Multicast (IGMP and PIM)

Protocol independent multicast (PIM) is a collection of protocols that deal with efficient delivery of IP multicast (MC) data. Those protocols are published in the series of RFCs and define different ways and aspects of multicast data distribution. PIM protocol family includes Internet Group Management protocol (IGMP), IGMP Snooping, Bootstrap router (BSR) protocol, and PIM variations: Sparse mode (PIM-SM), Source-Specific mode (PIM-SSM), Dense mode (PIM-DM) and Bidirectional mode (PIM-BIDIR). PIM-DM is not supported on NVIDIA Onyx.

PIM builds and maintains multicast routing tables based on the unicast routing information provided by unicast routing tables that can be maintained statically or dynamically by IP routing protocols like OSPF and BGP.

Basic PIM-SM

PIM relies on the underlying topology gathering protocols that collect unicast routing information and build multicast routing information base (MRIB). The primary role of MRIB is to determine the next hop for PIM messages. MC data flows along with the reverse path of the PIM control.

MC tree construction contains three phases:

1. Construction of a shared distribution tree. This tree is built around a special router called the rendezvous point (RP).
2. Establishing a native forwarding path from MC sources to the RP.
3. Building an optimized MC distribution tree directly from each MC source to all MC targets.

The first stage of the multicast tree establishment starts when the MC receiver expresses desire to start receiving MC data. It can happen as a result of using one of the L3 protocols like MLD or IGMP, or by static configuration. When such request is received by the last hop router (a designated router) this router starts to build a distribution path from the RP. It starts to send periodic "Join" messages to the nearest PIM neighbor router towards the RP. The next router continues to do the same. Eventually the process converges when Join messages reach RP or a router that has already created that distribution tree. Usually that tree is called a shared tree because it is created for any source for specific MC group G and is noted as (*,G).

At that stage, MC senders can start sending MC data. The RP decapsulates the packets and distributes them to all MC receivers along with the share tree.

On the second stage the RP switches from tunneling of multicast packets from MC sources to forwarding native traffic. When the RP identifies that a new MC source started to send packets, it initiates an establishment of a native forwarding path from the DR of that source to itself. For this purpose it starts to send Join messages towards MC source to nearest neighbor to that source according the MRIB. This is a source specific Join and is noted as (S,G). When data path is established up to the DR, the DR switches from tunneling MC packets to their native forwarding, so the RP does not need to decapsulate MC packets anymore, but still continue to distribute the packets along with shared tree.

On the third phase multicast receivers will try to switch from shared tree to source specific tree by creating a direct distribution path from a multicast source. When last hop router of the multicast receiver identifies multicast traffic coming from any multicast source it will start to send Join messages towards the source with purpose to create a direct source specific path to that source. Once such path will be established and Designated router that is attached to the source L2 network will start to distribute the multicast traffic directly bypassing shared tree, the last hop router will detach its receivers from shared tree for that data and will switch to the shortest path tree distribution.

Source-Specific Multicast (SSM)

Source-Specific Multicast (SSM) is a method of delivering multicast packets in which the only packets that are delivered to a receiver are those originating from a specific source address requested by the receiver. By so limiting the source, SSM reduces demands on the network and improves security.

SSM requires that the receiver specify the source address and explicitly excludes the use of the (*,G) join for all multicast groups in RFC 3376, which is possible only in IPv4's IGMPv3 and IPv6's MLDv2.

Source-specific multicast is best understood in contrast to any-source multicast (ASM). In the ASM service model a receiver expresses interest in traffic to a multicast address. The multicast network must discover all multicast sources sending to that address, and route data from all sources to all interested receivers.

This behavior is particularly well suited for groupware applications where all participants in the group want to be aware of all other participants, and the list of participants is not known in advance.

The source discovery burden on the network can become significant when the number of sources is large.

In the SSM service model, in addition to the receiver expressing interest in traffic to a multicast address, the receiver expresses interest in receiving traffic from only one specific source sending to that multicast address. This relieves the network of discovering many multicast sources and reduces the amount of multicast routing information that the network must maintain.

SSM requires support in last-hop routers and in the receiver's operating system. SSM support is not required in other network components, including routers and even the sending host. Interest in multicast traffic from a specific source is conveyed from hosts to routers using IGMPv3 as specified in RFC 4607.

By default SSM destination addresses defined in the ranges 232.0.0.0/8 for IPv4 or FF3x::/96 for IPv6. This range may be configured by user.

Source-specific multicast delivery semantics are provided for a datagram sent to an SSM address. That is, a datagram with source IP address S and SSM destination address G is delivered to each upper-layer "socket" that has specifically requested the reception of datagrams sent to address G by source S, and only to those sockets.
Bidirectional PIM

Bidirectional PIM (PIM-BIDIR) is a variant of PIM-SM that builds bidirectional distribution trees that connect multicast senders and receivers. It differs from PIM-SM by eliminating a need to tunnel multicast packets to RP and to keep a state for each (S,G) pair. It also eliminates a need in data driven protocol events. PIM-BIDIR achieves it by defining a new role, Designated Forwarder (DF), and by defining new forwarding rules and keeping all other PIM-SM mechanisms intact.

DF is a PIM enabled router that is the closest router to RP among all PIM routers residing on specific L2 network. It is dynamically elected by all PIM routers on that network. DF is required on each L2 multicast capable network for each RP. DF serves all multicast groups that share the same RP and has following duties:

- It is an only router that is responsible to receive and forward upstream multicast packets on that L2 segment
- It is a router that should collect all Join requests from the routers on that L2 segment
- It is an only router that will distribute downstream multicast packets on that segment.

Once Designated forwarders are elected and forwarding rules are established, PIM routers can start to issue (*,G) Join messages and build shared distribution trees. When shared tree is created, multicast sources can start to exchange data with receivers and it doesn’t require any additional maintenance of the multicast states.

Compared to PIM-SM, in bidirectional PIM:

- Each router will keep only (*,G) state and not (*,G) and (S,G) like in PIM-SM
- Multicast traffic from the beginning is forwarded naturally - no need to tunnel data to RP
- Resulting multicast tree is not shortest path optimal and converges around selected Rendezvous point, but is shared among all participants in that multicast group

In BIDIR-PIM, the packet forwarding rules have been improved over PIM-SM, allowing traffic to be passed up the shared tree toward the RP. To avoid multicast packet looping, bidir-PIM introduces a new mechanism called designated forwarder (DF) election, which establishes a loop-free SPT rooted at the RP.

PIM Load-Sharing

PIM load-sharing improves network efficiency in IP multicast applications especially in cases when we have multiple equal-cost paths to the same destination. There are two methods which enhance IP multicast bandwidth capacity consumption: rendezvous point load sharing and next-hop load sharing.

Routers should be connected via router port interfaces and not VLAN interfaces. Connecting two routers via VLAN interface with PIM load-sharing causes loops in the network.

Rendezvous Point Load-Sharing

IP multicast routing is facilitated by use of rendezvous points (RPs) which are anchors in IP multicast distribution trees, and, in case of PIM-BIDIR, are central points that perform IP multicast packet forwarding. Therefore, they can get heavily loaded.

When multiple RPs serve the same multicast IP addresses and are located at an equal distance from a traffic source or receiver, data streams can be shared between those RPs. This enhances switching performance, improves network bandwidth consumption and increases reliability. Data packets based on the packet flow parameters are equally shared between all RPs located at an equal-distance.

Next Hop Load-Sharing

Another way to improve network capacity consumption and increase the amount of IP multicast data carried by the network, is to utilize multiple equal-cost paths from RPs to IP multicast receivers. A network usually selects a single path to carry specific multicast group data packets from a source to a specific multicast destination. But when enabling next hop load-sharing, multiple paths between RP and multicast group receivers may be utilized, and based on traffic flow parameters, the data stream may be split to multiple flows that go through several equal-cost paths to the same destination.

Bootstrap Router

For correct operation each PIM router requires a capability to map a multicast group that it needs to serve to a Rendezvous point for that group. This mapping can be done manually or the mapping can be distributed dynamically in the network. BSR protocol serves for this purpose.

This protocol introduces new role in the multicast network – Bootstrap router. That router is responsible to flood multicast group to RP mapping through the multicast routing domain. Bootstrap router is elected dynamically among bootstrap router candidates (C-BSR) and once elected will collect from Rendezvous point candidate (C-RP) mapping information and distribute it in the domain.

Bootstrap activity contains 4 steps. First each C-BSR configured in the network originates floods into the network bootstrap messages that express the router desire to become BSR and also its BSR priority. Any C-BSR that receives that information and has lower priority will suspend itself, so eventually only one router will send BSR messages and become BSR.
When BSR is elected all RP candidates start to advertise to BSR a list of groups that this RP can serve. On the next step, after BSR learns the group mapping proposals, it forms a final group to RP mapping in the domain and starts to distribute it among PIM routers in the multicast routing domain. When PIM router receives BSR message with the group to RP mapping, it installs that mapping in the router local cache and uses that information to create multicast distribution trees.

Configuring Multicast

Precondition steps:

1. Enable IP routing functionality. Run:

   switch (config)# ip routing

2. Enable the desired VLAN. Run:

   switch (config)# vlan 10

3. Add this VLAN to the desired interface. Run:

   switch (config)# interface ethernet 1/1
   switch (config interface ethernet 1/1)# switchport access vlan 10

4. Create a VLAN interface. Run:

   switch (config)# interface vlan 10

5. Apply IP address to the VLAN interface. Run:

   switch (config interface vlan 10)# ip address 10.10.10.10 /24

6. Enable the interface. Run:

   switch (config interface vlan 10)# no shutdown

Configuring IGMP

IGMP is enabled when IP multicast is enabled and static multicast or PIM is enabled on the interface.

Verifying IGMP

1. Display a brief IGMP interface status. Run:

   switch (config)# show ip igmp interface brief

   VRF "default":

   ==================================
   Interface      IP Address      IGMP Querier      Membership Count  Version
   Vlan10         10.10.10.1      10.10.10.1       1                  v2

2. Display detailed IGMP interface status. Run:
2.

switch (config)# show ip igmp interface vlan 10

Interface vlan10
Status: protocol-down/link-down/admin-up
VRF: "vrf-default"
IP address: 10.10.10.1/24
Active querier: 10.10.10.1
Version: 2
Next query will be sent in: 00:01:45
Membership count: 0
IGMP version: 2
IGMP query interval: 125 secs
IGMP max response time: 10 secs
IGMP startup query interval: 31 secs
IGMP last member query interval: 1 secs
IGMP last member query count: 2
IGMP group timeout: 260 secs
IGMP querier timeout: 0 secs
IGMP unsolicited report interval: 10 secs
IGMP robustness variable: 2
IGMP interface immediate leave: Disabled
Multicast routing status on interface: Enabled
Multicast TTL threshold: 0

IGMP interface statistics:
General (sent/received):
v2-queries: 2/0
v2-reports: 0/0
v2-leaves : 0/0
v3-queries: 0/0
v3-reports: 0/0

Errors:
Checksum errors                          : 0
Packet length errors                     : 0
Packets with Local IP as source          : 0
Source subnet check failures             : 0
Query from non-querier                   : 0
Report version mismatch                  : 0
Query version mismatch                   : 0
Unknown IGMP message type                : 0
Invalid v2 reports                       : 0
Invalid v3 reports                       : 0
Invalid leaves                           : 0
Packets dropped due to router-alert check: 0

3. Display the list of IGMP groups and their status. Run:

switch (config)# show ip igmp groups
IGMP Connected Group Membership
Type: S – Static, D – Dynamic

<table>
<thead>
<tr>
<th>Group Address</th>
<th>Type</th>
<th>Interface</th>
<th>Uptime</th>
<th>Expires</th>
</tr>
</thead>
<tbody>
<tr>
<td>226.0.1.0</td>
<td>D</td>
<td>vlan10</td>
<td>00:00:05</td>
<td>N/A</td>
</tr>
<tr>
<td>10.10.10.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>226.0.1.1</td>
<td>D</td>
<td>vlan10</td>
<td>00:00:04</td>
<td>N/A</td>
</tr>
<tr>
<td>10.10.10.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Configuring PIM
Prerequisites:

1. If not enabled, enable IP routing. Run:

   ```
   switch (config)# ip routing
   ```

2. Globally enable multicast routing. Run:

   ```
   switch (config)# ip multicast-routing
   ```

To configure PIM:

1. Enable PIM. Run:

   ```
   switch (config)# protocol pim
   ```

2. Enable PIM on any IP interface (router port or VLAN interface) facing an L3 multicast source or L3 multicast receiver including transit interfaces. For example, run:

   ```
   switch (config)# interface ethernet 1/4 ip pim sparse-mode
   ```

   > The interface's primary address is always used in PIM.

3. Configure IGMP version on any IP interface (router port or VLAN interface) facing multicast receivers. For example, run:

   ```
   switch (config)# interface ethernet 1/4 ip igmp version {2|3}
   ```

   If IGMP must be enabled on a VLAN interface, IP IGMP snooping must also be enabled (globally and on the relevant VLAN interface):

   ```
   switch (config)# interface vlan 50 ip igmp version {2|3}
   switch (config)# ip igmp snooping
   switch (config)# vlan 50 ip igmp snooping
   ```

4. Configure a rendezvous point. Run:

   ```
   switch (config)# ip pim rp-address 10.10.10.10
   ```

   > A good practice is to configure the RP on the loopback interface. Although RP may be configured on the any interface with enabled PIM sparse mode. Note that a loopback interface does not require enabling PIM sparse mode to configure RP.

   > The RP address must be reachable to all switches.

5. Configure a group mapping for a static RP. Run:

   ```
   switch (config)# ip pim rp-address 192.168.0.1
   ```

   > You may also specify a “group-list <ip-address> <prefix>” parameter (ip pim rp-address 192.168.0.1 group-list 224.0.0.0/4) if you want different RPs for different groups.

Additional Reading and Use Cases

For more information about this feature and its potential applications, please refer to the following community post:
IGMP and PIM Commands

- IGMP and PIM Commands
- IGMP Snooping