CONFIGURATION GUIDE

Extreme SLX-OS IP Multicast Configuration Guide, 18r.2.00

Supporting the ExtremeRouting SLX 9850 and SLX 9640 and the ExtremeSwitching SLX 9540 Devices

9035755
November 2018

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

Open Source Declarations

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

Contents

Preface

Document conventions

The document conventions describe text formatting conventions, command syntax conventions, and important notice formats used in Extreme technical documentation.

Notes, cautions, and warnings

Notes, cautions, and warning statements may be used in this document. They are listed in the order of increasing severity of potential hazards.

NOTE
A Note provides a tip, guidance, or advice, emphasizes important information, or provides a reference to related information.

ATTENTION
An Attention statement indicates a stronger note, for example, to alert you when traffic might be interrupted or the device might reboot.

CAUTION
A Caution statement alerts you to situations that can be potentially hazardous to you or cause damage to hardware, firmware, software, or data.

DANGER
A Danger statement indicates conditions or situations that can be potentially lethal or extremely hazardous to you. Safety labels are also attached directly to products to warn of these conditions or situations.

Text formatting conventions

Text formatting conventions such as boldface, italic, or Courier font may be used to highlight specific words or phrases.

Format	Description
bold text	Identifies command names.
	Identifies keywords and operands.
	Identifies the names of GUI elements.
	Identifies text to enter in the GUI.
italic text	Identifies emphasis.
	Identifies variables.
	Identifies document titles.
`Courier font`	Identifies CLI output.

Format	Description
	Identifies command syntax examples.

Command syntax conventions

Bold and italic text identify command syntax components. Delimiters and operators define groupings of parameters and their logical relationships.

Convention	Description
bold text	Identifies command names, keywords, and command options.
italic text	Identifies a variable.
[ ]	Syntax components displayed within square brackets are optional. Default responses to system prompts are enclosed in square brackets.
{ **x** | **y** | **z** }	A choice of required parameters is enclosed in curly brackets separated by vertical bars. You must select one of the options.
x | **y**	A vertical bar separates mutually exclusive elements.
< >	Nonprinting characters, for example, passwords, are enclosed in angle brackets.
...	Repeat the previous element, for example, *member*[*member*...].
\	Indicates a "soft" line break in command examples. If a backslash separates two lines of a command input, enter the entire command at the prompt without the backslash.

Extreme resources

Visit the Extreme website to locate related documentation for your product and additional Extreme resources.

White papers, data sheets, and the most recent versions of Extreme software and hardware manuals are available at www.extremenetworks.com. Product documentation for all supported releases is available to registered users at www.extremenetworks.com/support/documentation.

Document feedback

Quality is our first concern at Extreme, and we have made every effort to ensure the accuracy and completeness of this document. However, if you find an error or an omission, or you think that a topic needs further development, we want to hear from you.

You can provide feedback in two ways:

- Use our short online feedback form at https://www.extremenetworks.com/documentation-feedback/.
- Email us at documentation@extremenetworks.com.

Provide the publication title, part number, and as much detail as possible, including the topic heading and page number if applicable, as well as your suggestions for improvement.

Contacting Extreme Technical Support

As an Extreme customer, you can contact Extreme Technical Support using one of the following methods: 24x7 online or by telephone. OEM customers should contact their OEM/solution provider.

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.
- GTAC Knowledge - Get on-demand and tested resolutions from the GTAC Knowledgebase, or create a help case if you need more guidance.
- The Hub - A forum for Extreme customers to connect with one another, get questions answered, share ideas and feedback, and get problems solved. This community is monitored by Extreme Networks employees, but is not intended to replace specific guidance from GTAC.
- Support Portal - Manage cases, downloads, service contracts, product licensing, and training and certifications.

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

About This Document

Supported hardware

In those instances in which procedures or parts of procedures documented here apply to some devices but not to others, this guide identifies exactly which devices are supported by this release and which are not.

Although many different software and hardware configurations are tested and supported by this release, documenting all possible configurations and scenarios is beyond the scope of this document.

The following hardware platforms are supported by this release:

- ExtremeRouting SLX 9850-4 router
- ExtremeRouting SLX 9850-8 router
- ExtremeRouting SLX 9640 router
- ExtremeSwitching SLX 9540 switch

To obtain information about other releases, refer to the documentation specific to that release.

Interface module capabilities

The following table lists the supported capabilities for the following SLX 9850 interface modules:

- BR-SLX9850-10Gx72S-M
- BR-SLX9850-100Gx36CQ-M
- BR-SLX9850-10Gx72S-D
- BR-SLX9850-100Gx36CQ-D
- BR-SLX9850-100Gx12CQ-M

TABLE 1 SLX 9850 interface modules capabilities

Capability	Modular interface module
MPLS	Yes
Packet buffer memory per interface module	12GB (BR-SLX9850-10Gx72S-M) 36GB (BR-SLX9850-100Gx36CQ-M) 8GB (BR-SLX9850-10Gx72S-D) 24GB (BR-SLX9850-100Gx36CQ-D) 8GB (BR-SLX9850-100Gx12CQ-M)

What's new in this document

The following table includes description of changes in functionality for the current release.

TABLE 2 Changes for the current release

Feature	Description	Described in
IGMP snooping restrict unknown multicast	Restrict unknown multicast avoids flooding of unknown multicast traffic on the member ports of VLAN.	IGMP snooping restrict unknown multicast on page 17

NOTE

On October 30, 2017, Extreme Networks, Inc. acquired the SLX-OS product line from Brocade Communications Systems, Inc. This transitional release includes references to both companies.

IP Multicast Overview

IP multicast overview

Multicast protocols allow a group or channel to be accessed over different networks by multiple stations (clients) for the receipt and transmission of multicast data. Distribution of stock quotes, video transmissions such as news services and remote classrooms, and video conferencing are all examples of applications that use multicast routing.

Extreme devices support the Protocol-Independent Multicast (PIM) protocol, along with the Internet Group Management Protocol (IGMP).

The Internet Group Management Protocol (IGMP) is used by IP hosts to report their multicast group memberships to any immediately-neighboring multicast routers.

PIM is a broadcast and pruning multicast protocol that delivers IP multicast datagrams. This protocol employs reverse path lookup check and pruning to allow source-specific multicast delivery trees to reach all group members. PIM builds a different multicast tree for each source and destination host group.

IPv4 Multicast Traffic Reduction

IGMP snooping overview

The forwarding of multicast control packets and data through a Layer 2 device configured with VLANs is most easily achieved by the Layer 2 forwarding of received multicast packets on all the member ports of the VLAN interfaces. However, this simple approach is not bandwidth efficient, because only a subset of member ports may be connected to devices interested in receiving those multicast packets. In a worst-case scenario, the data would get forwarded to all port members of a VLAN with a large number of member ports, even if only a single VLAN member is interested in receiving the data. Such scenarios can lead to loss of throughput for a device that gets hit by a high rate of multicast data traffic.

Internet Group Management Protocol (IGMP) snooping is a mechanism by which a Layer 2 device can effectively address this issue of inefficient multicast forwarding to VLAN port members. Snooping involves "learning" forwarding states for multicast data traffic on VLAN port members from the IGMP control (join/leave) packets received on them. The Layer 2 device also provides for a way to configure forwarding states statically through the CLI.

Multicast routing and IGMP snooping

Multicast routers use IGMP snooping to learn which groups have members on each of their attached physical networks. A multicast router keeps a list of multicast group memberships for each attached network, and a timer for each membership.

NOTE

"Multicast group memberships" means that at least one member of a multicast group on a given attached network is available.

There are two ways that hosts join multicast routing groups:

- By sending an unsolicited IGMP join request.
- By sending an IGMP join request as a response to a general query from a multicast router.

In response to the request, the device creates an entry in its Layer 2 forwarding table for that VLAN. When other hosts send join requests for the same multicast, the device adds them to the existing table entry. Only one entry is created per VLAN in the Layer 2 forwarding table for each multicast group.

VLANs can be configured as snooping only or routing with snooping. When Layer 3 multicast routing is enabled on a particular VE, snooping for the underlying VLAN is enabled implicitly. Explicit snooping can be enabled on a VLAN in addition to implicit snooping. Implicit snooping is by default IGMP snooping. With routing enabled on a VE, when explicit snooping is disabled, snooping reverts back to implicit snooping. This does not change the functionality in any way, but only removes the configuration. When routing is disabled on a VE where explicit snooping is configured, the routing side of the programming stops and the snooping side programming takes over.

When routing is enabled, the Layer 3 IGMP querier takes precedence on that VLAN. When routing is disabled, and if the snooping querier is configured, then the snooping querier takes effect.

PIM multicast router presence detection

The PIM hello-based multicast router presence detection feature scans the network traffic for incoming PIM hellos.

This feature is enabled when multicast routing or snooping is enabled.

When a PIM hello is detected, that port is marked for the presence of a multicast router and the information is saved. This prevents unnecessary flooding if the PIM designated router (DR) goes offline, as IGMP reports are forwarded to the multicast routers and not only the snooping-enabled router.

Enabling IGMP snooping

Use the following procedure to enable IGMP snooping on a VLAN.

1. Enter the **configure terminal** command to access global configuration mode.

```
device# configure terminal
```

2. Enter the VLAN configuration mode.

```
device(config)# vlan 1
device(config-vlan-1)
```

3. Enable IGMP snooping.

```
device(config-vlan-1)# ip igmp snooping enable
```

Configuring the IGMP snooping querier

If your multicast traffic is not routed because Protocol-Independent Multicast (PIM) is not configured, use the IGMP snooping querier in a VLAN.

The IGMP snooping querier sends out IGMP queries to trigger IGMP responses from devices that are to receive IP multicast traffic. The IGMP snooping querier listens for these responses to map the appropriate forwarding addresses.

NOTE
Snooping querier is suspended if Layer 3 IGMP is enabled on any of the cluster nodes.

Use the following procedure to configure the IGMP snooping querier.

1. Enter the **configure terminal** command to access global configuration mode.

```
device# configure terminal
```

2. Enter the **vlan** command with the VLAN number.

```
device(config)# interface vlan 25
```

3. Set the IGMP query interval for the VLAN.

```
device(config-Vlan-25)# ip igmp snooping query-interval 125
```

   The valid range is from 1 through 18000 seconds. The default is 125 seconds.

4. Set the last member query interval.

```
device(config-Vlan-25)# ip igmp snooping last-member-query-interval 1000
```

   The valid range is from 1000 through 25500 milliseconds. The default is 1000 milliseconds.

5. Configure the static Mrouter port.

```
device(config-Vlan-25)# ip igmp snooping mrouter interface ethernet 3/2
```

6. Configure a static IGMP group.

```
device(config-vlan-25)# ip igmp snooping static-group 225.0.0.1 interface ethernet 6/15
```

7. Configure the IGMP version.

```
device(config-vlan-25)# ip igmp snooping version v3
```

   NOTE
   Version 2 is enabled by default. When you change the version of IGMP snooping any existing static or dynamic group will get deleted. These groups will are relearnt at the next query interval when the query is sent out.

8. Set the snooping robustness variable.

```
device(config-Vlan-25)# ip igmp snooping robustness-variable 5
```

   The valid range is from 2 through 10. The default is 2.

9. Activate the IGMP snooping querier functionality for the VLAN.

```
device(config-Vlan-25)# ip igmp snooping querier enable
```

   NOTE
   The IGMP snooping querier and the static mrouter can be configured together on a VLAN interface.

Monitoring IGMP snooping

Monitoring the performance of your IGMP traffic allows you to diagnose any potential issues on your device. This helps you utilize bandwidth more efficiently by setting the device to forward IP multicast traffic only to connected hosts that request multicast traffic.

Use the following commands to monitor IGMP snooping on the device; the commands do not need to be entered in any specific order.

1. Enter the **show ip igmp groups** command to display all information on IGMP multicast groups for the device. Use this command to display the IGMP database, including configured entries for all groups on all interfaces, all groups on specific interfaces, or specific groups on specific interfaces.

```
device# show ip igmp groups
Total Number of Groups: 2
IGMP Connected Group Membership
Group Address  Interface Uptime      Expires      Last Reporter   Version
225.1.1.1      vlan25    00:05:27     00:02:32     25.1.1.1202
Member Ports: eth 2/24
```

2. Enter the **show ip igmp snooping** command specifying the VLAN ID to view snooping configuration information such as snooping querier enable, snooping query interval, IGMP operation mode, PIM snooping configuration, and IGMP snooping configuration.

```
device# show ip igmp snooping

 Vlan ID: 10
 Multicast Router ports:  eth1/1
 Querier - Disabled
 IGMP Operation mode: IGMPv3
 Is Fast-Leave Enabled : Enabled
 Max Response time = 10
 Last Member Query Interval = 1
 Query interval = 125
 Number of Multicast Groups: 0
```

3. Enter the **show ip multicast snooping mcache** command to view snooping configuration and PIM snooping configuration information.

```
device# show ip multicast snooping mcache
Flags : V2|V3 : IGMP Receiver, P_G : PIM (*,G) Join, P_SG: PIM (S,G) Join
                BR : PIM Blocked RPT
 Vlan ID : 10
 -------------
1     (20.20.20.20, 232.0.0.10 )  22:37:48    NumOIF: 1
       Outgoing Ports:
               eth1/34        Flags: 0x24 ( V3)  00:00:08/252s
```

4. Enter the **show ip igmp statistics interface** command to display the IGMP statistics for a VLAN or interface.

```
device# show ip igmp statistics interface vlan 1

IGMP packet statistics for all interfaces in vlan 1:
IGMP Message type     Edge-Received   Edge-Sent   Edge-Rx-Errors   ISL Received
Membership Query                  0           0                0              0
V1 Membership Report              0           0                0              0
V2 Membership Report              0           0                0              0
Group Leave                       0           0                0              0
V3 Membership Report              0           0                0              0
PIM hello                         0           0                0              0

IGMP Error Statistics:
Unknown types            0
Bad Length               0
Bad Checksum             0
```

5. Enter the **show ip igmp interface** command to display the Layer 3 IGMP interface configuration information.

```
device# show ip igmp interface
Interface Ve100
IGMP enabled
IGMP query interval 30 seconds
IGMP other-querierinterval 65 seconds
IGMP query response time 10 seconds
IGMP last-member query interval 1 seconds
IGMP immediate-leave disabled
IGMP querier100.0.0.1(this system)
IGMP version 2
```

6. Enter the **show ip igmp snooping mrouter vlan** command to display mrouter port-related information.

```
device# show ip igmp snooping mrouter vlan 10
Vlan     Interface    Expires (Sec)
10       eth1/4       250
10       eth1/1       238
```

7. Enter the **show ip igmp ssm-map** command to display the SSM mapping with the prefix list name and source address details.

```
device# show ip igmp ssm-map

+-------------------------------------+-------------------+
|         PrefixList Name             |   Source Address  |
+-------------------------------------+-------------------+
   ssm-map-230-to-232                  203.0.0.10
  ssm-map-233-to-234                  204.0.0.11
```

IGMP snooping restrict unknown multicast

Restrict unknown multicast avoids flooding of unknown multicast traffic on the member ports of VLAN.

IGMP snooping floods the unknown multicast data packets on the member ports of VLAN. The unknown multicast data traffic is the data traffic sent to multicast groups which are not learnt by means of IGMP membership reports/static IGMP group configuration.

Restricting the unknown multicast is achieved by redirecting the unknown multicast traffic (traffic destined to multicast MAC 01:00:5E:XX:XX:XX) to MG ID which has no replication ports. As there are no ports to send the traffic, the unknown multicast traffic is dropped.

Multicast traffic forwarding behavior with restrict unknown multicast

Different scenarios of multicast data traffic forwarding with restrict unknown multicast feature enabled.

Multicast traffic forwarding learnt via IGMP reports/static groups

IGMP v2 groups learnt via IGMP reports/static configuration are programmed in LEM table as (*,G,V) entries.

IGMP v3 groups learnt via IGMP report messages are programmed in KAPS as (S,G,V) entries.

The multicast data traffic is forwarded based on a match from LEM/KAPS table which provides a MG ID. The multicast traffic is replicated to the MG ID having egress ports.

Unknown multicast traffic forwarding

All the unknown multicast data traffic is restricted/dropped instead of flooding onto the VLAN members.

Mrouter port behavior

Mrouter ports which are learnt by means of receiving IGMP queries/PIM hello messages/static configuration are added as part of the restrict unknown multicast MG ID. With this all the unknown multicast data traffic is flooded only to the Mrouter ports of the VLAN.

Enable restrict unknown multicast

These are the platform dependent changes required for supporting this feature.

When restrict unknown multicast is enabled the following ACL entry is installed in control protocol classifier DB 10 for redirecting the unknown multicast to a MG ID.

If Mrouter ports are not present then the unknown multicast traffic is dropped by MG ID as there no ports.

If Mrouter ports are present then the traffic is forwarded to the MG ID with Mrouter ports.

Database	How to identify	Classification method	Action	Per port, VLAN or chip
10 (pmf_grp_id_l3V4CtrlProtoA)	(01-00-5E-XX-XX-XX) && (VSI_match) && (L2DestHit ==0)	PMF	Redirect to MG ID	Per Pure L2 VLAN

Restrict Unknown Multicast message flow

The message flow between modules during restrict unknow multicast configuration.

1. MC_HMS - When **ip igmp snooping restrict-unknown-multicast** is enabled/disabled, MC_HMS sends a **no_flood_enable / no_flood_disable** message to MCAST-SS.
2. MCAST-SS - Currently, MCAST-SS allocates MG ID for each VLAN. When **no_flood_enable** message is received from MC_HMS. The MG ID is allocated from IGMP group MG ID database. MCAST_IGMP_MRT_X_X_V route add message is posted from MCAST-SS to MCAGT for further processing. MCAST_IGMP_MRT_X_X_V route delete message is posted from MCAST-SS to MCAGT incase of **no_flood_disable/snoop_disable** message is received.
3. MCAGT - On receiving MCAST_IGMP_MRT_X_X_V route add/delete message, MCAGT updates the local VLAN database about the restrict unknown multicast status. MCAGT sends MC_DNLD_MSG_RESTRICT_UNK_MCAST message to HSLUA by filling up the MG ID, VLAN, AFI, enable/disable status of restrict unknown multicast details.
4. HSLUA - On receiving MC_DNLD_MSG_RESTRICT_UNK_MCAST enable message, the ACL entry in Hardware for corresponding VSI/VLAN ID. The ACL entry id is stored in HSLUA-MCAST VLAN database for future references. On receiving MC_DNLD_MSG_RESTRICT_UNK_MCAST disable message, the ACL entry is fetched from VLAN database and deleted from hardware. Currently, the restrict unknown multicast ACLs are programmed in the control protocol qualifier database 10.

TCAM profiles to support this feature

Restrict unknown multicast is supported in default and IPv6 multicast profiles.

Limitations

- The MG IDs used for redirecting the unknown multicast traffic is allocated from IGMP MG ID space. The maximum scale number for IGMP groups without restrict unknown multicast is 16384.
- If we enable restrict unknown multicast on a VLAN/BD, the IGMP groups maximum scaling number get reduced by 1.

> NOTE
> Maximum of 15872 IGMP groups can be learnt, if restrict unknown multicast is enabled on 512 VLANs/BDs.

PIM SM traffic snooping

By default, when a Extreme device receives an IP multicast packet, the device does not examine the multicast information in the packet. Instead, the device simply forwards the packet out all ports except the port that received the packet. In some networks, this method can cause unnecessary traffic overhead in the network. For example, if the Extreme device is attached to only one group source and two group receivers, but has devices attached to every port, the device forwards group traffic out all ports in the same broadcast domain except the port attached to the source, even though there are only two receivers for the group.

PIM SM traffic snooping eliminates the superfluous traffic by configuring the device to forward IP multicast group traffic only on the ports that are attached to receivers for the group.

PIM SM traffic snooping requires IP multicast traffic reduction to be enabled on the device. IP multicast traffic reduction configures the device to listen for IGMP messages. PIM SM traffic snooping provides a finer level of multicast traffic control by configuring the device to listen specifically for PIM SM join and prune messages sent from one PIM SM router to another through the device.

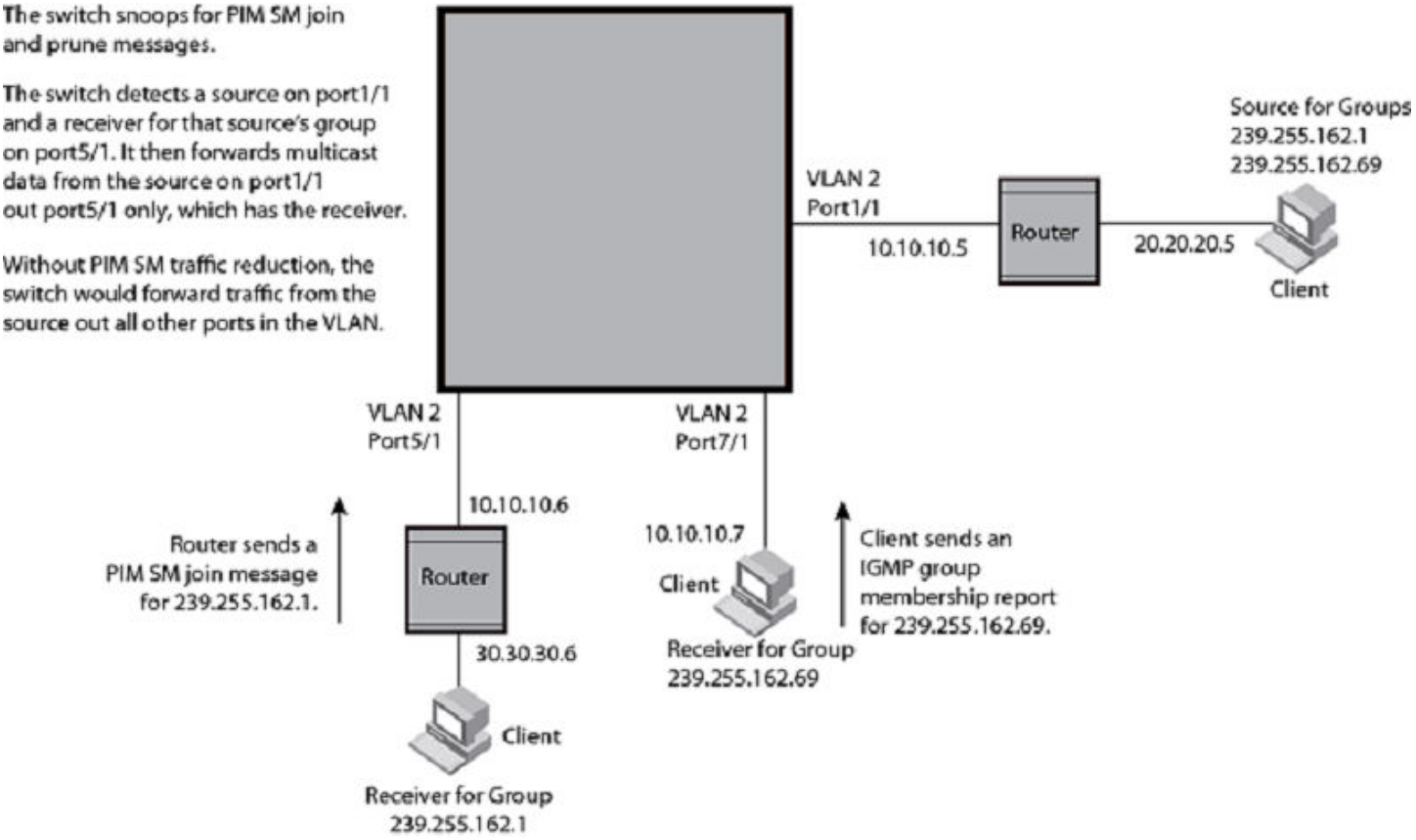
> NOTE
> This feature applies only to PIM SM version 2 (PIM V2).

Application examples

The figure below shows an example application of the PIM SM traffic snooping feature.

In this example, a device is connected through an IP router to a PIM SM group source that is sending traffic for a multicast group. The device also is connected to a receiver for the group.

FIGURE 1 PIM SM traffic reduction in an enterprise network



When PIM SM traffic snooping is enabled, the device starts listening for PIM SM join and prune messages and IGMP group membership reports. Until the device receives a PIM SM join message or an IGMP group membership report, the device forwards IP multicast traffic out on all ports. Once the device receives a join message or group membership report for a group, the device forwards subsequent traffic for that group only on the ports from which the join messages or IGMP reports were received.

In this example, the router connected to the receiver for group 239.255.162.1 sends a join message toward the group's source. Because PIM SM traffic snooping is enabled on the device, the device examines the join message to learn the group ID, then makes a forwarding entry for the group ID and the port connected to the receiver's router. The next time the device receives traffic for 239.255.162.1 from the group's source, the device forwards the traffic only on port 5/1, since that is the only port connected to a receiver for the group.

Supported configurations

- Standalone PIM snooping is not supported.
- IGMP snooping must be enabled in order to enable PIM snooping on the VLAN.

- Enabling Layer 3 multicast on a VE interface implicitly enables IGMP snooping and PIM snooping by default.
- PIM snooping gets disabled upon disabling IGMP snooping.
- A global PIM snooping CLI is not available.
- Disabling Layer 3 multicast will disable IGMP and PIM snooping unless configured by the user.
- You cannot disable IGMP or PIM snooping when Layer 3 multicast is configured.
- Clearing the IP IGMP groups clears the whole snooping database.

Assumptions and dependencies

- IPv6 PIM snooping is not supported.

Considerations for PIM snooping in SSM range

SSM is required only at the last-hop router (LHR) and is not required for the intermediate switch or router. At the LHR, if the SSM map is enabled in the IGMP, the v2 report falling in the SSM group range will be converted to (S,G) join and sent to PIM. If a host sends a v3 report falling in the same SSM range, and the router is V3 enabled with an SSM map configured on this router, then this report should be dropped. From PIM snooping perspective, the (S,G) join will be only created in the software.

High availability considerations

PIM snooping will sync its database with the standby dynamically as well as during bulk sync. Once the standby becomes active, it will use the synced database.

Enabling PIM snooping

You can enabled PIM snooping on a VLAN.

1. Enter global configuration mode.

```
device# configure terminal
```

2. Enter VLAN configuration mode.

```
device(config)# vlan 1
```

3. Enable PIM snooping.

```
device(config-vlan-1)# ip pim snooping enable
```

The following example enables PIM on a VLAN.

```
device# configure terminal
device(config)# vlan 1
device(config-vlan-1)# ip pim snooping enable
```

IPv4 Multicast Routing

IGMP

The Internet Group Management Protocol (IGMP) allows an IPv4 system to communicate IP multicast group membership information to its neighboring routers. The routers, in turn, limit the multicast of IP packets with multicast destination addresses to only those interfaces on the router that are identified as IP multicast group members.

In IGMPv2, when a router sends a query to the interfaces, the clients on the interfaces respond with a membership report of multicast groups to the router. The router can then send traffic to these groups, regardless of the traffic source. When an interface no longer needs to receive traffic from a group, it sends a leave message to the router, which in turn sends a group-specific query to that interface to see if any other clients on the same interface are still active.

In contrast, IGMPv3 provides selective filtering of traffic based on the traffic source. A router running IGMPv3 sends queries to every multicast-enabled interface at the specified interval. These general queries determine if any interface wants to receive traffic from the router.

There are different types of query messages:

- A "General Query" is sent by a multicast router to learn the complete multicast reception state of the neighboring interfaces. In a General Query, both the Group Address field and the Number of Sources (N) field are zero.
- A "Group-Specific Query" is sent by a multicast router to learn the reception state, with respect to a *single* multicast address, of the neighboring interfaces. In a Group-Specific Query, the Group Address field contains the multicast address of interest, and the Number of Sources (N) field contains zero.
- A "Group-and-Source-Specific Query" is sent by a multicast router to learn if any neighboring interface desires reception of packets sent to a specified multicast address, from any of a specified list of sources. In a Group-and-Source-Specific Query, the

Group Address field contains the multicast address of interest, and the Source Address [i] fields contain the source addresses of interest.

The interfaces respond to these queries by sending a membership report that contains one or more of the following records that are associated with a specific group:

- The current-state record indicates from which sources the interface wants to receive and not receive traffic. The record contains the source address of the interfaces and whether or not traffic will be received or included (IS_IN) or not received or excluded (IS_EX) from that source.
- The filter-mode-change record indicates that if the interface changes its current state from IS_IN to IS_EX, a TO_EX record is included in the membership report. Likewise, if the interface changes its current status from IS_EX to IS_IN, a TO_IN record appears in the membership report.
- The IGMPv2 Leave report is equivalent to a TO_IN (empty) record in IGMPv3. This record indicates that no traffic from this group will be received regardless of the source.
- The IGMPv2 group report is equivalent to an IS_EX (empty) record in IGMPv3. This record indicates that all traffic from this group will be received regardless of the source.
- The source-list-change record indicates that If the interface wants to add or remove traffic sources from its membership report, the membership report can have an ALLOW record, which contains a list of new sources from which the interface wishes to receive traffic. It can also contain a BLOCK record, which lists current traffic sources from which the interface wants to stop receiving traffic.

In response to membership reports from the interfaces, the router sends a Group-Specific Query or a Group-and-Source Specific Query to the multicast interfaces. For example, a router receives a membership report with a source-list-change record to block old sources from an interface. The router sends Group-and-Source Specific Queries to the source-group pair (S,G) identified in the record. If none of the interfaces is interested in the (S,G), it is removed from the (S,G) list for that interface on the router.

Each IGMPv3-enabled router maintains a record of the state of each group and each physical port within a virtual routing interface. This record contains the group, group-timer, filter mode, and source records information for the group or interface. Source records contain information on the source address of the packet and source timer. If the source timer expires when the state of the group or interface is in include mode, the record is removed.

Configuring IGMP SSM-MAP

To configure the IGMP SSM-MAP, follow the below procedure.

1. Enter the **configure terminal** command to access global configuration mode.

```
device# configure terminal
```

2. Enter the **vlan** command with the VLAN number.

```
device (config)# interface vlan 101
```

3. Set the IGMP query interval for the VLAN.

```
device (config-vlan-101)# ip igmp snooping query-interval 101
```

4. Configure prefix-list.

```
device (config)# ip igmp ssm-map ?
Possible completions:
               <Word:1-32>   IP prefix-list name
               enable        Enables IGMPv2 SSM Mapping
```

5. Configure ssm-map which converts associates igmpv1 or igmpv2 report packet with the configured source address 203.0.0.1.

```
device (config-vlan-101)# ip igmp ssm-map enab
device (config-vlan-101)# ip igmp ssm-map prefix-list1 203.0.0.1
device (config-vlan-101)# show ip igmp ssm-map
Fri Jul  21 11:30:06.878 UTC-07:00
+-----------------------------------+-------------------+
|         PrefixList Name           |   Source Address  |
+-----------------------------------+-------------------+
  prefix-list1                        203.0.0.1
```

6. Specify 238.0.0.0/8 as SSM group range.

```
device (config-router-pim-vrf-default-vrf)# router pim
device (config-router-pim-vrf-default-vrf)# ssm-enable
device (config-router-pim-vrf-default-vrf)# ssm-enable range prefix-list1
device# show ip pim settings
Fri Jul  21 11:31:45.333 UTC-07:00
vrf : default-vrf
  Maximum mcache              : 32768     Current Count               : 0
  Hello interval              : 30        Neighbor timeout            : 105
  Join/Prune interval         : 60        Inactivity interval         : 180
  Hardware drop enabled       : 1         Prune wait interval         : 3
  Register Suppress Time      : 60        Register Probe Time         : 10
  Register Stop Delay         : 0         Register Suppress interval  : 0
  SSM Enabled                 : Yes       SPT Threshold               : 1
  SSM Group Range             : 232.0.0.0/8
  SSM Range Prfix_name        : prefix-list1
  Route Precedence            : uc-non-default uc-default none
```

7. Ixia sends igmpv2 report for group 238.0.0.1.

```
device # sh ip igmp group
Total Number of Groups: 1
IGMP Connected Group Membership
Group Address    Interface            Uptime      Expires    Last Reporter     Version
238.0.0.1        vlan101              00:04:40    00:03:31   101.0.0.10        3
  Member Ports:  eth7/2

device # show ip igmp group detail
Group : 238.0.0.1
        Interface        vlan101
        Uptime           00:04:53
        Expires:         00:03:18
        Last Reporter:   101.0.0.10
        Member Ports:    eth7/2
        Last Reporter Mode: 3
        Interface : eth7/2
                INCL_SRC_LIST: 203.0.0.1
                EXCL_SRC_LIST: Nil

device # sh ip pim mc
Fri Jul  21 11:58:17.278 UTC-07:00
Total entries in mcache: 1
1    (203.0.0.1, 238.0.0.1) in Eth 7/16, Uptime 00:05:23
     SSM=1, RPT=0 SPT=1 Reg=0 RegSupp=0 RegProbe=0 JDUp=1 LSrc=0 LRcv=1
     upstream neighbor=91.0.0.2
     AgeSltMsk: 0 KAT timer: Expired
     num_oifs = 1
          Ve101(00:05:23/0) Flags: MI
     Flags (0x080684d4)
         ssm=1 needRte=0
```

Default IGMP version

IGMP v2 is enabled by default only when snooping or multicast routing are enabled on the system.

Also, you can specify what version of IGMP you want to run on a device on a per-VLAN basis. You can change the IGMP version for router ports, but not for Ve interfaces. If you do not specify an IGMP version, IGMPv2 is used.

Compatibility with IGMPv1 and IGMPv2

Different multicast groups, interfaces, and routers can run their own versions of IGMP. The version of IGMP is reflected in the membership reports that the hosts send to the router. Routers and interfaces must be configured to recognize the version of IGMP you want them to process.

An interface or router sends the queries and reports that include its IGMP version specified on it. The interface may recognize a query or report that has a different version. For example, an interface running IGMPv2 can recognize IGMPv3 packets, but cannot process them. When the router sends out IGMP queries over an IGMPv2 interface, the equal or lower version of reports are supported, but higher version of reports are not supported.

Reports sent by interfaces to routers that contain different versions of IGMP do not trigger warning messages; however, you can see the versions of the packets by using the **show ip igmp traffic** command.

The version of IGMP can be specified per interface (physical port or virtual routing interface), and per physical port within a virtual routing interface.

The IGMP version set on a Layer 3 physical interface or under a VLAN of the virtual routing interface supersedes the version set on a physical or virtual routing interface.

Likewise, the version on a physical or virtual routing interface supersedes the version set globally on the device.

Enabling the IGMP version

You can enable or change the IGMP version per interface or VLAN setting.

1. Enter global configuration mode.

```
device# configure terminal
```

2. Enter the interface configuration mode.

```
device(config)# interface ethernet 1/5
```

3. Enter the **ip igmp version** command.

```
device(config-if-1/5)# ip igmp version 3
```

Configuring RA option disable

RA (router alert) option disable can be configured at the global level.

The router alert disable option disables the snooping check for the presence of the router alert option. By default, IGMP snooping checks for the presence of the router alert option in the IP packet header of the IGMP message. Packets that do not include this option are dropped.

1. Enter the global configuration mode.

```
device# configure terminal
```

2. To disable the RA option, enter the **router-alert-check-disable** command.

```
device(config)# ip igmp router-alert-check-disable
```

Support for multicast multi VRF

Multi-VRF enables Multiple VPN routing instances and supports IP Multicast.

With multi VRF support all L3 multicast protocols operate as separate instances per VRF depending on the VRF specific multicast configuration. All the required configuration and mcast routing table will have multiple instances per VRF and function simultaneously allowing network paths to be segmented without using multiple routers.

The purpose of multi VRF is to support multiple instances of L3 multicast protocols on the same router at the same timeframe.

Configuring multicast over multi VRF for IPv4

To configure virtual routing and forwarding instances, complete the below procedure.

1. Enter global configuration.

```
device# configure terminal
```

2. Enter the **router pim vrf <vrf name>** command to enter the PIM router configuration mode and configure a variety of options.

```
device(config)# router pim vrf red
device(config-router-pim-vrf-red)#
```

3. Enter the **rp-address** command followed by the IP address to be configured as the RP for the PIM Sparse domain.

```
device(config-router-pim-vrf-red)# rp-address 100.1.1.1
```

4. Enter the **anycast-rp** command followed by the RP address and the anycast-rp-set parameter, which specifies a host based simple prefix list name used to specify the address of the Anycast RP set, including a local address.

```
device(config-router-pim-vrf-red)# anycast-rp 100.1.1.1 anycast-rp-set
```

5. Enter the **rp-address**command to specify the IP address of the RP.

```
device(config-router-pim-vrf-red)# rp-address 4.4.4.4
```

6. Enter the **bsr-candidate** command to configure the BSR candidate.

```
device(config-router-pim-vrf-red)# bsr-candidate interface loopback 11 mask 32
```

7. Enter the **hello-interval** command to configure the PIM hello timeout.

```
device(config-router-pim-vrf-red)# hello-interval 40
```

8. Enter the **message-interval** command to configure the PIM join or prune interval.

```
device(config-router-pim-vrf-red)# message-interval 180
```

9. Enter the **nbr-timeout** command to configure the PIM neighbor timeout.

```
device(config-router-pim-vrf-red)# nbr-timeout 160
```

10. Enter the **prune-wait** command to configure the PIM prune pending timeout.

```
device(config-router-pim-vrf-red)# prune-wait 5
```

11. For static RP configuration with specific group ranges, enter the following commands.

```
device(config-router-pim-vrf-red)# rp-address 4.4.4.4 static-rp-list
device(config)# ip prefix-list static-rp-list permit 225.1.1.0/24
```

    The following commands configure the RP candidate.

```
device(config-router-pim-vrf-red)# rp-candidate interface loopback 11
device(config-router-pim-vrf-red)# rp-candidate prefix my-rp-cand-list
device(config)# ip prefix-list my-rp-cand-list permit 226.1.1.0/24
device(config)# ip prefix-list my-rp-cand-list permit 228.1.1.0/24
```

12. Enter the **rpf ecmp rebalance** command to enable load sharing with dynamic rebalance.

```
device(config-router-pim-vrf-red)# rpf ecmp rebalance
```

13. Enter the **ssm-enable** range command to set the multicast address range to use for SSM.

```
device(config-router-pim-vrf-red)#ssm-enable range PL_ssm_range-230-to-234
```

14. Enter the **prune-wait** command to configure the PIM prune pending timeout.

```
device(config-router-pim-vrf-red)# prune-wait 5
```

IGMPv2 SSM mapping

The PIM-SSM feature requires all IGMP hosts to send IGMPv3 reports. Where you have an IGMPv2 host, this can create a compatibility problem. In particular, the reports from an IGMPv2 host contain a Group Multicast Address but do not contain source addresses. The IGMPv3 reports contain both the Group Multicast Address and one or more source addresses. This feature converts IGMPv2 reports into IGMPv3 reports through use of the ip igmp ssm-map commands and a properly configured prefix list.

The following sections describe how to configure the ACL and the **ip igmp ssm-map** commands to use the IGMPv2 SSM mapping feature:

- Configuring an ACL for IGMPv2 SSM mapping
- Configuring the IGMPv2 SSM Mapping Commands

Configuring IGMPv2 SSM mapping

The **ip ssm-map** commands can be used to enable the IGMPv2 mapping feature and to define the maps between IGMPv2 group addresses and multicast source addresses.

The PIM-SSM feature requires all IGMP hosts to send IGMPv3 reports. Where you have an IGMPv2 host, this can create a compatibility problem. In particular, the reports from an IGMPv2 host contain a group multicast address but do not contain source addresses. The IGMPv3 reports contain both the group multicast address and one or more source addresses. This feature converts IGMPv2 reports into IGMPv3 reports through use of the ip igmp ssm-map commands and a configured prefix list.

The prefix list used with this feature filters for the group multicast address. The prefix list is then associated with one or more source addresses. When the **ip igmp ssm-map enable** command is configured, IGMPv3 reports are sent for IGMPv2 hosts.

1. Enter global configuration mode.

```
device# configure terminal
```

2. Enter the **ip igmp ssm-map enable** command to enable the IGMPv2 mapping.

```
device(config)# ip igmp ssm-map enable
```

The following example configures the SSM map at the global configuration level.

```
device(config)# ip igmp ssm-map enable
device(config)# ip igmp ssm-map ssm-map-230-to-232 203.0.0.10
device(config)# ip igmp ssm-map ssm-map-233-to-234 204.0.0.10
```

The following example configures the prefix list for an SSM range.

```
device(config)# ip prefix-list ssm-map-230-to-232 seq 5 permit 230.0.0.0/8
device(config)# ip prefix-list ssm-map-230-to-232 seq 10 permit 231.0.0.0/8
device(config)# ip prefix-list ssm-map-230-to-232 seq 15 permit 232.0.0.0/8

device(config)# ip prefix-list ssm-map-233-to-234 seq 5 permit 233.0.0.0/8
device(config)# ip prefix-list ssm-map-233-to-234 seq 10 permit 234.0.0.0/8
device(config)# ip prefix-list ssm-map-230-to-232 seq 15 permit 232.0.0.0/8
```

PIM-sparse overview

PIM-sparse is most effective in large networks sparsely populated with hosts interested in multicast traffic, with most hosts not interested in all multicast data streams.

PIM-sparse devices are organized into domains. A PIM-sparse domain is a contiguous set of devices that all implement PIM and are configured to operate within a common boundary.

PIM-sparse creates unidirectional shared trees that are rooted at a common node in the network called the rendezvous point (RP). The RP acts as the messenger between the source and the interested hosts or routers. There are various ways of identifying an RP within a network. An RP can be configured either statically per PIM router, or by means of a bootstrap router (BSR). Within a network, the RP must always be upstream from the destination hosts.

Once the RP is identified, interested hosts and routers send join messages to the RP for the group in which they are interested. To reduce the number of Join messages incoming to an RP, the local network selects one of its upstream routers as the designated router (DR). All hosts below a DR send IGMP join messages to the DR. The DR sends only one join message to the RP on behalf of all its interested hosts.

PIM-sparse also provides the option of creating a source-based tree rooted at a router adjacent to the tree. This provides the destination hosts with an option of switching from the shared tree to the source-based tree if the latter has a shorter path between the source and the destination.

Bootstrap Router Protocol

For PIM Sparse Mode to function, every PIM router must know the RP in the network, so that it can map multicast groups to the available RP addresses. Bootstrap Router (BSR) Protocol is a mechanism by which a PIM router learns the RP information.

The RP addresses are used as the root of a multicast group-specific distribution tree, the branches of which extend to all the nodes interested in receiving the traffic for that particular multicast group. For multicast sources to reach all receivers, the RP information is crucial so that all PIM routes use the same group-to-RP address mapping. Each node learns the same RP information using the following methods:

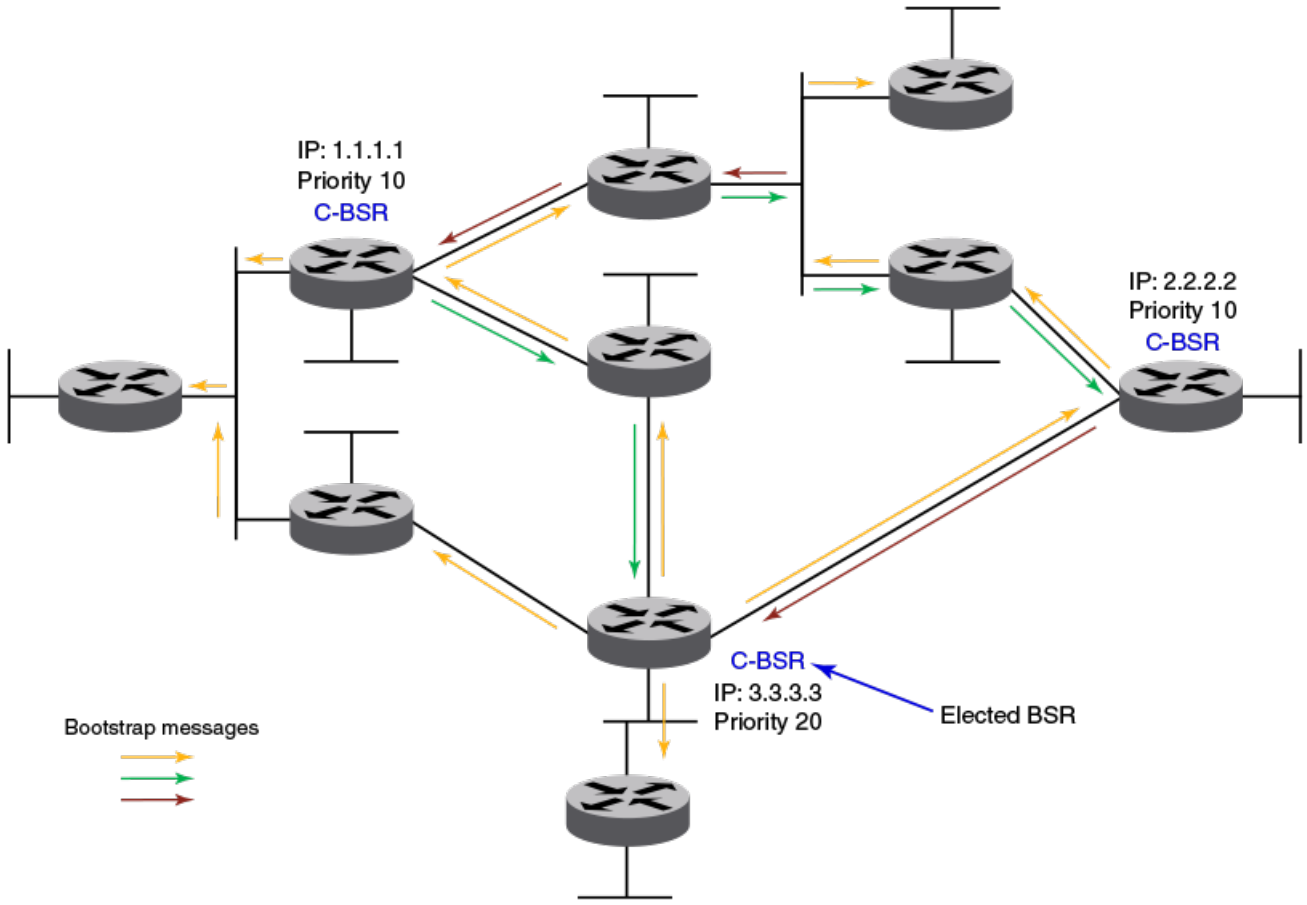
- Statically configuring the RP information on each PIM router.
- Using the BSR protocol, which distributes the RP information to each PIM router.

Some of the PIM routers act as Candidate RPs (C-RPs), out of which one C-RP gets elected and acts as RP for a particular group range. In addition, some PIM routers are configured as Candidate BSRs (C-BSRs), and one of these routers will be elected to act as the Bootstrap Router. All PIM routers learn the elected BSR through Bootstrap Messages (BSMs). All Candidate RPs will then report to the elected BSR, which will form the RP-set available in the network and distribute it to all the PIM routers. Therefore all PIM routers eventually have the same RP-set information.

The BSR protocol mechanism converges in the following phases:

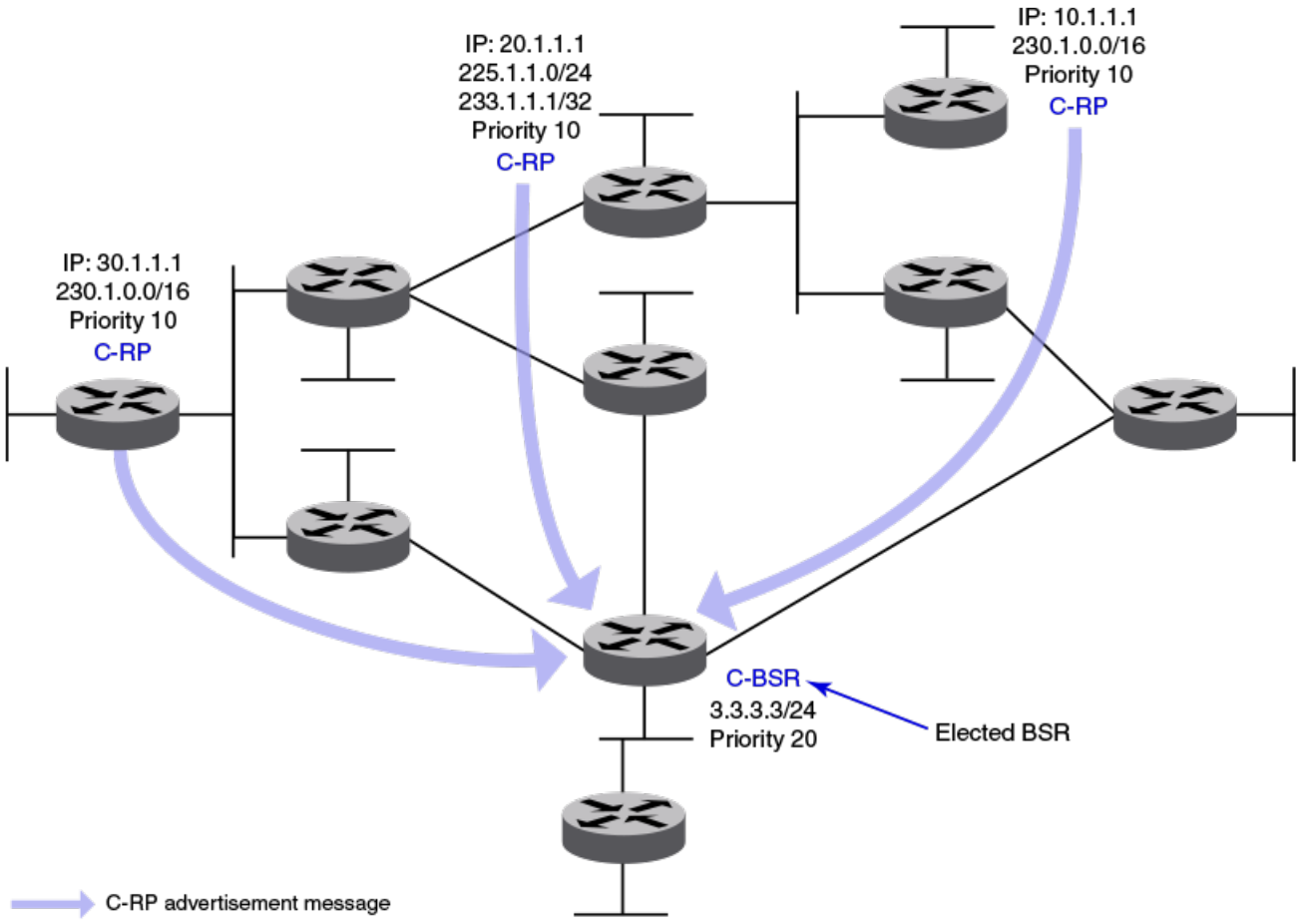
- BSR election
- Candidate RP Advertisement and RP set formation
- RP-set distribution

FIGURE 2 BSR Election



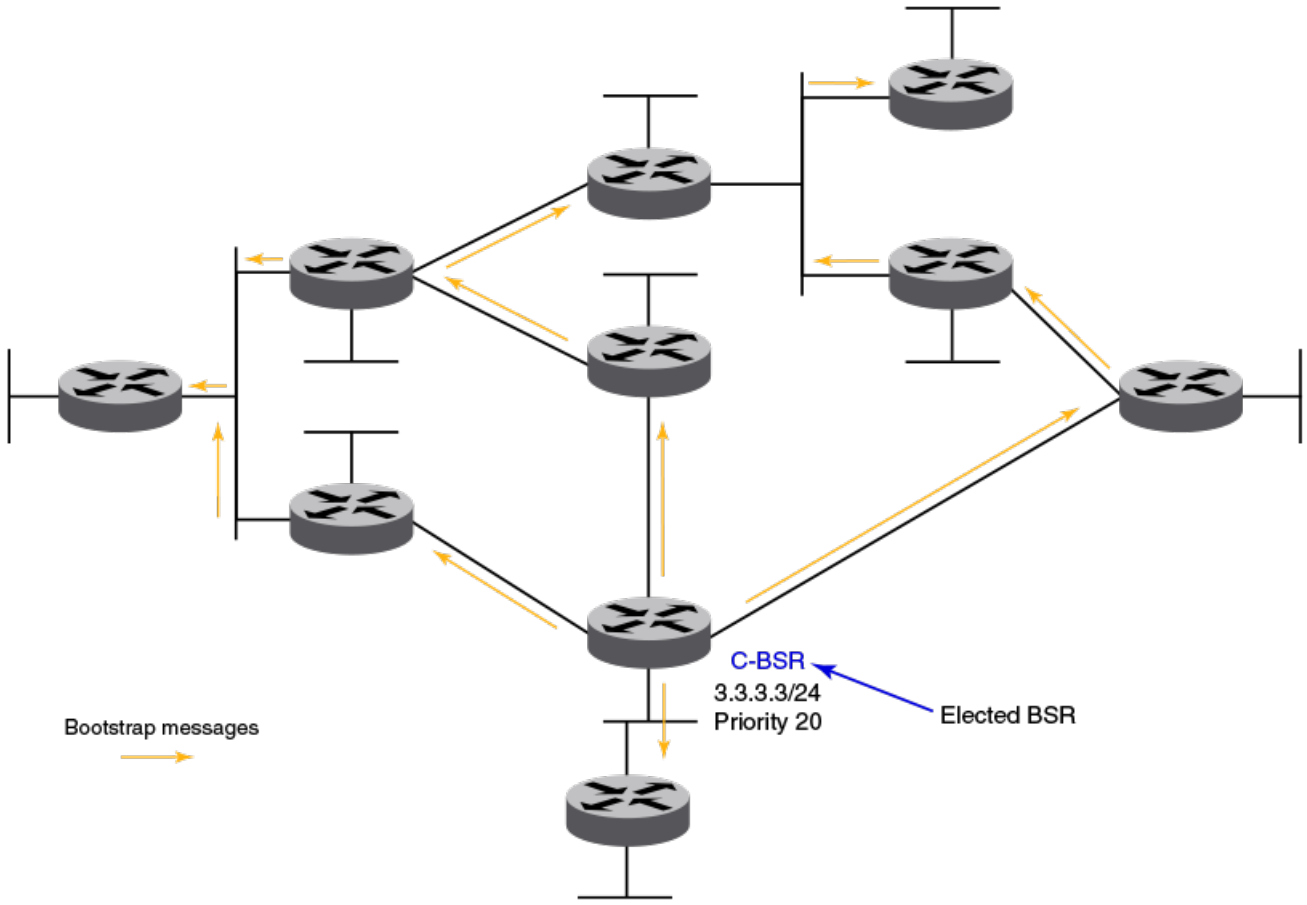
BSR election – Each candidate BSR periodically generates a Bootstrap message (BSM), which carries the configured BSR priority. Every PIM router in the domain floods these BSMs. Other C-BSRs that receive a BSM with higher priority suppress their own BSMs. Eventually, there will be only one C-BSR with BSMs that flood periodically into the network. This single C-BSR becomes the elected Bootstrap Router and its BSM informs all routers that it is the elected BSR.

FIGURE 3 Candidate RP advertisement and RP-set formation



Candidate RP advertisement and RP-set formation: Each candidate RP sends out periodic candidate RP advertisements (C-RP-Adv) messages to the elected BSR. These advertisement messages contain the candidate's priority and a list of multicast group ranges for which this C-RP would like to act as the RP. In addition, it also carries a hold time, for the BSR to discard this C-RP if the hold time expires. In this way, the elected BSR learns about all C-RPs up and reachable. As soon as the BSR starts receiving C-RP advertisements, it builds the RP-set information. This RP-set contains the list for multicast group ranges and C-RP addresses available for each of these group ranges, along with their respective priorities and hold times.

FIGURE 4 RP-set distribution



RP-set distribution: The RP-set built by the BSR is set through the same BSM message. Because these BSMs are flooded, the RP-set information rapidly reaches each PIM router. When a PIM router receives the RP-set, it adds all group-to-RP mappings to its pool of mappings, created from static RP configurations as well. Every PIM router runs the same RP hash algorithm to ensure the same C-RP is elected for a particular multicast group throughout the domain. In this way, all PIM routers can build the multicast group-specific distribution tree rooted to the same RP.

BSR timers and values

The BSR mechanism uses timers listed in the following table to ensure the protocol provides reliability and faster convergence. These timers can be configured.

TABLE 3 BSR timers and values

Timer	Default value	Description
Bootstrap message interval	60 seconds	The periodic interval after which a BSM is generated by a BSR.
Bootstrap timeout	130 seconds	The interval after which a BSR is timed out if no BSM is received from it.
Bootstrap minimum interval	10 seconds	The minimum interval after which a BSM should be sent out by a BSR.

TABLE 3 BSR timers and values (continued)

Timer	Default value	Description
C-RP mapping expiry timer	From message	Hold time from C-RP advertisement message. The hold time for C-RP is 2.5 times the RP advertisement interval.
RP mapping expiry timer	From message	Hold time from BSM.
Candidate RP advertisement interval	60 seconds	Periodic interval after which a C-RP generates an advertisement message to the BSR.

RP election algorithm (group-to-RP hashing)

The RP-set information received from the BSR is stored locally and updated by each PIM router periodically upon receiving BSMs. This RP-set contains the list for group prefixes and the corresponding list for C-RP for each group prefix.

The following steps list the RP election procedure for a particular multicast group address:

1. A longest match look-up is performed on all the group prefixes in the RP-set.
2. If more than one C-RP is found by a longest group prefix match, the C-RP with the lowest priority is elected.
3. If more than one C-RP has the same lowest priority, the BSR hash function is used to elect the RP.
4. If the hash functions return the same hash value for more than one C-RP, the highest IP address C-RP is elected.

Using loopback interfaces as an RP

Because loopback interfaces are operationally always up, it is preferable to use them as RPs. Beginning with Network OS 7.1.0, all existing PIM-SM protocol features are also supported on loopback interfaces. Layer 3-enabled loopback interfaces can act as static RP or Candidate-RP. They can also be configured as candidate-BSRs.

PIM-sparse device types

Devices configured with PIM-sparse interfaces also can be configured to fill one or more of the following roles:

- Bootstrap router (BSR): A router that distributes rendezvous point (RP) information to the other PIM-sparse devices within the domain. Each PIM-sparse domain has one active BSR. For redundancy, you can configure ports on multiple devices as candidate BSRs. The PIM-sparse protocol uses an election process to select one of the candidate BSRs as the BSR for the domain. The BSR with the highest BSR priority (a user-configurable parameter) is elected. If the priorities result in a tie, then the candidate BSR interface with the highest IP address is elected.

  The BSR must be configured as part of the Layer 3 core network.

- Rendezvous point (RP): The meeting point for PIM-sparse sources and receivers. A PIM-sparse domain can have multiple RPs, but each PIM-sparse multicast group address can have only one active RP. PIM-sparse devices learn the addresses of RPs and the groups for which they are responsible from messages that the BSR sends to each of the PIM-sparse devices.

  The RP must be configured as part of the Layer 3 core network.

  NOTE
  Extreme recommends that you configure the same ports as candidate BSRs and RPs.

- PIM designated router (DR): Once the RP has been identified, each interested host or router sends join messages to the RP for the group in which they are interested. The local network selects one of its upstream routers as the DR. All hosts below a DR send IGMP join messages to the DR. The DR sends only one join message to the RP on behalf of all its interested hosts. The

RP receives the first few packets of the multicast stream, encapsulated in the PIM register message, from the source hosts. These messages are sent as a unicast to the RP. The RP de-encapsulates these packets and forwards them to the respective DRs.

NOTE
DR election is based first on the router with the highest configured DR priority for an interface (if DR priority has been configured), and based next on the router with the highest IP address. To configure DR priority, use the **ip pim dr-priority** command.

PIM multinet

Extreme devices support PIM over secondary addresses in an IPv4 environment by configuring an IPv4 address with a secondary keyword.

Whenever a secondary address is configured on a interface, all the secondary addresses configured on the interface are sent out on the PIM Hello using the secondary address option.

Whenever a receiver uses a secondary address as its source and sends a IGMP group report, the PIM join and prunes are propagated up the network.

Whenever a secondary address is configured as a RP, the packets are processed appropriately

Displaying the secondary address

In this example the PIM neighbor on Ve10 has multiple IP addresses configured on the interface.

```
device# show ip pim neighbor
Total Number of Neighbors : 1
Port         Phy_Port     Neighbor        Holdtime Age           UpTime  Priority
                                          sec      sec      Dd HH:MM:SS
Ve10         Ve10          10.10.10.17     105      10          00:26:10         1
                          +20.20.20.21
```

Displaying PIM information

You can use several show commands to view information about PIM.

NOTE
Non-default VRFs can be configured with VRF name. For more information, refer Configuring multicast over multi VRF for IPv4 on page 25.

Use one of the following commands to view PIM information. The commands do not need to be entered in the specified order.

1. Enter the **show ip pim settings** command.

```
device# show ip pim settings
  Maximum mcache              : 24576     Current Count               : 0
  Hello interval              : 30        Neighbor timeout            : 105
  Join/Prune interval         : 60        Inactivity interval         : 180
  Hardware drop enabled       : 1         Prune wait interval         : 3
  Register Suppress Time      : 60        Register Probe Time         : 10
  Register Stop Delay         : 0         Register Suppress interval  : 0
  SSM Enabled                 : No        SPT Threshold               : 1
  Route Precedence            : uc-non-default uc-default none
```

2. Enter the **show ip pim mcache** command.

```
device# show ip pim mcache 50.1.1.101 230.1.1.1
IP Multicast Mcache Table
Entry Flags    : sm  - Sparse Mode, ssm - Source Specific Multicast
                 RPT - RPT Bit, SPT - SPT Bit, LSrc - Local Source
                 LRcv - Local Receiver, RegProbe - Register In Progress
                 RegSupp - Register Suppression Timer, Reg - Register Complete
                 needRte - Route Required for Src/RP
Interface Flags: IM - Immediate, IH - Inherited, WA - Won Assert
                 MJ - Membership Join, BR - Blocked RPT, BA - Blocked Assert
                 BF - Blocked Filter
Total entries in mcache: 8
1    (50.1.1.101, 230.1.1.1) in Ve 40, Uptime 00:03:29
     Sparse Mode, RPT=0 SPT=1 Reg=0 RegSupp=0 RegProbe=0 LSrc=0 LRcv=1
     upstream neighbor=40.1.1.3
     num_oifs = 2
           Ve 2(00:03:29/181) Flags: IM
           Ve 10(00:03:29/0) Flags: MJ
     Flags (0x400784d1)
         sm=1 ssm=0 needRte=0
```

The output of this command displays the multicast Mcache table.

3. Enter the **show ip pim traffic** command to display IPv4 traffic statistics.

```
device# show ip pim traffic
Port      |HELLO  |JOIN  |PRUNE  |ASSERT |GRAFT/REGISTER |REGISTER-STOP |BSR-MSGS  |RPC-MSGS
          |Rx     |Rx    |Rx     |Rx     |Rx             |Rx            |Rx        |Rx
----------+-------+------+-------+-------+---------------+--------------+----------+---------+
Ve10         54       0       0       0        0           0              0          0
Lo 1         0        0       0       0        0           0              0          0

device# show ip pim traffic
Port        |HELLO |JOIN  |PRUNE |ASSERT |GRAFT/REGISTER |REGISTER-STOP |BSR-MSGS |RPC-MSGS
            | Tx   |Tx    | Tx   |Tx     |Tx             |Tx
------------+------+------+------+-------+---------------+--------------+---------+---------+---
Ve10         29        0     0        0         0            0              0         0
Lo 1         28        0     0        0         0            0              0         0
```

The output of this command displays the Protocol Independent Multicast (PIM) traffic statistics categorized by each PIM enabled interface.

4. Enter the **show ip pim neighbor** command to display PIM neighbor information.

```
device(config)# show ip pim neighbor
-----+--------+------------+--------+---+---------+---------+-----+----------+-----------+----
Port |PhyPort |Neighbor    |Holdtime|T  |PropDelay|Override |Age  |UpTime    |VRF        |Prio
     |        |            |sec     |Bit|msec     |msec     |sec  |          |           |
-----+--------+------------+--------+---+---------+---------+-----+----------+-----------+----
v2     e1/1     2.1.1.2     105      1   500       3000      0     00:44:10   default-vrf  1
v4     e1/2     4.1.1.2     105      1   500       3000      10    00:42:50   default-vrf  1
v5     e1/1     5.1.1.2     105      1   500       3000      0     00:44:00   default-vrf  1
v22    e1/1     22.1.1.1    105      1   500       3000      0     00:44:10   default-vrf  1
Total Number of Neighbors : 4
```

5. Enter the **show ip pim bsr** command to display the bootstrap router information.

```
device# show ip pim bsr
PIMv2 Bootstrap information for Vrf Instance : default-vrf
------------------------------------------------------------------------------
  This system is the Elected BSR
  BSR address: 1.51.51.1. Hash Mask Length 32. Priority 255.
  Next bootstrap message in 00:01:00
  Configuration:
    Candidate loopback 2 (Address 1.51.51.1). Hash Mask Length 32. Priority 255.
Next Candidate-RP-advertisment in 00:01:00
  RP: 1.51.51.1
    group prefixes:
    224.0.0.0 / 4
  Candidate-RP-advertisement period: 60
```

6. Enter the **show ip pim rp-candidate** to display the rendezvous point (RP) information.

```
device# show ip pim rp-candidate
Next Candidate-RP-advertisement in 00:00:10
  RP: 207.95.7.1
    group prefixes:
    224.0.0.0 / 4
  Candidate-RP-advertisement period: 60
```

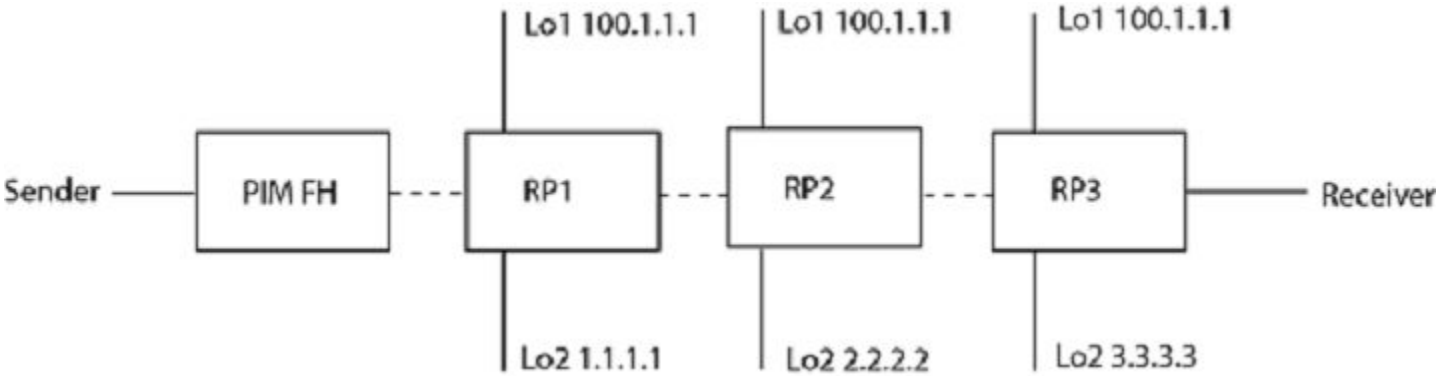
PIM Anycast RP

PIM Anycast RP is a method of providing load balancing and fast convergence to PIM RPs in an IPv4 multicast domain. The RP address of the Anycast RP is a shared address used among multiple PIM routers, known as PIM RP. The PIM RP routers create an Anycast RP set. Each router in the Anycast RP set is configured using two IP addresses; a shared RP address in their loopback address and a separate, unique ip address. The loopback address must be reachable by all PIM routers in the multicast domain. The separate, unique ip address is configured to establish static peering with other PIM routers and communication with the peers.

When the source is activated in a PIM Anycast RP domain**,** the PIM First Hop (FH) will register the source to the closet PIM RP. The PIM RP follows the same MSDP Anycast RP operation by decapsulating the packet and creating the (s,g) state. If there are external peers in the Anycast RP set, the router will re-encapsulate the packet with the local peering address as the source address of the encapsulation. The router will unicast the packet to all Anycast RP peers. The re-encapsulation of the data register packet to Anycast RP peers ensures source state distribution to all RPs in a multicast domain.

The example shown in the figure below is a PIM Anycast-enabled network with 3 RPs, 1 PIM-FH router connecting to its active source and local receiver. Loopback 1 in RP1, RP2, and RP3 have the same IP addresses 100.1.1.1. Loopback 2 in RP1, RP2, and RP3 each have separate IP addresses configured to communicate with their peers in the Anycast RP set.

FIGURE 5 Example of a PIM Anycast RP network



Configuring PIM Anycast RP

The PIM CLI specifies mapping of the RP and the Anycast RP peers.

1. Enter the **configure terminal** command to enter the global configuration mode.

   ```
   device# configure terminal
   ```

2. Enter the **router pim** command to enter the router PIM configuration mode.

   ```
   device(config)# router pim
   ```

3. Enter the **rp-address** command followed by the IP address to be configured as the RP for the PIM Sparse domain.

   ```
   device(config-pim-router)# rp-address 100.1.1.1
   ```

4. Enter the **anycast-rp** command followed by the RP address and the **anycast-rp-set** parameter, which specifies a host based simple prefix list name used to specify the address of the Anycast RP set, including a local address.

   ```
   device(config-pim-router)# anycast-rp 100.1.1.1 anycast-rp-set
   ```

The following example is a configuration of PIM Anycast RP 100.1.1.1. The example avoids using loopback 1 interface when configuring PIM Anycast RP because the loopback 1 address could be used as a router-id. A PIM First Hop router will register the source with the closest RP. The first RP that receives the register will re-encapsulate the register to all other Anycast RP peers.

The RP shared address 100.1.1.1 is used in the PIM domain. IP addresses 1.1.1.1, 2.2.2.2, and 3.3.3.3 are listed in the ACL that forms the self inclusive Anycast RP set. Multiple anycast-rp instances can be configured on a system; each peer with the same or different Anycast RP set.

```
device(config)# interface loopback 2
device(config-lbif-2)# ip address 100.1.1.1/24
device(config-lbif-2)# ip pim-sparse
device(config-lbif-2)# interface loopback 3
device(config-lbif-3)# ip address 1.1.1.1/24
device(config-lbif-3)# ip pim-sparse
device(config-lbif-3)# router pim
device(config-pim-router)# rp-address 100.1.1.1
device(config-pim-router)# anycast-rp 100.1.1.1 anycast-rp-set
device(config)# ip prefix-list anycast-rp-set permit 1.1.1.1/32
device(config)# ip prefix-list anycast-rp-set permit 2.2.2.2/32
device(config)# ip prefix-list anycast-rp-set permit 3.3.3.3/32
```

Enabling ECMP dynamic rebalance

Enabling ECMP dynamic rebalance configures the hash based distribution among the ECMP paths.

The **rebalance** option enables redistributing the load when a new next-hop is added. The redistribution is based on the hash function.

1. Enter the **configure terminal** command to enter the global configuration mode.

```
device# configure terminal
```

2. Enter the **router pim** command to enter the router PIM configuration mode.

```
device(config)# router pim
```

3. Enter the **rpf ecmp** command to enable ECMP load sharing.

```
device(config-pim-router)# rpf ecmp
```

4. Enter the **rpf ecmp rebalance** command to enable load sharing with dynamic rebalance.

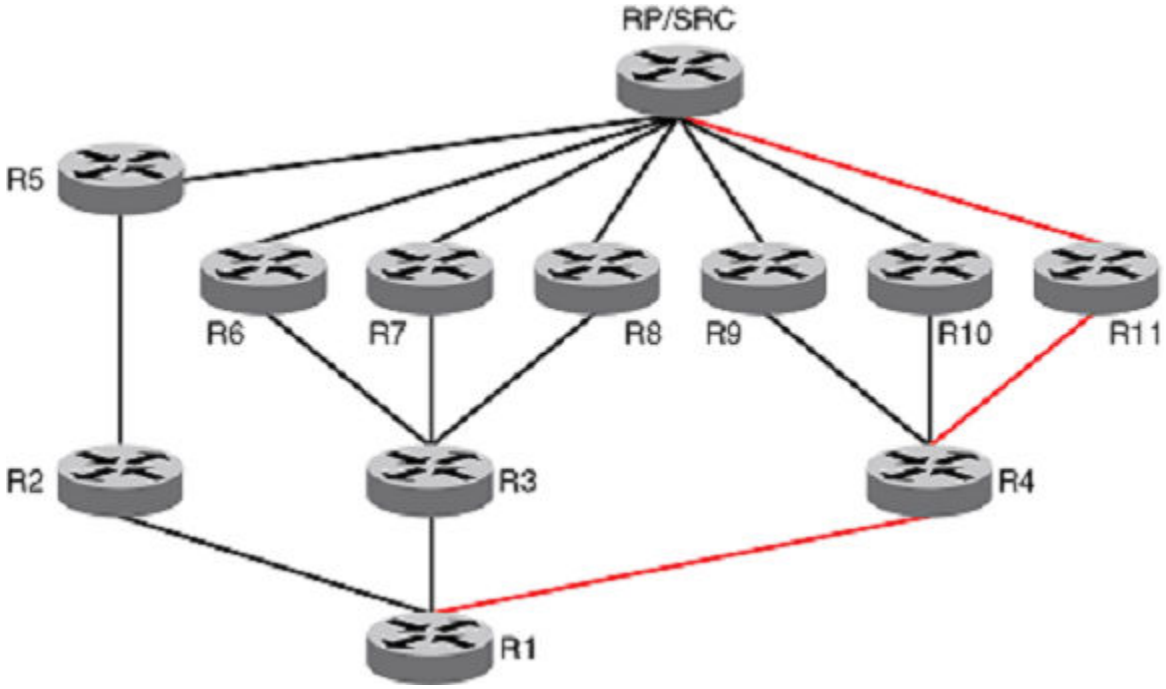
```
device(config-pim-router)# rpf ecmp rebalance
```

Multicast ECMP support

If there are multiple equal cost paths between PIM routers to reach the source or the RP, the multicast RPF algorithms distribute the load across available paths to take advantage of those paths.

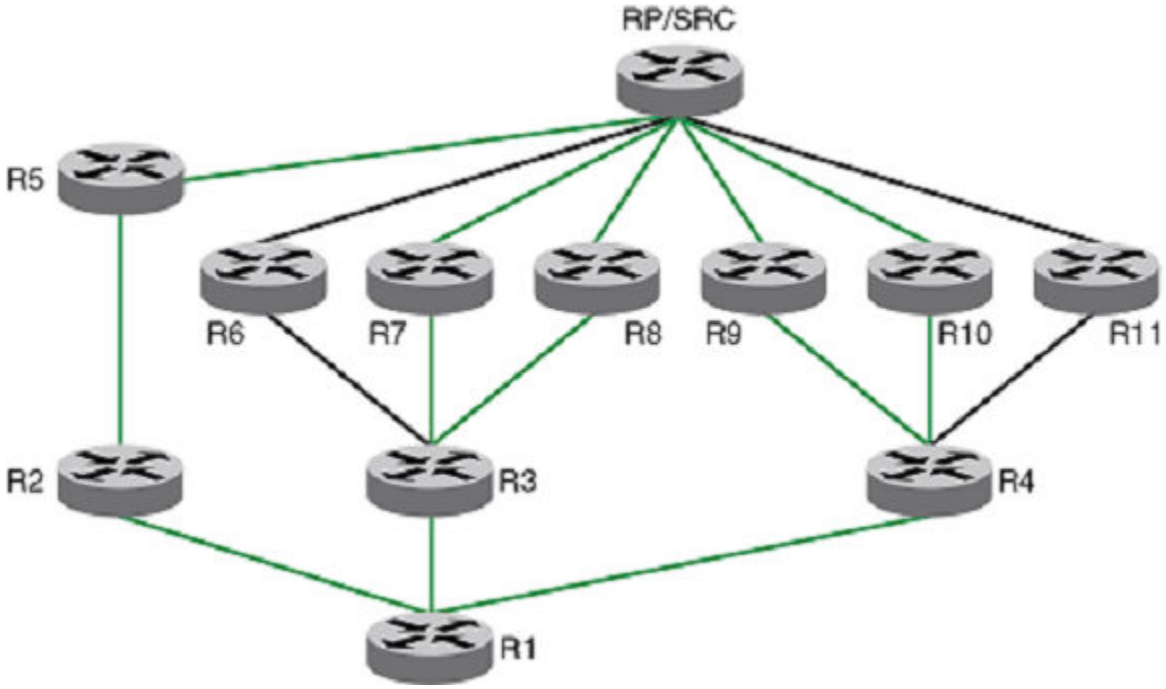
Figure 6 shows a topology in which R1 through R11 have IP addresses in ascending order (R1 having the lowest IP address and R11 having the highest). All the routers are PIM-enabled routers. The links emanating from each router are equal-cost multi-path (ECMP) links. The existing behavior path utilization is indicated in red. With the highest IP address neighbor chosen for the ECMP paths available, the multicast cache entries utilize only the R1-R4-R11-SRC/RP path.

FIGURE 6 Path utilization without multicast ECMP support



In the following figure, with the ECMP support turned on, the multicast entries will be distributed among the equal-cost next hops as indicated in green for better utilization of the available paths.

FIGURE 7 Path utilization with multicast ECMP support



The load distribution is achieved by distributing the multicast cache entries (*,G or S,G) to the available paths, and thus distributing the traffic. Two different methods are widely used to achieve this distribution:

- Hash based - Load splitting
- Least used path based - Load balancing

Extreme devices support the Hash based method of load distribution for multicast ECMP.

Hash based load distribution

The hash based load distribution depends on a hash function to distribute the multicast cache entries. The S, G, next-hop addresses are hash function based. This method splits the cache entries by choosing a different RPF neighbor and splits the traffic. Load balancing is based on the distribution of the keys S, G, next-hop. This method of distribution is the least disruptive as the hashing redistributes only those cache entries that are affected during link flaps. Some paths may not be utilized for the distribution of the multicast entries. For example, for the ECMP paths from R3 to R6, R7 and R8, only paths R3 to R7 and R3 to R8 are being utilized.

Deleting a path

When an ECMP path goes down, all the multicast entries using that path get redistributed among the other available paths.

Adding a path

When a new path is added to the ECMP set, there is no redistribution (default behavior without rebalance option) of the cache entries. Here optimal utilization of the paths is traded off in favor of not disturbing the existing flow. This method also requires a full branch setup towards the source or RP of the multicast distribution tree sometimes. When a path flaps (goes down and comes back up), the multicast entries which had been using this path would not be using this path anymore and it becomes worse if a subset of paths go down and come back up one by one, resulting in only the paths that did not flap to carry all entries.

Dynamic rebalancing

This option rebalances the traffic immediately on a new next-hop or path addition and helps in both new next-hop and path addition and path flap cases. There is least disruption in existing flows by using the hashing method.

Limitations and prerequisites

The following limitations and prerequisites apply to the configuration of ECMP path load balancing:

- The hash method is a load splitting method and hence traffic load balancing is not supported.
- S-based and S,G based hashing is not supported.
- The hash method is a load splitting method and not a load balancing method and hence the load balancing effect due to load splitting the multicast entries is only a best effort and the splitting is actually based on the number of S, G flows and the number of next-hops and the actual distribution of the S,G and the next-hop addresses.
- If the rebalancing is not configured, then link flap results in sub-optimal utilization of the ECMP links.
- The number of paths supported by multicast ECMP would be the same as unicast ECMP which is 32.

Mtrace overview

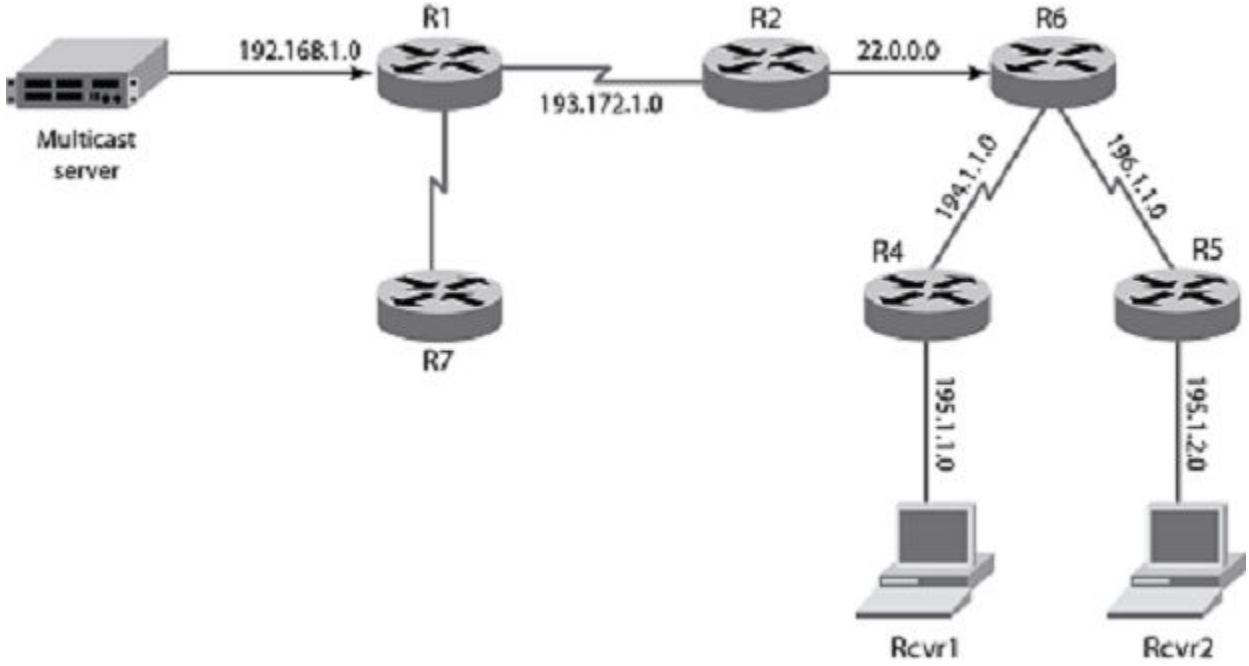
Mtrace is a diagnostic tool to trace the multicast path from a specified destination to a source for a multicast group. It runs over IGMP protocol. Mtrace uses any information available to it to determine a previous hop to forward the trace towards the source.

There are three main components in an Mtrace implementation. They are mtrace query, mtrace request, and mtrace response.

The unicast traceroute program allows the tracing of a path from one machine to another. The key mechanism for unicast traceroute is the ICMP TTL exceeded message, which is specifically excluded as a response to multicast packets. The multicast traceroute facility allows the tracing of an IP multicast routing path. Multicast traceroute also requires special implementations on the part of routers.

Multicast traceroute uses any information available to it in the router to determine a previous hop to forward the trace towards the source. Multicast routing protocols vary in the type and amount of state they keep; multicast traceroute endeavors to work with all of them by using whatever is available. For example, if a PIM-SM router is on the (*,G) tree, it chooses the parent towards the RP as the previous hop. In these cases, no source/group-specific state is available, but the path may still be traced.

FIGURE 8 Network topology



Mtrace components

There are 3 main components in a multicast traceroute implementation. They are:

1. Mtrace Query
2. Mtrace Request
3. Mtrace Response

- Mtrace Query

  The party requesting the traceroute sends a traceroute query packet to the last-hop multicast router for the given destination. The query and request have the same opcode, the receiving router can distinguish between a query and a request by checking the size of the packet. A query is a request packet with none of the response fields filled up.

- Mtrace Request

  The last-hop router turns the Query packet into a Request packet by adding a response data block containing its interface addresses and packet statistics, and then forwards the Request packet via unicast to the router that it believes is the proper previous hop for the given source and group. Each hop adds its response data to the end of the Request packet, then unicast forwards it to the previous hop.
- Mtrace Response

  The first hop router (the router that believes that packets from the source originate on one of its directly connected networks) changes the packet type to indicate a Response packet and sends the completed response to the response destination address. The response may be returned before reaching the first hop router if a fatal error condition such as "no route" is encountered along the path.

Configuring mtrace

The mtrace can be started on any router on the network.

Assume that the destination is 195.1.2.1, source is 192.168.1.1 and group is 225.1.1.1. The mtrace query is initially sent from R7. The initial header is not to be modified by any of the routers. R5 adds a response block based on the (S, G) or the (*, G) entry and adds its incoming interface, outgoing interface and other information specified in the draft and sends it to its upstream neighbor which is R6. R6 similarly adds a response block and sends it to its upstream neighbor R2, likewise till it reaches R1. Once it reaches R1, R1 determines that it is the first hop router and completes the response block and sends the response back to R7. R7 now reads the information from the packet and prints it out.

Enter Privileged EXEC mode and enter the **mtrace** command followed by the source, destination and group IP address.

```
device# mtrace source 20.1.1.2 destination 155.1.1.1 group 225.0.0.1
```

The following output displays:

```
Mtrace handle query from src 20.1.1.2  to dest 155.1.1.1  through group 225.0.0.1

Collecting Statistics, waiting time 5 seconds.....

Type Control-c to abort 0 12::1 PIM thresh^ 1 MTRACE_NO_ERR 1 13::1 PIM thresh^ 1 MTRACE_NO_ERR 2 102::2
PIM thresh^ 1 MTRACE_REACHED_RP
```

Protocol-Independent Multicast overview

The Protocol-Independent Multicast (PIM) protocol is a family of IPv4 multicast protocols. PIM does not rely on any particular routing protocol for creating its network topology state. Instead, PIM uses routing information supplied by other traditional routing protocols, such as Open Shortest Path First, Border Gateway Protocol, and Multicast Source Discovery Protocol.

PIM messages are sent encapsulated in an IP packet with the IP protocol field set to 103. Depending on the type of message, the packet is either sent to the PIM All-Router-Multicast address (224.0.0.13) or sent as unicast to a specific host.

As with IP multicast, the main use case of PIM is for the source to be able to send the same information to multiple receivers by using a single stream of traffic. This helps minimize the processing load on the source, as the source needs to maintain only one session irrespective of the number of actual receivers. It also minimizes the load on the IP network, because the packets are sent only on links that lead to an interested receiver.

Several types of PIM exist, but Extreme supports only PIM sparse mode (PIM-SM, and PIM-SSM). PIM-sparse explicitly builds unidirectional shared trees rooted at a rendezvous point (RP) per group, and optionally creates shortest-path trees per source.

Enabling PIM on a router

Use the following procedure to enable PIM globally.

1. Enter global configuration mode.

```
device# configure terminal
```

2. Enter the **router pim** command to enter the PIM router configuration mode and configure a variety of options.

```
device(config)# router pim
device(config-pim-router)#
```

Configuring PIM

Once you enable PIM on a device, you can configure a variety of options in the router PIM configuration mode.

1. Enter the **hello-interval** command to configure the PIM hello timeout.

```
device(config-router-pim-vrf-default-vrf)# hello-interval 40
```

2. Enter the **nbr-timeout** command to configure the PIM neighbor timeout.

```
device(config-router-pim-vrf-default-vrf)# nbr-timeout 160
```

3. Enter the **bsr-candidate** command to configure the BSR candidate.

```
device(config-router-pim-vrf-default-vrf)# bsr-candidate interface loopback 11 mask 32
```

4. Enter the **prune-wait** command to configure the PIM prune pending timeout.

```
device(config-router-pim-vrf-default-vrf)# prune-wait 5
```

5. Enter the **message-interval** command to configure the PIM join or prune interval.

```
device(config-router-pim-vrf-default-vrf)# message-interval 180
```

6. Enter the **spt-threshold infinity** command which allows router to use the PIM shared tree (RP tree) instead of Shortest Path Tree (SPT).

```
device(config)# router pim
device(config-router-pim-vrf-default-vrf)# spt-threshold infinity
```

7. Enter the **no spt-threshold** command to configure the PIM Shortest Path Tree (SPT) threshold. The default value of spt-threshold value is 1 which causes the router to switch to Shortest Path Tree on receiving the first multicast packet in LHR (Last Hop Router).

```
device(config)# router pim
device(config-router-pim-vrf-default-vrf)# no spt-threshold
```

8. Enter the **ssm-enable range** command to set the multicast address range to use for SSM.

```
device(config-router-pim-vrf-default-vrf)# ssm-enable range PL_ssm_range-230-to-234
```

Entering only the **ssm-enable** command applies the 232.0.0.0/8 default SSM range. This default range is displayed in the **show ip pim settings** output.

Enabling PIM-sparse on routed interfaces

The following procedure enables PIM-sparse and other options on supported interfaces.

You must enable PIM globally before enabling PIM sparse on the interface.

1. To enable PIM-sparse on an interface (Ethernet, loopback, or VE), enter the global configuration mode.

```
device# configure terminal
```

2. In global configuration mode, specify an interface.

```
device(config)# interface ethernet 1/1
```

3. Enter the **ip pim-sparse** command in the interface configuration mode.

```
device(conf-if-eth-1/1)# ip pim-sparse
```

4. (Optional) To change the designated router (DR) priority from the default, enter the **ip pim dr-priority** command in interface subtype configuration mode and specify a non-default value:

```
device(conf-if-eth-1/1# ip pim dr-priority 200
```

5. (Optional) To set the TTL threshold, enter the **ip pim ttl-threshold** command in the interface configuration mode.

```
device(conf-if-eth-1/1# ip pim ttl-threshold 50
```

Configuring PIM RP

You can use the PIM Sparse protocol's RP election process so that a backup RP can automatically take over if the active RP router becomes unavailable.

However, if you do not want the RP to be selected by the RP election process but want to explicitly identify the RP by address, use the **rp-address** command.

If you explicitly specify the RP, the device uses the specified RP for all group-to-RP mappings and overrides the set of candidate RPs supplied by the BSR.

1. Enter the **configure terminal** command to enter global configuration mode.

```
device# configure terminal
```

2. Enter the **router pim** command to enter the router PIM configuration mode.

```
device(config)# router pim
```

3. Enter the **rp-address** command to specify the IP address of the RP.

```
device(config-pim-router)# rp-address 4.4.4.4
```

   The command in this example identifies the device interface at IP address 4.4.4.4 as the RP for the PIM-sparse domain. The device uses the specified RP and ignores group-to-RP mappings received from the BSR.

4. For static RP configuration with specific group ranges, enter the following commands.

```
device(config-pim-router)# rp-address 4.4.4.4 static-rp-list
device(config)# ip prefix-list static-rp-list permit 225.1.1.0/24
```

The following commands configure the RP candidate.

```
device(config-pim-router)# rp-candidate interface loopback 11
device(config-pim-router)# rp-candidate prefix my-rp-cand-list
device(config)# ip prefix-list my-rp-cand-list permit 226.1.1.0/24
device(config)# ip prefix-list my-rp-cand-list permit 228.1.1.0/24
```

Multicast on bridge domain

Bridge domain interface is a logical interface that allows bidirectional flow of traffic.

Multiple service end points like port-vlan or port-vlan-vlan can be made part of a single broadcast domain where we can achieve any-to-any bridging called Bridge Domain. The service end points can be of different types like Pseudo wire, VxLAN tunnel/VNI endpoints and other. Multicast IGMP snooping and IP Multicast PIM are supported on Bridge Domain.

IGMP on bridge domain

IGMP snooping on bridge domain is used to learn the particular multicast group on specific ports associated with the bridge domain by trapping the IGMP control packets and programing the hardware entries with learned Multicast groups and list of interested ports that are part of the bridge domain. When multicast traffic comes from a source the traffic is sent to the interested receivers instead of flooding the multicast traffic on all ports of the bridge domain. IGMP on bridge domain internally works same as it works on VLAN. Bridge domain contains logical interfaces (LIFs) so corresponding multicast groups contain the LIFs as the outgoing interfaces.

PIM support on VE bind to bridge domain

Layer3 interface VE can be bound to VLAN or the bridge domain. PIM on bridge domain works similar to VLAN case where the VE is bound to VLAN or bridge domain. PIM-SM can be configured on the VE. PIM snooping is also supported on Bridge domain. PIM snooping is used to send the data traffic to only interested PIM routers rather than all PIM routers connected in the Bridge domain.

To enable PIM snooping on bridge domain, use enable command.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip pim-sparse
```

Configuring IGMP snooping on a bridge domain

Follow these steps to configure IGMP snooping on a bridge domain.

1. To enable IGMP snooping on bridge domain, use enable command. To disable the function, use the no form of this command.

```
device # ip igmp snooping enable
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp snooping enable
```

Syntax: [no] ip igmp snooping enable

2. To enable IGMP querier on bridge domain, use enable command. To disable the function, use the no form of this command.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp snooping querier enable
```

   Syntax: [no] ip igmp snooping querier enable

3. IGMP version.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp version v2
```

   Syntax: [no] ip igmp snooping version <v1/v2/v3>

4. To enable IGMP fast leave on bridge domain, use enable command. To disable the function, use the no form of this command.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp snooping fast-leave
```

   Syntax: [no] ip igmp snooping fast-leave

5. Querier interval. The default is 125 seconds.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp snooping query interval 30
```

   Syntax: [no] ip igmp snooping query interval <1-18000>

6. Query max response time.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# ip igmp snooping query-max-response-time 20
```

   Syntax: [no] ip igmp snooping query-max-response-time <1-25>

7. IGMP last membership query interval.

```
device(config)# bridge-domain 10
device(config-bridge-domain-10)# [no] ip igmp snooping last-member-query-interval  150
```

   Syntax:[no] ip igmp snooping last-member-query-interval <100-25500>

The following example is the steps in the previous configuration:

```
device(config-BD-id)# ip igmp snooping ?
Possible completions:
  enable                        IGMP Enable
  fast-leave                    Fast Leave Processing
  last-member-query-interval    Last Member Query Interval
  mrouter                       Multicast Router
  querier                       Querier
  query-interval                Query Interval
  query-max-response-time       IGMP Max Query Response Time
  static-group                  Static Group to be Joined
  version                       IGMP Snooping Version
```

Multi-Chasis Trunk (MCT)

A Multi-Chassis Trunk (MCT) is a trunk that initiates at a single MCT-unaware server or switch and terminates at two MCT-aware switches.

Link Aggregation (LAG) trunks provide link level redundancy and increased capacity. However, LAG trunks do not provide switch-level redundancy. If the switch to which the LAG trunk is attached fails, the entire LAG trunk loses network connectivity. With MCT, member links of the LAG are connected to two chassis. The MCT switches may be directly connected using an Inter-Chassis Link (ICL) to enable data flow and control messages between them. In this model, if one MCT switch fails, a data path will remain through the other switch.

In an MCT scenario, all links are active and can be load shared to increase bandwidth. In addition, traffic restoration can be achieved in milliseconds after an MCT link failure or MCT switch failure. MCT is designed to increase network resilience and performance.

MP-BGP EVPN

Multi-protocol BGP is an extension to BGP that enables BGP to carry routing information for multiple network layers. Ethernet VPN EVPN) connect a group of customer sites using a virtual bridge. Treats MAC addresses as routable addresses and distribute them in Border Gateway Protocol (BGP). Uses Multi-protocol BGP (MP-BGP)

Advantages of MP-BGP EVPN

The advantages of MP-BGP EVPN are -

- It is standard based.
- Scalable and reliable because of Border Gateway Protocol base.
- Policies can be applied.
- Support exchange of IP addresses and IP prefixes.

Layer 2 Multicast Snooping over MCT

Multicast control packets behavior on Multi-Chasis Trunk (MCT).

Multicast state information is synced between the MCT peers using MP-BGP EVPN transport. Multicast protocol packets will not be sent on the peer link unless required.

IGMP/MLD protocol packets are of two types:

1. Membership query
    - General query - In a query message, the multicast address field is set to 0 when MLD sends a general query. The general query learns which multicast addresses have listeners on an attached link.
    - Group specific query - A group address is a multicast address.
2. Membership reports - In a report, the multicast address field is that of the specific multicast address to which the sender is listening.
    - Version 1 membership report - MLD version 1 is based on version 2 of the Internet Group Management Protocol (IGMP).
    - Version 2 membership report - MLD version 2 is based on version 3 of the IGMP. MLD version 2 is fully backward-compatible with MLD version 1.
    - Leave group

IGMP/MLD query packet processing on MCT

For IGMP queries, each EVI has BUM suppressed MGID associated, IGMP/MLD query packets need to be transmitted on ICL to address the following scenarios:

- The querier connected to only one of the MCT peer switch becomes elected querier.
- Only one of the peer switch is configured as querier.
- The switch would age out IGMP/MLD routes if memberships are not confirmed within time out interval.Although query packet is received on MCT peer link, mrouter port is not learnt/considered on that peer link.

IGMP/MLD membership reports

For IGMP reports and leaves,

- Traditionally, each peer switch learns about L2 Multicast memberships by snooping the IGMP/MLD membership reports. The membership reports are then flooded on mrouter ports.
- For MCT, since mrouter port is not learnt on the peer link, membership reports are not flooded between the peer switches. Peer switches exchange learnt routes using EVPN NLRI messages between BGP peers running on MCT cluster control vlan.
- BGP on the peer switch would communicate with the IGMP/MLD routes to Multicast module to add/delete group memberships and allocates/rejects the MGIDs and installs/un-installs routes in hardware.
- Also, each peer switch would generate proxy report for the IGMP/MLD routes learnt from EVPN NLRI, if general query or group-specific query is received from any port, other than peer-link.

Leave membership report

When fast-leave is not configured and MCT peer receives leave membership report from one of its clients for group G, the switch/router informs other MCT peers about the group specific query and latency using IGMP leave synch route. The peer switch, which runs IGMP querier sends group specific query and group query.

mRouter synchronization

mRouter synchronization helps in achieving optimal path selection for unknown multicast traffic and optimal MP-BGP message exchange between MCT Peers.

BGP handling of EVPN IGMP routes

In DC applications, EVPN is used as way of standard inter-POD communication for both intra-DC and inter-DC.

A subnet can span across multiple PODs and DCs. EVPN provides robust multi-tenant solution with extensive multi-homing capabilities to stretch a subnet (e.g., VLAN) across multiple PODs and DCs. There can be many hosts/VMs (e.g., several hundreds) attached to a subnet that is stretched across several PODs and DCs. These hosts/virtual machines express their interests in multicast groups on a given subnet/VLAN by sending IGMP membership reports (Joins) for their interested multicast group(s). Also, an IGMP router (e.g., IGMPv1) periodically sends membership queries to find out if there are hosts on that subnet still interested in receiving multicast traffic for that group. Just like ARP/ND suppression mechanism in EVPN to reduce the flooding of ARP messages over EVPN, it is also desired to have a mechanism to reduce the flood of IGMP messages (both Queries and Reports) in EVPN. This is achieved through IGMP Join Sync and IGMP Leave Sync EVPN routes specified in the draft "draft-ietf-bess-evpn-igmp-mld-proxy-00".

TABLE 4 EVPN Routes

EVPN route type	Route type name	Usage
7	IGMP Join Synch Route	To exchange (S,G)/(*,G) learnt on BGP EVPN peers.
8	IGMP Leave Synch Route	To exchange group leave between BGP EVPN peers.

IGMP Join Synch Route

IGMP allows a network host to inform a router that it is interested in receiving a particular multicast stream.

To begin, the multicast group is assigned a multicast address (that is, an IP address in the 224.0.0.0/4 class D address space). Hosts register to receive the stream join the group by sending an IGMP Report to the upstream multicast router. The router then adds that group to the list of multicast groups that should be forwarded onto the local subnet.

The router does not maintain state about which hosts on the subnet are to receive traffic for the group. Instead, the router continues to send traffic to the subnet until either a timeout value expires or there are no more hosts in that group on the subnet.

TABLE 5 IGMP Join Synch Route

Packets	Usage
RD (8 Octets)	
Ethernet Segment Identifier (10 Octets)	0 if (S, G) / (*, G) is learnt on CEP
Ethernet Tag ID (4 octets)	0 or optionally, set to Vlan ID over which Multicast Route is learnt
Multicast Source Length (1 octet)	32 for IPv4 address, 128 for IPv6 address and 0 for Wildcard address (*)
Multicast Source Address (variable)	IP address of multicast source
Multicast Group Length (1 octet)	32 for IPv4 address, 128 for IPv6 address
Multicast Group Address (variable)	IP address of multicast group
Originator Router Length (1 octet)	32 for IPv4 address, 128 for IPv6 address
Originator Router Address (variable)	The IP address of the originating router.
Flags (1 octets) (optional)	The flag fields are defined below in the table

NOTE

For,
Querier Config Sync - Multicast Group is set to 224.0.0.2.
Mrouter Sync - Multicast Group is set to 224.0.0.1.

TABLE 6 Flags

0	1	2	3	4	5	6	7
reserved	Querier config synch	mRouter synch	PIM snooping	IE	V3	V2	V1
	Bit 1 indicates QuerierConfigSync.	Bit 2 indicates MrouterSync.	Bit 3 indicates PIM Snooping.	Bit 4 indicates whether the (S, G) information carries within the route-type of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit should be ignored if bit 5 is set in case of IGMP Version 2 & IGMP Version 1 & MLD Version 1.	Bit 5 indicates support for IGMP/MLD version 3.	Bit 6 indicates support for IGMP/MLD version 2.	Bit 7 indicates support for IGMP/MLD version 1.

NOTE

Bits 1, 2, 3 are proprietary fields.

IGMP Leave Synch Route

When a host no longer wants to receive multicast traffic, it sends the router an IGMP Leave message.

After receiving this message, the router sends a query to the local subnet to determine whether any group members remain, sending the message to all hosts on the subnet, at the multicast All-Hosts address (224.0.0.1). If any host responds, the router continues to send to the group; if not, the router removes the multicast group from its forwarding list and stops sending to the group.

TABLE 7 IGMP Join Synch Route

Packets	Usage
RD (8 Octets)	
Ethernet Segment Identifier (10 Octets)	0 if (S, G) / (*, G) is learnt on CEP
Ethernet Tag ID (4 octets)	0 or optionally, set to Vlan ID over which Multicast Route is learnt
Multicast Source Length (1 octet)	32 for IPv4 address, 128 for IPv6 address and 0 for Wildcard address (*)
Multicast Source Address (variable)	IP address of multicast source
Multicast Group Length (1 octet)	32 for IPv4 address, 128 for IPv6 address
Multicast Group Address (variable)	IP address of multicast group
Originator Router Length (1 octet)	32 for IPv4 address, 128 for IPv6 address
Originator Router Address (variable)	The IP address of the originating router.
Flags (1 octets) (optional)	The flag fields are defined below in the table

TABLE 8 Flags

0	1	2	3	4	5	6	7
reserved	reserved	reserved	reserved	IE	V3	V2	V1
				Bit 4 indicates whether the (S, G) information carries within the route-type of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit should be ignored if bit 5 is set in case of IGMP Version 2 & IGMP Version 1 & MLD Version 1.	Bit 5 indicates support for IGMP/MLD version 3.	Bit 6 indicates support for IGMP/MLD version 2.	Bit 7 indicates support for IGMP/MLD version 1.

NOTE

Leave Group Synchronization & Maximum Response time are used during in Leave group synch procedures.

IGMP join and leave procedure

IGMP Join and Leave Group IGMP routes sent to BGP through RibLib and transported to the BGP EVPN Peers.

These routes are carried in BGP EVPN NLRI as Type 7 and Type 8 routes. BGP adds the EVI-RT extended community to the EVPN NLRI and transports the route to the EVPN peers. An EVPN configuration must be present for the specified EVI.

EVPN IGMP routes received from the remote peers are validated against import rules and added to VPN table in BGP only if the validation passes. The routes are later imported into the MAC VRF table, if the RT in the route matches the RTs configured for a given EVI (VLAN/BD). Consolidation of the routes from different sources (RDs) takes place in BGP after the route passes the route target and import filtering checks.

The routes are then installed in BGP and also forwarded to IGMP process through RibLib. Installed routes are also forwarded to other EVPN peers. BGP EVPN should be configured to support the IGMP routes.

NOTE
No new configuration is needed in BGP to support the IGMP routes.

EVI Route Target extended community

The EVI Route Target (RT) is a new EVPN extended community of Type 6 and a subtype yet to be defined by IANA. However, EVI-RT extended community is NOT supported for IGMP routes.

Instead, the Route Target extended community with Type 0x00 and Subtype 0x02 is supported. This extended community carries the RT associated with the EVI (VLAN/BD), so that the receiving PE can identify the EVI properly.

ES-Import Route Target extended community

ES-Import (ES-I) Route Target (RT) is another EVPN extended community of Type 0x06 and Subtype 0x02. The 6-byte value calculated is based on the ES-Import.

The IGMP Join and Leave Synch routes carry the ES-Import RT for the ES on which the IGMP membership report was received. Thus, it may go only to the PEs attached to that ES (and not to any other PEs).

Encap Type support

Only NSH tunnel encapsulation type is supported. MPLS and VXLAN types are not supported.

Traffic Forwarding Path for L2 Multicast

Both peers are updated with (*, G) membership for CCEP.

When a receiver connected on CCEP sends membership report to join group G. The DF election of the CCEP and MGID updation by multicast module prevents duplication of multicast data packets destined to group G on CCEP as well as peer-link.

Always honor DF election while programming receivers on CCEP in MGID. However, the path given by DF election may not be optimal as it might direct the multicast data traffic originating at one peer switch to receiver on CCEP via peer-link even though (*,G) membership does not include CEP on the peer-switch that does not host the source.

When a group is learnt on member VLANs, the DF for IGMP/MLD route is elected by hashing on IVID of the member VLAN, Source-IP and Group-IP.

The OIF list of IGMP/MLD route on DF includes:

- Receivers connected via CCEP.
- ICL if its MCT peer has receivers connected via CEP.
- Receivers connected via local CEP.

The OIF list of IGMP/MLD route on non-DF includes:

- ICL to redirect the multicast stream to DF of the stream.
- Receivers connected via CEP.

Data Encapsulation of L2 Multicast Traffic on ICL

Data Encapsulation of L2 Multicast from CEP/CCEP received on member vlan is similar to L2 Flooding traffic.

- If the CCP is down, it will forward locally.
- If the remote CCEP is down, it will forward locally.
- If the local CCEP is down, it will not forward locally.
- If the ingress is the CEP, it will forward locally.
- If the ingress is the ICL, it will not forward locally.
- If the ingress is a different CCEP, it will forward locally.