Multicast

In This Chapter

This chapter provides information about IPv6, Internet Group Management Protocol (IGMP) and Protocol Independent Multicast (PIM).

Topics in this chapter include:

Introduction to Multicast

IP multicast provides an effective method of many-to-many communication. Delivering unicast datagrams is fairly simple. Normally, IP packets are sent from a single source to a single recipient. The source inserts the address of the target host in the IP header destination field of an IP datagram, intermediate routers (if present) simply forward the datagram towards the target in accordance with their respective routing tables.

Sometimes distribution needs individual IP packets be delivered to multiple destinations (like audio or video streaming broadcasts). Multicast is a method of distributing datagrams sourced from one (or possibly more) host(s) to a set of receivers that may be distributed over different (sub) networks. This makes delivery of multicast datagrams significantly more complex.

Multicast sources can send a single copy of data using a single address for the entire group of recipients. The routers between the source and recipients route the data using the group address route. Multicast packets are delivered to a multicast group. A multicast group specifies a set of recipients who are interested in a particular data stream and is represented by an IP address from a specified range. Data addressed to the IP address is forwarded to the members of the group. A source host sends data to a multicast group by specifying the multicast group address in the datagram's destination IP address. A source does not have to register in order to send data to a group nor do they need to be a member of the group.

Routers and Layer 3 switches use the Internet Group Management Protocol (IGMP) to manage membership for a multicast session. When a host wants to receive one or more multicast sessions it will send a join message for each multicast group it wants to join. When a host wants to leave a multicast group, it will send a leave message.

To extend multicast to the Internet, the multicast backbone (Mbone) is used. The Mbone is layered on top of portions of the Internet. These portions, or islands, are interconnected using tunnels. The tunnels allow multicast traffic to pass between the multicast-capable portions of the Internet. As more and more routers in the Internet are multicast-capable (and scalable) the unicast and multicast routing table will converge.

The original Mbone was based on Distance Vector Multicast Routing Protocol (DVMRP) and was very limited. The Mbone is, however, converging around the following protocol set:

- IGMP
- Protocol Independent Multicast (Sparse Mode) (PIM-SM)
- Border Gateway Protocol with multi-protocol extensions (MBGP)
- Multicast Source Discovery Protocol (MSDP)

Multicast Models

Alcatel-Lucent routers support two models to provide multicast:

Any-Source Multicast (ASM)

Any-Source Multicast (ASM) is the IP multicast service model defined in RFC 1112, *Host extensions for IP Multicasting*. An IP datagram is transmitted to a host group, a set of zero or more end-hosts identified by a single IP destination address (224.0.0.0 through 239.255.255.255 for IPv4). End-hosts can join and leave the group any time and there is no restriction on their location or number. This model supports multicast groups with arbitrarily many senders. Any end-host can transmit to a host group even if it is not a member of that group.

To combat the vast complexity and scaling issues that ASM represents, the IETF is developing a service model called Source Specific Multicast (SSM).

Source Specific Multicast (SSM)

The Source Specific Multicast (SSM) service model defines a channel identified by an (S,G) pair, where S is a source address and G is an SSM destination address. In contrast to the ASM model, SSM only provides network-layer support for one-to-many delivery.

The SSM service model attempts to alleviate the following deployment problems that ASM has presented:

- Address allocation — SSM defines channels on a per-source basis. For example, the channel (S1,G) is distinct from the channel (S2,G), where S1 and S2 are source addresses, and G is an SSM destination address. This averts the problem of global allocation of SSM destination addresses and makes each source independently responsible for resolving address collisions for the various channels it creates.
- Access control — SSM provides an efficient solution to the access control problem. When a receiver subscribes to an (S,G) channel, it receives data sent only by the source S. In contrast, any host can transmit to an ASM host group. At the same time, when a sender picks a channel (S,G) to transmit on, it is automatically ensured that no other sender will be transmitting on the same channel (except in the case of malicious acts such as address spoofing). This makes it harder to spam an SSM channel than an ASM multicast group.

- Handling of well-known sources — SSM requires only source-based forwarding trees. This eliminates the need for a shared tree infrastructure. In terms of the IGMP, PIM-SM, MSDP, MBGP protocol suite, this implies that neither the RP-based shared tree infrastructure of PIM-SM nor the MSDP protocol is required. Thus, the complexity of the multicast routing infrastructure for SSM is low, making it viable for immediate deployment. Note that MBGP is still required for distribution of multicast reachability information.
- Anticipating that point-to-multipoint applications such as Internet TV will be significant in the future, the SSM model is better suited for such applications.

Multicast in IP-VPN Networks

Multicast can be deployed as part of IP-VPN networks. For details on multicast support in IP-VPNs see SROS Services Guide.

Multicast Features

This section describes the multicast requirements when an Alcatel-Lucent router is deployed as part of the user's core network.

The required protocol set is as follows:

- Internet Group Management Protocol (Internet Group Management Protocol on page 27)
- Source Specific Multicast Groups (SSM on page 28)
- Protocol Independent Multicast (Sparse Mode) (PIM-SM on page 30)
- Multicast Extensions to MBGP (Multicast Extensions to MBGP on page 51)

Internet Group Management Protocol

Internet Group Management Protocol (IGMP) is used by IPv4 hosts and routers to report their IP multicast group memberships to neighboring multicast routers. A multicast router keeps a list of multicast group memberships for each attached network, and a timer for each membership.

Multicast group memberships include at least one member of a multicast group on a given attached network, not a list of all of the members. With respect to each of its attached networks, a multicast router can assume one of two roles, querier or non-querier. There is normally only one querier per physical network.

A querier issues two types of queries, a general query and a group-specific query. General queries are issued to solicit membership information with regard to any multicast group. Group-specific queries are issued when a router receives a leave message from the node it perceives as the last group member remaining on that network segment.

Hosts wanting to receive a multicast session issue a multicast group membership report. These reports must be sent to all multicast enabled routers.

IGMP Versions and Interoperability Requirements

If routers run different versions of IGMP, they will negotiate the lowest common version of IGMP that is supported on their subnet and operate in that version.

Version 1 — Specified in RFC-1112, *Host extensions for IP Multicasting*, was the first widely deployed version and the first version to become an Internet standard.

Version 2 — Specified in RFC-2236, *Internet Group Management Protocol*, added support for "low leave latency", that is, a reduction in the time it takes for a multicast router to learn that there are no longer any members of a particular group present on an attached network.

Version 3 — Specified in RFC-3376, *Internet Group Management Protocol*, adds support for source filtering, that is, the ability for a system to report interest in receiving packets only from specific source addresses, as required to support Source-Specific Multicast (See Source Specific Multicast (SSM)), or from all but specific source addresses, sent to a particular multicast address.

IGMPv3 must keep state per group per attached network. This group state consists of a filter-mode, a list of sources, and various timers. For each attached network running IGMP, a multicast router records the desired reception state for that network.

IGMP Version Transition

Alcatel-Lucent's routers are capable of interoperating with routers and hosts running IGMPv1, IGMPv2, and/or IGMPv3. RFC 5186, *Internet Group Management Protocol Version 3 (IGMPv3)/ Multicast Listener Discovery Version 2 (MLDv2) and Multicast Routing Protocol Interaction* explores some of the interoperability issues and how they affect the various routing protocols.

IGMP version 3 specifies that if at any point a router receives an older version query message on an interface that it must immediately switch into a compatibility mode with that earlier version. Since none of the previous versions of IGMP are source aware, should this occur and the interface switch to Version 1 or 2 compatibility mode, any previously learned group memberships with specific sources (learned via the IGMPv3 specific INCLUDE or EXCLUDE mechanisms) MUST be converted to non-source specific group memberships. The routing protocol will then treat this as if there is no EXCLUDE definition present.

Source-Specific Multicast Groups

IGMPv3 permits a receiver to join a group and specify that it only wants to receive traffic for a group if that traffic comes from a particular source. If a receiver does this, and no other receiver on

the LAN requires all the traffic for the group, then the designated router (DR) can omit performing a (*,G) join to set up the shared tree, and instead issue a source-specific (S,G) join only.

The range of multicast addresses from 232.0.0.0 to 232.255.255.255 is currently set aside for source-specific multicast in IPv4. For groups in this range, receivers should only issue source-specific IGMPv3 joins. If a PIM router receives a non-source-specific join for a group in this range, it should ignore it.

An Alcatel-Lucent router PIM router must silently ignore a received (*,G) PIM join message where G is a multicast group address from the multicast address group range that has been explicitly configured for SSM. This occurrence should generate an event. If configured, the IGMPv2 request can be translated into IGMPv3. The router allows for the conversion of an IGMPv2 (*,G) request into a IGMPv3 (S,G) request based on manual entries. A maximum of 32 SSM ranges is supported.

IGMPv3 also permits a receiver to join a group and specify that it only wants to receive traffic for a group if that traffic does not come from a specific source or sources. In this case, the DR will perform a (*,G) join as normal, but can combine this with a prune for each of the sources the receiver does not wish to receive.

Query Messages

The IGMP query source address is configurable at two hierarchal levels. It can be configured globally at each router instance IGMP level and can be configured at individual at the group-interface level. The group-interface level overrides the src-ip address configured at the router instance level.

By default, subscribers with IGMP policies send IGMP queries with an all zero SRC IP address (0.0.0.0). However, some systems only accept and process IGMP query messages with non-zero SRC IP addresses. This feature allows the BNG to inter-operate with such systems.

Protocol Independent Multicast (PIM)

PIM-SM leverages the unicast routing protocols that are used to create the unicast routing table, OSPF, IS-IS, BGP, and static routes. Because PIM uses this unicast routing information to perform the multicast forwarding function it is effectively IP protocol independent. Unlike DVMRP, PIM does not send multicast routing tables updates to its neighbors.

PIM-SM uses the unicast routing table to perform the Reverse Path Forwarding (RPF) check function instead of building up a completely independent multicast routing table.

PIM-SM only forwards data to network segments with active receivers that have explicitly requested the multicast group. PIM-SM in the ASM model initially uses a shared tree to distribute information about active sources. Depending on the configuration options, the traffic can remain on the shared tree or switch over to an optimized source distribution tree. As multicast traffic starts to flow down the shared tree, routers along the path determine if there is a better path to the source. If a more direct path exists, then the router closest to the receiver sends a join message toward the source and then reroutes the traffic along this path.

As stated above, PIM-SM relies on an underlying topology-gathering protocol to populate a routing table with routes. This routing table is called the Multicast Routing Information Base (MRIB). The routes in this table can be taken directly from the unicast routing table, or it can be different and provided by a separate routing protocol such as MBGP. Regardless of how it is created, the primary role of the MRIB in the PIM-SM protocol is to provide the next hop router along a multicast-capable path to each destination subnet. The MRIB is used to determine the next hop neighbor to whom any PIM join/prune message is sent. Data flows along the reverse path of the join messages. Thus, in contrast to the unicast RIB that specifies the next hop that a data packet would take to get to some subnet, the MRIB gives reverse-path information, and indicates the path that a multicast data packet would take from its origin subnet to the router that has the MRIB.

PIM-SM Functions

PIM-SM functions in three phases:

- Phase One on page 31
- Phase Two on page 31
- Phase Three on page 32

Phase One

In this phase, a multicast receiver expresses its interest in receiving traffic destined for a multicast group. Typically it does this using IGMP or MLD, but other mechanisms might also serve this purpose. One of the receiver's local routers is elected as the DR for that subnet. When the expression of interest is received, the DR sends a PIM join message towards the RP for that multicast group. This join message is known as a (*,G) join because it joins group G for all sources to that group. The (*,G) join travels hop-by-hop towards the RP for the group, and in each router it passes through the multicast tree state for group G is instantiated. Eventually the (*,G) join either reaches the RP or reaches a router that already has (*,G) join state for that group. When many receivers join the group, their join messages converge on the RP and form a distribution tree for group G that is rooted at the RP. This is known as the RP tree and is also known as the shared tree because it is shared by all sources sending to that group. Join messages are resent periodically as long as the receiver remains in the group. When all receivers on a leaf-network leave the group, the DR will send a PIM (*,G) prune message towards the RP for that multicast group. However if the prune message is not sent for any reason, the state will eventually time out.

A multicast data sender starts sending data destined for a multicast group. The sender's local router (the DR) takes those data packets, unicast-encapsulates them, and sends them directly to the RP. The RP receives these encapsulated data packets, removes the encapsulation, and forwards them onto the shared tree. The packets then follow the (*,G) multicast tree state in the routers on the RP tree, being replicated wherever the RP tree branches, and eventually reaching all the receivers for that multicast group. The process of encapsulating data packets to the RP is called registering, and the encapsulation packets are known as PIM register packets.

At the end of phase one, multicast traffic is flowing encapsulated to the RP, and then natively over the RP tree to the multicast receivers.

Phase Two

In this phase, register-encapsulation of data packets is performed. However, register-encapsulation of data packets is unsuitable for the following reasons:

- Encapsulation and de-encapsulation can be resource intensive operations for a router to perform depending on whether or not the router has appropriate hardware for the tasks.
- Traveling to the RP and then back down the shared tree can cause the packets to travel a relatively long distance to reach receivers that are close to the sender. For some applications, increased latency is unwanted.

Although register-encapsulation can continue indefinitely, for these reasons, the RP will normally switch to native forwarding. To do this, when the RP receives a register-encapsulated data packet from source S on group G, it will normally initiate an (S,G) source-specific join towards S. This join message travels hop-by-hop towards S, instantiating (S,G) multicast tree state in the routers along the path. (S,G) multicast tree state is used only to forward packets for group G if those

packets come from source S. Eventually the join message reaches S's subnet or a router that already has (S,G) multicast tree state, and then packets from S start to flow following the (S,G) tree state towards the RP. These data packets can also reach routers with (*,G) state along the path towards the RP - if so, they can short-cut onto the RP tree at this point.

While the RP is in the process of joining the source-specific tree for S, the data packets will continue being encapsulated to the RP. When packets from S also start to arrive natively at the RP, the RP will be receiving two copies of each of these packets. At this point, the RP starts to discard the encapsulated copy of these packets and it sends a register-stop message back to S's DR to prevent the DR unnecessarily encapsulating the packets. At the end of phase 2, traffic will be flowing natively from S along a source-specific tree to the RP and from there along the shared tree to the receivers. Where the two trees intersect, traffic can transfer from the shared RP tree to the shorter source tree.

Note that a sender can start sending before or after a receiver joins the group, and thus, phase two may occur before the shared tree to the receiver is built.

Phase Three

In this phase, the RP joins back towards the source using the shortest path tree. Although having the RP join back towards the source removes the encapsulation overhead, it does not completely optimize the forwarding paths. For many receivers the route via the RP can involve a significant detour when compared with the shortest path from the source to the receiver.

To obtain lower latencies, a router on the receiver's LAN, typically the DR, may optionally initiate a transfer from the shared tree to a source-specific shortest-path tree (SPT). To do this, it issues an (S,G) Join towards S. This instantiates state in the routers along the path to S. Eventually this join either reaches S's subnet or reaches a router that already has (S,G) state. When this happens, data packets from S start to flow following the (S,G) state until they reach the receiver.

At this point the receiver (or a router upstream of the receiver) will be receiving two copies of the data - one from the SPT and one from the RPT. When the first traffic starts to arrive from the SPT, the DR or upstream router starts to drop the packets for G from S that arrive via the RP tree. In addition, it sends an (S,G) prune message towards the RP. The prune message travels hop-by-hop instantiating state along the path towards the RP indicating that traffic from S for G should NOT be forwarded in this direction. The prune message is propagated until it reaches the RP or a router that still needs the traffic from S for other receivers.

By now, the receiver will be receiving traffic from S along the shortest-path tree between the receiver and S. In addition, the RP is receiving the traffic from S, but this traffic is no longer reaching the receiver along the RP tree. As far as the receiver is concerned, this is the final distribution tree.

Encapsulating Data Packets in the Register Tunnel

Conceptually, the register tunnel is an interface with a smaller MTU than the underlying IP interface towards the RP. IP fragmentation on packets forwarded on the register tunnel is performed based upon this smaller MTU. The encapsulating DR can perform path-MTU discovery to the RP to determine the effective MTU of the tunnel. This smaller MTU takes both the outer IP header and the PIM register header overhead into consideration.

PIM Bootstrap Router Mechanism

For proper operation, every PIM-SM router within a PIM domain must be able to map a particular global-scope multicast group address to the same RP. If this is not possible, then black holes can appear (this is where some receivers in the domain cannot receive some groups). A domain in this context is a contiguous set of routers that all implement PIM and are configured to operate within a common boundary.

The bootstrap router (BSR) mechanism provides a way in which viable group-to-RP mappings can be created and distributed to all the PIM-SM routers in a domain. Each candidate BSR originates bootstrap messages (BSMs). Every BSM contains a BSR priority field. Routers within the domain flood the BSMs throughout the domain. A candidate BSR that hears about a higher-priority candidate BSR suppresses its sending of further BSMs for a period of time. The single remaining candidate BSR becomes the elected BSR and its BSMs inform the other routers in the domain that it is the elected BSR.

It is adaptive, meaning that if an RP becomes unreachable, it will be detected and the mapping tables will be modified so the unreachable RP is no longer used and the new tables will be rapidly distributed throughout the domain.

PIM-SM Routing Policies

Multicast traffic can be restricted from certain source addresses by creating routing policies. Join messages can be filtered using import filters. PIM join policies can be used to reduce denial of service attacks and subsequent PIM state explosion in the router and to remove unwanted multicast streams at the edge of the network before it is carried across the core. Route policies are created in the **config>router>policy-options** context. Join and register route policy match criteria for PIM-SM can specify the following:

- Router interface or interfaces specified by name or IP address.
- Neighbor address (the source address in the IP header of the join and prune message).
- Multicast group address embedded in the join and prune message.

- Multicast source address embedded in the join and prune message.

Join policies can be used to filter PIM join messages so no *,G or S,G state will be created on the router.

Table 3: Join Filter Policy Match Conditions

Match Condition	Matches the:
Interface	RTR interface by name
Neighbor	The neighbors source address in the IP header
Group Address	Multicast Group address in the join/prune message
Source Address	Source address in the join/prune message

PIM register message are sent by the first hop designated router that has a direct connection to the source. This serves a dual purpose:

- Notifies the RP that a source has active data for the group
- Delivers the multicast stream in register encapsulation to the RP and its potential receivers.
- If no one has joined the group at the RP, the RP will ignore the registers.

In an environment where the sources to particular multicast groups are always known, it is possible to apply register filters at the RP to prevent any unwanted sources from transmitting multicast stream. You can apply these filters at the edge so that register data does not travel unnecessarily over the network towards the RP.

Table 4: Register Filter Policy Match Conditions

Match Condition	Matches the:
Interface	RTR interface by name
Group Address	Multicast Group address in the join/prune message
Source Address	Source address in the join/prune message

Reverse Path Forwarding Checks

Multicast implements a reverse path forwarding check (RPF). RPF checks the path that multicast packets take between their sources and the destinations to prevent loops. Multicast requires that an incoming interface is the outgoing interface used by unicast routing to reach the source of the multicast packet. RPF forwards a multicast packet only if it is received on an interface that is used by the router to route to the source.

If the forwarding paths are modified due to routing topology changes then any dynamic filters that may have been applied must be re-evaluated. If filters are removed then the associated alarms are also cleared.

Anycast RP for PIM-SM

The implementation of Anycast RP for PIM-SM environments enable fast convergence when a PIM rendezvous point (RP) router fails by allowing receivers and sources to rendezvous at the closest RP. It allows an arbitrary number of RPs per group in a single shared-tree protocol Independent Multicast-Sparse Mode (PIM-SM) domain. This is, in particular, important for triple play configurations that opt to distribute multicast traffic using PIM-SM, not SSM. In this case, RP convergence must be fast enough to avoid the loss of multicast streams which could cause loss of TV delivery to the end customer.

Anycast RP for PIM-SM environments is supported in the base routing/PIM-SM instance of the service router. In the 7750 SR product lines, this feature is supported in Layer 3-VPRN instances that are configured with PIM.

Implementation

The Anycast RP for PIM-SM implementation is defined in *draft-ietf-pim-anycast-rp-03*, *Anycast-RP using PIM*, and is similar to that described in RFC 3446, *Anycast RP Mechanism Using PIM and MSDP*, and extends the register mechanism in PIM so Anycast RP functionality can be retained without using Multicast Source Discovery Protocol (MSDP) (see Multicast in Virtual Private Networks on page 49).

The mechanism works as follows:

- An IP address is chosen to use as the RP address. This address is statically configured, or distributed using a dynamic protocol, to all PIM routers throughout the domain.
- A set of routers in the domain are chosen to act as RPs for this RP address. These routers are called the Anycast-RP set.
- Each router in the Anycast-RP set is configured with a loopback interface using the RP address.
- Each router in the Anycast-RP set also needs a separate IP address to be used for communication between the RPs.
- The RP address, or a prefix that covers the RP address, is injected into the unicast routing system inside of the domain.
- Each router in the Anycast-RP set is configured with the addresses of all other routers in the Anycast-RP set. This must be consistently configured in all RPs in the set.

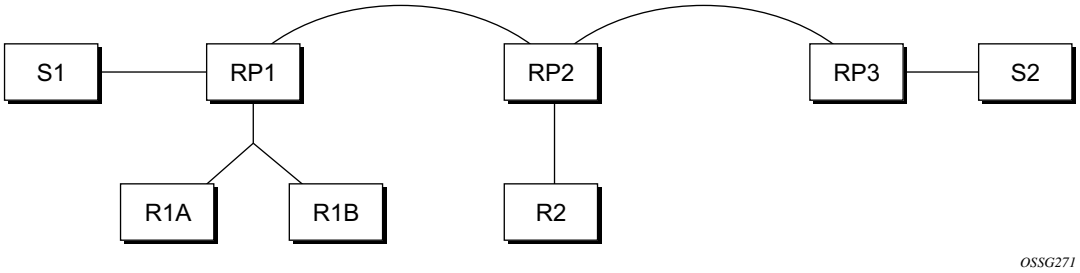

Figure 1: Anycast RP for PIM-SM Implementation Example

Assume the scenario in Figure 1 is completely connected where R1A, R1B, and R2 are receivers for a group, and S1 and S2 send to that group. Assume RP1, RP2, and RP3 are all assigned the same IP address which is used as the Anycast-RP address (for example, the IP address is RPA).

Note, the address used for the RP address in the domain (the Anycast-RP address) must be different than the addresses used by the Anycast-RP routers to communicate with each other.

The following procedure is used when S1 starts sourcing traffic:

- S1 sends a multicast packet.
- The DR directly attached to S1 will form a PIM register message to send to the Anycast-RP address (RPA). The unicast routing system will deliver the PIM register message to the nearest RP, in this case RP1A.
- RP1 will receive the PIM register message, de-encapsulate it, send the packet down the shared-tree to get the packet to receivers R1A and R1B.
- RP1 is configured with RP2 and RP3's IP address. Since the register message did not come from one of the RPs in the anycast-RP set, RP1 assumes the packet came from a DR. If the register message is not addressed to the Anycast-RP address, an error has occurred and it should be rate-limited logged.
- RP1 will then send a copy of the register message from S1's DR to both RP2 and RP3. RP1 will use its own IP address as the source address for the PIM register message.
- RP1 may join back to the source-tree by triggering a (S1,G) Join message toward S1. However, RP1 must create (S1,G) state.
- RP2 receives the register message from RP1, de-encapsulates it, and also sends the packet down the shared-tree to get the packet to receiver R2.
- RP2 sends a register-stop message back to the RP1. RP2 may wait to send the register-stop message if it decides to join the source-tree. RP2 should wait until it has received data from the source on the source-tree before sending the register-stop message. If RP2

decides to wait, the register-stop message will be sent when the next register is received. If RP2 decides not to wait, the register-stop message is sent now.

- RP2 may join back to the source-tree by triggering a (S1,G) Join message toward S1. However, RP2 must create (S1,G) state.
- RP3 receives the register message from RP1, de-encapsulates it, but since there are no receivers joined for the group, it can discard the packet.
- RP3 sends a register-stop message back to the RP1.
- RP3 creates (S1,G) state so when a receiver joins after S1 starts sending, RP3 can join quickly to the source-tree for S1.
- RP1 processes the register-stop message from each of RP2 and RP3. RP1 may cache on a per-RP/per-(S,G) basis the receipt of register-stop message messages from the RPs in the anycast-RP set. This option is performed to increase the reliability of register message delivery to each RP. When this option is used, subsequent register messages received by RP1 are sent only to the RPs in the Anycast-RP set which have not previously sent register-stop message messages for the (S,G) entry.
- RP1 sends a register-stop message back to the DR the next time a register message is received from the DR and (when the option in the last bullet is in use) if all RPs in the Anycast-RP set have returned register-stop messages for a particular (S,G) route.

The procedure for S2 sending follows the same as above but it is RP3 which sends a copy of the register originated by S2's DR to RP1 and RP2. Therefore, this example shows how sources anywhere in the domain, associated with different RPs, can reach all receivers, also associated with different RPs, in the same domain.

Distributing PIM Joins over Multiple ECMP Paths

Commonly used multicast load-balancing method is per bandwidth/round robin, but the interface in an ECMP set can also be used for a particular channel to be predictable without knowing anything about the other channels using the ECMP set.

The **mc-ecmp-hashing-enabled** command enables PIM joins to be distributed over the multiple ECMP paths based on a hash of S and G. When a link in the ECMP set is removed, the multicast streams that were using that link are re-distributed over the remaining ECMP links using the same hash algorithm. When a link is added to the ECMP set, new joins may be allocated to the new link based on the hash algorithm, but existing multicast streams using the other ECMP links stay on those links until they are pruned.

The default is **no mc-ecmp-hashing-enabled**, which means that the use of multiple ECMP paths (if enabled at the config>service>vprn context) is controlled by the existing implementation and CLI commands, that is, **mc-ecmp-balance**.

The **mc-ecmp-hasing-enabled** command is mutually exclusive with the **mc-ecmp-balance** command in the same context.

To achieve distribution of streams across the ECMP links, following are the hashings steps:

1. For a given S, G get all possible nHops.
2. Sort these nHops based on nhops address.
3. xor S and G addresses.
4. Hash the xor address over number of pim next hops.
5. Use the hash value obtained in stip 4, and get that element, in the sorted list, we obtained in step 2 as the preferred nHop.
6. If this element is not available/is not a pim Next hop (pim neighbor), the next available next hop is chosen.

The following example displays pim status indicating ECMP Hashing is disabled

```
*B:BB# show router 100 pim status

===============================================================================
PIM Status ipv4
===============================================================================
Admin State                       : Up
Oper State                        : Up

IPv4 Admin State                  : Up
IPv4 Oper State                   : Up

BSR State                         : Accept Any

Elected BSR
    Address                       : None
    Expiry Time                   : N/A
    Priority                      : N/A
    Hash Mask Length              : 30
    Up Time                       : N/A
    RPF Intf towards E-BSR        : N/A

Candidate BSR
    Admin State                   : Down
    Oper State                    : Down
    Address                       : None
    Priority                      : 0
    Hash Mask Length              : 30

Candidate RP
    Admin State                   : Down
    Oper State                    : Down
    Address                       : 0.0.0.0
    Priority                      : 192
    Holdtime                      : 150

SSM-Default-Range                 : Enabled
```

```
SSM-Group-Range
    None

MC-ECMP-Hashing                  : Disabled

Policy                           : None

RPF Table                        : rtable-u

Non-DR-Attract-Traffic           : Disabled
===============================================================================

-----------------------------------------------
*B:BB>config>service>vprn>pim# no mc-ecmp-balance mc-ecmp-balance mc-ecmp-balance-hold
*B:BB>config>service>vprn>pim# no mc-ecmp-balance
*B:BB>config>service>vprn>pim# mc-ecmp-mc-ecmp-balance mc-ecmp-balance-hold mc-ecmp-hash-
ing-enabled
*B:BB>config>service>vprn>pim# mc-ecmp-hashing-enabled
*B:BB>config>service>vprn>pim# info
-----------------------------------------------
                apply-to all
                rp
                    static
                        address 3.3.3.3
                            group-prefix 224.0.0.0/4
                        exit
                    exit
                    bsr-candidate
                        shutdown
                    exit
                    rp-candidate
                        shutdown
                    exit
                exit
                no mc-ecmp-balance
                mc-ecmp-hashing-enabled
-----------------------------------------------
*B:BB>config>service>vprn>pim#
apply-to        - Create/remove interfaces in PIM
 [no] import          - Configure import policies
 [no] interface       + Configure PIM interface
 [no] mc-ecmp-balance - Enable/Disable multicast balancing of traffic over ECMP links
 [no] mc-ecmp-balanc* - Configure hold time for multicast balancing over ECMP links
 [no] mc-ecmp-hashin* - Enable/Disable hash based multicast balancing of traffic over ECMP
links
 [no] non-dr-attract* - Enable/disable attracting traffic when not DR
      rp              + Configure the router as static or Candidate-RP
 [no] shutdown        - Administratively enable or disable the operation of PIM
 [no] spt-switchover* - Configure shortest path tree (spt tree) switchover threshold for a
group prefix
 [no] ssm-default-ra* - Enable the disabling of SSM Default Range
 [no] ssm-groups      + Configure the SSM group ranges
```

The following example shows distribution of PIM joins over multiple ECMP paths.

```
*A:BA# show router 100 pim group
```

```
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address                          Type     Spt Bit Inc Intf       No.Oifs
   Source Address                        RP
-------------------------------------------------------------------------------
225.1.1.1                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.1.1.2                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6
225.1.1.3                              (S,G)    spt     to_C2          1
   170.0.100.33                          10.20.1.6
225.1.1.4                              (S,G)    spt     to_C1          1
   170.0.100.33                          10.20.1.6
225.1.1.5                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.1.1.6                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6

225.2.1.1                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.2.1.2                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6
225.2.1.3                              (S,G)    spt     to_C2          1
   170.0.100.33                          10.20.1.6
225.2.1.4                              (S,G)    spt     to_C1          1
   170.0.100.33                          10.20.1.6
225.2.1.5                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.2.1.6                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6

225.3.1.1                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.3.1.2                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6
225.3.1.3                              (S,G)    spt     to_C2          1
   170.0.100.33                          10.20.1.6
225.3.1.4                              (S,G)    spt     to_C1          1
   170.0.100.33                          10.20.1.6
225.3.1.5                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.3.1.6                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6

225.4.1.1                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.4.1.2                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6
225.4.1.3                              (S,G)    spt     to_C2          1
   170.0.100.33                          10.20.1.6
225.4.1.4                              (S,G)    spt     to_C1          1
   170.0.100.33                          10.20.1.6
225.4.1.5                              (S,G)    spt     to_C0          1
   170.0.100.33                          10.20.1.6
225.4.1.6                              (S,G)    spt     to_C3          1
   170.0.100.33                          10.20.1.6
-------------------------------------------------------------------------------
```

```
Groups : 24
===============================================================================
```

PIM Interface on IES Subscriber Group Interfaces

PIM on a subscriber group interface allows for SAP-level replication over an ESM Group interface by establishing PIM adjacency to a downstream router. The following picture depicts the model:

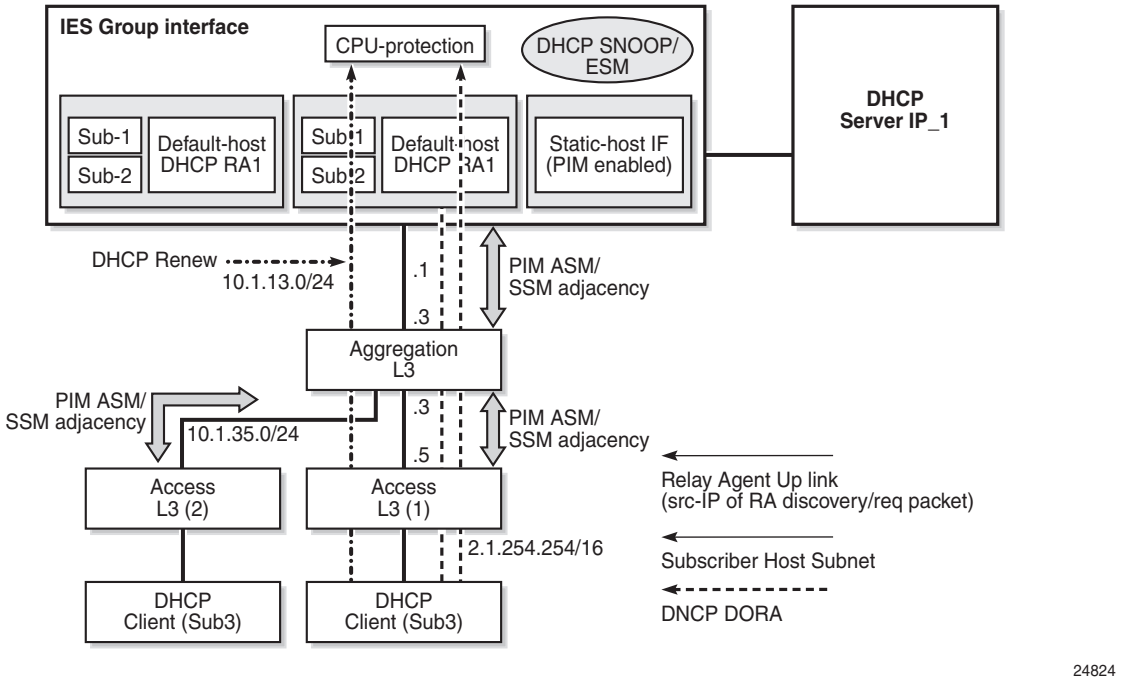

Figure 2: PIM Interface on IES Subscriber Group Interface

On an IES subscriber-interface, an Ethernet SAP is configured (LAG or physical port). On the SAP, a static-host is configured for connectivity to downstream Layer 3 aggregation devices (including PIM adjacency) while multiple default-hosts can be configured for subscriber traffic. Single SAP with a single static-host per group interface is supported to establish PIM adjacency on a given subscriber group interface. Both IPv4 PIM ASM and SSM are supported.

Feature caveats:

- Only IPv4 PIM is supported with a single static host used to form a PIM interface under a group interface. Using multiple hosts or non-static hosts is not supported. Configuring IPv6-related parameters in **configure**>**router**>**pim**>**interface** *group-ift* is not blocked, but takes no effect.

- **config**>**router**>**pim**>**apply-to** configuration does not apply to PIM interfaces on IES subscriber group interfaces.
- PIM on group interfaces is not supported in VPRN context.
- Extranet is not supported.
- Locally attached receivers are not supported (no IGMP/MLD and PIM mix in OIF list).
- Default anti-spoofing must be configured (IP+MAC).
- A subscriber profile with pim-policy enabled cannot combine with the following policies (**config**>**subscr-mgmt**>**sub-prof**):
  - → [**no**] **host-tracking** — Apply a host tracking policy
  - → [**no**] **igmp-policy** — Apply an IGMP policy
  - → [**no**] **mld-policy** — Apply an MLD policy
  - → [**no**] **nat-policy** — Apply a NAT policy
  - → [**no**] **sub-mcac-policy** — Apply a subscriber MCAC policy (MCAC policy can be used when configured in PIM interface context)
- The feature is supported on IOM3-XP or newer line cards. When enabling the feature on older hardware, joins may be accepted and an outgoing interface may be created for the group, but traffic will not be sent out on egress because no OIF is created in forwarding.

Multicast Only Fast Reroute (MoFRR)

With large scale multicast deployments, a link or nodal failure impacts multiple subscribers or a complete region/segment of receivers. This failure interrupts the receiver client experience. Besides the impact on user experience, though multicast client applications may buffer streams for short period of time, the loss of stream data may trigger unicast request for the missing stream data to the source in certain middleware implementations. Those requests can overload the network resources, if a traffic loss persists for a prolonged period.

To minimize service interruption to end-users and protect the network from sudden surge of unicast requests, SROS implements a fast failover scheme for native IP networks. SROS MoFRR implementation is based on http://tools.ietf.org/html/draft-karan-mofrr-02 and relies on:

- Sending a JOIN to a primary and a single standby upstream nodes over disjoined paths.
- Fast failover to a standby stream upon detection of a failure.

The functionality relies on failure detection on the primary path to switch to forwarding the traffic from the standby path. The traffic failure can happen with or without physical links or nodes going down. Various mechanisms for link/node failure detections are supported; however, to achieve best performance and resilience, it is recommended to enable MoFRR on every node in the network and use hop-by-hop BFD for fast link failure or data plane failure detection on each

upstream link. Without BFD, the PIM adjacency loss or route change could be used to detect traffic failure. Figure 3 and Figure 4 depict MoFRR behavior.

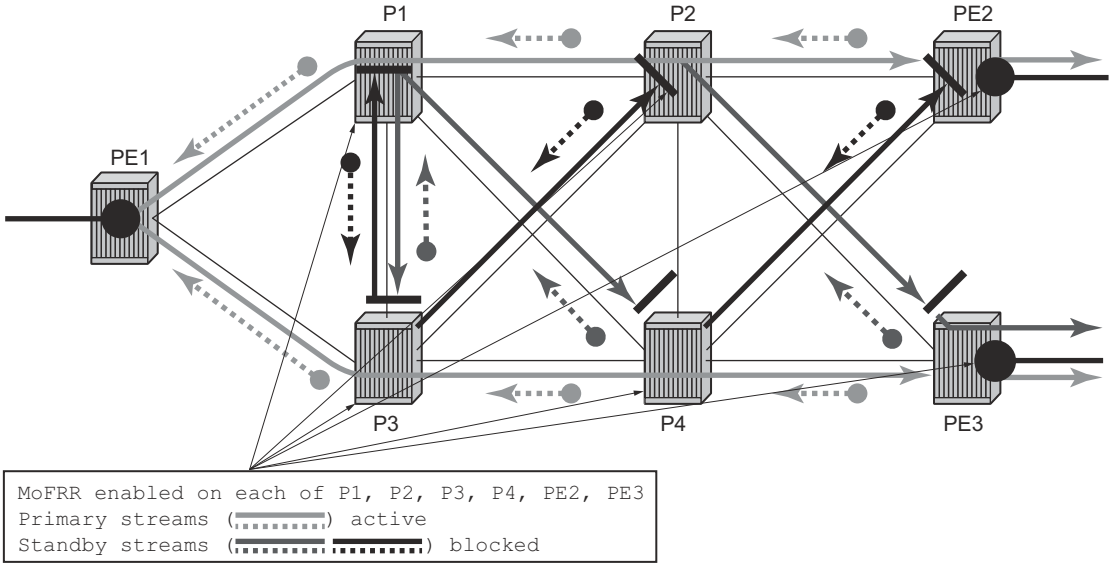

Figure 3: MoFRR Steady State No Failure

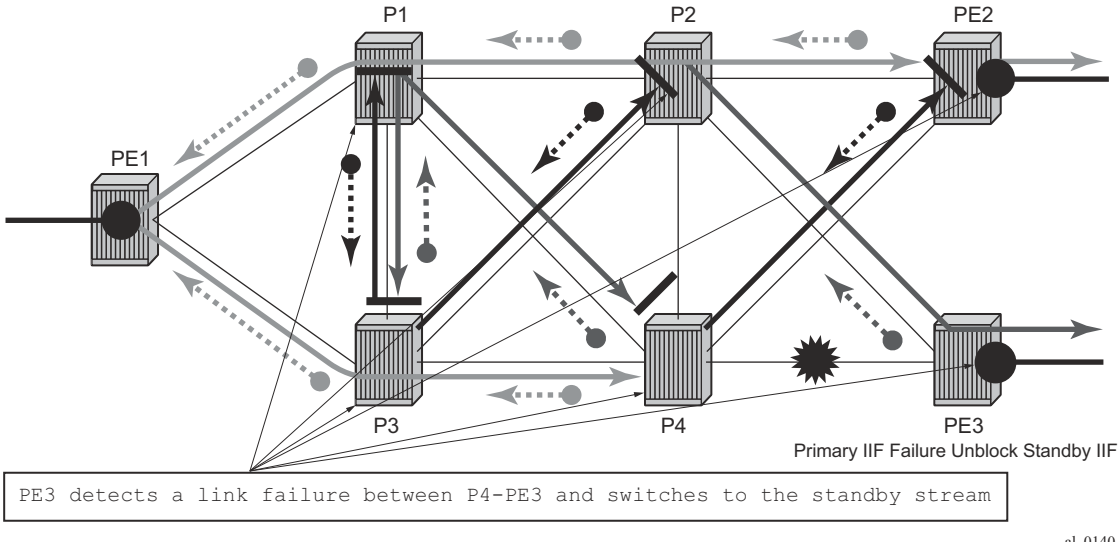

Figure 4: MoFRR Switch to Standby Stream on a Link Failure

The MoFRR functionality on SROS routers supports the following:

- IPv4 link/node failure protection in global routing instance.
- Rosen PIM SSM with MDT SAFI

- Active and a single standby stream JOINs L3 over disjoint ECMP paths
- Active and a single standby stream JOINs over ISIS/OSPF Loop Free Alternative paths.
- When enabled, MoFRR is enabled on all regular PIM interfaces supporting MoFRR for all multicast streams. Tunnel interfaces are ignored.

Multicast Source Discovery Protocol (MSDP)

MSDP-speaking routers in a PIM-SM (RFC 2362, *Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification*) domain have MSDP peering relationship with MSDP peers in another domain. The peering relationship is made up of a TCP connection in which control information is exchanged. Each domain has one or more connections to this virtual topology.

When a PIM-SM RP learns about a new multicast source within its own domain from a standard PIM register mechanism, it encapsulates the first data packet in an MSDP source-active message and sends it to all MSDP peers.

The source-active message is flooded (after an RPF check) by each peer to its MSDP peers until the source-active message reaches every MSDP router in the interconnected networks. If the receiving MSDP peer is an RP, and the RP has a (*.G) entry (receiver) for the group, the RP creates state for the source and joins to the shortest path tree for the source. The encapsulated data is de-encapsulated and forwarded down the shared tree of that RP. When the packet is received by the last hop router of the receiver, the last hop router also may join the shortest path tree to the source.

The MSDP speaker periodically sends source-active messages that include all sources.

Anycast RP for MSDP

MSDP is a mechanism that allows rendezvous points to share information about active sources. When RPs in remote domains hear about the active sources, they can pass on that information to the local receivers and multicast data can be forwarded between the domains. MSDP allows each domain to maintain an independent RP that does not rely on other domains but enables RPs to forward traffic between domains. PIM-SM is used to forward the traffic between the multicast domains.

Using PIM-SM, multicast sources and receivers register with their local RP by the closest multicast router. The RP maintains information about the sources and receivers for any particular group. RPs in other domains do not have any knowledge about sources located in other domains.

MSDP is required to provide inter-domain multicast services using Any Source Multicast (ASM). Anycast RP for MSDP enables fast convergence when should an MSDP/PIM PR router fail by allowing receivers and sources to rendezvous at the closest RP.

MSDP Procedure

When an RP in a PIM-SM domain first learns of a new sender, for example, by PIM register messages, it constructs a source-active (SA) message and sends it to its MSDP peers. The SA message contains the following fields:

- Source address of the data source
- Group address the data source sends to
- IP address of the RP

Note that an RP that is not a designated router on a shared network do not originate SAs for directly-connected sources on that shared network. It only originates in response to receiving register messages from the designated router.

Each MSDP peer receives and forwards the message away from the RP address in a peer-RPF flooding fashion. The notion of peer-RPF flooding is with respect to forwarding SA messages. The Multicast RPF Routing Information Base (MRIB) is examined to determine which peer towards the originating RP of the SA message is selected. Such a peer is called an RPF peer.

If the MSDP peer receives the SA from a non-RPF peer towards the originating RP, it will drop the message. Otherwise, it forwards the message to all its MSDP peers (except the one from which it received the SA message).

When an MSDP peer which is also an RP for its own domain receives a new SA message, it determines if there are any group members within the domain interested in any group described by an (S,G) entry within the SA message. That is, the RP checks for a (*,G) entry with a non-empty outgoing interface list. This implies that some system in the domain is interested in the group. In this case, the RP triggers an (S,G) join event toward the data source as if a join/prune message was received addressed to the RP. This sets up a branch of the source-tree to this domain. Subsequent data packets arrive at the RP by this tree branch and are forwarded down the shared-tree inside the domain. If leaf routers choose to join the source-tree they have the option to do so according to existing PIM-SM conventions. If an RP in a domain receives a PIM join message for a new group G, the RP must trigger an (S,G) join event for each active (S,G) for that group in its SA cache.

This procedure is called flood-and-join because if any RP is not interested in the group, the SA message can be ignored, otherwise, they join a distribution tree.

MSDP Peering Scenarios

Draft-ietf-mboned-msdp-deploy-nn.txt, *Multicast Source Discovery Protocol (MSDP) Deployment Scenarios,* describes how protocols work together to provide intra- and inter-domain ASM service.

Inter-domain peering:

- Peering between PIM border routers (single-hop peering)
- Peering between non-border routers (multi-hop peering)
- MSDP peering without BGP
- MSDP peering between mesh groups
- MSDP peering at a multicast exchange

Intra-domain peering:

- Peering between routers configured for both MSDP and MBGP
- MSDP peer is not BGP peer (meaning, no BGP peer)

MSDP Peer Groups

MSDP peer groups are typically created when multiple peers have a set of common operational parameters. Group parameters not specifically configured are inherited from the global level.

MSDP Mesh Groups

MSDP mesh groups are used to reduce source active flooding primarily in intra-domain configurations. When a number of speakers in an MSDP domain are fully meshed they can be configured as a mesh group. The originator of the source active message forwards the message to all members of the mesh group. Because of this, forwarding the SA between non-originating members of the mesh group is not necessary.

MSDP Routing Policies

MSDP routing policies allow for filtering of inbound and/or outbound active source messages. Policies can be configured at different levels:

- Global level — Applies to all peers
- Group level — Applies to all peers in peer-group
- Neighbor level — Applies only to specified peer

The most specific level is used. If multiple policy names are specified, the policies are evaluated in the order they are specified. The first policy that matches is applied. If no policy is applied source active messages are passed.

Match conditions include:

- Neighbor — Matches on a neighbor address is the source address in the IP header of the source active message.
- Route filter — Matches on a multicast group address embedded in the source active message
- Source address filter — Matches on a multicast source address embedded in the source active message

Auto-RP (discovery mode only) in Multicast VPN

Auto-RP is a vendor proprietary protocol to dynamically learn about availability of Rendezvous Point (RP) in network. Auto-RP protocol consists of announcing, mapping and discovery functions. SROS supports the discovery mode of Auto-RP that includes mapping and forwarding of RP-mapping and RP-candidate messages. Discovery mode also includes receiving RP-mapping messages locally to learn and maintain RP-candidate database.

Auto-RP protocol is supported with multicast VPN and global routing instance. Either BSR or Auto-RP is allowed to be configured per routing instance. Both mechanisms cannot be enabled together.

Multicast in Virtual Private Networks

Draft Rosen

RFC2547bis, *BGP/MPLS IP VPNs*, describes a method of providing a VPN service. A VPN provides secure connections to the network, allowing more efficient service to remote users without compromising the security of firewalls. The Rosen draft specifies the protocols and procedures which must be implemented in order for a service provider to provide a unicast VPN. The draft extends that specification by describing the protocols and procedures which a service provider must implement in order to support multicast traffic in a VPN, assuming that PIM [PIMv2] is the multicast routing protocol used within the VPN, and the SP network can provide PIM as well.

IGMP is not supported for receivers or senders directly attached to the PE.

For further information, refer to the Virtual Private Routed Network Service section of the Services Guide.

Dynamic Multicast Signaling over P2MP in GRT Instance

This feature provides a flexible multicast signaling solution to connect native IP multicast source and receivers running PIM multicast protocol via an intermediate MPLS (P2MP LDP LSP) network. The feature allows each native IP multicast flow to be connected via an intermediate P2MP LSP by dynamically mapping each PIM multicast flow to a P2MP LDP LSP.

The feature uses procedures defined in RFC 6826: Multipoint LDP In-Band Signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths. On the leaf node of a P2MP LSP, PIM signaling is dynamically mapped to P2MP LDP tree setup. On the root node of P2MP LSP, P2MP LDP signaling is handed back to PIM. Due to dynamic mapping of multicast IP flow to P2MP LSP, provisioning and maintenance overhead is eliminated as multicast distribution services are added and removed from the network. Per (S, G) IP multicast state is also removed from the network where P2MP LSPs are used to transport multicast flows.

Figure 5 illustrates dynamic mLDP signaling for IP multicast in GRT.

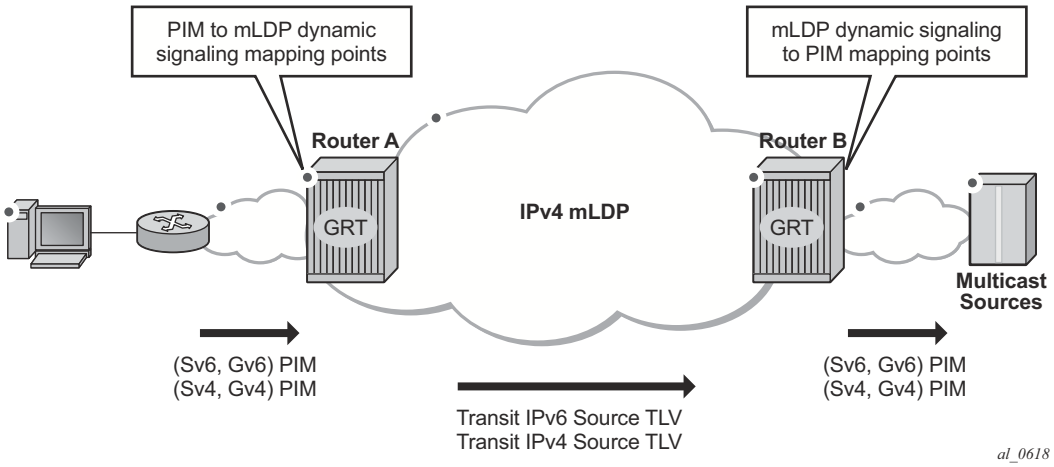

Figure 5: Dynamic mLDP Signaling for IP Multicast in GRT

As illustrated in Figure 5, P2MP LDP LSP signaling is initiated from the router that receives PIM JOIN from a downstream router (Router A). To enable dynamic multicast signaling, p2mp-ldp-tree-join must be configured on PIM outgoing interface of Router A. This enables handover of multicast tree signaling from PIM to P2MP LDP LSP. Being a leaf node of P2MP LDP LSP, Router A selects the upstream-hop as the root node of P2MP LDP FEC based on routing table lookup. If an ECMP path is available for a given route, then the number of trees are equally balanced towards multiple root nodes. The PIM Joins are carried in Transit IPv4 (IPv4 PIM SSM) or IPv6 (IPv6 PIM SSM) mLDP TLVs. On the root node of P2MP LDP LSP (Router B), multicast tree signaling is handed back to PIM and propagated upstream as native-IP PIM JOIN.

The feature is supported with IPv4 and IPv6 PIM SSM and IPv4 mLDP. Directly connected IGMP/MLD receivers are also supported with PIM enabled on outgoing interfaces and SSM mapping configured if required.

If multiple criteria exist to setup a multicast flow, the following priority is given:

1. Multicast (statically provisioned) over P2MP LSP (RSVP-TE or LDP)
2. Dynamic multicast signaling over P2MP LDP
3. PIM native-IP multicast

The following are feature caveats:

- A single instance of P2MP LDP LSP is supported between the root and leaf nodes per multicast flow; there is no stitching of dynamic trees.
- Extranet functionality is not supported.
- The router LSA link ID or the advertising router ID must be a routable IPv4 address (including IPv6 into IPv4 mLDP use cases).
- IPv6 PIM with dynamic IPv4 mLDP signaling is not supported with e-BGP or i-BGP with IPv6 next-hop.
- Inter-AS and IGP inter-area scenarios where the originating router is altered at the ASBR and ABR respectively, (hence PIM has no way to create the LDP LSP towards the source), are not supported.
- The feature requires chassis mode C.

Multicast Extensions to MBGP

This section describes the implementation of extensions to MBGP to support multicast. Rather than assuming that all unicast routes are multicast-capable, some routed environments, in some cases, some ISPs do not support or have limited support for multicast throughout their AS.

BGP is capable of supporting two sets of routing information, one set for unicast routing and the other for multicast routing. The unicast and multicast routing sets either partially or fully overlay one another. To achieve this, BGP has added support for IPv4 and mcast-IPv4 address families. Routing policies can be imported or exported.

The multicast routing information can subsequently be used by the Protocol Independent Multicast (PIM) protocol to perform its Reverse Path Forwarding (RPF) lookups for multicast-capable sources. Thus, multicast traffic can only be routed across a multicast topology and not a unicast topology.

MBGP Multicast Topology Support

Recursive Lookup for BGP Next Hops

The next hop for multicast RPF routes learned by MBGP is not always the address of a directly-connected neighbor. For unicast routing, a router resolves the directly-connected next-hop by repeating the IGP routes. For multicast RPF routes, there are different ways to find the real next-hops.

- Scanning to see if a route encompasses the BGP next hop. If one exists, this route is used. If not, the tables are scanned for the best matching route.
- Check to see if the recursed next hop is taken from the protocol routing table with the lowest administrative distance (protocol preference). This means that the operating system algorithm must preform multiple lookups in the order of the lowest admin distance. Note that unlike recursion on the unicast routing table, the longest prefix match rule does not take effect; protocol preference is considered prior to prefix length. For example, the route 12.0.0.0/14 learned via MBGP will be selected over the route 12.0.0.0/16 learned via BGP.

IPv6 Multicast

IPv6 multicast enables multicast applications over native IPv6 networks. There are two service models: Any Source Multicast (ASM) and Source Specific Multicast (SSM) which includes PIM SSM and MLD (v1 and v2). SSM does not require source discovery and only supports single source for a specific multicast stream. As a result, SSM is easier to operate in a large scale deployment that uses the one-to-many service model.

Multicast Listener Discovery (MLD v1 and v2)

MLD is the IPv6 version of IGMP. The purpose of MLD is to allow each IPv6 router to discover the presence of multicast listeners on its directly attached links, and to discover specifically which multicast groups are of interest to those neighboring nodes.

MLD is a sub-protocol of ICMPv6. MLD message types are a subset of the set of ICMPv6 messages, and MLD messages are identified in IPv6 packets by a preceding Next Header value of 58. All MLD messages are sent with a link-local IPv6 source address, a Hop Limit of 1, and an IPv6 Router Alert option in the Hop-by-Hop Options header.

Similar to IGMPv2, MLDv1 reports only include the multicast group addresses that listeners are interested in, and don't include the source addresses. In order to work with PIM SSM model, a similar SSM translation function is required when MLDv1 is used.

SSM translation allows an IGMPv2 device to join an SSM multicast network through the router that provides such a translation capability. Currently SSM translation can done at a box level, but this does not allow a per-interface translation to be specified. SSM translation per interface offers the ability to have a same (*,G) mapped to two different (S,G) on two different interfaces to provide flexibility.

MLDv2 is backward compatible with MLDv1 and adds the ability for a node to report interest in listening to packets with a particular multicast group only from specific source addresses or from all sources except for specific source addresses.

PIM SSM

The IPv6 address family for SSM model is supported. This includes the ability to choose which RTM table to use (unicast RTM, multicast RTM, or both). OSPF3, IS-IS and static-route have extensions to support submission of routes into the IPv6 multicast RTM.

IPv6 PIM ASM

IPv6 PIM ASM is supported. All PIM ASM related functions such as bootstrap router, RP, etc., support both IPv4 and IPv6 address-families. IPv6 specific parameters are configured under **configure>router>pim>rp>ipv6**.

Embedded RP

The detailed protocol specification is defined in RFC 3956, *Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address*. This RFC describes a multicast address allocation policy in which the address of the RP is encoded in the IPv6 multicast group address, and specifies a PIM-SM group-to-RP mapping to use the encoding, leveraging, and extending unicast-prefix-based addressing. This mechanism not only provides a simple solution for IPv6 inter-domain ASM but can be used as a simple solution for IPv6 intra-domain ASM with scoped multicast addresses as well. It can also be used as an automatic RP discovery mechanism in those deployment scenarios that would have previously used the Bootstrap Router protocol (BSR).

Multicast Connection Admission Control (MCAC)

Multicast Connection Admission Control (MCAC) allows a router to limit bandwidth used by multicast channels, either on a router or on access links, by controlling the number of channels that are accepted. When a pre-configured limit is reached, the router prevents receivers from joining any new channels not currently established. As result, running the MCAC function might cause some channels to be temporarily unavailable to receivers under overload. However, by rejecting new channel establishment during an overload condition, the degradation of the quality of the existing multicast service offering is avoided.

Operators can configure one or more MCAC policies (configure>router>mcac) to specify multicast channel admission rules and then reference a required policy on multicast-enabled IPv4 and IPv6 interfaces or group-interfaces. In addition operators can configure per-interface MCAC behavior.

Multicast CAC is supported on ESM subscriber interfaces as well as multicast interfaces in base router instance and in MVPNs. MCAC is supported for IGMP, IGMP-snooping, MLD, and PIM. When a MCAC policy is applied to a split horizon group, then member SAPs do not permit policy enforcement configurations.

Feature caveats:

- MCAC is not supported with PIM snooping and MLD snooping

MCAC Policy Overview

MCAC policy is used to define MCAC rules to be applied on an interface when receivers are trying to join multicast channels. Within each policy, an operator can define:

- Multicast channel:
  - → A channel can be defined using multicast group address only or both source and group addresses. Ranges can be used to group multiple multicast channels into a single MCAC channel. When ranges are used, each multicast channel within range will use the same channel BW, class, and priority configuration.
  - → Channel BW: a bandwidth value to be used for a channel in MCAC.
  - → Channel type (mandatory or optional): mandatory channels have BW pre-reserved on interfaces as soon as they are defined in MCAC policy, while optional channels consume BW on-demand; only when there are active receivers for that channel and the remaining BW allows for channels to be admitted.

→ Channel class: two classes are supported: high and low. For LAG interfaces, the class parameter allows further prioritizing of the mandatory or optional channels. This brings the number of priority levels to four during reshuffles of the joined channels when LAG ports are changing state.

NOTE: Multicast channels not specified in an MCAC policy applicable on a given interface are not subject to MCAC. Treatment of such channels is configurable as either accept or discard.

- Multicast channel bundle:
  - Multicast bundle defines multicast channels as described above. A channel can only be part of one bundle.
  - Maximum bundle BW – the maximum bandwidth the channels forming a given bundle can consume on an interface.
  - MCAC constraints – set of rules governing available BW for multicast channels over LAG as LAG ports are changing state.

MCAC Algorithm

It is important to point out that the MCAC algorithm is based on configured BW values. The configured channel BW based on MCAC policy is CAC-ed against pre-configured maximum bundle BW and pre-configured interface multicast BW limits. A channel must pass all levels of CAC before it is accepted. The statements outline the CAC algorithm for a multicast channel defined in MCAC policy:

A join for a particular multicast channel is accepted if:

1. Mandatory channels:

   A sufficient bandwidth exists on the interface according to the policy settings for the interface (Interface-level MCAC) and BW setting for a channel (Bundle-level MCAC). Note, there is always sufficient BW available on the bundle level, because mandatory channels get pre-reserved bandwidth.

2. Optional channels:

   A sufficient BW exists on both interface (Interface-level MCAC) and bundle level (Bundle-Level MCAC) based on channel configured BW and currently available BW on both interface and bundle.

When a policy is evaluated over a set of existing channels (adding policy MCAC on LAG), the channels are evaluated and admitted/dropped based on the following priority order: mandatory-high, mandatory-low, optional-high, optional low.

This method does not guarantee that all bundles are fully allocated while others are not. However it does ensure that all mandatory high channels are allocated before any mandatory lows ones are allocated.

Interface-level MCAC details

Interface-level MCAC constraints are applied to the interface on which the join was received. The channel is allowed when:

- If it is defined as mandatory and the bandwidth for the already accepted mandatory channels plus the bandwidth of this channel is not greater than the configured mandatory bandwidth on this interface.
- If it is defined optional and the bandwidth for the already accepted optional channels plus the bandwidth of this channel is not greater than the configured amount of unconstrained bandwidth less the configured amount of mandatory bandwidth on this interface.

Bundle-Level MCAC details

Bundle-level CAC is applied to the bundle to which the channel belongs that triggered the MCAC algorithm. The channel is allowed when:

- If it is defined as mandatory – always.
- If it is defined as optional, then the allocated bundle bandwidth cannot exceed the configured bandwidth. The allocated bandwidth equals the bandwidth of all the mandatory channels belonging to that bundle plus the bandwidth of the optional channels already accepted plus the bandwidth of this optional channel.

MCAC on Link Aggregation Group Interfaces

When MCAC enabled interfaces reside on a LAG, SROS allows operators to change MCAC behavior when the number of active ports in a LAG changes. Both MCAC policy bundle and MCAC interface allows operators to define multiple MCAC levels per LAG based on the number of active ports in the LAG. For each level, operators can configure corresponding BW limits.

When MCAC LAG constraints are enabled, the level to use is selected automatically based on the configuration and a currently active number of LAG ports. In a case of the available bandwidth reduction (for example, a LAG link failure causes change to a level with smaller BW configured), MCAC attempts first to fit all mandatory channels (in an arbitrary order). If there is no sufficient capacity to carry all mandatory channels in the degraded mode, some channels are dropped and all optional channels are dropped. If after evaluation of mandatory channels, there remains available bandwidth, then all optional channels are re-evaluated (in an arbitrary order). Channel re-

evaluation employs the above-described MCAC algorithm applied at the interface and bundle levels that use the constraints for the degraded mode of operation.

Multicast Debugging Tools

This section describes multicast debugging tools requirement for the router family of products.

The debugging tools for multicast consist out of three elements; mtrace, mstat, and mrinfo.

Mtrace

Assessing problems in the distribution of IP multicast traffic can be difficult. The **mtrace** feature utilizes a tracing feature implemented in multicast routers that is accessed via an extension to the IGMP protocol. The **mtrace** feature is used to print the path from the source to a receiver; it does this by passing a trace query hop-by-hop along the reverse path from the receiver to the source. At each hop, information such as the hop address, routing error conditions and packet statistics should be gathered and returned to the requestor.

Data added by each hop includes:

- Query arrival time
- Incoming interface
- Outgoing interface
- Previous hop router address
- Input packet count
- Output packet count
- Total packets for this source/group
- Routing protocol
- TTL threshold
- Fowarding/error code

The information enables the network administrator to determine:

- Where multicast flows stop
- the flow of the multicast stream

When the trace response packet reaches the first hop router (the router that is directly connected to the source’s net), that router sends the completed response to the response destination (receiver) address specified in the trace query.

If some multicast router along the path does not implement the multicast traceroute feature or if there is some outage, then no response is returned. To solve this problem, the trace query includes

a maximum hop count field to limit the number of hops traced before the response is returned. This allows a partial path to be traced.

The reports inserted by each router contain not only the address of the hop, but also the TTL required to forward and some flags to indicate routing errors, plus counts of the total number of packets on the incoming and outgoing interfaces and those forwarded for the specified group. Taking differences in these counts for two traces separated in time and comparing the output packet counts from one hop with the input packet counts of the next hop allows the calculation of packet rate and packet loss statistics for each hop to isolate congestion problems.

Finding the Last Hop Router

The trace query must be sent to the multicast router which is the last hop on the path from the source to the receiver. If the receiver is on the local subnet (as determined using the subnet mask), then the default method is to multicast the trace query to all-routers.mcast.net (224.0.0.2) with a TTL of 1. Otherwise, the trace query is multicast to the group address since the last hop router will be a member of that group if the receiver is. Therefore, it is necessary to specify a group that the intended receiver has joined. This multicast is sent with a default TTL of 64, which may not be sufficient for all cases.

When tracing from a multihomed host or router, the default receiver address may not be the desired interface for the path from the source. In that case, the desired interface should be specified explicitly as the receiver.

Directing the Response

By default, mtrace first attempts to trace the full reverse path, unless the number of hops to trace is explicitly set with the hop option. If there is no response within a 3 second timeout interval, a "*" is printed and the probing switches to hop-by-hop mode. Trace queries are issued starting with a maximum hop count of one and increasing by one until the full path is traced or no response is received. At each hop, multiple probes are sent. The first attempt is made with the unicast address of the host running mtrace as the destination for the response. Since the unicast route may be blocked, the remainder of attempts request that the response be multicast to mtrace.mcast.net (224.0.1.32) with the TTL set to 32 more than what’s needed to pass the thresholds seen so far along the path to the receiver. For the last attempts the TTL is increased by another 32.

Alternatively, the TTL may be set explicitly with the TTL option.

For each attempt, if no response is received within the timeout, a "*" is printed. After the specified number of attempts have failed, mtrace will try to query the next hop router with a DVMRP_ASK_NEIGHBORS2 request (as used by the mrinfo program) to determined the router type.

The output of mtrace is a short listing of the hops in the order they are queried, that is, in the reverse of the order from the source to the receiver. For each hop, a line is printed showing the hop number (counted negatively to indicate that this is the reverse path); the multicast routing protocol; the threshold required to forward data (to the previous hop in the listing as indicated by the up-arrow character); and the cumulative delay for the query to reach that hop (valid only if the clocks are synchronized). The response ends with a line showing the round-trip time which measures the interval from when the query is issued until the response is received, both derived from the local system clock.

Mtrace/mstat packets use special IGMP packets with IGMP type codes of 0x1E and 0x1F.

Mstat

The **mstat** command adds the capability to show the multicast path in a limited graphic display and provide drops, duplicates, TTLs and delays at each node. This information is useful to the network operator because it identifies nodes with high drop & duplicate counts. Duplicate counts are shown as negative drops.

The output of **mstat** provides a limited pictorial view of the path in the forward direction with data flow indicated by arrows pointing downward and the query path indicated by arrows pointing upward. For each hop, both the entry and exit addresses of the router are shown if different, along with the initial ttl required on the packet in order to be forwarded at this hop and the propagation delay across the hop assuming that the routers at both ends have synchronized clocks. The output consists of two columns, one for the overall multicast packet rate that does not contain lost/sent packets and a column for the (S,G)-specific case. The S,G statistics do not contain lost/sent packets.

Mrinfo

mrinfo is a simple mechanism based on the **ask_neighbors igmp** to display the configuration information from the target multicast router. The type of information displayed includes the Multicast of the router, code version, metrics, ttl-thresholds, protocols and status. This information, for instance, can be used by network operators to verify if bi-directional adjacencies exist. Once the specified multicast router responds, the configuration is displayed.