Virtual Private Routed Network Service

In This Chapter

This chapter provides information about the Virtual Private Routed Network (VPN) service and implementation notes.

Topics in this chapter include:

(Continued)

VPRN Service Overview

RFC 2547b is an extension to the original RFC 2547, *BGP/MPLS VPNs*, which details a method of distributing routing information using BGP and MPLS forwarding data to provide a Layer 3 Virtual Private Network (VPN) service to end customers.

Each Virtual Private Routed Network (VPRN) consists of a set of customer sites connected to one or more PE routers. Each associated PE router maintains a separate IP forwarding table for each VPRN. Additionally, the PE routers exchange the routing information configured or learned from all customer sites via MP-BGP peering. Each route exchanged via the MP-BGP protocol includes a Route Distinguisher (RD), which identifies the VPRN association and handles the possibility of IP address overlap.

The service provider uses BGP to exchange the routes of a particular VPN among the PE routers that are attached to that VPN. This is done in a way which ensures that routes from different VPNs remain distinct and separate, even if two VPNs have an overlapping address space. The PE routers peer with locally connected CE routers and exchange routes with other PE routers in order to provide end-to-end connectivity between CEs belonging to a given VPN. Since the CE routers do not peer with each other there is no overlay visible to the CEs.

When BGP distributes a VPN route it also distributes an MPLS label for that route. On an SR series router, the method of allocating a label to a VPN route depends on the VPRN label mode and the configuration of the VRF export policy. SR series routers support three label allocation methods: label per VRF, label per next hop, and label per prefix.

Before a customer data packet travels across the service provider's backbone, it is encapsulated with the MPLS label that corresponds, in the customer's VPN, to the route which best matches the packet's destination address. The MPLS packet is further encapsulated with one or additional MPLS labels or GRE tunnel header so that it gets tunneled across the backbone to the proper PE router. Each route exchanged by the MP-BGP protocol includes a route distinguisher (RD), which identifies the VPRN association. Thus the backbone core routers do not need to know the VPN routes. Figure 7 displays a VPRN network diagram example.

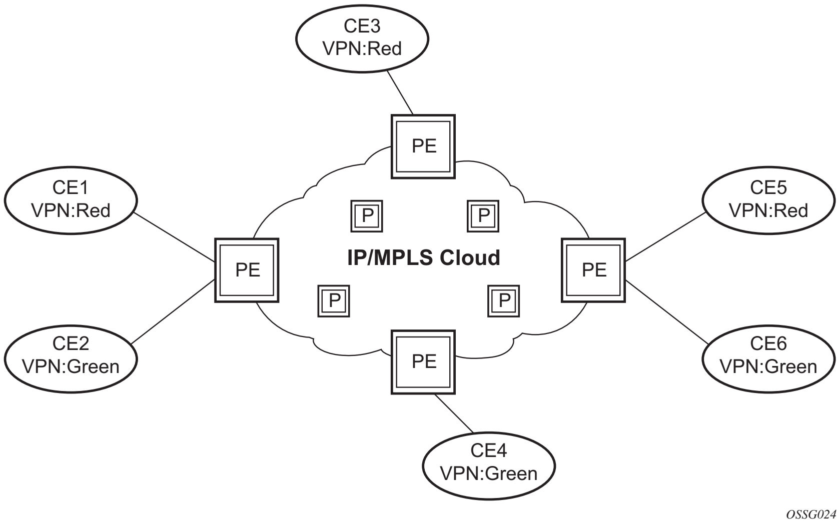


Figure 7: Virtual Private Routed Network

Routing Prerequisites

RFC4364 requires the following features:

- Multi-protocol extensions to BGP
- Extended BGP community support
- BGP capability negotiation

Tunneling protocol options are as follows:

- Label Distribution Protocol (LDP)
- MPLS RSVP-TE tunnels
- Generic Router Encapsulation (GRE) tunnels
- BGP route tunnel (RFC3107)

Core MP-BGP Support

BGP is used with BGP extensions mentioned in Routing Prerequisites on page 227 to distribute VPRN routing information across the service provider's network.

BGP was initially designed to distribute IPv4 routing information. Therefore, multi-protocol extensions and the use of a VPN-IP address were created to extend BGP's ability to carry overlapping routing information. A VPN-IPv4 address is a 12-byte value consisting of the 8-byte route distinguisher (RD) and the 4-byte IPv4 IP address prefix. A VPN-IPv6 address is a 24-byte value consisting of the 8-byte RD and 16-byte IPv6 address prefix. Service providers typically assign one or a small number of RDs per VPN service network-wide.

Route Distinguishers

The route distinguisher (RD) is an 8-byte value consisting of two major fields, the **Type** field and **Value** field. The **Type** field determines how the **Value** field should be interpreted.   The 7750 SR implementation supports the three (3) **Type** values as defined in the standard.

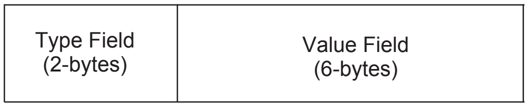


Figure 8: Route Distinguisher

The three Type values are:

- Type 0: Value Field —   Administrator subfield (2 bytes)
  Assigned number subfield (4 bytes)

  The administrator field must contain an AS number (using private AS numbers is discouraged). The Assigned field contains a number assigned by the service provider.
- Type 1: Value Field —   Administrator subfield (4 bytes)
  Assigned number subfield (2 bytes)

  The administrator field must contain an IP address (using private IP address space is discouraged). The Assigned field contains a number assigned by the service provider.
- Type 2: Value Field —   Administrator subfield (4 bytes)
  Assigned number subfield (2 bytes)

  The administrator field must contain a 4-byte AS number (using private AS numbers is discouraged). The Assigned field contains a number assigned by the service provider.

eiBGP Load Balancing

eiBGP load balancing allows a route to have multiple nexthops of different types, using both IPv4 nexthops and MPLS LSPs simultaneously.

Figure 9 displays a basic topology that could use eiBGP load balancing. In this topology CE1 is dual homed and thus reachable by two separate PE routers. CE 2 (a site in the same VPRN) is also attached to PE1. With eiBGP load balancing, PE1 will utilize its own local IPv4 nexthop as well as the route advertised by MP-BGP, by PE2.

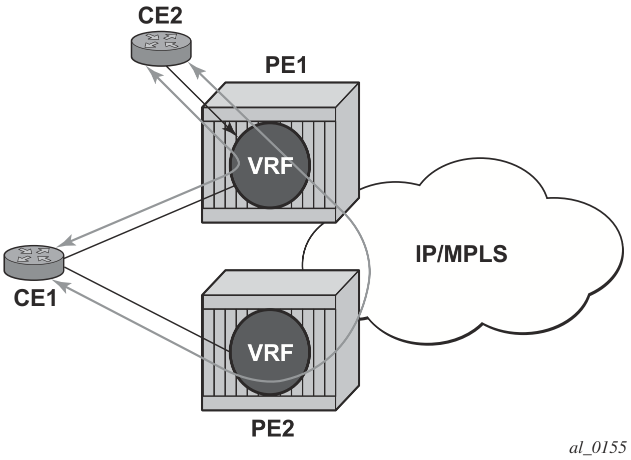


Figure 9: Basic eiBGP Topology

Another example displayed in Figure 10 shows an extra net VPRN (VRF). The traffic ingressing the PE that should be load balanced is part of a second VPRN and the route over which the load balancing is to occur is part of a separate VPRN instance and are leaked into the second VPRN by route policies.

Here, both routes can have a source protocol of VPN-IPv4 but one will still have an IPv4 nexthop and the other can have a VPN-IPv4 nexthop pointing out a network interface. Traffic will still be load balanced (if eiBGP is enabled) as if only a single VRF was involved.

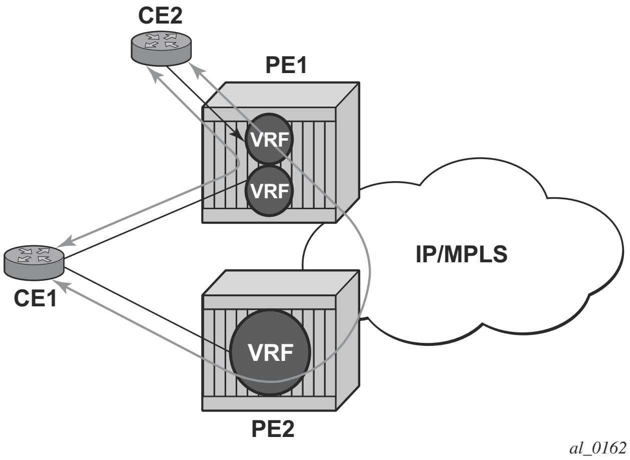


Figure 10: Extranet Load Balancing

Traffic will be load balanced across both the IPv4 and VPN-IPv4 next hops. This helps to use all available bandwidth to reach a dual-homed VPRN.

Route Reflector

The use of Route Reflectors is supported in the service provider core. Multiple sets of route reflectors can be used for different types of BGP routes, including IPv4 and VPN-IPv4 as well as multicast and IPv6.

CE to PE Route Exchange

Routing information between the Customer Edge (CE) and Provider Edge (PE) can be exchanged by the following methods:

- Static Routes
- E-BGP
- RIP
- OSPF
- OSPF3

Each protocol provides controls to limit the number of routes learned from each CE router.

Route Redistribution

Routing information learned from the CE-to-PE routing protocols and configured static routes should be injected in the associated local VPN routing/forwarding (VRF). In the case of dynamic routing protocols, there may be protocol specific route policies that modify or reject certain routes before they are injected into the local VRF.

Route redistribution from the local VRF to CE-to-PE routing protocols is to be controlled via the route policies in each routing protocol instance, in the same manner that is used by the base router instance.

The advertisement or redistribution of routing information from the local VRF to or from the MP-BGP instance is specified per VRF and is controlled by VRF route target associations or by VRF route policies.

VPN-IP routes imported into a VPRN, have the protocol **type bgp-vpn** to denote that it is an VPRN route. This can be used within the route policy match criteria.

CPE Connectivity Check

Static routes are used within many IES and VPRN services. Unlike dynamic routing protocols, there is no way to change the state of routes based on availability information for the associated CPE. CPE connectivity check adds flexibility so that unavailable destinations will be removed from the VPRN routing tables dynamically and minimize wasted bandwidth.

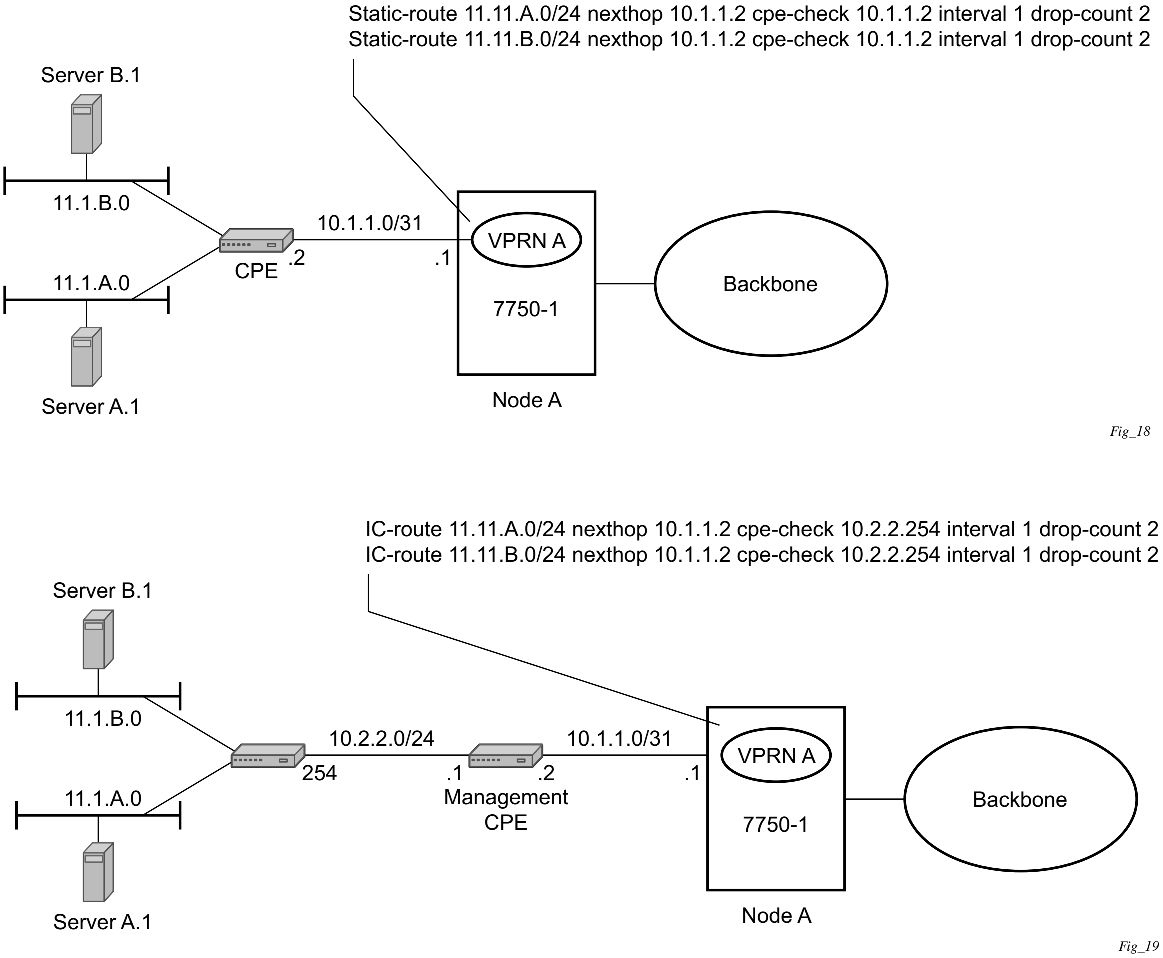

Static-route 11.11.A.0/24 nexthop 10.1.1.2 cpe-check 10.1.1.2 interval 1 drop-count 2
Static-route 11.11.B.0/24 nexthop 10.1.1.2 cpe-check 10.1.1.2 interval 1 drop-count 2
Server B.1
11.1.B.0
11.1.A.0
Server A.1
CPE
.2
10.1.1.0/31
.1
VPRN A
7750-1
Node A
Backbone
Fig_18
IC-route 11.11.A.0/24 nexthop 10.1.1.2 cpe-check 10.2.2.254 interval 1 drop-count 2
IC-route 11.11.B.0/24 nexthop 10.1.1.2 cpe-check 10.2.2.254 interval 1 drop-count 2
Server B.1
11.1.B.0
11.1.A.0
Server A.1
10.2.2.0/24
254
.1
.2
Management
CPE
10.1.1.0/31
.1
VPRN A
7750-1
Node A
Backbone
Fig_19

Figure 11: Multiple Hops to IP Target

The availability of the far-end static route is monitored through periodic polling. The polling period is configured. If the poll fails a specified number of sequential polls, the static route is marked as inactive.

Either ICMP ping or unicast ARP mechanism can be used to test the connectivity. ICMP ping is preferred.

If the connectivity check fails and the static route is deactivated, the SR-Series router will continue to send polls and re-activate any routes that are restored.

Constrained Route Distribution (RT Constraint)

Constrained VPN Route Distribution Based on Route Targets

Constrained Route Distribution (or RT Constraint) is a mechanism that allows a router to advertise Route Target membership information to its BGP peers to indicate interest in receiving only VPN routes tagged with specific Route Target extended communities. Upon receiving this information, peers restrict the advertised VPN routes to only those requested, minimizing control plane load in terms of protocol traffic and possibly also RIB memory.

The Route Target membership information is carried using MP-BGP, using an AFI value of 1 and SAFI value of 132. In order for two routers to exchange RT membership NLRI they must advertise the corresponding AFI/SAFI to each other during capability negotiation. The use of MP-BGP means RT membership NLRI are propagated, loop-free, within an AS and between ASes using well-known BGP route selection and advertisement rules.

ORF can also be used for RT-based route filtering, but ORF messages have a limited scope of distribution (to direct peers) and therefore do not automatically create pruned inter-cluster and inter-AS route distribution trees.

Configuring the Route Target Address Family

RT Constraint is supported only by the base router BGP instance. When the **family** command at the BGP router group or neighbor CLI context includes the **route-target** keyword, the RT Constraint capability is negotiated with the associated set of EBGP and IBGP peers.

ORF is mutually exclusive with RT Constraint on a particular BGP session. The CLI will not attempt to block this configuration, but if both capabilities are enabled on a session, the ORF capability will not be included in the OPEN message sent to the peer.

Originating RT Constraint Routes

When the base router has one or more RTC peers (BGP peers with which the RT Constraint capability has been successfully negotiated), one RTC route is created for each RT extended community imported into a locally-configured L2 VPN or L3 VPN service. These imported route targets are configured in the following contexts:

- config>service>vprn
- config>service>vprn>mvpn
- config>service>vpls>bgp
- config>service>epipe>bgp

By default, these RTC routes are automatically advertised to all RTC peers, without the need for an export policy to explicitly “accept” them. Each RTC route has a prefix, a prefix length and path attributes. The prefix value is the concatenation of the origin AS (a 4 byte value representing the 2- or 4-octet AS of the originating router, as configured using the **config>router>autonomous-system** command) and 0 or 16-64 bits of a route target extended community encoded in one of the following formats: 2-octet AS specific extended community, IPv4 address specific extended community, or 4-octet AS specific extended community.

A, 7750 SR, or may be configured to send the default RTC route to any RTC peer. This is done using the new **default-route-target** group/neighbor CLI command. The default RTC route is a special type of RTC route that has zero prefix length. Sending the default RTC route to a peer conveys a request to receive all VPN routes (regardless of route target extended community) from that peer. The default RTC route is typically advertised by a route reflector to its clients. The advertisement of the default RTC route to a peer does not suppress other more specific RTC routes from being sent to that peer.

Receiving and Re-Advertising RT Constraint Routes

All received RTC routes that are deemed valid are stored in the RIB-IN. An RTC route is considered invalid and treated as withdrawn, if any of the following applies:

- The prefix length is 1-31.
- The prefix length is 33-47.
- The prefix length is 48-96 and the 16 most-significant bits are not 0x0002, 0x0102 or 0x0202.

If multiple RTC routes are received for the same prefix value then standard BGP best path selection procedures are used to determine the best of these routes.

The best RTC route per prefix is re-advertised to RTC peers based on the following rules:

- The best path for a default RTC route (prefix-length 0, origin AS only with prefix-length 32, or origin AS plus 16 bits of an RT type with prefix-length 48) is never propagated to another peer.
- A PE with only IBGP RTC peers that is neither a route reflector or an ASBR does not re-advertise the best RTC route to any RTC peer due to standard IBGP split horizon rules.
- A route reflector that receives its best RTC route for a prefix from a client peer re-advertises that route (subject to export policies) to all of its client and non-client IBGP peers (including the originator), per standard RR operation. When the route is re-advertised to client peers, the RR (i) sets the ORIGINATOR_ID to its own router ID and (ii) modifies the NEXT_HOP to be its local address for the sessions (for example, system IP).

- A route reflector that receives its best RTC route for a prefix from a non-client peer re-advertises that route (subject to export policies) to all of its client peers, per standard RR operation. If the RR has a non-best path for the prefix from any of its clients, it advertises the best of the client-advertised paths to all non-client peers.
- An ASBR that is neither a PE nor a route reflector that receives its best RTC route for a prefix from an IBGP peer re-advertises that route (subject to export policies) to its EBGP peers. It modifies the NEXT_HOP and AS_PATH of the re-advertised route per standard BGP rules. No aggregation of RTC routes is supported.
- An ASBR that is neither a PE nor a route reflector that receives its best RTC route for a prefix from an EBGP peer re-advertises that route (subject to export policies) to its EBGP and IBGP peers. When re-advertised routes are sent to EBGP peers, the ABSR modifies the NEXT_HOP and AS_PATH per standard BGP rules. No aggregation of RTC routes is supported.

Note: These advertisement rules do not handle hierarchical RR topologies properly. This is a limitation of the current RT constraint standard.

Using RT Constraint Routes

In general (ignoring IBGP-to-IBGP rules, Add-Path, Best-external, etc.), the best VPN route for every prefix/NLRI in the RIB is sent to every peer supporting the VPN address family, but export policies may be used to prevent some prefix/NLRI from being advertised to specific peers. These export policies may be configured statically or created dynamically based on use of ORF or RT constraint with a peer. Note that ORF and RT Constraint are mutually exclusive on a session.

When RT Constraint is configured on a session that also supports VPN address families using route targets (that is: vpn-ipv4, vpn-ipv6, l2-vpn, mvpn-ipv4, mvpn-ipv6, mcast-vpn-ipv4 or evpn), the advertisement of the VPN routes is affected as follows:

- When the session comes up, the advertisement of the VPN routes is delayed for a short while to allow RTC routes to be received from the peer.
- After the initial delay, the received RTC routes are analyzed and acted upon. If S1 is the set of routes previously advertised to the peer and S2 is the set of routes that should be advertised based on the most recent received RTC routes then:
  - → Set of routes in S1 but not in S2 should be withdrawn immediately (subject to MRAI).
  - → Set of routes in S2 but not in S1 should be advertised immediately (subject to MRAI).
- If a default RTC route is received from a peer P1, the VPN routes that are advertised to P1 is the set that:
  - → (a) are eligible for advertisement to P1 per BGP route advertisement rules AND
  - → (b) have not been rejected by manually configured export policies AND
  - → (c) have not been advertised to the peer

Note: This applies whether or not P1 advertised the best route for the default RTC prefix.

In this context, a default RTC route is any of the following:

1. a route with NLRI length = zero
2. a route with NLRI value = origin AS and NLRI length = 32
3. a route with NLRI value = {origin AS+0x0002 | origin AS+0x0102 | origin AS+0x0202} and NLRI length = 48

- If an RTC route for prefix A (origin-AS = A1, RT = A2/n, n > 48) is received from an IBGP peer I1 in autonomous system A1, the VPN routes that are advertised to I1 is the set that:
  - → (a) are eligible for advertisement to I1 per BGP route advertisement rules AND
  - → (b) have not been rejected by manually configured export policies AND
  - → (c) carry at least one route target extended community with value A2 in the n most significant bits AND
  - → (d) have not been advertised to the peer

Note: This applies whether or not I1 advertised the best route for A.

- If the best RTC route for a prefix A (origin-AS = A1, RT = A2/n, n > 48) is received from an IBGP peer I1 in autonomous system B, the VPN routes that are advertised to I1 is the set that:
  - → (a) are eligible for advertisement to I1 per BGP route advertisement rules AND
  - → (b) have not been rejected by manually configured export policies AND
  - → (c) carry at least one route target extended community with value A2 in the n most significant bits AND
  - → (d) have not been advertised to the peer

  Note: This applies only if I1 advertised the best route for A.
- If the best RTC route for a prefix A (origin-AS = A1, RT = A2/n, n > 48) is received from an EBGP peer E1, the VPN routes that are advertised to E1 is the set that:
  - → (a) are eligible for advertisement to E1 per BGP route advertisement rules AND
  - → (b) have not been rejected by manually configured export policies AN
  - → (c) carry at least one route target extended community with value A2 in the n most significant bits AND
  - → (d) have not been advertised to the peer

Note: This applies only if E1 advertised the best route for A.

BGP Fast Reroute in a VPRN

BGP fast reroute is a feature that brings together indirection techniques in the forwarding plane and pre-computation of BGP backup paths in the control plane to support fast reroute of BGP traffic around unreachable/failed next-hops. In a VPRN context BGP fast reroute is supported using unlabeled IPv4, unlabeled IPv6, VPN-IPv4, and VPN-IPv6 VPN routes. The supported VPRN scenarios are outlined in Table 9.

Note that BGP fast reroute information specific to the base router BGP context is described in the BGP Fast Reroute section of the 7x50 SR OS Routing Protocols Guide.

Table 9: BGP Fast Reroute Scenarios (VPRN Context)

Ingress Packet	Primary Route	Backup Route	Prefix Independent Convergence
IPv4 (ingress PE)	IPv4 route with next-hop A resolved by an IPv4 route	IPv4 route with next-hop B resolved by an IPv4 route	Yes
IPv4 (ingress PE)	VPN-IPv4 route with next-hop A resolved by a GRE, LDP, RSVP or BGP tunnel	VPN-IPv4 route with next-hop A resolved by a GRE, LDP, RSVP or BGP tunnel	Yes, but if the VPN-IP routes are label-per-prefix the ingress card must be FP2 or better
MPLS (egress PE)	IPv4 route with next-hop A resolved by an IPv4 route	IPv4 route with next-hop B resolved by an IPv4 route	Yes
MPLS (egress PE)	IPv4 route with next-hop A resolved by an IPv4 rout	VPN-IPv4 route* with next-hop B resolved by a GRE, LDP, RSVP or BGP tunnel	Yes, but if the VPN-IP routes are label-per-prefix the ingress card must be FP2 or better for PIC
IPv6 (ingress PE)	IPv6 route with next-hop A resolved by an IPv6 route	IPv6 route with next-hop B resolved by an IPv6 route	Yes
IPv6 (ingress PE)	VPN-IPv6 route with next-hop A resolved by a GRE, LDP, RSVP or BGP tunnel	VPN-IPv6 route with next-hop B resolved by a GRE, LDP, RSVP or BGP tunnel	Yes, but if the VPN-IP routes are label-per-prefix the ingress card must be FP2 or better
MPLS (egress)	IPv6 route with next-hop A resolved by an IPv6 route	IPv6 route with next-hop B resolved by an IPv6 route	Yes

Table 9: BGP Fast Reroute Scenarios (VPRN Context)

Ingress Packet	Primary Route	Backup Route	Prefix Independent Convergence
MPLS (egress)	IPv6 route with next-hop A resolved by an IPv6 route	Yes, but if the VPN-IP routes are label-per-prefix the ingress card must be FP2 or better for PIC	VPRN label mode must be VRF. VPRN must export its VPN-IP routes with RD ≠ y. For the best performance the backup next-hop must advertise the same VPRN label value with all routes (e.g. per VRF label).

BGP Fast Reroute in a VPRN Configuration

In a VPRN context, BGP fast reroute is optional and must be enabled. Fast reroute can be applied to all IPv4 prefixes, all IPv6 prefixes, all IPv4 and IPv6 prefixes, or to a specific set of IPv4 and IPv6 prefixes.

If all IP prefixes require backup path protection, use a combination of the BGP instance-level **backup-path** and VPRN-level **enable-bgp-vpn-backup** commands. The VPRN BGP **backup-path** command enables BGP fast reroute for all IPv4 prefixes and/or all IPv6 prefixes that have a best path through a VPRN BGP peer. The VPRN-level **enable-bgp-vpn-backup** command enables BGP fast reroute for all IPv4 prefixes and/or all IPv6 prefixes that have a best path through a remote PE peer.

If only some IP prefixes require backup path protection, use route policies to apply the **install-backup-path** action to the best paths of the IP prefixes requiring protection. See BGP Fast Reroute section of the 7750 SR OS Routing Protocols Guide for more information.

VPRN Features

This section describes various VPRN features and any special capabilities or considerations as they relate to VPRN services.

- IP Interfaces on page 242
- Subscriber Interfaces on page 255
- SAPs on page 256
  - → Encapsulations on page 256
  - → QoS Policies on page 258
  - → Filter Policies on page 258
- CE to PE Routing Protocols on page 263
  - → PE to PE Tunneling Mechanisms on page 263
  - → Per VRF Route Limiting on page 263
  - → Using OSPF in IP-VPNs on page 267
- Spoke SDPs on page 264
  - → Carrier Supporting Carrier (CsC) on page 273
- Multicast in IP-VPN Applications on page 290

IP Interfaces

VPRN customer IP interfaces can be configured with most of the same options found on the core IP interfaces. The advanced configuration options supported are:

- VRRP
- Cflowd
- Secondary IP addresses
- ICMP Options

Configuration options found on core IP interfaces not supported on VPRN IP interfaces are:

- NTP broadcast receipt

QoS Policy Propagation Using BGP (QPPB)

This section discusses QPPB as it applies to VPRN, IES, and router interfaces. Refer to the QoS Policy Propagation Using BGP (QPPB) on page 19 section on page 15 and the IP Router Configuration section in the 7x50 OS Router Configuration Guide.

QoS policy propagation using BGP (QPPB) is a feature that allows a route to be installed in the routing table with a forwarding-class and priority so that packets matching the route can receive the associated QoS. The forwarding-class and priority associated with a BGP route are set using BGP import route policies. In the industry this feature is called QPPB, and even though the feature name refers to BGP specifically. On SR routers, QPPB is supported for BGP (IPv4, IPv6, VPN-IPv4, VPN-IPv6), RIP and static routes.

While SAP ingress and network QoS policies can achieve the same end result as QPPB, assigning a packet arriving on a particular IP interface to a specific forwarding-class and priority/profile based on the source IP address or destination IP address of the packet □the effort involved in creating the QoS policies, keeping them up-to-date, and applying them across many nodes is much greater than with QPPB. In a typical application of QPPB, a BGP route is advertised with a BGP community attribute that conveys a particular QoS. Routers that receive the advertisement accept the route into their routing table and set the forwarding-class and priority of the route from the community attribute.

QPPB Applications

There are two typical applications of QPPB:

1. Coordination of QoS policies between different administrative domains.
2. Traffic differentiation within a single domain, based on route characteristics.

Inter-AS Coordination of QoS Policies

The operator of an administrative domain A can use QPPB to signal to a peer administrative domain B that traffic sent to certain prefixes advertised by domain A should receive a particular QoS treatment in domain B. More specifically, an ASBR of domain A can advertise a prefix XYZ to domain B and include a BGP community attribute with the route. The community value implies a particular QoS treatment, as agreed by the two domains (in their peering agreement or service level agreement, for example). When the ASBR and other routers in domain B accept and install the route for XYZ into their routing table, they apply a QoS policy on selected interfaces that classifies traffic towards network XYZ into the QoS class implied by the BGP community value.

QPPB may also be used to request that traffic sourced from certain networks receive appropriate QoS handling in downstream nodes that may span different administrative domains. This can be achieved by advertising the source prefix with a BGP community, as discussed above. However, in this case other approaches are equally valid, such as marking the DSCP or other CoS fields based on source IP address so that downstream domains can take action based on a common understanding of the QoS treatment implied by different DSCP values.

In the above examples, coordination of QoS policies using QPPB could be between a business customer and its IP VPN service provider, or between one service provider and another.

Traffic Differentiation Based on Route Characteristics

There may be times when a network operator wants to provide differentiated service to certain traffic flows within its network, and these traffic flows can be identified with known routes. For example, the operator of an ISP network may want to give priority to traffic originating in a particular ASN (the ASN of a content provider offering over-the-top services to the ISP's customers), following a certain AS_PATH, or destined for a particular next-hop (remaining on-net vs. off-net).

Figure 12 shows an example of an ISP that has an agreement with the content provider managing AS300 to provide traffic sourced and terminating within AS300 with differentiated service appropriate to the content being transported. In this example we presume that ASBR1 and ASBR2 mark the DSCP of packets terminating and sourced, respectively, in AS300 so that other nodes within the ISP's network do not need to rely on QPPB to determine the correct forwarding-class to use for the traffic. Note however, that the DSCP or other COS markings could be left unchanged in the ISP's network and QPPB used on every node.

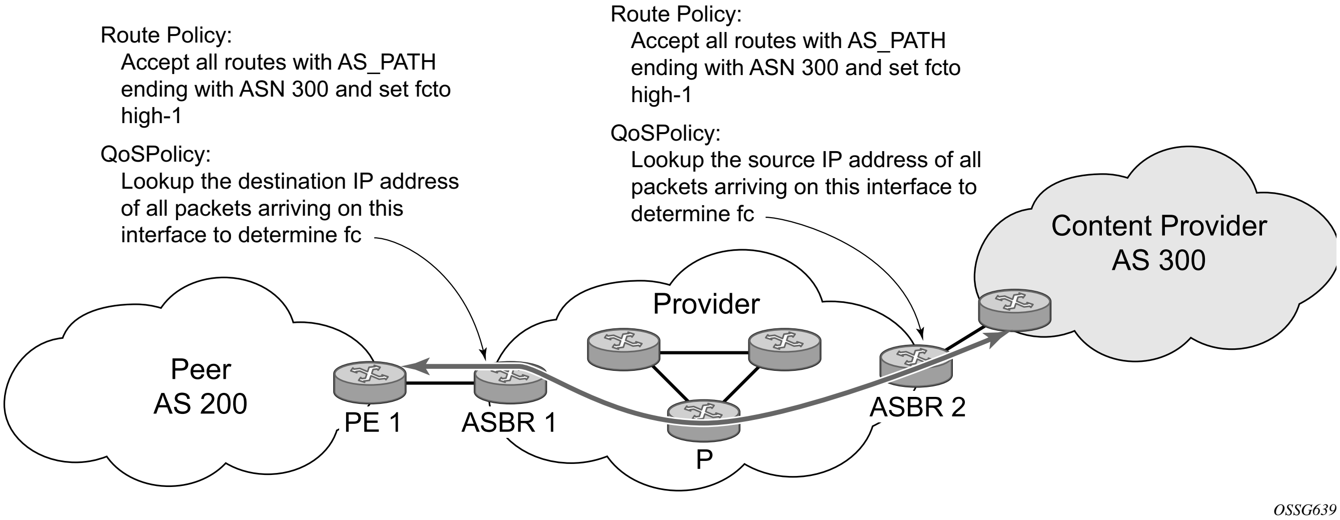

Route Policy:
Accept all routes with AS_PATH ending with ASN 300 and set fcto high-1
QoSPolicy:
Lookup the destination IP address of all packets arriving on this interface to determine fc
Route Policy:
Accept all routes with AS_PATH ending with ASN 300 and set fcto high-1
QoSPolicy:
Lookup the source IP address of all packets arriving on this interface to determine fc
Content Provider
AS 300
Provider
Peer
AS 200
PE 1
ASBR 1
P
ASBR 2
OSSG639

Figure 12: Use of QPPB to Differentiate Traffic in an ISP Network

QPPB

There are two main aspects of the QPPB feature:

- The ability to associate a forwarding-class and priority with certain routes in the routing table.
- The ability to classify an IP packet arriving on a particular IP interface to the forwarding-class and priority associated with the route that best matches the packet.

Associating an FC and Priority with a Route

This feature uses a command in the route-policy hierarchy to set the forwarding class and optionally the priority associated with routes accepted by a route-policy entry. The command has the following structure:

```
 fc fc-name [priority {low | high}]
```

The use of this command is illustrated by the following example:

```
config>router>policy-options
     begin
     community gold members 300:100
     policy-statement qppb_policy
          entry 10
               from
                    protocol bgp
                    community gold
               exit
               action accept
                    fc h1 priority high
               exit
          exit
     exit
     commit
```

The **fc** command is supported with all existing from and to match conditions in a route policy entry and with any action other than reject, it is supported with next-entry, next-policy and accept actions. If a next-entry or next-policy action results in multiple matching entries then the last entry with a QPPB action determines the forwarding class and priority.

A route policy that includes the **fc** command in one or more entries can be used in any import or export policy but the **fc** command has no effect except in the following types of policies:

- VRF import policies:
  - → config>service>vprn>vrf-import
- BGP import policies:
  - → config>router>bgp>import
  - → config>router>bgp>group>import
  - → config>router>bgp>group>neighbor>import
  - → config>service>vprn>bgp>import
  - → config>service>vprn>bgp>group>import
  - → config>service>vprn>bgp>group>neighbor>import
- RIP import policies:
  - → config>router>rip>import
  - → config>router>rip>group>import
  - → config>router>rip>group>neighbor>import
  - → config>service>vprn>rip>import
  - → config>service>vprn>rip>group>import
  - → config>service>vprn>rip>group>neighbor>import

As evident from above, QPPB route policies support routes learned from RIP and BGP neighbors of a VPRN as well as for routes learned from RIP and BGP neighbors of the base/global routing instance.

QPPB is supported for BGP routes belonging to any of the address families listed below:

- IPv4 (AFI=1, SAFI=1)
- IPv6 (AFI=2, SAFI=1)
- VPN-IPv4 (AFI=1, SAFI=128)
- VPN-IPv6 (AFI=2, SAFI=128)

Note that a VPN-IP route may match both a VRF import policy entry and a BGP import policy entry (if vpn-apply-import is configured in the base router BGP instance). In this case the VRF import policy is applied first and then the BGP import policy, so the QPPB QoS is based on the BGP import policy entry.

This feature also introduces the ability to associate a forwarding-class and optionally priority with IPv4 and IPv6 static routes. This is achieved using the following modified versions of the static-route commands:

- static-route {*ip-prefix*/*prefix-length*|*ip-prefix netmask*} [fc *fc-name* [priority {low | high}]] next-hop *ip-int-name*|*ip-address*

- static-route {*ip-prefix*/*prefix-length*|*ip-prefix netmask*} [fc *fc-name* [priority {low | high}]] indirect *ip-address*

Priority is optional when specifying the forwarding class of a static route, but once configured it can only be deleted and returned to unspecified by deleting the entire static route.

Displaying QoS Information Associated with Routes

The following commands are enhanced to show the forwarding-class and priority associated with the displayed routes:

- show router route-table
- show router fib
- show router bgp routes
- show router rip database
- show router static-route

This feature uses a **qos** keyword to the **show>router>route-table** command. When this option is specified the output includes an additional line per route entry that displays the forwarding class and priority of the route. If a route has no fc and priority information then the third line is blank. The following CLI shows an example:

show router route-table [**family**] [*ip-prefix*[/*prefix-length*]] [**longer** | **exact**] [**protocol** *protocol-name*] **qos**

An example output of this command is shown below:

```
A:Dut-A# show router route-table 10.1.5.0/24 qos
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix                                   Type    Proto    Age         Pref
       Next Hop[Interface Name]                                    Metric
       QoS
-------------------------------------------------------------------------------
10.1.5.0/24                                   Remote  BGP      15h32m52s   0
       PE1_to_PE2                                                  0
       h1, high
-------------------------------------------------------------------------------
No. of Routes: 1
===============================================================================
A:Dut-A#
```

Enabling QPPB on an IP interface

To enable QoS classification of ingress IP packets on an interface based on the QoS information associated with the routes that best match the packets the **qos-route-lookup** command is necessary in the configuration of the IP interface. The **qos-route-lookup** command has parameters to indicate whether the QoS result is based on lookup of the source or destination IP address in every packet. There are separate qos-route-lookup commands for the IPv4 and IPv6 packets on an interface, which allows QPPB to enabled for IPv4 only, IPv6 only, or both IPv4 and IPv6. Note however, current QPPB based on a source IP address is not supported for IPv6 packets nor is it supported for ingress subscriber management traffic on a group interface.

The qos-route-lookup command is supported on the following types of IP interfaces:

- base router network interfaces (config>router>interface)
- VPRN SAP and spoke SDP interfaces (config>service>vprn>interface)
- VPRN group-interfaces (config>service>vprn>sub-if>grp-if)
- IES SAP and spoke SDP interfaces (config>service>ies>interface)
- IES group-interfaces (config>service>ies>sub-if>grp-if)

When the qos-route-lookup command with the destination parameter is applied to an IP interface and the destination address of an incoming IP packet matches a route with QoS information the packet is classified to the fc and priority associated with that route, overriding the fc and priority/ profile determined from the sap-ingress or network qos policy associated with the IP interface (see section 5.7 for further details). If the destination address of the incoming packet matches a route with no QoS information the fc and priority of the packet remain as determined by the sap-ingress or network qos policy.

Similarly, when the qos-route-lookup command with the source parameter is applied to an IP interface and the source address of an incoming IP packet matches a route with QoS information the packet is classified to the fc and priority associated with that route, overriding the fc and priority/profile determined from the sap-ingress or network qos policy associated with the IP interface. If the source address of the incoming packet matches a route with no QoS information the fc and priority of the packet remain as determined by the sap-ingress or network qos policy.

Currently, QPPB is not supported for ingress MPLS traffic on network interfaces or on CsC PE'-CE' interfaces (config>service>vprn>nw-if).

QPPB When Next-Hops are Resolved by QPPB Routes

In some circumstances (IP VPN inter-AS model C, Carrier Supporting Carrier, indirect static routes, etc.) an IPv4 or IPv6 packet may arrive on a QPPB-enabled interface and match a route A1 whose next-hop N1 is resolved by a route A2 with next-hop N2 and perhaps N2 is resolved by a route A3 with next-hop N3, etc. In release 9.0 the QPPB result is based only on the forwarding-class and priority of route A1. If A1 does not have a forwarding-class and priority association then the QoS classification is not based on QPPB, even if routes A2, A3, etc. have forwarding-class and priority associations.

QPPB and Multiple Paths to a Destination

When ECMP is enabled some routes may have multiple equal-cost next-hops in the forwarding table. When an IP packet matches such a route the next-hop selection is typically based on a hash algorithm that tries to load balance traffic across all the next-hops while keeping all packets of a given flow on the same path. The QPPB configuration model described in Associating an FC and Priority with a Route on page 245 allows different QoS information to be associated with the different ECMP next-hops of a route. The forwarding-class and priority of a packet matching an ECMP route is based on the particular next-hop used to forward the packet.

When BGP FRR is enabled some BGP routes may have a backup next-hop in the forwarding table in addition to the one or more primary next-hops representing the equal-cost best paths allowed by the ECMP/multipath configuration. When an IP packet matches such a route a reachable primary next-hop is selected (based on the hash result) but if all the primary next-hops are unreachable then the backup next-hop is used. The QPPB configuration model described in Associating an FC and Priority with a Route on page 245 allows the forwarding-class and priority associated with the backup path to be different from the QoS characteristics of the equal-cost best paths. The forwarding class and priority of a packet forwarded on the backup path is based on the **fc** and priority of the backup route.

QPPB and Policy-Based Routing

When an IPv4 or IPv6 packet with destination address X arrives on an interface with both QPPB and policy-based-routing enabled:

- There is no QPPB classification if the IP filter action redirects the packet to a directly connected interface, even if X is matched by a route with a forwarding-class and priority
- QPPB classification is based on the forwarding-class and priority of the route matching IP address Y if the IP filter action redirects the packet to the indirect next-hop IP address Y, even if X is matched by a route with a forwarding-class and priority.

QPPB and GRT Lookup

Source-address based QPPB is not supported on any SAP or spoke SDP interface of a VPRN configured with the **grt-lookup** command.

QPPB Interaction with SAP Ingress QoS Policy

When QPPB is enabled on a SAP IP interface the forwarding class of a packet may change from **fc1**, the original **fc** determined by the SAP ingress QoS policy to fc2, the new fc determined by QPPB. In the ingress datapath SAP ingress QoS policies are applied in the first P chip and route lookup/QPPB occurs in the second P chip. This has the implications listed below:

- Ingress remarking (based on profile state) is always based on the original fc (fc1) and sub-class (if defined).
- The profile state of a SAP ingress packet that matches a QPPB route depends on the configuration of **fc2** only. If the de-1-out-profile flag is enabled in **fc2** and **fc2** is not mapped to a priority mode queue then the packet will be marked out of profile if its DE bit = 1. If the profile state of **fc2** is explicitly configured (in or out) and **fc2** is not mapped to a priority mode queue then the packet is assigned this profile state. In both cases there is no consideration of whether or not **fc1** was mapped to a priority mode queue.
- The priority of a SAP ingress packet that matches a QPPB route depends on several factors. If the de-1-out-profile flag is enabled in **fc2** and the DE bit is set in the packet then priority will be low regardless of the QPPB priority or **fc2** mapping to profile mode queue, priority mode queue or policer. If **fc2** is associated with a profile mode queue then the packet priority will be based on the explicitly configured profile state of **fc2** (in profile = high, out profile = low, undefined = high), regardless of the QPPB priority or **fc1** configuration. If **fc2** is associated with a priority mode queue or policer then the packet priority will be based on QPPB (unless DE=1), but if no priority information is associated with the route then the packet priority will be based on the configuration of **fc1** (if **fc1** mapped to a priority mode queue then it is based on DSCP/IP prec/802.1p and if **fc1** mapped to a profile mode queue then it is based on the profile state of **fc1**).

Table 10 summarizes these interactions.

Table 10: QPPB Interactions with SAP Ingress QoS

Original FC object mapping	New FC object mapping	Profile	Priority (drop preference)	DE=1 override	In/out of profile marking
Profile mode queue	Profile mode queue	From new base FC unless overridden by DE=1	From QPPB, unless packet is marked in or out of profile in which case follows profile. Default is high priority	From new base FC	From original FC and sub-class
Priority mode queue	Priority mode queue	Ignored	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then from original dot1p/ exp/DSCP mapping or policy default.	From new base FC	From original FC and sub-class
Policer	Policer	From new base FC unless overridden by DE=1	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then from original dot1p/ exp/DSCP mapping or policy default.	From new base FC	From original FC and sub-class
Priority mode queue	Policer	From new base FC unless overridden by DE=1	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then from original dot1p/ exp/DSCP mapping or policy default.	From new base FC	From original FC and sub-class
Policer	Priority mode queue	Ignored	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then from original dot1p/ exp/DSCP mapping or policy default.	From new base FC	From original FC and sub-class

Table 10: QPPB Interactions with SAP Ingress QoS (Continued)

Original FC object mapping	New FC object mapping	Profile	Priority (drop preference)	DE=1 override	In/out of profile marking
Profile mode queue	Priority mode queue	Ignored	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then follows original FC's profile mode rules.	From new base FC	From original FC and sub-class
Priority mode queue	Profile mode queue	From new base FC unless overridden by DE=1	From QPPB, unless packet is marked in or out of profile in which case follows profile. Default is high priority	From new base FC	From original FC and sub-class
Profile mode queue	Policer	From new base FC unless overridden by DE=1	If DE=1 override then low otherwise from QPPB. If no DEI or QPPB overrides then follows original FC's profile mode rules.	From new base FC	From original FC and sub-class
Policer	Profile mode queue	From new base FC unless overridden by DE=1	From QPPB, unless packet is marked in or out of profile in which case follows profile. Default is high priority	From new base FC	From original FC and sub-class

Object Grouping and State Monitoring

This feature introduces a generic operational group object which associates different service endpoints (pseudowires and SAPs) located in the same or in different service instances. The operational group status is derived from the status of the individual components using certain rules specific to the application using the concept. A number of other service entities, the monitoring objects, can be configured to monitor the operational group status and to perform certain actions as a result of status transitions. For example, if the operational group goes down, the monitoring objects will be brought down.

VPRN IP Interface Applicability

This concept is used by an IPv4 VPRN interface to affect the operational state of the IP interface monitoring the operational group. Individual SAP and spoke SDPs are supported as monitoring objects.

The following rules apply:

- An object can only belong to one group at a time.
- An object that is part of a group cannot monitor the status of a group.
- An object that monitors the status of a group cannot be part of a group.
- An operational group may contain any combination of member types: SAP or Spoke-SDPs.
- An operational group may contain members from different VPLS service instances.
- Objects from different services may monitor the oper-group.

There are two steps involved in enabling the functionality:

1. Identify a set of objects whose forwarding state should be considered as a whole group then group them under an operational group using the **oper-group** command.
2. Associate the IP interface to the oper-group using the **monitor-group** command

The status of the operational group (oper-group) is dictated by the status of one or more members according to the following rule:

- The oper-group goes down if all the objects in the oper-group go down. The oper-group comes up if at least one of the components is up.
- An object in the group is considered down if it is not forwarding traffic in at least one direction. That could be because the operational state is down or the direction is blocked through some validation mechanism.
- If a group is configured but no members are specified yet then its status is considered up.

- As soon as the first object is configured the status of the operational group is dictated by the status of the provisioned member(s).

The simple configuration below shows the oper-group g1, the VPLS SAP that is mapped to it and the IP interfaces in VPRN service 2001 monitoring the oper-group g1. This is example uses an R-VPLS context. The VPLS instance includes the **allow-ip-int-bind** and the **service-name v1**. The VPRN interface links to the VPLS using the **vpls v1** option. All commands are under the configuration service hierarchy.

To further explain the configuration. Oper-group g1 has a single SAP (1/1/1:2001) mapped to it and the IP interfaces in the VPRN service 2001 will derive its state from the state of oper-group g1.

```
oper-group g1 create

vpls 1 customer 1 create
       allow-ip-int-bind
       stp
           shutdown
       exit
       service-name "v1"
       sap 1/1/1:2001 create
           oper-group g1
           eth-cfm
              mep domain 1 association 1 direction down
     ccm-enable
      no shutdown
           exit
       exit
       sap 1/1/2:2001 create
       exit
       sap 1/1/3:2001 create
       exit
no shutdown


vprn 2001 customer 1 create
       interface "i2001"  create
           address 21.1.1.1/24
           monitor-oper-group "g1"
           vpls "v1"
       exit
     no shutdown
     exit
```

Subscriber Interfaces

Subscriber interfaces are composed of a combination of two key technologies, subscriber interfaces and group interfaces. While the subscriber interface defines the subscriber subnets, the group interfaces are responsible for aggregating the SAPs.

- Subscriber interface — An interface that allows the sharing of a subnet among one or many group interfaces in the routed CO model.
- Group interface — Aggregates multiple SAPs on the same port.
- Redundant interfaces — A special spoke-terminated Layer 3 interface. It is used in a Layer 3 routed CO dual-homing configuration to shunt downstream (network to subscriber) to the active node for a given subscriber.

SAPs

Encapsulations

The following SAP encapsulations are supported on the 7750 SR VPRN service:

- Ethernet null
- Ethernet dot1q
- SONET/SDH IPCP
- SONET/SDH ATM
- ATM - LLC SNAP or VC-MUX
- Cisco HDLC
- QinQ
- LAG
- Tunnel (IPSec or GRE)
- Frame Relay

ATM SAP Encapsulations for VPRN Services

The SR-Series series supports ATM PVC service encapsulation for VPRN SAPs. Both UNI and NNI cell formats are supported. The format is configurable on a SONET/SDH path basis. A path maps to an ATM VC. All VCs on a path must use the same cell format.

The following ATM encapsulation and transport modes are supported:

- RFC 2684, *Multiprotocol Encapsulation over ATM Adaptation Layer 5*:
  - → AAL5 LLC/SNAP IPv4 routed
  - → AAL5 VC mux IPv4 routed
  - → AAL5 LLC/SNAP IPv4 bridged
  - → AAL5 VC mux IPv4 bridged

Pseudowire SAPs

Pseudowire SAPs are supported on VPRN interfaces in the same way as on IES interfaces. For details of pseudowire SAPs, see Pseudowire SAPs on page 37.

QoS Policies

When applied to a VPRN SAP, service ingress QoS policies only create the unicast queues defined in the policy if PIM is not configured on the associated IP interface; if PIM is configured, the multipoint queues are applied as well.

With VPRN services, service egress QoS policies function as with other services where the class-based queues are created as defined in the policy.

Note that both Layer 2 or Layer 3 criteria can be used in the QoS policies for traffic classification in an VPRN.

Filter Policies

Ingress and egress IPv4 and IPv6 filter policies can be applied to VPRN SAPs.

DSCP Marking

Specific DSCP, forwarding class, and Dot1P parameters can be specified to be used by every protocol packet generated by the VPRN. This enables prioritization or de-prioritization of every protocol (as required). The markings effect a change in behavior on ingress when queuing. For example, if OSPF is not enabled, then traffic can be de-prioritized to best effort (be) DSCP. This change de-prioritizes OSPF traffic to the CPU complex.

DSCP marking for internally generated control and management traffic by marking the DSCP value should be used for the given application. This can be configured per routing instance. For example, OSPF packets can carry a different DSCP marking for the base instance and then for a VPRN service. ISIS and ARP traffic is not an IP-generated traffic type and is not configurable.

When an application is configured to use a specified DSCP value then the MPLS EXP, Dot1P bits will be marked in accordance with the network or access egress policy as it applies to the logical interface the packet will be egressing.

The DSCP value can be set per application. This setting will be forwarded to the egress linecard. The egress linecard does not alter the coded DSCP value and marks the LSP-EXP and IEEE 802.1p (Dot1P) bits according to the appropriate network or access QoS policy.

Table 11: DSCP/FC Marking

Protocol	IPv4	IPv6	DSCP Marking	Dot1P Marking	Default FC
ARP				Yes	NC
BGP	Yes	Yes	Yes	Yes	NC
BFD	Yes		Yes	Yes	NC
RIP	Yes	Yes	Yes	Yes	NC
PIM (SSM)	Yes	Yes	Yes	Yes	NC
OSPF	Yes	Yes	Yes	Yes	NC
SMTP	Yes				AF
IGMP/MLD	Yes	Yes	Yes	Yes	AF
Telnet	Yes	Yes	Yes	Yes	AF
TFTP	Yes		Yes	Yes	AF
FTP	Yes				AF

Table 11: DSCP/FC Marking (Continued)

Protocol	IPv4	IPv6	DSCP Marking	Dot1P Marking	Default FC
SSH (SCP)	Yes	Yes	Yes	Yes	AF
SNMP (get, set, etc.)	Yes	Yes	Yes	Yes	AF
SNMP trap/log	Yes	Yes	Yes	Yes	AF
syslog	Yes	Yes	Yes	Yes	AF
OAM ping	Yes	Yes	Yes	Yes	AF
ICMP ping	Yes	Yes	Yes	Yes	AF
Traceroute	Yes	Yes	Yes	Yes	AF
TACPLUS	Yes	Yes	Yes	Yes	AF
DNS	Yes	Yes	Yes	Yes	AF
SNTP/NTP	Yes				AF
RADIUS	Yes				AF
Cflowd	Yes				AF
DHCP	Yes	Yes	Yes	Yes	AF
Bootp	Yes				AF
IPv6 Neighbor Discovery	Yes				NC

Default DSCP Mapping Table

```
DSCP NameDSCP ValueDSCP ValueDSCP ValueLabel
     DecimalHexadecimalBinary
==========================================================
Default0  0x00 0b000000be
nc1  48   0x30 0b110000h1
nc2  56   0x38 0b111000nc
ef   46   0x2e 0b101110ef
af11 10   0x0a 0b001010assured
af12 12   0x0c 0b001100assured
af13 14   0x0e 0b001110assured
af21 18   0x12 0b010010l1
af22 20   0x14 0b010100l1
af23 22   0x16 0b010110l1
af31 26   0x1a 0b011010l1
af32 28   0x1c 0b011100l1
af33 30   0x1d 0b011110l1
af41 34   0x22 0b100010h2
af42 36   0x24 0b100100h2
af43 38   0x26 0b100110h2

default*0
```

*The default forwarding class mapping is used for all DSCP names/values for which there is no explicit forwarding class mapping.

Configuration of TTL Propagation for VPRN Routes

This feature allows the separate configuration of TTL propagation for in transit and CPM generated IP packets, at the ingress LER within a VPRN service context. The following commands are supported:

- `config router ttl-propagate vprn-local [none | vc-only | all]`
- `config router ttl-propagate vprn-transit [none | vc-only | all]`

You can enable TTL propagation behavior separately as follows:

- for locally generated packets by CPM (vprn-local)
- for user and control packets in transit at the node (vprn-transit)

The following parameters can be specified:

- The **all** parameter enables TTL propagation from the IP header into all labels in the stack, for VPN-IPv4 and VPN-IPv6 packets forwarded in the context of all VPRN services in the system.

- The **vc-only** parameter reverts to the default behavior by which the IP TTL is propagated into the VC label but not to the transport labels in the stack. You can explicitly set the default behavior by configuring the vc-only value.
- The **none** parameter disables the propagation of the IP TTL to all labels in the stack, including the VC label. This is needed for a transparent operation of UDP traceroute in VPRN inter-AS option B such that the ingress and egress ASBR nodes are not traced.

This command does not use a no version.

The user can override the global configuration within each VPRN instance using the following commands:

- `config service vprn ttl-propagate local [inherit | none | vc-only | all]`
- `config service vprn ttl-propagate transit [inherit | none | vc-only | all]`

Note the default behavior for a VPRN instance is to inherit the global configuration for the same command. You can explicitly set the default behavior by configuring the inherit value.

This command does not have a no version.

The commands do not apply when the VPRN packet is forwarded over GRE transport tunnel.

If a packet is received in a VPRN context and a lookup is done in the Global Routing Table (GRT), (when leaking to GRT is enabled for example), the behavior of the TTL propagation is governed by the LSP shortcut configuration as follows:

- when the matching route is an RSVP LSP shortcut:

  → `configure router mpls shortcut-transit-ttl-propagate`
- when the matching route is an LDP LSP shortcut:

  → `configure router ldp shortcut-transit-ttl-propagate`

When the matching route is a RFC 3107 label route or a 6PE route, It is governed by the BGP label route configuration

When a packet is received on one VPRN instance and is redirected using Policy Based Routing (PBR) to be forwarded in another VPRN instance, the TTL propagation is governed by the configuration of the outgoing VPRN instance.

Note that packets that are forwarded in different contexts can use different TTL propagation over the same BGP tunnel, depending on the TTL configuration of each context. An example of this might be VPRN using a BGP tunnel and an IPv4 packet forwarded over a BGP label route of the same prefix as the tunnel.

CE to PE Routing Protocols

The 7750 SR VPRN supports the following PE to CE routing protocols:

- BGP
- Static
- RIP
- OSPF

PE to PE Tunneling Mechanisms

The 7750 SR supports multiple mechanisms to provide transport tunnels for the forwarding of traffic between PE routers within the 2547bis network.

The 7750 SR VPRN implementation supports the use of:

- RSVP-TE protocol to create tunnel LSP's between PE routers
- LDP protocol to create tunnel LSP's between PE routers
- GRE tunnels between PE routers.

These transport tunnel mechanisms provide the flexibility of using dynamically created LSPs where the service tunnels are automatically bound (the “autobind” feature) and the ability to provide certain VPN services with their own transport tunnels by explicitly binding SDPs if desired. When the autobind is used, all services traverse the same LSPs and do not allow alternate tunneling mechanisms (like GRE) or the ability to craft sets of LSP's with bandwidth reservations for specific customers as is available with explicit SDPs for the service.

Per VRF Route Limiting

The 7750 SR allows setting the maximum number of routes that can be accepted in the VRF for a VPRN service. There are options to specify a percentage threshold at which to generate an event that the VRF table is near full and an option to disable additional route learning when full or only generate an event.

Spoke SDPs

Distributed services use service distribution points (SDPs) to direct traffic to another SR-Series router via service tunnels. SDPs are created on each participating SR-Series and then bound to a specific service. SDP can be created as either GRE or MPLS. Refer to the *Services Overview Guide* for information about configuring SDPs.

This feature provides the ability to cross-connect traffic entering on a spoke SDP, used for Layer 2 services (VLLs or VPLS), on to an IES or VPRN service. From a logical point of view, the spoke SDP entering on a network port is cross-connected to the Layer 3 service as if it entered by a service SAP. The main exception to this is traffic entering the Layer 3 service by a spoke SDP is handled with network QoS policies not access QoS policies.

Figure 13 depicts traffic terminating on a specific IES or VPRN service that is identified by the *sdp-id* and VC label present in the service packet.

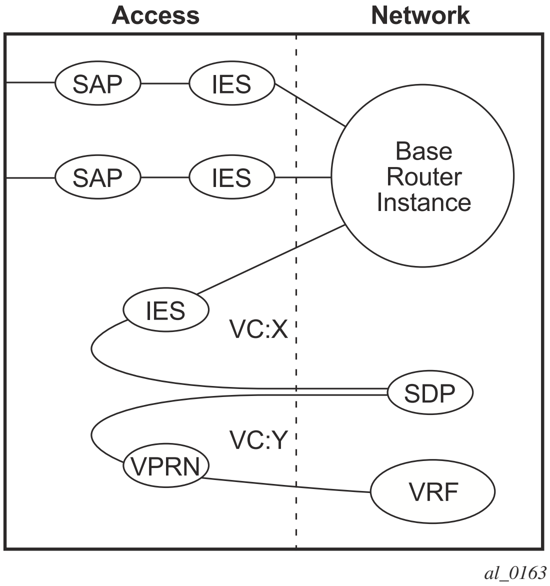


Figure 13: SDP-ID and VC Label Service Identifiers

T-LDP Status Signaling for Spoke-SDPs Terminating on IES/VPRN

T-LDP status signaling and PW active/standby signaling capabilities are supported on ipipe and epipe spoke SDPs.

Spoke SDP termination on an IES or VPRN provides the ability to cross-connect traffic entering on a spoke SDP, used for Layer 2 services (VLLs or VPLS), on to an IES or VPRN service. From a logical point of view the spoke SDP entering on a network port is cross-connected to the Layer 3 service as if it had entered using a service SAP. The main exception to this is traffic entering the Layer 3 service using a spoke SDP is handled with network QoS policies instead of access QoS policies.

When a SAP down or SDP binding down status message is received by the PE in which the Ipipe or Ethernet spoke-sdp is terminated on an IES or VPRN interface, the interface is brought down and all associated routes are withdrawn in a similar way when the spoke-sdp goes down locally. The same actions are taken when the standby T-LDP status message is received by the IES/VPRN PE.

This feature can be used to provide redundant connectivity to a VPRN or IES from a PE providing a VLL service, as shown in Figure 14.

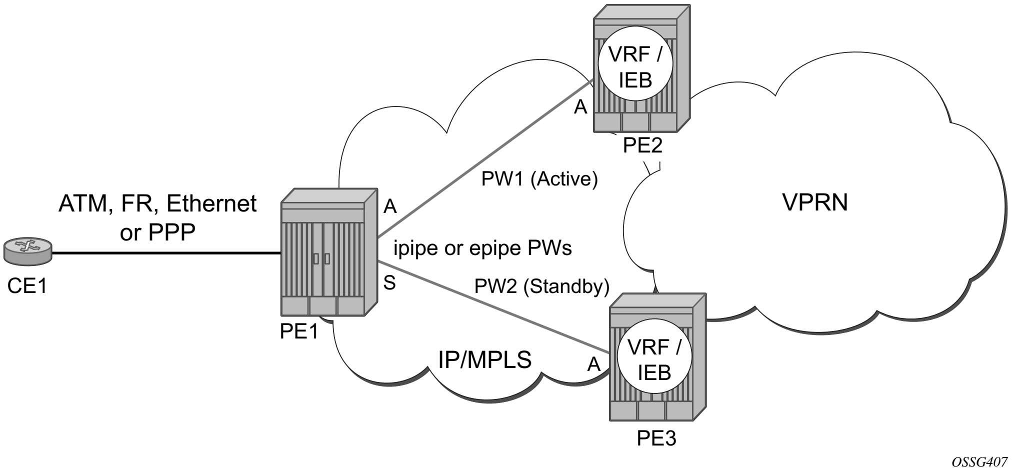


Figure 14: Active/Standby VRF Using Resilient Layer 2 Circuits

Spoke SDP Redundancy into IES/VPRN

This feature can be used to provide redundant connectivity to a VPRN or IES from a PE providing a VLL service, as shown in Figure 14, using either epipe or ipipe spoke-SDPs.

In Figure 14, PE1 terminates two spoke-SDPs that are bound to one SAP connected to CE1. PE1 chooses to forward traffic on one of the spoke SDPs (the active spoke-SDP), while blocking traffic on the other spoke-SDP (the standby spoke-SDP) in the transmit direction. PE2 and PE3 take any spoke-SDPs for which PW forwarding standby has been signaled by PE1 to an operationally down state.

Note that 7x50, 7710 routers are expected to fulfill both functions (VLL and VPRN/IES PE), while the 7705 must be able to fulfill the VLL PE function. Figure 15 illustrates the model for spoke-SDP redundancy into a VPRN or IES.

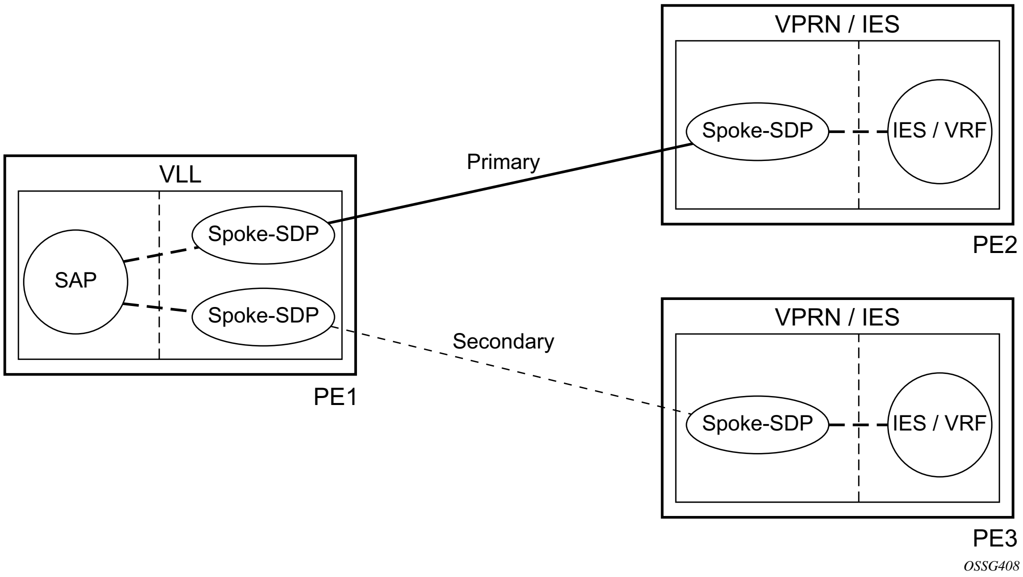


Figure 15: Spoke-SDP Redundancy Model

IP-VPNs

Using OSPF in IP-VPNs

Using OSPF as a CE to PE routing protocol allows OSPF that is currently running as the IGP routing protocol to migrate to an IP-VPN backbone without changing the IGP routing protocol, introducing BGP as the CE-PE or relying on static routes for the distribution of routes into the service providers IP-VPN. The following features are supported:

- Advertisement/redistribution of BGP-VPN routes as summary (type 3) LSAs flooded to CE neighbors of the VPRN OSPF instance. This occurs if the OSPF route type (in the OSPF route type BGP extended community attribute carried with the VPN route) is not external (or NSSA) and the locally configured domain-id matches the domain-id carried in the OSPF domain ID BGP extended community attribute carried with the VPN route.
- OSPF sham links. A sham link is a logical PE-to-PE unnumbered point-to-point interface that essentially rides over the PE-to-PE transport tunnel. A sham link can be associated with any area and can therefore appear as an intra-area link to CE routers attached to different PEs in the VPN.

IPCP Subnet Negotiation

This feature enables negotiation between Broadband Network Gateway (BNG) and customer premises equipment (CPE) so that CPE is allocated both ip-address and associated subnet.

Some CPEs use the network up-link in PPPoE mode and perform dhcp-server function for all ports on the LAN side. Instead of wasting 1 subnet for p2p uplink, CPEs use allocated subnet for LAN portion as shown in Figure 16.

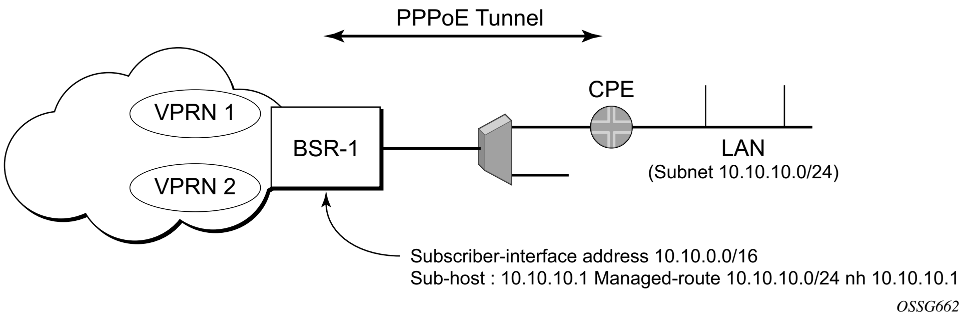

PPPoE Tunnel
CPE
VPRN 1
BSR-1
VPRN 2
LAN
(Subnet 10.10.10.0/24)
Subscriber-interface address 10.10.0.0/16
Sub-host : 10.10.10.1 Managed-route 10.10.10.0/24 nh 10.10.10.1
OSSG662

Figure 16: CPEs Network Up-link Mode

From a BNG perspective, the given PPPoE host is allocated a subnet (instead of /32) by RADIUS, external dhcp-server, or local-user-db. And locally, the host is associated with managed-route. This managed-route will be subset of the subscriber-interface subnet, and also, subscriber-host ip-address will be from managed-route range. The negotiation between BNG and CPE allows CPE to be allocated both ip-address and associated subnet.

Cflowd for IP-VPNs

The cflowd feature allows service providers to collect IP flow data within the context of a VPRN. This data can used to monitor types and general proportion of traffic traversing an VPRN context. This data can also be shared with the VPN customer to see the types of traffic traversing the VPN and use it for traffic engineering.

This feature should not be used for billing purposes. Existing queue counters are designed for this purpose and provide very accurate per bit accounting records.

Inter-AS VPRNs

Inter-AS IP-VPN services have been driven by the popularity of IP services and service provider expansion beyond the borders of a single Autonomous System (AS) or the requirement for IP VPN services to cross the AS boundaries of multiple providers. Three options for supporting inter-AS IP-VPNs are described in RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*.

The first option, referred to as Option-A (Figure 17), is considered inherent in any implementation. This method uses a back-to-back connection between separate VPRN instances in each AS. As a result, each VPRN instance views the inter-AS connection as an external interface to a remote VPRN customer site. The back-to-back VRF connections between the ASBR nodes require individual sub-interfaces, one per VRF.

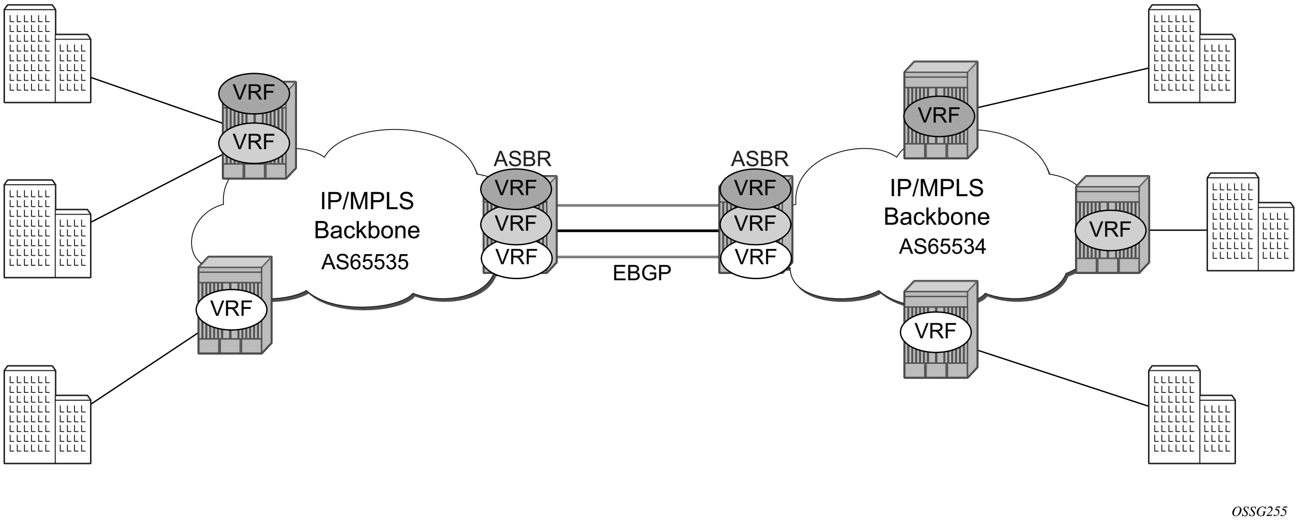


Figure 17: Inter-AS Option-A: VRF-to-VRF Model

The second option, referred to as Option-B (Figure 18), relies heavily on the AS Boundary Routers (ASBRs) as the interface between the autonomous systems. This approach enhances the scalability of the eBGP VRF-to-VRF solution by eliminating the need for per-VPRN configuration on the ASBR(s). However it requires that the ASBR(s) provide a control plan and forwarding plane connection between the autonomous systems. The ASBR(s) are connected to the PE nodes in its local autonomous system using iBGP either directly or through route reflectors. This means the ASBR(s) receive all the VPRN information and will forward these VPRN updates, VPN-IPV4, to all its EBGP peers, ASBR(s), using itself as the next-hop. It also changes the label associated with the route. This means the ASBR(s) must maintain an associate mapping of labels received and labels issued for those routes. The peer ASBR(s) will in turn forward those updates to all local IBGP peers.

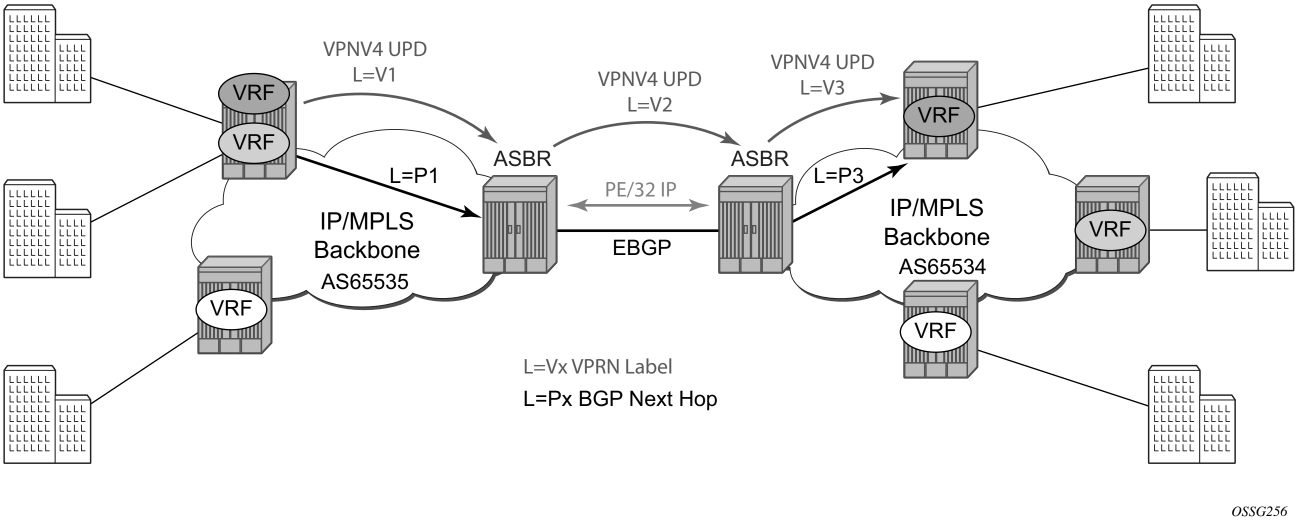


Figure 18: Inter-AS Option-B

This form of inter-AS VPRNs does not require instances of the VPRN to be created on the ASBR, as in option-A, as a result there is less management overhead. This is also the most common form of Inter-AS VPRNs used between different service providers as all routes advertised between autonomous systems can be controlled by route policies on the ASBRs by the **config>router>bgp>transport-tunnel** CLI command.

The third option, referred to as Option-C (Figure 19), allows for a higher scale of VPRNs across AS boundaries but also expands the trust model between ASNs. As a result this model is typically used within a single company that may have multiple ASNs for various reasons.

This model differs from Option-B, in that in Option-B all direct knowledge of the remote AS is contained and limited to the ASBR. As a result, in option-B the ASBR performs all necessary mapping functions and the PE routers do not need perform any additional functions then in a non-Inter-AS VPRN.

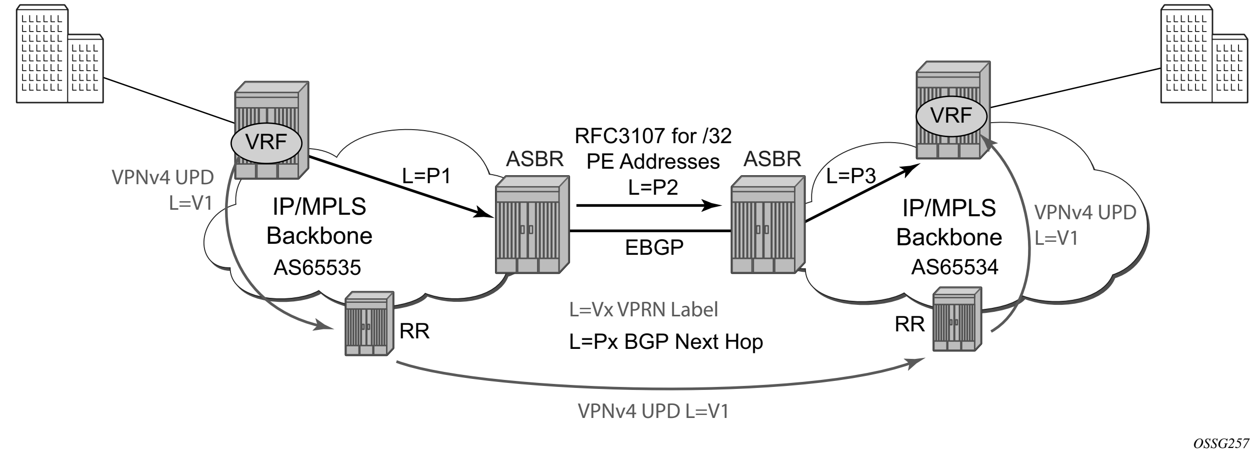


Figure 19: Option C Example

With Option-C, knowledge from the remote AS is distributed throughout the local AS. This distribution allows for higher scalability but also requires all PEs and ASBRs involved in the Inter-AS VPRNs to participate in the exchange of inter-AS routing information.

In Option-C, the ASBRs distribute reachability information for remote PE's system IP addresses only. This is done between the ASBRs by exchanging MP-eBGP labeled routes, using RFC 3107, *Carrying Label Information in BGP-4* . Either RSVP-TE or LDP LSP can be selected to resolve next-hop for multi-hop eBGP peering by the **config>router>bgp>transport-tunnel** CLI command.

Distribution of VPRN routing information is handled by either direct MP-BGP peering between PEs in the different ASNs or more likely by one or more route reflectors in ASN.

Carrier Supporting Carrier (CsC)

Carrier Supporting Carrier (CSC) is a solution that allows one service provider (the “Customer Carrier”) to use the IP VPN service of another service provider (the “Super Carrier”) for some or all of its backbone transport. RFC 4364 defines a Carrier Supporting Carrier solution for BGP/MPLS IP VPNs that uses MPLS on the interconnections between the two service providers in order to provide a scalable and secure solution.

CsC support in SROS allows a 7x50 to be deployed as any of the following devices shown in Figure 20:

- PE1 (service provider PE)
- CSC-CE1, CSC-CE2 and CSC-CE3 (CE device from the point of view of the backbone service provider)
- CSC-PE1, CSC-PE2 and CSC-PE3 (PE device of the backbone service provider)
- ASBR1 and ASBR2 (ASBR of the backbone service provider)

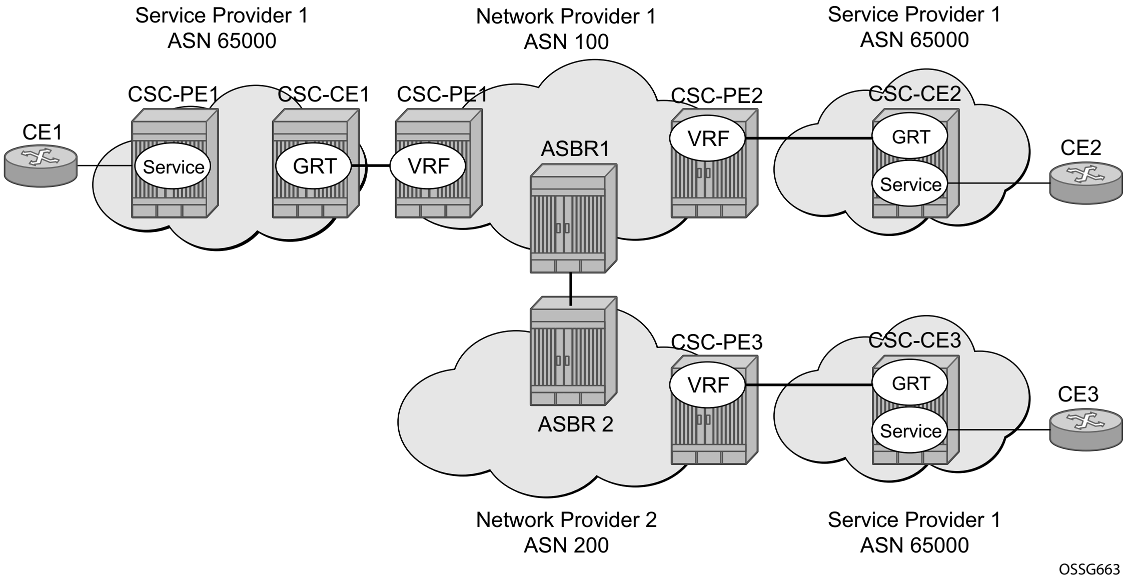


Figure 20: Carrier Supporting Carrier Reference Diagram

Terminology

CE — Customer premises equipment dedicated to one particular business/enterprise.

PE — Service provider router that connects to a CE to provide a business VPN service to the associated business/enterprise.

CSC-CE — An ASBR/peering router that is connected to the CSC-PE of another service provider for purposes of using the associated CsC IP VPN service for backbone transport.

CSC-PE — A PE router belonging to the backbone service provider that supports one or more CSC IP VPN services.

CSC Connectivity Models

A PE router deployed by a customer service provider to provide Internet access, IP VPNs, and/or L2 VPNs may connect directly to a CSC-PE device, or it may back haul traffic within its local “site” to the CSC-CE that provides this direct connection. Here, “site” means a set of routers owned and managed by the customer service provider that can exchange traffic through means other than the CSC service. The function of the CSC service is to provide IP/MPLS reachability between isolated sites.

The CSC-CE is a “CE” from the perspective of the backbone service provider. There may be multiple CSC-CEs at a given customer service provider site and each one may connect to multiple CSC-PE devices for resiliency/multi-homing purposes.

The CSC-PE is owned and managed by the backbone service provider and provides CSC IP VPN service to connected CSC-CE devices. In many cases, the CSC-PE also supports other services, including regular business IP VPN services. A single CSC-PE may support multiple CSC IP VPN services. Each customer service provider is allocated its own VRF within the CSC-PE; VRFs maintain routing and forwarding separation and permit the use of overlapping IP addresses by different customer service providers.

A backbone service provider may not have the geographic span to connect, with reasonable cost, to every site of a customer service provider. In this case, multiple backbone service providers may coordinate to provide an inter-AS CSC service. Different inter-AS connectivity options are possible, depending on the trust relationships between the different backbone service providers.

CSC-PE Configuration and Operation

This section applies to CSC-PE1, CSC-PE2 and CSC-PE3 in Figure 20 on page 273.

CSC Interface

From the point of view of the CSC-PE, the IP/MPLS interface between the CSC-PE and a CSC-CE has these characteristics:

1. The CSC interface is associated with one (and only one) VPRN service. Routes with the CSC interface as next-hop are installed only in the routing table of the associated VPRN.
2. The CSC interface supports EBGP or IBGP for exchanging labeled IPv4 routes (RFC 3107). The BGP session may be established between the interface addresses of the two routers or else between a loopback address of the CSC-PE VRF and a loopback address of the CSC-CE. In the latter case, the BGP next-hop is resolved by either a static or OSPFv2 route.
3. An MPLS packet received on a CSC interface is dropped if the top-most label was not advertised over a BGP (RFC 3107) session associated with one of the VPRN's CSC interfaces.
4. The CSC interface supports ingress QoS classification based on 802.1p or MPLS EXP. It is possible to configure a default FC and default profile for the CSC interface.
5. The CSC interface supports QoS (re)marking for egress traffic. Policies to remark 802.1p or MPLS EXP based on forwarding-class and profile are configurable per CSC interface.
6. By associating a port-based egress queue group instance with a CSC interface, the egress traffic can be scheduled/shaped with per-interface, per-forwarding-class granularity.
7. By associating a forwarding-plane based ingress queue group instance with a CSC interface, the ingress traffic can be policed to per-interface, per-forwarding-class granularity.
8. Ingress and egress statistics and accounting are available per CSC interface. The exact set of collected statistics depends on whether a queue-group is associated with the CSC interface, the traffic direction (ingress vs. egress), and the stats mode of the queue-group policers.

CSC interfaces are only supported on Ethernet ports and LAGs residing on FP2 or better cards and systems. An Ethernet port or LAG with a CSC interface can be configured in hybrid mode or network mode. The port or LAG supports null, dot1q or qinq encapsulation. To create a CSC interface on a port or LAG in null mode, the following commands are used:

config>service>vprn>network-interface>port *port-id*
config>service>vprn>network-interface>lag *lag-id*

To create a CSC interface on a port or LAG in dot1q mode, the following commands are used:

config>service>vprn>network-interface>port *port-id:qtag1*
config>service>vprn>network-interface>lag *port-id:qtag1*

To create a CSC interface on a port or LAG in qinq mode, the following commands are used:

config>service>vprn>network-interface>port *port-id:qtag1.qtag2*
config>service>vprn>network-interface>port *port-id:qtag1.**
config>service>vprn>network-interface>lag *port-id:qtag1.qtag2*
config>service>vprn>network-interface>lag *port-id:qtag1.**

A CSC interface supports the same capabilities (and supports the same commands) as a base router network interface except it does not support:

- IPv6
- LDP
- RSVP
- Proxy ARP (local/remote)
- Network domain configuration
- DHCP
- Ethernet CFM
- Unnumbered interfaces

QoS

Egress

Egress traffic on a CSC interface can be shaped and scheduled by associating a port-based egress queue-group instance with the CSC interface. The steps for doing this are summarized below:

1. Create an egress queue-group-template.
2. Define one or more queues in the egress queue-group. For each one specify scheduling parameters such as CIR, PIR, CBS and MBS and, if using H-QoS, the parent scheduler.
3. Apply an instance of the egress queue-group template to the network egress context of the Ethernet port with the CSC interface. When doing so, and if applicable, associate an accounting policy and/or a scheduler policy with this instance.
4. Create a network QoS policy.
5. In the egress part of the network QoS policy define EXP remarking rules, if necessary.
6. In the egress part of the network QoS policy map a forwarding-class to a queue-id using the port-redirect-group command. For example:

   ```
   config>qos>network>egress>fc$ port-redirect-group queue 5
   ```
7. Apply the network QoS policy created in step 4 to the CSC interface and specify the name of the egress queue-group created in step 1 and the specific instance defined in step 3.

Ingress

Ingress traffic on a CSC interface can be policed by associating a forwarding-plane based ingress queue-group instance with the CSC interface. The steps for doing this are summarized below:

1. Create an ingress queue-group-template.
2. Define one or more policers in the ingress queue-group. For each one specify parameters such as CIR, PIR, CBS and MBS and, if using H-Pol, the parent arbiter.
3. Apply an instance of the ingress queue-group template to the network ingress context of the forwarding plane (FP2 or better) with the CSC interface. When doing so, and if applicable, associate an accounting policy and/or a policer-control-policy with this instance.
4. Create a network QoS policy.
5. In the ingress part of the network QoS policy define EXP classification rules, if necessary.
6. In the ingress part of the network QoS policy map a forwarding-class to a policer-id using the fp-redirect-group policer command. For example:

   ```
   config>qos>network>ingress>fc$ fp-redirect-group policer 5
   ```
7. Apply the network QoS policy created in step 4 to the CSC interface and specify the name of the ingress queue-group created in step 1 and the specific instance defined in step 3.

MPLS

BGP-3107 is used as the label distribution protocol on the CSC interface. When BGP in a CSC VPRN needs to distribute a label corresponding to a received VPN-IPv4 route, it takes the label from the global label space. The allocated label will not be re-used for any other FEC regardless of the routing instance (base router or VPRN). If a label L is advertised to the BGP peers of CSC VPRN A then a received packet with label L as the top most label is only valid if received on an interface of VPRN A, otherwise the packet is discarded.

To use BGP-3107 as the label distribution protocol on the CSC interface, add the **advertise-label ipv4** command to the BGP neighbor configuration. This command causes the capability to send and receive labeled-IPv4 routes {AFI=1, SAFI=4} to be negotiated with the CSC-CE peers.

CSC VPRN Service Configuration

To configure a VPRN to support CSC service, the **carrier-carrier-vpn** command must be enabled in its configuration. The command will fail if the VPRN has any existing SAP or spoke-SDP interfaces. A CSC interface can be added to a VPRN (using the **network-interface** command) only if the **carrier-carrier-vpn** command is present.

A VPRN service with the **carrier-carrier-vpn** command may be provisioned to use auto-bind, configured spoke SDPs or some combination. All SDP types are supported except for:

- GRE SDPs

- LDP over RSVP-TE tunnel SDPs

Other aspects of VPRN configuration are the same in a CSC model as in a non-CSC model.

Traffic Leaking to GRT

Traffic leaking to Global Route Table (GRT) allows service providers to offer VPRN and Internet services to their customers over a single VRF interface. This currently supports IPv4 and requires the customer VRPN interfaces to terminate on a minimum of IOM3-XP and IMM hardware. It is also supported on the 7750 SR-C12.

Packets entering a local VRF interface can have route processing results derived from the VPRN forwarding table or the GRT.   The leaking and preferred lookup results are configured on a per VPRN basis. Configuration options can be general (for example, any lookup miss in the VRPN forwarding table can be resolved in the GRT), or specific (for example, specific route(s) should only be looked up in the GRT and ignored in the VPRN). In order to provide operational simplicity and improve streamlining, the CLI configuration is all contained within the context of the VPRN service.

This feature is enabled within the VPRN service context under **config>service>vprn>grt-lookup**. This is an administrative context and provides the container under which all specific commands can be entered, except for policy definition. Policy definitions remain unchanged but are referenced from this context.

The **enable-grt** command establishes the basic functionality. When it is configured, any lookup miss in the VRF table will be resolved in the GRT, if available. By itself, this only provides part of the solution. Packet forwarding within GRT must understand how to route packets back to the proper node and to the specific VPRN from which the destination exists. Destination prefixes must be leaked from the VPRN to the GRT through the use of policy. Policies are created under the **config>router>policy-options** hierarchy. By default, the number of prefixes leaked from the VPRN to the GRT is limited to five. The **export-limit** command under the **grt-lookup** hierarchy allows the service provider to override the default, or remove the limit.

When a VPRN forwarding table consists of a default route or an aggregate route, the customer may require the service provider to poke holes in those, or provide more specific route resolution in the GRT. In this case, the service provider may configure a “static-route”, under the “enable-grt” context. The lookup result will prefer any successful lookup in the GRT that is equal to or more specific than the static route, bypassing any successful lookup in the local VPRN.

This feature and Unicast Reverse Path Forwarding (uRPF) are mutually exclusive. When a VPRN service is configured with either of these functions, the other cannot be enabled.   Also, prefixes leaked from any VPRN should never conflict with prefixes leaked from any other VPRN or existing prefixes in the GRT. Prefixes should be globally unique with the service provider network and if these are propagated outside of a single providers network, they must be from the public IP space and globally unique. Network Address Translation (NAT) is not supported as part of this feature.The following type of routes will not be leaked from VPRN into the Global Routing Table (GRT):

- Aggregate routes
- Blackhole routes

- BGP VPN extranet routes

Traffic Leaking from VPRN to GRT for IPv6

This feature allows IPv6 destination lookups in two distinct routing tables. IPv6 packets within a Virtual Private Routed Network (VPRN) service is able to perform a lookup for IPv6 destination against the Global Route Table (GRT) as well as within the local VPRN.

Currently, VPRN to VPRN routing exchange is accomplished through the use of import and export policies based on Route Targets (RTs), the creation of extranets. This new feature allows the use of a single VPRN interface for both corporate VPRN routing and other services (for example, Internet) that are reachable outside the local routing context. This feature takes advantage of the dual lookup capabilities in two separate routing tables in parallel.

This feature enables IPv6 capability in addition to the existing IPv4 VPRN-to-GRT Route Leaking feature.

RIP Metric Propagation in VPRNs

When RIP is used as the PE-CE protocol for VPRNs (IP-VPNs), the RIP metric is only used by the local node running RIP with the Customer Equipment (CE). The metric is not used to or encoded into and MP-BGP path attributes exchanged between PE routers.

The RIP metric can also be used to exchanged between PE routers so if a customer network is dual homed to separate PEs the RIP metric learned from the CE router can be used to choose the best route to the destination subnet. By using the learned RIP metric to set the BGP MED attribute, remote PEs can choose the lowest MED and in turn the PE with the lowest advertised RIP metric as the preferred egress point for the VPRN.

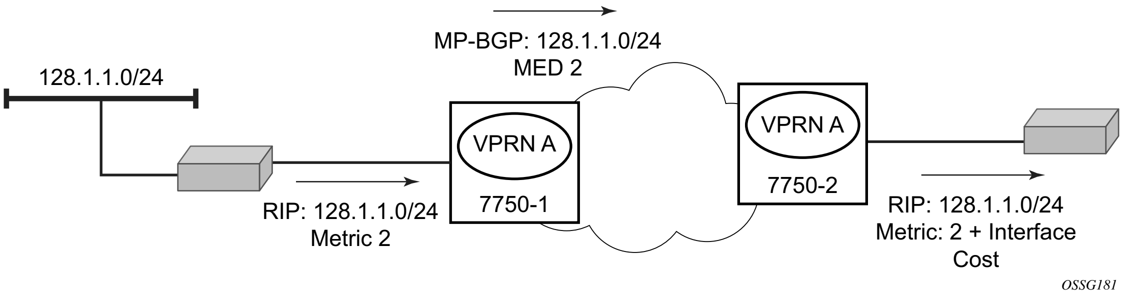


Figure 21: RIP Metric Propagation in VPRNs

NTP Within a VPRN Service

The NTP within a VPRN service enables the service router to act as the NTP server to CPE devices on a VPRN interface. Individual VPRN interfaces may be configured to listen to and respond to client requests, or additionally may be configured to send NTP broadcast messages. Authentication keys are configurable on a per-VPRN basis.

Only a single instance of NTP remains in the node that is time sourced to as many as five NTP servers attached to the “base” or “management” network.

The NTP show command displays NTP servers and all known clients. Because NTP is UDP based only, no state is maintained. As a result, the show command output only displays when the last message from the client was received.

PTP Within a VPRN Service

The PTP within a VPRN service provides access to the PTP clock within the 7750 SR through one or more VPRN services. Only one VPRN or the base routing instance may have configured peers, but all may have discovered peers. If desired, a limit on the maximum number of dynamic peers allowed may be configured on a per routing instance basis.

For more detail on PTP see the Basic System Configuration Guide.

VPN Route Label Allocation

The method used for allocating a label value to an originated VPN-IP route (exported from a VPRN) depends on the configuration of the VPRN service and its VRF export policies. SR OS supports three 3 label modes:

- label per VRF
- label per next hop (LPN)
- label per prefix (LPP)

Label per VRF is the label allocation default. It is used when the label mode is configured as VRF (or not configured) and the VRF export policies do not apply an advertise-label per-prefix action. All routes exported from the VPRN with the per-VRF label have the same label value. When the PE receives a terminating MPLS packet with a per-VRF label, the label value selects the VRF context in which to perform a forwarding table lookup and this lookup determines the outgoing interface (or set of interfaces if ECMP applies).

Label per next hop is used when the exported route is not a local or aggregate route, the label mode is configured as next-hop, and the VRF export policies do not apply an advertise-label per-prefix override. It is also used when an inactive (backup path) BGP route is exported by effect of the export-inactive-bgp command if there is no advertise-label per-prefix override. All LPN-exported routes with the same primary next hop have the same per-next-hop label value. When the PE receives a terminating MPLS packet with a per-next-hop label, the label lookup selects the outgoing interface for forwarding, without any FIB lookup that might cause problems with overlapping prefixes. LPN does not support ECMP, BGP fast reroute, QPPB, or policy accounting features that might otherwise apply.

Label per-prefix is used when a qualifying IP route is exported by matching a VRF export policy action with advertise-label per-prefix. Any IPv4 or IPv6 route that is not a local route, aggregate route, BGP-VPN route, or GRT lookup static route qualifies. With LPP, every prefix is associated with its own unique label value that does not change while the route is present in the route table. When the PE receives a terminating MPLS packet with a per-prefix label value, the packet is forwarded as if the FIB lookup found only the matching prefix route and not any of the more specific prefix routes that would normally be selected. LPP supports ECMP, QPPB, and policy accounting as part of the egress forwarding decision. It does not support BGP fast reroute or BGP sticky ECMP.

Table 12 summarizes the logic that determines the label allocation method for an exported route.

Table 12: Label allocation logic

1	If the IP route is LOCAL, AGGREGATE, or BGP-VPN always use the VRF label.
2	If the IP route is accepted by a VRF export policy with the advertise-label per-prefix action, use LPP.
3	If the IP (BGP) route is exported by the export-inactive-bgp command (VPRN best external), use LPN.
4	If the IP route is exported by a VPRN configured for label-mode next-hop, use LPN.
5	Else, use the per-VRF label.

Configuring the Service Label Mode

To change the service label mode of the VPRN the **config>service>vprn>label-mode** {**vrf** | **next-hop**} command is used:

The default mode (if the command is not present in the VPRN configuration) is vrf meaning distribution of one service label for all routes of the VPRN. When a VPRN X is configured with the label-mode next-hop command the service label that it distributes with an IPv4 or IPv6 route that it exports depends on the type of route as summarized in Table 13.

Table 13: Service Labels Distributed in Service Label per Next-Hop Mode

Route Type	Distributed Service Label
Remote route with IP A (associated with a SAP) as resolved next-hop	Platform-wide unique label allocated to next-hop A
Remote route with IP B (associated with a spoke SDP) as resolved next-hop	Platform-wide unique label allocated to next-hop B
Local route	Platform-wide unique label allocated to VPRN X
Aggregate route	Platform-wide unique label allocated to VPRN X
ECMP route	Platform-wide unique label allocated to next-hop A (the lowest next-hop address in the ECMP set)
BGP route with a backup next-hop (BGP FRR)	Platform-wide unique label allocated to next-hop A (the lowest next-hop address of the primary next-hops)

A change to the label-mode of a VPRN requires the VPRN to first be shutdown.

Restrictions and Usage Notes

The service label per next-hop mode has the following restrictions:

- ECMP — The VPRN label mode should be set to VRF if distribution of traffic across the multiple PE-CE next-hop interfaces of an ECMP route is desired.
- Hub and spoke VPN — The VPRN label mode should not be set to next-hop if the operator does not want the hub-connected CE to be involved in the forwarding of spoke-to-spoke traffic.
- BGP next-hop indirection — BGP next-hop indirection has no benefit in service label per next-hop mode. When the resolved next-hop interface of a BGP next-hop changes all of the affected BGP routes must be re-advertised to VPRN peers with the new service label corresponding to the new resolved next-hop.
- BGP anycast — When a PE failure results in redirection of MPLS packets to the other PE in a dual-homed pair, the service label mode is forced to VRF, for example, FIB lookup will determine the next-hop even if the label mode of the VPRN is configured as next-hop.
- U-turn routing — U-turn routing is effectively disabled by service-label per next-hop.
- Carrier Supporting Carrier — The label-mode configuration of a VPRN with CSC interfaces is ignored for BGP-3107 routes learned from connected CSC-CE devices.

QoS on Ingress Bindings

Traffic is tunneled between VPRN service instances on different PEs over service tunnels bound to MPLS LSPs or GRE tunnels. The binding of the service tunnels to the underlying transport is achieved either automatically (using the **auto-bind-tunnel** command) or statically (using the **spoke-sdp** command; not that under the VPRN IP interface). QoS control can be applied to the service tunnels for traffic ingressing into a VPRN service, see Figure 22.

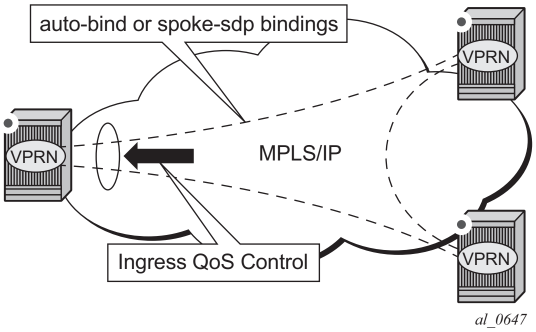


Figure 22: Ingress QoS Control on VPRN Bindings

An ingress queue group must be configured and applied to the ingress network FP where the traffic is received for the VPRN. All traffic received on that FP for any binding in the VPRN (either automatically or statically configured) which is redirected to a policer in the FP queue group (using **fp-redirect-qroup** in the network QoS policy) will be controlled by that policer. As a result, the traffic from all such bindings is treated as a single entity (per forwarding class) with regard to ingress QoS control. Any **fp-redirect-group multicast-policer, broadcast-policer** or **unknown-policer** commands in the network QoS policy are ignored for this traffic (IP multicast traffic would use the ingress network queues or queue group related to the network interface).

Ingress classification is based on the configuration of the ingress section of the specified network QoS policy, noting that the dot1p and exp classification is based on the outer Ethernet header and MPLS label whereas the DSCP applies to the outer IP header if the tunnel encapsulation is GRE, or the DSCP in the first IP header in the payload if **ler-use-dscp** is enabled in the ingress section of the referenced network QoS policy.

Ingress bandwidth control does not take into account the outer Ethernet header, the MPLS labels/ control word or GRE headers, or the FCS of the incoming frame.

The following command configures the association of the network QoS policy and the FP queue group and instance within the network ingress of a VPRN:

```
configure
    vprn
        network
```

```
ingress
    qos <network-policy-id> fp-redirect-group <queue-group-name>
                            instance <instance-id>
```

When this command is configured, it overrides the QoS applied to the related network interfaces for unicast traffic arriving on bindings in that VPRN. The IP and IPv6 criteria statements are not supported in the applied network QoS policy

This is supported for all available transport tunnel types and is independent of the label mode (**vrf** or **next-hop**) used within the VPRN. It is also supported for Carrier-Supporting-Carrier VPRNs.

The ingress network interfaces on which the traffic is received must be on FP2- and higher-based hardware. The above command is ignored on FP1-based hardware.

Multicast in IP-VPN Applications

This section and its subsections focuses on Multicast in IP VPN functionality. A reader should familiarize itself first with Multicast section in SROS Routing Protocols Guide where multicast protocols (PIM, IGMP, MLD, MSDP) are described.

Applications for this feature include enterprise customer implementing a VPRN solution for their WAN networking needs, customer applications including stock-ticker information, financial institutions for stock and other types of trading data and video delivery systems.

Implementation of multicast in IP VPNs entails the support and separation of the providers core multicast domain from the various customer multicast domains and the various customer multicast domains from each other.

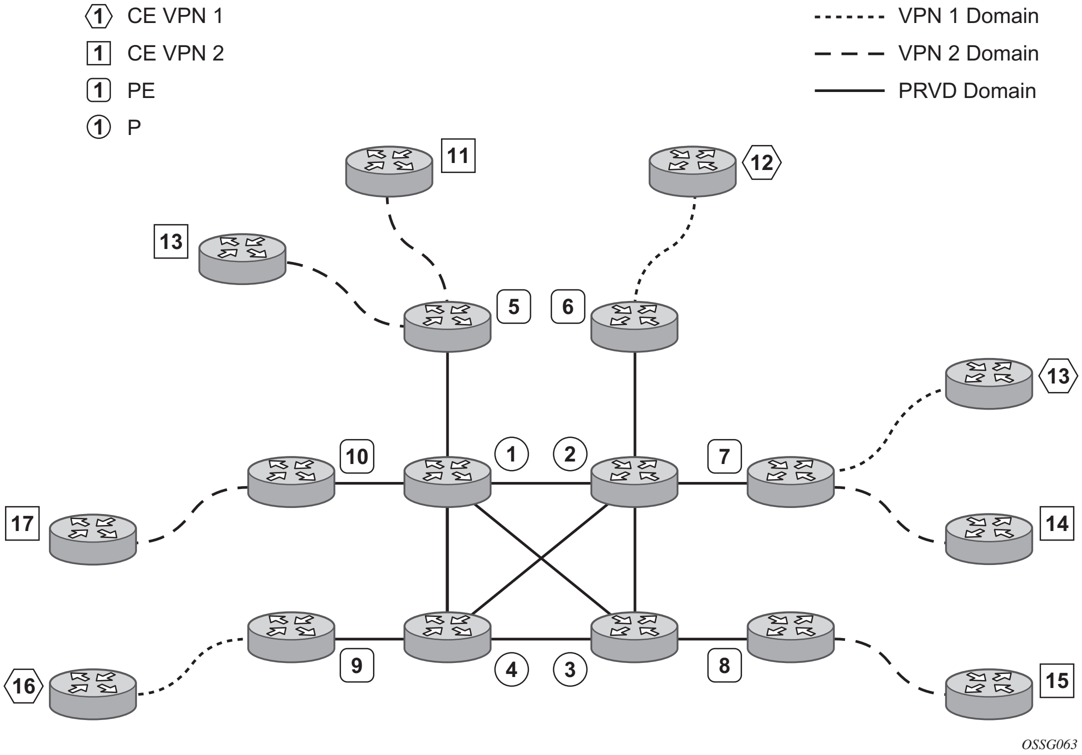


Figure 23: Multicast in IP-VPN Applications

Figure 23 depicts an example of multicast in an IP-VPN application. The provider's domain encompasses the core routers (1 through 4) and the edge routers (5 through 10). The various IP-VPN customers each have their own multicast domain, VPN-1 (CE routers 12, 13 and 16) and VPN-2 (CE Routers 11, 14, 15, 17 and 18). Multicast in this VPRN example, the VPN-1 data generated by the customer behind router 16 will be multicast only by PE 9 to PE routers 6 and 7 for delivery to CE routers 12 and 13 respectively. Data generated for VPN-2 generated by the customer behind router 15 will be forwarded by PE 8 to PE routers 5, 7 and 10 for delivery to CE routers 18, 11, 14 and 17 respectively.

The demarcation of these domains is in the PE's (routers 5 through 10). The PE router participates in both the customer multicast domain and the provider's multicast domain. The customer's CEs are limited to a multicast adjacency with the multicast instance on the PE specifically created to support that specific customer's IP-VPN. This way, customers are isolated from the provider's core multicast domain and other customer multicast domains while the provider's core routers only participate in the provider's multicast domain and are isolated from all customers' multicast domains.

The PE for a given customer's multicast domain becomes adjacent to the CE routers attached to that PE and to all other PE's that participate in the IP-VPN (or customer) multicast domain. This is achieved by the PE who encapsulates the customer multicast control data and multicast streams inside the provider's multicast packets. These encapsulated packets are forwarded only to the PE nodes that are attached to the same customer's edge routers as the originating stream and are part of the same customer VPRN. This prunes the distribution of the multicast control and data traffic to the PEs that participate in the customer's multicast domain. The Rosen draft refers to this as the default multicast domain for this multicast domain; the multicast domain is associated with a unique multicast group address within the provider's network.

Use of Data MDTs

Using the above method, all multicast data offered by a given CE is always delivered to all other CEs that are part of the same multicast. It is possible that a number of CEs do not require the delivery of a particular multicast stream because they have no downstream receivers for a specific multicast group. At low traffic volumes, the impact of this is limited. However, at high data rates this could be optimized by devising a mechanism to prune PEs from the distribution tree that although forming part of the customer multicast have no need to deliver a given multicast stream to the CE attached to them. To facilitate this optimization, the Rosen draft specifies the use of data MDTs. These data MDTs are signaled once the bandwidth for a given SG exceeds the configurable threshold.

Once a PE detects it is transmitting data for the SG in excess of this threshold, it sends an MDT join TLV (at 60 second intervals) over the default MDT to all PEs. All PEs that require the SG specified in the MDT join TLV will join the data MDT that will be used by the transmitting PE to send the given SG. PEs that do not require the SG will not join the data MDT, thus pruning the multicast distribution tree to just the PEs requiring the SG. After providing sufficient time for all PEs to join the data MDT, the transmitting PE switches the given multicast stream to the data MDT.

PEs that do not require the SG to be delivered, keep state to allow them to join the data MDT as required.

When the bandwidth requirement no longer exceeds the threshold, the PE stops announcing the MDT join TLV. At this point the PEs using the data MDT will leave this group and transmission resumes over the default MDT.

Sampling to check if an s,g has exceeded the threshold occurs every ten seconds. If the rate has exceeded the configured rate in that sample period then the data MDT is created. If during that period the transmission rate has not exceeded the configured threshold then the data MDT is not created. If the data MDT is active and the transmission rate in the last sample period has not exceeded the configured rate then the data MDT is torn down and the multicast stream resumes transmission over the default MDT.

Multicast Protocols Supported in the Provider Network

When MVPN auto-discovery is disabled, PIM-SM can be used for I-PMSI, and PIM-SSM or PIM-SM (Draft-Rosen Data MDT) can be used for S-PMSI; When MVPN S-PMSI auto-discovery is enabled, both PIM-SM and PIM SSM can be used for I-PMSI, and PIM-SSM can be used for S-PMSI. In the customer network, both PIM-SM and PIM-SSM are supported.

An MVPN is defined by two sets of sites: sender sites set and receiver sites set, with the following properties:

- Hosts within the sender sites set could originate multicast traffic for receivers in the receiver sites set.
- Receivers not in the receiver sites set should not be able to receive this traffic.
- Hosts within the receiver sites set could receive multicast traffic originated by any host in the sender sites set.
- Hosts within the receiver sites set should not be able to receive multicast traffic originated by any host that is not in the sender sites set.

A site could be both in the sender sites set and receiver sites set, which implies that hosts within such a site could both originate and receive multicast traffic. An extreme case is when the sender sites set is the same as the receiver sites set, in which case all sites could originate and receive multicast traffic from each other.

Sites within a given MVPN may be either within the same, or in different organizations, which implies that an MVPN can be either an intranet or an extranet. A given site may be in more than one MVPN, which implies that MVPNs may overlap. Not all sites of a given MVPN have to be connected to the same service provider, which implies that an MVPN can span multiple service providers.

Another way to look at MVPN is to say that an MVPN is defined by a set of administrative policies. Such policies determine both sender sites set and receiver site set. Such policies are established by MVPN customers, but implemented by MVPN service providers using the existing BGP/MPLS VPN mechanisms, such as route targets, with extensions, as necessary.

MVPN Membership Auto-discovery using BGP

BGP-based auto-discovery is performed by a multicast VPN address family. Any PE that attaches to an MVPN must issue a BGP update message containing an NLRI in this address family, along with a specific set of attributes.

The PE router uses route targets to specify MVPN route import and export. The route target may be the same as the one used for the corresponding unicast VPN, or it may be different. The PE router can specify separate import route targets for sender sites and receiver sites for a given MVPN.

The route distinguisher (RD) that is used for the corresponding unicast VPN can also be used for the MVPN.

When BGP auto-discovery is enabled, PIM peering on the I-PMSI is disabled, so no PIM hellos are sent on the I-PMSI. C-trees to P-tunnels bindings are also discovered using BGP S-PMSI AD routes, instead of PIM join TLVs. Configure PIM join TLVs when **c-mcast-signaling** is set to **pim** in the **config>service>vprn>mvpn>provider-tunnel>selective>auto-discovery-disable** context.

Table 14 and Table 15 describe the supported configuration combinations. If the CLI combination is not allowed, the system returns an error message. If the CLI command is marked as “ignored” in the table, the configuration is not blocked, but its value is ignored by the software.

Table 14: Supported Configuration Combinations

Auto-Discovery	Inclusive PIM SSM	Action
Yes	Yes	Allowed
No	Yes	Not Allowed
Yes or No	No	Allowed

Table 15: Supported Configuration Combinations

Auto-Discovery	C-Mcast-Signaling	s-PMSI A-D	Action
Yes	BGP	Ignored	Allowed
Yes	PIM	Yes	Allowed
Yes	PIM	No	Allowed
No	BGP	Ignored	Not Allowed
No	PIM	Ignored	Allowed

For example, if auto-discovery is disabled, the **c-mcast-signaling bgp** command will fail with an error message stating:

C-multicast signaling in BGP requires auto-discovery to be enabled

If **c-mcast-signaling** is set to **bgp** then **no auto-discovery** will fail with an error message stating

C-multicast signaling in BGP requires auto-discovery to be enabled

When **c-mcast-signaling** is set to **bgp**, S-PMSI A-D is always enabled (its configuration is ignored);

When **auto-discovery** is disabled, S-PMSI A-D is always disabled (its configuration is ignored).

When auto-discovery is enabled and **c-multicast-signaling** is set to **pim**, S-PMSI A-D configuration value is used.

MVPN (Rosen) Membership Auto-Discovery using BGP MDT-SAFI

MVPN implementation based on the draft *-Rosen* can support membership auto discovery using BGP MDT-SAFI. A CLI option is provided per MVPN instance to enable auto discovery either using BGP MDT-SAFI or NG-MVPN. Only PIM-MDT is supported with BGP MDT-SAFI method.

PE-PE Transmission of C-Multicast Routing using BGP

MVPN c-multicast routing information is exchanged between PEs by using c-multicast routes that are carried using MCAST-VPN NLRI.

VRF Route Import Extended Community

VRF route import is an IP address-specific extended community, of an extended type, and is transitive across AS boundaries (RFC 4360, *BGP Extended Communities Attribute*.

To support MVPN, in addition to the import/export route target extended communities used by the unicast routing, each VRF on a PE must have an import route target extended community that controls imports of C-multicast routes into a particular VRF.

The c-multicast import RT uniquely identifies a VRF, and is constructed as follows:

- The Global Administrator field of the c-multicast import RT must be set to an IP address of the PE. This address should be common for all the VRFs on the PE (this address may be the PE's loopback address).
- The Local Administrator field of the c-multicast import RT associated with a given VRF contains a 2 octets long number that uniquely identifies that VRF within the PE that contains the VRF.

A PE that has sites of a given MVPN connected to it communicates the value of the c-multicast import RT associated with the VRF of that MVPN on the PE to all other PEs that have sites of that MVPN. To accomplish this, a PE that originates a (unicast) route to VPN-IP addresses includes in the BGP updates message that carries this route the VRF route import extended community that has the value of the c-multicast import RT of the VRF associated with the route, except if it is known a priori that none of these addresses will act as multicast sources and/or RP, in which case the (unicast) route need not carry the VRF Route Import extended community.

All c-multicast routes with the c-multicast import RT specific to the VRF must be accepted. In this release, vrf-import and vrftraget policies don't apply to C-multicast routes.

The decision flow path is shown below.

```
if (route-type == c-mcast-route)
  if (route_target_list includes C-multicast_Import_RT){
    else
      drop;
  else
  Run vrf_import and/or vrf-target;
```

Provider Tunnel Support

Point-to-Multipoint Inclusive (I-PMSI) and Selective (S-PMSI) Provider Multicast Service Interface

BGP C-multicast signaling must be enabled for an MVPN instance to use P2MP RSVP-TE or LDP as I-PMSI (equivalent to ‘Default MDT’, as defined in draft Rosen MVPN) and S-PMSI (equivalent to ‘Data MDT’, as defined in draft Rosen MVPN).

By default, all PE nodes participating in an MVPN receive data traffic over I-PMSI. Optionally, (C-*, C-*) wildcard S-PMSI can be used instead of I-PMSI. For more details, see Wildcard (C-*, C-*) P2MP LSP S-PMSI. For efficient data traffic distribution, one or more S-PMSIs can be used, in addition to the default PMSI, to send traffic to PE nodes that have at least one active receiver connected to them. For more details, see P2MP LSP S-PMSI.

Only one unique multicast flow is supported over each P2MP RSVP-TE or P2MP LDP LSP S-PMSI. Number of S-PMSI that can be initiated per MVPN instance is restricted by CLI command **maximum-p2mp-spmsi**. P2MP LSP S-PMSI cannot be used for more than one (S,G) stream (that is, multiple multicast flow) as number of S-PMSI per MVPN limit is reached. Multicast flows that cannot switch to S-PMSI remain on I-PMSI.

P2MP RSVP-TE I-PMSI and S-PMSI

Point-to-Multipoint RSVP-TE LSP as inclusive or selective provider tunnel is available with BGP NG-MVPN only. P2MP RSVP-TE LSP is dynamically setup from root node on auto discovery of leaf PE nodes that are participating in multicast VPN. Each RSVP-TE I-PMSI or S-PMSI LSP can be used with a single MVPN instance only.

RSVP-TE LSP template must be defined (see MPLS user guide) and bound to MVPN as inclusive or selective (S-PMSI is for efficient data distribution and is optional) provider tunnel to dynamically initiate P2MP LSP to the leaf PE nodes learned via NG-MVPN auto-discovery signaling. Each P2MP LSP S2L is signaled based on parameters defined in LSP template.

P2MP LDP I-PMSI and S-PMSI

Point-to-Multipoint LDP LSP as inclusive or selective provider tunnel is available with BGP NG-MVPN only. P2MP LDP LSP is dynamically setup from leaf nodes on auto discovery of leaf node PE nodes that are participating in multicast VPN. Each LDP I-PMSI or S-PMSI LSP can be used with a single MVPN instance only.

The **multicast-traffic** CLI command must be configured per LDP interface to enable P2MP LDP setup. P2MP LDP must also be configured as inclusive or selective (S-PMSI is for efficient data distribution and is optional) provider tunnel per MVPN to dynamically initiate P2MP LSP to leaf PE nodes learned via NG-MVPN auto-discovery signaling.

Wildcard (C-*, C-*) P2MP LSP S-PMSI

Wildcard S-PMSI allows usage of selective tunnel as a default tunnel for a given MVPN. By using wildcard S-PMSI, operators can avoid full mesh of LSPs between MVPN PEs, reducing related signaling, state, and BW consumption for multicast distribution (no traffic is sent to PEs without any receivers active on the default PMSI).

The SR OS allows an operator to configure wildcard S-PMSI for ng-MVPN (**config>service>vprn>mvpn>pt>inclusive>wildcard-spmsi**), using LDP and RSVP-TE in P-instance. Support includes:

- IPv4 and IPv6
- PIM ASM and SSM
- Directly attached receivers

The SR OS (C-*, C-*) wildcard implementation uses wildcard S-PMSI instead of I-PMSI for a given MVPN. To switch MVPN from I-PMSI to (C-*, C-*) S-PMSI a VPRN shutdown is required. ISSU and UMH redundancy can be used to minimize the impact.

To minimize outage, the following upgrade order is recommended:

1. Route Reflector
2. Receiver PEs
3. backup UMH
4. active UMH

RSVP-TE/mLDP configuration under inclusive provider tunnel (**config>service>vprn>mvpn>pt>inclusive**) apply to wildcard S-PMSI when enabled.

Wildcard C-S and C-G values are encoded as defined in RFC6625: using zero for Multicast Source Length and Multicast Group Length and omitting Multicast Source and Multicast Group values respectively in MCAST_VPN_NLRI. For example, a (C-*, C-*) will be advertised as: RD, 0x00, 0x00, and originating router's IP address.

OPERATIONAL NOTE: All SR OS routers with BGP peering session to the PE with RFC6625 support enabled must be upgraded to SR OS release 13.0 before the feature is enabled. Failure to do so will result in the following processing on a router with BGP peering session to an RFC6625-enabled PE:

- BGP peer running Release 12.0 version R4 or newer will accept 0-length address and it will keep encoding length 4 with all zeros for the address
- BGP peer running Release 12 version R3 or older will not accept 0-length address and will keep restarting BGP session

The procedures implemented by SR OS are compliant to section 3 and 4 of RFC, 6625. Wildcards encoded as described above are carried in NLRI field of MP_REACH_NLRF_ATTRIBUTE. Both IPv4 and IPv6 are supported: (AFI) of 1 or 2 and a Subsequent AFI (SAFI) of MCAST-VPN.

The (C-*, C-*) S-PMSI is established as follows:

- UMH PEs advertise I-PMSI A-D routes without tunnel information present (empty PTA) - encoded as per RFC6513/6514 prior to advertising wildcard S-PMSI. I-PMSI needs to be signaled and installed on receiver PEs, because (C-*, C-*) S-PMSI is only installed when a first receiver is added. However, no LSP is established for I-PMSI).
- UMH PEs advertise S-PMSI A-D route whose NLRI contains (C-*, C-*) with tunnel information encoded as per RFC 6625
- Receiver PEs join wildcard S-PMSI if there are any receivers present.

OPERATIONAL NOTE: If UMH PE does not encode I-PMSI/S-PMSI A-D routes as per the above, or advertises both I-PMSI and wildcard S-PMSI with the tunnel information present, no interoperability can be achieved.

To ensure proper operation of BSR between PEs with (C-*, C-*) S-PMSI signaling, SROS implements two modes of operations for BSR.

By default (bsr unicast):

- BSR PDUs are sent/forwarded as unicast PDUs to neighbor PEs when I-PMSI with Pseudo-tunnel interface is installed.
- At every BSR interval timer BSR Unicast PDU are sent to all IPMSI interfaces when this is an elected BSR.
- BSMs received as multicast from C-instance interfaces are flooded as unicast in the P-instance.
- All PEs process BSR PDU's received on I-PMSI Pseudo-tunnel interface as unicast packets.
- BSR PDU's are not forwarded to PE's management control interface.
- BSR unicast PDU's use PE's System IP address as destination IP and sender PE's System address as Source IP.
- The BSR unicast functionality ensures that no special state needs to be created for BSR when (C-*, C-*) S-PMSI is enabled, which is beneficiary considering low volume of BSR traffic.

OPERATIONAL NOTES:

- For bsr unicast, the base IPv4 system address (IPv4) or the mapped version of the base IPv4 system address (IPv6) must be configured under the VPRN to ensure bsr unicast messages can reach the VPRN
- For bsr spmsi, the base IPv4/IPv6 system address must be configured under the VPRN to ensure B-SR S-PMSI's are established

BSR S-PMSI mode can be enabled to allow interoperability with other vendors. In that mode full mesh S-PMSI is required and created between all PEs in MVPN to exchange BSR PDUs between all PEs in MVPN. To operate as expected, the BSR S-PMSI mode requires a selective P-tunnel configuration. For IPv6 support (including dual-stack) of BSR S-PMSI mode, the IPv6 default system interface address must be configured as a loopback interface address under the VPRN and VPRN PIM contexts. Changing BSR signaling requires a VPRN shutdown.

Other key feature interactions and caveats for (C-*, C-*) include the following:

- Extranet is fully supported with wildcard S-PMSI trees.
- (C-S, C-G) S-PMSIs are supported when (C-*, C-*) S-PMSI is configured (including both BW and receiver PE driven thresholds).
- Geo-redundancy is supported (deploying with geo-redundancy eliminates traffic duplication when geo-redundant source has no active receivers at a cost of slightly increased outage upon a switch since wildcard S-PMSI may need to be re-establish).
- PIM in P-instance is not supported.
- SR OS implementation requires wildcard encoding as per RFC6625 and I-PMSI/S-PMSI signaling as defined above (I-PMSI signaled with empty PTA then S-PMSI signaled with P-tunnel PTA) for interoperability. Implementations that do not adhere to RFC6625 encoding, or signal both I-PMSI and S-PMSI with P-tunnel PTA will not inter-operate with SR OS implementation).

P2MP LSP S-PMSI

NG-MVPN support P2MP RSVP-TE and P2MP LDP LSPs as selective provider multicast service interface (S-PMSI). S-PMSI is used to avoid sending traffic to PEs that participate in multicast VPN, but do not have any receivers for a given C-multicast flow. This allows more-BW efficient distribution of multicast traffic over the provider network, especially for high bandwidth multicast flows. S-PMSI is spawned dynamically based on configured triggers as described in S-PMSI trigger thresholds section.

In MVPN, the head-end PE firstly discovers all the leaf PEs via I-PMSI A-D routes. It then signals the P2MP LSP to all the leaf PEs using RSVP-TE. In the scenario of S-PMSI:

1. The head-end PE sends an S-PMSI A-D route for a specific C-flow with the “Leaf Information Required” bit set.
2. The PEs who are interested in the C-flow respond with Leaf A-D routes.
3. The head-end PE then signals the P2MP LSP to all the leaf PEs using RSVP-TE.

Also, because the receivers may come and go, the implementation supports dynamically adding and pruning leaf nodes to and from the P2MP LSP.

When the tunnel type in the PMSI attribute is set to RSVP-TE P2MP LSP, the tunnel identifier is <Extended Tunnel ID, Reserved, Tunnel ID, P2MP ID>, as carried in the RSVP-TE P2MP LSP SESSION Object.

The PE can also learn via an A-D route that it needs to receive traffic on a particular RSVP-TE P2MP LSP before the LSP is actually setup. In this case, the PE needs to wait until the LSP is operational before it can modify its forwarding tables as directed by the A-D route.

Because of the way that LDP normally works, mLDP P2MP LSPs are setup without solicitation from the leaf PEs towards the head-end PE. The leaf PE discovers the head-end PE via I-PMSI or S-PMSI A-D routes. The tunnel identifier carried in the PMSI attribute is used as the P2MP FEC element. The tunnel identifier consists of the head-end PE’s address, along with a Generic LSP identifier value. The Generic LSP identifier value is automatically generated by the head-end PE.

Dynamic Multicast Signaling over P2MP LDP in VRF

This feature provides a multicast signaling solution for IP-VPNs, allowing the connection of IP multicast sources and receivers in C-instances, which are running PIM multicast protocol using Rosen MVPN with BGP SAFI and P2MP mLDP in P-instance. The solution dynamically maps each PIM multicast flow to a P2MP LDP LSP on the source and receiver PEs.

The feature uses procedures defined in RFC 7246: *Multipoint Label Distribution Protocol In-Band Signaling in Virtual Routing and Forwarding (VRF) Table Context*. On the receiver PE, PIM signaling is dynamically mapped to the P2MP LDP tree setup. On the source PE, signaling is handed back from the P2MP mLDP to the 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 VRF. Per (C-S, C-G) IP multicast state is also removed from the network, since P2MP LSPs are used to transport multicast flows.

Figure 24 illustrates dynamic mLDP signaling for IP multicast in VPRN.

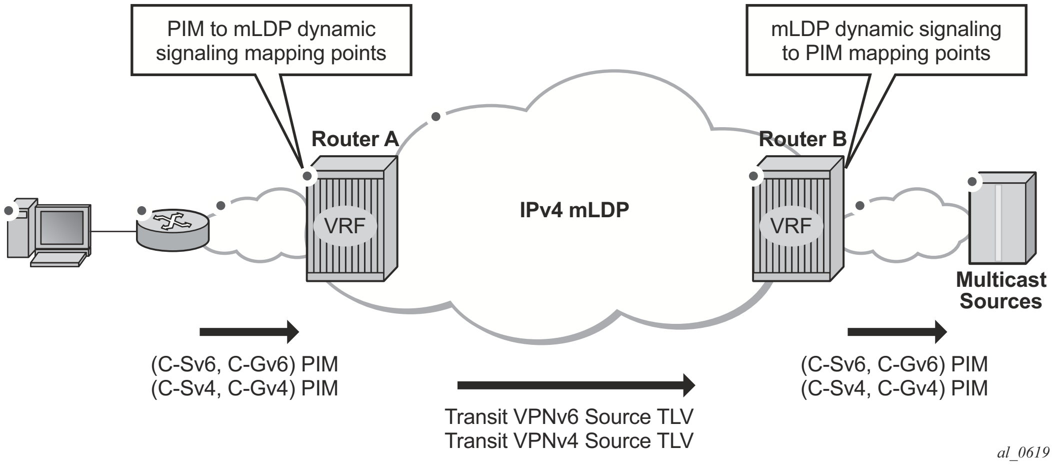


Figure 24: Dynamic mLDP Signaling for IP Multicast in VPRN

As illustrated in Figure 24, P2MP LDP LSP signaling is initiated from the receiver PE that receives PIM JOIN from a downstream router (Router A). To enable dynamic multicast signaling, the **p2mp-ldp-tree-join** must be configured on PIM customer-facing interfaces for the given VPRN of Router A. This enables handover of multicast tree signaling from the PIM to the P2MP LDP LSP. Being a leaf node of the P2MP LDP LSP, Router A selects the upstream-hop as the root node of P2MP LDP FEC, based on a 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 the Transit Source PE (Router B), and multicast tree signaling is handed back to the PIM and propagated upstream as native-IP PIM JOIN toward C-instance multicast source.

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.

The following are feature caveats:

- Dynamic mLDP signaling in a VPRN instance is mutually exclusive with Rosen or NG-MVPN.
- A single instance of P2MP LDP LSP is supported between the receiver PE and Source PE 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.
- When dynamic mLDP signaling is deployed, a change in Route Distinguisher (RD) on the Source PE is not acted upon for any (C-S, C-G)s until the receiver PEs learn about the new RD (via BGP) and send explicit delete and create with the new RD.
- Procedures of Section 2 of RFC 7246 for a case where UMH and the upstream PE do not have the same IP address are not supported.
- The feature requires chassis mode C.

MVPN Sender-only/Receiver-only

In multicast MVPN, by default, if multiple PE nodes form a peering with a common MVPN instance then each PE node originates a multicast tree locally towards the remaining PE nodes that are member of this MVPN instance. This behavior creates a mesh of I-PMSI across all PE nodes in the MVPN. It is often a case, that a given VPN has many sites that will host multicast receivers, but only few sites that either host both receivers and sources or sources only.

MVPN Sender-only/Receiver-only allows to optimize control and data plane resources by preventing unnecessary I-PMSI mesh when a given PE hosts multicast sources only or multicast receivers only for a given MVPN.

For PE nodes that host only multicast sources for a given VPN, operators can now block those PEs, through configuration, from joining I-PMSIs from other PEs in this MVPN. For PE nodes that host only multicast receivers for a given VPN, operators can now block those PEs, through configuration, to set-up a local I-PMSI to other PEs in this MVPN.

MVPN Sender-only/Receiver-only is supported with ng-mVPN using IPv4 RSVP-TE or IPv4 LDP provider tunnels for both IPv4 and IPv6 customer multicast. Figure 25 depicts 4-site MVPN with sender-only, receiver-only and sender-receiver (default) sites.

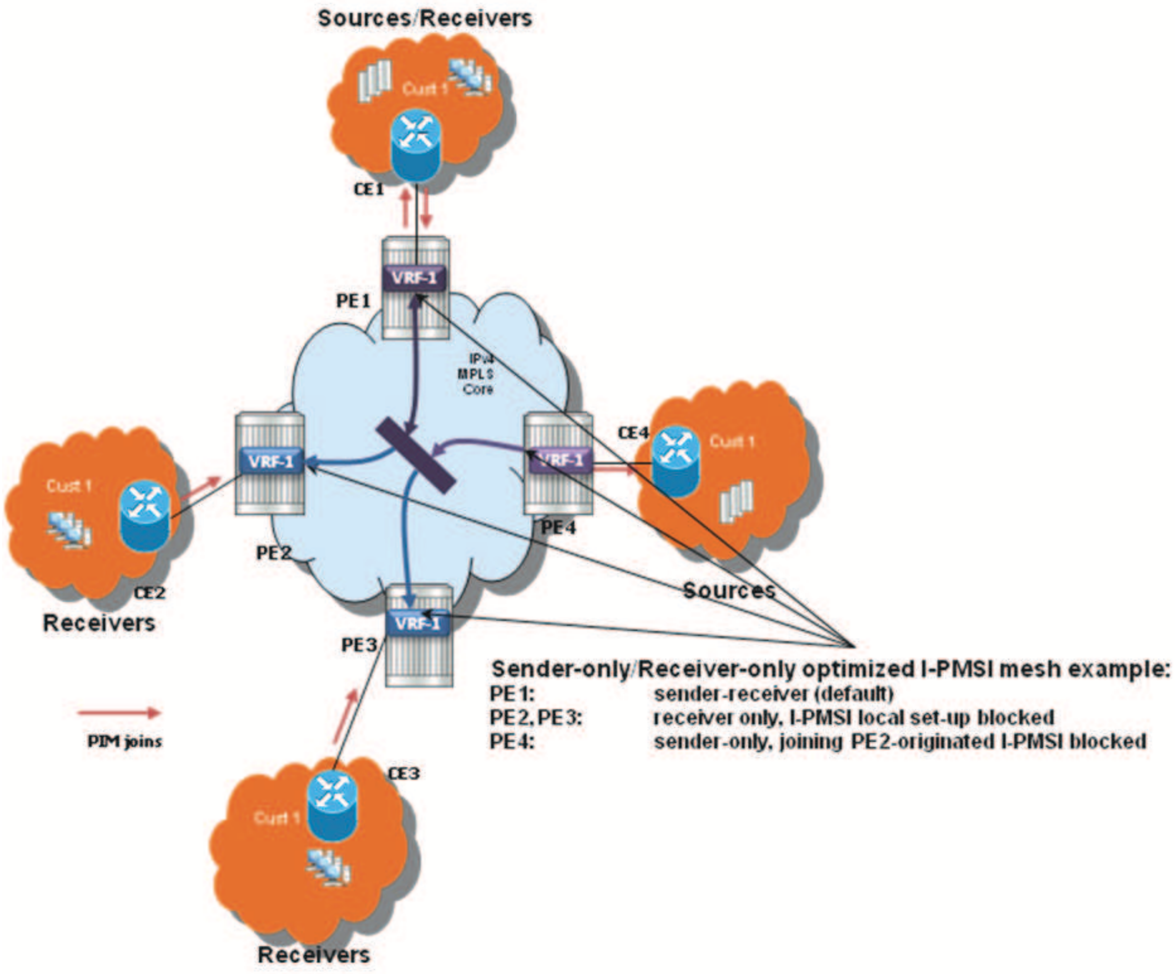

Sources/Receivers
Cust 1
CE1
VRF-1
PE1
IPv4
MPLS
Core
CE4
Cust 1
VRF-1
PE4
Sources
Cust 1
VRF-1
PE2
CE2
Receivers
VRF-1
PE3
Sender-only/Receiver-only optimized I-PMSI mesh example:
PE1: sender-receiver (default)
PE2, PE3: receiver only, I-PMSI local set-up blocked
PE4: sender-only, joining PE2-originated I-PMSI blocked
PIM joins
CE3
Cust 1
Receivers

Figure 25: MVPN Sender-only/Receiver-only Example

Extra attention needs to be paid to BSR/RP placement when Sender-only/Receiver-only is enabled. Source DR sends unicast encapsulated traffic towards RP, therefore, RP shall be at sender-receiver or sender-only site, so that *G traffic can be sent over the tunnel. BSR shall be deployed at the sender-receiver site. BSR can be at sender-only site if the RPs are at the same site. BSR needs to receive packets from other candidate-BSR and candidate-RPs and also needs to send BSM packets to everyone.

S-PMSI Trigger Thresholds

The mLDP and RSVP-TE S-PMSIs support two types of data thresholds: bandwidth-driven and receiver-PE-driven. The threshold evaluation and bandwidth driven threshold functionality are described in Use of Data MDTs.

In addition to the bandwidth threshold functionality, an operator can enable receiver-PE-driven threshold behavior. Receiver PE thresholds ensure that S-PMSI is only created when BW savings in P-instance justify extra signaling required to establish a new S-PMSI. For example, the number of receiver PEs interested in a given C-multicast flow is meaningfully smaller than the number of receiver PEs for default PMSI (I-PMSI or wildcard S-PSMI). To ensure that S-PMSI is not constantly created/deleted, two thresholds need to be specified: receiver PE add threshold and receiver PE delete threshold (expected to be significantly higher).

When a (C-S, C-G) crosses a data threshold to create S-PMSI, instead of regular S-PMSI signaling, sender PE originates S-PMSI explicit tracking procedures to detect how many receiver PEs are interested in a given (C-S, C-G). When receiver PEs receive an explicit tracking request, each receiver PE responds, indicating whether there are multicast receivers present for that (C-S, C-G) on the given PE (PE is interested in a given (C-S, C-G)). Note that if the geo-redundancy feature is enabled, receiver PEs do not respond to explicit tracking requests for suppressed sources and therefore only Receiver PEs with an active join are counted against the configured thresholds on Source PEs.

Upon regular sampling and check interval, if the previous check interval had a non-zero receiver PE count (one interval delay to not trigger S-PMSI prematurely) and current count of receiver PEs interested in the given (C-S, C-G) is non-zero and is less than the configured receiver PE add threshold, Source PE will set-up S-PMSI for this (C-S, C-G) following standard ng-MVPN procedures augmented with explicit tracking for S-PMSI being established.

Note that data threshold timer should be set to ensure enough time is given for explicit tracking to complete (for example, setting the timer to value that is too low may create S-PMSI prematurely).

Upon regular data-delay-interval expiry processing, when BW threshold validity is being checked, a current receiver PE count is also checked (for example, explicit tracking continues on the established S-PMSI). If BW threshold no longer applies or the receiver PEs exceed receiver PE delete threshold, the S-PMSI is torn down and (C-S, C-G) joins back the default PMSI.

Changing of thresholds (including enabling disabling the thresholds) is allowed in service. The configuration change is evaluated at the next periodic threshold evaluation.

The explicit tracking procedures follow RFC6513/6514 with clarification and wildcard S-PMSI explicit tracking extensions as described in IETF Draft: draft-dolganow-l3vpn-expl-track-00.

Migration from Existing Rosen Implementation

The existing Rosen implementation is compatible to provide an easy migration path.

The following migration procedure are supported:

- Upgrade all the PE nodes that need to support MVPN to the newer release.
- The old configuration will be converted automatically to the new style.

- Node by node, MCAST-VPN address-family for BGP is enabled. Enable auto-discovery using BGP.
- Change PE-to-PE signaling to BGP.

MVPN (NG-MVPN) Upstream Multicast Hop Fast Failover

MVPN upstream PE or P node fast failover detection method is supported with RSVP P2MP I-PMSI only. A receiver PE achieves fast upstream failover based on the capability to subscribe multicast flow from multiple UMH nodes and the capability to monitor the health of the upstream PE and intermediate P nodes using an unidirectional multi-point BFD session running over the provider tunnel.

A receiver PE subscribes multicast flow from multiple upstream PE nodes to have active redundant multicast flow available during failure of primary flow. Active redundant multicast flow from standby upstream PE allows instant switchover of multicast flow during failure of primary multicast flow.

Faster detection of multicast flow failure is achieved by keeping track of unidirectional multi-point BFD sessions enabled on the provider tunnel. Multi-point BFD sessions must be configured with 10 ms transmit interval on sender (root) PE to achieve sub-50ms fast failover on receiver (leaf) PE.

UMH **tunnel-status** selection option must be enabled on the receiver PE for upstream fast failover. Primary and standby upstream PE pairs must be configured on the receiver PE to enable active redundant multicast flow to be received from the standby upstream PE.

Multicast VPN Extranet

Multicast VPN extranet distribution allows multicast traffic to flow across different routing instances. A routing instance that received a PIM/IGMP JOIN but cannot reach source of multicast source directly within its own instance is selected as receiver routing instance (receiver C-instance). A routing instance that has source of multicast stream and accepts PIM/IGMP JOIN from other routing instances is selected as source routing instance (source C-instance). A routing instance that does not have either source or receivers but is used in the core is selected as a transit instance (transit P-instance). The following subsections detail supported functionality.

Multicast Extranet for Rosen MVPN for PIM SSM

Multicast extranet is supported for Rosen MVPN with MDT SAFI. Extranet is supported for IPv4 multicast stream for default and data MDTs (PIM and IGMP). The following extranet cases are supported:

- Local replication into a receiver VRF from a source VRF on a source PE.
- Transit replication from a source VRF onto a tunnel of a transit core VRF on a source PE. A source VRF can replicate its streams into multiple core VRFs as long as any given stream from source VRF uses a single core VRF (the first tunnel in any core VRF on which a join for the stream arrives). Streams with overlapping group addresses (same group address, different source address) are supported in the same core VRF.
- Remote replication from source/transit VRF into one or more receiver VRFs on receiver PEs.
- Multiple replications from multiple source/transit VRFs into a receiver VRF on receiver PEs.

Rosen MVPN extranet requires routing information exchange between the source VRF and the receiver VRF based on route export/import policies:

- Routing information for multicast sources must be exported using an RT export policy from the source VRF instance.
- Routing information must be imported into the receiver/transit VRF instance using an RT import policy.

Caveats:

- Multicast extranet functionality requires all network ports to be on FP2 or newer line cards.
- The source VRF instance and receiver VRF instance of extranet must exist on a common PE node (to allow local extranet mapping).

- SSM translation is required for IGMP (C-*, C-G).

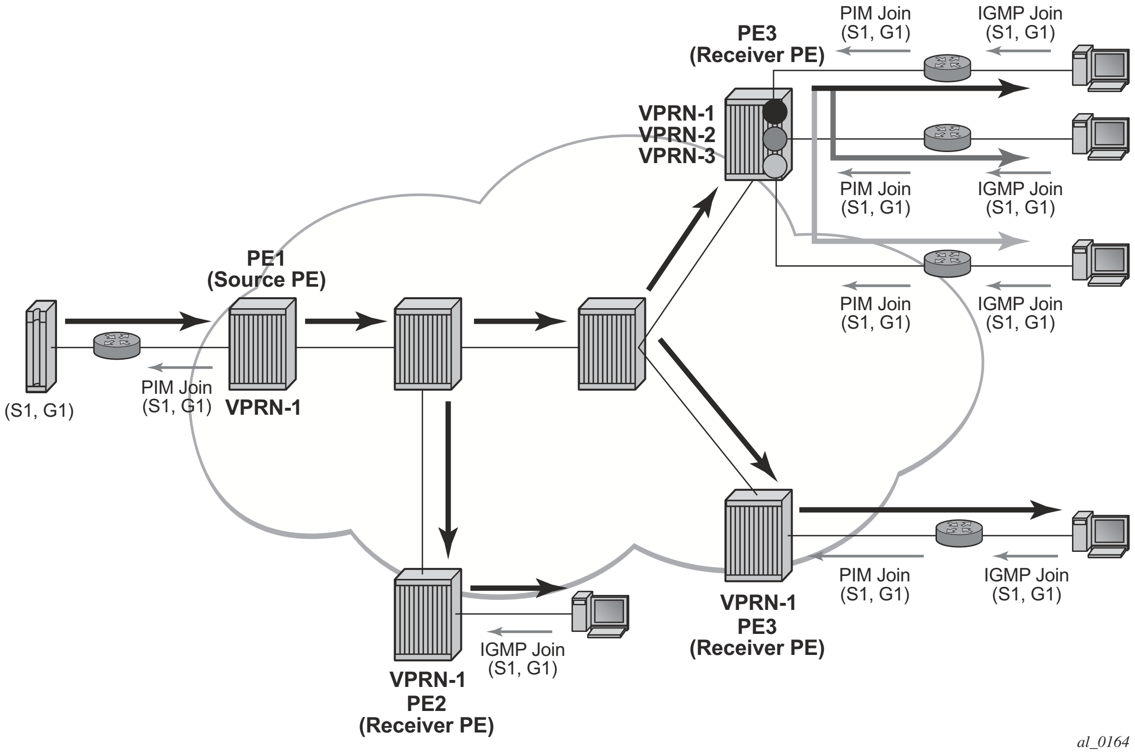


Figure 26: Multicast VPN Traffic Flow

In Figure 26, VPRN-1 is the source VPRN instance and VPRN-2 and VPRN-3 are receiver VPRN instances. The PIM/IGMP JOIN received on VPRN-2 or VPRN-3 is for (S1, G1) multicast flow. Source S1 belongs to VPRN-1. Due to the route export policy in VPRN-1 and the import policy in VPRN-2 and VPRN-3, the receiver host in VPRN-2 or VPRN-3 can subscribe to the stream (S1, G1).

Multicast Extranet for NG-MVPN for PIM SSM

Multicast extranet is supported for ng-MVPN with IPv4 RSVP-TE and mLDP I-PMSIs and S-PMSIs including (C-*, C-*) S-PMSI support where applicable. Extranet is supported for IPv4 C-multicast traffic (PIM/IGMP joins).

The following extranet cases are supported:

- Local replication into a receiver C-instance MVPN(s) on a source PE from a source P-instance MVPN.
- Remote replication from P-instance MVPN into one or more receiver C-instance MVPNs on receiver PEs.

- Multiple replications from multiple source/transit P-instance MVPNs into a receiver C-instance MVPN on receiver PEs.
- Transit replication on Source PE is not supported.

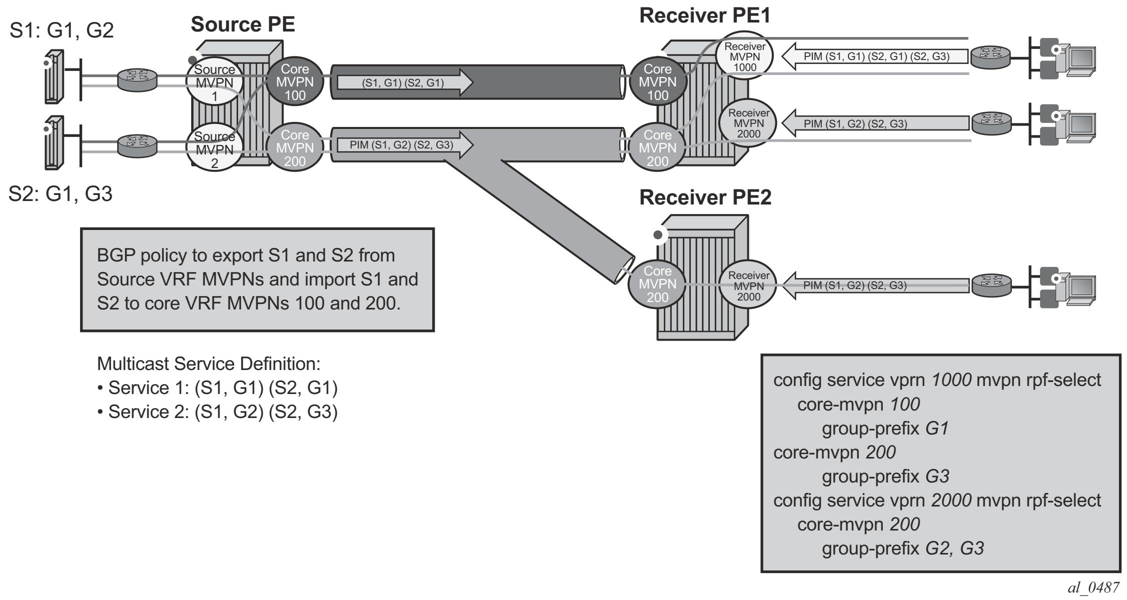
Multicast extranet for ng-MVPN, similarly to extranet for Rosen MVPN, requires routing information exchange between source ng-MVPN and receiver ng-MVPN based on route export and import policies. Routing information for multicast sources is exported using an RT export policy from a source ng-MVPN instance and imported into a receiver ng-MVPN instance using an RT import policy. S-PMSI/I-PMSI establishment and C-multicast route exchange occurs in a source ng-MVPN P-instance only (import and export policies are not used for MVPN route exchange). Note that sender-only functionality must not be enabled for the source/transit ng-MVPN on the receiver PE. It is recommended to enable receiver-only functionality on a receiver ng-MVPN instance.

Caveats:

- Multicast extranet functionality requires all network ports to be on FP2 or newer line cards.
- Source P-instance MVPN and receiver C-instance MVPN must reside on the receiver PE (to allow local extranet mapping).
- SSM translation is required for IGMP (C-*, C-G).

Multicast Extranet with Per-group Mapping for PIM SSM

In some deployments, such as IPTV or wholesale multicast services, it may be desirable to create one or more transit MVPNs to optimize delivery of multicast streams in the provider core. Figure 27 represents a sample deployment model.



BGP policy to export S1 and S2 from Source VRF MVPNs and import S1 and S2 to core VRF MVPNs 100 and 200.

Multicast Service Definition:
- Service 1: (S1, G1) (S2, G1)
- Service 2: (S1, G2) (S2, G3)

```
config service vprn 1000 mvpn rpf-select
    core-mvpn 100
        group-prefix G1
core-mvpn 200
        group-prefix G3
config service vprn 2000 mvpn rpf-select
    core-mvpn 200
        group-prefix G2, G3
```

Figure 27: Source PE Transit Replication and Receiver PE Per-group SSM Extranet Mapping

The architecture displayed in Figure 27 requires a source routing instance MVPN to place its multicast streams into one or more transit core routing instance MVPNs (each stream mapping to a single transit core instance only). It also requires receivers within each receiver routing instance MVPN to know which transit core routing instance MVPN they need to join for each of the multicast streams. To achieve this functionality, transit replication from a source routing instance MVPN onto a tunnel of a transit core routing instance MVPN on a source PE (see earlier sub-sections for MVPN topologies supporting transit replication on source PEs) and per-group mapping of multicast groups from receiver routing instance MVPNs to transit core routing instance MVPNs (as defined below) are required.

For per-group mapping on a receiver PE, the operator must configure a receiver routing instance MVPN per-group mapping to one or more source/transit core routing instance MVPNs. The mapping allows propagation of PIM joins received in the receiver routing instance MVPN into the core routing MVPN instance defined by the map. All multicast streams sourced from a single multicast source address are always mapped to a single core routing instance MVPN for a given receiver routing instance MVPN (multiple receiver MVPNs can use different core MVPNs for groups from the same multicast source address). If per-group map in receiver MVPN maps multicast streams sourced from the same multicast source address to multiple core routing instance MVPNs, then the first PIM join processed for those streams selects the core routing instance MVPN to be used for all multicast streams from a given source address for this receiver MVPN. PIM joins for streams sourced from the source address not carried by the selected core VRF MVPN instance will remain unresolved. When a PIM join/prune is received in a receiver routing instance MVPN with per-group mapping configured, if no mapping is defined for the PIM join's group address, non-extranet processing applies when resolving how to forward the PIM join/prune.

The main attributes for per-group SSM extranet mapping on receiver PE include support for:

- Rosen MVPN with MDT SAFI. * RFC6513/6514 NG-MVPN with IPv4 RSVP-TE/mLDP in P-instance (a P-instance tunnel must be in the same VPRN service as multicast source)
- IPv4 PIM SSM.
- IGMP (C-S, C-G), and for IGMP (C-*, C-G) using SSM translation.
- A receiver routing instance MVPN to map groups to multiple core routing instance MVPNs.
- In-service changes of the map to a different transit/source core routing instance (this is service affecting).

Caveats:

- When a receiver routing instance MVPN is on the same PE as a source routing instance MVPN, basic extranet functionality and not per-group (C-S, C-G) mapping must be configured (extranet from receiver routing instance to core routing instance to source routing instance on a single PE is not supported).
- Local receivers in the core routing instance MVPN are not supported when per-group mapping is deployed.
- Receiver routing instance MVPN that has per-group mapping enabled cannot have tunnels in its OIF lists.
- Per-group mapping is blocked if GRT/VRF extranet is configured.
- Per-group mapping requires network interfaces to be on FP2-based or newer line cards.

Multicast GRT-source/VRF-receiver Extranet with Per Group Mapping for PIM SSM

Multicast GRT-source/VRF-receiver (GRT/VRF) extranet with per-group mapping allows multicast traffic to flow from GRT into VRF MVPN instances. A VRF routing instance that received a PIM/IGMP join but cannot reach the source multicast stream directly within its own instance is selected as receiver routing instance. A GRT instance that has sources of multicast streams and accepts PIM joins from other VRF MVPN instances is selected as source routing instance.

Figure 28 depicts a sample deployment.

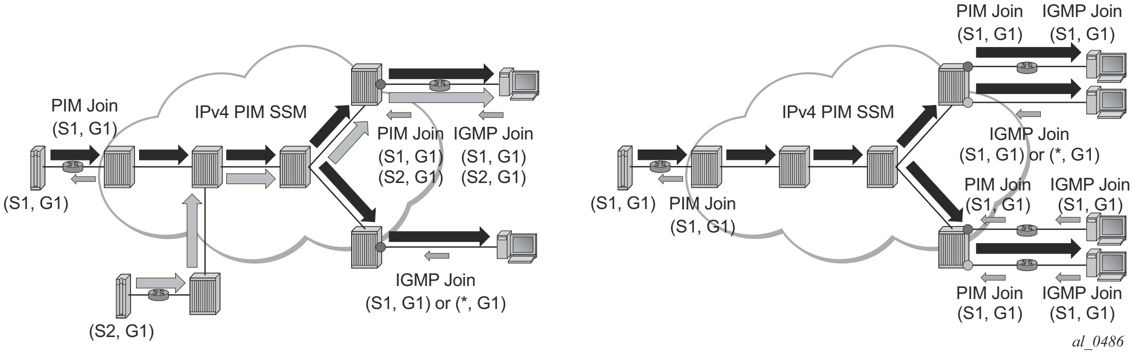


Figure 28: GRT/VRF Extranet

Routing information is exchanged between GRT and VRF receiver MVPN instances of extranet by enabling grt-extranet under a receiver MVPN PIM configuration for all or a subset of multicast groups. When enabled, multicast receivers in a receiver routing instance(s) can subscribe to streams from any multicast source node reachable in GRT source instance.

The main feature attributes are:

- GRT/VRF extranet can be performed on all streams or on a configured group of prefixes within a receiver routing instance.
- GRT instance requires Classic Rosen multicast.
- IPv4 PIM joins are supported in receiver VRF instances.
- Local receivers using IGMP: (C-S, C-G) and (C-*, C-G) using SSM translation are supported.
- The feature is blocked if a per-group mapping extranet is configured in receiver VRF.
- The feature requires FP2-based line cards or newer for all network ports.

Multicast Extranet with Per-group Mapping for PIM ASM

Multicast extranet with per-group mapping for PIM ASM allows multicast traffic to flow from a source routing instance to a receiver routing instance when a PIM ASM join is received in the receiver routing instance.

Figure 29 depicts PIM ASM extranet map support.

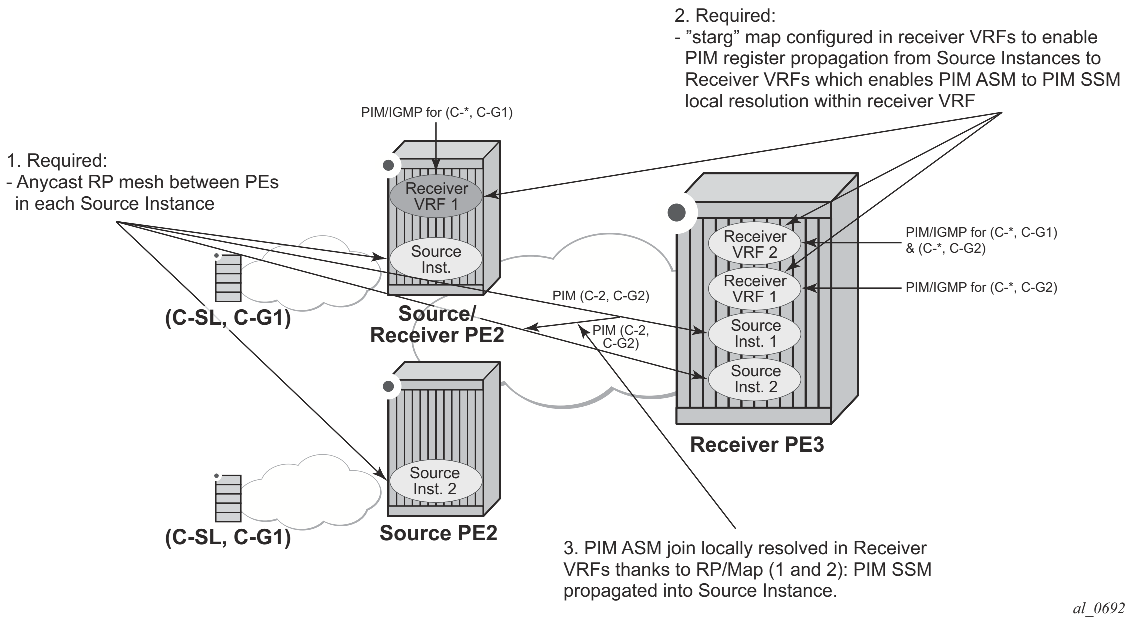

2. Required:
- "starg" map configured in receiver VRFs to enable PIM register propagation from Source Instances to Receiver VRFs which enables PIM ASM to PIM SSM local resolution within receiver VRF
PIM/IGMP for (C-*, C-G1)
1. Required:
- Anycast RP mesh between PEs in each Source Instance
Receiver VRF 1
Source Inst.
Receiver VRF 2
Receiver VRF 1
Source Inst. 1
Source Inst. 2
PIM/IGMP for (C-*, C-G1) & (C-*, C-G2)
PIM/IGMP for (C-*, C-G2)
PIM (C-2, C-G2)
PIM (C-2, C-G2)
(C-SL, C-G1)
Source/ Receiver PE2
Receiver PE3
Source Inst. 2
(C-SL, C-G1)
Source PE2
3. PIM ASM join locally resolved in Receiver VRFs thanks to RP/Map (1 and 2): PIM SSM propagated into Source Instance.
al_0692

Figure 29: Multicast Extranet with per group PIM ASM mapping

PIM ASM extranet is achieved by local mapping from the receiver to source routing instances on a receiver PE. The mapping allows propagation of anycast RP PIM register messages between the source and receiver routing instances over an auto-created internal extranet interface. This PIM register propagation allows the receiver routing instance to resolve PIM ASM joins to multicast source(s) and to propagate PIM SSM joins over an auto-created extranet interface to the source routing instance. PIM SSM joins are then propagated towards the multicast source within the source routing instance.

The following MVPN topologies are supported:

- Rosen MVPN with MDT SAFI: a local replication on a source PE and multiple-source/multiple-receiver replication on a receiver PE
- RFC 6513/6514 NG-MVPN (including RFC 6625 (C-*, C-*) wildcard S-PMSI): a local replication on a source PE and a multiple source/multiple receiver replication on a receiver PE
- Extranet for GRT-source/VRF receiver with a local replication on a source PE and a multiple-receiver replication on a receiver PE
- Locally attached receivers are supported without SSM mapping

To achieve the extranet replication, the operator must configure:

- Local PIM ASM mapping on a receiver PE from a receiver routing instance to a source routing instance (**configure>service>vprn>mvpn>rpf-select>core-mvpn** or **configure>service>vprn>pim>grt-extranet** as applicable).
- An anycast RP mesh between source and receiver PEs in the source routing instance.

Caveats:

- This feature is supported for IPv4 multicast.
- The multicast source must reside in the source routing instance the ASM map points to on a receiver PE (the deployment of transit replication extranet from source instance to core instance on Source PE with ASM map extranet from receiver instance to core instance on a receiver PE is not supported).
- A given multicast group can be mapped in a receiver routing instance using either PIM SSM mapping or PIM ASM mapping but not both.
- A given multicast group cannot map to multiple source routing instances.
- This feature requires network chassis mode D.

Non-Congruent Unicast and Multicast Topologies for Multicast VPN

Operators that prefer to keep unicast and multicast traffic on separate links in network have option to maintain two separate instances of route table (unicast and multicast) per VPRN.

Multicast BGP can be used to advertise separate multicast routes using Multicast NLRI (SAFI 2) on PE-CE link within VPRN instance. Multicast routes maintained per VPRN instance can be propagated between PE-PE using BGP Multicast-VPN NLRI (SAFI 129).

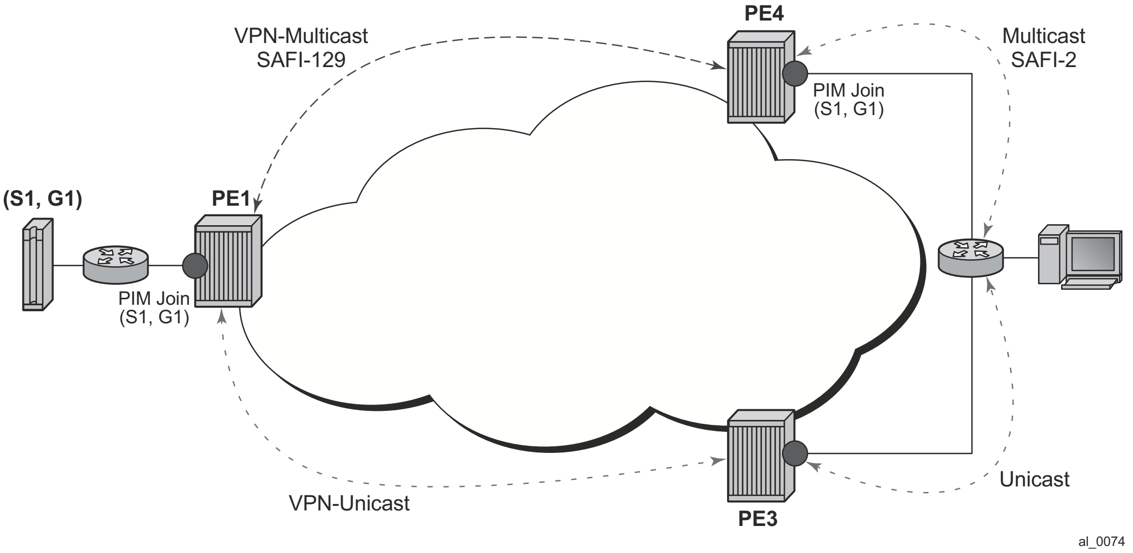


Figure 30: Incongruent Multicast and Unicast Topology for Non-Overlapping Traffic Links

SR OS supports option to perform RPF check per VPRN instance using multicast route table, unicast route table or both.

Non-congruent unicast and multicast topology is supported with NG-MVPN. Draft Rosen is not supported.

Multicast Auto RP Discovery

Auto-RP is a third-party proprietary implementation of a protocol to dynamically learn about availability of Rendezvous Point (RP) in network. Auto-RP protocol consists of announcing, mapping and discovery functions. SR OS only supports interoperability with the discovery mode of Auto-RP that includes mapping and forwarding of RP-mapping and RP-candidate messages using Alcatel-Lucent proprietary control processing. Auto-RP support is only available with multicast VPN (NG-MVPN and Rosen MVPN with BGP SAFI).

Main caveats:

- Support IPv4 multicast only.
- Provides interoperability with Auto-RP implementations without implementing PIM Dense mode processing.
- Does not support dual-homing of RP agent or RP (on same or different PEs).

IPv6 MVPN Support

IPv6 multicast support in SR OS allows operators to offer customers IPv6 multicast MVPN service. An operator utilizes IPv4 mLDP or RSVP-TE core to carry IPv6 c-multicast traffic inside IPv4 mLDP or RSVPE-TE provider tunnels (p-tunnels). The IPv6 customer multicast on a given MVPN can be blocked, enabled on its own or in addition to IPv4 multicast per PE or per interface. When both IPv4 and IPv6 multicast is enabled for a given MVPN, a single tree is used to carry both IPv6 and IPv4 traffic. Figure 31 shows an example of an operator with IPv4 MPLS backbone providing IPv6 MVPN service to Customer 1 and Customer 2.

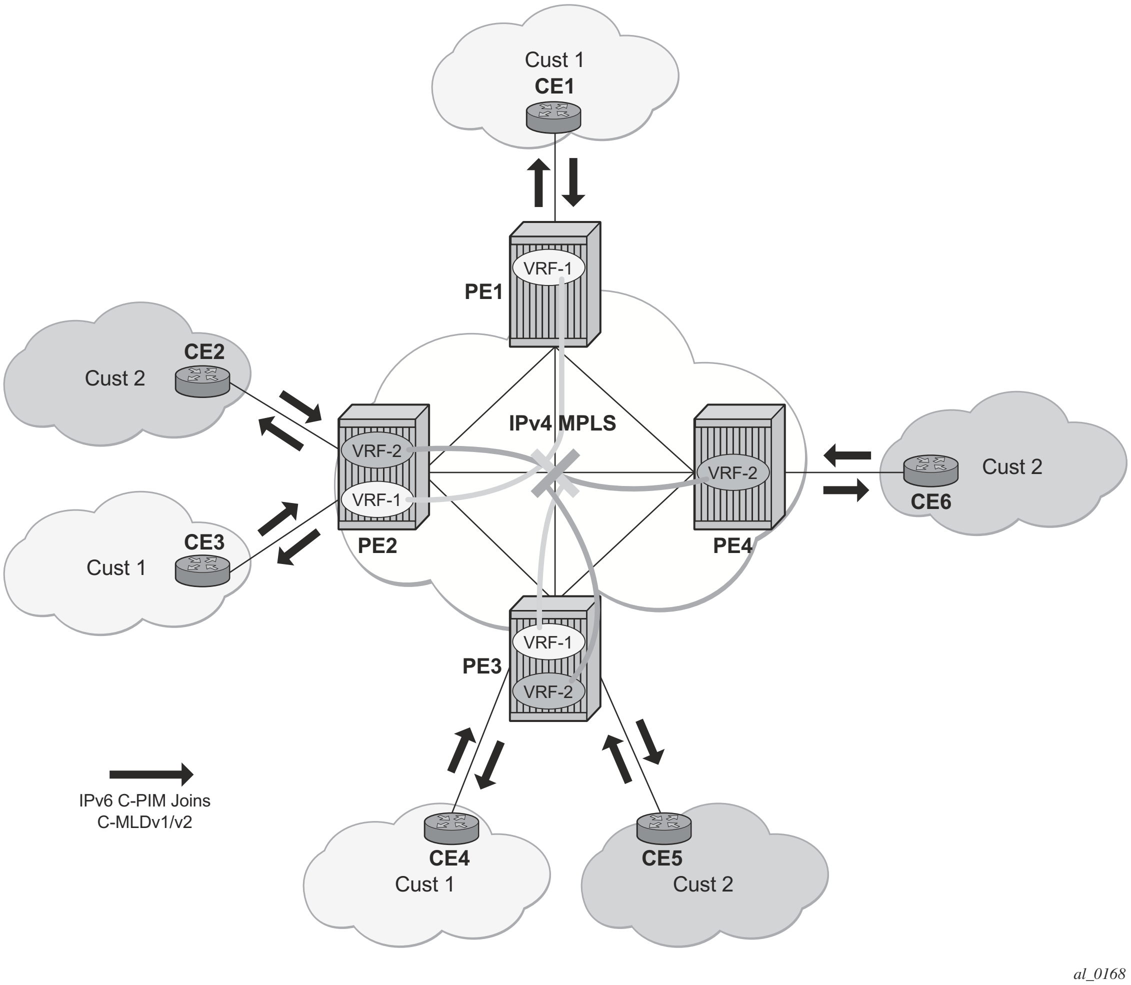


Figure 31: IPv6 MVPN Example

SROS IPv6 MVPN multicast implementation provides the following functionality:

- IPv6 C-PIM-SM (ASM and SSM)
- MLDv1 and MLDv2
- SSM mapping for MLDv1
- I-PMSI and S-PMSI using IPv4 P2MP mLDP p-tunnels
- I-PMSI and S-PMSI using IPv4 P2MP RSVP p-tunnels
- BGP auto-discovery
- PE-PE transmission of C-multicast routing using BGP mvpn-ipv6 address family

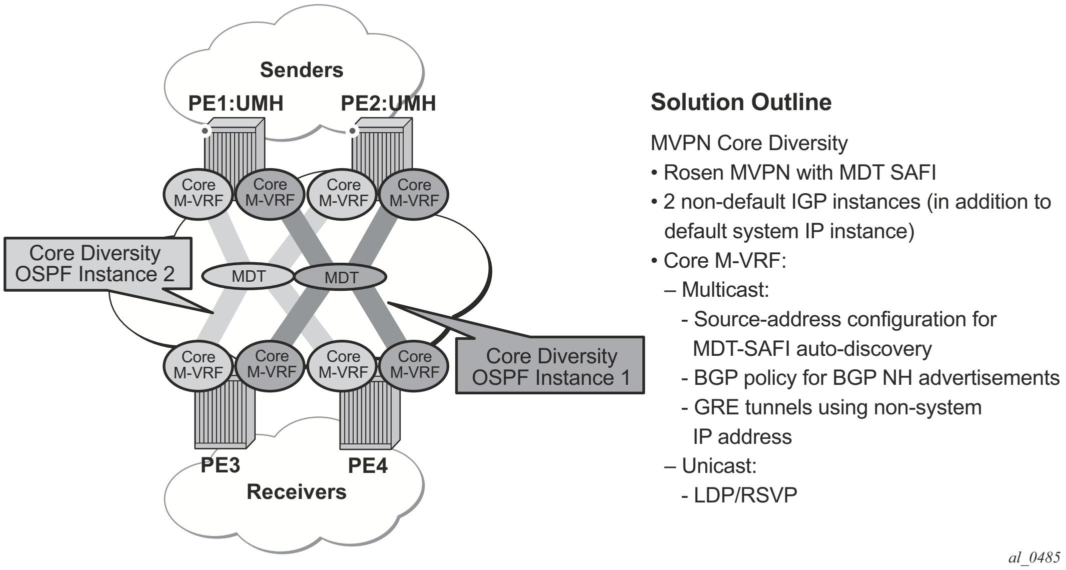
- IPv6 BSR/RP functions on functional par with IPv4 (auto-RP using IPv4 only)
- Embedded RP
- Inter-AS Option A

The following known caveats exist for IPv6 MVPN support:

1. IPv6 MVPN requires chassis mode D
2. Non-congruent topologies are not supported
3. IPv6 is not supported in MCAC
4. If IPv4 and IPv6 multicast is enabled, per-MVPN multicast limits apply to entire IPv4 and IPv6 multicast traffic as it is carried in a single PMSI. For example IPv4 AND IPv6 S-PMSIs are counted against a single S-PMSI maximum per MVPN.
5. IPv6 Auto-RP is not supported

Multicast Core Diversity for Rosen MDT_SAFI MVPNs

Figure 32 depicts Rosen MVPN core diversity deployment:



Solution Outline

MVPN Core Diversity

- Rosen MVPN with MDT SAFI
- 2 non-default IGP instances (in addition to default system IP instance)
- Core M-VRF:
  - Multicast:
    - Source-address configuration for MDT-SAFI auto-discovery
    - BGP policy for BGP NH advertisements
    - GRE tunnels using non-system IP address
  - Unicast:
    - LDP/RSVP

Figure 32: Multicast Core Diversity

Core diversity allows operator to optionally deploy multicast MVPN in either default IGP instance or one of two non-default IGP instances to provide, for example, topology isolation or different level of services. The following describes main feature attributes:

1. Rosen MVPN IPv4 multicast with MDT SAFI is supported with default and data MDTs.
2. Rosen MVPN can use a non-default OSPF or ISIS instance (using their loopback addresses instead of a system address).
3. Up to 3 distinct core instances are supported: system + 2 non-default OSPF instances – referred as “red” and “blue” below.
4. Note that the BGP Connector also uses non-default OSPF loopback as NH, allowing Inter-AS option B/C functionality to work with Core diversity as well.
5. The feature is supported with CSC-VPRN.

On source PEs (PE1: UMH, PE2: UMH in Figure 32), an MVPN is assigned to a non-default IGP core instance as follows:

1. MVPN is statically pointed to use one of the non-default “red”/“blue” IGP instances loopback addresses as source address instead of system loopback IP.

2. MVPN export policy is used to change unicast route next-hop VPN address (no longer required as of SROS Release 12.0.R4 - BGP Connector support for non-default instances).

The above configuration ensures that MDT SAFI and IPVPN routes for the non-default core instance use non-default IGP loopback instead of system IP. This ensures PIM advertisement/joins run in the proper core instance and GRE tunnels for multicast can be set-up using and terminated on non-system IP.

Note that if BGP export policy is used to change unicast route next-hop VPN address, unicast traffic must be forwarded in non-default “red” or “blue” core instance LDP or RSVP (terminating on non-system IP) must be used. GRE unicast traffic termination on non-system IP is not supported, and any GRE traffic arriving at the PE in “blue”, “red” instances destined to non-default IGP loopback IP will be forwarded to CPM (ACL or CPM filters can be used to prevent the traffic from reaching the CPM). This limitation does not apply if BGP connector attribute is used to resolve the multicast route.

No configuration is required on non-source PEs.

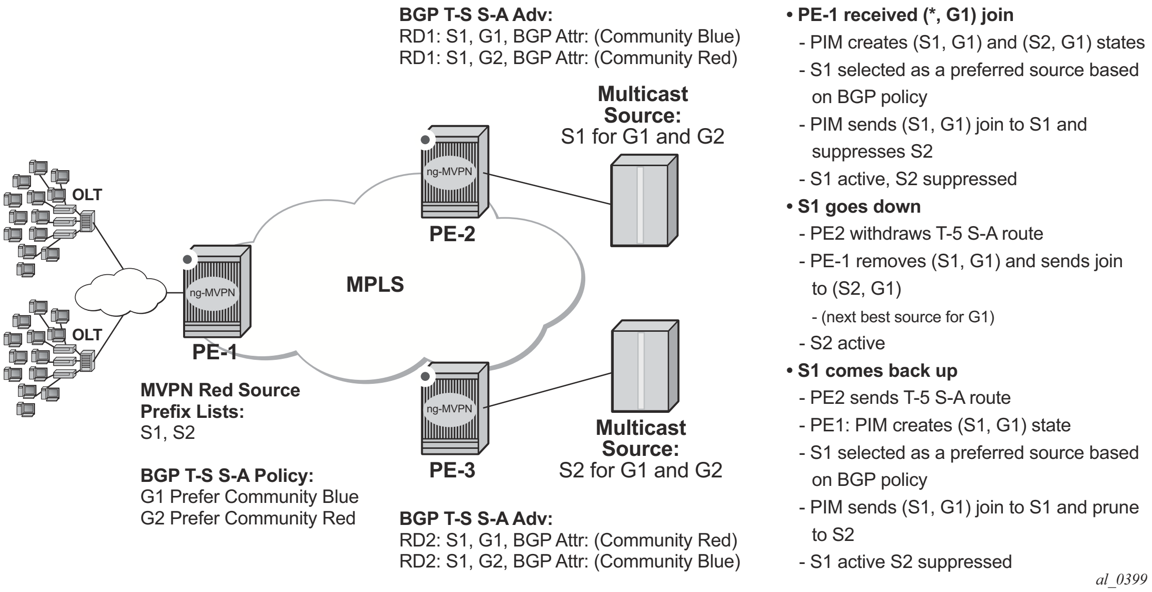
Feature caveats:

- VPRN instance must be shutdown to change the mdt-safi source-address. Note that CLI rollback that includes change of the auto-discovery is thus service impacting.
- To reset mdt-safi source-address to system IP, operator must first execute no auto-discovery (or auto-discovery default) then auto-discovery mdt-safi
- Configuring system IP as a source-address will consume one of the 2 IP addresses allowed, thus it should not be done.
- Operators must configure proper IGP instance loopback IP addresses within Rosen MVPN context and must configure proper BGP policies (prior to release 12.0.R4) for the feature to operate as expected. There is no verification that the address entered for MVPN provider tunnel source-address is such an address or is not a system IP address.
- The feature requires all ports to be present on IOM3-XP or newer (Chassis Mode D). The restriction is not enforced. Failing to observe this restriction will prevent the feature to operate properly in the network.

NG-MVPN Multicast Source Geo-Redundancy

Multicast source geo-redundancy is targeted primarily for MVPN deployments for multicast delivery services like IPTV. The solutions allows operators to configure a list of geographically dispersed redundant multicast sources (with different source IPs) and then, using configured BGP policies, ensure that each Receiver PE (a PE with receivers in its C-instance) selects only a single, most-preferred multicast source for a given group from the list. Although the data may still be replicated in P-instance (each multicast source sends (C-S, C-G) traffic onto its I-IPMSI tree or S-PMSI tree), each Receiver PE only forwards data to its receivers from the preferred multicast source. This allows operators to support multicast source geo-redundancy without the replication of traffic for each (C-S, C-G) in the C-instance while allowing fast recovery of service when an active multicast source fails.

Figure 33 shows an operational example of multicast source geo-redundancy:



- **PE-1 received (*, G1) join**
  - PIM creates (S1, G1) and (S2, G1) states
  - S1 selected as a preferred source based on BGP policy
  - PIM sends (S1, G1) join to S1 and suppresses S2
  - S1 active, S2 suppressed
- **S1 goes down**
  - PE2 withdraws T-5 S-A route
  - PE-1 removes (S1, G1) and sends join to (S2, G1)
    - (next best source for G1)
  - S2 active
- **S1 comes back up**
  - PE2 sends T-5 S-A route
  - PE1: PIM creates (S1, G1) state
  - S1 selected as a preferred source based on BGP policy
  - PIM sends (S1, G1) join to S1 and prune to S2
  - S1 active S2 suppressed

Figure 33: Preferred Source Selection for Multicast Source Geo-Redundancy

Operators can configure a list of prefixes for multicast source redundancy per MVPN on Receiver PEs:

- Up to 8 multicast source prefixes per VPRN are supported.
- Any multicast source that is not part of the source prefix list is treated as a unique source and automatically joined in addition to joining the most preferred source from the redundant multicast source list.

A Receiver PE selects a single, most-preferred multicast source from the list of pre-configured sources for a given MVPN during (C-*, C-G) processing as follows:

- A single join for the group is sent to the most preferred multicast source from the operator-configured multicast source list. Joins to other multicast sources for a given group are suppressed. Operator can see active and suppressed joins on a Receiver PE. Although a join is sent to a single multicast source only, (C-S, C-G) state is created for every source advertising Type-5 S-A route on the Receiver PE.
- The most preferred multicast source is a reachable source with the highest local preference for Type-5 SA route based on the BGP policy, as described later in this section.
- On a failure of the preferred multicast source or when a new multicast source with a better local preference is discovered, Receiver PE will join the new most-preferred multicast source. The outage experienced will depend on how quickly Receiver PE receives Type-5 S-A route withdrawal or looses unicast route to multicast source, and how quickly the network can process joins to the newly selected preferred multicast source(s).
- Note: Local multicast sources on a Receiver PE are not subject to the most-preferred source selection, regardless of whether they are part of redundant source list or not.

BGP policy on Type-5 SA advertisements is used to determine the most preferred multicast source based on the best local preference as following:

- Each Source PE (a PE with multicast sources in its C-instance) tags Type-5 SA routes with a unique standard community attribute using global BGP policy or MVPN vrf-export policy. Depending on multicast topology, the policy may require source-aware tagging in the policy. Either all MVPN routes or Type 5 SA routes only can be tagged in the policy (new attribute **mvpn-type 5**).
- Each receiver PE has a BGP VRF import policy that sets local preference using match on Type-5 SA routes (new attribute **mvpn-type 5**) and standard community attribute value (as tagged by the Source PEs). Using policy statements that also include group address match, allows receiver PEs to select the best multicast source per group. The BGP VRF import policy must be applied as **vrf-import** under *config>service>vprn>mvpn* context. It must have default-action **accept** specified, or all MVPN routes other than those matched by specified entries will be rejected. In addition, it must have **vrf-target** as a community match condition, because **vrf-target mvpn** configuration is ignored when **vrf-import** policy is defined.

Operators can change redundant source list or BGP policy affecting source selection in service. If such a change of the list/policy results in a new preferred multicast source election, make-before-break is used to join the new source and prune the previously best source.

For the proper operations, MVPN multicast source geo-redundancy requires the router:

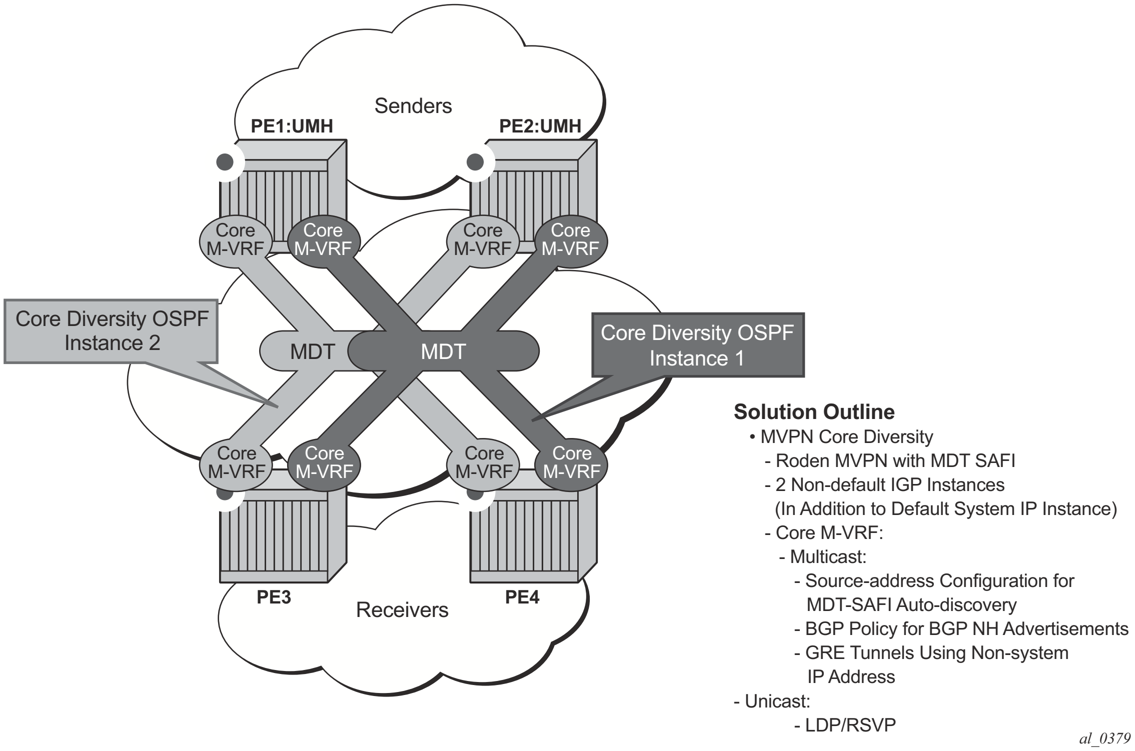
- To maintain the list of eligible multicast sources on Receiver PEs, Source PE routers must generate Type-5 S-A route even if the Source PE sees no active joins from any receiver for a given group.
- To trigger a switch from a currently active multicast source on a Receiver PE, Source PE routers must withdraw Type-5 S-A route when the multicast source fails or alternatively unicast route to multicast source must be withdrawn or go down on a Receiver PE.

MVPN multicast source redundancy solutions is supported for the following configurations only. Enabling the feature in unsupported configuration must be avoided:

1. NG-MVPN with RSVP-TE or mLDP or PIM with BGP c-multicast signaling in P-instance. Both I-PMSI and S-PMSI trees are supported.
2. IPv4 and IPv6 (C-*, C-G) PIM ASM joins in the C-instance.
3. Both **intersite-shared enabled** and **disabled** are supported. For **intersite-shared enabled**, operators must enable generation of Type-5 S-A routes even in the absence of receivers seen on Source PEs (intersite-shared persistent-type5-adv must be enabled).
4. The Source PEs must be configured as a sender-receiver, the Receiver PEs can be configured as a sender-receiver or a receiver-only.
5. The RP(s) must be on the Source PE(s) side. Static RP, anycast-RP, embedded-RP types are supported.
6. UMH redundancy can be deployed to protect Source PE to any multicast source. When deployed, UMH selection is executed independently of source selection after the most preferred multicast source had been chosen. Supported **umh-selection** options include: **highest-ip**, **hash-based**, and **unicast-rt-pref** (no support for tunnel-status).

Multicast Core Diversity for Rosen MDT SAFI MVPNs

Figure 34 shows a Rosen MVPN core diversity deployment.



Solution Outline

- MVPN Core Diversity
  - Roden MVPN with MDT SAFI
  - 2 Non-default IGP Instances (In Addition to Default System IP Instance)
  - Core M-VRF:
    - Multicast:
      - Source-address Configuration for MDT-SAFI Auto-discovery
      - BGP Policy for BGP NH Advertisements
      - GRE Tunnels Using Non-system IP Address
- Unicast:
  - LDP/RSVP

Figure 34: Multicast Core Diversity

Core diversity allows operators to optionally deploy multicast MVPN in either default IGP instance. or one of two non-default IGP instances to provide; for example, topology isolation or different level of services. The following describes the main feature attributes:

- Rosen MVPN IPv4 multicast with MDT SAFI is supported with default and data MDTs.
- Rosen MVPN can use a non-default OSPF or ISIS instance (using their loopback addresses instead of a system address).
- Up to 3 distinct core instances are supported: system + 2 non-default OSPF instances shown in Figure 34.
- Note that the BGP Connector also uses non-default OSPF loopback as NH, allowing Inter-AS option B/C functionality to work with Core diversity as well.
- The feature is supported with CSC-VPRN.

On source PEs (PE1: UMH, PE2: UMH in the above picture), an MVPN is assigned to a non-default IGP core instance as follows:

- MVPN is statically pointed to use one of the non-default IGP instances loopback addresses as source address instead of system loopback IP.
- MVPN export policy is used to change unicast route next-hop VPN address.
- BGP Connector support for non-default instances.

The configuration shown above ensures that MDT SAFI and IPVPN routes for the non-default core instance use non-default IGP loopback instead of system IP. This ensures PIM advertisement/joins run in the proper core instance and GRE tunnels for multicast can be set-up using and terminated on non-system IP. Note that if BGP export policy is used to change unicast route next-hop VPN address instead of BGP Connector attribute-based processing and unicast traffic must be forwarded in non-default core instances 1 or 2, LDP or RSVP (terminating on non-system IP) must be used. GRE unicast traffic termination on non-system IP is not supported and any GRE traffic arriving at the PE in instances 1 or 2, destined to non-default IGP loopback IP will be forwarded to CPM (ACL or CPM filters can be used to prevent the traffic from reaching the CPM).

No configuration is required on non-source PEs.

Known feature caveats include:

- VPRN instance must be shutdown to change the mdt-safi source-address. Note that CLI rollback that includes change of the auto-discovery is thus service impacting.
- To reset mdt-safi source-address to system IP, operator must first execute no auto-discovery (or auto-discovery default) then auto-discovery mdt-safi
- Configuring system IP as a source-address will consume one of the 2 IP addresses allowed, thus it should not be done.
- Operators must configure proper IGP instance loopback IP addresses within Rosen MVPN context and must configure proper BGP policies (prior to release R12.0R4) for the feature to operate as expected. There is no verification that the address entered for MVPN provider tunnel source-address is such an address or is not a system IP address.
- The feature requires all ports to be present on IOM3-XP or newer (Chassis Mode D). The restriction is not enforced. Failing to observe this restriction will prevent the feature to operate properly in the network.

Inter-AS MVPN

The Inter-AS MVPN feature allows set-up of Multicast Distribution Trees (MDTs) that span multiple Autonomous Systems (ASes).This section focuses on multicast aspects of the Inter-AS MVPN solution.

To support Inter-AS option for MVPNs, a mechanism is required that allows setup of Inter-AS multicast tree across multiple ASes. Due to limited routing information across AS domains, it is not possible to setup the tree directly to the source PE. Inter-AS VPN Option A does not require anything specific to inter-AS support as customer instances terminate on ASBR and each customer instance is handed over to the other AS domain via a unique instance. This approach allows operators to provide full isolation of ASes, but the solution is the least scalable case, as customer instances across the network have to exist on ASBR.

Inter-AS MVPN Option B allows operators to improve upon the Option A scalability while still maintaining AS isolation, while Inter-AS MVPN option C further improves Inter-AS scale solution but requires exchange of Inter-AS routing information and thus is typically deployed when a common management exists across all ASes involved in the Inter-AS MVPN. The following sub-sections provide further details on Inter-AS Option B and Option C functionality.

BGP Connector Attribute

BGP connector attribute is a transitive attribute (unchanged by intermediate BGP speaker node) that is carried with VPNv4 advertisements. It specifies the address of source PE node that originated the VPNv4 advertisement.

With Inter-AS MVPN Option B, BGP next-hop is modified by local and remote ASBR during re-advertisement of VPNv4 routes. On BGP next-hop change, information regarding the originator of prefix is lost as the advertisement reaches the receiver PE node.

BGP connector attribute allows source PE address information to be available to receiver PE, so that a receiver PE is able to associate VPNv4 advertisement to the corresponding source PE.

PIM RPF Vector

In case of Inter-AS MVPN Option B, routing information towards the source PE is not available in a remote AS domain, since IGP routes are not exchanged between ASes. Routers in an AS other than that of a source PE, have no routes available to reach the source PE and thus PIM JOINs would never be sent upstream. To enable setup of MDT towards a source PE, BGP next-hop (ASBR) information from that PE's MDT-SAFI advertisement is used to fake a route to the PE. If the BGP next-hop is a PIM neighbor, the PIM JOINs would be sent upstream. Otherwise, the PIM JOINs would be sent to the immediate IGP next-hop (P) to reach the BGP next-hop. Since the IGP next-hop does not have a route to source PE, the PIM JOIN would not be propagated forward unless it carried extra information contained in RPF Vector.

In case of Inter-AS MVPN Option C, unicast routing information towards the source PE is available in a remote AS PEs/ASBRs as BGP 3107 tunnels, but unavailable at remote P routers. If the tunneled next-hop (ASBR) is a PIM neighbor, the PIM JOINs would be sent upstream. Otherwise, the PIM JOINs would be sent to the immediate IGP next-hop (P) to reach the tunneled next-hop. Since the IGP next-hop does not have a route to source PE, the PIM JOIN would not be propagated forward unless it carried extra information contained in RPF Vector.

To enable setup of MDT towards a source PE, PIM JOIN thus carries BGP next hop information in addition to source PE IP address and RD for this MVPN. For option-B, both these pieces of information are derived from MDT-SAFI advertisement from the source PE. For option-C, both these pieces of information are obtained from the BGP tunneled route.

The RPF vector is added to a PIM join at a PE router when configure router **pim rpfv** option is enabled. P routers and ASBR routers must also have the option enabled to allow RPF Vector processing. If the option is not enabled, the RPF Vector is dropped and the PIM JOIN is processed as if the PIM Vector were not present.

For further details about RPF Vector processing please refer to [RFCs 5496, 5384 and 6513]

Inter-AS MVPN Option B

Inter-AS Option B is supported for Rosen MVPN PIM SSM using BGP MDT SAFI, PIM RPF Vector and BGP Connector attribute. The Figure 35 depict set-up of a default MDT:

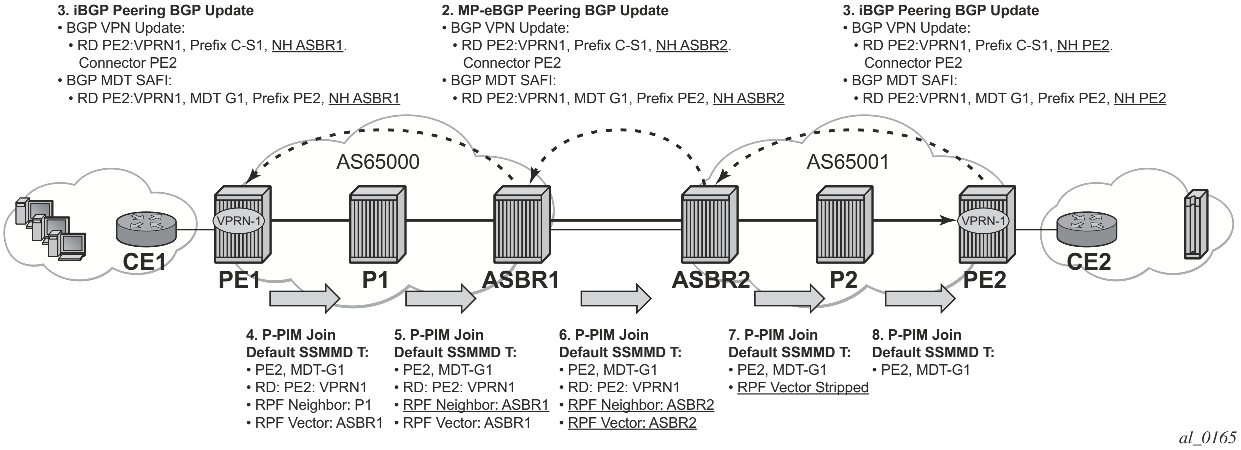


Figure 35: Inter-AS Option B Default MDT Setup

SROS inter-AS Option B is designed to be standard compliant based on the following RFCs:

- RFC 5384 - The Protocol Independent Multicast (PIM) Join Attribute Format
- RFC 5496 - The Reverse Path Forwarding (RPF) Vector TLV
- RFC 6513 - Multicast in MPLS/BGP IP VPNs

The SROS implementation was designed also to interoperate with older routers Inter-AS implementations that do not comply with the RFC 5384 and RFC 5496.

Inter-AS MVPN Option C

Inter-AS Option C is supported for Rosen MVPN PIM SSM using BGP MDT SAFI and PIM RPF Vector. Figure 36 depicts a default MDT setup:

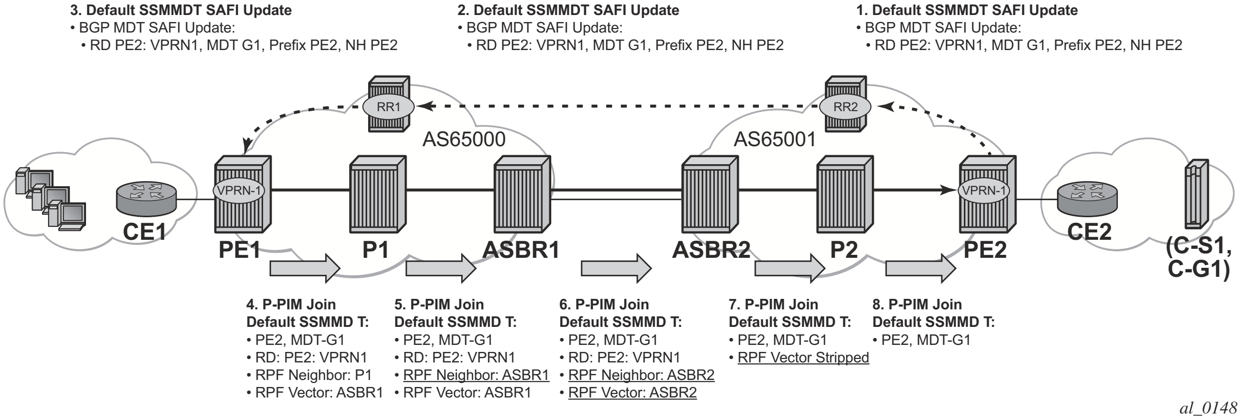


Figure 36: Inter-AS Option C Default MDT Setup

Additional caveats for Inter-AS MVPN Option B and C support are the following:

1. Inter-AS MVPN option B is not supported with duplicate PE addresses.
2. For Inter-AS Option C, BGP 3107 routes are installed into unicast rtm (rtable-u), unless routes are installed by some other means into multicast rtm (rtable-m), and option C will not build core MDTs, therefore, rpf-table is configured to rtable-u or both.
3. Additional Cisco interoperability notes are the following:
   → RFC 5384 - The Protocol Independent Multicast (PIM) Join Attribute Format
   → RFC 5496 - The Reverse Path Forwarding (RPF) Vector TLV
   → RFC 6513 - Multicast in MPLS/BGP IP VPNs

   The SROS implementation was designed to inter-operate with Cisco routers Inter-AS implementations that do not comply with the RFC5384 and RFC5496.

When **configure router pim rpfv mvpn** option is enabled, Cisco routers need to be configured to include RD in an RPF vector using the following command: **ip multicast vrf vrf-name rpf proxy rd vector** for interoperability. When Cisco routers are not configured to include RD in an RPF vector, operator should configure SROS router (if supported) using **configure router pim rpfv core mvpn**: PIM joins received can be a mix of core and mvpn RPF vectors.

VPRN Off-Ramp

This feature enhances the operation and deployment options service providers will have with the TMS-ISA DDoS and threat mitigation.

To enhance the operational flexibility of the DDoS scrubbing capabilities supported by MS-ISA TMS, this section discusses support for VPRN off-ramp to MS-ISA Threat Management Server (TMS) adding to the GRT (Global Routing Table) off-ramp already supported.

The current deployment model works as follows:

1. Off-ramping suspect traffic — Upon detection of a destination IP prefix under attack, the Collector Platform (CP) will communicate with a version of TMS running on MS-ISA (if licensed) instructing TMS-ISA to communicate with the 7750 Control Processing Module (CPM) to issue a BGP route advertisement setting a 7750 SR interface as the next-hop for traffic destined for the IP under attack. At the same time, the 7750 SR sets up an internal route across one or more TMS-ISAs so when traffic arrives destined for the route under attack, those packets will be forwarded to the TMS-ISA for DDoS scrubbing.
3. On-ramping of clean traffic — Once traffic is forwarded to the TMS-ISA, the attack traffic is separated from the legitimate traffic destined to the address or address range under attack, the "clean" traffic is forced into a clean VRF that must contain the original PE serving the IP under attack. The traffic is then tunneled (by LDP or GRE) to the originally intended PE where it can be routed to its destination.

This features offer the ability to off-ramp traffic directly from VPN context where MS-ISA TMS may communicate with may communicate with CPM to insert internal re-direct routes within specific VPNs used as either DDoS scrubbing VRFs (where an operator designates one or more VPRNs for carrying traffic to be scrubbed) or where the VPRN contains the global routes using the MPLS network to transport Internet traffic.

In this case, rather than having the MS-ISA TMS trigger an internal re-direct route from GRT to one or more MS-ISA TMS instances, we allow the operator to configure a VPRN associated with carrying off-ramp traffic to be scrubbed so that prefixes under attack are added to that VPRN VRF ensuring traffic coming in on that VPN is off-ramped to MS-ISA TMS for scrubbing.

This model allows operational flexibility and simplicity where GRTs throughout the network do not need to be constantly changed for off-ramping of traffic, but instead off-ramp traffic can be isolated into a "dirty VRF" for cleaning after which it is returned to an unmodified global routing table or another operator defined VPN used for on-ramping clean traffic toward customers.

In addition, allowing VRF-based redirection to MS-ISA TMS provides a critical piece of a full flowspec directed DDoS scrubbing solution where flowspec policies are used to separate attack flows into "dirty VRF" where 7750 SRs running MS-ISA TMS scan then be configured to forward traffic from these DDoS VRFs to the TMS running on MS-ISA for scrubbing. Clean traffic would then be returned to the network by either a VPN or the GRT.

DDoS Off-Ramping Through Flowspec

Flowspec policy can be used to further simplify DDoS scrubbing architecture resulting in more scalable and operationally simplified topologies.

In this case, CP detectors will detect the destination under attack, including the flows (src/dest pairs) involved in the attack and either directly (or through interaction with an operator's scripting engine) advertise flowspec extended BGP updates to peering and edge routers.

Peering/edge routes with flowspec enabled can redirect traffic to a VPN ensuring that traffic signaled for off-ramp by a CP is placed into a "dirty VPN" which will then be used to forward traffic to MS-ISA TMS for scrubbing.

Rather than altering the GRT table (as in the routed off-ramp model), flowspec policy allows granular policy redirection to take place to off-ramp VRF.

Since the GRT has not been altered, clean traffic is forwarded back to either a clean VRF or the GRT for routing to its originally intended destination.

Flowspec forwarding requires the IOM-3.

FIB Prioritization

The RIB processing of specific routes can be prioritized through the use of the **rib-priority** command. This command allows specific routes to be prioritized through the protocol processing so that updates are propagated to the FIB as quickly as possible.

The **rib-priority** command can be configured within the VPRN instance of the OSPF or IS-IS routing protocols. For OSPF, a prefix list can be specified that identifies which route prefixes should be considered high priority. If the rib-priority high command is configured under an **VPRN>OSPF>area>interface** context then all routes learned through that interface is considered high priority. For the IS-IS routing protocol, RIB prioritization can be either specified though a prefix-list or an IS-IS tag value. If a prefix list is specified than route prefixes matching any of the prefix list criteria will be considered high priority. If instead an IS-IS tag value is specified then any IS-IS route with that tag value will be considered high priority.

The routes that have been designated as high priority will be the first routes processed and then passed to the FIB update process so that the forwarding engine can be updated. All known high priority routes should be processed before the routing protocol moves on to other standard priority routes. This feature will have the most impact when there are a large number of routes being learned through the routing protocols.