP using the following command:

   `create mpls static lsp` *lsp_name* **destination** *ipaddress*
2. Configure the appropriate labels for the LSP using the following command:

   `configure mpls static lsp` *lsp_name* [{**egress** [*egress_label* | **implicit-null**] **egress-vlan** *evlan_name* **next-hop** *ipaddress*} {**ingress** *ingress_label* {**ingress-vlan** *ivlan_name*}}]
3. Configure optional traffic restrictions for IP or VPN traffic as needed using the following command:

   `configure mpls static lsp` *lsp_name* **transport** [**ip-traffic** [**allow** | **deny**] | **vpn-traffic** [**allow** {**all** | **assigned-only**} | **deny**]]
4. Enable the static LSP for operation using the following command:

   `enable mpls static lsp {`*lsp_name* | **all** }
5. When the configuration is complete, you can view the static LSP configuration with the following command:

   `show mpls static lsp {`**summary** | {*lsp_name*} {**detail**}}
6. Clear the counters for the static LSP using the command:

   `clear counters mpls static lsp {`*lsp_name* | **all** }
7. Once the static LSP is created on all path nodes, you can configure a default route, an IP route, or a VPN route to use the LSP. To configure a default or IP route to use the LSP, use the following command:

   `configure iproute add default [{`*gateway* {*metric*} {**vr** *vr_name*} {**unicast-only** | **multicast-only**}} | {**lsp** *lsp_name* {*metric*}}]

8. To configure a VPN route to use the LSP, use the following command:

   `configure l2vpn [vpls vpls_name | vpws vpws_name] peer ipaddress [add | delete] mpls lsp lsp_name`

9. To disable a static LSP, use the following command:

   `disable mpls static lsp {lsp_name | all }`

10. To delete a static LSP, use the following command:

    `delete mpls static lsp [lsp_name | all]`

## Configuring Penultimate Hop Popping

- To enable or disable PHP, use the following command:

  `enable mpls php [{vlan} vlan_name | vlan all]`

  This command enables or disables whether PHP is requested by the egress LER. If vlan all is selected, PHP is enabled on all VLANs on which *MPLS* has been configured. By default, PHP is disabled.

  When PHP is enabled, PHP is requested on all LSPs for which the switch is the egress LER.

  PHP is requested by assigning the Implicit Null Label in an advertised mapping. PHP is always performed when requested by a peer (for example, when acting as an intermediate LSR).

## Configuring QoS Mappings

The ExtremeXOS software provides examination and replacement services for the EXP field.

These services behave like the dot1p *QoS* commands. When EXP examination is enabled, the EXP value from the received frame is used to assign the packet to a QoS profile. Once a packet is assigned to a QoS profile, the EXP value may be overwritten in hardware before being transmitted. By default, the QoS profile EXP value is equivalent to the QoS profile number one (QP1). To enable QoS for *MPLS* LSPs, use the following command: `enable mpls exp examination`

This command enables EXP examination for MPLS received packets. When EXP examination is enabled, the EXP field in the outer or top label of the label stack is used to assign the received packet to an internal switch qosprofile. If enabled, all MPLS packets are mapped to a qosprofile based on the configured EXP bit value. That is, a packet with an EXP value of 0 is mapped to qosprofile QP1, a packet with an EXP value of 1 is mapped to qosprofile QP2, and so on. By default, EXP examination is disabled and all received MPLS packets are sent to QP1.

Each EXP value can be assigned a qosprofile. Multiple EXP values can be assigned to the same qosprofile. The following command is used to configure the switch to route packets with a received EXP value to a specific qosprofile:

`configure mpls exp examination {value} value {qosprofile} qosprofile`

The switch can overwrite the EXP value in the outer label of an MPLS packet.

The following command is used to enable the switch to replace the EXP value:

`enable mpls exp replacement`

This command enables EXP replacement for MPLS transmitted packets. When EXP replacement is enabled, the EXP field in the outer or top label of the label stack is replaced with the EXP value configured for the QoS profile for which the packet was assigned. By default, EXP replacement is disabled and packets are propagated without modifying the EXP field value.

Each qosprofile can be assigned an EXP value used to overwrite the EXP bit field in the outer label of the transmitted packet.

All qosprofiles can be configured to overwrite the EXP bit field using the same EXP value. The following command is used to configure the switch to overwrite MPLS packets transmitted from a specific qosprofile with a new EXP value:

```
configure mpls exp replacement {qosprofile} qosprofile {value} value
```

**Note**
On the X480, X460, E4G200, E4G400, X670, X670-G2, X770, and BlackDiamond X8, when exp examination is enabled, the dot1p value of VPLS/VPWS terminated frames are set from the internal priority. The enable/disable dot1p replacement commands do not have any effect in this situation. A future release will eliminate this restriction.

## Mapping Dot1p to EXP Bits

The priority of Ethernet tagged packets can be mapped into the *MPLS* network and vice versa using the switch fabric qosprofile.

Ethernet packets are assigned to a qosprofile by enabling dot1p examination. MPLS packets are assigned to a qosprofile by enabling exp examination. When the packets egress the switch, the dot1p and exp bit fields can be overwritten.

Enabling exp replacement instructs the switch to overwrite both the dot1p field and the exp field in the outer most label.

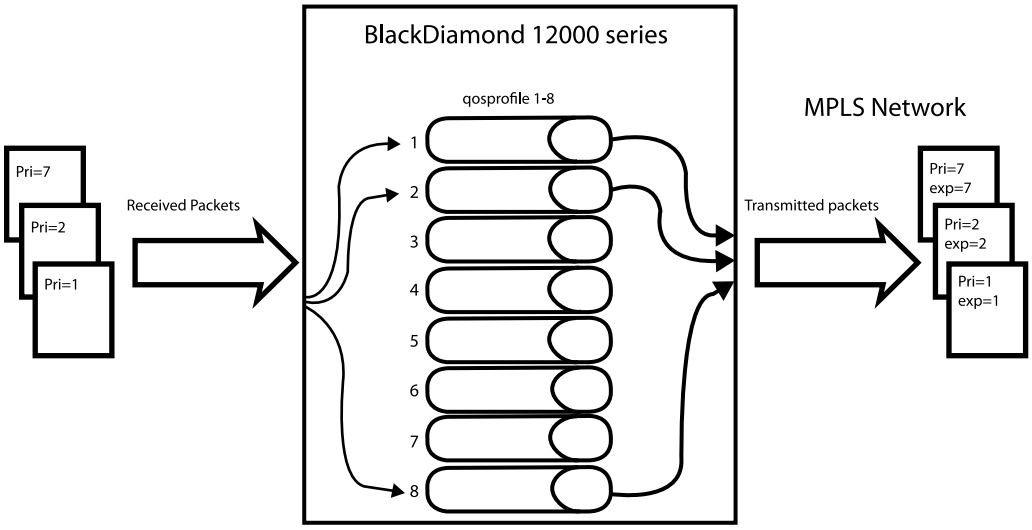



**Figure 197: Mapping Dot1p to EXP Bits**

By default, when dot1p examination and exp replacement are not enabled, all received packets are routed through qosprofile qp1 and the packets are transmitted with the dot1p and exp value of zero. As

shown in the figure above, the switch can be configured to route packets to a specific qosprofile based on the dot1p value in the 802.1p bit field. In this example dot1p examination is enabled. By default, when dot1p examination is enabled, packets are sent to the qosprofile that corresponds dot1p type to qosprofile mapping. This mapping can be viewed using `show dot1p type`. By default, MPLS exp replacement is disabled. By enabling MPLS exp replacement, MPLS packets are transmitted with the configured qosprofile dot1p and exp value. By default, these values correspond to the qosprofile. Qosprofile 1 overwrites the dot1p and exp fields with 0, qosprofile 2 overwrites the dot1p and exp fields with 1, and so on.

To configure the reverse packet flow, mpls exp examination and dot1p replacement must be configured. Enabling MPLS exp examination instructs the switch to route received MPLS packets to a qosprofile based on the EXP value. Enabling dot1p replacement instructs the switch to write the dot1p value in the transmitted packet.

## Enabling and Disabling LDP Loop Detection

There are two types of LDP loop detection. By default both are disabled. Enabling loop detection allows the switch to detect if a label advertisement loop exists within the network. Both hop count and path vector loop detection methods are used if loop detection is enabled.

- To enable LDP loop detection, use the following command:
  `enable mpls ldp loop-detection`
- To disable LDP loop detection, use the following command:
  `disable mpls ldp loop-detection`

These commands affect the LDP loop detection behavior for all LDP enabled VLANs.

## Configuring an LDP Label Advertisement Filter

- To configure an LDP label advertisement filter, use the following command:
  `configure mpls ldp advertise [{direct [all | lsr-id | none]} | {rip [all | none] | {static [all | none]}`

  This command configures a filter to be used by LDP when originating unsolicited label mapping advertisements to LDP neighbors.

  You can configure how the direct advertisement filter is applied, as follows:

  - direct—The advertisement filter is applied to the FECs associated with direct routes.
  - rip—The advertisement filter is applied to the FECs associated with *RIP* routes.
  - static—The advertisement filter is applied to the FECs associated with static routes.

  You can configure the advertisement filter, as follows:

  - all—Unsolicited label mappings are originated for all routes of the specified type (direct, RIP, or static).
  - lsr-id—An unsolicited label mapping is originated if a /32 direct route exists that matches the *MPLS* LSR ID. This filter is the default setting for direct routes and is only available for direct routes.
  - none—No unsolicited label mappings are originated for all routes of the specified type. This is the default setting for RIP and static routes.

You can control the number of labels advertised using the configure mpls ldp advertise command. Advertising labels for a large number of routes can increase the required number of labels that must be allocated by LSRs. Care should be used to insure that the number of labels advertised by LERs does not overwhelm the label capacity of the LSRs.

## Configuring LDP Session Timers

- To configure LDP session timers, use the following command:

  `configure mpls ldp timers [`**`targeted`**` | `**`link`**`] [{`**`hello-time`**` `*`hello_hold_seconds`*`} {`**`keep-alive-time`**` `*`keep_alive_hold_seconds`*`}]`

  This command configures the LDP peer session timers for the switch. The LDP peer session timers are separately configurable for link and targeted LDP hello adjacencies.

  The **`hello-time`** *`hello_hold_seconds`* parameter specifies the amount of time (in seconds) that a hello message received from a neighboring LSR remains valid. The rate at which hello messages are sent is one third the configured hello-time. If a hello message is not received from a particular neighboring LSR within the specified hello-time *`hello_hold_seconds`* then the hello-adjacency is not maintained with that neighboring LSR. Should two peers have different configured hello-time values, they negotiate to use the lower value.

  The **`session keep-alive time`** *`keep_alive_hold_seconds`* parameter specifies the time (in seconds) during which an LDP message must be received for the LDP session to be maintained. The rate at which keep alive messages are sent, provided there are no LDP messages transmitted, is one sixth the configured keep-alive-time. If an LDP PDU is not received within the specified **`session keep-alive time`** *`keep_alive_hold_seconds`* interval, the corresponding LDP session is torn down. Should two peers have different configured keep-alive-time values, they negotiate to use the lower value.

  In the event that two peers have both a link and a targeted hello adjacency, the hello-time values for the two hello adjacencies are negotiated separately. The keep-alive-time value is established based on negotiations occurring when the LDP session is established following the first hello adjacency to be established.

  The minimum and maximum values for both the **`hello-time`** *`hello_hold_seconds`* and **`keep-alive time`** *`keep_alive_hold_seconds`* are 6 and 65,534, respectively. Changes to targeted timers only affect newly created targeted peers. Disabling and then enabling all VPLS instances causes all current targeted peers to be re-created.

  The default values are as follows:

  - link ldp hello-time *`hello_hold_seconds`* – 15
  - targeted-ldp hello-time *`hello_hold_seconds`* – 45
  - link ldp hello-time *`interval_time`* – auto set to 1/3 the configured hello-time
  - targeted-ldp hello-time *`interval_time`* – auto set to 1/3 the configured hello-time
  - link ldp keep-alive *`keep-alive hold-seconds`* – 40
  - targeted-ldp keep-alive *`keep-alive hold-seconds`* – 60
  - link ldp keep-alive *`interval_time`* – auto set to 1/6 the configured keep-alive time
  - targeted-ldp keep-alive *`interval_time`* – auto set to 1/6 the configured keep-alive time

*Restoring LDP Session Timers*

- To restore the default values for LDP session timers, use the following command:
  `unconfigure mpls`

  This command can only be executed when *MPLS* is disabled, and it restores all MPLS configuration settings.

## Clearing LDP Protocol Counters

- To clear the LDP control protocol error counters, use the following command:
  `clear counters mpls ldp {{{vlan} vlan_name} | lsp all}`

  Omitting the optional *vlan* parameter clears counters for all LDP enabled interfaces.

## Resetting MPLS Configuration Parameter Values

- To reset *MPLS* configuration parameters to their default values, disable MPLS (see Enabling and Disabling MPLS on an LSR on page 1205) and then use the following command:
  `unconfigure mpls`

  This command affects the following configuration parameters:
  - All VLANs are removed from MPLS.
  - All EXP (examination and replacement) QOS mappings are reset.
  - LSR-ID is reset.
  - LDP and RSVP-TE are globally disabled.
  - MPLS Traps are disabled.
  - LDP and RSVP-TE timers are reset.
  - LDP Advertisement settings are reset.
  - LDP Loop Detection settings are reset.
  - All RSVP-TE LSPs are deleted.
  - All RSVP-TE Paths are deleted.
  - All RSVP-TE Profiles are deleted.
  - The default RSVP-TE Profile is reset.

## Managing the MPLS BFD Client

Bidirectional Forwarding Detection (BFD) is introduced in Bidirectional Forwarding Detection (BFD) on page 411.

The *MPLS* BFD client enables rapid detection of failures between MPLS neighbors on specific *VLAN*s. BFD detects forwarding path failures at a uniform rate, which makes the re-convergence time consistent and predictable and makes network profiling and planning easier for network administrators.

When BFD detects a communication failure, it informs MPLS, which treats the indication as an interface (VLAN) failure. This allows the MPLS protocols to quickly begin using alternate paths to affected neighbors (the methodology for selecting alternate paths is dependent upon the MPLS protocol in use and how it reacts to interface failure conditions). As MPLS connections (LSPs) are removed from the

interface, BFD sessions are removed as well, and the interface returns to a state without BFD protection. The MPLS protocol might continue to attempt to reestablish LSP connections across the interface, and if successful, also attempt to establish a BFD session with the corresponding neighbor. MPLS does not process BFD state changes until the BFD session is fully active in the UP state, at which point state changes are processed and the state for LSPs which cross the interface becomes BFD protected.

Note
BFD sessions are established only when both peers select the same LSP route. We recommend that BFD operate only on interfaces that have one peer.

- To enable the MPLS BFD client, use the following command:

  `enable mpls bfd [{vlan} vlan_name | vlan all]`

  Note
  BFD must be enabled on the interface before sessions can be established. To enable BFD, use the command: `[enable | disable] bfd vlan vlan_name` .

- To disable the MPLS BFD client, use the following command:

  `disable mpls bfd [vlan all | {vlan} vlan_name] {delete-sessions}`

- To display the MPLS BFD client information, use the following commands:

  `show mpls interface {{vlan} vlan_name} {detail}`

  `show mpls bfd [{vlan} vlan_name | ip_addr]`

# Displaying MPLS Configuration Information

## Displaying MPLS Basic Configuration Information

- To display basic *MPLS* configuration information, use the following command:

  `show mpls`

  This command displays the general configuration of all the MPLS components and system wide configuration.

The output, shown below, displays the switch MPLS, RSVP-TE, and LDP configuration status. It also shows the configuration status of *SNMP* trap, EXP examination, and EXP replacement settings. The configured LSR ID are also shown.

```
# show mpls
Virtual Router Name                    : VR-Default
MPLS Admin Status                      : Enabled
MPLS Oper Status                       : Enabled
RSVP-TE Admin Status                   : Enabled
RSVP-TE Oper Status                    : Enabled
LDP Admin Status                       : Enabled
LDP Oper Status                        : Enabled
MPLS SNMP Traps                        : Disabled
L2VPN SNMP Traps                       : Disabled
EXP Examination                        : Enabled
EXP Replacement                        : Disabled
LSR ID                                 : 192.99.1.5
```

## Displaying LDP Basic Configuration Information

- To display basic LDP configuration information, use the following command:
  `show mpls ldp`

  This command displays the general configuration of LDP.

  Some settings are not configurable. These fields are identified with an asterisk (*). The remaining fields can be modified using LDP configuration commands to change the behavior of LDP. A list of VLANs that have LDP enabled is shown at the bottom of the display output.

```
# show mpls ldp
LDP Status                : Enabled
Protocol Version          : v1*
Label Retention Mode      : Liberal*
Label Distribution Method : Downstream Unsolicited*
LDP Loop Detection
Status            : Disabled
Hop-Count Limit   : 255
Path-Vector Limit : 255
LDP Targeted Timers
Hello Hold      : 45 seconds
Keep Alive Hold : 60 seconds
LDP Link Timers
Hello Hold      : 15 seconds
Keep Alive Hold : 40 seconds
Label Advertisement
Direct : All
Rip    : None
Static : None
LDP VLANs  : loopback
: blowingrock
: boone
: asheville
* Indicates parameters that cannot be modified
```

## Displaying MPLS Interface Information

- To display the *MPLS* interface information, use the following command:
  `show mpls interface {{vlan} vlan_name} {detail}`

  When the optional parameters are omitted, this command displays information for all the configured MPLS *VLAN* interfaces.

  The summary MPLS interface information displayed includes the configured IP address, approximate time RSVP-TE and LDP have been up, number of neighbors or adjacencies, and a set of status flags.

  When the *vlan_name* parameter is specified, this command displays the current MPLS interface summary configuration and status for only the specified VLAN. If the optional detail keyword is specified, the summary information is displayed in the detail format.

## Displaying LDP Interface Information

- To display LDP interface information, use the following command:
  `show mpls ldp interface {{vlan} vlan_name} {detail | counters}`

  This command displays the operational LDP interface information.

All the *VLAN*s that have LDP enabled are listed. The negotiated LDP hello hold time is displayed. This is not the configured LDP hello hold time but represents the value that is negotiated between the local switch and the connected LDP peer. Hello messages are transmitted every 1/3 the negotiated hello hold time. In the following example, that would be every 5 seconds. The hello timer status information is displayed in milliseconds. The approximate LDP uptime and status flags are also shown.

```
# show mpls ldp interface
VLAN Name         #Adj  NegHHldTm  NxtHello  UpTm  Flags
----------------  ----  ---------  --------  ----  -----
loopback             0      15000      2230   16h  MLU
blowingrock          1      15000      2200   14h  MLU
boone                1      15000      2210   16h  MLU
asheville            1      15000      2210   16h  MLU
Flags: (M) MPLS Enabled, (L) LDP Enabled,
(U) LDP Operational
```

## Displaying MPLS Label Information

- To display *MPLS* label information, use the following commands:

  **show mpls** {**rsvp-te** | **static**} **label** {**summary** | *label_num* | [**advertised** | **received**] {*label_num*} | **received implicit-null**}

  **show mpls label l3vpn** {**summary** | *label_num* | [**advertised** | **received**] {*label_num*}}

  **show mpls** {**ldp**} **label** {**lsp**} {**summary** | *label_num* | [**advertised** | **received**] {*label_num*} | **received implicit-null**}

  **show mpls ldp label** {**lsp**} **advertised implicit-null** {*ipNetmask*}

  **show mpls** {**ldp**} **label l2vpn** {**summary** | *label_num* | [**advertised** | **received**] {*label_num*}}

  **show mpls ldp label l2vpn retained** {*ipaddress*}

  **show mpls ldp label lsp retained** {*ipNetmask*}

  **show mpls ldp label retained** [**l2vpn** {*ipaddress*} | **lsp** {*ipNetmask*}]

## Displaying MPLS Label Mapping Information

- To display *MPLS* label mapping information, use the following command:

  **show mpls ldp lsp** {**prefix** *ipNetmask*} {**ingress** | **egress** | **transit**} {**detail**}

  **show mpls rsvp-te lsp** [egress | transit] {**fast-reroute**} {{*lsp_name*} {[**destination** | **origin**] *ipaddress*} {**detail**} | **summary**}

  This command displays information about how to forward packets that arrive labeled as MPLS packets. As such, it shows how labels advertised to upstream peers are mapped to labels received from downstream peers. This mapping is sometimes referred to as the Incoming Label Map (ILM).

  When the label_number parameter is omitted, summary information is displayed for all incoming label assignments that have been made by the switch. When the label_number is specified, summary information is displayed for the specified label.

As can be seen below, the output display differs for label mappings signaled by LDP and RSVP-TE. Please see the respective sections for additional information.

```
# show mpls ldp lsp
Prefix            Adv Label  Peer Label  Next Hop        VLAN
192.168.0.4/32     0x80402   --          --               lpbk
192.168.0.2/32     0x00039   0x80400     11.0.2.2         vlan2
11.0.4.0/24        0x0003A   0x80401     11.0.2.2         vlan2
192.168.0.4/32     0x0003B   0x00013     11.0.2.2         vlan2
* BD-10K.6 # show mpls rsvp-te lsp
Ingress LSP Name Path Name        Destination     Transmit I/F     UpTm
Flags
---------------- ---------------- --------------- ---------------- ----
--------
LSR1-LSR4        path1-2-4        192.168.0.4     vlan2             47m
UEP--OIV
Egress LSP Name  Source IP       Destination     Receive I/F      UpTm
---------------- --------------- --------------- ---------------- ----
LSR4-LSR1        192.168.0.4     192.168.0.1     vlan1             47m
Transit LSP Name Source IP       Destination     Receive I/F   Transmit I/F  UpTm
---------------- --------------- --------------- ------------- ------------- ----
LSR2-LSR3        192.168.0.2     192.168.0.3     vlan2          vlan1        47m
Flags: (U) Up, (E) Enabled, (P) Primary LSP, (S) Secondary LSP,
(R) Redundant Paths, (B) Bandwidth Requested, (O) ERO Specified,
(I) IP Traffic Allowed, (V) VPN Traffic Allowed,
(v) VPN Assigned Traffic Allowed
```

## Displaying MPLS QoS Mapping Information

- To display *MPLS QoS* mapping information, use the following command:

```
* BD-10K.10 # show mpls exp examination
EXP --> QoS Profile mapping:
00 --> QP1
01 --> QP2
02 --> QP3
03 --> QP4
04 --> QP5
05 --> QP6
06 --> QP7
07 --> QP8
EXP Examination is disabled
* BD-10K.11 # show mpls exp replacement
QoS Profile --> EXP mapping:
QP1 --> 00
QP2 --> 01
QP3 --> 02
QP4 --> 03
QP5 --> 04
QP6 --> 05
QP7 --> 06
QP8 --> 07
EXP Replacement is disabled
* BD-10K.12 #
```

Configured mappings for both dot1p-to-exp and exp-to-dot1p are displayed.

## Displaying LDP Peer Session Information

- To display *MPLS* LDP peer session information, use the following command:
  `show mpls ldp peer {ipaddress} {detail}`

This command displays information about the status of LDP peer sessions. Summary information is displayed for all known LDP peers and LDP peer sessions. If you specify the *ipaddress* of an LDP peer, information for the single LDP peer is displayed. If you specify the **detail** keyword, additional information is displayed in a comprehensive detailed format.

By default the information displayed includes:

- Peer sessions
- Peer state
- Uptime
- Number of hello adjacencies

If you specify the **detail** keyword, the following additional information is displayed:

- Discontinuity time
- Negotiated label distribution
- Next hop address
- Keep-Alive hold timer
- Hello adjacency details

## Displaying LDP Protocol Counters

LDP control protocol packet error counters are maintained per interface.

- To view these counters, use the following command:

  `show mpls ldp interface {{vlan} vlan_name} {detail | counters}`

  These counters may be useful in determining LDP issues between peers.

  Error counters that continually increment should be investigated.

```
# show mpls ldp interface vlan blowingrock counters
VLAN: blowingrock (192.60.40.5)
Link     Targeted
Counter                          Adjacencies  Adjacencies
-------------------------------  -----------  -----------
Shutdown Notifications (Rcvd)              0            0
Shutdown Notifications (Sent)              0            0
Failed Session Attempts (NAKs)             0            0
Hello Errors                               0            0
Parameters Advertised Errors               0            0
Max PDU Length Errors                      0            0
Label Range Errors                         0            0
Bad LDP ID Errors                          0            0
Bad PDU Length Errors                      0            0
Bad Msg Length Errors                      0            0
Bad TLV Length Errors                      0            0
Bad TLV Value Errors                       0            0
Keep-Alive Timeout Errors                  0            0
```

  Omitting the optional **vlan** parameter displays counters for all LDP enabled interfaces.

## Displaying LDP LSP Forwarding Database

- To display information about LDP LSPs, use the following command:

  `show mpls ldp lsp {**prefix** *ipNetmask*} {**egress** | **ingress** | **transit**} {**detail**}`

  This command displays the LDP LSPs established to, from, and through this switch. By default, ingress, egress, and transit LSPs are all displayed. By optionally specifying the LSP type, the output display is filtered to show only the type of LSPs specified.

  When all LSP types are being displayed, LSPs that show only an advertised label represent egress LSPs from the network perspective or incoming LSPs from the switch perspective. LSPs that show only a received label represent ingress LSPs from the network perspective or outgoing LSPs from the switch perspective. LSPs that show both an incoming and an outgoing label represent transit LSPs. As Extreme switches are merge-capable, all LDP transit LSPs can also be used as ingress LSPs.

  The significance of the *VLAN* information shown depends on the LSP type. For ingress and transit LSPs, the indicated VLAN is the *MPLS* interface used to reach the next hop peer. For egress LSPs, there is no associated MPLS next hop interface. When the prefix being advertised is associated with a local (direct) VLAN, that VLAN name is displayed. When the advertised prefix is associated with a static or an *RIP* route, the VLAN field is empty.

  Advertised labels have switch-wide significance and are generally advertised out multiple interfaces.

```
* BD-10K.15 # show mpls ldp lsp
Prefix            Adv Label  Peer Label  Next Hop        VLAN
192.99.1.5/32      0x80402    --          --              loopback
192.80.40.0/24     0x80403    --          --              asheville
192.100.40.0/24    0x80404    --          --              boone
192.60.40.0/24     0x8040b    --          --              blowingrock
192.24.100.0/24    0x00015    0x80401     192.100.40.3    boone
192.99.1.3/32      0x00016    0x80403     192.100.40.3    boone
192.10.50.0/24     0x00018    0x80405     192.100.40.3    boone
10.20.30.0/24      0x0001c    0x00013     192.100.40.3    boone
11.136.96.0/24     0x0001d    0x00014     192.100.40.3    boone
```

  Specifying the optional detail keyword displays each LSP in detail format.

  Additionally, received packets and bytes are maintained for transit and egress (or incoming) LSPs. Specifying the keyword prefix and a matching ipNetmask restricts the display to a single entry.

```
*BD-X8.17 # show mpls ldp lsp prefix 11.108.96.0/24 detail
FEC IP/Prefix: 11.108.96.0/24:
Advertised Label: 0 (0)
Received Label  : 0x18 (24)
Next Hop IP     : 192.100.40.3
VLAN Name       : boone
Packets received  :        489
Bytes received    :      46944
```

## Displaying RSVP-TE LSP Configuration Information

- Tp display RSVP-TE LSP configuration information, use the following command:

  `show mpls rsvp-te`

This command displays summary configuration and status information for RSVP-TE. Global status of RSVP-TE and the configured standard and rapid-retry LSP timer values are included in the display output.

## Displaying the RSVP-TE Paths

- To display RSVP-TE paths, use the following command:

  `show mpls rsvp-te path {path_name} {detail}`

  This command displays the configuration and status information for *MPLS* RSVP-TE paths. Information is listed in tabular format and includes the path name, number of configured ERO objects, number of LSPs configured to use this path, the list of EROs and their type. Optionally specifying the **detail** keyword displays the path information in verbose format, including all LSPs that are configured to use the path.

## Displaying the RSVP-TE Path Profile

- To display RSVP-TE path profiles, use the following command:

  `show mpls rsvp-te profile {profile_name} {detail}`

  By default, this command displays all configured profile parameters for the specified profile. If the profile name is omitted, the profile parameter values for all configured LSP profiles are displayed. Optionally specifying the keyword **detail** displays the profile information is verbose format. When the **detail** keyword is specified, all LSPs that are configured to use this profile are also displayed.

## Displaying the RSVP-TE LSP

- To display the RSVP-TE LSP, use the following command:

  `show mpls rsvp-te lsp {{[destination | origin] ipaddress} {fast-reroute} {detail} | summary}`

  This command displays the LSP information associated with RSVP-TE that is used to forward packets within the *MPLS* network. If no options are specified, summary information for all RSVP-TE LSPs is displayed. This information includes the LSP name, LSP direction relative to the MPLS network (i.e., ingress, egress, or transit), incoming and or outgoing interface, configured destination, and uptime. If the optional LSP name parameter is specified, only the LSP information for the specified ingress LSP name is displayed. Optionally, the LSPs displayed can be further qualified by the keywords **ingress**, **egress**, and **transit**. These keywords qualify the LSPs displayed from the perspective of the switch.

  Ingress LSPs identify LSPs that originate from the switch into the MPLS network. Egress LSPs identify LSPs that terminate at the switch from the MPLS network. Transit LSPs represent LSPs that traverse the switch.

  The optional **destination** keyword limits the display to only those LSPs that terminate at the specified IP address. The optional **origin** keyword limits the display to only those LSPs that originate at the specified IP address. Optionally specifying the **detail** keyword displays the LSP information in verbose format. Additional information displayed includes the configured path and

profile, transmit and/or receive labels, failure and retry counts, packet and byte counters, and recorded routes.

## MPLS Configuration Example

The network configuration, shown in the following figure, illustrates how to configure a BlackDiamond switch to support a routed *MPLS* network.

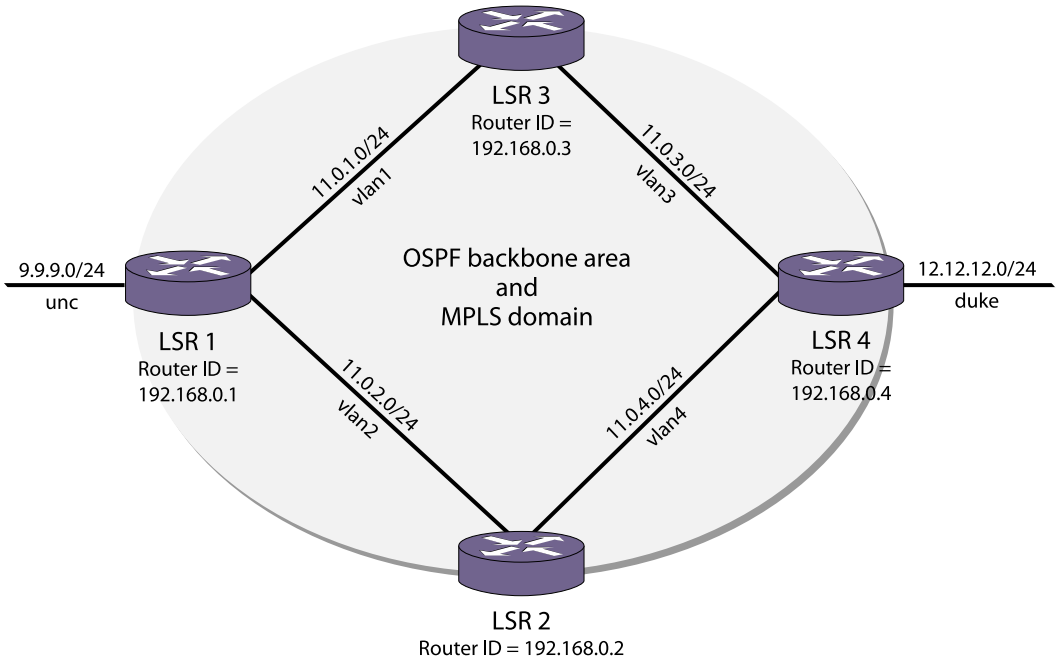



**Figure 198: MPLS Configuration Example**

The four switches, labeled LSR 1, LSR 2, LSR 3, and LSR 4, have the same physical hardwareconfiguration. Each switch contains two 8900-G48T-xl and an 8900-MSM128 module. The switches are all interconnected via Gigabit Ethernet to form the *OSPF* backbone area and the MPLS domain. In this example, two directly connected OSPF-disabled VLANs are shown: unc and duke. Traffic between unc and duke follows routed paths over calculated LSPs established between LSR 1 and LSR 4.

The commands used to configure LSR 1 are described below. The remaining LSRs are configured similarly.

The following commands configure the module types for specific slots:

```
configure slot 2 module 8900-G48T-xl
configure slot 2 module 8900-G48T-xl
```

The following command sets the maximum jumbo frame size for the switch chassis to 1600:

```
configure jumbo-frame-size size 1600
enable jumbo-frame ports all
```

The following commands create the VLANs:

```
create vlan lpbk
create vlan vlan1
create vlan vlan2
create vlan unc
```

The following commands configure the *VLAN* IP address and assign ports participating in each VLAN:

```
configure vlan lpbk ipaddress 192.168.0.1/32
enable ipforwarding lpbk
enable loopback-mode lpbk
configure vlan default delete ports all
configure vlan vlan1 ipaddress 11.0.1.1/24
configure vlan vlan1 add port 3:2 untagged
configure vlan vlan2 ipaddress 11.0.2.1/24
configure vlan vlan2 add port 3:3 untagged
configure vlan unc ipaddress 9.9.9.1/24
configure vlan unc add port 3:4 untagged
```

The following commands enable IP forwarding on the configured VLANs.

The MTU size is increased on the MPLS VLANs to accommodate the MPLS shim header:

```
enable ipforwarding vlan vlan1
configure ip-mtu 1550 vlan vlan1
enable ipforwarding vlan vlan2
configure ip-mtu 1550 vlan vlan2
enable ipforwarding vlan unc
```

The following command configures the MPLS LSR ID:

```
configure mpls lsr-id 192.168.0.1
```

The following commands add MPLS support to VLANs lpbk, vlan1, and vlan2:

```
configure mpls add vlan lpbk
configure mpls add vlan vlan1
configure mpls add vlan vlan2
```

The following commands enable MPLS on VLANs lpbk, vlan1, and vlan2 and LDP on VLANs vlan1 and vlan2:

```
enable mpls lpbk
enable mpls vlan1
enable mpls vlan2
enable mpls ldp vlan1
enable mpls ldp vlan2
```

The following command allows LDP to advertise a label mapping for the LSR ID:

```
configure mpls ldp advertise direct lsr-id
```

The following commands globally enable LDP and MPLS on the switch:

```
enable mpls protocol ldp
enable mpls
```

The following commands add lpbk, vlan1, and vlan2 to the backbone area.

The 0.0.0.0 (backbone) area does not need to be created because it exists by default:

```
configure ospf add vlan lpbk area 0.0.0.0
configure ospf add vlan vlan2 area 0.0.0.0
configure ospf add vlan vlan1 area 0.0.0.0
```

The following command enables distribution of local (direct) interfaces into the OSPF area:

```
enable ospf export direct cost 10 type ase-type-1
```

The following command configures the OSPF router ID on the switch and enables the distribution of a route for the OSPF router ID in the router LSA.

Originating the router ID as a host route allows other routers in the same OSPF area to establish calculated LSPs for external routes to this router:

```
configure ospf routerid 192.168.0.1
```

It also allows this router to be a peer in a L2 VPN.

The following command enables OSPF:

```
enable ospf
```

# Configuring MPLS Layer-2 VPNs (VPLS and VPWS)

**Note**
ELRP should not be used to protect VPLS service VLANs because ELRP is not aware of VPLS pseudowires.

## Configuring MPLS for Establishing Layer 2 VPN Instances

As described in Using LDP to Signal PW Label Mappings on page 1162, *MPLS* and LDP must be properly configured in order to establish the PWs that make up a Layer 2 VPN.

MPLS should be enabled using the enable mpls command and LDP should be enabled using the enable mpls protocol ldp command.

Since the MPLS LSR ID is used as the local endpoint for PWs, it is highly desirable to create a loopback *VLAN* whose associated IP address is that of the MPLS LSR ID. The configured prefix length should be 32. As described in Establishing LDP LSPs to PW Endpoints on page 1161, configuring this loopback VLAN for MPLS causes the address to be included in LDP address messages. Use the `configure mpls add vlan` *`vlan_name`* command for this. It is not required that LDP or MPLS be enabled on the VLAN for the address to be advertised by LDP. Use the configure mpls ldp advertise direct lsr-id command to initiate an LDP label mapping for an LDP LSP to the local endpoint.

## Creating or Deleting a Layer 2 VPN Domain

- To create a Layer 2 VPN, use the following command:

  `create l2vpn [`**`vpls`** *`vpls_name`* `|` **`vpws`** *`vpws_name`*`]` **`fec-id-type pseudo-wire`** *`pwid`*

  This command creates a named Layer 2 VPN instance. Multiple domains can be created, each representing a different L2 VPN. The pwid is a number that is used to signal and identify which Layer 2 VPN is associated with each pseudowire. All of the pseudowires carrying traffic for a specific Layer 2 VPN are signaled with the same pwid. No Layer 2 VPN traffic is forwarded over the Layer 2 VPN until at least one peer is added to the Layer 2 VPN and a service is associated with the Layer 2 VPN. The configured pwid also doubles as the Layer 2 VPN ID.
- To delete a Layer 2 VPN, use the following command:

  `delete l2vpn [`**`vpls`** `[`*`vpls_name`* `|` **`all`**`] |` **`vpws`** `[`*`vpws_name`* `|` **`all`**`]]`

This command deletes the named Layer 2 VPN instance or all Layer 2 VPN instances, depending on the keyword. All Layer 2 VPN peers and services associated with the deleted Layer 2 VPN instance(s) are also deleted.

## Enabling or Disabling a Layer 2 VPN Domain

- To enable a Layer 2 VPN, use the following command:
  `enable l2vpn [vpls [vpls_name | all] | vpws [vpws_name | all]]`

  This command enables a named Layer 2 VPN instance. By default, a newly created Layer 2 VPN is enabled.
- To disable a Layer 2 VPN, use the following command:
  `disable l2vpn [vpls [vpls_name | all] | vpws [vpws_name | all]]`

  This command disables the named Layer 2 VPN instance. When a Layer 2 VPN is disabled, no traffic flows across the Layer 2 VPN. The pseudowires connecting this peer to all other configured peers are also terminated, so the remote peers no longer see this LSR as an active peer.

## Adding or Deleting a Layer 2 VPN Peer

- To add a peer to the Layer 2 VPN, use the following command:
  `configure l2vpn [vpls vpls_name | vpws vpws_name] add peer ipaddress {{core {full-mesh | primary | secondary} | spoke}`

  For each new peer added, a pseudowire is signaled to carry traffic for this Layer 2 VPN. Up to 64 peers can be added to a VPLS; and only one peer can be added to a VPWS. For each peer added, that remote peer must also configure this local LSR as a peer for the Layer 2 VPN. For VPLS configurations, this insures that the VPLS core is configured as a full mesh of VPLS peers.

  The Layer 2 VPN names on each peer do not have to match since the pseudowire ID is used to define the Layer 2 VPN to which each pseudowire is associated.
- Delete a peer from the Layer 2 VPN, use the following command:
  `configure l2vpn [vpls vpls_name | vpws vpws_name] delete peer [ipaddress | all]`

  Once the peer is deleted, that specified peer is no longer a member of the Layer 2 VPN. For VPLS configurations, the peer must also be removed from all other VPLS peers to insure a proper full mesh and to prevent connectivity issues.

## Add or Delete a Layer 2 VPN Service

- To add a service to a Layer 2 VPN, use the following command:
  `configure l2vpn [vpls vpls_name | vpws vpws_name] add service [{vlan} vlan_name | {vman} vman_name]`

  Only one service can be added to each Layer 2 VPN. Traffic associated with the service is transported over the Layer 2 VPN. Three basic types of services are supported: *VLAN*, VMAN, and port. Both the VLAN and VMAN services are specified by adding the VLAN or VMAN name to the Layer 2 VPN. The port service is configured by adding a VMAN name to the Layer 2 VPN, configuring

the Layer 2 VPN to strip the VMAN tag, and adding the port as untagged to the VMAN. This allows incoming service traffic to be transported across the Layer 2 VPN exactly as it was received. See Managing Layer 2 VPN Packet Forwarding Options on page 1224 for information about configuring a Layer 2 VPN to strip the VMAN tag.

- To delete a service from a Layer 2 VPN, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] delete service [{vlan} vlan_name | {vman} vman_name]`

  Since there is no local service that needs to be connected to the Layer 2 VPN, the pseudowires to each of the configured peers for this Layer 2 VPN are terminated.

  Note

  Ports added to Layer 2 VPN service VPLS/VMAN should not be part of PVLAN or VLAN aggregation or VLAN translation.

## Enabling or Disabling a Layer 2 VPN Service

- To enable a Layer 2 VPN service, use the following command:

  `enable l2vpn [vpls [vpls_name | all] | vpws [vpws_name | all]] service`

  By default, any configured Layer 2 VPN service is enabled.
- To disable a service from the Layer 2 VPN, use the following command:

  `disable l2vpn [vpls [vpls_name | all] | vpws [vpws_name | all]] service`

  When the service is disabled, the service is disconnected from the Layer 2 VPN and disabled such that no packets are sent to or received from the service. The pseudowires to each of the configured peers for this Layer 2 VPN are terminated.

## Managing Layer 2 VPN Packet Forwarding Options

- To configure Layer 2 VPN packet forwarding options, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] {dot1q [ethertype hex_number | tag [include | exclude]]} {mtu number}`

  The options should be configured the same for every LSR for this Layer 2 VPN in order to prevent connectivity issues. Specifying the dot1q ethertype forces the switch to overwrite the dot1q ethertype value in the service packet. This can be used to interconnect two customer segments over the Layer 2 VPN that are using different configured ethertype values. By default, the dot1q tag in the service packet is not included. The switch can be configured to strip or exclude the dot1q tag. This can be used to emulate port services or for interoperability with equipment that may require tags.
- To unconfigure Layer 2 VPN packet forwarding options, use the following command:

  `unconfigure l2vpn [vpls vpls_name | vpws vpws_name] dot1q ethertype`

  This command resets the dot1q ethertype for the specified Layer 2 VPN to the default ethertype configured for the switch.

## Configuring the Layer 2 VPN MTU

The Maximum Transmission Unit (MTU) is the maximum packet size (excluding the data link header) that an interface can transmit. The Layer 2 VPN MTU defines the MTU for the customer facing interface. By default, the MTU size is set to 1500 bytes, which is the MTU size for standard Ethernet frames.

The frame size of the customer packet also includes the data link header and FCS field. The Ethernet data link header includes a minimum of a destination MAC address (six bytes), a source MAC address (six bytes), and an ethertype field (two bytes). The FCS field is four bytes. For a 1500-byte customer payload, the minimum customer frame size is 1518 bytes.

The Layer 2 frame, minus the FCS field, is encapsulated to form an *MPLS* packet to be transmitted to a Layer 2 VPN peer. This encapsulation adds another Ethernet header, an MPLS shim header, and a FCS field. The MPLS shim header usually includes two four-byte MPLS labels. Therefore, this encapsulation adds a minimum of 26 bytes.

The total frame size includes the customer payload, the Ethernet header on the customer facing interface (minimum 14 bytes), and the MPLS encapsulation (minimum 26 bytes). For a 1500-byte customer payload, the minimum Layer 2 VPN frame size is 1540 bytes.

The total frame size is restricted by the jumbo frame size. The maximum jumbo frame size that can be configured is 9216 bytes. If the customer payload in the above example is increased to 9176 bytes, the resulting Layer 2 VPN encapsulated frame size is 9216 bytes. Therefore, the maximum practical value of the Layer 2 VPN MTU is 9176 bytes.

There are additional considerations that can lower the maximum practical value of the Layer 2 VPN MTU setting. If the Layer 2 VPN is configured to include the *VLAN* tag of the customer interface, this increases the total frame size by four bytes. If the MPLS interface used to transmit the encapsulated Layer 2 VPN packet is a tagged VLAN, this also increases the total frame size by four bytes. If either of these configuration options is present, the maximum practical Layer 2 VPN MTU size is reduced accordingly.

- To configure the Layer 2 VPN MTU size, use the **mtu** option in the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] {dot1q [ethertype hex_number | tag [include | exclude]]} {mtu number}`

  The Layer 2 VPN MTU value is signaled to peers as part of the VC FEC information during PW setup. The local value must match the value received from the PW peer. If the two values do not match, the PW is not established.

## Managing VPLS Redundancy Options

- To configure VPLS redundancy options, use the following command:

  `configure {l2vpn} vpls vpls_name redundancy [esrp esrpDomain | eaps | stp]`
- To unconfigure VPLS redundancy options, use the following command:

  `unconfigure {l2vpn} vpls vpls_name redundancy [eaps | esrp | stp]`
- To display the configured VPLS redundancy option, use the following command:

  `show l2vpn vpls {vpls_name} {peer ipaddress} detail`

### Displaying Layer 2 VPN Status

- To display the status of all Layer 2 VPNs, a configured Layer 2 VPN, or a specific peer within a specific Layer 2 VPN, use the following command:

  `show [ {l2vpn} vpls {vpls_name} | l2vpn vpws {vpws_name} | l2vpn ] {peer ipaddress} {detail} | summary }`

### Displaying Layer 2 VPN Statistics

- To display statistics for Layer 2 VPNs, use the following command:

  `show mpls statistics l2vpn {vpls_name | vpws_name } {detail}`

### Managing Layer 2 VPN SNMP Traps

- To enable or disable Layer 2 VPN *SNMP* traps, use the following commands:

  `enable snmp traps l2vpn`

  `disable snmp traps l2vpn`

- To configure or remove a text-string identification for traps from a Layer 2 VPN, use the following command:

  `configure vpls vpls_name snmp-vpn-identifier identifier`

  `unconfigure vpls vpls_name snmp-vpn-identifier`

- To view the configured state for Layer 2 VPN SNMP traps, use the following command:

  `show mpls`

## VPLS VPN Configuration Examples

### Basic Point-to-Point VPLS Configuration Example

This *MPLS* VPLS network configuration shown in the following figure, builds upon the routed MPLS network configuration example shown in the following figure.

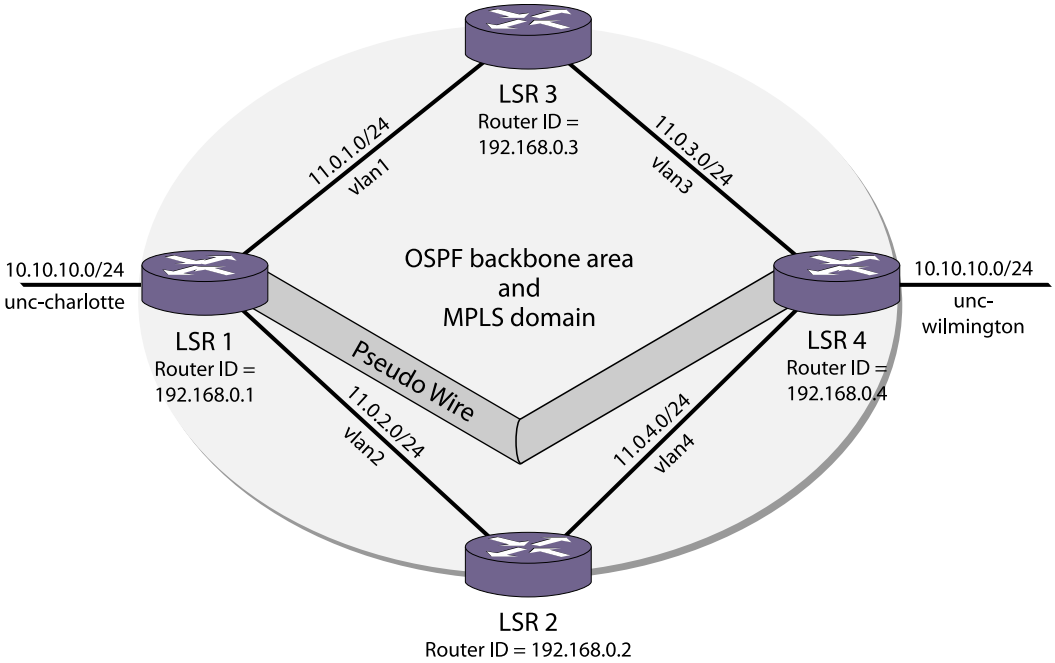



**Figure 199: MPLS VPLS Configuration Example**

Assuming that the routed MPLS network has already been configured as in the previous example, the commands used to create the VPLS on LSR1 and LSR4 follow. The nc-university-vpn is a point-to-point VPLS, since only two peers are participating.

The following command creates the VPLS with VPN ID 35. This command must be issued on both LSR1 and LSR4:

```
create l2vpn vpls nc-university-vpn fec-id-type pseudo-wire 35
```

The following command enables the VPLS instance specified by the *vpls_name*:

```
enable l2vpn [vpls [ vpls_name | all ]
```

On LSR1, configure the local *VLAN* service and the VPLS peer:

```
configure l2vpn vpls nc-university-vpn add service vlan unc-charlotte
configure l2vpn vpls nc-university-vpn add peer 192.168.0.4
```

On LSR4, configure the local VLAN service and the VPLS peer:

```
configure l2vpn vpls nc-university-vpn add service vlan unc-wilmington
configure l2vpn vpls nc-university-vpn add peer 192.168.0.1
```

## Multipoint Full Mesh VPLS Configuration Example

The example shown in the following figure configures a four node full-mesh *MPLS* VPLS configuration by adding two additional peers to the previous example.

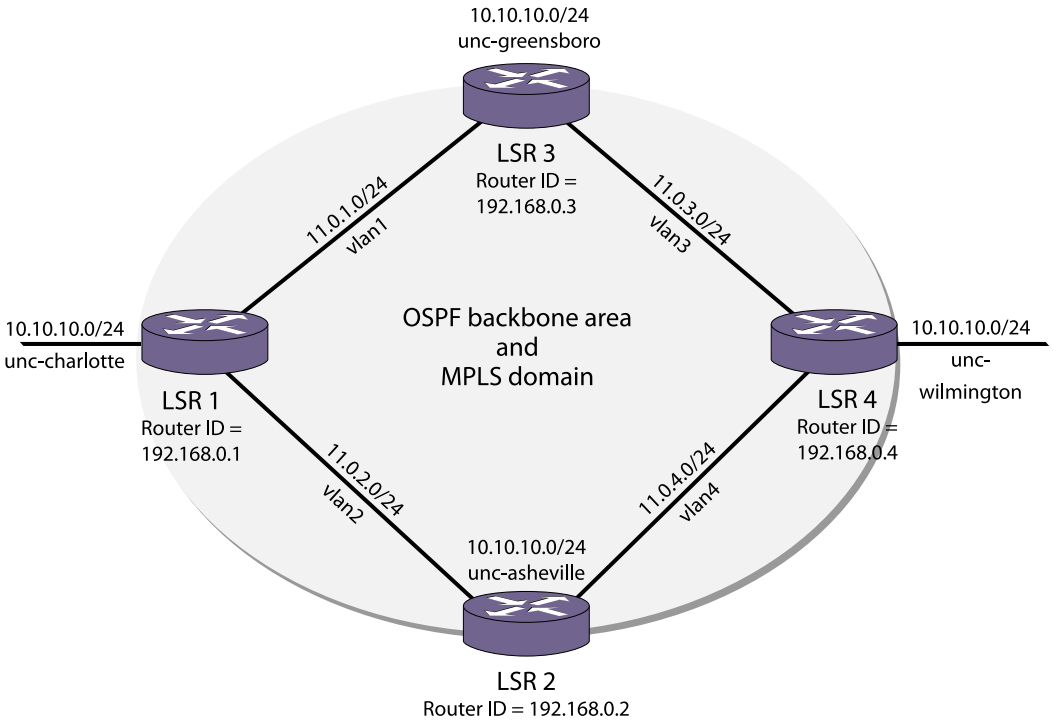



**Figure 200: Full Mesh Configuration Example**

*LSR1*

```
configure vpls nc-university-vpn add peer 192.168.0.2
configure vpls nc-university-vpn add peer 192.168.0.3
```

*LSR2*

```
create vpls nc-university-vpn fec-id-type pseudo-wire 35
configure vpls nc-university-vpn add peer 192.168.0.1
configure vpls nc-university-vpn add peer 192.168.0.3
configure vpls nc-university-vpn add peer 192.168.0.4
configure vpls nc-university-vpn add service vlan unc-asheville
```

*LSR3*

```
create vpls nc-university-vpn fec-id-type pseudo-wire 35
configure vpls nc-university-vpn add peer 192.168.0.1
configure vpls nc-university-vpn add peer 192.168.0.2
configure vpls nc-university-vpn add peer 192.168.0.4
configure vpls nc-university-vpn add service vlan unc-greensboro
```

*LSR4*

```
configure vpls nc-university-vpn add peer 192.168.0.2
configure vpls nc-university-vpn add peer 192.168.0.3
```

## VPLS with Redundant EAPS Configuration Example

The following sections provide examples of how to configure switches to support the configuration described in VPLS EAPS Redundancy Overview on page 1181:

- Core 1 Router Configuration on page 1229
- Core 2 Router Configuration on page 1229

### *Core 1 Router Configuration*

The following example shows the EAPS and VPLS configuration for the Core 1 router.

```
//create the EAPS protected VLAN and the VPLS service vlan
create vlan v1
configure v1 add ports 2:3  // port 2:3 is the port connected to dist 1
create eaps eaps2
configure eaps eaps2 mode transit
configure eaps eaps2 primary port 2:1  // port going to core 2
configure eaps eaps2 secondary port 2:3
configure eaps eaps2 add control vlan ctrl2   // eaps control vlan
configure eaps eaps2 add protected vlan v1
create eaps shared-port 2:1
configure eaps shared-port 2:1 mode controller
configure eaps shared-port 2:1 link-id 888
enable eaps2
create vpls vpls1 fec-id-type pseudo-wire 2
configure vpls vpls1 add service vlan v1
configure vpls vpls1 redundancy eaps
configure vpls vpls1 add peer <core2>
configure vpls vpls1 add peer <core3>
configure vpls vpls1 add peer <core4>
```

### *Core 2 Router Configuration*

The following example shows the EAPS and VPLS configuration for the Core 2 router.

```
//create the EAPS protected VLAN and the VPLS service vlan
create vlan v1
configure v1 add ports 2:9  // port 2:9 is the port connected to dist 5
create eaps eaps2
configure eaps eaps2 mode transit
configure eaps eaps2 primary port 2:1 // port going to core 1
configure eaps eaps2 secondary port 2:9
configure eaps eaps2 add control vlan ctrl2
configure eaps eaps2 add protected vlan v1
create eaps shared-port 2:1
configure eaps shared-port 2:1 mode partner
configure eaps shared-port 2:1 link-id 888
enable eaps2
create vpls vpls1 fec-id-type pseudo-wire 2
configure vpls vpls1 add service vlan v1
configure vpls vpls1 redundancy eaps
configure vpls vpls1 add peer <core1>
configure vpls vpls1 add peer <core3>
configure vpls vpls1 add peer <core4>
```

# Configuring H-VPLS

H-VPLS is configured at the edge of the network. The core of the network supports H-VPLS and is configured as described in Configuring MPLS Layer-2 VPNs (VPLS and VPWS) on page 1222. To

configure H-VPLS, you need to configure the H-VPLS spoke nodes and the PE core nodes that peer with the spoke nodes.

## Configuring H-VPLS Spoke Nodes

- To configure an MTU as an H-VPLS spoke node, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] add peer ipaddress {{core {full-mesh | primary | secondary} | spoke}`

  Use the **core primary** and **core secondary** command options as needed. The **core primary** option specifies that the spoke node peer is a core node and that the link between the peers is the primary spoke. The **core secondary** option specifies that the spoke node peer is a core node and that the link between the peers is the secondary spoke.
- To delete an H-VPLS spoke node, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] delete peer [ipaddress | all]`

## Configuring H-VPLS Core Nodes

- To configure a VPLS core node as an H-VPLS core node, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] add peer ipaddress {{core {full-mesh | primary | secondary} | spoke}`

  Use the spoke command option to specify that the peer is an H-VPLS spoke node. When the H-VPLS spoke and core peers are configured, VPLS communications can be established between them.

  Note
  To enable communications from the H-VPLS spoke across the VPLS network, the H-VPLS core node must also be configured to peer with the other VPLS nodes.

- To delete an H-VPLS core node, use the following command:

  `configure l2vpn [vpls vpls_name | vpws vpws_name] delete peer [ipaddress | all]`

## Configuring the MAC Address Withdrawal Feature

The MAC address withdrawal feature is enabled by default.

- To disable this feature, use the following command:

  `disable l2vpn vpls peer [ipaddress | all] fdb send-mac-withdrawal`
- To enable this feature after it has been disabled, use the following command:

  `enable l2vpn vpls peer [ipaddress | all] fdb send-mac-withdrawal`

### Displaying H-VPLS Configuration Information

- To display H-VPLS configuration information, use the following command:

  `show [ {l2vpn} vpls {vpls_name} | l2vpn vpws {vpws_name} | l2vpn ] {peer ipaddress} {detail} | summary }`

## Configuring Protected VPLS

- To configure a protected VPLS, use the following command:

  `create esrp esrp_domain {type [vpls-redundancy | standard]}`

  `configure {l2vpn} vpls vpls_name redundancy [esrp esrpDomain | eaps | stp]`
- To unconfigure a protected VPLS, use the following command:

  `unconfigure {l2vpn} vpls vpls_name redundancy [eaps | esrp | stp]`
- To display information about the protected VPLS, use the following command:

  `show esrp { {name} | {type [vpls-redundancy | standard]} }`

## Configuring RSVP-TE

### Enabling and Disabling RSVP-TE on the Switch

- To enable RSVP-TE on a switch, use the following command:

  `enable mpls protocol rsvp-te`
- To disable RSVP-TE on a switch, use the following command:

  `disable mpls protocol rsvp-te`

Note

*MPLS* must be globally enabled before RSVP-TE can become operational. For more information, see Enabling and Disabling MPLS on an LSR on page 1205.

### Enabling and Disabling RSVP-TE on a VLAN

- To enable RSVP-TE on one or all *VLAN*s, use the following command:

  `enable mpls rsvp-te [{vlan} vlan_name | vlan all]`
- To disable RSVP-TE on one or all VLANs, use the following command:

  `disable mpls rsvp-te te [{vlan} vlan_name | vlan all]`

  Disabling RSVP-TE on a VLAN causes all TE LSPs using that interface to be released, and prevents TE LSPs from being established or accepted on the specified VLAN.

### Configuring RSVP-TE Protocol Parameters

- To configure RSVP-TE protocol parameters, use the following command:

  `configure mpls rsvp-te timers session[{bundle-message-time bundle_message_milliseconds} {hello-keep-multiplier hello_keep_number}`

{**hello-time** *hello_interval_seconds*}{**refresh-keep-multiplier** *refresh_keep_number*} {**refresh-time** *refresh_seconds*}{**summary-refresh-time** *summary_refresh_milliseconds*}] [{**vlan**} *vlan_name* | **vlan all**]

This command configures the RSVP-TE protocol parameters for the specified *VLAN*. The RSVP-TE keyword **all** indicates that the configuration changes apply to all RSVP-TE enabled VLANs.

The hello-interval time specifies the RSVP hello packet transmission interval. The RSVP hello packet is used by the switch to detect when an RSVP-TE peer is no longer reachable. If an RSVP hello packet is not received from a peer within [hello-interval * keep-multiplier] seconds, the peer is declared down and all RSVP sessions to and from that peer are torn down. The default hello-interval time is zero, indicating that RSVP hellos are not enabled. The hello-interval can be set to any value between zero and 60 seconds.

The **refresh-time** parameter specifies the interval for sending refresh path messages. RSVP refresh messages provide soft state link-level keep-alive information for previously established paths and enable the switch to detect when an LSP is no longer active. RSVP sessions are torn down if an RSVP refresh message is not received from a neighbor within [(keep-multiplier + 0.5) * 1.5 * **refresh-time**] seconds. The default **refresh-time** is 30 seconds and the default keep-multiplier value is three. The minimum and maximum **refresh-time** values are one and 36,000 seconds (or ten hours), respectively. The minimum and maximum keep-multiplier values are one and 255, respectively.

The bundle-time, specified in tenths of a second, indicates the maximum amount of time a transmit buffer is held so that multiple RSVP messages can be bundled into a single PDU. The default bundle-time is zero, indicating that RSVP message bundling is not enabled. The bundle-time value can be set to any value between zero and 30 (or 3 seconds).

The summary-refresh-time, specified in tenths of a second, indicates the time interval for sending summary refresh RSVP messages. The summary-refresh-time must be less than the configured **refresh-time**. The default summary-refresh-time is zero, indicating that no summary refresh RSVP messages are sent. The summary-refresh-time value may be set to any value between zero to 100 (or 10 seconds).

If configured, the bundled and summary refresh RSVP messages are only sent to RSVP-TE peers supporting RSVP refresh reduction.

## Creating or Deleting an RSVP-TE LSP

- To create an RSVP-TE LSP, use the following command:

  `create mpls rsvp-te lsp` *lsp_name* **destination** *ipaddress*

  The *lsp_name* parameter is a character string that identifies the LSP within the switch. The *lsp_name* string must begin with an alphabetic character and can contain up to 31 additional alphanumeric characters. The LSP is not signaled until at least one path is added. See Adding a Path to an RSVP-TE LSP on page 1236.
- To delete an RSVP-TE LSP, use the following command:

  `delete mpls rsvp-te lsp [`*lsp_name* | **all**]

Deleting an LSP name disassociates all configured paths with this LSP and all configuration information for the LSP name is deleted. If the LSP has been specified for use by a static route or a VPLS, that configuration information is also deleted. If you specify the **all** keyword, all LSPs are deleted.

## Creating an RSVP-TE Path

- To create an RSVP-TE routed path resource, use the following command:

  ```
  create mpls rsvp-te path path_name
  ```

  The *path_name* parameter is a character string that is to used to identify the path within the switch. The *path_name* string must begin with an alphabetic character, and can contain up to 31 additional alphanumeric characters. The RSVP-TE LSP is not signaled along the path until an LSP adds the specified path name. The maximum number of configurable paths is 1000.
- To delete an RSVP-TE path, use the following command:

  ```
  delete mpls rsvp-te path [path_name | all]
  ```

  This command deletes a configured *MPLS* RSVP-TE routed path with the specified *path_name*. All associated configuration information for *path_name* is deleted. A path cannot be deleted as long as the *path_name* is associated with an LSP. If the **all** keyword is specified, all paths not associated with an LSP are deleted.

## Configuring an Explicit Route

The routed path for an RSVP-TE LSP can be described by a configured sequence of the LSRs and/or subnets traversed by the path. Each defined LSR or subnet represents an ERO subobject. Up to 64 subobjects can be added to each path name. LSRs and/or subnets can be either included or excluded.

- To add an RSVP-TE explicit route, use the following commands:
  - To specify an LSR or subnet to be included in the path calculation:

    ```
    configure mpls rsvp-te path path_name add ero {include} ipNetmask
    [strict|loose] {order number}
    ```
  - To specify an LSR or subnet to be excluded from the path calculation:

    ```
    configure mpls rsvp-te path path_name add ero exclude ipNetmask
    {order number}
    ```

These commands add an IP address to the explicit route object (ERO) for the specified path name. The include keyword is optional and the default behavior is to define an ERO that must be included in the path. The exclude keyword allows a path to be created that must avoid certain subnets. This can be useful when defining redundant LSPs or paths that must avoid the path of other LSPs or paths.

The ipaddress keyword identifies an LSR using either a /32 address, which may represent an LSR router ID, loopback address, or direct router interface, or an IP prefix, which represents a directly connected subnet. Each IP address or prefix is included in the ERO as an IPv4 subobject.

For EROs that are configured to be included in the path calculation, if the IP address is specified as strict, the strict subobject must be topologically4 adjacent to the previous subobject as listed in the ERO. If the IP address is specified as loose, the loose subobject is not required to be topologically

adjacent to the previous subobject as listed in the ERO. If omitted, the default subobject attribute is loose.

For EROs that are configured to be excluded in the path calculation, a given subnet is avoided if any address on that subnet is specified.

If the subobject matches a direct router interface or a directly attached subnet, the switch verifies that the path message is received on the matching router interface. The LSR path order is optionally specified using the order keyword. The order number parameter is an integer value from 1 to 65535. IP prefixes with a lower order number are sequenced before IP prefixes with a higher number. You can specify multiple addresses and assign them an order number. The order number determines the path that the LSP follows. Thus, the LSP follows the configured path of the IP prefix with the order value from low to high. If the order keyword is not specified, the number value for the LSR defaults to a value 100 higher than the current highest number value. For excluded nodes, the order is not important. Any excluded ERO will be always be avoided no matter where it is in the list of ERO objects.

If the list of IP prefixes added to the path does not reflect an actual path through the network topology, the path message is returned with an error from a downstream LSR and the LSP is not established.

The order of a configured subobject can not be changed. The ERO subobject must be deleted and re-added using a different order. If a subobject is added to or deleted from the ERO while the associated LSP is established, the path is torn down and is resignaled using the new ERO. Duplicate ERO subobjects are not allowed. Defining an ERO for the path is optional. If you do not configure an ERO, the path is signaled along the best CSPF calculated path and the ERO is not included in the path message. When the last subobject in the ERO of the path message is reached and the egress IP node of the path has not been reached, the remaining path to the egress node is signaled along the best CSPF calculated path. Specification of an ERO could lead to undesirable routed paths, so care should be taken when terminating the ERO routed-path definition prior to the configured path egress node.

To delete an RSVP-TE explicit route, use the following command:

```
configure mpls rsvp-te path path_name delete ero [all | ipNetmask |
order number]
```

This command deletes an LSR or subnet from the ERO for the specified path name. The LSR is specified using the *ipaddress*, or *order* parameter. If an LSR is deleted from an ERO while the associated LSP is established, the path is torn down and is resignaled using a new ERO. Use the all keyword to delete the entire ERO from the path name. When there is no configured ERO, the path is no longer required to take an explicit routed path. The path is then signaled along the best CSPF calculated path and no ERO is included in the path message.

## Reserving Bandwidth for MPLS

In order to manage the bandwidth used by *MPLS* RSVP-TE LSPs, a configured amount of bandwidth must be reserved for MPLS. Bandwidth in the transmit direction must be reserved on all MPLS interfaces which will be used as egress interfaces for RSVP-TE LSPs requesting bandwidth. Bandwidth in the receive direction can be optionally reserved on all MPLS interfaces that will be used as ingress interfaces

for RSVP-TE LSPs requesting bandwidth. Note that the receive bandwidth reservation can be exceeded, and is provided only for tracking and informational purposes.

- The following command is used to reserve bandwidth for MPLS:

  `configure mpls rsvp-te bandwidth committed-rate` *committed_bps* [**Kbps** | **Mbps** | **Gbps**] [{**vlan**} *vlan_name* | **vlan all**] {**receive** | **transmit** | **both**}

  The committed_rate_unit must be specified in Kbps, Mbps, or Gbps. Choosing the **vlan all** option reserves the specified amount of bandwidth on all MPLS interfaces.

  The default reserved value is zero. Therefore, no LSPs requesting bandwidth can be established until the bandwidth has been configured.

## Creating and Deleting an RSVP-TE Profile

- To create a traffic engineered LSP profile, use the following command:

  `create mpls rsvp-te profile` *profile_name* {**standard**}

  This command creates a configured RSVP-TE profile with the specified profile name. The default profile cannot be deleted. If a profile is associated with a configured LSP, the profile cannot be deleted.
- To delete a traffic engineered LSP profile, use the following command:

  `delete mpls rsvp-te profile` [*profile_name* | **all**]

## Configuring an RSVP-TE Profile

- To configure an RSVP-TE profile, use the following command:

  `configure mpls rsvp-te` **profile** *profile_name* {**bandwidth** [**best-effort** | [{**committed-rate** *committed_bps* [**Kbps** | **Mbps** | **Gbps**]} {**max-burst-size** *burst_size* [**Kb** | **Mb**]} {**peak-rate** *peak_bps* [**Kbps** | **Mbps** | **Gbps**]}]} {**hold-priority** *hold_priority*} {**mtu** [**number** | **use-local-interface**]} {**path-computation** [**full** | **partial**]} {**record** [**enabled** {**route-only**} | **disabled**]} {**setup-priority** *setup_priority*}

  A profile is a set of attributes that are applied to the LSP when the LSP is configured using the `create mpls rsvp-te lsp` command. A default profile is provided which cannot be deleted, but can be applied to any configured LSP. The profile name for the default profile is default. The default profile parameter values are initially set to their respective default values. The maximum number of configurable profiles is 1000 (one of which is reserved for the default profile).

  The **bandwidth** parameter specifies the desired reserved bandwidth for the LSP. Any positive integer value is valid. You must append the characters k for kilobits, m for megabits, or g for gigabits, to the bps value to specify the unit of measure. The default bandwidth bps value is zero, which indicates that the *QoS* for the LSP is best effort.

  The **max-burst-size** and **peak-rate** parameters are signaled in the sender Tspec object and add further definition of the expected traffic. The **mtu** parameter is also signaled in the sender Tspec object and defines the expected maximum packet size that is sent over the LSP.

The **setup-priority** and **hold-priority** are optional parameters indicating the LSP priority. During path set up, if the requested bandwidth cannot be reserved through the LSR, the **setup-priority** parameter is compared to the **hold-priority** of existing LSPs to determine if any of the existing LSPs need to be preempted to allow a higher priority LSP to be established. Lower numerical values represent higher priorities. The **setup-priority** range is 0 to 7 and the default value is 7. The **hold-priority** range is also 0 to 7 and the default value is 0.

Starting with ExtremeXOS 16.2, the **path-computation** keyword specifies computation strategy for calculating a path to the LSP destination. "full" requires the ingress node to full calculate a path to the LSP destination. "partial" allows the ingress node to calculate only part of the path to the LSP destination. In this case, it would be to the ABR, and the ABR would calculate the rest of the way (into the other area).

Note
Inter area RSVP fast reroute (FRR) is not supported.

The **record** keyword is used to enable hop-by-hop path recording. The enabled keyword causes the record route object (RRO) to be inserted into the path message. The RRO is returned in the reserve message and contains a list of IPv4 subobjects that describe the RSVP-TE path. Path recording by default is disabled. When disabled, no RRO is inserted into the path message.

- To delete an RSVP-TE path profile, use the following command:
  `delete mpls rsvp-te profile [profile_name | all]`

  This command deletes a configured RSVP-TE profile with the specified profile name. The default profile cannot be deleted. If a profile is associated with a configured LSP, the profile cannot be deleted. If you specify the **all** keyword, all profiles not associated with an LSP are deleted (except for the default profile).

## Adding a Path to an RSVP-TE LSP

- To add a path to an RSVP-TE LSP, use the following command:
  `configure mpls rsvp-te lsp lsp_name add path [path_name | any] {profile profile_name} {primary {frr_profile_name} | secondary}`

  This command adds a configured path to the specified RSVP-TE LSP. The LSP name parameter is a character string that is to be used to identify the LSP within the switch and must have been created previously. The LSP is not signaled until a path is added to the LSP. Up to three paths can be defined for the LSP: one primary and two secondary. All paths are signaled, but only one path is used to forward traffic at any one time. The switch chooses the local *MPLS VLAN* interface from which to signal the LSP. To force an LSP to use a specific local MPLS interface, configure the peer's interface IP address as the first ERO in the associated path. The profile name is optional. If omitted, the default profile is applied to the LSP. The path name can be specified or the LSP can be configured to take any path. For a given LSP, only one path can be configured to take any path through the MPLS network.

  The specified path defaults to primary when no primary path has been configured for the LSP and defaults to secondary if the primary path has been previously configured for the LSP.

  Each *path_name* added to an *lsp_name* must be unique, but a *path_name* can be associated with multiple LSP names.

All configured primary and secondary paths for the *lsp_name* must have the same endpoint IP address. For example, three paths can be configured for the *lsp_name*, but all paths should represent different topological paths through the network to the same LSP endpoint.

Adding a secondary *path_name* designates a path as a hot-standby redundant path, used in the event that the primary or the other secondary path cannot be established or fails. Provided the *path_name* has not already been established, all paths are signaled as soon as they are associated with an *lsp_name*. If the primary *path_name* fails, is not configured, or cannot be established after the specified LSP retry-timeout, one of the configured secondary paths becomes the active path for *lsp_name*. All of the secondary paths have equal preference; the first one available is chosen. If at any time the primary path is reestablished, *lsp_name* immediately switches to using the primary path. If a secondary path fails while in use, the remaining configured secondary path can become the active path for *lsp_name*.

- To delete a path from an RSVP-TE LSP, use the following command:

  `configure mpls rsvp-te lsp` *lsp_name* **`delete path`** [*path_name* | **`any`** | **`all`**]

  When you issue this command, the LSP associated with the path is immediately torn down. If the deleted path represents the in-use LSP for *lsp_name* and another secondary path is configured, the LSP immediately fails over to an alternate LSP. Because at least one path must be defined for each LSP, the last configured path cannot be deleted from the LSP.

## Setting up Fast-Reroute Protection for an LSP

To create a protected LSP, do the following:

1. Create the LSP as described in Creating or Deleting an RSVP-TE LSP on page 1232.
2. Enable the fast-reroute feature on the LSP using the following command: `configure mpls rsvp-te lsp` *lsp_name* **`fast-reroute`** [**`enable`** | **`disable`**]
3. Create a path for the LSP as described in Creating an RSVP-TE Path on page 1233.
4. Define the path route as described in .
5. If you want to use a custom profile instead of the default profile, create a profile for the protected LSP as described in Creating and Deleting an RSVP-TE Profile on page 1235.
6. If you want to configure the custom profile created in the previous step, configure the profile as described in Configuring an RSVP-TE Profile on page 1235.
7. If you want to use a custom fast-reroute profile instead of using the default fast-reroute profile, create the profile using the following command at the ingress LER:

   `create mpls rsvp-te profile` *profile_name* **`fast-reroute`**
8. If you want to configure the custom fast-reroute profile created in the previous step, use the following command at the ingress LER:

   `configure mpls rsvp-te profile` *frr_profile_name* {**`bandwidth`***bandwidth_rate_bps**bandwidth_rate_unit*} {**`detour`** {**`hop-limit`***hop_limit_value*} {**`bandwidth-protection`** [**`enabled`** | **`disabled`**]} {**`node-protection`** [**`enabled`** | **`disabled`**]}} {**`hold-priority`***hold_priority_value*} {**`setup-priority`***setup_priority_value*}

9. Add the path to the protected LSP and select the standard and fast-reroute profiles using the following command:

   ```
   configure mpls rsvp-te lsp lsp_nameaddpath [path_name | any]
   {profileprofile_name} {primary {frr_profile_name} | secondary}
   ```

## RSVP-TE Configuration Example

RSVP-TE LSPs comprise profiles, paths, and the actual LSP. This section describes how to configure an RSVP-TE LSP.

Configuring RSVP LSPs is a multi-step process with some optional steps, depending on the specific requirements of the LSP. Conceptually, a number of mandatory elements must be configured to create an RSVP-TE LSP. In addition, you can also configure optional elements. In certain configurations, there are also order dependencies.

The profile contains constraints that you might wish to apply to the LSP. These constraints can affect the path selected across the *MPLS* domain in order to meet those constraints. Examples of profile parameters include bandwidth, setup, and hold priority relative to other configured LSPs.

The path can be used to specify the explicit path across the MPLS domain that the LSP should follow. This is done using EROs. An ERO is an object, sent as part of the LSP setup request (path message) that explicitly specifies the part of the path across the MPLS domain the setup request should follow. You can configure both loose and strict EROs in a path.

Certain elements of configuration are order dependent. For example if you specify a profile or path when creating an LSP, those path or profile definitions must already exist. Similarly a path must exist before an ERO is created, as the ERO is added explicitly to the path.

The typical steps used to configure and verify an RSVP-TE LSP are as follows:

1. Create and configure a path (optional).
2. Reserve bandwidth for the LSP (optional).
3. Create and configure a profile (optional).
4. Create an LSP (mandatory).
5. Add a primary/secondary path to the LSP (mandatory).
6. Add a secondary path to the LSP (optional).

7. Verify LSP status (recommended).

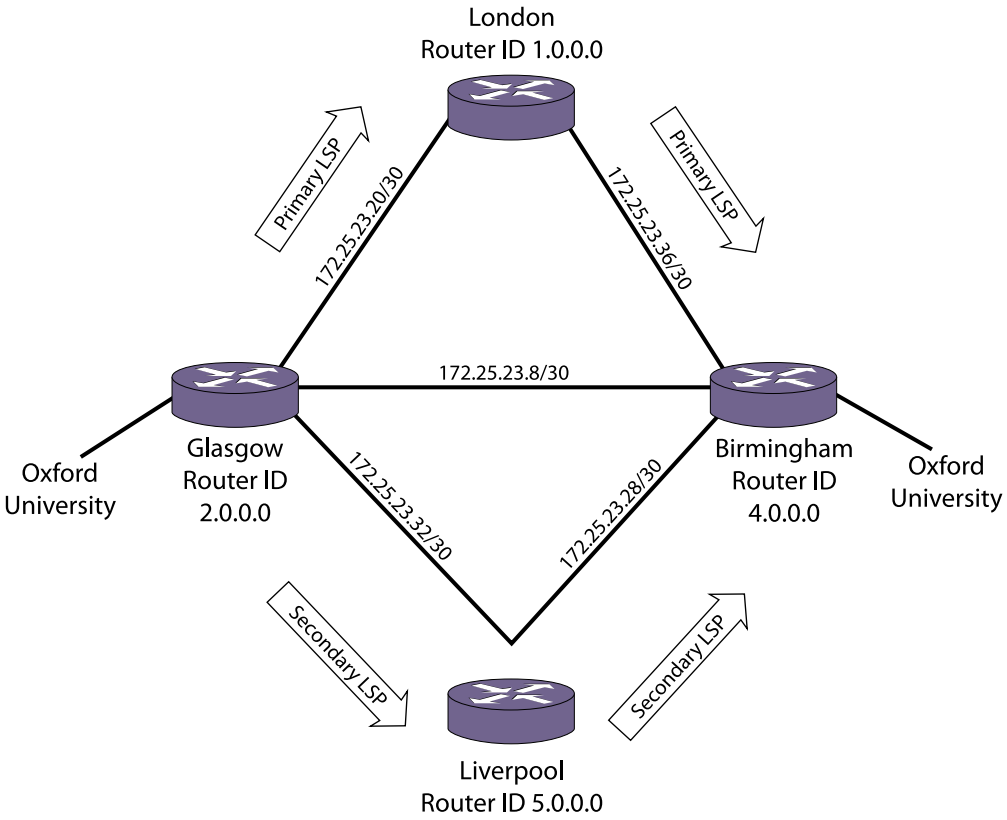



**Figure 201: RSVP-TE Configuration Example**

The configuration example, shown in the following figure, creates primary and secondary LSPs between the node Glasgow and the node Birmingham. The steps specifically create an LSP between Glasgow and Birmingham based on an explicitly routed path via London with bandwidth, setup priority, and hold priority profile requirements. A secondary path is also created which, in the event of failure of a link or node on the primary path, activates the secondary LSP from Glasgow to Liverpool to Birmingham.

Note

Before configuring RSVP-TE LSPs, you need to enable the protocol on the switch, and an initial step of adding RSVP-TE to a *VLAN* must be carried out for all VLANs over which the user wishes RSVP-TE LSPs to be signaled. This is a one-time operation.

A loopback VLAN with the LSR-ID should be added to MPLS to allow RSVP-TE LSPs to be established to the LSR-ID.

The following commands configure RSVP-TE for the switch and add RSVP signaling capabilities to the specified VLANs:

```
enable mpls
enable mpls protocol rsvp-te
configure mpls add vlan loopback
configure mpls add vlan gla-lon
enable mpls rsvp-te vlan gla-lon
enable mpls vlan gla-lon
configure mpls add vlan gla-liv
```

```
enable mpls rsvp-te vlan gla-liv
enable mpls vlan gla-liv
```

The following commands reserve bandwidth for RSVP-TE LSPs on these MPLS interfaces:

```
configure mpls rsvp-te bandwidth committed-rate 20 Mbps gla-lon
configure mpls rsvp-te bandwidth committed-rate 20 Mbps gla-liv
```

The following commands create and configure an LSP profile named Glasgow-Birmingham-pro.

LSPs that use the Glasgow-Birmingham-pro profile are signaled with a reserved bandwidth of 10 Mbps and an LSP setup and hold priority of 5.

```
create mpls rsvp-te profile Glasgow-Birmingham-pro
configure mpls rsvp-te profile Glasgow-Birmingham-pro bandwidth committed-rate 10 m
configure mpls rsvp-te profile Glasgow-Birmingham-pro setup-priority 5 hold-priority 5
```

The following commands define the primary and secondary paths between Glasgow and Birmingham:

```
create mpls rsvp-te path Glasgow-Birmingham-pri-path
create mpls rsvp-te path Glasgow-Birmingham-sec-path
```

The following commands pin each path to an LSR, such that each path takes a different route to the endpoint 4.0.0.0.

Path Glasgow-Birmingham-pri-path is routed through LSR 1.0.0.0 and path Glasgow-Birmingham-sec-path is routed through LSR 5.0.0.0.

```
configure mpls rsvp-te path Glasgow-Birmingham-pri-path add ero 1.0.0.0/32 loose
configure mpls rsvp-te path Glasgow-Birmingham-sec-path add ero 5.0.0.0/32 loose
```

The following commands create one RSVP-TE LSP with one primary and one secondary or backup path.

Each path uses the same profile.

```
create mpls rsvp-te lsp Glasgow-Birmingham-lsp destination 4.0.0.0
configure mpls rsvp lsp Glasgow-Birmingham-lsp add path Glasgow-Birmingham-pri-path
profile Glasgow-Birmingham-pro primary
configure mpls rsvp lsp Glasgow-Birmingham-lsp add path Glasgow-Birmingham-sec-path
profile Glasgow-Birmingham-pro secondary
```

Note

The secondary LSP is signaled, however it remains in a standby state unless the primary path becomes unavailable.

By default, a VPLS pseudowire flows over any available LSP.

However, a VPLS pseudowire can be specifically directed to use a configured RSVP-TE based LSP. Configuration is no different from configuring an LDP-based VPLS pseudowire, except that the RSVP-TE LSP is explicitly specified. The following command specifically directs a VPLS pseudowire to use a previously configured RSVP-TE LSP:

```
configure vpls Glasgow-Birmingham-cust1 peer 4.0.0.0 add mpls lsp Glasgow-Birmingham-
lsp
```

# Troubleshooting MPLS

The ExtremeXOS software includes multiple mechanisms for detecting and reporting problems.

Many failures generate an *SNMP* trap or log an error message. To find out more about a problem, you can enter show commands and review the flag states in the command output. The software also includes some *MPLS* troubleshooting tools, which are described in the following sections.

## Using LSP Ping

To assist with problem determination in an *MPLS* network, LSP ping support is included as a CLI ping option in the ExtremeXOS software. The LSP ping support is based on draft-ietf-mpls-lsp-ping-13.txt. This draft includes support for both connectivity verification and fault isolation for transport LSPs. Connectivity verification is supported using a modified ping packet that is sent over the specified transport LSP.

LSP ping is designed to catch failures where a transport LSP appears to be operational but is actually not functioning correctly. LSP data plane corruption is far less likely to occur than an LSP control plane failure, but the LSP ping is also useful for detecting possible latency issues.

- To send MPLS ping packets over an LSP, enter the following command:

  ping mpls lsp [*lsp_name* | **any** *host* | **prefix** *ipNetmask*] {**reply-mode** [**ip** | **ip-router-alert**]} {**continuous** | **count** *count*} {**interval** *interval*} {**start-size** *start-size* {**end-size** *end-size*}} {**ttl** *ttl*} {{**from** *from*} {**next-hop** *hopaddress*}}

  MPLS pings are sent to the well-known UDP port number 3503 with an IP in the 127.0.0.0/8 IP subnet. The source IP address is set to the sender.

  The time stamp field is supported for calculating round trip times and is accurate to 1/100 of a second. When replying to a ping, the LSP ping response (MPLS echo reply) sequence number and time-stamp fields are set to the LSP ping request (MPLS echo request) values. One MPLS echo response is sent for each MPLS echo request received. An MPLS echo reply is sent out-of-band as a natively IP routed IPv4 UDP packet. The normal IP routed path might or might not use an LSP.

  To reduce the possibility of fragmentation problems on the return path, MPLS echo reply packets do not include any padding that was sent in the MPLS echo request. Because each LSP is unidirectional, the return path is not directly relevant for verification of the LSP's functionality. What is important is that the LSP ping results are returned to the source of the MPLS echo request.

## Using LSP Trace

Transport LSP fault isolation is supported using the LSP trace feature. When the control plane detects a transport LSP failure, the LSR can switch to another LSP if one is available. When the failure is detected by LSP ping, you can use LSP trace to try to isolate the fault.

- To start an LSP trace, use the following command:

  traceroute mpls lsp [*lsp_name* | **any** *host* | **prefix** *ipNetmask*] {**reply-mode** [**ip** | **ip-router-alert**]} {{**from** *from*} {**ttl** *ttl*} {**next-hop** *hopaddress*}}

## Using the Health Check VCCV Feature

Health check Virtual Circuit Connectivity Verification (VCCV) can be used as a network diagnostic tool or as a network fault-alert tool, and can be configured on up to 16 VPLS domains. Connectivity between VPLS peers can be verified using the health check VCCV feature, which is defined in RFC 5085, Pseudowire Virtual Circuit Connectivity Verification (VCCV).

This implementation uses the following components defined in that RFC:

- VCCV Control Channel Type: *MPLS* Router Alert Label
- VCCV Control Verification Type: LSP Ping

Health check uses the LSP ping capability to verify connectivity. When the PW is set up, the two peers negotiate the VCCV capabilities, and if they establish a common set, health check becomes operational. If the VCCV capabilities do not match, health check cannot operate.

Health check operates in a single direction, so health checking should be enabled on the LSRs at both ends of the pseudowire in order to verify that traffic can flow bi-directionally between two VPLS peers. For multi-peer full-mesh or hierarchical VPLS networks, VCCV should be enabled on all VPLS peers to verify the entire VPLS network.

VCCV sends health check packets at regular intervals (the default interval is 5 seconds). If health check reaches the threshold for missed responses (the default fault-multiplier is 4), health check logs a message in *EMS* at the Warning level. Note that this log is not seen with the default log settings and no *SNMP* traps are sent. A health check failure does not change the state of the PW, which could remain operationally up and continue to support traffic, depending on the actual problem.

- To enable health check, use the following commands:

  `enable l2vpn [`**`vpls`** *`vpls_name`* `|` **`vpws`** *`vpws_name`*`]` **`health-check vccv`**

- To configure health check, use the following command:

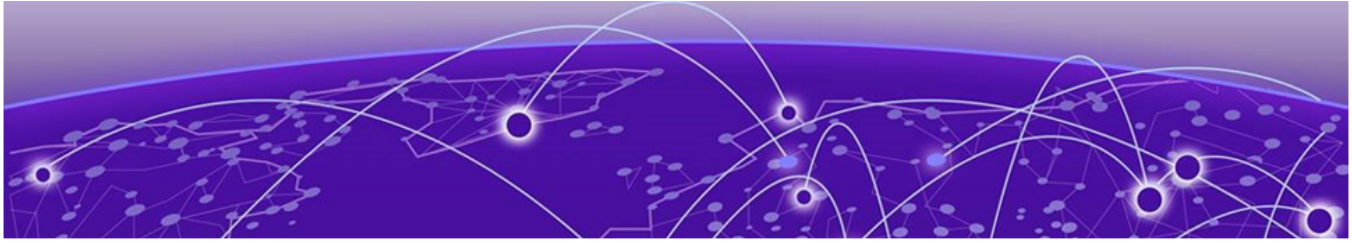
  `configure l2vpn [`**`vpls`** `[`*`vpls_name`* `|` **`all`**`] |` **`vpws`** `[`*`vpws_name`* `|` **`all`**`]]` **`health-check vccv`** `{`**`interval`** *`interval_seconds`*`}` `{`**`fault-multiplier`** *`fault_multiplier_number`*`}`

- View VPLS configuration information, including the health check feature configuration.

  `show [ {`**`l2vpn`**`} vpls {`*`vpls_name`*`} |` **`l2vpn vpws`** `{`*`vpws_name`*`} |` **`l2vpn`** `] {`**`peer`** *`ipaddress`*`} {`**`detail`**`} |` **`summary sharing`** `}`

- To disable health check, use the following command:

  `disable l2vpn [`**`vpls`** `[`*`vpls_name`* `|` **`all`**`] |` **`vpws`** `[`*`vpws_name`* `|` **`all`**`]]` **`health-check vccv`**



# IPv4 Unicast Routing



This chapter assumes that you are already familiar with IP unicast routing. If not, refer to the following publications for additional information:

- RFC 1256—*ICMP (Internet Control Message Protocol)* Router Discovery Messages
- RFC 1812—Requirements for IP Version 4 Routers

Note

For more information on interior gateway protocols, see:

- RIP on page 1330
- OSPF on page 1341
- IS-IS on page 1368

For information on exterior gateway protocols, see BGP on page 1389.

For more information on switch support for IPv6, IPv6 Unicast Routing on page 1294.

## IPv4 Unicast Overview

The switch provides full Layer 3, IPv4 unicast routing to all switches that run the Edge, Advanced Edge, and Core licenses (see the *Feature License Requirements* document.). It exchanges routing information with other routers on the network using one of the following routing protocols:

- Routing Information Protocol (*RIP (Routing Information Protocol)*).
- *OSPF (Open Shortest Path First)*.

- *BGP (Border Gateway Protocol)*.
- Intermediate System-Intermediate System (ISIS).

The switch dynamically builds and maintains a set of routing tables and determines the best path for each of its routes. Each host using the IP unicast routing functionality of the switch must have a unique IP address assigned. In addition, the default gateway assigned to the host must be the IP address of the router interface.

The ExtremeXOS software can provide both IPv4 and IPv6 routing at the same time. Separate routing tables are maintained for the two versions. Most commands that require you to specify an IP address can accept either an IPv4 or IPv6 address and act accordingly. Additionally, many of the IP configuration, enabling, and display commands have added tokens for IPv4 and IPv6 to clarify the version required. For simplicity, existing commands affect IPv4 by default and require you to specify IPv6, so configurations from an earlier release still correctly configure an IPv4 network.

## Router Interfaces

The routing software and hardware routes IP traffic between router interfaces. A router interface is simply a virtual LAN (*VLAN (Virtual LAN)*) that has an IP address assigned to it.

As you create VLANs with IP addresses belonging to different IP subnets, you can also choose to route between the VLANs. Both the VLAN switching and IP routing function occur within the switch.

Note

Each IP address and mask assigned to a VLAN must represent a unique IP subnet. You cannot configure the same IP address and subnet on different VLANs.

The figure below shows an example BlackDiamond switch configuration with two VLANs defined; Finance and Personnel. All ports on slots 1 and 3 are assigned to Finance, and all ports on slots 2 and 4 are assigned to Personnel. The figure shows the subnet address and interface address for each VLAN. Traffic within each VLAN is switched using the Ethernet MAC addresses. Traffic between the two VLANs is routed using the IP addresses.

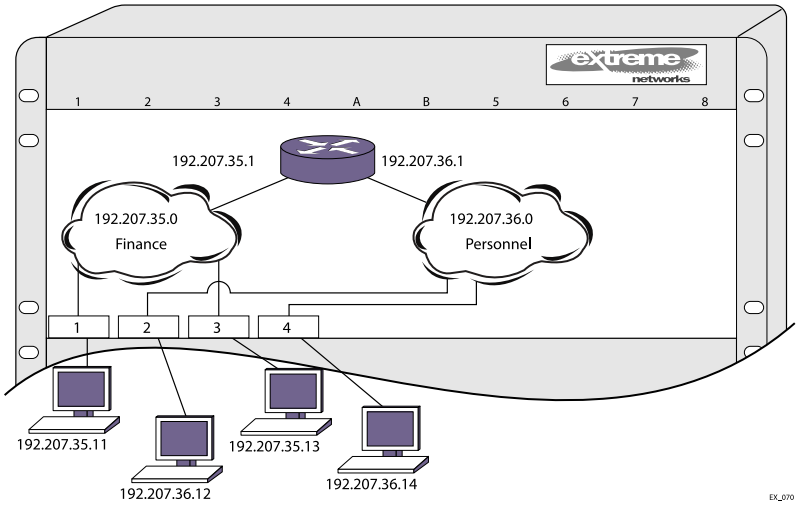



**Figure 202: Routing Between VLANs**

## GRE Tunnel

ExtremeXOS 15.4 and above supports creating a GRE-based IPv4 tunnel, and routing IPv4 traffic over it. This feature is supported on all platforms that have GRE tunneling support.

The switch administrator can configure a GRE tunnel by supplying the local source IPv4 address and the destination IPv4 address. Once configured, traffic can be redirected through the GRE tunnel. The TTL value of the outer IPv4 header will be set to 255 and the DSCP value is copied from the inner IPv4 header, the same as for the IPv6 tunnels. The encapsulated packets do not include the GRE checksum option, however if received with a checksum they are verified by the software, and then dropped if incorrect. The GRE module is capable of dealing with RFC 1701 neighbor options, with exception of the router option. Packets with this option set are dropped. However hardware does not support any options in the GRE header. If any of these options are set, the packet is either dropped, or sent to the CPU for processing. Since the key option of GRE tunnel is not configured, the GRE module only accepts GRE packets with a key value of 0, if present, and drops packets with other key values.

In ExtremeXOS Release 15.5, the following hardware is supported:

- Summit X460, X460-G2 X480, X670, X770, and E4G.
- SummitStack
- BD8900 (G96T-c, 10G24X-c, G48T-XL, G48X-XL 10G8X-XL, 40G6X-xm)
- BDX8 (all I/O cards)

From ExtremeXOS Release 15.6, the Summit X460-G2 and X670-G2 series is supported.

Note

GRE tunnels are IP tunnels which require L3 Function. L3 features are supported with EDGE license and above. All of the supported platforms' default license is EDGE or above, which include L3 features. In a stack all of the nodes must be GRE capable. For GRE in a stack, all stack nodes must be GRE hardware capable.

All blades in the chassis, or nodes in a stack need to support GRE tunnels, or else the feature cannot be configured/enabled. When all blades/nodes in the system are capable of running GRE, new GRE tunnels can be created. If a new blade is added to a chassis that is not capable of running GRE the blade will not be brought up. The “show slot” command displays this. If a new node is added to a stack, it will be powered on, and a log message is logged that the node is not compatible with GRE, and should be removed. This is done to prevent the node from continuously rebooting. If a system boots up with both a GRE configuration, and an incapable blade/node in the system, all GRE tunnels we be put in system disabled state. The “show tunnels” command displays this.

### *GRE Tunnel Example Configuration*

This example shows how the GRE tunnel feature could be configured.

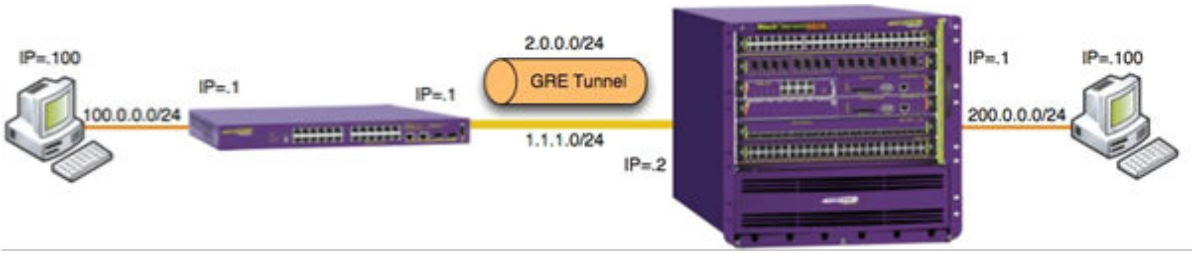



**Figure 203: GRE Tunnel Configuration**

**Summit Configuration**

```
configure default del port all
create vlan inet
configure vlan inet add port 24
configure vlan inet ipa 1.1.1.1/24
create vlan users
configure vlan users add port 1
configure vlan users ipa 100.0.0.1/24
create tunnel mytunnel gre destination 1.1.1.2 source 1.1.1.1
configure tunnel "mytunnel" ipaddress 2.0.0.1/24
configure iproute add 200.0.0.0/24 2.0.0.2
enable ipforwarding
```

**BlackDiamond 8800 Configuration**

```
configure default del port all
create vlan inet
configure vlan inet add port 10:1
configure vlan inet ipa 1.1.1.2/24
create vlan users
configure vlan users add port 10:2
configure vlan users ipa 200.0.0.1/24
create tunnel mytunnel gre destination 1.1.1.1 source 1.1.1.2
configure tunnel "mytunnel" ipaddress 2.0.0.2/24
configure iproute add 100.0.0.0/24 2.0.0.1
enable ipforwarding
```

## Populating the Routing Tables

The switch maintains a set of IP routing tables for both network routes and host routes. Some routes are determined dynamically from routing protocols, and some routes are manually entered. When multiple routes are available to a destination, configurable options such as route priorities, route sharing, and compressed routes are considered when creating and updating the routing tables.

### *Dynamic Routes*

Dynamic routes are typically learned by enabling the *RIP*, *OSPF*, IS-IS or *BGP* protocols, and are also learned from *ICMP* redirects exchanged with other routers. These routes are called dynamic routes because they are not a permanent part of the configuration. The routes are learned when the router starts up and are dynamically updated as the network changes. Older dynamic routes are aged out of the tables when an update for the network is not received for a period of time, as determined by the routing protocol.

Once a routing protocol is configured, dynamic routes require no configuration and are automatically updated as the network changes.

### *Static Routes*

Static routes are routes that are manually entered into the routing tables and are not advertised through the routing protocols. Static routes can be used to reach networks that are not advertised by routing protocols and do not have dynamic route entries in the routing tables. Static routes can also be used for security reasons, to create routes that are not advertised by the router.

Static routes are configured in the ExtremeXOS software, remain part of the configuration when the switch is rebooted, and are immediately available when the switch completes startup. Static routes are never aged out of the routing table, however, the Bidirectional Forwarding Detection (BFD) feature, can be used to bring down static routes when the host link fails.

Without BFD, static routes always remain operationally active because there is no dynamic routing protocol to report network changes. This can lead to a black hole situation, where data is lost for an indefinite duration. Because upper layer protocols are unaware that a static link is not working, they cannot switch to alternate routes and continue to use system resources until the appropriate timers expire.

With BFD, a static route is marked operationally inactive if the BFD session goes down. Upper layer protocols can detect that the static route is down and take the appropriate action.

A default route is a type of static route that identifies the default router interface to which all packets are routed when the routing table does not contain a route to the packet destination. A default route is also called a default gateway.

#### ExtremeXOS Resiliency Enhancement for IPv4 Static Routes

The ExtremeXOS Resiliency Enhancement feature provides a resilient way to use *ECMP (Equal Cost Multi Paths)* to load balance IPv4 traffic among multiple servers or other specialized devices. ExtremeXOS automatically manages the set of active devices using ECMP static routes configured with ping protection to monitor the health of these routes. Such servers or specialized devices do not require special software to support Bidirectional Forwarding Detection (BFD), or IP routing protocols such as *OSPF*, or proprietary protocols to provide keepalive messages. ExtremeXOS uses industry-standard and required protocols *ICMP*/ARP for IPv4 to accomplish the following automatically:

- Initially verify devices and activate their static routes, without waiting for inbound user traffic, and without requiring configuration of device MAC addresses.
- Detect silent device outages and inactivate corresponding static routes.
- Reactivate static routes after device recovery, or hardware replacement with a new MAC address.

ExtremeXOS currently supports similar protection and resiliency using Bidirectional Forwarding Detection (BFD) on IPv4 static routes. However, BFD can only be used when the local and remote device both support BFD.

### *Multiple Routes*

When there are multiple, conflicting choices of a route to a particular destination, the router picks the route with the longest matching network mask. If these are still equal, the router picks the route using the following default criteria (in the order specified):

- Directly attached network interfaces
- Static routes
- *ICMP* redirects

- Dynamic routes
- Directly attached network interfaces that are not active.

You can also configure blackhole routes—traffic to these destinations is silently dropped.

The criteria for choosing from multiple routes with the longest matching network mask is set by choosing the relative route priorities.

### *Relative Route Priorities*

The following table lists the relative priorities assigned to routes depending on the learned source of the route.

**Note**
You can change the order of the relative priorities, but we recommend that you only make changes if you are aware of the possible consequences.

**Table 128: Relative Route Priorities**

| Route Origin | Priority |
|---|---|
| Direct | 10 |
| *MPLS (Multiprotocol Label Switching)* | 20 |
| BlackHole | 50 |
| Static | 1100 |
| *ICMP* | 1200 |
| EBGP | 1700 |
| IBGP | 1900 |
| OSPFIntra | 2200 |
| OSPFInter | 2300 |
| IS-IS | 2350 |
| IS-IS L1 | 2360 |
| IS-IS L2 | 2370 |
| *RIP* | 2400 |
| OSPFASEXT | 3100 |
| OSPFExtern1 | 3200 |
| OSPFExtern2 | 3300 |
| IS-IS L1Ext | 3400 |
| IS-IS L2Ext | 3500 |
| BOOTP | 5000 |

### *IP Route Sharing and ECMP*

IP route sharing allows a switch to communicate with a destination through multiple equal-cost routes. In *OSPF*, *BGP*, and IS-IS, this capability is referred to as *ECMP* routing.

Without IP route sharing, each IP route entry in the routing tables lists a destination subnet and the next-hop gateway that provides the best path to that subnet. Every time a packet is forwarded to a particular destination, it uses the same next-hop gateway.

With IP route sharing, an additional ECMP table lists up to 2, 4, 8, 16, 32, or 64 next-hop gateways (depending on the platform and feature configuration) for each route in the routing tables. When multiple next-hop gateways lead to the same destination, the switch can use any of those gateways for packet forwarding. IP route sharing provides route redundancy and can provide better throughput when routes are overloaded.

The gateways in the ECMP table can be defined with static routes (up to 64-way), or they can be learned through the OSPF, BGP, or IS-IS protocols (16-way for OSPFv2 and *OSPFv3 (Open Shortest Path First version 3)*, and 8-way for BGP and IS-IS). For more information on the ECMP table, see ECMP Hardware Table on page 1263.

Note

BGP does not use ECMP by default, so if you require that functionality you must explicitly issue the command `configure bgp maximum-paths max-paths` with a value greater than 1.

### *Compressed Routes*

Compressed routes allow you to reduce the number of routes that are installed in the hardware routing tables. The switch uses hardware routing tables to improve packet forwarding performance. The switch can use both hardware and software to forward packets, but packet forwarding without software processing is faster. The hardware routing tables have less storage space than the software, so compressed routes conserve resources and improve scaling

The compressed route feature allows you to install less specific routes in the table, when overlapping routes with same next-hop exist. This route pruning technique is implemented as part of the Route Manager (RtMgr) process.

When a route is added, deleted or updated, the pruning algorithm is applied to see if the new route and/or its immediate children can be compressed or uncompressed as follows:

- If the parent node (immediate less specific route) of the newly added IP prefix has the same gateway as the new IP prefix, the newly added prefix is compressed.
- If the gateways of the newly added IP prefix and its immediate children are the same, the child nodes are compressed.
- If the gateways of the newly added IP prefix and its immediate children are not the same, and the child nodes had been previously compressed, the child nodes are uncompressed.

#### Event Log Messages

Event log messages are given in the following circumstances:

- When compression or uncompression start and end.

```
[ Severity level: Debug -Summary ]
```

- During each chunking start and end

```
    [ Severity level: Debug -Verbose ]
```

- When a route is compressed or uncompressed.

```
    [ Severity level: Debug -Verbose ]
```

#### Exceptional Scenarios

This section explains instances of exceptional route compression behavior.

- When a node does not have any best route.

  Consider the routing table shown in Table 129. When a node does not have any best route, children are uncompressed, if they were already compressed. Also this node is uncompressed, if it had previously been compressed.

**Table 129: Route Manager's Table When There is No Best Route for a Node**

| Prefix | Gateway | Number of best paths | Compressed? |
|---|---|---|---|
| 192.0.0.0/8 | 10.203.174.68 | 1 | No |
| 192.168.0.0/16 | 10.203.174.68 | 0 | No |
| 192.168.224.0/24 | 10.203.174.68 | 1 | No |
| 192.168.225.0/24 | 10.203.174.68 | 1 | No |

- When a node contains only a multicast route.

  Route compression is applied to unicast routes only. If a node contains only a multicast route, the compression algorithm is not applied to the node. Therefore multicast nodes are considered as nodes with no best unicast routes as shown in Table 129.

**Table 130: Route Manager's Table When a Node Contains Only a Multicast Route**

| Prefix | Gateway | Unicast/Multicast | Compressed? |
|---|---|---|---|
| 192.0.0.0/8 | 10.203.174.68 | Unicast Route | No |
| 192.168.0.0/16 | 10.203.174.68 | Multicast Route | No |
| 192.168.224.0/24 | 10.203.174.68 | Unicast Route | No |
| 192.168.225.0/24 | 10.203.174.68 | Unicast Route | No |

#### ECMP Handling When IP Route Sharing Is Enabled

The nodes that have *ECMP* table entries are compressed only if the following conditions are met; otherwise, potential sub-optimal forwarding occurs:

- The number of ECMP gateways for a given node must match the number of ECMP gateways in its parent node.

- A given node's set of gateways must match its parent's set of gateways.

The following table shows how compression is applied for the nodes with ECMP table entries when IP route sharing is enabled. Sample routes with ECMP table entries are taken for illustration. The Reason field in the table provides information about why the compression is applied or not applied for the node.

**Table 131: Route Manager's Table When IP Route Sharing is Enabled**

| Prefix | Gateways | Compressed? | Reason |
|---|---|---|---|
| 20.0.0.0/8 | Gw1: 30.1.10.1,<br>Gw2: 50.1.10.1 | NO | This is the top node. |
| 20.1.10.0/24 | Gw1: 30.1.10.1 | NO | Number of gateways did not match. This node has only one gateway, while the parent node has two. |
| 20.2.10.0/24 | Gw1: 30.1.10.1,<br>Gw2: 60.1.10.1 | NO | Number of gateways match. But one of the ECMP paths (gateway 60.1.10.1) does not match with its parent's ECMP paths. |
| 20.3.10.0/24 | Gw1: 30.1.10.1,<br>Gw2: 50.1.10.1 | YES | Number of gateways matches with its parent. Also all the gateways match with parent. |
| 20.4.10.0/24 | Gw1: 30.1.10.1,<br>Gw2: 50.1.10.1<br>Gw3: 60.1.10.1 | NO | Number of gateways does not match with its parent. |
| 20.1.10.44/32 | Gw1: 30.1.10.1<br>Gw2: 50.1.10.1 | NO | Number of gateways did not match. [This node has ECMP table entries, but parent node 20.1.10.0 does not have an ECMP table entry.] |

The following table shows only uncompressed routes.

**Table 132: HAL(TCAM)/Kernel Routing Table When IP Route Sharing is Enabled**

| Prefix | Gateway |
|---|---|
| 20.0.0.8/16 | Gw1: 30.1.10.1, Gw2: 50.1.10.1 |
| 20.1.10.0/24 | Gw1: 30.1.10.1 |
| 20.2.10.0/24 | Gw1: 30.1.10.1, Gw2: 60.1.10.1 |
| 20.4.10.0/24 | Gw1: 30.1.10.1, Gw2: 50.1.10.1 Gw3: 60.1.10.1 |
| 20.1.10.44/32 | Gw1: 30.1.10.1 Gw2: 50.1.10.1 |

Sample output is shown below:

```
* (debug) BD-12804.9 # enable iproute sharing
* (debug) BD-12804.10 # show iproute
Ori Destination      Gateway       Mtr  Flags       VLAN   Duration
#s  Default Route    12.1.10.10    1    UG---S-um--f v1     0d:20h:1m:3s
#s  Default Route    12.1.10.12    1    UG---S-um--f v1     0d:19h:14m:58s
#d  12.1.10.0/24     12.1.10.62    1    U------um--f v1     0d:20h:1m:3s
d   16.1.10.0/24     16.1.10.62    1    -------um--- v16    0d:20h:1m:4s
#d  22.1.10.0/24     22.1.10.62    1    U------um--f v2     0d:20h:1m:4s
#s  33.33.33.0/24    12.1.10.25    1    UG---S-um--f v1     0d:20h:1m:3s
#s  55.0.0.0/8       12.1.10.10    1    UG---S-um--f v1     0d:20h:1m:3s
```

```
#s  55.0.0.0/8       22.1.10.33    1    UG---S-um--f v2    0d:20h:1m:3s
#s  55.2.1.1/32      12.1.10.22    1    UG---S-um--f v1    0d:20h:1m:3s
#s  55.5.5.1/32      12.1.10.44    1    UG---S-um--f v1    0d:20h:1m:3s
#s  66.0.0.0/8       12.1.10.12    1    UG---S-um--f v1    0d:20h:1m:3s
#s  66.0.0.0/16      12.1.10.12    1    UG---S-um--c v1    0d:20h:1m:3s
#d  70.1.10.0/24     70.1.10.62    1    U------um--f v7    0d:20h:1m:4s
#s  78.0.0.0/8       12.1.10.10    1    UG---S-um--f v1    0d:20h:1m:3s
#s  79.0.0.0/8       12.1.10.10    1    UG---S-um--c v1    0d:20h:1m:3s
#s  79.0.0.0/8       12.1.10.12    1    UG---S-um--c v1    0d:20h:1m:3s
#s  80.0.0.0/8       12.1.10.10    1    UG---S-um--f v1    0d:20h:1m:3s
#d  80.1.10.0/24     80.1.10.62    1    U------um--f v8    0d:20h:1m:4s
#s  81.0.0.0/8       12.1.10.10    1    UG---S-um--f v1    0d:20h:1m:3s
#s  81.0.0.0/8       12.1.10.12    1    UG---S-um--f v1    0d:20h:1m:3s
#s  81.0.0.0/8       12.1.10.13    1    UG---S-um--f v1    0d:20h:1m:3s
#s  82.0.0.0/8       12.1.10.10    1    UG---S-um--f v1    0d:20h:1m:3s
#s  83.0.0.0/8       12.1.10.10    1    UG---S-um--f v1    0d:20h:1m:3s
#d  91.1.10.0/24     91.1.10.62    1    U------um--f v9    0d:20h:1m:4s
#d  92.1.10.0/24     92.1.10.62    1    U------um--f v10   0d:20h:1m:6s
#d  93.1.10.0/24     93.1.10.62    1    U------um--f v11   0d:20h:1m:6s

Origin(Ori): (b) BlackHole, (be) EBGP, (bg) BGP, (bi) IBGP, (bo) BOOTP
(ct) CBT, (d) Direct, (df) DownIF, (dv) DVMRP, (e1) ISISL1Ext
(e2) ISISL2Ext, (h) Hardcoded, (i) ICMP, (i1) ISISL1 (i2) ISISL2
(is) ISIS, (mb) MBGP, (mbe) MBGPExt, (mbi) MBGPInter, (mp) MPLS Lsp
(mo) MOSPF (o) OSPF, (o1) OSPFExt1, (o2) OSPFExt2
(oa) OSPFIntra, (oe) OSPFAsExt, (or) OSPFInter, (pd) PIM-DM, (ps) PIM-SM
(r) RIP, (ra) RtAdvrt, (s) Static, (sv) SLB_VIP, (un) UnKnown
(*) Preferred unicast route (@) Preferred multicast route
(#) Preferred unicast and multicast route

Flags: (B) BlackHole, (D) Dynamic, (G) Gateway, (H) Host Route
(L) Matching LDP LSP, (l) Calculated LDP LSP, (m) Multicast
(P) LPM-routing, (R) Modified, (S) Static, (s) Static LSP
(T) Matching RSVP-TE LSP, (t) Calculated RSVP-TE LSP, (u) Unicast, (U) Up
(f) Provided to FIB (c) Compressed Route

Mask distribution:
2 default routes               12 routes at length 8
1 routes at length 16           9 routes at length 24
2 routes at length 32

Route Origin distribution:
8 routes from Direct              18 routes from Static

Total number of routes = 26
Total number of compressed routes = 3
```

#### ECMP Handling When IP Route Sharing Is Disabled

If IP route sharing is disabled, the first best route is installed in the hardware table, if multiple best routes are available. Hence the compression algorithm considers the first best route for *ECMP* cases. As shown in the following table, when IP route sharing is disabled, all routes are compressed, except the first one in this case.

**Table 133: Route Manager's Table When IP Route Sharing is Disabled**

| Prefix | Gateways | Compressed? |
|---|---|---|
| 20.0.0.0/8 | Gw1: 30.1.10.1, Gw2: 50.1.10.1 | NO |
| 20.1.10.0/24 | Gw1: 30.1.10.1 | YES |
| 20.2.10.0/24 | Gw1: 30.1.10.1, Gw2: 60.1.10.1 | YES |

**Table 133: Route Manager's Table When IP Route Sharing is Disabled (continued)**

| Prefix | Gateways | Compressed? |
|---|---|---|
| 20.3.10.0/24 | Gw1: 30.1.10.1, Gw2: 50.1.10.1 | YES |
| 20.4.10.0/24 | Gw1: 30.1.10.1, Gw2: 50.1.10.1<br>Gw3: 60.1.10.1 | YES |
| 20.1.10.44/32 | Gw1: 30.1.10.1 Gw2: 50.1.10.1 | YES |

**Table 134: HAL(TCAM)/Kernel Routing Table When IP Route Sharing is Disabled**

| Prefix | Gateway |
|---|---|
| 20.0.0.8/16 | Gw1: 30.1.10.1, |

Sample output is shown below:

```
* (debug) # disable iproute sharing
* (debug) # show iproute
Ori Destination        Gateway         Mtr  Flags        VLAN       Duration
#s  Default Route      12.1.10.10      1    UG---S-um--f v1         0d:19h:58m:58s
#s  Default Route      12.1.10.12      1    UG---S-um--- v1         0d:19h:12m:53s
#d  12.1.10.0/24       12.1.10.62      1    U------um--f v1         0d:19h:58m:59s
d  16.1.10.0/24       16.1.10.62      1    -------um--- v16        0d:19h:58m:59s
#d  22.1.10.0/24       22.1.10.62      1    U------um--f v2         0d:19h:58m:59s
#s  33.33.33.0/24      12.1.10.25      1    UG---S-um--f v1         0d:19h:58m:58s
#s  55.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#s  55.0.0.0/8         22.1.10.33      1    UG---S-um--- v2         0d:19h:58m:58s
#s  55.2.1.1/32        12.1.10.22      1    UG---S-um--f v1         0d:19h:58m:58s
#s  55.5.5.1/32        12.1.10.44      1    UG---S-um--f v1         0d:19h:58m:58s
#s  66.0.0.0/8         12.1.10.12      1    UG---S-um--f v1         0d:19h:58m:58s
#s  66.0.0.0/16        12.1.10.12      1    UG---S-um--c v1         0d:19h:58m:58s
#d  70.1.10.0/24       70.1.10.62      1    U------um--f v7         0d:19h:58m:59s
#s  78.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#s  79.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#s  79.0.0.0/8         12.1.10.12      1    UG---S-um--- v1         0d:19h:58m:58s
#s  80.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#d  80.1.10.0/24       80.1.10.62      1    U------um--f v8         0d:19h:58m:59s
#s  81.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#s  81.0.0.0/8         12.1.10.12      1    UG---S-um--- v1         0d:19h:58m:58s
#s  81.0.0.0/8         12.1.10.13      1    UG---S-um--- v1         0d:19h:58m:58s
#s  82.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#s  83.0.0.0/8         12.1.10.10      1    UG---S-um--c v1         0d:19h:58m:58s
#d  91.1.10.0/24       91.1.10.62      1    U------um--f v9         0d:19h:58m:59s
#d  92.1.10.0/24       92.1.10.62      1    U------um--f v10        0d:19h:59m:2s
#d  93.1.10.0/24       93.1.10.62      1    U------um--f v11        0d:19h:59m:2s

Origin(Ori): (b) BlackHole, (be) EBGP, (bg) BGP, (bi) IBGP, (bo) BOOTP
(ct) CBT, (d) Direct, (df) DownIF, (dv) DVMRP, (e1) ISISL1Ext
(e2) ISISL2Ext, (h) Hardcoded, (i) ICMP, (i1) ISISL1 (i2) ISISL2
(is) ISIS, (mb) MBGP, (mbe) MBGPExt, (mbi) MBGPInter, (mp) MPLS Lsp
(mo) MOSPF (o) OSPF, (o1) OSPFExt1, (o2) OSPFExt2
(oa) OSPFIntra, (oe) OSPFAsExt, (or) OSPFInter, (pd) PIM-DM, (ps) PIM-SM
(r) RIP, (ra) RtAdvrt, (s) Static, (sv) SLB_VIP, (un) UnKnown
(*) Preferred unicast route (@) Preferred multicast route
(#) Preferred unicast and multicast route

Flags: (B) BlackHole, (D) Dynamic, (G) Gateway, (H) Host Route
(L) Matching LDP LSP, (l) Calculated LDP LSP, (m) Multicast
(P) LPM-routing, (R) Modified, (S) Static, (s) Static LSP
```

```
(T) Matching RSVP-TE LSP, (t) Calculated RSVP-TE LSP, (u) Unicast, (U) Up
(f) Provided to FIB (c) Compressed Route

Mask distribution:
2 default routes               12 routes at length 8
1 routes at length 16           9 routes at length 24
2 routes at length 32
Route Origin distribution:
8 routes from Direct              18 routes from Static

Total number of routes = 26
Total number of compressed routes = 8
```

LSP Route Handling

An LSP route is compressed only in the following circumstances:

- If the parent node of the LSP route is also an LSP route
- If the LSP nexthop of the parent node matches with this node

## Hardware Routing Table Management

The switch hardware can route traffic based on information stored in hardware and software routing tables. Routing tasks are completed much faster when the routes are already stored in hardware routing tables. When packets are routed using the hardware routing tables, this is called fast-path routing, because the packets take the fast path through the switch.

The switch hardware provides the fast path, and the ExtremeXOS software provides the routing management capabilities, including management of the hardware routing tables. The software collects all routing information, manages the routes according to the switch configuration, and stores the appropriate routes in the hardware routing tables. When an IP unicast packet arrives and the destination IP address is not found in the hardware route tables, it is routed by software. The software processing takes more time than the fast path, so routing based on the software tables is called slow-path routing.

BlackDiamond X8 and 8000 series modules, SummitStack, and Summit X440, X460, X480, X670, X670-G2, and X770 switches allow you to customize the software management of the hardware routing tables using the following hardware components:

- Extended IPv4 Host Cache on page 1254
- ECMP Hardware Table on page 1263

### *Extended IPv4 Host Cache*

The extended IPv4 host cache feature provides additional, configurable storage space on L3-capable switches to store additional IPv4 hosts in the hardware routing tables. This feature is supported on all switches, except those with the L2 Edge license.

All switches, except those with the L2 Edge license, support slow-path routing (using software routing tables), so adding more entries in the hardware routing table is a performance feature, which allows more hosts to benefit from fast-path routing. To use the extended IPv4 host cache feature effectively, it

helps to understand how the hardware tables operate on the switches that support this feature. The hardware forwarding tables controlled by this feature store entries for the following:

- IPv4 local and remote hosts
- IPv4 routes
- IPv6 local hosts
- IPv6 routes
- IPv4 and IPv6 multicast groups

The extended IPv4 host cache feature works by customizing the forwarding table space allotted to these components.

#### Introduction to Hardware Forwarding Tables

The extended IPv4 host cache feature relates to the four hardware forwarding tables shown in the following figure.

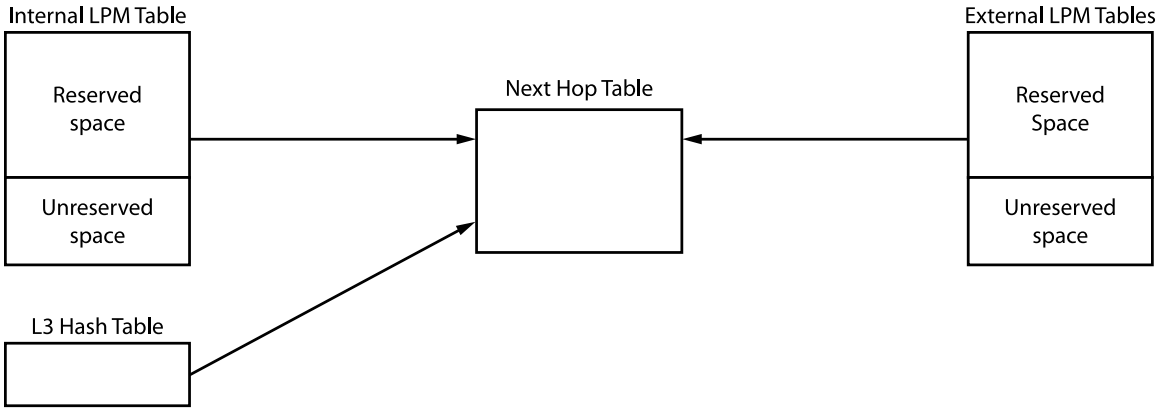



**Figure 204: Hardware Forwarding Tables**

The Longest Prefix Match (LPM) and Layer 3 (L3) Hash tables store host and route information for fast-path forwarding. When the switch locates a route or host in one of these tables, it follows a table index to the Next Hop table, which contains MAC address and egress port information that is shared by the hosts and routes in the other tables. The hardware routing table capacity is partly determined by the capacity of the Next Hop table. The Next Hop table capacity is smaller than the combined capacity of the other tables because typically, multiple routes and hosts share each Next Hop table entry. When the other tables map to many different next hop entries, the Next Hop table can limit the total number of hosts and routes stored in hardware.

On most platforms, the L3 Hash table is smaller than the LPM tables. Because the L3 Hash table is the only table that can store IPv4 and IPv6 multicast entries and IPv6 local hosts, and because of the way the L3 Hash table is populated, forwarding table capacity and forwarding performance can be improved by allocating space for storing IPv4 local and remote host entries in the LPM tables.

The extended IPv4 host cache feature specifically allows you to define the number of entries that are reserved in the LPM tables for IPv4 and IPv6 routes. The unreserved entries are available for IPv4 local and remote hosts. IPv4 hosts can also occupy unused areas of the L3 Hash table, and when necessary, unused space in the reserved section of the LPM tables. The maximum number of hosts that can be stored in the hardware routing tables depends on the configuration and usage of the tables, but the number of local IPv4 hosts and gateways is ultimately limited to the size of the Next Hop table minus three reserved entries.

#### LPM Table Management

The internal LPM tables are provided on all platforms. The external LPM tables are available only on BlackDiamond 8900 xl-series modules and Summit X480 switches, and they are supported only when external tables on those switches are configured to support external Layer 3 LPM entries. Because the external tables can be configured to support Layer 2 *FDB (forwarding database)* entries, Layer 3 LPM entries, or *ACL (Access Control List)* entries (or a combination of these), you must be aware of the external table configuration when managing LPM entries.

The ExtremeXOS software manages the content of the hardware tables based on the configuration specified by the following commands:

`configure iproute reserved-entries [`*num_routes_needed*`|`**maximum**`|`**default**`] `**slot** `[`**all**`|`*slot_num*`]`

`configure forwarding external-tables [`**l3-only** `{`**ipv4** `|` **ipv4-and-ipv6** `|` **ipv6**`}` `|` **l2-only** `|` **acl-only** `|` **l2-and-l3** `|` **l2-and-l3-and-acl** `|` **l2-and-l3-and-ipmc** `|` **none**`]`

The `configure iproute reserved-entries` command configures the LPM tables. The configure forwarding external-tables command is available only on BlackDiamond 8900 xl-series modules and Summit X480 switches, and configures the use of the external tables.

The `configure forwarding internal-tables [`**l2-and-l3** `|` **more** `[`**l2** `|` **l3-and-ipmc**`]]` command provides the ability to support additional IPv4 and IPv6 hosts and multicast table entries on Summit X450-G2, X460-G2, X670-G2, X770 and BlackDiamond X8 B-series switches.

#### IPv6 Routes and Hosts in External Tables

ExtremeXOS allows you to store IPv6 routes and hosts in external LPM tables. You can configure the external LPM to contain IPv4 or IPv6 routes, or both. Internal LPM tables can store IPv4 or IPv6 routes, both, or neither, based on the configuration setting for external-tables.

The `configure forwarding external-tables l3-only` command using the **ipv6** and **ipv4-and-ipv6** variables supports larger IPv6 route and host scaling in external LPM tables.

When an external LPM table is configured for l3-only ipv6, no IPv6 routes or IPv6 hosts are stored in any of the internal hardware tables. This provides the highest IPv6 scale, and avoids contention with IP Multicast in the L3 Hash hardware table.

IPv6 hardware and slowpath forwarding are supported on user-created Virtual Routers, and IPv6 tunnels are only supported on *VR-Default*.

The size of the internal LPM tables, and the size of the L3 Hash and Next Hop tables are fixed for some platforms. The following tables list the hardware capacity for each of the tables shown in Figure 204 on page 1255

**Table 135: Hardware Routing Table Configuration Capacities, Platforms without External Tables**

| Table | BlackDiamond 8900 xm-Series Modules and Summit X670 Switches | BlackDiamond 8000 a- and c-Series Modules | BlackDiamond 8000 e-Series Modules | BlackDiamond X8 Series Switches | Summit X440 Switches | Summit X460 Switches | Summit X670 Switches |
|---|---|---|---|---|---|---|---|
| Internal LPM | 16352 | 12256 | 480 | 16352 | 32 | 12256 | 16352 |
| External LPM | N/A | N/A | N/A | N/A | N/A | N/A | N/N |
| L3 Hash | 8192 | 8192 | 2048 | 16384 | 512 | 16384 | 8192 |
| Next Hop | 8192 | 8192 | 2048 | 16384 | 512 | 16384 | 8192 |

**Table 136: Hardware Routing Table Configuration Capacities, Platforms with External Tables**

| | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches [a] | | | | | |
|---|---|---|---|---|---|---|
| Table | l2-only, acl-only, and none Options | l2-and-l3-and-acl and l2-and-l3-and-ipmc Options | l2-and-l3 Option | l3-only Option | l3-only ipv4-and-ipv6 Option | l3-only ipv6 Option |
| Internal LPM IPv4 | 16352 | N/A | N/A | N/A | N/A | 16352[c] |
| Internal LPM IPv6 | 8176 | 8192[b] | 8192[b] | 8192[b] | N/A | N/A |
| External LPM IPv4 | N/A | 131040 | 262112 | 524256 | 475104 | N/A |
| External LPM IPv6 | N/A | N/A | N/A | N/A | 49152 | 245760 |
| L3 Hash | 16384 | 16384 | 16384 | 16384 | 16384 | 16384 |
| Next Hop | 16384 | 16384 | 16384 | 16384 | 16384 | 16384 |

a These platforms use additional external LPM tables and the actual value depends on the configuration set with the configure forwarding external-tables command.

b In this configuration, the internal LPM table stores only IPv6 routes. All IPv4 routes are stored in the external LPM tables.

c In this configuration, the internal LPM table stores only IPv4 routes. All IPv6 routes are stored in the external LPM tables.

The Summit X670-G2, X770 and BlackDiamond X8-100G4X have hardware forwarding tables internal to the switch chips that can be partitioned in a flexible manner. The Summit X670-G2 and X770 switches have the following configurable internal tables:

**Table 137: Summit X670-G2 and X770 Hardware Routing Table Configuration Capacities for Platforms with Configurable L2/L3 Internal Tables**

| L3 Characteristic | l2-and-l3 | more l3-and-ipmc | more l2 |
|---|---|---|---|
| L3 Hash IPv4 Unicast | 96K | 128K | 16K |
| L3 Hash IPv6 Unicast | 48K | 48K | 8K |
| Next Hop | 48K | 48K | 48K |
| Internal IPv4 LPM | 16K | 16K | 16K |
| Internal IPv6 LPM | 8K | 8K | 8K |
| IPv4 hosts with min LPM routes (assumes 75% utilization of L3 hash table) | 82K | 106K | 28K |
| IPv4 hosts with max LPM routes (assumes 75% utilization of L3 hash table) | 72K | 96K | 12K |
| Remote IPv4 Host Entries (assumes 75% utilization of L3 hash table) | 124K | 172K | 28K |
| IPv6 Host Entries (assumes 75% utilization of L3 hash table) | 36K | 36K | 6K |
| IP multicast groups | 4K | 4K | 4K |
| IP-multicast (s,v,g) entries (will depend on hash utilization) | 72K | 104K | 16K |

The Summit X460-G2, has hardware forwarding tables internal to the switch chips that can be partitioned in a flexible manner. The Summit X460-G2 has the following configurable internal tables:

**Table 138: Summit X460-G2 Hardware Routing Table Configuration Capacities for Platforms with Configurable L2/L3 Internal Tables**

| L3 Characteristic | l2-and-l3 | more l3-and-ipmc | more l2 |
|---|---|---|---|
| L3 Hash IPv4 Unicast | 40K | 56K | 16K |
| L3 Hash IPv6 Unicast | 24K | 32K | 8K |
| Next Hop | 32K | 32K | 32K |
| Internal IPv4 LPM | 12K | 12K | 12K |
| Internal IPv6 LPM | 6K | 6K | 6K |

**Table 138: Summit X460-G2 Hardware Routing Table Configuration Capacities for Platforms with Configurable L2/L3 Internal Tables (continued)**

| L3 Characteristic | l2-and-l3 | more l3-and-ipmc | more l2 |
|---|---|---|---|
| IPv4 hosts with min LPM routes (assumes 75% utilization of L3 hash table) | 38K | 50K | 24K |
| IPv4 hosts with max LPM routes (assumes 75% utilization of L3 hash table) | 30K | 42K | 12K |
| Remote IPv4 Host Entries (assumes 75% utilization of L3 hash table) | 48K | 72K | 24K |
| IPv6 Host Entries (assumes 75% utilization of L3 hash table) | 18K | 24K | 6K |
| IP multicast groups | 4K | 4K | 4K |
| IP-multicast (s,v,g) entries (will depend on hash utilization) | 24K | 40K | 8K |

The BlackDiamond X8-100G4X switch has the following configurable internal tables:

**Table 139: BlackDiamond X8-100G4X Hardware Routing Table Configuration Capacities for Platforms with Configurable L2/L3 Internal Tables**

| L3 Characteristic | l2-and-l3 | more l3-and-ipmc | more l2 |
|---|---|---|---|
| L3 Hash IPv4 Unicast | 160K | 224K | 96K |
| L3 Hash IPv6 Unicast | 64K | 64K | 48K |
| Next Hop | 64K | 64K | 64K |
| Internal IPv4 LPM | 16K | 16K | 16K |
| Internal IPv6 LPM | 8K | 8K | 8K |
| IPv4 hosts with min LPM routes (assumes 75% utilization of L3 hash table) | 130K | 178K | 64K |
| IPv4 hosts with max LPM routes (assumes 75% utilization of L3 hash table) | 120K | 168K | 72K |
| Remote IPv4 Host Entries (assumes 75% utilization of L3 hash table) | 208K | 304K | 88K |

**Table 139: BlackDiamond X8-100G4X Hardware Routing Table Configuration Capacities for Platforms with Configurable L2/L3 Internal Tables (continued)**

| L3 Characteristic | l2-and-l3 | more l3-and-ipmc | more l2 |
|---|---|---|---|
| IPv6 Host Entries (assumes 75% utilization of L3 hash table) | 48K | 48K | 36K |
| IP multicast groups | 16K | 16K | 16K |
| IP-multicast (s,v,g) entries (will depend on hash utilization) | 64K | 64K | 16K |

In addition to configuring the number of reserved entries in the LPM tables, the `configure iproute reserved-entries` command configures which entries are stored in which tables. The following table shows the hardware routing table contents for several configurations.

**Table 140: Hardware Routing Table Contents**

| Table | All platforms except BlackDiamond 8900 xl- Series Modules and Summit X480 Switches | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Maximum IPv4 Capacity Configuration | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Maximum IPv6 Capacity Configuration | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Default Configuration |
|---|---|---|---|---|
| Internal LPM—Reserved space | Entries for IPv4 and IPv6 routes. | Entries for IPv6 routes.[d] | Entries for IPv4 routes. | Entries for IPv6 routes. |
| Internal LPM—Unreserved space | Entries for IPv4 local and remote hosts. | N/A[e] | Entries for IPv4 local and remote hosts. | N/A |
| External LPM—Reserved space | N/A | Entries for IPv4 routes. | N/A | Entries for IPv4 routes. |
| External LPM—Unreserved space | N/A | Entries for IPv4 local and remote hosts. | Entries for IPv6 routes. | Entries for IPv4 local and remote hosts. |

20 d

**Table 140: Hardware Routing Table Contents (continued)**

| Table | All platforms except BlackDiamond 8900 xl- Series Modules and Summit X480 Switches | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Maximum IPv4 Capacity Configuration | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Maximum IPv6 Capacity Configuration | BlackDiamond 8900 xl-Series Modules and Summit X480 Switches, Default Configuration |
|---|---|---|---|---|
| L3 Hash | Entries for IPv4 local and remote hosts, IPv4 and IPv6 multicast entries, and IPv6 local hosts.a | Entries for IPv4 local and remote hosts, IPv4 and IPv6 multicast entries, and IPv6 local hosts. | Entries for IPv4 local and remote hosts, and IPv4 and IPv6 multicast entries. | IPv4 and IPv6 multicast entries, and IPv6 local hosts |
| Next Hop | MAC address and egress port information for the entries in the LPM and L3 Hash tables. | MAC address and egress port information for the entries in the LPM and L3 Hash tables. | MAC address and egress port information for the entries in the LPM and L3 Hash tables. | MAC address and egress port information for the entries in the LPM and L3 Hash tables. |

d IPv6 routes and hosts consume two entries.

e In this configuration, all space in the internal LPM table is reserved for IPv6 routes.

#### Extended IPv4 Host Cache Management Guidelines

When configuring the extended IPv4 host cache feature, consider the following guidelines:

- The **default** option configures the switch to store entries for local and remote IPv4 hosts in the LPM tables. On BlackDiamond 8000 a-, c-, and, e--series modules and Summit X440, and X460 switches, the default setting creates room for 48 local and remote IPv4 host entries. On BlackDiamond 8900-40GX-xm modules and Summit X670 and X770 switches, the default setting creates room for 4112 local and remote IPv4 host entries. On BlackDiamond 8900 xl-series modules and Summit X480 switches, the default setting creates room for 16384 local and remote IPv4 host entries. This option provides more room for IPv4 and IPv6 multicast and IPv6 hosts in the L3 Hash table.
- The **maximum** option reserves all space in the LPM tables for IPv4 and IPv6 routes. This option provides the maximum storage for IPv4 and IPv6 routes when you do not expect to store many IPv4 and IPv6 multicast and IPv6 hosts in the L3 Hash table.
- On BlackDiamond 8900 xl-series modules and Summit X480 switches, the **default** and **maximum** options automatically select a configuration that is compatible with the configuration defined by the configure forwarding external-tables command.
- On BlackDiamond 8900 xl-series modules and Summit X480 switches, the number you specify for the **num_routes_needed** must be compatible with the configuration defined by the configure

forwarding external-tables command. If you specify a number that is greater than the number of routes specified by the current configuration, an error message appears.

Note

If no IPv4 route is found in the LPM table and IPv4 unicast packets are slow-path forwarded for a given remote host, an IPv4 entry is created for the remote host in either the L3 hash table or LPM table. The hardware does not cache entries for remote IPv6 hosts, so IPv6 routes take precedence over IPv4 routes when both IPv4 and IPv6 routes are stored in the Internal LPM table.

**IPv4 Host Entry Population Sequence**

The ExtremeXOS software populates the hardware tables with IPv4 host entries by searching for available space in the following sequence:

1. Unreserved space in the LPM tables.
2. Available space in an L3 Hash table bucket.
3. Available space in the reserved section of the LPM table.
4. Space used by the oldest host entries in the LPM and L3 Hash tables.

The L3 Hash table is named for the hash function, which stores host and multicast entries based on an algorithm applied to the host IP address or multicast tuple (Source IP, Group IP, *VLAN* ID). The hash table stores entries in groups of 8 or 16 (depending on the hardware), and these groups are called buckets. When a bucket is full, any additional host or multicast addresses that map or hash to that bucket cannot be added. Another benefit of the extended IPv4 host cache feature is that you can reduce these conflicts (or “hash table collisions”), by making room for IPv4 hosts in the LPM table and reducing demand for the L3 Hash table.

A hardware-based aging mechanism is used to remove any remote IPv4 host entries that have not had IPv4 unicast packets forwarded to them in the previous hour. (Note that remote IPv4 hosts only need to be cached when all IPv4 routes do not fit within the number of routes reserved.) Aging helps to preserve resources for the hosts that are needed most. In a BlackDiamond 8800 chassis or SummitStack, aging is performed independently for each I/O module or stack node based on the ingress traffic for that module or node. Depending on the IPv4 unicast traffic flows, independent IPv4 host caches for each I/O module or stack node can provide increased hardware fast-path forwarding compared with ExtremeXOS releases prior to 12.1. Even with aging, it is still possible that the Next Hop table, LPM table, or L3 Hash bucket do not have space to accept a new host. In those cases, a least-recently used algorithm is used to remove the oldest host to make space for the new host in hardware.

Local IPv4 host entries are only subject to hardware-based aging if there has been a large amount of least-recently used replacement, indicating severe contention for HW table resources. Otherwise, local IPv4 host entries are retained just as in ExtremeXOS releases prior to 12.1, based on whether IP ARP refresh is enabled or disabled, and the value for the configure iparp timeout command.

Note

Gateway entries are entries that represent routers or tunnel endpoints used to reach remote hosts. Gateway entries are not aged and are not replaced by IPv6 hosts or multicast entries in the L3 Hash table or by any entries requiring space in the Next Hop table. The software can move gateway entries from the LPM table to the L3 Hash table to make room for new reserved routes.

#### Calculating the Number of Routes Needed

Guidelines for calculating the number of routes to reserve are provided in the ExtremeXOS Command Reference description for the following command:

```
configure iproute reserved-entries [ num_routes_needed | maximum |
default ] slot [all | slot_num]
```

#### Coexistence of Higher- and Lower-Capacity Hardware

The BlackDiamond X8 BDXB-XL series, 8900 xl-series modules and Summit X480 switches are considered higher-capacity hardware because they provide external LPM tables, additional memory, and greater processing power, which allows this hardware to support a large number of IP routes. In comparison, other BlackDiamond X8 and 8000 series modules and Summit family switches are considered lower-capacity hardware.

The ExtremeXOS software supports the coexistence of higher- and lower-capacity hardware in the same BlackDiamond X8 or BlackDiamond 8800 chassis or Summit family switch stack. To allow for coexistence and increased hardware forwarding, when the number of IPv4 routes exceeds 25,000, the lower-capacity hardware automatically transitions from using LPM routing to forwarding of individual remote hosts, also known as IP *FDB* (IP FDB) mode. Higher-capacity hardware continues using LPM routing. Lower capacity hardware operating in IP FDB mode is indicated with a d flag in the output of show iproute reserved-entries statistics command, indicating that only direct routes are installed.

Note
If you require a large number of IPv6 routes, you should use only xl-series modules, or a Summit X480 standalone, or a SummitStack comprised only of the X480. SummitStacks, or a BD8800 containing a mix of high- and low-capability hardware (slots without External TCAM) does not support more than 100,000 IPv6 routes present.

### *ECMP Hardware Table*

IP route sharing and the *ECMP* hardware table are introduced in IP Route Sharing and ECMP on page 1248. The following sections provide guidelines for managing the ECMP hardware table:

- ECMP Table Configuration Guidelines
- Troubleshooting: ECMP Table-Full Messages on page 1264

Note
Summit X440 series switches do not support ECMP.

#### ECMP Table Configuration Guidelines

The ECMP table contains gateway sets, and each gateway set defines the equal-cost gateways that lead to a destination subnet. When IP route sharing is enabled, subnet entries in the LPM table can be mapped to gateway set entries in the ECMP table, instead of to a single gateway within the LPM table.

Note
ExtremeXOS does not support configuration of the ECMP tables on Summit X440 series switches.

For improved ECMP scaling, each LPM table entry points to a gateway set entry in the ECMP table.Each gateway set entry is unique and appears only once in the ECMP table.

Each gateway set entry for the platforms listed above is unique and appears only once in the ECMP table. Multiple LPM table entries can point to the same gateway set entry. This efficient use of the ECMP table creates more room in the ECMP table for additional gateway set entries. It also makes IP route sharing available to every entry in the LPM table.

The following command allows you to configure the maximum number of next-hop gateways for gateway sets in the ECMP table:

```
configure iproute sharing max-gateways max_gateways
```

Each gateway entry in a gateway set consumes ECMP table space. As the **max_gateways** value decreases, the ECMP table supports more gateway sets. If you configure the **max_gateways** value to 64, the switch supports route sharing through up to 64 gateways per subnet, but supports the smallest number of gateway sets. If you do not need to support up to 64 different gateways for any subnet, you can decrease the **max_gateways** value to support more gateway sets.

To determine which gateways might be added to the ECMP table, consider how many local gateways are connected to the switch and can be used for ECMP, and consider the **max_gateways** value.

For example, suppose that you have four ECMP gateway candidates connected to the switch (labeled A, B, C, and D for this example) and the **max_gateways** option is set to 4. For platforms that allow a gateway set entry to support multiple subnets, this configuration could result in up to 11 gateway sets in the ECMP table: ABCD, ABC, ABD, ACD, BCD, AB, AC, AD, BC, BD, and CD.

If there are 4 gateways and you set **max-gateways** to 4, you can use the choose mathematical function to calculate the total number of gateway set possibilities as follows:

```
(4 choose 4) + (4 choose 3) + (4 choose 2) = 11
```

To calculate the number of gateway set possibilities for a given number of total gateways and a specific **max-gateways** value, use the choose function in the following formula:

```
(TGW choose MGW) + (TGW choose MGW-1) + ... + (TGW choose 2) = TGWsets
```

In the formula above, TGW represents the total local gateways, MGW represents the **max_gateways** value, and TGWsets represents the total gateway sets needed to support all possible shared paths.

To see if your platform supports the total gateway sets needed, do the following:

- Calculate the total ECMP gateway sets possible as described above.
- Compare your result to the IP route sharing (total combinations of gateway sets) capacities listed in the *ExtremeXOS Release Notes* to verify that the switch can support the desired number of gateway sets.

#### Troubleshooting: ECMP Table-Full Messages

If the ECMP table is full, no new gateway sets can be added, and IP forwarding is still done in hardware through one of the following:

- For platforms that allow a gateway set entry to support multiple subnets, forwarding can be done using an existing gateway set that is a partial subset of the unavailable gateway set. If the unavailable gateway set consists of N gateways, the subset used could include a range of gateways from N-1 gateways down to a single gateway. For example, if the ECMP table does not have room for a new gateway set using gateways E, F, G, and H, a partial subset such as EFG, EF, or E will be used.

- For platforms that require one gateway set entry per subnet, forwarding is done through a single gateway.

On BlackDiamond 8000 and X series modules and Summit family switches, an ECMP table-full condition produces the following message:

```
<Info:Kern.IPv4FIB.Info> Slot-1: IPv4 route can not use sharing on all its gateways.
Hardware ECMP Table full.  Packets are HW forwarded across a subset of gateways.
(Logged at most once per hour.)
```

If the ECMP table-full message appears, consider the following remedies:

- If the message source is a BlackDiamond 8000 e-series module, the ECMP table capacity is lower than for the following hardware: BlackDiamond X8 series and 8000 a-, c-, xl-, and xm-series hardware and Summit X460, X480, X670, X670-G2, and X770 series switches. Consider upgrading your hardware to support the greater ECMP table capacity. See the *ExtremeXOS Release Notes* for information on the total combinations of gateway sets supported for IP route sharing on different platforms.
- Reduce the number of gateways adjacent to the switch used for IP route sharing.
- Monitor the switch to see if the condition is transient. For example, if the number of ECMP table entries temporarily increases due to a network event and then returns to within the supported range, a permanent change might not be required.
- Determine if IP route sharing to all gateways is required. Since traffic is still being forwarded in hardware using one or more gateways, the condition may be acceptable.

# Configuring Unicast Routing

## Configuring Basic Unicast Routing

To configure IP unicast routing on the switch:

1. Create and configure two or more *VLAN*s.
2. For each VLAN that participates in IP routing, assign an IP address, use the following command:

   `configure {vlan} vlan_nameipaddress [ipaddress {ipNetmask} | ipv6-link-local | {eui64} ipv6_address_mask]`

   `Ensure that each VLAN has a unique IP address.`
3. Configure a default route using the following command:

   `configure iproute add default gateway {metric} {multicast |multicast-only |unicast |unicast-only} {vr vrname}`

   `Default routes are used when the router has no other dynamic or static route to the requested destination.`
4. Turn on IP routing for one or all VLANs using the following command:

   `enable ipforwarding {ipv4 | broadcast | ignore-broadcast | fast-direct-broadcast} {vlan vlan_name}`
5. Configure the routing protocol, if required. For a simple network using RIP, the default configuration may be acceptable.

6. Turn on *RIP* or OSPF using one of the following commands:

   ```
   enable rip
   enable ospf
   ```

   **Note**
   Unicast reverse path forwarding is available to help prevent Distributed Denial of Service attacks. For complete information, see "The protocol anomaly detection security functionality is supported by a set of anomaly-protection enable, disable, configure, clear, and show CLI commands. For further details, see the *ExtremeXOS 16.2 Command Reference Guide*.

## Adding a Default Route or Gateway

A default route or gateway defines a default interface to which traffic is directed when no specific routes are available. To add a default route, use the command:

```
configure iproute add default gateway {metric} {multicast | multicast-only | unicast | unicast-only} {vr vrname}
```

**Note**
If you define a default route and subsequently delete the *VLAN* on the subnet associated with the default route, the invalid default route entry remains. You must manually delete the configured default route.

## Configuring Static Routes

- To configure a static route, use the command:

  ```
  configure iproute add [ipNetmask | ip_addr mask] gateway {bfd} {metric} {multicast | multicast-only | unicast | unicast-only} {vlan egress_vlan} {vr vrname}
  ```

  **Note**
  Tracert might not always work if inter-VRF routing is configured in one of the hops to the destination.

  **Note**
  When inter-vr routing is configured the gateway address should be different from *VLAN* IP of the switch and it should be reachable (ARP resolved) from the switch.

  The gateway address cannot be loop back address or any local address. A static route's next-hop (gateway) must be associated with a valid IP subnet and cannot use the same IP address as a local VLAN. An IP subnet is associated with a single VLAN by its IP address and subnet mask. If the VLAN is subsequently deleted, the static route entries using that subnet must be deleted manually.

  For Inter-VR routing, the egress VLAN name specified in the static route command may be a VLAN belonging to a VR different from the VR of the static route itself. When the VRs differ, Inter-VR routing of hardware and software forwarded packets is performed. This command can enable or disable BFD protection for one static route. However, this protection is not provided until the BFD client is enabled at each end of the route with the following command:

```
enable iproute bfd {gateway} ip_addr {vr vrname}
```

- To disable BFD protection for a static route, use the following command:

```
disable iproute bfd {gateway} ip_addr {vr vrname}
```

## Configuring the Relative Route Priority

To change the relative route priority, use the following command:

```
configure iproute {ipv4} priority [blackhole | bootp | ebgp | host-
mobility | ibgp | icmp | isis | isis-level-1 | isis-level-1-external |
isis-level-2 | isis-level-2-external | mpls | ospf-as-external | ospf-
extern1 | ospf-extern2 | ospf-inter | ospf-intra | rip | static host-
mobility] priority {vr vrname}
```

## Configuring Hardware Routing Table Usage

Note

This procedure applies only to BlackDiamond X8 and 8000 series modules and Summit X440, X460, X480, X670, X670-G2, and X770 switches.

Allowing individual local and remote IPv4 unicast hosts to occupy the unused portion of the Longest Prefix Match (LPM) table helps reduce Layer 3 hardware hash table collisions, and reduces slowpath forwarding of IP unicast and multicast traffic. For more information, see “Hardware Routing Table Management” on page 1262.

- To configure the number of IP routes to reserve in the LPM hardware table, use the command:

```
configure iproute reserved-entries [ num_routes_needed | maximum |
default ] slot [all | slot_num]
```

- To display the current configuration for IP route reserved entries, use the command:.

```
show iproute reserved-entries {slot slot_num}
```

- To display the hardware table usage for IP routes, unicast and multicast, use the command:

```
show iproute reserved-entries statistics {slot slot_num }
```

## Configuring IP Route Sharing

IP route sharing is introduced in IP Route Sharing and ECMP on page 1248. The following sections describe how to manage IP route sharing:

- Managing IP Route Sharing on page 1267
- Viewing the IP Route Sharing Configuration on page 1268

### *Managing IP Route Sharing*

For BlackDiamond X8 and 8000 series modules, SummitStack, and Summit family switches that support Layer 3 routing, the ExtremeXOS software supports route sharing across up to 2, 4, 8, 16, 32, or 64 next-hop gateways.

- To configure the maximum number of *ECMP* gateways, use the following command:

```
configure iproute sharing max-gateways max_gateways
```

For guidelines on managing the number of gateways, see ECMP Hardware Table on page 1263.

- To enable route sharing, use the following command:

  `enable iproute {ipv4} sharing {{vr}vrname`

- To disable route sharing, use the following command:

  `disable iproute {ipv4} sharing {{vr} vrname`

*Viewing the IP Route Sharing Configuration*

To view the route sharing configuration, use the following command:

`show ipconfig {ipv4} {vlan vlan_name}`

## Configuring Route Compression

- To enable route compression for IPv4 routes, use the following command:

  `enable iproute compression {vr vrname}`

- To disable route compression for IPv4 routes, use the following command:

  `disable iproute compression {vr vrname}`

When you enable or disable route compression, that process is performed in chunks, rather than as one single processing event. Because the ExtremeXOS Route Manager processes a limited number of IP prefixes per second, route compression should not have any significant impact on performance. Likewise, when IP route compression is enabled, incremental route addition or deletion should not have a significant impact on performance.

Note

IP route compression is enabled by default.

## Configuring Static Route Advertisement

- To enable or disable advertisement of all static routes, use one of the following commands:

  `enable rip export [bgp | direct | e-bgp | i-bgp | ospf | ospf-extern1 | ospf-extern2 | ospf-inter | ospf-intra | static | isis | isis-level-1| isis-level-1-external | isis-level-2 | isis-level-2-external ] [cost number {tag number} | policy policy-name]`

  or

  `disable rip export [bgp | direct | e-bgp | i-bgp | ospf | ospf-extern1 | ospf-extern2 | ospf-inter | ospf-intra | static | isis | isis-level-1| isis-level-1-external | isis-level-2| isis-level-2-external ]`

  `enable ospf export [bgp | direct | e-bgp | i-bgp | rip | static | isis | isis-level-1 | isis-level-1-external | isis-level-2 | isis-level-2-external] [cost cost type [ase-type-1 | ase-type-2] {tag number} | policy-map]`

  or

```
disable ospf export [bgp | direct | e-bgp | i-bgp | rip | static |
isis | isis-level-1| isis-level-1-external | isis-level-2 | isis-
level-2-external]
```

## Configuring Distributed IP ARP Mode

The distributed IP ARP feature is available only on BlackDiamond X8 and 8800 series switches. The distributed IP ARP feature provides higher IP ARP scaling by distributing IP ARP forwarding information to only the I/O module to which each IP host is connected. This feature is off by default to match the operation in ExtremeXOS releases prior to 12.5. When this feature is off, complete IP ARP information for all destinations is stored on all modules, reducing the available space for unique destinations.

- To activate or deactivate the distributed IP ARP feature, use the following command:

  ```
  configure iparp distributed-mode [on | off]
  ```

  Note

  The switch does not use the specified feature configuration until the next time the switch boots. If you are using load sharing or access-lists with action "redirect-port" or "redirect-port-list", refer to the ExtremeXOS Command Reference command description for the above command for information on these restrictions.

- To display the configured and current states for this feature, use the following command:

  ```
  show iparp {ip_addr | mac | vlan vlan_name | permanent} {vr vr_name}
  ```

- To display distributed IP ARP mode statistics when this feature is activated, use the following command:

  ```
  show iparp distributed-mode statistics { slot [ slot| all ] }
  ```

# Displaying the Routing Configuration and Statistics

## Viewing IP Routes

Use the `show iproute` command to display the current configuration of IP unicast routing for the switch and for each *VLAN*. The `show iproute` command displays the currently configured routes and includes how each route was learned.

## Viewing the IP ARP Table

To view the IP ARP table entries and configuration, use the `show iparp` command.

## Viewing IP ARP Statistics

To view IP ARP table statistics, use the following commands:

```
show iparp distributed-mode statistics { slot [ slot | all ] }
```

```
show iparp stats [[ vr_name | vr {all | vr_name} ] {no-refresh} | {vr}
summary]
```

```
show iparp stats [vlan {all {vr vr_name}} | {vlan} vlan_name] {no-
refresh}

show iparp stats ports {all | port_list} {no-refresh}
```

## Viewing the IP Configuration for a VLAN

To view the IP configuration for one or more *VLAN*s, use the `show ipconfig` command.

## Viewing Compressed Routes

- View a compressed route using the following command: `show iproute`.

  Sample output:

```
Ori Destination        Gateway         Mtr  Flags       VLAN       Duration
#be 3.0.0.0/8          111.222.0.5     7    UG-D---um--- feed       0d:19h:52m:49s
#be 4.0.0.0/8          111.222.0.5     5    UG-D---um--- feed       0d:19h:52m:49s
#be 4.0.0.0/9          111.222.0.5     5    UG-D---um--c feed       0d:19h:52m:49s
#be 4.23.84.0/22       111.222.0.5     7    UG-D---um--c feed       0d:19h:52m:49s
#be 4.23.112.0/22      111.222.0.5     7    UG-D---um--c feed       0d:19h:52m:49s
.....................................
Origin(Ori): (b) BlackHole, (be) EBGP, (bg) BGP, (bi) IBGP, (bo) BOOTP
(ct) CBT, (d) Direct, (df) DownIF, (dv) DVMRP, (e1) ISISL1Ext
(e2) ISISL2Ext, (h) Hardcoded, (i) ICMP, (i1) ISISL1 (i2) ISISL2
(mb) MBGP, (mbe) MBGPExt, (mbi) MBGPInter
(mo) MOSPF (o) OSPF, (o1) OSPFExt1, (o2) OSPFExt2
(oa) OSPFIntra, (oe) OSPFAsExt, (or) OSPFInter, (pd) PIM-DM, (ps) PIM-SM
(r) RIP, (ra) RtAdvrt, (s) Static, (sv) SLB_VIP, (un) UnKnown
(*) Preferred unicast route (@) Preferred multicast route
(#) Preferred unicast and multicast route

Flags: (B) BlackHole, (D) Dynamic, (G) Gateway, (H) Host Route
(m) Multicast, (P) LPM-routing, (R) Modified, (S) Static
(u) Unicast, (U) Up (c) Compressed

Mask distribution:
19 routes at length  8           9 routes at length 9
9 routes at length 10          28 routes at length 11

Route Origin distribution:
7 routes from Direct          184816 routes from EBGP

Total number of routes = 184823
Total number of compressed routes = 93274
```

- Display an iproute summary using the following command: `show iproute summary`

  Sample output:

```
=================ROUTE SUMMARY=================
Mask distribution:
1 routes at length  8           7 routes at length 24
1 routes at length 32

Route origin distribution:
6    Static        3    Direct

Total number of routes = 9
Total number of compressed routes = 4
```

- Display a Route Manager configuration using the following command: `show configuration rtmgr`

  Sample output:

```
#
# Module rtmgr configuration.
#
disable iproute sharing
configure iproute priority mpls 20
.......
disable icmp timestamp vlan "to62"
enable ip-option loose-source-route
enable iproute compression ipv4 vr "VR-Default"
```

## Routing Configuration Example

The following figure illustrates a BlackDiamond switch that has three *VLAN*s defined as follows:

- Finance
  - All ports on slots 1 and 3 have been assigned.
  - IP address 192.207.35.1.
- Personnel
  - Protocol-sensitive VLAN using the IP protocol.
  - All ports on slots 2 and 4 have been assigned.
  - IP address 192.207.36.1.
- MyCompany
  - Port-based VLAN.
  - All ports on slots 1 through 4 have been assigned.

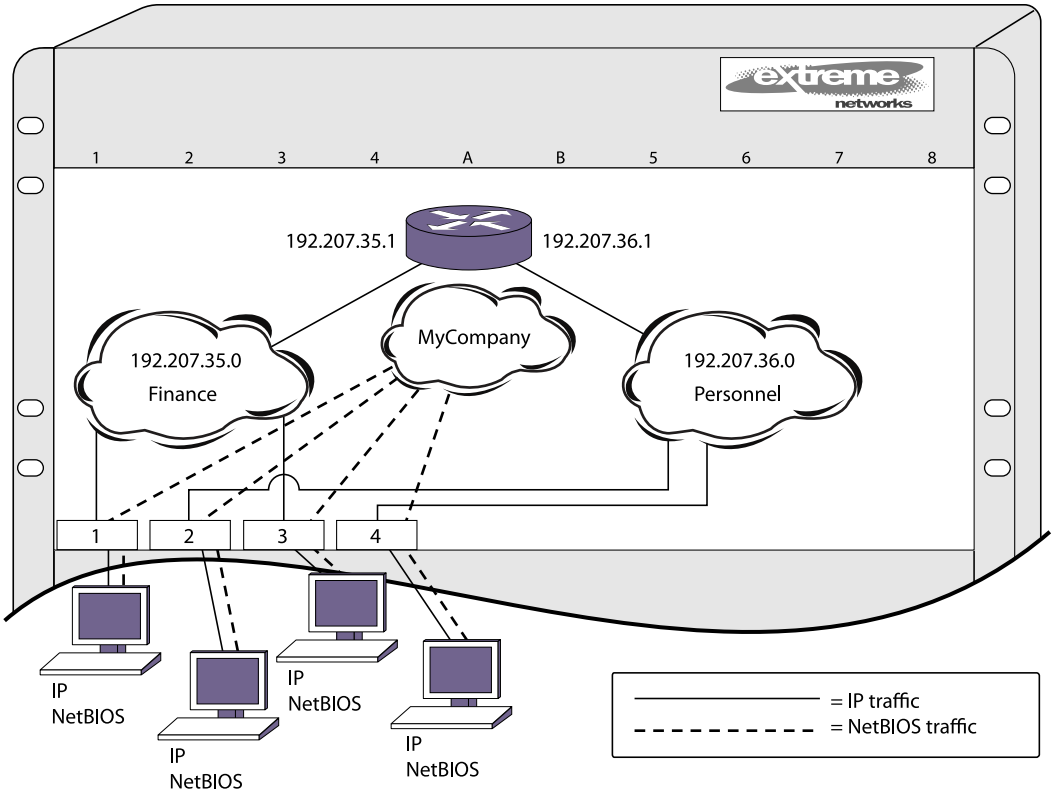



**Figure 205: Unicast Routing Configuration Example**

The stations connected to the system generate a combination of IP traffic and NetBIOS traffic. The IP traffic is filtered by the protocol-sensitive VLANs. All other traffic is directed to the VLAN MyCompany.

In this configuration, all IP traffic from stations connected to slots 1 and 3 have access to the router by way of the VLAN Finance. Ports on slots 2 and 4 reach the router by way of the VLAN Personnel. All other traffic (NetBIOS) is part of the VLAN MyCompany.

The example in the above figure is configured as follows:

```
create vlan Finance tag 10
create vlan Personnel tag 11
create vlan MyCompany

configure Finance protocol ip
configure Personnel protocol ip

configure Finance add port 1:*,3:* tag
configure Personnel add port 2:*,4:* tag
configure MyCompany add port all

configure Finance ipaddress 192.207.35.1
configure Personnel ipaddress 192.207.36.1
configure rip add vlan Finance
configure rip add vlan Personnel

enable ipforwarding
enable rip
```

# Duplicate Address Detection

The Duplicate Address Detection (DAD) feature checks networks attached to a switch to see if IP addresses configured on the switch are already in use on an attached network.

## DAD Overview

When enabled on a user VR or *VR-Default*, the DAD feature checks all IP addresses configured on the DAD-enabled VRs to determine if there are duplicate IP addresses on the networks connected to the switch. If a duplicate address is discovered, the switch does one of the following:

- Marks the IP address as valid
- Marks the IP address as duplicate and generates *EMS (Event Management System)* events to advertise this

At the completion of the DAD check for each interface, the interface is marked valid or duplicate. A valid IP interface can be used by all switch processes for IP communications. There are no restrictions on a valid IP address. If no duplicate address is detected, the IP address is marked valid.

A duplicate IP address cannot be used for IP communications. If a duplicate IP address is detected, the marking depends on the action that initiated the test and can depend on a previous marking for the IP address. For some events, a duplicate IP address generates an EMS event, and for some other events, a duplicate IP address results in a disabled IP address and corresponding EMS events.

### *Prerequisites for a DAD Check*

To successfully test an IP interface, at least one port in the host *VLAN* must be in the Up state. If all ports in the host VLAN are Down, the DAD check does not run.

The DAD check does not run on loopback VLANs; an IP address for a loopback VLAN is marked valid and the address is identified in the `show ip dad` command display with the L flag.

### *The DAD Check*

The DAD feature checks IP addresses by sending an ARP request to each IP address it checks. The source IP address in the ARP request is 0, and the destination IP address is the IP address being checked. If another device replies to the ARP request, a duplicate IP address is detected.

The DAD check is repeated a configurable number of times for each IP interface. During the IPv4 DAD check, the status for an interface under test is tentative, and this status is shown with the T flag when the `show ip dad` command is entered. The DAD check is very fast, so it might be hard to view the tentative state for an address. If the address had previously been marked duplicate, the status remains duplicate while the DAD check runs. If no duplicate address is detected when the DAD check runs at interface startup, the interface IP address is declared valid.

If the DAD check feature is not enabled at startup, you can enable it after startup with a CLI command. Once enabled at the switch prompt, a DAD check runs on all IP interfaces when you enter the `run ip dad`, and it automatically runs on a single interface when an interface is initialized.

Note

When you enable the DAD feature at the CLI prompt, no DAD check is performed until you enter the `run ip dad` command or an interface is initialized.

An interface initialization can be triggered by enabling a disabled *VLAN* that has an IP configuration, or you can initialize an interface by adding an IP address to a VLAN and enabling IP forwarding. The DAD check runs only on the interface being initialized, and it does not run again until another interface is initialized.

When a duplicate IP address is detected, an *EMS* event is generated and the IP address is marked as follows:

- Valid—The interface remains valid if it was marked valid as a result of a previous DAD check. This treatment prevents the switch from disabling an interface that was working and now has an address conflict with another device.
- Duplicate—The duplicate IP address is disabled and cannot be used by switch processes. This treatment is appropriate for an interface that is just joining a network and should not conflict with pre-established services.

You can use the `show ip dad` command to display duplicate IP addresses, which are marked with the D flag.

### *Switch Impact for DAD State Changes*

When an IP address is in a duplicate or tentative state, the normal behavior of the switch may change since that IP address isn't usable, even though it does exist. The following are some examples of what can happen when an IP address is marked duplicate or tentative:

- Routes may be withdrawn or marked inactive
- Dynamic IP ARP entries may get flushed
- *VRRP (Virtual Router Redundancy Protocol) virtual router (VR)* instances may be disabled and put into init state causing the backup VRRP router to take over mastership.
- Ping and traceroute commands may fail.
- The *DHCP (Dynamic Host Configuration Protocol)* client will send a DHCP decline to the DHCP server if the IP address for a DHCP client on a *VLAN* is not Valid
- The DHCP scope IP address range configuration might fail when the DHCP enabled VLAN IP address becomes duplicate.
- *SNMP (Simple Network Management Protocol)* requests may fail
- SNMP traps will not be sent if the if the configured source IP address is not Valid

### *Fixing a Duplicate IP Address*

If a DAD check declares an IP address duplicate, the address remains duplicate until a later DAD check declares the IP address valid, or until the affected *VLAN* is configured as a loopback VLAN. To make the interface IP address valid, do either of the following:

- Correct the duplicate address situation and enter the `run ip dad` command.
- Disable the host VLAN, correct the duplicate address situation, and bring up the interface.

## Guidelines and Limitations

The following guidelines and limitations apply to the DAD check feature:

- IPv6 and GRE tunnels are not supported on IP addresses that are validated by a DAD check.
- The DAD check does not run on loopback *VLANs*; an IP address for a loopback VLAN is marked valid and the address is identified in the show ip dad command display with the L flag.
- The switch MAC address is installed for a VLAN if needed.
- DAD detects duplicate IPv4 address configured on a VLAN that spans *MLAG (Multi-switch Link Aggregation Group)* peer switches only when the solicitation attempts using `configure ip dad attempts max_solicitations` is more than 1.

## Configuring DAD

To enable or disable the DAD feature and configure feature operation, use the following command:

```
configure ip dad [off | on | {on} attempts max_solicitations] {{vr}
vr_name | vr all}
```

## Running a DAD Check

To initiate a DAD check, use the following command:

```
run ip dad [{vlan} vlan_name | {{vr} vr_name} ip_address]
```

## Displaying DAD Configuration and Statistics

To display DAD configuration and statistics information, use the following command:

```
show ip dad {[{{vr} vr_name {ip_address} | vr all | {vlan} vlan_name}
```

## Clearing the DAD Counters

To clear the DAD feature statistics counters, use the following command:

```
clear ip dad {{vr} vr_name {ip_address} | vr all | {vlan} vlan_name}
{counters}
```

# Proxy ARP

Proxy Address Resolution Protocol (ARP) was first invented so that ARP-capable devices could respond to ARP request packets on behalf of ARP-incapable devices. Proxy ARP can also be used to achieve router redundancy and to simplify IP client configuration. The switch supports proxy ARP for this type of network configuration.

## ARP-Incapable Devices

- Configure the switch to respond to ARP requests on behalf of devices that are incapable of doing so. Configure the IP address and MAC address of the ARP-incapable device using the following command: .

  `configure iparp add proxy [ipNetmask | ip_addr {mask}] {vr vr_name} {mac | vrrp} {always}`

After it is configured, the system responds to ARP requests on behalf of the device as long as the following conditions are satisfied:

- The valid IP ARP request is received on a router interface.
- The target IP address matches the IP address configured in the proxy ARP table.
- The proxy ARP table entry indicates that the system should answer this ARP request, based on the ingress *VLAN* and whether the **always** parameter is set. When the **always** option is set, the switch always responds to the ARP request even when the ARP requester is on the same subnet as the requested host. If the **always** option is not set, the switch only answers if the ARP request comes in from a VLAN that is not on the same subnet as the requested host.

When all the proxy ARP conditions are met, the switch formulates an ARP response using one of the following MAC addresses:

- A specific MAC address specified with the *mac* variable.
- The *VRRP* virtual MAC address when the **vrrp** option is specified and the request is received on a VLAN that is running VRRP.
- The switch MAC address when neither of the above options applies.

## Proxy ARP Between Subnets

In some networks, it is desirable to configure the IP host with a wider subnet than the actual subnet mask of the segment. You can use proxy ARP so that the router answers ARP requests for devices outside of the subnet. As a result, the host communicates as if all devices are local. In reality, communication with devices outside of the subnet are proxied by the router.

For example, an IP host is configured with a class B address of 100.101.102.103 and a mask of 255.255.0.0. The switch is configured with the IP address 100.101.102.1 and a mask of 255.255.255.0. The switch is also configured with a proxy ARP entry of IP address 100.101.0.0 and mask 255.255.0.0, without the **always** parameter.

When the IP host tries to communicate with the host at address 100.101.45.67, the IP host communicates as if the two hosts are on the same subnet, and sends out an IP ARP request. The switch answers on behalf of the device at address 100.101.45.67, using its own MAC address. All subsequent data packets from 100.101.102.103 are sent to the switch, and the switch routes the packets to 100.101.45.67.

# IPv4 Multinetting

IP multinetting refers to having multiple IP networks on the same bridging domain (or *VLAN*). The hosts connected to the same physical segment can belong to any one of the networks, so multiple subnets can overlap onto the same physical segment. Any routing between the hosts in different networks is

done through the router interface. Typically, different IP networks are on different physical segments, but IP multinetting does not require this.

Multinetting can be a critical element in a transition strategy, allowing a legacy assignment of IP addresses to coexist with newly configured hosts. However, because of the additional constraints introduced in troubleshooting and bandwidth, Extreme Networks recommends that you use multinetting as a transitional tactic only, and not as a long-term network design strategy.

Multinetting was not supported in ExtremeXOS 10.1, but versions of ExtremeWare before that supported a multinetting implementation that required separate VLANs for each IP network. The implementation introduced in ExtremeXOS is simpler to configure, does not require that you create a dummy multinetting protocol, and does not require that you create VLANs for each IP network. This implementation does not require you to explicitly enable IP multinetting. Multinetting is automatically enabled when a secondary IP address is assigned to a VLAN.

## Multinetting Topology

For an IP multinetted interface, one of the IP networks on the interface acts as the transit network for the traffic that is routed by this interface. The transit network is the primary subnet for the interface. The remaining multinetted subnets, called the secondary subnets, must be stub networks. This restriction is required because it is not possible to associate the source of the incoming routed traffic to a particular network. IP routing happens between the different subnets of the same *VLAN* (one arm routing) and also between subnets of different VLANs.

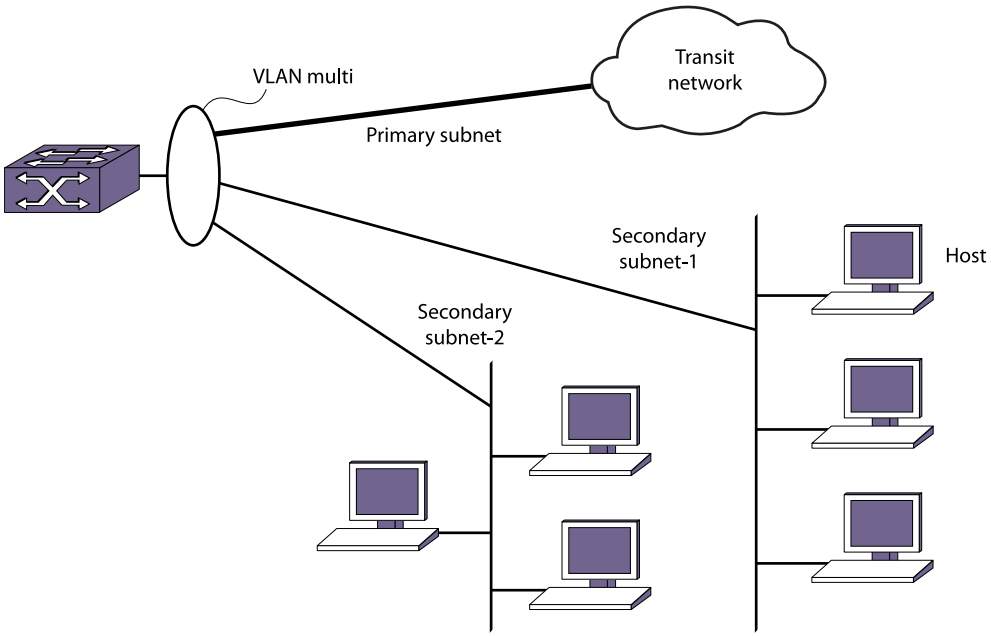



**Figure 206: Multinetted Network Topology**

Figure 210 shows a multinetted VLAN named multi. VLAN multi has three IP subnets so three IP addresses have been configured for the VLAN. One of the subnets is the primary subnet and can be connected to any transit network (for example, the Internet). The remaining two subnets are stub networks, and multiple hosts such as management stations (such as user PCs and file servers) can be connected to them. You should not put any additional routing or switching devices in the secondary

subnets to avoid routing loops. In Figure 210 the subnets are on separate physical segments, however, multinetting can also support hosts from different IP subnets on the same physical segment.

When multinetting is configured on a VLAN, the switch can be reached using any of the subnet addresses (primary or secondary) assigned to VLAN. This means that you can perform operations like ping, Telnet, Trivial File Transfer Protocol (TFTP), Secure Shell 2 (SSH2), and others to the switch from a host residing in either the primary or the secondary subnet of the VLAN. Other host functions (such as traceroute) are also supported on the secondary interface of a VLAN.

## How Multinetting Affects Other Features

*ARP*

ARP operates on the interface and responds to every request coming from either the primary or secondary subnet. When multiple subnets are configured on a *VLAN*and an ARP request is generated by the switch over that VLAN, the source IP address of the ARP request must be a local IP address of the subnet to which the destination IP address (which is being ARPed) belongs.

For example, if a switch multinets the subnets 10.0.0.0/24 and 20.0.0.0/24 (with VLAN IP addresses of 10.0.0.1 and 20.0.0.1), and generates an ARP request for the IP address 10.0.0.2, then the source IP address in the ARP packet is set to 10.0.0.1 and not to 20.0.0.1.

*Route Manager*

The Route Manager installs a route corresponding to each of the secondary interfaces. The route origin is direct, is treated as a regular IP route, and can be used for IP data traffic forwarding.

These routes can also be redistributed into the various routing protocol domains if you configure route redistribution.

*IRDP*

Some functional changes are required in Internet Router Discovery Protocol (IRDP) to support IP multinetting. When IRDP is enabled on a Layer 3 *VLAN*, the ExtremeXOS software periodically sends *ICMP* router advertisement messages through each subnet (primary and secondary) and responds to ICMP router solicitation messages based on the source IP address of the soliciting host.

*Unicast Routing Protocols*

Unicast routing protocols treat each IP network as an interface. The interface corresponding to the primary subnet is the active interface, and the interfaces corresponding to the secondary subnet are passive subnets.

For example, in the case of *OSPF*, the system treats each network as an interface, and hello messages are not sent out or received over the non-primary interface. In this way, the router link state advertisement (LSA) includes information to advertise that the primary network is a transit network and the secondary networks are stub networks, thereby preventing any traffic from being routed from a source in the secondary network.

Interface-based routing protocols (for example, OSPF) can be configured on per *VLAN* basis. A routing protocol cannot be configured on an individual primary or secondary interface. Configuring a protocol parameter on a VLAN automatically configures the parameter on all its associated primary and secondary interfaces. The same logic applies to configuring IP forwarding, for example, on a VLAN.

Routing protocols in the multinetted environment advertise the secondary subnets to their peers in their protocol exchange process. For example, for OSPF the secondary subnets are advertised as stub networks in router LSAs. *RIP* also advertises secondary subnets to its peers residing on the primary subnet.

This section describes the behavior of OSPF in an IPv4 multinetting environment:

- Each network is treated as an interface, and hello messages are not sent out or received over the non-primary interface. In this way, the router LSA includes information to advertise that the primary network is a transit network and the secondary networks are stub networks, thereby preventing any traffic from being routed from a source in the secondary network.
- Any inbound OSPF control packets from secondary interfaces are dropped.
- Direct routes corresponding to secondary interfaces can be exported into the OSPF domain (by enabling export of direct routes), if OSPF is not enabled on the container VLAN.
- When you create an OSPF area address range for aggregation, you must consider the secondary subnet addresses for any conflicts. That is, any secondary interface with the exact subnet address as the range cannot be in another area.
- The automatic selection algorithm for the OSPF router ID considers the secondary interface addresses also. The numerically highest interface address is selected as the OSPF router-id.

This section describes the behavior of the Routing Information Protocol (RIP) in an IP multinetting environment:

- RIP does not send any routing information update on the secondary interfaces. However, RIP does advertise networks corresponding to secondary interfaces in its routing information packet to the primary interface.
- Any inbound RIP control packets from secondary interfaces are dropped.
- Direct routes corresponding to secondary interfaces can be exported into the RIP domain (by enabling export of direct routes), if RIP is not enabled on the container VLAN.

There are no behavioral changes in the *BGP* in an IP multinetting environment.

This section describes a set of recommendations for using BGP with IP multinetting:

- Be careful of creating a BGP neighbor session with a BGP speaker residing in secondary subnet. This situation can lead to routing loops.
- All secondary subnets are like stub networks, so you must configure BGP in such a way that the BGP next hop becomes reachable using the primary subnet of a VLAN.
- When setting the BGP next hop using an inbound or outbound policy, ensure that the next hop is reachable from the primary interface.
- A BGP static network's reachability can also be resolved from the secondary subnet.
- Secondary interface addresses can be used as the source interface for a BGP neighbor.
- Direct routes corresponding to secondary interfaces can be exported into the BGP domain (by enabling export of direct routes).

This section describes the behavior of IS-IS in an IPv4 multinetting environment:

- IS-IS includes all the interface addresses in its reachability information. Adjacency is established only based on the primary interface address. If the adjacency-check option is disabled by the disable isis adjacency-check command, then IS-IS adjacency is established irrespective of the subnet address match.

### *IGMP Snooping and IGMP*

*IGMP (Internet Group Management Protocol)* snooping and IGMP treat the *VLAN* as an interface. Only control packets with a source address belonging to the IP networks configured on that interface are accepted.

IGMP accepts membership information that originates from hosts in both the primary and secondary subnets. The following describes the changes in behavior of IGMP in an IP multinetting environment:

- A Layer 3 VLAN always uses the primary IP address as the source address to send out an IGMP query, and querier election is based on the primary IP address of interface. Because the RFC dictates that there is only one querier per physical segment, the querier may be attached to any of configured IP interfaces, including secondary interfaces, although this is not a recommended configuration.
- For a static IGMP group, the membership report is also sent out using the primary IP address.
- For local multicast groups such as 224.0.0.X, the membership report is sent out using the first IP address configured on the interface, which is the primary IP address in ExtremeXOS.
- The source IP address check is disabled for any IGMP control packets (such as IGMP query and IGMP membership report). Source IP address checking for IGMP control packet is disabled for all VLANs, not just the multinetted VLANs.

### *Multicast Routing Protocols*

For Protocol-Independent Multicast (PIM), the following behavior changes should be noted in a multinetting environment:

- PIM does not peer with any other PIM router on a secondary subnet.
- PIM also processes data packets from the hosts secondary subnets.
- PIM also accepts membership information from hosts on secondary subnets.

### *EAPS, ESRP, and STP*

Control protocols like Ethernet Automatic Protection Switching (EAPS), *ESRP (Extreme Standby Router Protocol)*, and the *STP (Spanning Tree Protocol)* treat the *VLAN* as an interface. If the protocol control packets are exchanged as Layer 3 packets, then the source address in the packet is validated against the IP networks configured on that interface.

### *DHCP Server*

The *DHCP* server implementation in ExtremeXOS only supports address allocation on the primary IP interface of the configured *VLAN*. That is, all DHCP clients residing on a bridging domain have IP addresses belonging to the primary subnet. To add a host on secondary subnet, you must manually configure the IP address information on that host.

### *DHCP Relay*

When the switch is configured as a *DHCP* relay agent, it forwards the DHCP request received from a client to the DHCP server. When doing so, the system sets the GIADDR field in the DHCP request packet to the primary IP address of the ingress *VLAN*. This means that the DHCP server that resides on a remote subnet allocates an IP address for the client in the primary subnet range.

**Note**
DHCP Relay is not supported on Summit X430 or X440L2.

*VRRP*

Virtual Router Redundancy Protocol (VRRP) protection can be provided for the primary as well as for the secondary IP addresses of a *VLAN*. For multinetting, the IP address assigned to an VRRP *VR* identifier (VRID) can be either the primary or the secondary IP addresses of the corresponding VLAN.

For example, assume a VLAN v1 with two IP addresses: a primary IP address of 10.0.0.1/24, and a secondary IP address of 20.0.0.1/24.

To provide VRRP protection to such a VLAN, you must configure one of the following:

- Configure VRRP in VLAN v1 with two VRRP VRIDs. One VRID should have the virtual IP address 10.0.0.1/24, and the other VRID should have the virtual IP address 20.0.0.1/24. The other VRRP router, the one configured to act as backup, should be configured similarly.

—OR—

- Configure VRRP in VLAN v1 with two VRRP VRIDs. One VRID should have the virtual IP address as 10.0.0.1/24, and the other VRID should have the virtual IP address as 20.0.0.1/24.

It is possible for a VRRP VR to have additional virtual IP addresses assigned to it.

In this case, the following conditions must be met:

- Multiple virtual IP addresses for the same VRID must be on the same subnet.
- Multiple virtual IP addresses must all not be owned by the switch.

Assuming a VLAN v1 that has IP addresses 1.1.1.1/24 and 2.2.2.2/24, here are some more examples of valid configurations:

- VRRP VR on v1 with VRID of 99 with virtual IP addresses of 1.1.1.2 and 1.1.1.3
- VRRP VR on v1 with VRID of 100 with virtual IP addresses of 2.2.2.3 and 2.2.2.4
- VRRP VR on v1 with VRID of 99 with virtual IP addresses of 1.1.1.98 and 1.1.1.99
- VRRP VR on v1 with VRID of 100 with virtual IP addresses of 2.2.2.98 and 2.2.2.99

Given the same VLAN v1 as above, here are some invalid configurations:

- VRRP VR on v1 with VRID of 99 with virtual IP addresses of 1.1.1.1 and 2.2.2.2 (the virtual IP addresses are not on the same subnet)
- VRRP VR on v1 with VRID of 100 with virtual IP addresses of 2.2.2.2 and 1.1.1.1 (the virtual IP addresses are not on the same subnet)
- VRRP VR on v1 with VRID of 99 with virtual IP addresses of 1.1.1.1 and 1.1.1.99 (one virtual IP address is owned by the switch and one is not)
- VRRP VR on v1 with VRID of 100 with virtual IP addresses of 2.2.2.2 and 2.2.2.99 (one virtual IP address is owned by the switch and one is not).

## Configuring IPv4 Multinetting

You configure IP multinetting by adding a secondary IP address to a *VLAN*.

- Use the following command to add a secondary IP address:

  `configure vlan` *`vlan_name`* **`add secondary-ipaddress`** `[`*`ip_address {netmask}`* `|` *`ipNetmask`*`]`

After you have added a secondary IP address, you cannot change the primary IP address of a VLAN until you first delete all the secondary IP addresses.

- To delete secondary IP addresses, use the following command:

  `configure vlan vlan_name delete secondary-ipaddress [ip_address | all]`

## IP Multinetting Examples

The following example configures a switch to have one multinetted segment (port 5:5) that contains three subnets (192.168.34.0/24, 192.168.35.0/24, and 192.168.37.0/24).

Note
The secondary IP address of the super *VLAN* cannot be used for the sub VLAN IP address range.

```
configure default delete port 5:5
create vlan multinet
configure multinet ipaddress 192.168.34.1/24
configure multinet add secondary-ipaddress 192.168.35.1/24
configure multinet add secondary-ipaddress 192.168.37.1/24
configure multinet add port 5:5
enable ipforwarding
```

The following example configures a switch to have one multinetted segment (port 5:5) that contains three subnets (192.168.34.0, 192.168.35.0, and 192.168.37.0). It also configures a second multinetted segment consisting of two subnets (192.168.36.0 and 172.16.45.0). The second multinetted segment spans three ports (1:8, 2:9, and 3:10). *RIP* is enabled on both multinetted segments.

```
configure default delete port 5:5
create vlan multinet
configure multinet ipaddress 192.168.34.1
configure multinet add secondary-ipaddress 192.168.35.1
configure multinet add secondary-ipaddress 192.168.37.1
configure multinet add port 5:5
configure default delete port 1:8, 2:9, 3:10
create vlan multinet_2
configure multinet_2 ipaddress 192.168.36.1
configure multinet_2 add secondary-ipaddress 172.16.45.1
configure multinet_2 add port 1:8, 2:9, 3:10
configure rip add vlan multinet
configure rip add vlan multinet_2
enable rip
enable ipforwarding
```

# DHCP/BOOTP Relay

The following sections describe how to use the *DHCP*/BOOTP Relay feature:

Note
DHCP Relay is not supported on Summit X430 or X440L2.

- Managing DHCP/BOOTP Relay on page 1283.
- Configuring the DHCP Relay Agent Option (Option 82) at Layer 3 on page 1283.

- Viewing the DHCP/BOOTP Relay Statistics and Configuration on page 1285.

**Note**
This section discusses DHCP/BOOTP relay operation at Layer 3. For information on DHCP/BOOTP relay operation at Layer 2, see DHCP Snooping and Trusted DHCP Server on page 881.

BOOTP Relay agent of DHCPv6 relays the DHCPv6 messages between the server/client across subnets of a larger IPv6 network.

The DHCPv6 server/BOOTP relay agent listens to UDP port 547.

A relay agent relays both messages from clients and Relay-forward messages from other relay agents. When a relay agent receives a valid message, it constructs a new Relay-forward message and option from the DHCP message received, then forwards it to the next hop/server. The server responds with the corresponding IP address or configuration through a Relay-Reply message to its peer, and thus to the client.

The ExtremeXOS implementation of DHCPv6 takes reference from ISC DHCPv6.

## Managing DHCP/BOOTP Relay

After IP unicast routing has been configured, you can configure the switch to forward *DHCP* or BOOTP requests coming from clients on subnets being serviced by the switch and going to hosts on different subnets. This feature can be used in various applications, including DHCP services between Windows NT servers and clients running Windows 95.

**Note**
If DHCP/BOOTP Relay is enabled on a per *VLAN* basis, make sure it is enabled on both the client-side and server-side VLANs.

You can enable the use of LSP next hops, or you can enable DHCP/BOOTP relay. The software does not support both features at the same time.

**Note**
BootPrelay is not supported in VRF.

## Configuring the DHCP Relay Agent Option (Option 82) at Layer 3

After configuring and enabling the DHCP/BOOTP relay feature, you can enable the *DHCP* relay agent option feature. This feature inserts a piece of information, called option 82, into any DHCP request packet that is to be relayed by the switch. Similarly, if a DHCP reply received by the switch contains a valid relay agent option, the option is stripped from the packet before it is relayed to the client.

When DHCP option 82 is enabled, two types of packets need to be handled:

DHCP Request

When the switch (relay agent) receives a DHCP request, option 82 is added at the end of the packet. If the option has already been enabled, then the action taken depends on the configured policy (drop packet, keep existing option 82 value, or replace the existing option). If the incoming DHCP request is tagged, then that *VLAN* ID is added to the circuit ID sub option of option 82; otherwise, the default VLAN ID is added.

DHCP Reply

When the option 82 information check is enabled, the packets received from the DHCP server are checked for option 82 information. If the remote ID sub-option is the switch's MAC address, the packet is sent to the client; if not, the packet is dropped. If the check is not enabled. The packets are forwarded as-is.

The DHCP relay agent option consists of two pieces of data, called sub-options. The first is the agent circuit ID sub-option, and the second is the agent remote ID sub-option. When the DHCP relay agent option is enabled on switches running ExtremeXOS, the value of these sub-options is set as follows:

Agent circuit ID sub-option

The full circuit ID string uses the format <vlan_info>-<port_info>. You can use the default values for vlan_info and port_info, or you can configure these values as described later in this section.

Agent remote ID sub-option

Always contains the Ethernet MAC address of the relaying switch. You can display the Ethernet MAC address of the switch by issuing the show switch command.

Note

For more general information about the DHCP relay agent information option, refer to RFC 3046.

*Enabling and Disabling the DHCP Relay Agent Option*

- To enable the *DHCP* relay agent option, use the following command after configuring the DHCP/BOOTP relay function:

  ```
  configure bootprelay dhcp-agent information option
  ```

- To disable the DHCP relay agent option, use the following command:

  ```
  unconfigure bootprelay dhcp-agent information option
  ```

*Enabling and Disabling DHCP Packet Checking*

In some instances, a *DHCP* server may not properly handle a DHCP request packet containing a relay agent option.

- To prevent DHCP reply packets with invalid or missing relay agent options from being forwarded to the client, use this command:

  ```
  configure bootprelay dhcp-agent information check
  ```

- To disable checking of DHCP replies, use this command:

  ```
  unconfigure bootprelay dhcp-agent information check
  ```

*Configuring the DHCP Packet Handling Policy*

A *DHCP* relay agent may receive a client DHCP packet that has been forwarded from another relay agent. If this relayed packet already contains a relay agent option, then the switch handles this packet according to the configured DHCP relay agent option policy. The possible actions are to replace the option information, to keep the information, or to drop packets containing option 82 information.

- To configure this policy, use the following command:

  ```
  configure bootprelay dhcp-agent information policy [drop | keep |
  replace]
  ```

- The default relay policy is replace. Configure the policy to the default, use this command:

  ```
  unconfigure bootprelay dhcp-agent information policy
  ```

*Configuring the DHCP Agent Circuit ID Suboption*

- To configure the values used to create the agent circuit ID suboption, use the following commands:

  ```
  configure bootprelay dhcp-agent information circuit-id port-information port_info port port
  configure bootprelay dhcp-agent information circuit-id vlan-information vlan_info {vlan} [vlan_name |all]
  ```

## Viewing the DHCP/BOOTP Relay Statistics and Configuration

- To view the *DHCP*/BOOTP relay statistics and configuration, use the following command:

  ```
  show bootprelay
  ```

- To view the BOOTP relay enable/disable configuration, use the following command:

  ```
  show bootprelay configuration {{vlan} vlan_name | {vr} vr_name}
  ```

- To view the DHCP relay agent option (Option 82) configuration, use the following commands:

  ```
  show bootprelay dhcp-agent information circuit-id port-information ports all
  show bootprelay dhcp-agent information circuit-id vlan-information
  ```

# DHCP Smart Relay

A *DHCP* Relay agent relays DHCP requests from the client to the DHCP server, and relays the DHCP replies from the server to the client. It acts as a proxy, and can reduce the number of DHCP servers required in the network. The DHCP relay agent inserts its own IP address in the giaddr field (gateway address) of the DHCP request. The DHCP server looks into this IP address, identifys the DHCP client's subnet, and assigns an IP address accordingly.

The EXOS BOOTPRelay module is Extreme Networks' DHCP Relay agent. It is now enhanced to optionally insert the secondary addresses of the interface.

DHCP clients can now be dynamically assigned both public and private addresses in an effort to reduce administrative overhead. EXOS can now configure the BOOTPRelay agent to insert both primary and secondary address(es) of the client-facing interface as a gateway address in the DHCP packet. At any given point in time, the DHCP client can be assigned one IP address.

You can insert different addresses in the giaddr field (gateway address) in two different ways:

- Parallel mode: the switch simultaneously relays multiple DHCP Discover packets, each containing a different IP address as the gateway address. The relay agent receives a DHCP Discover request from the DHCP client. The relay agent makes multiple copies of this DHCP DISCOVER request, inserts each IP address of the client-facing interface in the giaddr field (gateway address) in each one of these copies, and relays all these DHCP DISCOVER packets simultaneously to the DHCP server.

The DHCP server has the responsibility to assign the correct IP address in the correct subnet by choosing the DHCP DISCOVER packet to respond to. Similarly, the DHCP client is responsible for accepting the appropriate DHCP OFFER from the DHCP server.

Note

Only the DHCP DISCOVER request is sent in multiple copies, with different IP addresses as the gateway address in each. All other DHCP packets are relayed normally.

- Sequential mode: the switch relays a DHCP DISCOVER packet for each IP address sequentially.

  The relay agent receives a DHCP DISCOVER request and inserts the primary address of the client-facing interface as the gateway address, and relays this packet to the server. The switch counts the number of times a DHCP client sends out a DHCP request without receiving a DHCP OFFER message. After three retries, the relay agents sets the secondary address as the gateway address in the next DHCP Discover request that gets relayed. If the DHCP server still does not respond after three retries, the next secondary address is used as the gateway address, and so on cyclically.

## Configuring DHCP Smart Relay

Issue the following commands to configure te DHCP Smart Relay feature:

- To configure DHCP smart relay mode to include the secondary IP address as giaddr at VR level: `configure bootprelay {ipv4 | ipv6} include-secondary {sequential | parallel | off} {vr vr_name}`
- To configure DHCP smart relay mode to include the secondary IP address as giaddr at *VLAN* level: `configure bootprelay {ipv4 | ipv6} {vlan} vlan_name include-secondary {sequential | parallel | off}`
- To unconfigure any smart relay configuration that was specified at the VLAN level: `unconfigure bootprelay {ipv4 | ipv6} {vlan} vlan_name include-secondary`
- To display various DHCP BOOTP Relay statistics:
  - `show bootprelay`
  - `show bootprelay ipv6`
  - `show bootprelay configuration ipv4`
  - `show bootprelay configuration ipv6`

## Supported Platforms

This feature is supported on all platforms.

# Broadcast UDP Packet Forwarding

UDP Forwarding is a flexible and generalized routing utility for handling the directed forwarding of broadcast UDP packets. UDP Forwarding enables you to configure your switch so that inbound broadcast UDP packets on a *VLAN* are forwarded to a particular destination IP address or VLAN. UDP Forwarding allows applications, such as multiple *DHCP* relay services from differing sets of VLANs, to be directed to different DHCP servers.

The following rules apply to UDP broadcast packets handled by this feature:

- If the UDP profile includes BOOTP or DHCP, it is handled according to guidelines specified in RFC 1542.
- If the UDP profile includes other types of traffic, these packets have the IP destination address modified as configured, and changes are made to the IP and UDP checksums and TTL field (decrements), as appropriate.

If UDP Forwarding is used for BOOTP or DHCP forwarding purposes, do not configure or use the existing bootprelay function. However, if the previous bootprelay functions are adequate, you may continue to use them.

Note

When using udp-profile to forward dhcp request, the behavior will be different from bootprelay. Where bootprelay will forward the dhcp packet with the client vlan IP as source IP, udp-profile will forward the dhcp packet with the source IP of the egress vlan towards the destination server.

UDP Forwarding only works across a Layer 3 boundary and currently, UDP Forwarding can be applied to IPv4 packets only, not to IPv6 packets.

## Configuring UDP Forwarding

- To configure UDP Forwarding, create a policy file for your UDP profile, and then associate the profile with a *VLAN* using the following command:
  `configure vlan vlan_name udp-profile [profilename | none]`

  You can apply a UDP Forwarding policy only to an L3 VLAN (a VLAN having at least one IP address configured on it). If no IP address is configured on the VLAN, the command is rejected.

  UDP profiles are similar to *ACL* policy files.UDP profiles use a subset of the match conditions allowed for ACLs. Unrecognized attributes are ignored. A UDP forwarding policy must contain only the following attributes:
  - Match attributes
    - Destination UDP port number (destination-port)
    - Source IP address (source-ipaddress)
  - Action modified (set) attributes
    - Destination IP address (destination-ipaddress)
    - VLAN name (vlan)

  Policy files used for UDP forwarding are processed differently from standard policy files. Instead of terminating when an entry's match clause becomes true, each entry in the policy file is processed and the corresponding action is taken for each true match clause.

  For example, if the following policy file is used as a UDP forwarding profile, any packets destined for UDP port 67 are sent to IP address 20.0.0.5 and flooded to VLAN to7:

```
entry one {
if match all {
    destination-port 67 ;
} then {
```

```
    destination-ipaddress 20.0.0.5 ;
}
}
entry two {
if match all {
    destination-port 67 ;
} then {
    vlan "to7" ;
}
}
```

If you include more than one VLAN set attribute or more than one destination-ipaddress set attribute in one policy entry, the last one is accepted and the rest are ignored.

Note

Although the XOS policy manager allows you to set a range for the destination-port, you should not specify the range for the destination-port attribute in the match clause of the policy statement for the UDP profile. If a destination-port range is configured, the last port in the range is accepted and the rest are ignored.

You can have two valid set statements in each entry of a UDP forwarding policy; one a destination-ipaddress and one a VLAN. The ExtremeXOS software currently allows a maximum of eight entries in a UDP forwarding policy, so you can define a maximum of 16 destinations for one inbound broadcast UDP packet: eight IP addresses and eight VLANs.

Note

It is strongly advised to have no more than eight entries in a UDP forwarding profile. The UDP forwarding module processes those entries even if the entries do not contain any attributes for UDP forwarding. Having more than eight entries drastically reduces the performance of the system. If the inbound UDP traffic rate is very high, having more than eight entries could cause the system to freeze or become locked.

If you rename a VLAN referred to in your UDP forwarding profile, you must manually edit the policy to reflect the new name, and refresh the policy.

You can also validate whether the UDP profile has been successfully associated with the VLAN by using the show policy command. UDP Forwarding is implemented as part of the netTools process, so the command does display netTools as a user of the policy.

- To remove a policy, use the **none** form of the following command:

  `configure vlan` *`vlan_name`* **`udp-profile`** [*`profilename`* | **`none`**]

  or use the following command:

  `unconfigure vlan` *`vlan_name`* **`udp-profile`**

For more information about creating and editing policy files, see Chapter 17, “Policy Manager.” For more information about ACL policy files, see Chapter 18, “ACLs.”

### Configuring UDP Echo Server Support

You can use UDP echo packets to measure the transit time for data between the transmitting and receiving ends.

- To enable UDP echo server support, use the following command:

  `enable udp-echo-server {vr vrid}{udp-port port}`
- To disable UDP echo server support, use the following command:

  `disable udp-echo-server {vr vrid}`

## IP Broadcast Handling

The ExtremeXOS software supports IP subnet directed broadcast forwarding. In the ExtremeXOS software, IP subnet directed broadcast forwarding is done in the software by default; if you want to perform forwarding in the hardware, see the command reference pages on IP forwarding in the *ExtremeXOS 16.2 Command Reference Guide*.

### IP Broadcast Handling Overview

To understand how IP broadcast handling functions in the ExtremeXOS software, consider the following two examples.

For the first example, a system sends an IP packet (such as the IP packet generated by the ping command) to an IP subnet directed broadcast address which is directly connected to that system. In this case, the IP packet goes out as a Layer 2 broadcast with the destination media access control (DMAC) addresses all set to FF, while the source media access control (SMAC) is set to the system MAC. This packet is sent out of all the ports of the *VLAN*.

In the second example, a system sends a packet (such as the IP packet generated by the ping command) to an IP subnet directed broadcast address which is remotely connected through a gateway. In this case, the IP packet goes out as a Layer 2 unicast packet with the DMAC equal to the gateway's MAC address, while the SMAC is set to the system MAC. At the gateway router, the existing IP packet forwarding mechanism is sufficient to send the packet out of the correct interface if the router is not the final hop router.

When the packet reaches the final hop router, which is directly connected to the target IP subnet, IP directed broadcast forwarding needs to be turned on.

The IP broadcast handling feature is applicable only at the final hop router directly attached to the target subnet. At the final hop router, when IP subnet directed broadcast forwarding is enabled on an IP VLAN via the command line, the following happens:

- Some basic validity checks are performed (for example, a check to see if the VLAN has IP enabled)
- A subnet broadcast route entry for the subnet is installed. For example, consider a system with the following configuration:

  VLAN-A = 10.1.1.0/24, ports 1:1, 1:2, 1:3, 1:4

  VLAN-B = 20.1.1.0/24, ports 1:5, 1:6, 1:7, 1:8

  VLAN-C = 30.1.1.0/24, ports 1:9, 1:10, 1:11

If you enable IP directed broadcast forwarding on VLAN-A, you should install a route entry for 10.1.1.255 on this system.

- A packet arriving on port 1:5 VLAN-B with destination IP (DIP) set to 10.1.1.255, the source IP (SIP) set to 20.1.1.3, the DMAC set to the router MAC, and the SMAC set to the originating system MAC, arrives at the installed route entry and is sent out on all the ports of VLAN-A, with DMAC set to be all FF and the SMAC set to the router's system MAC.
- An IP packet arriving on port 1:1 VLAN-A with the DIP set to 10.1.1.255, the SIP set to 10.1.1.3, the DMAC set to all FF, and the SMAC set to the originator's MAC, causes Layer 2 flooding on all ports of VLAN-A.

When IP subnet directed broadcast is disabled on an IP VLAN, it is disabled on all VLAN ports and all IP subnet directed broadcast entries are deleted.

Note
IP subnet directed broadcast uses fast-path forwarding.

## VLAN Aggregation

*VLAN* aggregation is a feature aimed primarily at service providers.

Note
This feature is supported only on the platforms listed for this feature in the license tables in the *Feature License Requirements* document.

The purpose of VLAN aggregation is to increase the efficiency of IP address space usage. It does this by allowing clients within the same IP subnet to use different broadcast domains while still using the same default router.

Using VLAN aggregation, a superVLAN is defined with the desired IP address. The subVLANs use the IP address of the superVLAN as the default router address. Groups of clients are then assigned to subVLANs that have no IP address, but are members of the superVLAN. In addition, clients can be informally allocated any valid IP addresses within the subnet. Optionally, you can prevent communication between subVLANs for isolation purposes. As a result, subVLANs can be quite small, but allow for growth without re-defining subnet boundaries.

Without using VLAN aggregation, each VLAN has a default router address, and you need to use large subnet masks. The result of this is more unused IP address space.

Multiple secondary IP addresses can be assigned to the superVLAN.

The following figure illustrates VLAN aggregation.

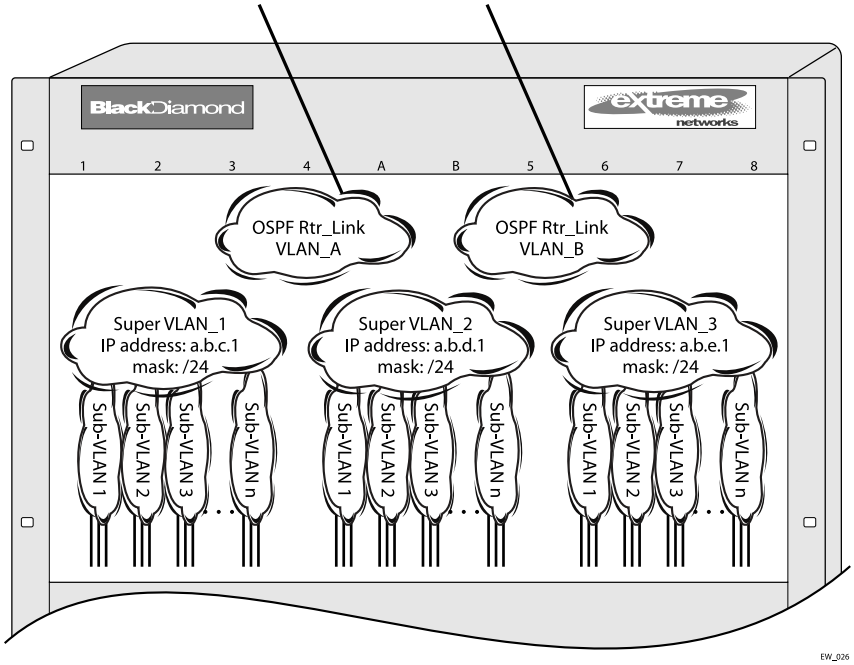



**Figure 207: VLAN Aggregation**

In the preceding figure, all stations are configured to use the address 10.3.2.1 for the default router.

## VLAN Aggregation Properties

*VLAN* aggregation is a very specific application, and the following properties apply to its operation:

- All broadcast and unknown traffic remains local to the subVLAN and does not cross the subVLAN boundary. All traffic within the subVLAN is switched by the subVLAN, allowing traffic separation between subVLANs (while using the same default router address among the subVLANs).
- Hosts can be located on the superVLAN or on subVLANs. Each host can assume any IP address within the address range of the superVLAN router interface. Hosts on the subVLAN are expected to have the same network mask as the superVLAN and have their default router set to the IP address of the superVLAN.
- All IP unicast traffic between subVLANs is routed through the superVLAN. For example, no *ICMP* redirects are generated for traffic between subVLANs, because the superVLAN is responsible for subVLAN routing. Unicast IP traffic across the subVLANs is facilitated by the automatic addition of an ARP entry (similar to a proxy ARP entry) when a subVLAN is added to a superVLAN. This feature can be disabled for security purposes.

## VLAN Aggregation Limitations

The following limitations apply to *VLAN* aggregation:

- No additional routers may be located in a subVLAN. This feature is only applicable for “leaves” of a network.
- A subVLAN cannot be a superVLAN, and vice-versa.
- SubVLANs are not assigned an IP address.
- A subVLAN should belong to only one superVLAN.

- A subVLAN or superVLAN should not be added to a private VLAN
- Before you can delete a superVLAN, you must delete all subVLANs in that superVLAN.
- When configuring a subVLAN address range, all addresses in the range must belong to the superVLAN subnet.

## SubVLAN Address Range Checking

You can configure subVLAN address ranges on each subVLAN to prohibit the entry of IP addresses from hosts outside of the configured range.

- To configure a subVLAN range, use the following command:

  `configure vlan` *`vlan_name`* **`subvlan-address-range`** *`ip address1 - ip address2`*

- To remove a subVLAN address range, use the following command:

  `unconfigure vlan` *`vlan_name`* **`subvlan-address-range`**

- To view the subVLAN address range, use the following command:

  `show vlan {`**`virtual-router`** *`vr-name`*`}`

## Isolation Option for Communication Between SubVLANs

To facilitate communication between subVLANs, by default, an entry is made in the IP ARP table of the superVLAN that performs a proxy ARP function. This allows clients on one subVLAN to communicate with clients on another subVLAN. In certain circumstances, intra-subVLAN communication may not be desired for isolation reasons.

- To prevent normal communication between subVLANs, disable the automatic addition of the IP ARP entries on the superVLAN using the following command:

  `disable subvlan-proxy-arp vlan [`*`vlan-name`* `|` **`all`**`]`

Note

The isolation option works for normal, dynamic, ARP-based client communication.

## VLAN Aggregation Example

The follow example illustrates how to configure *VLAN* aggregation. The VLAN vsuper is created as a superVLAN, and subVLANs, vsub1, vsub2, and vsub3 are added to it.

1. Create and assign an IP address to a VLAN designated as the superVLAN. Be sure to enable IP forwarding and any desired routing protocol on the switch.

```
create vlan vsuper
configure vsuper ipaddress 192.201.3.1/24
enable ipforwarding
enable ospf
configure ospf add vsuper area 0
```

2. Create and add ports to the subVLANs.

```
create vlan vsub1
configure vsub1 add port 10-12
create vlan vsub2
configure vsub2 add port 13-15
```

```
create vlan vsub3
configure vsub3 add port 16-18
```

3. Configure the superVLAN by adding the subVLANs.

```
configure vsuper add subvlan vsub1
configure vsuper add subvlan vsub2
configure vsuper add subvlan vsub3
```

4. Optionally, disable communication among subVLANs.

```
disable subvlan-proxy-arp vlan all
```

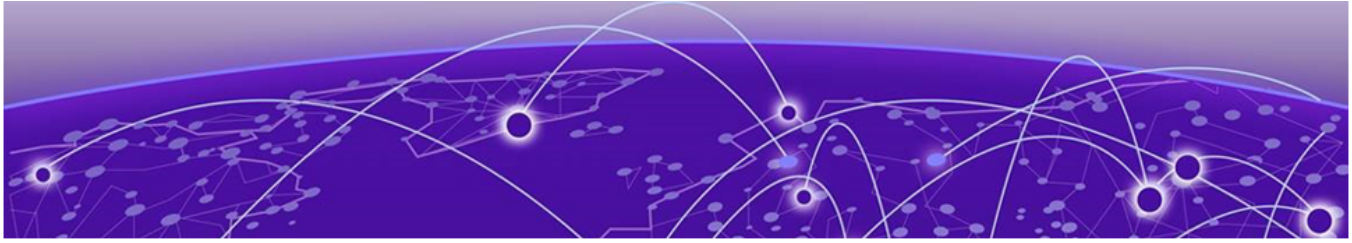
Note

The above command has no impact on Layer 3 traffic.

## Verify the VLAN Aggregation Configuration

The following commands can be used to verify proper *VLAN* aggregation configuration:

- `show vlan`—Indicates the membership of subVLANs in a superVLAN.
- `show iparp`—Indicates an ARP entry that contains subVLAN information. Communication with a client on a subVLAN must have occurred in order for an entry to be made in the ARP table.



# IPv6 Unicast Routing



This chapter assumes that you are already familiar with IPv6 unicast routing. If not, refer to the following publications for additional information:

- RFC 2460—Internet Protocol, Version 6 (IPv6) Specification
- RFC 4291—IP Version 6 Addressing Architecture

Note
For more information on interior gateway protocols, see RIPng or IPv6 Unicast Routing.

## IPv6 Unicast Overview

The switch provides full Layer 3, IPv6 unicast routing. It exchanges routing information with other routers on the network using the IPv6 versions of the following protocols:

- *RIPng (Routing Information Protocol Next Generation)*
- Open Shortest Path First (*OSPFv3 (Open Shortest Path First version 3)*)
- Intermediate System-Intermediate System (IS-IS)
- *BGP (Border Gateway Protocol)*

The switch dynamically builds and maintains a routing table and determines the best path for each of its routes.

The ExtremeXOS software can provide both IPv4 and IPv6 routing at the same time. Separate routing tables are maintained for the two protocols. Most commands that require you to specify an IP address can now accept either an IPv4 or IPv6 address and act accordingly. Additionally, many of the IP configurations, enabling, and display commands have added tokens for IPv4 and IPv6 to clarify the version for which the command applies. For simplicity, existing commands affect IPv4 by default and require you to specify IPv6, so configurations from an earlier release will still correctly configure an IPv4 network.

ACLs and routing policies also support IPv6. Use of an IPv6 address in a rule entry will automatically use IPv6.

## Router Interfaces

The routing software and hardware routes IPv6 traffic between router interfaces. A router interface is either a *VLAN (Virtual LAN)* that has an IP address assigned to it, or a Layer 3 tunnel.

As you create VLANs and tunnels with IP addresses, you can also choose to route (forward traffic) between them. Both the VLAN switching and IP routing function occur within the switch.

IPv4 and IPv6 interfaces can coexist on the same VLAN, allowing both IPv4 and IPv6 networks to coexist on the same Layer 2 broadcast domain.

Note

Each IP address and mask assigned to a VLAN must represent a unique IP subnet. You cannot configure the same IP address and subnet on different VLANs within the same *virtual router (VR)*.

## Tunnels

The ExtremeXOS software supports Layer 3 tunnels, which serve as a transition option, as networks change over from IPv4 to IPv6. The software supports these tunnels in Default-VR.

Note

IPv6 tunnels are supported only in Default-VR.

The ExtremeXOS software supports the use of IPv6-in-IPv4 tunnels (known as configured tunnels or 6in4 tunnels) and IPv6-to-IPv4 tunnels (known as 6to4 tunnels). Both types of tunnels are used to connect regions of IPv6 routing across a region of IPv4 routing. From the perspective of the router, the tunnel across the IPv4 region is one hop, even if multiple IPv4 routers are traversed during transport.

A 6in4 tunnel connects one IPv6 region to one other IPv6 region.

Multiple 6in4 tunnels can be configured on a single router to connect with multiple IPv6 regions. Dynamic and static routing can be configured across a 6in4 tunnel. Hosts in the IPv6 regions need not know anything about the configured tunnel, since packets destined for remote regions are sent over the tunnel like any other type of routing interface.

A 6to4 tunnel connects one IPv6 region with multiple IPv6 regions. Only one 6to4 tunnel can be configured on a single router.

GRE (Generic Router Encapsulation) is a tunneling mechanism where the customer network IP packet is encapsulated by a IP+GRE header across a network infrastructure. The decapsulation is done at the remote gateway providing the original IP datagram delivered to the peers local network. The GRE tunnel is a point-to-point path between two sites, in most cases the tunnel is used to enable communication between sites using private address spaces across a service provider network or third-party network.

## Specifying IPv6 Addresses

IPv6 Addresses are 128 bits (16 bytes) when compared to the 32 bit IPv4 addresses. The ExtremeXOS software accepts two standard representations for IPv6 addresses, as described in RFC 3513, section 2.2, items 1, 2, and 3.

For example, the 128 bits of the address can be represented by eight, four-digit hexadecimal numbers separated by colons:

```
2000:af13:ee10:34c5:800:9192:ba89:2311
3f11:5655:2300:304:0000:0000:7899:acde
```

Leading zeros in a four-digit group can be omitted.

There is a special use of a double colon (::) in an address. The double colon stands for one or more groups of 16 bits of zeros and can only be used once in an address. For example, the following addresses:

```
fe80:0:0:0:af34:2345:4afe:0
fe80:0:0:111:0:0:0:fe11
3c12:0:0:0:0:89:ff:3415
```

can be represented as:

```
fe80::af34:2345:4afe:0
fe80:0:0:111::fe11
3c12::89:ff:3415
```

Additionally, you can specify an address in a mixed IPv4/IPv6 mode that uses six, four-digit hexadecimal numbers for the highest-order part of the address, and uses the IPv4 dotted decimal representation for the lowest-order remaining portion.

For example:

```
0:0:0:0:0:0:192.168.1.1
0:0:0:0:0:ffff:10.0.14.254
```

These can be represented as:

```
::192.168.1.1
::ffff:10.0.14.254
```

Both global and link-local IP addresses can be configured on a *VLAN* or tunnel interface, using the following commands:

```
configure [{vlan} vlan_name | {tunnel} tunnel_name] ipaddress [ {eui64}
ipv6_address_mask | ipv6-link-local
```

```
configure tunnel tunnel_name ipaddress [ipv6-link-local | {eui64}
ipv6_address_mask ]
```

where *ipaddress* refers to the address specified in the above format.

The IPv6 address configuration can be verified using the following commands:

```
show vlan vlan_name}
```

```
show ipconfig ipv6 {vlan vlan_name | tunnel tunnelname}
```

```
show ipconfig ipv6 {vlan vlan_name | tunnel tunnelname}
```

### *Scoped Addresses*

IPv6 uses a category of addresses called link-local addresses that are used solely on a local subnet. Every IPv6 *VLAN* must have at least one link-local address. If a global IP address is configured on a VLAN that has no link-local address, one is assigned automatically by the switch. The link-local addresses start with the prefix fe80::/64. As a result, a switch can have the same link local address assigned to different VLANs, or different neighbors on different links can use the same link local address. Because of this, there are cases where you need to specify an address and a VLAN/tunnel to indicate which interface to use. For those cases, you can indicate the interface by using a scoped address. To scope the address, append the VLAN/tunnel name to the end of the address, separated by a percent sign (%). For example, to indicate the link local address fe80::2 on VLAN finance, use the following form:

```
fe80::2%finance
```

Scoped addresses also appear in the output of display commands.

### *IC_Pv6 Addresses Used in Examples*

For the purposes of documentation, we follow RFC 3849, which indicates that the prefix 2001:db8::/32 can be used as a global unicast address prefix and will not be assigned to any end party.

## Neighbor Discovery Protocol

The Neighbor Discovery Protocol (NDP), as defined in RFC 2461, defines mechanisms for the following functions:

- Resolving link-layer addresses of the IPv6 nodes residing on the link.
- Locating routers residing on the attached link.
- Locating the address prefixes that are located on the attached link.

- Learning link parameters such as the link MTU, or Internet parameters such as the hop limit value that has to be used in the outgoing packets.
- Automatic configuration of the IPv6 address for an interface.
- Detecting whether the address that a node wants to use is already in use by another node, also known as Duplicate Address Detection (DAD).
- Redirecting the traffic to reach a particular destination through a better first-hop.

In IPv4, MAC address resolution is done by ARP. For IPv6, this functionality is handled by NDP. The router maintains a cache of IPv6 addresses and their corresponding MAC addresses and allows the system to respond to requests from other nodes for the MAC address of the IPv6 addresses configured on the interfaces.

Also supported is router discovery—the ability to send out router advertisements that can be used by a host to discover the router. The advertisements sent out contain the prefixes and configuration parameters that allow the end nodes to auto-configure their addresses. The switch also responds to requests from nodes for router advertisements.

The following settings can be configured on an interface to manage router advertisements:

- Settings to control the sending of router advertisements over the interface periodically and to control responding to router solicitations
- The maximum time between sending unsolicited router advertisements
- The minimum time between sending unsolicited router advertisements

You can configure the following values, that are advertised by the switch:

- Managed address configuration flag
- Other stateful configuration flag
- Link MTU
- Retransmit timer
- Current hop limit
- Default lifetime
- Reachable time

Additionally, you can configure the following values for each prefix on the prefix list associated with an interface:

- Valid lifetime of the prefix
- On-link flag
- Preferred lifetime of the prefix
- Autonomous flag

Note

Unlike ExtremeWare, the ExtremeXOS software does not support host processing of neighbor router advertisements.

## Managing Neighbor Discovery

*Create and Delete Static Entries*

- Statically configure the MAC address of IPv6 destinations on the attached links.

  ```
  configure neighbor-discovery cache {vr vr_name} add [ipv6address |
  scoped_link_local] mac
  ```

  ```
  configure neighbor-discovery cache {vr vr_name} delete [ipv6address |
  scoped_link_local]
  ```

*Configure the Neighbor-Discovery Cache Size*

- Configure the maximum number of entries for the neighbor-discovery cache.

  ```
  configure neighbor-discovery cache {vr <vr_name>} max_entries
  <max_entries>
  ```

- Configure the maximum number of pending entries for the neighbor-discovery cache.

  ```
  configure neighbor-discovery cache {vr vr_name} max_pending_entries
  max_pending_entries
  ```

*Manage Neighbor-Discovery Cache Updates*

- Configure the timeout period for dynamic entries in the neighbor-discovery cache.

  ```
  configure neighbor-discovery cache {vr vr_name} timeout timeout
  ```

- Enable the refresh of dynamic entries in the neighbor-discovery cache before the timeout period ends.

  ```
  enable neighbor-discovery {vr vr_name} refresh
  ```

- Disable the refresh of dynamic entries in the neighbor-discovery cache before the timeout period ends.

  ```
  disable neighbor-discovery {vr vr_name} refresh
  ```

*Clear the Neighbor-Discovery Cache*

- The neighbor-discovery entries that are learned dynamically can be cleared using the following command:

  ```
  clear neighbor-discovery cache ipv6 {ipv6address {vr vr_name} | vlan
  vlan_name | vr vr_name} refresh
  ```

  The above CLI command is the IPv6 version of the command to clear IPv6 neighbor discovery entries in the IPv6 neighbor table. An enhancement added in ExtremeXOS 15.7 adds a refresh option so that when “clear neighbor-discovery refresh” is executed, neighbor solicitation will be sent out for each neighbor in the IPv6 neighbor discovery table and only active neighbors will be kept in the neighbor discovery table after the command is completed,

*Return to the Neighbor-Discovery Cache Default Configuration*

- Return to the neighbor-discovery cache default configuration.

  ```
  unconfigure neighbor-discovery cache {vr vr_name}
  ```

*Display Neighbor-Discovery Cache Entries*

- Both statically configured and dynamic neighbor-discovery entries can be viewed using the following command:

  ```
  show neighbor-discovery {cache {ipv6}} {[ipv6_addr | mac | permanent]
  {vr vr_name}} | vlan vlan_name | vr vr_name}
  ```

## IPv6 Router Advertisement Options for DNS

Neighbor Discovery (ND) for IP version 6 and IPv6 stateless address autoconfiguration provide ways to configure either fixed or mobile nodes with one or more IPv6 addresses, default routers, and other parameters [RFC4861][RFC4862]. ExtremeXOS now supports two RA options that provide the DNS information needed for an IPv6 host to reach Internet services.

The Recursive DNS Server (RDNSS) option contains the addresses of recursive DNS servers, and the DNS Search List (DNSSL) option for the Domain Search List that maintains parity with the DHCPv6 options, and ensures the necessary functionality to determine the search domains.

The RDNSS option contains one or more IPv6 addresses of recursive DNS servers. All of the addresses share the same lifetime value. If you wish to have different lifetime values, you can use multiple RDNSS options.

The DNSSL option contains one or more domain names of DNS suffixes. All of the domain names share the same lifetime value. If you wish to have different lifetime values, you can use multiple DNSSL options.

The existing ND message (i.e., Router Advertisement) is used to carry this information. An IPv6 host can configure the IPv6 addresses of one or more RDNSSs through RA messages. Using the RDNSS and DNSSL options, along with the prefix information option based on the ND protocol, an IPv6 host can perform the network configuration of its IPv6 address and the DNS information simultaneously without needing DHCPv6 for the DNS configuration. The RA options for RDNSS and DNSSL can be used on any network that supports the use of ND.

For IPv6-only networks that rely only on IPv6 stateless Autoconfiguration as a deployment model, these two options allow a host to configure its DNS information. This is useful when there is no DHCPv6 infrastructure, or hosts do not have a DHCPv6 client. For networks where DHCPv6 is deployed, you might not need an RA-based DNS configuration.

You can configure the RA options for DNS using the following commands:

*Default RDNSS Lifetime*

```
configure {vlan} vlan_name> router-discovery {ipv6} rdnss-lifetime
[<rdnss_lifetime> | infinity | auto]
```

*Add RDNS Server with Optional RDNSS Lifetime*

```
configure {vlan} vlan_name router-discovery {ipv6} add rdnss ipaddress
{{rdnss-lifetime} [rdnss_lifetime | infinity]}
```

*Default DNSSL Lifetime*

```
configure {vlan} vlan_name router-discovery {ipv6} dnssl-lifetime
[dnssl_lifetime | infinity | auto]
```

*Add a DNS suffix to DNSSL*

```
configure {vlan} vlan_name router-discovery {ipv6} add dnssl dns_suffix
{{dnssl-lifetime} [dnssl_lifetime | infinity]}
```

## IPv6 Router Advertisement Filtering

Newly connected IPv6 hosts can automatically discover configuration details by sending link-local router solicitation messages. Once received, an IPv6 router can send IPv6 router advertisements (RAs) back to the host, that include network configuration parameters. Since IPv6 router advertisement packets are used to configure, among other things, the host's gateway address, unauthorized users employ various methods to spoof IPv6 router advertisements to redirect, or deny service.

The IPv6 Router Advertisement Filtering feature exposes the existing "icmp-type" match criteria to allow IPv6 packets to match RAs and other IPv6 packets that use the ICMPv6 protocol. This functionality provides the ability to flexibly detect certain conditions and take appropriate actions based on network design and expectations.

*Limitations*

This feature has the following limitations:

- Only ingress *ACL (Access Control List)*s support the "icmp-type" match criteria for IPv6 packets. This match criteria cannot be used with egress ACLs.
- The IPv6 extension header parsing varies per platform - see "Platforms Supported" section for more detail
- The IPv6 source-address and destination-address, and the ethernet-source-address and ethernet-destination-address fields cannot be matched in the same rule without enabling "double-wide" mode. Double wide mode is not available on all of the supported platforms and causes a 50% reduction of ACL hardware resources.

*Supported Platforms*

All Summit, BD8K, and BDX8 platforms are supported. However, there are per-platform limitations on how many IPv6 extension headers can be parsed while still matching the supplied *ICMP (Internet Control Message Protocol)* type:

Up to 2 extension headers: X460, X670, X670-G2, X770, 8900-40G6X-xm, BD8X

0 extension headers: All other BDX8, BD8K, and Summit platforms

Newer chipsets include the X460, X670, 8900-40G6X-xm, and BDX8 series. Older chipsets include all other series (including BD8K xl-series).

The exact field compatibility with this match criteria depends on the platform, but all platforms are able to match the port and protocol (ICMPv6) in single wide mode. Using double wide mode provides access to a 128-bit source address, or source MAC address, for example. All of the above platforms support

double wide mode at the expense of reducing ACL scale by 50%. The XGS3 platforms do not support double wide mode at all.

On platforms that support double wide mode, if the layer-2 device is unable to identify whether the packet is an ICMPv6 Router Advertisement message, and the IPv6 Source Address of the packet is a link-local address or is unspecified, the packet is blocked.

You can also use the new “icmp-type” to match other protocol cases such as MLD and MLDv2.

### *CLI Commands*

The existing ACL match criteria `icmp-type` *`type`* is exposed in dynamic ACLs and static ACL policies on the target platforms. This same match criteria is already supported for rules that specify IPv4 criteria on the target platforms.

### *Example Policy*

Here is a policy to detect and log a “simple” RA attack, only allow TCP/UDP/ICMP/xyz protocol traffic that can be parsed (i.e., has up to 2 extension headers and, if fragmented, the L4 NH is contained in the first fragment), and deny everything else including “complex” RA attacks:

```
entry disallow_and_log_RA_attacks
{
if
{protocol icmpv6;icmp-type 134;}
then
{ deny; mirror-cpu; log; count RA_attack;}}
entry allow_tcp
{
if {protocol tcp; first-fragments;}
then {permit;}}
entry allow_udp
{
if {protocol udp; first-fragments;}
then {permit;}}
entry allow_icmp
{
if {protocol icmpv6; first-fragments;}
then {permit;}}
entry allow_xyz…entry denyall
{
if{first-fragments; }
then
{
deny;}}
```

The above policy works for newer chipsets, but leaves (at least) the following exposure for older chipsets: a malicious user could send an RA with a single extension header and it would be allowed to pass due to rule “allow_icmp” (newer chipsets would block this packet through the first rule). To mitigate this exposure on older chipsets, you could call out each “icmp-type” that is supported (ND, MLD, etc.), and then drop any ICMPv6 with an extension header.

For more information, please refer to http://tools.ietf.org/html/draft-ietf-v6ops-ra-guard-implementation-04.

# Managing Duplicate Address Detection

The Duplicate Address Detection (DAD) feature checks networks attached to a switch to see if IP addresses configured on the switch are already in use on an attached network. The following sections provide additional information on the DAD feature:



When you configure an active interface with an IPv6 address, the interface must send out an advertisement containing its address.

All other interfaces on the subnet have the opportunity to respond to the newly configured interface, and inform it that the address is a duplicate. Only after this process occurs, can the interface use the newly configured address. If the interface receives a message that the newly configured address is a duplicate, it cannot use the address.

Until the Duplicate Address Detection (DAD) process completes, the new address is considered tentative, and will be shown as such in any display output. If the address is a duplicate, it will also be labeled as such, and must be reconfigured. On an active interface, the DAD process should occur so quickly that you would not see the address labeled as tentative. However, if you are configuring an interface before enabling it, and you display the configuration, you will see that the address is currently tentative. As soon as you enable the interface, the address should be ready to use, or labeled as duplicate and must be reconfigured.

See RFC 2462, IPv6 Stateless Address Autoconfiguration, for more details.

## DAD Overview

When enabled on a user VR or *VR-Default*, the Duplicate Address Detection (DAD) feature runs when an IP interface is initialized or when a CLI command initiates a DAD check. The DAD check tests IP addresses by sending a neighbor solicitation to each IP address it checks. If another device replies to the neighbor solicitation, a duplicate IP address is detected.

If a duplicate address is detected, the IP interface remains or becomes inactive, and a warning message is logged. If no duplicate address is detected, the IP interface transitions to or remains in the active state. The switch does not automatically repeat the DAD check after the first check is complete. To manually run a DAD test on an interface or IP address, enter the run ipv6 dad command.

Because the automatic DAD check only runs when an interface is initialized, the switch does not detect a duplicate IP address if that address becomes active after the switch interface is initialized. However, if the switch is rebooted or the interface in brought down and then up, the automatic DAD check runs and sets to inactive any interface for which a duplicate IP address is detected.

To successfully test an IPv6 interface, at least one port in the host *VLAN* must be in the Up state. If all ports in the host VLAN are Down, the DAD check does not run. If a port is later added to the host VLAN, or if a port in the host VLAN transitions to Up after the DAD check at initialization is complete, a duplicate IP address can be detected and logged, but the IP interface on the host VLAN remains active.

## Configure DAD

- Enable or disable the DAD feature and configure feature operation.

  ```
  configure ipv6 dad [off | on | {on} attempts max_solicitations] {{vr} vr_name | vr all}
  ```

  **Note**
  For IPv6 interfaces, the DAD feature is automatically enabled on all platforms that support the Edge, Advanced Edge, and Core licenses.

*Run a DAD Test*

- Initiate a DAD test.

  ```
  run ipv6 dad [{vlan} vlan_name | {{vr} vr_name} ipaddress]
  ```

*Display DAD Configuration and Statistics*

- Display DAD configuration and statistics information.

  ```
  show ipv6 dad {[{{vr} vr_name | vr all | {vlan} vlan_name | {tunnel} tunnel_name} {tentative | valid | duplicate} | {{vr} vr_name} ipaddress]} {detail}
  ```

*Clear the DAD Counters*

- Clear the DAD feature statistics counters.

  ```
  clear ipv6 dad {{vr} vr_name {ipaddress} | vr all | {vlan} vlan_name} {counters}
  ```

## Populating the Routing Table

The switch maintains an IP routing table for both network routes and host routes. The table is populated from the following sources:

- Dynamically, by way of routing protocol packets or by *ICMP* redirects exchanged with other routers
- Statically, by way of routes entered by the administrator:
  - Default routes, configured by the administrator
  - Locally, by way of interface addresses assigned to the system
  - By other static routes, as configured by the administrator

Once routes are populated using the above method, IPv6 forwarding needs to be enabled on the *VLAN* using the following command:

```
enable ipforwarding ipv6 {vlan vlan_name | tunnel tunnel_name | vr vr_name}
```

**Note**
If you define a default route and subsequently delete the VLAN on the subnet associated with the default route, the invalid default route entry remains. You must manually delete the configured default route.

*Dynamic Routes*

Dynamic routes are typically learned by way of *RIPng*, *OSPFv3*, *BGP*, or IS-IS, and routers that use these protocols use advertisements to exchange information in their routing tables. Using dynamic routes, the routing table contains only networks that are reachable.

Dynamic routes are aged out of the table when an update for the network is not received for a period of time, as determined by the routing protocol. For details on the configuration and behavior of IPv6 dynamic routes, please refer to the specific Chapters on RIPng, OSPFv3, and IS-IS in this guide.

*Static Routes*

Static routes are manually entered into the routing table. Static routes are used to reach networks not advertised by routers.

- Static IPv6 routes can be created using the following command:

  `configure iproute add` *`ipv6Netmask`* [*`ipv6Gateway`* | *`ipv6ScopedGateway`*] {*`metric`*} {**`vr`** *`vr_name`*} {**`multicast`** | **`multicast-only`** | **`unicast`** | **`unicast-only`**}

  You can configure IPv6 default and blackhole routes. Traffic to blackhole routes is silently dropped.

  The IPv6 gateway can be a global address or a scoped link-local address of an adjacent router.
- You can create static routes, for security reasons, to control which routes you want advertised by the router.

  If you want all static routes to be advertised, configure static routes using one of the following commands:

  `enable ripng export` [**`direct`** | **`ospfv3`** | **`ospfv3-extern1`** | **`ospfv3-extern2`** | **`ospfv3-inter`** | **`ospfv3-intra`** | **`static`** | **`isis`** | **`isis-level-1`**| **`isis-level-1-external`** | **`isis-level-2`**| **`isis-level-2-external`** | **`bgp`**] [**`cost`** *`number`* {**`tag`** *`number`*} | **`policy`** *`policy-name`*]

  or

  `disable rip export` [**`bgp`** | **`direct`** | **`e-bgp`** | **`i-bgp`** | **`ospf`** | **`ospf-extern1`** | **`ospf-extern2`** | **`ospf-inter`** | **`ospf-intra`** | **`static`** | **`isis`** | **`isis-level-1`**| **`isis-level-1-external`** | **`isis-level-2`**| **`isis-level-2-external`** ]

  `enable ospfv3` {**`domain`** *`domainName`*} **`export`** [**`direct`** | **`ripng`** | **`static`** | **`isis`** | **`isis-level-1`** | **`isis-level-1-external`** | **`isis-level-2`** | **`isis-level-2-external`**| **`bgp`**] [**`cost`** *`cost`* **`type`** [**`ase-type-1`** | **`ase-type-2`**] | *`policy_map`*]

  or

  `disable ospfv3` {**`domain`** *`domainName`*} **`export`** [**`direct`** | **`ripng`** | **`static`** | **`isis`** | **`isis-level-1`** | **`isis-level-1-external`** | **`isis-level-2`** | **`isis-level-2-external`** | **`bgp`**]

  The default setting is disabled. Static routes are never aged out of the routing table.

- A static route's nexthop (gateway) must be associated with a valid IP subnet. An IP subnet is associated with a single *VLAN* by its IP address and subnet mask. If the VLAN is subsequently deleted, the static route entries using that subnet must be deleted manually.

  The IPv6 routes can be viewed using the following command:

  `show iproute ipv6 {priority | vlan vlan_name | tunnel tunnel_name | ipv6Netmask | summary {multicast | unicast}} {vr vr_name}}`

- You can view the IPv6 routes based on the type of the route using the following command:

  `show iproute ipv6 origin [direct | static | blackhole | bgp| ibgp | ebgp | ripng | ospfv3 | ospfv3-intra | ospv3-inter | ospfv3-extern1 | ospfv3-extern2 | isis | isis-leve1-1 | isis-level-2 | isis-level-1-external | isis-level-2-external] {vr vr_name}`

#### ExtremeXOS Resiliency Enhancement for IPv6 Static Routes

The ExtremeXOS Resiliency Enhancement feature provides a resilient way to use *ECMP (Equal Cost Multi Paths)* to load balance IPv6 traffic among multiple servers or other specialized devices. ExtremeXOS automatically manages the set of active devices using ECMP static routes configured with ping protection to monitor the health of these routes. Such servers or specialized devices do not require special software to support Bidirectional Forwarding Detection (BFD), or IP routing protocols such as *OSPF (Open Shortest Path First)*, or proprietary protocols to provide keepalive messages. ExtremeXOS uses industry-standard and required protocols ICMPv6/Neighbor Discovery for IPv6 to accomplish the following automatically:

- Initially verify devices and activate their static routes, without waiting for inbound user traffic, and without requiring configuration of device MAC addresses.
- Detect silent device outages and inactivate corresponding static routes.
- Reactivate static routes after device recovery, or hardware replacement with a new MAC address.

ExtremeXOS currently supports similar protection and resiliency using Bidirectional Forwarding Detection (BFD) on IPv4 static routes. However, BFD can only be used when the local and remote device both support BFD.

### *Multiple Routes*

When there are multiple, conflicting choices of a route to a particular destination, the router picks the route with the longest matching network mask. If these are still equal, the router picks the route using the following default criteria (in the order specified):

- Directly attached network interfaces
- Static routes
- *ICMP* redirects
- Dynamic routes
- Directly attached network interfaces that are not active.

Note

If you define multiple default routes, the route that has the lowest metric is used. If multiple default routes have the same lowest metric, the system picks one of the routes.

The criteria for choosing from multiple routes with the longest matching network mask is set by choosing the relative route priorities.

### *Relative Route Priorities*

The following table lists the relative priorities assigned to routes depending on the learned source of the route.

**Note**
Although these priorities can be changed, do not attempt any manipulation unless you are expertly familiar with the possible consequences.

**Table 141: Relative Route Priorities**

| Route Origin | Priority |
|---|---|
| Direct | 10 |
| BlackHole | 50 |
| Static | 1100 |
| *ICMP* | 1200 |
| OSPF3Intra | 2200 |
| OSPF3Inter | 2300 |
| IS-IS | 2350 |
| IS-IS L1 | 2360 |
| IS-IS L2 | 2370 |
| *RIPng* | 2400 |
| *OSPFv3* ASExt | 3100 |
| OSPFv3 Extern1 | 3200 |
| OSPFv3 Extern2 | 3300 |
| IS-IS L1Ext | 3400 |
| IS-IS L2Ext | 3500 |
| EBGP | 1700 |
| IBGP | 1900 |

To change the relative route priority, use the following command:

`configure iproute ipv6 priority [`**`ripng`** | **`blackhole`** | **`icmp`** | **`static`** | **`ospfv3-intra`** | **`ospfv3-inter`** | **`ospfv3-as-external`** | **`ospfv3-extern1`** | **`ospfv3-extern2`** | **`isis-level-1`** | **`isis-level-2`** | **`isis-level-1-external`** | **`isis-level-2-external`** |**`host-mobility`**] *priority* {**`vr`** *vr_name*}

## Unique Local Address (ULA) for IPv6

RFC 4193 defines a globally unique address that is used for local communications. For instance, they are routable within a specified site or between sites. These ULAs are similar to private addresses in IPv4 (RFC 1918).

Of the 128 bits available in IPv6 address, the last 64-bits are used as the interface ID.

The remaining 64 bits are used as follows:

First 8 bits are used to define the known prefix FC00::/7 or FD00::/7

40 bits - global ID - Generated using a random algorithm specified in the RFC 4193. EXOS expects the operator to specify the 40-bit Global ID as ULA address management becomes easier, especially a mult-vendor environment.

16 bits - Used to create subnets within the site.

ULA prefixes should not be accepted by the border routers.

Additionally, ULA prefixes should not be advertised by the border routers. In ExtremeXOS, for *BGP* and *OSPFv3*, you must specify the policies to filter ULA prefixes.

There are no new CLI commands introduced to configure ULAs in ExtremeXOS. You can use the following existing command to configure a ULA:

```
configure {vlan} vlan_name ipaddress [ipaddress {ipNetmask} |
ipv6-link-local | {eui64} ipv6_address_mask]
```

Here is an example of the command:

```
configure vlan v1 ipaddress fd21:0941:2c55::/48
```

The scope of ULA is global by default in RFC 4193.

All applications treat these addresses in a similar manner as any other type of global IPv6 unicast addresses.

## Managing IPv6 Unicast Routing

You should be familiar with IP Route Sharing for both IPv4 and IPv6. For information on IP Route Sharing, please refer to the following section: Configuring IP Route Sharing on page 1267

For BlackDiamond X8 and 8000 series modules, SummitStack, and Summit family switches that support Layer 3 routing, the ExtremeXOS software supports route sharing across up to 2, 4, 8, 16, or 32 next-hop gateways.

### Enable Route Sharing for IPv6

To enable route sharing for IPv6, use the following command:

```
enable iproute ipv6 sharing
```

### Configure Basic IP Unicast Routing

To configure basic IP unicast routing, do the following:

1. Create and configure two or more *VLAN*s.

2. Assign each VLAN that will be using routing an IP address using the following command:

   ```
   configure {vlan} vlan_name ipaddress [ipaddress {ipNetmask} | ipv6-link-local | {eui64} ipv6_address_mask]
   ```

   Ensure that each VLAN has a unique IP address.
3. Configure a static route using the following command:

   ```
   configure iproute add ipv6Netmask [ipv6Gateway | ipv6ScopedGateway] {metric} {vr vr_name} {multicast | multicast-only | unicast | unicast-only}
   ```

   or

   Configure a default route using the following command:

   ```
   configure iproute add default [{gateway {metric} {vr vr_name} {unicast-only | multicast-only}} | {lsp lsp_name {metric}}]
   ```

   Default routes are used when the router has no other dynamic or static route to the requested destination.
4. Turn on IP routing for one or all VLANs using the following command:

   ```
   enable ipforwarding ipv6 {vlan vlan_name | tunnel tunnel_name | vr vr_name}
   ```
5. Configure the routing protocol, if required. For a simple network using *RIPng*, the default configuration might be acceptable.
6. Turn on RIPng or *OSPFv3* using one of the following commands:

   ```
   enable rip
   ```

   ```
   enable ospfv3 {domain domainName}
   ```

   Note

   *BGP* does not use *ECMP* by default, so if you require that functionality you must explicitly issue the `configure bgp maximum-paths max-paths` command with a value greater than 1.

## Managing Router Discovery

*Enable and Disable Router Discovery*

- Enable or disable router discovery on a *VLAN*.

  ```
  enable router-discovery {ipv6} vlan vlan_name
  ```

  ```
  disable router-discovery {ipv6} vlan vlan_name
  ```

*Add and Delete Prefixes for Router Discovery*

- Add or delete prefixes for advertisement by router discovery.

  ```
  configure vlan vlan_name router-discovery {ipv6} add prefix prefix
  ```

```
configure vlan vlan_name router-discovery {ipv6} delete prefix [prefix
| all]
```

*Configure Router Discovery Settings*

- Configure the router discovery settings using the following commands:
  - `configure vlan vlan_name router-discovery {ipv6} default-lifetime defaultlifetime`
  - `configure vlan vlan_name router-discovery {ipv6} link-mtu linkmtu`
  - `configure vlan vlan_name router-discovery {ipv6} managed-config-flag on_off`
  - `configure vlan vlan_name router-discovery {ipv6} max-interval maxinterval`
  - `configure vlan vlan_name router-discovery {ipv6} min-interval mininterval`
  - `configure vlan vlan_name router-discovery {ipv6} other-config-flag on_off`
  - `configure vlan vlan_name router-discovery {ipv6} reachable-time reachabletime`
  - `configure vlan vlan_name router-discovery {ipv6} retransmit-time retransmittime`
  - `configure vlan vlan_name router-discovery {ipv6} set prefix prefix [autonomous-flag auto_on_off | onlink-flag onlink_on_off | preferred-lifetime preflife |valid-lifetime validlife]`
- Reset all router discovery settings to their default values using the command: `unconfigure vlan vlan_name router-discovery {ipv6}`
- To reset an individual router discovery setting to its default value, enter one of the following commands:
  - `unconfigure vlan vlan_name router-discovery {ipv6} default-lifetime`
  - `unconfigure vlan vlan_name router-discovery {ipv6} hop-limit`
  - `unconfigure vlan vlan_name router-discovery {ipv6} link-mtu`
  - `unconfigure vlan vlan_name router-discovery {ipv6} managed-config-flag`
  - `unconfigure vlan vlan_name router-discovery {ipv6} max-interval`
  - `unconfigure vlan vlan_name router-discovery {ipv6} min-interval`
  - `unconfigure vlan vlan_name router-discovery {ipv6} other-config-flag`
  - `unconfigure vlan vlan_name router-discovery {ipv6} reachable-time`
  - `unconfigure vlan vlan_name router-discovery {ipv6} retransmit-time`

*Display Router Discovery Configuration Settings*

To display router discovery settings, use the command: `show router-discovery {ipv6} {vlan vlan_name}`

## Managing Tunnels

IPv6-in-IPv4 and IPv6-to-IPv4 tunnels are introduced in Tunnels on page 1295.

*Create an IPv6-in-IPv4 Tunnel*

- Create an IPv6-in-IPv4 tunnel.

  `create tunnel tunnel_name ipv6-in-ipv4 destination destination-address source source-address`

  The source-address refers to an existing address in the switch, and the destination-address is a remote destination accessible from the switch. A maximum of 255 IPv6-in-IPv4 tunnels can be configured.

*Create an IPv6-to-IPv4 Tunnel*

A 6to4 tunnel connects one IPv6 region with multiple IPv6 regions. Only one 6to4 tunnel can be configured on a single router.

- To create an IPv6-to-IPv4 tunnel, use the command:

  `create tunnel tunnel_name 6to4 source source-address`

  The source-address is an existing address in the switch.

*Delete a Tunnel*

To delete a tunnel, use the command: `delete tunnel tunnel_name`

*Configure an IPv6 Address for a Tunnel*

To configure or unconfigure IPv6 addresses for the tunnels, use the commands:

`configure tunnel tunnel_name ipaddress [ipv6-link-local | {eui64} ipv6_address_mask ]`

`unconfigure tunnel tunnel_name ipaddress ipv6_address_mask`

*Display Tunnel Information*

- Display tunnel information.

  `show [{tunnel} {tunnel_name}]`

## Verify the IP Unicast Routing Configuration

- To display the currently configured routes, which includes how each route was learned, use the command: `show iproute ipv6`

  Additional verification commands include:

  - `show neighbor-discovery cache ipv6`—Displays the neighbor discovery cache of the system.
  - `show ipconfig ipv6`—Displays configuration information for one or more VLANs.
  - `show ipstats ipv6`—Displays the IPv6 statistics for the switch or for the specified VLANs.

## Managing IPv6 Routes and Hosts in External Tables

When external LPM tables are supported and configured on a switch, the configuration setting applies only to the external LPM tables. You can configure the external LPM to contain both IPv4 and, or, IPv6 routes. Internal LPM tables only store IPv4 routes.

The configure forwarding external-tables command using the ipv6 and ipv4-and-ipv6 variables supports larger IPv6 route and host scaling in external LPM tables.

When an external LPM table is configured for l3-only ipv6, no IPv6 routes or IPv6 hosts are configured in any of the internal hardware tables. This provides the highest IPv6 scale, and avoids contention with IP Multicast in the L3 Hash hardware table.

IPv6 hardware and slowpath forwarding are supported on user-created Virtual Routers, and IPv6 tunnels are only supported on *VR-Default*.

The size of the internal LPM tables, and the size of the L3 Hash and Next Hop tables are fixed. For specific information on hardware capacity, refer to the following table, the following table, and the following table in IPv4 Unicast Routing on page 1243.

Note

If you require a large number of IPv6 routes, you should use only xl-series modules, or a Summit X480 standalone, or a SummitStack comprised only of the X480. SummitStacks, or a BD8800 containing a mix of high- and low-capability hardware (slots without External TCAM) does not support more than 100,000 IPv6 routes present.

# IPv6 ECMP and 32-Way ECMP

This feature adds IPv6 *ECMP* support on several new platforms. Additionally, it adds support for 16-way and 32-way ECMP (for both IPv4 and IPv6), using static routes, and up to 8-way ECMP for IPv4 and IPv6 using routing protocols .

Sharing ECMP gateway sets now applies to IPv6 as well as IPv4. Sharing ECMP gateway sets for IPv6 means the entire IPv6 Longest-Prefix Match (LPM) hardware capacity can use ECMP, across up to 32 gateways.

## Feature Description

This feature adds IPv6 ECMP support on the following platforms:

- Summit stack and standalone models, except for the Summit X440 platform.
- BlackDiamond 8000, all I/O modules.
- BlackDiamond X8, all I/O modules.

Additionally, ExtremeEXOS supports 16-way and 32-way ECMP for both IPv4 and IPv6 using static routes, and up to 8-way ECMP for IPv4 and IPv6 using routing protocols. Now, the maximum number of gateways in each IPv4 or IPv6 gateway set in the ECMP hardware table can be configured as 16 or 32.

The following platforms support ECMP only for IPv6 routes whose mask length is <= 64-bits:

- BlackDiamond 8K I/O modules: G48Te2, G24Xc, G48Xc, G48Tc, 10G4Xc, 10G8Xc, S-G8Xc, S-10G1Xc, S-10G2Xc.

## DHCPv6 Relay Remote-ID Option

Available in ExtremeXOS 15.5, the remote-id option is added to the relay forward IPv6 message, if the specific *VLAN* has only the link local IPv6 address. This remote-id is added to all the request packets received from the client on that VLAN, if it is configured or not on the VLAN. The remoteID option added to the packet after the relay message header has the following format, as defined in RFC 4649:

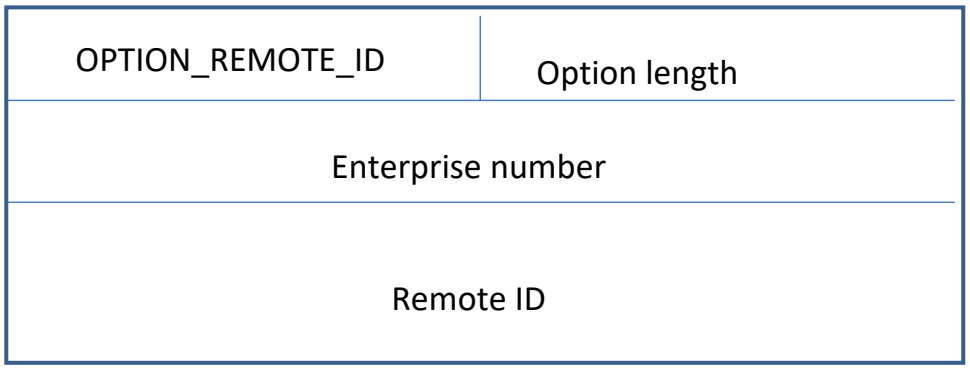



**Figure 208: DHCPv6 Relay Remote-ID Option**

If the remote-id is configured through the CLI on a specified VLAN, it can have either a user defined value, or the system-name. If the user does not configure the value or marks “none” as the remote-id, the switch MAC address will be used as the remote-id. The value of the remote-id present in the packet will have the following format:`vlan_port_<remoteId>`, where *`remoteId`* will be either the configured string, or the system name or the switch MAC address. The Enterprise number field will be populated with the Extreme Networks enterprise number, “1916”.

The remote-id is added to the packets on a specified VLAN, only if:

- IPv6 bootprelay is enabled on the VLAN.
- rhe vlan has only the link local address.

Once a relay agent receives one of the following messages from the client or relay agent, the relay agent will verify if the specific interface has only the link local address:

- DHCPV6_SOLICIT
- DHCPV6_REQUEST
- DHCPV6_CONFIRM
- DHCPV6_RENEW
- DHCPV6_REBIND
- DHCPV6_RELEASE
- DHCPV6_DECLINE
- DHCPV6_INFORMATION_REQUEST
- DHCPV6_RELAY_FORW
- DHCPV6_LEASEQUERY

If so, it verifies if the specific VLAN has a remote-id configured. If a remote-id is present, it is added to the packet along with the VLAN and port, in the format mentioned above. If it is not configured, the default (switch MAC address) will be added as the remote id, along with the VLAN and port.

On the reverse path, if it receives any of the following messages from the server or any other relay agent, and if the interface has only a link local address, it will verify for the presence of remote-id in the packet:

- DHCPV6_ADVERTISE
- DHCPV6_REPLY
- DHCPV6_RECONFIGURE
- DHCPV6_RELAY_FORW
- DHCPV6_LEASEQUERY_REPLY

If it is not present, the agent will check for the interface-id. If one of these is present, it will verify if the value matches the local configuration. If it matches, after removing the remote-id and/or interface-id, the packet will be forwarded to the respective client. If the remote-id is present in the packet, and it does not match the configured value, the packet will be dropped. If none of them are present, the packet will be forwarded to the client, based on the IPv6 address of the VLAN.

## DHCPv6 Relay Agent Prefix Delegation

DHCPv6 Prefix Delegation options provide a mechanism for automated delegation of IPv6 prefixes using *DHCP (Dynamic Host Configuration Protocol)*. This mechanism is intended for delegating a long-lived prefix from a delegating router to a requesting router, across an administrative boundary, where the delegating router does not have/require knowledge about the topology of the networks to which the requesting router is attached, and the delegating router does not require other information aside from the identity of the requesting router to choose a prefix for delegation.

For example, these options would be used by a service provider to assign a prefix to a Customer Premise Equipment (CPE) device acting as a router between the subscriber's internal network and the service provider's core network.

This prefix delegation mechanism is appropriate for use by an ISP to delegate a prefix to a subscriber, where the delegated prefix would possibly be sub-netted and assigned to the links within the subscriber's network.

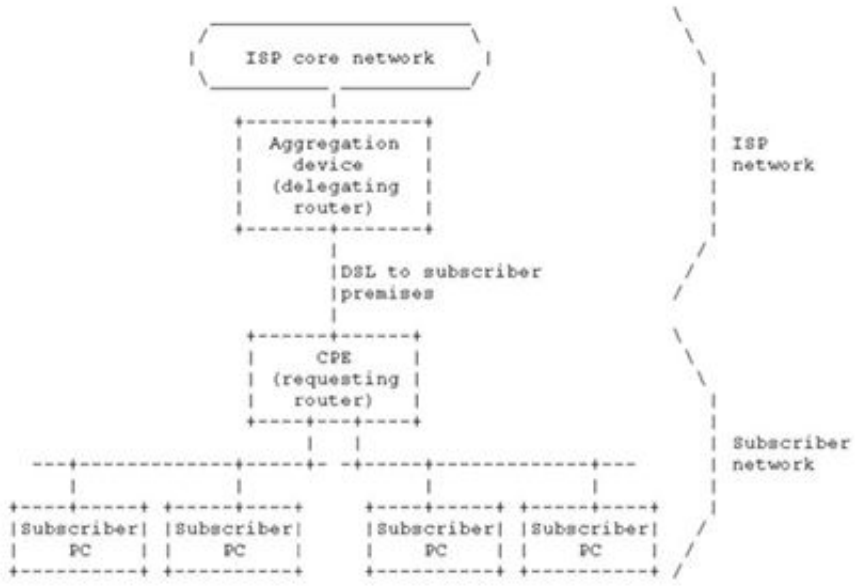



**Figure 209: Prefix Delgation Example**

The prefix delegation process begins when the requesting router requests configuration information through DHCP. When the delegating router receives the request, it selects and returns an available prefix or prefixes for delegation to the requesting router.

The requesting router now acts as a DHCP server for the subscriber network. It subnets the delegated prefix and assigns the longer prefixes to links in the subscriber's network. In a typical scenario based on the network shown in Figure 1, the requesting router subnets a single delegated /48 prefix into /64 prefixes and assigns one /64 prefix to each of the links in the subscriber network.

Prefix delegation with DHCP is independent of address assignment with DHCP. A requesting router can use DHCP for just prefix delegation or for prefix delegation along with address assignment and other configuration information.

## Relay Agent Behavior in Prefix Delegation

A relay agent forwards *DHCP* messages containing Prefix Delegation options in the same way as other DHCP messages.

RFC 3633 says “If a delegating router communicates with a requesting router through a relay agent (sitting in between the Aggregation device (PE) and the CPE), the delegating router may need a protocol or other out-of-band communication to add routing information for delegated prefixes into the provider edge router.”

The requesting router acts as a DHCPv6 server behind the CPE and assigns IPv6 subnets to the customer devices. Since there is no mechanism to notify the delegating router about the IPv6 subnets assigned by the requesting router, the core routers have no clue about these IPv6 subnets and therefore have no dynamic route added into the routing tables in the core. RFC 3633 leaves it to the implementations to device a mechanism to solve this problem.

When the ExtremeXOS DHCPv6 relay agent relays the DHCP messages with IA_PD options and detects that a successful prefix delegation operation has been completed and a prefix or a set of prefixes have been delegated, it installs one route for each of prefixes that has been delegated. This ensures reachability between the customer devices and the service provider network.

When ExtremeXOS DHCPv6 relay agent detects a successful DHCP prefix delegation transaction (Solicit-Advertise-Request-Reply), it looks into the details of the prefix(es) that are being delegated. Each DHCP message with IA_PD option can have multiple IAPREFIX options in it. Each IAPREFIX option represents a prefix that is being delegated.

ExtremeXOS DHCPv6 relay agent installs one route for each of the delegated prefixes. The route is installed on the *VLAN* on which the requesting router (CPE) is reachable and with the requesting router (CPE) as the gateway for the route, if the requesting router is directly connected to our EXOS DHCPv6 Relay agent. If the DHCPv6 messages from the requesting router have been received via another DHCPv6 Relay agent, the route is installed with the next hop DHCPv6 Relay agent as the gateway. If the DHCPv6 packets from the next hop DHCPv6 Relay agent have been IPv6 forwarded via other routers in between (which did not do DHCPv6 Relay, but just IPv6 forwarding), the next hop router to reach the next hop DHCPv6 Relay agent is used as the gateway for the route. EXOS DHCPv6 Relay uses Route Manager (rtMgr) client library APIs for adding and deleting these routes.

To ensure that the routes are retained across reboots (or restart process netTools), ExtremeXOS DHCPv6 relay agent stores these routes along with their validity times in an internal file. After a reboot, ExtremeXOS DHCPv6 relay agent reads this file and installs the routes once again. This file will be a flat file with the following format: `vlanName ipv6Prefix ipv6GatewayAddr startTime validTime`.

Only those IPv6 prefixes which are still valid after the reboot are added again. The expired IPv6 prefixes are discarded. As the file format denotes, this file contains the details about the VLAN, the actual delegated IPv6 prefix, the IPv6 Gateway Address, the start time when it got delegated and the time for which the delegated prefix is deemed to be valid (in seconds). This file is internal to EXOS DHCPv6 relay agent and cannot/should not be edited by anyone else.

ExtremeXOS DHCPv6 relay agent checkpoints the prefix delegations that have been detected.

*Basic Snooping Configuration*

```
create vlan v1

enable bootprelay ipv6
conf bootprelay ipv6 prefix-del snooping on vlan v1

con bootprelay ipv6 prefix-delegation snooping add 5001:db8:3553:bf00::/56
fe80::a440:cfd5:c05b:d324 vlan v1 valid-time 300

* (Engineering) Slot-1 Sharmila_DUT1_Stack-->.13 # sh bootprelay ipv6 prefix-delegation
snooping

Delegated Prefix                              Interface
        Gateway                                       Valid For
-------------------------------------------------------------------------

5001:db8:3553:bf00::/56                       v1
       fe80::a440:cfd5:c05b:d324                           300 secs

* (Engineering) Slot-1 Sharmila_DUT1_Stack-->.14 # show iproute ipv6
Ori Destination                                  Mtr  Flags         Duration
```

```
    Gateway                                     Interface
 bo 5001:db8:3553:bf00::/56                     1    -G-D---um---- 0d:0h:0m:32s
    fe80::a440:cfd5:c05b:d324                   [VR-Default]

Origin(Ori):(b) BlackHole, (be) EBGP, (bg) BGP, (bi) IBGP, (bo) BOOTP,
            (ct) CBT, (d) Direct, (df) DownIF, (dv) DVMRP, (e1) ISISL1Ext,
            (e2) ISISL2Ext, (h) Hardcoded, (i) ICMP, (i1) ISISL1 (i2) ISISL2,
            (is) ISIS, (mb) MBGP, (mbe) MBGPExt, (mbi) MBGPInter, (ma) MPLSIntra,
            (mr) MPLSInter, (mo) MOSPF (o) OSPFv3, (o1) OSPFv3Ext1, (o2) OSPFv3Ext2,
            (oa) OSPFv3Intra, (oe) OSPFv3AsExt, (or) OSPFv3Inter, (pd) PIM-DM, (ps) PIM-
SM,
            (r) RIPng, (ra) RtAdvrt, (s) Static, (sv) SLB_VIP, (un) UnKnown,
            (*) Preferred unicast route (@) Preferred multicast route,
            (#) Preferred unicast and multicast route.

Flags: (b) BFD protection requested, (B) BlackHole, (c) Compressed Route,
       (D) Dynamic, (f) Provided to FIB, (G) Gateway, (H) Host Route,
       (l) Calculated LDP LSP, (L) Matching LDP LSP, (m) Multicast,
       (p) BFD protection active, (P) LPM-routing, (R) Modified, (s) Static LSP,
       (S) Static, (t) Calculated RSVP-TE LSP, (T) Matching RSVP-TE LSP,
       (u) Unicast, (U) Up, (3) L3VPN Route.

Mask distribution:
     1 routes at length  56

Route Origin distribution:
     1 routes from Bootp

Total number of routes = 1
```

## DHCPv6 Client

ExtremeXOS supports Dynamic Host Configuration Protocol version 6 (DHCPv6) clients, receiving its IPv6 address and corresponding parameters for any *VLAN* interface from an external DHCPv6 server. For the ExtremeXOS device to operate as a DHCPv6 client, we need to enable DHCPv6 client on a VLAN interface to obtain an IPv6 address from the DHCPv6 server in the network.

ExtremeXOS DHCPv6 client supports only the stateful DHCPv6 mode of operation in this release. This means an external DHCPv6 server is required to configure the IPv6 address and its related configurations.

Note
Stateless Auto configuration and stateless DHCPv6 modes are not supported in this release.

The mode of operation of the DHCPv6 client is based on the autonomous (A), managed (M) and other configuration (O) flags in the received router advertisement (RA) messages. If the managed bit is 1 and the other configuration bit is 0, the DHCPv6 client will act as a stateful client. In stateful mode, the client receives IPv6 addresses from the DHCPv6 server, based on the identity association for nontemporary addresses (IA_NA) assignment.

In other words ExtremeXOS DHCPv6 client sends IA_NA option as part of the DHCPv6 message.

Note
Identity association for temporary address (IA_TA) and Identity association for prefix delegation (IA_PD) is not supported in this release.

If the incoming RA points to act the interface in stateless DHCPv6 mode (managed bit is 0 and the other configuration bit is 1) or DHCPv6 auto configuration mode (autonomous bit is 1 and managed, other configuration bit is 0) ExtremeXOS will release the assigned (if already assigned using stateful DHCPv6 client mode) DHCPv6 IP and other parameters for this VLAN interface on which this RA is received.

This is because the mode of operation needs to be changed because ExtremeXOS did not support stateless DHCPv6 and auto configuration DHCPv6 client modes and there will not be any subsequent operation. In the stateful DHCPv6 client mode, the DHCPv6 client requests global addresses from the DHCPv6 server. Based on the DHCPv6 server's response, the DHCPv6 client assigns the global addresses to interfaces and sets a lease time for all valid responses. When the lease time expires, the DHCPv6 client renews the lease from the DHCPv6 server. To configure stateful DHCPv6 client based on identity association of non-temporary (IA_NA) address assignment for a VLAN interface use the following command.

```
enable dhcp ipv6 vlan vlan_name
```

To unconfigure the stateful DHCPv6 client for an interface use:

```
disable dhcp ipv6 vlan vlan_name
```

To display the current status of the DHCPv6 client VLAN state use:

```
show dhcp-client ipv6 state {vlan}
```

## DHCP Unique Identifier

Each *DHCP* client and server has a DHCP unique identifier (DUID). A DHCP server uses DUIDs to identify clients for the selection of configuration parameters. DHCP clients use DUIDs to identify a server in messages where a server needs to be identified. Clients and servers treat DUIDs as opaque values and use it compare for equality.

The DUID is not required in all the DHCP messages. It has to be unique across all DHCP clients and servers. It has to be stable for any specific client or server - for example, a device's DUID should not change as a result of a change in the *VLAN*'s port.

## DHCP Unique Identifier Content

A *DHCP* unique identifier (DUID) consists of a two-octet type code followed by a variable number of octets that make up the actual identifier. A DUID can be no more than 128 octets long (not including the type code). The following types are currently defined:

- Link-layer address plus time
- Vendor-assigned unique ID based on Enterprise Number
- Link-layer address

ExtremeXOS uses Link-layer address plus time as its default client DUID. To configure DHCPv6 client identifier type for the client use:

```
configure dhcp ipv6 client identifier-type [ link-layer | link-layer-
plus-time | vendor-specific ]
```

## Client Requested DHCPv6 Options

The following default options are encapsulated by the DHCPv6 client when requesting DHCPv6 server.

1. OPTION_CLIENTID (code 1). The Client Identifier option is used to carry a DUID identifying a client between a client and a server.
2. OPTION_SERVERID (code 2). The Server Identifier option is used to carry a DUID identifying a server between a client and a server.

   Note
   OPTION_SERVERID option is included if DHCPv6 client is responding to a particular server.

3. OPTION_IA_NA (code 3). The client uses IA_NA options to request the assignment of non-temporary address assignment.
4. OPTION_IAADDR (code 5). The IA_ADDR is used to specify the IPv6 addresses associated with IA_NA or IA_TA option.

   Note
   OPTION_IAADDR is included only if the client lease is present.

   Note
   IAID is the last 4 bytes of the hardware address

5. OPTION_ORO (code 6). The Option Request option is used to identify a list of options in a message between client and server. A client includes an Option Request option in a Solicit, Request, Renew, Rebind, Confirm or Information-request message to inform the server about options the client is interested in.

The following are the Default requested options included by the ExtremeXOS DHCPv6 client as part of Option Request Option (OPTION_ORO) option:

1. OPTION_DNS_SERVERS (code 23).
2. OPTION_DOMAIN_LIST (code 24). Reference for above ORO options: http://tools.ietf.org/html/rfc3646.
3. OPTION_ELAPSED_TIME(code 8). A client MUST include an Elapsed Time option in messages to indicate how long the client has been trying to complete a *DHCP* message exchange.

# Configuring DHCPv6 BOOTP Relay

To configure the relay function:

1. Configure *VLAN*s and IP unicast routing.

   Note
   As in DHCPv4, when you create an IPV6 VLAN interface, the corresponding disabledV6VlanList has an entry. The VLAN interface entry is removed whenever the bootpRelayv6 for the respective VLAN is enabled, and vice versa.

2. Enable the *DHCP* or BOOTP relay function using the following commands:

   ```
   enable bootprelay {ipv4|ipv6} {vlan[vlan_name] |{vr} vr_name}|all
   [{vr} vr_name]}
   enable bootprelay {vlan[vlan_name] |{vr} vr_name}|all [{vr} vr_name]}
   ```

3. Configure the addresses to which DHCP or BOOTP requests should be directed using the following command:

   ```
   configure bootreplay {ipv4} | {{vlan}vlan_name} [[add ip_address] |
   delete [ip_address | all]]] | ipv6 [[add ipv6_address] | [delete
   ipv6_address | all]]]]] {vr vrid}
   configure bootprelay add ip_address {vr vrid}
   ```

   **Note**

   Use the `configure bootprelay ipv6 option interface-id InterfaceIDName vlan {vlan_name}` command to set up a unique interface-id. It can be MAC-ID, or port-vlan combination. You can also use this command to set up dhcpv6 server/next hop for each VLAN interface, or across VR. A configuration applied to the VR level is populated to all VLAN V6 interfaces.

4. To delete an entry, use the following command:

   ```
   configure bootprelay delete ip_address {vr vrid}
   ```

5. To disable BOOTP relay on one or more VLANs, use the following command

   ```
   disable bootprelay {ipv4 | ipv6} {vlan [vlan_name] | {{vr} vr_name} |
   all [{vr} vr_name]}
   disable bootprelay {{vlan} [vlan_name] | {{vr} vr_name} | all [{vr}
   vr_name]}
   ```

   **Note**

   When *VRRP (Virtual Router Redundancy Protocol)* and BOOTP/DHCP relay are both enabled on the switch, the relayed BOOTP agent IP address is the actual switch IP address, not the virtual IP address.

## Configure Route Compression

This helps with route optimization and scaling. This feature allows you to install only less specific routes in the table when overlapping routes with the same nexthop exist. For detailed information about route compression, see Hardware Routing Table Management.

**Note**

This feature is enabled by default starting from ExtremeXOS 15.6.

- Enable this feature.

  ```
  enable iproute ipv6 compression {vr vrname}
  ```

- Disable this feature.

  ```
  disable iproute ipv6 compression {vr vrname}
  ```

## Hardware Forwarding Behavior

The following table shows the Extreme switch platforms and the ExtremeXOS software versions that provide hardware forwarding support for IPv6 unicast features.

**Table 142: IPv6 Unicast Features**

| Switch Model | ExtremeXOS Release | Features |
|---|---|---|
| Summit X450-G2 series | ExtremeXOS 16.1 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X460 series | ExtremeXOS 12.5 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X460-G2 | ExtremeXOS 15.6 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X480 series | ExtremeXOS 12.4 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X670 series | ExtremeXOS 12.6 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X670-G2 series | ExtremeXOS 15.6 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| Summit X770 series | ExtremeXOS 15.4 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| BlackDiamond 8000 e-series | ExtremeXOS 11.6 and later | IPv6 Unicast forwarding |
| BlackDiamond 8000 e-series | ExtremeXOS 12.0 and later | IPv6 tunneling |
| BlackDiamond 8000 c-series | ExtremeXOS 12.1 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| BlackDiamond 8000 xl-series | ExtremeXOS 12.4 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| BlackDiamond 8000 xm-series | ExtremeXOS 12.6 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |
| BlackDiamond X8 series | ExtremeXOS 15.1 and later | IPv6 Unicast forwarding<br>IPv6 tunneling |

### Hardware Forwarding Limitations

Summit family switches, BlackDiamond 8000 series modules, and BlackDiamond X8 series modules can use hardware forwarding when the route mask is 64 bits or less. If the route mask is greater that 64 bits, limitations apply based on the hardware platform.

BlackDiamond 8000 e-series modules, BlackDiamond 8800 c-series modules, and Summit X440, and X460 switches support hardware forwarding for up to 256 routes with masks greater than 64 bits.

This support was added in ExtremeXOS Release 12.4 by using some of the slices previously used for *ACL* support to create a Greater Than 64 Bit (GT64B) table. The GT64B table stores only those routes with a mask greater than 64 bits. When IPv6 forwarding is enabled, the switch behavior is as follows:

- Fewer slices are available for ACLs. The GT64B table consumes 1 slice on BlackDiamond 8800 c-series modules and Summit X440, and 2 slices on BlackDiamond 8000 a- and e-series modules and Summit X460 switches.

  To use the GT64B table on X440, one ACL slice must be free. Use show access-list usage acl-slice to check if any slice is unused. If no slice is available, consider disabling a feature that is consuming ACL slices if that feature is not required. Features that are enabled by default such as *IGMP (Internet Group Management Protocol)* Snooping or MLD Snooping can be disabled to free up ACL resources if not required.
- Table-full messages appear when there is no more space in the GT64B table.
- If an eligible route cannot be added to the GT64B table (because the table is full), there is no guarantee that traffic for that route will be properly routed.
- If enabled, route compression for IPv6 can make room for additional routes by reducing the number of entries in the GT64B table.
- When an IPv6 address with a mask greater that 64 bits is configured on a *VLAN* or tunnel, that address is automatically added to the GT64B table.
- BlackDiamond 8800 c-series modules do not support hardware forwarding for routes with masks greater than 64 bits on user virtual routers.

BlackDiamond 8900 xl-series modules, and Summit X480 switches support hardware forwarding for up to 245,760 routes with masks greater than 64 bits, depending on the configured setting for external-tables.

BlackDiamond 8900 c- and xm-series modules, and Summit X480, X670, X670-G2, and X770 switches support hardware forwarding for up to 256 routes with masks greater than 64 bits.

This support was added in ExtremeXOS Release 12.4 by using a hardware table designed for this purpose. When IPv6 forwarding is enabled, the switch behavior is as follows:

- If no space is available in the hardware table, there is no guarantee that traffic for that route will be properly routed.
- If enabled, route compression for IPv6 can make room for additional routes by reducing the number of entries in the hardware table.
- When an IPv6 address with a mask greater that 64 bits is configured on a VLAN or tunnel, that address is automatically added to the hardware table.

## Hardware Tunnel Support

The platforms and software versions that provide hardware forwarding support for IPv6 traffic for both IPv6-in-IPv4 and 6to4 tunnels for various platforms are shown in the following table. In all the other platforms, tunnel traffic is forwarded in software and the statistics can be viewed from the show ipstats ipv6 command.

Note

The MTU for IPv6 Tunnels is set to 1480 bytes, and is not user-configurable. Configuring jumbo frames has no effect on the MTU size for IPv6 tunnels.

## Routing Configuration Example

The figure below illustrates a BlackDiamond switch that has three VLANs defined as follows:

- Finance
  - Protocol-sensitive *VLAN* using IPv6 protocol
  - All ports on slots 1 and 3 have been assigned.
  - IP address 2001:db8:35::1/48.
- Personnel
  - Protocol-sensitive VLAN using the IPv6 protocol.
  - All ports on slots 2 and 4 have been assigned.
  - IP address 2001:db8:36::1/48.
- MyCompany
  - Port-based VLAN.
  - All ports on slots 1 through 4 have been assigned.

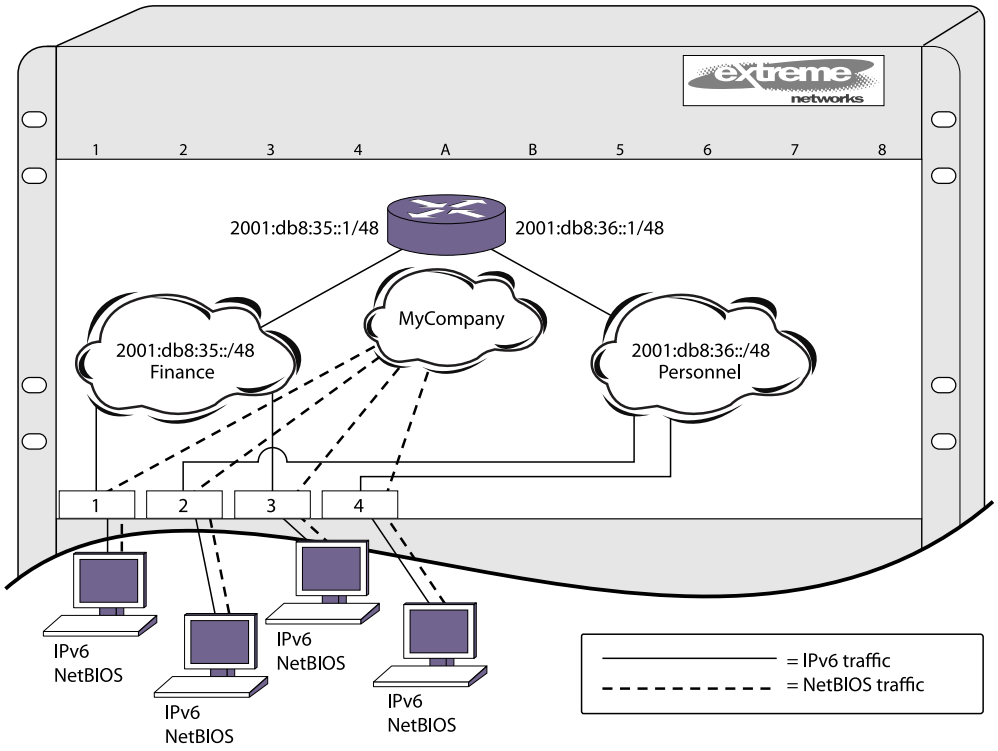



**Figure 210: IPv6 Unicast Routing Configuration Example**

The stations connected to the system generate a combination of IPv6 traffic and NetBIOS traffic. The IPv6 traffic is filtered by the protocol-sensitive VLANs. All other traffic is directed to the VLAN MyCompany.

In this configuration, all IPv6 traffic from stations connected to slots 1 and 3 have access to the router by way of the VLAN Finance. Ports on slots 2 and 4 reach the router by way of the VLAN Personnel. All other traffic (NetBIOS) is part of the VLAN MyCompany.

The example is configured as follows:

```
create vlan Finance tag 10
create vlan Personnel tag 11
create vlan MyCompany
configure Finance protocol ipv6
configure Personnel protocol ipv6
configure Finance add port 1:*,3:* tag
configure Personnel add port 2:*,4:* tag
configure MyCompany add port all
configure Finance ipaddress 2001:db8:35::1/48
configure Personnel ipaddress 2001:db8:36::1/48
configure ripng add vlan Finance
configure ripng add vlan Personnel
enable ipforwarding ipv6
enable ripng
```

# Tunnel Configuration Examples

## 6in4 Tunnel Configuration Example

The following figure illustrates a 6in4 tunnel configured between two IPv6 regions across an IPv4 region.

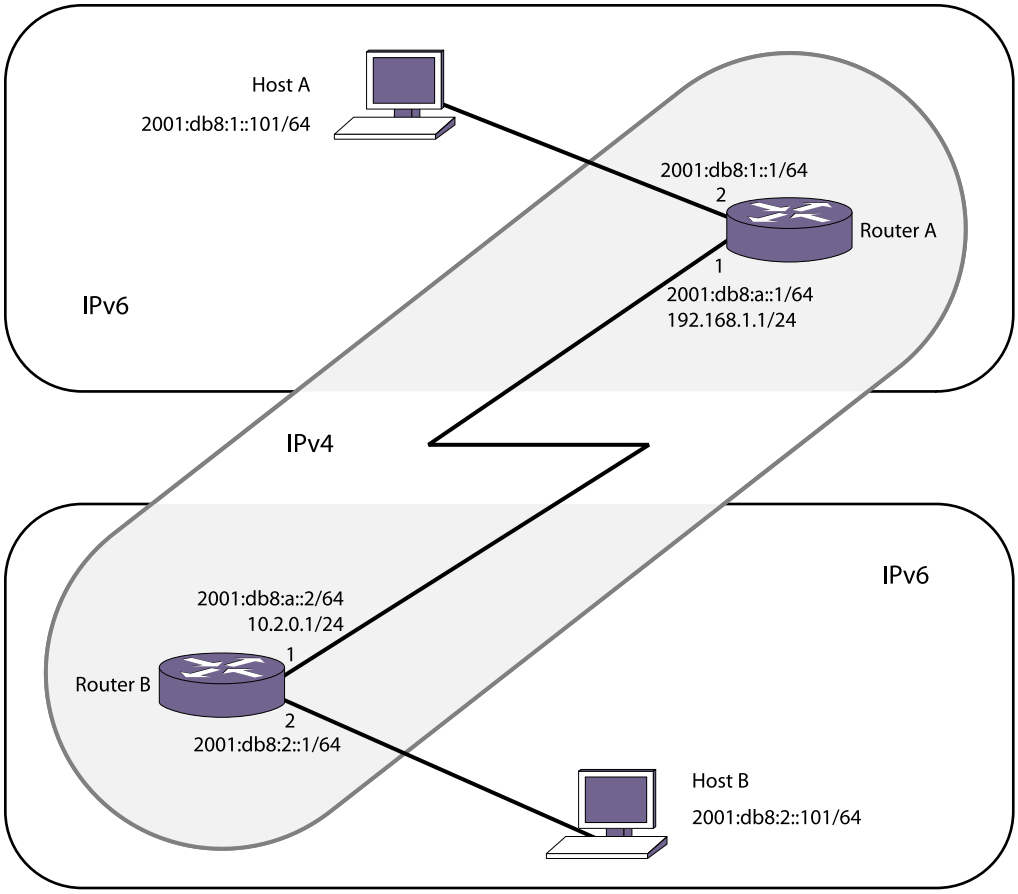



**Figure 211: 6in4 Tunnel Example**

In the following figure, Router A has an interface to an IPv4 region with the address 192.168.1.1 (for this example we are using private IPv4 addresses, but to tunnel across the Internet, you would use a public address). Router B has an IPv4 interface of 10.2.0.1. The IPv4 interface must be created before the tunnel is configured and cannot be deleted until the tunnel is deleted.

This example has one subnet in each IPv6 region, 2001:db8:1::/64 for Router A and 2001:db8:2::/64 for Router B. Hosts A and B are configured to use IPv6 addresses 2001:db8:1::101 and 2001:db8:2::101 respectively.

For traffic to move from one region to the other, there must be a route. In this example, a static route is created, but you could enable *RIPng* or *OSPFv3* on the tunnel interface.

In this example, we assume that the IPv4 network can route from Router A to Router B (in other words, some IPv4 routing protocol is running on the public-ipv4 interfaces). For platforms on which hardware based tunneling is supported (See the following table), IPv4 forwarding needs to be enabled on the tunnel source *VLAN*. However, in platforms on which IPv6-in-IPv4 tunnels are supported in software only, you do not need to enable IPv4 forwarding on the public interfaces in this example unless you are also routing IPv4 traffic on them (in this example, it is assumed you are running no IPv4 traffic inside your respective IPv6 networks, although you could).

When Host A needs to send a packet to 2001:db8:2::101 (Host B), it forwards it to Router A. Router A receives an IPv6 packet from the IPv6 source address 2001:db8:1::101 to the destination 2001:db8:2::101. Router A has the static route, for the route 2001:db8:2::/64 with next hop 2001:db8:a::2 (Router B)

through the tunnel interface. So Router A encapsulates the IPv6 packet inside an IPv4 header with the source address 192.168.1.1 and destination address 10.2.0.1. The encapsulated IPv6 packet passes through the IPv4 network and reaches the other end of the tunnel—Router B. Router B decapsulates the packet and removes the IPv4 header. Router B then forwards the IPv6 packet to the destination host—Host B.

Note

Each IPv6 packet is encapsulated inside an IPv4 header (20 bytes) before it is forwarded via a IPv6-in-IPv4 tunnel. For example, a 66-byte packet from Host A will be encapsulated and forwarded as a 86-byte packet by Router A.

*Router A*

```
configure vlan default delete port all
create vlan public-ipv4
configure vlan public-ipv4 add port 1 untagged
configure vlan public-ipv4 ipaddress 192.168.1.1/24
create tunnel public6in4 ipv6-in-ipv4 destination 10.2.0.1 source 192.168.1.1
configure tunnel public6in4 ipaddress 2001:db8:a::1/64
enable ipforwarding ipv6 public6in4
create vlan private-ipv6
configure vlan private-ipv6 add port 2 untagged
configure vlan private-ipv6 ipaddress 2001:db8:1::1/64
enable ipforwarding ipv6 private-ipv6
configure iproute add 2001:db8:2::/64 2001:db8:a::2
enable ipforwarding public-ipv4
```

*Router B*

```
configure vlan default delete port all
create vlan public-ipv4
configure vlan public-ipv4 add port 1 untagged
configure vlan public-ipv4 ipaddress 10.2.0.1/24
create tunnel public6in4 ipv6-in-ipv4 destination 192.168.1.1 source 10.2.0.1
configure tunnel public6in4 ipaddress 2001:db8:a::2/64
enable ipforwarding ipv6 public6in4
create vlan private-ipv6
configure vlan private-ipv6 add port 2 untagged
configure vlan private-ipv6 ipaddress 2001:db8:2::1/64
enable ipforwarding ipv6 private-ipv6
configure iproute add 2001:db8:1::/64 2001:db8:a::1
enable ipforwarding public-ipv4
```

## 6to4 Tunnel Configuration Example

The following figure illustrates a 6to4 tunnel configured between two IPv6 regions across an IPv4 region.

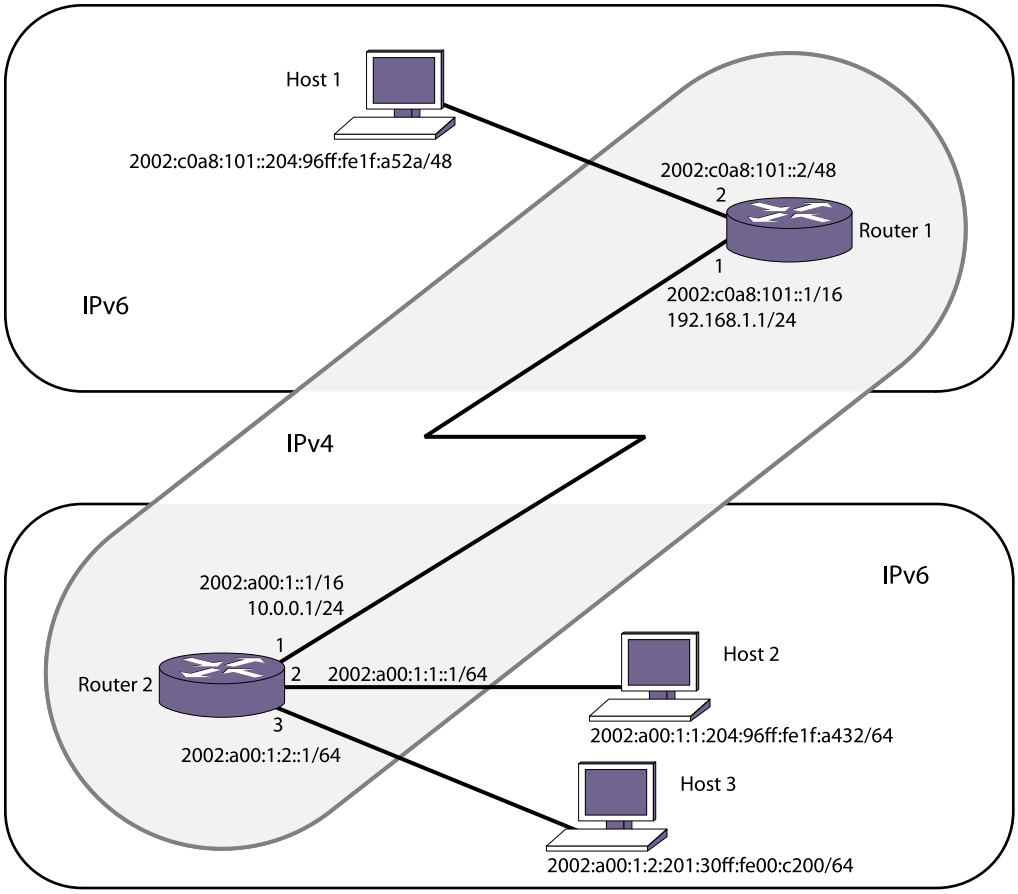



**Figure 212: 6to4 Tunnel Configuration Example**

In the following figure, Router 1 has an interface to an IPv4 region with the address 192.168.1.1 (for this example we are using private IPv4 addresses, but to tunnel across the Internet, you would use a public address). Router 2 has an IPv4 interface of 10.0.0.1. The IPv4 interface must be created before the tunnel is configured and cannot be deleted until the tunnel is deleted.

The IPv6 endpoints of 6to4 tunnels must follow the standard 6to4 address requirement. The address must be of the form 2002:<IPv4_source_endpoint>::/16, where <IPv4_source_endpoint> is replaced by the IPv4 source address of the endpoint, in hexadecimal, colon separated form. For example, for a tunnel endpoint located at IPv4 address 10.20.30.40, the tunnel address would be 2002:0a14:1e28::/16. In hex, 10 is 0a, 20 is 14, 30 is 1e and 40 is 28.

This example shows a simple setup on the Router 1 side (one big /48 IPv6 routing domain with no subnets), and a slightly more complex setup on the Router 2 side (two subnets :0001: and :0002: that are /64 in length). Hosts 1, 2, and 3 will communicate using their global 2002: addresses.

The hosts in this example configure themselves using the EUI64 interface identifier derived from their MAC addresses. Refer to your host OS vendor's documentation for configuring IPv6 addresses and routes.

In this example, we assume that the IPv4 network can route from Router 1 to Router 2 (in other words, some IPv4 routing protocol is running on the public-ipv4 interfaces). However, you do not need to enable IPv4 forwarding on the public interfaces in this example unless you are also routing IPv4 traffic

on them (in this example, it is assumed you are running no IPv4 traffic inside your respective IPv6 networks, although you could).

*Router 1*

```
configure vlan default delete port all
create vlan public-ipv4
configure vlan public-ipv4 add port 1 untagged
configure vlan public-ipv4 ipaddress 192.168.1.1/24
create tunnel public6to4 6to4 source 192.168.1.1
configure tunnel public6to4 ipaddress 2002:c0a8:0101::1/16
enable ipforwarding ipv6 public6to4
create vlan private-ipv6
configure vlan private-ipv6 add port 2 untagged
configure vlan private-ipv6 ipaddress 2002:c0a8:0101::2/48
enable ipforwarding ipv6 private-ipv6
```

*Router 2*

```
configure vlan default delete port all
create vlan public-ipv4
configure vlan public-ipv4 add port 1 untagged
configure vlan public-ipv4 ipaddress 10.0.0.1/24
create tunnel public6to4 6to4 source 10.0.0.1
configure tunnel public6to4 ipaddress 2002:0a00:0001::1/16
enable ipforwarding ipv6 public6to4
create vlan private-ipv6-sub1
configure vlan private-ipv6-sub1 add port 2 untagged
configure vlan private-ipv6-sub1 ipaddress 2002:0a00:0001:0001::1/64
enable ipforwarding ipv6 private-ipv6-sub1
create vlan private-ipv6-sub2
configure vlan private-ipv6-sub2 add port 3 untagged
configure vlan private-ipv6-sub2 ipaddress 2002:0a00:0001:0002::1/64
enable ipforwarding ipv6 private-ipv6-sub2
```

### *Host Configurations*

The IPv6 addresses of these hosts are based on their MAC address-derived EUI64 interface identifiers and the address prefixes for their subnets. Each host must also have a static route configured on it for 6to4 addresses.

#### Host 1:

- MAC address—00:04:96:1F:A5:2A
- IPv6 address—2002:c0a8:0101::0204:96ff:fe1f:a52a/48
- Static route—destination 2002::/16, gateway 2002:c0a8:0101::2

#### Host 2:

- MAC address—00:04:96:1F:A4:32
- IP address—2002:0a00:0001:0001:0204:96ff:fe1f:a432/64
- Static route—destination 2002::/16, gateway 2002:0a00:0001:0001::1

#### Host 3:

- MAC address—00:01:30:00:C2:00
- IP address—2002:0a00:0001:0002:0201:30ff:fe00:c200/64

- Static route—destination 2002::/16, gateway 2002:0a00:0001:0002::1

# GRE Tunnel Configuration Example

## Router 1 Configuration

### Summit Configuration

```
unconfigure switch all
configure default del port all
create vlan inet
configure vlan inet add port 24
configure vlan inet ipa 1.1.1.1/24
create vlan users
configure vlan users add port 1
configure vlan users ipa 100.0.0.1/24
create tunnel mytunnel gre destination 1.1.1.2 source 1.1.1.1
configure tunnel "mytunnel" ipaddress 2.0.0.1/24
configure iproute add 200.0.0.0/24 2.0.0.2
enable ipforwarding
```

### BD8K Configuration

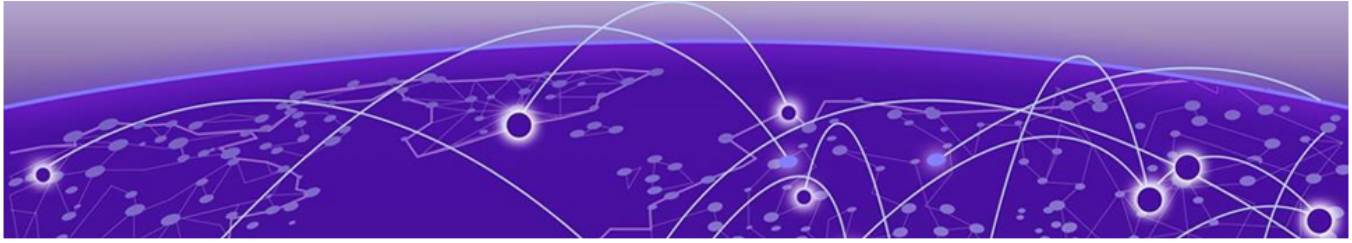
```
unconfigure switch all
configure default del port all
create vlan inet
configure vlan inet add port 10:1
configure vlan inet ipa 1.1.1.2/24
create vlan users
configure vlan users add port 10:2
configure vlan users ipa 200.0.0.1/24
create tunnel mytunnel gre destination 1.1.1.1 source 1.1.1.2
configure tunnel "mytunnel" ipaddress 2.0.0.2/24
configure iproute add 100.0.0.0/24 2.0.0.1
enable ipforwarding
```

## Router 2 Configuration

```
create vlan inet
configure vlan inet add port 10:1
configure vlan inet ipa 1.1.1.2/24
create vlan users
configure vlan users add port 10:2
configure vlan users ipa 200.0.0.1/24
create tunnel mytunnel gre destination 1.1.1.1 source 1.1.1.2
configure tunnel "mytunnel" ipaddress 2.0.0.2/24
configure iproute add 100.0.0.0/24 2.0.0.1
enable ipforwarding

Host IP connected to router 1-100.0.0.1
Host IP connected to router 2-200.0.0.1
```

**Note**

If a transit node is present between the tunneling sites, you should configure static or enable IBGP between the sites for network connectivity.



# RIP



This chapter assumes that you are already familiar with IP unicast routing.

If not, refer to the following publications for additional information:

- RFC 1058—Routing Information Protocol (RIP)
- RFC 1723—RIP Version 2
- *Interconnections: Bridges and Routers* , by Radia Perlman. ISBN 0-201-56332-0. Published by Addison-Wesley Publishing Company.

Note
RIP is available on platforms with an Edge, Advanced Edge or Core license. For specific information regarding RIP licensing, see the *Feature License Requirements* document.

## IGPs Overview

The switch supports the use of the following interior gateway protocols (IGPs):

- Routing Information Protocol (*RIP (Routing Information Protocol)*)
- *OSPF (Open Shortest Path First)*
- Intermediate System-Intermediate System (IS-IS)

RIP is a distance-vector protocol, based on the Bellman-Ford (or distance-vector) algorithm. The distance-vector algorithm has been in use for many years and is widely deployed and understood.

OSPF and IS-IS are link-state protocols, based on the Dijkstra link-state algorithm. OSPF and IS-IS are newer IGPs and solve a number of problems associated with using RIP on today's complex networks.

Note
RIP can be enabled on a *VLAN (Virtual LAN)* with either OSPF or IS-IS. OSPF and IS-IS cannot be enabled on the same VLAN.

RIP is described in this chapter, OSPF is described in OSPF on page 1341 and IS-IS is described in IS-IS on page 1368.

## RIP Versus OSPF and IS-IS

The distinction between *RIP* and the *OSPF* and IS-IS link-state protocols lies in the fundamental differences between distance-vector protocols and link-state protocols.

Using a distance-vector protocol, each router creates a unique routing table from summarized information obtained from neighboring routers. Using a link-state protocol, every router maintains an identical routing table created from information obtained from all routers in the autonomous system (AS). Each router builds a shortest path tree, using itself as the root. The link-state protocol ensures that updates sent to neighboring routers are acknowledged by the neighbors, verifying that all routers have a consistent network map.

## Advantages of RIP, OSPF, and IS-IS

The biggest advantage of using *RIP* is that it is relatively simple to understand and to implement, and it has been the de facto routing standard for many years.

RIP has a number of limitations that can cause problems in large networks, including the following:

- A limit of 15 hops between the source and destination networks
- A large amount of bandwidth taken up by periodic broadcasts of the entire routing table
- Slow convergence
- Routing decisions based on hop count; no concept of link costs or delay
- Flat networks; no concept of areas or boundaries

*OSPF* and IS-IS offer many advantages over RIP, including the following:

- No limitation on hop count
- Route updates multicast only when changes occur
- Faster convergence
- Support for load balancing to multiple routers based on the actual cost of the link
- Support for hierarchical topologies where the network is divided into areas

The details of RIP are explained later in this chapter.

# Overview of RIP

*RIP* is an IGP first used in computer routing in the Advanced Research Projects Agency Network (ARPAnet) as early as 1969. It is primarily intended for use in homogeneous networks of moderate size.

To determine the best path to a distant network, a router using RIP always selects the path that has the least number of hops. Each router that data must traverse is considered to be one hop.

## Routing Table

The routing table in a router using *RIP* contains an entry for every known destination network.

Each routing table entry contains the following information:

- IP address of the destination network
- Metric (hop count) to the destination network

- IP address of the next router
- Timer that tracks the amount of time since the entry was last updated

The router exchanges an update message with each neighbor every 30 seconds (default value), or when there is a change to the overall routed topology (also called triggered updates). If a router does not receive an update message from its neighbor within the route timeout period (180 seconds by default), the router assumes the connection between it and its neighbor is no longer available.

## Split Horizon

Split horizon is a scheme for avoiding problems caused by including routes in updates sent to the router from which the route was learned.

Split horizon omits routes learned from a neighbor in updates sent to that neighbor.

## Poison Reverse

Like split horizon, poison reverse is a scheme for eliminating the possibility of loops in the routed topology.

In this case, a router advertises a route over the same interface that supplied the route, but the route uses a hop count of 16, which defines that router as unreachable.

## Triggered Updates

Triggered updates occur whenever a router changes the metric for a route.

The router is required to send an update message immediately, even if it is not yet time for a regular update message to be sent. This generally results in faster convergence, but may also result in more *RIP*-related traffic.

## Route Advertisement of VLANs

Virtual LANs (VLANs) that are configured with an IP address but are configured to not route IP or are not configured to run *RIP*, do not have their subnets advertised by RIP.

RIP advertises only those VLANs that are configured with an IP address, are configured to route IP, and run RIP.

## RIP Version 1 Versus RIP Version 2

A new version of *RIP*, called RIP version 2, expands the functionality of RIP version 1 to include the following:

- Variable-length subnet masks (VLSMs).
- Support for next-hop addresses, which allows for optimization of routes in certain environments.
- Multicasting.

RIP version 2 packets can be multicast instead of being broadcast, reducing the load on hosts that do not support routing protocols.

Note

If you are using RIP with supernetting/Classless Inter-Domain Routing (CIDR), you must use RIPv2 only.

## Route Redistribution

More than one routing protocol can be enabled simultaneously on the switch.

Route redistribution allows the switch to exchange routes, including static routes, between the routing protocols. The following figure is an example of route redistribution between an *OSPF* AS and a *RIP* AS.

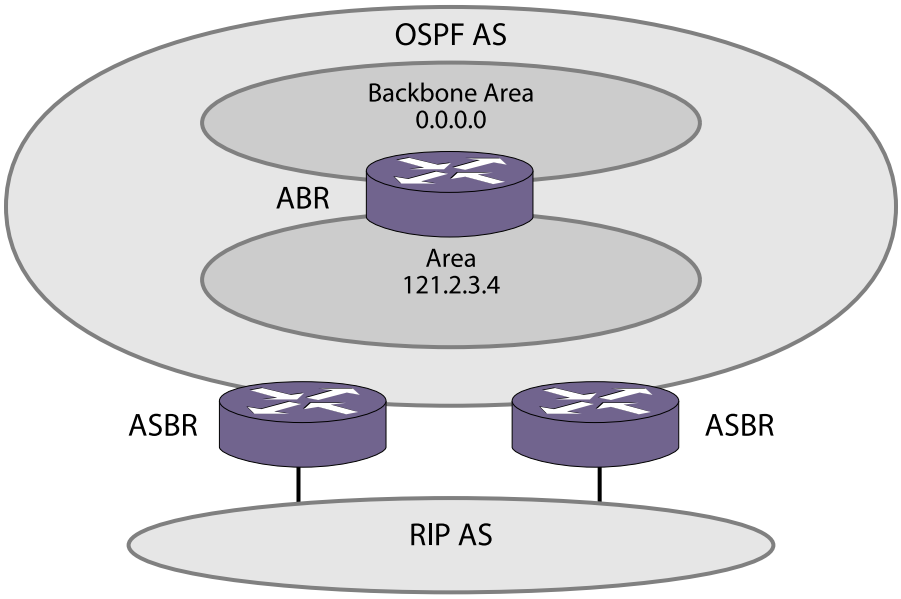



**Figure 213: Route Redistribution**

### Configuring Route Redistribution

Exporting routes from one protocol to another and from that protocol to the first one are discrete configuration functions. For example, to run *OSPF* and *RIP* simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from OSPF to RIP and the routes to export from RIP to OSPF. Likewise, for any other combinations of protocols, you must separately configure each to export routes to the other.

*Redistribute Routes into RIP*

- Enable or disable the exporting of static, direct, *BGP (Border Gateway Protocol)*-learned, and *OSPF*-learned routes into the *RIP* domain.

  `enable rip export [bgp | direct | e-bgp | i-bgp | ospf | ospf-extern1 | ospf-extern2 | ospf-inter | ospf-intra | static | isis | isis-level-1| isis-level-1-external | isis-level-2 | isis-level-2-external ] [cost number {tag number} | policy policy-name]`

```
disable rip export [bgp | direct | e-bgp | i-bgp | ospf | ospf-extern1
| ospf-extern2 | ospf-inter | ospf-intra | static | isis | isis-
level-1| isis-level-1-external | isis-level-2 | isis-level-2-
external ]
```

These commands enable or disable the exporting of static, direct, and OSPF-learned routes into the RIP domain.

- You can choose which types of OSPF routes are injected, or you can simply choose ospf, which will inject all learned OSPF routes regardless of type.

  The default setting is disabled.

## RIP Configuration Example

The following figure illustrates a BlackDiamond switch that has three VLANs defined as follows:

- Finance
  - Protocol-sensitive *VLAN* using the IP protocol.
  - All ports on slots 1 and 3 have been assigned.
  - IP address 192.207.35.1.
- Personnel
  - Protocol-sensitive VLAN using the IP protocol.
  - All ports on slots 2 and 4 have been assigned.
  - IP address 192.207.36.1.
- MyCompany
  - Port-based VLAN.
  - All ports on slots 1 through 4 have been assigned.

This example does use protocol-sensitive VLANs that admit only IP packets. This is not a common requirement for most networks. In most cases, VLANs will admit different types of packets to be forwarded to the hosts and servers on the network.

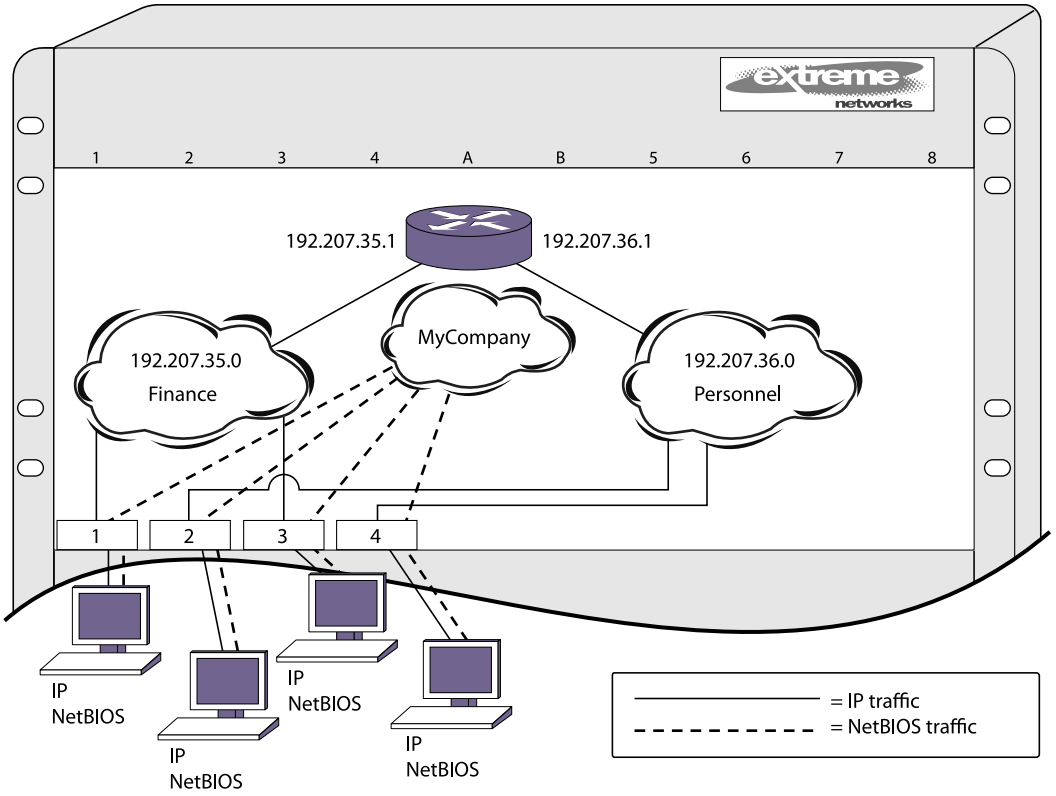



**Figure 214: RIP Configuration Example**

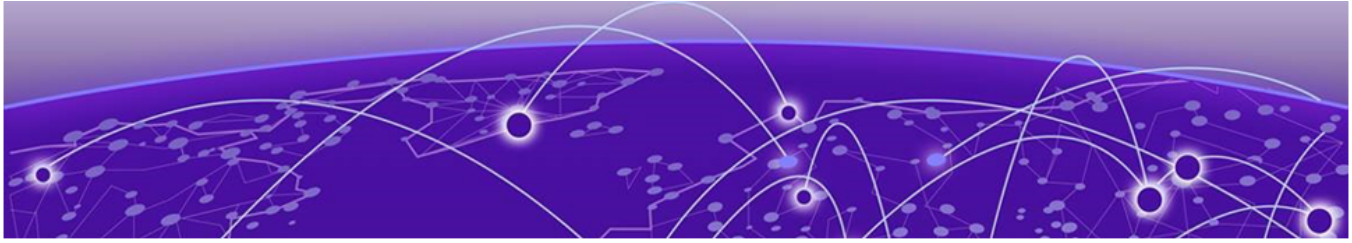
The stations connected to the system generate a combination of IP traffic and NetBIOS traffic. The IP traffic is filtered by the protocol-sensitive VLANs. All other traffic is directed to the VLAN MyCompany.

In this configuration, all IP traffic from stations connected to slots 1 and 3 have access to the router by way of the VLAN Finance. Ports on slots 2 and 4 reach the router by way of the VLAN Personnel. All other traffic (NetBIOS) is part of the VLAN MyCompany.

The example in the following figure is configured as follows:

```
create vlan Finance
create vlan Personnel
create vlan MyCompany
configure Finance protocol ip
configure Personnel protocol ip
configure Finance add port 1:*,3:*
configure Personnel add port 2:*,4:*
configure MyCompany add port all
configure Finance ipaddress 192.207.35.1
configure Personnel ipaddress 192.207.36.1
enable ipforwarding
configure rip add vlan all
enable rip
```

More commonly, NetBIOS traffic would be allowed on the Finance and Personnel VLANs, but this example shows how to exclude that traffic. To allow the NetBIOS traffic (or other type of traffic) along with the IP traffic, remove the configure finance protocol ip and configure Personnel protocol ip commands from the example.



# RIPng



This chapter describes the RIPng interior gateway protocol for IPv6 networks. It provides the various protocol schemes available, commands for configuring redistribution, and an example for configuring multiple protocols on one switch. This chapter assumes that you are already familiar with IP unicast routing. If not, refer to the following publication for additional information:

- RFC 2080—RIPng for IPv6

Note
RIPng is available on platforms with an Edge, Advanced Edge or Core license. For specific information regarding RIPng licensing. see the *Feature License Requirements* document.

## RIPng Overview

*RIPng (Routing Information Protocol Next Generation)* is an interior gateway protocol (IGP) developed for IPv6 networks.

The analogous protocol used in IPv4 networks is called Routing Information Protocol (*RIP (Routing Information Protocol)*). Like RIP, RIPng is a relatively simple protocol for the communication of routing information among routers. Many concepts and features of RIPng are directly parallel to those same features in RIP.

RIPng is a distance-vector protocol, based on the Bellman-Ford (or distance-vector) algorithm. The distance-vector algorithm has been in use for many years and is widely deployed and understood. The other common IGPs for IPv6 are (*OSPFv3 (Open Shortest Path First version 3)*) and Intermediate System-Intermediate System (IS-IS), which are link-state protocols.

Note
RIPng can be enabled on a *VLAN (Virtual LAN)* with either OSPFv3 or IS-IS. OSPFv3 and ISIS cannot be enabled on the same VLAN.

### RIPng versus OSPFv3 and IS-IS

The distinction between *RIPng* and the link-state protocols (*OSPFv3* and IS-IS) lies in the fundamental differences between distance-vector protocols (RIPng) and link-state protocols.

Using a distance-vector protocol, each router creates a unique routing table from summarized information obtained from neighboring routers. Using a link-state protocol, every router maintains an identical routing table created from information obtained from all routers in the autonomous system. Each router builds a shortest path tree, using itself as the root. The link-state protocol ensures that updates sent to neighboring routers are acknowledged by the neighbors, verifying that all routers have a consistent network map.

### Advantages of RIPng, OSPFv3, and IS-IS

The biggest advantage of using *RIPng* is that it is relatively simple to understand and implement, and it has been the de facto routing standard for many years.

RIPng has a number of limitations that can cause problems in large networks, including the following:

- A limit of 15 hops between the source and destination networks
- A large amount of bandwidth taken up by periodic broadcasts of the entire routing table
- Slow convergence
- Routing decisions based on hop count; no concept of link costs or delay
- Flat networks; no concept of areas or boundaries

*OSPFv3* and IS-IS offer many advantages over RIPng, including the following:

- No limitation on hop count
- Route updates multicast only when changes occur
- Faster convergence
- Support for load balancing to multiple routers based on the actual cost of the link
- Support for hierarchical topologies where the network is divided into areas

The details of RIPng are explained later in this chapter.

## RIPng Routing

*RIPng* is primarily intended for use in homogeneous networks of moderate size.

To determine the best path to a distant network, a router using RIPng always selects the path that has the least number of hops. Each router that data must traverse is considered to be one hop.

### Routing Table

The routing table in a router using *RIPng* contains an entry for every known destination network.

Each routing table entry contains the following information:

- IP address and prefix length of the destination network
- Metric (hop count) to the destination network
- IP address of the next hop router, if the destination is not directly connected
- Interface for the next hop
- Timer that tracks the amount of time since the entry was last updated

- A flag that indicates if the entry is a new one since the last update
- The source of the route, for example, static, RIPng, *OSPFv3*, etc.

The router exchanges an update message with each neighbor every 30 seconds (default value), or when there is a change to the overall routed topology (also called *triggered updates*). If a router does not receive an update message from its neighbor within the route timeout period (180 seconds by default), the router assumes the connection between it and its neighbor is no longer available.

### Split Horizon

Split horizon is a scheme for avoiding problems caused by including routes in updates sent to the router from which the route was learned.

Split horizon omits routes learned from a neighbor in updates sent to that neighbor.

### Poison Reverse

Like split horizon, poison reverse is a scheme for eliminating the possibility of loops in the routed topology.

In this case, a router advertises a route over the same interface that supplied the route, but the route uses a hop count of 16, which defines that router as unreachable.

### Triggered Updates

Triggered updates occur whenever a router changes the metric for a route.

The router is required to send an update message immediately, even if it is not yet time for a regular update message to be sent. This generally results in faster convergence, but may also result in more *RIPng*-related traffic.

### Route Advertisement of VLANs

*RIP* advertises only those VLANs that are configured with an IP address, are configured to route IP, and run RIP.

## Route Redistribution

More than one routing protocol can be enabled simultaneously on the switch. Route redistribution allows the switch to exchange routes, including static routes, between the routing protocols. Route redistribution is also called *route export*.

### Configuring Route Redistribution

Exporting routes from one protocol to another and from that protocol to the first one are discrete configuration functions.

For example, to run *OSPFv3* and *RIPng* simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from OSPFv3 to

RIPng and the routes to export from RIPng to OSPFv3. Likewise, for any other combinations of protocols, you must separately configure each to export routes to the other.

*Redistributing Routes into RIPng*

- Eable or disable the exporting of static, direct, or other protocol-learned routes into the *RIPng* domain using the following commands:

  `enable ripng export [direct | ospfv3 | ospfv3-extern1 | ospfv3-extern2 | ospfv3-inter | ospfv3-intra | static | isis | isis-level-1| isis-level-1-external | isis-level-2| isis-level-2-external | bgp] [cost number {tag number} | policy policy-name]`

  `disable rip originate-default`

  `enable ripng originate-defalut {always} {cost value}`

  These commands enable or disable the exporting of static, direct, and *OSPF (Open Shortest Path First)*-learned routes into the RIPng domain.
- You can choose which types of OSPF routes are injected, or you can simply choose ospf, which will inject all learned OSPF routes regardless of type. The default setting is disabled.

## RIPng Configuration Example

The following configuration is similar to the example in the *RIP* chapter, but uses IPv6 addresses. It illustrates a BlackDiamond switch that has three VLANs defined as follows:

- Finance
  - All ports on slots 1 and 3 have been assigned.
  - IP address 2001:db8:35::1/48.
- Personnel
  - All ports on slots 2 and 4 have been assigned.
  - IP address 2001:db8:36::1/48.
- MyCompany
  - Port-based *VLAN*.
  - All ports on slots 1 through 4 have been assigned.

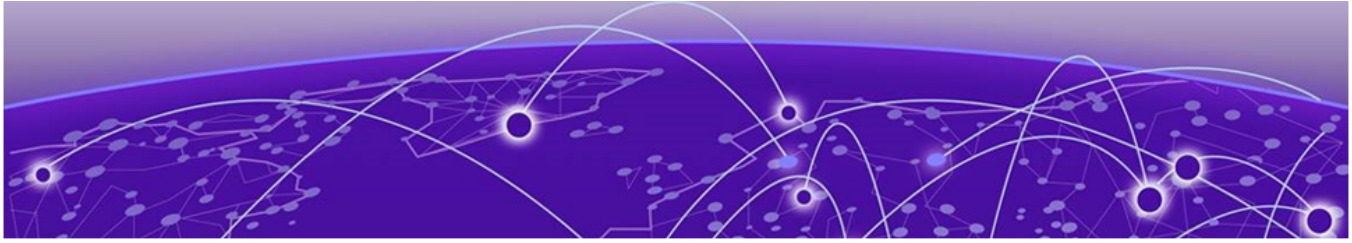
The stations connected to the system generate a combination of IPv6 traffic and NetBIOS traffic.

In this configuration, all traffic from stations connected to slots 1 and 3 have access to the router by way of the VLAN Finance. Ports on slots 2 and 4 reach the router by way of the VLAN Personnel. All traffic (NetBIOS and IPv6) is part of the VLAN MyCompany.

The example is configured as follows:

```
create vlan Finance
create vlan Personnel
create vlan MyCompany
configure Finance add port 1:*,3:*
configure Personnel add port 2:*,4:*
configure MyCompany add port all
configure Finance ipaddress 2001:db8:35::1/48
configure Personnel ipaddress 2001:db8:36::1/48
enable ipforwarding ipv6
configure ripng add vlan Finance
```

```
configure ripng add vlan Personnel
enable ripng
```



# OSPF



This chapter discusses the Open Shortest Path First (OSPF) protocol for distributing routing information between routers belonging to an autonomous system. This chapter provides an overview of the protocol's features and example configuration commands.

This chapter assumes that you are already familiar with IP unicast routing. If not, refer to the following publications for additional information:

- RFC 2328—OSPF Version 2
- RFC 1765—OSPF Database Overflow
- RFC 2370—The OSPF Opaque LSA Option
- RFC 3101—The OSPF Not-So-Stubby Area (NSSA) Option
- RFC 3623—Graceful OSPF Restart
- *Interconnections: Bridges and Routers* by Radia Perlman. Published by Addison-Wesley Publishing Company (ISBN 0-201-56332-0).

Note
OSPF is available on platforms with an Advanced Edge or Core license. For specific information regarding OSPF licensing, see the *Feature License Requirements* document.

## OSPF Overview

*OSPF (Open Shortest Path First)* is a link state protocol that distributes routing information between routers belonging to a single IP domain; the IP domain is also known as an autonomous system (AS).

In a link-state routing protocol, each router maintains a database describing the topology of the AS. Each participating router has an identical database maintained from the perspective of that router.

From the LSDB (link state database), each router constructs a tree of shortest paths, using itself as the root. The shortest path tree provides the route to each destination in the AS. When several equal-cost routes to a destination exist, traffic can be distributed among them. The cost of a route is described by a single metric.

OSPF is an IGP (interior gateway protocol), as is *RIP (Routing Information Protocol)*, the other common IGP. OSPF and RIP are compared in RIP on page 1330.

Note

Two types of OSPF functionality are available and each has a different licensing requirement. One is the complete OSPF functionality and the othesee the *Feature License Requirements* document.r is OSPF Edge Mode, a subset of OSPF that is described below. For specific information regarding OSPF licensing,

## OSPF Edge Mode

*OSPF* Edge Mode is a subset of OSPF available on platforms with an Advanced Edge license.

There are two restrictions on OSPF Edge Mode:

- At most, four Active OSPF *VLAN (Virtual LAN)* interfaces are permitted. There is no restriction on the number of Passive interfaces.
- The OSPF Priority on VLANs is 0, and is not configurable. This prevents the system from acting as a DR or BDR.

## BFD for OSPF

The BFD for *OSPF* feature gives OSPF routing protocol the ability to utilize BFD's fast failure detection to monitor OSPF neighbor adjacencies. CLI commands are provided to configure BFD protection for OSPF, so that as a registered BFD client, OSPF can request BFD protection for interested OSPF neighbors, and receive notifications about BFD session setup status and BFD session status updates (after establishing) from the BFD server. When BFD detects a communication failure between neighbors, it informs OSPF, which causes the OSPF neighbor state be marked as "down." This allows OSPF protocol to quickly begin network convergence and use alternate paths to the affected neighbor.

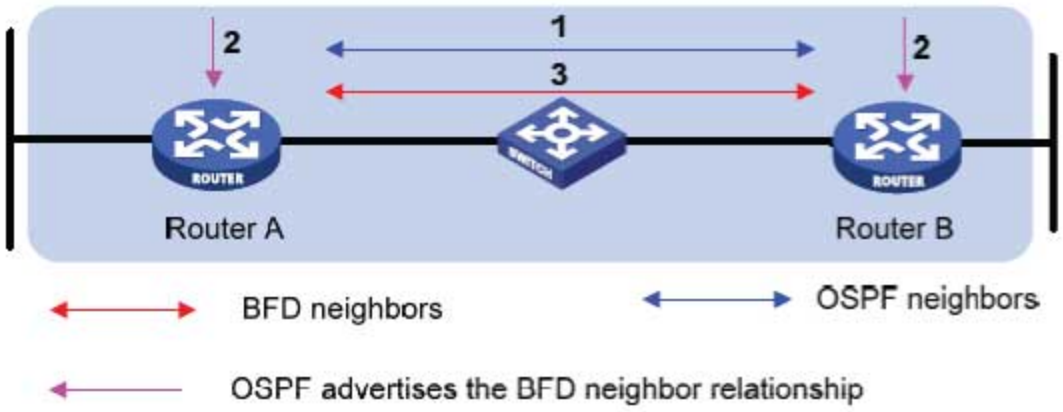



**Figure 215: Basic Operational Flow of OSPF BFD**

*Establishing a BFD Session for OSPF Neighbor*

1. OSPF discovers a neighbor.
2. If BFD for OSPF is configured, OSPF on both routers sends a request to the local BFD server to initiate a BFD neighbor session with the OSPF neighbor router.

3. The BFD neighbor session with the OSPF neighbor router is established on both sides if BFD session limit is not reached.

If the BFD session limit is reached, the OSPF neighbor will be marked as BFD session failed if synchronous request is used, or pending if asynchronous request is used, and the BFD server will send an asynchronous notification when the session registration passes later. (The asynchronous request is not available until the BFD client session create API is enhanced.)

*Eliminating the OSPF Neighbor Relationship by BFD Fault Detection*

1. A failure occurs in the network.
2. The BFD neighbor session with the OSPF neighbor router is removed because the BFD timer expired.
3. On both Router A and Router B, BFD notifies the local OSPF process that the BFD neighbor is DOWN.
4. The local OSPF process tears down the OSPF neighbor relationship by marking neighbor state DOWN. (If an alternative path is available, the routers will immediately start converging on it.)

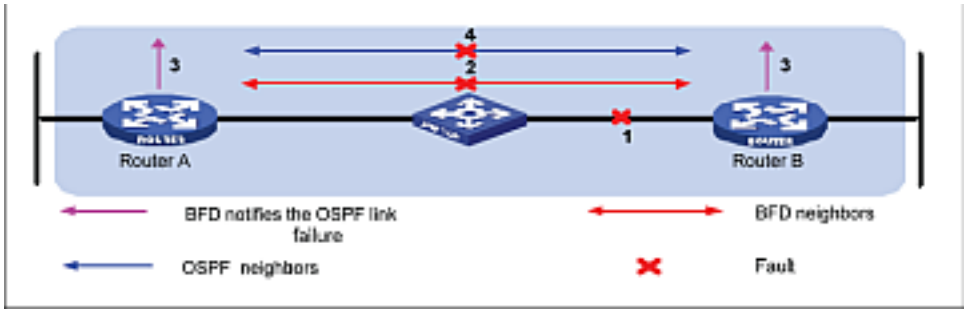



**Figure 216: OSPF Neighbor Relationship Eliminated by BFD Fault Detection**

OSPF will request the BFD server to delete the BFD session for OSPF neighbor moved to DOWN state. When the link failure is later resolved, OSPF needs to register the re-discovered neighbor to BFD again to initiate BFD session creation.

*Removing BFD Protection for OSPF*

1. BFD protection is removed from OSPF interface on Router A, OSPF process requests BFD server to delete all BFD sessions for neighbors learned on that interface.
2. BFD server will stop sending session status update to local OSPF process on Router A.
3. Before actually deleting any sessions, BFD server on Router A will first notify Router B to mark those session status as "Admin Down," which will cause Router B to stop using BFD protection for those OSPF neighbors.
4. When BFD protection is removed from OSPF interface on Router B, BFD sessions can be deleted immediately since they are already in "Admin Down" state.

When BFD for OSPF is configured on broadcast interface, the default behavior is to register only OSPF neighbors in FULL state with the BFD server. Separate BFD sessions are created for each neighbor learned on the same interface. If multiple clients ask for the same neighbor on the same interface, then single BFD sessions are established between the peers.

*OSPF Neighbor State Determination*

With active BFD protection, OSPF combines the BFD session state with the associated interface admin and operational states to determine the OSPF neighbor adjacency discovered on that OSPF interface. Regarding OSPF neighbor relationships, OSPF reacts directly to BFD session state changes only in the following circumstances:

- If BFD is enabled on the interface, and
- If BFD for OSPF is configured on the OSPF interface, and
- If a BFD session has been established to the neighbor, and
- If the BFD session has passed the INIT_COMPLETE state then:

1. The OSPF neighbor relationship will remain as FULL if the operational state of the BFD session is "UP" and the operational state of the associated *VLAN* interface is UP.
2. The OSPF neighbor relationship will be considered as DOWN if the operational state of the BFD session is DOWN or the operational state of the associated VLAN interface is DOWN.

In all other cases, the BFD session state is not considered as part of the reported OSPF neighbor state, and the OSPF neighbor state reverts to the operational state of the OSPF interface only. When the BFD session is in ADMIN_DOWN state, OSPF ignores BFD events and OSPF neighbor adjacency is not be affected by the BFD session state change.

## Link State Database

Upon initialization, each router transmits a link state advertisement (LSA) on each of its interfaces. LSAs are collected by each router and entered into the LSDB of each router. After all LSAs are received, the router uses the link state database (LSDB) to calculate the best routes for use in the IP routing table. *OSPF* uses flooding to distribute LSAs between routers. Any change in routing information is sent to all of the routers in the network. All routers within an area have the exact same LSDB.

The following table describes LSA type numbers.

**Table 143: LSA Type Numbers**

| Type Number | Description |
|---|---|
| 1 | Router LSA |
| 2 | Network LSA |
| 3 | Summary LSA |
| 4 | AS summary LSA |
| 5 | AS external LSA |
| 7 | NSSA external LSA |
| 9 | Link local—Opaque |
| 10 | Area scoping—Opaque |
| 11 | AS scoping—Opaque |

### *Database Overflow*

The *OSPF* database overflow feature allows you to limit the size of the LSDB and to maintain a consistent LSDB across all the routers in the domain, which ensures that all routers have a consistent view of the network.

Consistency is achieved by:

- Limiting the number of external LSAs in the database of each router.
- Ensuring that all routers have identical LSAs.
- To configure OSPF database overflow, use the following command:

  ```
  configure ospf ase-limit number {timeout seconds}
  ```

  Where:

  - *number*—Specifies the number of external LSAs that the system supports before it goes into overflow state. A limit value of 0 disables the functionality. When the LSDB size limit is reached, OSPF database overflow flushes LSAs from the LSDB. OSPF database overflow flushes the same LSAs from all the routers, which maintains consistency.
  - **timeout**—Specifies the timeout, in seconds, after which the system ceases to be in overflow state. A timeout value of 0 leaves the system in overflow state until OSPF is disabled and re-enabled.

### *Opaque LSAs*

Opaque LSAs are a generic *OSPF* mechanism used to carry auxiliary information in the OSPF database. Opaque LSAs are most commonly used to support OSPF traffic engineering.

Normally, support for opaque LSAs is autonegotiated between OSPF neighbors.

- In the event that you experience interoperability problems, you can disable opaque LSAs across the entire system using the following command:

  ```
  disable ospf capability opaque-lsa
  ```

- Re-enable opaque LSAs across the entire system:

  ```
  enable ospf capability opaque-lsa
  ```

  If your network uses opaque LSAs, we recommend that all routers on your OSPF network support opaque LSAs. Routers that do not support opaque LSAs do not store or flood them. At minimum a well interconnected subsection of your OSPF network must support opaque LSAs to maintain reliability of their transmission.

### *Graceful OSPF Restart*

RFC 3623 describes a way for *OSPF* control functions to restart without disrupting traffic forwarding.

Without graceful restart, adjacent routers will assume that information previously received from the restarting router is stale and will not be used to forward traffic to that router. However, in many cases, two conditions exist that allow the router restarting OSPF to continue to forward traffic correctly. The first condition is that forwarding can continue while the control function is restarted. Most modern router system designs separate the forwarding function from the control function so that traffic can still be forwarded independent of the state of the OSPF function. Routes learned through OSPF remain in the routing table and packets continue to be forwarded. The second condition required for graceful restart is that the network remain stable during the restart period. If the network topology is not

changing, the current routing table remains correct. Often, networks can remain stable during the time for restarting OSPF.

#### Restarting and Helper Mode

Routers involved with graceful restart fill one of two roles: the restarting router or the helper router.

With graceful restart, the router that is restarting sends out Grace-LSAs informing its neighbors that it is in graceful restart mode, how long the helper router should assist with the restart (the grace period), and why the restart occurred. If the neighboring routers are configured to help with the graceful restart (helper-mode), they will continue to advertise the restarting router as if it was fully adjacent. Traffic continues to be routed as though the restarting router is fully functional. If the network topology changes, the helper routers will stop advertising the restarting router. The helper router will continue in helper mode until the restarting router indicates successful termination of graceful restart, the Grace-LSAs expire, or the network topology changes. A router can be configured for graceful restart, and for helper-mode separately. A router can be a helper when its neighbor restarts, and can in turn be helped by a neighbor if it restarts.

#### Planned and Unplanned Restarts

Two types of graceful restarts are defined: planned and unplanned.

A planned restart would occur if the software module for *OSPF* was upgraded, or if the router operator decided to restart the OSPF control function for some reason. The router has advance warning, and is able to inform its neighbors in advance that OSPF is restarting.

An unplanned restart would occur if there was some kind of system failure that caused a remote reboot or a crash of OSPF, or an MSM/MM failover occurs. As OSPF restarts, it informs its neighbors that it is in the midst of an unplanned restart.

You can decide to configure a router to enter graceful restart for only planned restarts, for only unplanned restarts, or for both. Also, you can separately decide to configure a router to be a helper for only planned, only unplanned, or for both kinds of restarts.

#### Configuring Graceful OSPF Restart

- Configure a router to perform graceful *OSPF* restart:

  `configure ospf restart [`**`none`** | **`planned`** | **`unplanned`** | **`both`** | **`aware-only`**`]`

  Since a router can act as a restart helper router to multiple neighbors, you will specify which neighbors to help.
- Configure a router to act as a graceful OSPF restart helper:

  `configure ospf [`**`vlan`** `[`**`all`** | *`vlan-name`*`] | `**`area`** *`area-identifier`* | **`virtual-link`** *`router-identifier area-identifier`*`]` **`restart-helper`** `[`**`none`** | **`planned`** | **`unplanned`** | **`both`**`]`
- The graceful restart period sent out to helper routers can be configured with the following command:

  `configure ospf restart grace-period` *`seconds`*

  By default, a helper router will terminate graceful restart if received LSAs would affect the restarting router.

This will occur when the restart-helper receives an LSA that will be flooded to the restarting router or when there is a changed LSA on the restarting router's retransmission list when graceful restart is initiated.

- Disable this behavior:
  `disable ospf [vlan [all {vr vrf_name}| vlan-name] | area area-identifier | virtual-link router-identifier area-identifier] restart-helper-lsa-check`

## Areas

*OSPF* allows parts of a network to be grouped together into areas.

The topology within an area is hidden from the rest of the AS. Hiding this information enables a significant reduction in LSA traffic and reduces the computations needed to maintain the LSDB. Routing within the area is determined only by the topology of the area.

The three types of routers defined by OSPF are as follows:

- Internal router (IR)—An internal router has all of its interfaces within the same area.
- Area border router (ABR)—An ABR has interfaces in multiple areas. It is responsible for exchanging summary advertisements with other ABRs.
- Autonomous system border router (ASBR)—An ASBR acts as a gateway between OSPF and other routing protocols, or other autonomous systems.

### *Backbone Area (Area 0.0.0.0)*

Any *OSPF* network that contains more than one area is required to have an area configured as area 0.0.0.0, also called the backbone. All areas in an AS must be connected to the backbone. When designing networks, you should start with area 0.0.0.0 and then expand into other areas.

Note
Area 0.0.0.0 exists by default and cannot be deleted or changed.

The backbone allows summary information to be exchanged between area border routers (ABRs). Every ABR hears the area summaries from all other ABRs. The ABR then forms a picture of the distance to all networks outside of its area by examining the collected advertisements and adding in the backbone distance to each advertising router.

When a *VLAN* is configured to run OSPF, you must configure the area for the VLAN.

- If you want to configure the VLAN to be part of a different OSPF area, use the following command:
  `configure ospf vlan vlan-name area area-identifier`
- If this is the first instance of the OSPF area being used, you must create the area first using the following command:
  `create ospf area area-identifier`

### *Stub Areas*

*OSPF* allows certain areas to be configured as stub areas. A stub area is connected to only one other area. The area that connects to a stub area can be the backbone area. External route information is not distributed into stub areas. Stub areas are used to reduce memory consumption and computational requirements on OSPF routers.

Use the following command to configure an OSPF area as a stub area:

```
configure ospf area area-identifier stub [summary | nosummary] stub-default-cost cost
```

*Not-So-Stubby-Areas*

Not-so-stubby-areas (NSSAs) are similar to the existing *OSPF* stub area configuration option but have the following two additional capabilities:

- External routes originating from an ASBR connected to the NSSA can be advertised within the NSSA.
- External routes originating from the NSSA can be propagated to other areas, including the backbone area.

The CLI command to control the NSSA function is similar to the command used for configuring a stub area, as follows:

```
configure ospf area area-identifier nssa [summary | nosummary] stub-default-cost cost {translate}
```

The **translate** option determines whether type 7 LSAs are translated into type 5 LSAs. When configuring an OSPF area as an NSSA, **translate** should only be used on NSSA border routers, where translation is to be enforced. If **translate** is not used on any NSSA border router in a NSSA, one of the ABRs for that NSSA is elected to perform translation (as indicated in the NSSA specification). The option should not be used on NSSA internal routers. Doing so inhibits correct operation of the election algorithm.

*Normal Area*

A normal area is an area that is not:

- Area 0
- Stub area
- NSSA

Virtual links can be configured through normal areas. External routes can be distributed into normal areas.

*Virtual Links*

In the situation when a new area is introduced that does not have a direct physical attachment to the backbone, a virtual link is used.

A virtual link provides a logical path between the ABR of the disconnected area and the ABR of the normal area that connects to the backbone. A virtual link must be established between two ABRs that have a common area, with one ABR connected to the backbone. The following figure illustrates a virtual link.

Note

Virtual links cannot be configured through a stub or NSSA area.

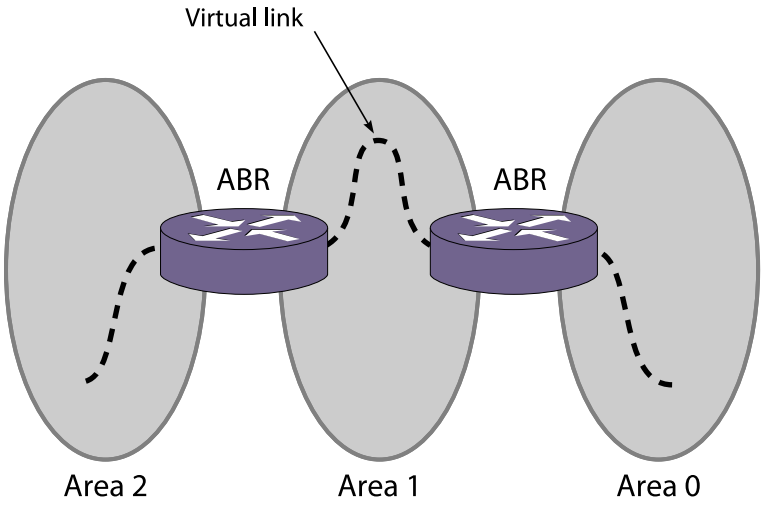



**Figure 217: Virtual Link Using Area 1 as a Transit Area**

Virtual links are also used to repair a discontiguous backbone area. For example, in the following figure, if the connection between ABR 1 and the backbone fails, the connection using ABR 2 provides redundancy so that the discontiguous area can continue to communicate with the backbone using the virtual link.

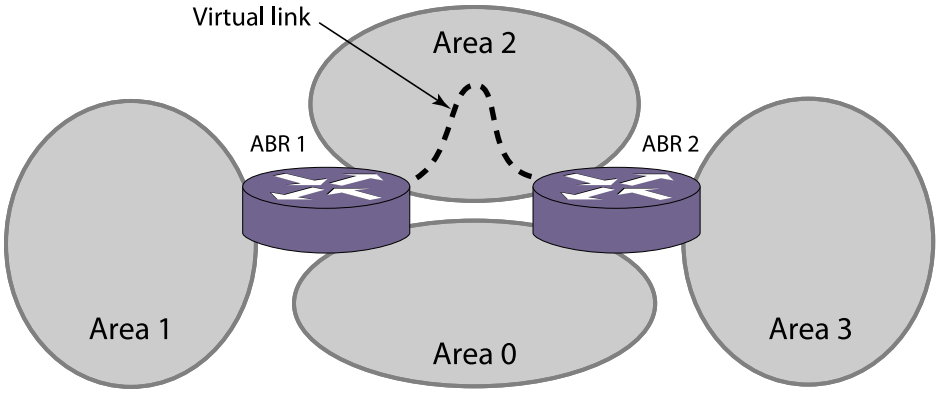



**Figure 218: Virtual Link Providing Redundancy**

## Point-to-Point Support

You can manually configure the *OSPF* link type for a *VLAN*. The following table describes the link types.

**Table 144: OSPF Link Types**

| Link Type | Number of Routers | Description |
|---|---|---|
| Auto | Varies | ExtremeXOS automatically determines the OSPF link type based on the interface type. This is the default setting. |
| Broadcast | Any | Routers must elect a designated router (DR) and a backup designated router (BDR) during synchronization. Ethernet is an example of a broadcast link. |

**Table 144: OSPF Link Types (continued)**

| Link Type | Number of Routers | Description |
|---|---|---|
| Point-to-point | Up to 2 | This type synchronizes faster than a broadcast link because routers do not elect a DR or BDR. It does not operate with more than two routers on the same VLAN. The Point-to-Point Protocol (PPP) is an example of a point-to-point link. An OSPF point-to-point link supports only zero to two OSPF routers and does not elect a designated router (DR) or backup designated router (BDR). If you have three or more routers on the VLAN, OSPF fails to synchronize if the neighbor is not configured. |
| Passive | | A passive link does not send or receive OSPF packets. |

Note

The number of routers in an OSPF point-to-point link is determined per VLAN, not per link.

All routers in the VLAN must have the same OSPF link type. If there is a mismatch, OSPF attempts to operate, but it may not be reliable.

## Route Redistribution

More than one routing protocol can be enabled simultaneously on the switch.

Route redistribution allows the switch to exchange routes, including static routes, between the routing protocols. The following figure is an example of route redistribution between an *OSPF* AS and a *RIP* AS.

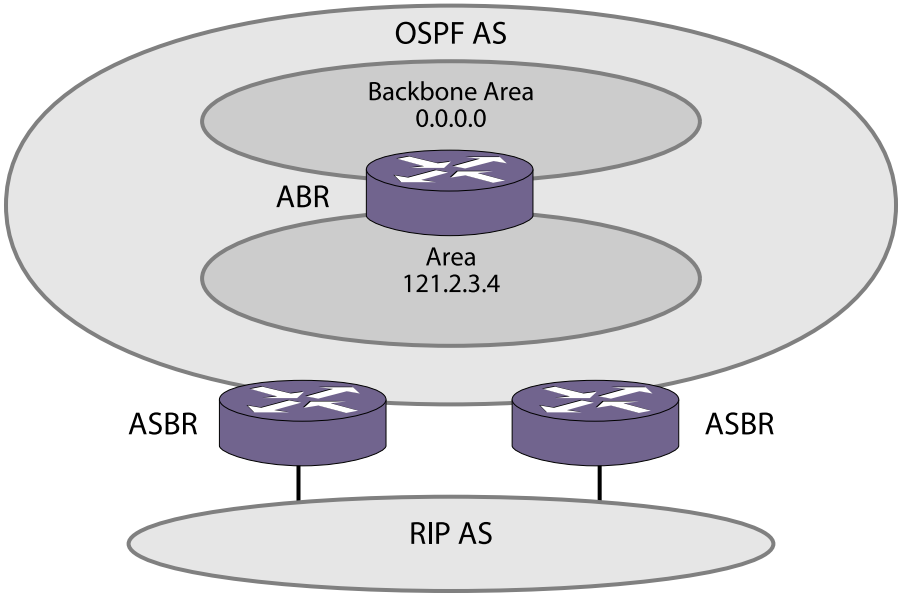



**Figure 219: OSPF Route Redistribution**

### Import Policy

Prior to ExtremeXOS 15.7, routing protocol OSPFv2 applied routing policies with keyword “import-policy”, which can only be used to change the attributes of routes installed into the switch routing table.

ExtremeXOS 15.7 provides the flexibility of using import policy to determine the routes to be added to or removed from the routing table.

To prevent a route being added to the routing table, the policy file must contain a matching rule with action “deny”. If there is no matching rule for a particular route, or the keyword “deny” is missing in the rule, the default action is “permit”, which means that route will be installed into the routing table. Refer to the following policy file example:

```
entry entry-one {
if {
nlri 11.22.0.0/16;
}
then {
cost 100;
}
}
entry entry-two {
if {
nlri 22.33.0.0/16;
}
then {
deny;
}
}
```

In the above policy example, entry-one is used to change the cost of any matching routes, and entry-two is used to remove those matching routes from the routing table.

Note

Only “Network Layer Reachability Information” (NLRI) and "route origin" can be used as matching criteria in policy rules; using "next_hop" as a matching criteria is not supported. Any other policy attribute is not recognized and is ignored.

## Configuring Route Redistribution

Exporting routes from one protocol to another and from that protocol to the first one are discrete configuration functions.

For example, to run *OSPF* and *RIP* simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from OSPF to RIP and the routes to export from RIP to OSPF. Likewise, for any other combinations of protocols, you must separately configure each to export routes to the other.

### *Redistribute Routes into OSPF*

To enable or disable the exporting of *BGP (Border Gateway Protocol)*, *RIP*, static, and direct (interface) routes to *OSPF*, use the following commands:

`enable ospf export [bgp | direct | e-bgp | host-mobility |i-bgp | rip | static | isis | isis-level-1 | isis-level-1-external | isis-level-2 | isis-level-2-external] [cost cost type [ase-type-1 | ase-type-2] {tag number} | policy-map]`

`disable ospf export [`**`bgp`** | **`direct`** | **`e-bgp`** | **`host-mobility`** |**`i-bgp`** | **`rip`** | **`static`** | **`isis`** | **`isis-level-1`**| **`isis-level-1-external`** | **`isis-level-2`** | **`isis-level-2-external`**`]`

These commands enable or disable the exporting of RIP, static, and direct routes by way of LSA to other OSPF routers as AS-external type 1 or type 2 routes. The default setting is disabled.

The cost metric is inserted for all Border Gateway Protocol (BGP), RIP, static, and direct routes injected into OSPF. If the cost metric is set to 0, the cost is inserted from the route. For example, in the case of BGP export, the cost equals the multiple exit discriminator (MED) or the path length. The tag value is used only by special routing applications. Use 0 if you do not have specific requirements for using a tag. (The tag value in this instance has no relationship with IEEE 802.1Q *VLAN* tagging.)

The same cost, type, and tag values can be inserted for all the export routes, or policies can be used for selective insertion. When a policy is associated with the export command, the policy is applied on every exported route.

The exported routes can also be filtered using policies.

Note

For routes exported to OSPF via a policy file, any refresh applied on that policy may result in temporary withdrawal and then immediate readvertising of those routes.

Verify the configuration using the command `show ospf`.

## OSPF Timers and Authentication

Configuring *OSPF* timers and authentication on a per-area basis is a shortcut to applying the timers and authentication to each *VLAN* in the area at the time of configuration. If you add more VLANs to the area, you must configure the timers and authentication for the new VLANs explicitly.

Use the command:

`configure ospf vlan [`*`vlan-name`* | **`all`**`] `**`timer`** *`retransmit-interval transit-delay hello-interval dead-interval`* `{`*`wait-timer-interval`*`}`

# Configuring OSPF

Each switch that is configured to run *OSPF* must have a unique router ID.

We recommend that you manually set the router ID of the switches participating in OSPF, instead of having the switch automatically choose its router ID based on the highest interface IP address. Not performing this configuration in larger, dynamic environments could result in an older LSDB remaining in use.

## Configuring OSPF Wait Interval

ExtremeXOS allows you to configure the *OSPF* wait interval, rather than using the router dead interval.

**Caution**
Do not configure OSPF timers unless you are comfortable exceeding OSPF specifications. Non-standard settings may not be reliable under all circumstances.

To specify the timer intervals, enter the following commands:

```
configure ospf area area-identifier timer retransmit-interval transit-
delay hello-interval dead-interval {wait-timer-interval}
```

```
configure ospf virtual-link router-identifier area-identifier timer
retransmit-interval transit-delay hello-interval dead-interval
```

```
configure ospf vlan [vlan-name | all] timer retransmit-interval transit-
delay hello-interval dead-interval {wait-timer-interval}
```

## OSPF Wait Interval Parameters

You can configure the following parameters:

- Retransmit interval—The length of time that the router waits before retransmitting an LSA that is not acknowledged. If you set an interval that is too short, unnecessary retransmissions result. The default value is 5 seconds.
- Transit delay—The length of time it takes to transmit an LSA packet over the interface. The transit delay must be greater than 0.
- Hello interval—The interval at which routers send hello packets. Shorter times allow routers to discover each other more quickly but also increase network traffic. The default value is 10 seconds.
- Dead router wait interval (Dead Interval)—The interval after which a neighboring router is declared down because hello packets are no longer received from the neighbor. This interval should be a multiple of the hello interval. The default value is 40 seconds.
- Router wait interval (Wait Timer Interval)—The interval between the interface coming up and the election of the DR and BDR. This interval should be greater than the hello interval. If this time is close to the hello interval, the network synchronizes very quickly but might not elect the correct DR or BDR. The default value is equal to the dead router wait interval.

**Note**
The *OSPF* standard specifies that wait times are equal to the dead router wait interval.

# OSPF Configuration Example

The following figure is an example of an autonomous system using *OSPF* routers. The details of this network follow.

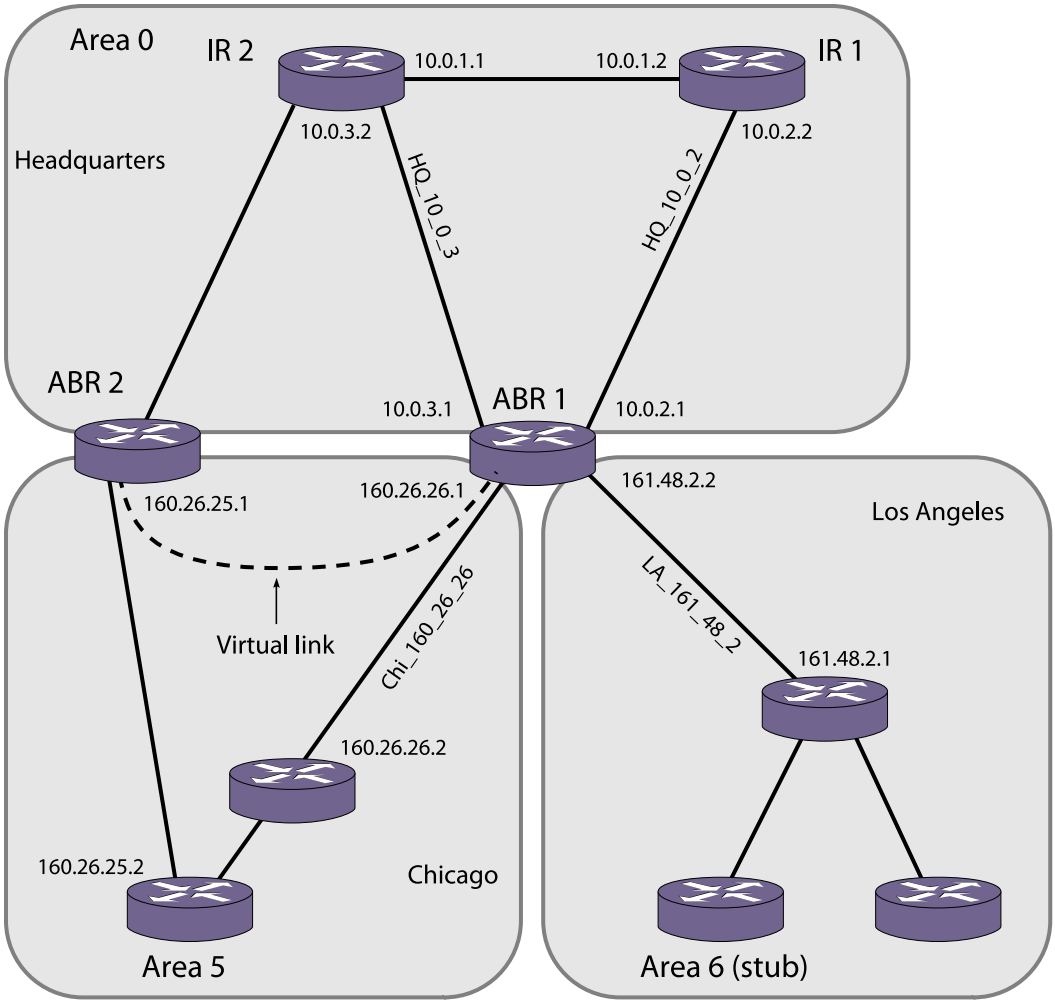



**Figure 220: OSPF Configuration Example**

Area 0 is the backbone area. It is located at the headquarters and has the following characteristics:

- Two internal routers (IR1 and IR2)
- Two area border routers (ABR 1 and ABR 2)
- Network number 10.0.x.x
- Two identified VLANs (HQ_10_0_2 and HQ_10_0_3)

Area 5 is connected to the backbone area by way of ABR 1 and ABR 2. It is located in Chicago and has the following characteristics:

- Network number 160.26.x.x
- One identified *VLAN* (Chi_160_26_26)
- Two internal routers

Area 6 is a stub area connected to the backbone by way of ABR 1. It is located in Los Angeles and has the following characteristics:

- Network number 161.48.x.x
- One identified VLAN (LA_161_48_2)

- Three internal routers
- Uses default routes for inter-area routing

The following section provides two router configurations for the example shown in the above figure.

## Configuration for ABR 1

The router labeled ABR 1 has the following configuration:

```
create vlan HQ_10_0_2
create vlan HQ_10_0_3
create vlan LA_161_48_2
create vlan Chi_160_26_26
configure vlan HQ_10_0_2 ipaddress 10.0.2.1 255.255.255.0
configure vlan HQ_10_0_3 ipaddress 10.0.3.1 255.255.255.0
configure vlan LA_161_48_2 ipaddress 161.48.2.2 255.255.255.0
configure vlan Chi_160_26_26 ipaddress 160.26.26.1 255.255.255.0
create ospf area 0.0.0.5
create ospf area 0.0.0.6
enable ipforwarding
configure ospf area 0.0.0.6 stub nosummary stub-default-cost 10
configure ospf add vlan LA_161_48_2 area 0.0.0.6
configure ospf add vlan Chi_160_26_26 area 0.0.0.5
configure ospf add vlan HQ_10_0_2 area 0.0.0.0
configure ospf add vlan HQ_10_0_3 area 0.0.0.0
configure ospf vlan LA_161_48_2 priority 10
configure ospf vlan Chi_160_26_26 priority 10
configure ospf vlan HQ_10_0_2 priority 5
configure ospf vlan HQ_10_0_3 priority 5
enable ospf
```

## Configuration for IR 1

The router labeled IR 1 has the following configuration:

```
configure vlan HQ_10_0_1 ipaddress 10.0.1.2 255.255.255.0
configure vlan HQ_10_0_2 ipaddress 10.0.2.2 255.255.255.0
enable ipforwarding
configure ospf add vlan all area 0.0.0.0
configure ospf area 0.0.0.0 priority 10
enable ospf
```

**Note**

In *OSPF* edge mode, the *VLAN* priority is “0” and cannot be set. (Refer to OSPF Edge Mode on page 1342.) When the license is upgraded to a Core license, the VLAN priority of "0" needs to be reset in order to participate in DR/BDR election.

# Displaying OSPF Settings

You can use a number of commands to display settings for *OSPF*.

- To show global OSPF information, use the `show ospf` command with no options.
- To display information about one or all OSPF areas, enter:

  `show ospf area {`**`detail`** `|` *`area-identifier`*`}`

  The **`detail`** option displays information about all OSPF areas in a detail format.

- To display information about OSPF interfaces for an area, a *VLAN*, or for all interfaces enter:

  `show ospf interfaces {vlan vlan-name | area area-identifier | enabled}`

  The **detail** option displays information about all OSPF interfaces in a detail format.

  ExtremeXOS provides several filtering criteria for the `show ospf lsdb` command.

  You can specify multiple search criteria, and only those results matching all of the criteria are displayed. This allows you to control the displayed entries in large routing tables.

- To display the current link-state database, enter:

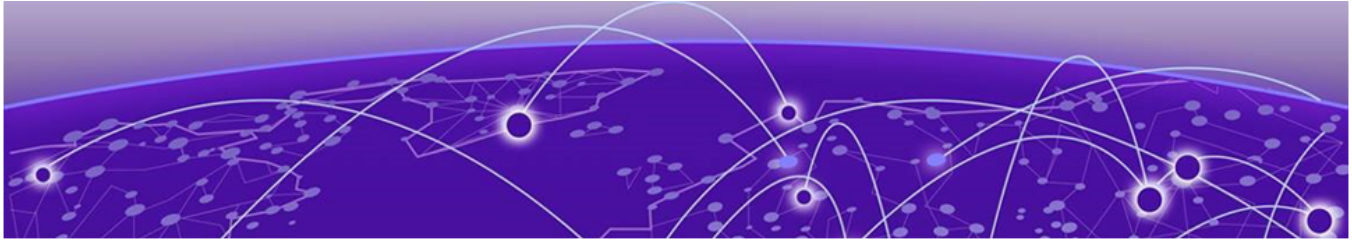
  `show ospf lsdb {detail | stats} {area [area-identifier | all]} {{lstype} [lstype | all]} {lsid lsid-address{lsid-mask}} {routerid routerid-address {routerid-mask}} {interface[[ip-address{ip-mask} | ipNetmask] | vlan vlan-name]}`

  The **detail** option displays all fields of matching LSAs in a multiline format. The **summary** option displays several important fields of matching LSAs, one line per LSA. The **stats** option displays the number of matching LSAs but not any of their contents. If not specified, the default is to display in the summary format.

  A common use of this command is to omit all optional parameters, resulting in the following shortened form:

  `show ospf lsdb`

  The shortened form displays LSAs from all areas and all types in a summary format.



# OSPFv3



This chapter discusses the *OSPF (Open Shortest Path First)* version 3 protocol for distributing routing information between routers belonging to an autonomous system. This chapter provides an overview of the protocol's features and example configuration commands.

Note
OSPFv3 is available on platforms with an Advanced Edge or Core license. For information about OSPFv3 licensing, see the *Feature License Requirements* document.

## OSPFv3 Overview

*OSPF* is a link state protocol that distributes routing information between routers belonging to a single IP domain; the IP domain is also known as an autonomous system (AS).

In a link-state routing protocol, each router maintains a database describing the topology of the AS. Each participating router has an identical database for an area maintained from the perspective of that router.

From the link state database (LSDB), each router constructs a tree of shortest paths, using itself as the root. The shortest path tree provides the route to each destination in the AS. When several equal-cost routes to a destination exist, traffic can be distributed among them. The cost of a route is described by a single metric.

*OSPFv3 (Open Shortest Path First version 3)* supports IPv6, and uses commands only slightly modified from that used to support IPv4. OSPFv3 has retained the use of the 4-byte, dotted decimal numbers for router IDs, LSA IDs, and area IDs.

OSPFv3 is an interior gateway protocol (IGP), as is the other common IGP for IPv6, *RIPng (Routing Information Protocol Next Generation)*. OSPFv3 and RIPng are compared in RIPng.

Note
Two types of OSPFv3 functionality are available and each has a different licensing requirement. One is the complete OSPFv3 functionality and the other is OSPFv3 Edge Mode, a subset of OSPFv3 that is described below. For specific information regarding OSPFv3 licensing, see the *Feature License Requirements* document.

## OSPFv3 Edge Mode

*OSPFv3* Edge Mode is a subset of OSPFv3 available on platforms with an Advanced Edge license.

There are two restrictions on OSPFv3 Edge Mode:

- At most, four Active OSPFv3 *VLAN (Virtual LAN)* interfaces are permitted. There is no restriction on the number of Passive interfaces.
- The OSPFv3 Priority on VLANs is 0, and is not configurable. This prevents the system from acting as a DR or BDR.

## BFD for OSPFv3

The BFD for *OSPFv3* feature gives OSPFv3 routing protocol the ability to utilize BFD's fast failure detection to monitor OSPFv3 neighbor adjacencies. CLI commands are provided to configure BFD protection for OSPFv3, so that as a registered BFD client, OSPFv3 can request BFD protection for interested OSPFv3 neighbors, and receive notifications about BFD session setup status and BFD session status updates (after establishing) from the BFD server. When BFD detects a communication failure between neighbors, it informs OSPFv3, which causes the OSPFv3 neighbor state be marked as "down." This allows OSPFv3 protocol to quickly begin network convergence and use alternate paths to the affected neighbor.

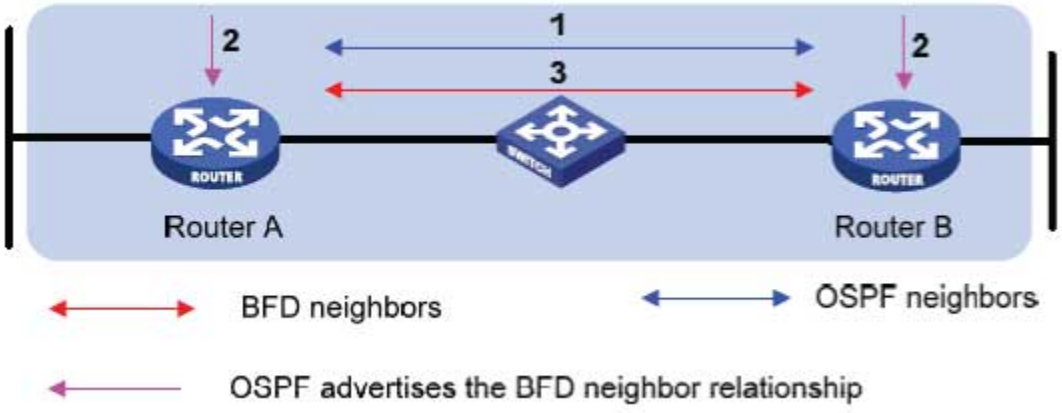



**Figure 221: Basic Operational Flow of OSPFv3 BFD**

*Establishing a BFD Session for OSPFv3 Neighbor*

1. OSPFv3 discovers a neighbor.
2. If BFD for OSPFv3 is configured, OSPFv3 on both routers sends a request to the local BFD server to initiate a BFD neighbor session with the OSPFv3 neighbor router.
3. The BFD neighbor session with the OSPFv3 neighbor router is established on both sides if BFD session limit is not reached.

If the BFD session limit is reached, the OSPFv3 neighbor will be marked as BFD session failed if synchronous request is used, or pending if asynchronous request is used, and the BFD server will send an asynchronous notification when the session registration passes later. (The asynchronous request is not available until the BFD client session create API is enhanced.)

*Eliminating the OSPFv3 Neighbor Relationship by BFD Fault Detection*

1. A failure occurs in the network.
2. The BFD neighbor session with the OSPFv3 neighbor router is removed because the BFD timer expired.
3. On both Router A and Router B, BFD notifies the local OSPFv3 process that the BFD neighbor is DOWN.
4. The local OSPFv3 process tears down the OSPFv3 neighbor relationship by marking neighbor state DOWN. (If an alternative path is available, the routers will immediately start converging on it.)

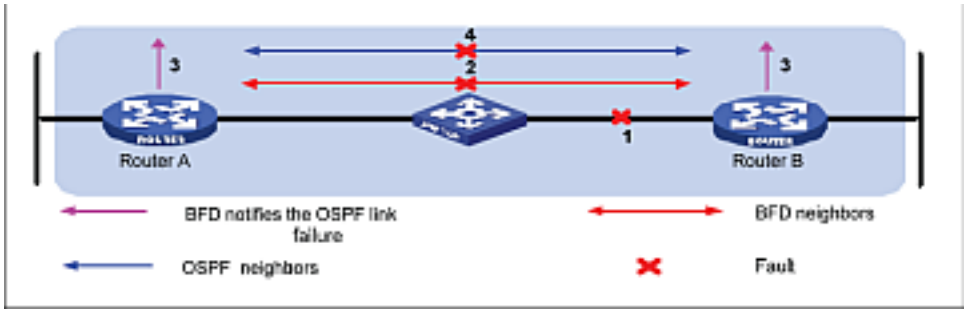



**Figure 222: OSPFv3 Neighbor Relationship Eliminated by BFD Fault Detection**

OSPFv3 will request the BFD server to delete the BFD session for OSPFv3 neighbor moved to DOWN state. When the link failure is later resolved, OSPFv3 needs to register the re-discovered neighbor to BFD again to initiate BFD session creation.

*Removing BFD Protection for OSPFv3*

1. BFD protection is removed from OSPFv3 interface on Router A, OSPFv3 process requests BFD server to delete all BFD sessions for neighbors learned on that interface.
2. BFD server will stop sending session status update to local OSPFv3 process on Router A.
3. Before actually deleting any sessions, BFD server on Router A will first notify Router B to mark those session status as "Admin Down," which will cause Router B to stop using BFD protection for those OSPFv3 neighbors.
4. When BFD protection is removed from OSPFv3 interface on Router B, BFD sessions can be deleted immediately since they are already in "Admin Down" state.

When BFD for OSPFv3 is configured on broadcast interface, the default behavior is to register only OSPFv3 neighbors in FULL state with the BFD server. Separate BFD sessions are created for each neighbor learned on the same interface. If multiple clients ask for the same neighbor on the same interface, then single BFD sessions are established between the peers.

*OSPFv3 Neighbor State Determination*

With active BFD protection, OSPFv3 combines the BFD session state with the associated interface admin and operational states to determine the OSPFv3 neighbor adjacency discovered on that OSPFv3 interface. Regarding OSPFv3 neighbor relationships, OSPFv3 reacts directly to BFD session state changes only in the following circumstances:

- If BFD is enabled on the interface, and
- If BFD for OSPFv3 is configured on the OSPFv3 interface, and

- If a BFD session has been established to the neighbor, and
- If the BFD session has passed the INIT_COMPLETE state then:

1. The OSPFv3 neighbor relationship will remain as FULL if the operational state of the BFD session is "UP" and the operational state of the associated *VLAN* interface is UP.
2. The OSPFv3 neighbor relationship will be considered as DOWN if the operational state of the BFD session is DOWN or the operational state of the associated VLAN interface is DOWN.

In all other cases, the BFD session state is not considered as part of the reported OSPFv3 neighbor state, and the OSPFv3 neighbor state reverts to the operational state of the OSPFv3 interface only. When the BFD session is in ADMIN_DOWN state, OSPFv3 ignores BFD events and OSPFv3 neighbor adjacency is not be affected by the BFD session state change.

## Link State Database

Upon initialization, each router transmits a link state advertisement (LSA) on each of its interfaces. LSAs are collected by each router and stored into the LSDB of each router. After all LSAs are received, the router uses the link state database (LSDB) to calculate the best routes for use in the IP routing table. *OSPFv3* uses flooding to distribute LSAs between routers.

Any change in routing information is sent to all of the routers in the network. All routers within an area have the exact same LSDB.

The following table describes LSA type numbers.

**Table 145: Selected OSPFv3 LSA Types**

| Type Number | Description |
|---|---|
| 0x0008 | Link LSA |
| 0x2001 | Router LSA |
| 0x2002 | Network LSA |
| 0x2003 | Inter-Area-Prefix LSA |
| 0x2004 | Inter-Area-Router LSA |
| 0x2009 | Intra-Area-Prefix LSA |
| 0x4005 | AS external LSA |

## Graceful OSPFv3 Restart

RFC 3623 describes a way for *OSPF* control functions to restart without disrupting traffic forwarding.

Without graceful restart, adjacent routers assume that information previously received from the restarting router is stale and does not use it to forward traffic to that router. However, in many cases, two conditions exist that allow the router restarting *OSPFv3* to continue to forward traffic correctly. The first condition is that forwarding can continue while the control function is restarted. Most modern router system designs separate the forwarding function from the control function so that traffic can still be forwarded independent of the state of the OSPFv3 function. Routes learned through OSPFv3 remain in the routing table and packets continue to be forwarded. The second condition required for graceful

restart is that the network remain stable during the restart period. If the network topology is not changing, the current routing table remains correct. Often, networks can remain stable during the time for restarting OSPFv3.

### *Restarting and Helper Mode*

Routers involved with graceful restart fill one of two roles: the restarting router or the helper router.

With graceful restart, the router that is restarting sends out Grace-LSAs informing its neighbors that it is in graceful restart mode, how long the helper router should assist with the restart (the grace period), and why the restart occurred. If the neighboring routers are configured to help with the graceful restart (helper-mode), they continue to advertise the restarting router as if it were fully adjacent. Traffic continues to be routed as though the restarting router is fully functional. If the network topology changes, the helper routers stop advertising the restarting router. The helper router continues in helper mode until the restarting router indicates successful termination of graceful restart, the Grace-LSAs expire, or the network topology changes. A router can be configured for graceful restart, and for helper-mode separately. A router can be a helper when its neighbor restarts, and can in turn be helped by a neighbor if it restarts.

### *Planned and Unplanned Restarts*

Two types of graceful restarts are defined: planned and unplanned.

A planned restart occurs if the software module for *OSPFv3* is upgraded, or if the router operator decides to restart the OSPFv3 control function for some reason. The router has advance warning, and is able to inform its neighbors in advance that OSPFv3 is restarting.

An unplanned restart occurs if there is some kind of system failure that causes a remote reboot or a crash of OSPFv3, or an MSM/MM failover occurs. As OSPFv3 restarts, it informs its neighbors that it is engaged in an unplanned restart.

You can configure a router to enter graceful restart for only planned restarts, for only unplanned restarts, or for both. Also, you can separately decide to configure a router to be a helper for only planned, only unplanned, or for both kinds of restarts.

### *Configuring Graceful OSPFv3 Restart*

- Configure a router to perform graceful *OSPFv3* restart:

  `configure ospfv3 restart [`**`none`** | **`planned`** | **`unplanned`** | **`both`**`]`

  Because a router can act as a restart helper router to multiple neighbors, you will specify which neighbors to help.
- Configure a router interface to act as a graceful OSPFv3 restart helper:

  `configure ospfv3 [[`**`vlan`** | **`tunnel`**`] `**`all`** | {**`vlan`**} *`vlan-name`* | {**`tunnel`**} *`tunnel-name`* | **`area`** *`area-identifier`*`] restart-helper [`**`none`** | **`planned`** | **`unplanned`** | **`both`**`]`
- Configure a virtual link to act as a OSPFv3 graceful restart helper:

  `configure ospf [`**`vlan`** `[`**`all`** | *`vlan-name`*`] | `**`area`** *`area-identifier`* | **`virtual-link`** *`router-identifier area-identifier`*`] `**`restart-helper`** `[`**`none`** | **`planned`** | **`unplanned`** | **`both`**`]`

- The graceful restart period sent out to helper routers can be configured with the following command:

  By default, a helper router will terminate graceful restart if received LSAs would affect the restarting router.

  This will occur when the restart-helper receives an LSA that will be flooded to the restarting router or when there is a changed LSA on the restarting router's retransmission list when graceful restart is initiated.
- Disable or enable helper router LSA check:

  **disable ospfv3 [[vlan | tunnel] all | vlan** *vlan-name* **| {tunnel}** *tunnel-name* **| area** *area-identifier*] **restart-helper-lsa-check**

  **enable ospfv3 [[vlan | tunnel] all | {vlan}** *vlan-name* **| {tunnel}** *tunnel-name* **| area** *area-identifier*] **restart-helper-lsa-check**

## Areas

*OSPFv3* allows parts of a network to be grouped together into areas.

The topology within an area is hidden from the rest of the AS. Hiding this information enables a significant reduction in LSA traffic and reduces the computations needed to maintain the LSDB. Routing within the area is determined only by the topology of the area.

The three types of routers defined by OSPFv3 are as follows:

- Internal router (IR)—An internal router has all of its interfaces within the same area.
- Area border router (ABR)—An ABR has interfaces in multiple areas. It is responsible for exchanging summary advertisements with other ABRs.
- Autonomous system border router (ASBR)—An ASBR acts as a gateway between OSPFv3 and other routing protocols, or other autonomous systems.

### *Backbone Area (Area 0.0.0.0)*

Any *OSPFv3* network that contains more than one area is required to have an area configured as area 0.0.0.0, also called the backbone.

All areas in an AS must be connected to the backbone. When designing networks, you should start with area 0.0.0.0 and then expand into other areas.

Note
Area 0.0.0.0 exists by default and cannot be deleted or changed.

The backbone allows summary information to be exchanged between area border routers (ABRs). Every ABR hears the area summaries from all other ABRs. The ABR then forms a picture of the distance to all networks outside of its area by examining the collected advertisements and adding in the backbone distance to each advertising router.

When a *VLAN* is configured to run OSPFv3, you must configure the area for the VLAN.

If you want to configure the VLAN to be part of a different OSPFv3 area, use the following command:

```
configure ospfv3 vlan vlan-name area area-identifier
```

Note
The only domain name currently supported is *OSPF*-Default.

If this is the first instance of the OSPFv3 area being used, you must create the area first using the following command:

```
create ospfv3 area area_identifier
```

*Stub Areas*

*OSPFv3* allows certain areas to be configured as stub areas.

A stub area is connected to only one other area. The area that connects to a stub area can be the backbone area. External route information is not distributed into stub areas. Stub areas are used to reduce memory consumption and computational requirements on OSPFv3 routers. To configure an OSPFv3 area as a stub area, use the following command:

```
configure ospfv3 area area_identifier stub [summary | nosummary] stub-default-cost cost
```

*Not-So-Stubby-Areas*

Not-so-stubby-areas (NSSAs) are similar to the existing *OSPFv3* stub area configuration option, but have the following two additional capabilities:

- External routes originating from an ASBR connected to the NSSA can be advertised within the NSSA.
- External routes originating from the NSSA can be propagated to other areas, including the backbone area.

The CLI command to control the NSSA function is similar to the command used for configuring a stub area, as follows:

```
configure ospfv3 area area-identifier nssa [summary | nosummary] stub-default-cost cost {translate}
```

The **translate** option in ABR determines whether it is either always translating type 7 LSAs are translated to type 5 LSAs or a candidate for election. When configuring an OSPFv3 area as an NSSA, **translate** should only be used on NSSA border routers, where translation is to be enforced. If **translate** is not used on any NSSA border router in a NSSA, one of the ABRs for that NSSA is elected to perform translation (as indicated in the NSSA specification). The option should not be used on NSSA internal routers. Doing so inhibits correct operation of the election algorithm.

*Normal Area*

A normal area is an area that is not:

- Area 0
- Stub area
- NSSA

Virtual links can be configured through normal areas. External routes can be distributed into normal areas.

*Virtual Links*

In a situation where a new area is introduced that does not have a direct physical attachment to the backbone, a virtual link is used.

A virtual link provides a logical path between the ABR of the disconnected area and the ABR of the normal area that connects to the backbone. A virtual link must be established between two ABRs that have a common area, with one ABR connected to the backbone. The following figure illustrates a virtual link.

Note
Virtual links cannot be configured through a stub or NSSA area.

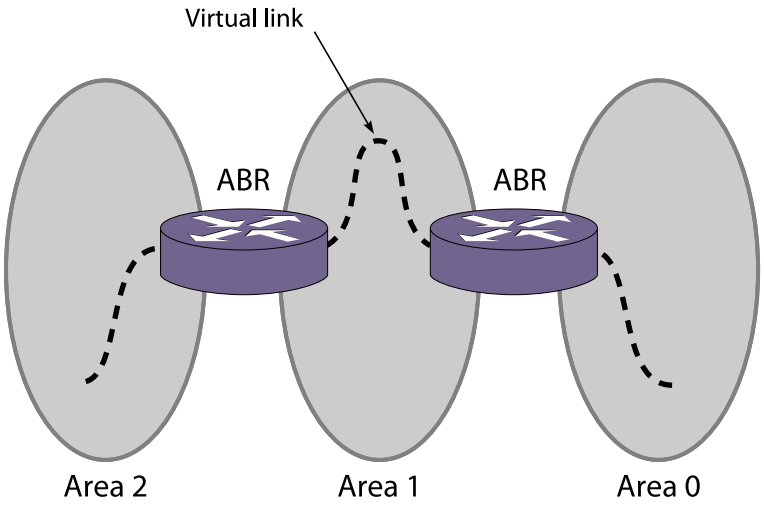



**Figure 223: Virtual Link Using Area 1 as a Transit Area**

Virtual links are also used to repair a discontiguous backbone area. For example, in the following figure, if the connection between ABR1 and the backbone fails, the connection using ABR2 provides redundancy so that the discontiguous area can continue to communicate with the backbone using the virtual link.

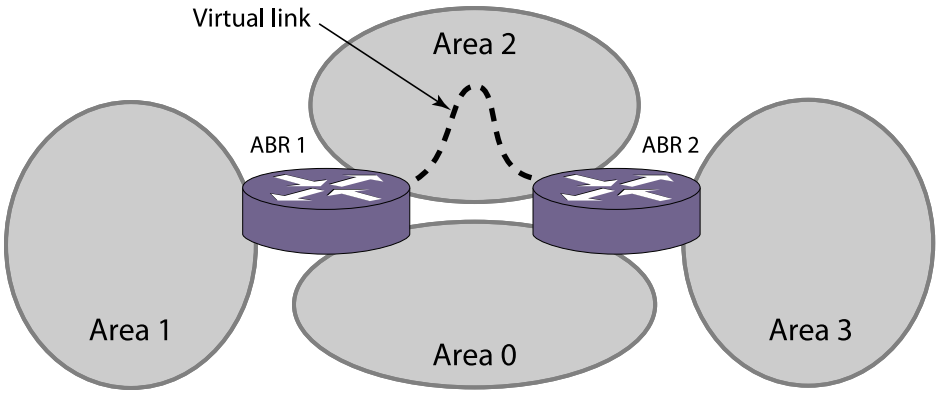



**Figure 224: Virtual Link Providing Redundancy**

## Link-Type Support

You can manually configure the *OSPFv3* link type for a *VLAN*. The following table describes the link types.

**Table 146: OSPFv3 Link Types**

| Link Type | Number of Routers | Description |
|---|---|---|
| Auto | Varies | ExtremeXOS automatically determines the OSPFv3 link type based on the interface type. This is the default setting. |
| Broadcast | Any | Routers must elect a designated router (DR) and a backup designated router (BDR) during synchronization. Ethernet is an example of a broadcast link. |
| Point-to-point | Up to 2 | This type synchronizes faster than a broadcast link because routers do not elect a DR or BDR. It does not operate with more than two routers on the same VLAN. The Point-to-Point Protocol (PPP) is an example of a point-to-point link. An OSPFv3 point-to-point link supports only zero to two OSPFv3 routers and does not elect a designated router (DR) or backup designated router (BDR). If you have three or more routers on the VLAN, OSPFv3 fails to synchronize if the neighbor is not configured. |
| Passive | | A passive link does not send or receive OSPFv3 packets. |

Note

The number of routers in an OSPFv3 point-to-point link is determined per VLAN, not per link.

All routers in the VLAN must have the same OSPFv3 link type. If there is a mismatch, OSPFv3 attempts to operate, but it may not be reliable.

# Import Policy

Prior to ExtremeXOS 15.7, routing protocol *OSPFv3* applied routing policies with keyword “import-policy”, which can only be used to change the attributes of routes installed into the switch routing table. ExtremeXOS 15.7 provides the flexibility of using import policy to determine the routes to be added to or removed from the routing table.

To prevent a route being added to the routing table, the policy file must contain a matching rule with action “deny”. If there is no matching rule for a particular route, or the keyword “deny” is missing in the rule, the default action is “permit”, which means that route will be installed into the routing table. Refer to the following policy file example:

```
entry entry-one {
if {
nlri 1001:0:0:1:0:0:0:0/64 exact;
}
then {
cost 100;
}
}
entry entry-two {
if {
nlri 2001:0:0:1:0:0:0:0/64 exact;
}
```

```
then {
deny;
}
}
```

In the above policy example, entry-one is used to change the cost of any matching routes, and entry-two is used to remove those matching routes from the routing table.

Note

Only “Network Layer Reachability Information” (NLRI) and "route origin" can be used as matching criteria in policy rules; using "next_hop" as a matching criteria is not supported. Any other policy attribute is not recognized and is ignored.

## Route Redistribution

More than one routing protocol can be enabled simultaneously on the switch.

Route redistribution allows the switch to exchange routes, including static routes, between the routing protocols. The following figure is an example of route redistribution between an *OSPFv3* AS and a *RIPng* AS.

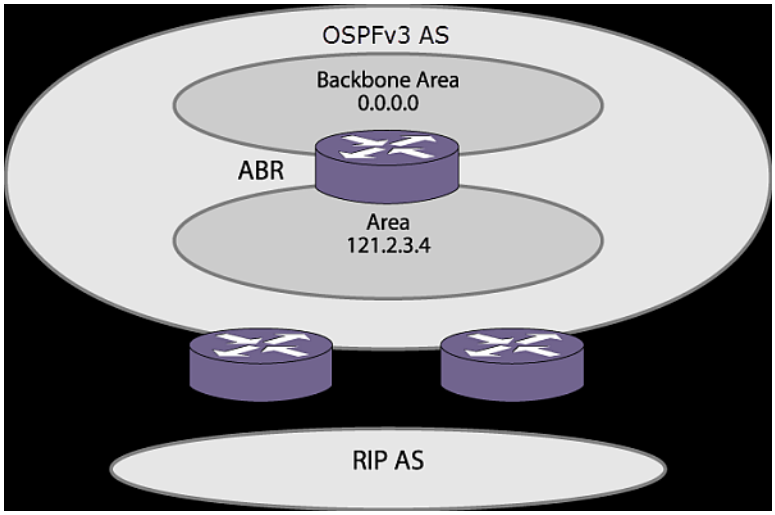



**Figure 225: OSPFv3 Route Redistribution**

### Configuring Route Redistribution

Exporting routes from one protocol to another and from that protocol to the first one are discrete configuration functions.

For example, to run *OSPFv3* and *RIPng* simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from OSPFv3 to RIPng and the routes to export from RIPng to OSPFv3. Likewise, for any other combinations of protocols, you must separately configure each to export routes to the other.

#### *Redistributing Routes into OSPFv3*

To enable or disable the exporting of *RIPng*, static, and direct (interface) routes to *OSPFv3*, use the following command:

```
enable ospfv3 export [direct | ripng | static | isis | isis-level-1 |
isis-level-1-external | isis-level-2 | isis-level-2-external| bgp |
host-mobility] [cost cost type [ase-type-1 | ase-type-2] | policy_map
{tag number}]
```

These commands enable or disable the exporting of RIPng, static, and direct routes by way of LSA to other OSPFv3 routers as AS-external type 1 or type 2 routes. The default setting is disabled.

The cost metric is inserted for all RIPng, static, and direct routes injected into OSPFv3. If the cost metric is set to 0, the cost is inserted from the route. The tag value is used only by special routing applications. Use 0 if you do not have specific requirements for using a tag. (The tag value in this instance has no relationship with IEEE 802.1Q *VLAN* tagging.)

The same cost, type, and tag values can be inserted for all the export routes, or policies can be used for selective insertion. When a policy is associated with the export command, the policy is applied on every exported route. The exported routes can also be filtered using policies.

Note

For routes exported to OSPFv3 via a policy file, any refresh applied on that policy may result in temporary withdrawal and then immediate readvertising of those routes.

Verify the configuration using the command:

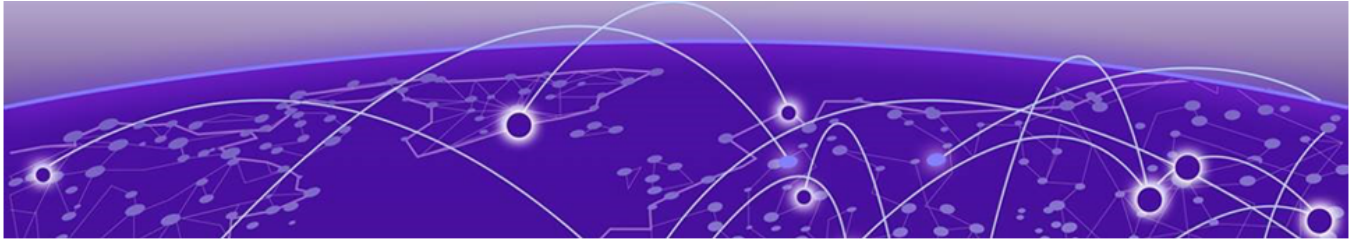
```
show ospfv3
```

## OSPFv3 Timers

Configuring *OSPFv3* timers on a per area basis is a shortcut to applying the timers to each *VLAN* in the area at the time of configuration.

If you add more VLANs to the area, you must configure the timers for the new VLANs explicitly. Use the command:

```
configure ospfv3 virtual-link {routerid} router_identifier {area}
area_identifier timer {retransmit-interval} retransmit_interval
{transit-delay} transit_delay {hello-interval} hello_interval {dead-
interval} dead_interval
```



# IS-IS



Note
The IS-IS feature is supported only at and above the license level listed for this feature in the license tables in the *Feature License Requirements* document.

This chapter introduces IS-IS, a link state protocol that distributes routing information between routers belonging to an autonomous system. It provides configuration commands and examples and information about configuring, displaying, and managing IS-IS on a network. This chapter assumes that you are already familiar with Intermediate System-Intermediate System (IS-IS) and IP routing.

If not, refer to the following publications for additional information:

- RFC 1195—Use of OSI IS-IS for Routing in TCP/IP and Dual Environments
- RFC 2763—Dynamic Hostname Exchange Mechanism for IS-IS
- RFC 2966—Domain-Wide Prefix Distribution with Two-Level IS-IS
- RFC 2973—IS-IS Mesh Groups
- RFC 3373—Three-Way Handshaking for IS-IS Point-to-Point Adjacencies
- RFC 3719—Recommendations for Interoperable Networks using Intermediate System to Intermediate System (IS-IS)
- RFC 3787—Recommendations for Interoperable IP Networks using Intermediate System to Intermediate System (IS-IS)
- draft-ieft-isis-ipv6-06—Routing IPv6 with IS-IS
- draft-ietf-isis-restart-02—Restart signaling for IS-IS
- draft-ietf-isis-wg-multi-topology-11—Multi-topology (MT) routing in IS-IS
- ISO 10589—OSI IS-IS Intra-Domain Routing Protocol (also available as RFC 1142)
- *Interconnections: Bridges and Routers* by Radia Perlman, ISBN 0-201-56332-0, Published by Addison-Wesley Publishing Company

# IS-IS Overview

IS-IS is a link state protocol that distributes routing information between routers belonging to a single IP domain; the IP domain is also known as an autonomous system (AS). In a link-state routing protocol, each router maintains a database describing the topology of the AS. Each participating router has an identical database maintained from the perspective of that router.

From the link state database (LSDB), each router constructs a tree of shortest paths, using itself as the root. The shortest path tree provides the route to each destination in the AS. When several equal-cost routes to a destination exist, traffic can be distributed among them. The cost of a route is described by a single metric.

IS-IS is an interior gateway protocol (IGP), as are *RIP (Routing Information Protocol)* and *OSPF (Open Shortest Path First)*. Unlike RIP and OSPF, IS-IS was not initially designed for IP. RFC 1195 specifies how IS-IS can run in an IP environment. The Extreme Networks implementation supports IS-IS only in IP environments. RIP, OSPF, and IS-IS are compared in RIP. The IPv6 versions of these protocols are compared in RIPng.

## Establishing Adjacencies

An adjacency is an acknowledged relationship between two IS-IS routers. An adjacency must be established before two IS-IS routers can exchange routing information.

IS-IS routers establish adjacencies by exchanging hello PDUs, which are also called Intermediate System to Intermediate System Hellos (IIHs). Hello PDUs contain some interface configuration and capability information. Once a pair of neighbors exchanges hello PDUs with acceptably matching configuration and capabilities, an adjacency is formed. Hello PDUs are sent periodically by each party to maintain the adjacency.

After an adjacency is formed, information about the adjacency is stored in a link state PDU (LSP), which is stored in the router link state database (LSDB). Routers periodically flood all of their LSPs to all other network nodes. When a router receives LSPs, it adds the LSPs to its LSDB, and uses the LSDB to calculate routes to other routers. These database maintenance operations are performed a little differently for the two adjacency types, point-to-point, and broadcast.

### *Point-to-Point Adjacency*

Point-to-point adjacencies can include no more than two routers in the same *VLAN (Virtual LAN)*. The following figure shows a point-to-point adjacency.

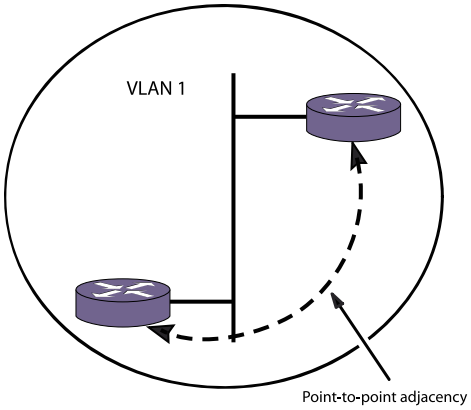



**Figure 226: Point-to-Point Adjacency**

Once a point-to-point adjacency is established, each router sends a CSNP (Complete Sequence Number PDU) listing a summary of all its LSPs. When a router receives its neighbor's CSNP, it sends any LSP it has that is either not present in the CSNP or is newer than the version in the CSNP. In a point-to-point adjacency, partial sequence number PDUs (PSNPs) are used to acknowledge each LSP a router receives from its neighbor. If a PSNP is not received within a configurable period of time, unacknowledged LSPs are re-sent.

A disadvantage to point-to-point adjacencies is that they do not scale well. The following figure shows a four-router network with point-to-point adjacencies.

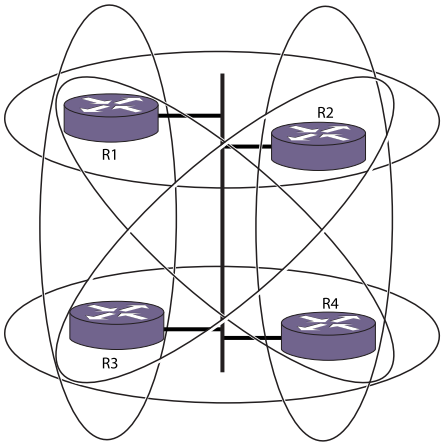



**Figure 227: Point-to-Point Adjacencies in a Four-Router Network**

In the network in the above figure, each of the ellipses represents a point-to-point adjacency. Each of the four routers periodically sends all of its LSPs to the other three routers. Each, in turn, will flood the received LSPs to the other two since they have no way of knowing which routers have already received them, generating N2 LSPs. This network routing traffic reduces the bandwidth available for data traffic. For networks that only have two routers and are not likely to grow, a point-to-point adjacency is appropriate.

### *Broadcast Adjacency*

Broadcast adjacencies can include two or more routers in a single adjacency. The following figure depicts a broadcast adjacency.

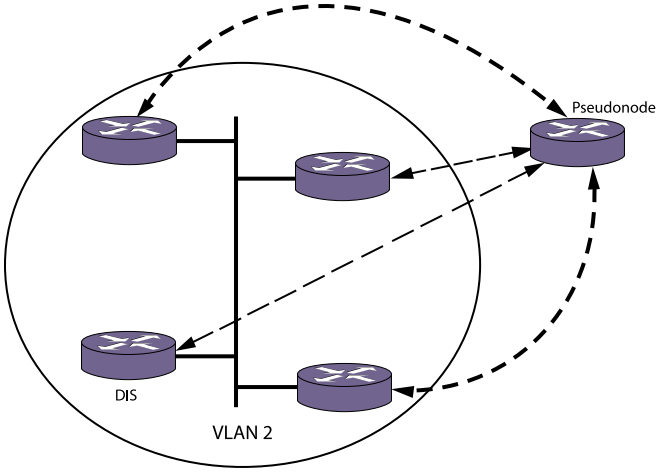



**Figure 228: Broadcast Adjacencies in a Four-Router Network**

When broadcast adjacencies are formed, one of the routers participating in the adjacency is elected designated intermediate system (DIS). The election is determined by a DIS priority configured for each router. In the event of a tie, the router with the numerically highest MAC address wins.

A broadcast network can be considered a virtual node, or pseudonode, to which all routers have a zero-cost adjacency.

The DIS acts on behalf of the pseudonode by advertising an LSP listing all routers in the adjacency with zero-cost metric. The DIS also periodically sends a complete sequence number PDU (CSNP), which lists all LSPs in the link-state database. If a router sees that it is missing one or more of the entries, it multicasts a request for them using a PSNP. Only the DIS responds to this request by sending the requested LSPs. All routers multicast their originated LSPs as they are refreshed, and multicast periodic hellos to the network.

The default configuration creates broadcast adjacencies.

## IS-IS Hierarchy

IS-IS has a two-level hierarchy. IS-IS routers may participate in either level or both. The following figure shows a basic IS-IS AS.

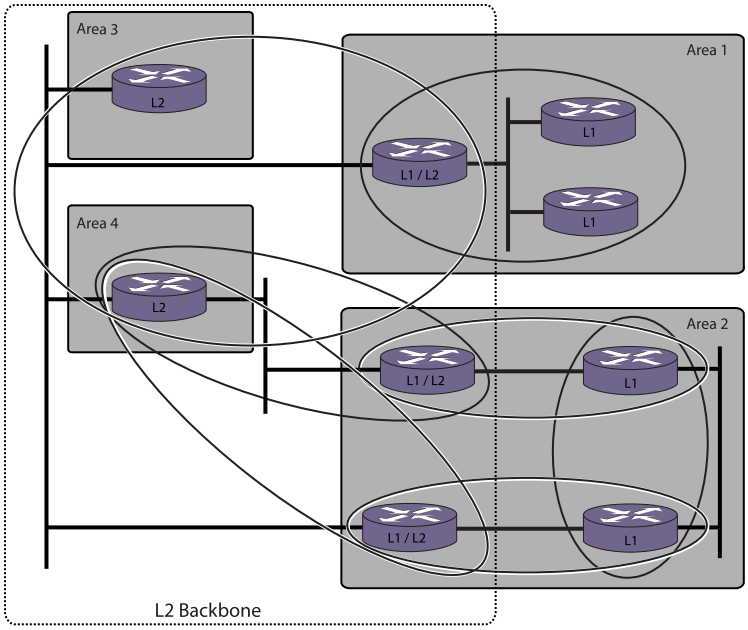



**Figure 229: Basic Autonomous System**

In the network in the above figure, each of the ellipses represents an adjacency. Level 1 (L1) routers within an area need only know the topology within that area as well as the location of the nearest L1/L2 attached router. An attached router is one that participates in both levels and can reach at least one other L1/L2 or L2 router with a different area address.

In an AS, there is only one L2 area, and it serves as a backbone area between L1 areas. This L2 area is also called a domain (not to be confused with the entire AS routing domain). L2 routers communicate with one another irrespective of L1 area membership or L2 area address.

Note

On any switch, the maximum number of L1/L2 adjacencies is one half the maximum number of L1 or L2 adjacencies because an L1/L2 adjacency defines both an L1 adjacency and an L2 adjacency. The supported adjacency limits are defined in the *ExtremeXOS Release Notes*.

For information on creating and managing IS-IS areas, see Managing an IS-IS Area Address on page 1384.

## IS-IS and IP Routing

IS-IS is not inherently an IP Routing protocol. Its control packets do not use IP. RFC 1195 specifies the use of IS-IS for IP Routing.

The ExtremeXOS software implementation of IS-IS requires that at least one IPv4 or IPv6 address be assigned to an interface.

IP addresses and subnets can be assigned independent of area structure provided that neighboring interfaces are on the same subnet. L1 routers exchange directly reachable IP address/mask/metric tuples. When routing in an L1 area, packets are routed via L1 if the destination address is reachable within the area. Otherwise packets are forwarded to the nearest L2 router. L2 routers advertise all IP

addresses reachable in their L1 area (if they are a member of one), as well as directly reachable addresses.

*Summary Addresses*

Routers can be manually configured to advertise abbreviated versions of reachable addresses, which are called summary addresses. By aggregating many routes into one summary address, route computation and lookup can be made more efficient. If packets arrive at the router as a result of the summary address but have no actual route (as a result of the summary address covering more address space than is actually used), the packets are discarded.

For instructions on managing summary addresses, see Managing IP Summary Addresses on page 1384.

*External Connectivity*

Externally reachable IP addresses are those learned or exported from other routing protocols like *RIP*, *BGP (Border Gateway Protocol)*, and *OSPF*. When using narrow metrics, external routes may use internal or external metrics. When calculating SPF, routes with internal metrics are preferred over routes with external metrics. When external metrics are used to compare routes, the internal cost is not considered unless the external metrics are equal.

## Authentication

Entire packets (as opposed to individual TLVs in a packet) can be authenticated. If authentication fails on a packet, the entire packet is discarded. Multi-part packets are authenticated separately. Routers are not required to support authentication. The ExtremeXOS software provides optional support for plain-text authentication.

## Dynamic Hostname

The dynamic hostname exchange extension helps address the usability issues of maintaining IS-IS routers. IS-IS system IDs (which default to the switch MAC address) are not very readable. When the dynamic hostname feature is enabled, learned IS-IS hostnames are used in place of system IDs in log entries where possible. In addition, hostname TLVs appear in the `show isis lsdb detail` command display.

For instructions on managing the dynamic hostname feature, see Configuring the Dynamic Hostname Feature on page 1380.

## Route Leaking

Route leaking allows L2 LSPs to be sent into L1 areas. Route leaking is configured with the command:

`configure isis area` *`area_name`* **`interlevel-filter level 2-to-1`** [*`policy`* | **`block-all`** | **`allow-all`**] {**`ipv4`** | **`ipv6`**}

The supplied policy defines what route or routes should be leaked.

**Caution**
Route leaking can reduce scalability and performance.

Note
Tracert does not work if the VRF leaked route is one of the hop(s) to destination.

## Metric Types

Interface metrics are available in two sizes: a 6-bit narrow metric and a 24-bit wide metric. By default only narrow metrics are used. Wide metrics allow for greater flexibility, however, and are required in order to use some extensions of IS-IS, such as IPv6. All routers in an IS-IS area must use the same metric style.

For instructions on configuring interface metric style and values, see Configuring VLAN Interface Metrics on page 1385. Refer to RFC 3787, Section 5.1, for migration details. Note that the ExtremeXOS software does not support section 5.2.

## IS-IS Restart

When an IS-IS router restarts, neighbors time out adjacencies and the network converges around it. IS-IS restart support, with the help of restart helper neighbors, can prevent this. A restarting router can send a restart request to indicate to its neighbors that it is restarting. Neighbors—provided they are helpers—send the restarting router a CSNP so that it can reconstruct its LSDB by tracking which LSPs it has and has not received. Neighbors can still time out the adjacency, so this might not work in environments with low hold timers.

IS-IS restart can be configured for planned and unplanned events. Planned events are user-initiated process restarts and user-initiated failovers. Unplanned events are process restarts or failovers that are not administratively initiated.

For information on configuring IS-IS restart, see Configuring the Graceful Restart Feature on page 1379.

## IPv4 and IPv6 Topology Modes

Interfaces can be IS-IS-enabled for IPv4, IPv6, or both. Within an IS-IS area, IPv4 and IPv6 can be supported in a single topology or in a multiple topology. A transition configuration is provide for migrating between the single and multiple topologies.

In a single topology, all interfaces in an area must be configured for IPv4, IPv6, or both IPv4 and IPv6. Adjacencies are denied if the topology configuration between two routers does not match. A single SPF calculation is performed for each route in a single topology area.

In a multiple topology area, each interface can be configured for IPv4, IPv6, or both IPv4 and IPv6. The router creates separate topologies for IPv4 and IPv6. Adjacencies are permitted between interfaces with different configurations, provided that the neighbors support at least one common protocol. Separate SPF calculations are computed for IPv4 and IPv6 routes.

Note
Although the IPv4 and IPv6 topologies can be different when multiple topologies are enabled, the topologies must be convex. That is, no IPv4 interface can be reachable only through traversal of IPv6 interfaces, and vice versa. All routers in an area must use the same topology mode.

When the transition topology mode is configured, the router allows adjacencies between any two interfaces, and distributes both single topology and multiple topology TLVs. Transition mode permits both single and multiple topologies to coexist, but it generates more traffic. Transition mode should be used only for transitioning between modes and should be disabled after the transition is complete.

By default, a single topology is used for both IPv4 and IPv6. As a result, all interfaces must be homogeneously ISIS-enabled for IPv4, IPv6, or both.

To change the configured topology mode, see Configure the Multi-Topology Feature on page 1381.

## Route Redistribution

More than one routing protocol can be enabled simultaneously on a network. Route redistribution allows the switch to exchange routes, including static routes, between the active routing protocols. The following figure shows a network that is running two routing protocols, IS-IS and *RIP*.

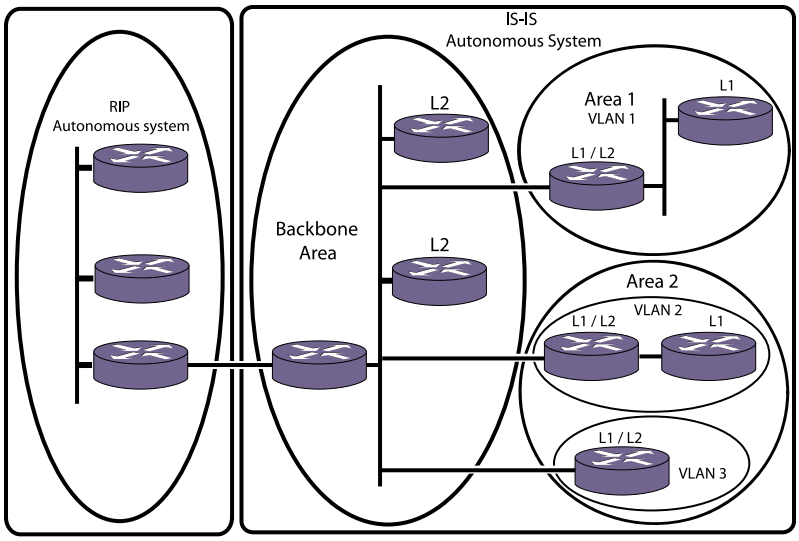



**Figure 230: AS External Route Redistribution**

By default, the IS-IS/RIP node does not inject its learned RIP routes into IS-IS—and it does not need to if originate-default feature is enabled using the command:

`enable isis area` *`area_name`* **`originate-default`** {**`ipv4`** | **`ipv6`**} .

If it does redistribute the RIP routes into IS-IS, all L2 IS-IS routers learn the RIP routes as type IS-IS level 2 external. The L2 IS-IS routers do not know that the routes actually originated from RIP. Also, if configured to do so, L1/L2 routers can leak these routes into the L1 areas and all IS-IS routers learn the RIP routes (without knowing that they actually came from RIP and without having to actually participate in RIP).

Route redistribution is configurable on both L2 and L1 using the `enable isis area export` commands. Redistributing routes into L1 generates L1 external routes. The export policy can choose to redistribute external routes with internal metrics into IS-IS.

## Configuring Route Redistribution

Exporting routes from one protocol to another and in the reverse direction are discrete configuration functions. For example, to run IS-IS and *RIP* simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from IS-IS to RIP and the routes to export from RIP to IS-IS. Likewise, for any other combinations of protocols, you must separately configure each to export routes to the other.

This section describes how to inject routing information learned from other IP routing protocols into an IS-IS domain, thereby advertising their reachability in the IS-IS domain. These routes are included in the locally originated LSPs. For information on exporting routes learned in IS-IS to another routing protocol, see the chapter that describes the destination routing protocol.

When you export routes from another protocol to IS-IS, the metric and the type of metric are configured. You can also configure a policy to filter out unwanted external routes.

- To enable or disable the exporting of *BGP*, OSPF, RIP, static, and direct (interface) IPv4 routes to IS-IS, use the following commands:

  `enable isis area` *area_name* **`export`** `{`**`ipv4`**`}` *route-type* `[`*policy* `|` **`metric`***mvalue* `{`**`metric-type`** `[`**`internal`** `|` **`external`**`]}] {`**`level`**`[`**`1`**`| 2 |`** **`both-1-and-2`**`]}`

  `disable isis area` *area_name* **`export`** `{`**`ipv4`**`}` *route-type*

- To enable or disable the exporting of *OSPFv3 (Open Shortest Path First version 3)*, *RIPng (Routing Information Protocol Next Generation)*, static, and direct (interface) IPv6 routes to IS-IS, use the following commands:

  `enable isis area` *area_name* **`exportipv6`***route-type* `[`*policy* `|` **`metric`***mvalue*`]` `{`**`level`**`[`**`1`**`| 2 |`** **`both-1-and-2`**`]}`

  `disable isis area` *area_name* **`exportipv6`***route-type*

  The cost metric is inserted for all

- *OSPF*, RIP, static, and direct routes injected into IS-IS.

  The same cost and type values can be inserted for all the export routes, or policies can be used for selective insertion. When a policy is associated with the export command, the policy is applied on every exported route. The exported routes can also be filtered using policies.

- Verify the configuration using the command:

  `show isis area [`*area_name* `|` **`all`**`]`

# Configuring IS-IS

## Configuring L1 Routers

To configure a switch to operate as a level 1 IS-IS router, do the following:

1. Prepare the IP interfaces that will support level 1 IS-IS routing as follows:
   a. Add an IPv4 or IPv6 address.
   b. Enable IP or IPv6 forwarding.
2. Create the IS-IS routing process, which is also called an area, using the following command:

   `create isis area` *area_name*

3. Configure the routing process to serve as a L1-only router using the following command:

   `configure isis area area_name is-type level [1 | 2 | both-1-and-2]`

   Specify 1 for the level option using the following command:
4. Add an IS-IS area level-1 address to the router using the following command:

   `configure isis area area_name add area-address area_address`
5. Add IS-IS-eligible interfaces to the area using the following command:

   `configure isis add [vlan all | {vlan} vlan_name] area area_name {ipv4 | ipv6}`

   An IS-IS-eligible interface is one that already has the appropriate IP address type (IPv4 or IPv6) address assigned to it.
6. The default IS-IS system ID is the switch MAC address. If you want to change the default IS-IS system ID, using the following command:

   `configure isis area area_name system-id [automatic | system_id]`
7. Enable the IS-IS router using the following command:

   `enable isis {area area_name}`

## Configuring L1/L2 Routers

To configure a switch to operate as a L1/L2 IS-IS router, do the following:

1. Prepare the IP interfaces that will support level 1 IS-IS routing as follows:
   a. Add an IPv4 or IPv6 address.
   b. Enable IP or IPv6 forwarding.
2. Create the IS-IS routing process, which is also called an area, using the following command:

   `create isis area area_name`
3. Configure the routing process to serve as an L1/L2 router using the following command:

   `configure isis area area_name is-type level [1 | 2 | both-1-and-2]`

   Specify both-1-and-2 for the level option using the following command:

   Note

   When no other L2 processes are defined on the router, the default IS type level is L1/L2, and this command can be omitted.
4. Add an IS-IS area level-1 address to the router using the following command:

   `configure isis area area_name add area-address area_address`
5. Add an IS-IS area level-2 address to the router using the following command:

   `configure isis area area_name add area-address area_address`
6. Add IS-IS-eligible interfaces to the area using the following command:

   `configure isis add [vlan all | {vlan} vlan_name] area area_name {ipv4 | ipv6}`

   An IS-IS-eligible interface is one that already has the appropriate IP address type (IPv4 or IPv6) address assigned to it.

7. If your topology requires interfaces to operate at a specific topology level, configure the appropriate interfaces with the following command:

   `configure isis [vlan all | {vlan} vlan_name] circuit-type level [1 | 2 | both-1-and-2]`

8. The default IS-IS system ID is the switch MAC address. If you want to change the default IS-IS system ID, use the following command:

   `configure isis area area_name system-id [automatic | system_id]`

   Note

   Although the IS-IS protocols manage IP routing, they use the Connectionless Network Protocol (CLNP) for IS-IS communications between routers. The IS-IS system ID is required to identify the router in the AS.

9. Enable the IS-IS router using the following command:

   `enable isis {area area_name}`

## Configuring L2 Routers

To configure a switch to operate as a level 2 IS-IS router, do the following:

1. Prepare the IP interfaces that will support level 2 IS-IS routing as follows:
   a. Add an IPv4 or IPv6 address.
   b. Enable IP or IPv6 forwarding.
2. Create the IS-IS routing process, which is also called an area, using the following command:

   `create isis area area_name`

3. Configure the routing process to serve as a L2-only router using the following command:

   `configure isis area area_name is-type level [1 | 2 | both-1-and-2]`

   Specify 2 for the level option.

4. Add an IS-IS area level-2 address to the router using the following command:

   `configure isis area area_name add area-address area_address`

5. Add IS-IS-eligible interfaces to level 2 using the following command:

   `configure isis add [vlan all | {vlan} vlan_name] area area_name {ipv4 | ipv6}`

   An IS-IS-eligible interface is one that already has the appropriate IP address type (IPv4 or IPv6) address assigned to it.

6. Add an IS-IS system ID to the router using the following command:

   `configure isis area area_name system-id [automatic | system_id]`

   Use the **automatic** option to assign the switch MAC address to the IS-IS system ID. The default option is **automatic**, so you can also enter the command without options to select the switch MAC address.

7. Enable the ISIS router using the following command:

   `enable isis {area area_name}`

## Configuring IS-IS Timers

IS-IS timers allow you to fine tune IS-IS operation. The following table lists the IS-IS timers and the command you can use to adjust them.

**Table 147: IS-IS Configuration Timers and Commands**

| Timer | Command |
|---|---|
| T1 hello restart interval | `configure isis [vlan all \| {vlan} vlan_name] timer restart-hello-interval seconds {level [1 \| 2]}` |
| T2 restart timer | `configure isis area area_name timer restart seconds {level [1 \| 2]}` |
| T3 restart grace period | `configure isis restart grace-period seconds` |
| LSP generation interval | `configure isis area area_name timer lsp-gen-interval seconds {level[1\| 2]}` |
| LSP maximum lifetime | `configure isis area area_name timer max-lsp-lifetime seconds` |
| LSP refresh interval | `configure isis area area_name timer lsp-refresh-interval seconds` |
| SPF interval | `configure isis area area_name timer spf-interval seconds {level[1\|2]}` |
| CSNP interval | `configure isis [vlan all \| {vlan} vlan_name] timer csnp-interval seconds {level [1 \| 2]}` |
| Hello interval | `configure isis [vlan all \| {vlan} vlan_name] timer hello-interval [seconds \| minimal] {level [1 \| 2]}` |
| Hello multiplier | `configure isis [vlan all \| {vlan} vlan_name] hello-multiplier multiplier {level [1 \| 2]}` |
| LSP retransmit interval | `configure isis [vlan all \| {vlan} vlan_name] timer retransmit-interval seconds` |
| LSP transmit interval | `configure isis [vlan all \| {vlan} vlan_name] timer lsp-interval milliseconds` |

## Configuring the Graceful Restart Feature

The graceful restart feature enables a router process to restart with minimal impact on the network. Without graceful restart, every other IS-IS router in the network must update its LSDB when the router goes down, and this generates a lot of traffic, and it introduces delays while new routes are established. With graceful restart, the router process requests restart support from its neighbors, and the neighbors wait for a predefined time before declaring routes dead. If the restart completes before the predefined time ends, the restart causes minimal impact on the network operation. For more information on graceful restart, see IS-IS Restart on page 1374.

- To enable or disable the graceful restart feature, use the following commands:

  ```
  enable isis restart-helper
  disable isis restart-helper
  ```

- To configure which events trigger a graceful restart, use the following commands:

  ```
  configure isis restart [ none | planned | unplanned | both ]
  ```

- To configure the timers used to control graceful restart operation, use the commands in the following table.

**Table 148: Timers that Control Graceful Restart**

| Timer | Configuration Command |
|---|---|
| T1 restart hello interval | `configure isis [vlan all \| {vlan} vlan_name] timer restart-hello-interval seconds {level [1 \| 2]}` |
| T2 restart timer | `configure isis area area_name timer restart seconds {level [1 \| 2]}` |
| T3 restart grace period | `configure isis restart grace-period seconds` |
| Hello interval | `configure isis [vlan all \| {vlan} vlan_name] timer hello-interval [seconds \| minimal] {level [1 \| 2]}` |
| Hello multiplier | `configure isis [vlan all \| {vlan} vlan_name] hello-multiplier multiplier {level [1 \| 2]}` |

## Configuring Hello Padding

Hello PDUs are padded to the MTU by default, and this is called hello padding. This is to help identify and correct cases where switches have mismatched MTUs. Because IS-IS packets cannot be fragmented in transit, maintaining consistent MTUs is important to ensure the reliability of packets traversing the network. However, this extra padding could be a waste of bandwidth in scenarios where the MTU is fixed throughout the network.

- To disable hello padding, use the following command:
  `disable isis [vlan all | {vlan} vlan_name] hello-padding`
- To enable hello padding, use the following command:
  `enable isis [vlan all | {vlan} vlan_name] hello-padding`

## Configuring Interlevel Filters

Interlevel filters allow you to control how routes are redistributed between IS-IS levels 1 and 2. The redistribution of routes from level 2 to level 1 is called route leaking. For more information, see Route Leaking on page 1373.

- To set an interlevel filter for redistribution from level 1 to level 2, use the following command:
  `configure isis area area_name interlevel-filter level 1-to-2 [policy | none] {ipv4 | ipv6}`
- To set an interlevel filter for redistribution from level 2 to level 1, use the following command:
  `configure isis area area_name interlevel-filter level 2-to-1 [policy | block-all | allow-all] {ipv4 | ipv6}`

## Configuring the Dynamic Hostname Feature

Dynamic Hostname on page 1373 introduces the dynamic hostname feature, which causes text names to replace numbers in some logs.

- To enable the dynamic hostname feature, use the following command:
  `enable isis area area_name dynamic-hostname [area-name | snmp-name]`

- To disable the dynamic hostname feature, use the following command:

  ```
  disable isis area area_name dynamic-hostname
  ```

### Configuring the Adjacency Check Feature

When enabled, the adjacency check feature permits an adjacency between two IS-IS interfaces only when both are configured for the same IP address type (IPv4 or IPv6) and are configured for the same subnet.

- To enable the adjacency check feature, use the following command:

  ```
  enable isis area area_name adjacency-check {ipv4 | ipv6}
  ```

- To disable the adjacency check feature, use the following command:

  ```
  disable isis area area_name adjacency-check {ipv4| ipv6}
  ```

### Configuring an Import Policy

When applied to a router process, an import policy controls how routes are imported to the FIB from all IS-IS processes on this *virtual router (VR)*.

- To apply a policy to a router process, use the following command:

  ```
  configure isis import-policy [policy-map | none]
  ```

Note

The import policy cannot be used to select which routes are added to the routing table. The import policy can only modify the route attributes as routes are added to the routing table.

### Configure the Multi-Topology Feature

The multi-topology feature is introduced in IPv4 and IPv6 Topology Modes on page 1374.

- To configure the multi topology feature, use the following command:

  ```
  configure isis area area_name topology-mode [single | multi |
  transition] {level [1 | 2]}
  ```

## Displaying IS-IS Information

### Displaying General Information for Global IS-IS

- To display general information for global IS-IS, use the following command:

  ```
  show isis
  ```

### Displaying Router-Specific Information

- To display router-specific information, use the following command:

  ```
  show isis area [area_name | all]
  ```

### Displaying Router Summary Addresses

- To display router summary addresses, use the following command:

  `show isis area area_name summary-addresses`

### Displaying IS-IS Interface Information

- To display IS-IS interface information, use the following command:

  `show isis vlan {enabled | { vlan_name | all} }`

### Displaying Link State Database Information

- To display link state database information, use the following command:

  `show isis lsdb {area area_name {lsp-id lsp_id}} {level [1|2]} {detail | stats}`

### Displaying IPv4 and IPv6 Topology Information

- To display IPv4 and IPv6 topology information, use the following command:

  `show isis topology {area area_name {level [1 | 2]}} {ipv4 | ipv6}`

### Displaying IS-IS Neighbors

- To display IS-IS neighbor information, use the following command:

  `show isis neighbors {area area_name} {vlan vlan_name} {ipv4 | ipv6} {detail}`

### Displaying IS-IS Counter Data

- To display IS-IS counter data, use the following command:

  `show isis counters {area [area_name | all] | vlan [vlan_name | all]}`

## Managing IS-IS

### Configuring Password Security

The ExtremeXOS software supports passwords for the following IS-IS AS components:

- Level 2 domains
- Level 1 areas
- *VLAN* interfaces

Domain and area authentication prevents intruders from injecting invalid routing information into the router. Because passwords must be configured to match at both ends of a connection, password

security also helps detect unconfigured and misconfigured interfaces. After configuration, the password is inserted into LSP, CSNP, and PSNP PDUs and are validated on the receiving end.

Note

The password configuration commands in this section provide an encrypted option, which controls how the passwords are saved and displayed on the switch. The encrypted option does not encrypt messages that are transmitted on the network. All passwords are transmitted in clear text.

- To configure password security for a level 2 domain, use the following command:

  `configure isis area area_name domain-password [none | {encrypted} simple password {authenticate-snp {tx-only}}]`

- To configure password security for a level 1 area, use the following command:

  `configure isis area area_name area-password [none | {encrypted} simple password {authenticate-snp {tx-only}}]`

  Interface authentication prevents unauthorized routers from forming an adjacency. This is achieved by inserting a password in hello PDUs and validating the password on the receiving end. You can configure password protection separately for level 1 and level 2.

- To configure password security for a VLAN interface, use the following command:

  `configure isis [vlan all | {vlan} vlan_name] password [none | {encrypted} simple password] level [1 | 2]`

## Managing Transit Traffic with the Overload Bit

If a router lacks the resources to maintain a complete link state database, routing packets to it could result in a blackhole or routing loop. The overload bit prevents this undesirable behavior by indicating to the other routers that the afflicted router should not be used as a transit router. Packets destined for directly attached routes can still be routed to a router with the overload bit set, but all other routing paths reconverge around it.

- To manually enable the overload bit feature, use the following command:

  `enable isis area area_name overload-bit {suppress [external | interlevel | all]}`

- To configure the router to automatically set the overload bit on startup, use the following command:

  `configure isis area area_name overload-bit on-startup [ off | {suppress [external | interlevel | all]} seconds]`

- To disable the overload bit feature, use the following command:

  `disable isis area area_name overload-bit`

Note

Enabling or disabling the overload bit feature does not modify the configuration of the command:

`configure isis area area_name overload-bit on-startup [ off | {suppress [external | interlevel | all]} seconds]`

## Clearing the IS-IS Counters

The `show isis counters {area [area_name | all] | vlan [vlan_name | all]}` command can display all counters or only those specific to the router process or the configured VLANs.

- To clear the IS-IS counters for the router and the VLANs, use the following command:

  `clear isis counters`
- To clear the IS-IS counters only for the router process, use the following command:

  `clear isis counters area [area_name | all]`
- To clear the IS-IS counters only for one or all VLANs, use the following command:

  `clear isis counters [vlan all | {vlan} vlan_name]`

## Originating an L2 Default Route

This feature injects a zero-cost route to 0.0.0.0 in LSPs originated by an L2 router, thereby advertising the router as the default gateway.

- To enable default route origination, use the following command:

  `enable isis area area_name originate-default {ipv4 | ipv6}`
- To disable default route origination, use the following command:

  `disable isis area area_name originate-default {ipv4 | ipv6}`

## Managing IP Summary Addresses

Summary addresses are introduced in Summary Addresses on page 1373. L2 routers include in their L2 LSPs a list of all destinations reachable in the L1 area attached to them by default. Summarization of the L1 destinations reduces the amount of information stored on each L2 router and helps in scaling to a large routing domain.

- To add summary addresses to an IS-IS router, use the following command:

  `configure isis area area_name add summary-address [ipv4_address_mask | ipv6_address_mask] {level [1 | 2]}`
- To delete summary addresses from an IS-IS router, use the following command:

  `configure isis area area_name delete summary-address [ipv4_address_mask | ipv6_address_mask] {level [1 | 2]}`
- To display the configured summary addresses for an IS-IS router, use the following command:

  `show isis area area_name summary-addresses`

## Managing an IS-IS Area Address

IS-IS was not originally designed for IP routing. Some aspects of CLNP routing, such as system IDs and area addresses, exist in the IP extension of IS-IS. The IS-IS area address is required to identify the area in which the router participates.

- Before you can configure the area address, you must create the IS-IS routing process, which is also called an area, using the following command:

  `create isis area area_name`

- To add an IS-IS area address to the router, use the following command:

  `configure isis area area_name add area-address area_address`
- To delete an IS-IS area address from a router, use the following command:

  `configure isis area area_name delete area-address area_address`

Note
The ExtremeXOS software implementation of IS-IS supports no more than three area addresses.

## Managing VLAN Interfaces

### *Adding a VLAN Interface*

- To add a *VLAN* interface to a router, use the following command:

  `configure isis add [vlan all | {vlan} vlan_name] area area_name {ipv4 | ipv6}`

Note
The interface is not enabled separately from the area. If the area is enabled, the interface will begin transmitting hellos as soon as this command is executed, provided the interface is in forwarding mode and has active links.

### *Setting the VLAN Interface Link Type*

The default link type is broadcast. You can change the link type to point-to-point.

- To set the link type, use the following command:

  `configure isis [vlan all | {vlan} vlan_name] link-type [broadcast | point-to-point]`

### *Setting the VLAN Interface Circuit Type*

- The circuit type can be level 1, level 2, or both level 1 and level 2. To set the circuit type, use the following command:

  `configure isis [vlan all | {vlan} vlan_name] circuit-type level [1 | 2 | both-1-and-2]`

### *Configuring VLAN Interface Metrics*

Normally, IS-IS metrics can have values up to 63. These metrics are narrow metrics. IS-IS generates two type, length, and value (TLV) codings, one for an IS-IS adjacency (code, length, and value (TLV) 2) and the second for an IP prefix (TLV 128 and TLV 130). During SPF, if the total cost of the path to a destination exceeds 1023, then according to ISO/IEC 10587, the path is ignored.

To overcome these restrictions, a second pair of TLVs is available, one for IP prefixes (TLV 135) and the second for IS-IS adjacency (TLV 22). With these TLVs, IS-IS metrics can have values up 16,777,215 and

the maximum path metric allowed is 4,261,412,864. This allows more flexibility while designing a domain. These metrics are wide metrics.

- To configure the metric style used on the router, use the following command:
  `configure isis area area_name metric-style [[narrow | wide] {transition}] | transition] {level [1 | 2]}`
- To configure the narrow metric used on one or all interfaces, use the following command:
  `configure isis [vlan all | {vlan} vlan_name] metric metric {level[1| 2]}`
- To configure the wide metric used on one or all interfaces, use the following command:
  `configure isis [vlan all | {vlan} vlan_name] wide-metric metric {level[1 | 2]}`

Note
Configured narrow and wide metrics for a particular interface must be identical in value while migrating from metric style narrow to metric style wide. Only if the metric style is "narrow" or "wide" (that is, no "transition") is it okay to have different values (because one of the values is not used).

### *Configuring the DIS Priority for Broadcast Interfaces*

The designated intermediate system (DIS) is introduced in Establishing Adjacencies on page 1369. When you configure the DIS priority, higher priority means higher election precedence (60 gets elected before 50).

- To configure the DIS priority, use the following command:
  `configure isis [vlan all | {vlan} vlan_name] priority priority {level[1 | 2]}`

Note
DIS priority 0 is the lowest priority value. Unlike an *OSPF* DR election, a DIS priority of 0 does not make a router ineligible for DIS election.

### *Configuring Interface Participation in a Mesh Environment*

Fully-meshed point-to-point topologies suffer from a potentially large amount of needless LSP flooding. At the trade-off of resiliency, interfaces can be restricted from flooding received LSPs.

- To configure an interface for a mesh environment, use the following command:
  `configure isis [vlan all | {vlan} vlan_name] mesh [block-none | block-all | block-group group_id]`
  a. To block the flooding of received LSPs altogether, use the **mesh block-all** option.
  b. To block the flooding of LSPs received on a specific group of interfaces (which are those with the same mesh group ID), use the **mesh block-group** *group_id* option.
  c. To remove blocking for an interface, use the **mesh block-none** option.

### *Resetting a VLAN Interface to the Default Values*

- To reset a *VLAN* interface to the default values, use the following command:
  `unconfigure isis [vlan all | {vlan} vlan_name] {level [1|2]}`

*Deleting a VLAN Interface*

- To delete a *VLAN* interface from a router, use the following command:
  `configure isis delete [vlan all | {vlan} vlan_name] {area area_name} {ipv4 | ipv6}`

## Managing IS-IS Routers

*Adding an IS-IS Router*

- To add an IS-IS router, use the following command:
  `create isis area area_name`

*Changing the IS-IS Level of a Router*

- To change the IS-IS level of a router, use the following command:
  `configure isis area area_name is-type level [1 | 2 | both-1-and-2]`

*Resetting an IS-IS Router to the Default Values*

- To reset an IS-IS router to the default values, use the following command:
  `unconfigure isis area area_name {level [1|2]}`

*Restarting All IS-IS Routers in a Virtual Router*

- To restart all IS-IS routers in a *VR*, use the following command and specify the IS-IS process:
  `restart process [class cname | name {msm slot}]`
  For example:
  ```
  restart process isis
  ```

*Disabling an IS-IS Router*

- To disable an IS-IS router, use the following command:
  `disable isis {area area_name}`

*Deleting an IS-IS Router*

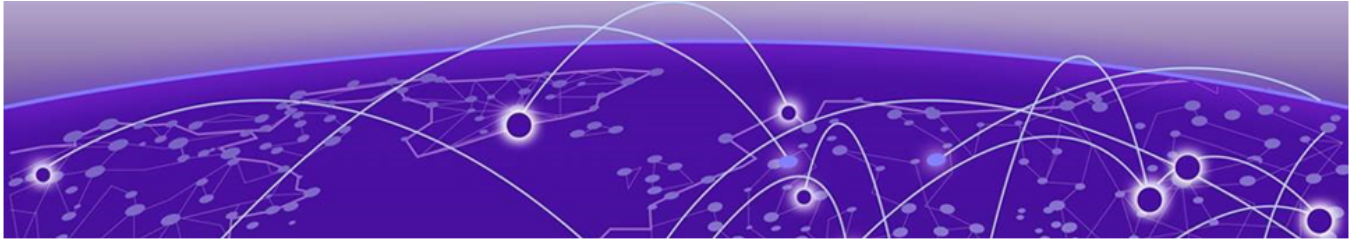
- To delete an IS-IS router, use the following command:
  `delete isis area [all | area_name]`

# Configuration Example

The following example shows the commands that configure an IS-IS router. Some commands apply to IPv4 or IPv6 and are labeled accordingly. Comments in parentheses identify commands that apply to specific applications.

```
create vlan v1
configure default delete ports 1
configure v1 add ports 1

IPv4:
configure v1 ipaddress 10.0.0.1/24
enable ipforwarding v1
```

```
IPv6:
configure v1 ipaddress fe80::204:96ff:fe20:b40a/128
enable ipforwarding ipv6 v1
create isis area a1
configure isis area a1 add area-address 01.0101.0202.0303.0404.0505.0606
configure isis area a1 system-id 11aa.22bb.33cc
configure isis area a1 is-type level 1 (For Level 1 Router)
configure isis area a1 is-type level 2 (For Level 2 Router)
configure isis area a1 is-type level both-1-and-2 (For Level 1/2 Router)
configure isis area a1 metric-style wide

IPv4 Mapping:
configure isis add vlan v1 area a1

IPv6 Mapping:
configure isis add vlan v1 area a1 ipv6

enable isis area a1 (or) enable isis
```



# BGP



The *BGP (Border Gateway Protocol)* chapter is intended to provide information on the border gateway protocol. In this chapter you will find content on configuring and managing BGP, displaying BGP information, in addition to comprehensive BGP configuration examples.

## BGP Overview

*BGP* is an exterior routing protocol that was developed for use in TCP/IP networks. The primary function of BGP is to allow different autonomous systems (ASs) to exchange network reachability information.

An AS is a set of routers that are under a single technical administration. This set of routers uses a different routing protocol, for example *OSPF (Open Shortest Path First)*, for intra-AS routing. One or more routers in the AS are configured to be border routers, exchanging information with other border routers (in different ASs) on behalf of all of the intra-routers.

BGP can be used as an exterior border gateway protocol (referred to as EBGP), or it can be used within an AS as an interior border gateway protocol (referred to as IBGP).

For more information on BGP, refer to the following documents:

- RFC 1745—BGP/IDRP for IP—OSPF Interaction
- RFC 1771—Border Gateway Protocol version 4 (BGP-4)
- RFC 1965—Autonomous System Confederations for BGP
- RFC 1966—BGP Route Reflection
- RFC 1997—BGP Communities Attribute
- RFC 2385—Protection of BGP Sessions via the TCP RSA Data Security, Inc. *MD5 (Message-Digest algorithm 5)* Message-Digest Algorithm Signature Option
- RFC 2439—BGP Route Flap Damping
- RFC 2545—Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing
- RFC 2796—BGP Route Reflection - An Alternative to Full Mesh IBGP
- RFC 2918—Route Refresh Capability for BGP-4
- RFC 3392—Capabilities Advertisement with BGP-4
- RFC 4271—Border Gateway Protocol 4 (BGP-4)

- RFC 4360—BGP Extended Communities Attribute
- RFC 4456—BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)
- RFC 4486—Subcodes for BGP Cease Notification Message
- RFC 4724—Graceful Restart Mechanism for BGP
- RFC 4760—Multiprotocol Extensions for BGP-4
- RFC 4893—BGP Support for Four-octet AS Number Space
- RFC 5396—Textual Representation of Autonomous System (AS) Numbers
- draft_ieft_idr_restart_10.txt—Graceful Restart Mechanism for BGP

Note

ExtremeXOS supports BGP version 4 only, and does not support connections to peers running older versions of BGP.

For complete information about software licensing, including how to obtain and upgrade your license and what licenses are appropriate for these features, see the *Feature License Requirements* document..

## BGP Four-Byte AS Numbers

The ExtremeXOS software supports 4-byte AS numbers, which can be entered and displayed in the ASPLAIN and ASDOT formats, which are described in RFC 5396, Textual Representation of Autonomous System (AS) Numbers.

Note

When entering an AS number in a policy file, you must enter a unique 2-byte or 4-byte AS number. The transition AS number, AS 23456, is not supported in policy files.

## BGP Attributes

The following *BGP* attributes are supported by the switch:

- Origin—Defines the origin of the route. Possible values are Interior Gateway Protocol (IGP), Exterior Gateway Protocol (EGP), and incomplete.
- AS_Path—The list of ASs that are traversed for this route. The local AS-path is added to the BGP update packet after the policy is applied.
- AS4_Path—This attribute is used by 4-byte peers when sending updates to 2-byte peers. This attribute carries AS-number information that can be represented only in 4-bytes.
- Next_hop—The IP address of the next hop BGP router to reach the destination listed in the NLRI field.
- Multi_Exit_Discriminator—Used to select a particular border router in another AS when multiple border routers exist.
- Local_Preference—Used to advertise this router's degree of preference to other routers within the AS.
- Atomic_aggregate—Indicates that the sending border router has used a route aggregate prefix in the route update.
- Aggregator—Identifies the BGP router AS number and router ID for the router that performed route aggregation.

- AS4_Aggregator: This attribute is used by 4-byte peers when sending updates to 2-byte peers. This attribute carries AS-number information that can be represented only in 4-bytes.
- Community—Identifies a group of destinations that share one or more common attributes.
- Cluster_ID—Specifies a 4-byte field used by a route reflector to recognize updates from other route reflectors in the same cluster.
- Originator_ID—Specifies the router ID of the originator of the route in the local AS.
- Extended Community—Provides a mechanism for labeling BGP-4 update messages that carry information
- Multiprotocol reachable NLRI—Used to advertise a feasible BGP route for the non IPv4-unicast address family
- Multiprotocol unreachable NLRI—This attribute is used to withdraw multiple unfeasible routes from service

## BGP Community Attributes

A *BGP* community is a group of BGP destinations that require common handling. ExtremeXOS supports the following well-known BGP community attributes:

- no-export
- no-advertise
- no-export-subconfed

## Extended Community Attributes

The extended community attribute provides a mechanism to label *BGP* routes. It provides two important enhancements over the standard community attribute:

- An expanded range. Extended communities are 8-bytes wide, whereas regular communities were only 4-bytes wide. So, this ensures that extended communities can be assigned for a plethora of uses, without the fear of overlap.
- The addition of a 'Type' field provides structure for the extended community

The following two types of extended communities are available:

- Route Target (RT)
- Site Of Origin (SOO)

Although these two community types are generally used in L3 VPN network setup, you can also use them in a non-L3 VPN network to control the distribution of BGP routes.

BGP does not send either the extended or standard community attributes to their neighbors by default; you must use the configuration command `configure bgp neighbor send-community`.

### *Extended Community Processing*

When *BGP* receives the extended community attribute in a route from its neighbor, it validates the community syntax. If the community is syntactically valid, the inbound neighbor route-policy is applied to the route. The inbound route-policy may contain extended-community statements in match block (in other words, an or/and set) of the policy. If the route is not rejected by the inbound route-policy, it is

added to the LocRIB of the BGP along with the extended community. The **detail** option of the show bgp routes command displays the routes with the extended community attribute if they are present in that route's path attribute.

### *Associating the Extended Community Attribute to the BGP Route*

The extended community attribute can be added to or removed from a *BGP* route using an ExtremeXOS policy in the same way this action is performed for a regular community attribute.

The extended-community keyword has been added in the Policy Manager, and can be used in the match as well as in the set block of a policy file.

#### Syntax in Match block

```
extended-community "<extended-community-1> <extended-community-2> …"
```

Where, the syntax of <extended-community-N> is

```
[rt|soo]:[<2-byte AS num>:<4-byte num> | <4-byte IP Address>:<2-byte num> | <4-byte AS
num>L:<2-byte num> | <first two bytes of AS num>.<last two bytes of AS num>:<2-byte num>]
```

The attributes are defined as follows:

- **rt**: route target extended community type
- **soo**: site of origin extended community type
- **<2-byte AS number>**: This is a 2-byte AS number; the use of private AS-number is not recommended
- **<4-byte num>**: a 4-byte unsigned number
- **<4-byte IP address>**: a valid host IP address; a network address is not accepted; use of private IP address is not recommended; class-D IP addresses are rejected
- **<4-byte AS num>**: This is a 4-byte AS number; the use of private AS-number is not recommended
- **<first two bytes of AS num>**: This is the number represented by the first two bytes of a four-byte AS number. The use of a private AS-number is not recommended.
- **<last two bytes of AS num>**: This is the number represented by the last two bytes of a four-byte AS number. The use of a private AS-number is not recommended

#### Syntax in Set block

```
extended-community  [set | add | delete] "<extended-community-1> <extended-community-2>
…."
extended-community remove
```

Where, the syntax of <extended-community-N> is the following:

```
[rt|soo]:[<2-byte AS num>:<4-byte num> | <4-byte IP Address>:<2-byte num>]
```

The attributes are defined as follows:

- **set**: Replaces the existing extended communities by the new ones as supplied in the policy statement.
- **add**: Adds new extended communities to the existing extended community attribute. If an extended community is already present, then policy will not add a duplicate extended community to the route.
- **delete**: Deletes some of the extended communities from the extended community attribute.
- **remove**: Removes extended community attribute from the route.
- **rt**: route target extended community type
- **soo**: site of origin extended community type
- **<2-byte AS number>**: This is 2-byte AS number. Use of private AS-number is not recommended
- **<4-byte num>**: 4-byte unsigned number
- **<4-byte IP address>**: A valid host IP address. Network address is not accepted. Use of private IP address is not recommended. Class-D IP address will be rejected.

*Examples of Extended Communities*

The following are examples of valid extended communities:

- rt:10.203.134.56:400
- soo:64500:1600
- rt:64511:2345678
- soo:172.168.45.10:500
- rt:1.15:20000
- soo:65551L:50000

The following are examples of invalid extended communities:

- rt:10.45.87.0:600: Invalid because the IP address is NOT a valid host IP address
- rt:239.1.1.1:400: Invalid because IP address belongs to class-D
- rt:100.200.300.400:200: Invalid because the IP address is invalid
- soo:12345678:500: Invalid because the AS number 12345678 is out of the 2-byte AS number range, 1 to 65535

*Extended Community Syntax*

Please note the following details with regard to extended community syntax:

- Only rt and soo extended community types are recognized in the policy file.
- The IP address MUST be a valid host address. Network address, Class-D and experimental IP address are not accepted.
- There should not be any blank spaces inside an extended community. For example, rt :100:200 is not a valid extended community because there are spaces between rt and :
- All three parameters of an extended community must be present, otherwise the extended community is rejected.

*Extended Community Match Rule in Policy*

Regular expressions are not supported for extended communities. In addition, an extended community match statement matches with a route's extended community if at least one of the extended

communities in the match statement matches with the route's extended community. There is no need to for all the extended communities in a single match statement to match with the route's extended community.

For example, suppose the policy file is the following:

```
entry one {
if {
extended-community "rt:64500:20000 rt:10.203.134.5:40 soo:64505:50000
soo:192.168.34.1:600";
} then {
permit;
}
}
```

The above community statement will match with all *BGP* routes that have at least one of the following extended communities in their extended community attribute:

- rt:64500:20000
- rt:10.203.134.5:40
- soo:64505:50000
- soo:192.168.34.1:600

### *Extended Community Set Rule in Policy*

A Policy set block can contain several extended community statements. Each set statement is applied to the matching route's extended community attribute in the top down order. That is, the first set is applied to the extended community attribute of the route, the second set is applied to the result of above, and so forth.

For example, assume that a policy is the following:

```
entry two {
if {
nlri 192.168.34.0/24;
} then {
extended-community set "rt:10.45.92.168:300";
extended-community add "rt:10.203.100.200:40 soo:64500:60000";
extended-community delete "rt:64505:10000 soo:72.192.34.10:70";
permit;
}
}
```

A *BGP* route 192.168.34.128/25 is received with extended community attribute rt:4567:100 soo:192.168.34.128.

When the above policy entry is applied to the route's extended community attribute, the following is true:

- After applying the 1st set (community set "rt:10.45.92.168:300"), the route's community becomes rt:10.45.92.168:300.
- After applying the 2nd set (community add "rt:10.203.100.200:40 soo:64500:60000"), the community becomes rt:10.45.92.168:300 rt:10.203.100.200:40 soo:64500:60000.
- After applying the 3rd set (community delete "rt:64505:10000 soo:72.192.34.10:70"), the community becomes rt:10.45.92.168:300 rt:10.203.100.200:40 soo:64500:60000. Please note that this delete

statement has no effect as none of the communities in the delete statement are present in the community attribute.

*Extended Communities and BGP Route Aggregation*

When *BGP* routes are aggregated with the as-match or as-set CLI option, all the component route's extended community attributes are aggregated and the resulting aggregated extended community attributes are attached to the aggregate network.

Aggregation of several extended community attributes is simply the set union of all the extended communities from all of the aggregated routes.

## Multiprotocol BGP

Multiprotocol *BGP* (MBGP), which is also known as BGP4+, is an enhanced BGP that supports more than just IPv4 unicast routes. In this release, MBGP supports:

- IPv4 unicast routing and IPv4 multicast routing on the platforms listed for this feature in the *Feature License Requirements* document.
- IPv6 unicast routing and IPv6 multicast routing on the platforms listed for this feature in the *Feature License Requirements* document.
- Layer 3 VPN routing on the platforms listed for this feature in in the *Feature License Requirements* document..

MBGP support for separate unicast and multicast routing tables allows BGP to have non-congruent topologies for unicast and multicast networks. The BGP multicast address-family routes are used by multicast protocols such as Protocol Independent Multicast (PIM) to build data distribution trees.

## Route Reflectors

One way to overcome the difficulties of creating a fully meshed AS is to use route reflectors. Route reflectors allow a single router to serve as a central routing point for the AS. All *BGP* speakers in the AS will peer with the route reflector to learn routes.

A cluster is formed by the route reflector and its client routers. Peer routers that are not part of the cluster must be fully meshed according to the rules of BGP.

A BGP cluster, including the route reflector and its clients, is shown in the following figure.

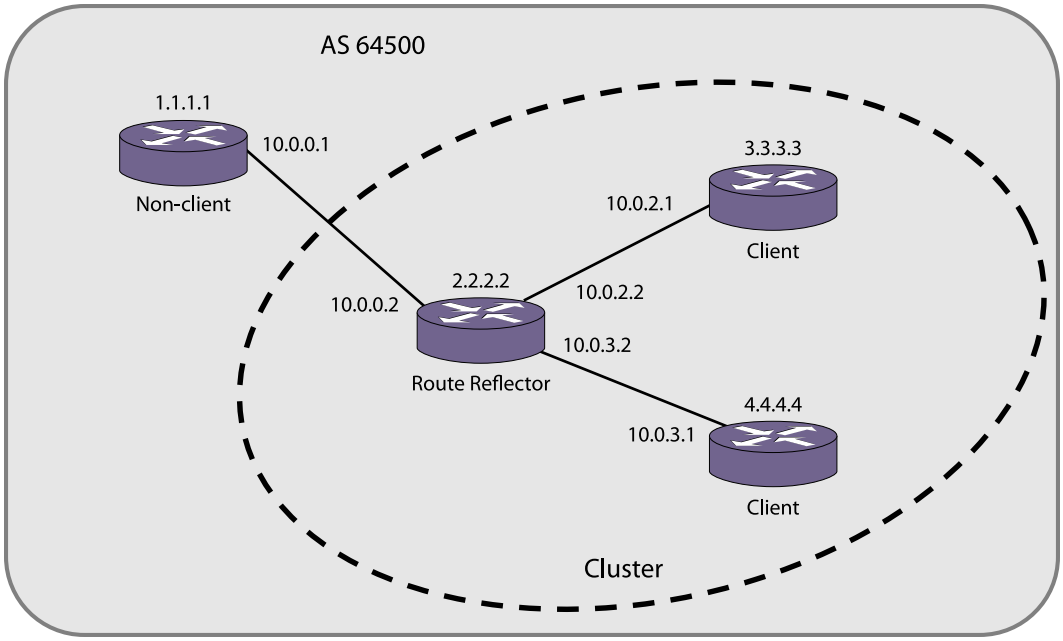



**Figure 231: Route Reflectors**

The topology shown in the figure above minimizes the number of BGP peering sessions required in an AS by using a route reflector.

In this example, although the BGP speakers 3.3.3.3 and 4.4.4.4 do not have a direct BGP peering session between them, these speakers still receive routes from each other indirectly through 2.2.2.2. The router 2.2.2.2 is called a route reflector and is responsible for reflecting routes between its clients. Routes received from the client 3.3.3.3 by the router 2.2.2.2 are reflected to 4.4.4.4 and vice-versa. Routes received from 1.1.1.1 are reflected to all clients.

## Route Confederations

*BGP* requires networks to use a fully meshed router configuration. This requirement does not scale well, especially when BGP is used as an IGP. One way to reduce the size of a fully meshed AS is to divide the AS into multiple sub-ASs and to group these sub-ASs into a routing confederation. Within the confederation, each sub-AS must be fully meshed. The confederation is advertised to other networks as a single AS.

The following figure shows an example of a confederation.

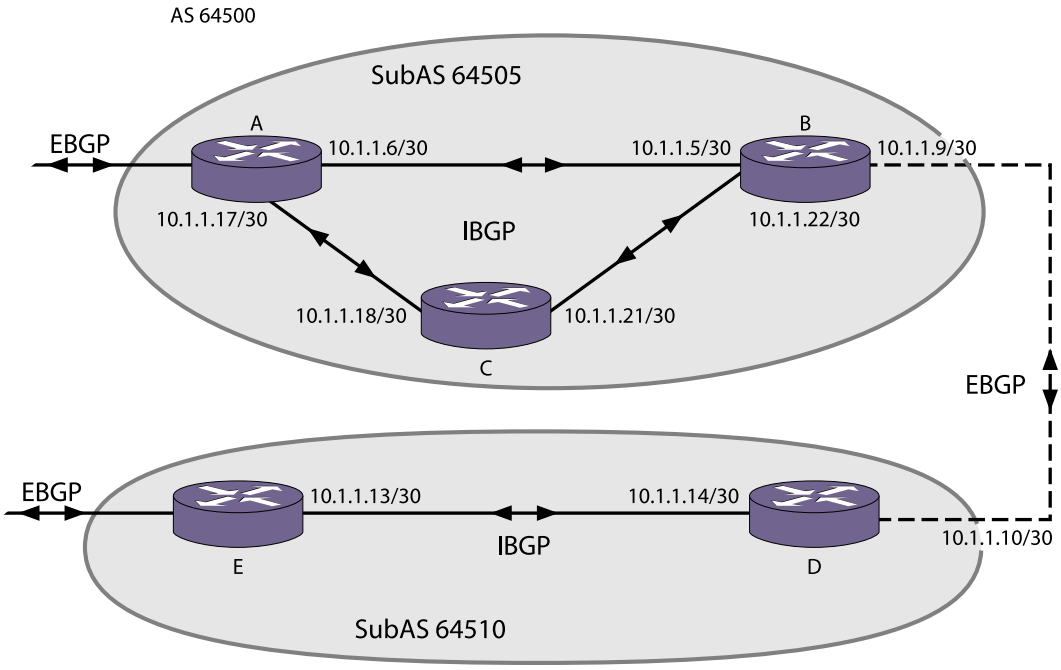



**Figure 232: Routing Confederation**

In this example, AS 64500 has five BGP speakers. Without a confederation, BGP would require that the routes in AS 64500 be fully meshed. Using the confederation, AS 64500 is split into two sub-ASs: AS 64505 and AS 64510. Each sub-AS is fully meshed, and IBGP is running among its members. EBGP is used between sub-AS 64505 and sub-AS 64510. Router B and router D are EBGP peers. EBGP is also used between the confederation and outside ASs.

## Inactive Route Advertisement

*BGP* inactive routes are defined as those routes that are rated best by BGP and not best in IP routing table. For example, an IGP route to the same destination may be best because it has a higher priority in the IP route table than the BGP best route. The default configuration of the ExtremeXOS software does not advertise BGP inactive routes to BGP neighbors.

The default configuration (no BGP inactive route advertisement) is more consistent with data traffic forwarding. However, when advertisement of inactive BGP routes is enabled, BGP need not depend upon the route manager module to know whether a BGP route is active or not. This actually improves the performance of BGP processing and advertisement.

When BGP inactive route advertising is enabled, inactive BGP routes are considered for BGP route aggregation. When this feature is disabled, inactive BGP routes are ignored while aggregating routes.

## Default Route Origination and Advertisement

The default route origination and advertisement feature allows you to originate and advertise a default route to a *BGP* neighbor (or to all neighbors in a peer group) even though no default route exists in the local IP routing table. It also allows you to associate policy rules to conditionally advertise a default route to BGP neighbors.

When default route origination becomes active, the default route is advertised to the specified BGP neighbors, overriding any previously sent default route. If a default route is added to the local IP routing table while default route origination is active, the default route defined by this feature takes precedence over the new regular default route. If default route origination becomes inactive, and a regular default route exists, the regular default route is advertised to BGP neighbors.

When you use a policy with default route origination, the default route is originated only if the local BGP RIB contains a route that matches the policy match conditions. You can use the following match conditions:

- NLRI
- AS-path
- Community
- Origin

You can also use the following policy actions in the policy to set the route attributes:

- AS-path
- Community
- Origin

After a policy is configured for default route origination, BGP must periodically scan the local BGP RIB to make sure that the policy rules evaluate to true for at least one route in local BGP RIB. If the rules evaluate to true, default origination remains active. If the rules evaluate to false, then default origination becomes inactive and the default routes must be withdrawn.

For more information on policy match conditions, actions, and configuration, see Routing Policies.

## Using the Loopback Interface

If you are using *BGP* as your IGP, you may decide to advertise the interface as available, regardless of the status of any particular interface. The loopback interface can also be used for EBGP multihop. Using the loopback interface eliminates multiple, unnecessary route changes.

## Looped AS_Path Attribute

When an EBGP speaker receives a route from its neighbor, it must validate the AS_Path attribute to ensure that there is no loop in the AS_Path. When an EBGP speaker finds it's own AS-number in the received EBGP route's AS_Path attribute, it is considered as a Looped AS Path and by default, the associated EBGP routes are silently discarded.

There are certain cases (such as in a spoke and hub topology) where it becomes necessary to accept and process EBGP routes with a looped AS_Path attribute.

This feature enables you to control the processing of EBGP routes with a looped AS_Path attribute. You can do the following:

- Allow the route with a looped AS_Path to be accepted
- Allow the route with a looped AS_Path to be accepted only if the number of occurrences of its own AS-Number is less than or equal to N, where N is an user-configurable unsigned integer configured

- Silently discard the route with looped AS_Path

Note

A looped AS path is always allowed for IBGP, irrespective of the *BGP* configuration.

## BGP Peer Groups

You can use *BGP* peer groups to group together up to 512 BGP neighbors. All neighbors within the peer group inherit the parameters of the BGP peer group. The following mandatory parameters are shared by all neighbors in a peer group:

- remote AS
- source-interface
- route-policy
- send-community
- next-hop-self

Each BGP peer group is assigned a unique name when it is created, and each peer supports either IPv4 or IPv6 address families (but not both).

When the first peer is added to a BGP peer group, the peer group adopts the IPv4 or IPv6 address family of the assigned peer. After the first peer is assigned, the peer group does not support any configuration options or capabilities for the IP address family not selected. For example, if an IPv6 peer is the first peer added to a peer group, the peer group no longer supports IPv4 configuration options or capabilities.

Changes made to the parameters of a peer group are applied to all neighbors in the peer group.

Modifying the following parameters will automatically disable and enable the neighbors before changes take effect:

- remote-as
- timer
- source-interface
- soft-in-reset
- password

## BGP Route Flap Dampening

Route flap dampening is a *BGP* feature designed to minimize the propagation of flapping routes across an internetwork. A route is considered to be flapping when it repeatedly alternates between being available and being unavailable. Without route flap dampening, each transition of a route produces a withdrawal or advertisement message, which is propagated throughout the network. Route flapping can generate large numbers of messages, which can impact network bandwidth and availability.

The route flap dampening feature minimizes the flapping problem by halting route advertising and withdrawal messages for the affected route for a period of time. To support route flap dampening, the ExtremeXOS software employs a combination of two techniques. The first technique uses fixed timers to reduce the frequency of route advertisement as specified in RFC 4271. The other technique uses

route flap damping algorithms. The software uses a combination of both techniques as specified in RFC 2439.

The fixed timers technique blocks all updates for the flapping route for a period defined by the internal MinRouteAdvertisementInterval (MRAI) timer (which is not configurable). For IBGP routes, this timer is set for 5 seconds. For EBGP routes, this timer is set to 30 seconds. The MRAI timer check is independent of the dampening configuration and is used to limit the frequency of route advertisements.

The route flap dampening algorithm uses configurable timers to manage route flap dampening. Suppose that the route to network 172.25.0.0 flaps. The router (in which route dampening is enabled) responds by doing the following:

- Assigns route 172.25.0.0 a penalty of 1000 and moves it to a history state in which the penalty value is monitored.
- Increments the penalty value by 1000 for each additional route flap.
- Accumulates penalties and compares them to the suppression limit, which is set to 2000 by default.

If the suppression limit is exceeded when the MRAI timer expires, the route is not advertised to neighbors.

A route remains suppressed or dampened until one of the following events occurs:

- The suppression limit is not met when the MRAI timer expires.
- The penalty placed on network 172.25.0.0 is decayed below the reuse limit.
- The maximum suppression timer expires.

The penalty is decayed by reducing the penalty value by one-half at the end of a configurable time period, called the half-life. Routes that flap many times may reach a maximum penalty level, or ceiling, after which no additional penalty is added. The ceiling value is not directly configurable, but the configuration parameter used in practice is the maximum route suppression time. No matter how often a route has flapped, after it stops flapping, it is advertised after the maximum route suppression time.

## BGP Route Selection

*BGP* selects routes based on the following precedence (from highest to lowest):

- Next-hop must be reachable.
- higher weight
- higher local preference
- shortest length (shortest AS path)
- lowest origin code
- lowest Multi Exit Discriminator (MED)
- route from external peer
- lowest cost to next hop
- lowest routerID
- lowest PeerID

## Private AS Number Removal from Route Updates

Private AS numbers are AS numbers in the range 64512 through 65534. You can remove private AS numbers from the AS path attribute in updates that are sent to external *BGP* (EBGP) neighbors. Possible reasons for using private AS numbers include:

- The remote AS does not have officially allocated AS numbers.
- You want to conserve AS numbers if you are multihomed to the local AS.

Private AS numbers should not be advertised on the Internet. Private AS numbers can be used only locally within an administrative domain. Therefore, when routes are advertised out to the Internet, the routes can be stripped out from the AS paths of the advertised routes using this feature.

## Route Redistribution

Multiple protocols, such as *BGP*, *OSPF*, and *RIP (Routing Information Protocol)*, can be enabled simultaneously on the switch. Route redistribution allows the switch to exchange routes, including static and direct routes, between any two routing protocols.

Exporting routes from another protocol to BGP and from BGP to another protocol are discrete configuration functions. For example, to run *OSPFv3 (Open Shortest Path First version 3)* and BGP simultaneously, you must first configure both protocols and then verify the independent operation of each. Then you can configure the routes to export from OSPFv3 to BGP and the routes to export from BGP to OSPFv3.

## BGP ECMP

The *BGP ECMP (Equal Cost Multi Paths)* feature supports load sharing by creating a multipath to a destination. This multipath contains multiple routes that are determined to have an equal cost because the following parameters are the same for each route:

- Weight
- Local preference (for IBGP multipaths)
- AS path (entire attribute, not just the length)
- Origin code
- Multi Exit Discriminator (MED)
- IGP distance to the next hop
- Source session (EBGP or IBGP)

**Note**

ECMP does not install an additional path if the next hop is the same as that of the best path. All paths within a multipath must have a unique next hop value.

BGP ECMP does not affect the best path selection. For example, the router continues to designate one of the paths as the best path and advertise this best path to its neighbors. EBGP paths are preferred to IBGP paths.

The BGP ECMP feature allows you to define the maximum number of equal cost paths (up to eight) in a multipath. A multipath for an IBGP destination is called an IBGP multipath, and the multipath for an EBGP destination is called an EBGP multipath.

If there are more equal cost paths for a destination than the configured maximum, the BGP identifier for the advertising BGP speaker is used to establish a path priority. The lower BGP identifier values have priority over the higher values. For example, if the configuration supports 4 paths in a multipath, only the four paths with the lowest BGP identifier values become part of the multipath.

## BGP Static Network

ExtremeXOS BGP allows users to add static networks in *BGP*, which will be redistributed (advertised) into the BGP domain if there is a corresponding active route in the IP routing table. Users can associate a policy with the static BGP network to change or to set the route attributes before the route is advertised to the BGP neighbors.

## Graceful BGP Restart

The graceful *BGP* restart feature enables the switch BGP process to restart without disrupting traffic forwarding. This feature also enables the BGP process to support the graceful BGP restart of a peer router. Without graceful restart, BGP routes within the restarting router are flushed, and BGP routes through the restarting router are flushed from BGP peers. A non-graceful restart interrupts traffic for the time it takes for BGP to restart and re-establish routes, and expends additional resources during the reconvergence.

### *Graceful Restart in the Restarting Switch*

During session startup, *BGP* peers indicate whether they have the graceful restart capability. When BGP restarts, the restarting router indicates during the new session startup that it is restarting and provides a restart-time value, which defines how long the receiving router will wait for the restart to complete.

In the restarting router, a graceful BGP restart preserves the BGP routes in the routing table, which allows the router to continue to use those routes until one of the following events occurs:

- The BGP restart is complete.
- The restart update delay timer expires.

During the restart, all pre-existing routes are marked stale. BGP receives new routes during the restart, but the Routing Information Base (RIB) is not updated and advertised until the restart is complete. An End of RIB message is sent when the restart is complete if the graceful restart feature is enabled.

An update-delay timer value determines how long the restarting switch will wait before updating the stale routes. If this timer expires before the restarting switch receives updates from the receiving routers, all stale routes are deleted. If the receiving routers provide updates before this timer expires, the time-stamps for any matching entries in the local RIB are preserved.

After the new BGP session is established, the new session uses the capabilities established with that session, which includes any updates to the graceful restart capability or timers.

### *Graceful Restart on the Receiving Switch*

During session startup, *BGP* peers indicate whether they have the graceful restart capability. When BGP restarts, the receiving router retains routes received from the restarting router and marks those routes

stale. The receiving router continues to advertise the restarting router as if it was fully functional until one of the following events occurs:

- The restarting router sends the EOR marker, indicating the end of a routing update and the end of the graceful restart.
- The restart timer defined by the restarting router expires.
- The stale-route-time timer defined on the receiving router expires.

If the receiving router receives RIB updates and the EOR marker before the timers expire, it updates the local RIB and deletes any stale entries. If either of the timers on the receiving router expires before the receiving switch receives the EOR marker from the restarting switch, all stale routes are deleted.

### *Planned and Unplanned Restarts*

Two types of graceful restarts are defined: planned and unplanned. A planned restart occurs when the software module for *BGP* is upgraded, or the router operator decides to restart the BGP control function. An unplanned restart occurs when there is some kind of system failure that causes a remote reboot or a crash of BGP, or when an MSM/MM failover occurs. You can decide to configure a router to enter graceful restart for only planned restarts, for only unplanned restarts, or for both. Also, you can decide to configure a router to be a receiver only, and not to do graceful restarts itself.

## Cease Subcodes

*BGP* uses the cease subcode in notification message to convey the reason for terminating the session. The cease subcodes currently supported are given in the following table.

**Table 149: Supported Cease Subcodes**

| Subcode | Description | Supported? |
|---|---|---|
| 1 | Maximum Number of Prefixes Reached | Yes |
| 2 | Administrative Shutdown | Yes |
| 3 | Peer De-configured | Yes |
| 4 | Administrative Reset | No |
| 5 | Connection Rejected | Yes |
| 6 | Other Configuration Change | No |
| 7 | Connection Collision Resolution | Yes |
| 8 | Out of Resources | No |

### *Maximum Number of Prefixes Reached*

This cease subcode is sent when the number of prefixes from a *BGP* neighbor exceeds the pre-configured limit. The notification message contains additional data to indicate the maximum prefix limit configured for the neighbor.

### *Administrative Shutdown*

This cease notification subcode is sent to a *BGP* neighbor in following two situations:

- BGP neighbor is disabled

- BGP protocol is globally disabled. All BGP neighbors that were in the established state send cease notifications with this subcode

*Peer De-configured*

This cease notification subcode is sent when *BGP* neighbor is deleted.

*Other Configuration Change*

This cease notification subcode is sent when the following configuration entities change:

- *BGP* neighbor is added to a peer group
- BGP neighbor is configured as a route-reflector client
- BGP neighbor is part of a peer group and the following configuration elements of the peer group are changed:
  - Password
  - Remote-as
  - Hold-time, keepalive-time
  - Source interface
  - Soft-in-reset

*Connection Collision Resolution*

This cease notification subcode is sent when there is a *BGP* connection collision.

## Fast External Fallover

*BGP* fast external fallover uses the BGP protocol to converge quickly in the event of a link failure that connects it to an EBGP neighbor.

When BGP fast external fallover is enabled, the directly-connected EBGP neighbor session is immediately reset when the connecting link goes down.

If BGP fast external fallover is disabled, BGP waits until the default hold timer expires (3 keepalives) to reset the neighboring session. In addition, BGP may tear down the session somewhat earlier than hold timer expiry if BGP detects that the TCP session and it's directly connected link is broken (BGP detects this while sending or receiving data from the TCP socket).

## Capability Negotiation

*BGP* supports the negotiation of the following capabilities between BGP peers:

- IPv4 unicast address family
- IPv4 multicast address family
- IPv6 unicast address family
- IPv6 multicast address family
- Route-refresh (code = 64 and Cisco-style code = 128)
- 4-byte-AS (code = 65)

BGP brings up peering with the minimal common capability for the both sides. For example, if a local router with both unicast and multicast capabilities peers with a remote router with unicast capability, the local router establishes the connection with unicast-only capability.

When there are no common capabilities, BGP sends an Unsupported Capability error and resets the connection. A manual intervention and configuration change might be required in order to establish a BGP peering session in this particular case.

Note

ExtremeXOS version 12.0 and earlier does not negotiate address families. By default, ExtremeXOS version 12.1 advertises MBGP options and is rejected by switches running previous versions, which can delay the establishment of a BGP session. We recommend that you disable the capability advertisement feature of MBGP while peering with switches running previous versions of ExtremeXOS for faster neighbor session establishment.

### *IPv4 Capability Negotiation*

For IPv4 peers, *BGP* supports the following capabilities by default:

- IPv4 unicast address family
- IPv4 multicast address family
- Route-refresh (code = 64 and Cisco-style code = 128)
- 4-byte-AS (code = 65)

By default, BGP sends those capabilities in its OPEN message. In addition, BGP supports graceful restart.

When BGP receives a notification 2/4 (Unsupported optional parameters) in response to an OPEN, it assumes that the peer does not support capability negotiation and MBGP and sends an OPEN message without any capability.

If the peer speaker sends no capabilities, but the local speaker is configured for the IPv4 unicast capability, the assumption is that the peer speaker is operating in legacy mode, and the session defaults to the exchange of IPv4 unicast NLRIs (a non MBGP session).

If the local speaker is configured explicitly with the IPv4 unicast family disabled, it cannot peer with legacy peers, and it will send the Optional Attribute Error whenever it receives an update packet. Because the IPv4 address family is enabled for Extreme Networks switches by default, it is recommended that you explicitly disable this address family when you desire non-standard behavior.

### *IPv6 Capability Negotiation*

For IPv6 peers, the route refresh capability is enabled by default, and no address family is enabled by default. You must enable a common set of capabilities on the local and neighbor peers before peering can be established. For IPv6 capability negotiation, IPv6 peering is set to idle if no common address families are negotiated. IPv6 *BGP* peering supports only the IPv6 unicast and IPv6 multicast address families; IPv4 address families are not supported.

## Route Refresh

Route Refresh helps minimize the memory footprint of *BGP* by not storing the original BGP route path attributes from a neighbor that advertises route refresh capability in an OPEN message. Whenever you execute the command `configure bgp neighbor [remoteAddr | all] {addressfamily [ipv4-unicast | ipv4-multicast]} soft-reset in`, BGP sends a route refresh message to its peer if that peer had advertised route refresh capability in its OPEN message. In response to the route refresh message, the neighbor sends its entire RIB-OUT database for the requested address family. This helps reapply the inbound neighbor policy, if there are any changes.

# Configuring BGP

## Clear Configuration Overview

The following procedure configures a basic *BGP* topology:

1. Configure the interfaces that will connect to BGP neighbors. For each interface, do the following:
   a. Create a *VLAN (Virtual LAN)*.
   b. Assign one or more ports to the VLAN.
   c. Configure a VLAN IP address.
   d. Enable IP forwarding on the VLAN.

   For more information on configuring VLANs, see VLANs on page 502
2. Configure the BGP router ID using the following command:

   `configure bgp routerid router identifier`
3. Configure the AS number to which the router should belong using the following command:

   `configure bgp AS-number number`
4. To add one or more IBGP neighbors, use the following command and specify the AS number to which the router belongs:

   `create bgp neighbor remoteaddr remote-AS-number as-number {multi-hop}`
5. To add one or more EBGP neighbors, use the following command and specify the AS number of the remote AS (which is different from the AS to which the router belongs):

   `create bgp neighbor remoteaddr remote-AS-number as-number {multi-hop}`
6. If you want to simultaneously configure BGP options for multiple neighbors, create and configure peer groups as described in Configuring BGP Peer Groups.
7. If the BGP network will support IPv4 traffic, you can skip this step. If the BGP network will support any other address family, you must enable support for that address family on BGP neighbors with either of the following commands:

   `enable bgp neighbor [all | remoteaddr] capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]`

   `enable bgp peer-group peer-group-name capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]`
8. To configure additional BGP neighbor options, see Configuring BGP Neighbors.
9. For instructions on configuring additional BGP features, see the list under Configuring BGP.
10. Enable BGP neighbors using the following command:

    `enable bgp neighbor [remoteaddr | all]`

11. Enable BGP using the following command:

    ```
    enable bgp
    ```

For instructions on displaying BGP information, see Displaying BGP Information.

## Configuring BGP Router Settings

*Configure the BGP Router ID*

- To configure the *BGP* router ID, use the following command:

  ```
  configure bgp routerid router identifier
  ```

*Configure the AS Number*

- To configure the AS number for the switch., use the following command:

  ```
  configure bgp AS-number number
  ```

*Configure the AS Number and Community Display Formats*

- To configure the AS number display format, use the following commands:

  ```
  configure bgp as-display-format [asdot | asplain]
  ```

  ```
  enable bgp community format AS-number : number
  ```

  ```
  disable bgp community format AS-number : number
  ```

*Configure the BGP Local Preference*

- To configure the *BGP* local preference, use the following command:

  ```
  configure bgp local-preference number
  ```

*Configure the BGP MED*

- To configure the *BGP* MED, use the following command:

  ```
  configure bgp med [none | bgp_med]
  ```

  ```
  enable bgp always-compare-med
  ```

  ```
  disable bgp always-compare-med
  ```

*Configure BGP ECMP*

- To enable or disable *BGP ECMP*, enter the following command:

  ```
  configure bgp maximum-paths max-paths
  ```

  The *max-paths* setting applies to BGP on the current VR. Specify more than 1 path to enable BGP ECMP and define the maximum number of paths for IBGP and EBGP multipaths. Specify 1 path to disable ECMP. To display BGP ECMP configuration information, use the `show bgp` command.

*Configure Graceful BGP Restart*

- To configure a router to perform graceful *BGP* restart, use the following command:

  `configure bgp restart [none | planned | unplanned | both | aware-only]`
- The address families participating in graceful restart are configured using the following command:

  `configure bgp restart [add | delete] address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]`
- There are three timers that can be configured with the following commands:

  `configure bgp restart restart-time seconds`

  `configure bgp restart stale-route-time seconds`

  `configure bgp restart update-delay seconds`
- You can use the following commands to verify the BGP graceful restart configuration:

  `show bgp`

  `show bgp neighbor remoteaddr {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} [accepted-routes | received-routes | rejected-routes | transmitted-routes] {detail} [all | as-path path-expression | community [no-advertise | no-export | no-export-subconfed | number community-number | autonomous-system-id | bgp-community] | network [any | netMaskLen | networkPrefixFilter] {exact}]`

*Configuring Fast External Fallover*

The fast external fallover module consists of two commands; one for enabling fallover (enable bgp fast-external-fallover) and one for disabling fallover (disable bgp fast-external-fallover). Fast external fallover is disabled by default.

These commands apply to all directly-connected external *BGP* neighbors.

## Configuring BGP Neighbors

*Create and Delete BGP Neighbors*

- To create or delete *BGP* neighbors, , use the following commands:

  `create bgp neighbor remoteaddr remote-AS-number as-number {multi-hop}`

  `create bgp neighbor remoteaddr peer-group peer-group-name {multi-hop}`

  `delete bgp neighbor [remoteaddr | all]`

*Configure a Description for a Neighbor*

- To configure a description for one or all *BGP* neighbors, use the following command:

  `configure bgp neighbor [all | remoteaddr] description {description}`

*Configure a Password for Neighbor Communications*

- To configure a password to use for communications with *BGP* neighbors, use the following command:

  ```
  configure bgp neighbor [all | remoteaddr] password [none | {encrypted} tcpPassword]
  ```

- To configure a password for the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name password [none | tcpPassword]
  ```

*Configure the Supported Address Families and Route Refresh*

- All *BGP* negotiated capabilities (except for the 4-Byte-AS capability) can be enabled and disabled using the following commands:

  ```
  enable bgp neighbor [all | remoteaddr] capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]

  disable bgp neighbor [all | remoteaddr] capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]
  ```

- To configure the capabilities for a peer group, use the following command:

  ```
  enable bgp peer-group peer-group-name capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]

  disable bgp peer-group peer-group-name capability [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4 | route-refresh]
  ```

*Configure Timers for BGP Neighbor Communications*

- Configure the timers that apply to communications with a neighbor.

  ```
  configure bgp neighbor [remoteaddr | all] timer keep-alive keepalive hold-time holdtime
  ```

- Configure the timers that apply to the neighbors in a peer group.

  ```
  configure bgp peer-group peer-group-name timer keep-alive seconds hold-time seconds
  ```

*Configure the Neighbor Shutdown Priority*

- To configure the neighbor shutdown priority, use the following command:

  ```
  configure bgp neighbor [all | remoteaddr] shutdown-priority number
  ```

*Configuring Route Acceptance*

**Assign a Weight Value to Routes Learned from a Neighbor**

- To configure the weight value that applies to routes learned from a neighbor, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] weight weight
  ```

- To configure the weight value that applies to routes learned from the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name weight number
  ```

#### Configure the Maximum Number of Prefixes

- To configure the maximum number of prefixes to accept from a *BGP* neighbor, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} maximum-prefix number {{threshold percent} {teardown {holddown-interval seconds}} {send-traps}
  ```

- To configure the maximum number of prefixes to accept from the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} maximum-prefix number {{threshold percent} {teardown {holddown-interval seconds}} {send-traps}
  ```

#### Configure Acceptance of Looped BGP Routes from Neighbors

- To enable and disable the acceptance of looped routes from one or all neighbors, use the following commands:

  ```
  configure bgp neighbor [all | remoteaddr] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} allowas-in {max-as-occurrence as-count}

  configure bgp neighbor [all | remoteaddr] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} dont-allowas-in
  ```

- To enable and disable the acceptance of looped routes from the neighbors in a peer group, use the following commands:

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} allowas-in {max-as-occurrence as-count}

  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} dont-allowas-in
  ```

### *Configuring Route Origination*

#### Configure the Source Interface Address

- To configure the source interface address to use for communications with a neighbor, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] source-interface [any | ipaddress ipAddr]
  ```

- To configure the source interface address to use for communications with neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name source-interface [any | ipaddress ipAddr]
  ```

#### Enable and Disable Default Route Origination

- To enable or disable *BGP* default route origination and advertisement for BGP neighbors, use the following commands:

  ```
  enable bgp [{neighbor} remoteaddr | neighbor all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} originate-default {policy policy-name}

  disable bgp [{neighbor} remoteaddr | neighbor all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} originate-default
  ```

- To enable or disable BGP default route origination and advertisement for a peer group, use the following commands:

  ```
  enable bgp {peer-group} peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} originate-default {policy policy-name}

  disable bgp {peer-group} peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} originate-default
  ```

#### Configure Inactive Route Advertisement

This command applies to the specified address family for all neighbors.

- To enable or disable *BGP* inactive route advertising, use the following commands:

  ```
  enable bgp {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} advertise-inactive-route

  disable bgp {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} advertise-inactive-route
  ```

#### Configure the Originating Next Hop Address for Outgoing Updates

- To configure outgoing updates to the specified neighbors to specify the address of the *BGP* connection originating the update as the next hop address, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} [next-hop-self | no-next-hop-self]
  ```

- To make a configuration change for the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} [next-hop-self | no-next-hop-self]
  ```

#### Include or Exclude the Community Path Attribute

- To configure neighbor communications to include or exclude the community path attribute, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} [send-community | dont-send-community] {both | extended | standard}
  ```

- To make a configuration change for the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} [send-community | dont-send-community] {both | extended | standard}
  ```

#### Remove Private AS Numbers from Route Updates

- To configure private AS numbers to be removed from updates for neighbors or peer groups, use the following commands:

  ```
  enable bgp neighbor [remoteaddr | all] remove-private-AS-numbers

  enable bgp peer-group peer-group-name remove-private-AS-numbers
  ```

- To disable this feature, use the following commands:

  ```
  disable bgp neighbor [remoteaddr | all] remove-private-AS-numbers

  disable bgp peer-group peer-group-name remove-private-AS-numbers
  ```

*Configure a Route Map Filter*

- To configure a route map filter for one or all *BGP* neighbors, use the following command:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} route-policy [in | out] [none | policy]
  ```

- To configure a route map filter for the neighbors in a peer group, use the following command:

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} route-policy [in |out] [none | policy]
  ```

*Enable and Disable the Soft Input Reset Feature for a Neighbor*

- To enable or disable the soft input reset feature for a neighbor, use the following commands:

  ```
  enable bgp neighbor [all | remoteaddr] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} soft-in-reset

  disable bgp neighbor [all | remoteaddr] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} soft-in-reset
  ```

- Enable or disable the soft input reset feature for a peer group.

  ```
  enable bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} soft-in-reset

  disable bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} soft-in-reset
  ```

*Configure Route Flap Dampening*

You can supply the dampening configuration parameters directly through a command line interface (CLI) command, or use the command to associate a policy that contains the desired parameters.

- To enable route flap dampening for neighbors or a peer group, use the following commands:

  ```
  configure bgp neighbor [all | remoteaddr] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} dampening {{half-life half-life-minutes {reuse-limit reuse-limit-number suppress-limit suppress-limit-number max-suppress max-suppress-minutes} | policy-filter [policy-name | none]}
  ```

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} dampening {{half-life half-life-minutes {reuse-limit reuse-limit-number suppress-limit suppress-limit-number max-suppress max-suppress-minutes} | policy-filter [policy-name | none]}
  ```

- To disable route flap dampening for a *BGP* neighbor or peer group, use the following commands:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} no-dampening
  ```

  ```
  configure bgp peer-group peer-group-name {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} no-dampening
  ```

  **Note**

  When you disable dampening, all the configured dampening parameters are deleted.

## Configuring BGP Peer Groups

*Create or Delete a BGP Peer Group*

To create or delete peer groups, use the following commands:

- `create bgp peer-group peer-group-name`
- `delete bgp peer-group peer-group-name`

**Note**

No capabilities are enabled at time of peer-group creation. If an IPv4 peer is the first peer added to the peer-group, the IPv4-unicast and IPv4-multicast capabilities are enabled by default. If the first peer assigned to a peer-group is an IPv6 peer, no capabilities are enabled.

*Add Neighbors to a BGP Peer Group*

- To create a new neighbor and add it to a *BGP* peer group, use the following command:

  ```
  create bgp neighbor remoteaddr peer-group peer-group-name {multi-hop}
  ```

  The new neighbor is created as part of the peer group and inherits all of the existing parameters of the peer group. The peer group must have remote AS configured.

- To add an existing neighbor to a peer group, use the following command:

  `configure bgp neighbor [all | remoteaddr] peer-group [peer-group-name | none] {acquire-all}`

  If you do not specify the **acquire-all** option, only the mandatory parameters are inherited from the peer group. If you specify the **acquire-all** option, all of the parameters of the peer group are inherited. This command disables the neighbor before adding it to the peer group.

  To remove a neighbor from a peer group, use the **peer-group none** option.

  When you remove a neighbor from a peer group, the neighbor retains the parameter settings of the group. The parameter values are not reset to those the neighbor had before it inherited the peer group values.

*Configure a Remote AS Number for a Peer Group*

- To configure a remote AS number for a peer group, use the following command:

  `configure bgp peer-group peer-group-name remote-AS-number number`

## Create and Delete BGP Static Networks

- To create a static *BGP* network, use the following command:

  `configure bgp add network {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} ipaddress/masklength {network-policy policy}`

  Note

  This command adds the route to BGP only if the route is present in the routing table.

- To delete a static BGP network, use the following command:

  `configure bgp delete network {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} [all | ipaddress/mask length]`

## Import Routes from Other Protocols to BGP

Before importing routes from another protocol to *BGP*, you must first configure both protocols and then verify the independent operation of each. When you import routes from another protocol, the imported routes are limited to those for the respective address family and protocol. For example, you can import IPv4 unicast routes from *RIP* to BGP, but you cannot import IPv6 unicast routes from RIP to BGP. IPv6 routes can be imported only from IPv6 routing protocols; and IPv4 routes can be imported only from IPv4 routing protocols.

You can use route maps to associate BGP attributes including Community, NextHop, MED, Origin, and Local Preference with the routes. Route maps can also be used to filter out exported routes.

For instructions on importing routes from another protocol to BGP, refer to the chapter for the protocol from which you want to import routes.

- To configure an import policy to apply to imported routes, use the following command:

  `configure bgp import-policy [policy-name | none]`

  For Layer 3 VPNs, a local PE stores the routes received from a remote PE in the VPN-VRF RIB.
- To enable or disable export of these routes as IPv4 unicast routes to the CE, use the following command:

  `enable bgp export remote-vpn {export-policy} policy-name} {address-family [ipv4-unicast | vpnv4]}`

  `disable bgp export remote-vpn {address-family [ipv4-unicast | vpnv4]}`

## Export BGP Routes to other Protocols

Before exporting routes from *BGP* to another protocol, you must first configure both protocols and then verify the independent operation of each. When you export routes from BGP, the exported routes are limited to those for the respective address family and protocol. For example, you can export IPv4 unicast routes from BGP to *RIP*, but you cannot export IPv6 unicast routes from BGP to RIP. IPv6 routes can be exported only from BGP to IPv6 routing protocols; and IPv4 routes can be exported only to IPv4 routing protocols.

You can use route maps to associate BGP attributes including Community, NextHop, MED, Origin, and Local Preference with the routes. Route maps can also be used to filter out exported routes.

- To enable or disable the export of routes into BGP from other routing sources like ospf and rip, use the following commands:

  `enable bgp export [blackhole | direct | isis | isis-level-1 | isis-level-1-external | isis-level-2 | isis-level-2-external | ospf | ospf-extern1 | ospf-extern2 | ospf-inter | ospf-intra | ospfv3 | ospfv3-extern1 | ospfv3-extern2 | ospfv3-inter | ospfv3-intra | rip | ripng | static {address-family [{ipv4-unicast |ipv4-multicast | ipv6-unicast | ipv6-multicast]} {export-policy policy-name}`

  `disable bgp export [blackhole | direct | isis | isis-level-1 | isis-level-1-external | isis-level-2 | isis-level-2-external | ospf | ospf-extern1 | ospf-extern2 | ospf-inter | ospf-intra | ospfv3 | ospfv3-extern1 | ospfv3-extern2 | ospfv3-inter | ospfv3-intra | rip | ripng | static {address-family [{ipv4-unicast |ipv4-multicast | ipv6-unicast | ipv6-multicast]}`

  Note

  When exporting BGP routes, static routes, configured with the `configure bgp add network` command, take precedence over BGP discovered routes.

  For Layer 3 VPNs, a local PE advertises the local customer routes to the remote PE by exporting the BGP routes in the VPN-VRF associated with that customer as VPNv4 routes.
- To enable or disable the export to the remote PE of BGP routes from the CE or static/direct routes in the CE VPN-VRF, use the following commands:

  `enable bgp export [bgp | direct | static] {export-policy} policy-name} {address-family [ipv4-unicast | vpnv4]}`

```
disable bgp export [bgp | direct | static] {address-family [ipv4-
unicast | vpnv4]}
```

## Configure Route Aggregation

Route aggregation is the process of combining the characteristics of several routes so that they are advertised as a single route. Aggregation reduces the amount of information that a *BGP* speaker must store and exchange with other BGP speakers. Reducing the information that is stored and exchanged also reduces the size of the routing table.

- To enable or disable BGP route aggregation, use the following commands:

  ```
  enable bgp aggregation

  disable bgp aggregation
  ```

- To create or remove aggregate routes, use the following commands:

  ```
  configure bgp add aggregate-address {address-family [ipv4-unicast |
  ipv4-multicast | ipv6-unicast | ipv6-multicast]} ipaddress/masklength
  {as-match | as-set} {summary-only} {advertise-policy policy}
  {attribute-policy policy}

  configure bgp delete aggregate-address {address-family [ipv4-unicast |
  ipv4-multicast | ipv6-unicast | ipv6-multicast]} [ip address/
  masklength | all]
  ```

## Configure Route Reflectors

The configuration for a router reflector topology takes place on the router or routers that serve as route reflectors. Route reflector clients must be configured for *BGP* participation, but no special configuration is required on route reflector clients to support the route reflector.

- To configure a route reflector to treat neighbors or peer group neighbors as route reflector clients, use the following commands:

  ```
  configure bgp neighbor [remoteaddr | all] [route-reflector-client |
  no-route-reflector-client]

  configure bgp peer-group peer-group-name [route-reflector-client | no-
  route-reflector-client]
  ```

- If multiple route reflectors are used in a cluster, you must configure the route reflector clients on each route reflector, and you must configure each route reflector with a common cluster ID using the following command:

  ```
  configure bgp cluster-id cluster-id
  ```

## Configure a Route Confederation

- To configure every router in the confederation with the same confederation ID, use the following command:

  ```
  configure bgp confederation-id number
  ```

- For each EBGP confederation peer, use the following command to configure the remote AS number as a confederation sub-AS-number:

  ```
  configure bgp add confederation-peer sub-AS-number number
  ```

- To remove the sub-AS-number configuration for an EBGP peer, use the following command:

  ```
  configure bgp delete confederation-peer sub-AS-number number
  ```

# Managing BGP

## Enable and Disable BGP Neighbors

- To enable or disable a *BGP* neighbor, use the following commands:

  ```
  enable bgp neighbor [remoteaddr | all]

  disable bgp neighbor [remoteaddr | all]
  ```

## Enable and Disable a Peer Group

- To enable or disable a peer group, use the following commands:

  ```
  enable bgp peer-group peer-group-name

  disable bgp peer-group peer-group-name
  ```

## Enable and Disable BGP

- To enable or disable *BGP* on the switch, use the following commands:

  ```
  enable bgp

  disable bgp
  ```

## Refresh BGP Routes

- To refresh the routes for a BGP neighbor, use the following commands:

  ```
  configure bgp neighbor [remoteaddr | all] {address-family [ipv4-
  unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]}
  soft-reset {in | out}

  configure bgp peer-group peer-group-name {address-family [ipv4-unicast
  | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} soft-reset
  {in | out}
  ```

## Reapply a Policy

- To reapply the route policy associated with the network command, aggregation, import, and redistribution, use the following command:

  ```
  configure bgp soft-reconfiguration
  ```

### Clear BGP Flap, Session, or Route Statistics

- To clear the *BGP* flap statistics, use the following command:

  ```
  clear bgp {neighbor} remoteaddr {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} flap-statistics [all | rd rd_value | as-path path expression | community [no-advertise | no-export | no-export-subconfed | number community_num | AS_Num:Num] | network [any / netMaskLen | networkPrefixFilter] {exact}]
  ```

  When clearing CE to PE peer sessions, select an IPv4 address family. When clearing PE to PE sessions, select the VPNv4 address family.

### Clear BGP Neighbor Counters

- To clear the counters for a *BGP* neighbor, use the following command:

  ```
  clear bgp neighbor [remoteaddr | all] counters
  ```

## Displaying BGP Information

### Display BGP Router Configuration and Route Statistics

- To display *BGP* router configuration and route statistics, use the following command:

  ```
  show bgp
  ```

### Display Peer Group Configuration Information

- To display peer group configuration information, use the following command:

  ```
  show bgp peer-group {detail | peer-group-name {detail}}
  ```

### Display BGP Route Information

- To display summary route information, enter the following command:

  ```
  show bgp routes {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast | vpnv4]} summary
  ```

- To display route information and statistics, enter the following command:

  ```
  show bgp routes {address-family [ipv4-unicast | ipv4-multicast | ipv6-unicast | ipv6-multicast]} {detail} [all | as-path path-expression | community [no-advertise | no-export | no-export-subconfed | number community-number | autonomous-system-id : bgp-community] | network [any / netMaskLen | networkPrefixFilter] {exact}]
  ```

  ```
  show bgp routes address-family vpnv4 {detail} [all | as-path <path-expression> | community [no-advertise | no-export | no-export-subconfed | number <community-number> | <autonomous-system-id> : <bgp-community>] | rd <rd> network [any / <netMaskLen> | <networkPrefixFilter>] {exact}]
  ```

- To display information about routes exchanged with a neighbor, enter the following command:

  ```
  show bgp neighbor <remoteaddr> {address-family [ipv4-unicast | ipv4-
  multicast | ipv6-unicast | ipv6-multicast]} [accepted-routes |
  received-routes | rejected-routes | transmitted-routes] {detail} [all
  | as-path <path-expression> | community [no-advertise | no-export |
  no-export-subconfed | number <community-number> | <autonomous-system-
  id> : <bgp-community>] | network [any / <netMaskLen> |
  <networkPrefixFilter>] {exact}]
  ```

  ```
  show bgp neighbor <remoteaddr> address-family vpnv4 [accepted-routes |
  received-routes | rejected-routes | transmitted-routes] {detail} [all
  | as-path <path-expression> | community [no-advertise | no-export |
  no-export-subconfed | number <community-number> | <autonomous-system-
  id> : <bgp-community>] | rd <rd> network [any / <netMaskLen> |
  <networkPrefixFilter>] {exact}]
  ```

- To display information about suppressed or dampened routes, enter the following command:

  ```
  show bgp neighbor <remoteaddr> {address-family [ipv4-unicast | ipv4-
  multicast | ipv6-unicast | ipv6-multicast]} [accepted-routes |
  received-routes | rejected-routes | transmitted-routes] {detail} [all
  | as-path <path-expression> | community [no-advertise | no-export |
  no-export-subconfed | number <community-number> | <autonomous-system-
  id> : <bgp-community>] | network [any / <netMaskLen> |
  <networkPrefixFilter>] {exact}]
  ```

  ```
  show bgp {neighbor} <remoteaddr> address-family vpnv4 [flap-statistics
  | suppressed-routes] {detail} [all | as-path <path-expression> |
  community [no-advertise | no-export | no-export-subconfed | number
  <community-number> | <autonomous-system-id> : <bgp-community>] | rd
  <rd> network [any / <netMaskLen> | <networkPrefixFilter>] {exact}]
  ```

## Display Layer 3 VPN Peer Session Information

Layer 3 VPN *BGP* session information can be configured with the same commands used for other BGP sessions. When displaying information for CE to PE peer sessions, select an IPv4 address family. When displaying information for PE to PE sessions, select the VPNv4 address family.

- Use the following commands to display Layer 3 VPN session information:

  ```
  show bgp routes vpnv4 [all | as-path | community | detail | rd |
  summary]
  ```

## Display BGP Memory Usage

- To display *BGP* memory usage information, enter the following command:

  ```
  show bgp memory {detail | memoryType}
  ```

# Configuration Examples

## BGP IPv6 Example

The following figure shows the network topology for this example.

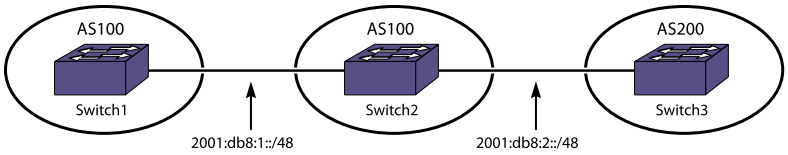



**Figure 233: BGP IPv6 Example**

*Switch1 Configuration*

```
create vlan v1
configure v1 ipaddress 2001:db8:1::1/48
configure v1 add ports 1
create vlan net
configure net add ports 3
configure net ipaddress 2001:db8:22::1/48
enable ipforwarding ipv6
configure bgp AS-number 100
configure bgp routerid 1.1.1.1
create bgp neighbor 2001:db8:2131::2 remote-AS-number 100
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
configure bgp add network address-family ipv6-unicast 2001:db8:22::/48
configure iproute add 2001:db8:2131::/48 2001:db8:1::2
configure ospfv3 routerid 1.1.1.1
configure ospfv3 add vlan v1 area 0.0.0.0
enable ospfv3
```

*Switch2 Configuration*

```
create vlan v1
create vlan v2
create vlan lpback
enable loopback lpback
configure v1 ipaddress 2001:db8:1::2/48
configure v2 ipaddress 2001:db8:2::1/48
configure lpback  ipaddress 2001:db8:2131::2/48
enable ipforwarding ipv6
configure v1 add ports 1:9
configure v2 add ports 1:11
configure bgp AS-number 100
configure bgp routerid 1.1.1.2
create bgp neighbor 2001:db8:1::1 remote-AS-number 100
create bgp neighbor 2001:db8:2::2 remote-AS-number 200
configure bgp neighbor 2001:db8:1::1 source-interface ipaddress 2001:db8:2131::2
enable bgp neighbor all capability ipv6-unicast
enable bgp adj-rib-out
enable bgp neighbor all
enable bgp
configure ospfv3 routerid 1.1.1.2
configure ospfv3 add vlan all area 0.0.0.0
enable ospfv3
```

*Switch3 Configuration*

```
create vlan v1
create vlan net
configure v1 add ports 2:11
configure net add ports 1:23
configure v1 ipaddress 2001:db8:2::2/48
configure net ipaddress 2001:db8:55::1/48
enable ipforwarding ipv6
configure bgp AS-number 200
configure bgp routerid 2.1.1.2
create bgp neighbor 2001:db8:2::1 remote-AS-number 100
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
enable bgp export static address-family ipv6-unicast
enable bgp export direct  address-family ipv6-unicast
configure iproute add 2001:db8:66::/48 2001:db8:55::100
```

*Configuration Displays (show Commands) for Switch2*

```
* Switch2.47 # show bgp neighbor
Peer                AS    Weight State        InMsgs OutMsgs(InQ)  Up/Down
------------------------------------------------------------------------------
Ie-- 2001:db8:1::1  100   1      ESTABLISHED  17     19    (0    ) 0:0:11:08
Ee-- 2001:db8:2::2  200   1      ESTABLISHED  28     24    (0    ) 0:0:01:53
Flags: (d) disabled, (e) enabled, (E) external peer, (I) internal peer
(m) EBGP multihop, (r) route reflector client
BGP Peer Statistics
Total Peers       : 2
EBGP Peers        : 1                IBGP Peers          : 1
RR Client         : 0                EBGP Multihop       : 0
Enabled           : 2                Disabled            : 0


* Switch2.50 # show bgp routes address-family ipv6-unicast all
Routes:
Destination             LPref Weight MED  Peer           Next-Hop       AS-Path
------------------------------------------------------------------------------
* ?  2001:db8:2::/48    100   1      0    2001:db8:2::2  2001:db8:2::2  200
*>?  2001:db8:55::/48   100   1      0    2001:db8:2::2  2001:db8:2::2  200
*>?  2001:db8:66::/48   100   1      0    2001:db8:2::2  2001:db8:2::2  200
*>i  2001:db8:22::/48   100   1      0    2001:db8:1::1  2001:db8:1::1
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 4
Feasible Routes   : 4
Active Routes     : 3
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 1
Routes from Ext Peer: 3


Switch2.57 # show bgp neighbor 2001:db8:1::1 address-family ipv6-unicast received-routes
all
Routes:
Destination            LPref Weight MED  Peer           Next-Hop      AS-Path
------------------------------------------------------------------------------
```

```
*>i  2001:db8:22::/48  100   1      0    2001:db8:1::1  2001:db8:1::1
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 1
Feasible Routes   : 1
Active Routes     : 1
Rejected Routes   : 0
Unfeasible Routes : 0

Switch2.59 # show bgp routes address-family ipv6-unicast detail all
Routes:
Route: 2001:db8:2::/48, Peer 2001:db8:2::2, BEST
Origin Incomplete, Next-Hop 2001:db8:2::2, LPref 100, MED 0
Weight 1,
As-PATH: 200

Route: 2001:db8:55::/48, Peer 2001:db8:2::2, BEST, Active
Origin Incomplete, Next-Hop 2001:db8:2::2, LPref 100, MED 0
Weight 1,
As-PATH: 200

Route: 2001:db8:66::/48, Peer 2001:db8:2::2, BEST, Active
Origin Incomplete, Next-Hop 2001:db8:2::2, LPref 100, MED 0
Weight 1,
As-PATH: 200

Route: 2001:db8:22::/48, Peer 2001:db8:1::1, BEST, Active
Origin IGP, Next-Hop 2001:db8:1::1, LPref 100, MED 0
Weight 1,
As-PATH:
BGP Route Statistics
Total Rxed Routes : 4
Feasible Routes   : 4
Active Routes     : 3
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 1
Routes from Ext Peer: 3

Switch2.72 # sh bgp neighbor 2001:db8:2::2 address-family ipv6-unicast transmitted-routes
all
Advertised Routes:
Destination           LPref Weight MED  Peer       Next-Hop         AS-Path
-------------------------------------------------------------------------------
>i  2001:db8:22::/48        0      0               2001:db8:2::1    100
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Advertised Routes : 1
```

## Graceful BGP Restart Configuration Example for IPv4

In the following IPv4 configuration example, EXOS-1 is the restarting BGP router, and EXOS-2 is the receiving *BGP* router.

To configure router EXOS-1, use the following commands:

```
create vlan bgp-restart
configure vlan bgp-restart add port 2:2
configure vlan bgp-restart ipaddress 10.0.0.1/24
enable ipforwarding

configure bgp as-number 100
configure bgp route-id 10.0.0.1
configure bgp restart both
create bgp neighbor 10.0.0.2 remote-as 64500
enable bgp neighbor all
enable bgp
```

To configure router EXOS-2, use the following commands:

```
create vlan bgp-restart
configure vlan bgp-restart add port 2:5
configure vlan bgp-restart ipaddress 10.0.0.2/24
enable ipforwarding

configure bgp as-number 64500
configure bgp route-id 10.0.0.2
configure bgp restart aware-only
create bgp neighbor 10.0.0.1 remote-as 100
enable bgp neighbor all
enable bgp
```

## Graceful BGP Restart Configuration Example for IPv6

In the following IPv6 configuration example, EXOS-1 is the restarting *BGP* router, and EXOS-2 is the receiving BGP router.

To configure router EXOS-1, use the following commands:

```
create vlan bgp-restart
configure vlan bgp-restart add port 2:2
configure vlan bgp-restart ipaddress 2001:db8:1::1/48
enable ipforwarding ipv6

configure bgp as-number 100
configure bgp routerid 10.0.0.1
configure bgp restart both
conf bgp restart add address-family ipv6-unicast
create bgp neighbor 2001:db8:1::2 remote-as 200
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

To configure router EXOS-2, use the following commands:

```
create vlan bgp-restart
configure vlan bgp-restart add port 2:5
configure vlan bgp-restart ipaddress 2001:db8:1::2/48
enable ipforwarding ipv6
configure bgp as-number 200
configure bgp routerid 10.0.0.2
configure bgp restart aware-only
create bgp neighbor 2001:db8:1::1 remote-as 100
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

## Route Reflector Example for IPv4

Figure 231 on page 1396 shows the network topology for this example. Router 1.1.1.1 in this example is a regular *BGP* peer.

To configure router 1.1.1.1 to connect to the route reflector as a regular BGP peer, use the following commands:

```
create vlan to_rr
configure vlan to_rr add port 1:1
configure vlan to_rr ipaddress 10.0.0.1/24
enable ipforwarding vlan to_rr

configure bgp router 1.1.1.1
configure bgp as-number 100
create bgp neighbor 10.0.0.2 remote-as 100
enable bgp
enable bgp neighbor all
```

To configure router 2.2.2.2, the route reflector, use the following commands:

```
create vlan to_nc
configure vlan to_nc add port 1:1
configure vlan to_nc ipaddress 10.0.0.1/24
enable ipforwarding vlan to_nc
create vlan to_c1
configure vlan to_c1 add port 1:2
configure vlan to_c1 ipaddress 10.0.2.2/24
enable ipforwarding vlan to_c1

create vlan to_c2
configure vlan to_c2 add port 1:2
configure vlan to_c2 ipaddress 10.0.3.2/24
enable ipforwarding vlan to_c2

configure bgp router 2.2.2.2
configure bgp as-number 100
create bgp neighbor 10.0.0.1 remote-as 100
create bgp neighbor 10.0.2.1 remote-as 100
create bgp neighbor 10.0.3.1 remote-as 100
configure bgp neighbor 10.0.2.1 route-reflector-client
configure bgp neighbor 10.0.3.1 route-reflector-client
enable bgp neighbor all
enable bgp
```

To configure router reflector client 3.3.3.3, use the following commands:

```
create vlan to_rr
configure vlan to_rr add port 1:1
configure vlan to_rr ipaddress 10.0.2.1/24
enable ipforwarding vlan to_rr

configure bgp router 3.3.3.3
configure bgp as-number 100
create bgp neighbor 10.0.2.2 remote-as 100
enable bgp neighbor all
enable bgp
```

To configure route reflector client 4.4.4.4, use the following commands:

```
create vlan to_rr
configure vlan to_rr add port 1:1
configure vlan to_rr ipaddress 10.0.3.1/24
```

```
enable ipforwarding vlan to_rr

configure bgp router 4.4.4.4
configure bgp as-number 100
create bgp neighbor 10.0.3.2 remote-as 100
enable bgp neighbor all
enable bgp
```

## Route Reflector Example for IPv6

The following figure shows the network topology for this example.

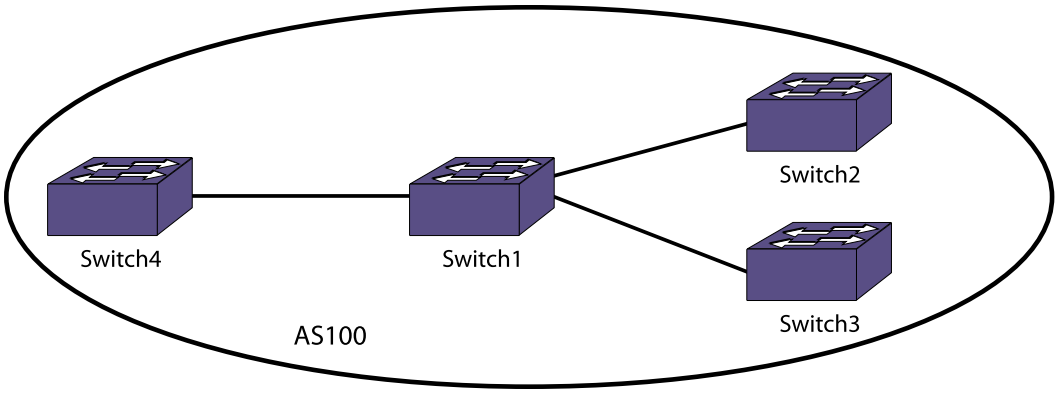



**Figure 234: Route Reflector Example**

Switch1 is the route reflector, and Switch2 and Switch3 are route reflector clients. Switch4 is a regular neighbor (non-route-reflector client).

### *Route Reflector (Switch1) Configuration*

To configure the route reflector, enter the following commands

```
create vlan "v3"
create vlan "v1"
create vlan "v2"
configure vlan v3 add ports 1:16
configure vlan v1 add ports 1:9
configure vlan v2 add ports 1:1
configure v1 ipaddress 2001:db8:1::1/48
configure v2 ipaddress 2001:db8:3::1/48
configure v3 ipaddress 2001:db8:5::1/484
enable ipforwarding ipv6
configure bgp AS-number 100
configure bgp routerid 1.1.1.1
create bgp neighbor 2001:db8:1::2 remote-AS-number 100
configure bgp neighbor 2001:db8:1::2 route-reflector-client
create bgp neighbor 2001:db8:3::2 remote-AS-number 100
configure bgp neighbor 2001:db8:3::2 route-reflector-client
create bgp neighbor 2001:db8:5::2 remote-AS-number 100
enable bgp neighbor 2001:db8:1::2 capability ipv6-unicast
enable bgp neighbor 2001:db8:3::2 capability ipv6-unicast
enable bgp neighbor 2001:db8:5::2 capability ipv6-unicast
enable bgp neighbor all
enable bgp
configure ospfv3 routerid 1.1.1.1
configure ospfv3 add vlan all area 0.0.0.0
enable ospfv3
```

*Switch2 Route Reflector Client Configuration*

To configure the Switch 2 route reflector client, enter the following commands:

```
create vlan "v1"
configure vlan v1 add ports 1
configure v1 ipaddress 2001:db8:1::2/48
enable ipforwarding ipv6 vlan v1
configure bgp AS-number 100
configure bgp routerid 1.1.1.2
create bgp neighbor 2001:db8:1::1 remote-AS-number 100
enable bgp neighbor 2001:db8:1::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:1::1
enable bgp
configure ospfv3 routerid 1.1.1.2
configure ospfv3 add vlan v1 area 0.0.0.0
enable ospfv3
```

*Switch3 Route Reflector Client Configuration*

To configure the Switch 3 route reflector client,e nter the following commands:

```
create vlan "v1"
configure vlan v1 add ports 1
configure v1 ipaddress 2001:db8:3::2/48
enable ipforwarding ipv6
configure bgp AS-number 100
configure bgp routerid 2.1.1.2
create bgp neighbor 2001:db8:3::1 remote-AS-number 100
enable bgp neighbor 2001:db8:3::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:3::1
enable bgp
configure ospfv3 routerid 2.1.1.2
configure ospfv3 add vlan v1 area 0.0.0.0
enable ospfv3
```

*Switch4 Configuration*

Switch4 is not a route reflector client. To configure this switch, enter the following commands:

```
create vlan "net"
enable loopback-mode vlan net
create vlan "v1"
configure vlan v1 add ports 9
configure v1 ipaddress 2001:db8:5::2/48
configure net ipaddress 2001:db8:2555::1/48
enable ipforwarding ipv6
configure bgp AS-number 100
configure bgp routerid 5.1.1.2
configure bgp add network address-family ipv6-unicast 2001:db8:2555::/48
create bgp neighbor 2001:db8:5::1 remote-AS-number 100
enable bgp neighbor 2001:db8:5::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:5::1
enable bgp
configure ospfv3 routerid 5.1.1.2
configure ospfv3 add vlan v1 area 0.0.0.0
enable ospfv3
```

*Configuration Display for Switch1*

The following display shows that the route reflector has two route reflector client peers (r flag), and one regular peer:

```
* Switch1.30 # show bgp neighbor
Peer            AS    Weight State        InMsgs OutMsgs(InQ)  Up/Down
------------------------------------------------------------------------------
Ier- 2001:db8:1::2   100   1      ESTABLISHED  6      8     (0    ) 0:0:04:03
Ier- 2001:db8:3::2   100   1      ESTABLISHED  5      7     (0    ) 0:0:02:51
Ie-- 2001:db8:5::2   100   1      ESTABLISHED  6      7     (0    ) 0:0:01:41
Flags: (d) disabled, (e) enabled, (E) external peer, (I) internal peer
(m) EBGP multihop, (r) route reflector client
BGP Peer Statistics
Total Peers       : 3
EBGP Peers        : 0                IBGP Peers          : 3
RR Client         : 2                EBGP Multihop       : 0
Enabled           : 3                Disabled            : 0
```

*Configuration Display for Switch2*

The following command displays the *BGP* routes for route reflector client Switch2:

```
* Switch2.10 # show bgp routes address-family ipv6-unicast detail all
Routes:
Route: 2001:db8:2555::/48, Peer 2001:db8:1::1, BEST, Active
Origin IGP, Next-Hop 2001:db8:5::2, LPref 100, MED 0
Weight 1, RR Orig ID 5.1.1.2
As-PATH:
RR Cluster IDs: 1.1.1.1
BGP Route Statistics
Total Rxed Routes : 1
Feasible Routes   : 1
Active Routes     : 1
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 1
Routes from Ext Peer: 0
```

## Route Confederation Example for IPv4

The following figure shows the network topology for this example.

To configure router A, use the following commands:

```
create vlan ab
configure vlan ab add port 1
configure vlan ab ipaddress 10.1.1.6/30
enable ipforwarding vlan ab
configure ospf add vlan ab area 0.0.0.0

create vlan ac
configure vlan ac add port 2
configure vlan ac ipaddress 10.1.1.17/30
enable ipforwarding vlan ac
configure ospf add vlan ac area 0.0.0.0
enable ospf

configure bgp as-number 64505
configure bgp routerid 10.1.1.17
```

```
configure bgp confederation-id 64500
enable bgp

create bgp neighbor 10.1.1.5 remote-AS-number 64505
create bgp neighbor 10.1.1.18 remote-AS-number 64505
enable bgp neighbor all
```

To configure router B, use the following commands:

```
create vlan ba
configure vlan ba add port 1
configure vlan ba ipaddress 10.1.1.5/30
enable ipforwarding vlan ba
configure ospf add vlan ba area 0.0.0.0

create vlan bc
configure vlan bc add port 2
configure vlan bc ipaddress 10.1.1.22/30
enable ipforwarding vlan bc
configure ospf add vlan bc area 0.0.0.0

create vlan bd
configure vlan bd add port 3
configure vlan bd ipaddress 10.1.1.9/30
enable ipforwarding vlan bd
configure ospf add vlan bd area 0.0.0.0
enable ospf

configure bgp as-number 64505
configure bgp routerid 10.1.1.22
configure bgp confederation-id 64500
enable bgp

create bgp neighbor 10.1.1.6 remote-AS-number 64505
create bgp neighbor 10.1.1.21 remote-AS-number 64505
create bgp neighbor 10.1.1.10 remote-AS-number 64510
configure bgp add confederation-peer sub-AS-number 64510
enable bgp neighbor all
```

To configure router C, use the following commands:

```
create vlan ca
configure vlan ca add port 1
configure vlan ca ipaddress 10.1.1.18/30
enable ipforwarding vlan ca
configure ospf add vlan ca area 0.0.0.0

create vlan cb
configure vlan cb add port 2
configure vlan cb ipaddress 10.1.1.21/30
enable ipforwarding vlan cb
configure ospf add vlan cb area 0.0.0.0
enable ospf

configure bgp as-number 64505
configure bgp routerid 10.1.1.21
configure bgp confederation-id 64500
enable bgp

create bgp neighbor 10.1.1.22 remote-AS-number 64505
create bgp neighbor 10.1.1.17 remote-AS-number 64505
enable bgp neighbor all
```

To configure router D, use the following commands:

```
create vlan db
configure vlan db add port 1
configure vlan db ipaddress 10.1.1.10/30
enable ipforwarding vlan db
configure ospf add vlan db area 0.0.0.0

create vlan de
configure vlan de add port 2
configure vlan de ipaddress 10.1.1.14/30
enable ipforwarding vlan de
configure ospf add vlan de area 0.0.0.0
enable ospf

configure bgp as-number 64510
configure bgp routerid 10.1.1.14
configure bgp confederation-id 64500
enable bgp

create bgp neighbor 10.1.1.9 remote-AS-number 64505
create bgp neighbor 10.1.1.13 remote-AS-number 64510
configure bgp add confederation-peer sub-AS-number 64505
enable bgp neighbor all
```

To configure router E, use the following commands:

```
create vlan ed
configure vlan ed add port 1
configure vlan ed ipaddress 10.1.1.13/30
enable ipforwarding vlan ed
configure ospf add vlan ed area 0.0.0.0
enable ospf

configure bgp as-number 64510
configure bgp routerid 10.1.1.13
configure bgp confederation-id 64500
enable bgp

create bgp neighbor 10.1.1.14 remote-AS-number 64510
enable bgp neighbor 10.1.1.14
```

## Route Confederation Example for IPv6

Figure 232 on page 1397 shows the network topology for this example.

To configure router A, use the following commands:

```
create vlan ab
configure vlan ab add port 23
configure vlan ab ipaddress 2001:db8:1::6/48
enable ipforwarding ipv6 vlan ab
configure ospf add vlan ab area 0.0.0.0
create vlan ac
configure vlan ac add port 16
configure vlan ac ipaddress 2001:db8:3::17/48
enable ipforwarding ipv6 vlan ac
configure ospfv3 routerid 10.1.1.17
configure ospfv3 add vlan ac area 0.0.0.0
enable ospfv3
configure bgp as-number 65001
configure bgp routerid 10.1.1.17
configure bgp confederation-id 200
```

```
create bgp neighbor 2001:db8:1::5 remote-AS-number 65001
create bgp neighbor 2001:db8:3::18 remote-AS-number 65001
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

To configure router B, use the following commands:

```
configure default delete port all
create vlan ba
configure vlan ba add port 5:33
configure vlan ba ipaddress 2001:db8:1::5/48
enable ipforwarding ipv6 vlan ba
configure ospf add vlan ba area 0.0.0.0
create vlan bc
configure vlan bc add port 5:8
configure vlan bc ipaddress 2001:db8:2::22/48
enable ipforwarding ipv6 vlan bc
configure ospf add vlan bc area 0.0.0.0
create vlan bd
configure vlan bd add port 5:4
configure vlan bd ipaddress 2001:db8:4::9/48
enable ipforwarding ipv6 vlan bd
configure ospfv3 add vlan bd area 0.0.0.0
configure ospfv3 routerid 10.1.1.22
enable ospfv3
configure bgp as-number 65001
configure bgp routerid 10.1.1.22
configure bgp confederation-id 200
create bgp neighbor 2001:db8:1::6 remote-AS-number 65001
create bgp neighbor 2001:db8:2::21 remote-AS-number 65001
create bgp neighbor 2001:db8:4::10 remote-AS-number 65002
configure bgp add confederation-peer sub-AS-number 65002
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

To configure router C, use the following commands:

```
configure default delete port all
create vlan ca
configure vlan ca add port 2:16
configure vlan ca ipaddress 2001:db8:3::18/48
enable ipforwarding ipv6 vlan ca
configure ospfv3 add vlan ca area 0.0.0.0
create vlan cb
configure vlan cb add port 2:5
configure vlan cb ipaddress 2001:db8:2::21/48
enable ipforwarding ipv6 vlan cb
configure ospfv3 add vlan cb area 0.0.0.0
configure ospfv3 routerid 10.1.1.21
enable ospfv3
configure bgp as-number 65001
configure bgp routerid 10.1.1.21
configure bgp confederation-id 200
create bgp neighbor 2001:db8:2::22 remote-AS-number 65001
create bgp neighbor 2001:db8:3::17 remote-AS-number 65001
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

To configure router D, use the following commands:

```
configure default delete port all
create vlan db
configure vlan db add port 13
configure vlan db ipaddress 2001:db8:4::10/48
enable ipforwarding ipv6 vlan db
configure ospf add vlan db area 0.0.0.0
create vlan de
configure vlan de add port 5
configure vlan de ipaddress 2001:db8:5::14/48
enable ipforwarding ipv6 vlan de
configure ospfv3 add vlan de area 0.0.0.0
configure ospfv3 routerid 10.1.1.14
enable ospfv3
configure bgp as-number 65002
configure bgp routerid 10.1.1.14
configure bgp confederation-id 200
create bgp neighbor 2001:db8:4::9 remote-AS-number 65001
create bgp neighbor 2001:db8:5::13 remote-AS-number 65002
configure bgp add confederation-peer sub-AS-number 65001
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

To configure router E, use the following commands:

```
configure default delete port all
create vlan ed
configure vlan ed add port 5
configure vlan ed ipaddress 2001:db8:5::13/48
enable ipforwarding ipv6 vlan ed
configure ospfv3 add vlan ed area 0.0.0.0
configure ospfv3 routerid 10.1.1.13
enable ospfv3
configure bgp as-number 65002
configure bgp routerid 10.1.1.13
configure bgp confederation-id 200
create bgp neighbor 2001:db8:5::14 remote-AS-number 65002
enable bgp neighbor all capability ipv6-unicast
enable bgp neighbor all
enable bgp
```

## Default Route Origination Example for IPv4

The following example configures the originate default route feature for *BGP* neighbor 10.203.134.5 using policy def_originate.pol.

```
def_originate.pol
entry prefix_matching {
if match any {
nlri 192.168.3.0/24;
} then {
as-path "64505";
permit;
}
}
enable bgp neighbor 10.203.134.5 originate-default policy def_originate
```

With this configuration, a default route is originated and sent to neighbor 10.203.134.5 only if there is a BGP route in the local RIB which matches the statement nlri 192.168.3.0/24. If a matching route exists,

the default route is sent to neighbor 10.203.134.5 with the 64505 as-path prepended. If this is an EBGP neighbor, then the local AS-Number is prepended after 64505.

If the route for the match statement nlri 192.168.3.0/24 goes away and there is no other matching route in the BGP RIB, the default route origination feature becomes inactive and BGP withdraws the 0.0.0.0/0 default route from neighbor 10.203.134.5. When a matching route becomes available again in the local BGP RIB, the default route origination feature becomes active again and the default route 0.0.0.0/0 is advertised to neighbor 10.203.134.5.

## Default Route Origination Example for IPv6

The following example configures the originate default route feature for *BGP* neighbor 2001:db8:1::2 using policy def_originate.pol.

```
def_originate.pol
entry prefix_matching {
if match any {
nlri 2001:db8:2::/48;
} then {
as-path "65001";
permit;
}
}
enable bgp neighbor 2001:db8:1::2 address-family ipv6-unicast originate-default policy
def_originate
```

With this configuration, a default route is originated and sent to neighbor 2001:db8:1::2 only if there is a BGP route in the local RIB which matches the statement nlri 2001:db8:2::/48. If a matching route exists, the default route is sent to neighbor 2001:db8:1::2 with the 65001 as-path prepended. If this is an EBGP neighbor, then the local AS-Number is prepended after 65001.

If the route for the match statement 2001:db8:2::/48 goes away and there is no other matching route in the BGP RIB, the default route origination feature becomes inactive and BGP withdraws the default route ::/0 from neighbor 2001:db8:1::2. When a matching route becomes available again in the local BGP RIB, the default route origination feature becomes active again and the default route ::/0 is advertised to neighbor 2001:db8:1::2.

## BGP Speaker Black Hole Example

Black hole routing is used to protect a service provider's internal network from distributed denial of service (DDoS) attacks. The strategy is to drop inbound DDoS attack traffic destined to a target network at the edge of the provider network as soon as the target is identified.

Since the attack traffic may enter the service provider network from any of its edge routers, it is not feasible to manually configure a static black hole route entry on each edge router. The problem is further complicated by the fact that the target network, and the need to block traffic to it, is dynamic. Also, the service provider may serve multiple customers and you don't want to drop traffic to a customer network until it is identified as a target for an ongoing attack.

Instead, a service provider can use *BGP* to distribute a black hole route (route entry for the target network with a black hole next-hop) from a single router to all its edge *BGP* speakers that will then drop

traffic destined to the victim's network right at the provider edge. The following figure shows an example topology to achieve this.

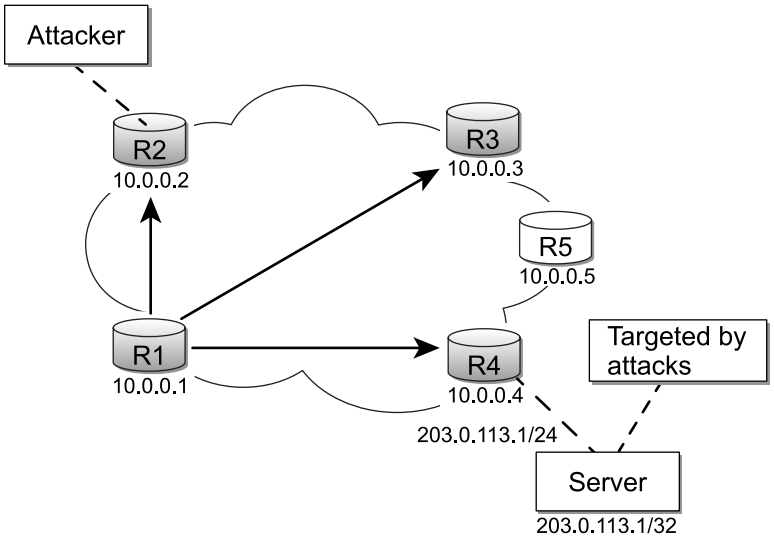



**Figure 235: Black Hole Routing Using BGP**

*Step 1*

Prior to the attack, select an address for the intended black hole next-hop. Configure the forwarding plane of each edge router so that packets forwarded to this next-hop are dropped:

1. Create a black hole *VLAN* with an IP address that is in the same subnet as the chosen black hole next-hop.
2. Add an active port to the black hole VLAN (uusually an unused port in the switch).
3. Create a static *FDB (forwarding database)* entry that maps a well-chosen, unused MAC address to the black hole VLAN and the active port added to that VLAN.
4. Create a static ARP entry that maps the black hole next-hop to the above MAC address.
5. Create an *ACL (Access Control List)* filter to deny packets that exit the blackhole VLAN.

   In the following example configuration, 192.168.2.0/24 is the subnet of the black hole VLAN, "BH_VLAN," and 192.168.2.66 is the chosen black hole next-hop. The active port 6:9 is added as the egress port for "BH_VLAN."

```
create vlan BH_VLAN
configure vlan BH_VLAN tag 666
enable loopback-mode vlan BH_VLAN
configure vlan BH_VLAN ipaddress 192.168.2.1 255.255.255.0
enable ipforwarding vlan BH_VLAN
disable igmp snooping vlan BH_VLAN
disable igmp vlan BH_VLAN
create fdb 00:02:03:04:05:06 vlan BH_VLAN port 6:9
configure iparp add 192.168.2.66 vr VR-Default 00:02:03:04:05:06
configure access-list BH_ACL vlan BH_VLAN egress
```

   When a packet arrives in the forwarding plane and looks up a route that has the above black hole next-hop as its next-hop, a subsequent ARP and FDB look-up occurs that forwards the packet to exit the switch using the above black hole VLAN, "BH_VLAN," and port "6:9." The packet is dropped due to the deny action in the egress ACL filter.

The following policy file discards any traffic that exits the black hole VLAN, “BH_VLAN.” Note that the match on “source-address 0.0.0.0/0” matches any egress packet ensuring that all packets exiting via the black hole VLAN are dropped:

```
edit policy BH_ACL
entry bh-acl {
if {
    source-address 0.0.0.0/0;
} then {
deny ;
}
}
```

*Step 2*

- Prior to the attack, configure inbound route-maps on all edge *BGP* speakers (R2 through R4 in the following figure).

  These inbound policies modify the next-hop of specifically marked BGP network layer reach-ability information (NLRIs) to point to the chosen black hole next-hop. We use BGP community or extended-community attributes to identify NLRIs that need to be black holed (ones whose next-hops have to be modified). The community values that are chosen should be reserved for this purpose within the provider network.

  In the following example, a community of 666:0 is chosen for identifying blackhole routes. The next-hop of BGP NLRIs with that community attribute is modified to use the blackhole next-hop.

```
R3.1 # edit policy BH_policy_NH
entry bh-nhset {
        if match any {
            community 666:0;
            nlri any/32 ;
        } then {
            next-hop 192.168.2.66 ;
            permit ;
        }
}
entry bh-default {
        if match any {
        } then {
            permit ;
        }
}
```

*Step 3*

Once the target network has been identified during a DDoS attack, apply an outbound policy or export policy to one router (in our example, R1) within the provider network so that the route to the target network is advertised to the other edge routers within the community 666:0.

The following example creates a static route on R1 to the target network 203.0.113.1/32 with a static export policy that applies to the community. When the attack targets change, you only need to create or delete static routes to the target networks. The policy exports them to the edge *BGP* speakers with the selected community attribute values attached.

```
R1.1 # edit policy BH_COMM_APPLY
entry bh-comm-apply {
        if match any  {
```

```
            nlri 203.0.113.0/24;
            nlri any/32;
        } then {
            community set “666:0”;
        }
}
R1.2 # configure iproute add 203.0.113.1/32 10.0.0.6
R1.3 # enable bgp export static export-policy BH_COMM_APPLY
```

Alternatively, you can apply the policy as an outbound policy as below:

```
R1.10 # configure bgp neighbor 10.0.0.2 route-policy out BH_COMM_APPLYR1.11
# configure bgp neighbor 10.0.0.3 route-policy out BH_COMM_APPLYR1.12
# configure bgp neighbor 10.0.0.4 route-policy out BH_COMM_APPLY
```

*show bgp route all Output*

```
R4.67 # show bgp route all
Routes:
```

| Destination | Peer | Next-Hop | LPref | Weight | MED | AS-Path |
|---|---|---|---|---|---|---|
| ---------- | | | | | | |
| *>i 192.51.100.0/28 | 10.0.0.1 | 10.0.0.1 | 100 | 1 | 0 | 64500 |
| *>i 192.51.100.16/28 | 10.0.0.1 | 10.0.0.1 | 100 | 1 | 0 | 64500 |
| *>i 192.51.100.32/28 | 10.0.0.1 | 10.0.0.1 | 100 | 1 | 0 | 64500 |
| *>i 192.51.100.48/28 | 10.0.0.1 | 10.0.0.1 | 100 | 1 | 0 | 64500 |
| *>i 192.51.100.64/28 | 10.0.0.1 | 10.0.0.1 | 100 | 1 | 0 | 64500 |
| *>i 203.0.113.1/32 | 10.0.0.1 | 192.168.2.66 | 100 | 1 | 0 | 64500 |

```
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
```

```
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 6
Feasible Routes : 6
Active Routes : 6
Rejected Routes : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 6
Switch.68 # rtlookup 203.0.113.1
Ori  Destination      Gateway         Mtr  Flags         VLAN      Duration
#be  203.0.113.1/32   192.168.2.66    1   UG-D---um--f BH_VLAN   0d:1h:5m:5s
```

**Note**

For the above solution, the edge routers, R1 through R4, may still export the route to the target network to external AS(s), but the traffic is dropped at the edge of the provider network.

An alternative solution for protecting the network is to perform step 1 only on a designated sink router, (R5 in the following figure) and redistributes the black hole next-hop using iBGP to R2 through R4. When traffic arrives at routers R2 through R4, it is forwarded to R5, since R2–R4 have iBGP routes that resolve the black hole next-hop to R5. Router R5 then discards the traffic.

## BGP Route Filtering Example for IPv4

You can use policy files with *BGP* attributes to filter IPv4 routes. The example in this section is for the topology shown in the following figure.

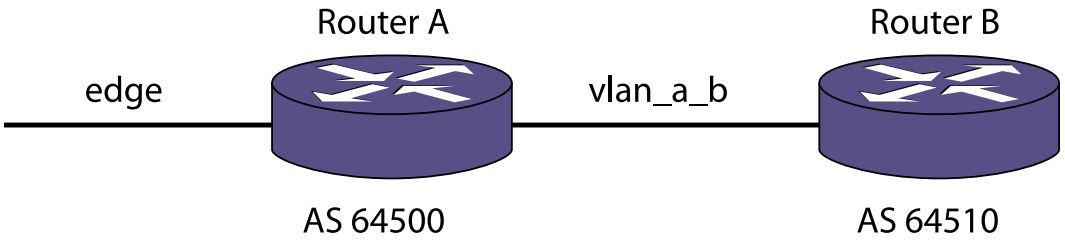



**Figure 236: BGP IPv4 Route Filtering Example**

*Router A Configuration*

```
configure vlan default delete ports all
create vlan "vlan_a_b"
create vlan "edge"
configure vlan vlan_a_b add ports 5 untagged
configure vlan edge add ports 23 untagged
configure vlan edge ipaddress 10.10.10.1/24
enable ipforwarding vlan edge
configure vlan vlan_a_b ipaddress 10.20.20.1/24
enable ipforwarding vlan vlan_a_b
configure bgp AS-number 64500
configure bgp routerid 10.10.10.1
enable bgp community format AS-number:number
create bgp neighbor 10.10.10.2 remote-AS-number 64505
enable bgp neighbor 10.10.10.2
create bgp neighbor 10.20.20.2 remote-AS-number 64510
enable bgp neighbor 10.20.20.2
```

```
enable bgp neighbor 10.10.10.2 soft-in-reset
configure bgp neighbor 10.20.20.2 send-community both
enable bgp
```

*Router B Configuration*

```
create vlan "vlan_B_A"
configure vlan vlan_B_A add ports 25 untagged
configure vlan vlan_B_A ipaddress 10.20.20.2/24
enable ipforwarding vlan vlan_B_A
configure bgp AS-number 64510
configure bgp routerid 10.20.20.2
enable bgp community format AS-number:number
create bgp neighbor 10.20.20.1 remote-AS-number 64500
enable bgp neighbor 10.20.20.1
enable bgp
```

*BGP Routes Before Policy Application*

The following example shows the routes in the *BGP* routing table at Router A after completing the configuration described in the previous two sections:

```
* Switch.52 # show bgp route all
Routes:
Destination       Peer           Next-Hop        LPref Weight MED    AS-Path
-------------------------------------------------------------------------------
*>i  172.16.1.0/24     10.10.10.2     10.10.10.2      100   1      0      64505
*>i  10.1.1.0/24       10.10.10.2     10.10.10.2      100   1      0      64505 64496 0
*>i  10.1.2.0/24       10.10.10.2     10.10.10.2      100   1      0      64505 64496 0
*>i  10.1.3.0/24       10.10.10.2     10.10.10.2      100   1      0      64505 64496 0
*>i  10.1.34.0/24      10.10.10.2     10.10.10.2      100   1      0      64505 64499
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 5
Feasible Routes   : 5
Active Routes     : 5
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 5
```

*Creating and Applying the Route Filter Policy*

The policy described in this section is applied to Router A and does the following:

- Denies routes in network 172.16.1.0/24
- Sets the values of community and MED for routes in network 10.1.0.0/16
- Sets the values of community and MED for routes that contains AS path 64499

The following is a route filter policy named custFilter:

```
entry et1 {
if match all {
nlri 172.16.1.0/24;
} then {
deny;
}
```

```
}
entry et2 {
if match any {
nlri 10.1.0.0/16;
as-path 64499;
} then {
med set 100;
community set "2342:6788";
permit;
}
}
```

- Apply the custFilter inbound policy:

  ```
  * (pacman) DUTA.53 # configure bgp neighbor 10.10.10.2 route-policy in
  custfilter
  ```

*BGP Routes After Policy Application*

The following example shows the routes in the *BGP* routing table at Router A after applying the custFilter inbound policy:

```
* Switch.55 # show bgp route all
Routes:
Destination       Peer           Next-Hop        LPref Weight MED    AS-Path
-------------------------------------------------------------------------------
*>i  10.1.1.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.2.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.3.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.34.0/24      10.10.10.2     10.10.10.2      100   1      100    64505 64499
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 5
Feasible Routes   : 4
Active Routes     : 4
Rejected Routes   : 1
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 4
```

Route 172.16.1.0/24 is not present in the BGP routing table shown above.

The next example shows that the MED and community values are set as defined in the policy.

```
* Switch.56 # show bgp route detail all
Routes:
Route: 10.1.1.0/24, Peer 10.10.10.2, BEST, Active
Origin IGP, Next-Hop 10.10.10.2, LPref 100, MED 100
Weight 1,
As-PATH: 64505 64496 0
Community: 2342:6788
Route: 10.1.2.0/24, Peer 10.10.10.2, BEST, Active
Origin IGP, Next-Hop 10.10.10.2, LPref 100, MED 100
Weight 1,
As-PATH: 64505 64496 0
Community: 2342:6788
Route: 10.1.3.0/24, Peer 10.10.10.2, BEST, Active
Origin IGP, Next-Hop 10.10.10.2, LPref 100, MED 100
```

```
Weight 1,
As-PATH: 64505 64496 0
Community: 2342:6788
Route: 10.1.34.0/24, Peer 10.10.10.2, BEST, Active
Origin IGP, Next-Hop 10.10.10.2, LPref 100, MED 100
Weight 1,
As-PATH: 64505 64499
Community: 2342:6788
BGP Route Statistics
Total Rxed Routes : 5
Feasible Routes   : 4
Active Routes     : 4
Rejected Routes   : 1
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 4
```

View the routes that are denied at Router A.

```
* Switch.57 # show bgp neighbor 10.10.10.2 rejected-routes all
Rejected Routes:
Destination       Peer           Next-Hop        LPref Weight MED    AS-Path
-------------------------------------------------------------------------------
i  172.16.1.0/24     10.10.10.2     10.10.10.2      0     1      0      64505
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 5
Rejected Routes   : 1
Unfeasible Routes : 0
```

The next command example shows the denied route as an inactive route.

The routes were updated because soft-reset is configured for this neighbor.

```
* Switch.61 # show bgp neighbor 10.10.10.2 received-routes all
Routes:
Destination       Peer           Next-Hop        LPref Weight MED    AS-Path
Destination       Peer           Next-Hop        LPref Weight MED    AS-Path
-------------------------------------------------------------------------------
i  172.16.1.0/24     10.10.10.2     10.10.10.2      0     1      0      64505
*>i  10.1.1.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.2.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.3.0/24       10.10.10.2     10.10.10.2      100   1      100    64505 64496 0
*>i  10.1.34.0/24      10.10.10.2     10.10.10.2      100   1      100    64505 64499
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 5
Feasible Routes   : 4
Active Routes     : 4
Rejected Routes   : 1
Unfeasible Routes : 0
```

The following command examples show that the denied routes are not transmitted to the neighbors:

```
* Switch.58 # show bgp neighbor 10.20.20.2 transmitted-routes all
Advertised Routes:
```

```
Destination                 Next-Hop        LPref Weight MED    AS-Path
-------------------------------------------------------------------------------
>i  10.1.1.0/24                 10.20.20.1      0            100    64500 64505 64496 0
>i  10.1.2.0/24                 10.20.20.1      0            100    64500 64505 64496 0
>i  10.1.3.0/24                 10.20.20.1      0            100    64500 64505 64496 0
>i  10.1.34.0/24                10.20.20.1      0            100    64500 64505 64499
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Advertised Routes : 4
* Switch.59 # show bgp neighbor 10.20.20.2 transmitted-routes detail all
Advertised Routes:
Route: 10.1.1.0/24, Active
Origin IGP, Next-Hop 10.20.20.1, MED 100
As-PATH: 64500 64505 64496 0
Route: 10.1.2.0/24, Active
Origin IGP, Next-Hop 10.20.20.1, MED 100
As-PATH: 64500 64505 64496 0
Route: 10.1.3.0/24, Active
Origin IGP, Next-Hop 10.20.20.1, MED 100
As-PATH: 64500 64505 64496 0
Route: 10.1.34.0/24, Active
Origin IGP, Next-Hop 10.20.20.1, MED 100
As-PATH: 64500 64505 64499
BGP Route Statistics
Advertised Routes : 4
```

## BGP Route Filtering Example for IPv6

You can use policy files with *BGP* attributes to filter IPv6 routes. The example in this section is for the topology shown in the following figure.

*Router A Configuration*

```
configure vlan default delete ports all
create vlan "a_b"
enable ipforwarding ipv6 vlan a_b
create vlan "edge"
enable ipforwarding ipv6 vlan edge
configure vlan a_b add ports 5 untagged
configure vlan edge add ports 23 untagged
configure edge ipaddress 2001:db8:2000::1/48
configure a_b ipaddress 2001:db8:3000::1/48
configure bgp AS-number 2100
configure bgp routerid 10.10.10.1
enable bgp community format AS-number:number
create bgp neighbor 2001:db8:2000::2 remote-AS-number 1100
enable bgp neighbor 2001:db8:2000::2
create bgp neighbor 2001:db8:3000::2 remote-AS-number 3300
enable bgp neighbor 2001:db8:3000::2
enable bgp neighbor 2001:db8:2000::2 capability ipv6-unicast
enable bgp neighbor 2001:db8:2000::2 address-family ipv6-unicast soft-in-reset
enable bgp neighbor 2001:db8:2000::2 capability ipv6-multicast
configure bgp neighbor 2001:db8:3000::2 send-community both
enable bgp neighbor 2001:db8:3000::2 capability ipv6-unicast
enable bgp neighbor 2001:db8:3000::2 capability ipv6-multicast
enable bgp
```

*Router B Configuration*

```
configure vlan default delete ports all
create vlan "b_a"
enable ipforwarding ipv6 vlan b_a
configure vlan b_a add ports 25 untagged
configure b_a ipaddress 2001:db8:3000::2/48
configure bgp AS-number 3300
configure bgp routerid 10.20.20.2
enable bgp community format AS-number:number
create bgp neighbor 2001:db8:3000::1 remote-AS-number 2100
enable bgp neighbor 2001:db8:3000::1
enable bgp neighbor 2001:db8:3000::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:3000::1 capability ipv6-multicast
enable bgp
```

*BGP Routes Before Policy Application*

The following example shows the routes in the *BGP* routing table at Router A after completing the configuration described in the previous two sections:

```
* Switch.122 # show bgp routes address-family ipv6-unicast all
Routes:
Destination                                LPref Weight MED
Peer                                    Next-Hop
AS-Path
------------------------------------------------------------------------
*>i  2001:db8:4001::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100 5623
*>i  2001:db8:4002::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100 5623
*>i  2001:db8:4003::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100 5623
*>i  2001:db8:5030::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100 4977
*>i  2001:db8:5031::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100 4977
*>i  2001:db8:a004::/48                          100   1      0
2001:db8:2000::2                        2001:db8:2000::2
1100
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 6
Feasible Routes   : 6
Active Routes     : 6
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 6
```

### *Creating and Applying the Route Filter Policy*

The policy described in this section is applied to Router A and does the following:

- Denies routes in network 2001:db8:a004::/48
- Sets the community and MED values for routes in network 2001:db8:4000::/44
- Sets the community and MED values for routes that contain AS path 4977

The following is a route filter policy named custFilter:

```
entry et1 {
if match all {
nlri 2001:db8:a004::/48;
} then {
deny;
}
}
entry et2 {
if match any {
nlri  2001:db8:4000::/44;
as-path 4977;
} then {
med set 100;
community set "2342:6788";
permit;
}
}
```

Apply the custFilter inbound policy.

```
* Switch.53 # configure bgp neighbor 2001:db8:2000::1 ipv6-unicast route-policy in
custfilter
```

### *BGP Routes After Policy Application*

The following example shows the routes in the *BGP* routing table at Router A after applying the custFilter inbound policy:

```
* Switch.127 # show bgp routes address-family ipv6-unicast all
Routes:
Destination                                LPref Weight MED
Peer                                  Next-Hop
AS-Path
------------------------------------------------------------------------------
*>i  2001:db8:4001::/48                          100   1      100
2001:db8:2000::2                      2001:db8:2000::2
1100 5623
*>i  2001:db8:4002::/48                          100   1      100
2001:db8:2000::2                      2001:db8:2000::2
1100 5623
*>i  2001:db8:4003::/48                          100   1      100
2001:db8:2000::2                      2001:db8:2000::2
1100 5623
*>i  2001:db8:5030::/48                          100   1      100
2001:db8:2000::2                      2001:db8:2000::2
1100 4977
*>i  2001:db8:5031::/48                          100   1      100
2001:db8:2000::2                      2001:db8:2000::2
1100 4977
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
```

```
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 6
Feasible Routes   : 5
Active Routes     : 5
Rejected Routes   : 1
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 5
```

Route 2001:db8:a004::/48 is not present in the BGP routing table shown above.

The routes were updated because soft-reset is configured for this neighbor.

The following command examples show that the denied routes are not transmitted to the neighbors:

```
* Switch.130 # show bgp neighbor 2001:db8:3000::2 address-family ipv6-unicast transmitted-
routes all
Advertised Routes:
Destination                                LPref Weight MED
Peer                                     Next-Hop
AS-Path
------------------------------------------------------------------------------
>i  2001:db8:4001::/48                          0            100
2001:db8:3000::1
2100 1100 5623
>i  2001:db8:4002::/48                          0            100
2001:db8:3000::1
2100 1100 5623
>i  2001:db8:4003::/48                          0            100
2001:db8:3000::1
2100 1100 5623
>i  2001:db8:5030::/48                          0            100
2001:db8:3000::1
2100 1100 4977
>i  2001:db8:5031::/48                          0            100
2001:db8:3000::1
2100 1100 4977
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Advertised Routes : 5
```

The next example shows another way to see that the MED values are set as defined in the policy.

```
* Switch.131 # show bgp neighbor 2001:db8:3000::2 address-family ipv6-unicast transmitted-
routes detail all
Advertised Routes:
Route: 2001:db8:4001::/48, Active
Origin IGP, Next-Hop 2001:db8:3000::1, MED 100
As-PATH: 2100 1100 5623
Route: 2001:db8:4002::/48, Active
Origin IGP, Next-Hop 2001:db8:3000::1, MED 100
As-PATH: 2100 1100 5623
Route: 2001:db8:4003::/48, Active
Origin IGP, Next-Hop 2001:db8:3000::1, MED 100
As-PATH: 2100 1100 5623
Route: 2001:db8:5030::/48, Active
Origin IGP, Next-Hop 2001:db8:3000::1, MED 100
```

```
As-PATH: 2100 1100 4977
Route: 2001:db8:5031::/48, Active
Origin IGP, Next-Hop 2001:db8:3000::1, MED 100
As-PATH: 2100 1100 4977
BGP Route Statistics
Advertised Routes : 5
```

## Route Aggregation Example for IPV4

The following figure shows the topology for this example.

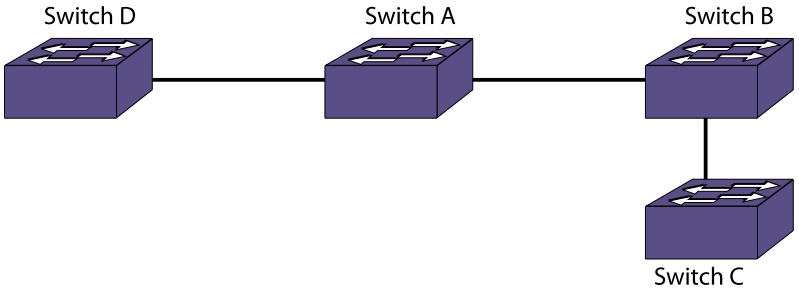



**Figure 237: Route Aggregation Example**

- Configure router A.

```
Configure router A.
create vlan "v3"
create vlan "v1"
create vlan "v2"
configure vlan v3 add ports 1
configure vlan v1 add ports 23
configure vlan v2 add ports 16
configure v1 ipaddress 10.1.1.1/24
configure v2 ipaddress 10.1.2.1/24
configure v3 ipaddress 10.1.4.1/24
enable ipforwarding
configure bgp AS-number 100
configure bgp routerid 1.1.1.1
create bgp neighbor 10.1.1.2 remote-AS-number 64500
create bgp neighbor 10.1.2.2 remote-AS-number 64505
create bgp neighbor 10.1.4.2 remote-AS-number 64510
enable bgp neighbor all
enable bgp
enable bgp aggregation
configure bgp add aggregate-address 172.16.0.0/16 as-set summary-only
```

- Configure router B.

```
Configure router B.
create vlan "v1"
create vlan "net"
configure vlan v1 add ports 5:33
configure vlan net add ports 5:20
configure v1 ipaddress 10.1.1.2/24
configure net ipaddress 172.16.1.1/24
enable ipforwarding
configure bgp AS-number 64500
configure bgp routerid 1.1.1.2
configure bgp add network address-family 172.16.1.0/24
create bgp neighbor 10.1.1.1 remote-AS-number 100
enable bgp neighbor 10.1.1.1
enable bgp
```

- Configure router C.

```
Configure router C.
create vlan "v1"
create vlan "net"
configure vlan net add ports 2:15
configure vlan v1 add ports 2:16
configure v1 ipaddress 10.1.2.2/24
configure net ipaddress 172.16.2.1/24
enable ipforwarding
configure bgp AS-number 64505
configure bgp routerid 2.1.1.2
configure bgp add network 172.16.2.0/24
create bgp neighbor 10.1.2.1 remote-AS-number 100
enable bgp neighbor 10.1.2.1
enable bgp
```

- Configure router D.

```
Configure router D.
create vlan "v1"
configure vlan v1 add ports 24
configure v1 ipaddress 10.1.4.2/24
enable ipforwarding
configure bgp AS-number 64510
configure bgp routerid 5.1.1.2
create bgp neighbor 10.1.4.1 remote-AS-number 100
enable bgp neighbor 10.1.4.1
enable bgp
```

- The following command displays the aggregated route at Router D:

```
* Switch.22 # show bgp routes all
Routes:
Destination         Peer            Next-Hop        LPref Weight MED        AS-Path
-------------------------------------------------------------------------------------
-
*>i  172.16.0.0/16         10.1.4.1        10.1.4.1        100   1      0          100
300
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 1
Feasible Routes   : 1
Active Routes     : 1
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 1
```

## Route Aggregation Example for IPv6

- Configure router A.

```
create vlan "v3"
create vlan "v1"
create vlan "v2"
configure vlan v3 add ports 1
configure vlan v1 add ports 23
configure vlan v2 add ports 16
configure v1 ipaddress 2001:db8:1::1/48
configure v2 ipaddress 2001:db8:3::1/48
```

```
configure v3 ipaddress 2001:db8:5::1/48
enable ipforwarding ipv6
configure bgp AS-number 100
configure bgp routerid 1.1.1.1
create bgp neighbor 2001:db8:1::2 remote-AS-number 200
create bgp neighbor 2001:db8:3::2 remote-AS-number 300
create bgp neighbor 2001:db8:5::2 remote-AS-number 400
enable bgp neighbor all
enable bgp neighbor all capability ipv6-unicast
enable bgp
enable bgp aggregation
configure bgp add aggregate-address address-family ipv6-unicast 2001::/16 as-set
summary-only
```

- Configure router B.

```
create vlan "v1"
create vlan "net"
configure vlan v1 add ports 5:33
configure vlan net add ports 5:20
configure v1 ipaddress 2001:db8:1::2/48
configure net ipaddress 2001:db8:2222::1/48
enable ipforwarding ipv6
configure bgp AS-number 200
configure bgp routerid 1.1.1.2
configure bgp add network address-family ipv6-unicast 2001:db8:2222::/48
create bgp neighbor 2001:db8:1::1 remote-AS-number 100
enable bgp neighbor 2001:db8:1::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:1::1
enable bgp
```

- Configure router C.

```
create vlan "v1"
create vlan "net"
configure vlan net add ports 2:15
configure vlan v1 add ports 2:16
configure v1 ipaddress 2001:db8:3::2/48
configure net ipaddress 2001:db8:2333::2/48
enable ipforwarding ipv6
configure bgp AS-number 300
configure bgp routerid 2.1.1.2
configure bgp add network address-family ipv6-unicast 2001:db8:2333::/48
create bgp neighbor 2001:db8:3::1 remote-AS-number 100
enable bgp neighbor 2001:db8:3::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:3::1
enable bgp
```

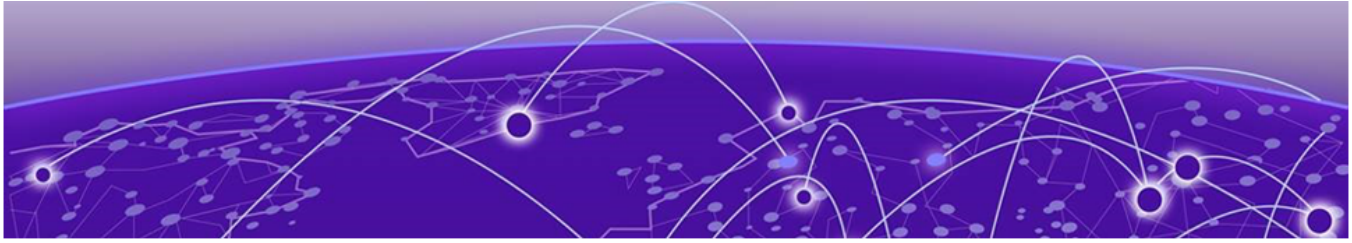
- Configure router D.

```
create vlan "v1"
configure vlan v1 add ports 24
configure v1 ipaddress 2001:db8:5::2/48
enable ipforwarding ipv6
configure bgp AS-number 400
configure bgp routerid 5.1.1.2
create bgp neighbor 2001:db8:5::1 remote-AS-number 100
enable bgp neighbor 2001:db8:5::1 capability ipv6-unicast
enable bgp neighbor 2001:db8:5::1
enable bgp
```

- The following command displays the aggregated route at Router D:

```
* switch # sh bgp routes address-family ipv6-unicast all
Routes:
Destination                                LPref Weight MED
Peer                                    Next-Hop
```

```
AS-Path
------------------------------------------------------------------------------
*>i  2001::/16                                    100   1      0
2001:db8:5::1                              2001:db8:5::1
100 { 200 300 }
Flags: (*) Preferred BGP route, (>) Active, (d) Suppressed, (h) History
(s) Stale, (m) Multipath, (u) Unfeasible
Origin: (?) Incomplete, (e) EGP, (i) IGP
BGP Route Statistics
Total Rxed Routes : 1
Feasible Routes   : 1
Active Routes     : 1
Rejected Routes   : 0
Unfeasible Routes : 0
Route Statistics on Session Type
Routes from Int Peer: 0
Routes from Ext Peer: 1
```



# Layer 3 Virtual Private Network



This chapter introduces Layer 3 VPN, a way to create a tunnel between customer sites through a Provider Backbone, and its features and options in a *BGP (Border Gateway Protocol)*/*MPLS (Multiprotocol Label Switching)* VPN environment. This chapter also provides an example showing L3VPN configuration on a BGP/MPLS setup.

## Overview of Layer 3 VPN

Layer 3 Virtual Private Networks (L3VPN) is a specific implementation of PPVPN (Provider Provisioned VPN). L3VPN is a way to create a tunnel between customer sites through a Provider Backbone, and that tunnel is established and maintained by the Service Provider.

Within a Layer 3 VPN, the customer advertises IP routing knowledge to the provider network. The ISP then advertises this routing information across its network to the other customer locations. This simple concept requires some coordination between the provider and the customer. Also, the provider must configure its network to support the advertisement of these routes and have a way to segregate the customer routes. This is accomplished with the help of a specially designed Network Layer Reachability Information (NLRI).

Within a L3VPN, several routers will play a different role. On the customer site, there is a CE router (Customer Edge). This router is the property of the customer and is managed by them. This router is outside of the Autonomous System (AS) of the Provider. On the Provider side there are some PE routers (Provider Edge) and P routers (Provider). The PE routers are facing the CE routers. They have all the customer's IP routing knowledge. The P routers do not have that knowledge; they are not facing any CE routers, and they serve only to transport the data from PE routers to other PE routers. Their knowledge is very limited, and they usually are intended only for swapping *MPLS* labels. In other words the Provider's core is pure MPLS and has no knowledge of L3VPN, while the edge is L3VPN aware.

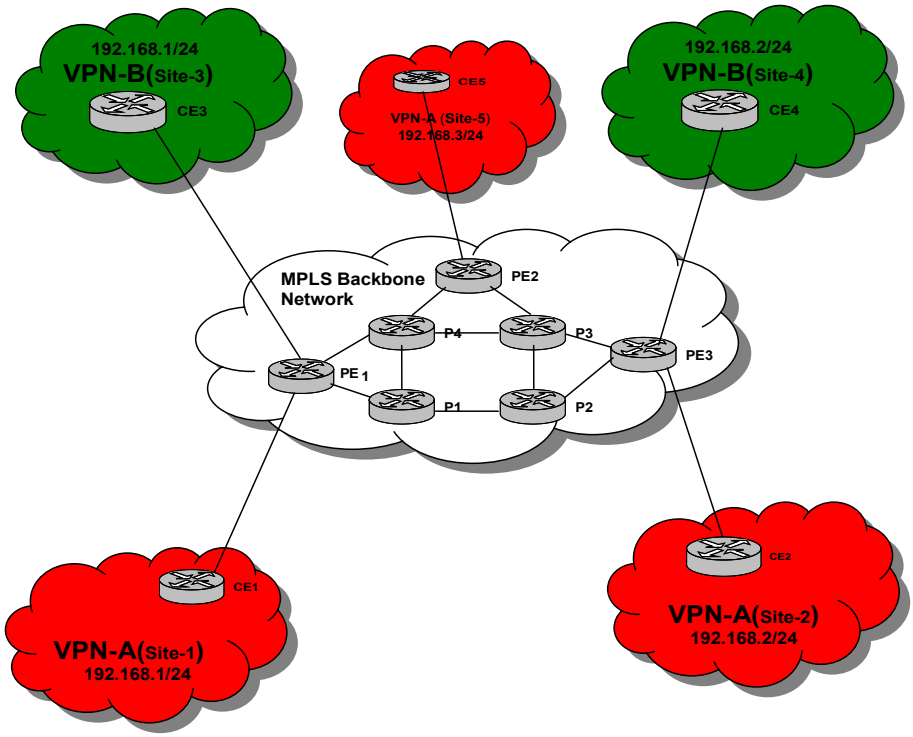



**Figure 238: Different Router Roles in L3 VPN**

A L3VPN requires an MPLS transport mechanism in the Provider Backbone for the data forwarding and Multiprotocol *BGP* (MBGP) between the PE routers to exchange VPN-IPv4 routing information.

## Overview of BGP/MPLS Network

This section introduces the basic concepts of *BGP*/*MPLS* VPNs as described in the Internet RFC 4364. The following figure illustrates an example BGP/MPLS VPN network configuration that includes two VPNs: VPN-A comprises three sites, and VPN-B comprises two. Both VPNs use the same set of IP addresses as defined in RFC 1918.

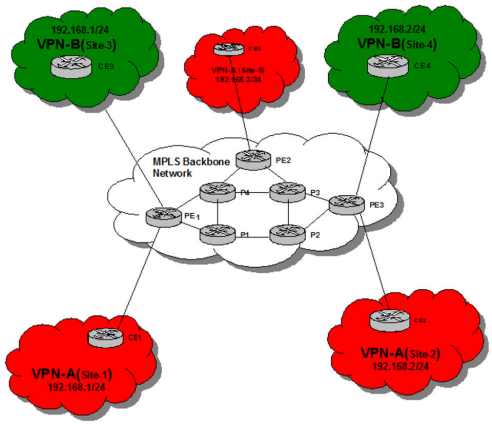



**Figure 239: BGP/MPLS VPN Example**

The Customer Edge boxes represent switches or routers, while the Provider Edge and P(rovider) boxes are routers. If the CE box is a router, then it only needs to peer with the adjacent PE router (s). The PE and P routers form the core MPLS network. The P routers only maintain routing information for the core MPLS network, while the PE routers only maintain routing information for the VPNs they support. This ensures that no single router maintains routing information for all the VPNs.

Since a PE router can support multiple VPNs that may have overlapping address spaces, each PE router must maintain multiple Virtual Routing and Forwarding tables (VRFs). In the example, each site is a member of one VPN, and the PE router is configured to associate a particular VPN with each physical interface to a CE router. The PE router maintains one VRF for each VPN, and packets received from a particular physical interface are forwarded using the VRF for the VPN associated with that interface.

VRFs for a specific VPN are populated in two ways: (1) when the PE router learns routes from a directly-attached CE router that is a member of the VPN, and (2) when the PE router receives routes for the VPN through MBGP from another PE router. The PE router can learn the routes that are reachable at a particular CE router's site through configuration (static routes), or by routing protocol exchanges with the CE router.

VPN route distribution uses BGP-4 Multiprotocol Extensions that enable BGP to carry routes from multiple address families. The VPN-IPv4 address family supports BGP/MPLS VPNs. A VPN-IPv4 address is a 12-byte entity, beginning with an 8-byte (RD) and ending with a 4-byte IPv4 address. The RD makes it possible to create distinct routes to a common IPv4 address prefix, which is necessary when the same IPv4 address prefix is used in two different VPNs.

The purpose of the RD is solely to allow one to create distinct routes to a common IPv4 address prefix. The Route target attribute is used to identify a set of VRFs. Every VRF is associated with one or more route targets. Export route targets identify the set of route targets that a PE router attaches to a route learned from a particular CE site. Import route targets identify the set of route targets that determine whether a route received from another PE router can be inserted in a particular VRF. A VPN-IPv4 route can only be installed in a particular VRF if there is a route target that is both one of the route's targets, and one of the VRF's import route targets. Route targets allow a PE router to only maintain routing information for the VPNs it is supporting. Using import and export route targets offers flexibility in constructing a variety of VPN topologies (such as a fully-meshed closed user group, or a hub-and-spoke architecture). Route Targets are encoded as BGP Extended Communities attributes.

When distributing a VPN route through BGP, a PE router includes its own IP address as the BGP next hop (next-hop-self). The PE router always assigns and distributes an MPLS label for each customer VPN VRF that it receives from directly connected sites (CEs). BGP-distributed MPLS labels require that there be a label switched path between the PE router that installs the BGP-distributed route and the PE router that is the BGP next hop of that route. This is necessary because a multi-label stack is used to forward VPN packets across the MPLS backbone.

The outer MPLS label gets the packet across the backbone. This label is obtained from the MPLS signaling protocols, and is associated with the best LSP to the BGP next hop address of the PE Router that advertised the VPN route. The inner MPLS label is obtained from BGP, and is associated with the best route to the VPN destination address. This label identifies the VRF that the egress PE Router uses to forward the packet to a CE device, and may indicate the outgoing interface that the packet should be forwarded over (along with the appropriate link layer header for that interface).

The use of a two-level MPLS label stack is an important scalability attribute of the model, because it is the two-level label stack that enables the P routers to operate without containing routes for any of the VPNs.

In summary, key aspects of the BGP/MPLS VPN model include:

- Direct peering of customer routers with service provider routers.
- Maintenance of multiple forwarding tables by PE routers.
- Introduction of the VPN-IPv4 address family.
- Constrained distribution of routing information via Route Targets.
- Use of MPLS label switching in the backbone network.

## VPNV4 NLRI

One of the core principles of operating a VPN is maintaining separation between the customer networks. L3VPN uses a special format to represent customer routes within the provider's network. This format allows each provider router to uniquely view routes from different customers as separate, even when they advertise the same IPv4 prefix. The format consists of these fields: Mask, MPLS label, route distinguisher, and IPv4 prefix.

## Limitations

The following list identifies the limitations of the L3VPN feature:

- *RIP (Routing Information Protocol)* is not supported for PE – CE peering routing protocol.
- IP Multicast BGP/MPLS VPN is not supported
- OSPFv2 and ISIS for PE – CE peering routing protocol.
- IPv6 VPN is not supported.
- Graceful restart mechanism for BGP with MPLS (RFC-4781) is not supported.
- Constraint Route distribution for BGP/MPLS VPN (RFC-4684) is not supported.
- Carrier of carriers BGP/MPLS VPN configuration is not supported.
- XML support to configure BGP/MPLS VPN parameters.
- Carrier's carrier (RFC 4364, Section 9).
- Inter-AS / inter-provider VPNs (RFC 4364, Section 10).

- Use of route reflectors with BGP ORFs.
- No other BGP related MIB other than RFC 1657 is supported.

## Overlapping Customer Address Spaces

VPN customers often manage their own networks and use the RFC-1918 private address space. If globally unique IPv4 addresses are not used, the same 32-bit IPv4 address can be used to identify different systems in different VPNs. This causes routing problems because *BGP* assumes that each IPv4 address that it carries is globally unique. To solve this problem, BGP/*MPLS* VPN converts non-unique IPv4 addresses into unique VPN-IPv4 address families. Each IPv4 address is prepended with an 8-bytes long Route Distinguisher (RD).

A VPN-IPv4 address is a 12-byte quantity composed of an 8-byte RD and 4 byte IPv4 address. The 8-byte RD is composed of a 2-bytes type field and a 6-bytes value field. The value of the Type field determines the lengths of the Value field's two subfields (Administrator and Assigned Number), as well as the semantics of the Administrator field.

Most commonly, the Admin field holds a 2-byte non-private AS number and the Assigned number field holds a 4-byte number assigned by the VPN service provider. Another common format for the RD is the Admin field holds a global 4-byte IPv4 address, and the Assigned number field holds 2 byte number assigned by the VPN service provider.

## Multi-protocol BGP Extension

*BGP* is originally designed to carry routing information for IPv4 address families only. Since VPN-IPv4 addresses are 12 bytes wide, BGP requires some kind of extension to carry VPN-V4 address family routing information. This is achieved by using MBGP extension which allows BGP to carry multiprotocol address families routing information. Therefore, to deploy BGP/*MPLS* VPNs and to support the distribution of VPN-IPv4 routes, PE routers are required to support the MP-BGP extensions with VPN-IPv4 address family support.

## Multiple Forwarding Tables

Each PE router needs to maintain one forwarding table per VPN that it is directly connected. These forwarding tables are called Virtual Routing and Forwarding tables (VRF). When a PE router is configured, each of it's VRF is associated with one or more *VLAN (Virtual LAN)*s, that is, there is a many to one mapping from VLAN to VPN/VRF. When receiving inbound data traffic from a directly attached CE router, PE router determines the VRF for the packet based on which VLAN the packet has been received. The ingress VLAN will uniquely map to a particular VRF in the system. Then, PE router performs a route lookup for the packet's destination IP address in the associated VRF. In the figure above, both PE1 and PE3 have two VRFs, one each for VPN-A and VPN-B, whereas, PE2 has only one VRF which is for VPN-A.

## Quality of Service in BGP/MPLS VPN

The L3VPN implementation uses the existing *ACL (Access Control List)* and *QoS (Quality of Service)* framework available in ExtremeXOS to achieve QoS on a PE. Please refer to the Quality of Service on page 724 section for more information.

## Virtual Routing and Forwarding Instances

The concept of Virtual Routing and Forwarding (VRF) is similar to the *virtual router (VR)* (VR) already present in Extreme Switches. Like a VR, each VRF also has a distinct set of VLANs and ports, as well as a unique name assigned to it across the system. VLANs are created, and ports are added to in the same manner as is done for VRs. Logically separate instances of the same routing protocol may run in the contexts of multiple VRFs of the same switch, simultaneously. In addition, a VRF may also contain a Route Distinguisher (RD), multiple import and export Router Targets (RTs), as well as a VPN ID. These three items allow a VRF to support L3VPNs.

The goal of VRF is to provide separation of overlapping address spaces belonging to separate routing domains. The objective of the VR is to split the switch into multiple routing domains so that traffic originating on one VR never enters another VR. The objective of VRFs is to control the flow of traffic across various VPN sites. As the PE-PE connection is always part of the default VR or a user VR, the forwarding table corresponding to a particular VRF has the incoming interface belonging to the VRF, and egress interfaces belonging to the default, or user VR. Similarly, the forwarding table (*MPLS* label table) corresponding to a user, or default VR, could have an interface belonging to a VRF as part of its egress interfaces. In this sense a VRF is "associated" with the user, or default VR (also refered to as parent VRs).

A VRF is not the same as a VR. A VRF always requires that you have a Parent VR, and this can either be the default VR or User VR. Protocols running in a VRF run as a separate logical instance within the context of the protocol process that is running in the Parent VR. EXOS VRFs come in two types:

- Non-VPN VRFs - Non-VPN VRFs are used for L3 Routing and Forwarding just like VRs. These provide the ability to scale protocol deployments. *BGP* and static routing are supported in the Non-VPN VRF.
- VPN VRFs - VPN VRFs enable configuration of RD and RT to realize L3VPN topologies.

  Note
  MPLS and BGP must be configured in the parent VR (PE-facing).

  BGP and static routing are supported in the VPN VRF.

Again, a VRF always requires that you have a Parent VR, and this can either be the default VR or User VR. Protocols running in a VRF run as a separate logical instance within the context of the protocol process that is running in the Parent VR. For example, the BGP process for a VR running MPLS will handle all PE to CE instances of BGP by virtualizing the data structures. This approach of allowing VRFs in a VR is more scalable than having a process for every instance of PE-to-CE routing protocol.

## L3VPN Configuration Example

Here is an example of a Layer 3 VPN. In this example CE1 and CE2 are in different AS. You can put them in the same AS by adding the `allowas-in` command.

In this example, you should be able to ping CE2 "foo1's" IP address from CE1.

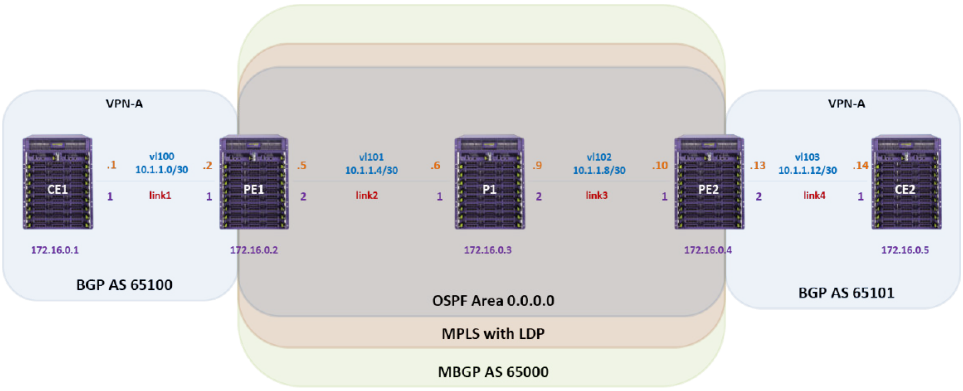


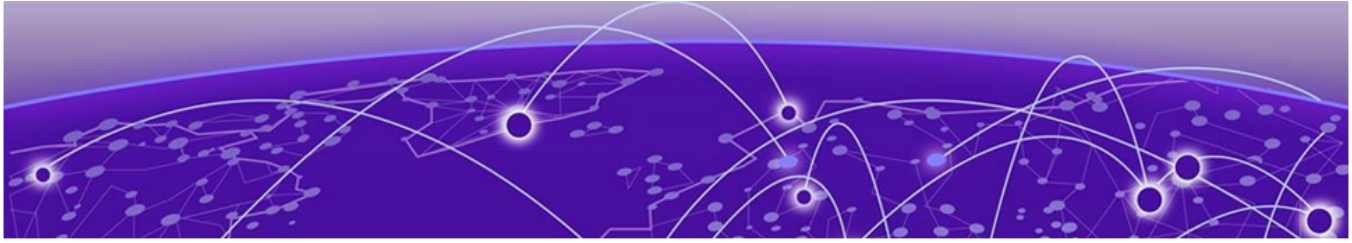

**Figure 240: Layer 3 Virtual Private Network**

```
CE1:
configure snmp sysName "CE1"
create vlan "lo0"
enable loopback-mode vlan lo0
create vlan "vl100"
configure vlan vl100 tag 100
enable jumbo-frame ports 1
configure vlan vl100 add ports 1 tagged
configure vlan Mgmt ipaddress 192.168.56.111 255.255.255.0
configure vlan lo0 ipaddress 172.16.0.1 255.255.255.255
configure vlan vl100 ipaddress 10.1.1.1 255.255.255.252
enable ipforwarding vlan vl100

configure bgp AS-number 65100
configure bgp routerid 172.16.0.1
create bgp neighbor 10.1.1.2 remote-AS-number 65000
enable bgp neighbor 10.1.1.2
enable bgp

PE1:
configure snmp sysName "PE1"
create vr "vpn-a" type vpn-vrf vr "VR-Default"
configure vr VR-Default delete ports 1
configure vr vpn-a add ports 1
create vlan vl100 vr vpn-a tag 100
configure vl100 add ports 1 tagged
create vlan "vl101"
configure vlan vl101 tag 101
enable jumbo-frame ports 1
enable jumbo-frame ports 2
configure vlan vl100 add ports 1 tagged
configure vlan vl101 add ports 2 tagged
configure vlan Mgmt ipaddress 192.168.56.102 255.255.255.0
configure vlan lo0 ipaddress 172.16.0.2 255.255.255.255
enable ipforwarding vlan lo0
configure vlan vl101 ipaddress 10.1.1.5 255.255.255.252
enable ipforwarding vlan vl101
configure vlan vl100 ipaddress 10.1.1.2 255.255.255.252
enable ipforwarding vlan vl100

configure vr vpn-a add protocol bgp
configure vr vpn-a rd 172.16.0.2:100
configure vr vpn-a route-target both add 65000:100

enable iproute mpls-next-hop

configure bgp AS-number 65000
```

```
configure bgp routerid 172.16.0.2
create bgp neighbor 172.16.0.4 remote-AS-number 65000
configure bgp neighbor 172.16.0.4 source-interface ipaddress 172.16.0.2
enable bgp neighbor 172.16.0.4
configure bgp neighbor 172.16.0.4 next-hop-self
configure bgp neighbor 172.16.0.4 address-family vpnv4 next-hop-self
enable bgp neighbor 172.16.0.4 capability vpnv4
enable bgp export vr vpn-a direct address-family vpnv4
enable bgp export vr vpn-a bgp address-family vpnv4
enable bgp

virtual-router vpn-a
configure bgp AS-number 65000
configure bgp routerid 172.16.0.2
create bgp neighbor 10.1.1.1 remote-AS-number 65100
enable bgp neighbor 10.1.1.1
disable bgp neighbor 10.1.1.1 capability ipv4-multicast
enable bgp export remote-vpn address-family ipv4-unicast
enable bgp
virtual-router VR-Default

configure mpls add vlan "lo0"
enable mpls vlan "lo0"
enable mpls ldp vlan "lo0"
configure mpls add vlan "vl101"
enable mpls vlan "vl101"
enable mpls ldp vlan "vl101"
configure mpls lsr-id 172.16.0.2
enable mpls protocol ldp
enable mpls

configure ospf routerid 172.16.0.2
enable ospf
configure ospf add vlan lo0 area 0.0.0.0 passive
configure ospf add vlan vl101 area 0.0.0.0 link-type point-to-point

P1:
configure snmp sysName "P1"

configure vlan default delete ports 1-2
create vlan "lo0"
enable loopback-mode vlan lo0
create vlan "vl101"
configure vlan vl101 tag 101
create vlan "vl102"
configure vlan vl102 tag 102
enable jumbo-frame ports 1
enable jumbo-frame ports 2
configure vlan vl101 add ports 1 tagged
configure vlan vl102 add ports 2 tagged
configure vlan Mgmt ipaddress 192.168.56.103 255.255.255.0
configure vlan lo0 ipaddress 172.16.0.3 255.255.255.255
enable ipforwarding vlan lo0
configure vlan vl101 ipaddress 10.1.1.6 255.255.255.252
enable ipforwarding vlan vl101
configure vlan vl102 ipaddress 10.1.1.9 255.255.255.252
enable ipforwarding vlan vl102

enable iproute mpls-next-hop

configure mpls add vlan "lo0"
enable mpls vlan "lo0"
enable mpls ldp vlan "lo0"
configure mpls add vlan "vl101"
```

```
enable mpls vlan "vl101"
enable mpls ldp vlan "vl101"
configure mpls add vlan "vl102"
enable mpls vlan "vl102"
enable mpls ldp vlan "vl102"
configure mpls lsr-id 172.16.0.3
enable mpls protocol ldp
enable mpls

configure ospf routerid 172.16.0.3
enable ospf
configure ospf add vlan lo0 area 0.0.0.0 passive
configure ospf add vlan vl101 area 0.0.0.0 link-type point-to-point
configure ospf add vlan vl102 area 0.0.0.0 link-type point-to-point

PE2:

configure snmp sysName "PE2"

configure vr vpn-a add ports 2
configure vr VR-Default delete ports 2
create vr "vpn-a" type vpn-vrf vr "VR-Default"
configure vlan default delete ports 2
create vlan "lo0"
enable loopback-mode vlan lo0
create vlan "vl102"
configure vlan vl102 tag 102
create vlan "vl103" vr vpn-a
configure vlan vl103 tag 103
enable jumbo-frame ports 1
enable jumbo-frame ports 2
configure vlan vl102 add ports 1 tagged
configure vlan vl103 add ports 2 tagged
configure vlan Mgmt ipaddress 192.168.56.104 255.255.255.0
configure vlan lo0 ipaddress 172.16.0.4 255.255.255.255
enable ipforwarding vlan lo0
configure vlan vl102 ipaddress 10.1.1.10 255.255.255.252
enable ipforwarding vlan vl102
configure vlan vl103 ipaddress 10.1.1.13 255.255.255.252
enable ipforwarding vlan vl103

configure vr vpn-a add protocol bgp
configure vr vpn-a rd 172.16.0.4:103
configure vr vpn-a route-target both add 65000:100

enable iproute mpls-next-hop

configure bgp AS-number 65000
configure bgp routerid 172.16.0.4
create bgp neighbor 172.16.0.2 remote-AS-number 65000
configure bgp neighbor 172.16.0.2 source-interface ipaddress 172.16.0.4
enable bgp neighbor 172.16.0.2
configure bgp neighbor 172.16.0.2 next-hop-self
configure bgp neighbor 172.16.0.2 address-family vpnv4 next-hop-self
enable bgp neighbor 172.16.0.2 capability vpnv4
enable bgp export vr vpn-a direct address-family vpnv4
enable bgp export vr vpn-a bgp address-family vpnv4
enable bgp

virtual-router vpn-a
configure bgp AS-number 65000
configure bgp routerid 172.16.0.4
create bgp neighbor 10.1.1.14 remote-AS-number 65101
enable bgp neighbor 10.1.1.14
```

```
disable bgp neighbor 10.1.1.14 capability ipv4-multicast
enable bgp export remote-vpn address-family ipv4-unicast
enable bgp
virtual-router VR-Default

configure mpls add vlan "lo0"
enable mpls vlan "lo0"
enable mpls ldp vlan "lo0"
configure mpls add vlan "vl102"
enable mpls vlan "vl102"
enable mpls ldp vlan "vl102"
configure mpls lsr-id 172.16.0.4
enable mpls protocol ldp
enable mpls

configure ospf routerid 172.16.0.4
enable ospf
configure ospf add vlan lo0 area 0.0.0.0 passive
configure ospf add vlan vl102 area 0.0.0.0 link-type point-to-point

CE2:

configure snmp sysName "CE2"

configure vr vpn-a add ports 1
configure vlan default delete ports all
configure vr VR-Default delete ports 1
configure vlan default delete ports 1
create vlan "foo1"
enable loopback-mode vlan foo1
create vlan "lo0"
enable loopback-mode vlan lo0
create vlan "vl103"
configure vlan vl103 tag 103
enable jumbo-frame ports 1
configure vlan vl103 add ports 1 tagged
configure vlan Mgmt ipaddress 192.168.56.105 255.255.255.0
configure vlan lo0 ipaddress 172.16.0.5 255.255.255.255
configure vlan vl103 ipaddress 10.1.1.14 255.255.255.252
configure vlan foo1 ipaddress 10.2.1.1 255.255.255.0
enable ipforwarding vlan foo1

configure bgp AS-number 65101
configure bgp routerid 172.16.0.5
configure bgp add network 10.2.1.0/24
create bgp neighbor 10.1.1.13 remote-AS-number 65000
enable bgp neighbor 10.1.1.13
enable bgp
```



# Multicast Routing and Switching



This chapter introduces the features and usage of IP multicasting, which allows a single host on a network to send a packet to a group of hosts. For more information on IP multicasting, refer to the following publications:

- RFC 1112—Host Extension for IP Multicasting
- RFC 2236—Internet Group Management Protocol, Version 2
- RFC 3569—SSM for IPv4/IPv6 (only for IPv4)
- PIM-SM Version 2—draft-ietf-pim-sm--v2-new-05
- RFC 4601—PIM SM (only for IPv4)
- RFC 2362 PIM-SM (Edge Mode)
- RFC 3973 PIM-DM (only for IPv4)
- RFC 3569—draft-ietf-ssm-arch-06.txt PIM-SSM PIM Source Specific Multicast
- PIM-DM Draft IETF Dense Mode—draft-ieft-idmr-pimdm-05.txt, draft-ieft-pim-dm-new-v2-04.txt
- RFC 3376—Internet Group Management Protocol, Version 3

The following URL points to the website for the IETF PIM Working Group: http://datatracker.ietf.org/wg/pim/documents/.

## Multicast Routing Overview

Multicast routing and switching is the functionality of a network that allows a single host (the multicast server) to send a packet to a group of hosts. With multicast, the server is not forced to duplicate and send enough packets for all the hosts in a group. Instead, multicast allows the network to duplicate packets for all of the hosts in a group. Multicast greatly reduces the bandwidth required to send data to a group of hosts. IP multicast routing is a function that allows multicast traffic to be forwarded from one subnet to another across a routing domain.

IP multicast routing requires the following functions:

- A router that can forward IP multicast packets
- A router-to-router multicast routing protocol (for example, Protocol Independent Multicast (PIM)) to discover multicast routes
- A method for the IP host to communicate its multicast group membership to a router (for example, *IGMP (Internet Group Management Protocol)*)

Note
You should configure IP unicast routing before you configure IP multicast routing.

## Multicast Table Management

The ExtremeXOS software uses the following tables to support IP multicast traffic:

- IPv4 multicast route table
- L3 hash table
- IP multicast group table
- *FDB (forwarding database)* table (L2 table)
- L2 multicast table (L2MC table)

### IP Multicast Hardware Lookup Modes

Extreme platforms support various hardware forwarding lookup modes by using a combination of L3 hash table and L2 (*FDB*) table. Refer to Multicast Table Management for details on these tables. The scalability limits vary based on the lookup mode used.

#### *Configuration Options*

Configuration options allow you to choose the hardware forwarding lookup mode for multicast forwarding. Here is a list of options:

- source-group-vlan -- Uses L3 hash table with S,G,V lookup. This is the default mode for all platforms except the Summit X430.
- group-vlan -- Uses L3 hash table with *,G,V lookup.
- mac-vlan -- Uses L2 table with DMAC, *VLAN (Virtual LAN)* lookup. This is the default mode for x430.
- mixed-mode -- Uses both L2 and L3 tables for multicast. In this mode, the following logic is applied on installing the cache entries in the hardware:

  Multicast cache entries requiring forwarding across VLANs would be installed in the L3 IP multicast table. This includes PIM, MVR, and PVLAN cache entries.

  Multicast cache entries requiring L2 forwarding within a VLAN are installed in the L2 table. This includes entries corresponding to *IGMP* Snooping, PIM snooping, and MLD snooping.

  Any IPv4/v6 reserved multicast addresses (for example, 224.0.0.x or IPv6 equivalent) are installed in the L3 IP multicast table as needed. These reserved addresses map to the following

multicast MAC addresses: 01:00:5e:00:00:xx, 33:33:00:00:00:xx, 33:33:00:00:01:xx, or 33:33:ff:xx:xx:xx.

Note

Any change in the lookup key configuration causes all cache entries to be cleared, and traffic is temorarily dropped until the system re-learns the multicast caches and associated subscriptions.

Note

mac-vlan mode helps increase scaling and is particularly useful on platforms like the Summit X440, which has limited L3 hardware table entries. This mode is also supported in other Summit platforms, and the BlackDiamond8K and BlackDiamond X.

Note

mac-vlan and mixed-mode are not supported prior to ExtremeXOS 15.3.1.

The EXOS multicast process continues to maintain the cache entries as "S,G,V", and interacts with HAL the same way as today. EXOS hardware abstraction layer (HAL) applies the logic explained above and installs the cache entries in the appropriate hardware table. If the cache entry needs to be installed in the L2 multicast table, HAL derives the MAC address based on the standard logic and installs the MAC entry in the L2 table.

The IP multicast address to MAC address mapping is not validated for the received/forwarded multicast packets in EXOS to date. If the lookup mode is configured either as "mac-vlan" or "mixed-mode", the multicast kernel module is modified to validate this mapping and, if a packet does not use the standard mapping, the packet is dropped.

*IPMC Compression*

In order to increase the scaling of multicast entries, EXOS implements a feature called IPMC compression which allows multiple <S,G,V> (or <*,G,V>) IP multicast FDB entries to utilize the same IP multicast group table entry when the associated egress port lists are the same. The base IP multicast compression implementation will be reused for achieving L2 multicast entry reuse. In this case, multiple <MAC,VLAN> multicast FDB entries can use a single L2MC index if the egress port lists of the cache entries are the same.

*Interactions with Static FDBs*

EXOS allows you to create FDB entries for multicast MAC address using:

`create fdb` *`mac_addr`* {**`vlan`**} *`vlan_name`* **`ports`** *`port_list`*

These entries also get installed in the L2 table and use the L2MC table for hardware forwarding. If there is a dynamic <MAC,VLAN> entry from MCMGR and a static entry from FDB manager, the static entry takes precedence and the dynamic entry would get deleted in hardware. Compression of L2MC indices is not supported on these types of entries. Each newly created static multicast FDB will cause the allocation of a new L2MC index.

*Interactions with Dynamic FDBs*

When IP multicast forwarding entries are utilizing the L2 MAC table, the multicast entries are installed as static in the hardware L2 table to avoid undesirable interactions with L2 protocol or user administered

FDB flushing. These multicast L2 entries also take precedence over dynamic unicast L2 MAC entries. If there is a hash bucket collision upon inserting an L2 multicast entry, it will replace another dynamic unicast L2 entry if one exists in the same hash bucket.

### *Platforms with External-Tables (TCAM)*

The X480 and BD8K xl-series have a large external TCAM that can be used to store MAC FDBs, L3 routes, ACLs, and/or IPMC forwarding entries based on configuration. Only the internal L2 table is used to store <MAC,VLAN> forwarding entries for IP multicast caches on these platforms due to a hardware limitation. Additionally, the `configure forwarding external-table l2-and-l3-and-ipmc` configuration option, which uses the external TCAM to store <S,G,V> entries, is not compatible with the "mac-vlan" and "mixed-mode" options of this feature.

### *Virtual Router Support*

Current IPMC cache hardware entries stored as <S,G,V> additionally include the VRID associated with the ingress *virtual router (VR)*. In this feature, <MAC, VLAN> cache entries are stored in the L2 table which does not additionally include the VRID. However, user VRs are still supported since the VLAN portion of the lookup key is unique across all VRs.

### *IPMC Cache Rate Limiting*

Based on the number of cache entries supported on each platform, there is a software cache limiting implementation present in EXOS multicast. The HAL module informs MCMGR about the supported limit, MCMGR creates cache entries up to MAX supported limit, and the remaining traffic is dropped in software.

### *Supported Platforms*

This feature is implemented on all Summit, BD8K, and BDX8 platforms.

Note
The mixed-mode setting is supported on all platforms except the BD8K "e2-series".

### *Limitations*

The following limitations exist for the L2MC table feature:

- The "mixed-mode" configuration option is not allowed on platforms using older chipsets. Please see the "Platforms Supported" section for details.
- When the "mixed-mode" configuration option is engaged on BD8K platforms, newly inserted slots which do not support "mixed-mode" will fail initialization.
- On SummitStack, this same condition causes the following log to be displayed repeatedly every 30 seconds:

  ```
  <HAL.IPv6Mc.Error> Stack slot %d is incompatible with the multicast forwarding lookup configuration.
  ```

  Either remove this node from the stack or change the multicast forwarding lookup configuration.
- When using the "mac-vlan" configuration option:

PIMv4/V6, MVR features cannot be used.
IGMPv3 should not be used in conjunction with this mode
Private VLAN multicast should not be used.
Issues with IP multicast address to MAC address mapping:

All IPv4 multicast frames use multicast mac addresses starting with 01:00:5e:xx:xx:xx. The lower order 23 bits of the IP multicast address is used in the MAC address derivation. As only 23 bits of MAC addresses are available for mapping layer 3 IP multicast addresses to layer 2 MAC addresses, this mapping results in 32:1 address ambiguity.

When traffic is received for 1 out of these 32 overlapping address, then the MAC, Vlan entry is installed in hardware based on the IGMP group membership of received traffic's destination multicast IP address. After this installation, traffic to any of the remaining 31 addresses is delivered based on the existing cache entry and the actual receiver list of the remaining 31 addresses will not be honored.

IPv6 multicast streams use multicast MAC addresses in the form 33:33:xx:xx:xx:xx. The lower 24 bits of the IPv6 multicast address are used to derive the MAC address. So, the address ambiguity issue is also applicable to IPv6 with more severity. Given this condition, we do not recommend using overlapping IP multicast addresses with this mode.

This limitation applies to "mixed-mode", too.

- IPv4 multicast addresses consist of a block of addresses (224.0.0.x) used for network control traffic. Packets having IP destination addresses from the LNCB are always flooded to all ports of the VLAN. The address ambiguity issue discussed above is applicable for the addresses in this block too. For example, 224.0.0.5 (address used for *OSPF (Open Shortest Path First)*) and 225.0.0.5 would use the same MAC address 01:00:5e:00:00:05. If a mac based multicast FDB entry is installed on the hardware for 01:00:5e:00:00:05 based on the 225.0.0.5 join list, it would break OSPF functionality. Hence, MAC addresses mapping to the LNCB block will not be installed in the L2 table, resulting in software forwarding for those streams. We recommend that you avoid useing multicast addresses that map to the 01:00:5e:00:00:xx MAC address range.
- As per RFC 3307, IANA assigned reserved IPv6 multicast addresses could be in the group Id range of 0x00000001 to 0x3FFFFFFF. As a result, EXOS switches flood traffic addressed to ff02::/98 to all ports of the VLAN. Since the lower 32 bits of IPv6 multicast addresses are mapped to the multicast mac address, not installing all of the addresses in this range would make it too restrictive. So, installing entries for 33:33:00:00:00:xx in hardware would be avoided, and the traffic would be software forwarded.

  In addition, the following important IPv6 multicast addresses cannot be installed as hardware forwarding entries:

  *DHCP (Dynamic Host Configuration Protocol)*: All-dhcp-agents address FF02:0:0:0:0:0:1:2
  All-dhcp-servers address FF05:0:0:0:0:0:1:3
  Neighbor Discovery (ND): Solicited-node-address FF02::1:FF00:0000/104

Therefore, the following multicast MAC addresses are not programmed in hardware, and the corresponding packets are handled in slowpath: 33:33:00:00:01:xx , 33:33:ff:xx:xx:xx

- Given the issues with IP multicast address to MAC address mapping, no attempt is made to merge subscriber lists of multiple overlapping IP groups.
- The following limitation regarding IPMC compression is also applicable for this feature, because this feature uses the same L2MC entry for multiple l2 multicast entries with same egress ports. All MAC-VLAN forwarding entries utilizing the same L2MC entry will be subject to a single BD8K backplane link (12Gbps).
- On those platforms supporting the "external-table" (X480, BD8K "xl-series"), any IP multicast caches installed in the L2 table will be only installed in the internal L2 table due to a hardware limitation which prevents L2MC access from the ESM (External Search Machine).
- When IP multicast forwarding entries are installed as <MAC,VLAN>, IGMP or MLD packets which have a MAC-DA=<group> will cause the refresh of the IP multicast cache, preventing timely entry age-out.
- The L2MC table is limited to 1K entries on all platforms. This means that only up to 1K unique port lists can be addressed from the <MAC,VLAN> IP multicast forwarding entries that are stored in the L2 table. Additionally, statically created multicast FDB entries do not perform L2MC index compression.

## IPv4 Multicast Route Table

Beginning with Release 12.1, all IP multicast routes are stored and maintained in the software multicast route table. Routes are added to the multicast route table from the following sources:

- Multicast static routes (configured manually by the network administrator)
- Multicast dynamic routes (learned through protocols such as MBGP and MISIS)

The multicast route table is used for reverse path forwarding (RPF) checks, not for packet forwarding. The switch uses RPF checks to avoid multicast forwarding loops. When a multicast packet is received, the switch does an RPF lookup, which checks the routing tables to see if the packet arrived on the interface on which the router would send a packet to the source. The switch forwards only those packets that arrive on the interface specified for the source in the routing tables.

The RPF lookup uses the multicast routing table first, and if no entry is found for the source IP address, the RPF lookup uses the unicast routing table.

Note

Because the multicast routing table is used only for RPF checks (and not for routing), IP route compression and *ECMP (Equal Cost Multi Paths)* do not apply to multicast routes in the multicast routing table.

Beginning with ExtremeXOS software version 12.1, the route metric is no longer used to select between multicast and unicast routes. If the RPF lookup finds a route in the multicast table, that route is used. The unicast routing table is used only when no route is found in the multicast table.

The advantage to having separate routing tables for unicast and multicast traffic is that the two types of traffic can be separated, using different paths through the network.

## L3 Hash Table

The L3 hash table is introduced in Introduction to Hardware Forwarding Tables on page 1255. The L3 hash table stores entries for IPv4 routes, IPv4 and IPv6 hosts, and IPv4 and IPv6 multicast groups. For multicast, L3 hash table supports <S,G,V> and <*,G,V> lookups. The entry from this table provides an index to IP Multicast Group table.

To make more space available in the L3 hash table for IPv4 and IPv6 multicast groups, you can do the following:

- Configure the extended IPv4 host cache feature to move IPv4 local and remote routes to the LPM table as described in Extended IPv4 Host Cache on page 1254.
- Configure BlackDiamond 8900 xl-series modules or Summit X480 series switches to do one of the following:
  - Move IPv4 local and remote hosts to the external LPM table.
  - Move IPv6 local hosts to the external LPM table.
  - Move IPv4 local and remote hosts to the external LPM table and support IPv4 multicast entries in the external LPM table.

  For more information, see the description for the #unique_1391 command.

**Note**
To benefit from the use of the external LPM tables, you must leave the IP multicast compression feature enabled, which is the default setting.

## IP Multicast Group Table

The IP multicast group table specifies the egress ports for Layer 2 and Layer 3 multicast traffic groups. To make more space available in the IP multicast group table, you can do the following:

- Leave IP multicast compression enabled (it is enabled by default). This allows multiple L3 hash entries to share the same IP multicast group entry if the the egress list is the same.
- Use the following I/O modules, which provide higher capacity tables:
  - BlackDiamond 8900 xl-series modules
  - BlackDiamond 8900-G96T-c

## Capacity Restrictions for Mixed Installations

A mixed installation is a switch configuration that contains I/O modules with different table sizes. The actual IP multicast group table capacity for the switch is set to that supported on the I/O module with the smallest tables. To increase the capacity of IP multicast tables, all I/O modules must support the minimum table size you want.

Multicast forwarding entries are programmed in all I/O modules. Only multicast traffic ingressing a given I/O module utilize these forwarding entries. Other egress-only I/O modules only require the multicast group table entry.

If you add a higher-capacity I/O module to a switch that has been running with lower capacity modules, the switch generates a message and adjusts the table capacity on the higher-capacity card to that of the lower-capacity card.

Compared to the L3 hash table that uses an IP address for forwarding, the L2 table uses a MAC address. The L2 table stores unicast and multicast MAC entries, and it supports <DMAC, *VLAN* lookup. The entry from this table provides an index to the L2MC table that specifies the egress ports.

# PIM Overview

Protocol Independent Multicast (PIM) is the de-facto standard for routing multicast traffic over the Internet. Other multicast routing protocols such as DVMRP and MOSPF are sometimes used in controlled environment, but are not widely deployed. PIM does not depend on a particular unicast routing protocol for its operation. Also, it does not have any mechanism of its own for route discovery. PIM operation is based on the routing table being populated by another routing protocol, or by the user. This provides flexibility in routing unicast and multicast traffic based on a common database.

PIM has two flavors, sparse and dense mode, that are deployed in different topologies. These two flavors, called PIM-SM and PIM-DM, are different in operation. PIM-SM is based on a "join protocol", where traffic is not forwarded on a segment unless an explicit request originates (typically through *IGMP*) from the network segment. PIM-DM is based on a "flood and prune" mechanism, where every one receives traffic until they explicitly inform (through the PIM-DM prune mechanism) that they do not want that particular stream. Thus, PIM-DM is typically deployed in topologies where listeners are densely populated. And PIM-SM is typically deployed where the receivers are sparsely populated over the network, so that most of the network segments' bandwidth is conserved.

You can configure dense mode or sparse mode on a per-interface basis. After they are enabled, some interfaces can run dense mode, while others run sparse mode. The switch supports both dense mode and sparse mode operation.

The switch also supports PIM snooping.

## PIM Edge Mode

PIM Edge Mode is a subset of PIM that operates with the following restrictions:

- The switch does not act as a candidate rendezvous point (CRP).
- The switch does not act as a candidate bootstrap router (CBSR).
- At most, four active PIM-SM interfaces are permitted. There is no restriction on the number of passive interfaces (within the limit of the maximum IP interfaces).
- Only PIM Sparse Mode (PIM-SM) is supported.

Note

This feature is supported at and above the license level listed for this feature in the license tables in the *Feature License Requirements* document.

Active PIM interfaces can have other PIM enabled routers on them. Passive interfaces should only have hosts sourcing or receiving multicast traffic. If another PIM router is connected to a multi-access *VLAN* then passive mode should not be enabled for that respective VLAN. *OSPF* passive mode should not be enabled for a VLAN when a PIM neighbor is present.

## PIM Dense Mode

Protocol-Independent Multicast - Dense Mode (PIM-DM) is a multicast routing protocol. PIM dense-mode is a flood and prune-based protocol. Convergence is based on the downstream routers' response for the traffic received. The downstream router in turn floods the traffic to its own downstream interfaces. Each router sends prune to the interface on which it received the traffic under the following conditions:

- Traffic was not received on RPF interface towards the source.
- The PIM router is a leaf router, and there are no *IGMP*/MLD members.
- All the downstream PIM routers have pruned the stream, and there are no IGMP/MLD members.

A new feature, called PIM-DM state refresh, creates two PIM-DM operating modes, which are described in the following sections:

**Note**
For additional information on PIM-DM, see RFC 3973, Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification.

### *PIM-DM Without State Refresh*

PIM-DM is a broadcast and prune protocol, which means that multicast servers initially broadcast traffic to all destinations, and then switches later prune paths on which there are no receivers. The following figure shows a dense mode multicast tree with an active branch and a pruned branch.

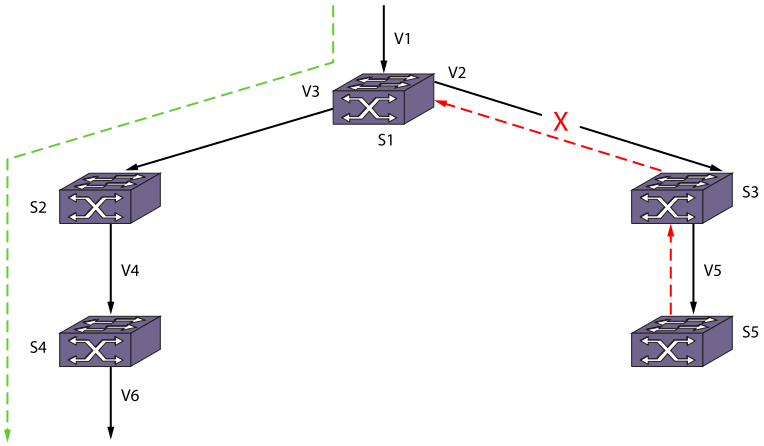



**Figure 241: PIM-DM Operation**

In the previous figure, multicast traffic is flowing from *VLAN* V1 connected to switch S1. S1 floods multicast traffic to both neighbors S2 and S3 which in turn flood multicast traffic to S4 and S5. S4 has *IGMP* members, so it floods multicast traffic down to VLAN V6. S5, which has no multicast members, sends a prune upstream towards the source. The green line shows the flow of traffic on the active branch, and the red line shows the prune sent upstream for the pruned branch. After outgoing interface V2 is pruned from the multicast tree, subsequent multicast traffic from S1 flows only through S2 and S4 and is not forwarded to S3.

After S3 sends a prune upstream, S3 starts a prune hold time timer on outgoing interface V5. When this timer expires, S3 adds V5 back to the multicast egress list and sends a graft upstream to pull multicast traffic down again. When multicast traffic arrives from S1, it is forwarded to S5, which repeats the upstream prune message because it still has no members. This prune, time-out, and flood process

repeats as long as the traffic flow exists and no members are on the pruned branch, and this process consumes bandwidth during every cycle.

Note
This feature is supported at and above the license level listed for this feature in the license tables in the *Feature License Requirements* document.

PIM-DM routers perform reverse path multicasting (RPM). However, instead of exchanging its own unicast route tables for the RPM algorithm, PIM-DM uses the existing unicast routing table for the reverse path. As a result, PIM-DM requires less system memory.

### *PIM-DM with State Refresh*

The PIM-DM State Refresh feature keeps the PIM-DM prune state from timing out by periodically sending a state refresh control message down the source tree. These control messages reset the prune hold time timer on each pruned interface and prevent the bandwidth waste that occurs with each prune, time-out, and flood cycle.

When a topology change occurs, the PIM-DM State Refresh feature improves network convergence. For example, suppose that an S, G entry on S5 in the following figure is removed due to non-availability of a route. Without PIM-DM State Refresh, multicast traffic is blocked for minutes (due to a time-out on the upstream routers). In the meantime if an *IGMP* member or a PIM-DM neighbor joins S5, there is no way to pull traffic down immediately because S5 does not have any S, G information. State refresh control messages solve this problem by indicating S, G state information periodically to all downstream routers. When S5 receives a state refresh from S3, it scans the S, G information and all pending requests from PIM-DM neighbors and IGMP members. If there are pending requests for the group in the state refresh message, S5 can immediately send a graft message upstream to circumvent the upstream timers and pull multicast traffic to its members and neighbors.

- To enable, configure, and disable the PIM-DM State Refresh feature, use the following commands:

```
configure pim state-refresh {vlan} [vlanname | all] [on | off]
configure pim state-refresh timer origination-interval interval
configure pim state-refresh timer source-active-timer interval
configure pim state-refresh ttl ttlvalue
```

## PIM Sparse Mode

Unlike PIM-DM, Protocol-Independent Multicast - Sparse Mode (PIM-SM) is an explicit join protocol, which means that multicast receivers, and the routers that support them, must join multicast groups before they receive multicast traffic. When all receivers on a network branch leave a multicast group, that branch is pruned so that the multicast traffic does not continue to consume bandwidth on that branch. PIM-SM supports shared trees as well as shortest path trees (SPTs). PIM-SM is beneficial for large networks that have group members that are sparsely distributed.

Note
This feature is supported at and above the license level listed for this feature in the license tables in the *Feature License Requirements* document.

Using PIM-SM, the router sends a join message to the rendezvous point (RP). The RP is a central multicast router that is responsible for receiving and distributing the initial multicast packets. You can configure a dynamic or static RP.

When a router has a multicast packet to distribute, it encapsulates the packet in a unicast message and sends it to the RP. The RP decapsulates the multicast packet and distributes it among all member routers.

When a router determines that the multicast rate has exceeded a configured threshold, that router can send an explicit join to the originating router. When this occurs, the receiving router gets the multicast directly from the sending router and bypasses the RP.

Note
You can run either PIM-DM or PIM-SM per virtual LAN (*VLAN*).

## PIM Mode Interoperation

An Extreme Networks switch can function as a PIM multicast border router (PMBR). A PMBR integrates PIM-SM and PIM-DM traffic.

When forwarding PIM-DM traffic into a PIM-SM network, the PMBR acts as a virtual first hop and encapsulates the initial traffic for the RP. The PMBR forwards PIM-DM multicast packets to the RP, which, in turn, forwards the packets to those routers that have joined the multicast group.

The PMBR also forwards PIM-SM traffic to a PIM-DM network, based on the (*.*.RP) entry. The PMBR sends a (*.*.RP) join message to the RP, and the PMBR forwards traffic from the RP into the PIM-DM network.

No commands are required to enable PIM mode interoperation. PIM mode interoperation is automatically enabled when a dense mode interface and a sparse mode interface are enabled on the same switch.

## PIM Source Specific Multicast

PIM-SM works well in many-to-many multicasting situations. For example, in video conferencing, each participating site multicasts a stream that is sent to all the other participating sites. However, PIM-SM is overly complex for one-to-many multicast situations, such as multimedia content distribution or streaming stock quotes. In these and similar applications, the listener is silent and can know the source of the multicast in advance, or can obtain it. In these situations, there is no need to join an RP, as the join request can be made directly towards the source.

Note
This feature is supported at and above the license level listed for this feature in the license tables in the *Feature License Requirements* document.

Note
(*;G)s are created for groups inside the SSM range if SSM is not enabled.

PIM Source Specific Multicast (PIM-SSM) is a special case of PIM-SM, in which a host explicitly sends a request to receive a stream from a specific source, rather than from any source.

IGMPv3 hosts can use PIM SSM directly, because the ability to request a stream from a specific source first became available with IGMPv3. The PIM-SSM capable router interprets the IGMPv3 message to initiate a PIM-SM join towards the source.

Note

IGMPv1 and IGMPv2 hosts can use PIM SSM if *IGMP*-SSM mapping is enabled and configured on the ExtremeXOS switch. For more information, see Using IGMP-SSM Mapping.

The following table describes PIM-SSM behavior while sending IGMPV3 joins in the SSM range and outside the SSM range for IPv4:

**Table 150: Using PIM-SSM While Sending IGMPV3 Joins (IPv4)**

| | | | | ExtremeXOS 15.4 | | ExtremeXOS 15.5 | |
|---|---|---|---|---|---|---|---|
| SSM Enabled | SSM range | Mode | Include Src | Action | Observation | Action | Observation |
| No | Yes | Incl | Yes | Send IGMPv3 join in SSM range | -the group is learned - (*;G) is not created | Send IGMPv3 join in SSM range | -the group is learned - (*;G) is created |
| No | Yes | Incl | Yes | Send IGMPv3 out of SSM range | -the group is learned -no (*;G) is created | Send IGMPv3 out of SSM range | -the group is learned - (*;G) is created |
| No | Yes | Excl | No | Send IGMPv3 join in SSM range | -the group is not learned (PD4-31387 92131) -no (*;G) is created | Send IGMPv3 join in SSM range | -the group is not learned (PD4-31387 92131) -no (*;G) is created |
| No | Yes | Excl | No | Send IGMPv3 join out of SSM range | -the group is learned - (*;G) is created | Send IGMPv3 join out of SSM range | -the group is learned - (*;G) is created |
| No | Yes | Excl | Yes | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created |
| No | Yes | Excl | Yes | Send IGMPv3 join out SSM range | -the group is learned -no (*;G) is created | Send IGMPv3 join out SSM range | -the group is learned -no (*;G) is created |
| No | No | Incl | Yes | Send IGMPv3 join | -the group is learned -no (*;G) is created | Send IGMPv3 join | -the group is learned - (*;G) is created |
| No | No | Excl | No | Send IGMPv3 join | -the group is learnt - (*;G) is created | Send IGMPv3 join | -the group is learnt - (*;G) is created |

**Table 150: Using PIM-SSM While Sending IGMPV3 Joins (IPv4) (continued)**

| | | | | ExtremeXOS 15.4 | | ExtremeXOS 15.5 | |
|---|---|---|---|---|---|---|---|
| SSM Enabled | SSM range | Mode | Include Src | Action | Observation | Action | Observation |
| No | No | Excl | Yes | Send IGMPv3 join | -the group is learned -no (*;G) is created | Send IGMPv3 join | -the group is learned -no (*;G) is created |
| Yes | Yes | Incl | Yes | Send IGMPv3 join in SSM range | -the group is learned -(S;G) is created | Send IGMPv3 join in SSM range | -the group is learned -(S;G) is created |
| Yes | Yes | Incl | Yes | Send IGMPv3 out of SSM range | -the group is learned -no (*;G) is created | Send IGMPv3 out of SSM range | -the group is learned -(*;G) is created |
| Yes | Yes | Excl | No | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created |
| Yes | Yes | Excl | No | Send IGMPv3 join out of SSM range | -the group is learned -(*;G) is created | Send IGMPv3 join out of SSM range | -the group is learned -(*;G) is created |
| Yes | Yes | Excl | Yes | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created | Send IGMPv3 join in SSM range | -the group is not learned -no (*;G) is created |
| Yes | Yes | Excl | Yes | Send IGMPv3 join out SSM range | -the group is learned -no (*;G) is created | Send IGMPv3 join out SSM range | -the group is learned -no (*;G) is created |
| Yes | No | Incl | Yes | Send IGMPv3 join | -the group is learned -no (*;G) is created | Send IGMPv3 join | -the group is learned -(*;G) is created |
| Yes | No | Excl | No | Send IGMPv3 join | -the group is learned -(*;G) is created | Send IGMPv3 join | -the group is learned -(*;G) is created |
| Yes | No | Excl | Yes | Send IGMPv3 join | -the group is learned -no (*;G) is created | Send IGMPv3 join | -the group is learned -no (*;G) is created |

The following table describes PIM-SSM behavior while sending MLDV2 joins in the SSM range and outside the SSM range for IPv6:

**Table 151: Using PIM-SSM While Sending MLDV2 Joins (IPv6)**

| | | | | ExtremeXOS 15.4 | | ExtremeXOS 15.5 | |
|---|---|---|---|---|---|---|---|
| SSM Enabled | SSM range | Mode | Include Src | Action | Observation | Action | Observation |
| No | Yes | Incl | Yes | Send MLDv2 join in SSM range | -the group is learned - no (*;G) is created | Send MLDv2 join in SSM range | -the group is learned - (*;G) is created |
| No | Yes | Incl | Yes | Send MLDv2 out of SSM range | -the group is learned - no (*;G) is created - (S;G) is created | Send MLDv2 out of SSM range | -the group is learned - (*;G) is created |
| No | Yes | Excl | No | Send MLDv2 join in SSM range | -the group is learned - (*;G) is created | Send MLDv2 join in SSM range | -the group is not learned - no (*;G) is created |
| No | Yes | Excl | No | Send MLDv2 join out of SSM range | -the group is learned - (*;G) is created | Send MLDv2 join out of SSM range | -the group is learned - (*;G) is created |
| No | Yes | Excl | Yes | Send MLDv2 join in SSM range | -the group is learned - (*;G) is created | Send MLDv2 join in SSM range | -the group is not learned - no (*;G) is created |
| No | Yes | Excl | Yes | Send MLDv2 join out SSM range | -the group is learned - (*;G) is created | Send MLDv2 join out SSM range | -the group is learned - (*;G) is created |
| No | No | Incl | Yes | Send MLDv2 join | -the group is learned - (S;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |
| No | No | Excl | No | Send MLDv2 join | -the group is learned - (*;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |
| No | No | Excl | Yes | Send MLDv2 join | -the group is learned - (*;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |
| Yes | Yes | Incl | Yes | Send MLDv2 join in SSM range | -the group is learned - (S;G) is created | Send MLDv2 join in SSM range | -the group is learned - (S;G) is created |

**Table 151: Using PIM-SSM While Sending MLDV2 Joins (IPv6) (continued)**

| | | | | ExtremeXOS 15.4 | | ExtremeXOS 15.5 | |
|---|---|---|---|---|---|---|---|
| SSM Enabled | SSM range | Mode | Include Src | Action | Observation | Action | Observation |
| Yes | Yes | Incl | Yes | Send MLDv2 out of SSM range | -the group is learned - no (*;G) is created - (S;G) is created | Send MLDv2 out of SSM range | -the group is learned - (*;G) is created |
| Yes | Yes | Excl | No | Send MLDv2 join in SSM range | -the group is learned - (*;G) is created | Send MLDv2 join in SSM range | -the group is not learned - no (*;G) is created |
| Yes | Yes | Excl | No | Send MLDv2 join out of SSM range | -the group is learned - (*;G) is created | Send MLDv2 join out of SSM range | -the group is learned - (*;G) is created |
| Yes | Yes | Excl | Yes | Send MLDv2 join in SSM range | -the group is learned - (*;G) is created | Send MLDv2 join in SSM range | -the group is not learned - no (*;G) is created |
| Yes | Yes | Excl | Yes | Send MLDv2 join out SSM range | -the group is learned - (*;G) is created | Send MLDv2 join out SSM range | -the group is learned - (*;G) is created |
| Yes | No | Incl | Yes | Send MLDv2 join | -the group is learned - (S;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |
| Yes | No | Excl | No | Send MLDv2 join | -the group is learned - (*;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |
| Yes | No | Excl | Yes | Send MLDv2 join | -the group is learned - (*;G) is created | Send MLDv2 join | -the group is learned - (*;G) is created |

PIM-SSM has the following advantages:

- No overhead of switching to the source-specific tree and waiting for the first packet to arrive
- No need to learn and maintain an RP
- Fewer states to maintain on each router
- No need for the complex register mechanism from the source to the RP
- Better security, as each stream is forwarded from sources known in advance

PIM-SSM has the following requirements:

- Any host that participates directly in PIM-SSM must use IGMPv3. For PIM IPv6 host must use MLDv2.
- To support IGMPv1 and IGMPv2 hosts, IGMP-SSM mapping must be enabled and configured. To support MLDv1 hosts, MLD-SSM mapping must be enabled and configured.

PIM-SSM is designed as a subset of PIM-SM and all messages are compliant with PIM-SM. PIM-SSM and PIM-SM can coexist in a PIM network; only the last hop router need to be configured for PIM-SSM if both source and receivers are present all the time. However, to avoid any JOIN delay, it is recommended that you enable all routers along the (s,g) path for PIM-SSM.

### *Configuring the PIM-SSM Address Range*

A range of multicast addresses is used for PIM-SSM. Within that address range, non-IGMPv3 messages are ignored, and any IGMPv3 exclude messages are ignored. These messages are ignored for all router interfaces, even those not configured for PIM-SSM. By default there is no PIM-SSM range specified on the router. If you choose the default keyword in the CLI when specifying the PIM-SSM range, you configure the range 232.0.0.0/8 and default PIM-SSM range for IPv6 is FF3x::/96. You can also choose to specify a different range for PIM-SSM by using a policy file.

To configure the PIM-SSM address range, use the following command:

```
configure pim ssm range [default | policy policy-name]
```

## PIM Snooping

PIM snooping provides a solution for handling multicast traffic on a shared media network more efficiently. In networks where routers are connected to a L2 switch, multicast traffic is essentially treated as broadcast traffic (see the following figure).

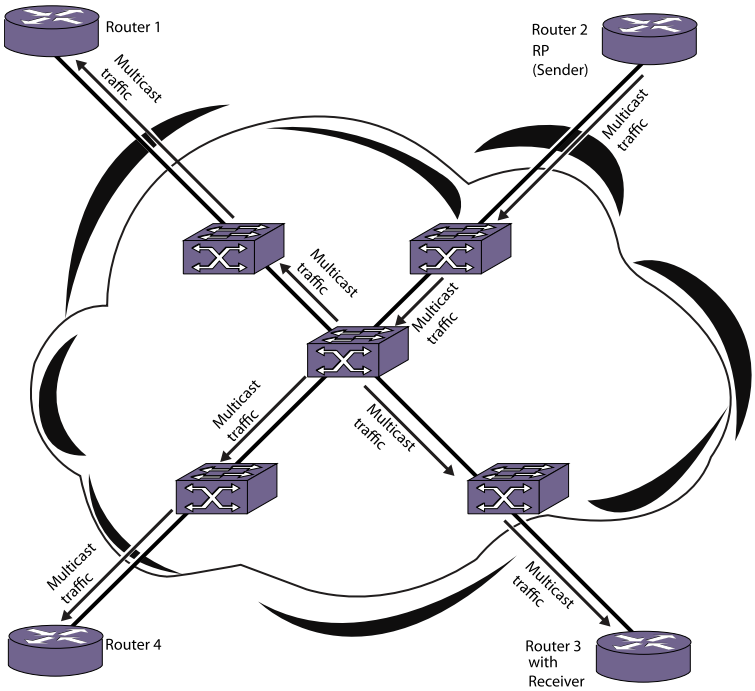



**Figure 242: Multicast Without PIM Snooping**

*IGMP* snooping does not solve this flooding issue when routers are connected to a L2 switch. Switch ports are flooded with multicast packets. PIM snooping addresses this flooding behavior by efficiently replicating multicast traffic only onto ports which routers advertise the PIM join requests (see the following figure).

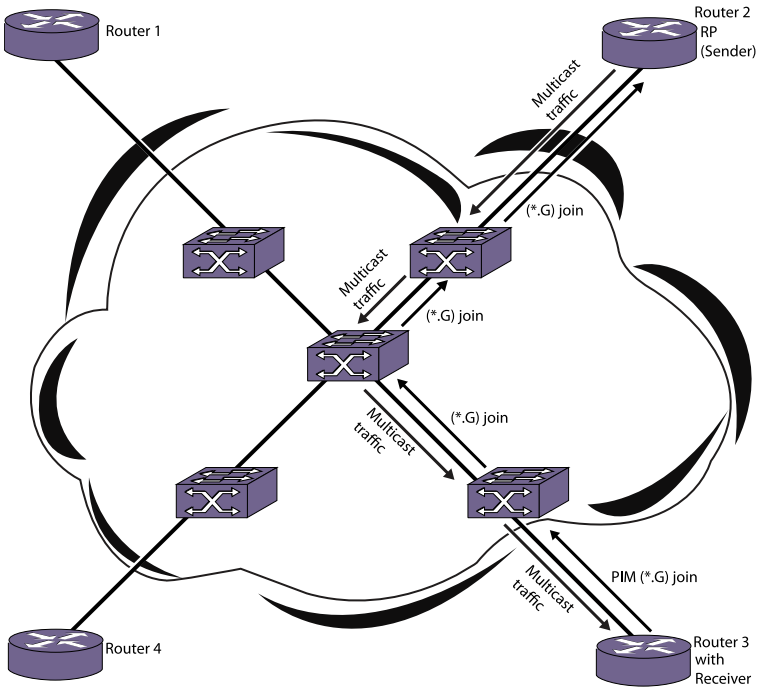



**Figure 243: Multicast With PIM Snooping**

PIM snooping does not require PIM to be enabled. However, IGMP snooping must be disabled on *VLAN*s that use PIM snooping. PIM snooping and MVR cannot be enabled on the same VLAN.

To enable PIM snooping on one or all VLANs, use the following command:

```
enable pim snooping {{vlan} name}
```

To disable PIM snooping on one or all VLANs, use the following command:

```
disable pim snooping {{vlan} name}
```

Note

PIM snooping can be enabled only between PIM SM enabled switches. It should not enabled between PIM DM enabled switches.

## PIM Register Policy

This feature allows you to filter register messages based on the policy file configured at the First Hop Router (FHR) and Rendezvous Point (RP) in PIM-SM domain. You can use the register policy to filter out specific PIM register messages that have encapsulated specific (S,G) packets. This feature allows you to detect and deny malicious multicast packets from flowing into a multicast shared tree, and causing a potential service blackout. The PIM Register Policy feature is supported in both the PIM IPV4 and PIM IPV6 mode .

*Filtering at FHR*

- FHR receives the source multicast packet and sends a register message towards the RP. Before it sends the register message to the RP, the FHR checks the configured register filter policy. If the (S,G) is denied by the policy, the register will not send a message to the RP. The FHR adds the L3 entries to stop the packet from arriving at the CPU. An *EMS (Event Management System)* message is logged.
- The FHR checks the register policy before generating a NULL register packet. If the policy is denied by the filter then the NULL register is not sent to the RP.
- If the cache's Group is in the SSM range, or is received in the PIM dense circuit, then this filtering is not applicable. The cache miss packet will go thru the normal processing.
- If a non-SSM (S,G) cache already exists but is denied by the filter policy, then (S,G) cache is removed. The cache miss comes to the CPU for register processing if the traffic is still flowing.

The PIM filtering policy is configured at the FHR using the `configure pim {ipv4 | ipv6} register-policy [policy | none]` command.

*Filtering at RP*

- When an encapsulated PIM register packet or PIM NULL register is received by the RP, and is denied by the registering filter policy, the register message is discarded. Additionally, no (S,G) cache is created in the PIM cache.
- The register drop counter is incremented, and the EMS message is logged.
- If a register is received from the *MSDP (Multicast Source Discovery Protocol)*, it also goes through the RP filtering policy.

The PIM filtering policy is configured at RP using the following command:

```
configure pim {ipv4 | ipv6} register-policy rp [rp_policy_name | none]
```

## PIM DR Priority

The DR_Priority option allows a network administrator to give preference to a particular router in the DR election process by giving it a numerically larger DR Priority. The DR_Priority option is included in every Hello message, even if no DR Priority is explicitly configured on that interface. This is necessary because priority-based DR election is only enabled when all neighbors on an interface advertise that they are capable of using the DR_Priority Option. The default priority is 1.

DR Priority is a 32-bit unsigned number, and the numerically larger priority is always preferred. A router's idea of the current DR on an interface can change when a PIM Hello message is received, when a neighbor times out, or when a router's own DR Priority changes. If the router becomes the DR or ceases to be the DR, this will normally cause the DR Register state machine to change state. Subsequent actions are determined by that state machine.

The DR election process on the interface consists of the following:

- If any one of the neighbors on the interface is not advertised, the DR priority (not DR capable) will not considered for the all the neighbors in the circuit and the primary IP address will be considered for all the neighbors.
- Higher DR priority or higher primary address will be elected as DR.

Use the following command to configure PIM DR Priority:

```
configure pim {ipv4 | ipv6} {vlan} [vlan_name] dr-priority priority
```

## PIM ECMP Load Splitting

The PIM *ECMP* feature allows downstream PIM routers to choose multiple ECMP paths to source via hash from one of following selections without affecting existing unicast routing algorithm:

- Source
- Group
- Source-Group
- Source-Group-Next Hop

This feature operates on a per (S,G) basis splitting the load onto available equal-cost paths by hashing according to the selection criteria configured by the user. It does not operate by counting the flows. Load splitting need not balance the traffic on the available paths. PIM ECMP load splitting uses a hash algorithm based on the selected criteria to pick up the path to use and will result in load-sharing the traffic when there are many multicast streams that utilize approximately the same amount of bandwidth.

### *PIM ECMP Load Splitting Based on Source Address*

When you enable PIM ECMP load splitting based on source address, the RPF interface for each (*, G) or (S,G) state is selected among the equal cost paths based on the hash derived from the source address. For an (S, G) state, the address considered for hashing is the source address of the state. For a (*, G) state, the address considered for hashing is the address of the RP that is associated with the state's group address. There is no randomization applied when calculating the hash value. The same hash value is generated on all the EXOS routers for a given source address. If there are two equal cost paths ("left" and "right") available at the last hop router and at each of the intermediate routers for a given source, each of these routers pick the same hash, and the traffic flows can get skewed (to either "left" or "right" paths).

### *PIM ECMP Load Splitting Based on Group Address*

When you enable PIM ECMP load splitting based on group address, the RPF interface for each (*, G) or (S,G) state is selected among the equal cost paths based on the hash derived from the group address. If multiple equal cost common paths exist to the multicast source and the RP that is associated with the state's group address, the same hash will be chosen for both (*, G) and (S, G) states as the same group address is used in deriving the hash. There is no randomization applied when calculating the hash value. The same hash value is generated on all the EXOS routers for a given group address. If there are two equal cost paths ("left" and "right") available at the last hop router and at each of the intermediate routers for a given group, each of these routers pick the same hash and the traffic flows can get skewed (to either "left" or "right" paths).

### *PIM ECMP Load Splitting Based on Source-Group Addresses*

When you enable PIM ECMP load splitting based on source-group address, the RPF interface for each (*, G) or (S,G) state is selected among the equal cost paths based on the hash derived from the source and group addresses. For an (S, G) state, the address considered for hashing is the source address of the state. For a (*, G) state, the address considered for hashing is the address of the RP that is associated with the state's group address. There is no randomization applied when calculating the hash

value. The same hash value is generated on all the EXOS routers for a given source-group address. If there are two equal cost paths ("left" and "right") available at the last hop router and at each of the intermediate routers for a given source-group, each of these routers pick the same hash and the traffic flows can get skewed (to either "left" or "right" paths).

*PIM ECMP Load Splitting Based on Source-Group-Next Hop Addresses*

When you enable PIM ECMP load splitting based on source-group-next hop address, the RPF interface for each (*, G) or (S,G) state is selected among the equal cost paths based on the hash derived from the source, group and next hop addresses. The hash value derived after introducing the next hop address is still predictable as there is no randomization applied when calculating the hash value. However, since the next hop address used at each of the routers vary, the hash value generated on each of the EXOS routers is different. As the hash value is different on each of the routers, the problem of traffic path skew present in the above mentioned schemes does not exist in this scheme.

*Reconvergence Due to Unicast Routing Changes*

When a unicast route to a source or RP address changes (when a path goes down or a new path becomes available), all the (*, G) and (S, G) states change based on the available unicast route information provided by Route Manager process. If one of the paths goes down and comes back up, multicast forwarding will reconverge to same RPF path that was used before the path went down. The hash function based on Source-Group-Next Hop avoids skewing of traffic flows because it introduces the actual next-hop IP address of PIM neighbors into the calculation resulting in different hash value being computed for each router. The Source-Group-Next Hop based hash function doesn't take the total number of available paths into consideration and so it increases stability of the paths chosen during path failures. During path failures, the multicast states that were using the failed path would need to reconverge onto the remaining paths. All other states using the unaffected paths are not affected.

*Limitations*

- Cannot be used along with static multicast routes.
- Not supported for ExtremeXOS Multicast Tools (mtrace and mrinfo) in current release.
- Load splitting is not applied for configured static multicast routes and multicast routes present in the multicast routing table.
- Load splitting is only effective when the equal cost paths are upstream PIM neighbors on different interfaces. When the equal cost paths are PIM neighbors on the same shared *VLAN*, PIM assert mechanism chooses one path to avoid traffic duplication. The path chosen by PIM assert mechanism overrides the path selected by Multicast ECMP load splitting.

## IPv6 Specific Features

Apart from adding support for IPv6 addresses, PIMv6 adds the following functionality to existing PIM implementation:

- Secondary address list - This is a new option which will be added to the V6 Hello messages sent. The list includes all addresses assigned to an interface, including the link local addresses. The receiving router must process these addresses and must associate the same with the neighbor that sent the message.
- Tunnel interface - This is similar to a *VLAN* interface. APIs are now added to get callbacks from VLAN manager client for IP address configuration for a tunnel, etc.

*Secondary address list*

The Address List Option, in a Hello message, advertises all the secondary addresses associated with the source interface of the router originating the message. These addresses are associated with the neighbor, and are used to compute the neighbor's primary address. The function NBR uses information gathered through PIM Hello messages to map the IP address A of a directly connected PIM neighbor on interface I to the primary IP address of the same router. The primary IP address of a neighbor is the address that it uses as the source of its PIM Hello messages.

*Tunnel interfaces*

Two PIMv6 domains can be connected through an IPv4 network. In this case, PIMv6 routers across the domains communicate over the IPv4 network by tunneling the IPv6 packets inside IPv4 headers. To enable such communication, PIMv6 provides support for Tunnel interfaces.

The following tunnel types are supported:

- 6-in-4
- 6-to-4

*Configuration details*

PIMv6 is incorporated into all CLI commands that currently support the PIM implementation. New keywords are added to support IPv6, and show command output is modified to display IPv6 related information. For specific configuration details, refer to the *ExtremeXOS 16.2 Command Reference Guide*.

## IGMP Overview

*IGMP* is a protocol used by an IP host to register its IP multicast group membership with a router. A host that intends to receive multicast packets destined for a particular multicast address registers as a member of that multicast address group. Periodically, the router queries the multicast group to see if the group is still in use. If the group is still active, a single IP host responds to the query, and group registration is maintained.

IGMPv2 is enabled by default on the switch, and the ExtremeXOS software supports IGMPv3. However, the switch can be configured to disable the generation of periodic IGMP query packets. IGMP should be enabled when the switch is configured to perform IP multicast routing.

IETF standards require that a router accept and process IGMPv2 and IGMPv3 packets only when the router-alert option is set in received IGMP packets.

By default, the ExtremeXOS software receives and processes all IGMP packets, regardless of the setting of the router-alert option within a packet. When the switch will be used with third-party switches that expect IETF compliant behavior, use the following command to manage this feature:

```
configure igmp router-alert receive-required [on | off] {{vlan}
vlan_name}

configure igmp router-alert transmit [on | off] {{vlan} vlan_name}
```

By default, IGMP report/leave message for the local multicast address (224.0.0.x/24 groups) will always have the router-alert option set, regardless of IGMP router-alert transmit option (on and off) setting by the user.

IGMPv3, specified in RFC 3376, adds support for source filtering. Source filtering is the ability for a system to report interest in receiving packets only from specific source addresses (filter mode include) or from all sources except for specific addresses (filter mode exclude). IGMPv3 is designed to be interoperable with IGMPv1 and IGMPv2.

Note
The ExtremeXOS software supports IGMPv3 source include mode filtering, but it does not support IGMPv3 specific source exclude mode filtering.

Note
It is not possible for the BlackDiamond X8 and Summit X670 series switches to have *ICMP (Internet Control Message Protocol)*/IGMP code and type fields on egress. ICMP/IGMP type requires UDF (user defined fields). Ingress Pipeline has UDF but Egress pipeline hardware does not have UDF. So it cannot match ICMP/IGMP types on egress pipeline.

## IGMP Snooping

*IGMP* snooping is a Layer 2 function of the switch; it does not require multicast routing to be enabled. In IGMP snooping, the Layer 2 switch keeps track of IGMP reports and only forwards multicast traffic to that part of the local network that requires it. IGMP snooping optimizes the use of network bandwidth and prevents multicast traffic from being flooded to parts of the local network that do not need it. The switch does not reduce any IP multicast traffic in the local multicast domain (224.0.0.x).

IGMP snooping is enabled by default on all *VLAN*s and VMANs in the switch. If IGMP snooping is disabled on a VLAN or VMAN, all IGMP and IP multicast traffic floods within the VLAN or VMAN. IGMP snooping expects at least one device on every VLAN to periodically generate IGMP query messages.

To enable or disable IGMP snooping, use the following command:

```
enable igmp snooping {forward-mcrouter-only | {vlan} name | with-proxy
vr vrname}
```

```
disable igmp snooping {forward-mcrouter-only | {vlan} name | with-proxy
vr vrname}
```

Note
IGMP snooping is not supported on SVLANs on any platform.

The IGMP snooping proxy feature represented by "with-proxy" in the above commands is enabled by default. This feature optimizes the forwarding of IGMPv1 and IGMPv2 reports. The following is true for each group:

- Only the first received IGMP join is forwarded upstream.
- Only the IGMP leave for last host is forwarded upstream.

When a switch receives an IGMP leave message on a port, it sends a group-specific query on that port if proxy is enabled (even if it is a non-querier). The switch removes the port from the group after leave timeout (The timeout value is configurable, with a default value of 1000 ms., and a range from 0 to 175000 ms). If all the ports are removed from the group, the group is deleted and the IGMP leave is

forwarded upstream. If IGMP snooping proxy is disabled, then all the IGMP reports are forwarded upstream.

Note
IGMP snooping proxy does not apply to IGMPv3 reports.

IGMP snooping is implemented primarily through *ACL (Access Control List)*s, which are processed on the interfaces. These special purpose ACLs are called IGMP snooping hardware filters. On Summit family switches and BlackDiamond 8800 series switches, the software allows you to choose between two types of IGMP snooping hardware filters: per-port filters (the default) and per-VLAN filters.

The two types of IGMP snooping hardware filters use switch hardware resources in different ways. The two primary hardware resources to consider when selecting the IGMP snooping hardware filters are the Layer 3 multicast forwarding table, and the interface ACLs. The size of both of these hardware resources is determined by the switch model. In general, the per-port filters consume more resources from the multicast table and less resources from the ACL table. The per-VLAN filters consume less space from the multicast table, and more from the ACL table.

In Summit family switches and BlackDiamond 8800 series switches, using the per-port filters can fill up the multicast table and place an extra load on the CPU. To avoid this, configure the switch to use the per-VLAN filters.

Note
The impact of the per-VLAN filters on the ACL table increases with the number of VLANs configured on the switch. If you have a large number of configured VLANs, we suggest that you use the per-port filters.

## IGMP Snooping Filters

*IGMP* snooping filters allow you to configure a policy file on a port to allow or deny IGMP report and leave packets coming into the port. (For details on creating policy files, see Policy Manager on page 635.) The IGMP snooping filter feature is supported by IGMPv2 and IGMPv3.

Note
Do not confuse IGMP snooping filters with IGMP snooping hardware filters explained in previous section. IGMP snooping filters are software filters, and the action is applied at the software level by the ExtremeXOS multicast manager.

For the policies used as IGMP snooping filters, all the entries should be IP address type entries, and the IP address of each entry must be in the class-D multicast address space but should not be in the multicast control subnet range (224.0.0.x/24).

1. Use the following template to create a snooping filter policy file that denies IGMP report and leave packets for the 239.11.0.0/16 and 239.10.10.4/32 multicast groups:

```
#
# Add your group addresses between "Start" and "end"
# Do not touch the rest of the file!!!!
entry igmpFilter {
if match any {
#------------------ Start of group addresses ------------------
nlri  239.11.0.0/16;
nlri  239.10.10.4/32;
```

```
#------------------- end of group addresses -------------------
} then {
deny;
}
}
entry catch_all {
if {
} then {
permit;
}
}
```

2. After you create a policy file, use the following command to associate the policy file and filter to a set of ports:

   `configure igmp snooping vlan` *vlanname* **ports** *portlist* **filter** [*policy* | **none**]

3. To remove the filter, use the **none** option.
4. To display the IGMP snooping filters, use the following command:

   `show igmp snooping {`**vlan**`}` *name* **filter**

## Static IGMP

To receive multicast traffic, a host must explicitly join a multicast group by sending an *IGMP* report; then, the traffic is forwarded to that host. In some situations, you might like multicast traffic to be forwarded to a port where a multicast-enabled host is not available (for example, when you test multicast configurations).

Static IGMP emulates a host or router attached to a switch port, so that multicast traffic is forwarded to that port, and the switch sends a proxy join for all the statically configured IGMP groups when an IGMP query is received. You can emulate a host to forward a particular multicast group to a port; and you may emulate a router to forward all multicast groups to a port. Static IGMP is only available with IGMPv2.

- To emulate a host on a port, use the following command:

  `configure igmp snooping {`**vlan**`}` *vlan_name* `{`**ports** *port_list*`}`**add static group** *ip_address*

- To emulate a multicast router on a port, use the following command:

  `configure igmp snooping {`**vlan**`}` *vlan_name* `{`**ports** *port_list*`}`**add static router**

- To remove these entries, use the corresponding commands:

  `configure igmp snooping {`**vlan**`}` *vlan_name* `{`**ports** *port_list* `}`**delete static group** [*ip_address* | **all**]

  `configure igmp snooping vlan` *vlan_name* `{`**ports** *port_list* `}` **delete static router**

- To display the IGMP snooping static groups, use the following command:

  `show igmp snooping {`**vlan**`}` *vlan_name* **static** [**group** | **router**]

## IGMP Loopback

Prior to ExtremeXOS 15.3.2, you could configure static groups, but it was necessary to specify port(s). As of ExtremeXOS Release 15.3.2, the configuration of dynamic groups is supported. The *IGMP*

Loopback feature, along with the existing static group feature, supports the configuration of static and/or dynamic groups with or without ports.

A *VLAN* in loopback mode may not have ports associated with it, but its operational status is up. However, it is not possible to have multicast receivers on a VLAN without having a port. Sometimes there is a need to pull the multicast traffic from upstream to the loopback VLAN for troubleshooting. The traffic need not always be forwarded to any ports/receivers. The IGMP Loopback feature allows you to configure groups on a VLAN without specifying a port, so the traffic is pulled from upstream but not forwarded to any port.

The loopback (Lpbk) port is the logical port on a VLAN in the application context. If you configure a group on a VLAN but do not specify the port, the switch forms an IGMPv2 join and assumes it to be received on the Lpbk port. A dynamic group ages out after the membership timeout if there are no other receivers. Membership joins refresh the dynamic group. The static group remains until it is removed from the configuration.

The following command is used to configure static or dynamic group entry :

```
configure igmp snooping {vlan} vlan_name {ports port_list} add [static | dynamic] group ip_address
```

## Limiting the Number of Multicast Sessions on a Port

- The default configuration places no limit on the number of multicast sessions on each *VLAN* port. To place a limit on the number of learned *IGMP* groups, use the following command:
  ```
  configure igmp snooping {vlan} vlan_name ports port_list set join-limit {num}
  ```
- To remove a session limit, use the following command:
  ```
  unconfigure igmp snooping {vlan} vlan_name ports port_list set join-limit
  ```

## Enabling and Disabling IGMP Snooping Fast Leave

When the fast leave feature is enabled and the last host leaves a multicast group, the router immediately removes the port from the multicast group. The router does not query the port for other members of the multicast group before removing the group, and/or the port.

The default setting for *IGMP* snooping fast leave is disabled.

- To enable the fast leave feature, use the following command:
  ```
  enable igmp snooping {vlan} name fast-leave
  ```
- To disable the fast leave feature, use the following command:
  ```
  disable igmp snooping {vlan} name fast-leave
  ```

## IGMP-SSM Mapping

The *IGMP*-SSM Mapping feature allows IGMPv1 and IGMPv2 hosts to participate in SSM functionality, and eliminates the need for IGMPv3. You can configure SSM map entries that specify the sources used for a group/group range for which SSM functionality has to be applied. You also have the option to configure the domain name and DNS server to use to obtain the source list.

When a IGMPv1 or IGMPv2 report is received, the configured sources are provided to PIM so that it can send source specific joins. When the host leaves or when the membership times out, PIM is informed so that it can consider sending prunes.

In a multi-access network (where more than one router is receiving IGMP messages from the hosts), only the designated router sends joins towards the source, so it is desirable to have same configuration for SSM group range and SSM Mapping range on all routers in a *VLAN*.

*Limitations*

- A single group address or range can be mapped to a maximum of 50 sources. If more than 50 sources are configured, the switch uses the 50 longest-matching prefixes.
- We recommend a maximum of 500 mappings per switch, but no limit is imposed by the software.

*Configuring IGMP-SSM Mapping*

To support PIM-SSM for IGMPv1 and IGMPv2 clients, a PIM-SSM range must be configured, and that range should include all groups to which the clients want access. If IGMPv1 and IGMPv2 clients request group addresses outside the PIM-SSM range, those addresses are ignored by PIM-SSM and forwarded to PIM as (*, G) requests.

- To enable *IGMP*-SSM mapping, first configure a PIM-SSM range, and then enable IGMP-SSM mapping using the following commands:
  `configure pim ssm range [default | policy policy-name]`
  `enable igmp ssm-map {vr vr-name}`
- To configure an IGMP-SSM mapping, use the following command:
  `configure igmp ssm-map add group_ip [prefix | mask] [source_ip | src_domain_name] {vr vr-name}`
- To remove a single IGMP-SSM mapping, use the following command:
  `configure igmp ssm-map delete group_ip [ prefix} | mask] [source_ip | all] vr vr-name }`
- To remove all IGMP-SSM mappings on a *VR*, use the following command:
  `unconfigure igmp ssm-map {vr vr-name}`
- To disable IGMP-SSM mapping on a virtual router, use the following command:
  `disable igmp ssm-map {vr vr-name}`

*Displaying IGMP-SSM Mappings*

To see whether or not *IGMP*-SSM mapping is enabled or disabled and to view the configured mappings for a multicast IP address, use the following command:

`show igmp ssm-map {group_ip} {vr vr-name}`

## Configuring EAPS Support for Multicast Traffic

The ExtremeXOS software provides several commands for configuring how EAPS supports multicast traffic after an EAPS topology change. For more information, see the descriptions for the following commands:

`configure eaps multicast add-ring-ports [on | off]`

`configure eaps multicast send-query [on | off]`

```
configure eaps multicast temporary-flooding [on | off]
```

These commands apply for both *IGMP* and MLD.

Note

Using the `configure eaps multicast send-query` command applies to both IGMP and MLD. This also replaces the `configure eaps multicast send-igmp-query` command.

# Configuring IP Multicast Routing

## Enabling Multicast Forwarding

To enable IP multicast forwarding:

1. Configure the system for IP unicast routing.
2. Enable multicast forwarding on the interface.

   ```
   enable ipmcforwarding {vlan name}
   ```

## Configuring PIM

To configure PIM multicast routing, enable multicast forwarding as described in Enabling Multicast Forwarding on page 1484 and do the following:

1. Configure PIM on all IP multicast routing interfaces using the following command:

   ```
   configure pim {ipv4 | ipv6} add vlan [vlan-name | all] {dense | sparse} {passive}
   ```

2. To enable and configure the PIM-DM state refresh feature on one or all VLANs, use the following commands:

   ```
   configure pim {ipv4 | ipv6} state-refresh {vlan} [vlan_name | all] [on | off]
   configure pim { ipv4 | ipv6 } state-refresh timer origination-interval interval
   configure pim {ipv4 | ipv6} state-refresh timer source-active-timer interval
   configure pim {ipv4 | ipv6} state-refresh ttl ttlvalue
   ```

3. For PIM-SSM, specify the PIM-SSM range, enable IGMPv3, and enable PIM-SSM on the interfaces using the following commands:

   ```
   configure pim {ipv4 | ipv6} ssm range [default | policy policy-name]
   enable igmp {vlan vlan name } {IGMPv1 | IGMPv2 | IGMPv3}
   enable pim {ipv4 | ipv6} ssm vlan [vlan_name | all]
   ```

4. Enable PIM on the router using the following command:

   ```
   enable pim
   ```

## Configuring Multicast Static Routes

Note

Multicast static routes are supported in the IPv4 address family, but not the IPv6 address family.

Static routes are used to reach networks not advertised by routers, and are manually entered into the routing table.

- You can use either of two commands to create multicast static routes. The recommended command is the following:

  `configure iproute add [ipNetmask | ip_addr mask] gateway {metric} {multicast | multicast-only | unicast | unicast-only} {vr vrname}`

  For example:

  ```
  configure iproute add 55.1.10.0/24 44.1.12.33 multicast
  ```

- The following command is still supported for backward compatibility with earlier ExtremeXOS software releases:

  `configure ipmroute add [source-net mask-len | source-net mask] {{protocol} protocol} rpf-address {metric} {vr vr-name}`

  In the following example, the configure ipmroute add command is used to specify protocol information for a route:

  ```
  configure ipmroute add 58.1.10.0/24 ospf 44.1.12.33
  ```

  When a static route is configured with protocol information, the route is shown as UP only when the protocol route is available. Otherwise, the route is Down. In the example above, the multicast static route 58.1.10.0/24 is shown as UP only when the *OSPF* route is available to reach the network 58.1.10.0/24.

  Static routes are stored in the switch configuration and can be viewed with the show configuration command. Static multicast routes that do not include protocol information are displayed using the configure iproute command format, even if they were created using the configure ipmroute command. Static routes that are created with a protocol field are displayed using the configure ipmroute command format.

## Disabling IP Multicast Compression

The IP multicast compression feature is available only on Summit family switches and BlackDiamond 8000 series modules, and is enabled by default. You should only disable this feature if you suspect that switch processing resources are limited or if you think this feature is causing problems on the switch.

- To disable or enable IP multicast compression, use the following command:

  `configure forwarding ipmc compression {group-table | off}`

## Multicast Over MLAG Configuration

Consider the following sample topology for *MLAG (Multi-switch Link Aggregation Group)*:

DUT-1(core lic)======ISC vlan========DUT-2(core lic)

| |

+----------------DUT-3(edge lic)---------------+

DUT-1 and DUT-2 are MLAG peers, DUT-3 is a L2 switch whose uplink is a *LAG (Link Aggregation Group)* up to the pair of MLAG switches.

- RP and BSR can be configured on same device along with MLAG config but we recommend to keep RP node away from MLAG peers. One MLAG peer will be Designated Router(DR) and another one will be elected as NON-DR for MLAG vlan. DR node will send *,G and try to pull the traffic from RP. Non-DR does not pull the traffic until DR is alive. If you config RP on Non-DR node then both MLAG peers will pull the traffic which will trigger the assert to avoid the traffic duplication. It is not recommended to setup RP on any vlan on MLAG peers.
- It is best to avoid the assert operation since a small amount of traffic duplication happens during the this operation. You can avoid assert in some, but not all the scenarios.
- DR priority configuration will help to make RP node as DR. (The DR priority feature is available from ExtremeXOS15.3.2 release onwards).
- It is recommended that for PIM-SM deployments that the RP is configured on loopback *VLAN*s instead of regular VLANs. This ensures continuous connectivity to the RP needing active ports present in that respective VLAN.

## PIM Configuration Examples

*PIM-DM Configuration Example*

In the following figure, the system labeled IR 1 is configured for IP multicast routing, using PIM-DM.

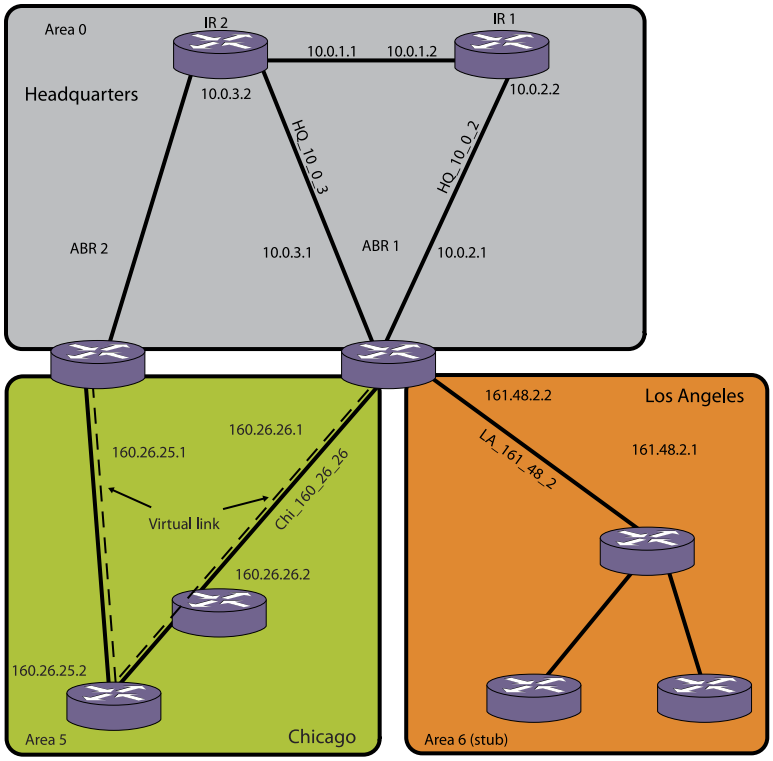



**Figure 244: IP Multicast Routing Using PIM-DM Configuration Example**

**Note**

The above figure is used in the *OSPF* chapter to describe the Open Shortest Path First (OSPF) configuration on a switch. See OSPF Overview on page 1341 for more information about configuring OSPF.

The router labeled IR1 has the following configuration:

```
configure vlan HQ_10_0_1 ipaddress 10.0.1.2 255.255.255.0
configure vlan HQ_10_0_2 ipaddress 10.0.2.2 255.255.255.0
configure ospf add vlan all area 0.0.0.0
enable ipforwarding
enable ospf
enable ipmcforwarding
configure pim add vlan all dense
enable pim
configure pim state-refresh vlan all on
```

### *PIM-SM Configuration Example*

In the following figure, the system labeled ABR1 is configured for IP multicast routing using PIM-SM.

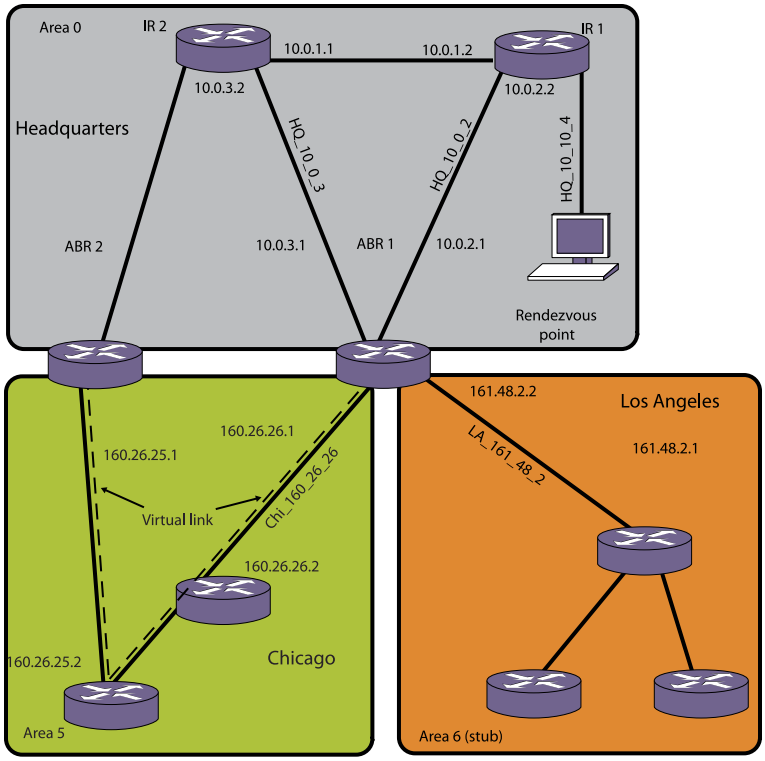



**Figure 245: IP Multicast Routing Using PIM-SM Configuration Example**

**Note**

The above figure is used in the *OSPF* section to describe the Open Shortest Path First (OSPF) configuration on a switch. See OSPF Overview on page 1341 for more information about configuring OSPF.

The router labeled ABR1 has the following configuration:

```
configure vlan HQ_10_0_2 ipaddress 10.0.2.1 255.255.255.0
configure vlan HQ_10_0_3 ipaddress 10.0.3.1 255.255.255.0
configure vlan LA_161_48_2 ipaddress 161.48.2.2 255.255.255.0
configure vlan CHI_160_26_26 ipaddress 160.26.26.1 255.255.255.0
configure ospf add vlan all area 0.0.0.0
enable ipforwarding
enable ipmcforwarding
configure pim add vlan all sparse
tftp TFTP_SERV -g -r rp_list.pol
configure pim crp HQ_10_0_3 rp_list 30
configure pim cbsr HQ_10_0_3 30
```

The policy file, rp_list.pol, contains the list of multicast group addresses serviced by this RP. This set of group addresses are advertised as candidate RPs. Each router then elects the common RP for a group address based on a common algorithm. This group to RP mapping should be consistent on all routers.

The following is a policy file that configures the CRP for the address ranges 239.0.0.0/24 and 232.144.27.0:

```
entry extreme1 {
    if match any {
    }
        then {
```

```
            nlri  239.0.0.0/24 ;
            nlri  232.144.27.0/24 ;
    }
}
```

### *PIM-SSM Configuration Example*

In the following example, the default PIM-SSM range of 232.0.0.0/8 is configured. For all interfaces, non-IGMPv3 messages and IGMPv3 exclude messages are ignored for addresses in this range. Hosts that use IGMPv3 on *VLAN* v13 can request and receive source specific multicast streams for addresses in the PIM-SSM range.

```
create vlan v12
create vlan v13
configure v12 add port 1
configure v13 add port 2
configure v12 ipaddress 12.1.1.1/24
configure v13 ipaddress 11.1.1.1/24
configure pim add vlan all sparse
enable ipforwarding
enable ipmcforwarding
enable igmp IGMPv3
configure pim ssm range default
enable pim ssm vlan v13
enable pim
```

**Note**

(*;G)s are created for groups outside the SSM range. SSM may not be enabled for the ingress vlan (see Table 150 on page 1469.

### *PIM Snooping Configuration Example*

The following figure shows a network configuration that supports PIM snooping.

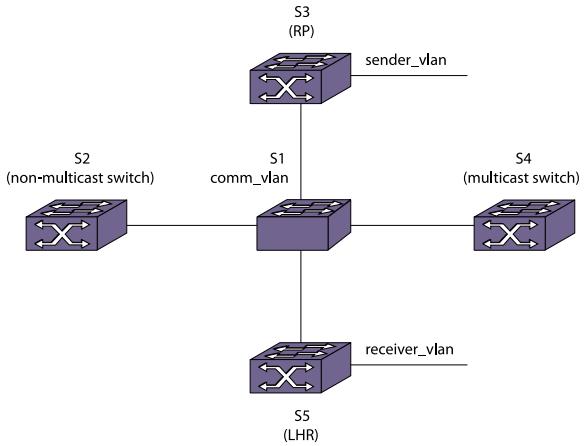



**Figure 246: PIM Snooping Configuration Example**

In the figure above, Layer 3 switches S2, S3, S4, and S5 are connected using the Layer 2 switch S1 through the *VLAN* comm_vlan. Switches S3, S4, and S5 are multicast capable switches, and switch S2 is a non-multicast capable switch, which has no multicast routing protocol running.

Without PIM snooping, any ingress multicast data traffic on comm_vlan is flooded to all the switches, including switch S2, which does not support multicast traffic. *IGMP* snooping does not reduce flooding because it floods the multicast traffic to all router ports.

The network design calls for most multicast traffic to egress switch S5. PIM snooping helps by intercepting the PIM joins received from the downstream routers and forwarding multicast traffic only to those ports that received PIM joins.

#### Switch S1 (PIM Snooping Switch) Configuration Commands

The following is an example configuration for the PIM snooping switch S1:

```
create vlan comm_vlan
configure vlan comm_vlan add port 1,2,3,4
disable igmp snooping
disable igmp_snooping comm_vlan
enable pim snooping
enable pim snooping comm_vlan
```

#### Switch S3 Configuration Commands

The following is an example configuration for switch S3, which also serves as an RP:

```
create vlan comm_vlan
configure vlan comm_vlan add port 1
configure comm_vlan ipa 10.172.168.4/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure pim add vlan comm_vlan sparse
configure ospf add vlan comm._vlan area 0.0.0.0

create vlan sender_vlan
configure vlan sender_vlan add port 2
configure sender_vlan ipa 10.172.169.4/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure pim add vlan comm._vlan sparse
configure ospf add vlan comm_vlan area 0.0.0.0

enable pim
enable ospf

configure pim crp static 10.172.169.4 pim_policy // RP is configured using the policy
pim_policy for the group 224.0.0.0/4
```

#### Switch S5 Configuration Commands

The following is an example configuration for switch S5, which serves as the last hop router for multicast traffic:

```
create vlan comm_vlan
configure vlan comm_vlan add port 1
configure comm_vlan ipa 10.172.168.2/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure pim add vlan comm_vlan sparse
configure ospf add vlan comm._vlan area 0.0.0.0

create vlan receiver_vlan
configure vlan sender_vlan add port 1
configure sender_vlan ipa 10.172.170.4/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure pim add vlan comm._vlan sparse
configure ospf add vlan comm_vlan area 0.0.0.0

enable pim
```

```
enable ospf

configure pim crp static 10.172.169.4 pim_policy // RP is configured using the policy
pim_policy for the group 224.0.0.0/4
```

#### Switch S4 Configuration Commands

The following is an example configuration for switch S4, which is neither an LHR nor a RP:

```
create vlan comm_vlan
configure vlan comm_vlan add port 1
configure comm_vlan ipa 10.172.168.3/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure pim add vlan comm_vlan sparse
configure ospf add vlan comm._vlan area 0.0.0.0

enable pim
enable ospf

configure pim crp static 10.172.169.4 pim_policy // RP is configured using the policy
pim_policy for the group 224.0.0.0/4
```

#### Switch S2 Configuration Commands

The following is an example configuration for switch S2, which is not enabled for PIM:

```
create vlan comm_vlan
configure vlan comm_vlan add port 1
configure comm_vlan ipa 10.172.168.6/24
enable ipforwarding comm_vlan
enable ipmcforwarding comm._vlan
configure ospf add vlan comm._vlan area 0.0.0.0

enable ospf
```

#### PIM Snooping Example Configuration Displays

After the example configuration is complete, multicast receivers connect to the network through switch S5 and multicast sources connect through switch S3.

When switch S5 receives an IGMP request from the receiver_vlan for group 225.1.1.1, it sends a PIM (*, G) join towards switch S3, which is the RP. The PIM snooping feature on switch S1 snoops the (*, G) join, and the resulting entry can be viewed by entering the following command at switch S1:

```
# show pim snooping vlan comm_vlan
PIM Snooping                          ENABLED
Vlan comm_vlan(3971)              Snooping ENABLED
Source          Group           RP               UpPort   DownPort   Age      HoldTime
*               225.1.1.1       10.172.169.4        1        2          15        210
Neighbor IP      DR     Port       Age         Hold Time
10.1272.168.4    YES     1          2            105
10.1272.168.2    NO      2          2            105
10.1272.168.3    NO      3          2            105
```

Once multicast traffic arrives from the sender_vlan, the LHR (switch S2) sends the (S, G) join message, which is snooped by the PIM snooping switch, switch S1. The resulting entries can be viewed by entering the following command at switch S1:

```
# show pim snooping vlan comm_vlan
PIM Snooping                          ENABLED
Vlan comm_vlan(3971)              Snooping ENABLED
```

```
Source          Group           RP              UpPort   DownPort   Age      HoldTime
*               225.1.1.1       10.172.169.4       1          2        15        210
10.172.169.10   225.1.1.1       10.172.169.4       1          2        15        210
Neighbor IP      DR     Port       Age         Hold Time
10.1272.168.4    YES     1          2           105
10.1272.168.2    NO      2          2           105
10.1272.168.3    NO      3          2           105
```

Multicast traffic is forwarded only to those ports that have received (*, G) or (S, G) joins and designated router (DR) ports.

## Multicast VLAN Registration

Multicast *VLAN* Registration (MVR) is designed to support distributing multicast streams for IPTV to subscribers over a Layer 2 network. In a standard Layer 2 network, a multicast stream received on a VLAN is not forwarded to another VLAN. The streams are confined to the Layer 2 broadcast domain. In an *IGMP* snooping environment, streams are forwarded only to interested hosts on a VLAN. For inter-VLAN forwarding (routing) a multicast routing protocol, such as PIM/DVMRP must be deployed.

MVR breaks this basic rule, so that a stream received over Layer 2 VLANs is forwarded to another VLAN, eliminating the need for a Layer 3 routing protocol. It simplifies the multicast stream distribution and is a better solution for IPTV-like services. With MVR, a multicast stream is forwarded to all VLANs containing interested hosts.

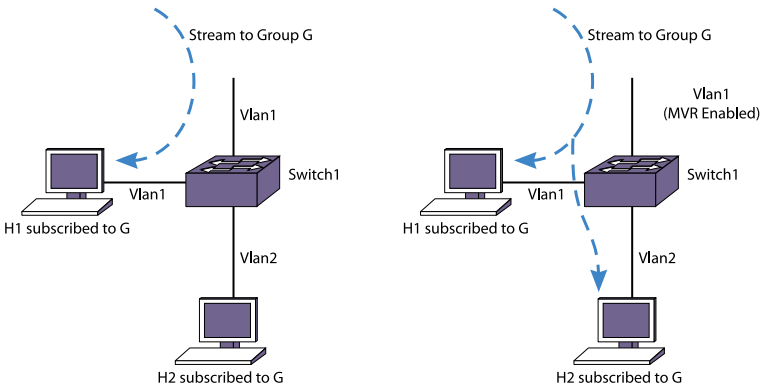



**Figure 247: Standard VLAN Compared to an MVR VLAN**

In the above figure, the left side shows a standard VLAN carrying a multicast stream.

The host H1 receives the multicast stream because it resides on VLAN Vlan1, but host H2 does not receive the multicast traffic because no IP multicast routing protocol forwards the stream to VLAN Vlan2. On the right side of the figure, H2 does receive the multicast stream. Because Vlan1 was configured as an MVR VLAN, the multicast stream is forwarded to the other VLANs on the switch. containing hosts that have requested the stream. To configure a VLAN as an MVR VLAN, use the following command: `configure mvr add vlan vlan-name`

Typically, IGMP snooping is enabled, so only hosts that have requested a stream can see the multicast traffic. For example, another host on VLAN2 cannot receive the traffic unless it has sent an IGMP request to be included in the group.

Notice that only VLAN1 is MVR enabled. Configure MVR only on the ingress VLAN. To enable MVR on the switch, use the following command: `enable mvr`

**Note**
MVR is not supported on the Summit X430.

## Basic MVR Deployment

Because MVR is primarily targeted for IPTV and similar applications, a basic deployment for that application is shown in the following figure. In the figure, an IPTV server is connected through a router to a network of switches. Switch 1 has three customer *VLAN*s, Vlan2, Vlan3, and Vlan4. The multicast streams are delivered through the network core (Metro Ethernets), which often use a ring topology and some kind of redundant protection to provide high availability. For example, McastVlan forms a ring through switches Switch1 through Switch4. The link from Switch2 to Switch4 is shown as blocked, as it would be if some form of protection (such as EAPS) is used.

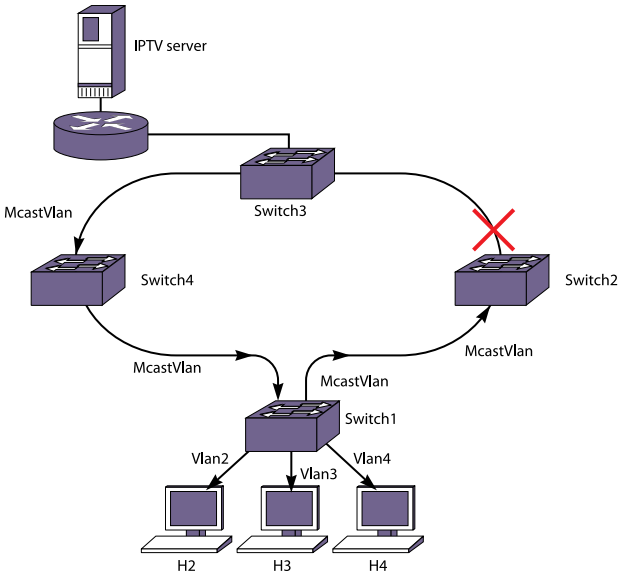



**Figure 248: Basic MVR Deployment**

Without MVR, there are two ways to distribute multicast streams in this topology:

- Extend subscriber VLANs (Vlan2, Vlan3, and Vlan4) to the network core, by tagging the ports connecting the switches.
- Configure all VLANS with an IP address and run PIM or DVMRP on each switch.

There are problems with both of these approaches. In the first approach, multiple copies of the same stream (IPTV channel) would be transmitted in the core, wasting bandwidth. In the second approach, all switches in the network core would have to be Layer 3 multicast aware, and would have to run a multicast protocol. Typical network cores are Layer 2 only.

MVR provides a simple solution to this problem If McastVlan in Switch1 is configured with MVR, it leaks the traffic into the local subscriber VLANs that contain hosts that request the traffic. For simple cases, perform these configuration steps:

- Configure MVR on McastVlan.
- Configure an IP address and enable *IGMP* and IGMP snooping on the subscriber VLANs (by default IGMP and IGMP snooping are enabled on Extreme Networks' switches).
- For all the multicast streams (IPTV channels), configure static IGMP snooping membership on the router on McastVlan.
- Enable MVR on the switches in the network.

Note
MVR works best with IGMPv1 and IGMPv2. We recommend that you do not use MVR with IGMPv3.

The strategy above conserves bandwidth in the core and does not require running PIM on the subscriber switches.

In this topology, a host (for example, a cable box or desktop PC) joins a channel through an IGMP join message. Switch1 snoops this message and adds the virtual port to the corresponding cache's egress list. This is possible because an MVR enabled VLAN can leak traffic to any other VLAN. When the user switches to another channel, the host sends an IGMP leave for the old channel and a join for the new channel. The corresponding virtual port is removed from the cache for the old channel and is added to the cache for the new channel.

As discussed in Static and Dynamic MVR on page 1494, McastVlan also proxies IGMP joins learned on other VLANs to the router. On an MVR network it is not mandatory to have a router to serve the multicast stream. All that is required is to have a designated IGMP querier on McastVlan. The IPTV server can also be directly connected to McastVlan.

## *Static and Dynamic MVR*

### Static MVR

In a typical IPTV network, there are several high demand basic channels. At any instant there is at least one viewer for each of these channels (streams), and they should always be available at the core (ring). When a user requests one of these channels, it is quickly pulled locally from the multicast *VLAN*. You have the option to use the static router configuration in each of the switches in the core. But this will cause all the channels to be available in the core, which may not be desired. For example, on an Extreme Networks router, you can use the following commands:

```
configure igmp snooping {vlan} vlan_name ports port_list add static
router
```

You can use the static MVR configuration and choose the groups for which the multicast stream should be flooded. If a multicast stream for a group in the static MVR range is received on an MVR enabled VLAN, it is always flooded on the MVR VLAN. This allows the neighbor switch in the ring to receive all the static MVR streams.

#### Dynamic MVR

In contrast, since a video content provider would like to provide a variety of on-demand and other premium channels, there are often many lower demand (fewer viewers) premium channels that cannot all be made available simultaneously at the core network. These should be streamed from the router only if requested by a host.

*IGMP* is the standard method used by a host to request a stream. However, IGMP packets are constrained to a VLAN. Thus, subscribers' IGMP join requests on the VLAN cannot be forwarded onto other VLANS. You can use a dynamic MVR configuration, and choose the groups for which the IGMP join requests should be proxied over the MVR VLAN. Thus, in Figure 248 on page 1493, McastVlan sends join and leave messages on behalf of Vlan2, Vlan3, and Vlan4. The router receives the messages on McastVlan and streams corresponding channels onto the core network. This provides on-demand service, and an administrator does not need to configure static IGMP on the router for each of these channels.

### Configuring MVR Address Range

By default, all multicast groups belong to MVR address range. Use the following command to specify the MVR address range:

```
configure mvr vlan vlan-name mvr-address {policy-name | none}
```

- Only the groups listed in the policy with "allow" action belong to MVR address range.
- For non-MVR groups, snoop cache is created and join requests are not proxied over MVR *VLAN*.

### Configuring Static and Dynamic MVR

By default, all MVR streams are static.

- Use the following command to specify which groups are static:

  ```
  configure mvr vlan vlan-name static group {policy-name | none}
  ```

  - Only the groups listed in policy with "allow" action are static. Any other groups in MVR address range are dynamic.
  - The groups in policy with "deny" action are dynamic. Any other groups in MVR address range are also dynamic, unless explicitly configured in the policy with "allow" action.
  - A group should not be configured to be both static and dynamic.

### MVR Configuration Example

The following example configuration is a two DUT scenario, in L2 mode , with no routing protocol or PIM configured.

- DUT-1 is sender
- DUT-2 is receiver
- *VLAN* v1 spans over DUT-1 and DUT-2 , DUT-2 also has v2 where *IGMP* joins are coming IN (225.1.1.1)
- DUT-2 has a VLAN v3, which also has a receiver connected sending IGMP join for same group as v2 (225.1.1.1)
- VLAN v1 in the DUT-2 has another port apart from the trunk port, no joins are being sent on this port.

Configure MVR on vlan v1 on DUT-2

```
* X460-48t.157 # show config "mcmgr"
#
# Module mcmgr configuration.
#
enable mvr vr VR-Default
configure mvr add vlan v1
configure mvr vlan v1 mvr-address none
configure mvr vlan v1 static group none*
 *X460-48t.158 #
```

The traffic will be flooded for the group only on MVR vlan (v1).

Since there are IGMP joins coming in on v2 and v3, v2, v3, and the second port in the MVR vlan v1 will receive traffic.

Configure the following policy file;

```
* X460-48t.155 # vi mvrPolicy.polentry policy1 {
if match any {
    nlri 225.1.1.1/24;
} then {
    permit;
}
}
-------------------------
```

When applying this policy file under static group on DUT-2

```
# configure mvr vlan v1 static group mvrPolicy
#configure mvr vlan v1 mvr-address none
```

When the policy file contains "permit", the traffic flows to v2, v3, and the second port in MVR VLAN.

When the policy file is changed to "deny", the second port in the MVR VLAN v1 will stop receiving the traffic.

If you configure static policy (by default – permit), traffic for that group range will always be available in the MVR VLAN, that is, it will be forwarded to all the ports in MVR VLAN.

When applying this policy file under mvr-address (Dynamic) on DUT-2:

```
# configure mvr vlan v1 static group none
# configure mvr vlan v1 mvr-address mvrPolicy
```

When the policy file contains "permit", the traffic flows to v2,v3 and the second port in MVR vlan.

When the policy file is changed to "deny", the second port in MVR VLAN v1 continues receiving traffic, but VLAN v2 and v3 stop receiving traffic, in spite of IGMP groups being learned. This is because the join on v2 and v3 will not be leaked to MVR VLAN.

Essentially, the dynamic policy does not directly apply on traffic, but it affects the joins, based on which traffic is forwarded or blocked.

Dynamic means only if the join is sent then traffic is forwarded.

The join is leaked to MVR VLAN so traffic from MVR VLAN will be received by other VLANs (v2 and v3).

To confirm if a join was leaked to MVR VLAN use the `show igmp group` command. It should have the group learned on MVR VLAN (v1) with port as "MVR".

*MVR Forwarding*

The goal for MVR is to limit the multicast traffic in the core Layer 2 network to only the designated multicast *VLAN*. If the backbone Layer 2 port is tagged with multiple VLANs, as shown in the following figure, a set of rules is needed to restrict the multicast streams to only one VLAN in the core.

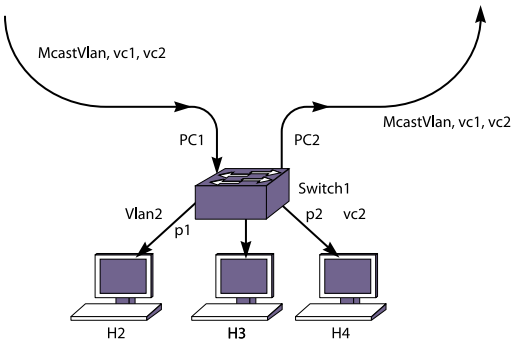



**Figure 249: Multiple VLANs in the Core Network**

In the above figure, the core network has 2 more VLANs, vc1 and vc2, to provide other services. With MVR, multicast traffic should be confined to McastVlan, and should not be forwarded to vc1 and vc2. It should be noted that MVR is configured only on the ingress VLAN (McastVlan). MVR is not configured on any other VLANs.

In the same way as the *IGMP* snooping forwarding rules, the multicast stream is forwarded onto member ports and router ports on the VLAN. For a stream received on MVR enabled ports, this rule is extended to extend membership and router ports to all other VLANs. This rule works well on the topology in the following figure. However, in a tagged core topology, this rule forwards traffic onto VLANs, such as vc1 and vc2, on ports PC1 and PC2. This results in multiple copies of same stream on the core network, thus reintroducing the problem that MVR was intended to solve.

To avoid multiple copies of the same stream, MVR forwards traffic with some special restrictions. MVR traffic is not forwarded to another VLAN unless a host is detected on the port. On the ingress MVR VLAN, packets are not duplicated to ports belonging to MVR VLANs. This is to prevent duplicate multicast traffic streams on ingress ports. Streams belonging to static MVR groups are always forwarded on MVR VLANs so that any host can join such channels immediately. However, dynamic groups are streamed from the server only if some host is interested in them. A command is provided to receive traffic on a port which is excluded by MVR. However, regular IGMP rules still apply to these ports, so the ports must have a router connected or an IGMP host to receive the stream.

These rules are to prevent multicast packets from leaking into an undesired virtual port, such as p2 on VLAN pc2 in the following figure. These rules also allow that, in most topologies, MVR can be deployed with minimal configuration. However, unlike EAPS and *STP (Spanning Tree Protocol)*, MVR is not intended to be a Layer 2 protocol to solve packet looping problems. Since multicast packets leak across

VLANs, one can misconfigure and end up with a multicast storm. MVR does not attempt to solve such problems.

**Note**
If a port is blocked by Layer 2 protocols, that port is removed from the egress list of the cache. This is done dynamically per the port state.

For most situations, you do not need to manually configure ports to receive the MVR multicast streams. But if one of the forwarding rules denies forwarding to a port that requires the streams, you can manually receive the MVR multicast streams by using the following command:

```
configure mvr vlan vlan_name add receiver port port-list
```

## Inter-Multicast VLAN Forwarding

In Basic MVR Deployment on page 1493, only simple topologies are considered, in which subscribers on different *VLAN*s access a multicast VLAN. There are topologies where streams need to be forwarded onto another multicast VLAN, as shown in the following figure. In this figure, a Multicast Service Provider (MSP) multicast VLAN is attached to ports 1:1-2 on both switches, SW1 and SW2. On the customer side, another multicast VLAN, delivers multicast streams to other switches around the ring.

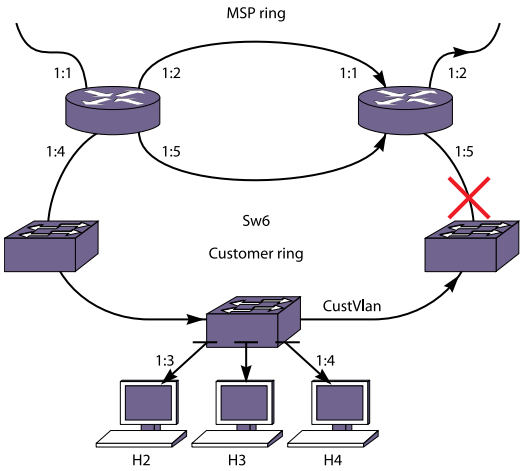



**Figure 250: Inter-Multicast VLAN Forwarding**

In this topology, a multicast stream can be leaked into the customer multicast network through either switch SW1 or SW2. However, as described in MVR Forwarding on page 1497, packets are not forwarded to router ports (ports 1:4 and 1:5 can be router ports if SW2 is an *IGMP* querier). To get around this, MVR needs to be configured on CustVlan either on SW1 or SW2. Since the forwarding rules apply only to non-MVR VLANs, traffic from one MVR VLAN is leaked into the router ports of another VLAN, if MVR is enabled on that.

In the topology above, the MSP multicast VLAN is carried on two switches that also carry the customer multicast VLAN. When multiple switches carry both multicast VLANs, it is imperative that MVR is configured on only one switch. Only that switch should be used as the transit point for multicast streams from one multicast ring into another. Otherwise, duplicate packets are forwarded. Also on the non-MVR switches, the ring ports should be configured as static router ports, so that ring ports are excluded from forwarding packets onto the customer ring. There is no mechanism to elect a designated MVR forwarder, so it must be configured correctly.

## MVR Configurations

MVR enables Layer 2 network installations to deliver bandwidth intensive multicast streams. It is primarily aimed at delivering IPTV over Layer 2 networks, but it is valuable in many existing EAPS or *STP* installations. This section explores a few possible deployment scenarios and configuration details. Of course, real world networks can be lot different from these examples. This section is meant to present some ideas on how to deploy MVR over existing networks, as well as to design new networks that support MVR.

### *MVR with EAPS*

Since MVR is designed with a Layer 2 ring topology in mind, it is strongly recommended that it should be deployed with EAPS. The MVR plus EAPS combination provides a superior solution for any triple play network, where the service provider intends to provide data, voice, and video services. EAPS is a proven solution for providing sub-second SONET-like protection to Layer 2 rings. For more detail on EAPS refer to EAPS on page 966.

Consider a typical EAPS topology in the following figure, where 3 *VLAN*s on the core ring serve various clients on each switch. To provide video service, one of the VLANs on the EAPS ring is designated as a multicast VLAN. MVR is enabled only on this VLAN (mcastvlan). V1 is the control VLAN, and V2 is another protected VLAN. A router serving the multicast feed would typically run PIM on mcastvlan, to support the static and dynamic *IGMP* membership on the VLAN.

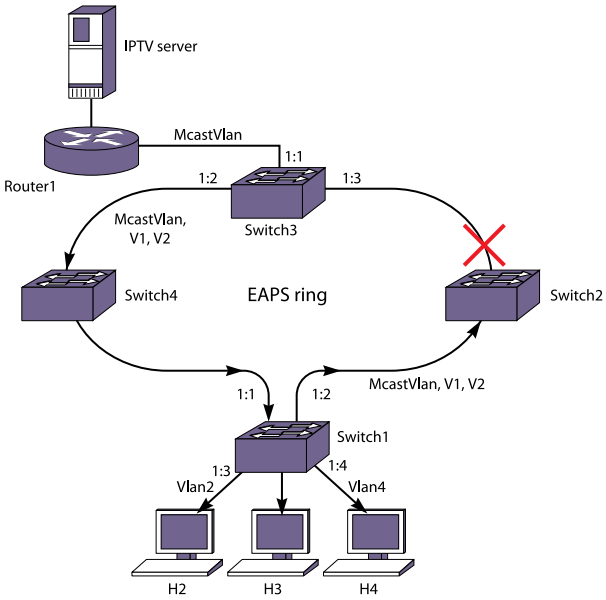



**Figure 251: MVR on an EAPS Ring**

The following is a typical configuration for the router and switches.

**Router1:**

```
create vlan mcastvlan
configure mcastvlan add port 1:1
create vlan server
configure server add port 1:2
configure mcastvlan ipaddress 10.1.1.1/24
configure server ipaddress 11.1.1.1/24
configure igmp snooping mcastvlan port 1:1 add static group 239.1.1.1
enable ipforwarding
```

```
enable ipmcforwarding
configure igmp snooping leave-timeout 2000
configure pim add vlan all
enable pim
```

**Switch1:**

```
create vlan mcastvlan
create vlan v1
create vlan v2
create vlan vlan2
configure vlan vlan2 add port 1:3
configure vlan vlan2 ipaddress 10.20.1.1/24
configure mcastvlan tag 20
configure mcastvlan add port 1:1,1:2 tag
configure mvr add vlan mcastvlan
configure vlan v1 tag 30
configure v1 add port 1:1,1:2 tag
configure vlan v2 tag 40
configure v2 add port 1:1,1:2 tag

create eaps e1
configure eaps e1 mode transit
configure eaps e1 add control vlan v1
configure eaps e1 add protect vlan mcastvlan
configure eaps e1 add protect vlan v2
configure eaps port primary port 1:1
configure eaps port secondary port 1:2
enable eaps
enable mvr
```

**Switch3:**

```
create vlan McastVlan
create vlan v1
create vlan v2
configure mcastvlan tag 20
configure mcastvlan add port 1:2,1:3 tag
configure mcastvlan add port 1:1
configure mvr add vlan mcastvlan
configure vlan v1 tag 30
configure v1 add port 1:2,1:3 tag
configure vlan v2 tag 40
configure v2 add port 1:2,1:3 tag

create eaps e1
configure eaps e1 mode master
configure eaps e1 add control vlan v1
configure eaps e1 add protect vlan mcastvlan
configure eaps e1 add protect vlan v2
configure eaps port primary port 1:3
configure eaps port secondary port 1:2
enable eaps
enable mvr
```

Note

In this example, Switch3 is the EAPS master, but any of the other switches in the ring could have been configured as the master.

### *MVR with STP*

In a Layer 2 ring topology, MVR works with *STP* as it works with EAPS. However, in other Layer 2 topologies, additional configuration steps may be needed to make sure that multicast feeds reach all network segments. Extra configuration is required because all ports in the *VLAN* are part of an STP domain, so that solely by examining the configuration it is not clear whether a port is part of bigger ring or is just serving a few hosts. In EAPS this problem is solved by distinguishing between configured primary or secondary ports from other VLAN ports. Consider a simplified Layer 2 STP network as shown in the following figure.

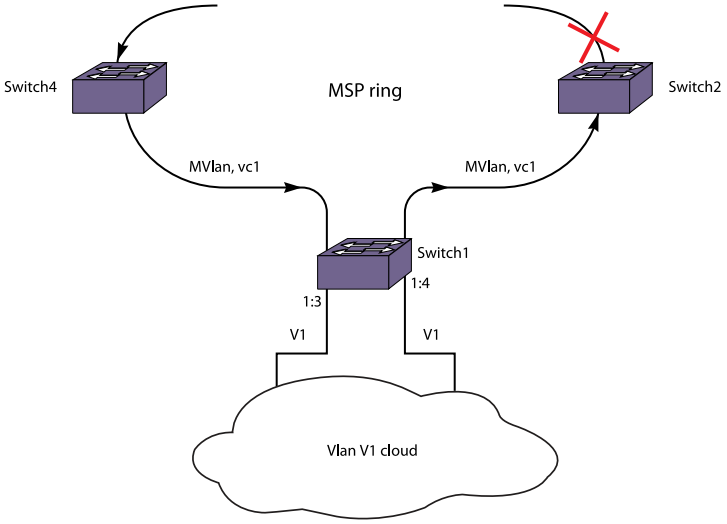



**Figure 252: MVR with STP**

In this topology, subscribers are in a Layer 2 cloud on VLAN V1.

STP is configured for all ports of V1. Since V1 spans on the ring as well, multicast cannot be forwarded on V1 blindly. Forwarding rules (described in MVR Forwarding on page 1497), dictate that multicast traffic is not forwarded on STP enabled ports. This is to make sure that multiple copies of multicast packets are not forwarded on the ring. However, since other STP enabled ports on V1 (1:3,1:4) are not part of the ring multicast stream, they need to be configured so that they get the packets. To configure the ports to receive packets, use the following command (mentioned in MVR Forwarding on page 1497):

```
configure mvr vlan vlan-name add receiver port port-list
```

**Note**
If the Layer 2 cloud is connected back to ring ports, traffic may end up leaking into VLAN V1 in the ring. There is no way to avoid that. So, such topologies must be avoided.

The following is a typical configuration for Switch 1 in the above figure:

```
create vlan v1
configure v1 tag 200
configure v1 add port 1:1, 1:2 tag
configure v1 add port 1:3, 1:4
create vlan mvlan
configure mvlan add port 1:1, 1:2
configure mvr add vlan mvlan
create stpd stp1
configure stp1 add vlan v1 port all
enable stpd stp1 port all
```

```
configure mvr vlan v1 add receiver port 1:3,1:4
enable mvr
```

*MVR in a VMAN Environment*

In the case of a VMAN, a packet is tagged with a VMAN tag in addition to a possible *VLAN* tag. This is to provide VLAN aggregation for all customer traffic in the VMAN ring. Each customer is given its own VLAN, and traffic from all customers can be tunneled on a single VMAN tag into the metro ring to an outside Broadband Remote Access Server (BRAS). In a VMAN network, multicast traffic can be distributed over a separate VLAN in the metro core. These packets are not subjected to VMAN tunneling. Thus, IPTV service can be provided on this multicast VLAN on a VMAN network.

MVR deployment in a VMAN environment is not any different from that in an EAPS environment, since a separate multicast VLAN on the metro ring is used for multicasting. However, it provides interesting capabilities to MSPs for their video offerings. Different service bundles can be offered on separate VLANs. Packets are not forwarded to any metro link segments where a stream is not required.

The following figure illustrates an example design for MVR in a VMAN environment. Any multicast packet entering on MVlan is forwarded on MVlan to the next switch. These multicast packets are not tunneled.

With MVR, switches on the VMAN do not have to run any routing protocol. If MVR is enabled on the multicast VLAN, MVlan, traffic is pulled from the IPTV server. Such multicast packets egressing from the CE port are always untagged. The downstream DSLAM distributes untagged multicast packets to the respective subscriber VLANs.

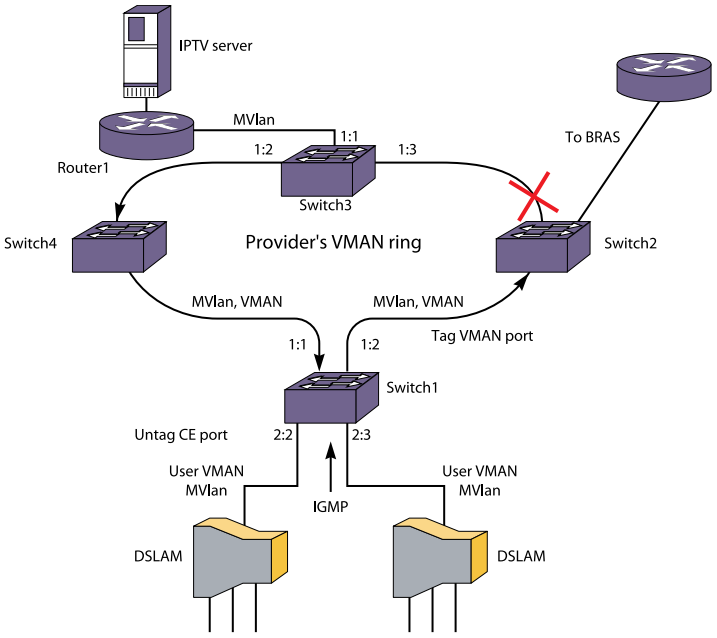



**Figure 253: MVR in a VMAN Environment**

The following is a typical configuration for Switch 1 in the above figure:

```
create vman vman2
configure vman vman2 tag 300
configure vman vman2 add port 2:2-2:3 untagged
configure vman vman2 add port 1:1,1:2 tagged
enable port 2:*
```

```
enable port 1:*
create vlan mvlan
configure vlan mvlan tag 200
configure vlan mvlan add port 1:1,1:2 tag
configure mvr add vlan mvlan
enable mvr
```

# Displaying Multicast Information

## Displaying the Multicast Routing Table

- To display part or all of the entries in the multicast routing table, use the following command:
  `show iproute {ipv4} {{vlan} name | [ipaddress netmask | ipNetmask] | origin [direct | static | mbgp | imbgp | embgp]} multicast {vr vr_name}`

## Displaying the Multicast Cache

- The multicast cache stores information about multicast groups. To display part or all of the entries in the multicast cache, use the following command:
  `show mcast cache {{vlan} vlan_name} {{[group grpaddressMask | grpaddressMask] {source sourceIP | sourceIP}} {type [snooping | pim | mvr]}| {summary}}`

## Looking Up a Multicast Route

- To look up the multicast route to a specific destination, use the following command with the **multicast** option:
  `rtlookup [ipv4_address | ipv6_address] { unicast | multicast | rpf } {vr vr_name}`

## Looking Up the RPF for a Multicast Source

- To look up the RPF for a multicast source, use the following command with the **rpf** option:
  `rtlookup [ipv4_address | ipv6_address] { unicast | multicast | rpf } {vr vr_name}`

## Displaying the PIM Snooping Configuration

- To display the PIM snooping configuration for a *VLAN*, use the following command:
  `show pim snooping {vlan} vlan_name`

# Troubleshooting PIM

## Multicast Trace Tool

The multicast trace tool is the multicast equivalent of unicast trace route mechanism and is an effective tool for debugging multicast reachability problems. This tool is based on an IETF draft and uses *IGMP*.

Because it is harder to trace a multicast path from the source to the destination, a multicast trace is run from the destination to the source. The multicast trace can be used to do the following:

- Locate where a multicast traffic flow stops
- Identify sub-optimal multicast paths

A multicast trace is used for tracing both potential and actual multicast forwarding tree paths. When the multicast tree is established and traffic is flowing, this tool traces the actual traffic flow. If there is no traffic while executing the command, this tool displays the potential path for the group and source being traced.

You can direct a multicast trace to any network destination, which can be a multicast source or destination, or a node located between a source and destination. After you initiate the trace, a multicast trace query packet is sent to the last-hop multicast router for the specified destination. The query packet contains the source address, group address, destination/receiver address, response address, maximum number of hops, and TTL to be used for the multicast trace response.

The previous hop router selection is based on the multicast routing protocol and the state for the S,G entry in the processing router.

For example:

- If there is no S,G state in the router, the parent closest to the RP is chosen as the previous hop.
- If the S,G state exists in the router, the parent closest to the source is chosen as the previous hop.

The last hop router converts the multicast trace query into a unicast traceroute request by appending response data (for the last hop router) into the received query packet, and the last hop router forwards the request packet to the router that it believes is the proper previous hop for the given source and group.

Each multicast router adds its response data to the end of the request packet, and then forwards the modified unicast request to the previous hop.

The first hop router (the router that determines that packets from the source originate on one of its directly connected networks) changes the packet type to response packet and sends the completed response to the query generator. If a router along the multicast route is unable to determine the forwarding state for a multicast group, that router sends a response back to the originator with NO ROUTE forwarding code.

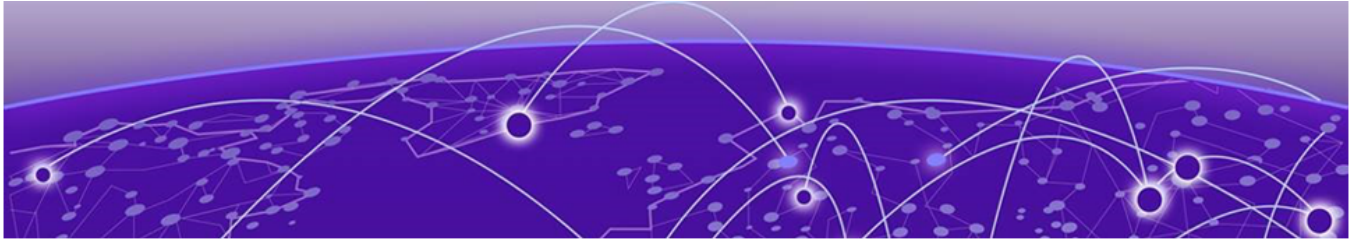
To initiate a multicast trace, use the following command:

```
mtrace source src_address {destination dest_address} {group grp_address}
{from from_address} {gateway gw_address} {timeout seconds} {maximum-hops
number} {router-alert [include | exclude]} {vr vrname}
```

## Multicast Router Information Tool

The multicast router information tool is an ExtremeXOS command that allows you to request information from a specific multicast router. For more information, see the command description for the following command:

```
mrinfo {router_address} {from from_address} {timeout seconds} {multiple-
response-timeout multi_resp_timeout} {vr vrname}
```



# IPv6 Multicast



This chapter introduces IPv6 multicast, which allows a single IPv6 host to send a packet to a group of IPv6 hosts, and the features and configuration of the Multicast Listener Discovery (MLD) protocol. For more information on IPv6 multicasting, refer to the following publications:

- RFC 2710—Multicast Listener Discovery (MLD) for IPv6
- RFC 3810—Multicast Listener Discovery Version 2 (MLDv2) for IPv6

## Multicast Listener Discovery (MLD) Overview

Multicast Listener Discovery (MLD) is a protocol used by an IPv6 host to register its IP multicast group membership with a router. To receive multicast traffic, a host must explicitly join a multicast group by sending an MLD report; then, the traffic is forwarded to that host. Periodically, the router queries the multicast group to see if the group is still in use. If the group is still active, a single IP host responds to the query, and group registration is maintained.

MLD is the IPv6 equivalent to *IGMP (Internet Group Management Protocol)*, and MLDv1 is equivalent to IGMPv2. The ExtremeXOS software supports both MLDv1 and MLDv2 protocol.

Note
This release does not support MVR, PVLAN, *VLAN (Virtual LAN)* Aggregation, and Multicast Troubleshooting tools for MLD/IPv6.

### ExtremeXOS Resiliency Enhancement for IPv6 Static Routes

The ExtremeXOS Resiliency Enhancement feature provides a resilient way to use *ECMP (Equal Cost Multi Paths)* to load balance IPv6 traffic among multiple servers or other specialized devices. ExtremeXOS automatically manages the set of active devices using ECMP static routes configured with ping protection to monitor the health of these routes. Such servers or specialized devices do not require special software to support Bidirectional Forwarding Detection (BFD), or IP routing protocols such as *OSPF (Open Shortest Path First)*, or proprietary protocols to provide keepalive messages. ExtremeXOS uses industry-standard and required protocols ICMPv6/Neighbor Discovery for IPv6 to accomplish the following automatically:

- Initially verify devices and activate their static routes, without waiting for inbound user traffic, and without requiring configuration of device MAC addresses.
- Detect silent device outages and inactivate corresponding static routes.
- Reactivate static routes after device recovery, or hardware replacement with a new MAC address.

ExtremeXOS currently supports similar protection and resiliency using Bidirectional Forwarding Detection (BFD) on IPv4 and IPv6 static routes. However, BFD can only be used when the local and remote device both support BFD.

# Managing MLD

## Enabling and Disabling MLD on a VLAN

MLD is disabled by default on all *VLAN*s. You can enable MLD using the enable `mld {vlan vlan name} {MLDv1 | MLDv2}` command.

This allows IPv6 hosts to register with IPv6 multicast groups and receive IPv6 multicast traffic.

- To disable MLD on a VLAN after it has been enabled, use the `disable mld {vlan name}` command.
- To enable MLD on a VLAN after it has been disabled, use the `enable mld {vlan vlan_name} {MLDv1 | MLDv2}` command.

## MLD Snooping

Similar to *IGMP* snooping, MLD snooping is a Layer 2 function of the switch; it does not require multicast routing to be enabled. In MLD snooping, the Layer 2 switch keeps track of MLD reports and only forwards multicast traffic to the part of the local network that requires it. MLD snooping optimizes the use of network bandwidth and prevents multicast traffic from being flooded to parts of the local network that do not need it.

MLD snooping is disabled by default on all *VLAN*s in the switch.

When MLD snooping is disabled on a VLAN, all MLD and IPv6 multicast traffic floods within the VLAN.

MLD snooping expects at least one device on every VLAN to periodically generate MLD query messages.

- Enable or disable MLD snooping:

  ```
  enable mld snooping
  ```

  ```
  disable mld snooping
  ```

  Multicast packets with a scope id less than 2 are not forwarded by the MLD snooping enabled switch. Kill entry is installed in the hardware for this traffic.

  Multicast packets with a scope id of 2 and group address in the range of FF02::/111 (Addresses allocated by IANA as per RFC 3307) are always flooded to all ports of the VLAN by hardware and a copy of the packet is provided to slow path. There are no cache entries in software or hardware for these addresses.
- Multicast packets with a scope id of 2 and group address as solicited multicast address (FF02::1:FFXX:XXXX/104) are flooded to all ports of VLAN for 135 seconds (Default MLD query interval + Maximum response time), if there are no members for this group.

Otherwise, the traffic is forwarded based on the snooping database. Multicast cache entries for these addresses are maintained only in the software and traffic is always slow path forwarded.

Multicast addresses with a scope id of 2 and that do not qualify in the above categories will be forwarded based on the snooping database.

Cache entries for these multicast addresses will be installed in hardware. Unregistered packets are dropped.

- In general, all multicast data traffic on a PIMv6 enabled VLAN is controlled by the PIMv6 protocol. However, multicast traffic with either the group address or source address as the link local address will not be controlled by PIMv6. Instead, it will be L2 forwarded based on the snooping database.

  For multicast packets with a scope id greater than 2 on PIMv6 enabled VLANs, cache entries are controlled by the PIMv6 protocol.

  PIMv6 provides a list of egress VLANs for which packets need to be forwarded. The snooping database is used to construct the set of ports for ingress VLANs as well as for each egress VLAN.

  On PIMv6 disabled VLANs, traffic is forwarded based on the snooping database on the ingress VLAN.

  In both cases, cache entry is installed in the hardware, and traffic is fast path forwarded.
- The MLD snooping proxy feature is enabled automatically when MLD snooping is enabled. This feature optimizes the forwarding of MLDv1 reports. The following conditions apply for each group:
  - Only the first received MLD join is forwarded upstream.
  - Only the MLD leave for last host is forwarded upstream.

    When a switch receives an MLD leave message on a port, it sends a group-specific query on that port if proxy is enabled (even if it is a non-querier). The switch removes the port from the group after the leave timeout (a configuable value from 0 - 175000ms with a default of 1000ms). If all the ports are removed from the group, the group is deleted and the MLD leave is forwarded upstream. If MLD snooping proxy is disabled, then all the MLD reports are forwarded upstream.

    Note
    MLD snooping proxy does not apply to MLDv2 reports.
- MLD snooping is implemented primarily through *ACL (Access Control List)*s, which are processed on the interfaces. These special purpose ACLs are called MLD snooping hardware filters. On Summit family switches and BlackDiamond 8800 series switches, the software allows you to choose between two types of MLD snooping hardware filters: per-port filters (the default) and per-VLAN filters.

  The two types of MLD snooping hardware filters use switch hardware resources in different ways.
- The two primary hardware resources to consider when selecting the MLD snooping hardware filters are the Layer 3 multicast forwarding table, and the interface ACLs. The size of both of these hardware resources is determined by the switch model. In general, the per-port filters consume more

resources from the multicast table and less resource from the ACL table. The per-VLAN filters consume less space from the multicast table and more from the ACL table.

- In Summit family switches and BlackDiamond 8800 series switches, since the multicast table size is smaller, using the per-port filters can fill up the multicast table and place an extra load on the CPU. To avoid this, configure the switch to use the per-VLAN filters.

**Note**
The impact of the per-VLAN filters on the ACL table increases with the number of VLANs configured on the switch. If you have a large number of configured VLANs, we suggest that you use the per-port filters.

### *MLD Snooping Filters*

MLD snooping filters allow you to configure a policy file on a port to allow or deny MLD report and leave packets coming into the port.

(For details on creating policy files, see Policy Manager on page 635.) The MLD snooping filter feature is supported by MLDv1 and MLDv2.

**Note**
Do not confuse MLD snooping filters with MLD snooping hardware filters explained in previous section. MLD snooping filters are software filters, and the action is applied at software level by the ExtremeXOS multicast manager.

For the policies used as MLD snooping filters, all the entries should be IPv6 address type entries, and the IPv6 address of each entry must not be in the range of FF02::/96.

Use the following template to create a snooping filter policy file that denies MLD report and leave packets for the FF03::1/128 and FF05::1/112 multicast groups:

```
#
# Add your group addresses between "Start" and "end"
# Do not touch the rest of the file!!!!
entry mldFilter {
if match any {
#------------------ Start of group addresses ------------------
nlri FF03::1/128;
nlri FF05::1/112;
#------------------- end of group addresses -------------------
} then {
deny;
}
}
entry catch_all {
if {
} then {
permit;
}
}
```

After you create a policy file, use the following command to associate the policy file and filter to a set of ports:

`configure mld snooping vlan` *`vlan_name`* **`ports`** *`port_list`* **`filter`** [*`policy`*]

To remove the filter, use the **none** option.

To display the MLD snooping filters, use the following command:

```
show mld snooping {vlan} name filter
```

*Limiting the Number of Multicast Sessions on a Port*

The default configuration places no limit on the number of multicast sessions on each *VLAN* port. To place a limit on the number of learned MLD reports, or remove the limit, use the command:

```
configure mld snooping {vlan} vlan_name ports port_list join-limit
[num_joins | no-limit]
```

## Configuring MLD Snooping

- To configure the timers that control MLD operation, use the command:

  ```
  configure mld query_interval query_response_interval
  last_member_query_interval {robustness}
  ```

- Similar to *IGMP* snooping, MLD snooping is a Layer 2 function of the switch. It does not require multicast routing to be enabled. MLD snooping keeps track of MLD reports, and only forwards multicast traffic to that part of the local network that requires it. MLD snooping is disabled by default on all *VLAN*s. If MLD snooping is disabled on a VLAN, all MLD and IPv6 multicast traffic floods within the VLAN. To enable MLD snooping, use the command:

  ```
  enable mld snooping {vlan name}
  ```

## Clearing MLD Group Registration

To clear a single group or all groups in a *VLAN* learned through MLD, use the command:

```
clear mld group {v6grpipaddress} {{vlan} name}
```

## Configuring Static MLD

In some situations, you might want multicast traffic to be forwarded to a port where a multicast-enabled host is not available (for example, when you test multicast configurations). Static MLD emulates a host or router attached to a switch port, so that multicast traffic is forwarded to that port, and the switch sends a proxy join for all the statically configured MLD groups when an MLD query is received. You can emulate a host to forward a particular multicast group to a port; and you may emulate a router to forward all multicast groups to a port.

- To emulate a host on a port, use the command:

  ```
  configure mld snooping {vlan} vlan_name{ ports port_list } add static
  group v6grpipaddress
  ```

- To emulate a multicast router on a port, use the command:

  ```
  configure mld snooping {vlan}vlan_name ports port_list add static
  router
  ```

- To remove these entries, use the corresponding commands:

  ```
  configure mld snooping {vlan} vlan_name {ports port_list } delete
  static group [all | v6grpipaddress]
  ```

```
configure mld snooping {vlan} vlan_name ports port_list delete static
router
```

## MLD Loopback

Prior to ExtremeXOS 15.3.2, you could configure static groups, but it was necessary to specify port(s). As of 15.3.2, the configuration of dynamic groups is now supported. The MLD Loopback feature along with the existing static group feature supports the configuration of static and/or dynamic groups with or without ports.

A *VLAN* in loopback mode may not have ports associated with it, but its operational status is up. However, it is not possible to have multicast receivers on a VLAN without having a port. Sometimes, there is a need to pull the multicast traffic from upstream to the loopback VLAN for troubleshooting. The traffic need not always be forwarded to any ports/receivers. The MLD Loopback feature allows you to configure groups on a VLAN without specifying a port, so the traffic is pulled from upstream but not forwarded to any port.

The loopback (Lpbk) port is the logical port on a VLAN in the application context. When you configure a group on a VLAN but do not specify the port, the switch forms an MLDv1 join and assumes it to be received on the Lpbk port. A dynamic group ages out after the membership timeout if there are no other receivers. Membership joins refresh the dynamic group. The static group remains until it is removed from the configuration.

The following command is used to configure static or dynamic group entry:

```
configure mld snooping {vlan} vlan_name {ports portlist} add [static |
dynamic] group ip_address
```

## Displaying MLD Information

- To display MLD configuration and operation information, use the command:

  ```
  show mld group {{vlan} {name} | {v6grpipaddress}} {MLDv2}
  ```

- To display the MLD static group information, use the command:

  ```
  show igmp snooping {vlan} name static [group | router]
  ```

## MLD SSM Mapping

The MLD-SSM Mapping feature allows MLDv1 hosts to participate in SSM functionality, and eliminates the need for MLDv2. You can configure SSM map entries that specify the sources used for a group/group range for which SSM functionality is applied. You also have the option to configure domain name and DNS server to use to obtain the source list.

When a MLDV1 report is received, the configured sources are provided to PIM so that it can send source specific joins. When the host leaves, or when the membership times out, PIM is informed so that it can consider sending prunes.

Note

The sources mapped to only the LPM group are used.

*Feature Implementation Information*

- This feature is implemented as an extension to existing *IGMP* SSM support.
- The CLI commands for this feature are applicable in VR context.
- PIM is completely unaware of existence of this feature, so there is no change in PIM processing.
- On last hop the (S, G) cache created through MLDv1 join is similar to the (S, G) cache created as a result of RPT to SPT switchover in PIM-SM. There is no indication that the cache is created as result of MLDv1 join, or MLDv2 report.

*Limitations*

- Only 50 sources (static or dynamic) are allowed for each group address, or group range.

  **Note**
  The DNS server may send only 14 IPv6 source addresses in its response thereby limiting the number of dynamic sources supported.

- Only one DNS name is allowed for each group address/group range.

*SSM Address Range*

The address prefix FF3x::/32 is reserved by IANA for SSM use. All SSM addresses must have P=1, T=1 and plen=0. RFC 3306 mandates that the network prefix is zero, which results in the SSM address range to be in FF3x::/96 range. Since future documents may allow a non-zero network prefix, this feature allows the addresses in range FF3x::/32 in SSM map configuration. The default SSM range is FF3x::/96.

SSM address range is configured from the PIM context using the following command:

```
configure pim ipv6 ssm range default | policy policy_name
```

When this command is issued, PIM notifies MCMGR with the details of the SSM address range. MCMGR applies this range for the MLD SSM feature.

*Handling MLD Reports*

The following table captures the enhanced functionality.

**Table 152: MLD Mapping**

| MLDv1 Join | MLD SSM Map Disabled | MLD SSM Map Enabled |
|---|---|---|
| Group in SSM range but no map entries | Dropped | Dropped |
| Group in SSM range with map entries | Dropped | SSM map sources included |
| Group not in SSM range | SSM map sources not included | SSM map sources not included |

MLDv1 reduction messages in the SSM range are accepted and processed normally. Multicast manager will send out a group-specific query and refresh the receivers on receiving joins.

When an MLDv2 report is received, following group records types are ignored if they refer to SSM group range:

- MODE_IS_EXCLUDE

- CHANGE_TO_EXCLUDE_MODE

Other group records in the same report are processed.

*SSM Data Forwarding*

L2 forwarding decision of SSM data traffic is taken by the Multicast manager. Sources configured in the SSM map are considered only when a MLDv1 receiver is connected to a port. For the same group, there may be another port on which only MLDv2 receiver is present. For this port, SSM map sources are not considered.

L3 forwarding decision of SSM data traffic is taken by PIM. PIM provides the list of egress VLANs to be included, and the multicast manager decides the egress ports on each of these VLANs.

*DNS Request/Response*

When a DNS name is configured for a group in an SSM map, a DNS request is sent to resolve the name. If static sources are already configured for the group, it takes precedence and the DNS request is not sent. This feature uses the AAAA record type. For example, if the multicast group is FF30::1234:5678 and the configured domain name is ext.foo.bar, then the DNS request is sent for 8.7.6.5.4.3.2.1.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.3.F.F.ext.xxx.xxx.

The DNS request is not sent every time a report is received. A DNS request is sent only in the following scenarios:

- When a DNS name is configured for a group.
- When the refresh ssm-map command is used.
- When the DNS age timer expires.

The DNS response may contain one or more addresses. These addresses are updated to the SSM mapping table against the specific group.

Note
LPM is used to match the group for a DNS response.

If a static source is configured for a group for which DNS resolved addresses are already present, the dynamic sources are flushed and the static sources are added. Static mapped addresses are preferred over DNS resolved addresses.

*DNS Timers*

The DNS age timer is a decrementing timer to zero and applies to a group/group range. The value of this timer is calculated by the following formula:

DNS age = (value of TTL in the DNS response) + (wait TTL)

wait TTL = 260 seconds

A DNS request is sent when receiving an MLDv1 report, but the TTL has expired. The source is not removed until the wait TTL time expires.

On receiving the DNS response, the existing DNS resolved addresses which are not in the DNS response are removed. Any new addresses in the DNS response are added and the DNS age timer is refreshed.

The L2 SSM data caches are modified for the addresses removed, or added. PIM is notified to take action on L3 SSM caches.

If the DNS response is not received and the DNS age timer expires, the mapping entry is removed (if there are no static addresses). The SSM data traffic forwarding is stopped immediately when the group is removed.

### *DNS Server*

This feature does not check or track DNS servers configured in the switch. You must correctly configure and administer the DNS server.

The following command is used to configure the DNS server:

`configure dns-client add name-server` *`ip_address`* {**`vr`** *`vr_name`*

The server(s) are displayed using the following command:

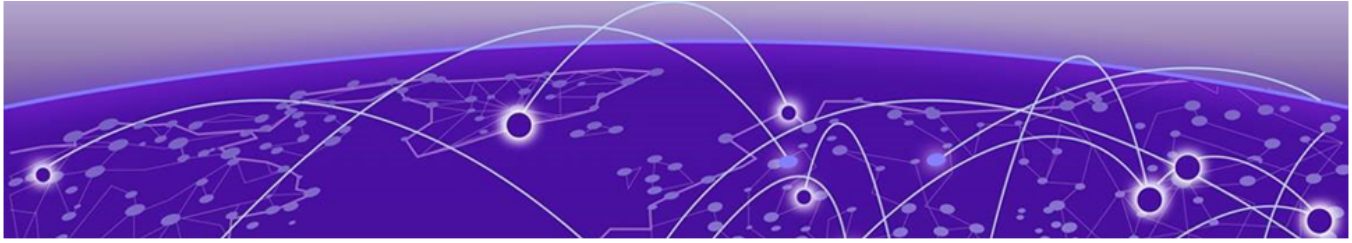
```
# show dns-client
Number of domain suffixes: 0
Number of domain servers: 1
Name Server 0: 10.120.89.75 VR-Default
```

MCMGR uses nettools library to perform DNS lookups. gethostbyname_c is used by specifying the callback function to be invoked when DNS response is received.

### *Configuring MLD SSM Mapping*

Use the following commands to configure MLD SSM Mapping in ExtremeXOS:

- Enable or disable MLD SSM Mapping on a VR: `[enable | disable] mld ssm-map { {vr}` *`vrname`* `}`
- Add an MLD SSM Mapping entry on a VR: `configure mld ssm-map add` *`v6groupnetmask`* `[v6sourceip |` *`src_domain_name`*`] { {vr}` *`vrname`*`}`
- Delete an MLD SSM Mapping entry on a VR:`configure mld ssm-map delete` *`v6groupnetmask`* `[v6sourceip |` *`src_domain_name`*`] { {vr}` *`vrname`*`}`
- Delete all MLD SSM Mapping entries on a VR: `unconfigure mld ssm-map { {vr}` *`vrname`*`}`
- Display the status of MLD-SSM mapping feature on a VR, and display the MLD-SSM mapping entries: `show mld ssm-map {` *`v6groupnetmask`* `} { {vr}` *`vrname`* `}`
- Send out a DNS request for a particular group. On receiving the DNS response, the "DNS Age" in the SSM mapping entry is refreshed: `refresh mld ssm-map` *`v6groupnetmask`* `{ {vr}` *`vrname`*`}`
- Configure the DNS server: `configure dns-client add name-server` *`ip_address`* `{vr` *`vr_name`*`}]`
- Display the DNS Servers: `show dns-client`



# MSDP



This chapter introduces MSDP (Multicast Source Discovery Protocol), an interdomain multicast protocol used to connect multiple multicast routing domains that run PIM-SM (Protocol Independent Multicast-Sparse Mode). This chapter discusses the features and configuration for PIM border, MSDP peers, mesh groups, anycast RP, SA cache, redundancy, and *SNMP (Simple Network Management Protocol)* MIBs.

Note
For more information about MSDP, refer to RFC 3618.

## MSDP Overview

*MSDP (Multicast Source Discovery Protocol)* is an interdomain multicast protocol used to connect multiple multicast routing domains that run Protocol Independent Multicast-Sparse Mode (PIM-SM).

MSDP speakers are configured on each PIM-SM domain. These speakers establish a peering relationship with other MSDP speakers through secured TCP connections. When the source sends traffic, the MSDP speaker learns about the source through its Rendezvous Point (RP). In turn, the RP advertises the source to its peers through Source Active (SA) messages. The peers receive these advertisements and inform their RPs about the presence of the active source in the other PIM-SM domain, which triggers the normal PIM operation in the corresponding domains.

For example, as businesses expand and networks grow in size, it might become necessary to connect PIM domains to allow multicast applications to reach other offices across the network. MSDP simplifies this process by providing a mechanism to connect those multicast routing domains without reconfiguring existing domains. Each PIM domain remains separate and has its own RP. The RP in each domain establishes an MSDP peering relationship over a TCP connection either with RPs in other domains or with border routers leading to other domains. When an RP learns about a new multicast source in its own domain (using the normal PIM registration process), it then sends a SA message to all of its MSDP peers, letting them know about the new stream. In this way, the network can receive multicast traffic from all over the network without having to reconfigure each existing PIM domain.

## Supported Platforms

*MSDP* is supported on all platforms running a minimum software version of ExtremeXOS 12.0 with the Core license.

Our implementation of MSDP is compliant with RFC 3618 and RFC 3446, and compatible with other devices that are compliant with these standards.

## Limitations

The limitations of *MSDP* are as follows:

- There is no support for MSDP operating with SA cache disabled (transit node). MSDP will always cache/store received SA messages.
- There is no support for logical RP.
- There is no support for MSDP on user-created virtual routers (VRs).
- *RIP (Routing Information Protocol)* routes are not used for peer-RPF checking. So, our implementation of MSDP does not exactly conform to rule (iii) in section 10.1.3 of RFC 3618. However, our implementation of MSDP uses *BGP (Border Gateway Protocol)*/*OSPF (Open Shortest Path First)* for peer-RPF checking as per rule (iii) in section 10.1.3.
- Read-write/read-create access is not supported on MSDP MIB objects.

# PIM Border Configuration

To create a PIM-SM domain for *MSDP*, you must restrict the reach of Bootstrap Router (BSR) advertisements by defining a *VLAN (Virtual LAN)* border. BSR advertisements are not sent out of a PIM interface configured as a VLAN border, thereby defining a PIM domain for MSDP.

To configure a PIM VLAN border, use the command:

```
configure pim vlan_name border
```

# MSDP Peers

*MSDP* peers exchange messages to advertise active multicast sources. The peer with the higher IP address passively listens to a well-known port number and waits for the side with the lower IP address to establish a Transmission Control Protocol (TCP) connection on port 639. When a PIM-SM RP that is running MSDP becomes aware of a new local source, it sends an SA message over the TCP connection to its MSDP peer. When the SA message is received, a peer-RPF check is performed to make sure the peer is toward the originating RP. If so, the RPF peer floods the message further. If not, the SA message is dropped and the message is rejected.

- Configure an MSDP peer using the command:

  ```
  create msdp peer remoteaddr {remote-as remote-AS} {vr vrname}
  ```

- Delete an MSDP peer using the command:

  ```
  delete msdp peer [all | remoteaddr] {vr vr_name}
  ```

- Display configuration and run-time parameters about an MSDP peer using the command:

  ```
  show msdp [peer {detail} | {peer} remoteaddr] {vr vr_name}
  ```

## MSDP Default Peers

You can configure a default peer to accept all SA messages. Configuring a default peer simplifies the peer-RPF checking of SA messages. If no policy is specified, the current peer is the default RPF peer for all SA messages.

When configuring a default peer, you can also specify an optional policy filter. If the peer-RPF check fails, and a policy filter is configured, the default peer rule is applied to see if the SA message should be accepted or rejected.

You can configure multiple default peers with different policies. However, all default peers must either be configured with a default policy or not. A mix of default peers, with a policy and without a policy, is not allowed.

- Configure an *MSDP* default peer, and optional policy filter using the command:

  `configure msdp peer [`*remoteaddr* `|` **all**`] default-peer {default-peer-policy` *filter-name*`} {vr` *vrname*`}`

- Remove the default peer using the command:

  `configure msdp peer [`*remoteaddr* `|` **all**`] no-default-peer {vr` *vrname*`}`

- Verify that a default peer is configured using the command:

  `show msdp [peer {detail} | {peer}` *remoteaddr*`] {vr` *vr_name*`}`

## Peer Authentication

*MSDP* supports TCP RSA Data Security, Inc. *MD5 (Message-Digest algorithm 5)* Message-Digest Algorithm authentication (RFC 2385) to secure control messages between MSDP peers. You must configure a secret password for an MSDP peer session to enable TCP RSA Data Security, Inc. MD5 Message-Digest Algorithm authentication. When a password is configured, MSDP receives only authenticated MSDP messages from its peers. All MSDP messages that fail TCP RSA Data Security, Inc. MD5 Message-Digest Algorithm authentication are dropped.

- Configure TCP RSA Data Security, Inc. MD5 Message-Digest Algorithm authentication on an MSDP peer using the command:

  `configure msdp peer [`*remoteaddr* `|` **all**`] password [none | {encrypted}` *tcpPassword*`] {vr` *vrname*`}`

- Remove the password using the command:

  `configure msdp peer {all |` *remoteaddr*`} password none`

  `The password displays in encrypted format and cannot be seen as simple text. Additionally, the password is saved in encrypted format.`

- Display the password in encrypted format using the command:

  `show msdp [peer {detail} | {peer}` *remoteaddr*`] {vr` *vr_name*`}`

## Policy Filters

You can configure a policy filter to control the flow of SA messages going to or coming from an *MSDP* peer. For example, policy filters can help mitigate state explosion during denial of service (DoS) or other attacks by limiting what is propagated to other domains using MSDP.

- Configure an incoming or outgoing policy filter for SA messages.

  `configure msdp peer [remoteaddr | all] sa-filter [in | out] [filter-name | none] {vr vr_name}`

- To remove a policy filter for SA messages, use the **none** keyword:

  `configure msdp [{peer} remoteaddr | peer all] sa-filter [in | out] none`

- Verify that a policy filter is configured on an MSDP peer.

  `show msdp [peer {detail} | {peer} remoteaddr] {vr vr_name}`

## SA Request Processing

You can configure the router to accept or reject SA request messages from a specified *MSDP* peer or all peers. If an SA request filter is specified, only SA request messages from those groups permitted are accepted. All others are ignored.

- Configure the router to accept SA request messages from a specified MSDP peer or all peers using the command:

  `enable msdp [{peer} remoteaddr | peer all] process-sa-request {sa-request-filter filter-name } {vr vr_name}`

- Configure the router to reject SA request messages from a specified MSDP peer or all peers using the command:

  `disable msdp [{peer} remoteaddr | peer all] process-sa-request {vr vrname}`

- Display configuration and run-time parameters about MSDP peers using the command:

  `show msdp [peer {detail} | {peer} remoteaddr] {vr vr_name}`

# MSDP Mesh-Groups

*MSDP* can operate in a mesh-group topology. A mesh-group limits the flooding of SA messages to neighboring peers. In a mesh-group, every MSDP peer must be connected to every other peer in the group. In this fully-meshed topology, when an SA message is received from a member of the mesh-group, the SA message is always accepted, but not flooded to other members of the same group. Because MSDP peers are connected to every other peer in the mesh-group, an MSDP peer is not required to forward SA messages to other members of the same mesh-group. However, SA messages are flooded to members of other mesh-groups. An MSDP mesh-group is an easy way to implement

inter-domain multicast, as it relaxes the requirement to validate looping of MSDP control traffic (that is, peer-RPF checking is not required). Consequently, SA messages do not loop in the network.

Note
We recommend that you configure anycast RP peers in a mesh topology.

- Configure an MSDP mesh-group using the command:

  `create msdp mesh-group` *mesh-group-name* {**vr** *vrname*}
- Remove an MSDP mesh-group using the command:

  `delete msdp mesh-group` *mesh-group-name* {**vr** *vrname*}
- Display information about an MSDP mesh-group using the command:

  `show msdp` [**mesh-group** {**detail**} | {**mesh-group**} *mesh-group-name*] {**vr** *vrname*}
- Configure a peer to be a member of an MSDP mesh-group using the command:

  `configure msdp peer` [*remoteaddr* | **all**] **mesh-group** [*mesh-group-name* | **none**] {**vr** *vrname*}
- Remove a peer from an MSDP mesh-group using the command:

  `configure msdp` [{**peer**} *remoteaddr* | **peer all**] **mesh-group none** {**vr** *vrname*}

## Anycast RP

Anycast RP is an application of *MSDP* that allows multiple RPs to operate simultaneously in a PIM-SM domain. Without anycast RP, multiple routers can be configured as candidate RPs, but at any point in time, only one router can serve as RP. Because anycast RP allows multiple RPs to be simultaneously active, anycast RP provides both load sharing and redundancy, as each RP serves the receivers that are closest to it in the network and can take over for additional receivers if another RP fails.

In an anycast RP topology, all RPs in a PIM-SM domain are configured with the same IP address on a loopback *VLAN*. The loopback VLAN IP address should have a 32-bit mask, so that it specifies a host address. All the routers within the PIM-SM domain select the nearest RP for each source and receiver. If the senders and receivers within the PIM-SM domain are distributed evenly, the number of senders that register with each RP is approximately equal.

Another requirement of the anycast RP topology is that MSDP must run on all RPs in the PIM-SM domain, so all RPs are also MSDP peers. We recommend that you configure an MSDP mesh connection between all MSDP peers in the domain.

Whenever any multicast source becomes active, this information is sent in an MSDP SA message to the other MSDP peers in the domain, announcing the presence of the source. If any RP within the domain fails, the IP routing protocol mechanism ensures that next available RP is chosen. If a sender registers with one RP and a receiver joins another RP, the information shared through MSDP enables PIM-SM to establish an SPT between the receiver and the source.

Note
We recommend that you configure anycast RP peers in a mesh topology.

The exchange of information in an anycast RP process works as follows:

- When the first-hop router sends a PIM Register message to the nearest RP, the PIM router checks to see if the nearest RP is the RP for the group.
- If the nearest RP is the RP for the group, an MSDP SA message is created and forwarded to the other MSDP peers.
- The MSDP SA message includes the configured originator ID, which is a mandatory configuration component.
- Each remote peer checks the RPF of the originator ID address and informs the PIM process on that remote router about active multicast sources.
- Remote receivers get data packets through the remote shared tree, and can then switch over to the SPT by sending join messages directly towards the source.

To configure anycast RP, do the following at each anycast RP router:

1. Create and configure a loopback VLAN using the commands:

   ```
   create vlan vlan_name {tag tag } {description vlan-description } {vr
   name }

   enable loopback-mode vlan vlan_name
   ```

2. Assign the anycast RP address to the loopback VLAN with a 32 bit subnet mask using the command:

   ```
   configure {vlan} vlan_name ipaddress [ipaddress {ipNetmask} | ipv6-
   link-local | {eui64} ipv6_address_mask]
   ```

   Note

   The anycast RP address must be unique to the loopback VLAN and be the same on all anycast RP peers. It must not match the router IP address, the PIM BSR address, or any other IP addresses used by the router or any other network devices.

3. Enable IP forwarding and IP multicast forwarding on the loopback VLAN using the commands:

   ```
   enable ipforwarding {ipv4 | broadcast} {vlan vlan_name}

   enable ipmcforwarding {vlan name}
   ```

4. Add the loopback VLAN into the unicast routing domain using the appropriate command for your unicast routing protocol:

   ```
   configure ospf add vlan vlan-name area area-identifier link-type [auto
   | broadcast | point-to-point] {passive}

   configure rip add vlan [vlan_name | all]

   configure isis add [vlan all | {vlan} vlan_name] area area_name {ipv4
   | ipv6}
   ```

5. Add the loopback VLAN into the PIM-SM domain and configure it as an RP using the commands:

   ```
   configure pim {ipv4 | ipv6} add vlan [vlan-name | all] {dense |
   sparse} {passive}

   configure pim {ipv4 | ipv6} crp static ip_address [none | policy]
   {priority [0-254]}
   ```

6. Enable MSDP and establish a peering relationship with similar anycast RP neighbors using the commands:

   ```
   create msdp peer remoteaddr {remote-as remote-AS} {vr vrname}

   configure msdp peer [remoteaddr | all] password [none | {encrypted}
   tcpPassword] {vr vrname}

   configure msdp peer remoteaddr description {peer-description} {vr
   vrname}

   enable msdp [{peer} remoteaddr | peer all] {vr vr_name}

   enable msdp {vr vrname}
   ```

7. Configure a unique originator ID for each anycast RP peer using the command:

   ```
   configure msdp originator-id ip-address {vr vrname}
   ```

## SA Cache

As an *MSDP* router learns of new sources either through a PIM-SM Source-Register (SR) message or SA message from its RPF peer, it creates an entry in SA cache (or refreshes the entry if it is already there) and forwards this information to its peers. These entries are refreshed by periodic SA messages received from the MSDP peers. If these entries are not refreshed within six minutes, they will time out. When a PIM-SM RP detects that the source is no longer available it informs MSDP, which in turn removes the SA information from the local database.

Caching makes it easy for local receivers to know immediately about inter-domain multicast sources and to initiate building a source tree towards the source. However, maintaining a cache is heavy both in CPU processing and memory requirements.

Note

Our implementation of MSDP does not support operating with local cache disabled.

- Remove an SA cache server.

  ```
  unconfigure msdp sa-cache-server {vr vrname}
  ```

  As MSDP uses the flood-and-join model to propagate information about sources, there is a restriction that no more than two advertisements per cache entry will be forwarded per advertisement interval. This is helpful in reducing an SA message storm and unnecessarily forwarding them to peers.

  By default, the router does not send SA request messages to its MSDP peers when a new member joins a group and wants to receive multicast traffic. The new member simply waits to receive SA messages, which eventually arrive.

- Configure the MSDP router to send SA request messages immediately to the MSDP peer when a new member becomes active in a group.

  ```
  configure msdp sa-cache-server remoteaddr {vr vr_name}
  ```

- Purge all SA cache entries.

  ```
  clear msdp sa-cache {{peer} remoteaddr | peer all} {group-address grp-addr} {vr vrname}
  ```

- Display the SA cache database.

  ```
  show msdp [sa-cache | rejected-sa-cache] {group-address grp-addr} {source-address src-addr} {as-number as-num} {originator-rp originator-rp-addr} {local} {peer remoteaddr} {vr vrname}
  ```

### Maximum SA Cache Entry Limit

You can configure a limit on the maximum number of SA cache entries that can be stored in the cache database. Once the number of SA cache entries exceeds the pre-configured limit, any newly received cache entries are discarded. You can configure the limit on a per-peer basis. By default, no SA message limit is set. The router can receive an unlimited number of SA entries from an *MSDP* peer.

- Configure a limit on the number of SA entries that can be stored in cache.

  ```
  configure msdp peer [remoteaddr | all] sa-limit max-sa {vr vr_name}
  ```

  To allow an unlimited number of SA entries, use 0 (zero) as the value for *max-sa*.

- Display the SA cache limit.

  ```
  show msdp [peer {detail} | {peer} remoteaddr] {vr vr_name}
  ```

## Redundancy

Because the peering relationship between *MSDP* peers is based on TCP connections, after a failover occurs the TCP connections need to be re-established again.

All SA cache entries learned from the old peering relationships must be flushed and relearned again on new TCP connections.

On a dual MSM system, MSDP runs simultaneously on both MSMs. During failover, the MSDP process on the active MSM receives and processes all control messages. MSDP on the standby MSM is in a down state, and doesn't receive, transmit, or process any control messages. If the active MSM fails, the MSDP process loses all state information and the standby MSM becomes active. However, the failover from the active MSM to the standby MSM causes MSDP to loses all state information and dynamic data, so it is not a hitless failover.

On fixed-configuration, stackable switches, an MSDP process failure brings down the switch.

## SNMP MIBs

*SNMP* MIB access is not supported for *MSDP*.

# Software Upgrade and Boot Options



This chapter provides instructions for upgrading ExtremeXOS software and firmware, including how to download and install a new image, perform a hitless upgrade, upload and download the configuration using TFTP, and synchronize nodes.

To upgrade your ExtremeXOS images and modules, start with ExtremeXOS Upgrade Process on page 1522.

To install modules on top of an existing ExtremeXOS image that will not be upgraded, see Installing a Modular Software Package on page 1535.

## ExtremeXOS Upgrade Process

Upgrading your ExtremeXOS image and modules uses the following process:

1. Backup all configuration files (see Creating a Backup Configuration File on page 1524).
2. Download ExtremeXOS and module package images from the Extreme Networks support portal (see Downloading a New Image on page 1528).

Note

Beginning with ExtremeXOS 16.1, there are two methods available for the upgrade process:

- `download url` *`filename`* - provides a means of downloading multiple files at the same time over TFTP, HTTP, or anonymouns FTP.
- `download image` *`filename`* - downloads a single .xos or .xmod file from a TFTP server.

3. Place the image files on a server that your switch can locate (see Installing a Core Image on page 1531).
4. Discover your inactive partition (see Finding the Inactive Partition on page 1530).
5. Use the `download image` command to install each image on the inactive partition (see Installing a Core Image on page 1531).
6. Reboot the switch (see Reboot Options on page 1532).

## Upgrade Example

We recommend upgrading ExtremeXOS images and any modular packages at the same time. The following example shows the ExtremeXOS image summitX-15.7.1.4.xos and the modular package summitX-15.7.1.4-ssh.xmod being installed together before rebooting the switch.

```
X460G2-48t-10G4.1 # show switch
SysName: X460G2-48t-10G4
SysLocation:
SysContact: support@extremenetworks.com, +1 888 257 3000
System MAC: 00:04:96:97:D1:84
System Type: X460G2-48t-10G4

SysHealth check: Enabled (Normal)
Recovery Mode: All
System Watchdog: Enabled

Current Time: Sat Mar 14 04:03:09 2015
Timezone: [Auto DST Disabled] GMT Offset: 0 minutes, name is UTC.
Boot Time: Wed Mar 11 06:00:50 2015
Boot Count: 296
Next Reboot: None scheduled
System UpTime: 2 days 22 hours 2 minutes 18 seconds

Current State: OPERATIONAL
Image Selected: secondary
Image Booted: secondary
Primary ver: 15.6.1.4
Secondary ver: 15.6.1.4

X460G2-48t-10G4.2 # download image 10.68.9.7 summitX-15.7.1.4.xos primary
Debug information files are present in internal-memory.
These files will be removed if you continue with download.
Do you want to continue with download and remove existing files from internal-memory?
(y/N) Yes
Do you want to install image after downloading? (y - yes, n - no, <cr> - cancel) Yes

Downloading to Switch.....................................................................
Installing to primary partition!

Installing to
Switch.....................................................................
Image installed successfully
This image will be used only after rebooting the switch!
X460G2-48t-10G4.4 # reboot
Are you sure you want to reboot the switch? (y/N) Yes
```

## Creating a Backup Configuration File

Before upgrading ExtremeXOS on a switch, it is highly recommended that you create a backup copy of the configuration. Creating a copy of the "Config Booted" configuration file, with the version number of the ExtremeXOS image included in the name of the backup copy, is the simplest method.

To do this:

1. Save the configuration to the database so the switch can reapply the configuration after the switch reboot using the `save` command.
2. Enter `y` at the prompt to save your changes.
3. Execute the `show switch` command to view the "Config Booted" file. The following output is displayed:

```
X460-24p.10 # show switch
SysName:          X460-24p
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:51:FE:E2
System Type:      X460-24p

SysHealth check:  Enabled (Normal)
Recovery Mode:    All
System Watchdog:  Enabled

Current Time:     Thu Sep  4 00:57:18 2014
Timezone:         [Auto DST Disabled] GMT Offset: 0 minutes, name is UTC.
Boot Time:        Wed Sep  3 20:07:11 2014
Boot Count:       402
Next Reboot:      None scheduled
System UpTime:    4 hours 50 minutes 7 seconds

Current State:    OPERATIONAL
Image Selected:   primary
Image Booted:     primary
Primary ver:      15.7.1.4
Secondary ver:    15.7.1.5

Config Selected:  ssh-privatekey.cfg
Config Booted:    ssh-privatekey.cfg
                  ssh-privatekey.cfg Created by ExtremeXOS version
15.7.1.5
            219131 bytes saved on Mon Jul 14 23:03:08 2014
```

4. Using the ExtremeXOS version shown in the "Config Booted:" line, enter the "copy" command as shown:

   `cp config selected config-booted-exos_ver`

   For example: `cp ssh-privatekey.cfg ssh-privatekey-15_7_1_5`

   Note

   For more information about version strings, see Understanding the Image Version String on page 1533.

For more information about ExtremeXOS configuration files, see Configuration Changes on page 1543.

In addition, ExtremeManagement and Ridgeline products have documented methods for backing up and restoring configuration for Extreme devices.

For ExtremeManagement procedures, visit https://extranet.extremenetworks.com/ (a valid customer account is required access this site) or access the online help available from the **Help** menu in the ExtremeManagement application.

For Ridgeline procedures, see the #unique_3189 section of the *Ridgeline Reference Guide*.

## Download URL Method

ExtremeXOS 16.1 provides an enhancement to the download CLI command so that it now accepts a URL as the name of the file to download.

Note
This feature is not available unless you are currently running ExtremeXOS 16.1 or higher.

URL protocols can be TFTP, HTTP, or anonymous FTP. The format of a URL is:

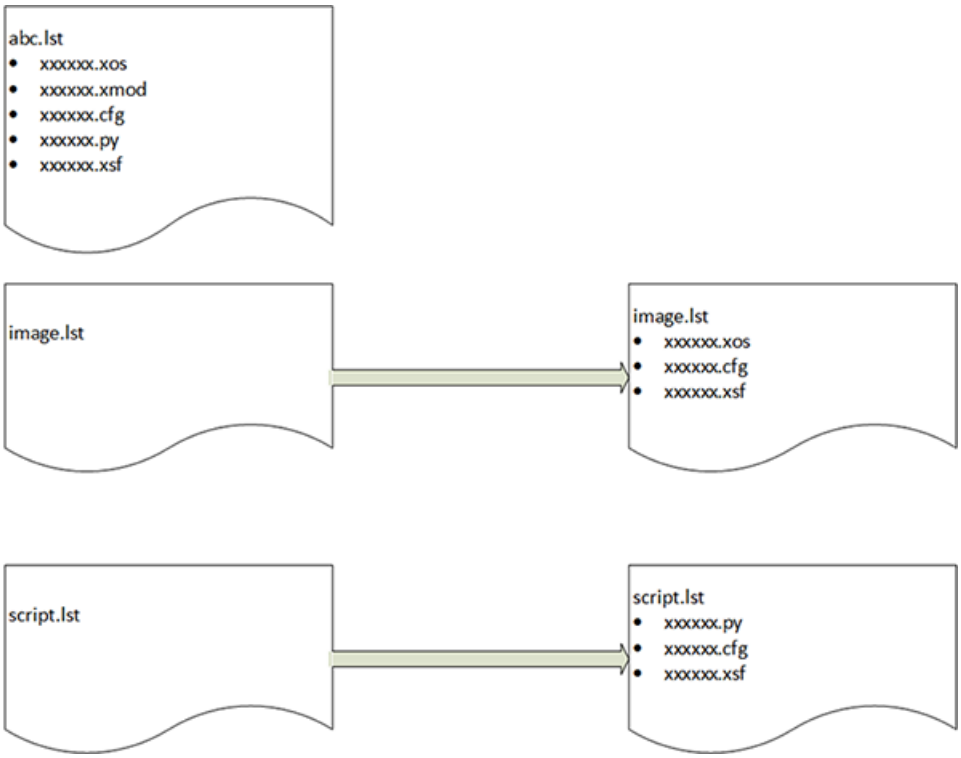
- http://10.10.10.1/filename.xos
- tftp://10.10.10.1/filename.xos
- ftp://10.10.10.1/filename.xmod

In addition to accepting a URL that ends in .xos or .xmod, the URL filename can now end in .lst.

A .lst file contains filenames at the same location as the .lst file URL and will be downloaded, and then installed one after the other. The .lst file method can define bundles of downloads for:

- Large image file sizes
- SSH installs with ExtremeXOS
- OpenFlow xmod and any additional .xmod modules install with ExtremeXOS
- Files ending in '.cfg', '.xsf', '.pol', '.xlic', '.py', '.ssh'
- Any additional .lst files so that you can create lists of lists
- Any bundling that makes it easier to download with a single command

The .lst file allows you to download .xos, .xmods, and other files at the same time. Once downloaded as a list, simply reboot and the files become active. This feature allows for better organization of downloads across switches.




All of the filenames in any .lst must be available at the same URL path provided in the download command.

Any .xos and .xmod files are installed to the inactive partition.

For any other file types, the /usr/local/cfg directory is used.

See the following for an example of .lst files

*Example*

An HTTP server on 10.68.9.7 for directory 16.1/xxxx/xxxx/release:

This example shows how the .lst file can contain filenames ending in .lst to get a list of lists (of lists etc...)

cat big.lst – big.lst contains other list file names:

xos.lst

xmod.lst

script.lst

cat xos.lst – xos.lst contains an EXOS image:

summitX-16.1.0.18.xos

cat xmod.lst – xmod.lst contains a number of .xmod filenames:

summitX-16.1.0.18-debug.xmod

summitX-16.1.0.18-esvt.xmod

summitX-16.1.0.18-LegacyCLI.xmod

summitX-16.1.0.18-openflow.xmod

summitX-16.1.0.18-reachnxt-1.8.1.8.xmod

summitX-16.1.0.18-techSupport.xmod

cat script.lst – script.lst contains a number of python scripts the user wants to download to a switch:

jsonrpc.py

jsontest.py

otst.py

ping.py

readvr.py

A single download command downloads all of the above files. Example uses big.lst:

```
X460G2-24t-10G4.1 # download url http://10.68.9.7/big.lst
http://10.68.9.7/xos.lst
Downloading http://10.68.9.7/summitX-16.1.0.18.xos

Downloading to
Switch.....................................................................................
.................................................
Installing to primary partition!

Installing to
Switch.....................................................................................
.....................................................................................
.....................
Image installed successfully
This image will be used only after rebooting the switch!

http://10.68.9.7:8080/xmod.lst
Downloading http://10.68.9.7/summitX-16.1.0.18-debug.xmod

Downloading to Switch.....
Installing to primary partition!

Installing to Switch.........................................
Image installed successfully
Downloading http://10.68.9.7/summitX-16.1.0.18-esvt.xmod

Downloading to Switch
Installing to primary partition!

Installing to Switch...
Image installed successfully
Downloading http://10.68.9.7/summitX-16.1.0.18-LegacyCLI.xmod
```

```
Downloading to Switch..
Installing to primary partition!

Installing to Switch.....................
Legacy CLI framework was Successfully Installed !!!

Image installed successfully
Downloading http://10.68.9.7/summitX-16.1.0.18-openflow.xmod

Downloading to Switch...
Installing to primary partition!

Installing to Switch.......................................
Image installed successfully
Downloading http://10.68.9.7/summitX-16.1.0.18-reachnxt-1.8.1.8.xmod

Downloading to Switch...
Installing to primary partition!

Installing to Switch....
Image installed successfully
Downloading http://10.68.9.7/summitX-16.1.0.18-techSupport.xmod

Downloading to Switch..
Installing to primary partition!

Installing to Switch..
Image installed successfully
http://10.68.9.7/script.lst
http://10.68.9.7/jsonrpc.py
http://10.68.9.7/jsontest.py
http://10.68.9.7/otst.py
http://10.68.9.7/ping.py
http://10.68.9.7/readvr.py
(pacman debug) X460G2-24t-10G4.2 #
```

## Downloading a New Image

The ExtremeXOS core image (.xos) file contains the executable code that runs on the switch and is preinstalled at the factory. An ExtremeXOS module image (.xmod) supplements the core image. Note that the version number of the core image and the module must match. As new versions of this image are released, you should upgrade both the software and modular packages running on your system.

On BlackDiamond 8800 series switches with two MSMs installed, you can upgrade the images without taking the switch out of service. Known as a hitless upgrade, this method of downloading and installing a new image minimizes network interruption, reduces the amount of traffic lost, and maintains switch operation. For more information, see Understanding Hitless Upgrade--Modular Switches Only on page 1537. For information about installing a new firmware image on a BlackDiamond 8800 series switch, see Upgrading the Firmware on page 1553.

To find and install image files:

1. Point your browser to http://esupport.extremenetworks.com and log in with your eSupport account.

If you do not have an eSupport account, you can request one with the **Request Web Login** link.

Note

You can also download release notes from the eSupport site. The *ExtremeXOS Release Notes* will always have the most current download and install instructions.

2. Find the ExtremeXOS image and all modules you wish to upgrade. Later you will install all files on the image at the same time.
3. Save the files to a TFTP server on the network. You can also store them on a compact flash card (installed in the compact flash slot of an MSM) or on a USB 2.0 device for BlackDiamond X chassis and Summit X460, X480, X670, X670G2, X670V, or X770 switches.

Note

Due to additional functionality and new platforms added in ExtremeXOS 15.6.1, the summitX .xos image is too large to be installed on Summit X480 switches. There is a new .xos image, summitX480-15.6.xx.yy.xos, which is to be used for Summit X480 and stacks that include one or more Summit X480 switches. This image unbundles operational diagnostics to provide an image that can be installed to Summit X480. There is a new xmod, summitX480-15.6.xx.yy-diagnostics.xmod, which can be used to update the operational diagnostics image to X480 and stacks that include X480 switches.

- Installing the summitX480 image over a previous release will leave the previous installation of the diagnostics image intact as it is stored separately from the main .xos image. This version can continue to be used to run diagnostics. The diagnostics xmod can be downloaded and installed in order to get the latest version. The diagnostics xmod can be installed to the active or standby partition and diagnostics can be used immediately. There is no need to reboot or any other action to complete the installation.
- If an X480 switch requires rescue recovery, the summitX-15.6.xx.yy xos package can be used and this will install the diagnostics code.

4. When you have all images downloaded and saved, proceed to Finding the Inactive Partition on page 1530.

### *Image Integrity Checking*

This feature adds digital signature verification in ExtremeXOS image and XMOD modules. Image integrity is checked against the digital signature before actual installation.

#### Digital Signature

Prior to ExtremeXOS 16.1, the ExtremeXOS image or XMOD modules downloaded to the switch were only protected by CRC check. This is sufficient for checking corrupted image. However, it cannot prevent malicious attacks. For example, an attacker can de-package the image, replace certain fields, recompute CRC and then repackage the image.

Digital signature is commonly used to demonstrate of the authenticity of a digital message, in this case, the image downloaded to the switch. Only images with digital signature validated on the switch can be installed. Otherwise, the installation should be aborted.

This features uses the Public Key Infrastructure (PKI) approach. Specifically, with the RSA algorithm, two keys, the private key and the public key are generated using openssl utility. The ExtremeXOS image

or XMOD module is digitally signed with the private key. The public key is installed on the swtich in the format of a X.509 certificate, which is verified before being used.

When building the image, the signature is computed for the ExtremeXOS image or XMOD module, then included in the final image provided to the customer. On the switch, during downloading process, the signature is verified against the image using the public key previously installed.

In order to deliver the public key to the customer securely, it is also digitally signed and it is distributed in the format of a X.509 certificate. In order to do so, another set of keys are generated to sign this certificate. A self-signed root certificate is also installed on the switch to verify the certifcate containing the image signing public key.

All these keys and certificates are generated offline and the private keys should be stored safely.

Transition from an image without signature to one with needs two steps, the first step is to download the EXOS image and install the public key certificates. At this time, the signature cannot be verified because there is no key to validate the image. But after the first installation, all subsequent downloaded images can be validated using the installed key.

The certificates are only included in the ExtremeXOS image; XMOD modules do not need to include certificates.

Downgrading from an ExtremeXOS version supporting digital signature to one that does not is allowed. No special handling is needed. A warning message is printed on the console to remind the user that the image is not digitally singed. The user does have the choice to either proceed with downgrading or not.

*Hash Verification*

Currently, each ExtremeXOS image or XMOD is posted along side with an *MD5 (Message-Digest algorithm 5)* hash checksum, which can be verified using any offline tools. As of ExtremeXOS 16.1, an enhancement now provides a stronger SHA256 hash checksum is generated for you to verify offline.

## Finding the Inactive Partition

A switch can store up to two core images: an active and inactive. When downloading a new image, you must select on which partition to install the new image. You must install the software image to the inactive partition, and must specify that partition while downloading the image to the switch.

To find the inactive partition:

1. From the CLI, enter the command `show switch` (or `show slot detail` on SummitStack).

```
X460-24p.10 # show switch
SysName:          X460-24p
SysLocation:
SysContact:       support@extremenetworks.com, +1 888 257 3000
System MAC:       00:04:96:51:FE:E2
System Type:      X460-24p

SysHealth check:  Enabled (Normal)
Recovery Mode:    All
System Watchdog:  Enabled

Current Time:     Thu Sep  4 00:57:18 2014
Timezone:         [Auto DST Disabled] GMT Offset: 0 minutes, name is UTC.
Boot Time:        Wed Sep  3 20:07:11 2014
```

```
Boot Count:       402
Next Reboot:      None scheduled
System UpTime:    4 hours 50 minutes 7 seconds

Current State:    OPERATIONAL
Image Selected:   primary
Image Booted:     primary
Primary ver:      15.7.1.4
Secondary ver:    15.7.1.5

Config Selected:  ssh-privatekey.cfg
Config Booted:    ssh-privatekey.cfg
                  ssh-privatekey.cfg Created by ExtremeXOS version
15.7.1.5
            219131 bytes saved on Mon Jul 14 23:03:08 2014
```

2. Locate the `Image Booted:` line. If it indicates `primary`, you will know to install the images on the secondary partition. If it indicates `secondary`, you will know to install the images on the primary partition.

   For help with versions and image filenames, refer to the following:

   - Understanding the Image Version String on page 1533
   - Image Filename Prefixes on page 1534
   - Software Signatures on page 1535

When you know the inactive partition, proceed to Installing a Core Image on page 1531.

## Installing a Core Image

Once .xos and .xmod files are stored on a network server (see Downloading a New Image on page 1528), download the image to the switch using the following procedure:

1. Use a login session on the master node (SummitStack or chassis) to verify which *virtual router (VR)* connects to your server. Use one of the following ping commands to confirm which virtual router reaches your server:

   ```
   ping vr vr-Mgmt host
   ping vr vr-Default host
   ```

   At least one of these commands must successfully reach your server for you to download the image. After verifying the virtual router that reaches your server, specify that virtual router when you download the image.

2. Download the new image to the switch using the command:

   ```
   download image [[hostname | ipaddress] filename {{vr} vrname} |
   memorycard filename] {partition} {msm slotid | slot slot number}
   ```

   **Note**

   If you configure the switch to write core dump (debug) files to the internal memory card and attempt to download a new software image, you might have insufficient space to complete the image download. If this occurs, you will need to move or delete the core dump files from the internal memory. You will be asked during the download process if you want to remove these files. For more information, see Understanding Core Dump Messages on page 1535.

3. Before the download begins, the switch asks if you want to install the image immediately after the download is finished. Enter `y` to install the image after download.

   If you enter `n` to install the image at a later time, the image is still downloaded and saved to the switch, but you must use the `install image` command when ready to install the image.
4. Install .xmod images to the inactive partition using the same command shown above.
5. After installing the .xos and all .xmod iamges, reboot the switch (see Reboot Options on page 1532).

For information about installing a new BootROM image on a BlackDiamond X or a Summit family switch, see Upgrading the BootROM on page 1552.

*Installing a Core Image with NMS*

Depending on your platform, you can use ExtremeManagement or Ridgeline to download the core image.

You can either load the new image to the NMS server on your network or configure it to automatically poll and download newly released images to the server. Follow the instructions described in the NMS documentation to appropriately configure the server for your network environment.

For more information about installing the NMS client and server, configuring it, and the platforms the NMS supports, refer to the NMS documentation that comes with the product or the documentation available from the Extreme Networks website at www.extremenetworks.com/documentation/.

## Reboot Options

*Reboot the Switch*

- To reboot the switch immediately, use the following command

  reboot

  **Note**
  The reboot occurs immediately and any previously schedule reboots are cancelled.

  On SummitStack, the `reboot` command reboots the *active* topology and can be run from the master node only.
- To schedule a time to reboot the switch, use the command:

  `reboot {time month day year hour min sec} {cancel} {msm slot_id} {slot slot-number }` (modular switches)

  `reboot {[time mon day year hour min sec] | cancel} {slot slot-number }` (SummitStack)

  The options use the following format:
  - **date** — in mm dd yyy format

- **time** — 24-hour clock in hh mm ss format

  Note
  When you configure a timed reboot of the switch, you can use the `show switch` command to see the scheduled time.

- To reboot the entire stack topology, use the command:

  ```
  reboot stack-topology {as-standby}
  ```

  Note
  This command can be run on any node and can be used to eliminate the dual master condition manually.

- To cancel a previously scheduled reboot, use the **cancel** option of the `reboot` command.

*Reboot the Management Module--Modular Switches Only*

To reboot a management module in a specific slot, rather than rebooting the switch, use the command:

```
reboot {time month day year hour min sec} {cancel} {msm slot_id} {slot slot-number }
```

Where the options are::

- **slot_id** — Specifies the slot where the module is installed
- **msm-a** — Specifies the MSM module installed in slot A
- **msm-b** — Specifies the MSM module installed in slot B

Note
When you configure a timed reboot of an MSM/MM, you can use the `show switch` command to see the scheduled time.

For more information about all of the options available with the reboot command, see the *ExtremeXOS 16.2 Command Reference Guide*.

*Reboot a Node in a SummitStack*

To reboot a single node in the SummitStack, use the command:

```
reboot {[time mon day year hour min sec] | cancel} {slot slot-number }
```

The `reboot slot command` works only on the active master node for other slots, and on an active non-master node for its own slot.

The reboot `stack-topology {as-standby}` command can also be used to reboot a single node by specifying the MAC address of the node.

The `reboot node-address` command reboots the specified node from any node.

## Understanding the Image Version String

The image version string contains build information for each version of ExtremeXOS.

You can use the `show version` and `show switch` commands to display the ExtremeXOS version running on your switch. The output will be structured as one of the following:

- ExtremeXOS Version <major>.<minor>.<sustaining>.<build>

  In the case of a patch release, the version structure is ExtremeXOS Version <major>.<minor>.<sustaining>.<build>-patchX-Y where X is the patch type and Y is the build number for the patch.
- <major>.<minor>.<sustaining>.<build>patchX-Y

The following table describes the image version fields.

**Table 153: Image Version Fields**

| Field | Description |
|---|---|
| major | Specifies the ExtremeXOS major version number. |
| minor | Specifies the ExtremeXOS minor version number. |
| sustaining | Specifies the ExtremeXOS sustaining version number. |
| build | Specifies the ExtremeXOS build number. |
| patch | Identifies a specific patch release. |

The `show version` command also displays information about the firmware (BootROM) images running on the switch. For more information, see Displaying the BootROM and Firmware Versions on page 1555.

## Image Filename Prefixes

The software image file can be an .xos file, which contains an ExtremeXOS core image, or an .xmod file, which contains an ExtremeXOS modular software package.

### *Filename Prefixes*

You can identify the appropriate image or module for your platform based on the filename of the image. The following table lists the filename prefixes for each platform.

| Platform | Filename Prefixes |
|---|---|
| BlackDiamond X8 series | bdX- |
| BlackDiamond 8810 | bd8800- |
| BlackDiamond 8806 | bd8800- |
| Summit family<br>EG4-200, EG4-400 | summitX- |
| Summit X430 | summitlite- |
| Summit X480 | summitX480-15.6.xx.yy.xos (X480 and stacks that include X480) |

For example, if you have a BlackDiamond X8 switch, download image filenames with the prefix bdX-. For additional installation requirements see the sections, Installing a Core Image on page 1531 and Installing a Modular Software Package on page 1535.

## Software Signatures

Each ExtremeXOS image contains a unique signature.

The BootROM checks for signature compatibility and denies an incompatible software upgrade. In addition, the software checks both the installed BootROM and software and also denies an incompatible upgrade.

## Understanding Core Dump Messages

If you configure the switch to write core dump (debug) files to the internal memory card and attempt to download a new software image, you might have insufficient space to complete the image download.

If this occurs, move or delete the core dump files from the internal memory. For example, if the switch supports a compact flash card or USB 2.0 storage device and space is available, transfer the files to the storage device. On switches without removable storage devices, transfer the files from the internal memory card to a TFTP server. This frees up space on the internal memory card while keeping the core dump files.

The switch displays a message similar to the following and prompts you to take action:

```
Core dumps are present in internal-memory and must be removed before this download can
continue.
(Please refer to documentation for the "configure debug core-dumps" command for
additional information)
Do you want to continue with download and remove existing core dumps? (y/n)
```

Enter `y` to remove the core dump files and download the new software image. Enter n to cancel this action and transfer the files before downloading the image.

For information about configuring and sending core dump information, see the `configure debug core-dumps` and `save debug tracefiles memorycard` commands.

# Installing a Modular Software Package

In addition to the functionality available in the ExtremeXOS core image, you can add functionality to your switch by installing modular software packages or feature packs. A complete listing of these feature packs can be found with their availability requirements in the *Feature License Requirements* document.

Modular software packages are contained in files named with the file extension .xmod, while the core images use the file extension .xos.

Modular software packages are built at the same time as core images and are designed to work in concert with the core image, so the <major>.<minor>.<patch> field of a modular software package must match the <major>.<minor>.<patch> field of the core image that it will be running with.

You can install a modular software package on the active or inactive partition. You would install on the active partition if you want to add the package functionality to the currently running core image without having to reboot the switch. You would install on the inactive partition if you want the functionality available after a switch reboot.

To install the package on the inactive partition, you use the same process that you use to install a new core image. Follow the process described in Installing a Core Image on page 1531. On BlackDiamond 8800 series switches, you can use hitless upgrade to install the package. See Understanding Hitless Upgrade--Modular Switches Only on page 1537 for more information.

Toctivate the installed modular software package either by rebooting the switch or by issuing the following command: `run update`

You can uninstall packages by issuing the following command:

`uninstall image` *`fname partition`* {**`msm`** *`slotid`*} {**`reboot`**}

**Note**

Do not terminate a process that was installed since the last reboot unless you have saved your configuration. If you have installed a software module and you terminate the newly installed process without saving your configuration, your module may not be loaded when you attempt to restart the process with the `start process` command.

## Upgrading a Modular Software Package

When Extreme Networks introduces a new core software image, a new feature pack or modular software package is also available. If you have a software module installed and upgrade to a new core image, you need to upgrade to the corresponding feature pack at the same time (see ExtremeXOS Upgrade Process on page 1522).

If you wish to upgrade your existing module software package *without* upgrading your ExtremeXOS image, two methods are available. Regardless of which method you choose, you must terminate and restart the processes associated with the software module.

1. Method One
   a. Terminate the processes associated with the software module using one of the following commands:

      `terminate process` *`name`* [**`forceful`** | **`graceful`**] {**`msm`** *`slot`*} (modular switches)

      `terminate process` *`name`* [**`forceful`** | **`graceful`**] {**`slot`** *`slot`*} (SummitStack)

   b. Download the software module from your TFTP server, compact flash card, or USB 2.0 storage device using the following command:

      `download` [**`url`** *`url`* {**`vr`** *`vrname`*} | `image` [**`active`** | **`inactive`**] [[*`hostname`* | *`ipaddress`*] *`filename`* {{**`vr`**} *`vrname`*} {**`block-size`** *`block_size`*} | **`memorycard`** *`filename`*] {*`partition`*} {**`msm`** *`slotid`*}

c. Activate the installed modular package, if installed on the active partition, using the following command: `run update`

d. Restart the processes associated with the software module using one of the following commands:

`start process` *name* {**msm** *slot*} (modular switches)

`start process` *name* {**slot** *slot*} (SummitStack)

2. Method Two

a. Download the software module from your TFTP server, compact flash card, or USB 2.0 storage device using the following command:

`download` [**url** *url* {**vr** *vrname*} | `image` [**active** | **inactive**] [[*hostname* | *ipaddress*] *filename* {{**vr**} *vrname*} {**block-size** *block_size*} | **memorycard** *filename*] {*partition*} {**msm** *slotid*}

b. Activate the installed modular package, if installed on the active partition, using the following command: `run update`

c. Terminate and restart the processes associated with the software module using one of the following commands:

`restart process` [**class** *cname* | *name* {**msm** *slot*}] (modular switches)

`restart process` [class *cname* | *name* {**slot** *slot*}] (SummitStack)

## Understanding Hitless Upgrade--Modular Switches Only

Hitless upgrade is a mechanism that allows you to upgrade the ExtremeXOS software running on the MSMs without taking the switch out of service.

Some additional benefits of using hitless upgrade include:

- Minimizing network downtime.
- Reducing the amount of traffic lost.

Although any method of upgrading software can have an impact on network operation, including interrupting Layer 2 network operation, performing a hitless upgrade can decrease that impact.

You must have two MSMs installed in your switch to perform a hitless upgrade. With two MSMs installed in the switch, one assumes the role of primary and the other assumes the role of backup. The primary MSM provides all of the switch management functions including bringing up and programming the I/O modules, running the bridging and routing protocols, and configuring the switch. The primary MSM also synchronizes its configurations with the backup MSM which allows the backup to take over the management functions of the primary.

Note

The software on the I/O modules is not updated during a hitless upgrade, only the software on the MSMs. The I/O module software is updated when the switch reboots or when a disabled slot is enabled.

If you download an image to the backup MSM, the image passes through the primary MSM before the image is downloaded to the backup MSM.

## Understanding the I/O Version Number

*BlackDiamond 8800 and BlackDiamond X Series Switches*

Each ExtremeXOS image comes bundled with an I/O module image and contains a unique upgrade compatibility version number, known as the I/O version number.

This number determines the relationship between the I/O module image and the ExtremeXOS image and their support for hitless upgrade. The I/O version number contains build information for each version of ExtremeXOS, including the major and minor version numbers, and the I/O version number.

Extreme Networks generates the I/O version number, and this number increases over time. Any modifications to the I/O module image after a major software release changes the I/O version number. For example, if Extreme Networks delivers a patch or service release that modifies the I/O module image, the I/O version number increases.

When you initiate a hitless upgrade by using the `run msm-failover {force}` command on the backup MSM, it checks the I/O version number to determine if a hitless upgrade is possible.

Depending on the currently running software, the switch performs, allows, or denies a hitless upgrade. The following describes the switch behavior:

- If the new ExtremeXOS image supports hitless upgrade and is compatible with the current running I/O module image, you can perform a hitless upgrade.
- If the new ExtremeXOS image supports hitless upgrade, is compatible with the current running I/O image but with a degradation of functionality, you can perform a hitless upgrade with caveats. The switch warns you that the upgrade is hitless; however, the downloaded software may result in a loss of new functionality. You can either continue the upgrade with the modified functionality or cancel the action.

  To prevent a loss in functionality, schedule time to take the switch offline to perform the upgrade; do not upgrade the software using hitless upgrade.
- If the new ExtremeXOS image supports hitless upgrade but is not compatible with the current running I/O module image (the I/O version numbers do not match), you cannot perform a hitless upgrade.

  The switch warns you that the upgrade may not be hitless. You can either continue the upgrade or cancel the action. If you continue the upgrade, the primary MSM downloads the new image to the I/O module and reboots.

  The following is a sample of the warning message displayed by the switch:

```
WARNING: Failover will not be hitless due to incompatible images. Traffic will be
interrupted.
Are you sure you want to failover? (y/n)
```

## Performing a Hitless Upgrade

The steps described in this section assume the following:

- You have received the new software image from Extreme Networks, and the image is on either a TFTP server, compact flash card, or USB 2.0 storage device. For more information, see Downloading a New Image on page 1528.
- You are running a version of ExtremeXOS that supports hitless upgrade.

Review the following list to confirm that your system supports hitless upgrade:

- BlackDiamond 8800 series switches—Both MSMs are running at least ExtremeXOS 11.4 and adhere to the minimum level of ExtremeXOS software required for all modules installed in the switch.

**Note**
Hitless upgrade for network login is not supported when upgrading from an earlier version of ExtremeXOS software to ExtremeXOS 12.1 or later. If a hitless upgrade is required, you must globally disable network login and re-enable it when the upgrade is complete.

### *Hitless Upgrade Caveats for BlackDiamond 8800 and BlackDiamond X Series Switches*

The following is a summary of hitless upgrade caveats for BlackDiamond 8800 and BlackDiamond X series switches:

- If you attempt a hitless upgrade between major releases, the switch warns you that the upgrade is not hitless. You can either continue the upgrade or cancel the action. If you continue the upgrade, the primary MSM downloads the new image to the I/O module and reboots them.

**Note**
If you are upgrading to a newer MSM module on a BlackDiamond 8800 or BlackDiamond X series switch, you must ensure you are running a version of ExtremeXOS that supports the newer MSM module before it is installed in the switch.

Hitless upgrade is only supported between versions of same major and minor releases (for instance, ExtremeXOS **15.5**.2.9 and **15.5**.3.4). Do not attempt to perform a hitless upgrade between different major or minor releases (for instance, ExtremeXOS 15.5.1 and 15.4.1).

### *Summary of Tasks*

To perform a hitless upgrade to install and upgrade the ExtremeXOS software on your system:

1. View the current switch information with the `show switch` command.
   a. Determine your selected and booted image partitions.
   b. Verify which MSM is the primary and which is the backup.
   c. Confirm that the MSMs are synchronized.
2. Select the inactive partition to download the image to (and the partition to boot from after installing the image).
3. Download and install the new ExtremeXOS core image on the backup MSM.
4. Reboot this MSM.
5. Verify that the backup MSM comes up correctly and that the MSMs are synchronized.
6. Initiate failover from the primary MSM to the backup MSM.
   The backup MSM now becomes the new primary MSM.
7. Verify that the MSMs come up correctly and that they are synchronized.
8. Download and install the new ExtremeXOS core image on the original primary MSM (new backup MSM).
9. Reboot this MSM.
10. Verify that the new backup MSM comes up correctly and that the MSMs are synchronized.
11. Initiate failover from the new primary MSM to the new backup MSM.
    This optional step restores the switch to the original primary and backup MSM.

12. Confirm that the failover is successful.

    This optional step confirms which MSM is the primary or the backup.

*Detailed Steps*

To perform a hitless upgrade to install and upgrade the ExtremeXOS software on your system:

1. View current switch information using the `show switch` command.

   Determine your selected and booted partition, verify which MSM is the primary and which is the backup, and confirm that the MSMs are synchronized.

   Output from this command indicates, for each MSM, the selected and booted images and if they are in the primary or the secondary partition. The selected image partition indicates which image will be used at the next reboot. The booted image partition indicates the image used at the last reboot. It is the active partition.

   The current state indicates which MSM is the primary (displayed as MASTER), which MSM is the backup (displayed as BACKUP), and if the backup MSM is synchronized with the primary MSM (displayed as In Sync).
2. Select the inactive partition to download the image to and download and install the new ExtremeXOS core image on the backup MSM using the following command:

   download [**url** *url* {**vr** *vrname*} | image [**active** | **inactive**] [[*hostname* | *ipaddress*] *filename* {{**vr**} *vrname*} {**block-size** *block_size*} | **memorycard** *filename*] {*partition*} {**msm** *slotid*}

   Note

   If the backup MSM is installed in slot B, specify `msm B`. If the backup MSM is installed in slot A, specify `msm A`.

   a. If you have an expired service contract and attempt to download a new image, you see the following message:

      ```
      Service contract expired, please renew it to be able to download the new software image.
      ```

      If you see this message, you must renew your service contract to proceed.
   b. When you have a current service contract, before the download begins the switch asks if you want to install the image immediately after the download is finished.
   c. After you download and install the software image on the inactive partition, you must reboot the MSM manually before you proceed. To reboot the switch, use the `reboot` command.
   d. Reboot only the backup MSM so the switch continues to forward traffic.
   e. If you install the image at a later time, use the `install image` command to install the software.
3. Verify that the backup MSM comes up correctly and that the MSMs are synchronized using the `show switch` command.

   The current state indicates which MSM is the primary (displayed as MASTER), which MSM is the backup (displayed as BACKUP), and if the backup MSM is synchronized with the primary MSM (displayed as In Sync).

4. Initiate failover from the primary MSM to the backup MSM using the following command:

   `run msm-failover {force}`

   When you failover from the primary MSM to the backup MSM, the backup becomes the new primary, runs the software on its active partition, and provides all of the switch management functions.

   If you have a BlackDiamond 8800 series switch and the new ExtremeXOS image supports hitless upgrade but is not compatible with the current running I/O module image (the I/O version numbers do not match), you cannot perform a hitless upgrade.

   The switch displays a warning message similar to the following:

   ```
   WARNING: The other MSM operates with a different version of I/O module image.
   If you continue with the MSM failover, all I/O modules will be reset.
   Are you sure you want to failover? (y/n)
   ```

   You can either continue the upgrade or cancel the action. If you continue the upgrade, the primary MSM downloads the new image to the I/O module and reboots.
5. Verify that the backup MSM comes up correctly and that the MSMs are synchronized using the `show switch` command.

   The current state indicates which MSM is the primary (displayed as MASTER), which MSM is the backup (displayed as BACKUP), and if the backup MSM is synchronized with the primary MSM (displayed as In Sync).
6. Select the inactive partition to download the image to and download and install the new ExtremeXOS core image on the new backup MSM (this was the original primary MSM) using the following command:

   `download [url url {vr vrname} | image [active | inactive] [[hostname | ipaddress] filename {{vr} vrname} {block-size block_size} | memorycard filename] {partition} {msm slotid}`

   **Note**

   If the new backup MSM is installed in slot B, specify `msm A`. If the new backup MSM is installed in slot A, specify `msm B`.

   a. Before the download begins, the switch asks if you want to install the image immediately after the download is finished.
   b. After you download and install the software image on the alternate partition, you need to reboot the MSM manually before you proceed. To reboot the switch, use the `reboot` command.
   c. Reboot only the backup MSM so the switch continues to forward traffic.
   d. If you install the image at a later time, use the `install image` command to install the software.
7. Verify that the new backup MSM comes up correctly and that the MSMs are synchronized using the `show switch` command.

   The current state indicates which MSM is the primary (displayed as MASTER), which MSM is the backup (displayed as BACKUP), and if the backup MSM is synchronized with the primary MSM (displayed as In Sync).

8. Optionally, initiate failover from the new primary MSM to the new backup MSM using the following command:

   ```
   run msm-failover {force}
   ```

   When you failover from the new primary MSM to the new backup MSM, this optional step restores the switch to the original primary and backup MSM.
9. Optionally, confirm that the failover is successful by checking the current state of the MSMs using the `show switch` command.

You can also perform a hitless upgrade on ExtremeXOS modular software packages (.xmod files). To perform a hitless upgrade of a software package, you must install the core software image first, and the version number of the modular software package must match the version number of the core image that it will be running with.

For more detailed information about modular software packages, see the *Feature License Requirements* document. To perform a hitless upgrade, follow the steps described in Performing a Hitless Upgrade on page 1538.

## Hitless Upgrade Examples

This section provides an example to perform a hitless upgrade on the BlackDiamond 8800 series switches.

Note

Before you begin, make sure you are running a version of ExtremeXOS that supports hitless upgrade. For more information, see the list Performing a Hitless Upgrade on page 1538.

### *Examples on the BlackDiamond 8800 Series Switches*

Using the assumptions described below, the following examples perform a hitless upgrade for a core software image on BlackDiamond 8800 series switches:

- You have received the new software image from Extreme Networks named bd8800-11.4.0.12.xos.
- You do not know your selected or booted partitions.
- You are currently using the primary partition.
- The image is on a TFTP server named tftphost.
- You are installing the new image immediately after download.
- The MSM installed in slot A is the primary.
- The MSM installed in slot B is the backup.
- You are running ExtremeXOS 11.4 or later on both MSMs.

#### Performing a Hitless Upgrade on the Alternate Partition

The following example shows the commands necessary to perform a hitless upgrade on the alternate partition.

In this example, the secondary partition is the inactive partition:

```
show switch
download image tftphost bd8800-11.4.0.12.xos secondary
show switch
```

```
reboot msm B
show switch
run msm-failover
show switch
```

After executing these commands, MSM B will be the master, and the secondary partition will be the active partition for both MSMs. The previously running software will reside on the inactive partition (now, the primary partition).

# Configuration Changes

## Image Configuration Overview

The configuration is the customized set of parameters that you have selected to run on the switch. As you make configuration changes, the new settings are stored in run-time memory. Settings that are stored in run-time memory are not retained by the switch when the switch is rebooted. To retain the settings and have them loaded when you reboot the switch, you must save the configuration to nonvolatile storage.

The switch can store multiple user-defined configuration files, each with its own filename. By default, the switch has two prenamed configurations: a primary and a secondary configuration. When you save configuration changes, you can select to which configuration you want the changes saved or you can save the changes to a new configuration file. If you do not specify a filename, the changes are saved to the configuration file currently in use. Or if you have never saved any configurations, you are asked to save your changes to the primary configuration.

Note
Configuration files have a .cfg file extension. When you enter the name of the file in the CLI, the system automatically adds the .cfg file extension.

If you have made a mistake or you must revert to the configuration as it was before you started making changes, you can tell the switch to use the backup configuration on the next reboot.

Each filename must be unique and can be up to 32 characters long. Filenames are also case sensitive. For information on filename restrictions, refer to the specific command in the *ExtremeXOS 16.2 Command Reference Guide*.

- First, save the configuration using the command:

  `save configuration {`**`primary`** `|` **`secondary`** `|` *`existing-config`* `|` *`new-config`*`}`

  Where the options are:

  - **`primary`** — Specifies the primary saved configuration
  - **`secondary`** — Specifies the secondary saved configuration
  - **`existing-config`** — Specifies an existing user-defined configuration (displays a list of available user-defined configuration files)
  - **`new-config`**—Specifies a new user-defined configuration
- You are prompted to save the changes. Enter `y` to save the changes or enter `n` to cancel the process.
- Next, use the configuration command:

  `use configuration [`**`primary`** `|` **`secondary`** `|` *`file_name`*`]`

Where the options are:

- **primary** — Specifies the primary saved configuration
- **secondary** — Specifies the secondary saved configuration
- **file_name** — Specifies an existing user-defined configuration (displays a list of available user-defined configuration files)

The configuration takes effect on the next reboot.

Note

If the switch is rebooted while in the middle of saving a configuration, the switch boots to factory default settings if the previously saved configuration file is overwritten. The configuration that is not in the process of being saved is unaffected.

## View a Configuration

- To view the current configuration, enter the command: on the switch.

  `show configuration {module-name} {detail}`

  You can also view just that portion of the configuration that applies to a particular module (for example, *SNMP (Simple Network Management Protocol)*) by using the *module-name* parameter.

  Beginning with ExtremeXOS 12.1, when you specify show configuration only, the switch displays configuration information for each of the switch modules excluding the default data.

  You can send output from the `show configuration {module-name} {detail}` command to the Extreme Networks Technical Support for problem-solving purposes. The output maintains the command line interface (CLI) format of the current configuration on the switch.

## Restore Factory Defaults

The following two procedures restore the switch to factory defaults and reboot the switch. Use the first procedure when you can log in to the switch and issue a CLI command and the second when you cannot.

- When you can log in to the switch, use the following command:

  `unconfigure switch`

  This command resets most of the configuration, with the exception of user-configured user accounts and passwords, the date, and the time. On SummitStack, the command also preserves stacking-specific parameters so the stack can be formed after reboot.

- Unset the currently selected configuration image, reset all switch parameters, and reboot the switch .

  `unconfigure switch all`

- If you cannot log in because the switch is in a continuous booting loop, use the following procedure:

  a. Reboot the switch while pressing the space bar. This puts the switch in Bootstrap mode.

  b. From the Bootstrap prompt, type `boot` and then press **[Enter]** and press the spacebar. This puts the switch in BootROM mode.

c. From the BootROM prompt, type `config none`.

   The following appears:
   ```
   Configuration selected: none
   ```

d. From the BootROM prompt, enter `reboot`. The switch reboots and restores the factory defaults.

## Uploading ASCII-Formatted Configuration Files

You can upload your current configuration in ASCII format to a TFTP server.

The uploaded ASCII file retains the CLI format and allows you do the following:

- View and modify the configuration using a text editor, and later download a copy of the file to the same switch or to one or more different switches.
- Send a copy of the configuration file to Extreme Networks Technical Support for problem-solving purposes.

### *Summary of Tasks*

The following summary describes only the CLI involved to transfer the configuration and load it on the switch; it is assumed that you know how to modify the configuration file with a text editor. As previously described, to use these commands, use the .xsf file extension, as these steps are not applicable to configurations that use the .cfg file extension.

To work with an ASCII-formatted configuration file, complete the following tasks:

1. Upload the configuration to a network TFTP server using the following command:

   `upload configuration [`*`hostname`* `|` *`ipaddress`*`] `*`filename`* `{`**`vr`** *`vr-name`*`} {`**`block-size`** *`block_size`*`}`

   After the configuration file is on the TFTP server, use a text editor to enter the desired changes, and rename the file if necessary so it has the .xsf extension.
2. Download the configuration from the TFTP server to the switch using one of the `tftp` and `tftp get` commands.
3. Verify the configuration file is on the switch using the `ls` command
4. Load and restore the new configuration file on the switch using the following command:

   `load script `*`filename`* `{`**`arg1`**`} {`**`arg2`**`} ... {`**`arg9`**`}`
5. Save the configuration to the configuration database so the switch can reapply the configuration after switch reboot using the following command:

   `save configuration {`**`primary`** `|` **`secondary`** `|` *`existing-config`* `|` *`new-config`*`}`

   When you save the configuration file, the switch automatically adds the .cfg file extension to the filename. This saves the ASCII configuration as an XML-based configuration file.

   **Note**
   Configuration files are forward compatible only and not backward compatible. That is, configuration files created in a newer release, such as ExtremeXOS 15.5, might contain commands that do not work properly in an older release, such as ExtremeXOS 15.3.

*Upload the ASCII Configuration File To a TFTP Server*

To upload the current switch configuration as an ASCII-based file to the TFTP server, use the `upload configuration` command and save the configuration with the .xsf file extension.

For example, to transfer the current switch configuration as an ASCII-based file named meg_upload_config1.xsf to the TFTP server with an IP address of 10.10.10.10, do the following:

```
upload configuration 10.10.10.10 meg_upload_config1.xsf
```

If you successfully upload the configuration to the TFTP server, the switch displays a message similar to the following:

```
Uploading meg_upload_config1.xsf to 10.10.10.10 ... done!
```

*Download the ASCII Configuration File to the Switch*

To download the configuration from the TFTP server to the switch, use the `tftp` or `tftp get` command.

For example, to retrieve the configuration file named meg-upload_config1.xsf from a TFTP server with an IP address of 10.10.10.10, you can use one of the following commands:

```
tftp 10.10.10.10 -g -r meg_upload_config1.xsf
tftp get 10.10.10.10 meg_upload_config1.xsf
```

If you successfully download the configuration to the switch, the switch displays a message similar to the following:

```
Downloading meg_upload_config1.xsf to switch... done!
```

*Verify that the ASCII Configuration File is on the Switch*

To confirm that the ASCII configuration file is on the switch, use the `ls` command. The file with an .xsf extension is the ASCII configuration.

The following sample output contains an ASCII configuration file:

```
-rw-r--r--    1 root     0           98362 Nov  2 13:53 Nov022005.cfg
-rw-r--r--    1 root     0          117136 Dec 12 12:56 ridgeline.cfg
-rw-r--r--    1 root     0              68 Oct 26 11:17 mcastgroup.pol
-rw-r--r--    1 root     0           21203 Dec 13 15:40 meg_upload_config1.xsf
-rw-r--r--    1 root     0          119521 Dec  6 14:35 primary.cfg
-rw-r--r--    1 root     0           96931 Nov 11 11:01 primary_11_11_05.cfg
-rw-r--r--    1 root     0           92692 Jul 19 16:42 secondary.cfg
```

*Load the ASCII Configuration File*

After downloading the configuration file, you must load the new configuration on the switch.

To load and restore the ASCII configuration file, use the command:

`load script` *`filename`* {**`arg1`**} {**`arg2`**} **`...`** {**`arg9`**}

After issuing this command, the ASCII configuration quickly scrolls across the screen.

The following is an example of the type of information displayed when loading the ASCII configuration file:

```
script.meg_upload_config1.xsf.389 # enable snmp access
script.meg_upload_config1.xsf.390 # enable snmp traps
```

```
script.meg_upload_config1.xsf.391 # configure mstp region purple
script.meg_upload_config1.xsf.392 # configure mstp revision 3
script.meg_upload_config1.xsf.393 # configure mstp format 0
script.meg_upload_config1.xsf.394 # create stpd s0
```

Instead of entering each command individually, the script runs and loads the CLI on the switch.

*Save the Configuration*

After you load the configuration, save it to the configuration database for use by the switch. This allows the switch to reapply the configuration after a switch reboot.

- To save the configuration, use the command;

  `save configuration {`**`primary`** | **`secondary`** | *`existing-config`* | *`new-config`*`}`

  When you save the configuration file, the switch automatically adds the .cfg file extension to the filename. This saves the ASCII configuration as an XML-based configuration file.

  You can use any name for the configuration.

  For example, after loading the file `meg_upload_config1.xsf`, you need to save it to the switch.

To save the configuration as configuration1.cfg, enter the command:

```
save configuration configuration1
```

*Using Autoconfigure and Autoexecute Files*

Two features allow you automatically execute scripts that can manage the switch configuration.

**Autoconfigure:**

Configuration commands placed in the default.xsf file are executed by the switch as it comes up and is unable to find its usual configuration file or if the switch is unconfigured or if the configuration file cannot be determined due to a corrupt NVRAM.

This returns the switch to some basic configuration. When default.xsf is executed, the `show switch` command shows default.xsf as the booted configuration file.

The default.xsf file can have any CLI commands as long as they are all executed within 500 seconds.

The script is aborted when the commands are not executed within that time. When the file is loaded, the results can be seen by executing the `show script output default` command.

**Autoexecute:**

Configuration commands placed in the autoexec.xsf file are executed after a switch loads its configuration.

The file is not executed when a default.xsf file has been executed. Use the file to execute commands after a switch is up and running and also to revert changes made to the configuration by UPM scripts that run persistent commands. The commands must be executed within 500 seconds or the script execution is aborted.

When an autoexec.xsf file is executed. the results can be seen by executing the `show script output autoexec` command.

## Using TFTP to Upload the Configuration

You can upload the current configuration to a TFTP server on your network. Using TFTP, the uploaded configuration file retains your system configuration and is saved in XML format. This allows you to send a copy of the configuration file to Extreme Networks Technical Support for problem-solving purposes.

- To view your current switch configuration, use the `show configuration {module-name} {detail}` command available on your switch. Do not use a text editor to view or modify your XML-based switch configuration files.

  To view your current switch configuration in ASCII-format, see Uploading ASCII-Formatted Configuration Files on page 1545 for more information about uploading and downloading ASCII-formatted configuration files.

  For more information about TFTP, see Using the Trivial File Transfer Protocol on page 52.
- To upload the configuration from the switch to a TFTP server, you can use either the `tftp` or the `tftp put` command:

  `tftp [ ip-address | host-name ] { -v vr_name } { -b block_size } [ -g | -p ] [ -l local-file { -r remote-file } | -r remote-file { -l local-file } ]`

  Where the options are:
  - **host-name** — Specifies the host name of the TFTP server
  - **ip-address** — Specifies the IP address of the TFTP server
  - **-p** — Puts the specified file from the local host and copies it to the TFTP server
  - **-l** *local-file* — Specifies the name of the configuration file that you want to save to the TFTP server
  - **-r** *remote-file* — Specifies the name of the configuration file on the TFTP server

  `tftp put [ ip-address | host-name] {vr vr_name} {block-size block_size}local-file { remote-file}`

  Where the options are:
  - **put** —P uts the specified file from the local host and copies it to the TFTP server
  - **host-name** — Specifies the host name of the TFTP server
  - **ip-address** — Specifies the IP address of the TFTP server
  - *local-file* — Specifies the name of the configuration file that you want to save to the TFTP server
  - *remote-file* — Specifies the name of the configuration file on the TFTP server

If you upload a configuration file and see the following message:

```
Error: No such file or directory
```

Check to make sure that you entered the filename correctly, including the .cfg extension, and that you entered the correct host name or IP address for the TFTP server.

If your upload is successful, the switch displays a message similar to the following:

```
Uploading megtest1.cfg to TFTPhost ... done!
```

You can also upload the current configuration in ASCII format from the switch to a TFTP server on your network. For more information, see Uploading ASCII-Formatted Configuration Files on page 1545.

## Using TFTP to Download the Configuration

You can download previously saved XML formatted XOS configuration files from a TFTP host to the switch to modify the switch configuration. Do not use a text editor to view or modify your switch configuration files; modify your switch configurations directly in the CLI.

To view your current switch configuration in ASCII-format, see Uploading ASCII-Formatted Configuration Files on page 1545 for more information about uploading and downloading ASCII-formatted configuration files.

For more information about TFTP, see Using the Trivial File Transfer Protocol on page 52.

Note

By default, if you transfer a file with a name that already exists on the system, the switch prompts you to overwrite the existing file. For more information, see the `tftp get` command in the *ExtremeXOS 16.2 Command Reference Guide*.

- To download the configuration from a TFTP host to the switch, you can use either the tftp or the tftp get command:

  `tftp [ ip-address | host-name ] { -v vr_name } { -b block_size } [ -g | -p ] [ -l local-file { -r remote-file } | -r remote-file { -l local-file } ]`

  Where the options are:

  - **host-name** — Specifies the host name of the TFTP server
  - **ip-address** — Specifies the IP address of the TFTP server
  - **-p** — Puts the specified file from the local host and copies it to the TFTP server
  - **-l** *local-file* — Specifies the name of the configuration file that you want to save to the TFTP server
  - **-r** *remote-file* — Specifies the name of the configuration file on the TFTP server

  `tftp get [ ip-address | host-name] { vr vr_name } { block-size block_size } remote-file local-file} {force-overwrite}`

  Where the optons are:

  - **get** — Gets the specified file from the TFTP server and copies it to the local host
  - *host-name* — Is the host name of the TFTP server
  - *ip-address* — Is the IP address of the TFTP server
  - *remote_file* — Specifies the name of the configuration file that you want to retrieve from the TFTP server
  - *local-file* — Specifies the name of the configuration file on the switch
  - **force-overwrite** — Specifies the switch to automatically overwrite an existing file

  If you download a configuration file and see the following message:

  ```
  Error: Transfer timed out
  ```

  Make sure that you entered the filename correctly, including the .cfg extension, and that you entered the correct host name or IP address for the TFTP server.

  If your download is successful, the switch displays a message similar to the following:

```
Downloading megtest2.cfg to switch... done!
```

Configurations are downloaded and saved into the switch nonvolatile memory. The configuration is applied after you reboot the switch.

If the configuration currently running in the switch does not match the configuration that the switch used when it originally booted, an asterisk (*) appears before the command line prompt when using the CLI.

You can also download the current configuration in ASCII format from a TFTP server on your network to the switch. For more information, see Uploading ASCII-Formatted Configuration Files on page 1545.

## Synchronizing Nodes--Modular Switches and SummitStack Only

Before synchronizing nodes on a modular chassis or nodes on a SummitStack, review the following list to confirm that your platform and both installed MSMs/MMs are running software that supports the `synchronize` command:

- BlackDiamond 8800 series switches with a mix of BlackDiamond 8000 a-, c-, e-, xl-, and xm-series modules installed—Both MSMs are running ExtremeXOS 11.5 or later.
- SummitStack—all nodes are active nodes and running ExtremeXOS 12.0 or later.

On a dual MSM system or a SummitStack with redundancy, you can take the primary node configurations and images and replicate them on the backup node using the `save configuration` command.

Caution
During a synchronization on a modular chassis, half of the switch fabric is lost. On a SummitStack, the active stack will briefly alternate between a ring and daisy-chain topology. When the primary node finishes replicating its configurations and images to the backup node, the full switch fabric or the stack ring is restored.

In addition to replicating the configuration settings and images, this command also replicates which configuration or image the node should use on subsequent reboots. This command does not replicate the run-time configuration. You must use the `save configuration` command to store the run-time configuration first.

On a SummitStack, you can synchronize an active node in the stack with the master node using the following command:

```
synchronize {slot slotid}
```

### Additional Behavior on the BlackDiamond 8800 Series Switches Only

On the BlackDiamond 8800 series switches, the I/O ports on the backup MSM go down when you synchronize the MSMs.

When the primary MSM finishes replicating its configurations and images to the backup MSM, the I/O ports on the backup MSM come back up.

## Automatic Synchronization of Configuration Files

On a dual MSM/MM (node) modular chassis or on a SummitStack where redundancy is in use, ExtremeXOS automatically synchronizes all of the configuration files from the primary node to the backup node if the switch detects that the backup node's configuration file contents are different from the primary node.

You cannot configure this behavior.

The switch deletes the old configuration files on the backup node only upon a successful file synchronization. If an error occurs, the switch does not delete the old configuration files on the backup node. For example, if you install a backup node that contains different configuration files from the primary node, the old configuration files are deleted after a successful bootup of the backup node.

To see a complete listing of the configuration files on your system, use the `ls` command.

For more detailed information, see Replicating Data Between Nodes on page 56.

# Accessing the Bootloader

The Bootloader of the switch initializes certain important switch variables during the boot process. In the event the switch does not boot properly, some boot option functions can be accessed through the Bootloader.

Interaction with the Bootloader is required only under special circumstances and should be done only under the direction of Extreme Networks Customer Support. The necessity of using these functions implies a nonstandard problem which requires the assistance of Extreme Networks Customer Support.

1. Attach a serial cable to the console port of the switch.
2. Attach the other end of the serial cable to a properly configured terminal or terminal emulator, power cycle the switch, and press the spacebar key on the keyboard of the terminal during the bootup process.

Note

On the BlackDiamond X series switches, press and hold the spacebar key to enter the bootROM (actually a BIOS) as soon as you see CF card tested OK on the screen.

On BlackDiamond 8800 series switches, when you see the BootROM banner, press the spacebar key to get into the Bootloader application.

On Summit family switches, when you see the Bootloader banner, press the spacebar key to get into the Bootloader application.

As soon as you see the `BOOTLOADER>` prompt (Summit family switches) or the `BootRom ->` prompt (BlackDiamond X series), release the spacebar. You can issue a series of commands to:

- View the installed images.
- Select the image to boot from.
- Load a recovery image over the management port.
- To see a list of available commands or additional information about a specific command, enter `h` or type `help`.

The following describes some ways that you can use the Bootloader:

- Viewing images — To display a list of installed images, use the `show image` command.
- Selecting an image — To change the image that the switch boots from in flash memory, use the `boot {image number}` command. If you specify **`image number`**, the specified image is booted. If you do not specify an image number, the default image is booted.

3. To exit the Bootloader, use the `boot` command. Specifying boot runs the currently selected ExtremeXOS image.

# Upgrading the BootROM

## Summit Family Switches and SummitStack Only

The Summit family switches have a two-stage BootROM. The first stage, called bootstrap, does basic initialization of the switch processor and will load one of two second-stage bootloaders (called primary and secondary).

If the switch does not boot properly, both the bootstrap and the bootloader allows the user to access the boot options using the CLI.

If necessary, the bootloader can be updated after the switch has booted, using TFTP. You can upgrade the BootROM from a TFTP server on the network after the switch has booted and only when asked to do so by an Extreme Networks technical representative. For information about loading an image to a TFTP server and verifying which *VR* connects to your TFTP server, see Installing a Core Image on page 1531.

- On a SummitStack, the BootROM can be centrally upgraded.

  Use the command above on the primary (Master) stack node to download a BootROM to all stacking nodes.
- To download to a single stacking node, you need to specify the *`slot`* parameter:

  `download bootrom [`*`ipaddress | hostname`*`] `*`filename`*` {`**`slot`**` `*`slot-number`*`} {{`**`vr`**`} `*`vrname`*`}`

Note

User-created VRs are supported only on the platforms listed for this feature in the *Feature License Requirements* document.

### *Accessing the Bootstrap CLI on the Summit Family Switches*

The bootstrap CLI contains commands to support the selection of which bootloader to use. Interaction with the bootstrap is required only under special circumstances and should be done only under the direction of Extreme Networks Customer Support.

To access the bootstrap CLI:

1. Attach a serial cable to the serial console port of the switch.
2. Attach the other end of the serial cable to a properly configured terminal or terminal emulator.
3. Power cycle or reboot the switch.

4. As soon as you see the Bootstrap Banner, press the spacebar.

   The `BOOTSTRAP>` prompt appears on the screen.

   Note

   If you accidentally enter the bootstrap CLI when you want to enter the Bootloader, at the `BOOTSTRAP>` prompt enter the `boot` command.

For detailed information and instructions on accessing the bootloader, see Accessing the Bootloader on page 1551.

# Upgrading the Firmware

## BlackDiamond 8800 Series Switches Only

Firmware images are bundled with ExtremeXOS software images. The bundled firmware images include Microcontroller binaries, System FPGA images and BootROM images for the MM, I/O, and Fabric modules.

ExtremeXOS automatically compares the existing firmware image flashed into the hardware with the firmware images bundled with the ExtremeXOS image when you download and install a new version of ExtremeXOS.

After a firmware image upgrade, messages are sent to the log. You can configure the switch to automatically upgrade the firmware when a different image is detected, or you can have the switch prompt you to confirm the upgrade process.

- Configure the switch's behavior during a firmware upgrade by using the command:

  `configure firmware [auto-install | install-on-demand]`

  Where the ptions are:

  - **`auto-install`** — Specifies ExtremeXOS to automatically upgrade the firmware if the software detects a newer firmware image is available. The switch does not prompt you to confirm the firmware upgrade.
  - **`on-demand`** — Specifies the switch to prompt you to upgrade the firmware when ExtremeXOS determines that a newer firmware image is available. This is the default behavior.

  You can use the `install firmware {force}` command to install the firmware bundled with the ExtremeXOS image. To install the new BootROM and firmware, wait until the `show slot` command indicates the MM, I/O, and Fabric modules are operational. When the modules are operational, use the `install firmware` command.

  If the bundled firmware image is newer than the existing firmware image, the switch prompts you to confirm the upgrade.
- Enter `y` to upgrade the firmware.

- Enter `n` to cancel the firmware upgrade for the specified hardware and continue scanning for other hardware that needs to be upgraded.
- Enter `cr` to cancel the upgrade.

  During the firmware upgrade, the switch also prompts you to save your configuration changes to the current, active configuration. Enter `y` to save your configuration changes to the current, active configuration. Enter `n` if you do not want to save your changes.

  The new BootROM and firmware overwrite the older versions flashed into the hardware. A reboot is required to load the newly installed firmware. However, this does not need to be done immediately after a firmware upgrade. Use the reboot command to reboot the switch and activate the new BootROM and firmware. During the firmware upgrade, do not turn off or disrupt the power to the switch. If a power interruption occurs, the firmware may be corrupted and need to be recovered. ExtremeXOS automatically attempts to recover corrupted firmware; however, in some situations user intervention is required.

  *PoE (Power over Ethernet)* firmware is always automatically upgraded or downgraded to match the operational ExtremeXOS code image. This configuration is not applicable to PoE firmware.

## BlackDiamond X8 Series Switches Only

Firmware images are bundled with ExtremeXOS software images. The bundled firmware images include microcontroller binaries, system FPGA images, and BootROM images for the MM, I/O, and Fabric modules. When you download and install a new version of ExtremeXOS, ExtremeXOS automatically compares the existing firmware image against the firmware images bundled with the ExtremeXOS image.

After a firmware image upgrade, messages are sent to the log. You can set the switch to automatically upgrade the firmware when a different image is detected or have the switch prompt you to confirm the upgrade.

- Set the switch's behavior during a firmware upgrade using the command:

  `configure firmware [`**`auto-install`**` | `**`install-on-demand`**`]`

  - **`auto-install`**—Automatically upgrade the firmware if a newer firmware image is available. You are not prompted to confirm the firmware upgrade.
  - **`install-on-demand`**—Prompts you to upgrade the firmware when a newer firmware image is available (default behavior).
- To install the firmware bundled with the ExtremeXOS image, use the `install firmware {`**`force`**`}` command.
- To install the new BootROM and firmware, wait until the show slot command indicates the MM, I/O and Fabric modules are operational. When the modules are operational, use the `install firmware` command.

  If the bundled firmware image is newer than the existing firmware image, the switch prompts you to confirm the upgrade.
- Type `y` to upgrade the firmware.
- Type `n` to cancel the firmware upgrade for the specified hardware, and continue scanning for other hardware that needs to be upgraded.

- Press **[Enter]** to cancel the upgrade.
  During the firmware upgrade, the switch also prompts you to save your configuration changes to the current, active configuration.
- To save your configuration changes to the current, active configuration, type `y`. To discard your changes, type `n`.
  The new BootROM and firmware overwrite the older versions.

You must reboot the switch to load the newly installed firmware. However, you do not need to do this immediately after a firmware upgrade. Use the `reboot` command to reboot the switch and activate the new BootROM and firmware. During the firmware upgrade, do not turn off or disrupt the power to the switch. If power is lost, the firmware may be corrupted and need to be recovered. ExtremeXOS automatically attempts to recover corrupted firmware; however, in some situations user intervention is required.

## Displaying the BootROM and Firmware Versions

To display the BootROM (firmware) version on the switch and on all of the modules and PSU controllers installed in a modular switch, use the `show version` command.

The following is sample output from the Summit series switch:

```
Switch    : 800132-00-02 0512G00636  Rev 2.0 BootROM: 1.0.0.6    IMG: 11.4.0.15
XGM-2xn-1 :
Image   : ExtremeXOS version 11.4.0.15 v1140b15 by release-manager
on Fri Dec 30 11:05:42 PST 2005
BootROM : 1.0.0.6
```

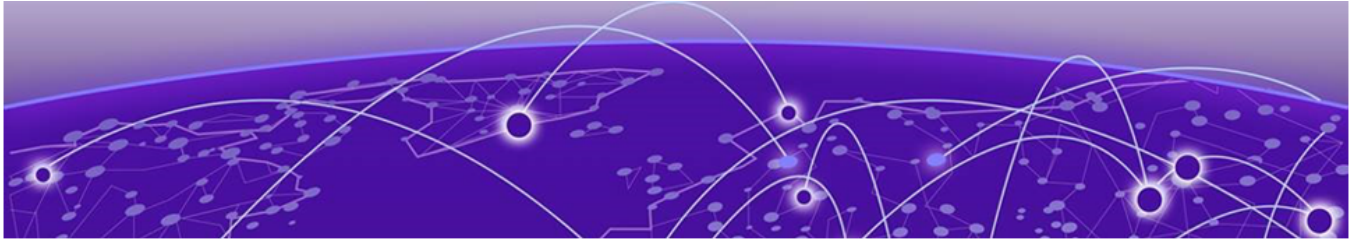
The following is sample output from a BlackDiamond X8 series switch:

```
BD-X8.8 # show version detail
Chassis     : 800427-00-00 00000000000 Rev 0.0
Slot-1      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-2      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-3      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-4      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-5      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-6      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-7      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
Slot-8      : 800439-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
FM-1        : 800433-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
FM-2        : 800433-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
FM-3        : 800433-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
FM-4        : 800433-00-00 00000000000 Rev 0.0 BootROM: 1.0.0.5    IMG: 15.1.0.23  FPGA :
0.0.32.0
MM-A        : 800432-00-00 1130G-00826 Rev 0.0 BootROM: 1.0.0.2    IMG: 15.1.0.23  FPGA :
0.1.18.0
MM-B        :
PSUCTRL-1   : 450357-00-00 00000000000 Rev 0.0
```

```
PSUCTRL-2   : 450357-00-00 00000000000 Rev 0.0
FanTray-1   : 450350-00-00 00000000000 Rev 0.0 BootROM: 1.0.2.5
FanTray-2   : 450350-00-00 00000000000 Rev 0.0 BootROM: 1.0.2.5
FanTray-3   : 450350-00-00 00000000000 Rev 0.0 BootROM: 1.0.2.5
FanTray-4   : 450350-00-00 00000000000 Rev 0.0 BootROM: 1.0.2.5
FanTray-5   : 450350-00-00 00000000000 Rev 0.0 BootROM: 1.0.2.5
PSU-1       : H2500A2-EX  4300-00212 1109X-88834 Rev 1.0
PSU-2       : H2500A2-EX  4300-00212 1109X-88758 Rev 1.0
PSU-3       :
PSU-4       :
PSU-5       : H2500A2-EX  4300-00212 1109X-88787 Rev 1.0
PSU-6       : H2500A2-EX  4300-00212 1109X-88831 Rev 1.0
PSU-7       :
PSU-8       :
Image   : ExtremeXOS version 15.1.0.23 v1510b23 by release-manager
on Fri Dec 16 12:05:18 EST 2011
BootROM : 1.0.0.2
Diagnostics : 1.8 (MM), 1.6 (I/O and FM)
```

The following is sample output from a BlackDiamond 8800 series switch:

```
Chassis   : 800129-00-02 04344-00039 Rev 2.0
Slot-1    : 800114-00-04 04364-00021 Rev 4.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
Slot-2    : 800115-00-02 04344-00006 Rev 2.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
Slot-3    : 800113-00-04 04354-00031 Rev 4.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
Slot-4    :
Slot-5    : 800112-00-03 04334-00040 Rev 3.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
Slot-6    : 800112-00-03 04334-00004 Rev 3.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
Slot-7    :
Slot-8    :
Slot-9    :
Slot-10   :
MSM-A     : 800112-00-03 04334-00040 Rev 3.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
MSM-B     : 800112-00-03 04334-00004 Rev 3.0 BootROM: 1.0.1.7    IMG: 11.4.0.23
PSUCTRL-1 : 450117-00-01 04334-00021 Rev 1.0 BootROM: 2.13
PSUCTRL-2 : 450117-00-01 04334-00068 Rev 1.0 BootROM: 2.13
Image   : ExtremeXOS version 11.4.0.23 v1140b23 by release-manager
on Thu Feb 16 12:47:41 PST 2006
BootROM : 1.0.1.7
```



# Troubleshooting



If you encounter problems when using the switch, this appendix may be helpful. If you have a problem not listed here or in the release notes, contact Extreme Networks Technical Support.

## Troubleshooting Checklists

This section provides simple troubleshooting checklists for Layer 1, Layer 2, and Layer 3. The commands and recommendations described are applicable to both IPv4 and IPv6 environments unless otherwise specified. If more detailed information about a topic is available, you are referred to the applicable section in this appendix.

### Layer 1

When troubleshooting Layer 1 issues, verify:

- The installation of cables and connectors.
- The behavior of LED status lights. For additional information about LEDs, see LEDs.

- That the port is enabled, the link status is active, and speed and duplex parameters match the port settings at the other end of the cable.
  a. To display the configuration of one or more ports, use the `show ports configuration` command.
- That the packets are being received and transmitted.
  a. To display the number of packets being received and transmitted, use the `show ports {port_list | stack-ports stacking-port-list} statistics {no-refresh}` command.
- That there are no packet errors.
  a. To display packet error statistics, use the following commands:

  `show ports {port_list | stack-ports stacking-port-list} rxerrors {no-refresh}`—Displays receive error statistics

  `show ports {port_list | stack-ports stacking-port-list} txerrors {no-refresh}`—Displays transmit error statistics

  `show ports {mgmt | port_list |tag tag} collisions {no-refresh}`—Displays collision statistics

## Layer 2

When troubleshooting Layer 2 issues, verify:

- That the MAC addresses are learned, in the correct *VLAN (Virtual LAN)*, and are not blackhole entries.
  a. To display *FDB (forwarding database)* entries, use the `show fdb` command.
- Your VLAN configuration, including the VLAN tag, ports in the VLAN, and whether or not the ports are tagged.
  a. To display detailed information for each VLAN configured on the switch, use the `show vlan detail` command.

  For additional VLAN troubleshooting tips, see VLANs.
- Your *STP (Spanning Tree Protocol)* configuration, including the STP domain (*STPD (Spanning Tree Domain)*) number, VLAN assignment, and port state.
  a. To display STP information, use the following commands:

  `show stpd` detail—Displays the STP settings on the switch

  `show stpd ports`—Displays the STP state of a port

  `show vlan stpd`—Displays the STP configuration of the ports assigned to a specific VLAN

  For additional STP troubleshooting tips, see STP.

## Layer 3

When troubleshooting Layer 3 issues, verify:

- The IP address assigned to each *VLAN* router interface.
  a. To display summary information for all of the VLANs configured on the device, use the `show vlan` command.
- That IP forwarding is enabled, the routing protocol is globally enabled, and the routing protocol is enabled for a VLAN.
  a. To display the configuration information for a specific VLAN, use one of the following commands:

  `show ipconfig {ipv4} {vlan vlan_name}`—IPv4 environment

  `show ipconfig ipv6 {vlan vlan_name | tunnel tunnelname}`—IPv6 environment
- Which destination networks are in the routing table and the source of the routing entry.
  a. To display the contents of the routing table or the route origin priority, use one of the following commands:

  `show iproute`—IPv4 environment

  `show iproute ipv6`—IPv6 environment

  b. To display the contents of the routing table only for routes of a specified origin, use one of the following commands:

  `show iproute origin`—IPv4 environment

  `show iproute ipv6 origin`—IPv6 environment
- That the IP Address Resolution Protocol (ARP) table has the correct entries.

  **Note**
  The ARP table is applicable only in IPv4 environments.

  a. To display the contents of the IP ARP table, use the `show iparp` command.
- That the Neighbor Discovery (ND) cache has the correct entries.

  **Note**
  The ND cache is applicable only in IPv6 environments.

  a. To display the contents of the ND cache, use the `show neighbor-discovery cache ipv6` command.
- IP routing protocol statistics for the CPU of the switch.
  a. Only statistics of the packets handled by the CPU are displayed. To display IP statistics for the CPU of the switch, use one of the following commands:

  `show ipstats`—IPv4 environment

  `show ipstats ipv6`—IPv6 environment

- Your *OSPF (Open Shortest Path First)* configuration, including the OSPF area ID, router state, link cost, OSPF timers, interface IP address, and neighbor list.

  Note
  OSPF is applicable only in IPv4 environments.

  a. To display OSPF information, use the following commands:

  `show ospf`—Displays global OSPF information for the switch

  `show ospf area`—Displays information related to OSPF areas

  `show ospf area`—Displays detailed information related to OSPF areas

  `show ospf interfaces detail`—Displays detailed information about OSPF interfaces

- Your *OSPFv3 (Open Shortest Path First version 3)* configuration, including the OSPFv3 area ID, router state, link cost, OSPFv3 timers, interface IP address, and neighbor list.

  OSPFv3 is applicable only in IPv6 environments.

  a. To display OSPFv3 information, use the following commands:

  `show ospfv3`—Displays global OSPFv3 information for the switch

  `show ospfv3 area`—Displays information related to OSPFv3 areas

  `show ospfv3 interfaces`—Displays detailed information about OSPFv3 interfaces

- Your *RIP (Routing Information Protocol)* configuration, including RIP poison reverse, split horizon, triggered updates, transmit version, and receive version.

  Note
  RIP is applicable only in IPv4 environments.

  a. To display detailed information about how you have RIP configured on the switch, use the `show rip` command.

- RIP activity and statistics for all VLANs on the switch.

  Note
  RIP is applicable only in IPv4 environments.

  a. To display RIP-specific statistics for all VLANs, use the `show rip interface` detail command.

- Your RIP next generation (*RIPng (Routing Information Protocol Next Generation)*) configuration, including RIPng poison reverse, split horizon, triggered updates, transmit version, and receive version.

  Note
  RIPng is applicable only in IPv6 environments.

  a. To display detailed information about how you have RIPng configured on the switch, use the `show ripng` command.

- RIPng activity and statistics for all VLANs on the switch.

  Note
  RIPng is applicable only in IPv6 environments.

  a. To display RIPng-specific statistics for all VLANs, use the `show ripng interface` command.
- End-to-end connectivity.
  a. To test for connectivity to a specific host, use the `ping` command.
- The routed path between the switch and a destination end station.
  a. To verify and trace the routed path, use the `traceroute` command.

## LEDs

### Power LED does not light:

Check that the power cable is firmly connected to the device and to the supply outlet.

### On powering-up, the MGMT LED lights yellow:

The device has failed its Power On Self Test (POST) and you should contact your supplier for advice.

### A link is connected, but the Status LED does not light:

Check that:

- All connections are secure.
- Cables are free from damage.
- The devices at both ends of the link are powered-up.
- Both ends of the Gigabit link are set to the same autonegotiation state.

  The Gigabit link must be enabled or disabled on both sides. If the two sides are different, typically the side with autonegotiation disabled will have the link LED lit, and the side with autonegotiation enabled will not be lit. The default configuration for a Gigabit port is autonegotiation enabled. Verify by entering the following command:

  ```
  show ports configuration
  ```

### On power-on, some I/O modules do not boot:

Check the output of the `show power budget` command to see if all power supplies display the expected input voltage. Also refer to the section Power Management Guidelines for more detailed information about power management.

### ERR LED on the Management Switch Fabric Module (MSM) turns amber:

Check the syslog message for “critical” software errors. To reset the ERR LED and clear the log, use the following command and reboot the switch:

```
clear log static
```

If you continue to see “critical” software errors or the ERR LED is still amber after issuing the clear log static command and a switch reboot, contact Extreme Networks Technical Support for further assistance.

## Status LED on the I/O module turns amber:

Check the syslog message for a related I/O module error. If the error is an inserted I/O module that conflicts with the software configuration, use one of the following commands to reset the slot configuration:

```
clear slot
```

```
configure slot slot module module_type
```

Otherwise, contact Extreme Networks Technical Support for further assistance.

## ENV LED on the MSM turns amber:

Check each of the power supplies and all of the fans. Additionally, you display the status in the `show power` and `show fans` displays.

## Predictive Failure LED on the AC power supply blinks amber:

Check the current status of the power supply. If the speed of both fans is above 2000 RPM, the AC power supply unit (PSU) is operating normally and no failure is imminent. To check and view the health of the installed PSU, use the following command:

```
show power {ps_num} {detail}
```

## Switch does not power up:

All products manufactured by Extreme Networks use digital power supplies with surge protection. In the event of a power surge, the protection circuits shut down the power supply.

To reset the power, unplug the switch for 1 minute, plug it back in, and attempt to power-up the switch. If this does not work, try using a different power source (different power strip/outlet) and power cord.

# Using the Command Line Interface

This section describes helpful information for using and understanding the command line interface (CLI).

## General Tips and Recommendations

*The initial welcome prompt does not display:*

Check that:

- Your terminal or terminal emulator is correctly configured
- Your terminal or terminal emulator has the correct settings:
  - 9600 baud
  - 8 data bits
  - 1 stop bit
  - no parity
  - XON/OFF flow control enabled

For console port access, you may need to press [Return] several times before the welcome prompt appears.

*The SNMP Network Manager cannot access the device:*

Check that:

- The *SNMP (Simple Network Management Protocol)* access is enabled for the system.
- The device IP address, subnet mask, and default router are correctly configured, and that the device has been reset.
- The device IP address is correctly recorded by the SNMP Network Manager (refer to the user documentation for the Network Manager).
- The community strings configured for the system and Network Manager are the same.
- The SNMPv3 USM, Auth, and VACM configured for the system and Network Manager are the same.

*The Telnet workstation cannot access the device:*

Check that:

- The device IP address, subnet mask, and default router are correctly configured, and that the device has been reset.
- You entered the IP address of the switch correctly when invoking the Telnet facility.
- Telnet access is enabled for the switch.

If you attempt to log in and the maximum number of Telnet sessions are being used, you should receive an error message indicating so.

*Traps are not received by the SNMP Network Manager:*

Check that the SNMP Network Manager's IP address and community string are correctly configured, and that the IP address of the Trap Receiver is configured properly on the system.

*The SNMP Network Manager or Telnet workstation can no longer access the device:*

Check that:

- Telnet access or SNMP access is enabled for the system.
- The port through which you are trying to access the device has not been disabled. If it is enabled, check the connections and network cabling at the port.
- The port through which you are trying to access the device is in a correctly configured *VLAN*.
- The community strings configured for the device and the Network Manager are the same.

Try accessing the device through a different port. If you can now access the device, a problem with the original port is indicated. Re-examine the connections and cabling.

A network problem may be preventing you from accessing the device over the network. Try accessing the device through the console port.

*Permanent entries remain in the FDB:*

If you have made a permanent entry in the *FDB* that requires you to specify the VLAN to which the entry belongs and then deleted the VLAN, the FDB entry remains. Although this does not harm the system, if you want to removed the entry, you must manually delete it from the FDB.

*Default and static routes:*

If you have defined static or default routes, those routes remain in the configuration independent of whether the VLAN and VLAN IP address that used them remains. You should manually delete the routes if no VLAN IP address is capable of using them.

*You forget your password and cannot log in:*

If you are not an administrator, another user having administrator access level can log in, delete your user name, and create a new user name for you, with a new password.

Alternatively, another user having administrator access level can log in and initialize the device. This will return all configuration information (including passwords) to the initial values.

In the case where no one knows a password for an administrator level user, contact your supplier.

## The Summit switch displays only the "(pending-AAA) login" prompt

It is possible that the switch has not yet been fully initialized in the SummitStack. Wait for the Authentication Service (AAA) on the master node is now available for login. message to appear and then log in normally.

It is also possible that the stack has no master-capable node. In this case, AAA authentication will not be possible. This can occur when a stack has been deliberately dismantled and stacking was not unconfigured beforehand.

In either case, you may log in to the node using the failsafe account. If you have forgotten this account, and the stack has no master-capable node or has been dismantled, see Rescuing a Stack that has No Master-Capable Node on page 172.

The "(Pending-AAA) login:" prompt on other switches.

This login prompt is discussed in Logging in to the Switch on page 14.

## MSM Prompt - Modular Switches Only

*You do not know which MSM you are connected to:*

If you use a console connection to access and configure the switch, you should connect to the console port of the primary MSM, not the backup MSM. To determine which console port you are connected to use the show switch command. The output displays both the primary and backup MSMs, if installed, and an asterisk (*) appears to the right of the MSM you are connected to.

The following truncated sample output indicates that you are connected to MSM-A, the primary MSM:

```
MSM: MSM-A * MSM-B
```

*You have user privileges, not administrator privileges, on the backup MSM:*

If you establish a console connection to access the backup MSM, only user privileges are available. This is true regardless of the privileges configured on the primary MSM. If you enter an administrator level command on the backup MSM, the switch displays a message stating that the command is only supported on the primary MSM.

## Node Prompt--SummitStack Only

*You do not know which Node you are connected to:*

If you use a console connection to access and configure the switch, you should connect to the console port of the master node.

The word "Slot" followed by a hyphen and a single decimal digit slot number will be inserted into the prompt in stacking mode. A sample of this prompt is:

```
* Slot-6 Stack.21 #
```

The sample indicates a changed configuration (*), the stackable is in stacking mode and is currently using the slot number 6 in the active topology ("Slot-6"), the system name is the default of "Stack", the command about to be executed is the 21st command, and the user is logged in as the administrator on the Master node (#).

There will be no specific prompt that indicates the node role. Run show switch command to discover the identities of the Master and Backup nodes. A successful login on a Standby node will show the ">" character instead of the "#" character at the end of the prompt.

## Command Prompt

*You do not know if the switch configuration has been saved:*

If an asterisk (*) precedes the command prompt, a new change to the switch configuration has not been saved. To save the configuration, use the `save configuration` command. After you save the configuration, the asterisk (*) no longer precedes the command prompt.

*You do not know if you are logged in as an administrator or a user:*

Observe the console prompt. If you are logged in as an administrator, the prompt ends with the hash symbol(#). If you are logged in as a user, the prompt ends with a greater than sign (>).

The following is sample output from an administrator-level account:

```
BD-10808.1 #
```

The following is sample output from a user-level account:

```
BD-10808.1 >
```

## Port Configuration

*No link light on 10/100 Base port:*

If patching from a switch to another switch, ensure that you are using a category 5 (CAT5) crossover cable. This is a CAT5 cable that has pins 1 and 2 on one end connected to pins 3 and 6 on the other end.

*Excessive RX CRC errors:*

When a device that has autonegotiation disabled is connected to an Extreme Networks switch with autonegotiation enabled, the Extreme Networks switch links at the correct speed, but in half-duplex mode. The Extreme Networks switch 10/100 physical interface uses a method called parallel detection to bring up the link. Because the other network device is not participating in autonegotiation (and does not advertise its capabilities), parallel detection on the Extreme Networks switch is able only to sense 10 Mbps versus 100 Mbps speed and not the duplex mode. Therefore, the switch establishes the link in half-duplex mode using the correct speed.

The only way to establish a full-duplex link is either to force it at both sides, or run autonegotiation on both sides (using full-duplex as an advertised capability, which is the default setting on the Extreme Networks switch).

Note

A mismatch of duplex mode between the Extreme switch and another network device causes poor network performance. Viewing statistics using the `show ports rxerrors` command on the Extreme Networks switch may display a constant increment of CRC errors. This is characteristic of a duplex mismatch between devices. This is NOT a problem with the Extreme Networks switch.

Always verify that the Extreme Networks switch and the network device match in configuration for speed and duplex.

*No link light on Gigabit fiber port:*

Check that:

- The transmit fiber goes to the receive fiber side of the other device and vice-versa. All Gigabit fiber cables are of the crossover type.

- The Gigabit ports are set to Auto Off (using the command `configure ports port_list {medium [copper | fiber]} auto off speed speed duplex [half | full]`) if you are connecting the Extreme Networks switch to devices that do not support autonegotiation.

  By default, the Extreme Networks switch has autonegotiation set to On for Gigabit ports and set to Off for 10 Gigabit ports.
- You are using multimode fiber (MMF) when using a 1000BASE-SX small form-factor pluggable (SFP), and single-mode fiber (SMF) when using a 1000BASE-LX SFP. 1000BASE-SX technology does not work with SMF. The 1000BASE-LX technology works with MMF but requires the use of a mode conditioning patchcord (MCP).

## Software License Error Messages

*You do not have the required software license:*

If you attempt to execute a command and you do not have the required license, the switch returns the following message:

```
Error: This command cannot be executed at the current license level.
```

*You have reached the limits defined by the current software license level:*

If you attempt to execute a command and you have reached the limits defined by the current license level the switch returns the following message:

```
Error: You have reached the maximum limit for this feature at this license level.
```

For more information about licensing requirements, see the *Feature License Requirements* document.

## VLANs

*You cannot add a port to a VLAN:*

If you attempt to add a port to a *VLAN* and get an error message similar to:

```
localhost:7 # configure vlan marketing add ports 1:1,1:2
Error: Protocol conflict when adding untagged port 1:1. Either add this port as tagged or
assign another protocol to this VLAN.
```

You already have a VLAN using untagged traffic on a port. Only one VLAN using untagged traffic can be configured on a single physical port.

You verify the VLAN configuration using the following command:

`show vlan {virtual-router vr-name}`

The solution for this error using this example is to remove ports 1 and 2 from the VLAN currently using untagged traffic on those ports.

If this were the “default” VLAN, the command would be:

```
localhost:23 # configure vlan default delete ports 1:1,1:2
```

You can now re-enter the previous command without error:

```
localhost:26 # configure vlan marketing add ports 1:1,1:2
```

### *VLAN names:*

There are restrictions on VLAN names. They cannot contain whitespaces and cannot start with a numeric value.

### *VLANs, IP addresses, and default routes:*

The system can have an IP address for each configured VLAN. You must configure an IP address associated with a VLAN if you intend to manage (Telnet, *SNMP*, ping) through that VLAN or route IP traffic.

You can also configure multiple default routes for the system. The system first tries the default route with the lowest cost metric.

## STP

### *You have connected an endstation directly to the switch and the endstation fails to boot correctly:*

The switch has the Spanning Tree Protocol (STP) enabled, and the endstation is booting before the STP initialization process is complete. Specify that STP has been disabled for that *VLAN*, or turn off STP for the switch ports of the endstation and devices to which it is attempting to connect; then, reboot the endstation.

### *Spanning Tree Domain names:*

There are restrictions on *STPD* names. They cannot contain whitespaces and cannot start with a numeric value.

### *You cannot add ports within a VLAN to the specified STPD:*

Check to ensure that you are adding ports that already exist in the carrier VLAN.

If you see an error similar to the following:

```
Error: Cannot add VLAN default port 3:5 to STP domain
```

You might be attempting to add:

- Another 802.1D mode STP port to a physical port that already contains an 802.1D mode STP port (only one 802.1D encapsulation STP port can be configured on a particular STP port).
- A carrier VLAN port to a different STP domain than the carrier VLAN belongs.
- A VLAN and/or port for which the carrier VLAN does not yet belong.

**Note**
This restriction is only enforced in an active STPD and when you enable STP to make sure you have a legal STP configuration.

*Only one carrier VLAN can exist in an STPD:*

Only one carrier VLAN can exist in a given STPD although some of the ports on the carrier VLAN can be outside the control of any STPD at the same time.

The StpdID must be identical to the VLANid of the carrier VLAN in that STPD.

*The switch keeps aging out endstation entries in the switch FDB:*

If the switch continues to age out endstation entries in the switch *FDB*:

- Reduce the number of topology changes by disabling STP on those systems that do not use redundant paths.
- Specify that the endstation entries are static or permanent.

## ESRP

*ESRP names:*

There are restrictions on *ESRP (Extreme Standby Router Protocol)* names. They cannot contain whitespaces and cannot start with a numeric value.

*You cannot enable an ESRP domain:*

Before you enable a specific ESRP domain, it must have a domain ID. A domain ID is either a user-configured number or the 802.1Q tag (VLANid) of the tagged master *VLAN*. The domain ID must be identical on all switches participating in ESRP for that particular domain. If you do not have a domain ID, you cannot enable ESRP on that domain.

Note the following on the interaction of tagging, ESRP, and ESRP domain IDs:

- If you have an untagged Master VLAN, you must specify an ESRP domain ID.
- If you have a tagged master VLAN, ESRP uses the 802.1Q tag (VLANid) of the master VLAN for the ESRP domain ID. If you do not use the VLANid as the domain ID, you must specify a different domain ID.

*You cannot delete the master VLAN from the ESRP domain:*

If you attempt to remove the master VLAN before disabling the ESRP domain, you see an error message similar to the following:

```
ERROR: Failed to delete master vlan for domain "esrp1" ; ESRP is enabled!
```

If this happens:

- Disable the ESRP domain using the disable esrp command.
- Remove the master VLAN from the ESRP domain using the configure esrp delete master command.

## VRRP

*You cannot define VRRP virtual router parameters:*

Before configuring any *virtual router (VR)* parameters for VRRP, you must first create the VRRP instance on the switch.

If you define VRRP parameters before creating the VRRP, you may see an error similar to the following:

```
Error: VRRP VR for vlan vrrp1, vrid 1 does not exist.
Please create the VRRP VR before assigning parameters.
Configuration failed on backup MSM, command execution aborted!
```

If this happens:

- Create a VRRP instance using the `create vrrp vlan vrid` command.
- Configure the VRRP instance's parameters.

# Using ELRP to Perform Loop Tests

A switch running ELRP transmits multicast packets with a special MAC destination address out of some or all of the ports belonging to a *VLAN*. All of the other switches in the network treat this packet as a regular, multicast packet and flood it to all of the ports belonging to the VLAN. When the packets transmitted by a switch are received back by that switch, this indicates a loop in the Layer 2 network.

After a loop is detected through ELRP, different actions can be taken such as blocking certain ports to prevent loop or logging a message to system log. The action taken is largely dependent on the protocol using ELRP to detect loops in the network.

You can use ELRP on a "standalone" basis or with other protocols such as *ESRP*, as described in Using ELRP with ESRP. Protocols such as Ethernet Automatic Protection Switching (EAPS) require that a network have a ring topology to operate. In this case you can use ELRP to ensure that the network has a ring topology.

ExtremeXOS software does not support ELRP and Network Login on the same port.

The "standalone" ELRP commands determine whether or not a network has an Layer 2 loop.

## About Standalone ELRP

Standalone ELRP gives you the ability to send ELRP packets, either periodically or on an ad hoc "one-shot" basis on a specified subset of *VLAN* ports. If any of these transmitted packets is received back then standalone ELRP can perform a configured action such as sending a log message to the system log file, sending a trap to the *SNMP* manager, and disabling the port where the looped packet arrived.

ELRP uses QP8 to send/receive control packets over the network to detect loops, so it is advisable to not use QP8 in for user traffic in the network.

Standalone ELRP allows you to:

- Configure ELRP packet transmission on specified VLANs.

- Specify some or all the ports of the VLAN for packet transmission. Each VLAN must be configured individually for ELRP.

  Note

  Reception of packets is not limited to any specific ports of the VLAN and cannot be configured.

- Save and restore standalone ELRP configuration across reboots.
- Request non-periodic or periodic transmission of ELRP packets on specified ports of a VLAN.

*Non-periodic ELRP Requests*

You can specify the number of times ELRP packets must be transmitted and the interval between consecutive transmissions. A message is printed to the console and logged into the system log file indicating detection of network loop when ELRP packets are received back or no packets are received within the specified duration.

*Periodic ELRP Requests*

You can configure the interval between consecutive transmissions. If ELRP packets are received back, a message is printed to the system log file and/or a trap is sent to the *SNMP* manager indicating detection of a network loop.

You have the option to configure the switch to automatically disable the port where the looped packet arrived as well as the length of time (in seconds) that the port should remain disabled. When this hold time expires, the port is automatically enabled.

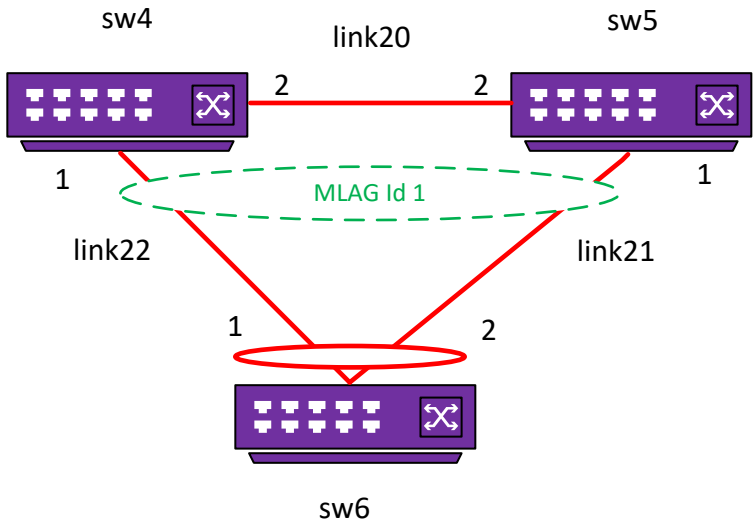
Should a loop occur on multiple ports, only the first port in the *VLAN* on which the PDU is received is disabled. The second port is ignored for 1 or 2 seconds and then if another PDU is received, that port is disabled until the loop is gone. This prevents shutting down all ports in the VLAN.

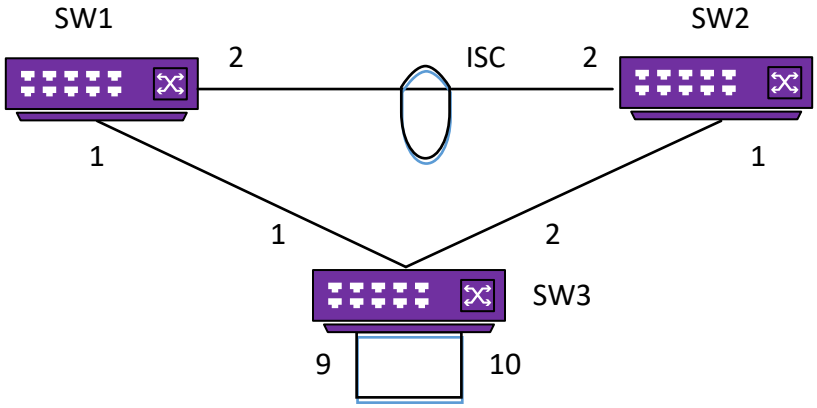
*ELRP Egress Port Disable*

When a loop is detected via ELRP, an option to disable the port where the packet egressed was added to suppress the loop in ExtremeXOS 16.1. All existing functionalities supported on disabling the port where the packet ingressed are supported on this enhancement as well. A log message or *SNMP* trap will be generated based on the configuration. The user must specify a holdtime or keep the port disabled until a user enables it.

ELRP egress blocking enhancement adds the following value:

- More efficient bandwidth utilization.
- Better loop protection in *MLAG (Multi-switch Link Aggregation Group)* environment. If an edge device is dual-homed to an MLAG pair and this edge device loses its *LAG (Link Aggregation Group)* configuration, ELRP will be able to prevent the loop by disabling egress port. This can be illustrated in the following three network topologies:




*Exclude Port List*

When you have configured the switch to automatically disable the port where the looped packet arrived, there may be certain ports that you do not want disabled. You can then create a list of ports that are excluded from this automatic disabling and that will remain enabled. This list can also contain EAPS ring ports. You can also specify that EAPS ring ports are excluded. When this option is selected, the actual EAPS ring ports do not have to be explicitly listed.

You can configure any port on the switch into the exclude port list. The list can also be edited to add or delete any port on the switch. Then when ELRP detects a loop and the disable feature is enabled, it checks to see if the port from which the PDU arrived is on the user defined exclude port list or is part of a trunk port that is defined on the list. When the port is on the list, ELRP logs the event and does not disable the port. When the port is not on the list, ELRP disables the port.

**Limitations**

The following are limitations to this feature:

- A specified port is added to the list regardless of its corresponding *VLAN*.
- Only ports on the local switch can be added.
- A loop detected on an excluded port may persist indefinitely until user action is taken.
- *MLAG* ISC links should not be blocked, therefore they should be added to the excluded port list.
- On Summit X480 platform, ELRP does not detect loop when enabled on VPLS service VLAN due to hardware limitation.

## Configure Standalone ELRP

The ELRP client (standalone ELRP) must be enabled globally in order for it to work on any *VLAN*.

- Globally enable the ELRP client.

  ```
  enable elrp-client
  ```

  The ELRP client can be disabled globally so that none of the ELRP VLAN configurations take effect.
- Globally disable the ELRP client.

  ```
  disable elrp-client
  ```

*Configuring Non-periodic Requests*

- To start one-time, non-periodic ELRP packet transmission on specified ports of a *VLAN* using a particular count and interval, use one of the following commands:

  ```
  configure elrp-client one-shot vlan_name ports [ports | all] interval
  sec retry count [log | print | print-and-log]
  ```

  (This command is backward compatible with Extreme Networks switches running the ExtremeWare software.)

  ```
  run elrp vlan_name {ports ports} {interval sec} {retry count}
  ```

  These commands start one-time, non-periodic ELRP packet transmission on the specified ports of the VLAN using the specified count and interval. If any of these transmitted packets is returned, indicating loopback detection, the ELRP client performs the configured actions of logging a message in the system log file and/or printing a log message to the console. There is no need to trap to the *SNMP* manager for non-periodic requests.

*Configuring Periodic Requests*

- Start periodic ELRP packet transmission on specified ports of a *VLAN* using a particular interval.

  ```
  configure elrp-client periodic vlan_name ports [ports | all] interval
  sec [log | log-and-trap | trap] {disable-port {egress | ingress}
  {duration {seconds } | permanent }}
  ```

  This command starts periodic ELRP packet transmission on the specified ports of the VLAN using a specified interval. If any of these transmitted packets is returned, indicating loopback detection, the ELRP client performs the configured action of logging a message in the system log file and/or sending a trap to the *SNMP* manager.

When the option of disabling a port is configured, you choose the duration, in seconds, as a hold time or you disable the port permanently. When ELRP disables the port, the operation is not persistent. When the switch is rebooted, the port is enabled when the switch comes up.

Note

ELRP detects loops on a per VLAN basis. When the **disable port** option is selected, keep in mind that the entire port will be disabled. This may affect connectivity for other VLANs configured on that port that did not have any data loop problems. ELRP also does not distinguish between uplink ports and host ports. When the **disable port** option is selected and ELRP detects a loop, any and all ports where the loop was detected will be disabled, including uplink ports.

- Disable a pending one-shot or periodic ELRP request for a specified VLAN.
  `unconfigure elrp-client vlan_name`

*Configuring Exclude Port List*

- Configure an ELRP exclude port list.
  `configure elrp-client disable-ports [exclude | include] [ ports | eaps-ring-ports]`
- Disable an ELRP exclude port list.
  #unique_3303

## Displaying Standalone ELRP Information

- Display summary ELRP status information.
  `show elrp`
- Display information about ELRP disabled ports.
  `show elrp disabled-ports`

## Example: ELRP on Protocol-based VLANs

For ELRP to detect loops on a protocol-based VLAN (other than the protocol any), you need to add the ethertype 0x00bb to the protocol.

```
# Create VLANs
create vlan v1
create vlan v2
# Protocol filter configuration
configure vlan v1 protocol IP
configure vlan v2 protocol decnet
# Add ports to the VLAN
configure vlan v1 add ports 1
configure vlan v2 add ports 2
# Enable ELRP on the create VLANs
enable elrp-client
configure elrp-client periodic v1 ports all interval 5 log
configure elrp-client periodic v2 ports all interval 5 log
# Add the ethertype to the protocol
configure protocol IP add snap 0x00bb
configure protocol decnet add snap 0x00bb
```

VLANs v1 and v2 can then detect the loop on their respective broadcast domains.

# Using the Rescue Software Image

**Warning**

The rescue image completely re-initializes the system. All data residing on the switch is cleared, including configuration files, policy files, and other system-related files. Use this feature only with the guidance of Extreme Networks Technical Support.

The rescue software image recovers a switch that does not boot up by initializing the internal memory card and installing the ExtremeXOS software on both primary and secondary images of that card. To use the rescue software image, you must be running ExtremeXOS 11.1 or later. Earlier versions of ExtremeXOS do not support the rescue software image.

BlackDiamond X8 series switches and BlackDiamond 8800 series switches support loading the rescue image to the compact flash card installed in the MSM. For more information see Obtain the Rescue Image from a Compact Flash Card.

Before you begin the recovery process, collect the following information:

- IP address, netmask, and gateway for the switch
- IP address of the Trivial File Transfer Protocol (TFTP) server that contains the ExtremeXOS image
- ExtremeXOS image filename (the image has a .xos filename extension)

**Note**

The rescue process initializes the primary and secondary images with the ExtremeXOS software image. No additional software packages or configuration files are preserved or installed. This process takes a minimum of 7 minutes to complete. To install additional modular software packages and configuration files, see Software Upgrade and Boot Options for more information.

## Obtaining the Rescue Image from a TFTP Server

To recover the switch, you must enter the Bootloader and issue a series of commands.

- To access the Bootloader:
  a. Attach a serial cable to the console port of the MSM.
  b. Attach the other end of the serial cable to a properly configured terminal or terminal emulator. The terminal settings are:

     The terminal settings are:
     - 9600 baud
     - 8 data bits
     - 1 stop bit
     - no parity
     - XON/OFF flow control enabled

c. Reboot the MSM and press the spacebar key on the keyboard of the terminal during the boot up process.

Note

You must press the spacebar key immediately after a power cycle of the MSM in order to get into the Bootloader application.

On BlackDiamond 8800 series switches, when you see the BootROM banner, press the spacebar key to get into the Bootloader application.

On the BlackDiamond X8 series switches, press and hold the spacebar key to enter the bootROM (actually a BIOS) as soon as you see CF card tested OK on the screen.

d. As soon as you see the BootRom -> prompt (BlackDiamond X8, 8800 series switches), release the spacebar. From here, you can begin the recovery process.

- To obtain the rescue image and recover the switch:

a. Provide the network information (IP address, netmask, and gateway) for the switch using the following command:

```
configip ipaddress ip-address[netmask] gateway gateway-address
```

Where the following is true:

- **ip-address**—Specifies the IP address of the switch
- **netmask**—Specifies the netmask of the switch
- **gateway-address**—Specifies the gateway of the switch

b. Download the ExtremeXOS image using the following command:

```
download image tftp-address filename
```

Where the following is true:

- **tftp-address**—Specifies the IP address of the TFTP server that contains the ExtremeXOS image
- **filename**—Specifies the filename of the ExtremeXOS image

If you attempt to download a non-rescue image, the switch displays an error message and returns you to the BOOTLOADER -> (BlackDiamond 10808 switch) or the BootRom -> (BlackDiamond 8800 and BlackDiamond 12800 series switches) command prompt.

After you download the ExtremeXOS image file, the switch installs the software and reboots. After the switch reboots, the switch enters an uninitialized state. At this point, configure the switch and save your configuration. In addition, if you previously had modular software packages installed, you must re-install the software packages to each switch partition. For more information about installing software packages, see Software Upgrade and Boot Options.

If you are unable to recover the switch with the rescue image, or the switch does not reboot, contact Extreme Networks Technical Support.

## Obtaining the Rescue Image from a Compact Flash Card

Before you remove or install any hardware, review the hardware documentation listed in Extreme Networks Documentation.

In addition to recovering the switch using the internal memory card and the management port, there is also support for loading the rescue image to the compact flash card installed in the MSM.

The compact flash card must be file allocation table (FAT) formatted. Use a PC with appropriate hardware such as a compact flash reader/writer and follow the manufacturer's instructions to access the compact flash card and place the image onto the card.

Note
This feature is supported only on BlackDiamond 8800 series switches.

To recover the switch, you must remove power from the switch, install an appropriate compact flash card into the MSM, and enter the Bootloader to issue a series of commands.

- To access the Bootloader:
  a. Remove all power cords from the power supplies switch. There should be no power to the switch.
  b. Insert the FAT formatted compact flash card into the compact flash slot of the MSM installed in slot 5/A.
  c. Remove the MSM installed in slot 6/B. Place the MSM in a safe location and do not re-install it until you finish recovering the switch.
  d. Attach a serial cable to the console port of the MSM installed in slot 5/A.
  e. Attach the other end of the serial cable to a properly configured terminal or terminal emulator.

     The terminal settings are:

     - 9600 baud
     - 8 data bits
     - 1 stop bit
     - no parity
     - XON/OFF flow control enabled

  f. Provide power to the switch by re-inserting the power cords into the power supplies.
  g. Immediately press the spacebar until the BootRom -> prompt appears.

     Note
     You must press the spacebar key immediately after a power cycle of the MSM in order to get into the Bootloader application.

  h. As soon as you see the BootRom -> prompt, release the spacebar. From here, you can begin the recovery process.
- To obtain the rescue image that you placed on the compact flash card and recover the switch:
  a. Download the ExtremeXOS image that is already on the compact flash card using the following command:

     ```
     boot file filename
     ```

     Where the filename specifies the image file for the BlackDiamond 8800 series switches.
  b. At the BootRom -> prompt, press **[Return]**.

     The following message appears: `ok to continue`
  c. Type `yes` to begin the recovery process. This takes a minimum of seven minutes.
  d. After the process runs, the BootRom -> prompt displays the following message:

     ```
     ****press enter to reboot****
     ```

e. Press **[Return]** to reboot the switch.

   The switch reboots and displays the login prompt. You have successfully completed the setup from the compact flash card.

f. Remove the compact flash card installed in the MSM.

After you download the ExtremeXOS image file, the switch installs the software and reboots. After the switch reboots, the switch enters an uninitialized state. At this point, configure the switch and save your configuration. In addition, if you previously had modular software packages installed, you must re-install the software packages to each switch partition. For more information about installing software packages, see Software Upgrade and Boot Options on page 1522.

If you are unable to recover the switch with the rescue image, or the switch does not reboot, contact Extreme Networks Technical Support.

## Performing Compact Flash Recovery Using a USB Memory Drive

In addition to recovering the switch using the network, there is also support on the BlackDiamond X8 for rescuing from a USB memory drive. The USB memory drive must be file allocation table (FAT) formatted.

Use a computer with appropriate hardware to place the rescue image onto the memory drive. Place the rescue image in the top level directory of the memory drive.

Note

This feature is supported only on BlackDiamond X8 series switches. Before you remove or install any hardware, review the hardware documentation which is listed in the preface.

### *Setup and Access to the BootROM*

1. Remove all power cords from the power supply switch. There should be no power to the switch.
2. Insert the FAT formatted USB memory drive into the bottom USB port (labeled USB-1) of the MM installed in the slot you wish to recover.
3. If present, remove the MM from the other management slot. Place the MM in a safe location and do not reinstall it until you finish recovering the switch.
4. Attach a rollover serial cable to the console port of the MM still installed.
5. Attach the other end of the serial cable to a properly configured terminal or terminal emulator.

   The terminal settings are:

   - 9600 baud
   - 8 data bits
   - 1 stop bit
   - no parityXON/OFF flow control enabled
6. Provide power to the switch by re-inserting the power cords into the power supplies.
7. Immediately press the spacebar until the `BootROM ->` prompt appears.

### *Determining the Current BootROM Image Version*

1. Use the “vers” command at the BootROM prompt.

```
BootRom > vers
```

```
EFI Specification Revision : 2.0
EFI Vendor                 : INSYDE Corp.
EFI Revision               : 4096.1
EFI Build Version          : Release8_6_1
Extreme Version            : 1.0.0.6
Build Date                 : 12/03/01
Build Time                 : 17:03:03
```

2. If the Extreme Version is 1.0.0.5 or earlier, you must recover the compact flash using TFTP.
3. If the Extreme Version is exactly 1.0.0.6, then you must first locate the drive letter that contains your rescue image using these BIOS commands:
4. If the Extreme Version is 1.0.0.1 or greater, proceed with the instructions in the following section directly from the BootROM -> prompt.

*Installing the ExtremeXOS Image*

1. Install the image from the memory drive.

   `download image usb` *`filename`*

   Where the *`filename`* specifies the desired image file for the BlackDiamond X series switches.

   Example for BIOS version 1.0.0.6:

   ```
   fs2:\> download image usb bdX-15.1.1.6.xos
   ```

   Example for BIOS version 1.0.0.7 or greater:

   ```
   BootRom> download image usb bdX-15.1.1.6.xos
   ```

2. When prompted to continue, enter `yes` to begin the recovery process. This takes a minimum of seven minutes.

   After the process runs, the BootRom prompts you to reboot.
3. Follow the on-screen instructions to reboot. The switch reboots, and then displays the login prompt. The recovery process from the memory drive is complete.
4. Remove the USB memory stick installed in the MM.
5. If applicable, re-insert the secondary MM and synchronize it via the `synchronize` command from the Master MM's console.

## Rescuing a Node in a SummitStack

You can use the rescue option on a node if it becomes unbootable due to a corrupt image. The rescue operation is independent of the SummitStack feature.

To rescue a node:

1. Establish a terminal session using the console port of the node.
2. Reboot this node. While rebooting enter the BootROM program.
   a. To enter the bootrom, wait until you see the message "Starting Default Bootloader..." and then press and hold the space bar until the BootROM prompt appears.

3. Provide the network information (IP address, netmask, and gateway) for the switch using the following command:
   `config ipaddress ip-address[netmask]gateway gateway-address`

   Where the:

   - *ip-address*—Specifies the IP address of the switch
   - *netmask*—Specifies the netmask of the switch
   - *gateway-address*—Specifies the gateway of the switch
4. Download the ExtremeXOS image using the following command:
   `download image tftp-address filename`

   Where the:

   - *tftp-address*—Specifies the IP address of the TFTP server that contains the ExtremeXOS image
   - *filename*—Specifies the filename of the ExtremeXOS image

   If you attempt to download a non-rescue image, the switch displays an error message and returns you to the BootROM command prompt.

After you download the ExtremeXOS image file, the switch installs the software and reboots. After the switch reboots, the switch enters an uninitialized state. At this point, configure the switch and save your configuration. In addition, if you previously had modular software packages installed, you must reinstall the software packages to each switch partition. For more information about installing software packages, see Software Upgrade and Boot Options

If you are unable to recover the switch with the rescue image, or the switch does not reboot, contact Extreme Networks Technical Support.

The rescue process:

- Does not affect stacking configuration parameters.
- Sets the security configuration to the factory defaults.

## Debug Mode

The *EMS (Event Management System)* provides a standard way to filter and store messages generated by the switch. With EMS, you must enable debug mode to display debug information. You must have administrator privileges to use these commands. If you do not have administrator privileges, the switch rejects the commands.

- Enable or disable debug mode for EMS:
  `enable log debug-mode`

After debug mode has been enabled, you can configure EMS to capture specific debug information from the switch. Details of EMS can be found in Status Monitoring and Statistics.

## Saving Debug Information

You can save switch data and statistics to a network TFTP server, an internal memory card that comes preinstalled in the switch, a removable compact flash card (BlackDiamond 8800 series switch), or a removable USB 2.0 storage device (Summit X460, X480, X670, X670V, and X770 switches).

With assistance from Extreme Networks Technical Support personnel, you can configure the switch to capture troubleshooting information, such as a core dump file, to these locations.

The switch only generates core dump files in the following situations:

- If an ExtremeXOS process fails.
- When forced under the guidance of Extreme Networks Technical Support.

The core dump file contains a snapshot of the process when the error occurred.

Note

Use the commands described in this section only under the guidance of Extreme Networks Technical Support personnel to troubleshoot the switch.

Before you can enable and save process core dump information to removable storage devices, you must install a compact flash card or USB 2.0 storage device. .

## Enabling the Send Debug Information Switch

This allows information to be sent to the internal memory card or a removable storage device.

- Enable the switch to save process core dump information.

  `configure debug core-dumps [internal-memory | memorycard | off]`

  A removable storage device is a compact flash card on a BlackDiamond 8800, or a USB 2.0 storage device on a BlackDiamond X8 or Summit X460, X480, X670, X670V, X670G2, and X770 series switch.

  Core dump files have a .gz file extension. The filename format is: core.<process-name.pid>.gz where process-name indicates the name of the process that failed and pid is the numerical identifier of that process. If you save core dump files to a switch with a compact flash card, the filename also includes the affected MSM: MSM-A or MSM-B.

  If you configure the switch to write core dump files to the internal memory card and attempt to download a new software image, you might have insufficient space to complete the image download.

  If this occurs, you must decide whether to continue the software download or move or delete the core dump files from the internal memory. For example, if your switch supports a removable storage device with space available, transfer the files to the storage device. On switches that do not have a removable storage device, transfer the files from the internal memory card to a TFTP server. This frees up space on the internal memory card while keeping the core dump files.

## Copy Debug Information to Removable Storage Devices

- Save and copy debug information to the specified compact flash card or USB 2.0 storage device.

  `save debug tracefiles memorycard`

  After the switch writes a core dump file or other debug information to the storage device, and before you can view the contents on the device, you must ensure it is safe to remove the device from the switch. Use the `eject memorycard` command to prepare the device for removal. After you

issue the `eject memorycard` command, you can manually remove the storage device from the switch and read the data on the device.

To access and read the data on a removable storage device, use a PC with the appropriate hardware, such as a compact flash reader/writer and follow the manufacturer's instructions to access the compact flash card and read the data.

## Copying Debug Information to a TFTP Server

SummitStack only—To get debug information from a node in the SummitStack that is not a master node, you must configure an alternate IP address on the node and have its management port connected. You may then use the tftp command to upload specific files. The upload debug command functions only on the master node.

1. Save and copy debug information to the specified TFTP server.
   `upload debug [ ipaddress | hostname ] {{vr} vrname} {block-size block_size}`

   Progress messages are displayed that indicate the file being copied and when the copying is finished.

   Depending on your platform, the switch displays a message similar to the following:

   ```
   The following files on have been uploaded:
   Tarball Name: TechPubsLab_C_09271428.tgz
   ./primary.cfg
   ```

   You can also use this command in conjunction with the `show tech` command.

   Prior to uploading debug information files, the switch prompts you with the following message to run the `show tech-support` command with the logto file option:

   ```
   Do you want to run show tech logto file first? (y/n)
   ```

2. Enter y to run the `show tech-support` command before uploading debug information.
   If you enter y, the show_tech.log.tgz file is included during the upload. Enter n to upload debug information without running the `show tech` command.

   After you upload the debug information, you should see a compressed TAR file on the TFTP server, which contains the debug information.

   The TAR file naming convention is

   ```
   <SysName>_<{<slot#>|A|B}I|C>_<Current Time>.tgz
   - Current Time = mmddhhmm ( month(01-12), date(01-31), hour(0-24), minute(00-59) ).
   ```

## Managing Debug Files

For the purposes of this section, it is assumed that you have configured the switch to send core dump information under the guidance of Extreme Networks Technical Support.

Managing the debug files might include any of the following tasks: renaming or copying a core dump file, displaying a comprehensive list of files including core dump files, transferring core dump files, and

deleting a core dump file. You can manage the debug files using the same commands that you use to manage other switch files.

Note
Filenames are case-sensitive. For information on filename restrictions, refer to the specific command in the *ExtremeXOS 16.2 Command Reference Guide*.

## Evaluation Precedence for ACLs

The ACLs on a port are evaluated in the following order:

- Persistent dynamic ACLs
- Host-integrity permit ACLs
- MAC source address deny ACLs
- Source IP lockdown source IP permit ACLs
- Source IP lockdown deny all ACLs
- ARP validation CPU ACLs
- ACLs created using the CLI
- DoS Protect-installed ACLs
- MAC-in-MAC installed ACLs
- ACLs applied with a policy file (see ACLs on page 640 for precedence among these ACLs)

## TOP Command

The `top` command is a UNIX-based command that displays real-time CPU utilization information by process.

The output contains a list of the most CPU-intensive tasks and can be sorted by CPU usage, memory usage, and run time. For more detailed information about the `top` command, refer to your UNIX documentation.

## TFTP Server Requirements

We recommend using a TFTP server that supports blocksize negotiation (as described in RFC 2348, TFTP Blocksize Option), to enable faster file downloads and larger file downloads.

## System Odometer

Each field replaceable component contains a system odometer counter in EEPROM.

The `show odometers` command displays an approximate days of service duration for an individual component since the component was manufactured.

### Monitored Components

On a modular switch, the odometer monitors the following components:

- Chassis
- MSMs

- I/O modules
- Power controllers

On Summit family switches, the odometer monitors the following components:

- Switch
- XGN-2xn card

## Recorded Statistics

The following odometer statistics are collected by the switch:

- Service Days—The amount of days that the component has been running
- First Recorded Start Date—The date that the component was powered-up and began running

Depending on the software version running on your switch, the modules installed in your switch, and the type of switch you have, additional or different odometer information may be displayed.

The following is sample output from a BlackDiamond 8800 series switch:

```
Service  First Recorded
Field Replaceable Units           Days     Start Date
-------------------------         -------  --------------
Chassis     : BD-8810                 209  Dec-07-2004
Slot-1      : G48T                    208  Dec-07-2004
Slot-2      : 10G4X                   219  Nov-02-2004
Slot-3      : G48T                    228  Oct-26-2004
Slot-4      : G24X                    226  Oct-19-2004
Slot-5      : G8X                     139  Dec-07-2004
Slot-6      :
Slot-7      : 10G4X                   160  Dec-16-2004
Slot-8      : 10G4X                   133  Dec-14-2004
Slot-9      : G48P                    111  Nov-04-2004
Slot-10     :
MSM-A       : MSM-48C                 137  Dec-07-2004
MSM-B       :
PSUCTRL-1   :                         209  Dec-07-2004
PSUCTRL-2   :                         208  Dec-07-2004
```

The following is sample output from a BlackDiamond X8 switch:

```
Service First Recorded
Field Replaceable Units        Days   Start Date
-------------------------------------------------------------
Chassis        : BD-X8         456   May-26-2012
Slot-1         : BDXA-10G48X     494   Oct-28-2011
Slot-2         : BDXA-10G48X      352   Jul-09-2012
Slot-3         : BDXA-10G48X      563   Oct-31-2011
Slot-4         : BDXA-40G24X      146   May-10-2012
Slot-5         :
Slot-6         :
Slot-7         :
Slot-8         :
FM-1           : BDXA-FM20T       453   May-28-2012
FM-2           : BDXA-FM20T       453   May-25-2012
FM-3           : BDXA-FM20T       386   May-23-2012
FM-4           : BDXA-FM20T       386   May-24-2012
```

```
MM-A            : BDX-MM1           414   Jun-03-2012
MM-B            : BDX-MM1           331   May-26-2012
```

The following is sample output from a Summit series switch:

```
Service  First Recorded
Field Replaceable Units           Days     Start Date
------------------------          -------  --------------
Switch     : SummitX4                   7  Dec-08-2004
XGM-2xn-1  :
```

# Temperature Operating Range

## Modular Switches

On modular switches, each I/O module and MSM/MM has its own temperature sensor, normal temperature range and minimum and maximum temperature threshold, and status. See the `show temperature` command for information about each module.

- If the temperature on any I/O module or MSM is out of its normal range, an error message is logged. The Status column in show temperature shows Warning.
- If the temperature on any I/O module or MSM was out of normal range, then returns back to its normal range, a Notice message is logged. The Status column in show temperature shows Normal.
- If the temperature on any I/O module is above its maximum allowed, or below its minimum allowed, that I/O module is powered down and marked Failed in show slot, and Error in show temperature.
- If the temperature on any MSM is above its maximum allowed, or below its minimum allowed, that MSM is marked FAIL (OverHeat) in show switch, Failed in show slot, and Error in show temperature. If that MSM was the primary, the other MSM becomes primary. MSMs out of allowed temperature range are not powered down (by hardware design), nor are they rebooted.
- Beyond the mechanisms above, you can use the UPM feature to take any further desired actions when an I/O module or MSM overheats.

## Summit Family Switches and SummitStack

On Summit family switches, if a switch runs outside the expected range, the switch logs an error message, generates a trap, and continues running. No components are shutdown. To verify the state of the switch, use either the `show switch` or `show temperature` commands. If the temperature exceeds the maximum limit, the `show switch` output indicates the switch in an OPERATIONAL (Overheat) mode, and the `show temperature` output indicates an error state due to overheat.

On a SummitStack, use the `show temperature` command to see the temperature information of all active nodes in the stack.

# Unsupported Module Type

## BlackDiamond 8800 Series Switches Only

To reduce the chances of ports fluctuating between powered and non-powered states, newly inserted powered devices (PDs) are not powered when the actual delivered power for the module is within approximately 19 W of the configured inline power budget for that slot.

However, actual aggregate power can be delivered up to the configured inline power budget for the slot (for example, when delivered power from ports increases or when the configured inline power budget for the slot is reduced).

# Corrupted BootROM on BlackDiamond 8800 Series Switches

If your default BootROM image becomes corrupted, you can force the MSM to boot from an alternate BootROM image, by inserting a pen into the Alternate (A) and Reset (R) holes on the BlackDiamond 8000 series MSM and applying pressure.

The alternate BootROM image also prints boot progress indicators, and you can later use this alternate image to re-install a new default BootROM image. Finally, a corrupted compact flash card can be recovered from either the Alternate or Default BootROM.

For more information, refer to the hardware documentation listed in Related Publications on page 5.

# Inserting Powered Devices in the PoE Module

## BlackDiamond 8800 Series Switches Only

To reduce the chances of ports fluctuating between powered and non-powered states, newly inserted powered devices (PDs) are not powered when the actual delivered power for the module is within approximately 19 W of the configured inline power budget for that slot.

However, actual aggregate power can be delivered up to the configured inline power budget for the slot (for example, when delivered power from ports increases or when the configured inline power budget for the slot is reduced).

# Modifying the Hardware Table Hash Algorithm

## BlackDiamond X8 Switches

A 40 Gb port must use multiple 20 Gb HGd links to carry its traffic. In the degenerate case, a particular link is always the target of the hash result. With a 40 Gb link hashing into a 20 Gb link, half the traffic is lost. ExtremeXOS attempts to configure the packet hash algorithm to accommodate most common situations; however, some customers may have uncommon situations and may wish to adjust the hash. For this reason, the user can specify values for some of the hash parameters in normal packet hash mode:

- Source port hash (cannot be used when 40 Gb ports are in use)

- Packet field hash calculation algorithm (CRC or XOR)

The packet hash is configurable.

A Dynamic Load Balance (DLB) algorithm allows distribution of flows to the links in the switch fabric trunk that are currently carrying the least load. DLB uses the packet field hash, but does not use the hash code that is reduced to a number modulo the number of links in the group. Instead the number used is 15 bits for a total of 32 K possible hash codes. This 15-bit hash code indexes into a 32 K-entry "flow table". Each time an unused entry is allocated to one or more "micro-flows" (i.e., to flows that generate the same 15-bit hash value), a load calculation is performed and the link with the most available bandwidth is assigned to the flow table entry. DLB offers two modes:

- Spray mode causes the link assignment to occur on every packet transmission. This is similar to a "round-robin" hash.Every packet goes to the link with the least load, but ordering within flows is not guaranteed. This mode is useful for Bandwidth Management Testing (BMT).
- Eligibility mode keeps the link fixed to the flow entry until a 32 ms inactivity timeout occurs. Ordering within flows is guaranteed.

For blades that provide 10 Gb ports only (and for 40 Gb blades that are entirely configured for 10 Gb operation), source port hashing will be used by default.

While source port hashing can be used for non-blocking operation, such operation depends on the switch fabric distribution. For example, suppose three 10 Gb ports on one BDXA-10G48X card hash to the same switch fabric "channel." Also suppose that the three 10 Gb ports' traffic is aggregated into the same 40 Gb port on a different I/O blade. Since each switch fabric channel can only provide 20 Gb of bandwidth to a single 40 Gb port, then the switch fabric channel will try to send 30 Gbps of Ethernet packets to the same 20 Gb switch fabric link, causing congestion in the channel on the FM blade.

To avoid this situation, the user can configure DLB Eligibility mode, which will cause per 10 Gb Ethernet port traffic to be distributed to all switch fabric channels instead of being sent to the same channel.

Use the following command: `disable elrp-client [default | source-port | packet {algorithm [crc |xor] | dynamic-mode [spray | eligibility | none]}] {slot slot-number}`

## BlackDiamond 8800 Series Switches, SummitStack, and Summit Family Switches Only

With hardware forwarding, the switch stores addresses in the hardware table to quickly forward packets to their destination.

The switch uses a hash algorithm to decide where to store the addresses in the hardware table. The standard, default hash algorithm works well for most systems; however, for some addresses with certain patterns, the hardware may attempt to store address information in the same section of the hardware. This can cause an overflow of the hardware table even though there is enough room to store addresses.

### *Error Messages Displayed With ExtremeXOS 11.4 and Earlier*

If you experience a full hardware table that affects Layer 2, IP local host, and IP multicast forwarding, you see messages similar to the following in the log:

```
<Info:HAL.IPv4Adj.Info> : adj 136.159.188.109: IP add error is Table full for new or
newly resolved ARP, egress valid
```

```
<Info:HAL.IPv4Adj.Info> : adj 136.159.188.109: returned -17 for L3 table bucket 181

<Warn:HAL.IPv4Mc.Warning> : Could not allocate a hardware S,G,V entry
(889f4648,effffffa,70) - hardware table resource exceeded (rv=-17).
```

*Error Messages Displayed With ExtremeXOS 11.5 and Later*

If you experience a full hardware table that affects Layer 2, IP local host, and IP multicast forwarding, you see messages similar to the following in the log:

```
<HAL.IPv4Adj.L3TblFull> MSM-A: IPv4 unicast entry not added.  Hardware L3 Table full.

<Card.IPv4Adj.Warning> Slot 4: IPv4 unicast entry not added.  Hardware L3 Table full.

<HAL.IPv4Mc.GrpTblFullEnt> MSM-A: IPv4 multicast entry (10.0.0.1,224.1.1.1,vlan 1) not
added.  Hardware Group Table full.

<Card.IPv4Mc.Warning> Slot-4: IPv4 multicast entry not added.  Hardware L3 Table full.
```

## Understanding the Error Reading Diagnostics Message

For more detailed infomation about error messages and their meanings, see the *ExtremeXOS 22.3 EMS Messages Catalog*.

### Summit Family Switches Only

If you have never run diagnostics on the switch or stack ports and use the `show diagnostics` command, the switch displays a message similar to the following:

```
Result: FAIL
Test date run is invalid. Please run Diagnostics.
Error reading diagnostics information.
```

This message is normal and expected if you have never run diagnostics on the switch. After running diagnostics, you see information about the executed test using the `show diagnostics` command.

## Proactive Tech Support

Proactive tech support functionality allows you to collect system information when support is required. System information is pushed into a cloud-hosted collector where the Extreme TAC can use it to identify problems to provide solutions. Proactive Tech Support provides the following functionality:

- Proactive tech support is implemented as a loadable application, called techSupport.xmod. The XMOD is bundled with EXOS image packages, but can also be upgraded independently at any time without restarting the switch.
- Proactive tech support allows the switch to proactively collect switch status and send the status report to up to four cloud-hosted collectors.
- Proactive tech support allows both automatic and manual status report.
- Proactive tech support allows each collector to have different report data set, report frequency, or report mode. A default collector is configured automatically to minimize the configuration required to enable the feature.

## Limitations

- Proactive tech support uses SSL to secure the switch information transmission on the internet. SSL functionality is provided by a separate XMOD, called ssh.xmod. If the feature is enabled and ssh.xmod is not installed, the switch information is transmitted as clear text.
- XMOD applications can be dynamically upgraded without restarting the switch . But the upgraded XMOD version needs to match the installed EXOS version.

## Locating a Collector

The switch attempts to locate a data collector using the DNS hostname, or the IP address of a collector. You specify the DNS hostname or IP address of a collector when you add a collector. When a DNS hostname is configured for a collector, the switch attempts to resolve the IP address for the configured DNS hostname before it attempts to collect to the collector. If the process of resolving IP address for a collector fails, the connection process stops and logs an error message.

When you start the feature a default collector is automatically configured with an IP address set to 12.38.14.200; this is a public IP address of a Splunk server hosted in Extreme's internal network. When you enable the feature, the switch automatically attempts to locate the default collector and send switch status information to the collector. You can configure up to four collectors on the switch. If you want to collect switch status information in your own collector (beside the default collector), you can use the `configure tech-support add collector` command to configure an additional collector. The switch will send independent status reports to each collector based on each collector's configuration.

By default, the primary IP address on the Mgmt *VLAN* interface is used for the discovery process of a collector. If an IP address is not found on the Mgmt VLAN, an error messaged is displayed. Optionally, you can specify a source IP address to connect to a collector. The connection between the switch and a collector is established and maintained with TCP.

## Data Collection

These sections discuss various ways that data is collected and presented.

### *Encrypted or Open Text*

Status reports are sent either in encrypted or clear text. By default, this is determined by the installation of the SSL module. If the SSL module is installed, the status report is sent using an SSL over TCP session in encrypted text. Otherwise, the status report is sent over a regular TCP session in clear text. You can disable the SSL transport even if the SSL module is installed by using the `configure tech-support collector [`*`hostname`* `|` *`ip_address`*`] tcp-port` *`port`* `{vr` *`vr_name`*`} {from` *`source_ip_address`*`} {ssl [on | off]}` command.

### *Push vs. Pull*

For security reasons, status reports are never pushed from a collector out to a switch. The switch always pushes status reports to a collector. The first release of the tech-support application supports local configuration only. The switch pushes status reports to a collector based solely on the local configuration.

*Report Format*

The status report format is the output of the `show tech-support` command. You can limit the amount of information in the report using the **`data-set`** configuration parameter. Two data-sets are available in the first release.

*Frequency*

You can also configure when the switch sends a status report. The switch can send a report based on a critical severity event occurrence, daily, at boot-up, or manually when directed by the network administrator. Because of the amount of information that can be transmitted, automatic status reports are sent no more frequently than one report per hour. When a series of critical severity events occurs, only the first one triggers the switch to send status reports. All other critical severity events that occur within one hour after the first critical severity event are ignored.

*Authentication and Verification*

When SSL is enabled, the switch can authenticate the collector during the SSL handshake by verifying whether the collector's certificate is issued by one of switch's trusted Certificate Authorities (CA). In this case, the switch has to know the CA that issued the collector certificate.

When SSL is enabled, the collector can also authenticate the switch during the SSL handshake by requesting that the switch send the certificate, and then verify that the switch's certificate is issued by one of collector's trusted CAs. In this case, the switch has to have a certificate, and the collector has to know the CA that issued the switch the certificate.

To keep the required configuration minimal, the ExtremeXOS 15.4 release of the tech-support application does not perform authentications. The switch does not verify that the collector is a valid, authorized server before the transmission. The collector does not verify that received status reports are transmitted from a valid, authorized device either.

You can discard or purge the status report from the database for the following reasons:

- The status report is not in a valid format.
- The status report does not contain the minimum required information.
- The serial number of the hardware in the status report is invalid or does not match any serial number in Extreme Network's manufacturing database.
- The status report is received from the same device more than once per hour.

## Configuring Proactive Tech Support

The examination and reporting of switch status is implemented as a loadable application, called techSupport.xmod. You can dynamcally upgrade the XMOD applications without restarting the switch by issuing `download image` *`ip_address app.xmod`*, `run update`, and then `restart process techSupport` commands.

*Safe Default Startup*

By default proactive tech support services are disabled. You can opt-out of this service either by answering "Yes", "Y", or hitting the <enter> key to answer the question that is posed when accessing the switch for the first time. You can also use the `disable tech-support` collector command. Here is the system output prompt for the tech support feature:

```
The switch will proactively attempt to send basic configuration and
operational switch information for the purpose of assisting Extreme
Networks resolve customer reported issues. Uploaded data is encrypted if
the ssh.xmod is installed. Otherwise, a reduced switch data set is sent
in clear text that contains no customer specific information.

Would you like to disable the automatic switch reporting service? [Y/n]:
```

## Service Verification Test Tool

The EXOS Service verification tool (ESVT) tests traffic throughput between two Extreme Network switches without having to use a traffic generator. The figure below shows a typical service verification test tool environment:

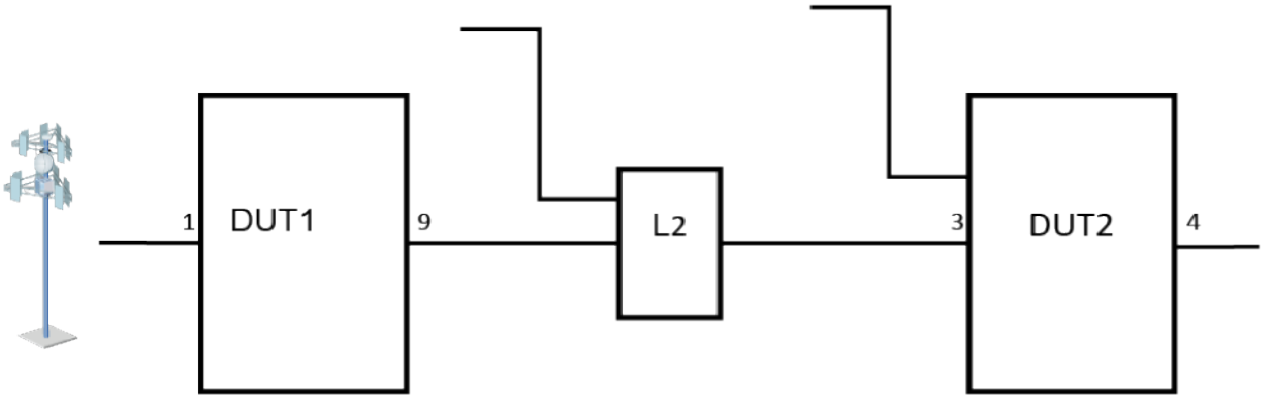



**Figure 254: Typical service verification test tool environment**

As shown in the figure above, you have an Extreme Networks switch DUT1 installed at a customer location, and customer service is defined to be from the antenna to the Extreme Networks switch DUT2 (and possibly beyond the DUT2). You can use the service verification test tool to test the portion of the service between the DUT1 and the DUT2. The connection between the antenna and the DUT1 is not covered by the test tool. Additionally, any continuation of the service beyond the DUT2 is not included in the test.

Prior to beginning the test, you must provision the part of the network to be tested by the service verification test tool. This includes the *VLAN* (tagged or untagged on port 9) on the DUT1, any L2 devices in the path through the network, and the VLAN (tagged or untagged on port 3) at the DUT2. Additionally, any L2/L3 protocol packets created in the CPU on any of the above devices, and sent over the same VLAN as the service being verified, will create errors in the test results. All L2/L3 protocols for the service VLAN must be disabled unless specifically included in the test tool instructions.

The most straightforward mode of operation is for the DUT1 to generate sufficient test packets to fill the allocated bandwidth of the service going towards the DUT2. The DUT1 counts the test packets sent towards the DUT2. On the DUT2, the simplest mode of operation is to wrap the received test packets back to the DUT1 using the reverse path of the service connection. The DUT1 counts the received test packets. At the conclusion of the test, the DUT1 displays the total number of test packets sent and received.

As line rate test traffic cannot be generated by the CPU in the DUT1, generating test packets must make use of some hardware resources. You can form an internal loop by placing a single port into loopback mode. You can generate line rate traffic by placing a test packet into this loop. You can then use a

metered (internal) connection to the service VLAN to send test packets into the service under test at the desired bandwidth. You can assign a loopback port using a CLI command that will mimic the QOS of egress port (port 9).

At the DUT2, it is best to update the source MAC address of the test packets as they are forwarded back to the DUT1. This prevents MAC learning problems in any L2 devices between the DUT1 and the DUT2 that might be caused by learning the same source MAC address on two different ports. By making a L3 forwarding decision at the DUT2, the source MAC address of the test packets is modified to be that of the DUT2.

To make an L3 forwarding decision at the DUT2, the destination MAC address of the test packets must be the address of the DUT2.The destination IP address lookup on the DUT2 must point to an IP next hop that is the DUT1. The DUT1 will also modify the destination MAC address of the test packets to be that of the DUT1.

To make the test as simple as possible, you should assign IP addresses to the service VLAN interfaces at the DUT1 and the DUT2. These IP addresses should be from private IP address space, and should be in the same IP subnet. For example, you could assign 15.15.15.14/24 to the DUT1 and 15.15.15.15/24 to the DUT2. These temporary IP addresses should be unconfigured as soon as the service verification test is completed.

Note

The service being verified is an L2 service only. The test network must not cross L2 boundaries.

## Supported Platforms

The Service Verification Tool is supported on the following platforms: X450-G2, X460, X480, E4G-200, E4G-400, X670, X670G2, X770, Stacking, BDX8 and BD8800.

## Configuring the Service Verification Test Tool

To run a service verification test, issue the following command:

**run esvt traffic-test {vlan}** *vlan_name* **loopback-port** *loopback-port* **peer-switch-ip** *ipaddress* **packet-size** *packet_size* **rate** *rate* **[Kbps|Mbps | Gbps] duration** *time* **[seconds | minutes | hours]**

Here is an example configuration to run ESVT for the environment shown in Figure 1:

DUT1

```
create vlan v1
configure vlan v1 tag 100
configure vlan v1 add port 9 tagged
configure vlan v1 ipaddress  2.2.2.3/24
enable ipforwarding v1
run esvt traffic-test v1 loopback-port 2 peer-switch-ip 2.2.2.2 packet-size 64 rate 200
Mbps duration 5 minutes
 show esvt traffic-test
```

DUT2

```
create vlan v1
configure vlan v1 tag 100
```

```
configure vlan v1 add port 3 tagged
configure vlan v1 ipaddress  2.2.2.2/24
enable ipforwarding v1
```

To stop the service verification test, issue the following command:

**stop esvt traffic-test {{vlan}** *vlan_name*}

To show measurements from the service verification test, issue the following command:

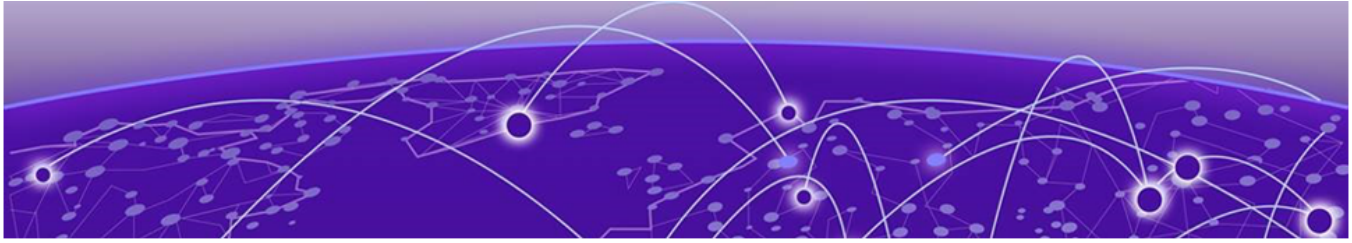
**show esvt traffic-test {{vlan}** *vlan_name* **}**

# Supported Standards, Protocols, and MIBs

This appendix provides information about the MIB support provided by the ExtremeXOS *SNMP (Simple Network Management Protocol)* agent residing on Extreme Networks devices running ExtremeXOS and Extreme Networks proprietary MIBS.

Note

All information for supported standards and protocols can be found in the following location: http://www.extremenetworks.com/resources. This appendix contains only the MIB support details.



# Extreme Networks Proprietary MIBs



The Extreme Networks MIBs are located on the eSupport web site under Download Software Updates, located at https://esupport.extremenetworks.com.

## EXTREME ACL MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeAclStatsTable | extremeAclStatsVlanIfIndex | The IfIndex of the *VLAN (Virtual LAN)* in which this policy/rule is applied |
| | extremeAclStatsPortIfIndex | The IfIndex of the the port in which this policy/rule is applied |
| | extremeAclDirection | The ingress/egress direction to which this policy/rule is applied (0 for ingress and 1 for Egress) |
| | extremeAclStatsCounterName | Name of the counter for which the stats is requested |
| | extremeAclStatsPktCount | The total number of packets that matches this rule |
| | extremeAclStatsByteCount | The total number of bytes that matches this rule |

## ExtremeXOS Configuration Management Enhancements

This feature addresses network management requirements identified for the Ridgeline product. The following enhancements are part of this release:

- Expose the time and date of the last configuration save operation.
- Issue a notification to NMS upon completion of a database save operation.
- Provide the ability for EXOS to report on demand the presence of unsaved configuration changes on the system.
- Issue a notification whenever any system configuration is altered, added, or deleted.
- Provide the ability for NMS to download an image file to an EXOS device. MIB objects reflecting time of next scheduled reboot.

A configuration management MIB, EXTREME-CFGMGMT-MIB, is introduced to help facilitate this functionality. The ability to query configuration activity and to perform switch management operations can be done by using the tables, scalars, and traps supported in the configure management, and the system (EXTREME-SYSTEM-MIB) MIBs.

## EXTREME-MAC-AUTH-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeMacAuthClientTable | All objects | This table contains client authentication based on the source MAC address of the traffic received on a port. |
| ExtremeMacAuthClientEntry | | An entry in the MAC authentication table. Each entry represents a MAC authentication client |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeMacAuthClientAddress | The MAC address of the client. |
| | extremeMacAuthClientInitialize | Remove the entry for the client from extremeMacAuthClientTable. |
| | extremeMacAuthClientReauthenticate | Re-authenticate the client on all ports on which the client is connected. |

## EXTREME-AUTOPROVISION-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeAutoProvisionConfig | | Enables or disables the Auto Provision feature on the switch. |

The following traps can be generated.

| Trap | Notification Objects | Comments |
|---|---|---|
| extremeAutoProvisionStatus | | This trap reports the auto provision result (success/failed). It contains the attributes it received from the *DHCP (Dynamic Host Configuration Protocol)* server. |
| | extremeAutoProvisionResult | Result of the Auto provision. |
| | extremAutoProvisionIpAddress | The IP address received from the DHCP server for this interface. |
| | extremeAutoProvisionGateway | The gateway address received from the DHCP server for this interface. |
| | extremeAutoProvisionTFTPServer | The IP address of the TFTP server received from the DHCP server. |
| | extremeAutoProvisionConfigFileName | The configuration file name received from the DHCP server which the auto-provision enabled switch will download from the TFTP server and apply. It can use either the cfg or xsf extension. |

## EXTREME-CFM-MIB

The ExtremeCFM MIB provides information about the Connectivity Fault Management (CFM) Group. This is an extension to IEEE8021- CFM-MIB.

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeCfmGroup | ExtremeCfmGroupOperStatus | This is a textual convention which, indicates the operational status of a group associated with a MEP on a port of an association in a given domain. |
| extremeCfmGroupNextIndexTable | | This object contains an unused value for extremeCfmGroupIndex in the extremeCfmGroupTable, or a zero to indicate that none exists. |
| | extremeCfmGroupNextIndexEntry | The conceptual row of extremeCfmGroupNextIndexTable. Not accessible. |
| | extremeCfmGroupNextIndex | Value used as the index of the Group table entries, for this Maintenance association End Point Identifier when the management entity wants to create a new row in that table. Read only. |
| extremeCfmGroupTable | | Includes configuration objects and operations for the Group function, mainly used by the registered clients like ERPS, *EAPS (Extreme Automatic Protection Switching)* to know link detection failure through CFM. Each row in the table represents a Group for the defined MEP. |
| | extremeCfmGroupEntry | This is a complex index with four indices to access any row in extremeCfmGroupTable. The first three are Maintenance Domain, MaNet, and MEP indices. The fourth index is the specific Group on the selected MEP. Not accessible. |
| | extremeCfmGroupName | The name of a CFM group. Group name must be alpha-numeric. |
| | extremeCfmGroupStatus | Whether the group is operational or not. Ready only. |
| | extremeCfmMepIfIndex | Interface index of the interface, either a bridge port, or an aggregated IEEE 802.1 link within a bridge port, to which the MEP and hence the group is attached. Upon a restart of the system, the system, if necessary, changes the value of this variable so that it indexes the entry in the interface table with the same value of ifAlias that it indexed before the system restart. If no such entry exists, then the system sets this variable to 0. Read only. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeCfmGroupRemoteMEPs | Lists the Remote MEPs associated with a group. Not all Remote MEPs of an MA may be associated with a group. Ready only. |
| | extremeCfmGroupClients | Lists all the registered clients with a group. The clients are informed with link failure or recovery through group status notifications. Ready only. |
| | extremeCfmGroupRowStatus | The status of the row. All columns must have a valid value before a row can be activated. |
| extremeCfmGroupMepDbTable | | Maintains information about other MEPs in that group. |
| | extremeCfmGroupMepDbEntry | Complex index including Maintenance Domain, MaNet, MEP and Group indices along with RMEP ID to identify a row in this table.<br>Not accessible. |
| | extremeCfmGroupMepDbRMepId | Maintenance association End Point Identifier of a remote MEP whose information from the group MEP database is to be returned.<br>Not accessible. |
| | extremeCfmGroupMepDbRowStatus | The status of the row. All columns must have a valid value before a row can be activated. |
| extremeCfmGroupStatusDownUpAlarm | extremeCfmGroupStatus | A notification (DownUpAlarm) is sent to the management entity with the OID of the group that has detected the status change. |

## EXTREME-CLEARFLOW-MIB

This MIB defines the following Extreme-specific CLEAR-Flow traps generated by Extreme Networks devices.

| Trap | Comments |
|---|---|
| extremeClearflowMessage | CLEAR-Flow message trap. |
| The varbinds supported in this trap are: | |
| extremeClearflowMsgId | User-defined message ID. |
| extremeClearflowMsg | User-defined message. |
| extremeClearflowPolicyName | Policy file name. |
| extremeClearflowRuleName | Rule name which triggered this message. |
| extremeClearflowRuleValue | Calculated rule value. |
| extremeClearflowRuleThreshold | Rule threshold value. |

| Trap | Comments |
|---|---|
| extremeClearflowRuleInterval | Rule sampling and evaluation interval. |
| extremeClearflowVlanName | *VLAN* name on which this policy is applied. |
| extremeClearflowPortName | Port name on which this policy is applied. |

## EXTREME-EAPS-MIB

Managed objects for *EAPS* MIB are defined proprietary on ExtremeXOS.

The following table list the existing traps supported in this MIB.

| Trap | Varbinds | Comments |
|---|---|---|
| extremeEapsStateChange | extremeEapsName, extremeEapsMode, extremeEapsPrevState, extremeEapsState, extremeEapsFailedFlag, extremeEapsPrimaryStatus extremeEapsSecondaryStatus | Send on Master/Transit Nodes |
| extremeEapsLastStatusChangeTime | extremeEapsLastStatusChange extremeEapsStatusTrapCount | Send on Master/Transit Nodes. General indication of a status change using 10 second timer. |
| extremeEapsPortStatusChange | extremeEapsName, extremeEapsPrimaryStatus extremeEapsSecondaryStatus extremeEapsLastStatusChange | Send on Master/Transit Nodes |
| extremeEapsConfigChange | extremeLastConfigurationChange | 30 second granularity Send on Master/Transit Nodes |
| extremeEapsSharedPortStateChange | extremeEapsSharedPortIfIndex extremeEapsSharedPortLinkId extremeEapsSharedPortState extremeEapsSharedPortNbrStatus extremeEapsSharedPortRootBlockerStatus extremeEapsLastStatusChange | Send on Controllerr/Partner Nodes |
| extremeEapsRootBlockerStatusChange | extremeEapsSharedPortIfIndex extremeEapsSharedPortRootBlockerStatus extremeEapsSharedPortRootBlockerId extremeEapsLastStatusChange | Send on Controller/Partner Nodes |

The following table lists the MIB entries.

| MIB Variable | Description |
|---|---|
| extremeEapsLastStatusChange | Time gets updated for any of the following changes... extremeEapsState, extremeEapsSharedPortState extremeEapsSharedPortSegmentStatus extremeEapsFailedFlag, extremeEapsPrimaryStatus extremeEapsSecondaryStatus extremeEapsSharedPortNbrStatus extremeEapsSharedPortRootBlockerStatus extremeEapsSharedPortSegmentFailedFlag |
| extremeEapsStatusTrapCount | Indicates the number of status traps sent out since the switch booted. Status traps include extremeEapsStateChange, extremeEapsPortStatusChange, extremeEapsSharedPortStateChange, extremeEapsRootBlockerStatusChange |

## EXTREME-EDP-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeEdpTable | All objects | This table contains *EDP (Extreme Discovery Protocol)* information of this device. |
| extremeEdpNeighborTable | All objects | This table contains EDP neighbor information. |
| extremeEdpPortTable | All objects | |

## EXTREME-ENTITY-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variable | Comments |
|---|---|---|
| extremeEntityFRUTable | entPhysicalIndex | A table containing information about each FRU in the chassis based on the Entity MIB. |
| | extremeEntityFRUStartTime | First recorded start time. |
| | extremeEntityFRUOdometer | Number of time units in service. |
| | extremeEntityFRUOdometer Unit | Time unit used to represent value reported by extremeEntityFRUOdometer. Depends on the underlying hardware capability. |

## EXTREME-ERPS-MIB

Managed objects for *ERPS (Ethernet Ring Protection Switching)* MIB are defined proprietary on ExtremeXOS. Groups and tables are implemented as read only.

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeErpsProtectedVlanTable | All objects | Contains the grouping of set of protected VLANs. |
| extremeErpsRingTable | All objects | Contains information for each ERPS ring present in the switch. |
| extremeErpsStatsTable | All objects | Contains statistics information for each of the ring present in the switch. |
| extremeErpsGlobalInfo | | Contains the information of ERPS configured globally in the switch. |
| extremeErpsNotification | | Contains two types of traps: extremeErpsStateChangeTrap and extremeErpsFailureTrap. extremeErpsStateChangeTrap is generated on the following events:<br>• Local SF is received for the ring.<br>• Local Clear SF is received for the ring.<br>• Remote failure is detected on this ring.<br>• Remote failure is cleared on this ring.<br>• Force Switch is issued for a ring.<br>• Manual Switch is issued for a ring. |

## EXTREME-ESRP-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeEsrpDomainTable | All objects | This table contains information for *ESRP (Extreme Standby Router Protocol)* domains on this device. |
| extremeEsrpDomainMemberTable | All objects | This table contains information for member VLANs of the ESRP domain. |
| extremeEsrpDomainNeighborTable | All objects | This table contains neighbor router information for ESRP domains on this device. |
| extremeEsrpDomainAwareTable | All objects | This table contains ESRP aware information for this device. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeEsrpDomainStatsTable | All objects | This table contains statistics on ESRP hello packets exchanged and ESRP state changes for this device. |
| extremeEsrpNotifications | extremeEsrpDomainStateChange | Signifies ESRP state change. |

## EXTREME-FDB-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeFdbIpFdbTable | All objects | Supports *SNMP (Simple Network Management Protocol)* get and get next operations only. |
| extremeFdbPermFdbTable | All objects | |
| extremeFdbMacExosFdbTable | extremeFdbMacExosFdbEntry | A table that contains information about the hardware MAC *FDB (forwarding database)* table. Supported only on switches running ExtremeXOS. Supports SNMP get and get next operations only. |
| extremeFdbMacFdbCounterTable | All objects | Supports SNMP get and get next operations only. |

| Trap | Comments |
|---|---|
| extremeMACTrackingAdd | The specified MAC address was added to the FDB on the mentioned port and *VLAN*. |
| extremeMACTrackingDel | The specified MAC address was deleted from the FDB on the mentioned port and VLAN. |
| extremeMACTrackingMove | The specified MAC address was moved from the previous port to the new port on the specified VLAN. |

## EXTREME-MPLS-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeMplsNotifications | extremepwUpDownNotificationEnable | If this object is set to true (1), then it enables the emission of pwUp and pwDown notifications; otherwise these notifications are not emitted. |
| | extremepwDeletedNotificationEnable | If this object is set to true (1), then it enables the emission of pwDeleted notification; otherwise this notification is not emitted. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | pwNotificationMaxRate | This variable indicates the maximum number of notifications issued per second. If events occur more rapidly, the implementation may simply fail to emit these notifications during that period, or might queue them until an appropriate time. A value of 0 means no throttling is applied and events might be notified at the rate at which they occur. |
| | extremepwNotificationPwIndex | This variable indicates the index of the PW that either went up, down, or was deleted as reported by the corresponding notification. |
| | extremepwNotificationPwOperStatus | This variable is used to report the value of pwOperStatus in pwTable (RFC 5601) associated with the PW that went up or down. |
| | extremepwNotificationPeerAddrType | Denotes the address type of the peer node. |
| | extremepwNotificationPeerAddr | This object contains the value of the peer node address. |
| | extremeMplsNotifTunnelIndex | Uniquely identifies a set of tunnel instances between a pair of ingress and egress LSRs. Contains part of the index of a TE tunnel that underwent state change. The same tunnel can also be looked up in the TE MIB. |
| | extremeMplsNotifTunnelInstance | Uniquely identifies a particular instance of a tunnel between a pair of ingress and egress LSRs. |
| | extremeMplsNotifTunnelIngressLSRId | Identity of the ingress LSR associated with this tunnel instance. |
| | extremeMplsNotifTunnelEgressLSRId | Identity of the egress LSR associated with this tunnel instance. |
| | extremeMplsNotifTunnelAdminStatus | Reports the desired operational status of this tunnel. |
| | extremeMplsNotifTunnelOperStatus | Reports the actual operational status of this tunnel, which is typically but not limited to, a function of the state of individual segments of this tunnel. |
| | extremeMplsNotifLdpEntityLdpId | The LDP identifier. This is the primary index to identify a row in the mplsLdpEntityTable (RFC 3815). |
| | extremeMplsNotifLdpEntityIndex | This index is used as a secondary index to uniquely identify a row in the mplsLdpEntityTable (RFC 3815).<br>This object is meaningful to some, but not all, LDP implementations. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeMplsNotifLdpPeerLdpId | The LDP identifier of this LDP Peer. |
| | extremeMplsNotifLdpSessionState | The current state of the session, all of the states 1 to 5 are based on the state machine for session negotiation behavior. |
| | extremeMplsNotifLdpSessionDiscontinuityTime | The initial value of this object is the value of sysUpTime when the entry was created in this table. Subsequent notifications report the time when the session between a given entity and peer goes away or a new session is established. |
| | extremeVplsNotifConfigIndex | Unique index for the conceptual row identifying a VPLS service in the vplsConfigTable. |
| | extremeVplsNotifConfigVpnId | This objects indicates the IEEE 802-1990 VPN ID of the associated VPLS service. This object has the same value as vplsConfigVpnId in the vplsConfigTable for an index value equal to extremeVplsNotifConfigIndex sent in the notification. |
| | extremeVplsNotifConfigAdminStatus | The administrative state of the VPLS service.<br>This object has the same value as vplsConfigAdminStatus in the vplsConfigTable for an index value equal to extremeVplsNotifConfigIndex sent in the notification. |
| | extremeVplsNotifStatusOperStatus | The current operational state of this VPLS service. A value of up (1) indicates that all PWs for this VPLS are up and the attachment circuit is up.<br>A value of degraded (2) indicates that at least one PW for this VPLS is up and the attachment circuit is up.<br>A value of down (3) indicates that all PWs for this VPLS are down or the attachment circuit is down. |
| | extremePwStatusChange | This notification is generated when the pwOperStatus object for a pseudowire transitions from up (1) to down (2) or from down (2) to up (1). |
| | extremePwDeleted | This notification is generated when a PW has been deleted. |
| | extremeMplsTunnelStatusChange | This notification is generated when the mplsTunnelOperStatus object for a TE-LSP transitions from up (1) to down (2) or from down (2) to up (1). This new state is indicated by the included value of mplsTunnelOperStatus. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeMplsLdpSessionStatusChange | This notification is generated when the value of 'mplsLdpSessionState' (RFC3815) enters or leaves the operational (5) state. |
| | extremeVplsStatusChange | This notification is generated to inform recipients of the state of the VPLS.<br>When all PWs in this VPLS are up or ready and the attachment circuit is up, extremeVplsNotifStatusOperStatus is set to vplsOperStatusUp (1) in the notification.<br>When at least one PW in this VPLS is up or ready and the attachment circuit is up, extremeVplsNotifStatusOperStatus is set to vplsOperStatusDegraded (2) in the notification.<br>When all PWs in this VPLS are down or the attachment circuit is down, extremeVplsNotifStatusOperStatus is set to vplsOperStatusDown (3) in the notification.<br>Once a notification has been sent with vplsOperStatusDegraded (2), no further notification will be sent until extremeVplsNotifStatusOperStatus transitions to vplsOperStatusUp (1) or vplsOperStatusDown (3). |
| extremeVplsConfigTable | | This table specifies information for configuring and monitoring VPLS. |
| | extremeVplsConfigEntry | A row in this table represents a VPLS in a packet network. It is indexed by extremeVplsConfigIndex, which uniquely identifies a single VPLS. |
| ExtremeVplsConfigEntry | extremeVplsConfigIndex | Unique index for the conceptual row identifying a VPLS service. |
| | extremeVplsConfigRedunType | This variable indicates the redundancy type for this VPLS. Redundancy states can be None, EAPS, *ESRP* and *STP (Spanning Tree Protocol)*. |
| | extremeVplsConfigEAPSStatus | This variable indicates the EAPS redundancy status for this VPLS. It displays Protected if the *VLAN* is protected, else Not Protected. EAPS status displays the following values: Waiting—if waiting on EAPS status.Connected—if EAPS ring status is connected to VLAN. PW is not installed in this state.Disconnected—if EAPS ring state is disconnected. PW is installed in this state.Unknown—if EAPS is not found. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVplsConfigESRPState | This variable indicates the ESRP redundancy state for this VPLS. ESRP states can have two values: master and slave. PW is installed only when ESRP is in the master state. |
| extremeVplsStatusTable | | This table specifies information for configuring and monitoring VPLS. |
| | extremeVplsStatusEntry | A row in this table represents a VPLS in a packet network. It is indexed by extremeVplsConfigIndex, which uniquely identifies a single VPLS. |
| | extremeVplsStatusIndex | Unique index for the conceptual row identifying a VPLS service. |
| | extremeVplsOperStatus | This variable indicates the VPLS operation status. This is different from the VPLS PW status shown in the standard MIB. Supported values are:<br>Up—all PWs are in the Up or Ready state.Degraded—At least one PW is in the Up or Ready state.Down—None of the PWs are in Up or Ready state. |
| extremeVplsPeerTable | | This table provides information for monitoring VPLS peer entries. |
| | extremeVplsPeerEntry | This entry contains information for all the peers belonging to a particular VPLS. |
| | extremeVplsIndex | Unique index for the conceptual row identifying a VPLS service. |
| | extremeVplsPwBindIndex | Secondary Index for the conceptual row identifying a PW within the PwEntry which must match an entry from the PW-STD-MIB's PwTable, which represents an already-provisioned PW that is then associated with this VPLS instance. |
| | extremePwInstalled | Boolean true or false, indicating if PW is installed. |
| | extremePwMode | The peer mode of this PW. It indicates its current mode and the mode of the switch it is connected to. It can have the following values: Core-to-core, Core-to-Spoke, and Spoke-to-Core. |
| | extremePwConfiguredRole | The configured role of this PW. It applies only in the case the PWMode is Core-to-core. In this case, the configured role can either be Primary or Secondary. For all the other PW modes, the configured role does not apply. |

The following traps can be generated.

| Trap | Comments |
|---|---|
| pwStatusChange | This notification is generated when the pwOperStatus object for a PW transitions from up (1) to down (2) or from down (2) to up (1). |
| pwDeleted | This notification is generated when a PW has been deleted. |
| extremeMPLSTrapsEnable | *MPLS (Multiprotocol Label Switching)* trap status. If enabled then all the MPLS related traps are sent out. |
| extremeL2VpnTrapsEnable | L2VPN trap status. If enabled then all the L2VPN related traps are sent out. |

## EXTREME-MPLS-TE-MIB

The following tables, groups, and variables are supported in this MIB:

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeMplsTunnelTable | mplsTunnelIndex | The mplsTunnelTable (see RFC3812) allows new *MPLS* tunnels to be created between an LSR and a remote endpoint, and existing tunnels to be reconfigured or removed. Note that only point-to-point tunnel segments are supported, although multipoint-to-point and point- to-multipoint connections are supported by an LSR acting as a cross-connect. Each MPLS tunnel can thus have one out-segment originating at this LSR and/or one in-segment terminating at this LSR. Extreme Networks MPLS implementation allows tunnel instances with a common endpoint to be grouped at the ingress LSR to provide redundancy. The role of each tunnel in the group must be configured and is indicated by extremeMplsTunnelRedundancyType. |
| | mplsTunnelInstance | Uniquely identifies a particular instance of a tunnel between a pair of ingress and egress LSRs. It is useful to identify multiple instances of tunnels for the purposes of backup and parallel tunnels. |
| | mplsTunnelIngressLSRId | Identity of the ingress LSR associated with this tunnel instance. |
| | mplsTunnelEgressLSRId | Identity of the egress LSR associated with this tunnel instance. |
| | mplsTunnelRedundancyType | Identifies the tunnel redundancy type associated with this tunnel instance. A value of primary (1) or secondary (2) MAY be assigned by the network administrator or by an *SNMP* manager at the time of setting up the tunnel. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | mplsTunnelRedundancyStatus | Indicates the actual redundancy status of this tunnel. When the status is active, the tunnel is the preferred tunnel in the group. |
| | mplsTunnelTransportStatus | Indicates the type of traffic the tunnel group can be used for sending.<br>When the status is allowAllIp (0), IP traffic destined for all IPv4 routes is allowed over any tunnel in the group marked active.When the status is allowAssignedIpOnly (1), IP traffic destined only for IPv4 static routes that have been explicitly configured to use this tunnel group is allowed.When the status is allowAllLayer 2Vpn (2), Layer 2 VPN traffic for all Layer 2 VPNs is allowed over any tunnel in the group marked active.When the status is allowAssignedLayer 2VpnOnly (3), Layer 2 VPN traffic destined only for PWs that have been explicitly configured to use this tunnel group are allowed. |

## EXTREME-OSPF-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeOspfInterfaceTable | All objects | This table contains Extreme Networks specific information about *OSPF (Open Shortest Path First)* interfaces. |

## EXTREME-PoE-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremePethSystemAdminEnable | | Objects are supported read-only. |
| extremePethSystemDisconnectPrecedence | | |
| extremePethSystemUsageThreshold | | |
| extremePethSystemPowerSupplyMode | | |
| extremePethPseSlotTable | All objects | |
| extremePethPsePortTable | All objects | |

## EXTREME-PORT-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremePortLoadshareTable | All objects | Not supported. |
| extremePortSummitlinkTable | All objects | Not supported. |
| extremePortLoadshare2Table | extremePortLoadshare2Entry | A table of bindings between a master port and its load-sharing slaves. Create/delete entries here to add/delete a port to/from a load-sharing group. Default is empty table. There are restrictions on what row creates will be accepted by each device. |
| | extremePortLoadshare2MasterIfIndex | The if Index value which identifies the port controlling a load-sharing group of ports that includes extremePortLoadshareSlaveIfIndex. |
| | extremePortLoadshare2SlaveIfIndex | The if Index value which identifies the port which is a member of a load-sharing group controlled by extremePortLoadshare2MasterIfIndex. |
| | extremePortLoadshare2Algorithm | This value identifies the load sharing algorithm to be used for this group of load shared ports. |
| | extremePortLoadshare2Status | The row status variable, used according to row installation and removal conventions. |
| | extremePortLoadshare2MinLinks | The minimum active links that must be up in order trunk to remain up. |
| extremePortRateShapeTable | All objects | Not supported |
| extremePortUtilizationTable | extremePortUtilizationEntry | Global *QoS (Quality of Service)* profiles are defined in the extremeQosProfileTable. This table contains a list of ports for which certain QoS parms are reported. |
| | extremePortUtilizationAvgTxBw | The reported average bandwidth in the transmit direction. When displayed, it shows as an integer value. For example, 99.99% is displayed as 9999. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremePortUtilizationAvgRxBw | The reported average bandwidth in the receive direction. When displayed, it shows as an integer value. For example, 99.99% is displayed as 9999. |
| | extremePortUtilizationPeakTxBw | The reported peak bandwidth in the transmit direction. When displayed it shows as an integer value. For example, 99.99% is displayed as 9999. |
| | extremePortUtilizationPeakRxBw | The reported peak bandwidth in the receive direction. When displayed it shows as an integer value. For example, 99.99% is displayed as 9999. |
| extremePortInfoTable | All objects | Not supported. |
| extremePortXenpakVendorTable | All objects | Not supported. |
| extremePortIngressStatsPortTable | All objects | Not supported. |
| extremePortIngressStatsQueueTable | All objects | Not supported. |
| extremePortEgressRateLimitTable | All objects | Not supported. |
| extremeWiredClientTable | All objects | Not supported. |
| extremePortUtilizationExtnTable | extremePortUtilizationExtnEntry | Global QoS profiles are defined in the extremeQosProfileTable. This table contains a list of ports for which certain QoS parameters are reported. |
| | extremePortUtilizationAvgTxPkts | The reported number of average packets in the transmit direction per second. |
| | extremePortUtilizationAvgRxPkts | The reported number of average packets in the receive direction per second. |
| | extremePortUtilizationPeakTxPkts | The reported number of peak packets in the transmit direction per second. |
| | extremePortUtilizationPeakRxPkts | The reported number of peak packets in the receive direction per second. |
| | extremePortUtilizationAvgTxBytes | The reported number of average bytes in the transmit direction per second. |
| | extremePortUtilizationAvgRxBytes | The reported number of average bytes in the receive direction per second. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremePortUtilizationPeakTxBytes | The reported number of peak bytes in the transmit direction per second. |
| | extremePortUtilizationPeakRxBytes | The reported number of peak bytes in the receive direction per second. |
| extremePortQosStatsTable | extremePortQosStatsEntry | This table lists Ports QoS information for either ingress or egress. |
| | extremePortQosIngress | Indicates whether the port is in ingress/egress. |
| | extremePortQP0TxBytes | The number of QoS 0 bytes that gets transmitted from this port. |
| | extremePortQP0TxPkts | The number of QoS 0 packets that gets transmitted from this port. |
| | extremePortQP1TxBytes | The number of QoS 1 bytes that gets transmitted from this port. |
| | extremePortQP1TxPkts | The number of QoS 1 packets that gets transmitted from this port. |
| | extremePortQP2TxBytes | The number of QoS 2 bytes that gets transmitted from this port. |
| | extremePortQP2TxPkts | The number of QoS 2 packets that gets transmitted from this port. |
| | extremePortQP3TxBytes | The number of QoS 3 bytes that gets transmitted from this port. |
| | extremePortQP3TxPkts | The number of QoS 3 packets that gets transmitted from this port. |
| | extremePortQP4TxBytes | The number of QoS 4 bytes that gets transmitted from this port. |
| | extremePortQP4TxPkts | The number of QoS 4 packets that gets transmitted from this port. |
| | extremePortQP5TxBytes | The number of QoS 5 bytes that gets transmitted from this port. |
| | extremePortQP5TxPkts | The number of QoS 5 packets that gets transmitted from this port. |
| | extremePortQP6TxBytes | The number of QoS 6 bytes that gets transmitted from this port. |
| | extremePortQP6TxPkts | The number of QoS 6 packets that gets transmitted from this port. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremePortQP7TxBytes | The number of QoS 7 bytes that gets transmitted from this port. |
| | extremePortQP7TxPkts | The number of QoS 7 packets that gets transmitted from this port. |
| extremePortMauTable | extremePortMauEntry | Port Optics Status Table. |
| | extremePortMauType | This object identifies the MAU type. |
| | extremePortMauVendorName | This object identifies the MAU vendor name. |
| | extremePortMauStatus | This object identifies the status of the MAU for this interface. |
| extremePortCongestionStatsTable | extremePortCongestionStatsEntry | This table lists ports congestion information. |
| | extremePortCongDropPkts | The number of packets dropped due to congestion on this port. |
| extremePortQosCongestionStatsTable | extremePortQosCongestionStats<br>Entry | This table lists ports per QoS congestion information. |
| | extremePortQP0CongPkts | The number of QoS 0 packets that got dropped due to congestion on this port. |
| | extremePortQP1CongPkts | The number of QoS 1 packets that got dropped due to congestion on this port. |
| | extremePortQP2CongPkts | The number of QoS 2 packets that got dropped due to congestion on this port. |
| | extremePortQP3CongPkts | The number of QoS 3 packets that got dropped due to congestion on this port. |
| | extremePortQP4CongPkts | The number of QoS 4 packets that got dropped due to congestion on this port. |
| | extremePortQP5CongPkts | The number of QoS 5 packets that got dropped due to congestion on this port. |
| | extremePortQP6CongPkts | The number of QoS 6 packets that got dropped due to congestion on this port. |
| | extremePortQP7CongPkts | The number of QoS 7 packets that got dropped due to congestion on this port. |

The following traps can be generated.

| Trap | Comments |
|---|---|
| extremePortMauChangeTrap | This trap is sent whenever a MAU is inserted or removed. When the MAU is inserted, the value of extremePortMauStatus is 'inserted' and extremePortMauType indicates the type of the MAU inserted. If MAU is removed, the value of extremePortMauStatus is empty and the type of the MAU will be NONE. |
| extremeRateLimitExceededAlarm | This notification indicates the first time a poll of a rate-limited port has a non-zero counter. |

## Pseudowire LSP Sharing MIB

By implementing the tables in PW LSP sharing MIB, the *SNMP* manager is able to observe the transmit packet counters over each LSP that is configured for used by the PW, and aggregated transmit and receive packet counters over PW itself. It consist three tables:

- extremePwPerfTable
- extremePwLspOutboundMappingTable
- extremePwLspPerfTable

There are no standard MIB tables and scalars objects that can be used for Pseudo Wire LSP Sharing. MIB table and object defined here are Extreme Networks proprietary. These three table are implemented in code/ l3protocol/mpls/src/extrememplsmib.my.

A new SNMP MIB object is introduced for PW LSP Load Sharing: extremePwObjects OBJECT IDENTIFIER ::= { extremeMplsMIB 5 }

### extremePwPerfTable

This table contains the aggregated transmit and receive packet counters for in-service PWs.

| MIB Object | Description | Variable Type | Support Status |
|---|---|---|---|
| pwPerfInPackets | Receive packet counter per PW | Read | Supported |
| pwPerfInBytes | Receive packet counter in bytes per PW | Read | Supported |
| pwPerfOutPackets | Transmit packet counters per PW | Read | Supported |
| pwPerfOutBytes | Transmit packet counter in bytes per PW | Read | Supported |
| pwPerfInHCPackets | Receive high capacitypacket counter per PW | Read | Supported |
| pwPerfInHCBytes | Receive packet counter in bytes per PW | Read | Supported |

| MIB Object | Description | Variable Type | Support Status |
|---|---|---|---|
| pwPerfOutHCPackets | Transmit packet counters per PW | Read | Supported |
| pwPerfOutHCBytes | Transmit packet counter in bytes per PW | Read | Supported |

## extremePwLspOutboundMappingTable

This table provides the mapping between PWs and LSPs by providing an LSP index. LSP indexes are assigned uniquely for each PW. Entries in this table indicate that an LSP is being used by an in-service PW. An SNMP notification will be sent when an entry is added or deleted.

| MIB Object | Description | Variable Type | Support Status |
|---|---|---|---|
| lspIndex | LSP index per LSP used by the PW. It is only unique per PW peer. | Index | Supported |
| pwLspOutboundLsrXcIndex | Corresponding entry in the *MPLS*-LSR-STD-MIB | Read | Supported |
| pwLspOutboundTunnelIndex | Corresponding RSVP-TE tunnel index based on LSP index used by the PW | Read | Supported |
| pwLspOutboundTunnelInstance | Corresponding RSVP-TE tunnel instance based on LSP index used by the PW | Read | Supported |
| pwLspOutboundTunnelLclLSR | Corresponding RSVP-TE tunnel ingress LSR ID based on LSP index used by the PW | Read | Supported |
| pwLspOutboundTunnelPeerLSR | Corresponding RSVP-TE tunnel egress LSR ID based on LSP index used by the PW | Read | Supported |
| pwLspOutboundTunnelTypeInUse | LSP Type used by the PW. mplsTe value (MIB definition is 1) is being used for RSVP-TE LSPs, and mplsNonTe value (MIB definition is 2) is being used for Static or LDP LSPs. | Read | Supported |

### extremePwLspPerfTable

This table contains the transmit packet and byte counters for traffic sent over a specific PW using a specific LSP.

| MIB Object | Description | Variable Type | Support Status |
|---|---|---|---|
| pwLspPerfOutPackets | Receive packet counter per LSP used by the PW. | Index | Supported |
| pwLspPerfOutBytes | Receive packet counter in bytes per LSP used by the PW. | Read | Supported |
| pwLspPerfOutHCPackets | Receive high capacity packet counter per LSP used by the PW. | Read | Supported |
| pwLspPerfOutHCBytes | Receive high capacity packet counter in bytes per LSP used by the PW. | Read | Supported |

## EXTREME-QOS-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeQosProfileTable | extremeQosProfileIndex | An index that uniquely identifies an entry in the *QoS* table. |
| | extremeQosProfileName | A unique QoS profile name. |
| | extremeQosProfileMinBw | The minimum percentage of bandwidth that this queue requires. The switch is required to provide the minimum amount of bandwidth to the queue. The lowest possible value is 0%. |
| | extremeQosProfileMaxBw | The maximum percentage of bandwidth that this queue is permitted to use. |
| | extremeQosProfilePriority | The level of priority at which this queue will be serviced by the switch. |
| | extremeQosProfileRowStatus | The status of the extremeQosProfile entry. This object can be set to: active (1)createAndGo (4)createAndWait(5)destroy (6) The following values may be read: active (1)notInService(2)notReady (3) |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeIQosProfileTable | extremeIQosProfileIndex | The ExtremeXOS software does not support global QoS profile settings in CLI; it supports per port settings only. Walks of this table display the default values with which the ports are initialized. |
| | extremeIQosProfileName | A unique ingress QoS profile name. |
| | extremeIQosProfileMinBwType | The type of the current minimum bandwidth setting. A value of 1 signifies that the minimum bandwidth value is a percentage of the configurable port bandwidth. A value of 2 or 3 signifies a guaranteed minimum available bandwidth in Kbps or Mbps, respectively. |
| | extremeIQosProfileMinBw | The guaranteed minimum bandwidth for this queue, expressed as either a percentage or a specific bandwidth value, as specified by the value of extremeIQosProfileMinBwType. |
| | extremeIQosProfileMaxBwType | The type of the current maximum bandwidth setting. A value of 1 signifies that the maximum bandwidth value is a percentage of the configurable port bandwidth. A value of 2 or 3 signifies a maximum allowed bandwidth in Kbps or Mbps, respectively. |
| | extremeIQosProfileMaxBw | The maximum allowed input bandwidth for this queue, expressed as either a percentage or a specific bandwidth value, as specified by the value of extremeIQosProfileMaxBwType. |
| | extremeIQosProfileRED | The random early drop threshold. When the input queue fill ratio exceeds this percentage, frames start to drop randomly with a linear increasing drop probability as the queue fill count approaches the maximum queue size. A value of 100 indicates that this feature is currently disabled. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeIQosProfileMaxBuf | The percentage of the total ingress queue size to use. Lower values can be used to reduce the maximum latency through this queue, but with potentially greater loss with bursty traffic. |
| extremePerPortQosTable | extremePerPortQosIndex | The value of this variable is the same as the value of extremeQosProfileIndex of the QoS profile which is overridden (for the port specified by ifIndex) by the definition in this table. |
| | extremePerPortQosMinBw | The minimum percentage of bandwidth that this queue on the specified port requires. The switch is required to provide the minimum amount of bandwidth to the queue. The lowest possible value is 0%. |
| | extremePerPortQosMaxBw | The maximum percentage of bandwidth that this queue on the specified port is permitted to use. |
| | extremePerPortQosPriority | The level of priority at which this queue will be serviced by the switch. |
| | extremePerPortQosRowStatus | The status of the extremePerPortQos entry. This object can be set to active (1) and createAndGo (4). The following value may be read: active (1). Note that a destroy (6) is not supported. A row will only be deleted from this table when the QoS profile indicated in that row is changed globally. |
| extremeQosByVlanMapping Table | extremeVlanIfIndex | Shows mapping of *VLAN* to queues for untagged packets. For tagged packets, the vpri field determines which queue the packet should be using. |
| | extremeQosByVlanMappingQosProfil e Index | Value of extremeQosProfileIndex that uniquely identifies a QoS profile entry in an extremeQosProfileTable. This indicates the QoS to be given to traffic for this VLAN in the absence of any other more specific configuration information for this traffic. |

## EXTREME-RMON-MIB

The following tables, groups, and variables are supported in this MIB

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeRtStatsTable | extremeRtStatsIndex | All objects are supported as read-only. |
| | extremeRtStatsIntervalStart | |
| | extremeRtStatsCRCAlignErrors | |
| | extremeRtStatsUndersizePkts | |
| | extremeRtStatsOversizePkts | |
| | extremeRtStatsFragments | |
| | extremeRtStatsJabbers | |
| | extremeRtStatsCollisions | All objects are supported as read-only. |
| | extremeRtStatsTotalErrors | |
| | extremeRtStatsUtilization | |

## EXTREME-PVLAN-MIB

Two MIB tables allow for GET/SET operations to view and configure ExtremeXOS PVLAN objects. The first is a PVLAN name table with PVLAN name as the index, and the second is the PVLAN subscriber table with PVLAN name, member *VLAN* type, and Subscriber VlanIfIndex as the index. Functionality includes supporting the current set of PVLAN CLI configuration options.

### PVLAN Table (extremePvlanTable)

| extremePvlanName (index) | extremePvlanVrName | extremePvlanNetworkVlanIfIndex |
|---|---|---|
| DisplayString | DisplayString | InterfaceIndexOrZero |
| p1 | VR-Default14 | 14 |
| foo | VR-Cust1 | 0 |

### PVLAN Subscriber Table (extremePvlanSubscriberTable)

| extremePvlanName (index) | extremePvlanSubscriberVlanIfIndex (index) | extremePvlanSubscriberType | extremePvlanSubscriberLoopBackPortIfIndex |
|---|---|---|---|
| DisplayString | InterfaceIndex | Integer | InterfaceIndexOrZero |
| foo | 9 | PVLAN_ISOLATED(2) | 0 |
| foo | 12 | PVLAN_ISOLATED(2) | 1002 |
| p1 | 1 | PVLAN_ISOLATED(2) | 0 |

| extremePvlanName (index) | extremePvlanSubscriberVlanIfIndex (index) | extremePvlanSubscriberType | extremePvlanSubscriberLoopBackPortIfIndex |
|---|---|---|---|
| p1 | 6 | PVLAN_NON_ISOLATED(1 ) | 1004 |
| p1 | 8 | PVLAN_ISOLATED(2) | 1010 |
| p1 | 10 | PVLAN_NON_ISOLATED(1 ) | 0 |

RowStatus is defined as: active(1), createAndGo(4), destroy(6)

## EXTREME-SNMPv3-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeTargetAddrExtTable | extremeTargetAddrExtIgnoreMPModel | When this object is set to TRUE, the version of the trap/notification sent to the corresponding management target (trap receiver) will be the same as in releases of ExtremeWare prior to 7.1.0. Thus, the value of the snmpTargetParamsMPModel object in the snmpTargetParamsTable is ignored while determining the version of the trap/notification to be sent. When a trap-receiver is created through the RMON trapDestTable or from the CLI command 'configure snmp add trapreceiver ....', the value of this object is set to TRUE for the corresponding entry in this table. |
| | extremeTargetAddrExtStandardMode | When this object is set to TRUE, the management target is treated as a 'standard mode' one, in that any Extreme Networks specific extra varbindspresent in a standards-based trap/notification is not sent to this management target. Only the varbinds defined in the standard are sent. |
| | extremeTargetAddrExtTrapDestIndex | This object contains the value of the trapDestIndex in the corresponding entry of the RMON trapDestTable. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeTargetAddrExtUseEventComm | This object is used only when sending RMON alarms as SNMPv3 traps. When it is set to TRUE and an RMON risingAlarm or fallingAlarm is being sent for an event, then the eventCommunity in the RMON event table is compared to the snmpTargetAddrName in the snmpTargetAddrTable. The alarm is sent to the target only when the two are the same. This behavior is exhibited only when the snmpTargetParamsMPModel corresponding to the target indicates an SNMPv3 trap. For SNMPv1/v2c traps, the community in the RMON trapDestTable is used for the comparison, which is the regular method, as per the standards. When this object is set to FALSE, then the RMON alarm (if being sent as an SNMPv3 trap) is sent without using the event community for any comparison. |
| | extremeTargetAddrExtTrapSrcIp | This object contains the source IP address from which traps have to be sent out. If this object is assigned an IP address that does not belong to the switch, an error is thrown. |
| | extremeTargetAddrExtVrName | This object contains the VR name through which the *SNMP* Traps are being sent out. |
| extremeUsm3DESPrivProtocol | | Supported from ExtremeXOS 12.3 |
| extremeUsmAesCfb192Protocol | | Supported from ExtremeXOS 12.3 |
| extremeUsmAesCfb256Protocol | | Supported from ExtremeXOS 12.3 |

## EXTREME-STP-EXTENSIONS-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeStpDomainTable | All objects | This table contains *STP* information per STP domain. |
| | extremeStpDomainStpdDescription | The description associated with this STP domain. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeStpPortTable | All objects | This table contains port-specific information per STP domain. |
| extremeStpVlanPortTable | All objects | This table contains information of the ports belonging to a STP domain on a per *VLAN* basis. |
| extremeStpNotifications | extremeStpEdgePortLoopDetected | This is a trap that would be sent when a loop has been detected on the network login edge safeguard port (the port will be disabled). |

## EXTREME-STPNOTIFICATIONS-MIB

This MIB defines the following Extreme-specific *STP* Notifications trap generated by Extreme Networks devices.

| Trap | Comments |
|---|---|
| extremeStpEdgePortLoopDetected | A loop has been detected on the network login edge safeguard port and the port will be disabled. |

## EXTREME-SYSTEM-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeSaveConfiguration | When this object is set, the device copies the contents of the configuration database to a buffer and saves it to the persistent store specified by the value of the object. The save is performed asynchronously, and the *SNMP* agent continues to respond to both gets and sets while the save is taking place. A network management application may use the extremeSaveStatus object to determine when the asynchronous save operation has completed. |
| | extremeSaveStatus | This object returns the status of a save operation invoked by setting the extremeSaveConfiguration object. A network management application can read this object to determine that a save operation has completed. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeCurrentConfigInUse | Shows which NVRAM configuration store was used at last boot. |
| | extremeConfigToUseOnReboot | Controls which NVRAM configuration store will be used on next reboot. |
| | extremeOverTemperatureAlarm | Alarm status of overtemperature sensor in device enclosure. |
| | extremePrimaryPowerOperational | Not supported. Always returns True. |
| | extremePowerStatus | Not supported. Always returns presentOK. |
| | extremePowerAlarm | Not supported. Always returns False. |
| | extremeRedundantPowerStatus | |
| | extremeRedundantPowerAlarm | Not supported. Always returns presentOK. |
| | extremeInputPowerVoltage | Contains the input voltage of the latest power-supply to power-on in a system with multiple power-supplies. |
| | extremePrimarySoftwareRev | This value indicates the revision number of the software image on the primary partition. |
| | extremeSecondarySoftwareRev | This value indicates the revision number of the software image on the secondary partition. |
| | ExtremeImageToUseOnReboot | This value indicates the partition where the software image is located and to be used on the next boot. |
| | extremeSystemID | This represents the system ID of a Summit switch. |
| | extremeSystemBoardID | Not supported. |
| | extremeSystemLeftBoardID | |
| | extremeSystemRightBoardID | |
| | extremeRmonEnable | |
| | extremeBootROMVersion | Returns information for the current MSM only. |
| | extremeDot1dTpFdbTableEnable | Not supported. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeHealthCheckErrorType | |
| | extremeHealthCheckAction | |
| | extremeHealthCheckMaxRetries | |
| | extremeCpuUtilRisingThreshold | |
| | extremeCpuTaskUtilPair | |
| | extremeCpuAggregateUtilization | |
| | extremeCpuUtilRisingThreshold | |
| | extremeAuthFailSrcAddr | |
| | extremeCpuTransmitPriority | |
| | extremeImageBooted | |
| | extremeMasterMSMSlot | This returns the internal slot number assigned to the MSM. Values supported are: 11,12 on the BlackDiamond 88107,8 on the BlackDiamond 8806 |
| | extremeChassisPortsPerSlot | Returns the maximum ports per slot for the system. |
| | ExtremeMsmFailoverCause | The cause of the last MSM failover: never (1) means an MSM Failover has not occurred since the last reboot.admin (2) means the failover was initiated by the user.exception (3) means the former master MSM encountered a software exception condition.removal (4) means the master MSM was physically removed from the chassis.hwFailure (5) means a diagnostic failure was detected in the master MSM.watchdog (6) means that the master MSM hardware watchdog timer expired.keepalive (7) means the master MSM failed to respond to slave keepalive requests. The MSM failover will have been hitless only in the admin (2) and exception (3) cases. |
| extremeFanStatusTable | extremeFanStatusEntry | Operational status of all internal cooling fans. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeFanNumber | Identifier of cooling fan, numbered from the front and/or left side of device. |
| | extremeFanOperational | Operational status of a cooling fan. |
| | extremeFanEntPhysicalIndex | The entity index for this fan entity in the entityPhysicalTable table of the entity MIB. |
| | extremeFanSpeed | The speed (RPM) of a cooling fan in the fantray. |
| extremeCpuTaskTable | All objects | Not supported. |
| extremeCpuTask2Table | All objects | Not supported. |
| extremeSlotTable | All objects | Cards are currently not configurable via SNMP. |
| extremePowerSupplyEntPhysicalIndex | extremePowerSupplyEntPhysicalIndex | The entity index for this psu entity in the entityPhysicalTable of the entity MIB. |
| extremePowerSupplyNumber | extremePowerSupplyNumber | Power supply number. |
| extremePowerSupplySerialNumber | extremePowerSupplySerialNumber | The serial number of the power supply unit. |
| extremePowerSupplySource | extremePowerSupplySource | The power supply unit input source. |
| extremePowerSupplyTable | extremePowerSupplyStatus | Status of the power supply. |
| | extremePowerSupplyInputVoltage | Input voltage of the power supply. |
| | extremePowerSupplyFan1Speed | The speed of Fan-1 in the power supply unit. |
| | extremePowerSupplyFan2Speed | The speed of Fan-2 in the power supply unit. |
| | extremePowerSupplyInputPowerUsage | Input power usage for the given psu slot. The value 0 in this field indicates the power usage is not supported or that there is a read failure. |
| | extremePowerMonSupplyNumOutput | Number of output sensors in the power supply unit. |
| | extremePowerSupplyInputPowerUsageUnitMultiplier | The magnitude of watts for the usage value in extremePowerSupplyInputPowerUsage. |
| extremePowerSupplyOutputPowerTable | extremePowerSupplyIndex | Power supply unit slot index. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremePowerSupplyOutputSensorIdx | Power supply Sensor index. |
| | extremePowerSupplyOutputVoltage | Output voltage per sensor for the current PSU slot number. A zero (0) in this field indicates that the PSU does not support output voltage reading or else there is an output voltage read error. |
| | extremePowerSupplyOutputCurrent | Output current per sensor for the current PSU slot number. A zero (0) in this field indicates that the PSU does not support output current reading or else there is an output current read error. |
| | extremePowerSupplyOutputUnitMultiplier | The magnitude of volts and amps for the usage value in extremePowerSupplyOutputVoltage and extremePowerSupplyOutputCurrent. |
| extremePowerSupplyUsageTable | extremeSlotIndex | Slot number in the chassis/ stack based system. |
| | extremePowerSupplyUsageValue | Power usage of the particular slot in the chassis or stack. The power usage is measured in milliwatts. |
| | extremePowerSupplyUnitMultiplier | The magnitude of watts for the usage value in extremePowerSupplyUsageValue. |
| extremeSystemPowerUsage | extremeSystemPowerUsageValue | The current power usage of the system. In stack mode. this variable tells the total power usage of the entire system. |
| | extremeSystemPowerUsageUnitMultiplier | The magnitude of watts for the usage value in extremeSystemPowerUsageValue. |
| extremeSystemPowerMonitorTable | extremeSystemPowerMonitorIndex1 | Reserved. Can be used for future expansion. Currently set to zero. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeSystemPowerMonitorPollInterval | Configure how often input power is measured. It is configured in seconds with a default value of 60 seconds. If zero (0) is configured, then the input power measurement is disabled. |
| | extremeSystemPowerMonitorReportChanges | Configure report-changes. Has none or log or trap or log-and-trap, with a default of none. |
| | extremeSystemPowerMonitorChangeThreshold | Configure the input power change threshold to initiate report-changes action. By default 2 watts is configured. This field is configured in watts. |
| extremeSystemPowerUsageNotification | sysUpTime, sysDescr, extremeSystemPowerUsageValue, extremeSystemPowerUsageUnitMultiplier | Whenever the power usage is increased/decreased by the configured threshold value, then the power usage trap is generated if the trap is enabled. |
| extremeImageTable | extremeImageNumber | This table contains image information for all images installed on the device. extremeImageNumber values are not compatible with EW releases. Current image has value 3 instead of 0. |
| | extremeMajorVersion | This is the first number within the software image version number which consists of 4 numbers separated by a period. |
| | extremeMinorVersion | This is the second number within the software image version number. |
| | extremeBuildNumber | This is the fourth number within the software image version number. |
| | extremeTechnologyReleaseNumber | The technology release version. This value is zero for all but TR releases. |
| | extremeSustainingReleaseNumber | The sustaining release number for the ExtremeXOS version. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeBranchRevisionNumber | This is the branch from where the software image was built. |
| | extremeImageType | This is the software image type (for example, ExtremeXOS core, ExtremeXOS module, ExtremeXOS firmware). |
| | extremeImageDescription | Description of image contains image version, including major version, submajor version, minor version, build version, build branch, build-master login, and build date. |
| | extremePatchVersion | The ExtremeXOS release patch version. This is the third number within the software image version number. |
| extremeImageFeatureTable | extremeImageFeatureEntry | A table containing information about the software features. |
| | extremeImageFeatureNumber | A unique integer identifying the particular software image. This indicates the partition on which the image is loaded: 0—Current partition1—Primary partition2—Secondary partition |
| | extremeImageSshCapability | Indicates whether image has SSH capability: 1=nossh—shown by default when the SSH license is not enabled.2=ssh—shown when the SSH license is enabled. |
| | extremeImageUAACapability | Not supported. |
| extremeCpuMonitorInterval | | This value determines how frequent CPU usage will be monitored. |
| extremeCpuMonitorTotalUtilization | | This value indicates the total CPU utilization. |
| extremeCpuMonitorTable | | This value indicates the total CPU utilization. |
| | extremeCpuMonitorSlotId | This value indicates the slot ID where the CPU is being monitored. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeCpuMonitorProcessName | This value indicates the process name for the process being monitored. |
| | extremeCpuMonitorProcessId | This value indicates the process ID. |
| | extremeCpuMonitorProcessState | This value indicates the current state of the process. |
| | extremeCpuMonitorUtilization5secs | This value indicates the CPU utilization in the past 5 seconds. |
| | extremeCpuMonitorUtilization10secs | This value indicates the CPU utilization in the past 10 seconds. |
| | extremeCpuMonitorUtilization30secs | This value indicates the CPU utilization in the past 30 seconds. |
| | extremeCpuMonitorUtilization1min | This value indicates the CPU utilization in the past 1minute. |
| | extremeCpuMonitorUtilization5mins | This value indicates the CPU utilization in the past 5 minutes. |
| | extremeCpuMonitorUtilization30mins | This value indicates the CPU utilization in the past 30 minutes. |
| | extremeCpuMonitorUtilization1hour | This value indicates the CPU utilization in the past 1 hour. |
| | extremeCpuMonitorUserTime | This value indicates the CPU usage under User Mode. |
| | extremeCpuMonitorSystemTime | This value indicates the CPU usage under system mode. |
| extremeCpuMonitorSystemTable | | This table contains CPU monitoring information for all system processes. |
| | extremeCpuMonitorSystemSlotId | This value indicates the slot ID where the CPU is being monitored. |
| | extremeCpuMonitorSystemUtilization5secs | This value indicates the CPU utilization in the past 5 seconds. |
| | extremeCpuMonitorSystemUtilization10secs | This value indicates the CPU utilization in the past 10 seconds. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeCpuMonitorSystemUtilization30secs | This value indicates the CPU utilization in the past 30 seconds. |
| | extremeCpuMonitorSystemUtilization1min | This value indicates the CPU utilization in the past 1 minute. |
| | extremeCpuMonitorSystemUtilization5mins | This value indicates the CPU utilization in the past 5 minutes. |
| | extremeCpuMonitorSystemUtilization30mins | This value indicates the CPU utilization in the past 30 minutes. |
| | extremeCpuMonitorSystemUtilization1hour | This value indicates the CPU utilization in the past 1 hour. |
| | extremeCpuMonitorSystemMaxUtilization | This value indicates the maximum CPU utilization so far. |
| extremeMemoryMonitorSystemTable | | This table contains system-level memory monitor information. |
| | extremeMemoryMonitorSystemSlotId | This value indicates the slot ID where the memory is being monitored. |
| | extremeMemoryMonitorSystemTotal | This value indicates the total memory installed on the switch. |
| | extremeMemoryMonitorSystemFree | This value indicates the amount of memory that is free. |
| | extremeMemoryMonitorSystemUsage | This value indicates the amount of memory consumed for kernel code. |
| | extremeMemoryMonitorUserUsage | This value indicates the amount of memory consumed by user processes as well as the kernel. |
| extremeMemoryMonitorTable | | This table contains memory monitor information for each user process. |
| | extremeMemoryMonitorSlotId | This value indicates the slot ID where the memory is being monitored. |
| | extremeMemoryMonitorProcessName | This value indicates the name of the process being monitored. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeMemoryMonitorUsage | This value indicates the amount of memory being consumed by this user process. |
| | extremeMemoryMonitorLimit | Not supported. |
| | extremeMemoryMonitorZone | |
| | extremeMemoryMonitorGreenZoneCount | |
| | extremeMemoryMonitorYellowZoneCount | |
| | extremeMemoryMonitorOrangeZoneCount | |
| | extremeMemoryMonitorRedZoneCount | |
| | extremeMemoryMonitorGreenZoneThreshold | |
| | extremeMemoryMonitorYellowZoneThreshold | |
| | extremeMemoryMonitorOrangeZoneThreshold | |
| | extremeMemoryMonitorRedZoneThreshold | |

## EXTREME-TRAP-MIB

This MIB defines the following Extreme-specific SNMPv1 traps generated by Extreme Networks devices.

| Trap | Comments |
|---|---|
| extremeOverheat | The on-board temperature sensor has reported an overheat condition. The system shuts down until the unit has sufficiently cooled such that operation can begin again. A cold start trap is issued when the unit has come back on line. |
| extremeFanfailed | One or more of the cooling fans inside the device has failed. A fanOK trap will be sent once the fan has attained normal operation. |
| extremeFanOK | A fan has transitioned out of a failure state and is now operating correctly. |
| extremeInvalidLoginAttempt | A user attempted to log into the console or by telnet but was refused access due to an incorrect username or password. |
| extremePowerSupplyGood | One or more previously bad sources of power to this agent has come back to life without causing an agent restart. |
| extremePowerSupplyFail | One or more sources of power to this agent has failed. Presumably a redundant power-supply has taken over. |
| extremeEdpNeighborAdded | This node discovers a new neighbor through *EDP*. |

| Trap | Comments |
|---|---|
| extremeEdpNeighborRemoved | No EDP updates are received from this neighbor within the configured timeout period and this neighbor entry is aged out by the device. |
| extremeModuleStateChanged | Signifies that the value of the extremeSlotModuleState for the specified extremeSlotNumber has changed. Traps are reported only for significant states. |

## EXTREME-TRAPPOLL-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeSmartTrapFlushInstanceTableIndex | This object acts as a flush control for the extremeSmartTrapInstanceTable. Setting this object can flush the matching entries from the extremeSmartTrapInstanceTable based on certain rules as defined in the MIB. |
| extremeSmartTrapRulesTable | | The entries created in the extremeSmartTrapRulesTable define the rules that are used to generate Extreme smart traps.<br>The object extremeSmartTrapRulesDesiredOID supports OID values whose prefix are among the following:<br>ipAddrTable, ifMauTable, extremeSlotTable, extremeVlanGroup, extremeVirtualGroup, extremeVlanProtocolTable, extremeVlanProtocolVlanTable, extremeVlanOpaqueTable, extremeStpDomainTable, extremeStpPortTable, extremeStpVlanPortTable, pethPsePortTable, extremePethSystem, extremePethPseSlotTable, extremePethPsePortTable |
| extremeSmartTrapInstanceTable | | extremeSmartTrapInstanceTable is a read-only table that stores the information about which variables have changed according to rules defined in the extremeSmartTrapRulesTable. |

## EXTREME-V2TRAP-MIB

This MIB defines the following Extreme-specific SNMPv2c traps generated by Extreme Networks devices.

| Trap | Comments |
|---|---|
| extremeHealthCheckFailed | CPU HealthCheck has failed. |
| extremeMsmFailoverTrap | MSM failover occurred. |
| extremeBgpM2PrefixReachedThreshold | This notification is generated when the number of prefixes received over this peer session reaches the threshold limit. |
| extremeBgpM2PrefixMaxExceeded | This notification is generated when the number of prefixes received over this peer session reaches the maximum configured limit. |
| extremeEapsStateChange | Send on master/transit nodes. |

| Trap | Comments |
|---|---|
| extremeEapsFailTimerExpFlagSet | This notification is generated when the EAPS domain fail timer expires for the first time, while its state is not in Fail state. |
| extremeEapsFailTimerExpFlagClear | This notification is generated when the EAPS domain fail timer expired flag is cleared. |
| extremeEapsLinkDownRingComplete | If a transit is in link-down state, and it receives a Health-Check-Pdu from the master indicating the ring is complete, there is some problem with the transit switch that has issued this trap message. |
| extremeEapsLastStatusChangeTime | Sent on master/transit nodes. Provides a general indication of a status change using a 10 second timer. |
| extremeEapsPortStatusChange | Sent on master/transit nodes. |
| extremeEapsConfigChange | Sent on master/transit nodes. This trap has a granularity of 30 seconds. |
| extremeEapsSharedPortStateChange | Sent on controller/partner nodes. |
| extremeEapsRootBlockerStatusChange | |
| extremeNMSInventoryChanged | These traps are not generated by the ExtremeXOS *SNMP* agent but by the Ridgeline NMS. |
| extremeNMSTopologyChanged | |
| extremeOverheatNormal | An overheat (return to) normal notification indicates that the on-board tempature sensor has reported a temperature that has returned to within the normal operating range from having been in an overheat (or overcold) condition. The temperature of the unit has sufficiently cooled to be below the maximum (or warmed to be above the minimum) of the normal operating range. |

## EXTREME-VLAN-MIB

The following tables, groups, and variables are supported in this MIB.

Note

*SNMP* Set operations are not allowed for "Mgmt" and "Default" *VLAN*s.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeVirtualGroup | extremeNextAvailableVirtIfIndex | |
| extremeVlanIfTable | extremeVlanIfIndex | While creating a new row in the extremeVlanIfTable, the value of the object extremeVlanIfDescr must be specified.<br>For all tables in this MIB that contain objects with RowStatus semantics, the only values supported are: {active, createAndGo, destroy}. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVlanIfDescr | This is a description of the VLAN interface. |
| | extremeVlanIfType | The VLAN interface type. |
| | extremeVlanIfGlobalIdentifier | Not supported. |
| | extremeVlanIfStatus | The status column for this VLAN interface.<br>This object can be set to: active (1); createAndGo (4); createAndWait (5); destroy (6). The following values may be read: active (1); notInService (2); notReady (3). |
| | extremeVlanIfIgnoreStpFlag | Not supported. |
| | extremeVlanIfIgnoreBpduFlag | Not supported. |
| | extremeVlanIfLoopbackModeFlag | Setting this object to true causes loopback mode to be enabled on this VLAN. |
| | extremeVlanIfVlanId | The VLAN ID of this VLAN. |
| extremeGlobalMappingTable | Not supported | |
| extremeVlanEncapsTable | Not supported | |
| extremeVlanIpTable | All objects | For all tables in this MIB that contain objects with RowStatus semantics, the only values supported are: {active, createAndGo, and destroy}. |
| extremeVlanProtocolTable | Not supported | |
| extremeVlanProtocolBindingTable | Partial | For all tables in this MIB that contain objects with RowStatus semantics, the only values supported are: {active, createAndGo, and destroy}.<br>New to the ExtremeXOS software: association of a protocol filter and VLAN. |
| extremeVlanProtocolVlanTable | Not supported | |
| extremeVlanProtocolDefTable | Partial | For all tables in this MIB that contain objects with RowStatus semantics, the only values supported are:<br>{ active, createAndGo, and destroy}<br>New to the ExtremeXOS software:<br>This is a new table introduced to add a protocol filter with etype values during creation of the filter itself. |
| extremeVlanOpaqueTable | All objects | This is a read only table. For adding ports to a VLAN or deleting ports from a VLAN, use extremeVlanOpaqueControlTable. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeVlanOpaqueControlTable | All objects | For all tables in this MIB that contain objects with RowStatus semantics, the only values supported are: {active, createAndGo, and destroy}.<br>New to the ExtremeXOS software: extremeVlanOpaqueControlTable is a write only table and cannot be used to read. This is used to add/delete ports on a VLAN. |
| extremeVlanStackTable | All objects | Not supported. |
| extremeVlanL2StatsTable | All objects | Not supported.<br>This table contains per VLAN information about the number of packets sent to the CPU, the number of packets learned, the number of *IGMP (Internet Group Management Protocol)* control packets snooped and the number of IGMP data packets switched. This is the same information that is available using the CLI command show l2stats. |
| extremePortVlanStatsTable | | VLAN statistics per port. |
| extremePortVlanStatsEntry | extremeStatsPortIfIndex | The index of this table. |
| | extremeStatsVlanNameIndex | The index of this table. |
| | extremePortVlanStatsCntrType | Read-only. |
| | extremePortVlanTotalReceivedBytesCounter | The total number of bytes received by a port for a particular VLAN. |
| | extremePortVlanTotalReceivedFramesCounter | The total number of frames received by a port for a particular VLAN. |
| | extremePortVlanTotalTransmittedBytesCounter | The total number of bytes transmitted by a port for a particular VLAN. |
| | extremePortVlanTotalTransmittedFramesCounter | The total number of frames transmitted by a port for a particular VLAN. |
| | extremePortConfigureVlanStatus | The row status variable, used according to row installation and removal conventions. |

## EXTREME-VM-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeVMFTPServerTable | extremeVMFTPServerEntry | |
| | extremeVMFTPServerType | The type of the FTP server. The backup server is contacted if the primary fails to respond. |
| | extremeVMFTPAddrType | The IP type of IP address. |
| | extremeVMFTPServer | The IP address of the FTP server used for transferring various management files. |
| | extremeVMFTPSynchInterval | The time in minutes between automatic synchronization attempts. A value of 0 indicates that automatic synchronizations are not performed. Note that each switch does not perform a synchronization at exactly the time configured, but varies the synchronization interval between 3/4 and 5/4 of the configured interval. This avoids the situation where all switches in a network attempt a synchronization at exactly the same moment. Automatic synchronization is disabled by default, and requests to enable them are rejected until the FTP server information (IPv4 or IPv6 address, username, and password) is configured. |
| | extremeVMFTPRowStatus | There can only be two entries in this table, one each for primary and secondary FTP servers. |
| | extremeVMFTPPathName | The FTP server directory name for the policies to be synchronized. A value of '/pub' is used by default. |
| | extremeVMFTPUsername | A valid username on the FTP server. |
| | extremeVMFTPPassword | The password associated with the FTP user. This object returns a zero length string when queried. |
| extremeVMGeneral | extremeVMFTPPolicyDir | The server directory name for the policies to be synchronized. A value of '/' is used by default. |
| | extremeVMLastSynch | The timestamp of the most recent synchronization attempt. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVMLastSynchStatus | The result of the most recent synchronization attempt:success (1)—indicates that the synchronization completed successfullyaccessDenied (2)—The username and password were not accepted by the serverserverTimeout (3)—Could not establish a file transfer session with the configured serverserverNotConfigured (4)—The server configuration is not complete. |
| | extremeVMSynchAdminState | Triggers a synchronization cycle on demand. A synchronization automatically downloads new or updated policies as well as delete policies to match those on the server: idle (1) is returned whenever this object is read.synchronizeNow (2) triggers an immediate synchronization, and is reflected in extremeVMSynchOperState. Attempts to set this variable to synchronizeNow (2) are rejected if a synchronization is currently in progress. |
| | extremeVMSynchOperState | Indicates if a synchronization is in progress, either on-demand or automatic. |
| | extremeVMTrackingEnabled | The virtual machine (VM) tracking feature is disabled by default and can be enabled using this object. |
| | extremeVMPortConfigTable | Configures the VM features on each port. |
| extremeVMPortConfigEntry | | An entry in the table for VM features on each port. |
| | extremeVMPortConfigIfIndex | The value of ifIndex of a physical port capable of supporting the VM tracking features. |
| | extremeVMPortConfigVMTrackingEnabled | Enables the VM tracking feature on a port. The VM tracking feature is disabled by default. |
| extremeVMVPPTable | | This table contains port policies contained within this switch. Port policies come in two variants: network and local. Network policies are downloaded from the FTP server; local policies reside only within a single switch. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| extremeVMVPPEntry | | An entry in the table of VM policy information of this device. This table is populated with two sets of policies, those downloaded from the policy server and those defined locally on this switch. |
| | extremeVMVPPType | The type of the port policy. Network port policies are obtained from a central policy store. Local policies are specific to this particular switch. |
| | extremeVMVPPName | The name of the port VPP. VPP names must be alpha-numeric and must start with an alpha character. |
| | extremeVMVPPControl | Performs the requested operation on this policy. synchronizeNow (1) downloads a copy of the policy from the FTP server. (Network policies only). This object returns noOperation (2) if read. |
| | extremeVMVPPRowStatus | Only local VPPs can be created or deleted. |
| extremeVMMappingTable | | This table contains the mapping of port policies to VM MAC addresses. |
| | extremeVMMappingEntry | An entry in the table of VM information of this device. |
| | extremeVMMappingType | The type of mapping for this entry. A local mapping exists only on this specific switch. A network mapping is one obtained through a download of a mapping file. |
| | extremeVMMappingMAC | The MAC address associated with the VM. Note that a VM can have multiple MAC addresses. |
| | extremeVMMappingIngressVPPName | The ingress policy associated with the VM/MAC address. Note that this may refer to a policy without a corresponding entry in the extremeVMVPPTable if a network policy mapping refers to a non-existent policy. This indicates an error in the policy mapping file that is consulted if network authentication fails.<br>When creating an entry in this table, this name must refer to an existing, valid, local policy. The creation of a mapping to a network policy is not permitted. Those mappings must be created at the central policy server. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVMMappingEgressVPPName | The egress policy associated with the VM/MAC address. Note that this might refer to a policy without a corresponding entry in the extremeVMVPPTable if a network policy mapping refers to a non-existent policy. This indicates an error in the policy mapping file that is consulted if network authentication fails.<br>When creating an entry in this table, this name must refer to an existing, valid, local policy. The creation of a mapping to a network policy is not permitted. Those mappings must be created at the central policy server. |
| | extremeVMMappingStatus | Indicates the virtual port profile mapping status:<br>vppValid (1)—A VPP mapped to this VM MAC address does not have any policies associated with it (or) all the policies associated with this VPP can be applied (policy validation is success). Policy validation will happen only when this VM MAC is detected.vppMissing (2)—This value is applicable only for network VM if the specified VPP name was missing.vppInvalid (3)—One of the polices mapped to VPP cannot be applied (policy validation failed) because the policy file contains one or more errors that prevent it from being appliedvppNotMaped (4)—The VM does not have any VPP mapped. |
| | extremeVMMappingRowStatus | Only local polices can be created or deleted. |
| | | The virtual port profile associated with the VM MAC address. When creating an entry in this table, this name must refer to an existing, valid, local profile. The creation of a mapping to a network profile is not permitted. Those mappings must be created at the central policy server. |
| extremeVMVPP2PolicyTable | | This table contains the mapping of a VPP to individual policies. |
| extremeVMVPP2PolicyEntry | | An individual mapping of VPP to policy. |
| | extremeVMVPP2PolicyVPPName | The name of the VPP. |
| | extremeVMVPP2PolicyPolicyName | The name of the local policy. |

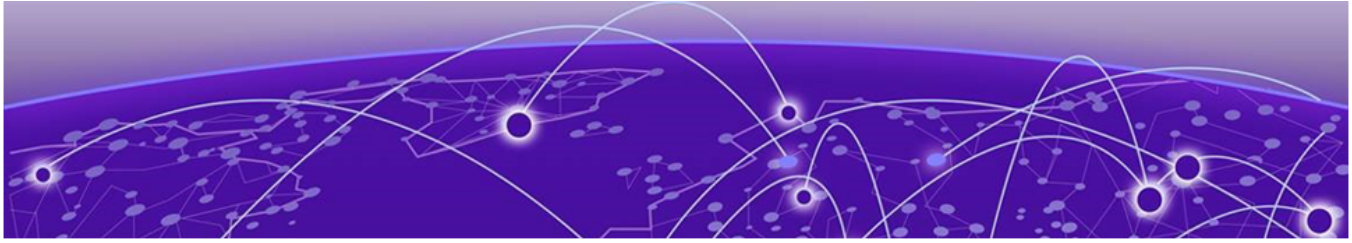
| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVMVPP2PolicyType | The type of policy |
| | extremeVMVPP2PolicyOrder | The order in which this policy is executed. |
| | extremeVMVPP2PolicyRowStatus | The row status for this mapping. |
| extremeVMDetected | extremeVMDetectedNumber | The number of VMs detected on this switch. |
| extremeVMDetectedTable | | This table contains the currently detected VMs on this switch. |
| extremeVMDetectedEntry | | An entry in the table of VM information of this device. |
| | extremeVMDetectedMAC | The MAC address associated with the VM. Note that a VM can have multiple MAC addresses. |
| | extremeVMDetectedVMName | The name of the VM. Note that a VM authenticated locally might not have a name. |
| | extremeVMDetectedIngressVPPName | The name of the policy applied (or attempted to apply) to this VM. |
| | extremeVMDetectedEgressVPPName | The name of the policy applied (or attempted to apply) to this VM. |
| | extremeVMDetectedIfIndex | The value of ifIndex on which this VM was detected. |
| | extremeVMDetectedAdminStatus | The administrative status of the VM authentication. Setting this variable to authenticating (1) will force the re-authentication of the VM. This variable always returns idle (2) when read. |
| | extremeVMDetectedOperStatus | The authentication status of the VM: authenticating (1)—an authentication is currently in progressauthenticatedNetwork (2)—the VM has been authenticated by a network sourceauthenticatedLocally (3)—the VM has been authenticated by the local databasedenied (4)—the VM was explicitly denied entrynotAuthenticated (5)—the authentication process timed out or was never attempted |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVMDetectedResultIngress | Indicates the result of a VM entry into the network. Only the two values below will be returned at any point of time:<br>policyApplied (1)—All the ingress policies in the VPP were successfully applied to the port.policyNotApplied (2) - One of the ingress policies in the VPP was not applied to the port. If this value is returned then refer to the extremeVMDetectedIngErrPolicies object for list of failed ingress policies. |
| | extremeVMDetectedResultEgress | Indicates the result of a VM entry into the network. Only the two values below will be returned at any point of time:<br>policyApplied (1) - All the egress policies in the VPP were successfully applied to the port.policyNotApplied (2) - One of the egress policies in the VPP was not applied to the port. If this value is returned then refer to the extremeVMDetectedEgrErrPolicies object for list of failed egress policies. |
| | extremeVMDetectedIngErrPolicies | Displays the list of failed ingress policies. |
| | extremeVMDetectedEgrErrPolicies | Displays the list of failed egress policies. |
| | extremeVMDetectedVPPName | The name of the VPP applied (or attempted to apply) to this VM. |
| | extremeVMDetectedVPPResult | Indicates the result of a VPP associated with a VM MAC:<br>vppMapped (1)—indicates that the named VPP was mapped.vppNotMapped (2)—indicates that the no VPP was mapped.vppInvalid (3)—indicates that the VPP mapped was invalid.vppMissing (4)—indicates that the VPP mapped was missing |
| extremeVMVPPDetailTable | | This table contains the mapping of a VPP to individual policies. |
| extremeVMVPPDetailEntry | | A set of mappings from a VPP to one or more policies. |
| | extremeVMVPPDetailVPPName | The name of the VPP. |
| | extremeVMVPPDetailDirection | The direction in which the policy is applied. |
| | extremeVMVPPDetailType | The type of policy. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | extremeVMVPPDetailOrder | The order in which this policy is executed. |
| | extremeVMVPPDetailPolicyName | The name of the local policy |
| | extremeVMVPPDetailRowStatus | The row status for this mapping. |
| extremeVMVPPSynchType | | Indicates the type of policy |

The following traps can be generated.

| Trap | Notification Objects | Comments |
|---|---|---|
| extremeVMNotificationPrefix | extremeVMVPPSyncFailed | A synchronization attempt failed. |
| | extremeVMVPPInvalid | A VPP definition is invalid, indicating it cannot be applied to a port. |
| | extremeVMMapped | This notification is generated whenever a MAC address is manually mapped to a local policy. |
| | extremeVMUnMapped | This notification is generated whenever a MAC address is manually unmapped to a local policy. |
| | extremeVMDetectResult | This notification is generated after a VM is detected on a port and reflects the result of that operation. |
| | extremeVMUnDetectResult | This notification is generated after a VM is undetected (removed) from a port. |



# MIB Support Details



The following sections describe the MIB support provided by the ExtremeXOS *SNMP (Simple Network Management Protocol)* agent residing on Extreme Networks devices running ExtremeXOS:

Where applicable, these sections note how the implementation differs from the standards, or from the private MIBs.

Note

Only entries for the default VR are supported.

ExtremeXOS software supports only the following “Row status” in the MIBs:

- Create and Go
- Active
- Destroy

The standard MIBs are described in the following sections.

## IEEE 802.1AB (LLDP-MIB)

All tables and variables of this MIB are supported.

## IEEE 802.1AB (LLDP-EXT-DOT1-MIB)

All tables and variables of this MIB are supported.

## IEEE 802.1AB (LLDP-EXT-DOT3-MIB)

All tables and variables of this MIB are supported.

## IEEE 802.1AG (CFM MIB)

This MIB contains objects for the 802.1ag protocol.

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1agCfmMdTable | dot1agCfmMdIndex | The index to the Maintenance Domain (MD) table. |
| | dot1agCfmMdFormat | Supported as read only. The type (and thereby format) of the MD name. |
| | dot1agCfmMdName | Supported as read only. The MD name. |
| | dot1agCfmMdMdLevel | Supported as read only. The MD Level. |
| | dot1agCfmMdMhfCreation | Enumerated value indicating whether the management entity can create MHFs (MIP Half Function) for this MD. Supported as read only. Configuration in CLI is not supported. The defMHFdefault (2) value is returned. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMdMhfIdPermission | Supported as read only. Configuration in CLI is not supported. The sendIdChassisManage (4) value is returned. |
| | dot1agCfmMdMaNextIndex | Value to be used as the index of the Maintenance Association (MA) table entries. |
| | dot1agCfmMdRowStatus | The status of the row. The writable columns in a row cannot be changed if the row is active. All columns must have a valid value before a row can be activated. |
| dot1agCfmMdTableNextIndex | dot1agCfmMdTableNextIndex | An unused value for dot1agCfmMdIndex that is used as the MD index for the next newly created domain. |
| dot1agCfmMaNetTable | dot1agCfmMdIndex | MD index. |
| | dot1agCfmMaIndex | Index of the MA table. dot1agCfmMdMaNextIndex needs to be inspected to find an available index for row-creation. |
| | dot1agCfmMaNetFormat | Supported as read only. The type of the MA name. |
| | dot1agCfmMaNetName | Supported as read only. MA name. |
| | dot1agCfmMaNetCcmInterval | Supported as read only. Transmission interval between CCMs to be used by all MEPs in the MA. |
| | dot1agCfmMaNetRowStatus | The status of the row. The writable columns in a row cannot be changed if the row is active. All columns must have a valid value before a row can be activated. |
| dot1agCfmMepTable | dot1agCfmMdIndex | MD Index. |
| | dot1agCfmMaIndex | MA Index. |
| | dot1agCfmMepIdentifier | A small integer, unique over a given MA identifying a specific MA end point. |
| | dot1agCfmMepIfIndex | Supported as read only. Interface index of the interface bridge port. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMepDirection | Supported as read only. The direction in which the MEP faces on the bridge port. |
| | dot1agCfmMepPrimaryVid | Supported as read only. VID of the MEP. |
| | dot1agCfmMepActive | Supported as read only. Administrative state of the MEP. |
| | dot1agCfmMepFngState | Current state of the MEP Fault Notification Generator State Machine. Can have any one of the following values: fngReset (1)fngDefect (2)fngReportDefect (3)fngDefectReported (4)fngDefectClearing (5) |
| | dot1agCfmMepCciEnabled | Supported as read only. If set to true, the MEP will generate CCM. |
| | dot1agCfmMepCcmLtmPriority | Supported as read only. The priority value for CCMs and LTMs transmitted by the MEP. Configuration in CLI is not supported. Default value 0 is returned. |
| | dot1agCfmMepMacAddress | MAC address of the MEP. In Extreme Networks switches, the MAC is returned. |
| | dot1agCfmMepLowPrDef | Supported as read only. An integer value specifying the lowest priority defect that is allowed to generate a Fault Alarm. Configuration in CLI is not supported. Default *VLAN (Virtual LAN)* macRemErrXon is returned. |
| | dot1agCfmMepFngAlarmTime | Supported as read only. The time that defects must be present before a Fault Alarm is issued. Configuration in CLI is not supported. Default value 2.5 seconds is returned. |
| | dot1agCfmMepFngResetTime | Supported as read only. The time that defects must be absent before resetting a Fault Alarm. Configuration in CLI is not supported. Default value 10 seconds is returned. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMepHighestPrDefect | The highest priority defect that has been present. Possible values are: none (0)defRDICCM (1)defMACstatus (2)defRemoteCCM (3)defErrorCCM (4)defXconCCM (5) |
| | dot1agCfmMepDefects | Error condition to be sent. The conditions can be any one of the following: bDefRDICCM (0)bDefMACstatus (1)bDefRemoteCCM (2)bDefErrorCCM (3)bDefXconCCM (4) |
| | dot1agCfmMepErrorCcmLastFailure | The last-received CCM that triggered a DefErrorCCM fault. |
| | dot1agCfmMepXconCcmLastFailure | The last-received CCM that triggered a DefXconCCM fault. |
| | dot1agCfmMepCcmSequenceErrors | The total number of out-of-sequence CCMs received from all remote MEPs. |
| | dot1agCfmMepCciSentCcms | The total number of Continuity Check messages transmitted. |
| | dot1agCfmMepNextLbmTransId | Next sequence number/ transaction identifier to be sent in a Loopback message. |
| | dot1agCfmMepLbrIn | The total number of valid, in-order Loopback Replies received. |
| | dot1agCfmMepLbrInOutOfOrder | The total number of valid, out-of-order Loopback Replies received. |
| | dot1agCfmMepLbrBadMsdu | The total number of LBRs received whose mac_service_data_unit did not match (except for the OpCode) that of the corresponding LBM. |
| | dot1agCfmMepLtmNextSeqNumber | Next transaction identifier/ sequence number to be sent in a Linktrace message. |
| | dot1agCfmMepUnexpLtrIn | The total number of unexpected LTRs received. |
| | dot1agCfmMepLbrOut | The total number of Loopback Replies transmitted. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMepTransmitLbmStatus | Supported as read only. A Boolean flag set to true by the bridge port to indicate that another LBM may be transmitted. |
| | dot1agCfmMepTransmitLbmDestMac Address | Supported as read only. The target MAC address field to be transmitted. |
| | dot1agCfmMepTransmitLbmDestMepId | Supported as read only. To transmit the LBM, destMEPID need not be given. |
| | dot1agCfmMepTransmitLbmDestIsMepId | Supported as read only. This always returns FALSE. |
| | dot1agCfmMepTransmitLbmMessages | Supported as read only. The number of Loopback messages to be transmitted. |
| | dot1agCfmMepTransmitLbmDataTlv | Supported as read only. An arbitrary amount of data to be included in the Data TLV if the Data TLV is selected to be sent. This returns 0. |
| | dot1agCfmMepTransmitLbmVlanPriority | Priority. A three-bit value to be used in the VLAN tag, if present in the transmitted frame. Configuration in CLI is not supported. Default value 0 is returned. |
| | dot1agCfmMepTransmitLbmVlanDropEn able | Not supported. Returns FALSE. |
| | dot1agCfmMepTransmitLbmResultOK | Indicates the result of the operation linked with MEP active state. |
| | dot1agCfmMepTransmitLbmSeqNumber | The Loopback Transaction Identifier of the first LBM (to be) sent. The value returned is undefined if dot1agCfmMepTransmitLbmR esultOK is false. |
| | dot1agCfmMepTransmitLtmStatus | Supported as read only. A Boolean flag set to true by the bridge port to indicate that another LTM may be transmitted. |
| | dot1agCfmMepTransmitLtmFlags | Supported as read only. The flags field for LTMs transmitted by the MEP. Currently, useFDBonly (0) is supported. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMepTransmitLtmTargetMacAddress | Supported as read only. The target MAC address field to be transmitted. |
| | dot1agCfmMepTransmitLtmTargetMepId | Not supported. To transmit the LTM destMEPID need not be given. Value 0 is returned. |
| | dot1agCfmMepTransmitLtmTargetIsMepId | Not supported. This always returns FALSE. |
| | dot1agCfmMepTransmitLtmTtl | Supported as read only. The LTM TTL field. |
| | dot1agCfmMepTransmitLtmResult | Supported as read only. Indicates the result of the operation. Linked with MEP active state. |
| | dot1agCfmMepTransmitLtmSeqNumber | The LTM transaction identifier. |
| | dot1agCfmMepTransmitLtmEgressIdentifier | Supported as read only. Identifies the MEP Linktrace Initiator that is originating or the Linktrace Responder that is forwarding this LTM. |
| | dot1agCfmMepRowStatus | The status of the row. The writable columns in a row cannot be changed if the row is active. All columns must have a valid value before a row can be activated. |
| dot1agCfmMepDbTable | dot1agCfmMdIndex | MD Index. |
| | dot1agCfmMaIndex | MA Index. |
| | dot1agCfmMepIdentifier | MA end point identifier. |
| | dot1agCfmMepDbRMepIdentifier | Maintenance association end point identifier of a remote MEP whose information from the MEP database is to be returned. |
| | dot1agCfmMepDbRMepState | The operational state of the remote MEP IFF state machines. The state can be any one of the following: rMepIdle (1)rMepStart (2)rMepFailed (3)rMepOk (4) |
| | dot1agCfmMepDbRMepFailedOkTime | The time (SysUpTime) at which the IFF remote MEP state machine last entered either the RMEP_FAILED or RMEP_OK state. |
| | dot1agCfmMepDbMacAddress | The MAC address of the remote MEP. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmMepDbRdi | State of the RDI bit in the last received CCM. |
| | dot1agCfmMepDbPortStatusTlv | An enumerated value of the Port status TLV received in the last CCM from the remote MEP or the default value. The value is one of the following: psNoPortStateTLV (0)psBlocked (1)psUp (2) |
| | dot1agCfmMepDbInterfaceStatusTlv | An enumerated value of the Interface status TLV received in the last CCM from the remote MEP. The value can be one of the following: isNoInterfaceStatusTLV (0)isUp (1)isDown (2)isTesting (3)isUnknown (4)isDormant (5)isNotPresent (6)isLowerLayerDown (7) |
| | dot1agCfmMepDbChassisIdSubtype | networkAddress (5) is returned if senderIDTLV is received. |
| | dot1agCfmMepDbChassisId | The first octet contains the IANA Address Family Numbers enumeration value for the specific address type, and octets 2 through N contain the network address value in network byte order. |
| | dot1agCfmMepDbManAddressDomain | Not supported. Value zero is returned. |
| | dot1agCfmMepDbManAddress | Not supported. Value zero is returned. |
| dot1agCfmLtrTable | dot1agCfmMdIndex | MD Index. |
| | dot1agCfmMaIndex | MA Index. |
| | dot1agCfmMepIdentifier | MA end point identifier. |
| | dot1agCfmLtrSeqNumber | Transaction identifier/ sequence number returned by a previous transmit linktrace message command. |
| | dot1agCfmLtrReceiveOrder | An index to distinguish among multiple LTRs with the same LTR Transaction Identifier field value. |
| | dot1agCfmLtrTtl | TTL field value for a returned LTR. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmLtrForwarded | Indicates if a LTM was forwarded by the responding MP. |
| | dot1agCfmLtrTerminalMep | Not supported. Value FALSE is returned. |
| | dot1agCfmLtrLastEgressIdentifier | An octet field holding the Last Egress Identifier returned in the LTR Egress Identifier TLV of the LTR. |
| | dot1agCfmLtrNextEgressIdentifier | An octet field holding the Next Egress Identifier returned in the LTR Egress Identifier TLV of the LTR. |
| | dot1agCfmLtrRelay | Value returned in the Relay Action field. |
| | dot1agCfmLtrChassisIdSubtype | networkAddress (5) is returned if senderIDTLV is received. |
| | dot1agCfmLtrChassisId | The first octet contains the IANA Address Family Numbers enumeration value for the specific address type, and octets 2 through N contain the network address value in network byte order. |
| | dot1agCfmLtrManAddressDomain | Not supported. Value zero is returned. |
| | dot1agCfmLtrManAddress | Not supported. Value zero is returned. |
| | dot1agCfmLtrIngress | The value returned in the Ingress Action Field of the LTM. |
| | dot1agCfmLtrIngressMac | MAC address returned in the ingress MAC address field. |
| | dot1agCfmLtrIngressPortIdSubtype | interfaceName (5) is returned if present. |
| | dot1agCfmLtrIngressPortId | interfaceName (5), then the octet string identifies a particular instance of the ifName object. If the particular ifName object does not contain any values, another port identifier type should be used. |
| | dot1agCfmLtrEgress | The value returned in the Egress Action Field of the LTM. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1agCfmLtrEgressMac | MAC address returned in the egress MAC address field. |
| | dot1agCfmLtrEgressPortIdSubtype | interfaceName (5) is returned if present. |
| | dot1agCfmLtrEgressPortId | interfaceName (5), then the octet string identifies a particular instance of the ifName object. If the particular ifName object does not contain any values, another port identifier type should be used. |
| | dot1agCfmLtrOrganizationSpecificTlv | All organization specific TLVs returned in the LTR. |
| dot1agCfmStackTable | dot1agCfmStackifIndex | Index object. This object represents the bridge port or aggregated port on which MEPs or MHFs might be configured. |
| | dot1agCfmStackVlanIdOrNone | Index object. VLAN ID to which the MP is attached. |
| | dot1agCfmStackMdLevel | Index object. MD Level of the Maintenance Point. |
| | dot1agCfmStackDirection | Index object. Direction in which the MP faces on the bridge port. |
| | dot1agCfmStackMdIndex | The index of the MD in the dot1agCfmMdTable to which the MP is associated. |
| | dot1agCfmStackMaIndex | The index of the MA in the dot1agCfmMaNetTable and dot1agCfmMaCompTable to which the MP is associated. |
| | dot1agCfmStackMepId | If an MEP is configured, the MEP ID. |
| | dot1agCfmStackMacAddress | MAC address of the MP. |
| dot1agCfmMaMepListTable | dot1agCfmMdIndex | MD Index. |
| | dot1agCfmMaIndex | MA Index. |
| | dot1agCfmMaMepListIdentifier | MEP identifier. |
| | dot1agCfmMaMepListRowStatus | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1agCfmFaultAlarm (NOTIFICATION) | dot1agCfmMepHighestPrDefect | A MEP has a persistent defect condition. A notification (fault alarm) is sent to the management entity with the OID of the MEP that has detected the fault.<br>The management entity receiving the notification can identify the system from the network source address of the notification, and can identify the MEP reporting the defect by the indices in the OID of the dot1agCfmMepHighestPrDefect variable in the notification: dot1agCfmMdIndex—Also the index of the MEPs.MD table entry (dot1agCfmMdTable).dot1agCfmMaIndex—Also an index (with the MD table index) of the MEP's MA network table entry (dot1agCfmMaNetTable), and (with the MD table index and component ID) of the MEP's MA component table entry.dot1agCfmMepIdentifier—MEP Identifier and final index into the MEP table (dot1agCfmMepTable). |
| dot1agCfmMaCompTable | | Not supported |
| dot1agCfmConfigErrorList Table | | Not supported |
| dot1agCfmVlanTable | | Not supported |
| dot1agCfmDefaultMdTable | | Not supported |

## IEEE8021-PAE-MIB

This MIB contains objects for the 802.1X protocol draft D10 of the 802.1X standard.

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1xPaeSystemAuthControl | | |
| dot1xPaePortTable | All objects | |
| dot1xAuthConfigTable | Not supported | In lieu of these tables, Extreme Networks supports the per-station based versions which are present in the IEEE8021X- EXTENSIONS-MIB. |
| dot1xAuthStatsTable | Not supported | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1xAuthDiagTable | Not supported | This table has been deprecated in the drafts subsequent to the 2001 version of the 802.1X standard. |
| dot1xAuthSessionStatsTable | Not supported | |
| dot1xSuppConfigTable | None | These tables are not applicable to the switch since they are for a supplicant. |
| dot1xSuppStatsTable | None | |

## IEEE8021X-EXTENSIONS-MIB

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1xAuthStationTable | All objects | |
| dot1xAuthConfigTable | All objects | |
| dot1xAuthStatsTable | All objects | |

## ISIS-MIB (draft-ietf-isis-wg-mib-10.txt)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| IsisSysTable | isisSysInstance | |
| | isisSysVersion | |
| | isisSysType | |
| | isisSysID | |
| | isisSysMaxPathSplits | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysMaxLSPGenInt | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysMaxAreaAddresses | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysPollESHelloRate | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysWaitTime | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysAdminState | |
| | isisSysLogAdjacencyChanges | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysNextCircIndex | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | isisSysL2toL1Leaking | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysMaxAge | |
| | isisSysReceiveLSPBufferSize | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysExistState | Only state destroy (6) is supported. |
| isisSysLevelTable | isisSysLevelIndex | |
| | isisSysLevelOrigLSPBuffSize | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisSysLevelMinLSPGenInt | Supported isisSysLevelMinLSPGenInt range is 1 - 120 (MIB is 1 - 65535). |
| | isisSysLevelOverloadState | |
| | isisSysLevelSetOverload | |
| | isisSysLevelSetOverloadUntil | |
| | isisSysLevelMetricStyle | |
| | isisSysLevelSPFConsiders | |
| isisManAreaAddrTable | isisManAreaAddr | |
| | isisManAreaAddrExistState | |
| isisAreaAddrTable | isisAreaAddr | isisAreaAddrTable only displays the contents of the isisManAreaAddrTable. |
| isisSysProtSuppTable | isisSysProtSuppProtocol | |
| | isisSysProtSuppExistState | Supported as read-only. |
| isisSummAddrTable | isisSummAddressType | |
| | isisSummAddress | |
| | isisSummAddrPrefixLen | |
| | isisSummAddrExistState | |
| | isisSummAddrMetric | Supported as read-only. |
| | isisSummAddrFullMetric | Supported as read-only. |
| isisCircTable | isisCircIndex | |
| | isisCircIfIndex | Supported as read-only. |
| | isisCircIfSubIndex | Supported as read-only. |
| | isisCircAdminState | Setting isisCircAdminState to off (2) deletes the entry. |
| | isisCircExistState | Only state destroy (6) is supported. |
| | isisCircType | Only broadcast (1) and ptToPt (2) are supported for isisCircType. |
| | isisCircExtDomain | Unsupported object. Always returns MIB default value if one exists for a GET. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | isisCircLevel | |
| | isisCircPassiveCircuit | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisCircMeshGroupEnabled | |
| | isisCircMeshGroup | |
| | isisCircSmallHellos | |
| | isisCircLastUpTime | |
| | isisCirc3WayEnabled | Supported as read-only. |
| isisCircLevelTable | isisCircLevelIndex | |
| | isisCircLevelMetric | Supported isisCircLevelMetric range is 1 - 63 (MIB is 0 - 63). |
| | isisCircLevelWideMetric | Supported isisCircLevelMetricWide range is 1 - 16777214 (MIB is 0 - 16777214). |
| | isisCircLevelISPriority | |
| | isisCircLevelIDOctet | Supported as read-only. |
| | isisCircLevelID | |
| | isisCircLevelDesIS | |
| | isisCircLevelHelloMultiplier | Default isisCircLevelHelloMultiplier is 3 (MIB is 10). |
| | isisCircLevelHelloTimer | Default isisCircLevelHelloTimer is 10000 ms (MIB is 3000 ms). Supported range is 1000 - 600000 ms (MIB is 10 - 600000 ms). Fractions of a second are rounded to the nearest whole second. |
| | isisCircLevelDRHelloTimer | Supported as read-only. isisCircLevelDRHelloTimer object is computed as 1/3 of isisCircLevelHelloTimer. |
| | isisCircLevelLSPThrottle | Default isisCircLevelLSPThrottle is 33 ms (MIB is 30 ms). |
| | isisCircLevelMinLSPRetransInt | |
| | isisCircLevelCSNPInterval | |
| | isisCircLevelPartSNPInterval | Unsupported object. Always returns MIB default value if one exists for a GET. |
| isisSystemCounterTable | isisSysStatLevel | |
| | isisSysStatCorrLSPs | |
| | isisSysStatAuthTypeFails | |
| | isisSysStatAuthFails | |
| | isisSysStatLSPDbaseOloads | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | isisSysStatManAddrDropFromAreas | |
| | isisSysStatAttmptToExMaxSeqNums | |
| | isisSysStatSeqNumSkips | |
| | isisSysStatOwnLSPPurges | |
| | isisSysStatIDFieldLenMismatches | |
| | isisSysStatMaxAreaAddrMismatches | |
| | isisSysStatPartChanges | |
| | isisSysSPFRuns | |
| isisCircuitCounterTable | isisCircuitType | |
| | isisCircAdjChanges | |
| | isisCircNumAdj | |
| | isisCircInitFails | |
| | isisCircRejAdjs | |
| | isisCircIDFieldLenMismatches | |
| | isisCircMaxAreaAddrMismatches | |
| | isisCircAuthTypeFails | |
| | isisCircAuthFails | |
| | isisCircLANDesISChanges | |
| isisPacketCounterTable | isisPacketCountLevel | |
| | isisPacketCountDirection | |
| | isisPacketCountIIHello | |
| | isisPacketCountISHello | |
| | isisPacketCountESHello | |
| | isisPacketCountLSP | |
| | isisPacketCountCSNP | |
| | isisPacketCountPSNP | |
| | isisPacketCountUnknown | |
| isisISAdjTable | isisISAdjIndex | |
| | isisISAdjState | |
| | isisISAdj3WayState | |
| | isisISAdjNeighSNPAAddress | |
| | isisISAdjNeighSysType | |
| | isisISAdjNeighSysID | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | isisISAdjExtendedCircID | |
| | isisISAdjUsage | |
| | isisISAdjHoldTimer | |
| | isisISAdjNeighPriority | |
| | isisISAdjLastUpTime | |
| isisISAdjAreaAddrTable | isisISAdjAreaAddrIndex | |
| | isisISAdjAreaAddress | |
| isisISAdjIPAddrTable | isisISAdjIPAddrIndex | |
| | isisISAdjIPAddressType | |
| | isisISAdjIPAddress | |
| isisISAdjProtSuppTable | isisISAdjProtSuppProtocol | |
| isisIPRATable | isisIPRAIndex | |
| | isisIPRADestType | |
| | isisIPRADest | |
| | isisIPRADestPrefixLen | |
| | isisIPRANextHopType | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisIPRANextHop | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisIPRAType | Supported as read-only. |
| | isisIPRAExistState | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisIPRAAdminState | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisIPRAMetric | Supported as read-only. |
| | isisIPRAMetricType | Supported as read-only. |
| | isisIPRAFullMetric | Supported as read-only. |
| | isisIPRASNPAAddress | Unsupported object. Always returns MIB default value if one exists for a GET. |
| | isisIPRASourceType | |
| isisLSPSummaryTable | isisLSPLevel | |
| | isisLSPId | |
| | isisLSPSeq | |
| | isisLSPZeroLife | |
| | isisLSPChecksum | |
| | isisLSPLifetimeRemain | |
| | isisLSPPDULength | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | isisLSPAttributes | |
| isisLSPTLVTable | isisLSPTLVIndex | |
| | isisLSPTLVSeq | |
| | isisLSPTLVChecksum | |
| | isisLSPTTLVType | |
| | isisLSPTLVLen | |
| | isisLSPTLVValue | |

## PIM-MIB (draft-ietf-pim-mib-v2-01.txt)

This MIB is superset of RFC 2934.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| pimInterfaceTable | pimInterfaceIfIndex | |
| | pimInterfaceAddress | |
| | pimInterfaceNetMask | |
| | pimInterfaceMode | |
| | pimInterfaceDR | |
| | pimInterfaceHelloInterval | |
| | pimInterfaceStatus | |
| | pimInterfaceJoinPruneInterval | |
| | pimInterfaceCBSRPreference | |
| | pimInterfaceTrigHelloInterval | Not supported. |
| | pimInterfaceHelloHoldtime | These objects are supported as read only. |
| | pimInterfaceLanPruneDelay | |
| | pimInterfacePropagationDelay | |
| | pimInterfaceOverrideInterval | |
| | pimInterfaceGenerationID | |
| | pimInterfaceJoinPruneHoldtime | |
| | pimInterfaceGraftRetryInterval | |
| | pimInterfaceMaxGraftRetries | |
| | pimInterfaceSRTTLThreshold | |
| | pimInterfaceLanDelayEnabled | |
| | pimInterfaceSRCapable | |
| | pimInterfaceDRPriority | This object is supported as read only. |
| pimNeighborTable | pimNeighborAddress | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | pimNeighborIfIndex | |
| | pimNeighborUpTime | |
| | pimNeighborExpiryTime | |
| | pimNeighborMode | Feature unsupported, only default value is returned. |
| | pimNeighborLanPruneDelay | Feature unsupported, only default value is returned. |
| | pimNeighborOverrideInterval | |
| | pimNeighborTBit | Feature unsupported so only default value is returned. |
| | pimNeighborSRCapable | Feature unsupported so only default value is returned. |
| | pimNeighborDRPresent | Feature unsupported so only default value is returned. |
| pimIpMRouteTable | pimIpMRouteUpstreamAssertTimer | |
| | pimIpMRouteAssertMetric | |
| | pimIpMRouteAssertMetricPref | |
| | pimIpMRouteAssertRPTBit | |
| | pimIpMRouteFlags | |
| | pimIpMRouteRPFNeighbor | |
| | pimIpMRouteSourceTimer | |
| | pimIpMRouteOriginatorSRTTL | Feature unsupported so only default value is returned. |
| pimIpMRouteNextHopTable | pimIpMRouteNextHopPruneReason | |
| | pimIpMRouteNextHopAssertWinner | |
| | pimIpMRouteNextHopAssertTimer | |
| | pimIpMRouteNextHopAssertMetric | Not supported. |
| | pimIpMRouteNextHopAssertMetricPref | Not supported. |
| | pimIpMRouteNextHopJoinPruneTimer | Not supported. |
| pimRPSetTable | pimRPSetGroupAddress | |
| | pimRPSetGroupMask | |
| | pimRPSetAddress | |
| | pimRPSetHoldTime | |
| | pimRPSetExpiryTime | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | pimRPSetComponent | |
| pimCandidateRPTable | pimCandidateRPGroupAddress | |
| | pimCandidateRPGroupMask | |
| | pimCandidateRPAddress | |
| | pimCandidateRPRowStatus | |
| pimComponentTable | pimComponentIndex | |
| | pimComponentBSRAddress | |
| | pimComponentBSRExpiryTime | |
| | pimComponentCRPHoldTime | This object is supported as read only. |
| | pimComponentStatus | |
| Scalars | pimJoinPruneInterval | |
| | pimSourceLifetime | State Refresh feature is not supported, so these variables are set to defaults. |
| | pimStateRefreshInterval | |
| | pimStateRefreshLimitInterval | |
| | pimStateRefreshTimeToLive | |
| PIM Traps | pimNeighborLoss | Not supported. |

## SNMPv3 MIBs

The ExtremeXOS *SNMP* stack fully supports the SNMPv3 protocol and therefore implements the MIBs in the SNMPv3 RFCs.

Specifically, the MIBs in following RFCs are fully supported:

- RFC 2576—Coexistence between Version 1, Version 2, and Version 3 of the Internet-standard Network Management Framework
- RFC 3410—Introduction and Applicability Statements for Internet-Standard Management Framework
- RFC 3411—An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks
- RFC 3412—Message Processing and Dispatching for the Simple Network Management Protocol (SNMP)
- RFC 3413—Simple Network Management Protocol (SNMP) Applications.
- RFC 3414—User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)
- RFC 3415—View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)
- RFC 3826—The Advanced Encryption Standard (AES) Cipher Algorithm in the SNMP User-based Security Model

## RFC 1213 (MIB-II)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| System group scalars | All objects | The sysServices object always returns the value 79. |
| Interfaces group | | Supported as per RFC 2233. |
| IP Group scalars | All objects | |
| ipAddrTable | All objects | |
| ipRouteTable | All objects | Supported as read only. Routes are indexed by prefix only. |
| ipNetToMediaTable | All objects | context support is available for the ipNetToMediaTable in MIB-2. In order to retrieve entries in various VRs for ipNetToMediaTable in MIB-2, VRs can be mentioned as contexts. For example, to get the entries in *VR-Mgmt*, the following net-snmp query needs to be used:<br>`snmpwalk -v3 -n "VR-Mgmt" -u <v3username> -a md5 -a <v3user-authentication password> <deviceIpAddress/Hostname> ipNetToMediaTable`<br>Here -n has the context value of "VR-Mgmt". If no context is specified in the *SNMP* query, then only the entries in the *VR-Default* are shown. |
| *ICMP (Internet Control Message Protocol)* group | All objects | |
| TCP group scalars | All objects | |
| tcpConnTable | All objects | |
| UDP group scalars | All objects | |
| udpTable | All objects | |
| EGP Group | Not supported | |
| SNMP group | All objects | |
| At group | All objects | Supported as read only. |

## RFC 1215

This MIB defines an SMI for SNMPv1 traps, and some traps themselves.

Of these, the following are supported.

| Traps | Comments |
|---|---|
| coldStart | |
| warmStart | |

| Traps | Comments |
|---|---|
| authenticationFailure | The authentication failure trap will have additional extreme proprietary varbinds (extremeAuthFailSrcAddr, extremeAuthFailSrcAddressType , extremeAuthFailSrcAddress, extremeAuthFailSrcAddressVrName ) |
| linkDown | |
| linkUp | |

## RFC 1493 (BRIDGE-MIB) and draft-ietf-bridge-rstpmib-03.txt

The BRIDGE-MIB has been augmented with draft-ietf-bridge-rstpmib-03.txt for 802.1w support.

Objects below that are defined in the latter are marked as such.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| dot1dBase group scalars | dot1dBaseBridgeAddress | |
| | dot1dBaseNumPorts | This object returns the number of ports in *STP (Spanning Tree Protocol)* domain s0, not the total number of ports on the switch. |
| | dot1dBaseType | |
| dot1dBasePortTable | dot1dBasePort<br>dot1dBasePortIfIndex<br>dot1dBasePortCircuit<br>dot1dBasePortMtuExceededDiscards | dot1BasePort is supported and returns the port number.<br>dot1dBasePortIfIndex is supported.<br>dot1dBasePortCircuit always has the value { 0 0 } since there is a one to one correspondence between a physical port and its ifIndex.<br>dot1dBasePortMtuExceededDiscards is supported and value can be checked if packets exceeds MTU configured on particular ports. |
| dot1dStp group scalars | dot1dStpProtocolSpecification | Values for these objects will be returned for the *STP* domain s0 only. For other domains, see the EXTREME-STPEXTENSTIONS-MIB. |
| | dot1dStpPriority | |
| | dot1dStpTimeSinceTopologyChange | |
| | dot1dStpTopChanges | |
| | dot1dStpDesignatedRoot | |
| | dot1dStpRootCost | |
| | dot1dStpRootPort | |
| | dot1dStpMaxAge | |
| | dot1dStpHelloTime | |
| | dot1dStpHoldTime | |
| | dot1dStpForwardDelay | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | dot1dStpBridgeMaxAge | |
| | dot1dStpBridgeHelloTime | |
| | dot1dStpBridgeForwardDelay | |
| | dot1dStpVersion | This object is not present in the original RFC1493, but is defined in the Internet draft draft-ietf-bridge-rstp-mib-03.txt. |
| | dot1dStpTxHoldCount | This object is not present in the original RFC1493, but is defined in the Internet draft draft-ietf-bridge-rstp-mib-03.txt. This object not supported; it always returns a value of (1). Attempting to set it yields an error. |
| | dot1dStpPathCostDefault | This object is not present in the original RFC1493, but is defined in the Internet draft draft-ietf-bridge-rstp-mib-03.txt. For this object only 8021d1998 (1) is supported at this time, not stp802112001 (2). Attempting to set (2) yields an error. |
| dot1dStpExtPortTable | All objects | This object is not present in the original RFC1493, but is defined in the Internet draft draft-ietf-bridge-rstp-mib-03.txt. The object 'dot1dStpPortProtocolMigration' is not supported; it always returns a value of (2). Attempting to set it yields an error. |
| dot1dStpPortTable | All objects | |
| STP Traps | newRoot | |
| | topologyChange | |
| dot1dTpFdbTable | Supported | The object dot1dTpFdbTable displays ports and *FDB (forwarding database)* MAC addresses. They include both the static and dynamic FDB entries on the switch. The MIB does not provide a way to identify the *VLAN* on which the entry was learned. The port numbers are assumed to be 1 to 128 on Slot 1, and 128 to 255 on Slot 2, etc. (that is, with a total of 128 ports on each of the slots on a chassis system). |
| dot1dTpPortTable | Supported | |
| dot1dStatic group | Supported | |
| | Dot1dStaticAllowedToGoTo | Not supported |

## RFC 4363 (Q-BRIDGE-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| RFC 4363 dot1qBasegroup | dot1qVlanVersionNumber<br>dot1qMaxVlanId<br>dot1qMaxSupportedVlans<br>dot1qNumVlans<br>dot1qGvrpStatus | |
| dot1qPortVlanTable | dot1qPvid<br>dot1qPortAcceptcableFrameTypes<br>dot1qPortIngressFiltering<br>dot1qPortGvrpStatus<br>dot1qPortGvrpFailedRegistrations<br>dot1qPortGvrpLastPduOrigin<br>dot1qPortRestrictedVlanRegistration | |
| dot1qVlanStaticTable | dot1qVlanStaticName<br>dot1qVlanStaticEgressPorts<br>dot1qVlanForbiddenEgressPorts<br>dot1qVlanStaticUntaggedPorts<br>dot1qVlanStaticRowStatus | |

## RFC 1724 (RIPv2-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| rip2Globals | rip2GlobalRouteChanges | |
| | rip2GlobalQueries | |
| rip2IfStatTable | All objects | |
| rip2IfConfTable | All objects | |
| rip2PeerTable | Not supported | |

## RFC 1757 (RMON-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| etherStatsTable | All objects, except etherStatsDropEvents | |
| historyControlTable | All objects | |
| etherHistoryTable | All objects, except etherHistoryDropEvents | |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| alarmTable | All objects | |
| eventTable | All objects | |
| logTable | All objects | |

## RFC 1850 (OSPF-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| ospfGeneralGroup | All objects | |
| ospfAreaTable | All objects | |
| ospfStubAreaTable | All objects | |
| ospfLsdbTable | All objects | |
| ospfAreaRangeTable | All objects | |
| ospfHostTable | All objects | |
| ospfIfTable | All objects | |
| ospfIfMetricTable | All objects | |
| ospfVirtIfTable | All objects | |
| ospfNbrTable | All objects | |
| ospfVirtNbrTable | All objects | |
| ospfExtLsdbTable | All objects | |
| ospfAreaAggregateTable | All objects | |
| ospfTrap | All traps | |

## RFC 2021 (RMON2-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | probeCapabilities | |
| | probeSoftwareRev | |
| | probeHardwareRev | |
| | probeDateTime | |
| | probeResetControl | |
| trapDestTable | All objects | |

## RFC 2233 (IF-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | IfNumber | |
| ifTable | ifIndex | The ifIndex for ports is calculated as ((slot * 1000) + port). For VLANs, the ifIndex starts from 1000001. |
| | ifDescr | |
| | ifType | Only the following values are supported: {other, ethernetCsmacd, softwareLoopback, propVirtual} |
| | ifMtu | |
| | ifSpeed | |
| | ifPhysAddress | |
| | ifAdminStatus | The testing state is not supported. |
| | ifOperStatus | |
| | ifLastChange | |
| | ifInOctets | Updated every time *SNMP* queries this counter. |
| | ifInUcastPkts | Updated every time SNMP queries this counter. |
| | IfInNUcastPkts(deprecated) | Though deprecated, this object returns a value if the system keeps a count. Updated every time SNMP queries this counter. |
| | ifInDiscards | No count is kept of this object, so it always returns 0. |
| | ifInErrors | Updated every time SNMP queries this counter. |
| | ifInUnknownProtos | No count is kept of this object, so it always returns 0. |
| | ifOutOctets | Updated every time SNMP queries this counter. |
| | ifOutUcastPkts | Updated every time SNMP queries this counter. |
| | IfOutNUcastPkts(deprecated) | Though deprecated, this object returns a value if the system keeps a count. Updated every time SNMP queries this counter. |
| | ifOutDiscards | No count is kept of this object, so it always returns 0. |
| | ifOutErrors | Updated every time SNMP queries this counter. |
| | IfOutQLen(deprecated) | Though deprecated, this object will returns a value if the system keeps a count. Updated every time SNMP queries this counter. |
| | IfSpecific | Not implemented. Always returns iso.org.dod.internet. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| ifXTable | All objects | Only port interfaces return non-zero values for the counter objects in this table.<br>The object ifPromiscuousMode is supported as read-only.<br>ifCounterDiscontinuityTime is not implemented.<br>All the statistics counters in the ifXTable are updated every time SNMP queries them. |
| ifStackTable | Not supported | |
| IfTestTable | Not supported | |
| ifRcvAddressTable | All objects | ifRcvAddressTable is supported as read-only. Also, only entries for physical ports appear in it. |
| snmpTraps | linkDown | |
| | linkUp | |

## RFC 2465 (IPV6 MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| ipv6Forwarding | All objects | |
| ipv6DefaultHopLimit | All objects | |
| ipv6Interfaces | All objects | |
| ipv6IfTableLastChange | All objects | |
| ipv6IfTable | All objects except ipv6IfEffectiveMtu | |
| ipv6IfStatsTable | All objects | |
| ipv6AddrPrefixTable | All objects | |
| ipv6AddrTable | All objects | |
| ipv6RouteNumber | All objects | |
| ipv6DiscardedRoutes | All objects | |
| ipv6RouteTable | All objects | |
| ipv6NetToMediaTable | All objects | |

## RFC 2466 (IPV6 ICMP MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| ipv6IfIcmpTable | All objects | |

## RFC 2613 (SMON)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| smonVlanStatsControlTable | smonVlanStatsControlIndex | A unique arbitrary index for this smonVlanStatsControlEntry. |
| | smonVlanStatsControlDataSource | The source of data for this set of *VLAN* statistics. This object MAY NOT be modified if the associated smonVlanStatsControlStatus object is equal to active (1). |
| | smonVlanStatsControlCreateTime | The value of sysUpTime when this control entry was last activated. This object allows a management station to detect deletion and recreation cycles between polls. |
| | smonVlanStatsControlOwner | Administratively assigned named of the owner of this entry. It usually defines the entity that created this entry and is therefore using the resources assigned to it, though there is no enforcement mechanism, nor assurance that rows created are ever used. |
| | smonVlanStatsControlStatus | The status of this row. An entry MAY NOT exist in the active state unless all objects in the entry have an appropriate value. If this object is not equal to active (1), all associated entries in the smonVlanIdStatsTable SHALL be deleted. |
| smonVlanIdStatsTable | smonVlanIdStatsId | The unique identifier of the VLAN monitored for this specific statistics collection. Tagged packets match the VID for the range between 1 and 4094. An external RMON probe MAY detect VID=0 on an Inter Switch Link, in which case the packet belongs to a VLAN determined by the PVID of the ingress port. The VLAN to which such a packet belongs can be determined only by a RMON probe internal to the switch. |
| | smonVlanIdStatsTotalPkts | The total number of packets counted on this VLAN. |
| | smonVlanIdStatsTotalOverflowPkts | The number of times the associated smonVlanIdStatsTotalPkts counter has overflowed. |
| | smonVlanIdStatsTotalHCPkts | The total number of packets counted on this VLAN. |
| | smonVlanIdStatsTotalOctets | The total number of octets counted on this VLAN. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | smonVlanIdStatsTotalOverflowOctets | The number of times the associated smonVlanIdStatsTotalOctets counter has overflowed. |
| | smonVlanIdStatsTotalHCOctets | The total number of octets counted on this VLAN. |
| | smonVlanIdStatsNUcastOverflowOctets | The number of times the associated smonVlanIdStatsNUcastOctets counter has overflowed. |
| | smonVlanIdStatsCreateTime | The value of sysUpTime when this entry was last activated. This object allows a management station to detect deletion and recreation cycles between polls. |
| dataSourceDapsTable | dataSourceCapsObject | Defines an object that can be a SMON data source or a source or a destination for a port copy operation. |
| | dataSourceRmonCaps | General attributes of the specified dataSource. Note that these are static attributes, which SHOULD NOT be adjusted because of current resources or configuration. |
| | dataSourceCopyCaps | PortCopy function capabilities of the specified dataSource. Note that these are static capabilities, which SHOULD NOT be adjusted because of current resources or configuration. |
| | dataSourceCapsIfIndex | This object contains the ifIndex value of the ifEntry associated with this smonDataSource. The agent MUST create 'propVirtual' ifEntries for each dataSourceCapsEntry of type VLAN or entPhysicalEntry. |
| portCopyConfigTable | portCopySource | The ifIndex of the source which has all packets redirected to the destination as defined by portCopyDest. |
| | portCopyDest | Defines the ifIndex destination for the copy operation. |
| | portCopyDestDropEvents | The total number of events in which port copy packets were dropped by the switch at the destination port due to lack of resources. Note that this number is not necessarily the number of packets dropped; it is just the number of times this condition has been detected. A single dropped event counter is maintained for each portCopyDest. Thus all instances associated with a given portCopyDest will have the same portCopyDestDropEvents value. The value for this field is “0” (zero) due to hardware limitation. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | portCopyDirection | This object affects the way traffic is copied from a switch source port, for the indicated port copy operation. If this object has the value copyRxOnly (1), then only traffic received on the indicated source port will be copied to the indicated destination port. If this object has the value copyTxOnly (2), then only traffic transmitted out the indicated source port will be copied to the indicated destination port. If this object has the value copyBoth (3), then all traffic received or transmitted on the indicated source port will be copied to the indicated destination port.<br>The creation and deletion of instances of this object is controlled by the portCopyRowStatus object. Note that there is no guarantee that changes in the value of this object performed while the associated portCopyRowStatus object is equal to active will not cause traffic discontinuities in the packet. |
| | portCopyStatus | Defines the status of the port copy entry. In order to configure a source to destination portCopy relationship, both source and destination interfaces MUST be present as an ifEntry in the ifTable and their respective ifAdminStatus and ifOperStatus values must be equal to up (1). If the value of any of those two objects changes after the portCopyEntry is activated, portCopyStatus transitions to notReady (3).<br>The capability of an interface to be source or destination of a port copy operation is described by the copySourcePort (0) and copyDestPort (1) bits in dataSourceCopyCaps. Those bits should be appropriately set by the agent, in order to allow for a portCopyEntry to be created. |
| smonPrioStatsControlTable | | Not supported due to hardware limitations. |
| smonPrioStatsTable | | |

## RFC 2665 (EtherLike-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| Dot3StatsTable | dot3StatsIndex | |
| | dot3StatsAlignmentErrors | |
| | dot3StatsFCSErrors | |
| | dot3StatsSingleCollisionFrames | |
| | dot3StatsMultipleCollisionFrames | |
| | dot3StatsSQETestErrors | Not supported |
| | dot3StatsDeferredTransmissions | |
| | dot3StatsLateCollisions | |
| | dot3StatsExcessiveCollisions | |
| | dot3StatsInternalMacTransmitErrors | |
| | dot3StatsCarrierSenseErrors | Not supported |
| | dot3StatsFrameTooLongs | |
| | dot3StatsInternalMacReceiveErrors | |
| | dot3StatsSymbolErrors | Not supported |
| | dot3StatsEtherChipSet | |
| | dot3StatsDuplexStatus | |
| dot3CollTable | dot3CollCount | |
| | dot3CollFrequencies | |
| dot3ControlTable | | Not supported |
| dot3PauseTable | | Not supported |

### Other Unsupported Tables and Nodes in EtherLike MIB:

dot3ControlTable, dot3PauseTable, dot3Tests – all nodes under this, dot3Errors, etherConformance, etherGroups, etherCompliance, dot3Compliance. RFC 1657 (Definitions of Managed Objects for the Fourth Version of the Border Gateway Protocol (*BGP (Border Gateway Protocol)*-4) using SMIv2).

All tables and variables of this MIB are supported with read-only access.

## R_RFC 2668 (MAU-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| ifMauTable | All objects | |
| ifJackTable | All objects | |
| ifMauAutoNegTable | All objects | Setting auto-negotiation through *SNMP* is not supported. |

The following new Extreme proprietary MAU types have been added to the ifMauType textual convention:

```
extremeMauType1000BaseWDMHD OBJECT IDENTIFIER

    ::= { extremeMauType 7 }

            "Gigabit WDM, half duplex"

extremeMauType1000BaseWDMFD OBJECT IDENTIFIER

    ::= { extremeMauType 8 }

            "Gigabit WDM, full duplex"

extremeMauType1000BaseLX70HD OBJECT IDENTIFIER

    ::= { extremeMauType 9 }

            "Gigabit LX70, half duplex"

extremeMauType1000BaseLX70FD OBJECT IDENTIFIER

    ::= { extremeMauType 10 }

            "Gigabit LX70, full duplex"

extremeMauType1000BaseZXHD OBJECT IDENTIFIER

    ::= { extremeMauType 11 }

            "Gigabit ZX, half duplex"

extremeMauType1000BaseZXFD OBJECT IDENTIFIER

    ::= { extremeMauType 12 }

            "Gigabit ZX, full duplex"

extremeMauType1000BaseLX100HD OBJECT IDENTIFIER

    ::= { extremeMauType 13 }

            "Gigabit LX100, half duplex"

extremeMauType1000BaseLX100FD OBJECT IDENTIFIER

    ::= { extremeMauType 14 }

            "Gigabit LX100, full duplex"

extremeMauType10GBaseCX4 OBJECT IDENTIFIER

    ::= { extremeMauType 15 }

            "10 Gigabit CX4"

extremeMauType10GBaseZR OBJECT IDENTIFIER

    ::= { extremeMauType 16 }

            "10 Gigabit ZR"

extremeMauType10GBaseDWDM OBJECT IDENTIFIER
```

```
    ::= { extremeMauType 17 }

            "10 Gigabit DWDM"

extremeMauType10GBaseCX OBJECT IDENTIFIER

    ::= { extremeMauType 18 }

            "10 Gigabit CX - SFP+ twin coax cable"

extremeMauType10GBaseT OBJECT IDENTIFIER

    ::= { extremeMauType 19 }

            "10 Gigabit BaseT"

extremeMauType40G OBJECT IDENTIFIER

    ::= { extremeMauType 20 }

            "40 Gigabit interface"
```

Corresponding MAU Type List Bits values have been added:

```
extreme_ifMauTypeListBits_b1000baseWDMHD    -- 64

extreme_ifMauTypeListBits_b1000baseWDMFD    -- 65

extreme_ifMauTypeListBits_b1000baseLX70HD    -- 66

extreme_ifMauTypeListBits_b1000baseLX70FD    -- 67

extreme_ifMauTypeListBits_b1000baseZXHD    -- 68

extreme_ifMauTypeListBits_b1000baseZXFD    -- 69
```

The following standards-based additions have been made as a 'Work in Progress', as per draft-ietf-hubmib-mau-mib-v3-02.txt:

A new enumeration 'fiberLC(14)' for the JackType textual convention.

New MAU types:

```
dot3MauType10GigBaseX OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "X PCS/PMA (per 802.3 section 48), unknown PMD."

    ::= { dot3MauType 31 }

dot3MauType10GigBaseLX4 OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "X fiber over WWDM optics (per 802.3 section 53)"

    ::= { dot3MauType 32 }
```

```
dot3MauType10GigBaseR OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "R PCS/PMA (per 802.3 section 49), unknown PMD."

    ::= { dot3MauType 33 }

dot3MauType10GigBaseER OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "R fiber over 1550 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 34 }

dot3MauType10GigBaseLR OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "R fiber over 1310 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 35 }

dot3MauType10GigBaseSR OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "R fiber over 850 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 36 }

dot3MauType10GigBaseW OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "W PCS/PMA (per 802.3 section 49 and 50), unknown PMD."

    ::= { dot3MauType 37 }

dot3MauType10GigBaseEW OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "W fiber over 1550 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 38 }

dot3MauType10GigBaseLW OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "W fiber over 1310 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 39 }

dot3MauType10GigBaseSW OBJECT-IDENTITY

    STATUS      current

    DESCRIPTION "W fiber over 850 nm optics (per 802.3 section 52)"

    ::= { dot3MauType 40 }
```

Corresponding new Mau Type List bit values:

```
b10GbaseX(31)                    — 10GBASE-X
b10GbaseLX4(32)                    — 10GBASE-LX4
b10GbaseR(33)                    — 10GBASE-R
b10GbaseER(34                    — 10GBASE-ER
b10GbaseLR(35)                     — 10GBASE-LR
b10GbaseSR(36)                     — 10GBASE-SR
b10GbaseW(37)                    — 10GBASE-W
b10GbaseEW(38)                     — 10GBASE-EW
b10GbaseLW(39)                     — 10GBASE-LW
b10GbaseSW(40)                     — 10GBASE-SW
```

## RFC 6933 (ENTITY-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| entityPhysicalTable | entPhysicalIndex | The index for this entry. |
| | entPhysicalDescr | A textual description of physical entity. This object should contain a string that identifies the manufacturer's name for the physical entity and should be set to a distinct value for each version or model of the physical entity. |
| | entPhysicalVendorType | An indication of the vendor-specific hardware type of the physical entity. Note that this is different from the definition of MIB-II's sysObjectID. An agent should set this object to an enterprise-specific registration identifier value indicating the specific equipment type in detail. The associated instance of entPhysicalClass is used to indicate the general type of hardware device. If no vendor-specific registration identifier exists for this physical entity, or the value is unknown by this agent, then the value { 0 0 } is returned. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | entPhysicalContainedIn | The value of entPhysicalIndex for the physical entity that 'contains' this physical entity. A value of zero indicates this physical entity is not contained in any other physical entity. Note that the set of 'containment' relationships define a strict hierarchy; that is, recursion is not allowed. In the event that a physical entity is contained by more than one physical entity (e.g., double-wide modules), this object should identify the containing entity with the lowest value of entPhysicalIndex. |
| | entPhysicalClass | An indication of the general hardware type of the physical entity. An agent should set this object to the standard enumeration value that most accurately indicates the general class of the physical entity, or the primary class if there is more than one entity. If no appropriate standard registration identifier exists for this physical entity, then the value 'other(1)' is returned. If the value is unknown by this agent, then the value 'unknown(2)' is returned. |
| | entPhysicalParentRelPos | An indication of the relative position of this 'child' component among all its 'sibling' components. Sibling components are defined as entPhysicalEntries that share the same instance values of each of the entPhysicalContainedIn and entPhysicalClass objects. |
| | entPhysicalName | The textual name of the physical entity. The value of this object should be the name of the component as assigned by the local device and should be suitable for use in commands entered at the device's 'console'. This might be a text name (e.g., 'console') or a simple component number (e.g., port or module number, such as '1'), depending on the physical component naming syntax of the device. If there is no local name, or if this object is otherwise not applicable, then this object contains a zero-length string. Note that the value of entPhysicalName for two physical entities will be the same in the event that the console interface does not distinguish between them, e.g., slot-1 and the card in slot-1. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | entPhysicalHardwareRev | The vendor-specific hardware revision string for the physical entity. The preferred value is the hardware revision identifier actually printed on the component itself (if present). Note that if revision information is stored internally in a non-printable (e.g., binary) format, then the agent must convert such information to a printable format in an implementation-specific manner. If no specific hardware revision string is associated with the physical component, or if this information is unknown to the agent, then this object will contain a zero-length string. |
| | entPhysicalFirmwareRev | The vendor-specific firmware revision string for the physical entity. Note that if revision information is stored internally in a non-printable (e.g., binary) format, then the agent must convert such information to a printable format in an implementation-specific manner. If no specific firmware programs are associated with the physical component, or if this information is unknown to the agent, then this object will contain a zero-length string. |
| | entPhysicalSoftwareRev | The vendor-specific software revision string for the physical entity. Note that if revision information is stored internally in a non-printable (e.g., binary) format, then the agent must convert such information to a printable format in an implementation-specific manner. If no specific software programs are associated with the physical component, or if this information is unknown to the agent, then this object will contain a zero-length string. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | entPhysicalSerialNum | The vendor-specific serial number string for the physical entity. The preferred value is the serial number string actually printed on the component itself (if present). On the first instantiation of a physical entity, the value of entPhysicalSerialNum associated with that entity is set to the correct vendor-assigned serial number, if this information is available to the agent. If a serial number is unknown or non-existent, the entPhysicalSerialNum will be set to a zero-length string instead. Note that implementations that can correctly identify the serial numbers of all installed physical entities do not need to provide write access to the entPhysicalSerialNum object. Agents that cannot provide non-volatile storage for the entPhysicalSerialNum strings are not required to implement write access for this object. Not every physical component will have a serial number, or even need one. Physical entities for which the associated value of the entPhysicalIsFRU object is equal to 'false(2)' (e.g., the repeater ports within a repeater module) do not need their own unique serial numbers. An agent does not have to provide write access for such entities and may return a zero-length string. If write access is implemented for an instance of entPhysicalSerialNum and a value is written into the instance, the agent must retain the supplied value in the entPhysicalSerialNum instance (associated with the same physical entity) for as long as that entity remains instantiated. This includes instantiations across all re-initializations/reboots of the network management system, including those resulting in a change of the physical entity's entPhysicalIndex value. |
| | entPhysicalMfgName | The name of the manufacturer of this physical component. The preferred value is the manufacturer name string actually printed on the component itself (if present). Note that comparisons between instances of the entPhysicalModelName, entPhysicalFirmwareRev, entPhysicalSoftwareRev, and the entPhysicalSerialNum objects are only meaningful amongst entPhysicalEntries with the same value of entPhysicalMfgName. If the manufacturer name string associated with the physical component is unknown to the agent, then this object will contain a zero-length string. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | entPhysicalModelName | The vendor-specific model name identifier string associated with this physical component. The preferred value is the customer-visible part number, which may be printed on the component itself. If the model name string associated with the physical component is unknown to the agent, then this object will contain a zero-length string. |
| | entPhysicalAlias | Not supported. |
| | entPhysicalAssetID | Not supported. |
| | entPhysicalIsFRU | This object indicates whether or not this physical entity is considered a 'field replaceable unit' by the vendor. If this object contains the value 'true(1)', then this entPhysicalEntry identifies a field replaceable unit. For all entPhysicalEntries that represent components permanently contained within a field replaceable unit, the value 'false(2)' should be returned for this object. |
| | entPhysicalMfgDate | This object contains the date of manufacturing of the managed entity. If the manufacturing date is unknown or not supported, the object is not instantiated. The special value '0000000000000000'H may also be returned in this case. |
| | entPhysicalUris | This object contains identification information about the physical entity. The object contains URIs; therefore, the syntax of this object must conform to RFC 3986. Multiple URIs may be present and are separated by white space characters. Leading and trailing white space characters are ignored.If no URI identification information is known about the physical entity, the object is not instantiated. A zero-length octet string may also be returned in this case.<br><br>Note: This object has read-write permission, but is implemented with read-only mode. Write operation is not supported. |
| | entPhysicalUUID | This object contains identification information about the physical entity. The object contains a Universally Unique Identifier, the syntax of this object must conform to RFC 4122, Section 4.1. A zero-length octet string is returned if no UUID information is known. |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | entAliasMappingIdentifier | The value of this object identifies a particular conceptual row associated with the indicated entPhysicalIndex and entLogicalIndex pair.Because only physical ports are modeled in this table, only entries that represent interfaces or ports are allowed. If an ifEntry exists on behalf of a particular physical port, then this object should identify the associated ifEntry. For repeater ports, the appropriate row in the 'rptrPortGroupTable' should be identified instead. |
| | entPhysicalChildIndex | The value of entPhysicalIndex for the contained physical entity. |
| | entConfigChange | An entConfigChange notification is generated when the value of entLastChangeTime changes. It can be utilized by an NMS to trigger logical/ physical entity table maintenance polls. An agent should not generate more than one entConfigChange 'notification-event' in a given time interval (five seconds is the suggested default). A 'notification-event' is the transmission of a single trap or inform PDU to a list of notification destinations. If additional configuration changes occur within the throttling period, then notification-events for these changes should be suppressed by the agent until the current throttling period expires. At the end of a throttling period, one notification-event should be generated if any configuration changes occurred since the start of the throttling period. In such a case, another throttling period is started right away. An NMS should periodically check the value of entLastChangeTime to detect any missed entConfigChange notification-events. |
| entityGeneral | entLastChangeTime | The value of sysUpTime at the time a conceptual row is created, modified, or deleted in any of these tables:<br>• entPhysicalTable<br>• entLogicalTable<br>• entLPMappingTable<br>• entAliasMappingTable<br>• entPhysicalContainsTable |

## RFC 2787 (VRRP-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| vrrpOperations | vrrpNodeVersion | |
| | vrrpNotificationCntl | |
| vrrpStatistics | vrrpRouterChecksumErrors | |
| | vrrpRouterVersionErrors | |
| | vrrpRouterVrIdErrors | |
| vrrpOperTable | All objects | Creation of a new row or modifying an existing row requires vrrpOperAdminState to be set to down; otherwise any kind of set will fail on this table.<br>vrrpOperAuthType does not support ipAuthenticationHeader. |
| vrrpAssoIpAddrTable | All objects | |
| vrrpRouterStatsTable | All objects | |
| vrrpNotifications | vrrpTrapNewMaster | |
| | vrrpTrapAuthFailure | |

## RFC 3621 (PoE-MIB)

The following tables, groups, and variables are supported in this MIB.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| pethPsePortTable | All objects | Objects in this table are read-only. |
| pethMainPseTable | All objects | Objects in this table are read-only. |
| pethNotifications | pethPsePortOnOffNotification | |
| | pethMainPowerUsageOnNotification | |
| | pethMainPowerUsageOffNotification | |

## RFC 5601 (PW-STD-MIB)

The following tables, groups, and variables are supported in this MIB.

All tables and variables of this MIB are supported as read only. The comments here are abbreviated versions of the description in the RFC documentation.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| pwTable | pwIndex | A unique index for the conceptual row identifying a PW within this table. |
| | pwPeerAddr | This object contains the value of the peer node address of the PW/PE maintenance protocol entity. This object should contain a value of all zeroes if not applicable (pwPeerAddrType is 'unknown'). |

| Table/Group | Supported Variables | Comments |
|---|---|---|
| | pwID | Pseudowire identifier. If the pwOwner object is pwIdFecSignaling or l2tpControlProtocol, then this object is signaled in the outgoing PW ID field within the Virtual Circuit FEC Element. For other values of pwOwner, this object is not signaled and it can be set to zero. |
| | pwLocalCapabAdvert | If a maintenance protocol is used, it indicates the capabilities the local node will advertise to the peer. |
| | pwRemoteGroupID | This object is obtained from the Group ID field as received through the maintenance protocol used for PW setup. Value of zero is reported if not used. Value of 0xFFFFFFFF is used if the object is yet to be defined by the PW maintenance protocol. |
| | pwOutboundLabel | The PW label used in the outbound direction (i.e., toward the PSN). It might be set manually if pwOwner is 'manual'; otherwise, it is set automatically. |
| | pwInboundLabel | The PW label used in the inbound direction (i.e., packets received from the PSN). It may be set manually if pwOwner is 'manual'; otherwise, it is set automatically. |
| | pwCreateTime | The value of sysUpTime at the time this PW was created. |
| | pwUpTime | Specifies the time since last change of pwOperStatus to Up (1). |
| | pwLastChange | The value of sysUpTime at the time the PW entered its current operational state. If the current state was entered prior to the last re-initialization of the local network management subsystem, then this object contains a zero value. |
| | pwAdminStatus | The desired operational status of this PW. This object can be set at any time. |
| | pwOperStatus | This object indicates the operational status of the PW; it does not reflect the status of the Customer Edge (CE) bound interface. |
| | pwLocalStatus | Indicates the status of the PW in the local node. |
| | pwRemoteStatus | Indicates the status of the PW as was advertised by the remote. |
| | pwRowStatus | For creating, modifying, and deleting this row. This object can be changed at any time. |
| | pwOamEnable | This variable indicates if OAM is enabled for this PW. It can be changed at any time. |
| pwIndexMappingTable | All objects | This table enables the reverse mapping of the unique PWid parameters [peer IP, PW type, and PW ID] and the pwIndex. The table is not applicable for PWs created manually or by using the generalized FEC. |

## RFC 5602 (PW-MPLS-STD-MIB)

The following tables, groups, and variables are supported in this MIB.

All tables and variables of this MIB are supported as read only. The comments here are abbreviated versions of the description in the RFC documentation.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| pwMplsTable | pwIndex | This table controls *MPLS (Multiprotocol Label Switching)*-specific parameters when the PW is going to be carried over MPLS PSN. |
| | pwMplsMplsType | This object is set by the operator to indicate the outer tunnel types, if existing. |
| | pwMplsTtl | This object is set by the operator to indicate the PW TTL value to be used on the PW shim label. |
| | pwMplsLocalLdpID | The LDP identifier of the LDP entity that creates this PW in the local node. |
| | pwMplsLocalLdpEntityIndex | The local node LDP Entity Index of the LDP entity creating this PW. |
| | pwMplsPeerLdpID | The peer LDP identifier of the LDP session. This object should return the value zero if LDP is not used or if the value is not yet known. |
| pwMplsOutboundTable | All objects | This table reports and configures the current outbound MPLS tunnels (i.e., toward the PSN) or the physical interface in the case of a PW label only that carries the PW traffic. It also reports the current outer tunnel and LSP that forward the PW traffic. |

## RFC 5603 (PW-ENET-STD-MIB)

The following tables, groups, and variables are supported in this MIB.

All tables and variables of this MIB are supported as read only. The comments here are abbreviated versions of the description in the RFC documentation.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| pwEnetTable | pwIndex | This table contains the index to the Ethernet tables associated with this Ethernet PW, the *VLAN* configuration, and the VLAN mode. |
| | pwEnetPwInstance | If multiple rows are mapped to the same PW, this index is used to uniquely identify the individual row. |
| | pwEnetPwVlan | This object defines the (service-delimiting) VLAN field value on the PW. |
| | pwEnetVlanMode | This object indicates the mode of VLAN handling between the port or the virtual port associated with the PW and the PW encapsulation. |

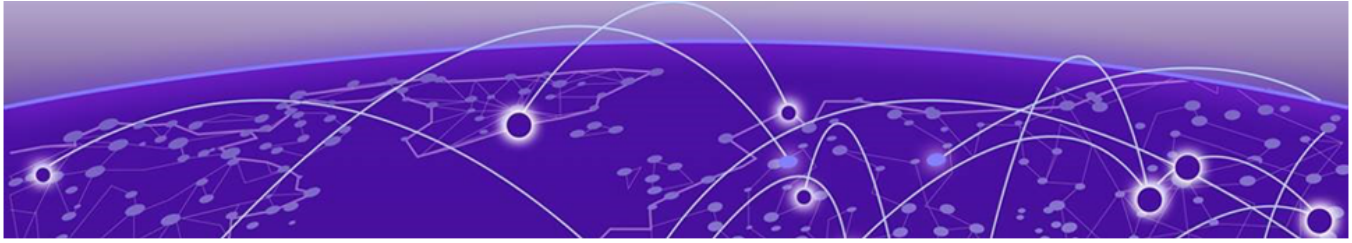
| Table/Group | Supported Variables | Comments |
|---|---|---|
| | pwEnetPortVlan | This object defines if the mapping between the original port (physical port or VPLS virtual port) to the PW is VLAN based or not. |
| | pwEnetPortIfIndex | This object is used to specify the ifIndex of the Ethernet port associated with this PW for point-to-point Ethernet service, or the ifIndex of the virtual interface of the VPLS instance associated with the PW if the service is VPLS. |
| | pwEnetRowStatus | This object enables creating, deleting, and modifying this row. |

## VPLS-MIB (draft-ietf-l2vpn-vpls-mib-02.txt)

The following tables, groups, and variables are supported in this MIB.

All tables and variables of this MIB are supported as read only. The comments here are abbreviated versions of the description in the RFC documentation.

| Table/Group | Supported Variables | Comments |
|---|---|---|
| vplsConfigTable | vplsConfigIndex | Unique index for the conceptual row identifying a VPLS service. |
| | vplsConfigName | A textual name of the VPLS. If there is no local name, or this object is otherwise not applicable, then this object MUST contain a zero-length octet string. |
| | vplsConfigAdminStatus | The desired administrative state of the VPLS service. |
| | vplsConfigRowStatus | For creating, modifying, and deleting this row. |
| | vplsConfigMtu | The value of this object specifies the MTU of this vpls instance. |
| | vplsConfigVpnId | This object indicates the IEEE 802-1990 VPN ID of the associated VPLS service. |
| vplsStatusTable | All objects | This table provides information for monitoring VPLS. |
| vplsPwBindTable | vplsConfigIndex | This table provides an association between a VPLS service and the corresponding PWs. A service can have more than one PW association. PWs are defined in the pwTable. |
| | vplsPwBindIndex | |
| | vplsPwBindType | The value of this object indicates whether the PW binding is of type mesh or spoke. |
| | vplsPwBindRowStatus | For creating, modifying, and deleting this row. |



# Software Licensing

Extreme Networks software may contain software from third party sources that must be licensed under the specific license terms applicable to such software. Applicable copyright information is provided below.

Copyright (c) 1995-1998 by Cisco Systems, Inc.

Permission to use, copy, modify, and distribute this software for any purpose and without fee is hereby granted, provided that this copyright and permission notice appear on all copies of the software and supporting documentation, the name of Cisco Systems, Inc. not be used in advertising or publicity pertaining to distribution of the program without specific prior notice be given in supporting documentation that modification, permission, and copying and distribution is by permission of Cisco Systems, Inc.

Cisco Systems, Inc. makes no representations about the suitability of this software for any purpose. THIS SOFTWARE IS PROVIDED "AS IS"' AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

MD5C.C - RSA Data Security, Inc., *MD5 (Message-Digest algorithm 5)*

Copyright (C) 1991-2, RSA Data Security, Inc. Created 1991. All rights reserved.

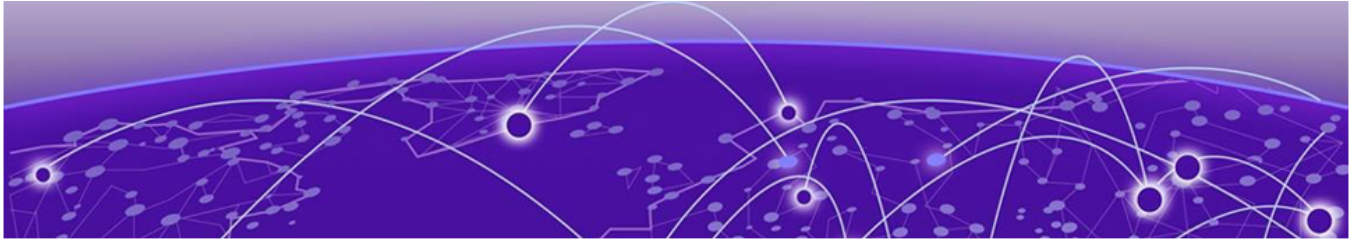
License to copy and use this software is granted provided that it is identified as the "RSA Data Security, Inc. *MD5* Message-Digest Algorithm" in all material mentioning or referencing this software or this function.

License is also granted to make and use derivative works provided that such works are identified as "derived from the RSA Data Security, Inc. MD5 Message-Digest Algorithm" in all material mentioning or referencing the derived work.

RSA Data Security, Inc. makes no representations concerning either the merchantability of this software or the suitability of this software for any particular purpose. It is provided "as is" without express or implied warranty of any kind.

These notices must be retained in any copies of any part of this documentation and/or software.

$Id: md5c.c,v 1.2.4880.1 2005/06/24 01:47:07 lindak Exp $ This code is the same as the code published by RSA Inc. It has been edited for clarity and style only.



# Glossary

**ACL**

An Access Control List is a mechanism for filtering packets at the hardware level. Packets can be classified by characteristics such as the source or destination MAC, IP address, IP type, or QoS queue. Once classified, the packets can be forwarded, counted, queued, or dropped.

**ad hoc mode**

An 802.11 networking framework in which devices or stations communicate directly with each other, without the use of an AP.

**ARP**

Address Resolution Protocol is part of the TCP/IP suite used to dynamically associate a device's physical address (MAC address) with its logical address (IP address). The system broadcasts an ARP request, containing the IP address, and the device with that IP address sends back its MAC address so that traffic can be transmitted.

**ATM**

Asynchronous Transmission Mode is a start/stop transmission in which each character is preceded by a start signal and followed by one or more stop signals. A variable time interval can exist between characters. ATM is the preferred technology for the transfer of images.

**BGP**

Border Gateway Protocol is a router protocol in the IP suite designed to exchange network reachability information with BGP systems in other autonomous systems. You use a fully meshed configuration with BGP.

BGP provides routing updates that include a network number, a list of ASs that the routing information passed through, and a list of other path attributes. BGP works with cost metrics to choose the best available path; it sends updated router information only when one host has detected a change, and only the affected part of the routing table is sent.

BGP communicates within one AS using Interior BGP (IBGP) because BGP does not work well with IGP. Thus the routers inside the AS maintain two routing tables: one for the IGP and one for IBGP. BGP uses exterior BGP (EBGP) between different autonomous systems.

**bridge**

In conventional networking terms, bridging is a Layer 2 function that passes frames between two network segments; these segments have a common network layer address. The bridged frames pass only to those segments connected at a Layer 2 level, which is called a broadcast domain (or VLAN). You must use Layer 3 routing to pass frames between broadcast domains (VLANs).

In wireless technology, bridging refers to forwarding and receiving data between radio interfaces on APs or between clients on the same radio. So, bridged traffic can be forwarded from one AP to another AP without having to pass through the switch on the wired network.

**BSS**

Basic Service Set is a wireless topology consisting of one access point connected to a wired network and a set of wireless devices. Also called an infrastructure network. See also *IBSS (Independent Basic Service Set)*.

**CEP**

Customer Edge Port, also known as Selective Q-in-Q or C-tagged Service Interface, is a role that is configured in software as a CEP VMAN port, and connects a VMAN to specific CVLANs based on the CVLAN CVID. The CNP role, which is configured as an untagged VMAN port, connects a VMAN to all other port traffic that is not already mapped to the port CEP role.

**Chalet**

Chalet is a web-based user interface for setting up and viewing information about a switch, removing the need to enter common commands individually in the CLI.

**CHAP**

Challenge-Handshake Authentication Protocol is one of the two main authentication protocols used to verify a user's name and password for PPP Internet connections. CHAP is more secure because it performs a three-way handshake during the initial link establishment between the home and remote machines. It can also repeat the authentication anytime after the link has been established.

**CIST regional root bridge**

Within an MSTP region, the bridge with the lowest path cost to the CIST root bridge is the CIST regional root bridge If the CIST root bridge is inside an MSTP region, that same bridge is the CIST regional root for that region because it has the lowest path cost to the CIST root. If the CIST root bridge is outside an MSTP region, all regions connect to the CIST root through their respective CIST regional roots.

**CIST root port**

In an NSTO environment, the port on the CIST regional root bridge that connects to the CIST root bridge is the CIST root port. The CIST root port is the master port for all MSTIs in that MSTP region, and it is the only port that connects the entire region to the CIST root bridge.

**CLI**

Command Line Interface. The CLI provides an environment to issue commands to monitor and manage switches and wireless appliances.

**CoS**

Class of Service specifies the service level for the classified traffic type.

**Data Center Connect**

DCC, formerly known as DCM (Data Center Manager), is a data center fabric management and automation tool that improves the efficiency of managing a large virtual and physical network. DCC provides an integrated view of the server, storage, and networking operations, removing the need to use multiple tools and management systems. DCC automates VM assignment, allocates appropriate network resources, and applies individual policies to various data objects in the switching fabric

(reducing VM sprawl). Learn more about DCC at http://www.extremenetworks.com/product/data-center-connect/.

**DCB**

Data Center Bridging is a set of IEEE 802.1Q extensions to standard Ethernet, that provide an operational framework for unifying Local Area Networks (LAN), Storage Area Networks (SAN) and Inter-Process Communication (IPC) traffic between switches and endpoints onto a single transport layer.

**DHCP**

Dynamic Host Configuration Protocol allows network administrators to centrally manage and automate the assignment of IP addresses on the corporate network. DHCP sends a new IP address when a computer is plugged into a different place in the network. The protocol supports static or dynamic IP addresses and can dynamically reconfigure networks in which there are more computers than there are available IP addresses.

**DoS attack**

Denial of Service attacks occur when a critical network or computing resource is overwhelmed so that legitimate requests for service cannot succeed. In its simplest form, a DoS attack is indistinguishable from normal heavy traffic. ExtremeXOS software has configurable parameters that allow you to defeat DoS attacks.

**DSSS**

Direct-Sequence Spread Spectrum is a transmission technology used in Local Area Wireless Network (LAWN) transmissions where a data signal at the sending station is combined with a higher data rate bit sequence, or chipping code, that divides the user data according to a spreading ratio. The chipping code is a redundant bit pattern for each bit that is transmitted, which increases the signal's resistance to interference. If one or more bits in the pattern are damaged during transmission, the original data can be recovered due to the redundancy of the transmission. (Compare with *FHSS (Frequency-Hopping Spread Spectrum)*.)

**EAP-TLS/EAP-TTLS**

EAP-TLS Extensible Authentication Protocol - Transport Layer Security. A general protocol for authentication that also supports multiple authentication methods, such as token cards, Kerberos, one-time passwords, certificates, public key authentication and smart cards.

IEEE 802.1x specifies how EAP should be encapsulated in LAN frames.
In wireless communications using EAP, a user requests connection to a WLAN through an access point, which then requests the identity of the user and transmits that identity to an authentication server such as RADIUS The server asks the access point for proof of identity, which the access point gets from the user and then sends back to the server to complete the authentication.

EAP-TLS provides for certificate-based and mutual authentication of the client and the network. It relies on client-side and server-side certificates to perform authentication and can be used to dynamically generate user-based and session-based WEP keys.

EAP-TTLS (Tunneled Transport Layer Security) is an extension of EAP-TLS to provide certificate-based, mutual authentication of the client and network through an encrypted tunnel, as well as to generate dynamic, per-user, per-session WEP keys. Unlike EAP-TLS, EAP-TTLS requires only server-side

certificates.
(See also *PEAP (Protected Extensible Authentication Protocol)*.)

**EAPS**

Extreme Automatic Protection Switching is an Extreme Networks-proprietary version of the Ethernet Automatic Protection Switching protocol that prevents looping Layer 2 of the network. This feature is discussed in RFC 3619.

**ECMP**

Equal Cost Multi Paths is a routing algorithm that distributes network traffic across multiple high-bandwidth OSPF, BPG, IS-IS, and static routes to increase performance. The Extreme Networks implementation supports multiple equal cost paths between points and divides traffic evenly among the available paths.

**edge safeguard**

Loop prevention and detection on an edge port configured for RSTP. Configuring edge safeguard on RSTP edge ports can prevent accidental or deliberate misconfigurations (loops) resulting from connecting two edge ports together or from connecting a hub or other non-STP switch to an edge port. Edge safeguard also limits the impact of broadcast storms that might occur on edge ports. This advanced loop prevention mechanism improves network resiliency but does not interfere with the rapid convergence of edge ports.

**EDP**

Extreme Discovery Protocol is a protocol used to gather topology information about neighboring Extreme Networks switches.

**EMS**

The Event Management System is an Extreme Networks-proprietary system that saves, displays, and filters events, which are defined as any occurrences on a switch that generate a log message or require action.

**ERPS**

Ethernet Ring Protection Switching provides fast protection and recovery switching for Ethernet traffic in a ring topology. It also ensures that the Ethernet layer remains loop-free. It is defined in ITU/T G.8032.

**ESRP**

Extreme Standby Router Protocol is an Extreme Networks-proprietary protocol that provides redundant Layer 2 and routing services to users.

**ESRP group**

An ESRP group runs multiple instances of ESRP within the same VLAN (or broadcast domain). To provide redundancy at each tier, use a pair of ESRP switches on the group. See *ESRP (Extreme Standby Router Protocol)*.

**event**

Any type of occurrence on a switch that could generate a log message or require an action. For more, see *syslog*.

**Extreme Defender for IoT**

Extreme Defender for IoT provides unique in-line security for mission critical and/or vulnerable IoT devices. Placed between the IoT device and the network, the Defender for IoT solution helps secure and isolate IoT devices protecting them from internal and external hacking attempts, viruses, malware and ransomware, DDoS attacks, and more. Designed to be simple and flexible, Defender for IoT can be deployed over any network infrastructure to enable secure IoT management without significant network changes.

The solution is comprised of the Extreme Defender Application Software and the Defender Adapter (SA201) or AP3912i access point. Edge Applications Engine is the supported platform for the Extreme Defender Application.

For more information, see https://www.extremenetworks.com/product/extreme-defender-for-iot/.

**Extreme Management Center**

Extreme Management Center (Management Center), formerly Netsight™, is a web-based control interface that provides centralized visibility into your network. Management Center reaches beyond ports, VLANs, and SSIDs and provides detailed control of individual users, applications, and protocols. When coupled with wireless and Identity & Access Management products, Management Center becomes the central location for monitoring and managing all the components in the infrastructure. Learn more about Management Center at http://www.extremenetworks.com/product/management-center/.

**ExtremeAnalytics**

ExtremeAnalytics™, formerly Purview™, is a network powered application analytics and optimization solution that captures and analyzes context-based application traffic to deliver meaningful intelligence about applications, users, locations, and devices. ExtremeAnalytics provides data to show how applications are being used. This can be used to better understand customer behavior on the network, identify the level of user engagement, and assure business application delivery to optimize the user experience. The software also provides visibility into network and application performance allowing IT to pinpoint and resolve performance issues in the infrastructure whether they are caused by the network, application, or server. Learn more about ExtremeAnalytics at http://www.extremenetworks.com/product/extremeanalytics/.

**ExtremeCloud Appliance**

The ExtremeCloud Appliance is a next generation orchestration application offering all the mobility services required for modern unified access deployments. The ExtremeCloud Appliance extends the simplified workflows of the ExtremeCloud public cloud application to on-prem/private cloud deployments.

The ExtremeCloud Appliance includes comprehensive critical network services for wireless and wired connectivity, wireless device secure onboarding, distributed and centralized data paths, role-based access control through the Application Layer, integrated location services, and IoT device onboarding through a single platform.

Built on architecture with the latest technology, the embedded operating system supports application containers that enable future expansion of value added applications for the unified access edge. Learn more about ExtremeCloud Appliance at https://www.extremenetworks.com/product/extremecloud-appliance/.

**ExtremeCloud**

ExtremeCloud is a cloud-based network management Software as a Service (SaaS) tool. ExtremeCloud allows you to manage users, wired and wireless devices, and applications on corporate and guest networks. You can control the user experience with smarter edges – including managing QoS, call admission control, secure access policies, rate limiting, multicast, filtering, and traffic forwarding, all from an intuitive web interface. Learn more about ExtremeCloud at http://www.extremenetworks.com/product/extremecloud/.

**ExtremeCloud™ IQ**

ExtremeCloud™ IQ is an industry-leading and visionary approach to cloud-managed networking, built from the ground up to take full advantage of the Extreme Networks end-to-end networking solutions. ExtremeCloud IQ delivers unified, full-stack management of wireless access points, switches, and routers and enables onboarding, configuration, monitoring, troubleshooting, reporting, and more. Using innovative machine learning and artificial intelligence technologies, ExtremeCloud IQ analyzes and interprets millions of network and user data points, from the network edge to the data center, to power actionable business and IT insights, and deliver new levels of network automation and intelligence. Learn more about ExtremeCloud IQ at https://www.extremenetworks.com/extremecloud-iq/.

**ExtremeControl**

ExtremeControl, formerly Extreme Access Control™ (EAC), is a set of management software tools that use information gathered by a hardware engine to control policy to all devices on the network. The software allows you to automate and secure access for all devices on the network from a central dashboard, making it easier to roll out security and identity policies across the wired and wireless network. Learn more about ExtremeControl at https://www.extremenetworks.com/product/extremecontrol/.

**ExtremeSwitching**

ExtremeSwitching is the family of products comprising different switch types: Modular (X8 and 8000 series [formerly BlackDiamond] and S and K series switches); Stackable (X-series and A, B, C, and 7100 series switches); Standalone (SSA, X430, and D, 200, 800, and ISW series); and Mobile Backhaul (E4G). Learn more about ExtremeSwitching at http://www.extremenetworks.com/products/switching-routing/.

**ExtremeWireless**

ExtremeWireless products and solutions offer high-density WiFi access, connecting your organization with employees, partners, and customers everywhere they go. The family of wireless products and solutions includes APs, wireless appliances, and software. Learn more about ExtremeWireless at http://www.extremenetworks.com/products/wireless/.

**ExtremeXOS**

ExtremeXOS, a modular switch operating system, is designed from the ground up to meet the needs of large cloud and private data centers, service providers, converged enterprise edge networks, and everything in between. Based on a resilient architecture and protocols, ExtremeXOS supports network virtualization and standards-based SDN capabilities like VXLAN gateway and OpenStack Cloud orchestration. ExtremeXOS also supports comprehensive role-based policy. Learn more about ExtremeXOS at http://www.extremenetworks.com/product/extremexos-network-operating-system/.

**FDB**

The switch maintains a database of all MAC address received on all of its ports and uses this information to decide whether a frame should be forwarded or filtered. Each forwarding database (FDB) entry

consists of the MAC address of the sending device, an identifier for the port on which the frame was received, and an identifier for the VLAN to which the device belongs. Frames destined for devices that are not currently in the FDB are flooded to all members of the VLAN. For some types of entries, you configure the time it takes for the specific entry to age out of the FDB.

**FHSS**

Frequency-Hopping Spread Spectrum is a transmission technology used in Local Area Wireless Network (LAWN) transmissions where the data signal is modulated with a narrowband carrier signal that 'hops' in a random but predictable sequence from frequency to frequency as a function of time over a wide band of frequencies. This technique reduces interference. If synchronized properly, a single logical channel is maintained. (Compare with *DSSS (Direct-Sequence Spread Spectrum)*.)

**gratuitous ARP**

When a host sends an ARP request to resolve its own IP address, it is called gratuitous ARP.

**hitless failover**

In the Extreme Networks implementation on modular switches and SummitStacks, hitless failover means that designated configurations survive a change of primacy between the two MSMs (modular switches) or master/backup nodes (SummitStacks) with all details intact. Thus, those features run seamlessly during and after control of the system changes from one MSM or node to another.

**IBSS**

An IBSS is the 802.11 term for an ad hoc network. See *ad hoc mode*.

**ICMP**

Internet Control Message Protocol is the part of the TCP/IP protocol that allows generation of error messages, test packets, and operating messages. For example, the ping command allows you to send ICMP echo messages to a remote IP device to test for connectivity. ICMP also supports traceroute, which identifies intermediate hops between a given source and destination.

**IGMP**

Hosts use Internet Group Management Protocol to inform local routers of their membership in multicast groups. Multicasting allows one computer on the Internet to send content to multiple other computers that have identified themselves as interested in receiving the originating computer's content. When all hosts leave a group, the router no longer forwards packets that arrive for the multicast group.

**LAG**

A Link Aggregation Group is the logical high-bandwidth link that results from grouping multiple network links in link aggregation (or load sharing). You can configure static LAGs or dynamic LAGs (using the LACP).

**LLDP**

Link Layer Discovery Protocol conforms to IEEE 802.1ab and is a neighbor discovery protocol. Each LLDP-enabled device transmits information to its neighbors, including chassis and port identification, system name and description, VLAN names, and other selected networking information. The protocol also specifies timing intervals in order to ensure current information is being transmitted and received.

**MD5**

Message-Digest algorithm is a hash function that is commonly used to generate a 128-bit hash value. It was designed by Ron Rivest in 1991. MD5 is officially defined in RFC 1321 - The MD5 Message-Digest Algorithm.

**MIC**

Message Integrity Check (or Code), also called 'Michael', is part of WPA and TKIP. The MIC is an additional 8-byte code inserted before the standard 4-byte ICV appended in by standard WEP to the 802.11 message. This greatly increases the difficulty in carrying out forgery attacks.
Both integrity check mechanisms are calculated by the receiver and compared against the values sent by the sender in the frame. If the values match, there is assurance that the message has not been tampered with.

**MLAG**

The Multi-switch Link Aggregation Group feature allows users to combine ports on two switches to form a single logical connection to another network device. The other network device can be either a server or a switch that is separately configured with a regular LAG (or appropriate server port teaming) to form the port aggregation.

**MPLS**

Multiprotocol Label Switching speeds up network traffic. When forwarding packets, the Layer 2 (Switching) label is used to avoid complex destination lookups in the routing table. MPLS uses Label Switched Paths (LSPs) to establish the network path. The packet will be labeled so that service providers can decide the best way to keep traffic flowing. The Multiprotocol Label Switching Transport Profile (MPLS-TP) extensions to MPLS are designed to meet service provider requirements and are used as a network layer technology in transport networks. MPLS-TP gives service providers a reliable packet-based technology that is based on circuit-based transport networking. MPLS-TP is expected to be a low cost level 2 technology (if the limited profile is implemented in isolation) that will provide QoS, end-to-end OAM and protection switching.

**MSDP**

Multicast Source Discovery Protocol is used to connect multiple multicast routing domains. MSDP advertises multicast sources across Protocol Independent Multicast-Sparse Mode (PIM-SM) multicast domains or Rendezvous Points (RPs). In turn, these RPs run MSDP over TCP to discover multicast sources in other domains.

**MSTI**

Multiple Spanning Tree Instances control the topology inside an MSTP region. An MSTI is a spanning tree domain that operates within a region and is bounded by that region; and MSTI does not exchange BPDUs or send notifications to other regions. You can map multiple VLANs to an MSTI; however, each VLAN can belong to only one MSTI. You can configure up to 64 MSTIs in an MSTP region.

**MSTP**

Multiple Spanning Tree Protocol, based on IEEE 802.1Q-2003 (formerly known as IEEE 892.1s), allows you to bundle multiple VLANs into one STP topology, which also provides enhanced loop protection and better scaling. MSTP uses RSTP as the converging algorithm and is compatible with legacy STP protocols.

**NetLogin**

Network login provides extra security to the network by assigning addresses only to those users who are properly authenticated. You can use web-based, MAC-based, or IEEE 802.1X-based authentication with network login. The two modes of operation are campus mode and ISP mode.

**netmask**

A netmask is a string of 0s and 1s that mask, or screen out, the network part of an IP address, so that only the host computer part of the address remains. A frequently-used netmask is 255.255.255.0, used for a Class C subnet (one with up to 255 host computers). The ".0" in the netmask allows the specific host computer address to be visible.

**OSPF**

An interior gateway routing protocol for TCP/IP networks, Open Shortest Path First uses a link state routing algorithm that calculates routes for packets based on a number of factors, including least hops, speed of transmission lines, and congestion delays. You can also configure certain cost metrics for the algorithm. This protocol is more efficient and scalable than vector-distance routing protocols. OSPF features include least-cost routing, ECMP routing, and load balancing. Although OSPF requires CPU power and memory space, it results in smaller, less frequent router table updates throughout the network. This protocol is more efficient and scalable than vector-distance routing protocols.

**OSPFv3**

Open Shortest Path First version 3 is one of the routing protocols used with IPV6 and is similar to OSPF.

**PEAP**

Protected Extensible Authentication Protocol is an IETF draft standard to authenticate wireless LAN clients without requiring them to have certificates. In PEAP authentication, first the user authenticates the authentication server, then the authentication server authenticates the user. If the first phase is successful, the user is then authenticated over the SSL tunnel created in phase one using EAP-Generic Token Card (EAP-GTC) or Microsoft Challenged Handshake Protocol Version 2 (MSCHAP V2). (See also *EAP-TLS/EAP-TTLS*.)

**PoE**

The Power over Ethernet standard (IEEE 802.3af) defines how power can be provided to network devices over existing Ethernet connections, eliminating the need for additional external power supplies.

**QoS**

Quality of Service is a technique that is used to manage network resources and guarantee a bandwidth relationship between individual applications or protocols. A communications network transports a multitude of applications and data, including high-quality video and delay-sensitive data such as real-time voice. Networks must provide secure, predictable, measurable, and sometimes guaranteed services. Achieving the required QoS becomes the secret to a successful end-to-end business solution.

**RADIUS**

RADIUS is a client/server protocol and software that enables remote access servers to communicate with a central server to authenticate dial-in users and authorize their access to the requested system or service. RADIUS allows a company to maintain user profiles in a central database that all remote servers can share. It provides better security, allowing a company to set up a policy that can be applied at a single administered network point. With RADIUS, you can track usage for billing and for keeping network statistics.

**rate limiting**

In QoS, rate limiting is the process of restricting traffic to a peak rate (PR).

**rate shaping**

In QoS, rate shaping is the process of reshaping traffic throughput to give preference to higher priority traffic or to buffer traffic until forwarding resources become available.

**RIP**

This IGP vector-distance routing protocol is part of the TCP/IP suite and maintains tables of all known destinations and the number of hops required to reach each. Using Routing Information Protocol, routers periodically exchange entire routing tables. RIP is suitable for use only as an IGP.

**RIPng**

Routing Information Protocol Next Generation is one of the routing protocols used with IPv6 and is similar to RIP.

**segment**

In Ethernet networks, a section of a network that is bounded by bridges, routers, or switches. Dividing a LAN segment into multiple smaller segments is one of the most common ways of increasing available bandwidth on the LAN.

**SNMP**

Simple Network Management Protocol is a standard that uses a common software agent to remotely monitor and set network configuration and runtime parameters. SNMP operates in a multivendor environment, and the agent uses MIBs, which define what information is available from any manageable network device. You can also set traps using SNMP, which send notifications of network events to the system log.

**SNTP**

Simple Network Time Protocol is used to synchronize the system clocks throughout the network. An extension of NTP, SNTP can usually operate with a single server and allows for IPv6 addressing.

**SSL**

Secure Socket Layer is a protocol for transmitting private documents using the Internet. SSL works by using a public key to encrypt data that is transferred over the SSL connection. SSL uses the public-and-private key encryption system, which includes the use of a digital certificate. SSL is used for other applications than SSH, for example, OpenFlow.

**standard mode**

Use ESRP standard mode if your network contains switches running ExtremeWare and switches running ExtremeXOS, both participating in ESRP.

**STP**

Spanning Tree Protocol, defined in IEEE 802.1d, used to eliminate redundant data paths and to increase network efficiency. STP allows a network to have a topology that contains physical loops; it operates in bridges and switches. STP opens certain paths to create a tree topology, thereby preventing packets from looping endlessly on the network. To establish path redundancy, STP creates a tree that spans all of the switches in an extended network, forcing redundant paths into a standby, or blocked, state.

STP allows only one active path at a time between any two network devices (this prevents the loops) but establishes the redundant links as a backup if the initial link should fail. If STP costs change, or if one

network segment in the STP becomes unreachable, the spanning tree algorithm reconfigures the STP topology and re-establishes the link by activating the standby path.

**STPD**

Spanning Tree Domain is an STP instance that contains one or more VLANs. The switch can run multiple STPDs, and each STPD has its own root bridge and active path. In the Extreme Networks implementation of STPD, each domain has a carrier VLAN (for carrying STP information) and one or more protected VLANs (for carrying the data).

**syslog**

A protocol used for the transmission of event notification messages across networks, originally developed on the University of California Berkeley Software Distribution (BSD) TCP/IP system implementations, and now embedded in many other operating systems and networked devices. A device generates a messages, a relay receives and forwards the messages, and a collector (a syslog server) receives the messages without relaying them.
syslog uses the UDP as its underlying transport layer mechanism. The UDP port that has been assigned to syslog is 514. (RFC 3164)

**virtual router**

In the Extreme Networks implementations, virtual routers allow a single physical switch to be split into multiple virtual routers. Each virtual router has its own IP address and maintains a separate logical forwarding table. Each virtual router also serves as a configuration domain. The identity of the virtual router you are working in currently displays in the prompt line of the CLI. The virtual routers discussed in relation to Extreme Networks switches themselves are not the same as the virtual router in VRRP.

In VRRP, the virtual router is identified by a virtual router (VRID) and an IP address. A router running VRRP can participate in one or more virtual routers. The VRRP virtual router spans more than one physical router, which allows multiple routers to provide redundant services to users.

**VLAN**

The term VLAN is used to refer to a collection of devices that communicate as if they are on the same physical LAN. Any set of ports (including all ports on the switch) is considered a VLAN. LAN segments are not restricted by the hardware that physically connects them. The segments are defined by flexible user groups you create with the CLI.

**VMAN**

In ExtremeXOS software, Virtual MANs are a bi-directional virtual data connection that creates a private path through the public network. One VMAN is completely isolated from other VMANs; the encapsulation allows the VMAN traffic to be switched over Layer 2 infrastructure. You implement VMAN using an additional 892.1Q tag and a configurable EtherType; this feature is also known as Q-in-Q switching.

**VR-Control**

This virtual router is part of the embedded system in Extreme Networks switches. VR-Control is used for internal communications between all the modules and subsystems in the switch. It has no ports, and you cannot assign any ports to it. It also cannot be associated with VLANs or routing protocols. (Referred to as VR-1 in earlier ExtremeXOS software versions.)

**VR-Default**

This virtual router is part of the embedded system in Extreme Networks switches. VR-Default is the default VR on the system. All data ports in the switch are assigned to this VR by default; you can add and delete ports from this VR. Likewise, VR-Default contains the default VLAN. Although you cannot delete the default VLAN from VR-Default, you can add and delete any user-created VLANs. One instance of each routing protocol is spawned for this VR, and they cannot be deleted. (Referred to as VR-2 in earlier ExtremeXOS software versions.)

**VR-Mgmt**

This virtual router is part of the embedded system in Extreme Networks switches. VR-Mgmt enables remote management stations to access the switch through Telnet, SSH, or SNMP sessions; and it owns the management port. The management port cannot be deleted from this VR, and no other ports can be added. The Mgmt VLAN is created VR-Mgmt, and it cannot be deleted; you cannot add or delete any other VLANs or any routing protocols to this VR. (Referred to as VR-0 in earlier ExtremeXOS software versions.)

**VRRP**

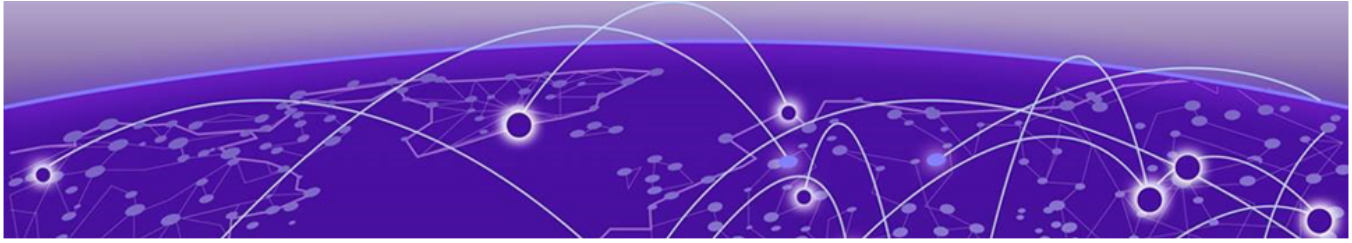
The Virtual Router Redundancy Protocol specifies an election protocol that dynamically assigns responsibility for a virtual router to one of the VRRP routers on a LAN. The VRRP router controlling the IP address(es) associated with a virtual router is called the master router, and forwards packets sent to these IP addresses. The election process provides dynamic failover in the forwarding responsibility should the master router become unavailable. In case the master router fails, the virtual IP address is mapped to a backup router's IP address; this backup becomes the master router. This allows any of the virtual router IP addresses on the LAN to be used as the default first-hop router by end-hosts. The advantage gained from using VRRP is a higher availability default path without requiring configuration of dynamic routing or router discovery protocols on every host. VRRP is defined in RFC 2338.

**WLAN**

Wireless Local Area Network.

**XNV**

Extreme Network Virtualization is an ExtremeXOS feature that enables the software to support VM port movement, port configuration, and inventory on network switches.



# Index

## Specials



## Numerics



## A

## B

## C

## D

## E

# F

## G



## H



## I

## J



## K



## L

## M

# N

## O



## P

# Q



# R

## S

# T

## U

# V

## W



## X