Highly Available Inter-Domain Multicast Transport over MPLS

Technical Whitepaper

Version 1.0

Authored by:

Nicholas Russo

CCDE #20160041

CCIE #42518 (EI/SP)

Change History

Version and Date	Change	Responsible Person
20210713 Version 0.1	Initial Draft	Nicholas Russo
20210927 Version 0.2	Technical corrections and EU MoFRR	Nicholas Russo
20220103 Version 1.0	Technical corrections and publication	Nicholas Russo

Contents

Figures

Tables

1. Solution Overview

This section explains the business problem to be solved and the high-level architecture of the solution described in this document.

1.1. Business Problem and Technical Challenges

A large US-based service provider was tasked with integrating with a slightly smaller provider in western Europe. This integration would enable inter-carrier connectivity within customer networks spread between both geographic locations. This included all combinations of IPv4/IPv6 and unicast/multicast traffic. Providing unicast connectivity is straightforward; both carriers used Multi-Protocol Label Switching (MPLS) for multi-tenancy, which offers three common options for integration. More complicated to solve was the multicast aspect for several reasons:

a. For Any Source Multicast (ASM), source discovery information must be shared between the carriers. Multicast Source Discovery Protocol (MSDP) exists for this purpose but has been defined only for IPv4 in RFC3618. Using MSDP for IPv4 would necessitate divergent architectures between IP versions, increasing the solution's overall complexity.
b. Most multicast traffic was mission-critical in nature and very sensitive to packet loss, much like "Broadcast Video" service class defined in RFC4594. Enabling fast-reroute (FRR) on multicast traffic is often more complicated than doing so for unicast traffic.
c. Overcoming ASM challenges using a pure Source-Specific Multicast (SSM) deployment was difficult since not all hosts in the network supported IPv4 Internet Group Management Protocol (IGMP) version 3 and IPv6 Multicast Listener Discovery (MLD) version 2. Even ignoring the inter-carrier ASM difficulties, intra-carrier ASM has its own challenges with respect to rendezvous point (RP) placement/availability, shortest-path tree (SPT) switchover, and operational complexity. Sometimes, even IGMPv3-capable and MLDv2-capable hosts ran applications that were unable to signal interest in multicast sources due to technical limitations with the software implementation.

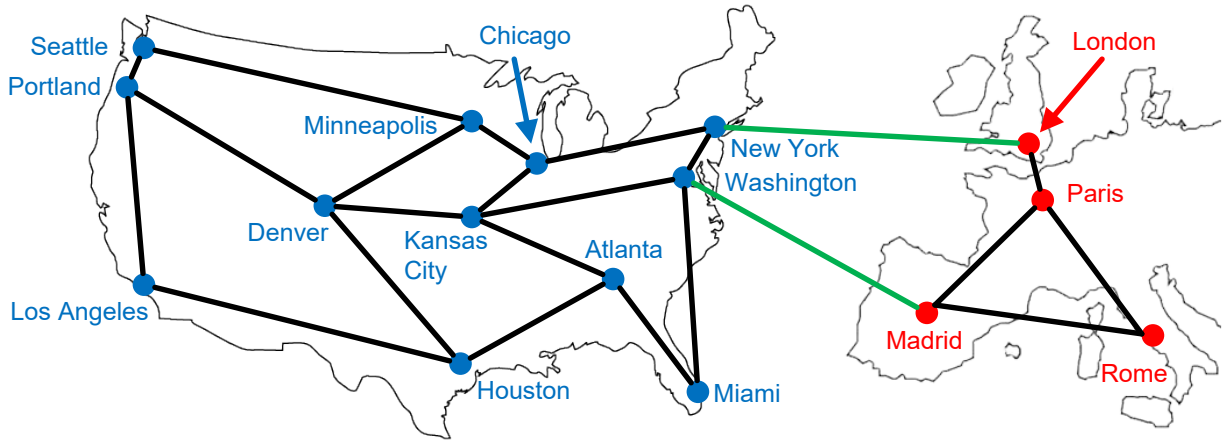
Note that this document expects readers to be technically skilled on the topics of general IP routing, MPLS, and multicast.

1.2. Architecture Overview

This document will describe all aspects of the inter-carrier (also called inter-AS or autonomous system) design but focuses specifically on the multicast design. In the interest of customer anonymity and security, the locations of carrier points of presence (POPs) have been changed in this document. Black links within a continent are intra-AS, unified by a single Interior Gateway Protocol (IGP) and Border Gateway Protocol (BGP) AS number. Green links that interconnect the continents use external BGP (eBGP) to exchange customer routes. Each is enabled with

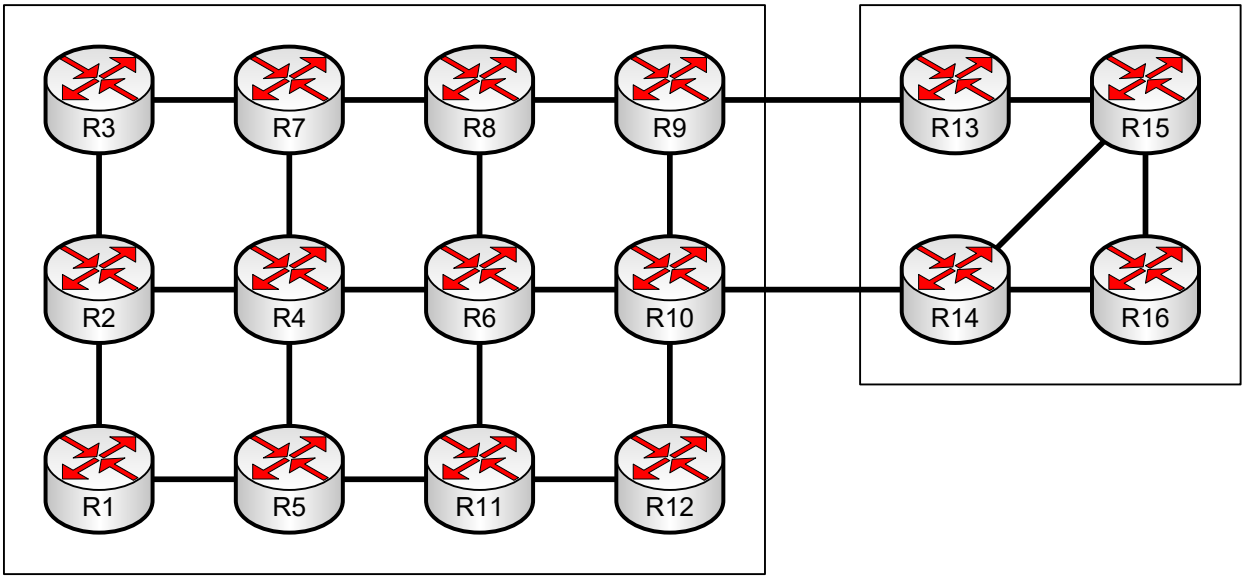
various MPLS-related technologies (discussed later), ultimately enabling end-to-end multi-tenancy between any pair of cities on the map below.

Figure 1 - World Map of Connected Cities



The diagram below translates the map above into a network diagram. This diagram below, and derivatives of it, will appear regularly throughout this document. The router numbers are used for brevity and map to the cities shown above.

Figure 2 - High-Level Device Interconnections



The solution relies exclusively on SSM for inter-AS multicast transport. This has many benefits:

a. Operationally simple to build and maintain; no need for RPs, MSDP, or tree switchovers
b. Offers a unified multicast transport solution for both IPv4 and IPv6
c. Integrates with Domain Name System (DNS) in mapping multicast groups to sources for clients that do not support IGMPv3 and/or MLDv2

d. Provides more label-switched multicast (LSM) options for core multicast transport, such as multicast Label Distribution Protocol (mLDP) in-band signaling
e. Easily scoped at customer and AS boundaries to regionalize traffic; no need for complex RP-related edge filtering based on join or register filters
f. Smaller attack surface; no RPs to defend from registration attacks or extensive (*,G) joins

On the topic of FRR for MPLS traffic, the solution protects both unicast and multicast flows. It works by combining one-hop "primary" MPLS traffic engineering (TE) tunnels with link-protecting "backup" MPLS TE tunnels. These TE tunnels are signaled using Resource Reservation Protocol (RSVP) MPLS TE extensions defined in RFC3209. This provides a relatively high-scale and fully dynamic FRR solution for traffic. The details of this design, along with many other supplemental aspects, are discussed in greater detail later in this document.

2. US-based Network Design

This section details how the US network was designed. Aspects irrelevant to this document, such as IPv4/v6 subnet allocation, device security, and operations management have been omitted for brevity.

2.1. IGP Design

This section explains the IGP design whereby Intermediate System to Intermediate System (IS-IS) has been deployed.

2.1.1. General IS-IS Design

The entire network uses a flat IS-IS level-2 (L2) network with all core interfaces participating in the same IS-IS process. Although irrelevant to IS-IS L2 operations, all routers are placed in the same IS-IS area of 49.0000. Areas beginning with 49 are identified as private by the International Organization for Standardization (ISO) and are commonly used in both documentary and production networks today. All core links are configured as IS-IS point-to-point, removing the Designated Intermediate System (DIS) election. This has many benefits:

a. Reduces IGP convergence time, both initially and after topology changes
b. Reduces topological complexity; no pseudo-nodes to traverse when running SPF
c. Reduces unnecessary control plane traffic as Complete Sequence Number Protocol Data Units (CSNP) can be sent less frequently, or not at all, after the initial neighbor establishment. The DIS must send them at regular intervals on multi-access networks

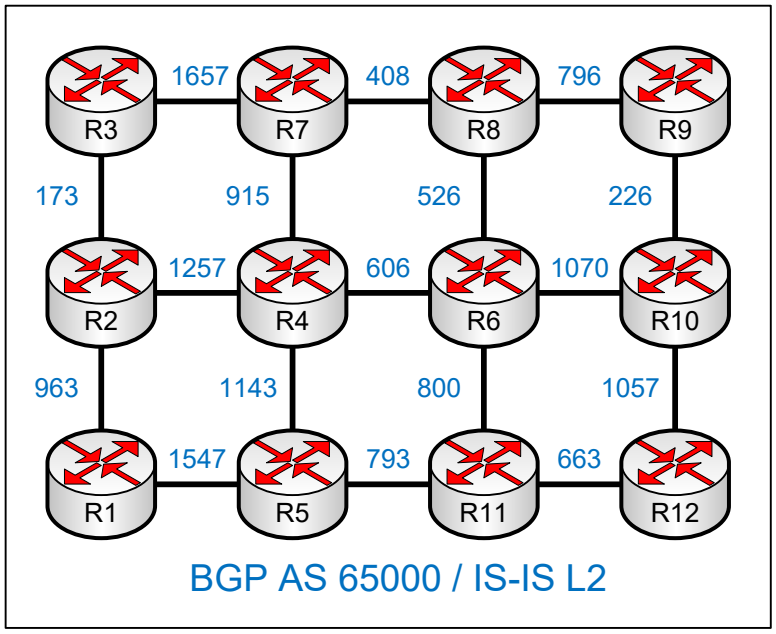
Because this is an MPLS network, there are two other useful IS-IS features available to network designers. First, we can reduce the size of the IPv4 routing table and slightly reduce the size of each IS-IS Link State Packet (LSP) by only advertising loopback prefixes within each LSP. The transit subnets between routers are not relevant for any MPLS services and can safely be omitted from advertisement. Operators troubleshooting a network should be aware that diagnostic tools like "ping" and "traceroute" will need to be sourced from device loopbacks as a result.

Second, IS-IS "wide metrics" should be enabled, which has multiple benefits. First, it enables a larger total end-to-end metric; the maximum narrow metric is 1023 (2^10-1) and the maximum wide metric is 16,777,215 (2^24-1). Additionally, wide metrics enables new types of Type Length Value mappings (TLV) to be communicated between devices. This enables IS-IS to advertise MPLS TE information between the routers so that TE tunnels can be built in the future. These TE-related extensions to IS-IS are defined in RFC3784.

Perhaps the most interesting aspect of this design is the IS-IS metric allocation. Each link has an IS-IS metric equal to the road mileage between the cities. Road distance is preferred over direct, as-the-crow-flies distance in this context because it often represents the path of physical links between cities. This provides a more accurate representation of the true "cost" of inter-city

transit across the national network. The diagram below shows the symmetrically configured IS-IS metrics configured on each link, colored in blue.

Figure 3 - IS-IS Metrics Based on Road Mileage



2.1.2. Fast Reconvergence and Optimization

This network relies heavily on MPLS TE-based FRR, which is discussed in depth later. This technology allows the network to quickly route around failed links, allowing IS-IS to convergence with minimal packet loss. Given that fact, tuning IS-IS to converge rapidly adds little value. The only advantage is that FRR tunnels would be used for shorter periods of time.

Tuning the many timers that go into an IS-IS design, such as the shortest-path first (SPF) and partial recalculation (PRC) timers, requires extensive research and testing. Configuring fast converging, aggressive timers often cause more flooding, which leads to slightly more bandwidth consumption and more computing expense on each device. This also increases operational complexity without any significant performance improvement for end users. Given that the network is a large, flat L2 flooding domain, additional flooding events should be minimized.

However, Cisco devices (and perhaps others) support a feature known as "fast-flood". When a router receives an LSP that indicates a topology change has occurred, it triggers an SPF run. The router then runs SPF, and when complete, floods the LSPs that caused SPF to run to other devices. Enabling "fast-flood" allows a router to flood SPF-causing LSPs before running SPF. This speeds up convergence by more rapidly distributing topology changes throughout the network without contributing to any additional flooding; it just speeds up flooding that was already scheduled. It also may reduce the number of total SPF runs in the network by ensuring more routers batch together more topology changes rather than reacting in isolation.

Since the execution of SPF and pacing of LSP updates is not relevant, the designers focused on link detection speed instead. Again, the reliance on FRR guarantees minimal traffic loss during

failover events, but only if the link failures can be detected quickly. Two different techniques are used to determine link failures depending on the type of core link.

When two routers shared a link whereby the layer-1 status is an accurate indication of up/down state, routers can rely on the optical or electrical carrier signal. On Cisco devices, the default "carrier-delay" is 10 milliseconds (ms) but was reduced to 5 ms in this particular network. A value of 0 ms is configurable but means that even a spurious up/down flap that causes 1 ms of loss would cause the entire network to react; FRR would be triggered, traffic would be redirected, and IS-IS would reconverge. This would cause an even greater outage than if the 1 ms of loss was silently tolerated. The precise value of 5 ms was not chosen arbitrarily, but rather as the result of extensive testing. Outages less than 5 ms were deemed acceptable with respect to packet loss (from a business perspective) whereas anything longer was considered a trigger for FRR to be triggered with subsequent IGP reconvergence.

On core links where the layer-1 status is not an accurate indication of up/down state, Bidirectional Forwarding Detection (BFD) can be used. This UDP-based protocol binds to clients, such as IS-IS and TE-FRR, and notifies them when bidirectional traffic is no longer flowing. There are two types of BFD packets:

a. Control packets which are sent slowly and are used to manage the session
b. Data packets (known as echoes) which are sent rapidly and determine the link's health

The data packets use a source IP and destination IP of the sender, ensuring that the packets are looped back to the sender at layer-3. This is how bidirectional connectivity is confirmed using the minimal amount of computing power between devices. In our environment, 50 ms was used as the BFD interval with a multiplier of 3, meaning that failures would be detected in 150 ms or less. Because most core links used direct fiber connections, relying on the carrier signal was much more common than relying on BFD. As such, most link failures were detected in 5 ms instead of 150 ms. Configuring BFD as a safeguard even on links where the layer-1 status is an accurate indicator of up/down status is often a good prevention against catastrophic failures.

2.2. MPLS Core Design

This section describes the MPLS design decision relating to core label switched path (LSP) construction between provider edge (PE) devices.

2.2.1. Primary TE Tunnels

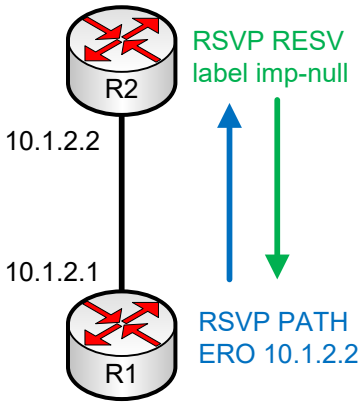
In Cisco parlance, the word "primary" in the context of MPLS TE tunnels often refers to the "one-hop auto-tunnel" feature. When enabled on a router, it constructs a one-hop TE tunnel from the router to every other directly connected device over point-to-point links. Because IS-IS distributes MPLS TE information, each router knows which peers are TE-capable; in the current design, this includes all devices. On Cisco devices, these tunnels automatically enable the following features:

a. Request for fast-reroute to protect against link failures. If the link over which the TE tunnel is built suffers a failure, it has the capability to be rerouted along another path.

b. Automatic announcement of IP prefixes reachable via the TE tunnel by modifying the IS-IS shortest path first (SPF) algorithm. PE loopbacks are therefore reachable through the tunnel, allowing all LSPs terminating on a given loopback to be protected. Cisco calls this "autoroute announce" but other vendors have comparable features.

At a protocol level, the headend sends an RSVP PATH message using an Explicit Route Object (ERO) towards the tailend, which is directly connected. Upon receipt, the tailend responds with an RSVP RESV (reservation) message containing an MPLS label mapping. Because these tunnels are only one-hop, the MPLS label assigned is implicit-null. Therefore, the primary tunnel does not add any MPLS encapsulation, making it inadequate for end-to-end MPLS transport between PEs that are not directly connected. The diagram below illustrates this RSVP signaling.

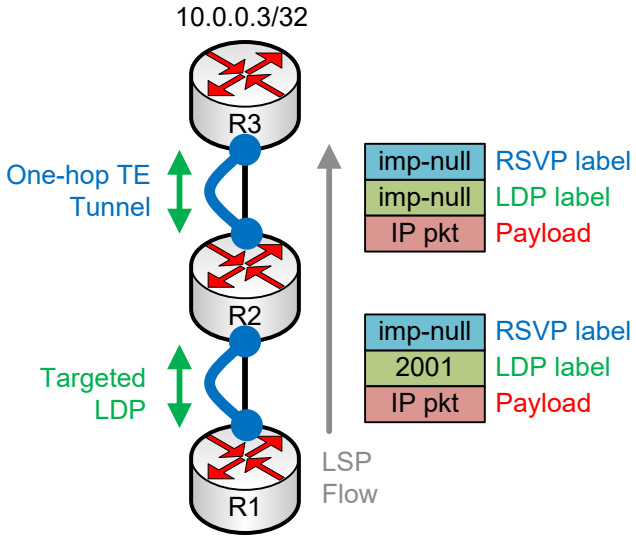
Figure 4 - RSVP Signaling for Primary One-Hop Tunnels



To overcome this, Label Distribution Protocol (LDP), defined in RFC5036, is enabled inside of each tunnel. Primary auto-tunnels do not support IP multicast traffic, such as LDP hello messages. The LDP sessions are "targeted" (sometimes called tLDP), using unicast for transport, and the LDP hello exchange is initiated by the tunnel headend. The tunnel tailend will dynamically accept the tLDP hellos, a TCP-based LDP session forms, and the routers begin exchanging label bindings as usual.

Note that the "auto-tunnel" aspect of this feature is simply a configuration convenience. The concept of a one-hop tunnel is generic. Primary one-hop tunnels can be configured manually on any vendor device that supports MPLS TE. The diagram below illustrates the how this feature works along with sample encapsulations along an LSP. R2's local label for the 10.0.0.3/32 prefix is 2001 in this example. The implicit-null labels are depicted only for illustrative purposes; there is only one MPLS label between R1 and R2 in this example. Future diagrams will omit these implicit-null depictions for clarity and technical accuracy.

Figure 5 - One-hop Primary TE Tunnels with Targeted LDP



2.2.2. Backup TE Tunnels

Primary one-hop tunnels with tLDP inside of them add no value when deployed in isolation. The data-plane behavior would be identical to regular LDP deployed directly to physical links rather than as targeted sessions inside of one-hop TE tunnels. One-hop tunnels are only useful when paired with backup tunnels.

Cisco supports the "backup auto-tunnel" feature which allows these backup tunnels to be dynamically created as they are needed. If an FRR-enabled TE tunnel traverses a router enabled for this feature, the device will build next-hop (NHOP) and next-next-hop (NNHOP) backup tunnels automatically. These protect against next-hop failures, which implies link protection, and failures of the second hop, which implies node protection.

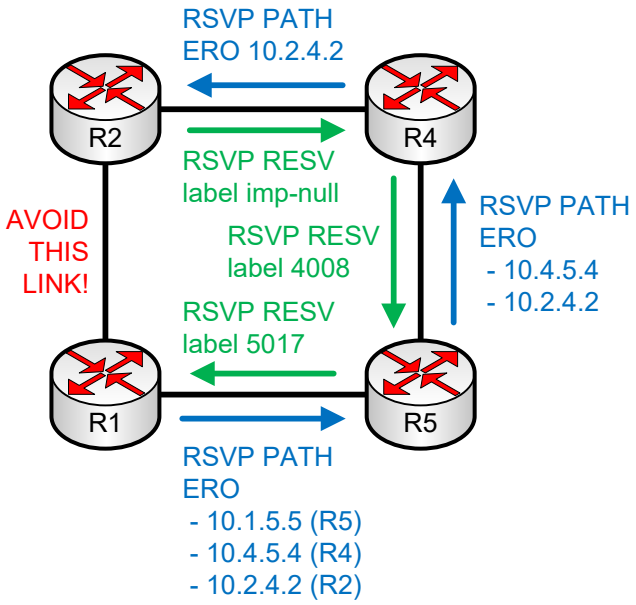
In the design discussion in this document, only NHOP (link protection) backup tunnels make sense, so NNHOP (node protection) tunnels are disabled. Because LDP is a hop-by-hop protocol that allows directly connected devices to exchange their local label mappings for a prefix, you cannot skip over entire nodes. NNHOP can therefore never be supported for LDP LSPs; this is true for unicast and multicast traffic. In our particular environment, node failures were exceedingly rare, so this trade-off was not a significant disadvantage.

When combined with one-hop tunnels, the enablement of backup tunnels guarantees that every link in the topology is backed up, provided each node has at least two links. The primary tunnels are one-hop, but the backup tunnels may traverse multiple nodes when forming a repair path. In those cases, the backup tunnel adds additional MPLS encapsulation to transport the (often short-lived) FRR traffic to the other node.

The constrained shortest path first (CSPF) algorithm using by MPLS TE dynamically determines a path through the network that avoids the link in question. If such a path is found, the headend

sends an RSVP PATH message downstream towards the tailend following the ERO generated for the path. The RSVP RESV messages flow back upstream towards the headend with label bindings at each hop. The diagram below illustrates the RSVP signaling for these tunnels.

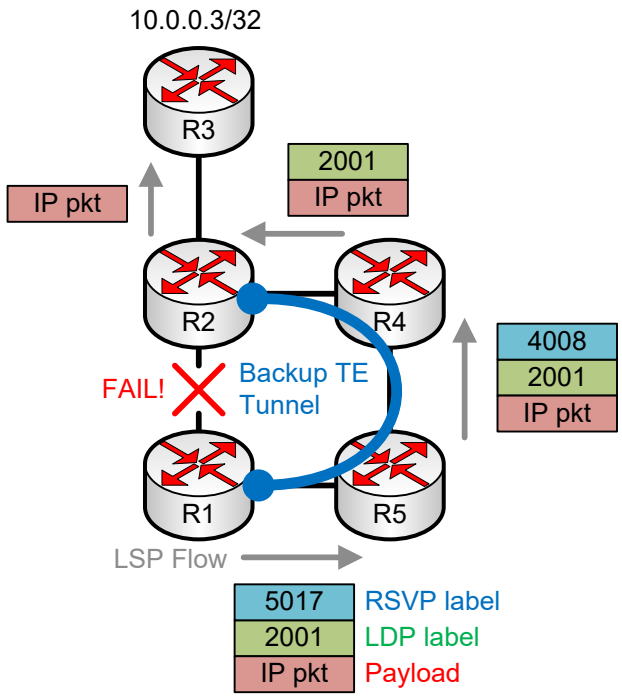
Figure 6 - RSVP Signaling for NHOP Backup Tunnels



TE backup tunnels are often active for short periods of time. When a link failure occurs in the network, the first devices to learn about it are the devices connected on that link. Assuming RSVP can detect the problem quickly (notified via layer-1 detection or BFD), the upstream router can immediately route MPLS traffic out of the failed one-hop tunnel and into the backup tunnel designed to protect that link. In modern devices, this switchover typically takes about 5 ms. The pre-built backup will take a longer path through the network, increasing latency and jitter for a short time, until the IS-IS reconverges around the failure. Once complete, a new series of one-hop tunnels will be used. The backup tunnel will no longer be used for traffic forwarding. Combining this 5 ms switchover with the 5 ms (carrier delay) or 150 ms (BFD) failure detection times, the total loss period was typically 10 ms or 155 ms. As mentioned earlier, BFD was used sparingly as most routers could rely on carrier sensing, so 10 ms was certainly the median time.

In this way, the combination of automatic primary and backup tunnels guarantees topology-independent link-protection on all network links with minimal configuration and maintenance. The diagram below illustrates the activation of a TE tunnel that protects the R1-R2 link. Traffic takes an alternative path through the network inside the TE tunnel via additional MPLS encapsulation, ultimately preserving R2's original local label for 10.0.0.3/32, which is 2001. Since R2 is the second-to-last hop, it performs penultimate hop popping (PHP) and exposes the plain IP packet to R3.

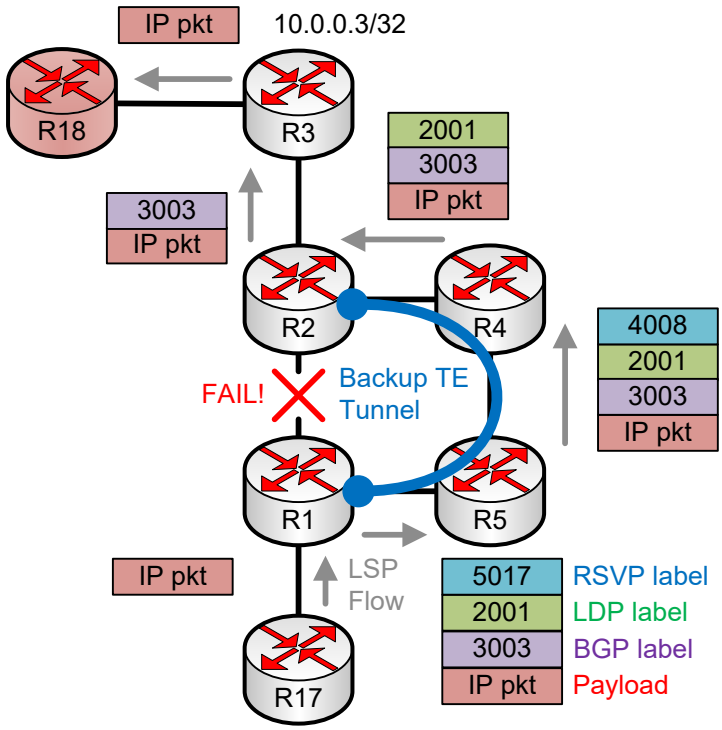
Figure 7 - Activation of TE-based NHOP Backup Tunnels for LDP LSPs



MPLS layer-3 VPN (L3VPN) services are discussed later in this document, but the diagram below illustrates how unicast LSPs between customer edge (CE) devices are transported across the core in this design. The forwarding equivalence class (FEC) from R1 to R3 in these examples remains the same, retaining the LDP and RSVP labels along the LSP. An additional BGP-allocated VPN label is added first, identifying the VPN prefix behind R3 to which the IP packet is destined. R17 through R20 represent CEs in this sample network (colored light red for clarity) and label 3003 is the end-to-end VPN label. TE-FRR is currently active, which illustrates that the maximum stack depth (MSD) in this network is 3. Multicast transported in mLDP has an MSD of 2, a TE-FRR label followed by an mLDP label, which is discussed later.

Given an MSD of 3, the MPLS Maximum Transmission Unit (MTU) should be greater than or equal to 1512 bytes on all core links. This ensures that customers can pass full 1500-byte IPv4/v6 packets through the network, assuming only MPLS L3VPN is used. For MPLS L2VPNs, the calculus changes, but this network did not support L2VPNs as there was no use-case for it.

Figure 8 - Activation of TE-based NHOP Backup Tunnels for MPLS L3VPNs



2.2.3. Label Switched Multicast (LSM) Transport

Unifying IPv4 and IPv6 service transport, along with unicast and multicast, was a high priority when designing this network. While there are various ways of transporting customer IP multicast traffic across MPLS networks, this document focuses on multicast LDP (mLDP) in-band signaling. This technology is not commonly deployed because it supports a narrow set of use cases unsuited to general-purpose customer multicast but was an excellent choice for this network. Because mLDP is an extension of LDP, it does not require any additional control-plane sessions and behaves much like a BGP address-family or negotiated capability.

All other multicast-over-MPLS technologies involve some form of overlay signaling, such as a GRE/IP tunnel carrying Protocol Independent Multicast (PIM) messages or an additional BGP address-family to signal VPN membership. These methods clearly separate underlay technologies, such as mLDP multicast delivery tree (MDT) construction, and overlay technologies, such as PIM and BGP. mLDP in-band signaling directly encodes the customer multicast VPN information into the mLDP opaque field. This field is a TLV-style value that allows mLDP to communicate different kinds of trees, VPN information, and more.

The major drawback of mLDP in-band signaling is that it only works for SSM traffic, never for ASM or bidirectional traffic. Another drawback is core state; since the core is aware of customer VPN flows, customers should be limited on how many sources they offer. A trivial denial of service (DOS) attack could consist of a client issuing endless IGMPv3 or MLDv2 membership reports containing many sources, negatively impacting the entire MPLS network. These security-related topics are addressed later using flow admission control and group scoping.

mLDP in-band signaling for multicast VPN contains four key pieces of information:

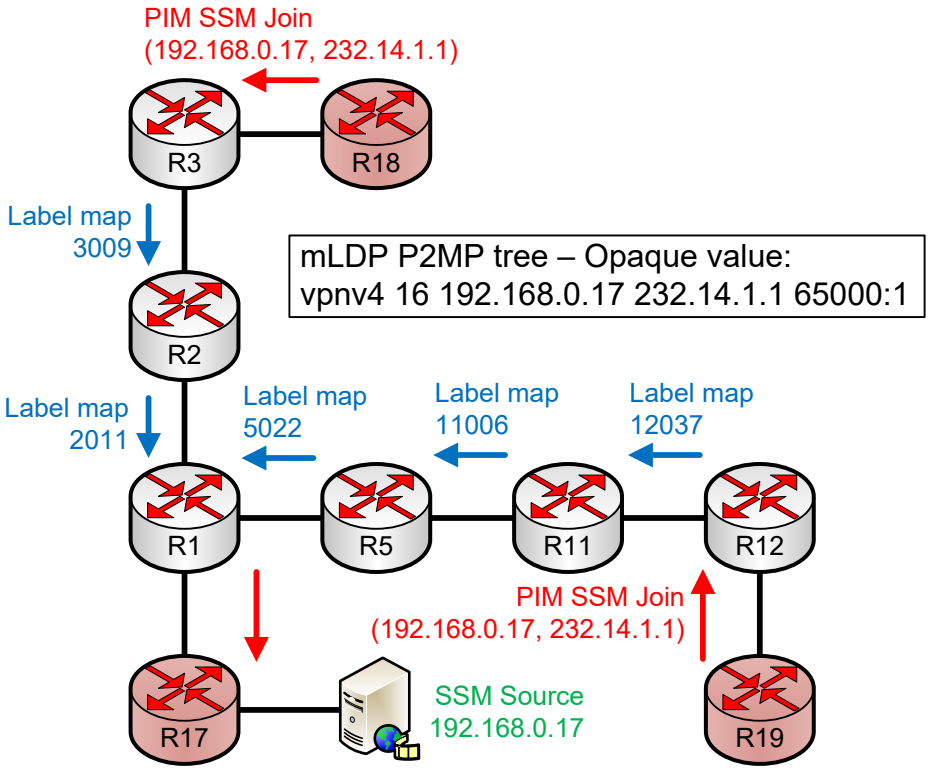
a. The 1-byte address-family identifier of 0xFA (VPNv4) or 0xFB (VPNv6)
b. The 2-byte VPN mapping length, which is 16 bytes (IPv4) or 40 bytes (IPv6)
c. The source address, which is 4 bytes (IPv4) or 16 bytes (IPv6)
d. The group address, which is 4 byte (IPv4) or 16 bytes (IPv6)
e. The BGP Route Distinguisher (RD), which is always 8 bytes

The diagram below illustrates example mLDP in-band FEC values for VPNv4 and VPNv6 (not drawn to scale).

Figure 9 - mLDP In-Band Signaling Opaque Value Format

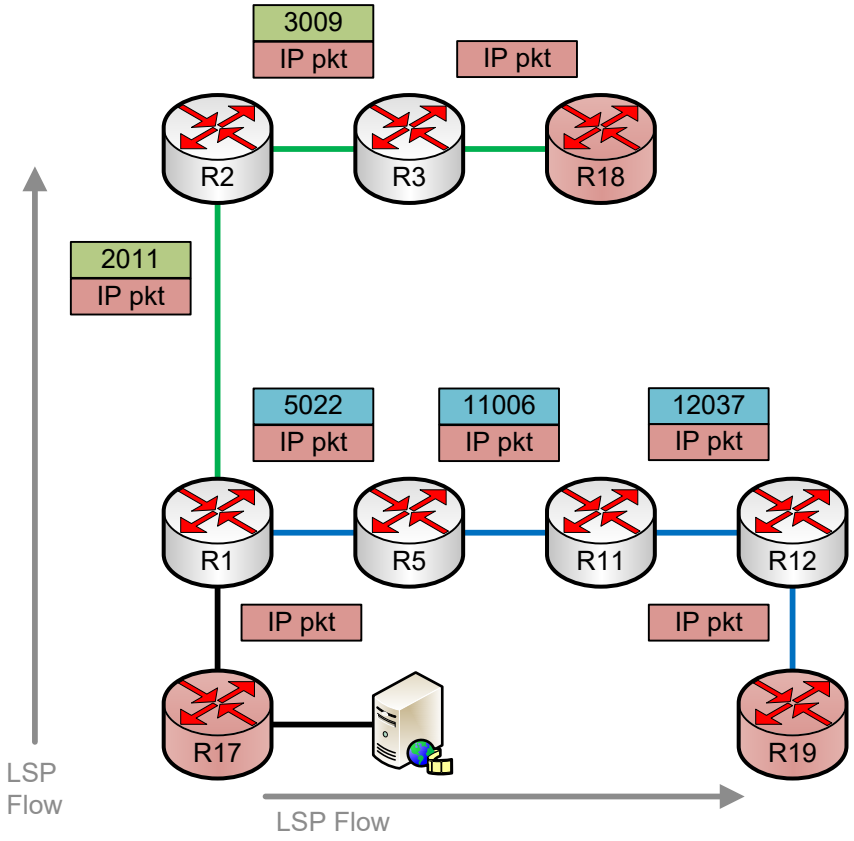
Because mLDP is directly carrying multicast mappings, no additional customer-oriented control-plane protocols are necessary. Additionally, enabling this feature does not require reforming BGP sessions with new capabilities, as is often to case when enabling more complex multicast VPN techniques. When an egress PE receives a PIM (S,G) join for an SSM group, it translates that join into an mLDP FEC mapping using the opaque format above. This message is passed upstream towards the BGP next-hop towards the source, effectively following a reverse path forwarding (RPF) process across the MPLS core towards the ingress PE. This RPF process follows the unicast routing table to reach the root of the mLDP delivery tree. The diagram below illustrates a point-to-multipoint (P2MP), downstream-only delivery tree from a single source behind R1 to two receivers behind R3 and R12. Routers R17 through R19 are customer devices and the VRF RD in which R17 is placed is 65000:1.

Figure 10 - mLDP In-Band Control-Plane Label Mapping Example



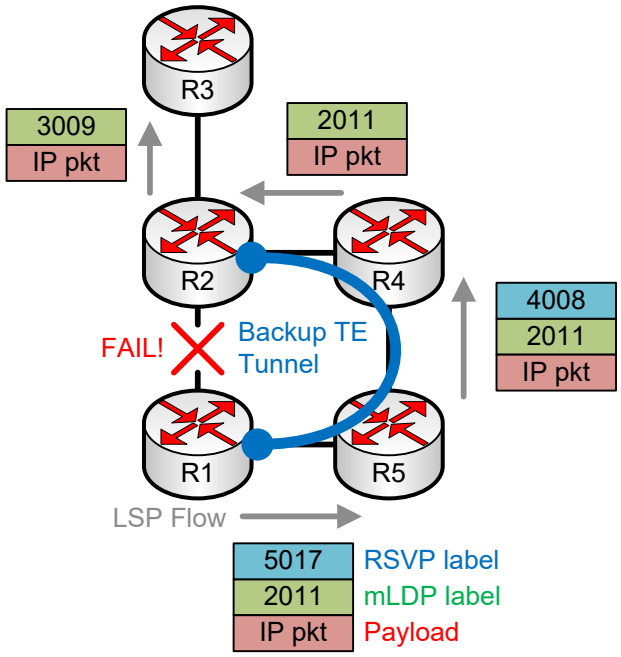
R1 will replicate multicast received from R17 towards both R2 and R5 using labels 2011 and 5022, respectively. Traffic is continuously label-switched from R1 towards the egress PEs along the mLDP-signaled LSP. Note that penultimate hop popping (PHP) is not used for mLDP LSPs; the mLDP label must be exposed to the egress PE so traffic is routed correctly. The two unique downstream replication paths are colored in green (to R2) and blue (to R12) for clarity. This matches the links in the diagram below to aid in understanding.

Figure 11 - mLDP In-Band Data-Plane Downstream Forwarding Example



For environments where all multicast flows are (or can be made into) SSM and where the number of sources is relatively small, mLDP in-band signaling is an excellent choice. Much like unicast LDP LSPs, mLDP LSPs can be protected by the same FRR mechanisms. The enablement of primary/backup tunnels, discussed in the previous section, also applies to multicast flows. Entire nodes cannot be skipped, but if a link between two nodes suffers a failure, mLDP traffic is rerouted into an NHOP backup tunnel. In accordance with design requirements, this minimizes packet loss for multicast traffic during IS-IS reconvergence while concurrently minimizing the introduction of new technologies. The diagram below illustrates such a failure and the MPLS labels involved. These tunnels are "facility" backups in that they backup an entire network resource, such as a link, and are therefore not LSP-specific. Any LSP, unicast or multicast, transiting between R1 and R2 would be backed up by this NHOP tunnel.

Figure 12 - Activation of TE-based NHOP Backup Tunnels for mLDP LSPs



2.2.4. Operations, Administration, and Maintenance (OAM)

Unlike IP-based technologies, MPLS technologies often support robust OAM methods. These techniques allow network engineers to test various LSPs and diagnose problems more rapidly. There are 3 types of LSP verification (LSPV) as defined in RFC4379 relevant in this network:

a. **IPv4 LSPV:** Used to test FECs following an IPv4 unicast transport path, typically signaled by LDP or BGP labeled-unicast. Since only LDP is deployed in this network, the FEC type is always set to LDP, and operators can quickly determine if LSPs are healthy. A simple ICMP-based ping is inadequate as this only tests IP reachability, not MPLS reachability. Both MPLS ping and MPLS traceroute tests are supported.
b. **mLDP LSPV:** Used to test FECs following an mLDP-signaled transport path. MPLS traceroute is not supported, but MPLS pings are multicast and will solicit replies from many receivers. More specifically for mLDP P2MP trees, the ping is initiated from the root (ingress PE) and targets the mLDP opaque value to identify a FEC. Every egress PE terminating the LSP will respond, allowing operators to quickly discover any problems.
c. **RSVP-TE LSPV**: MPLS-TE is used for primary one-hop tunnels between devices and for link-protecting backup tunnels. For primary tunnels, MPLS ping can ensure individual links are MPLS enabled, even if the encapsulation is implicit-null, which rules out any steady-state forwarding problems related to label binding. For backup tunnels, MPLS ping can measure the health of various backup tunnels, ensuring they are operational even when not active. MPLS traceroute can discover the backup tunnel's path before it is used, which may help estimate incurred latency when the tunnel is activated.

Note that other forms of MPLS OAM exist, such as for Segment Routing (SR), pseudowires, and Transport Profile (TP). These are irrelevant for this design and are omitted from this document.

One of SR's biggest benefits is that it obviates the need for protocols like LDP and RSVP-TE by performing a comparable function using IGP extensions. However, LDP provides LSM support via mLDP as previously explained. Multicast transport techniques over SR is still immature and not widely supported, making it a poor choice for this network at the time it was designed.

2.3. MPLS Edge Design

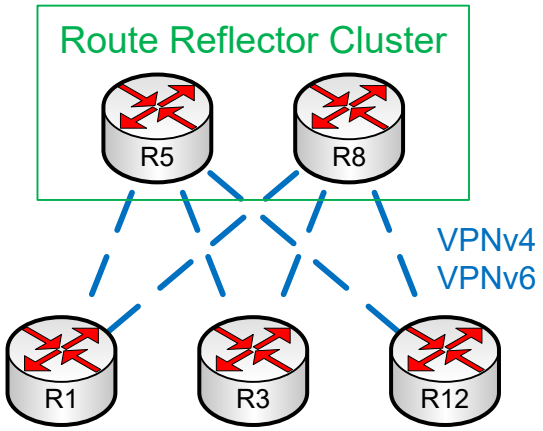
This section details how MPLS services are offered to customers in the network, which relies primary on BGP. This is not the major focus of this document. For a more detailed discussion on MPLS service design considerations, please read my whitepaper titled "Global MPLS Design Using Carrier Supporting Carrier (CSC)".

2.3.1. BGP VPN Topology

This network deployed two geographically separated route-reflectors to increase availability. R5 and R8 performed this role for both the VPNv4 and VPNv6 address-families. All PEs in the US-based network peer to both RRs, enabling any-to-any PE connectivity in the future. Thanks to mLDP in-band signaling, no additional BGP address-families are necessary, such as IPv4 MDT, IPv4 MVPN, or IPv6 MVPN. This simplifies the overall design and reduces the level of expertise required for operators maintaining the network.

A BGP route-reflection "cluster" is defined as the set of RRs that service the same set of clients. Because R5 and R8 service the same PEs, R5 and R8 are part of the same cluster, regardless of whether they have the same cluster ID or not. In most designs, it is not necessary or beneficial to peer RRs that exist in the same cluster. Some exceptions apply, such as when the RR is also a PE or when the BGP process on a given device is considered unstable. An unnecessary intra-RR-cluster peering causes the same routes to be reflected between PEs with little availability benefit. The iBGP sessions between PEs and RRs will remain up so long as IP reachability exists, so the dense core topology can tolerate multiple transit link and node failures. The diagram below depicts the BGP VPN topology, and only a subset of route-reflector clients is shown for brevity.

Figure 13 - BGP VPNv4/v6 Route Reflection



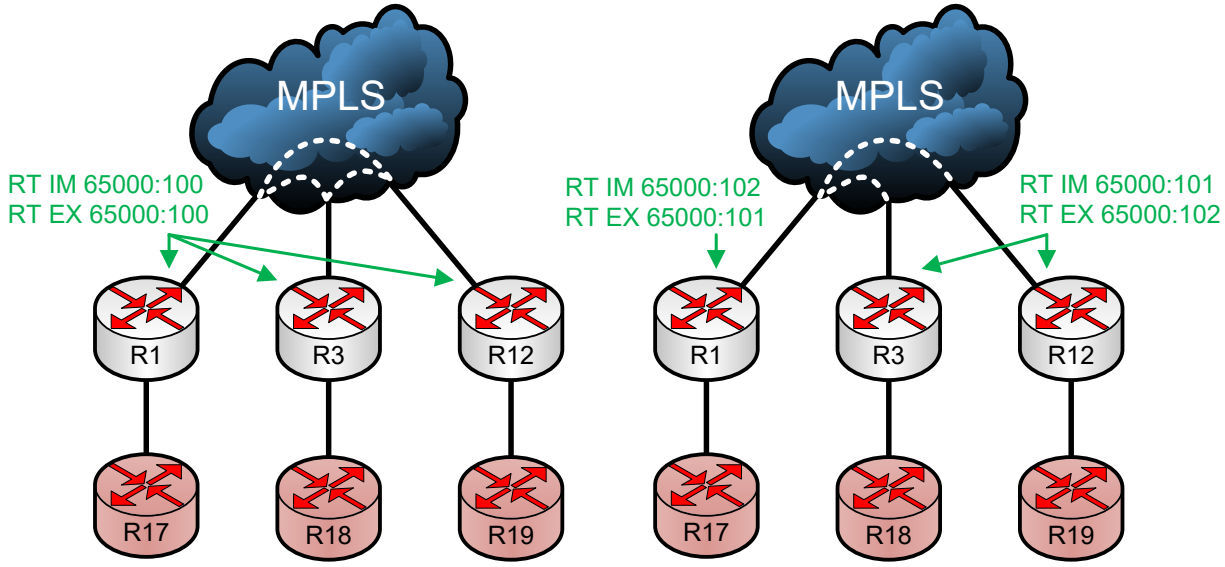
2.3.2. Layer-3 VPN Unicast Service

The network offers relatively basic MPLS layer-3 VPN services between any pair of PEs in the network. VPN membership is defined by Route Targets (RTs), which are BGP extended communities. For routes exported from a local PE's VPN Routing and Forwarding (VRF) into BGP, the exported RTs are appended to the route. To import routes into a local PE's VRF from BGP, a route must be carrying at least one RT that is imported by the VRF. The two most common designs are as follows, although any arbitrary topology can technically be built:

a. Any-to-any connectivity, whereby every PE servicing a specific customer/tenant will import and export the same RT. This is the simplest and most widely deployed.
b. Hub-spoke connectivity, whereby the hub sites will export RT 1 and import RT 2 while the spokes export RT 2 and import RT 1. This is particularly useful for centralized/extranet services, which is discussed more later.

The diagram below illustrates some examples of these VPNs along with their route targets. The topology on the left shows an any-to-any L3VPN and the topology on the right depicts a hub-spoke L3VPN where R1 is the hub.

Figure 14 - Example L3VPN Topologies; Any-to-any and Hub-spoke



2.3.3. Layer-3 VPN Multicast Service

First, we must identify the group addresses to be used. The approach is similar between IPv4 and IPv6 but varies due to incongruent feature sets between the two IP versions. To better unify the designs, the IPv4 multicast groups use the following format:

a. First octet is 232 to signify SSM traffic per RFC4607
b. Second octet is the scope, which follows RFC7346. This is an IPv6 concept but has been implemented in IPv4 for consistency and is a 4-bit number (0 to 15)
c. Third octet is type of traffic to simplify QoS and security classification (see table)
d. Fourth octet is the group ID, which is application-specific and is ignored by the network

The table below maps the third octet values to the types of traffic those values represent.

Table 1 - Multicast Traffic Types

Third Octet	Traffic Type
0	Transactional Data (chat, messaging, location beacons, etc.)
1	Standard Definition (SD) Video, typically 480 vertical pixels (480p) or less
2	High Definition (HD) Video, typically 720p or 1080p
3	Ultra-High Definition (UHD) Video, typically 2160p or greater

The table below enumerates some example multicast groups by combining all 4 octets together. While there are 16 possible scopes, only a subset has been formally defined. This network uses a non-standard, Cisco-defined value of 14 to indicate traffic that is not quite globally scoped but is broader than a single organization. In Cisco parlance, this is known as "VPN" scope, which is appropriate for inter-AS multicast traffic. Recall that the second octet represents the multicast group's scope.

Table 2 - Example IPv4 Multicast Groups with Scopes and Types

Group IP	Explanation
232.5.1.9	SD video with site-local scope. Limited to a customer site only.
232.8.2.9	HD video with organizational-local scope. Limited to single carrier only.
232.14.3.9	UHD video with "VPN" (unofficial) scope. Can traverse across carriers.
232.15.0.9	Transactional data with global scope. Can traverse the Internet, if ever supported.

The scopes and traffic types just described also apply to IPv6. The concept of scoping is intrinsic to IPv6 and has a dedicated 4-bit field (bits 13-16) in the group address, but the traffic type must be encoded manually. The table below shows examples of IPv6 multicast groups. Note that the first 12 bits of "ff3" are constant. The "ff" indicates IPv6 multicast and the "3" is conventionally used to indicate SSM traffic (P and T bits set to 1 per RFC4607). The full SSM range is ff3x::/96, allowing the last 32 bits for group allocation and where "x" is the 4-bit scope value. In this particular design, the last 16 bits carry the group ID while the second-to-last 16 bits carries the traffic type. 16 bits (4 bytes) is a natural boundary in IPv6 as shown by the examples below.

Table 3 - Example IPv6 Multicast Groups with Scopes and Types

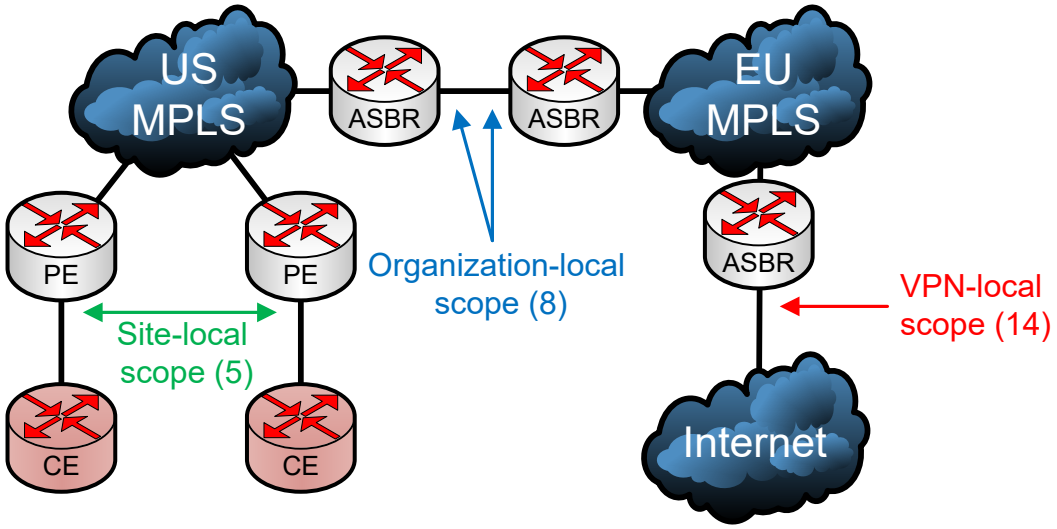
Group IP	Explanation

ff35::1:9	SD video with site-local scope. Limited to a customer site only.
ff38::2:9	HD video with organizational-local scope. Limited to single carrier only.
ff3d::3:9	UHD video with "VPN" (unofficial) scope. Can traverse across carriers.
ff3e::9	Transactional data with global scope. Can traverse the Internet, if ever supported.

The assignment of scopes can be used to enforce multicast boundaries. Feature sets vary by platform, but in Cisco IOS-based devices, IPv4 and IPv6 use different methods. In IPv4, an access control list (ACL) boundary is applied, matching sources, groups, or both. In IPv6, the ACL method is not available, but the numeric scope can be configured instead. Numerically, the scopes are an inclusive lower-bound on multicast traffic transiting an interface. A scope of N means that scopes greater than N are permitted while scopes less than or equal to N are denied.

At a high-level, the topology below illustrates the scoping boundaries irrespective of IP version or exact configuration. On PE-CE links connecting directly to customers, traffic is site-local scoped, allowing organizational-scoped or greater to transit. This ensures traffic local to a given customer site is not accidentally transported over MPLS, which would waste bandwidth, PIM (S,G) state in the VRF, and mLDP in-band state in the core network. On inter-AS links, traffic is organizational-scoped, allowing VPN-scoped or greater to transit. This enables customers to constrain some multicast traffic to only their US-based sites, reducing inter-continental traffic. To expand past the MPLS networks, multicast traffic must be scoped at the global level to cross the VPN-local scoped boundary at the Internet edge. These scoping boundaries also improve security by reducing the attack surface for state-exhaustion denial of service (DOS) attacks.

Figure 15 - Multicast Scoping Boundaries



As mentioned in the architecture overview, not all clients in the network supported IGMPv3 for IPv4 and MLDv2 for IPv6. Some of these clients were legacy systems that could not be upgraded. To ensure that all multicast traffic transiting the network was SSM-based, last-hop PIM routers need to map ASM-based IGMPv2 and MLDv1 membership reports into SSM

entries somehow. Cisco IOS can statically map multicast group ranges to a list of sources on the last-hop router, allowing that device to issue PIM SSM (S,G) join messages up the reverse path towards the source, even if the client did not specify the sources.

This static mapping solution has several major drawbacks for MPLS L3VPN service providers:

1. It scales poorly from a management-plane perspective as each PE must maintain complex group-range to source-list mappings. While network automation solutions (sometimes called "infrastructure as code") can simplify the daily management of these mappings, it requires an entirely new set of skills, training, procedures, and possibly new investments.
2. It only works on last-hop routers. In most MPLS L3VPN environments, the PE and CE exchange routes using a PE-CE routing protocol; this is commonly eBGP. This implies that the receivers are not directly connected to the PE and therefore do not exchange IGMP or MLD signaling with the PE. It then becomes a router within the customer's network that is responsible for this mapping, which is almost impossible to coordinate and manage at scale across different organizations.
3. For testing and discovery purposes, a customer has no way to determine which source addresses are mapped to which group ranges without a deeper network analysis ("show" commands on a device, packet captures, etc.) The customer's inability to examine the carrier's multicast control-plane will complicate troubleshooting and daily operations.

To overcome all these issues, the last-hop routers are configured to use DNS instead. It is not realistic to require individual customers to operate their own DNS servers to perform this mapping function. Instead, the carrier offers a centralized pair of DNS servers that have identical configurations. These are accessible to all customers in all multicast-enabled VPNs using the "centralized services" design of a hub-spoke L3VPN topology. Customer last-hop routers can target these servers for their DNS mappings which comes at no additional monetary cost and requires only minimal technical skills/training. All multicast-enabled VRFs will import the "DNS downstream" route-target and export the "DNS upstream" route-target. The PEs servicing the data centers where the DNS servers are hosted will perform the opposite action, creating the hub-spoke extranet.

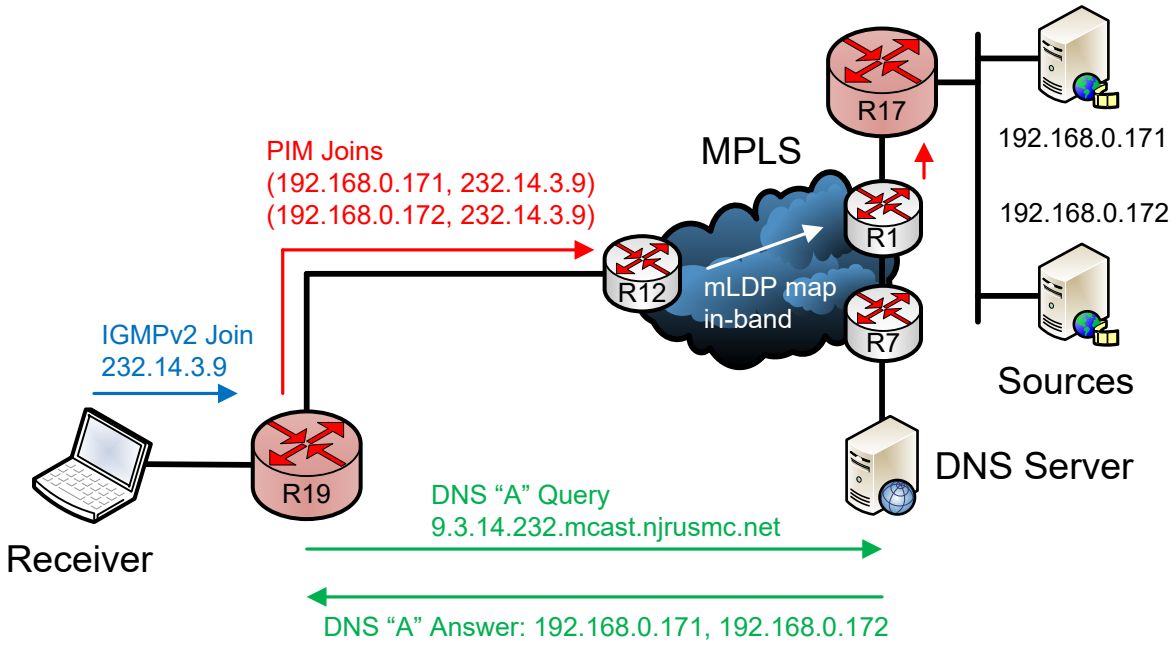
Consider IPv4 first. Upon receiving an IGMPv2 membership report for a group, the last-hop customer router consults the DNS server by sending an "A" query. The format of the hostname being resolved is as follows:

1. The multicast group in reverse octet order. For example, group 232.14.3.9 would be encoded as 9.3.14.232 at the beginning of the hostname
2. The multicast-specific domain name if one exists. If not configured, Cisco appends the string "in-addr.arpa" by default
3. The general-purpose domain name, which is appended last

As a complete example, a group of 232.14.3.9 with a multicast-specific domain of "mcast" and a general domain name of "njrusmc.net" would result in a hostname of "9.3.14.232.mcast.njrusmc.net" carried in the DNS query. This is the "A" record configured on the DNS server, and the mapped value must be at least one IPv4 unicast address. These addresses represent the sources that are sending traffic to group 232.14.3.9 as illustrated below. If multiple sources are returned, the router will issue PIM (S,G) joins for each source. This is a useful

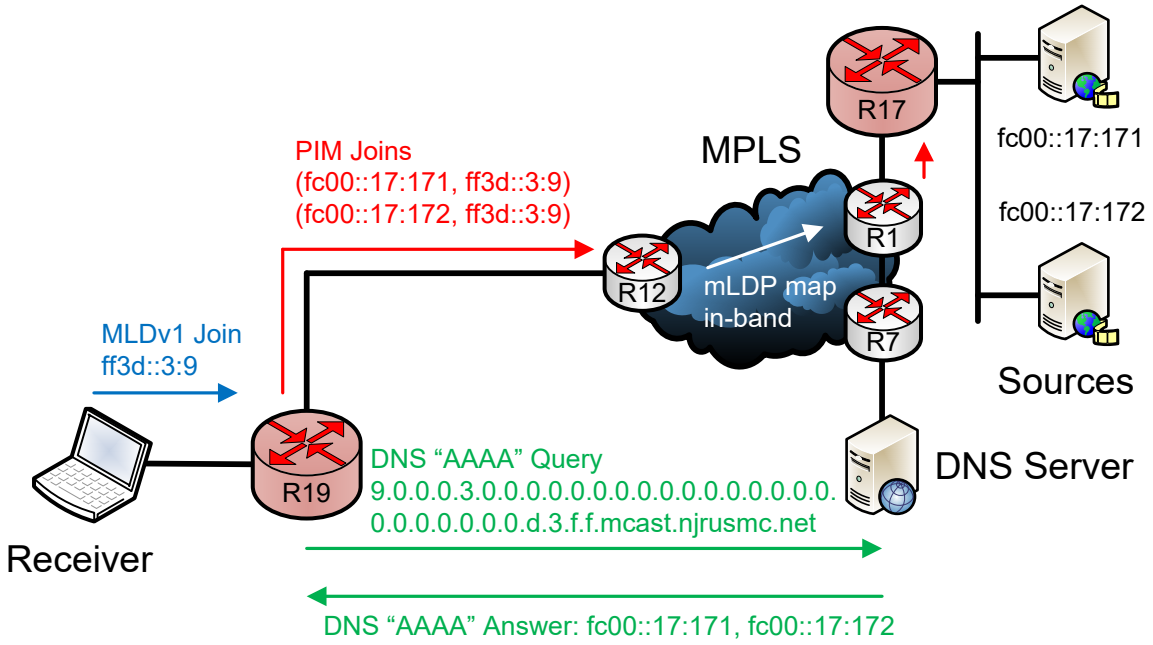
technique to build highly available multicast delivery systems, assuming the sources are delivering identical content.

Figure 16 - Mapping IGMP ASM Groups to IPv4 Sources



The process is similar for IPv6 except the encoding of the group address is different. In IPv4, each octet (8 bits) is kept intact but displayed in reverse order. In IPv6, only the hexadecimal digits (4 bits, sometimes called "nibbles") are kept intact and are also displayed in reverse order, separated by periods. For example, the IPv6 multicast group of ff3d::3:9 would be encoded as "9.0.0.0.3.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.d.3.f.f.mcast.njrusmc.net" using the domain names from earlier. This is the "AAAA" record configured on the DNS server, and like the IPv4 mappings, the values are IPv6 multicast sources for this group. The diagram below illustrates this process, and once again, operators can specify multiple multicast sources for high availability.

Figure 17 - Mapping MLD ASM Groups to IPv6 Sources



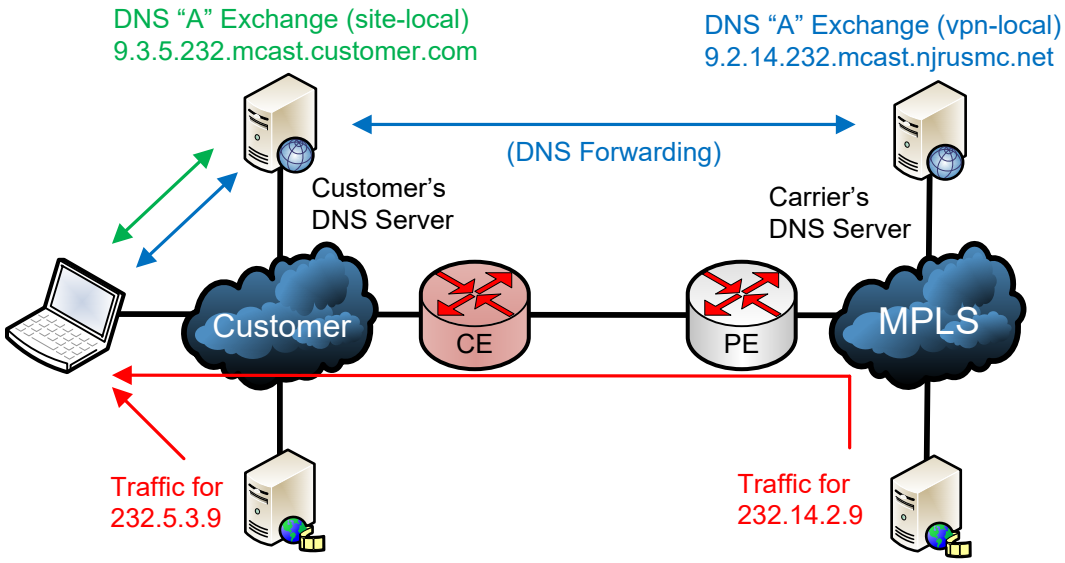
Some customers may choose to run their own DNS servers, and in so doing, will configure their last-hop routers to target those servers. The main drawback is the management burden of maintaining a functioning DNS domain, but some customers were already managing DNS anyway. There are two key advantages to this approach:

1. Minimal changes on the last-hop router; just enable IGMP/MLD SSM mapping and optionally configure a multicast domain name. Retain existing DNS domain name and server configurations.
2. Enables overlapping (site-local scoped) group resolution on a per-site or per-customer basis. Because multicast traffic is scoped to a site, the groups covered by that range cannot transit the MPLS network, and thus can be recycled within each customer's network and DNS domain, much like RFC1918 or IPv6 ULA addressing. This cannot be achieved using a centralized, network-wide DNS service. Customer DNS servers can reference the upstream carrier-hosted DNS servers if they are unable to resolve a specific mapping, creating a hierarchical DNS architecture.

The diagram below illustrates this architecture, which requires no special configuration in the carrier network and enables customers to gain additional control over their DNS mappings. A site-local group of 232.5.3.9, shown in green, can be resolved at the customer-managed DNS server since the source is local to that site. A VPN-local group of 232.14.2.9, shown in blue, probably requires resolution at the carrier-managed DNS server for transit between sites (or between continents). The customer's DNS server would be configured to forward unresolvable requests for the "njrusmc.net" domain to the carrier's DNS server. This hierarchical DNS architecture isn't anything new; recycling such a common and well-known technique makes a more robust multicast design.

Although not depicted, also note that modern clients running IGMPv3 or MLDv2 can simply use DNS to resolve multicast sources on their own. Be sure to add the proper "A" and "AAAA" records for those sources which are unrelated to the network-based mappings just described.

Figure 18 - Hierarchical DNS For Multicast Source Resolution



2.4. QoS Design

This section describes the Quality of Service (QoS) design within the organization. Because this network was primarily built to transport multicast traffic, the QoS policies are focused on differentiating between different multicast flows. For brevity, only IPv4 multicast groups are depicted as the concepts in IPv6 relating to scope and prioritization are identical.

2.4.1. MPLS Edge Ingress EXP Mapping

On all customer facing links, PEs must set an MPLS experimental (EXP) value on each MPLS-encapsulated packet on ingress. This enables the MPLS network to apply per-hop behaviors (PHBs) to individual packets based on their traffic type. As discussed previously, there are four types of multicast flows in the network. These are easily categorized by their multicast group addresses, simplifying the QoS classification process. The table below illustrates the group-to-EXP mappings applied on all ingress PEs. These mappings also apply equally to IPv4 and IPv6, and the "x" in the multicast group represents the scope value described earlier.

Table 4 - Ingress MPLS EXP Mappings

Traffic Type	Match Criteria	Imposed MPLS EXP
Multicast UHD Video	Groups 224.x.3.0/24	EXP 5

Multicast HD Video	Groups 224.x.2.0/24	EXP 4
Multicast SD Video	Groups 224.x.1.0/24	EXP 3
Multicast Transactional Data	Groups 224.x.0.0/24	EXP 2
All other traffic, including all unicast	Any	EXP 0

Note that any existing Differentiated Services Code Point (DSCP) value applied to any IP packet transiting the network is ignored. DSCP is never used for any classification decision; most customers in this environment used non-standard DSCP values that the carrier preferred to ignore. The DSCP values are retained from end-to-end so as not to inconvenience the customers. Note that unicast voice and voice signaling traffic does not receive any explicit treatment in this network. It was not a critical application and was present only in small quantities, making it difficult to justify special attention within the QoS design.

Other ingress QoS features, such as policing and remarking, are largely unnecessary in this design. Multicast state admission control techniques (discussed later) help manage the flow of traffic without needing data-plane traffic conditioners on ingress.

2.4.2. MPLS Core Queuing

Assuming customer traffic has been properly marked on ingress, the MPLS core devices should apply the proper treatment on individual MPLS packets. Because unicast and multicast traffic are both label-switched, DSCP can be largely ignored with one exception. Network control traffic between devices, such as LDP, RSVP, and BGP, will not be MPLS-encapsulated if the destination is one-hop away. To cover this case, DSCP CS6 is matched in addition to EXP 6 for network control. The table below illustrates the queuing policy and bandwidth allocations.

Table 5 - MPLS Core Queuing Policy

Traffic Type	**Match Criteria**	**Bandwidth Percentage**
Network Control	EXP 6 or DSCP CS6	2% BW reserve
Multicast UHD Video	EXP 5	40% BW reserve
Multicast HD Video	EXP 4	20% BW reserve
Multicast SD Video	EXP 3	10% BW reserve
Multicast Transactional Data	EXP 2	5% BW reserve + WRED
All other traffic, including all unicast	EXP 0	23% BW reserve + WRED

Weighted Random Early Detection (WRED) is an Active Queue Management (AQM) technique that preemptively discards packets to prevent queues from overflowing. This is only useful for

elastic flows that respond to packet loss by reducing their rate of transmission; all TCP-based applications behave this way. UDP-based applications may respond similarly, but it depends on the specific application behavior.

In this network, most of the transactional data applications responded to the receipt of a multicast message with a unicast reply to the sender. If the sender does not receive an acknowledgement, it retries its transmission following an exponential back-off algorithm. In this way, transactional data is elastic, as is most of the best effort traffic. WRED is therefore enabled for classes matching EXP 2 or EXP 0. The multicast video streams, regardless of quality, behave as "broadcast video" given their inelastic, unidirectional nature. WRED is not an effective tool to manage congestion for such traffic and is omitted from the multicast video queues.

As a final comment, notice that EXP 1 is not used in this design. This value is often used for scavenger traffic (explicitly low priority) and could be used in the future to deliberately mark-down unimportant flows. Since no such mark-down mechanism was required in this carrier network, a scavenger queue was omitted to avoid "gold-plating", but a future implementation remains possible. Had the design used EXP 1 for transactional data, as an example, implementing a scavenger queue would have been sloppier as a different, less standardized EXP value would be used instead.

2.4.3. MPLS Edge Queuing and Admission Control

Traffic egressing from the carrier towards customers must also receive the proper QoS treatment. Since this traffic lacks MPLS encapsulation, the carrier cannot match based on MPLS EXP. Because customer DSCP must be retained, there are two suitable MPLS QoS design options:

a. **Short-pipe mode:** Perform egress queuing based on egress IP characteristics, such as source, destination, DSCP, or layer-4 port information
b. **Long-pipe mode:** Perform egress queuing based on the topmost label's EXP value of the received MPLS packet from the core

The advantage of long-pipe over short-pipe is that it avoids a second round of classification. It would require enabling LDP explicit-null mappings for PE loopbacks, adding 4 bytes of encapsulation to every packet at the penultimate hop, along with a second MPLS lookup in the LFIB. Short-pipe was chosen to avoid these minor inconveniences, and since all the classification constructs already existed on every PE, they were easily recycled between ingress classification and egress edge queuing policies. Therefore, the carrier will match based on multicast group, ignoring customer DSCP.

The only exception is that DSCP CS6 is used to match network control traffic, such as BGP. This allows the classification-related constructs (in Cisco parlance, the access-lists and class-maps) to be recycled from the ingress edge policy, requiring only a small addition to handle network control. The table below describes the edge queuing policy and uses identical bandwidth and WRED configurations as the core queuing policy. Only the match criteria are different.

Table 6 - MPLS Edge Queuing Policy

Traffic Type	Match Criteria	Bandwidth Percentage

Network Control	DSCP CS6	2% BW reserve
Multicast UHD Video	Groups 224.x.3.0/24	40% BW reserve
Multicast HD Video	Groups 224.x.2.0/24	20% BW reserve
Multicast SD Video	Groups 224.x.1.0/24	10% BW reserve
Multicast Transactional Data	Groups 224.x.0.0/24	5% BW reserve + WRED
All other traffic, including all unicast	Any	23% BW reserve + WRED

To further strengthen this QoS policy while simultaneously improving security, the carrier deployed multicast state admission control at the network edge. The first step is identifying how much bandwidth each flow of a given type requires. The word "flow" in this context refers to an (S,G) state entry installed in the egress direction on an interface. The various multicast video types consume different levels of bandwidth given their varying levels of quality. Each time a single (S,G) entry is installed in the table, the "cost" of each flow, measured in kbps, is subtracted from the associated allowance configured on the interface for that given video class.

Because multicast transport was the biggest driver of this network design, 75% of the interface bandwidth on every egress PE was broadly allocated for multicast traffic. The allocation was evenly divided between IPv4 and IPv6 flows as well. Consider the following example of costs and allowances on a 1 Gbps interface, which was the most deployed PE-CE interface speed.

There are 750 Mbps available for multicast traffic to be evenly divided between IPv4 and IPv6, resulting in 375 Mbps for each version of IP. The following per-IP-version plan was developed:

a. Allow 4 flows of UHD at 50 Mbps each for a total of 200 Mbps
b. Allow 10 flows of HD at 10 Mbps each for a total of 100 Mbps
c. Allow 10 flows of SD at 5 Mbps each for a total of 50 Mbps
d. Allow 25 flows of transactional data at 1 Mbps each for a total of 25 Mbps

Putting this data into tabular form, costs can be assigned to each flow, measured in kbps. This will allow each flow to be weighed against a maximum state limit per video type. The columnar total of all "Total BW" numbers equals 750,000 kbps or 750 Mbps. Viewed another way, the "BW/Flow" column represents the "cost" of each flow, and the "Total BW" column represents the per-type allowance configured on the interface.

Table 7 - Flow Admission Control Allocations

IP Version	Flow Type	# Flows	BW/flow (kbps)	Total BW (kbps)
IPv4	UHD Video	4	50,000	200,000
IPv4	HD Video	10	10,000	100,000
IPv4	SD Video	10	5,000	50,000

IPv4	Transactional Data	25	1,000	25,000
IPv6	UHD Video	4	50,000	200,000
IPv6	HD Video	10	10,000	100,000
IPv6	SD Video	10	5,000	50,000
IPv6	Transactional Data	25	1,000	25,000

This solution has many advantages. First, the combination of control-plane admission control and data-plane egress queuing will minimize packet loss. Congestion is highly unlikely and would likely be caused by best-effort unicast traffic consuming more than its share of bandwidth. In those cases, congestion management (queuing) and congestion avoidance (WRED) are applied to the unicast traffic, making room for the multicast traffic in accordance with the policy.

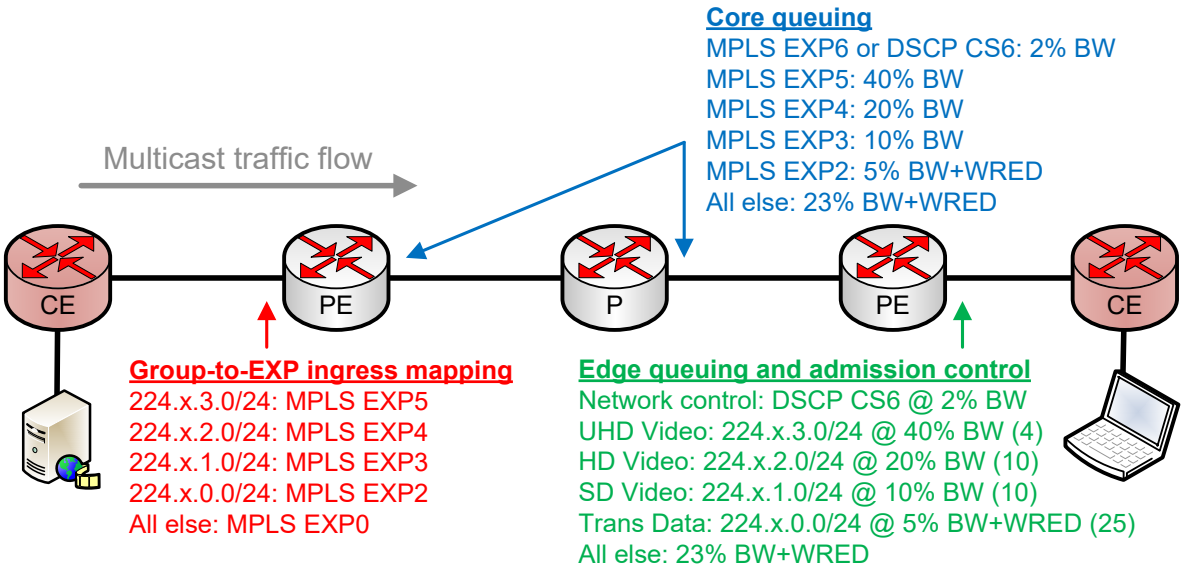
Second, implementing such a configuration allows for future flexibility. Suppose IPv6 continues to grow in relevance and popularity. Perhaps carving the 750 Mbps allocation for multicast into unequal portions, such as 600 Mbps for IPv6 and 150 Mbps for IPv4 (4:1 ratio), becomes the best business decision. Adjusting the allowances in a known-good, pre-built policy is easier than being pressured to re-engineer such a policy from scratch in an already-operational network.

Third, consider the user experience perspective. Suppose Alice prefers a UHD video stream that no one else at her site is currently watching. The UHD allowance has been exhausted because her colleague Bob is already watching 4 separate UHD feeds concurrently, the maximum allocated number. When the PIM (S,G) join message associated with Alice's group membership report reaches the egress PE, the join is rejected due to multicast state admission control. It is not installed into the multicast routing table and the egress PE does not map the PIM (S,G) join message into an mLDP mapping for the corresponding (S,G, RD) VPNv4/v6 opaque value. Alice will likely face a blank screen for a few seconds (or perhaps an error message depending on the application), then try the next best option, which is HD. This is a better outcome than Alice joining the UHD group and degrading all 5 of the UHD streams due to potential QoS congestion.

Last, the solution outlined above is very conservative. The quantity of flows permitted via admission control is mapped precisely to the bandwidth allocations in the queuing policies. This disables oversubscription and trades off "quantity" in favor of "quality". In bandwidth-constrained environments where additional multicast consumption is required, admission control limits can be raised. It's unlikely that a mix of precisely 4 UHD, 10 HD, and 10 SD flows exist concurrently, which implies some likelihood of unused bandwidth. It's important to hypothesize, test, and analyze the results of any oversubscription experimentation to find the optimal balance.

The diagram below summarizes the QoS strategy for the carrier. It includes the ingress group-to-EXP mapping, core queuing, and edge queuing with admission control.

Figure 19 - End-to-End Carrier QoS Design



3. Inter-Continental Network Design

This section discusses the European network and its integration with the US-based network.

3.1. European Network Summary

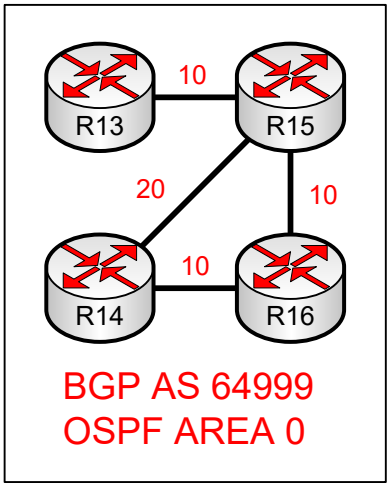
The European network was not a major focus of this design effort because it already existed while the US network was newly built. It uses many divergent technologies and is worth a brief discussion, nonetheless.

3.1.1. Unicast Routing and Forwarding Design

In stark contrast to the US-based MPLS network, the European MPLS network is based on Open Shortest Path First (OSPF) instead of IS-IS. All the links are point-to-point, but the link costs were fixed to the same value, with some exceptions for long-distance links. All the European PEs were fully meshed using MPLS TE tunnels; the network did not use LDP in any capacity. This provided maximum control and path optimization in the European theater at the cost of additional management complexity and RSVP-TE core state. As mentioned earlier, the European network was older than the newly designed US network and thus did not consider Segment Routing (SR) as an MPLS transport technology. Like the US network, TE-FRR was used to protect all links in the network, and where possible, node protection was also enabled.

Europe offers comparable MPLS L3VPN services as the US with a similar route-reflector design. Other edge technologies, such as QoS, flow admission control, and security techniques were applied in the European network once the integration with the US network was complete. These topics are not relevant to the inter-AS integration and are omitted for brevity. The diagram below illustrates the basic OSPF/BGP design with OSPF costs included.

Figure 20 - European Network OSPF Costs



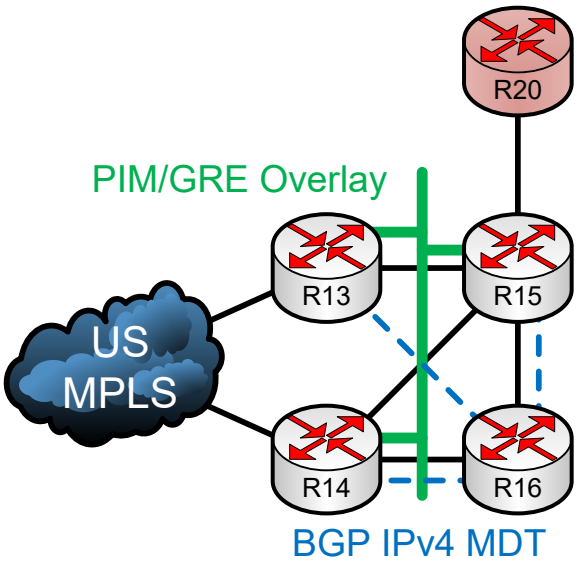
3.1.2. Multicast Routing and Forwarding Design

At the time of the European network's creation, only the "Draft Rosen" style of multicast VPN was widely available, defined in RFC6037. In summary, this solution tunnels customer multicast traffic inside of an IP-based Generic Routing Encapsulation (GRE) tunnel between PEs. This tunnel allows the PEs to exchange two kinds of traffic:

1. Control-plane overlay signaling, which are VRF-aware PIM messages between the PEs. The GRE tunnel is effectively an emulated LAN, providing any-to-any connectivity between PEs and behavior much like a virtual switch. The tunnel destination is a multicast address, avoiding headend replication to reduce bandwidth usage. Because PIM is supported from end-to-end, both SSM and ASM are supported for customers.
2. Data-plane traffic transport between customer sites. After the delivery trees are built across the VPN, customer traffic flows across, encapsulated in GRE packets with multicast destinations. The exact mechanics on precisely which multicast addresses are used as destinations for which flows is irrelevant to the design and is omitted for brevity.

This solution requires that the MPLS core be PIM-enabled. Most designs use PIM SSM in the core combined with the BGP Multicast Delivery Tree (MDT) IPv4 address-family to distribute PE loopbacks. Using BGP to distribute these loopbacks, which are the sources of the GRE tunnel, obviates the need for any ASM RP deployments in the core. When all the PEs know about one another's loopbacks, they can issue PIM SSM (S,G) joins towards each loopback (source) targeting the MDT group address. The diagram below illustrates the high-level design and R20 represents a CE within a VRF. The AS boundary routers (ASBRs) are also running VRFs, which is discussed more later.

Figure 21 - Draft Rosen IP/GRE MDT Design



This solution is very different than mLDP in-band signaling even though it solves a similar problem. The table below compares these two technologies. Some of these topics have been discussed earlier in different contexts and are summarized again.

Figure 22 - Comparing Draft Rosen IP/GRE MDT with mLDP In-Band Signaling

	Draft Rosen IP/GRE MDT	**mLDP in-band signaling**
Scalability	Medium; low core state but full mesh of PEs in overlay	Low; expansive core state when (S, G, RD) tuples are numerous
Flexibility	High; supports PIM ASM, SSM, bidirectional mode, and BSR	Low; only supports PIM SSM
Divergent Tech	High; separate underlay/overlay PIM topologies/groups, MTU issues	Low; same transport as unicast LSPs, no new MTU concerns
Protection	Low; multicast only FRR (MoFRR) only; duplicative and often wasteful	High; general purpose MPLS TE link (NHOP) protection
Configuration	Long/complex; additional BGP AFI or ASM RP design, new group allocations for core and per VRF/AFI	Short/easy; generic enablement per VRF/AFI, no special allocations between IPv4/v6 AFIs
Refresh Style	Poor; soft state, underlay and overlay PIM signaling is continuous	Medium; Hard state for LDP, soft state for RSVP-TE
OAM Toolset	Weak; IP-based ping, traceroute, and multicast traceroute	Strong; IPv4 (LDP), mLDP, and RSVP-TE ping/traceroute

3.2. Inter-AS MPLS Connectivity

This section discusses how the two continental networks exchange unicast and multicast routes for both IPv4 and IPv6. In summary, the networks are integrated using MPLS Inter-AS "Option A" defined in RFC4364, section 10A. Additional technical details are provided below in the two subsections that follow, each of which describes a different routing design.

3.2.1. Active/Standby Routing with Core MoFRR

At the time the network was originally designed, there were two inter-continental links between two separate pairs of routers. This created link-level and node-level resilience in the global network and both links used eBGP to exchange routes. Using MPLS Inter-AS "Option A", each customer VRF was configured using a different 802.1Q subinterface on the inter-AS links. This multiplexing technique enables one-to-one mapping between layer-3 VRFs and layer-2 VLANs which maintains multi-tenancy from end to end. The ASBRs behave exactly like regular PEs.

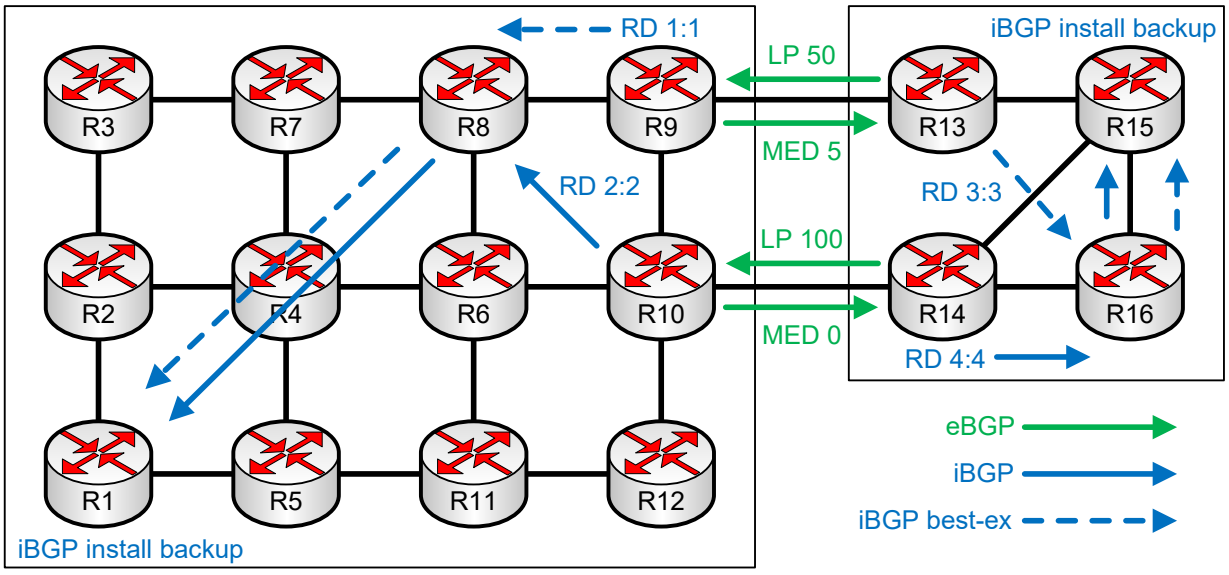
Equal-cost multipath (ECMP) techniques could be utilized by both networks to maximize the utilization on both links, but this complicates troubleshooting. First, ECMP hashing of MPLS L3VPN packets is based on the source and destination IPv4/v6 addresses of the encapsulated packet. Few engineers know this and even fewer know how to determine the exact path for a given source/destination pair. Second, it complicates RPF. To select RPF paths in an ECMP environment, routers will choose the neighbor with the higher IP address as a tie breaker. This implies that operators must discover this technical detail to examine all candidate RPF paths.

Instead, this design uses an active/standby technique on both sides of the routing exchange. For simplicity, all BGP policy configurations are centralized on one of the four devices. To prefer the R10-R14 link over the R9-R13 link, only two attributes must be adjusted on R9:

a. Set a local-preference of 50 on ingress for all routes received from R13. Any value less than 100 is adequate, assuming that 100 is the default local-preference value.
b. Set a multi-exit discriminator (MED) of 5 on egress for all routes advertised to R13. Any value greater than 0 is adequate, since all routers assume a missing MED means 0 MED.

To ensure a fast failover between links, R9 and R13 (the backup routers), are configured to advertise their best external routes to their RRs. These routers will choose an iBGP route from the other ASBR as their best-path due to local-preference. Advertising this best external path to the RRs allows the rest of the network to pre-install it as a backup path. Note that this assumes each route uses a unique RD so that the RR does not directly compare primary and backup paths; this simple technique guarantees that remote PEs will learn both routes for fast failover. The diagram below illustrates this design, assuming R5, R8, and R16 are BGP VPNv4/v6 route-reflectors. R1 and R15 are example PEs that receive both routes as designed. Because these circuits were provided by another carrier's L2VPN service, BFD was enabled on these links and BGP was registered to it. This enabled BGP to failover and reconverge in approximately 200 ms.

Figure 23 - Inter-AS Option A with Active/Standby and Unique RD



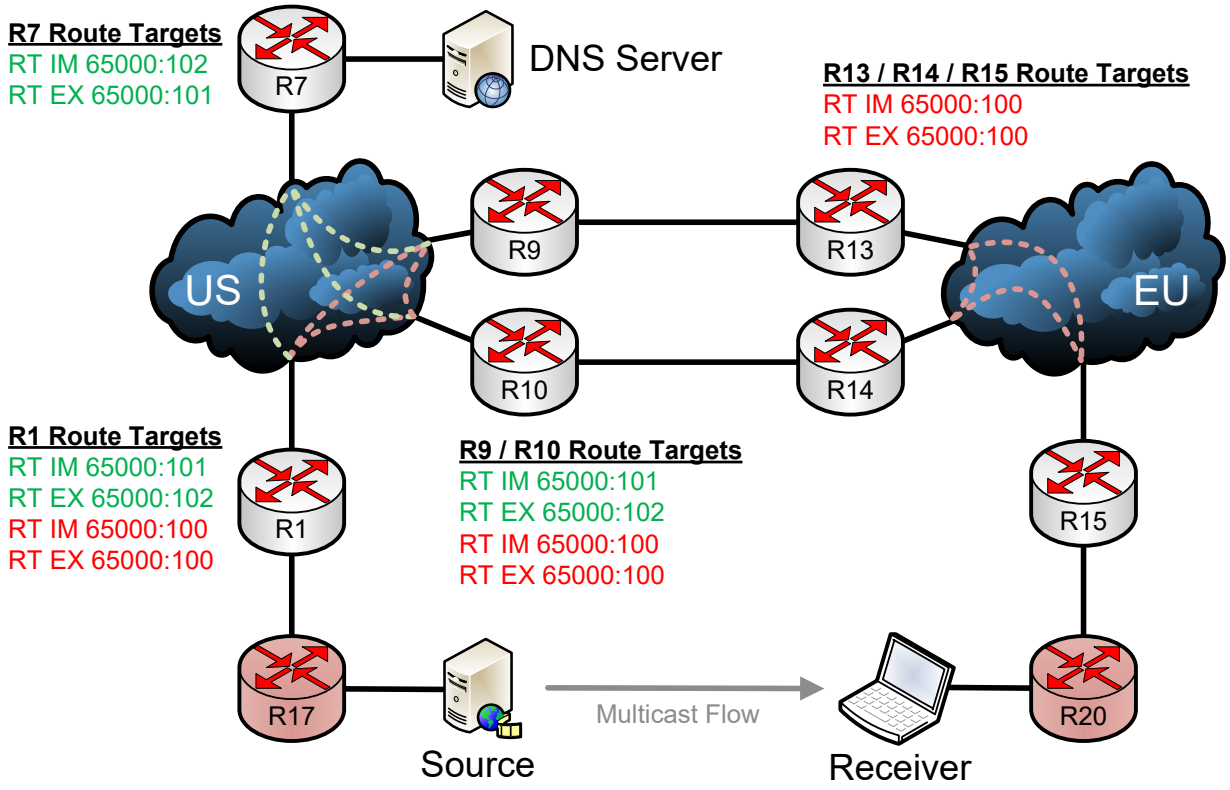
Because the ASBRs in an MPLS Inter-AS "Option A" design behave like regular PEs, PIM SSM must be enabled between them. This allows for IP multicast exchange between the tenant

networks that span multiple contents. There are two major topics to consider regarding this integration, both of which have been discussed to some degree already. Also note that "Option A" allows carriers to use completely different multicast VPN techniques.

First, recall that many multicast receivers are unaware of the SSM sources and must use DNS to discover them, or rely on the last-hop router to do so. A shared L3VPN extranet was built to offer these DNS resolution services to all multicast customers who required it within the US. The extranet is easily extended to Europe by importing/exporting the proper route targets on the ASBR VRFs of interest. This ensures CE routers like R20 can resolve SSM sources from different continents. The diagram below illustrates an example route target plan to build the connectivity just described. Note that European customers aren't aware of this extranet since RTs are not exchanged between continents; all routes received by R13 and R14 appear the same. There is no differentiation between extranet routes and intranet routes at this exchange.

The BGP modifications discussed earlier ensure that the R10-R14 inter-AS link will be used for all multicast exchange unless source-specific RPF modifications are made. This is seldom necessary.

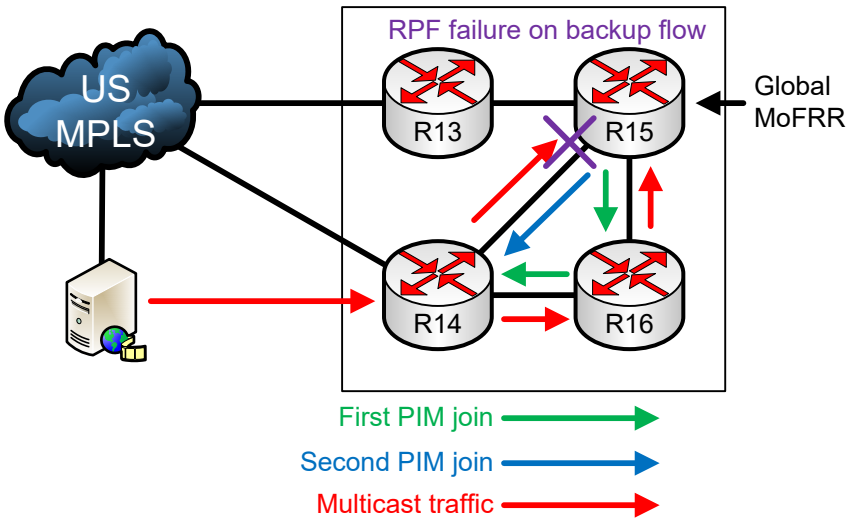
Figure 24 - Enabling Inter-AS DNS Exchange with Route Targets



Second, consider high availability. In the US-based network, label-switched multicast was combined with primary/backup RSVP-TE tunnels for FRR. These types of tunnels cannot carry IP multicast and thus cannot offer protection for them. Instead, multicast-only fast-reroute (MoFRR) can be deployed on the egress PE. Assuming this PE has ECMP-based RPF paths back to the source, it will issue a second PIM (S,G) join towards that source via an ECMP path.

The egress PE identifies the collection of sources and groups upon which MoFRR should be enabled. This feature is enabled in the global table for the MDT transport traffic; this provides protection for both IPv4 and IPv6 multicast flows within the tunnel. The diagram below illustrates this design. Assuming that R15 has two ECMP paths to R14, it will issue two PIM (S,G) joins towards R14's loopback, which is the source of the tunneled MVPN traffic. R14 will receive both joins and install both the R14-R15 and R14-R16 links in the outgoing interface list for the MDT group in question. This consumes additional bandwidth in the network as it provides two copies of each packet to R15. The backup flow will be dropped due to RPF failure so long as the primary flow remains intact, ensuring that only one copy is decapsulated for processing into the VRF, and ultimately towards the CE.

Figure 25 - Core MoFRR for Provider MDT Protection



The design just described was implemented in production because of its simplicity and conformance to present-day requirements. There are two main drawbacks:

1. Ability to use only one inter-AS link at a time due to BGP traffic policies
2. Inability to provide local (EU-based) ingress PE node protection for multicast traffic. Core MoFRR in Option A designs is limited to a single local ingress PE since the egress PE cannot target the remote (US-based) PE directly

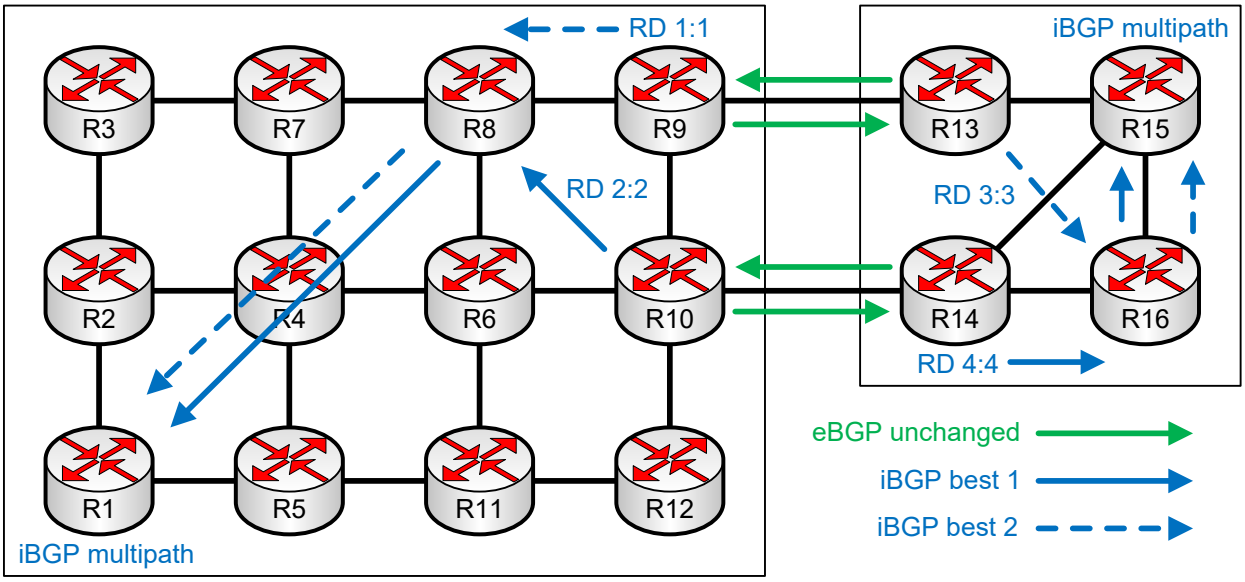
3.2.2. Active/Active Routing with Edge MoFRR

Future requirements for increased availability were being drafted at the time this design was delivered, and this section proposes a fully functional design to meet those more stringent requirements. Most significantly, these requirements demanded that multicast traffic tolerate node failures at the AS edge with minimal downtime. This cannot be achieved using the previous design given the active/standby BGP implementation and MoFRR limitations on some platforms.

To overcome these challenges, BGP should operate in an ECMP-based fashion. All local-preference and MED adjustments are removed so that the European AS learns two equally good routes. These routes still have unique RDs, allowing the RR to evaluate both as best-paths and subsequently advertise them to other PEs in the network. Egress PEs that learn these routes (such

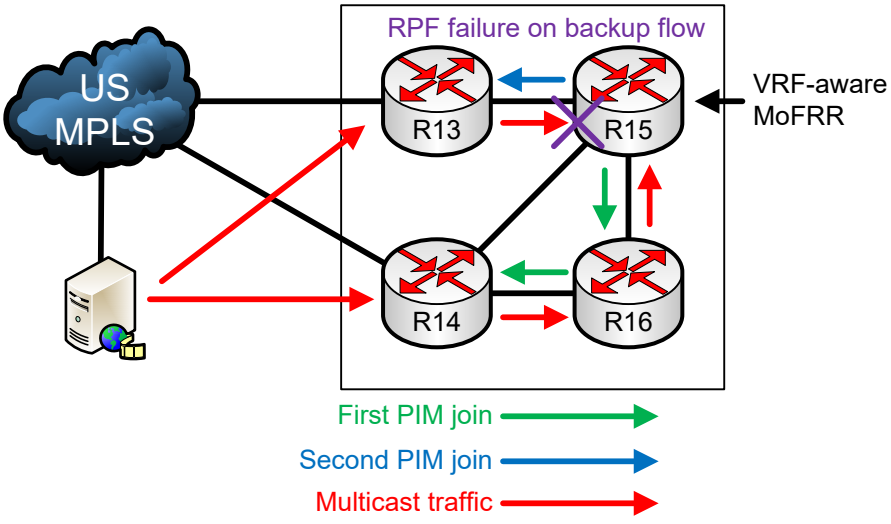
as R15) must install both using iBGP multipath. Note that the IGP cost to the BGP next-hop must be equal for both routes to be installed in the RIB/FIB. This may require reconfiguring IGP costs in the network or configuring BGP to ignore this best-path evaluation step. In our design, R15 had equal-cost paths to R14; a one-hop path with a cost of 20 and a two-hop path whereby each link has a cost of 10.

Figure 26 - Active/Active with iBGP Multipath



Theoretically, with both iBGP routes installed, R15 should be able to issue PIM (S,G) joins within the VRF towards each ASBR. Both customer PIM joins would be sent within the emulated LAN overlay, which would trigger the corresponding provider PIM joins towards each ASBR in the underlay. Unfortunately, this did not work as expected on lab devices despite being logically valid. The design is illustrated below for completeness as it may work on some platforms and in some contexts.

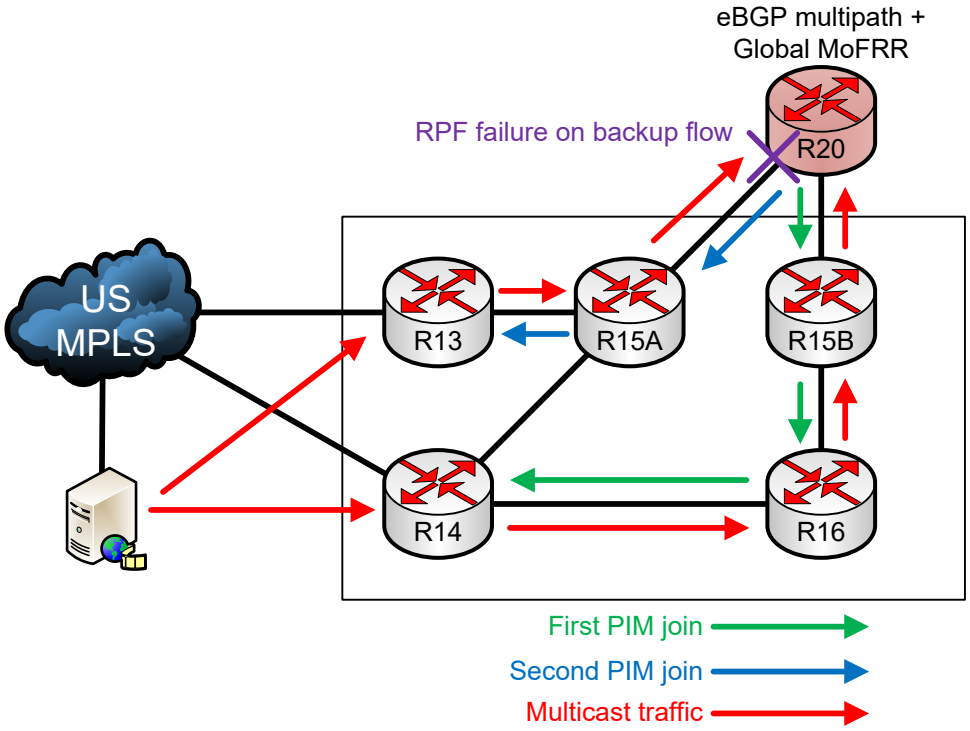
Figure 27 - MVPN-aware MoFRR for Ingress PE Protection (Theoretical)



A known-good design would require multiple egress PEs. Suppose that R15 was replaced with a pair of routers, R15A and R15B. The CE of R20 would be dual-homed to both PEs using eBGP. Each egress PE would be configured to prefer a different ingress PE; this might be automatic given the topology if IGP costs are conveniently configured. If not, egress-influencing attributes such as local-preference or Cisco weight can be used on the egress PE, applied inbound.

The ingress PEs (ASBRs) then advertise the eBGP route for the multicast source to the egress PEs (R15A and R15B). The CE must enable eBGP multipath so that it installs both paths in the RIB/FIB. Rather than trying to configure VRF-aware MoFRR on the egress PE, this design requires MoFRR to be configured on the CE in the global table. This simpler design offloads the origination of secondary PIM (S,G) joins to the CE, allowing the egress PEs to behave normally. Because each join targets a different ingress PE, they'll build disjointed delivery trees to different ASBRs. The US-based network will receive PIM joins over each inter-AS link and trigger the proper mLDP label mapping signaling to build the P2MP delivery trees accordingly.

Figure 28 - MVPN-aware MoFRR for Ingress PE Protection with Dual-homed CE



Interestingly, this design is easier to implement for the carrier as there are no BGP policy adjustments nor any MoFRR configurations. The drawback is that it's only valid for dual-homed sites and that the responsibility of MoFRR configuration is a customer responsibility. Additionally, not all vendors support MoFRR for IPv6. If the CE supports IPv6 MoFRR but the egress PEs do not, this is operationally valid. The PEs will receive two ordinary looking PIM joins and aren't aware of MoFRR at all. If the CE does not support IPv6 MoFRR, then there is no obvious workaround; choose your CEs carefully if you require IPv6 MoFRR support.

4. Complexity Assessment

This section objectively addresses the complexity of each solution using the State/Optimization/Surface (SOS) model. This model was formalized by White and Tantsura *("Navigating Network Complexity: Next-generation routing with SDN, service virtualization, and service chaining", R. White / J. Tantsura Addison-Wesley 2016)* and is used as a quantifiable measurement of network complexity. This section is relevant when comparing this solution to alternatives designs which solve a similar set of problems.

4.1. State

State quantifies the amount of control-plane data present and the rate at which state changes in the network. While generally considered something to be minimized, some network state is always required. The manner in which a solution scales, typically with respect to time and/or space complexity, is a good measurement of network state.

First, consider the general MPLS transport strategy that combines primary and backup auto-tunnels. Given N MPLS-enabled links on a router, there are N primary tunnels and N backup tunnels. Each new link added to the device adds 2 new tunnels. This results in relatively low core state even on tunnel midpoint routers. Unlike end-to-end MPLS TE designs, the addition of faraway PE routers will not impact all the remaining PEs.

To compute the total number of TE tunnels in any arbitrary network using this design, count the number of links in the network and multiple by 4. Consider that there are N primary tunnels and N backup tunnels, which evaluates to 2N. MPLS TE tunnels are unidirectional, to multiplying 2N by 2 yields 4N to cover tunnels in both directions. Overall, this scales linearly with respect to a single device, which is very good.

Next, consider the multicast VPN strategy of mLDP in-band signaling. Like most SSM-based technologies, scale is inherently lower since every customer-signaled (S,G) entry is exposed to the core using the BGP RD to differentiate the state entries. Theoretically, this would scale parabolically (cubic) by counting the unique (S,G, RD) tuples. As discussed earlier, it's important to deploy a variety of techniques to constrain the (S, G) state in the core, such as using DNS, multicast scoping/boundaries, and flow admission control.

Last, consider the inclusion of DNS for SSM mapping. The number of DNS entries scales in parabolic time (quadratic) by counting the unique (S, G) tuples. The DNS server is not aware of different BGP RDs and therefore cannot retain state for it. It is possible that the DNS server could host multiple domains for different customers, and the domain name is a rough proxy for RD, leading to cubic scale since S, G, and domain all count as tuple components.

4.2. Optimization

Unlike state and surface, optimization has a positive connotation and is often the target of any design. Optimization is a general term that represents the process of meeting a set of design goals to the maximum extent possible; certain designs will be optimized against certain criteria. Common optimization designs will revolve around minimizing cost, convergence time, and network overhead while maximizing utilization, manageability, and user experience.

The solution is optimized for multicast transport across continents with minimum packet loss. The lack of ASM support is a sizable trade-off that requires the introduction of DNS services, a centralized extranet, and cooperative customers (i.e., those willing and able to transition to SSM). Since there are no shared trees in SSM, traffic always takes the shortest path from the source to the receivers. This is true in the customer/European networks using PIM and in the US-based network using mLDP in-band signaling. This is always considered a positive optimization.

The underlaying primary/backup auto-tunnel design was also deployed specifically to support mLDP in-band traffic, although it protects unicast traffic as well. It only provides link (NHOP) protection, not node (NNHOP) protection, which is a trade-off of using mLDP of any flavor. Overall, the design provides topology-independent TE-FRR along with a coordinated QoS/admission control design to maximize the performance of multicast applications. All of these technologies are standards-based and are supported on many commercial vendors at the time of this writing.

4.3. Surface

Surface defines how tightly intertwined components of a network interact. Surface is a two-dimensional attribute that measures both breadth and depth of interactions between said components. The breadth of interaction is typically measured by the number of places in the network some interaction occurs, whereas the depth of interaction helps describe how closely coupled two components operate.

Evaluating the breadth of the MPLS transport design, note that the primary/backup auto-tunnels are configured on every MPLS device. This is maximally broad as it is uniformly configured, which also includes targeted LDP sessions over each primary auto-tunnel which supports unicast and multicast LDP mappings. RSVP-TE and LDP work closely together and are tightly integrated, which is a somewhat deep surface interaction.

At the PEs, three different protocols are tightly integrated into two pairs: mLDP+PIM and mLDP+BGP. On the PE-CE link, PIM (S,G) joins are translated directly into mLDP state entries. To complete the opaque value, the RD from BGP is included as well, allowing an egress PE to issue mLDP mapping messages up the reverse path towards the ingress PE connected to the source. Computing this reverse path requires performing a VPNv4/v6 loopback on the egress PE to search for a unicast route, deepening the surface interaction between mLDP and BGP.

DNS is also tightly integrated with PIM. When receiving ASM-based IGMP and MLD membership reports, routers query a DNS server to learn the sources, which must be pre-

configured. DNS effectively inserts these sources into the multicast control-plane, ultimately ending up in PIM (S,G) join and mLDP label mapping messages upstream.

Overall, this solution has relatively broad and deep surface interactions spreading across many protocols. A malfunction in one protocol, say BGP VPN route advertisement, can completely break the design: inability to perform DNS lookups, inability to find an RPF route, etc.

Appendix A – Acronyms

Acronym	Definition
AS	Autonomous System (BGP)
ASBR	Autonomous System Boundary Router
ASM	Any Source Multicast
ASN	Autonomous System Number (BGP)
BFD	Bidirectional Forwarding Detection (BFD)
BGP	Border Gateway Protocol
BSR	Bootstrap Router
BW	Bandwidth
CE	Customer Edge router
CSC	Carrier Supporting Carrier
CSNP	Complete Sequence Number PDU
CSPF	Constrained Shortest Path First
DIS	Designated Intermediate System
DNS	Domain Name System
DOS	Denial Of Service
DSCP	Differentiated Services Code Point
eBGP	External BGP
ERO	Explicit Route Object
EXP	MPLS Experimental bits
FEC	Forward Equivalence Class
FRR	Fast ReRoute

Acronym	Definition
GRE	Generic Routing Encapsulation
HD	High Definition video
ICMP	Internet Control Message Protocol
IGMP	Internet Group Management Protocol
IGP	Interior Gateway Protocol
IP	Internet Protocol
IS-IS	Intermediate System to Intermediate System
ISO	International Organization for Standardization
L2	IS-IS Level 2
L2VPN	Layer-2 Virtual Private Network
L3VPN	Layer-3 Virtual Private Network
LAN	Local Area Network
LDP	Label Distribution Protocol
LSM	Label Switched Multicast
LSP	Label Switched Path (MPLS)
LSP	Link State Packet (IS-IS)
LSPV	Label Switched Path Verification
MAC	Media Access Control (Ethernet)
MDT	Multicast Delivery Tree
MED	Multi-Exit Discriminator (BGP)
MLD	Multicast Listener Discovery
mLDP	Multicast Label Distribution Protocol
MoFRR	Multicast Only Fast ReRoute

Acronym	Definition
MPLS	Multi-Protocol Label Switching
ms	Millisecond
MSD	Maximum Stack Depth
MSDP	Multicast Source Discovery Protocol
MTU	Maximum Transmission Unit
MVPN	Multicast Virtual Private Network
NHOP	Next Hop (Link Protection)
NNHOP	Next Hop (Node Protection)
OAM	Operations, Administration, and Maintenance
OSPF	Open Shortest Path First
P	Provider (core) router
PDU	Protocol Data Unit
PE	Provider Edge router
PHP	Penultimate Hop Popping
PIM	Protocol Independent Multicast
POP	Point Of Presence
PRC	Partial ReCalculation
QoS	Quality of Service
RD	Route Distinguisher
RESV	RSVP Reservation message
RP	Rendezvous Point
RR	BGP Route Reflector
RSVP-TE	Resource Reservation Protocol - Traffic Engineering

Acronym	Definition
RT	Route Target
SD	Standard Definition video
SOS	State Optimization Surface
SPT	Shortest Path Tree
SR	Segment Routing
SSM	Source Specific Multicast
TCP	Transmission Control Protocol
tLDP	Targeted Label Distribution Protocol
TLV	Type Length Value
TP	MPLS Transport Profile
UDP	User Datagram Protocol
UHD	Ultra High Definition video
VPN	Virtual Private Network
VRF	Virtual Routing and Forwarding
WAN	Wide Area Network
WLAN	Wireless LAN
WRED	Weighted Random Early Detection

Appendix B – References

Border Gateway Protocol (BGP) - IETF RFC 4271

DiffServ Classification for QoS - IETF RFC4594

Draft Rosen PIM/GRE Multicast VPN - IETF RFC6037

MPLS Layer-3 VPNs - IETF RFC4364

IPv6 Multicast Scoping - IETF RFC7346

Intermediate System to Intermediate System (IS-IS) - ISO Standard 10589

IS-IS MPLS-TE Extensions - IETF RFC3784

Label Distribution Protocol (LDP) - IETF RFC5036

Label Switch Path Verification (LSPV) - IETF RFC4379

Multicast Source Discovery Protocol (MSDP) - IETF RFC3618

RSVP MPLS-TE Extensions - IETF RFC3209

Source Specific Multicast (SSM) - IETF RFC4607

Navigating Network Complexity (White and Tantsura)

Global MPLS Design Using Carrier Supporting Carrier (CSC)