Global MPLS Design Using Carrier Supporting Carrier (CSC)

Technical Whitepaper

Version 1.2

Authored by:

Nicholas Russo
CCDE #20160041
CCIE #42518 (EI/SP)

THE INFORMATION HEREIN IS PROVIDED ON AN "AS IS" BASIS, WITHOUT ANY WARRANTIES OR REPRESENTATIONS, EXPRESS, IMPLIED OR STATUTORY, INCLUDING WITHOUT LIMITATION, WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Change History

Version and Date	Change	Responsible Person
20200914 Version 0.1	Initial Draft	Nicholas Russo
20200917 Version 0.2	Spelling/grammar corrections	Nicholas Russo
20201002 Version 0.3	Technical clarifications	Nicholas Russo
202001102 Version 1.0	Initial Release	Nicholas Russo
202001205 Version 1.1	Legal disclaimers and cleanup	Nicholas Russo
202101005 Version 1.2	Technical clarifications	Nicholas Russo

Contents

Figures

Tables

1. Overview

1.1. Problem Statement

Working for a large service provider, we struggled to find a way to connect disparate sites in a secure, multi-tenant way across continents. We lacked both the financial and political resources to build a global transport infrastructure ourselves, which was exacerbated by concerns surrounding the initial capital investments and long-term operating expenses. Some of our sites were deployed in developed countries where a wide variety of Wide Area Network (WAN) connectivity options were available. Others were in developing countries where the connectivity options were few and poor performing. Due to security and cost concerns, using the public Internet as transport was not an option at the time this network was designed.

In addition to providing transport connectivity between regions, our diverse collection of customers required a variety of different services. Some required basic IPv4/v6 connectivity, others had non-IP applications requiring layer-2 transport, and still others needed IP multicast transport across the world. Almost all customers required some combination of high scalability, high availability, rapid provisioning, and low packet loss.

Once we identified a transport provider, we learned that it may not be accessible to all locations where we needed a point of presence (POP). Our solution would also have to account for contingency connections, such as one-off direct circuits or additional service providers. These other transports should fit into the design as seamlessly as possible and serve as alternative paths where possible. Furthermore, our primary provider could not guarantee the availability of Ethernet access media, which implied our last-mile design had to be transport-independent.

1.2. Solution Summary

We selected Multi-Protocol Label Switching (MPLS) as the core technology used in the solution. Unlike modern alternatives, MPLS is well-known, widely supported, and has enjoyed decades of success in production. Additionally, much of our network equipment did not support the newest multi-tenancy VPN technologies such as Ethernet Virtual Private Network (EVPN) and Virtual eXtensible Local Area Network (VXLAN).

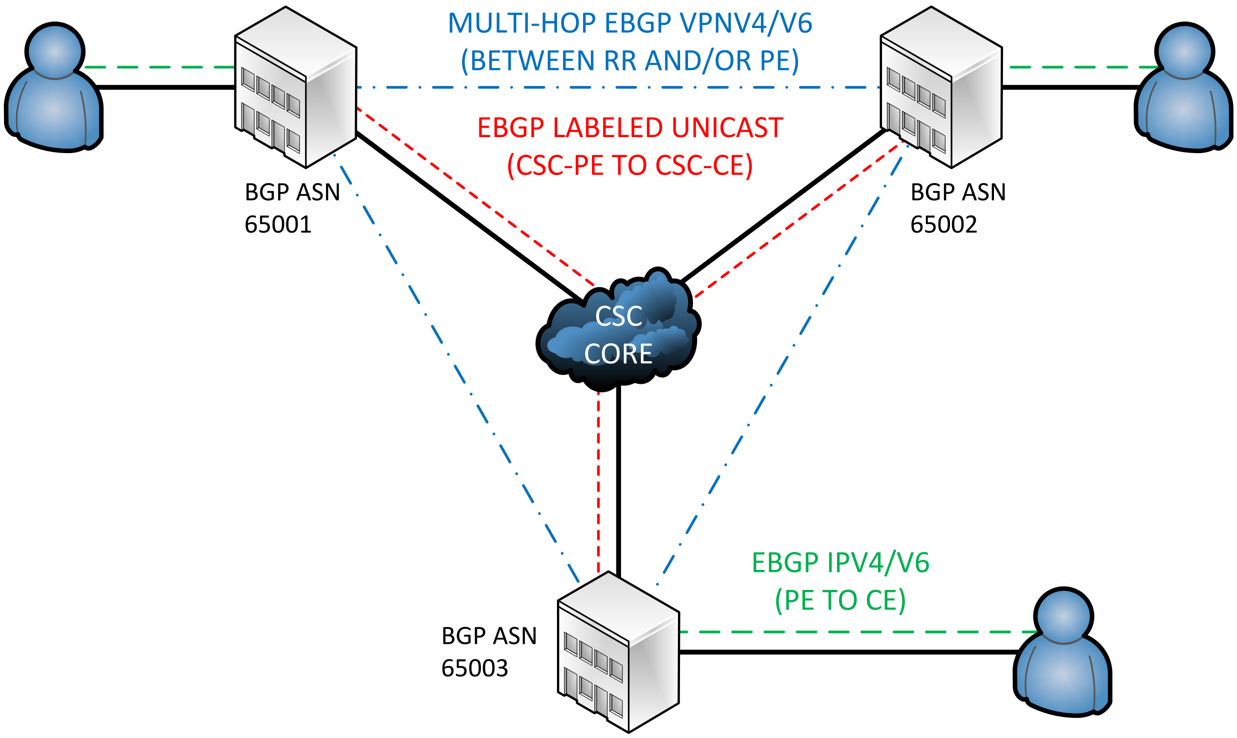
Because we were not able to build a global transport network, we relied on an existing Tier 1 service provider that offered a variety of transport services globally. The most accessible, scalable, and flexible solution available was Carrier Supporting Carrier (CSC). This solution extends the concept of a traditional MPLS Layer-3 VPN (VPN) by allowing the customer to run their own MPLS network within the VPN. As such, our remote POPs could offer a wide array of network services to our customers and the Tier 1 service provider would act as an MPLS transport network only.

CSC is seldom used in real life because other options, such as Ethernet LAN (E-LAN) services, make it easy to connect remote POPs at layer-2. Smaller carriers can run their regular interior gateway protocols (IGP) and MPLS label exchange protocols without any layer-3 interactions with the core carrier. However, such technologies require Ethernet last-mile connectivity (notwithstanding sloppy layer-2 interworking designs) which could not be guaranteed in every country in which we had a POP. CSC provides last-mile circuit flexibility/independence while also improving scale as the customer and core carriers exchange routes using Border Gateway Protocol (BGP). In this context, BGP is extended to include an MPLS label for every prefix and is known as BGP labelled unicast (BGP-LU).

What makes this design truly unique is not only the rare deployment of a production, global scale CSC network, but the inclusion of Inter-AS MPLS Option C. This relatively complex integration allows two different BGP autonomous systems (AS) to exchange BGP VPN routing information in a highly scalable way. Rather than exchanging such information through the AS boundary routers (ASBRs) as Options A and B do, Option C peers the BGP VPN route-reflectors (RR) instead. This allows the ASBRs to be unaware of any VPN routing, serving only as CSC customer edge (CSC-CE) devices connecting to the core carrier's CSC provider edge (CSC-PE) devices. The justification for this design, instead of the more traditional internal BGP (iBGP) VPN sessions, comes later in this document.

The term "BGP VPN" is a generic statement that represents any BGP address-family used to carry customer VPN information, whether it is IPv4/v6 routes, MAC addresses, Virtual Private LAN Service (VPLS) discovery/signalling messages, multicast VPN (VPN) discovery/signalling messages, and more. This highly generic combined design leveraging CSC and Option C allows any service to be extended between any pair of POPs in the world, regardless of their manner of connectivity. Some exceptions apply, often with multicast VPN transport, which is discussed later. The diagram below illustrates a high-level design L3VPN design.

Figure 1 - High-level CSC/Option C Architecture



2. Architecture

This section describes the solution in greater technical detail. It examines each individual component in depth, adding new components as it progresses. This document is not a training tutorial on the technologies, but does explain how they work within the context of the design.

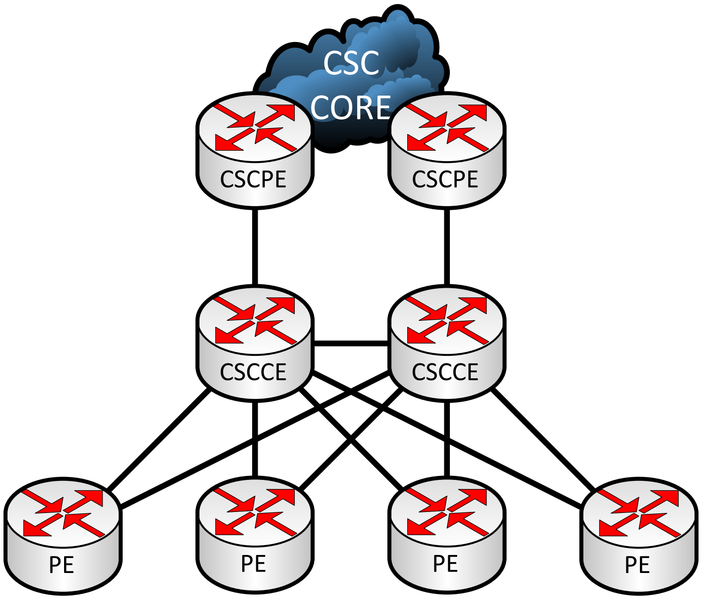
2.1. Point of Presence (POP) Design

Individual POPs within the architecture do not have to be identical, but there are some common design constraints that apply to all of them. This section explores the design of the POPs themselves without focusing on inter-POP communications. In my particular customer, the POPs operated autonomously for about a year before we decided to tie them together. During that first year, they only served their regional customers with no inter-POP/global connectivity available.

2.1.1. Physical Connectivity

We developed two conceptual POP designs, each of which had two options for BGP route-reflector (RR) placement to service the BGP VPN address-families. The first design was a traditional aggregation block with two distribution/core routers on top. Every customer-facing PE device would dual-home to each distribution/core router (typically CSC-CEs or dedicated P routers) using a directly connected Ethernet connection. Such designs are decades old and are commonly seen in campus access networks and traditional data centers where the vast majority of traffic is north/south. In our case, north/south means inter-POP, and this was indeed the main traffic pattern for most customers once global connectivity was established. Very little traffic traveled east/west, meaning intra-POP, although this was certainly supported. The diagram below illustrates the traditional aggregated POP design.

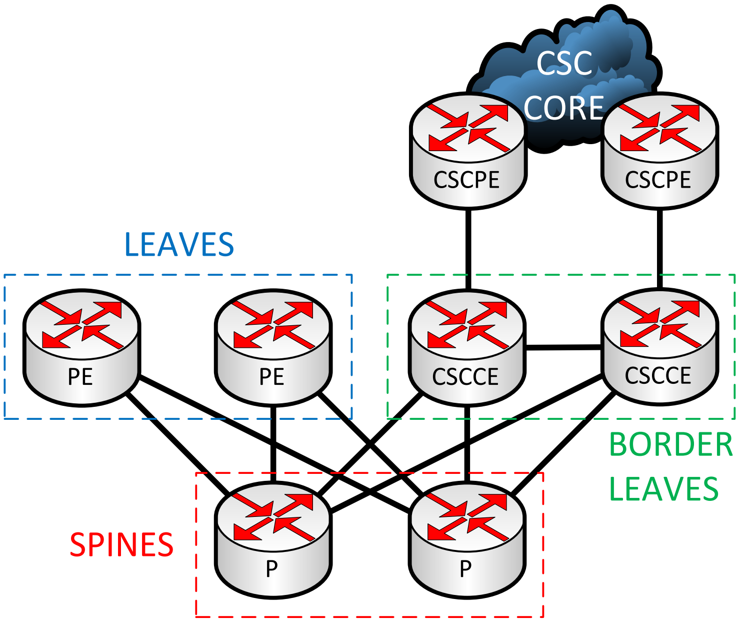
Figure 2 - Traditional POP Physical Design



The second design was based on a leaf/spine design, effectively adding another pair of routers between the customer facing PEs and the CSC-CEs. Both the PEs and CSC-CEs are "leaves" in this design, with the CSC-CEs being classified as "border leaves" given their integration with an external network. The middle tier consisted of the "spines" whereby every leaf is connected to every spine. Leaves never connect to leaves and spines never connect to spines within the same tier, with one exception. The border leaves can optionally be interconnected because shuttling ingress/egress traffic between edge devices is useful to improve availability or implement ingress/egress traffic engineering in the future. The main technical advantage of leaf/spine over the traditional design is the ability to improve scale for east/west traffic. Simply add more spines to increase availability, capacity, or both.

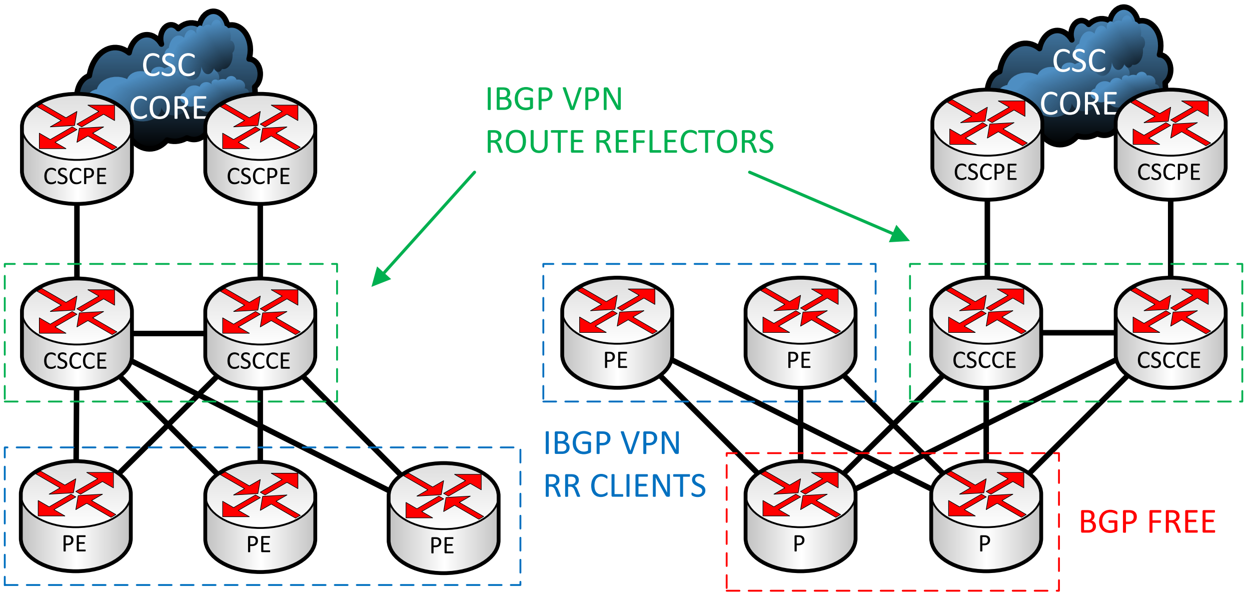
This can also be viewed as a disadvantage, since the only purpose of a spine is to forward traffic. This incurs additional cost and management burden. In real life, we never deployed leaf/spine POPs as there was no compelling operational justification, despite their popularity at the time. This document will discuss the details surrounding its deployment nonetheless. The diagram below illustrates the conceptual leaf/spine POP physical design.

Figure 3 - Leaf/Spine POP Physical Design



We overlaid two different BGP RR strategies atop these POP designs. The first was a low-cost approach that repurposed the CSC-CEs, whether they were aggregation routers or border leaves, to serve as BGP RRs for the POP. Because these devices were already quite powerful in terms of computing capacity, using them to serve as BGP RRs was a low-risk, cost-effective choice. Each PE in the POP would peer to these RRs using internal BGP (iBGP) which is detailed later in this document. This is the design we selected in real-life as cost concerns governed many of our decisions. The diagram below illustrates the intra-POP iBGP VPN sessions overlaid on both the traditional and leaf/spine physical designs. Note that the precise details regarding the iBGP topology are discussed later in the document.

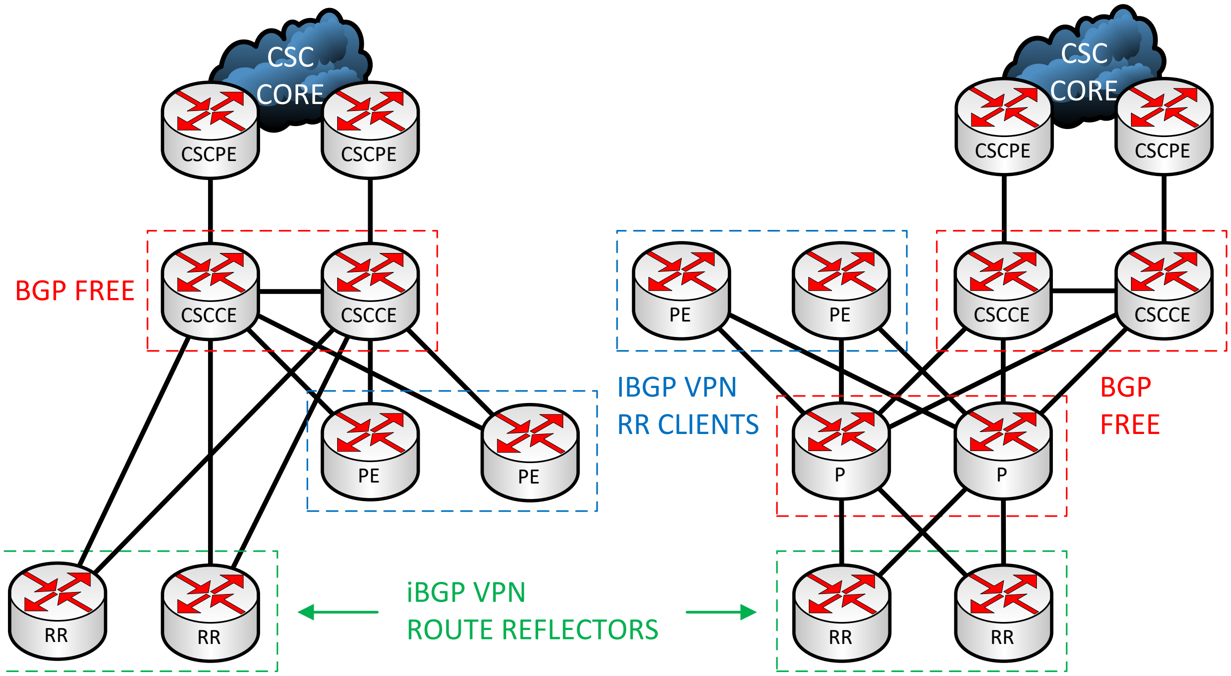
Figure 4 - Using CSC-CEs as BGP VPN Route Reflectors



The second design involved a pair of dedicated RRs outside of the forwarding path of customer traffic. These routers would look like PEs from a physical connectivity perspective, but would not service any customers and would never be used for traffic forwarding. This non-transit behavior can be implemented by manipulating IGP (discussed later). In modern designs, these BGP RRs are often low-cost virtual routers with large memory allocations, medium CPU allocations, and low network bandwidth allocations. Additionally, we considered using a different pair of BGP RRs for all the different VPN services we offered, such as IPv4 VPN, IPv6 VPN, multicast VPN, etc. This incurs even greater cost and management burden, but reduces fate sharing and slightly improves availability.

Some of the largest carriers manage risk by spreading different BGP address-families across different RRs to the maximum extent economically possible. In our environment, we did not have a general-purpose computing environment immediately available. When including the capital investment needed to build and maintain it, this solution was prohibitively expensive and not at all worth doing. The diagram below illustrates conceptual examples of adding dedicated RRs to the traditional and leaf/spine POP designs at a high-level. Note that the term “BGP free” means that there are no VPN capabilities on those devices. Some devices, like the CSC-CE, may run BGP for a different purpose later.

Figure 5 - Using Dedicated Out-of-band BGP Route Reflectors



2.1.2. IGP Routing

Because each regional POP is relatively small (consisting of 10 to 30 devices), any IGP would scale adequately without much concern. Although our organization had no need for any MPLS traffic engineering (TE) given the tiny size of our POPs and lack of a long-haul infrastructure, we agreed that choosing a link-state protocol was necessary. This makes future TE integration easier, along with support for emerging technologies like Segment Routing (SR). This reduced our choices to Open Shortest Path First (OSPF) and Intermediate System to Intermediate System (IS-IS), the two most popular link-state IGPs.

OSPF was the more appropriate choice for our network because our operators were already extensively trained in this protocol. Some network OS implementations, like Cisco IOS, IOS-XE, and IOS-XR, will ignore OSPF external routes when redistributing OSPF into BGP by default. This is useful because any BGP routes redistributed into OSPF will not be considered for redistribution from OSPF back into BGP. In short, this prevents routing loops with no additional design or implementation effort. IS-IS has no such default behavior, and this will become relevant later in the document when discussing CSC integration. To prevent potential routing loops, IS-IS would require manual configuration to match/filter these routes at the point of redistribution (CSC-CE). For network implementation experts, this is inconsequential, but it is avoidable complexity that adds no value. In both cases, the scale of each POP is small enough that a flat OSPF area 0 or IS-IS level-2 design is adequate, with the exception of dedicated RRs in OSPF environments (discussed later).

First, consider basic OSPF optimizations. All transit links should use the point-to-point (P2P) network type to speed convergence, reduce link-state database bloating, and reduce the topological graph complexity. P2P links do not have a designated router (DR) and thus no DR election. A link interconnecting exactly two OSPF speakers is not a multi-access network and therefore does not benefit from a DR, which is represented as a Link State Advertisement type 2 (LSA2) in the LSDB. As such, no LSA2 should be present anywhere in the network, reducing the number of total graph vertices by almost half. It is advisable to retain "stub networks" within the router LSA for OSPFv2 (LSA1) or the intra-area prefix LSA within OSPFv3 (LSA9) to simplify troubleshooting. This allows operators to ping transit links, to source pings from transit links, and to see at a glance which links might be experiencing problems by checking the routing table. Given the small network and the rarity with which these IP subnets change, there is little operational benefit to suppressing these prefixes.

Next, consider OSPF security. Modern OSPF implementations allow for SHA-256 authentication (some platforms offer even stronger hashes) which should be preferred instead of the older MD5 option. In addition to authentication, OSPFv3 also offers IPsec encryption, which in the author's experience, is overly complex, prone to breaking, and not worth deploying. OSPF TTL-security ensures that neighbors are directly connected, preventing any long-range hijacking attacks from external networks, such as those accessible over CSC. Protecting the OSPF link-state database (LSDB) itself can be accomplished by setting maximum LSA limits to prevent accidental LSA injection at scale, perhaps due to unfiltered BGP to OSPF redistribution. Such concerns were irrelevant in our environment given that our global Internet connections were placed in a VPN, but some customers may prefer to transport Internet traffic in the global routing table. This topic is discussed in greater detail later in the document.

In less symmetric networks, some operators deploy loop free alternate (LFA) technologies to allow OSPF to inspect the LSDB in greater detail to determine if backup paths exist. When they do, the router can preemptively install these backup paths in hardware for faster failover. In our case, POPs are perfectly symmetric with the same IGP cost used on all links (10 in our case), automatically resulting in equal-cost multi-path (ECMP). This feature allows for load sharing between devices based on various hashing algorithms which are out of scope for this document. More important than the load sharing is the high availability; because both routes are used for forwarding, they are both programmed in hardware already. This obviates the need for complex LFA techniques and given our early-in-career network operators, ECMP was the best choice. Note that some hardware platforms may benefit from LFA enabled even in ECMP environments. This depends on how the platform maintains its forwarding tables and is likewise out of scope for this document.

The first step in the convergence process is failure detection. Because all devices in the POP were directly connected (i.e. no intermediate Ethernet switches), the Ethernet interface line status was an accurate indication of a link's up/down status. This raises the question of "carrier delay"; how long after a failure is detected should the control-plane mark the interface as down? In our first three years of operation, we observed only two false-negative micro-flaps whereby an interface loses electrical or optical signal for a brief period of time (a few milliseconds at most), but immediately returns. Marking this as a link flap and starting the convergence process is more detrimental than just waiting, so we used a relatively aggressive carrier-delay of 5 milliseconds. This delay helps the control-plane ignore rare microflaps rather than starting the convergence process prematurely.

Note that Bidirectional Forwarding Detection (BFD) is generally unnecessary in the POP because of the direct Ethernet connections. If, for example, Ethernet switches (or other transit devices such as media converters) were present, using BFD with echoes enabled would be a good design decision. Various protocols, including all IGPs, can register to BFD, which notifies them when links go down. While BFD is typically slower than using link status for failure detection, they can be used together; if line status stays up after a link fails, BFD will detect it soon enough. This is a "belt and suspenders" approach that some carriers use to maximally reduce risk, but we saw it as introducing unnecessary, low-value complexity.

Both OSPF and IS-IS have many tunable convergence timers, but in the author's experience, two of them have an outsized impact and should be optimized first when optimization is deemed necessary. Note that often times such optimization is unnecessary, but our customers had strict performance requirements that heavily influenced our routed convergence design. These timers are the OSPF LSA generation and SPF throttle timers.

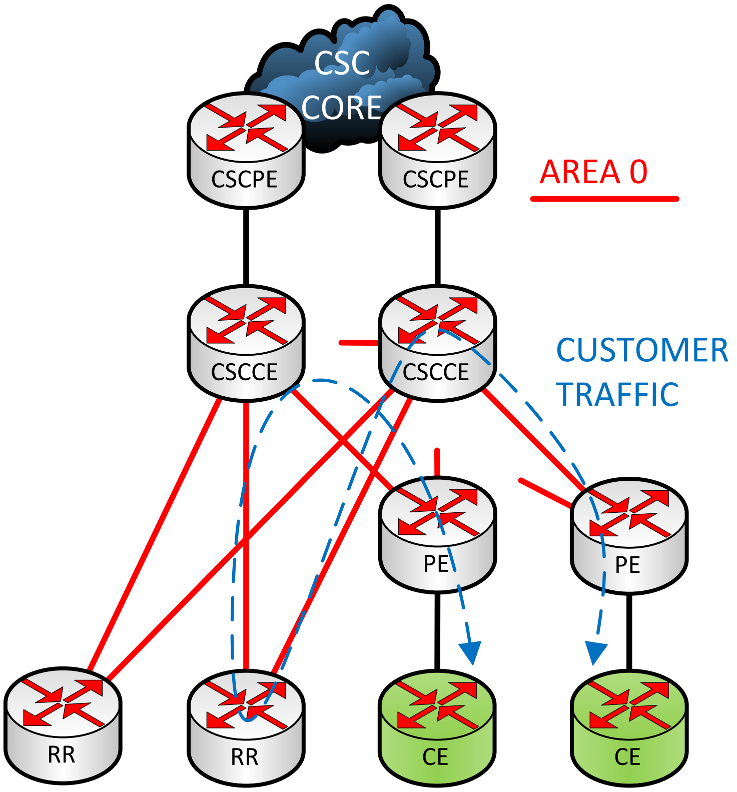
First, we adjusted the OSPF LSA generation throttle timers. This controls how long to wait between originating the same LSA after observing a change in the network. We selected 50 ms to better group multiple concurrent link failures. Degrees of interface disjointedness varied widely based on the hardware, which was not consistent network-wide due to budget limitations. For example, some PEs had their distribution/core uplinks spread across two linecards, while other devices did not. Rather than try to "point optimize" individual devices or POPs, we used a relatively conservative initial LSA delay timer discussed above. Generation of successive LSAs began at 250 ms after the initial LSA, doubling each time up to a maximum of 1000 ms. This exponential back-off prevents excessive IGP flooding when successive changes keep occurring.

Next, we adjusted the OSPF SPF throttle timers. SPF is the algorithm run each time a change to the OSPF topology is detected. In this design, full SPF runs whenever a change to an LSA1 is detected, while a partial SPF is run for changes in an LSA3 (inter-area routes) or LSA5 (external routes). These timers are tuned for more rapid SPF calculations to speed convergence within the POP. Given the relatively small OSPF topology with low prefix count (and modern routers), SPF runtimes are not a concern. Our testing indicated that all LSA flooding can complete in less than 50 ms within any POP. Therefore, the SPF initial way timer was set to 50 ms, capturing all of the LSA changes and running SPF only once. In the unlikely and unobserved event that SPF doesn't capture all the LSAs, it will run again after 300 ms, doubling up to a 1000 ms maximum.

Some readers might be curious about incremental SPF (iSPF). While academically clever, the author's operational experience with iSPF is largely negative. It defines various "shortcuts" that OSPF can take in specific topologies to skip steps in the SPF process. For example, a singly-connected router is the gateway to other routers fails, iSPF can summarily discard everything behind it. This may have a positive impact in large networks, but simply stated, the technology is buggy, hard to troubleshoot, and uncommonly deployed. Modern Cisco devices don't even support it anymore. We opted not to deploy iSPF.

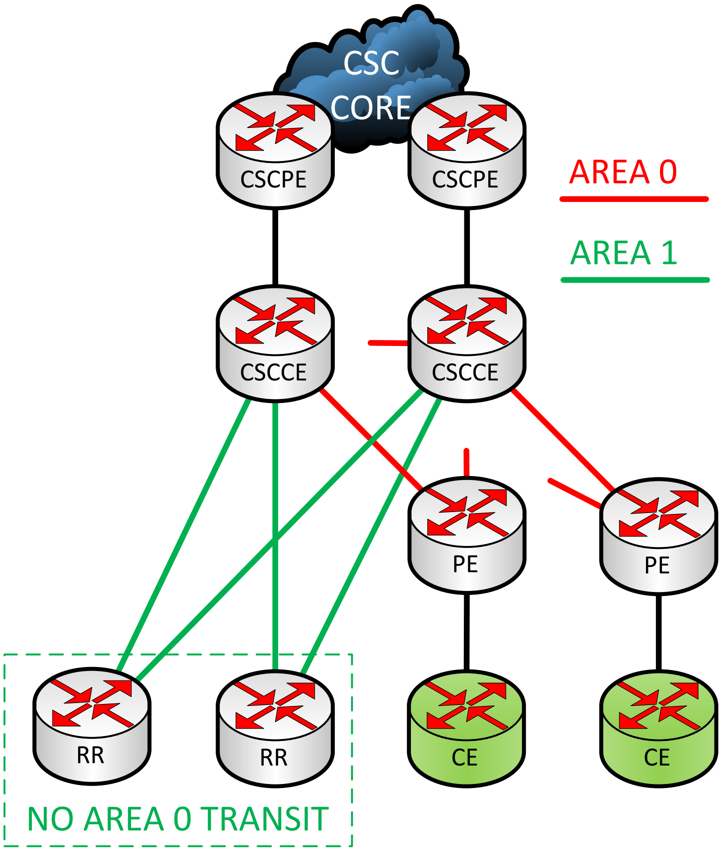
Lastly, there is one case where using OSPF areas within the POP makes sense. When dedicated RRs are used, they should never be used for core transport forwarding. It would be better to blackhole traffic entirely than to crash the RRs which might be servicing other satellite POPs across the network (discussed later). If the RRs are placed in area 0, the diagram illustrates what might happen if enough link failures occur within a POP.

Figure 6 - Using RRs for Transit in a POP with Link Failures



While it is exceedingly unlikely, the impact is severe, and is worth protecting against. Take advantage of one of OSPF's many loop control prevention mechanisms by putting these RRs into a different area, perhaps area 1. No special area types or LSA filtering/summarization is necessary. The area assignment alone will prevent two PEs in area 0 from communicating across RRs in area 1. Now, when the link failures occur, intra-POP PE traffic simply fails as the MPLS label switched path (LSP) between the PEs is broken due to BGP next-hop inaccessibility. Put another way, we can leverage the dreaded "disjoint area 0" as a good thing, preferring to have broken connectivity rather than causing damage to our BGP VPN infrastructure which is likely servicing satellite POPs. The diagram below summarizes the high-level OSPF design using multiple areas to ensure dedicated RRs are not used for core transport forwarding.

Figure 7 - Preventing Transit RRs in a POP using Areas



Although we did not deploy IS-IS, it is worth a brief discussion. Like OSPF, the same recommendations for graph optimization, security, and performance tuning exist, with the exception of TTL-security. Because IS-IS is not based on IP, it is inherently insulated from IP-based hijacking attacks. Such attacks can never target IS-IS, so TTL-security is unnecessary. As it relates to dedicated RRs and keeping them out of the transit path, IS-IS has a specific feature named the overload bit (OL). When set on a router, the OL-bit signals to all other routers that the device is “overloaded” and should never be used for forwarding, even if no other paths exist. While some OSPF implementations have a “max-metric” feature, this is just a cost adjuster and does not prevent transit traffic, but merely discourages it. In contrast, the IS-IS OL-bit is both authoritative and effective on these dedicated route reflectors. IS-IS routers with the OL-bit set can never be used for transit, even as a last resort.

2.1.3. Multicast Routing

Multicast routing within each POP is relatively simple. Protocol Independent Multicast version 2 (PIMv2) is enabled everywhere that IGP is, minimizing any possibility of a reverse path forwarding (RPF) failure. RFC4607 describes PIM source-specific multicast (SSM) which is the only operating mode of PIM supported in the core network. This uses the 232.0.0.0/8 multicast group range. This document will discuss multicast VPNs in greater detail later, but in summary, PE loopbacks can discover one another using various BGP address-families. Since the only purpose of a PIM any-source multicast (ASM) rendezvous point (RP) is to discover multicast sources, no RPs are needed in this network. Eliminating the presence of PIM RPs significantly simplifies the multicast design, implementation, and maintenance complexity network-wide.

As discussed earlier, BFD was not enabled in our POPs as it was unnecessary. However, if BFD is used, PIM should be registered to BFD for fast failover on par with IGP.

2.1.4. BGP VPN Services Routing

This section discusses the BGP VPN design for a variety of address-families pertaining to customer services. Specifically, we offered the following services in our environment:

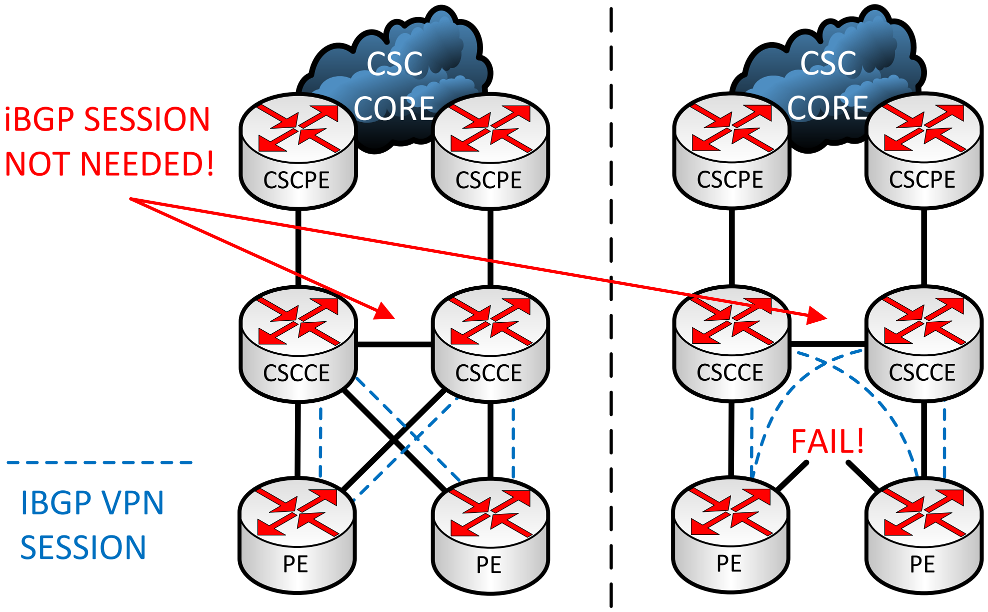
a. IPv4 VPN (VPNv4): Multi-tenant IPv4 connectivity across MPLS L3VPN
b. IPv6 VPN (VPNv6): Multi-tenant IPv6 connectivity across MPLS L3VPN
c. VPLS: Multi-tenant, multi-access Ethernet connectivity using BGP discovery and LDP signaling via RFC4762
d. IPv4 MVPN: Multi-tenant IPv4 multicast VPN discovery/signaling in tandem with MPLS L3VPN
e. IPv6 MVPN: Multi-tenant IPv6 multicast VPN discovery/signaling in tandem with MPLS L3VPN

This document discusses each of these services in greater depth later in the document, but for now, just know that these 5 BGP address-families were configured on every PE and RR in the global network. Also note that each service uses the word “multi-tenant”. While it is possible to offer global IPv4/v6 services outside of a VPN context, neither our organization nor our customers had any use for this (discussed in greater detail later). Lastly, any other VPN service, such as EVPN, could also be offered. Given our hardware limitations, we opted for more traditional service offerings.

As previous sections have suggested, each POP will house a pair of BGP RRs that service all 5 of these address-families. Whether those RRs are built into the CSC-CEs (as we did) or deployed as dedicated devices is generally irrelevant to the design specifications that follow. Each RR in the POP would peer to each PE using iBGP, making each PE an RR client. In effect, VPN routes from one PE were reflected to every other PE twice, once from each RR, providing high availability. This design is common, intuitive, and easy to troubleshoot.

Some may find this next point controversial; we did not peer the two RRs together using any BGP address-family. Such a peering introduces needless complexity in several dimensions: configuration, troubleshooting, and BGP bestpath evaluation, as well as additional consumption of network and computing resources. In BGP parlance, RRs that service the exact same set of clients are in the same “cluster”. This is true regardless of their cluster ID configuration, which is irrelevant in this design. Within a cluster, there is little benefit to peering RRs. Even with multiple concurrent link failures, the iBGP sessions between loopbacks will remain up within the POP thanks to IGP. IGP is enabled between the CSC-CEs for this reason. On less mature BGP implementations, the intra-cluster-RR peering can be useful if iBGP session stability is questionable, but this is rare. Frankly, I would not recommend deploying immature BGP implementations on globally-significant RRs in the first place. In the past decade, the author has never observed otherwise stable BGP sessions failing after having been properly established. The diagram below illustrates this design, as well as a common misconception regarding multiple link failures within a POP and its impact on iBGP sessions.

Figure 8 - Intra-POP iBGP VPN Sessions and Link Failure Tolerance



In addition to reducing BGP complexity, the decision to omit this inter-RR iBGP session allowed us to create two disjoint BGP VPN meshes. One was named “mesh A” and the other was named “mesh B”. The meshes only converged at the PEs, which were not configured as RRs, and thus were not able to reflect iBGP routes between meshes. Such a design is conceptually similar to storage area networks (SAN) where the SAN A/B transport networks are completely independent for availability purposes. If one SAN becomes corrupted or otherwise fails, it would be contained only to that SAN, and the same is likewise true for these disjoint iBGP meshes.

This A/B mesh design applies to all 5 BGP address-families, and also note that the RRs did not have any customer VPNs configured. This improved their memory utilization as there was no import/copy process from BGP into local routing tables on a per VPN basis. The relevance of the A/B mesh design is explained more later in the document as it relates to inter-POP connectivity.

2.1.5. MPLS Label Advertisement

There are a variety of ways to distribute MPLS label information within the POP. This section explores three varieties and explains why only one was deployed in our production environment.

2.1.5.1. Label Distribution Protocol (LDP)

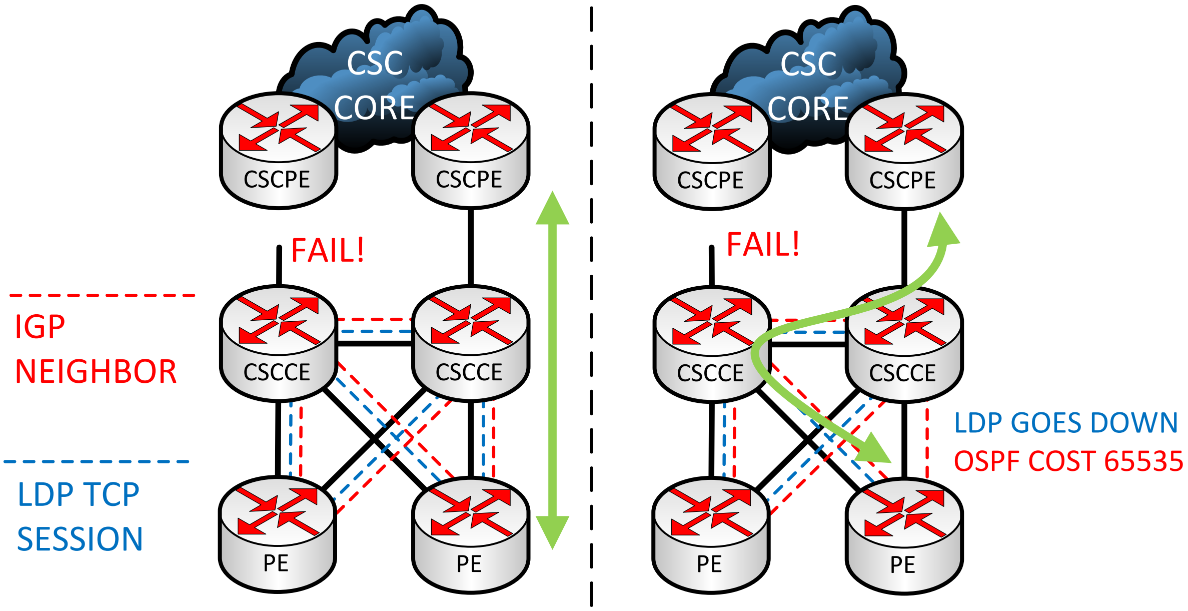
LDP is a protocol-independent, hard-state label allocation and distribution option for MPLS networks. Generally speaking, it behaves like, and runs alongside, IGP. It uses link-local UDP multicast hellos for discovery and multi-hop TCP sessions between loopbacks (typically) for label advertisement and withdrawal signaling. LDP sessions can be authenticated using MD5 via the TCP header, which is a security technique we implemented. Of greater significance to our customers are two other LDP features. These are designed to speed up convergence and prevent forwarding black holes in MPLS networks.

First, LDP/IGP synchronization protects against two common cases: where IGP converges before LDP can exchange label bindings or an LDP session is closed but traffic continues to forward along the original path.

In both cases, it is a synchronization issue between IGP and LDP where the two protocols converge at different times. In our case, LDP/IGP synchronization was enabled on all IGP-enabled core interfaces, much like PIM. When OSPF has an adjacency on a link but LDP does not, this feature raises the OSPF link cost to the maximum value of 65535.

This makes the link highly undesirable and, assuming other paths exist to the same destination, forces OSPF to re-route around the LDP-incapable link. The diagram illustrates the OSPF/LDP re-routing concept just described.

Figure 9 - LDP/IGP Synchronization with LDP Session Failures

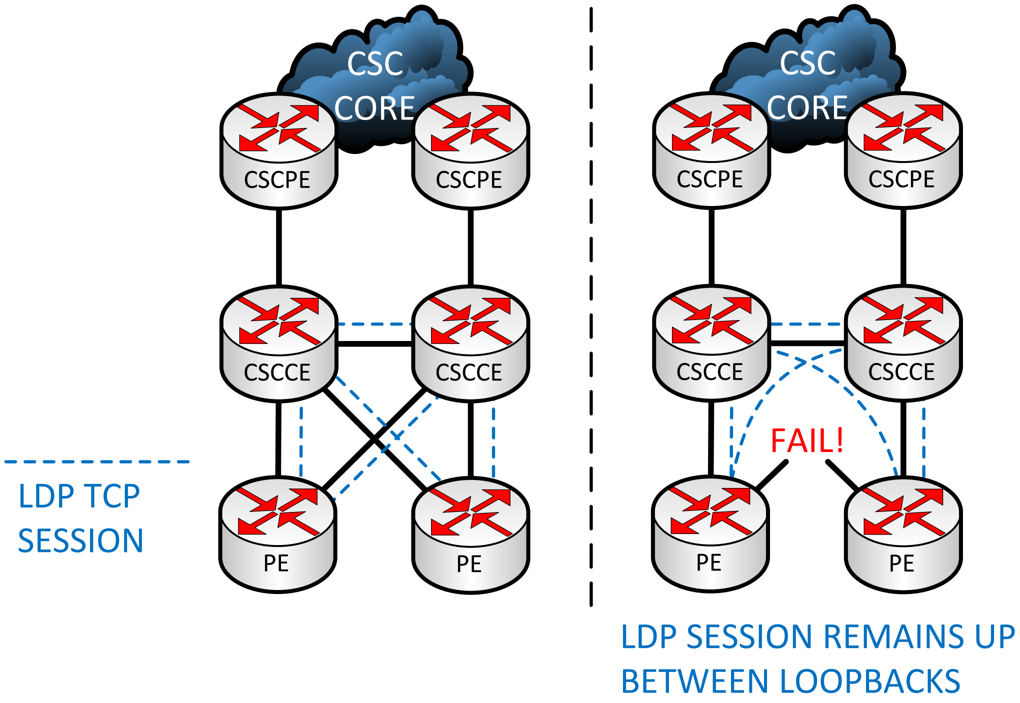


Next, we enabled LDP session protection. This feature is enabled for all peers with a 10 minute hold down timer. This sends targeted hellos to all neighbor's LDP router IDs so that if a link fails while the neighbor's loopback is reachable via IP, the session stays up.

Although IGP will determine the forwarding path, this cuts down on LDP convergence and Label Information Base (LIB) refresh times. If the LDP peer is not reachable after 10 minutes, the label bindings for that peer are flushed and the preserved LDP session is torn down. Conceptually, LDP session protection is similar to carrier-delay.

Instead of being a small time period designed to tolerate short microflaps, it is a larger time period designed to tolerate actual link failures. The diagram below illustrates how LDP sessions remain up even after links fail. You'll observe that the behavior is very similar to iBGP.

Figure 10 - LDP Session Protection with Link Failures



2.1.5.2. Resource Reservation Protocol for MPLS Traffic Engineering (MPLS-TE)

Although RSVP was originally invented for Integrated Services (IntServ) with respect to Quality of Service (QoS), it was repurposed to signal MPLS TE tunnels. RSVP is a connectionless, soft-state protocol, much like PIM. RSVP-TE typically relies on link-state topology database information from OSPF or IS-IS to view the network and compute paths through the network given a variety of constraints.

Unlike LDP, which runs alongside IGP and never changes the way traffic is actually forwarded, RSVP-TE redirects traffic into arbitrary tunnels using MPLS encapsulation. This “source routing” approach is often used by carriers to better utilize all transmissions links and to provide Fast ReRoute (FRR) services when primary transmission links fail.

During our analysis, we found no suitable use-case for RSVP-TE in our environment as our POPs were relatively small and did not require intricate intra-POP TE forwarding. Using RSVP-TE for FRR was considered but ultimately rejected as configuring IGP for ECMP and tuning IGP convergence timers was sufficient to meet our availability requirements. RSVP-TE is not discussed any more in this document as a result.

2.1.5.3. Segment Routing (SR)

SR for MPLS is a relatively new concept that allows link-state IGPs like OSPF and IS-IS to distribute label information directly. Because LDP and PIM are both protocol-independent and run alongside IGP, they can be combined with any IGP without much nuance. However, SR requires IGPs to be extended to support label distribution, making support more limited. Other drawbacks include hardware support, as SR is relatively new, and our current equipment did not uniformly support it. We intended to migrate to SR from LDP at some unspecified point in the future when SR supportability was universal.

There are several advantages to deploying SR over LDP. As it relates to convergence, there is only one protocol, so problems surrounding IGP synchronization and session protection don't exist. Simply tune IGP to converge at the desired pace and the label switched paths will be immediately available. Additionally, SR traffic engineering (SR-TE) is stateless in that transit routers do not retain information for each TE tunnel as they do with RSVP-TE. In addition to providing a scale advantage, this makes SR much more flexible. LDP has no mechanism for point-to-point (P2P) TE-style LSPs and RSVP has no mechanism for multipoint-to-point (MP2P) IGP-style LSPs. SR can support both without any special configuration.

Given the CSC design, each POP is an independent IGP domain. This implies that the label distribution method used in each POP is also independent. It is possible to use LDP, RSVP-TE, and SR at the same time but in different locations. Operationally, there is little advantage to strategically designing the network in this manner. However, using a migratory example, transitioning POPs from LDP to SR on a per-POP basis, organized regionally, can be a smart approach.

2.1.6. Customer Services

As is true in any MPLS network, the PEs in each POP can support a variety of customer services. In our environment, we offered three types of VPN services which are detailed in this section.

2.1.6.1. Layer-3 VPN

MPLS Layer-3 VPNs (L3VPN) provides routed connectivity across the network which logically collapses the entire MPLS core into a single router from the customer's perspective. Customer edge (CE) devices will use some kind of routing protocol to exchange routes with the Provider edge (PE) devices. BGP is most frequently used in commercial networks, but in our environment, OSPF was frequently used. Technically, any routing protocol can work, including static routes.

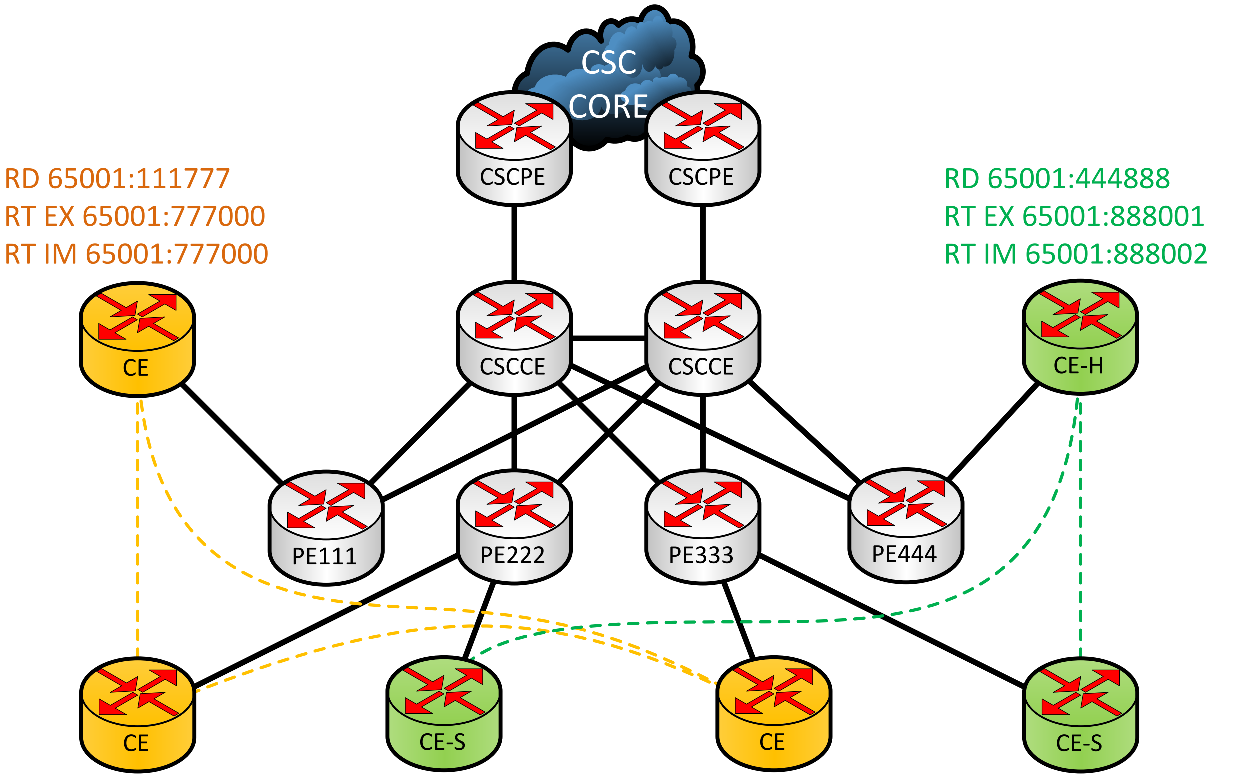
There are two other important design components to consider. First, each VRF must be assigned a Route Distinguisher (RD). Each RD is a 64-bit value which differentiates prefixes that are carried inside of IPv4/v6 BGP VPNs. Because MPLS L3VPN provides multitenancy, different customers may use overlapping IP networks, which is commonly observed with RFC1918 address space for IPv4. A simple and common approach to allocating RDs is to use the BGP ASN as the first 32 bits and a unique value for the last 32 bits representing a specific PE+VRF combination. While RDs may be duplicated across VRFs on different PEs, doing so may negatively impact convergence time. In our environment, every VRF in the network had a different RD. For example, suppose each PE can have a maximum of 1,000 VRFs. A given PE has an ID of 321 and that PE is within AS 65001. The first VRF on that POP would use RD 65001:321000 and the 1000th VRF would use RD 65001:321999. Such a design scales to 1,000 PEs per POP with 1,000 VRFs per PE. For greater scale, one can leverage additional digits in the low-order 32 bits of the RD, but six digits is adequate for most networks.

Second, consider the Route Target (RT) design. Like the RD, each RT is 64-bits and is it common for the first 32 bits to represent the BGP ASN. Regarding the value, these should be determined on a per-customer basis. For a customer that needs only any-to-any connectivity between all sites in a VPN, a rather common design, the value 000 can be used. An example RT within AS 65001 and a customer ID of 654 would be 65001:654000. A hub/spoke VPN would require at least two RTs:

1. 65001:654001 upstream connectivity exported by the hubs and imported by the spokes
2. 65001:654002 downstream connectivity exported by the spokes and imported by the hubs

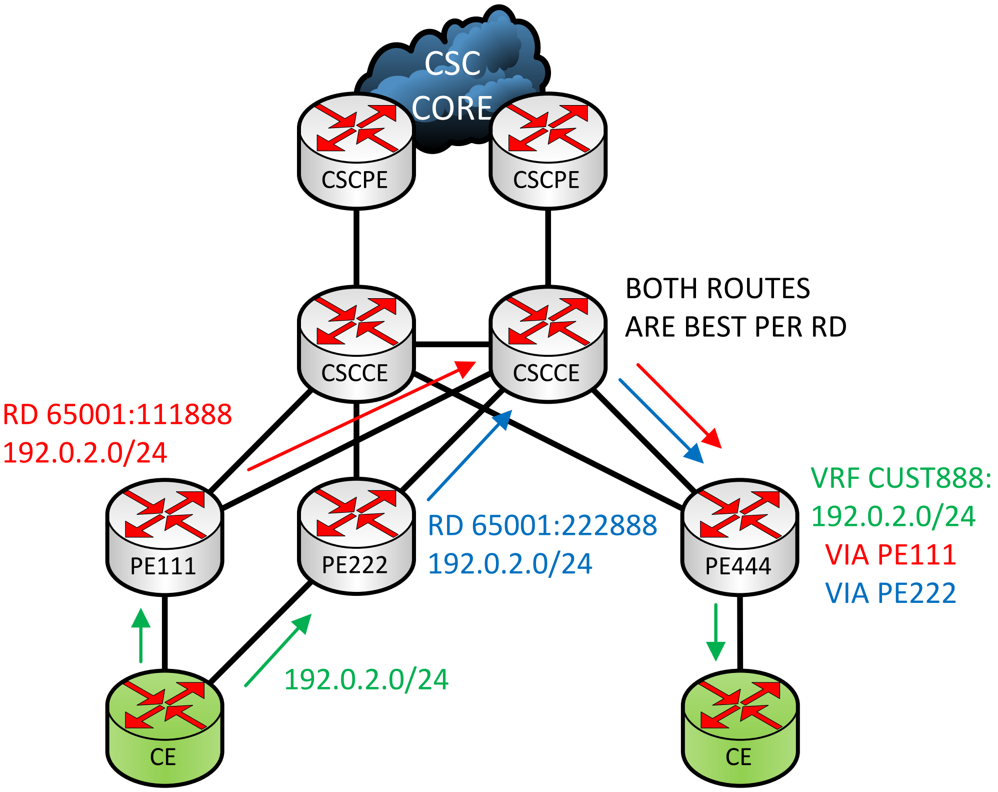
Like the RD allocation, this allows for up to 1,000 customers and 1,000 RTs per customer. As a brief foreshadow, this document uses inter-AS connectivity extensively, so using the BGP ASN as part of the RT for all customers may not be suitable. For some environments, this may lead to unnecessary RT configurations and other administrative burdens (i.e., needing to import N-1 RTs just to form a basic inter-AS any-to-any VPN). Consider using the customer ID for the first 32 bits, or using a generic BGP ASN, such as 65000, for all RTs in the greater network. The diagram below shows how RDs and RTs work together to form MPLS L3VPNs within a POP. Sometimes it's a good idea to use the customer ID from the RT as the VRF ID in the RD (orange 777 and green 888) as shown here. Assume the POP is in BGP ASN 65001.

Figure 11 - Building MPLS L3VPNs within a POP



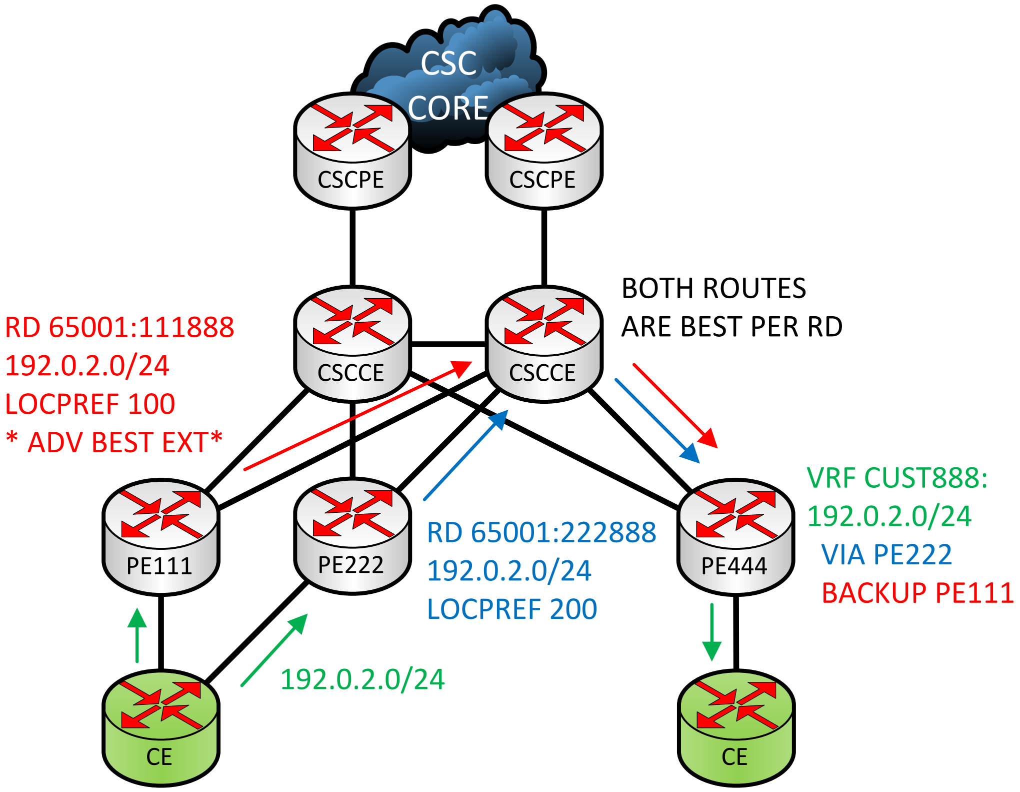
As it relates to availability, using unique RDs on every PE allows the BGP RRs to retain all of the routes. This is useful for multi-homed sites because all of the egress PEs will receive the same routes from the customer and advertise them to all BGP RRs in the cluster. Because the routes are distinguished (different), BGP best-path on the RR does not compare them. All of the paths will be best in their own RD-indexed tables, so the RR can reflect all of the routes towards the ingress PE. The ingress PE can install all of them, either for active-active load sharing or active/standby fast failover. While there are other, more complicated solutions to this problem (shadow session, shadow RR, and BGP additional-paths capability), years of operational experience suggest that using unique RDs is a reliable and effective choice for MPLS L3VPNs. The ingress PE simply needs to import both routes and install them into the routing table using ECMP. The diagram below illustrates the active/active design.

Figure 12 - Unique L3VPN RD for Active/Active Forwarding



Perhaps counter intuitively, the active/standby design takes more effort to design and implement. First, a primary link must be chosen, typically by setting the BGP local-preference inbound on the primary egress PE to be greater than the BGP local-preference applied on the alternate egress PEs. The alternate egress PE needs to be configured to advertise its best external route to the RRs. The eBGP route from the customer won't be the best path as the high local-preference iBGP route will win. The BGP RR doesn't care, because the unique RDs ensure that these routes are separate and thus are not compared. Both are advertised to the ingress PE, which imports both into the VPN routing table. BGP best-path runs on the ingress PE, and the device chooses the route through the primary egress PE with the higher BGP local-preference. The second best-path is installed as a repair-route, a pre-programmed backup that can be switched on if the primary route is withdrawn or is otherwise unreachable from a control-plane perspective (e.g. BGP next-hop is inaccessible). The diagram below illustrates the active/standby design.

Figure 13 - Unique L3VPN RD for Active/Standby Forwarding



2.1.6.2. Layer-2 VPN

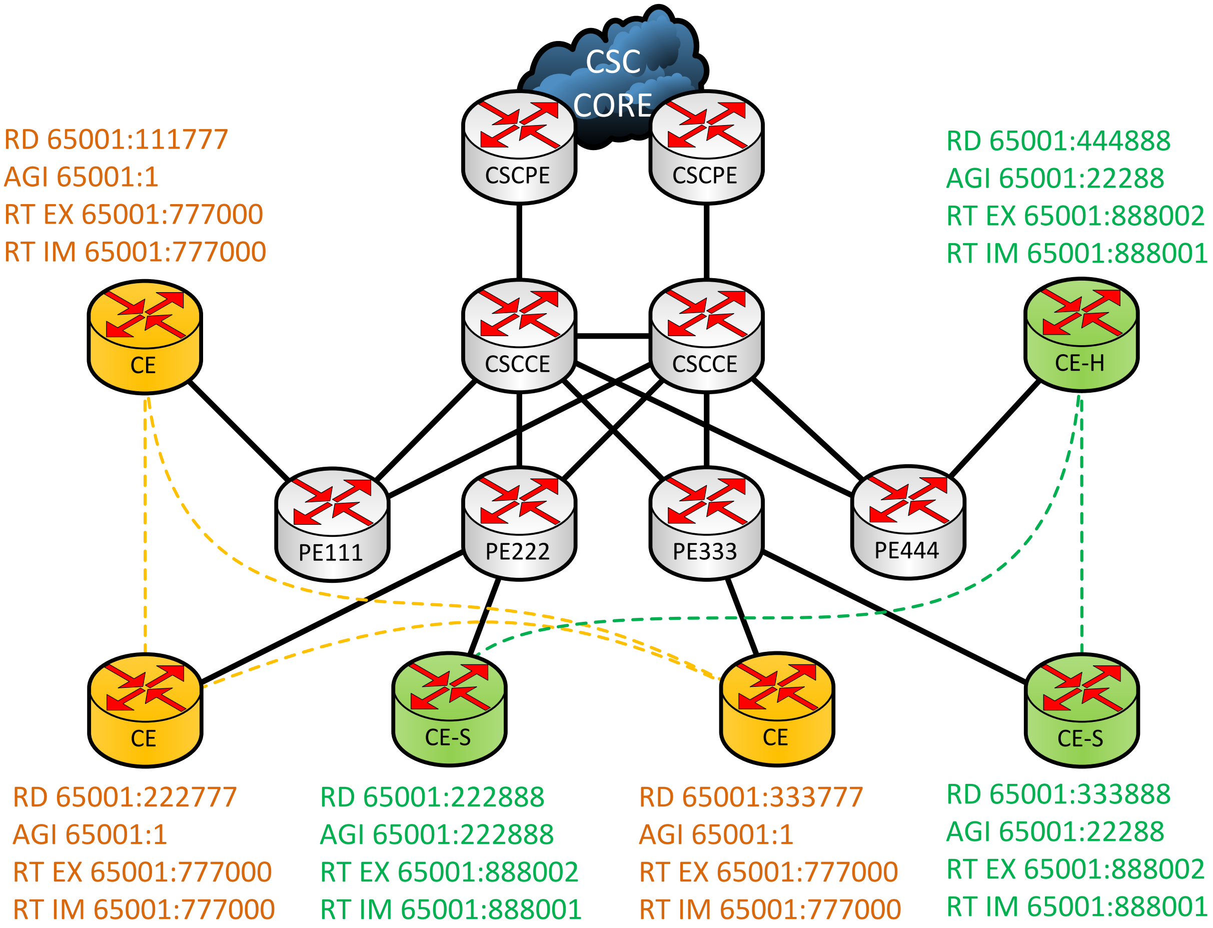
MPLS Layer-2 VPNs (L2VPN) provides switched/bridged connectivity across the network which logically collapses the entire MPLS core into a single switch/bridge from the customer's perspective. CE devices do not exchange any routing information with the PE devices. As a result, L2VPNs are popular because customers can use any routing protocol they want (including non-IP protocols such as IS-IS). Customers can also roll out new services internally, such as multicast or IPv6, without needing to coordinate anything with the service provider.

As discussed in the BGP VPN services design section, we opted to use BGP for VPN endpoint discovery and LDP for pseudowire signaling. This allowed us to create Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS) networks for our customers with minimal static configuration. The concepts of RD and RT also exist in the context of BGP-based VPLS.

The concept of RD remains the same, although L2VPN doesn't have "routes" per se. The "route" is the MPLS endpoint IP address, typically the BGP next-hop, prefixed by the 64-bit RD. The concept of RT, although "route" is a misnomer, is also similar to L3VPN. When RTs are imported, a specific VPN endpoint establishes an LDP-signaled pseudowire to all PEs that exported that RT.

An additional extended community, known as the L2VPN attachment group identifier (AGI) is also included. This is based on the BGP ASN and the operator-specified VPN ID, and must match in order for RTs to be imported. It's a way of controlling high-level VPN membership while the RT determines the precise connectivity within a given VPN. Note that the only difference between VPWS and VPLS is the number of endpoints. VPWS is a point-to-point connection and would likely be configured as an independent AGI with the same RT imported and exported by both nodes. Extending this design to 3 or more nodes, using the same AGI/RT strategy, would create a full mesh of pseudowires between all nodes in the VPN. Adjusting the RTs to create a hub/spoke VPN or other custom topology is also possible and may provide improvements in security and scale. The diagram below illustrates the high-level design and operation of customer L2VPNs. It is common for the AGI to be the same as the RD, but it doesn't have to be. More importantly, the AGI will need to be manually adjusted for inter-AS VPNs because the AS number (first half of the AGI) will cause a mismatch, and the VPN cannot form.

Figure 14 - Building MPLS L2VPNs within a POP



To better explain the services offered, the table below explains the three main characteristics. Each category can operate in either “wire mode” (VLANs are transparent) or “VLAN mode” (VLANs are mapped to service instances). Also, note that VPWS is roughly synonymous with

Ethernet Private Line (EPL) style services and VPLS is roughly synonymous with E-LAN style services.

Figure 15 - L2VPN Services Offered

	EPL / EVPL	EP-LAN / EVP-LAN	EP-TREE / EVP-TREE
Number of nodes	2	At least 3	At least 3
Connectivity style	Point-to-point bridge	Any-to-any switch	Hub/spoke, similar to private VLANs
Route-target design	One RT, import and export on all nodes	One RT, import and export on all nodes	Two RTs, swapped import and export on hubs and spokes

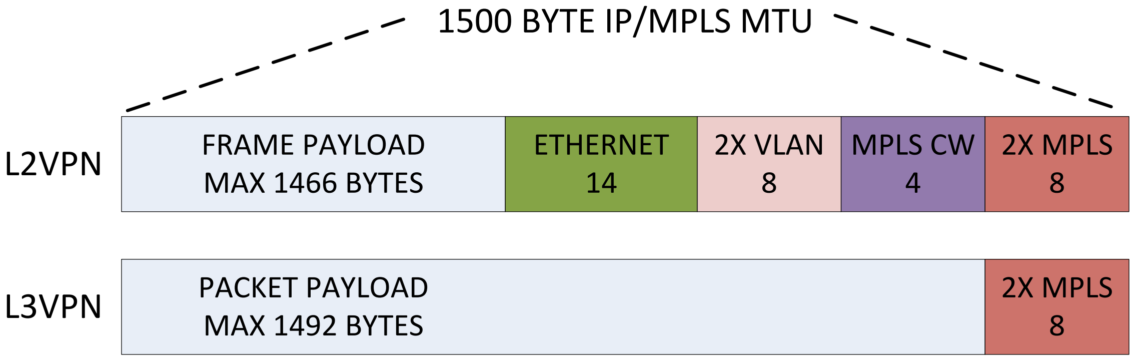
Some readers may be wondering why we chose LDP-signaled over BGP-signaled VPLS. The former is operationally simpler to understand and has better OAM capabilities (at least on Cisco IOS) than the latter. Understanding how label blocks and virtual offsets are computed in BGP-signaled VPLS requires expert-level networking skills, which was in short supply within our organization. On a technical level, not all vendors support setting the C-bit, signaling the inclusion of an L2VPN control word (CW). The lack of a control word has several well-known drawbacks: no ability to include sequence numbers, frames within a given pseudowire taking different paths in the network, and more. LDP-signaled VPLS avoids these issues entirely.

Although MTU is worth considering in any network, it is especially important for L2VPNs. In L3VPNs, by contrast, the only additional MTU overhead is the MPLS encapsulation which is entirely predictable and is typically two 4-byte shim headers for a total of 8 bytes. With L2VPNs, there are many encapsulation layers for which to account:

a. The MPLS encapsulation within the POP: 8 bytes for two MPLS shim headers
b. The pseudowire control-word: 4 bytes
c. Any customer VLANs retained (i.e. not popped) over the VPN: 8 bytes for up to 2 VLAN headers. This may not be relevant if you only offer VLAN-based services whereby all VLANs are removed at ingress.
d. Customer standard Ethernet header: 14 bytes

The total additional overhead becomes 34 bytes for L2VPN compared to 8 bytes for L3VPN. In our environment, we provided a full 1500 byte MTU to our customers over both L3VPN and L2VPN by using jumbo frames both intra-POP and inter-POP over CSC. If jumbo support is not available in your network, it is imperative that your customers know the precise MTU that is available. Ignoring the upper-most layer-2 encapsulation (Ethernet in our case) within the POP, the diagram below illustrates how these two services differ with respect to MTU. The diagram also assumes the more difficult (and worse) case of only having a 1500 byte MPLS MTU.

Figure 16 - Calculating MTU for VPN Services



2.1.6.3. Multicast VPN

This section details the multicast VPN design for the network with a focus on intra-POP services and design constraints. Note that within a POP, most of these design challenges are meaningless, and any MVPN profile could be reasonably implemented. The difficulty arises when extending MVPNs over CSC, which is discussed later in this document. This section focuses primarily on the technical decision making process regarding the service offerings and implementation.

While there are many solutions for delivering multicast across MPLS networks in general, there are many constraints when performing both inter-AS MVPN and MVPN over CSC at the same time. Some technologies, like multicast LDP (mLDP) and point-to-multipoint RSVP-TE (P2MP RSVP-TE), can technically work in these cases, but not on all platforms, and certainly not without challenges. The most suitable solution is to use an MVPN design that has the following attributes:

a. Uses a default, non-partitioned MDT for full intra-VPN connectivity.
b. Uses BGP, at a minimum, for discovery of the MDT PEs across the VPN.
c. Does not require PIM RPs anywhere in the customer carrier global table.
d. Accomplishes one of the following:
    a. Uses PIM for SP multicast signaling and GRE for SP multicast encapsulation.
    b. Re-uses existing unicast LSPs if the core carrier doesn't support MVPN at all.

Given these constraints, the following options are available. Note that the "Profile ID" column is a Cisco-specific identifier, which is a shorthand for identifying MVPN design options.

Table 1 - Plausible MVPN Profile Options

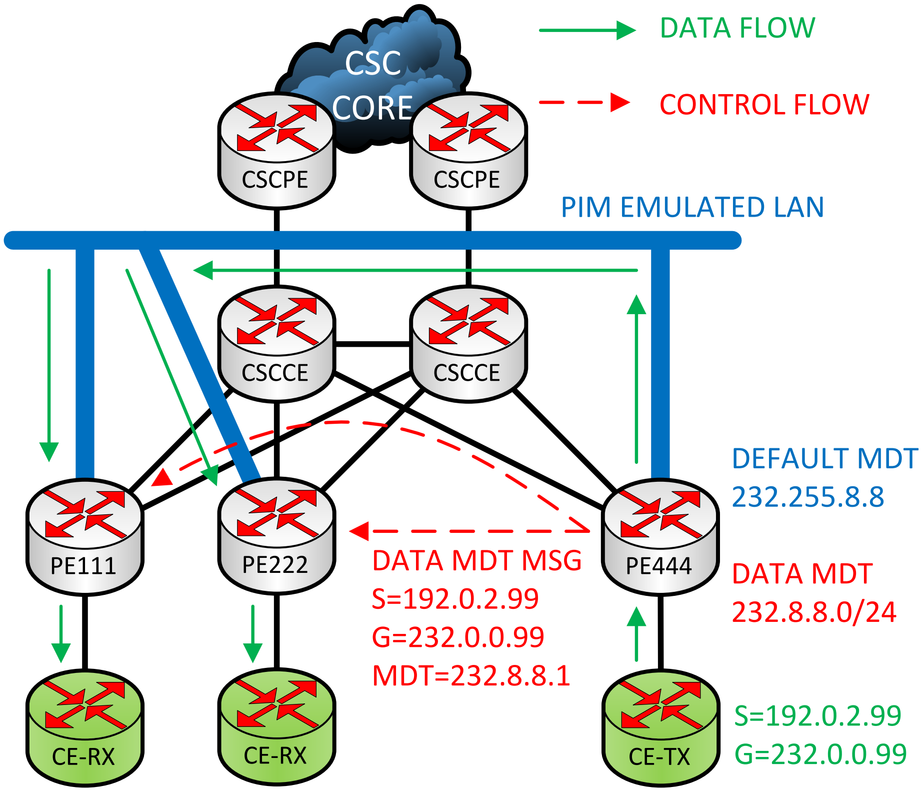
Profile ID	PE Discovery	Core signaling/encap	Customer signaling
0	BGP IPv4 MDT	PIM/GRE	PIM overlay
3	BGP IPv4/v6 MVPN	PIM/GRE	PIM overlay
11	BGP IPv4/v6 MVPN	PIM/GRE	BGP IPv4/v6 MVPN

19	BGP IPv4/v6 MVPN	Ingress Replication	PIM overlay
21	BGP IPv4/v6 MVPN	Ingress Replication	BGP IPv4/v6 MVPN

This document will use the Cisco-specific profile numbers for the sake of brevity. Profile 0 is the classic "Draft Rosen" technique that uses a dedicated, inflexible BGP address-family to advertise PE loopbacks between devices in a given VPN. This source discovery process obviates the need for PIM RPs in the network and enables SSM to be exclusively deployed for customer multicast transport, even for default MDTs. In real life, this is the option we chose, as it was widely supported, well-documented, and commonly used in production networks for years. The main drawback of this approach is that, assuming no other multicast-related BGP sessions are established, all VPNs must use profile 0 regardless of their connectivity requirements. This can be limiting for future operations.

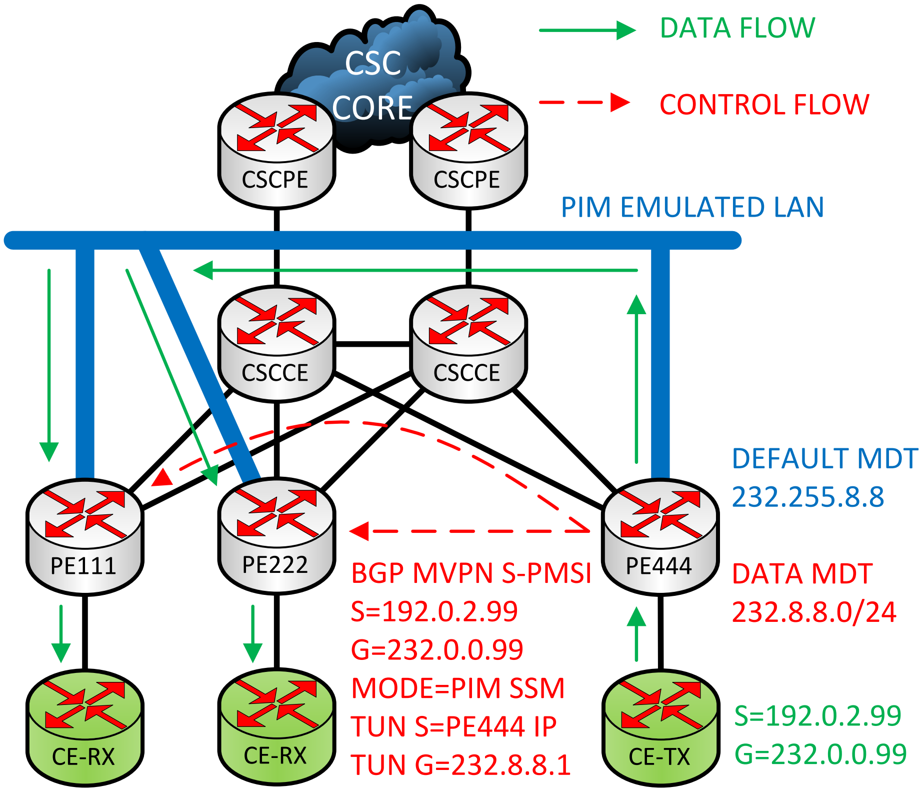
Because BGP is only used for remote PE discovery, switching over to more optimal MDTs for high-bandwidth flows (called "data MDTs" in Cisco parlance) was handled within the PIM overlay. These selective MDTs are a subset of the larger inclusive, default MDT. The ingress PE (the one connected to the source) signals this using a special PIM message named "Data MDT Join". Each tree describes a different provider multicast service interface (PMSI). The Inclusive PMSI (I-PMSI) represents the default MDT and the Selective PMSI (S-PMSI) represents the individual data MDTs. The diagram below represents how they components fit together within profile 0.

Figure 17 - MVPN Profile 0 Design



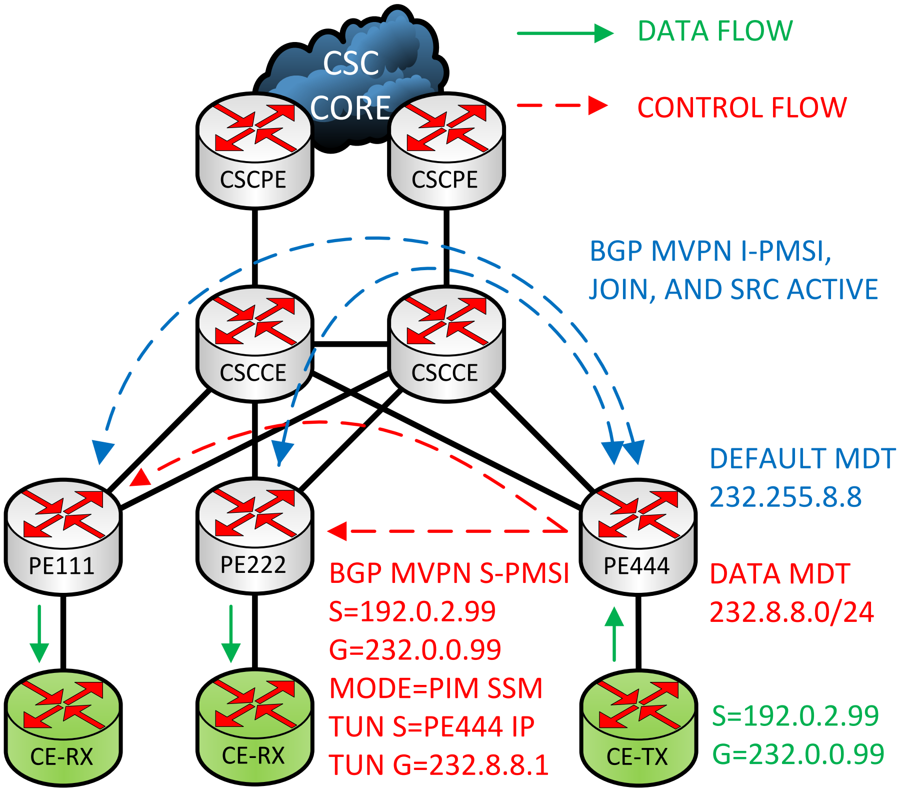
Profile 3 operates almost identically to profile 0 except it uses a different BGP address-family. Its purpose is still limited to source discovery, but BGP is now capable of signaling more than just IPv4 addresses to serve as SSM sources. By building the network using profile 3 in the first place, the designer can retain the benefits of a stable, known-good solution while also being prepared for future requirements. If we were rebuilding the network today, this is the profile we would have likely chosen for the vast majority of customer VPNs. All of the customer multicast signaling still uses PIM over the inter-PE emulated LAN, and while this limits scale, it is simple to understand and operate. In this context, there are two BGP messages. One is used for I-PMSI endpoint discovery and is used to construct the default MDT. The other is used to signal S-PMSI switchover events as they occur. The S-PMSI message is comparable to the "Data MDT Join" message, but contains a bit more contextual data about the tunnel itself. The diagram below illustrates how they messages work with MVPN profile 3.

Figure 18 - MVPN Profile 3 Design



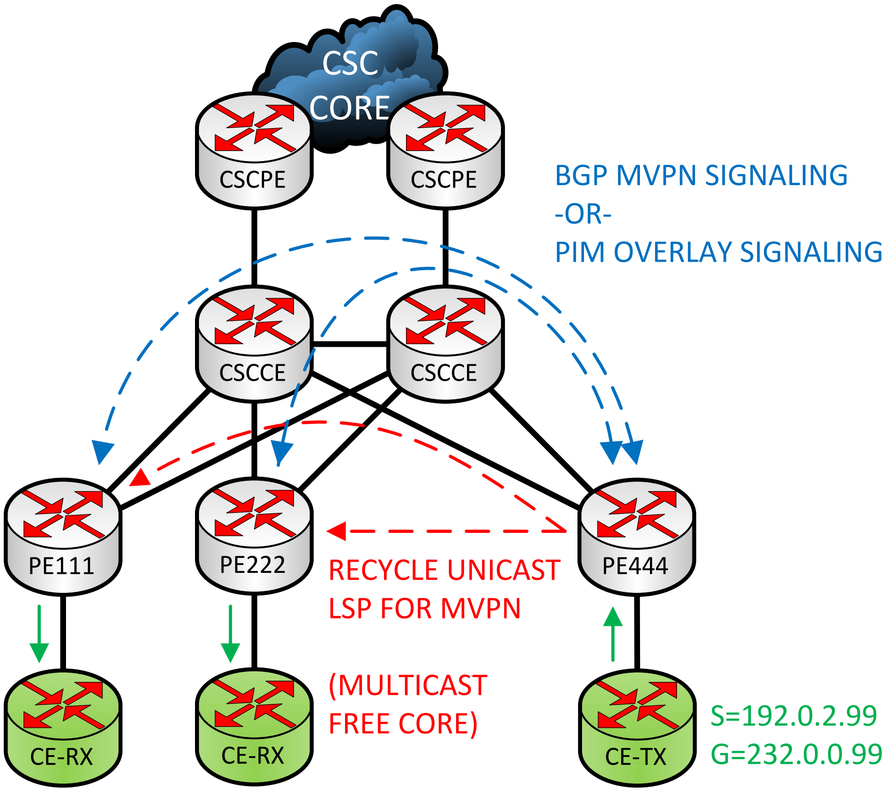
Suppose a customer has hundreds of routers in a multicast VPN, most of which are senders and receivers. Having all of these neighbors on an emulated LAN exchanging soft-state PIM messages would scale poorly. Profile 11 addresses this by using the modern BGP MVPN address-families to signal customer multicast information. This obviates the need for a PIM overlay, relying on BGP both for PE discovery and customer multicast signaling. This solution is complicated and not commonly deployed, but the solution is at least technically possible when BGP IPv4/v6 MVPN is used instead of BGP IPv4 MDT. The technical nuances behind how this signaling works is beyond the scope of this whitepaper, but in summary, BGP uses different messages that roughly correspond with PIM (*,G) join, PIM (S,G) join, and PIM register messages. Withdrawing a BGP MVPN NLRI relating to (*,G) or (S,G) state is comparable to a sending PIM prune. The diagram below illustrates the high-level operation of profile 11.

Figure 19 - MVPN Profile 11 Design



Some carriers do not support MVPN at all, providing only unicast transport. This could be true for the customer carrier POPs or the broader CSC transport network. Ingress Replication (IR) allows MVPNs to use existing unicast LSPs for multicast transport by replicating multicast traffic at the ingress PE. While this is highly inefficient and defeats the purpose of multicast in general, it can be useful for low-bandwidth applications. For example, in our environment, we had an application that dynamically discovered its peers using multicast, not DNS like most applications would use. This was a very low-bandwidth flow, and if our core carrier did not support MVPN, it would have been an appropriate choice for some of our customers. Profiles 19 and 21 differ in that one uses PIM overlay emulation while one uses BGP MVPN for customer multicast signaling. The diagram below illustrates the high-level operation of IR MVPN without differentiating between customer multicast signaling types.

Figure 20 - MVPN Ingress Replication Design (Profiles 19 and 21)



It is useful to briefly consider a future where CSC no longer exists. Suppose the entire transport network is converted to E-LAN because all sites suddenly have access to Ethernet last-mile uplinks. Now, the possibilities for MVPN are broadened, assuming BGP MVPN IPv4/v6 have been deployed, providing even more service offerings for customers. This is yet another reason to deploy the modern BGP control-plane, even if there is no immediate operational benefit.

2.2. Carrier Supporting Carrier (CSC) Design

This section details how our POPs (customer carrier) integrated with the CSC network (core carrier). There are endless ways to achieve this, but this document will focus primarily on the decisions we made in real life.

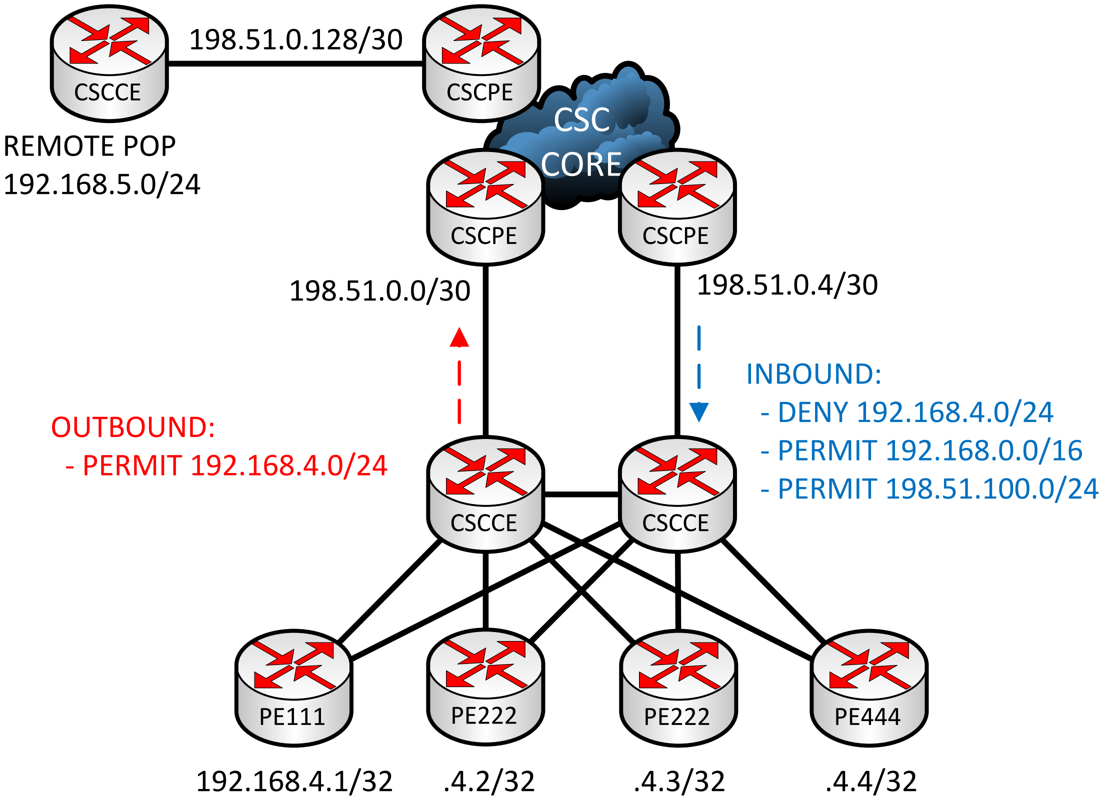
2.2.1. BGP Labeled-unicast (BGP-LU) Connectivity

To connect the CSC-CE (our device) to the CSC-PE, we used BGP-LU for IPv4. The purpose of this connection is to exchange transport prefixes between POPs. This is sometimes limited to PE and RR loopbacks, but in our design, we allowed all prefixes, including point-to-point transit networks, to be exchanged. This simplifies troubleshooting for operators when using simple tools like “ping” and “traceroute”, providing full reachability between all customer carrier IP networks. Global IPv4 routing table bloating was not a concern in our network of a few hundred sites, each of which contributed only a handful of IPv4 subnets. The CSC carrier did not seem to care, either. Some CSC carrier's limit the number of prefixes received from a CSC-CE, however.

To secure the BGP control-plane, we applied BGP route filters on these CSC-CE to CSC-PE sessions. Applied inbound, we denied any local POP networks. For example, if a POP was using the 192.168.0.0/24 address space for its various transit links, device loopbacks, and global management networks, we could block that entire range, including longer matches. This guarantees that even if the BGP AS-path loop control mechanism breaks down due to a core carrier misconfiguration, the POP will never learn its own prefixes via eBGP. Then, to prevent any route leakage from other customers that the core carrier may be servicing, we permitted the remaining POP networks, say 192.168.0.0/16, capturing all the other sites. We also permitted the CSC-CE to CSC-PE transit links, which were subnets provided by the core carrier, but aggregated nicely into an easily-matched prefix, such as 198.51.100.0/24. Again, this provided full connectivity but significantly reduced any remote possibility of a routing problem.

Applied outbound, we permitted only local POP networks. Using the prefix example above, that would be 192.168.0.0/24. POPs are never meant to be transit sites except in uncommon situations where a satellite is "tethered" to a regional POP temporarily. We used this strategy when needing to get a POP online that was physically near a region POP, but for which the CSC circuit was not yet provisioned. In that rare case, the satellite POP's transport prefix would be added to the permitted outbound filter. The diagram below illustrates these simple BGP filters and their positive impact on BGP's stability.

Figure 21 - eBGP-LU Inbound and Outbound Filters



2.2.2. Interaction Between IGP and BGP-LU

Our CSC provider only offered eBGP-LU as a CSC-PE to CSC-PE protocol, so extending IGP to the carrier was not an option. Although this would have simplified our device configurations, it would have increased operational complexity for both the core and customer carriers given the lack of control between peers. However, there are two broad design strategies for extending CSC-learned transport routes from the CSC-CEs to the PEs in each POP.

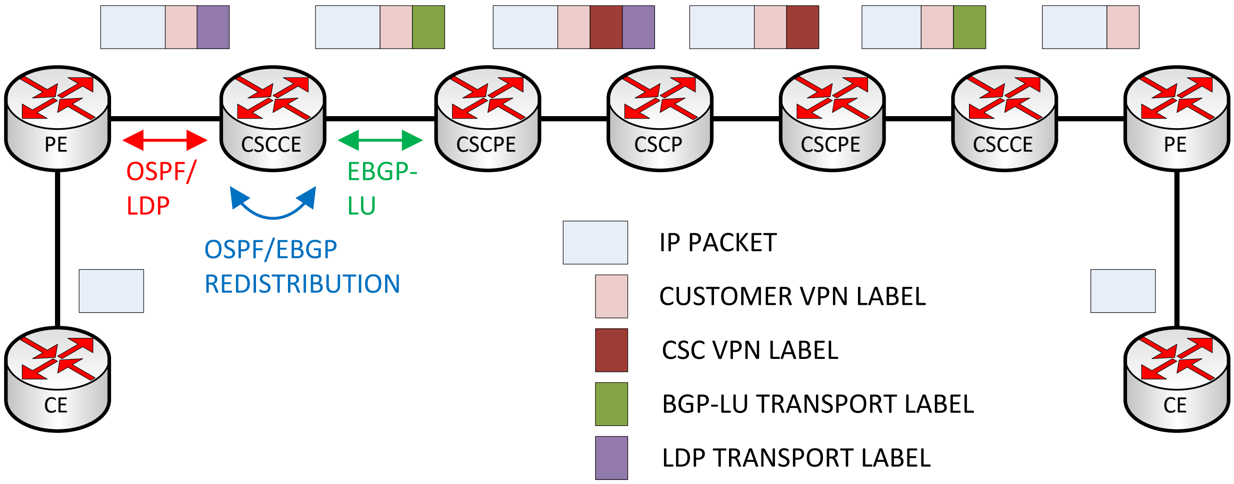
a. Mutual redistribution between IGP and BGP-LU at the CSC-CEs
b. Run iBGP-LU from CSC-CEs to PEs

In terms of implementation difficulty and technical understanding, mutual redistribution is far simpler. Because our POPs were all Cisco IOS-based and OSPF was our IGP of choice, the redistribution between BGP and OSPF was especially easy. By default, OSPF external routes are not redistributed from OSPF into BGP, which is a strong deterrent against routing loops. Because each of our regional POPs had two CSC-CEs with physically disjointed fiber uplinks to two separate CSC-PEs, this loop prevention was important. Configuration-wise, the customer carrier would not need to add any loop prevention filters. The PEs within the POP would learn all inter-POP transport routes as OSPF external routes. The label stack depth for intra-POP and inter-POP customer is remains 2. When a packet arrives at the ingress PE, the router:

a. Performs routing lookup on the VPN destination prefix and pushes BGP VPN label
b. Next-hop is IGP-learned, performs routing lookup to push IGP transport label

The diagram below illustrates this high-level flow for a sample prefix across CSC.

Figure 22 - Inter-POP Flow with IGP/eBGP Redistribution



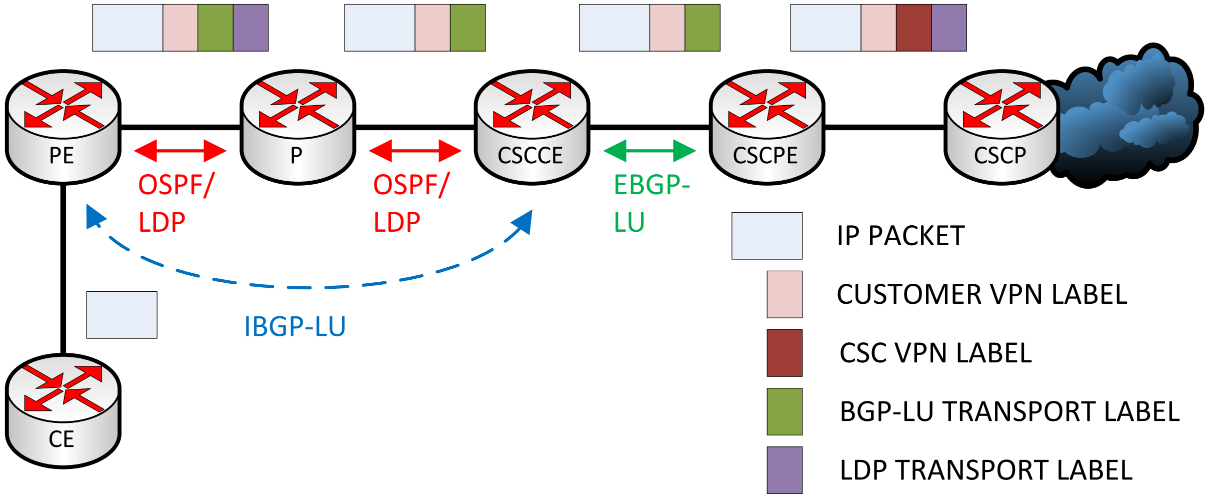
Extending BGP-LU to the PEs using iBGP sessions is a more advanced, complex solution that may offer benefits for some customers. First, there is no redistribution, so there is no possibility of a routing loop. Also, it helps separate intra-POP routing from inter-POP routing from the perspective of each PE, using IGP for the former and iBGP for the latter. This separation may simplify deploying a non-LDP labeling method, such as Segment Routing and RSVP-TE. Minor BGP tuning, such as whether to use next-hop-self on the CSC-CEs towards the PEs has its own set of trade-offs. Such a design is similar to seamless/unified MPLS. The benefits end here, but there are many drawbacks.

Because there is an additional level of indirection (i.e., another routing lookup) on each ingress PE for inter-POP customer traffic, a third label must be imposed in addition to the standard “transport” and “VPN” labels. When a packet arrives at the ingress PE, the router:

c. Performs routing lookup on the VPN destination prefix and pushes BGP VPN label
d. Next-hop is iBGP-learned; performs routing lookup to push iBGP-LU transport label
e. Next-hop is IGP-learned, performs routing lookup to push IGP transport label

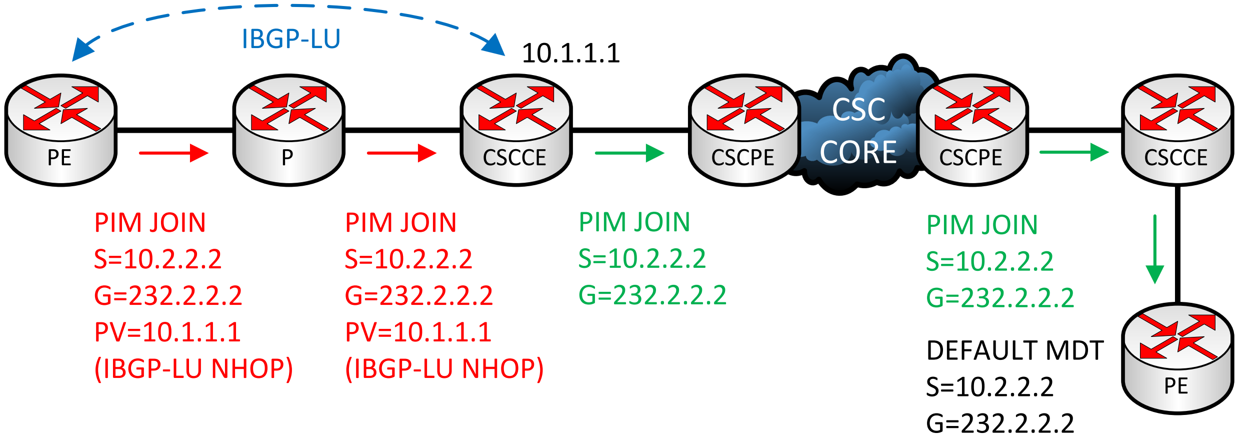
Consider POPs that have dedicated P routers, such as those arranged in a leaf/spine fashion. The P routers (spines) only have IGP routes for the local POP and are unaware of CSC’s existence entirely. It’s basically a BGP-free core within the customer POP, but this raises new problems. The first and most obvious issue revolves around Maximum Transmission Unit (MTU). Although an extra MPLS shim header only adds 4 bytes of encapsulation, architects must take care to account for this between PEs (leaves) and Ps (spines). The diagram below illustrates the label stacking process for this design.

Figure 23 - Inter-POP Flow with iBGP-LU from CSC-CE to PE



More significantly, inter-POP multicast-VPN traffic (discussed in depth later) is more challenging when PIM is used for provider multicast signaling because these P routers cannot perform RPF lookups for remote PEs. Instead, the PEs must insert a proxy vector into their join messages which specifies the CSC-CE to which the P router should continue building the tree. This is a destination for which the P router can perform an RPF lookup. Not all platforms can originate or interpret PIM proxy vectors. The diagram below illustrates how an inter-AS MVPN might work at a high-level, using the PIM proxy vector.

Figure 24 - Originating the PIM Proxy Vector for P Router RPF



Lastly, using iBGP-LU is significantly more difficult to design, implement, operate, and troubleshoot than CSC-CE eBGP-LU/IGP redistribution. It was not a difficult choice for us; we chose to redistribute, knowing that the likelihood of loops was infinitesimally small.

While we had many regional POPs with dedicated CSC-CEs, Ps, and PEs, some POPs were just a single router performing the CSC-CE and PE functionality together. These routers did not run IGP as there was no reason, but each router had at least one eBGP-LU uplink to at least one CSC-PE for connectivity to the regional POPs. The precise integration of these remote POPs and their BGP VPN connectivity is explained later.

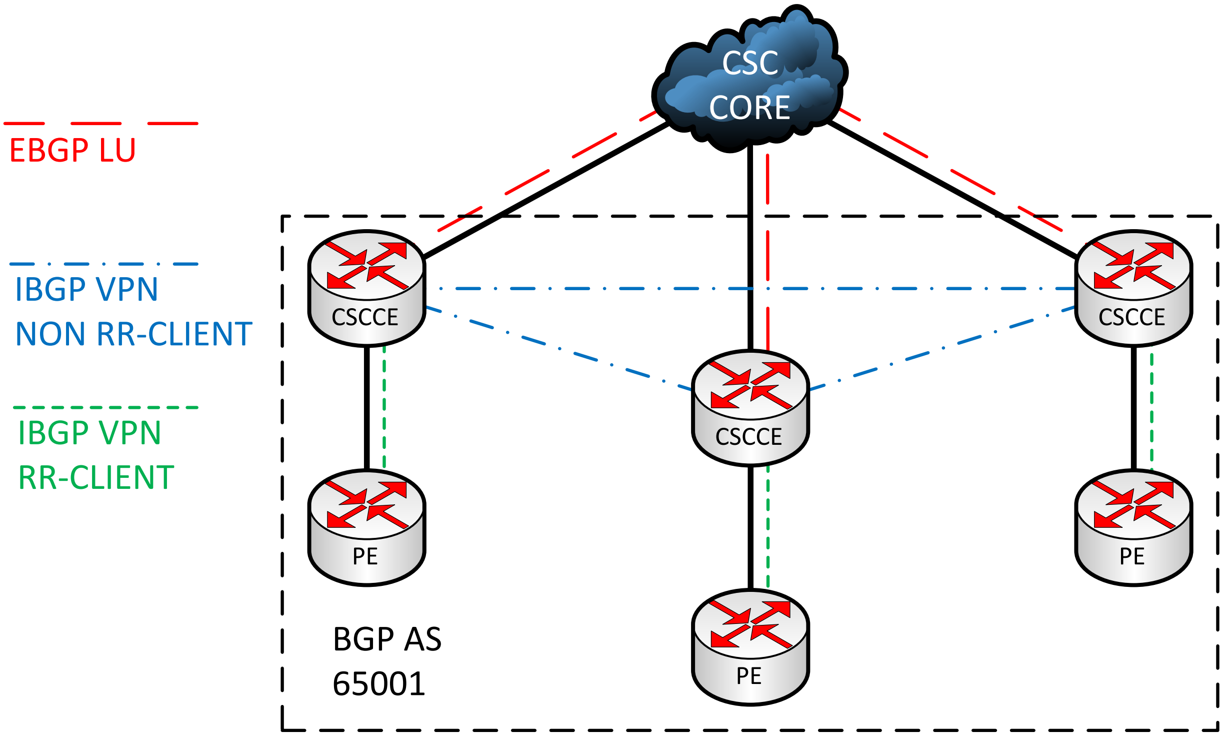
2.2.3. Inter-AS BGP VPN Services Routing

To provide customer connectivity across CSC, the PEs need to somehow share customer routing information. This section details the BGP VPN design and how the previously described services are extended over CSC. Although CSC is a rare design in the first place, most production designs use some form of iBGP connectivity between customer carrier POPs. Using iBGP has some advantages:

a. Only one AS number to manage, both for the customer and core carrier.
b. Simplified route-reflector configuration; only one peer AS.
c. Easily to conceptualize, troubleshoot, and expand without automation.

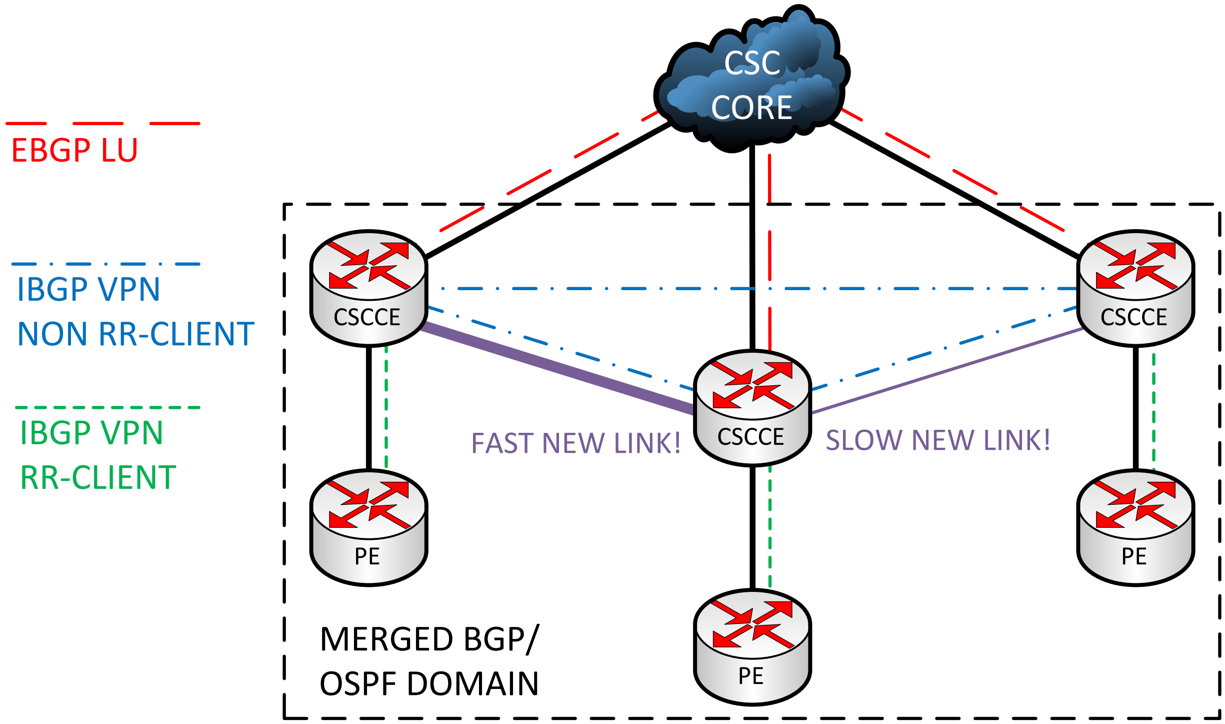
In production, we choose to use iBGP for these reasons. At the time, all operating processes were manual (no automation existed) and the BGP skill level of our engineers was low. The diagram below illustrates a basic iBGP VPN design whereby all RRs are fully-meshed over CSC.

Figure 25 - Basic iBGP Non RR-Client Mesh over CSC



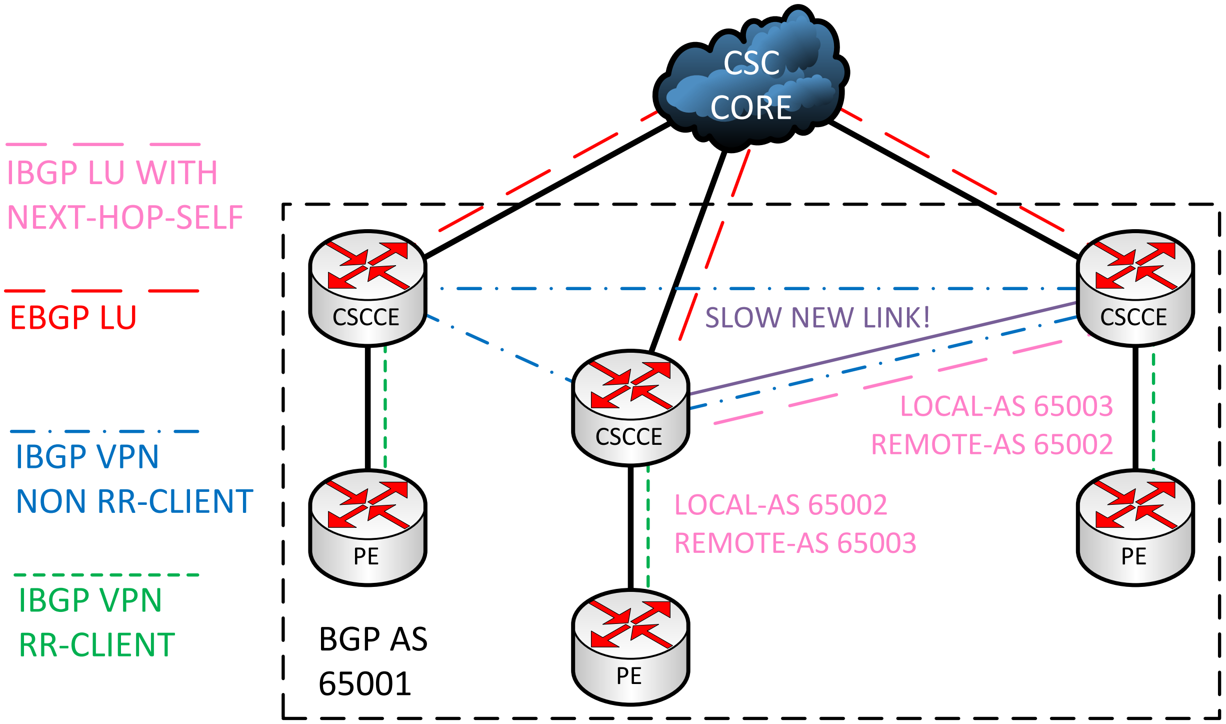
Over time, we observed a number of significant drawbacks of using iBGP over CSC in our environment as depicted above. First, any additional transport links between POPs, such as E-LINE/E-LAN services or dark fiber circuits, significantly complicated the design. Running IGP between the POPs is sloppy as routers now must decide between eBGP-LU and IGP routes for transport between PEs. This complicates redistribution (if used), filtering, flooding/failure domain boundaries, and more. Even more complications occur if the different POPs use different IGPs and different label distribution methods. More complex still is when multiple links exist between sites at various speeds, whereby some are faster than CSC and some are slower. The diagram below illustrates what this confusing situation might look like.

Figure 26 - Introducing Backdoor Links with Merged IGP Domains



Using eBGP between the POPs is clearly the superior approach for selecting the arbitrary "best" link when given a multiple options. Since all the POPs are in the same AS, this becomes complicated. While iBGP-LU can technically be transformed to work like eBGP-LU using a combination of local-AS and next-hop-self adjustments, it's a sloppy and unscalable workaround. The diagram below illustrates such an implementation, which we deployed in real-life to overcome urgent, uncommon circumstances. This was the catalyst for considering a new design.

Figure 27 - Introducing Backdoor Links with iBGP-LU and Local AS



When iBGP is used, the core carrier will typically replace all instances of the customer carrier's BGP ASN with their own BGP ASN when advertising transport routes between POPs. If POPs have direct connectivity between themselves, perhaps using backdoor links or alternative layer-2 transports, this could cause routing loops. Such loops would be rare, even with a mesh of backdoor links. Even if the AS-path length prevents the actual routing loop, the looped prefixes are still available where they should not be unless they are explicitly filtered elsewhere.

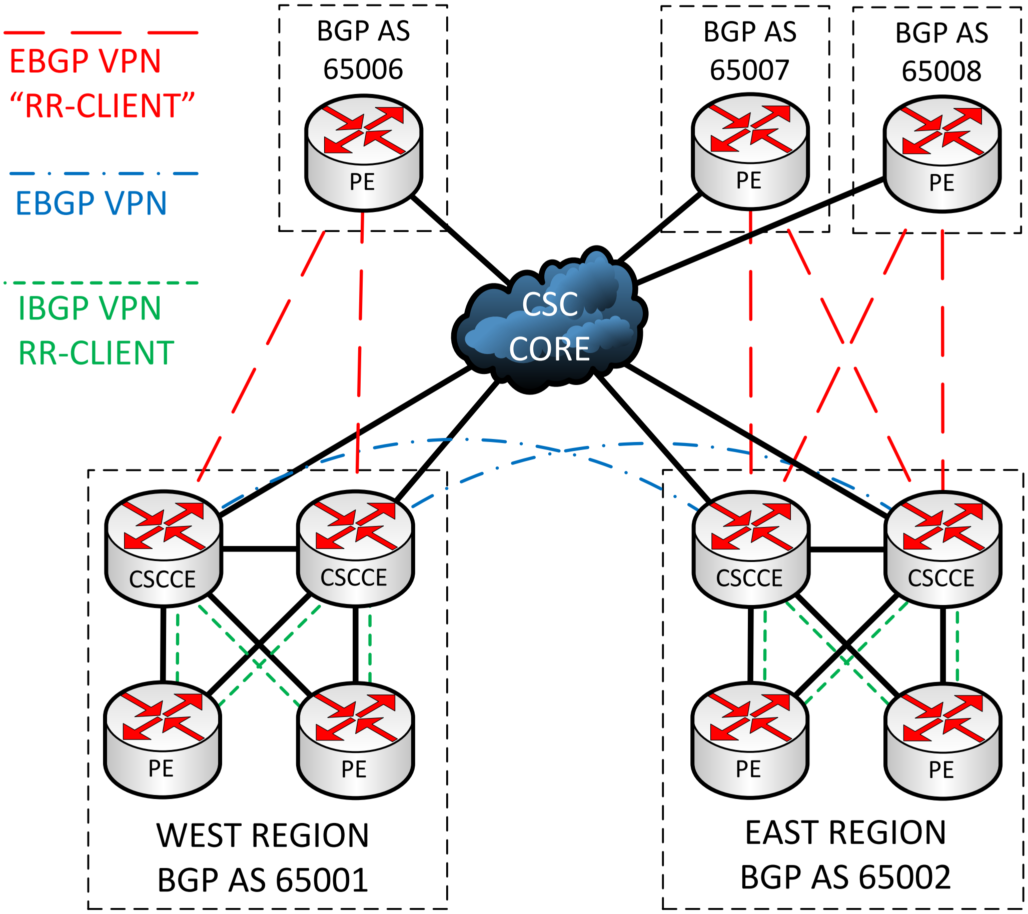
Placing each POP (regardless of size) into its own BGP AS alleviates all of these problems. While it does require some additional AS number management and slightly different CSC-PE configurations by the core carrier, the operational benefits far outweigh these administrative inconveniences. Because each POP has its own AS, the BGP VPN sessions between POPs will use eBGP. One of the main features of MPLS Inter-AS Option C is that it allows these BGP VPN speakers, often route-reflectors, to exchange eBGP routes without updating the BGP next-hop. The assumption is that the different AS' are already exchanging, at a minimum, all of the prefixes necessary for MPLS transport.

It is important to understand that a BGP "autonomous system" is just a logical construct that governs BGP behavior. It is not necessarily a different administrative or operational domain; all of the POPs in the network remained under the control of a single organization regardless of the BGP ASNs assigned. BGP confederations could also be used in the case where the core carrier demands that all POPs be in the same BGP AS (for operational simplicity on their part) while also gaining the advantage of confed-external peers between POPs. This is true for both labeled transport on backdoor links and VPN services between RRs.

Consider a network that has a mix of large POPs, both traditional and leaf/spine, and small, single-router POPs. All of these POPs are connected via CSC with full connectivity over functional LSPs. The small POPs should connect back to the closest regional POP that hosts a pair of RRs to service the PEs in that regional POP. The small POPs logically act like "satellite PEs", accessible over CSC, by connecting via BGP VPN to the RRs using eBGP. Because the RRs will not change the BGP next-hops when advertising VPN routes to these satellite PEs, they'll behave just like any other PE in the regional POP. The design mimics iBGP with respect to optimal MPLS forwarding as the RRs are not forced into the transit path while also overcoming the iBGP limitations.

The diagram below illustrates how satellite POPs connect back into their parent regions. For the sake of brevity, this section uses the quoted phrase "route-reflection" to describe eBGP behavior with respect to remote sites. This term is explained in greater detail later.

Figure 28 - Satellite POP Connectivity to Regional POP Using eBGP VPN

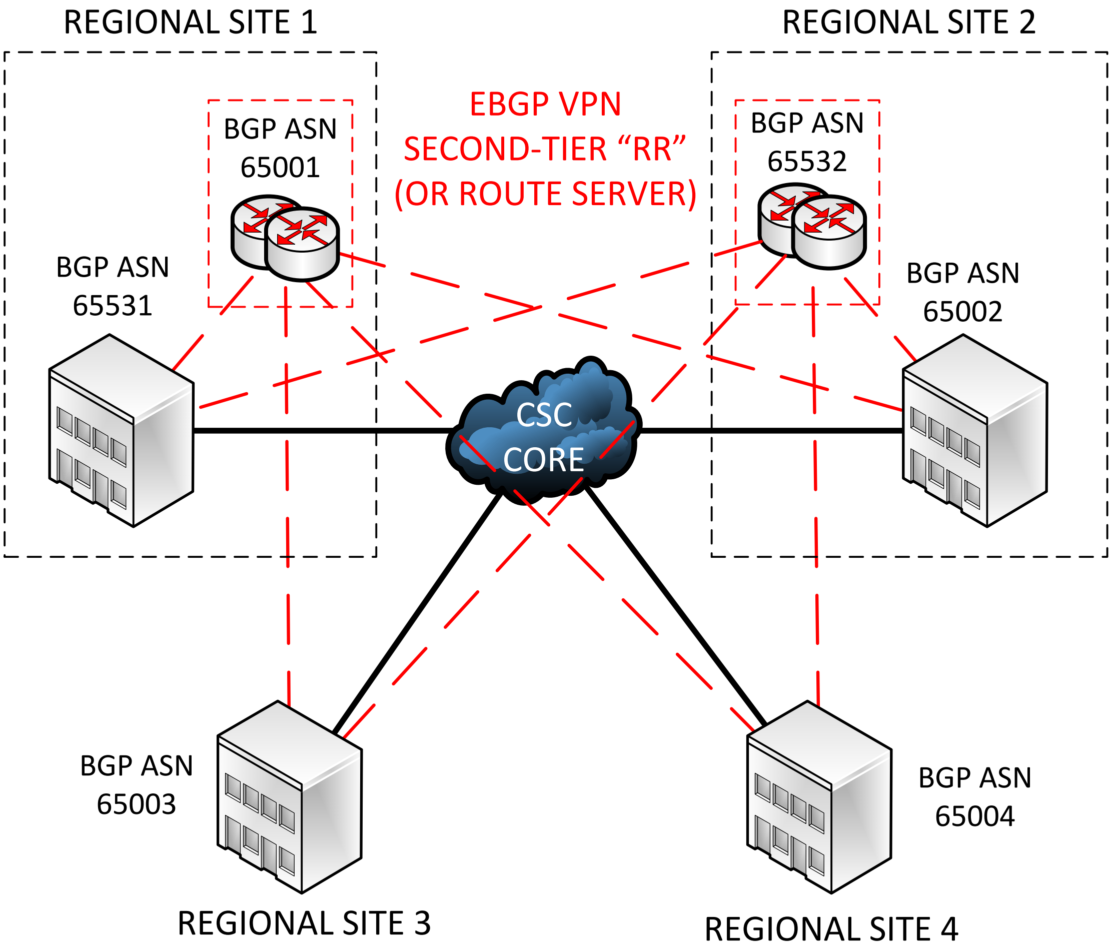


Inter-regional BGP VPN connectivity can be designed in one of two main ways:

a. Add a second tier of "route-reflection", except using eBGP instead of iBGP
b. Directly connecting RRs between regions using eBGP

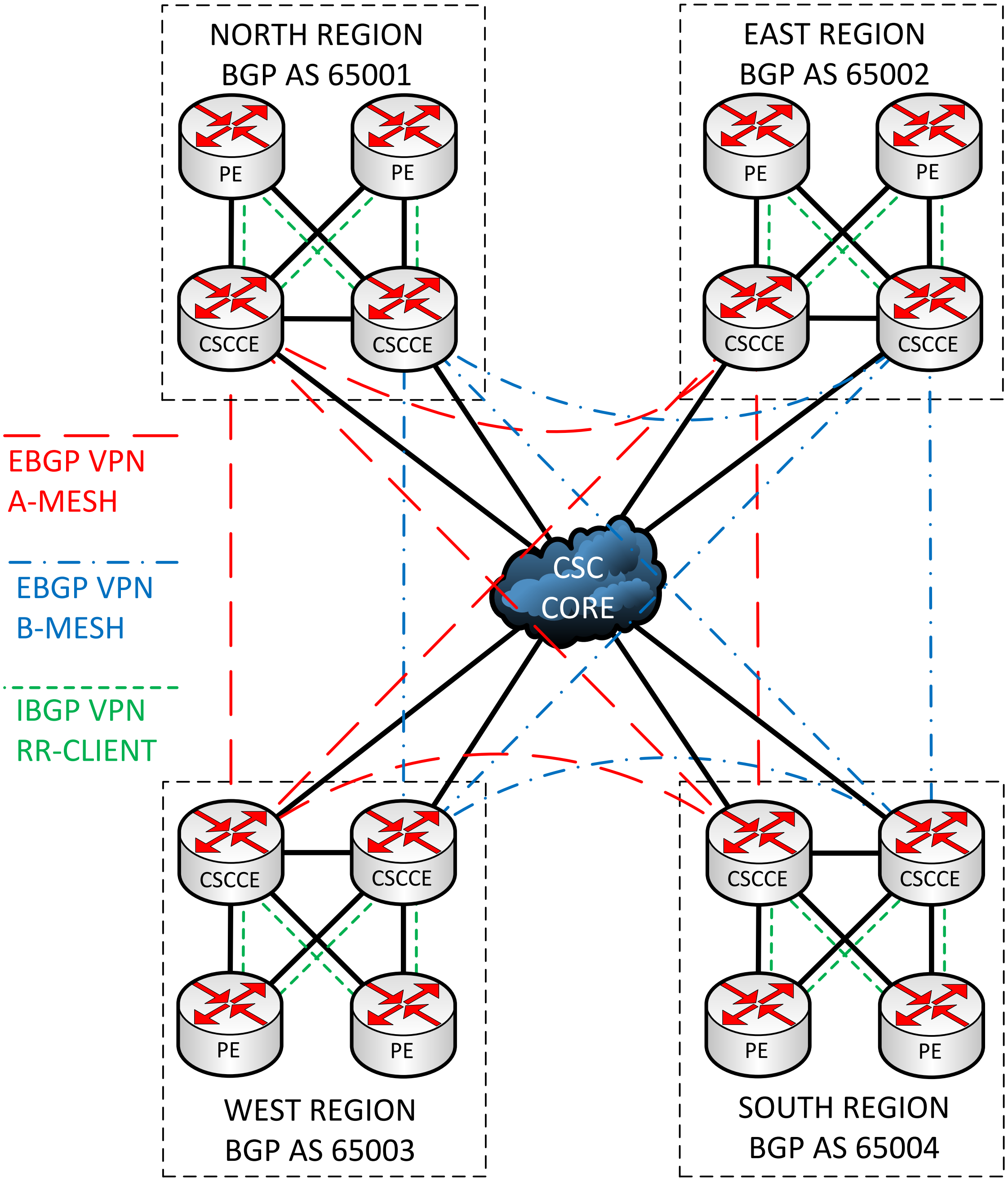
The advantage of a second tier of "route-reflection" is improved scale when the number of regions is very large. In our case, we only had 6 regions, so scale in this context was not a large concern. Additionally, these second tier "RRs" would have to be hosted somewhere, presumably in 2 of the 6 regions, creating points of failure. If those 2 regions went offline, the remaining 4 regions would not be able to exchange any BGP VPN routes, which was an unacceptable trade-off to gain a level of scale we didn't need. Additionally, these routers are not technically route-reflectors because all of their peers will be eBGP. As is common in Internet Exchange Points (IXP), these routers can be route-servers, which operationally act like route-reflectors with some minor technical differences regarding AS path recording. This design does, in fact, work with eBGP VPN address-families and is technically valid, although it strains credulity and would not be advisable to deploy in production without a compelling reason.

Figure 29 - Second-Tier "Route Reflection" with eBGP VPN Sessions



The density of the mesh between regional POPs is often dependent on the level of availability required balanced with the increased complexity and management burdens required to implement it. As discussed previously, there is a strict separation between the "A" mesh and "B" mesh within each POP. Each PE in the regional POP and every satellite PE that connects back to that regional POP forms a BGP VPN session to both the "A" and "B" RRs. These two meshes never connect directly together as it would add almost no improvements in availability but comes at a steep cost in terms of design conceptualization. This same logic extends from intra-POP to inter-POP. All of the "A" RRs can be fully-meshed over eBGP to implement Inter-AS Option C. Likewise, the same design applies to the "B" RRs, effectively creating two separate, parallel meshes for improved availability and fault domain isolation. Given that these eBGP sessions are multi-hop, just like iBGP sessions, individual CSC-PE to CSC-CE uplink failures won't affect these meshes because each "side" can still route across alternative uplinks. Only when a POP is completely cut off from the core carrier will these sessions fail. Unlike the second-tier of "route reflection" (really, route-servers or just multi-hop/next-hop-unchanged eBGP behavior), any pair of regions, including their satellites, can communicate across CSC provided they have connectivity as there are no inter-regional dependencies. The diagram below illustrates the BGP VPN mesh design.

Figure 30 - eBGP VPN Inter-region Mesh Design



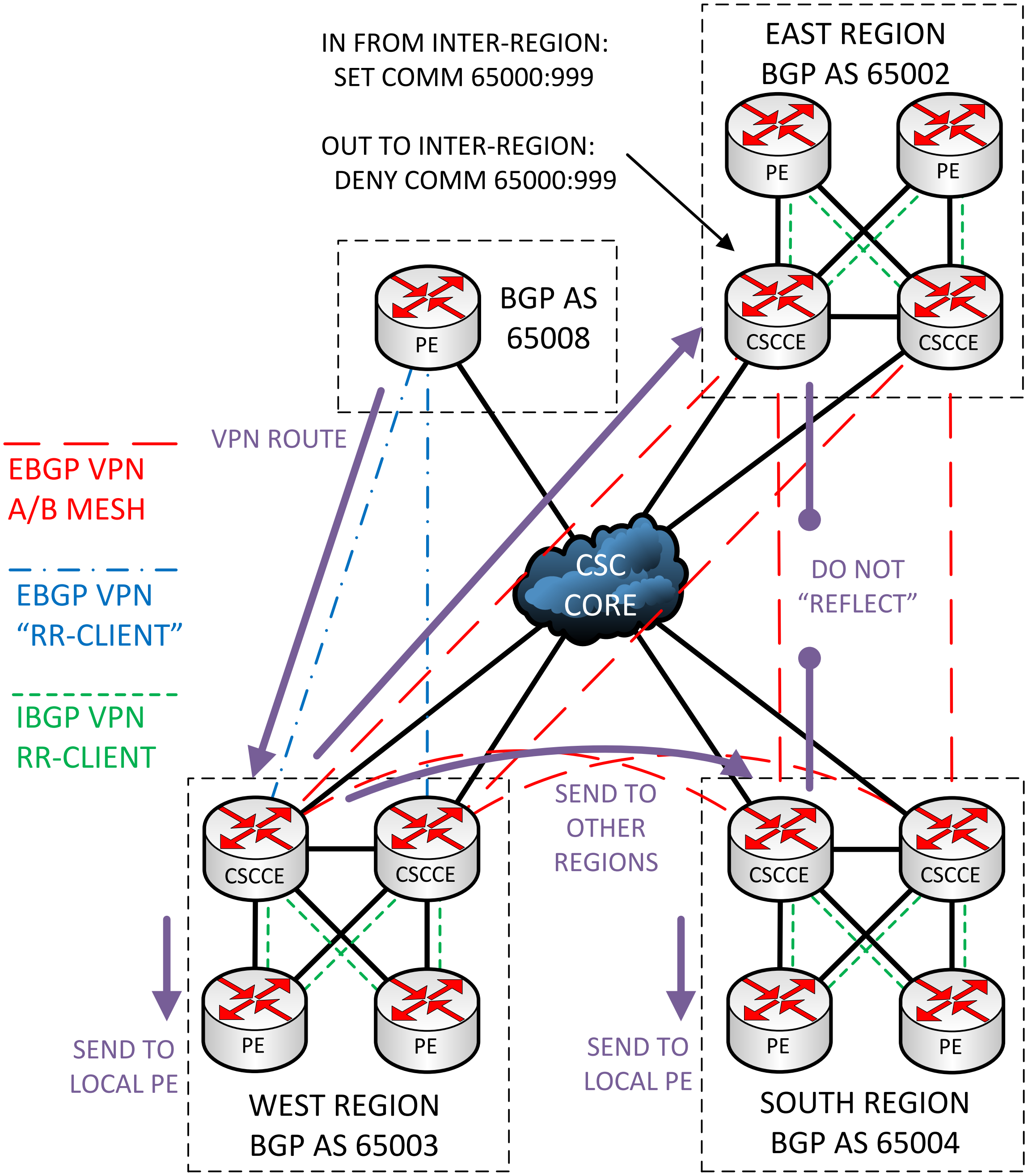
As it relates to prefix filtering on these inter-regional POP BGP VPN sessions, it only makes sense to advertise regional routes to other regions. The word "regional" in this context means all customer routes learned from PEs within a regional POP, plus all the customer routes learned from satellite PEs within that region. Some of these PEs are truly iBGP (intra-POP), while others are eBGP (satellite POPs). To simulate the iBGP advertisement rules on inter-regional connections, the following two configuration steps are sufficient:

1. Match all incoming inter-regional VPN routes and apply a community (e.g. 65000:999)

2. Deny all outgoing inter-regional VPN routes with that community; permit all others

Filtering these routes outbound prevents over-reflection. All of the regions are fully meshed anyway, and reflecting routes between regions creates unnecessary BGP table bloating and general confusion among network operations. Again, this technique is a simple and scalable way to simulate regular iBGP non-RR client advertisement rules over eBGP. The diagram below illustrates how this filtering works in the global network.

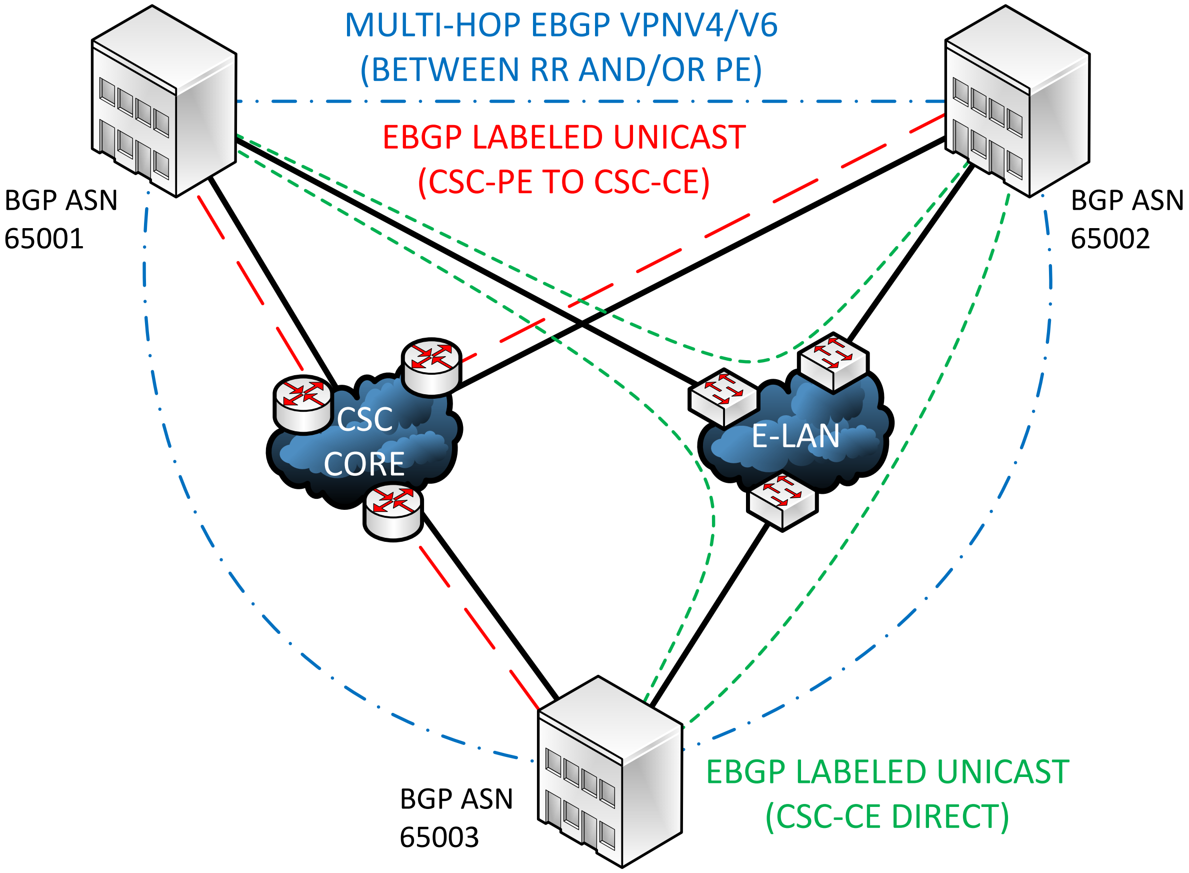
Figure 31 - Controlling Inter Region eBGP VPN Advertisements



2.2.4. Non-CSC Transport Supplementation

As discussed previously, one of the main drivers to deploy Option C with eBGP over iBGP between POPs for VPN service enablement was to more easily allow non-CSC transport links to be added. Taking this idea to its logical conclusion, the customer carrier may consider having one set of CSC-CEs connected to CSC and the other set of (now-misnamed) CSC-CEs connected over an Ethernet LAN services. Indeed, we wanted to do exactly that, but time and budgetary constraints slowed our progress. The diagram below illustrates the high-level design, which has some obvious benefits. A system-wide failure in either carrier would not cause a loss of network connectivity for dual-connected POPs.

Figure 32 - High-level Non-CSC Auxiliary Transport Design

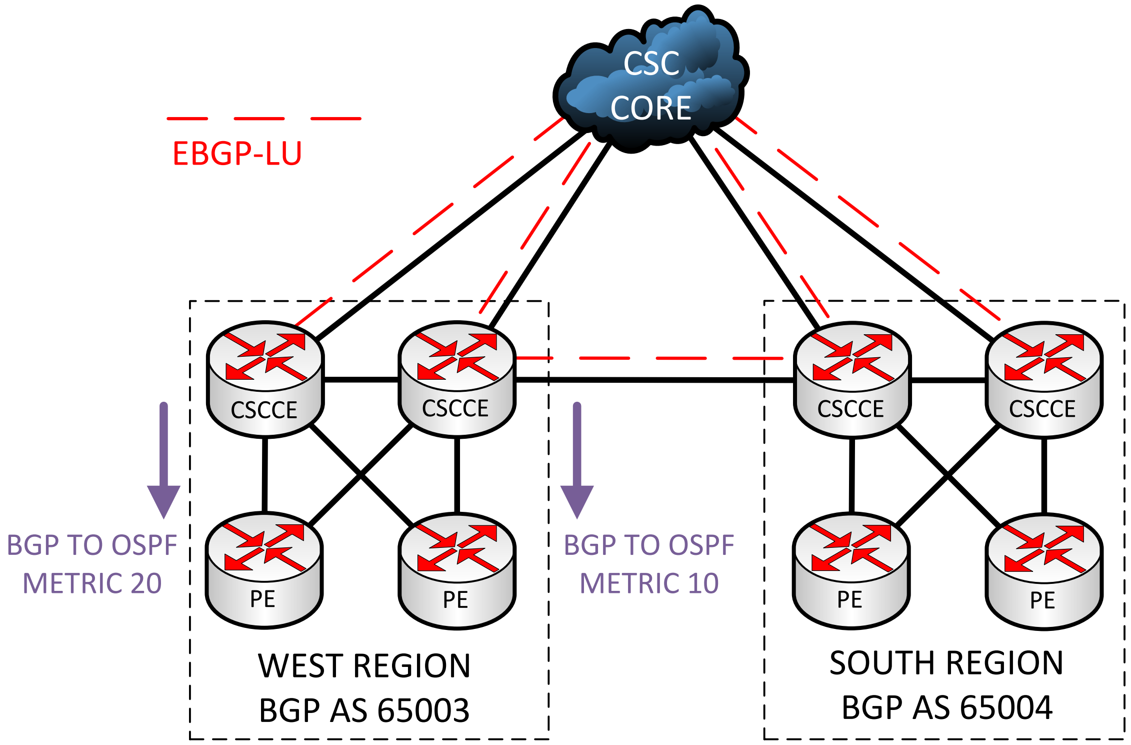


Consider the simpler case of a point-to-point link (say, VPWS or dark fiber) between two POPs. Using eBGP means that we don't have to get creative with local AS spoofing and can simply peer the POPs directly. BGP will prefer the direct link by default as the AS path to the peer POP will be shorter over the direct link than over CSC. As it relates to IGP redistribution, the same design concepts apply here as they did earlier when discussing the CSC-CEs. Mutual redistribution can occur on both ends of the link without any complexities because, assuming OSPF is used, only internal routes are candidate for redistribution into BGP. Therefore, these

POPs will never act as transit nodes for one another, making this a "peering" link and not a "transit" link, to user Internet terminology.

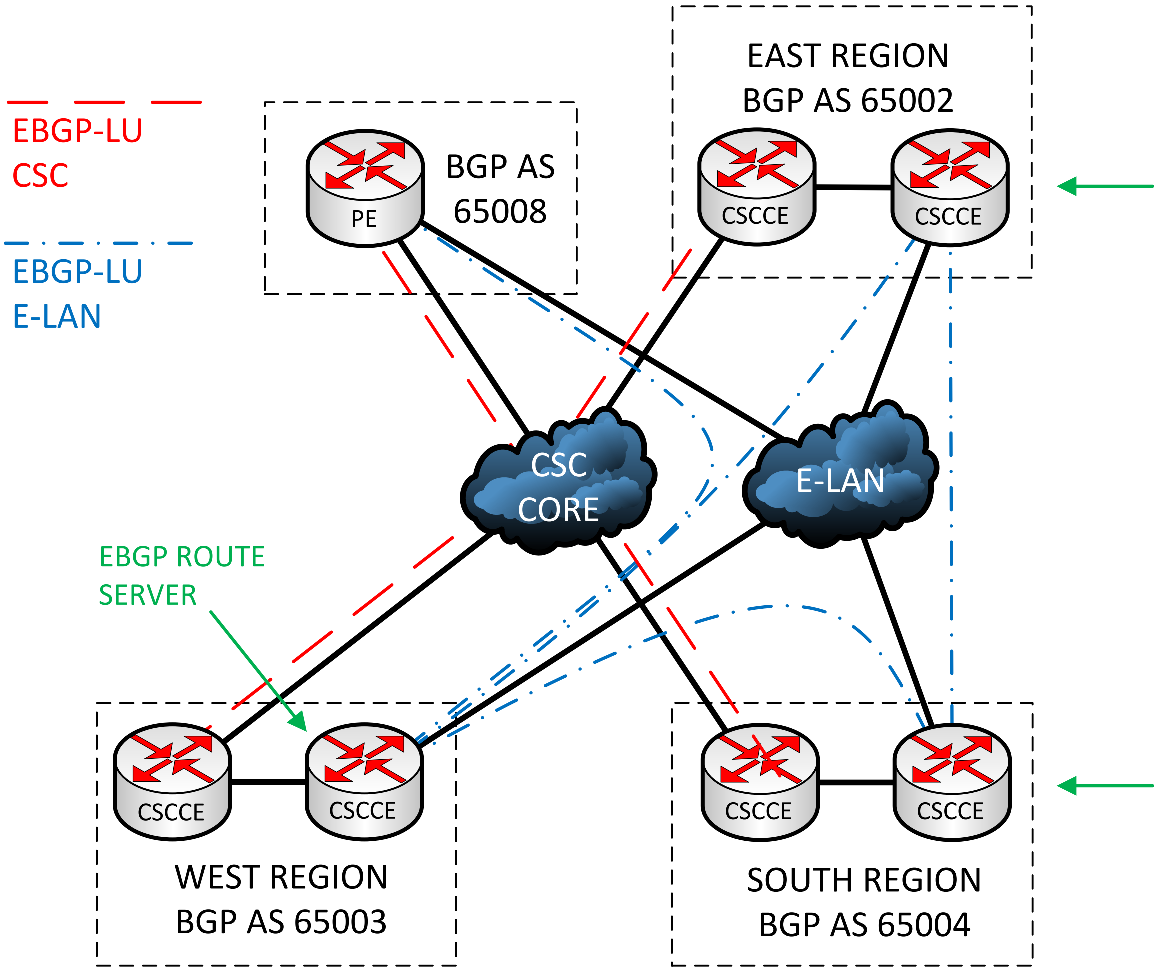
In order for the PEs to choose the direct link over CSC, use OSPF external type-2 routes with a lower seed metric on the ASBR terminating the direct link. This document won't detail all the different BGP and IGP configuration changes relating to forwarding policy, but engineering traffic to flow over either transport is not challenging nor is it the focus of this whitepaper. Just ensure that routes being redistributed into a POP's IGP have the proper metric based on the desired traffic patterns in your environment. The diagram below illustrates this design.

Figure 33 - Non-CSC Direct Links Between POPs



Now, consider the "logical conclusion" design where every POP is connected to both CSC and an E-LAN/VPLS service. The E-LAN connectivity is very similar to an Internet Exchange Point (IXP) design whereby many different BGP AS' share a LAN segment. While a full-mesh of eBGP-LU peers is possible, it scales poorly, both in terms of the control-plane and management-plane. Instead, consider using route-servers (discussed earlier) hosted in the regional POPs on the CSC-CEs. Satellite POPs within a region will connect to these route-servers as clients, and the regional POPs will peer to one another as route-server clients, too. This is necessary to avoid any BGP next-hop changes along the way. The diagram below illustrates this design.

Figure 34 - Non-CSC E-LAN Service Between POPs



Admittedly, this second design is quite rare and has likely never been deployed in production. The design is mostly conceptual and should be thoroughly vetted before actual deployment. In our environment, we only consumed point-to-point links for tactical, “quick fix” reasons which were quickly decommissioned when no longer needed. This obviated the need for any creativity with BGP route-servers combined with labeled-unicast. Be sure to use the same community-based filtering method described for eBGP VPN sessions with these eBGP-LU sessions between route-servers and their clients to prevent over-advertisement and potential loops.

Instead of using route-servers, some vendors implement eBGP in such a way that the next-hop for a given NLRI is in the same subnet as a remote peer. This makes sense for IXP connections or any other fully-meshed layer-2 network, as is the case here. Beware that the introduction of labeled-unicast may change this behavior, requiring some configuration workarounds (e.g. enabling multi-hop eBGP and next-hop-unchanged despite the session being single-hop) to make MPLS forwarding and label allocation work correctly. Cisco IOS-based devices appear to require this workaround; be sure to test your specific platforms extensively.

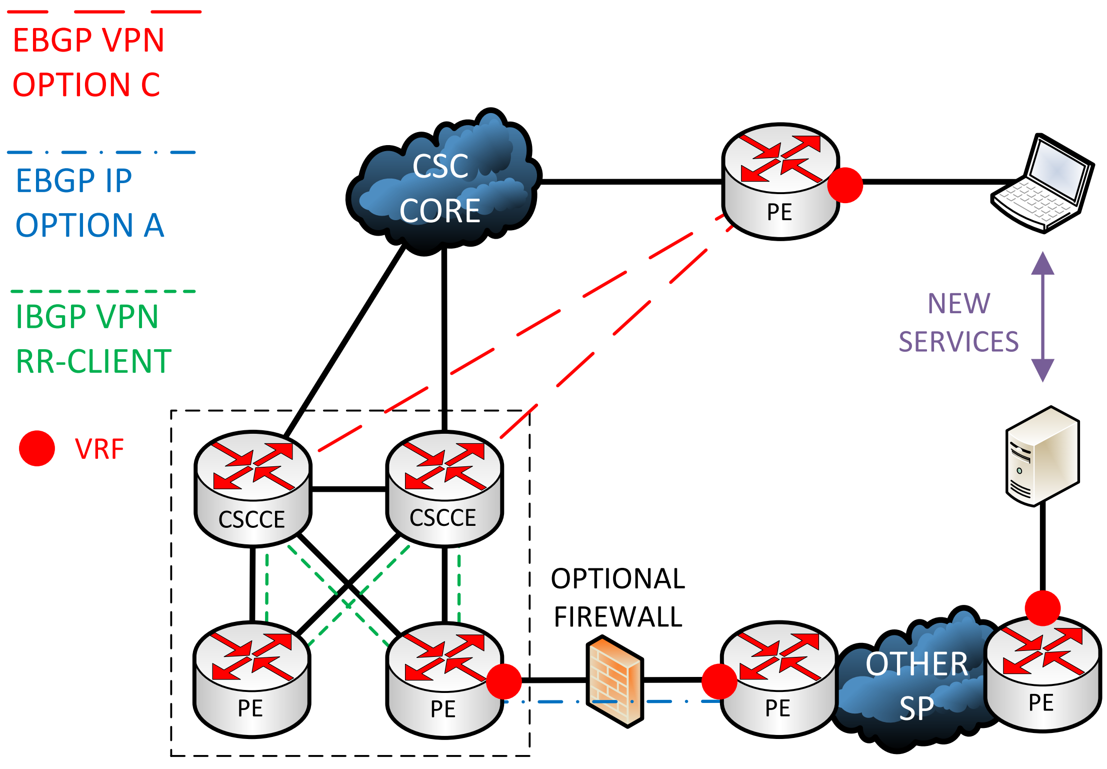
2.3. Extranet Integration

Our transport network was not the only one of its kind. Often times, we integrated with other carriers to gain access to new geographic regions and/or unique services that we could not offer ourselves. Such network extensions with external partners are known as extranets.

As is common in most production networks, we deployed Inter-AS MPLS Option A for all extranet interactions. This simple approach treats the ASBR like a standard PE, using basic layer-2/layer-3 multiplexing techniques across the inter-AS link to create many sessions. For example, to extend four VPNs between two service providers using a single Ethernet link, each carrier would create four VLANs, assigned each to a different VRF, and import/export the proper RTs to build the required VPN topologies for each. VRF-aware BGP for IPv4/v6 runs over each logical link to exchange the routing information. For non-IPv4/v6 VPNs, such as layer-2 VPNs, BGP is not necessary. Other protocols/techniques used in standard PE-CE connections relevant to those services will work just the same.

From an operational perspective, the solution is simple to understand, troubleshoot, and maintain. Challenges arise with scale, both due to computing limitations (e.g. limits on number of interfaces, VRFs, and BGP sessions) and configuration management. In our case, we never had more than 20 extranet VRFs on a single ASBR. This was manageable for humans as the RT assignments rarely changed once configured. All the usual design considerations regarding high availability and ASBR security still apply. The diagram below illustrates an extranet example.

Figure 35 - Extranet Integration with MPLS Inter-AS Option A



2.4. Quality of Service (QoS) Design

In service provider networks, there are many different aspects to QoS, and this document explains core, edge, and CSC components in depth.

2.4.1. Queuing and Shaping

This section details how traffic is treated across the MPLS transport network. In CSC, most of a packet's life (at least with respect to elapsed time and distance traveled) takes place in the core carrier's network. The customer carrier has little insight beyond the information supplied by the core carrier. In our case, we knew which MPLS experimental bits (EXP) were mapped to which per-hop behaviors (PHB) and could design our QoS solution accordingly.

In our case, the Differential Services Code Point (DSCP) values used by our customers were assigned to PHBs that did not directly align with our core carrier. This is true in most carrier environments. As an intermediary between these two networks, we needed to design some kind of value mapping/translation mechanism to achieve end-to-end QoS. Additionally, we had strict requirements to be transparent from a QoS perspective, meaning only pipe-mode models were acceptable. That is to say, we were not allowed to modify customer DSCP values for any reason.

In developing our queuing design, we examined the bandwidth allocations and queue types used by the core carrier. This carrier used an 8-queue design, one for each EXP value. Two of these EXP values, 7 and 3, were not applicable to our environment at all, so a 6-queue design was more appropriate for our network. For example, we had very little elastic video data (EXP3) and felt it would be simpler to treat all video as inelastic leveraging Admission Control (AC) without Active Queue Management (AQM) techniques like Weighted Random Early Detection (WRED).

To the maximum extent possible, we tried to match our internal bandwidth allocations with those used by the carrier, although the slightly divergent policies introduced some inconsistencies. The table below illustrates both policies, showing all 8 EXP values and the bandwidth allocated to each queue within our POPs compared to the CSC network. Note that the policy, at least within the customer carrier, should match IP precedence (IPP) and/or DSCP Class Selector (CS) values to capture any non-MPLS traffic, such as BGP sessions, OAM traffic, or GRE-encapsulated multicast VPN traffic.

Table 2 - Core Queuing Allocations

Match criteria	CSC carrier purpose	Bandwidth POP %	Bandwidth CSC %
EXP 7 IPP 7 / DSCP CS7	Asynchronous Transfer Mode (ATM) Pseudowires	0 (unused)	5
EXP 6 IPP 6 / DSCP CS6	Network control	2	5
EXP 5	Voice bearer	23 LLQ + AC	10 LLQ

IPP 5 / DSCP CS5	Voice/video signaling		
EXP 4 IPP 4 / DSCP CS4	Inelastic video bearer	15 + AC	15
EXP 3 IPP 3 / DSCP CS3	Elastic video bearer	0 (unused)	10 + WRED
EXP 2 IPP 2 / DSCP CS2	Elastic data (transactional and bulk)	25 + WRED	30 + WRED
EXP 1 IPP 1 / DSCP CS1	Internet and Scavenger (low priority) data	10 + WRED	10 + WRED
EXP 0 IPP 0 / DSCP DF	Default data	25 + WRED	15 + WRED

In our network, the CSC uplinks were typically 1 Gbps Ethernet links, but the core carrier could not guarantee this, as many circuits still used SONET/SDH. The carrier typically provisioned (and policed) circuits at 150 Mbps, roughly the same speed as an OC-3 or STM-1 (155 Mbps), and applied ingress policers on their CSC-PEs to enforce the contracted rate. As such, traffic conditioning via shaping on the CSC-CE was necessary to slow down traffic to this rate, despite the line rate of the interface being much faster.

In most cases, the customer carrier should use the minimum possible time committed (Tc) to improve the user experience for real-time and transactional applications, such as voice, teleconferencing, and multi-media services. Given a committed information rate (CIR) of 150 Mbps and a target Tc of 4 ms (the minimum value on our hardware), the burst committed (Bc) would be 600 kilobits (kb). This means that every 4 ms, the interface can physically send 600 kb sent at the physical rate of 1 Gbps, then wait until the next 4 ms interval to send more traffic.

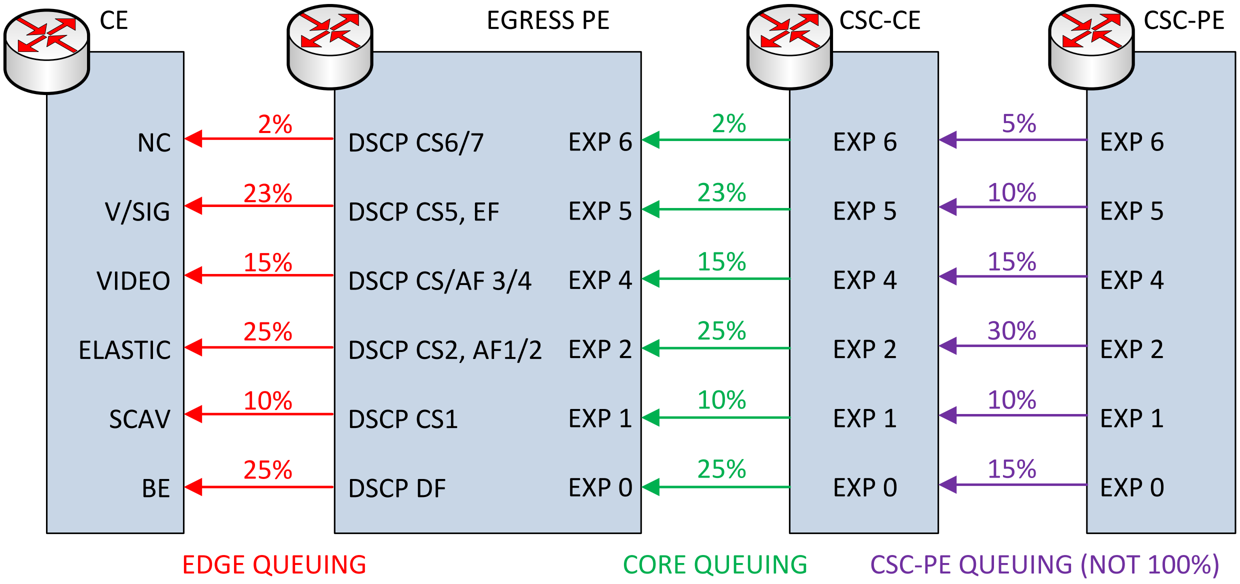
Because carriers sometimes do not provide low-level details about the configuration of their policers, we made moderately conservative assumptions regarding burst excess (Be). We used 600 kb for this value as well, matching Bc. If the shaper does not send any traffic for an entire Tc, the shaper is allowed to burst up to an additional 600 kb in the next Tc, for a total of 600 kb and a peak information rate (PIR) of 300 Mbps. In effect, this allows the customer carrier to reclaim up to one lost Tc due to inactivity. In our experience, we did not observe any negative effects from this assumption, although an even more conservative approach would be using a Be value of 0 kb. This effectively creates a peak shaper, setting the PIR to 150 Mbps.

Most customer-facing PE-CE links in our environment tended to be the slowest, least stable links, and thus deserving of the most precise QoS. At the same time, customers used a variety of DSCP values within their own networks and their ability to remark was limited. On the CE, they can implement their own queuing and shaping for outbound traffic. On the ingress PE receiving that traffic, the classification and marking process preserves those values and maps them to

MPLS EXP or DSCP tunnel and described earlier. The queuing and shaping on the egress PE is very similar to that used in the core, except exclusively matches DSCP values as the PE-CE links never run MPLS. To keep things consistent, we opted for a 6-queue policy with similar PHBs.

The diagram below illustrates the complete queuing design from the CSC-PE to the CE.

Figure 36 - Queuing and PHB Design



2.4.2. Classification, Marking, and Policing

This section details edge QoS actions relating to customer ingress. As mentioned earlier, our customer DSCP schemes seldom matched our design or the core carrier's design. Customers sometimes used different DSCP schemes which needed to be preserved across the network, although most were functionally similar. As such, we selected the short-pipe QoS design, allowing customer values to be retained end-to-end and allowing egress queuing actions to be based on those DSCP values. This is in contrast to the long-pipe and uniform strategies that evaluate MPLS EXP received by the egress PE for egress queuing decisions.

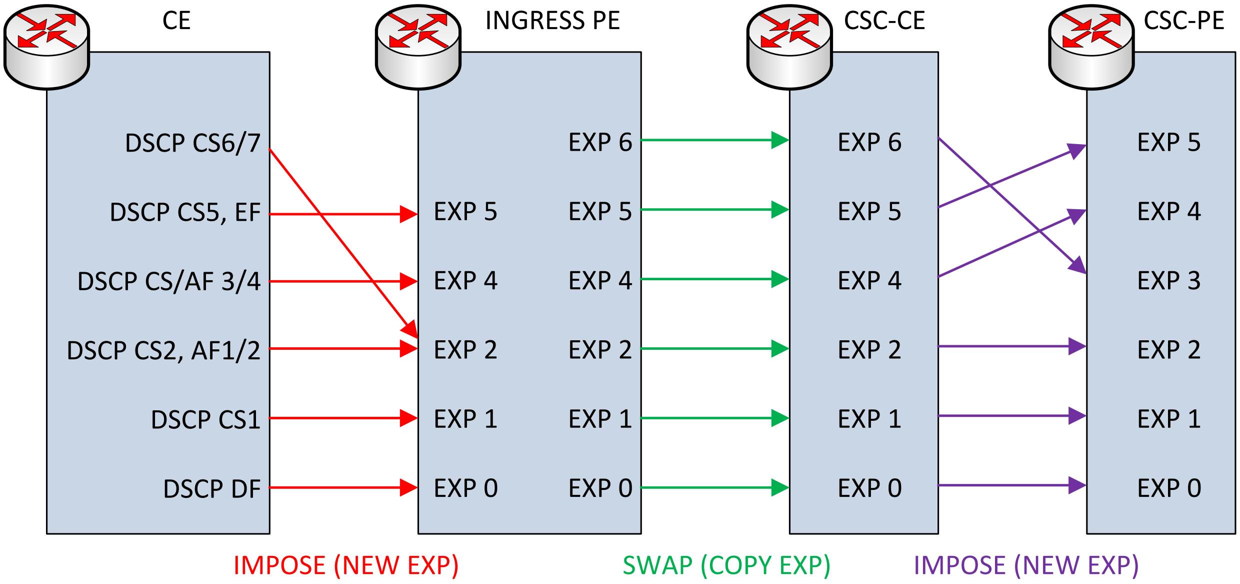
Focusing on the classification and marking first, the table below illustrates the DSCP values that customers must use if they want traffic to be treated in accordance with the correct PHBs. This policy is RFC4594-compliant with respect to the DSCP values themselves. For those operating in environments with extensive Cisco collaboration equipment (e.g. IP phones, call control systems, voicemail/contact management systems, etc.), it is common to use DSCP CS5 for broadcast video and DSCP CS3 for voice signaling. These values can be swapped without significant impact, but our customers were mostly RFC4594-compliant. Note that because MVPN traffic is GRE-encapsulated, imposing EXP values has no impact. Instead, the outermost GRE/IP header will receive the value specified in the table below. This still operates in pipe mode because the original customer DSCP values are preserved.

Table 3 - Ingress PE Classification and Marking

Inbound DSCP	Purpose	Imposed at PE	Imposed at CSC-PE
DSCP CS6, CS7	Network control	EXP 2 / DSCP CS2	EXP 3 / DSCP CS3
DSCP CS5, EF	Voice bearer traffic	EXP 5 / DSCP CS5	EXP 5 / DSCP CS5
DSCP CS3, CS4, AF3x, AF4x	All video traffic	EXP 4 / DSCP CS4	EXP 4 / DSCP CS4
DSCP CS2	OAM traffic	EXP 2 / DSCP CS2	EXP 2 / DSCP CS2
DSCP AF1x, AF2x	Transactional/bulk data	EXP 2 / DSCP CS2	EXP 2 / DSCP CS2
DSCP CS1	Scavenger data	EXP 1 / DSCP CS1	EXP 1 / DSCP CS1
Anything else	Unmarked/default data	EXP 0 / DSCP DF	EXP 0 / DSCP DF

The diagram below illustrates this process from CE all the way to CSC-PE. It is worth noting that both the customer and core carriers protect their own network control traffic by never allowing customer traffic to compete directly with it. For example, an end customer's DSCP CS6 traffic has MPLS EXP 2 imposed by the customer carrier. A customer carrier's MPLS EXP 6 traffic has MPLS EXP 3 imposed by the core carrier (according to them, at least). This treatment is imperfect as it mixes inelastic customer network control with elastic customer data. This trade-off allowed us to use a simpler queuing strategy by introducing only a small risk to customer network stability. The diagram below illustrates how DSCP and EXP are handled for upstream flows from CE to CSC-PE.

Figure 37 - DSCP to EXP Mapping on Ingress



Additionally, customer carriers should consider policing traffic on ingress from customers. Because packet loss must be minimized, these policies can simply impose different, lower priority MPLS EXP or DSCP tunnel values when traffic limits are exceeded. For example, a G.711 phone call, assuming it is encapsulated in Ethernet, consumes about 90 kbps per call. Each customer is allowed 100 calls, for a total of 9 Mbps of bandwidth. Customers can manage this in their telephony control-plane using various call admission control technologies, which is commonly deployed. Assuming a single-rate, three-color policer is used, it should mark conforming and exceeding traffic as EXP 5 per the table above. This provides low-latency treatment to conforming and permissible excess burst traffic. Violating traffic beyond the CIR for extended periods of time is marked as EXP 0. Put another way, customer voice is never dropped (unless the aggregate link CIR is overwhelmed using hierarchical policers), but will stop receiving low-latency treatment beyond 9 Mbps. This allows customers to make their own risk/reward decisions regarding admission control and voice oversubscription, although this is strongly discouraged.

At a minimum, policing voice traffic this way is particularly helpful to prevent saturation in the core, because most LLQ implementations will stop providing LLQ treatment to traffic in excess of the allocated percentage. Without the policer/remarker, one customer could flood DSCP EF (and subsequently, EXP 5) into the network beyond what the carriers have forecasted. This selfish act would harm all customers and likely have a negative business outcome for the carrier. This same policing strategy makes sense for other queues as well, but in our network, we limited the policer to customer voice traffic for simplicity. As a final technical point, using hierarchical policers can be useful as well. A generic CIR policer could encapsulate the entire ingress police, with subrate policers on a per-class basis as described above. We opted to skip this approach for operational simplicity.

2.5. Management, Security, and Automation

To provide improve maintainability and visibility, we maintained two management layers. These layers provided different views of the network, allowing us to more quickly isolate faults by having additional information readily available at the Network Operations Center (NOC).

2.5.1. Global Management View (GMV) Design

The Global Management View (GMV) reveals the reachability status of POPs from the perspective of the NOC. Hosted from two of the PEs for high availability, these hosts are in the global routing table, alongside the core IGP/LDP protocols. The GMV consisted of only one user VLAN with troubleshooting/monitoring clients present. This VLAN was advertised into OSPF, making it reachable throughout the network via redistribution into BGP at the CSC-CE.

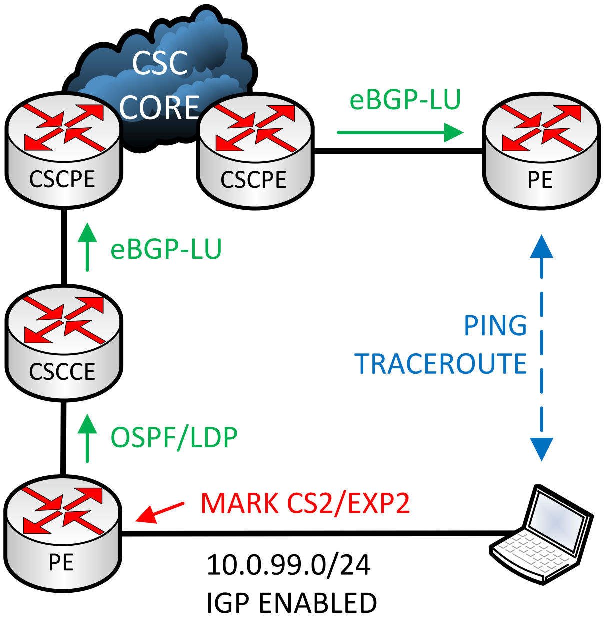
For security reasons, none of the routers in the network are actually “managed” via the global routing table. This guarantees that other hosts within the CSC VPN cannot access the customer carrier’s network. While it is possible to somewhat securely manage devices in the global table, we decided to limit such device management to a VPN, which is discussed later. However, in the global context, two activities were permitted:

1. ICMP echo-request and echo-reply between the GMV NOC VLAN and each device.
2. MPLS LSP verification (LSPV) messaging anywhere in the network.

The GMV used a simple "ping" to each device's loopback to measure up/down status, and displayed it using commercial network visualization software. CSC-PE uplinks were also measured in this way, because the GMV clients were at the same layer in the network as the VRF-enabled CSC-PE interfaces. Whenever outages were observed in the GMV, it was immediately obvious that a data-plane fault occurred, such as a link-down event or broken LSP.

The ingress PEs marked all GMV traffic with DSCP CS2 and MPLS EXP2 to signify OAM traffic. This provided the proper treatment across the network consistent with the QoS design discussed earlier. The diagram below illustrates how GMV operations were designed. Note that all MPLS routers will allocate MPLS labels for these non-loopback routers so that traffic is MPLS encapsulated for most of the journey, ensuring that correct QoS treatment is applied.

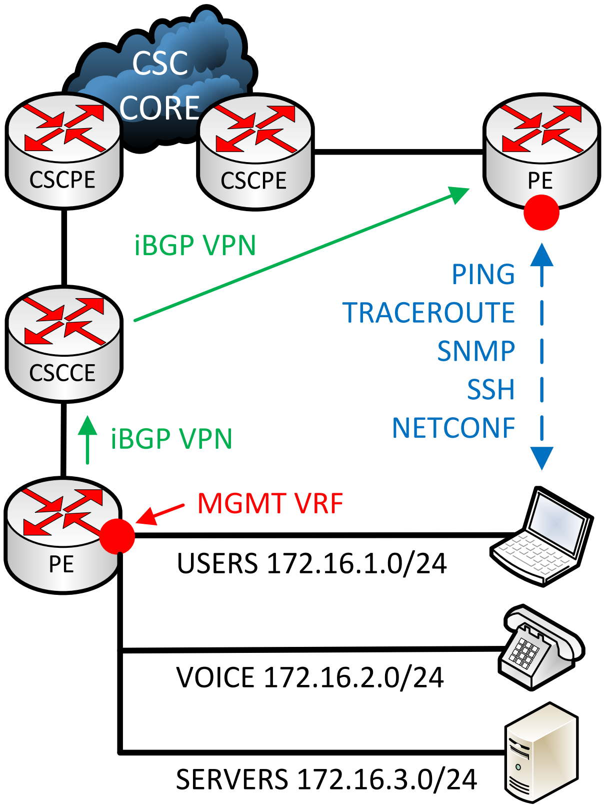
Figure 38 - Global Management View (GMV) Design



2.5.2. VPN Management View (VMV) Design

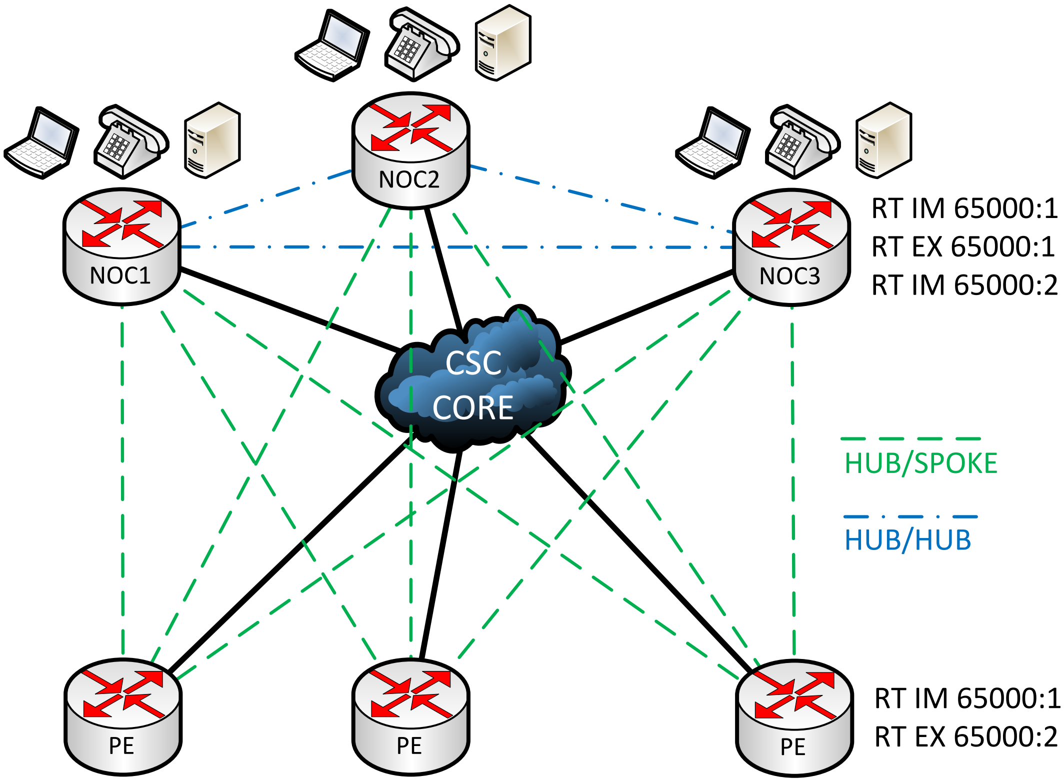
In contrast to the GMV, the VPN Management View (VMV) was similar in concept but operated inside of an MPLS L3VPN. NOC machines were placed in a VLAN and connected to two PEs for the purpose of network management. However, these PE interfaces were placed inside of a VRF designed exclusively for VPN management only. Remote POPs used an additional, VRF-enabled loopback for management in addition to the transport/BGP loopback discussed earlier. The NOCs each had three VLANs in the VMV: users, servers, and voice. NOC administrator laptops were placed in the user VLAN (similar to GMV) while management servers were placed in the server VLAN. The voice VLAN hosted IP phones and call control systems for inter-NOC communication. Also, routers within a POP connected their out-of-band management physical interfaces to the VMV user VLAN for local management. This allowed intra-POP routers to avoid needing management VRF loopbacks (which would require VRFs on spines and CSC-CEs, for example). The diagram below illustrates how these components fit together.

Figure 39 - VPN Management View (VMV) Design



Using a pair of route-targets, this VPN followed a tree design. All three of the NOCs would import and export the “root” RT. This allowed all the NOCs to form a full mesh between one another. Additionally, the NOCs would import the “leaf” RT, allowing them to access all of the remote POPs. The non-NOC sites would export the “leaf” RT and import the “root” RT. This ultimately creates hub/spoke style of network with all of the hubs being fully-meshed. The diagram below illustrates the high-level connectivity between sites using dummy RTs.

Figure 40 - VMV Connectivity with Hub/Spoke Route Targets



A full VMV mesh everywhere was avoided because it adds unnecessary security risk with no operational benefit. Remote sites did not have firewalls or other security appliances (discussed later), and many were located on customer premises. If one was compromised, we did not want attackers to perform leap-frog attacks by traversing laterally through the VMV to attack other remote sites. Because all of the NOCs had extensive security defenses, any attacks due to compromise would have to fight through those upstream defenses first. These defenses are discussed later in this document.

The VMV has the additional advantage of being completely inaccessible via the global routing table and the CSC carrier. Even if the CSC carrier accidentally adds the wrong site into your VPN, the worst possible outcome would be a compromise of the GMV, which has no management access to any device. As mentioned earlier, the GMV exists for visibility and troubleshooting correlation only. The best security plans often mix control-plane, data-plane, and management-plane techniques into a unified defense.

Because both the GMV and VMV were tracked concurrently using network visualization software, operators could form hypotheses about network issues before touching a keyboard. This helped reduce mean time to repair (MTTR) for common outages. The matrix below illustrates how the global and VPN management views intersect. For example, if the GMV reports that a remote POP is reachable, but the VMV does not, the cause is likely related to the BGP VPN control-plane. Perhaps the BGP sessions are down completely or a route-target has not been properly imported/exported. The data-plane and label switched path are known-good, otherwise the GMV would also report an outage.

Table 4 - Global and VPN Management Outage Matrix

	VMV up	**VMV down**
GMV up	Fully operational	BGP VPN control plane issue
GMV down	Impossible case; misconfig	Transport control or data plane issue

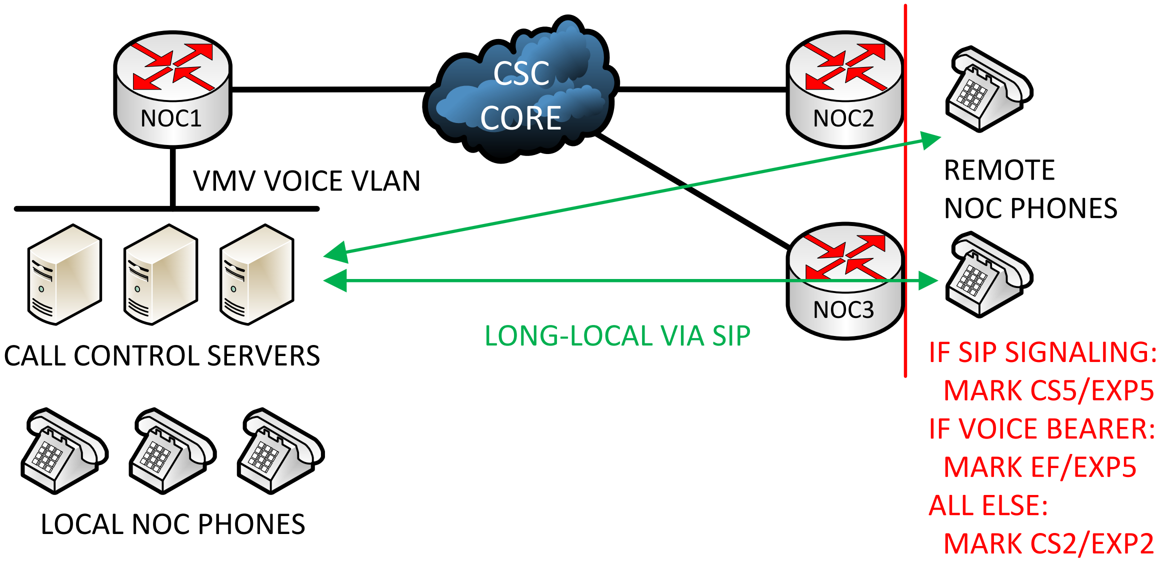
The VMV contained a variety of network management services:

1. Simple Network Management Protocol (SNMP) version 3 with auth/priv security
   a. Polling conducted to determine reachability in VMV
   b. Standard SNMP traps used for event notification
   c. Additionally, syslog messages were encapsulated in traps for secure transport
2. Authentication/ Authorization, and Accounting (AAA) services
   a. RADIUS: 802.1X authentication and backup login authentication/authorization
   b. TACACS: Primary login authentication/authorization, command authorization
3. Network Time Protocol (NTP) servers operating in client/server unicast mode. All devices reached back to these centralized NTP servers to flatten strata architecture
4. NetFlow: Captured IPv4/v6 PE-CE traffic and MPLS CSC-CE to CSC-PE traffic

With respect to the voice over IP (VoIP) design, one NOC hosted the call management services. All other NOCs used a long-locals to register their phones over CSC back to the main NOC. While this centralized design did reduce availability to some extent, outages were rare, and the operational complexity/cost of managing a distributed call control system was undesirable.

Furthermore, voice traffic and signaling used DSCP EF and CS5 respectively per the QoS design discuss earlier. For MPLS transport, EXP5 is used for all labels imposed at the ingress PE. This includes Session Initiation Protocol (SIP) for phone registration and call placement. All other network services discussed above were remarked to DSCP CS2/EXP2 at the ingress PEs. The diagram below illustrates the high-level voice and management connectivity and QoS design.

Figure 41 - Voice over IP (VoIP) and Voice QoS Design



2.5.3. Management LAN Security and High Availability

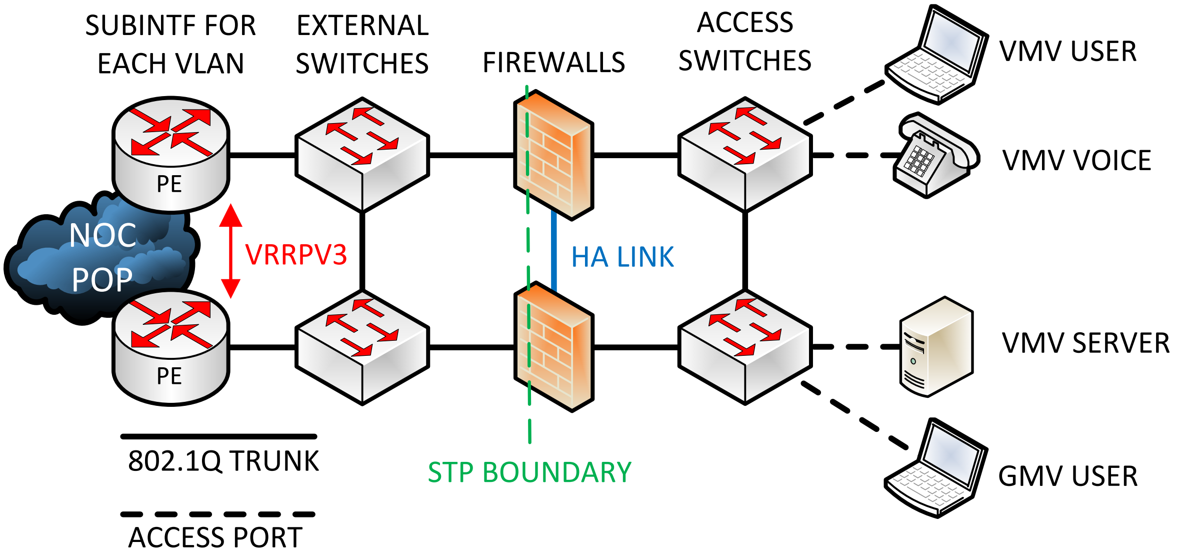
In order to protect the GMV and VMV from external attacks, we deployed a variety of layer-2 through layer-4 security techniques. As discussed earlier, the high-level routing separation between GMV and VMV is a necessary defense technique, but it is not sufficient.

All GMV and VMV VLANs connected to two PEs using Virtual Router Redundancy Protocol (VRRPv3) to provide IPv4 gateway failover for management hosts. VRRPv3 supports IPv6, and while our management networks were not running IPv6 at the time, we wanted to future-proof the design and implementation. The VRRP endpoints exist on the PEs, defined for each of the 4 NOC VLANs (1 GMV and 3 VMV). One PE was the VRRP master for all VLANs for simplicity as there was no operational benefit for load-sharing. The other PE was the VRRP standby device.

In between the clients and the PEs, there were 6 additional devices. First, a pair of internal access switches, with half of the devices connected to each one. These switches connected to a pair of stateful firewalls, deployed in a state-sharing active/standby configuration, operating in transparent (bridged) mode. On the outside of the firewalls were two more external switches, which connected up to the PEs.

The layer-2 control and data plane design in this security stack was kept as simple as possible. The firewall blocked Bridge Protocol Data Units (BPDUs) from transiting across it, creating two separate spanning-tree protocol (STP) domains. Each domain contained two switches with a single link between them, so loops were impossible. STP could technically be disabled on all four switches, but for extra safety against rogue device attachment or firewall BPDU misconfiguration, we retained STP in the network. The firewall bridged all GMV and VMV VLANs across the outside and inside networks, which implies all inter-VLAN VMV traffic was inspected by the firewall. All of these new devices have their out-of-band management interfaces connected to the access switches in the VMV user VLAN alongside every other network device in the POP. The diagram below illustrates the LAN security stack that was present at each NOC.

Figure 42 - NOC Security Stack Design



It would not be a good use of time to detail the firewall policies in depth, but to summarize:

1. These management protocols were allowed from inside to outside:
   a. GMV users: Ping and traceroute
   b. VMV users: Ping, traceroute, SSH, HTTPS, and NETCONF
   c. VMV servers: Ping, traceroute, SSH, HTTPS, SNMP poll requests, NTP, TACACS, RADIUS, and HTTPS
   d. VMV voice: Voice bearer and SIP signaling

2. These management protocols were allowed from outside to inside (exceptions):
   a. GMV users: Nothing
   b. VMV users: Nothing
   c. VMV servers: Ping, traceroute, SSH, HTTPS, NetFlow, NTP, TACACS, RADIUS, SNMP poll responses, and SNMP traps
   d. VMV voice: Voice bearer and SIP signaling

Other common network protocols like DHCP, DNS, and TFTP were permitted in various places as well, but that isn't explained in this document as this level of detail is unnecessary.

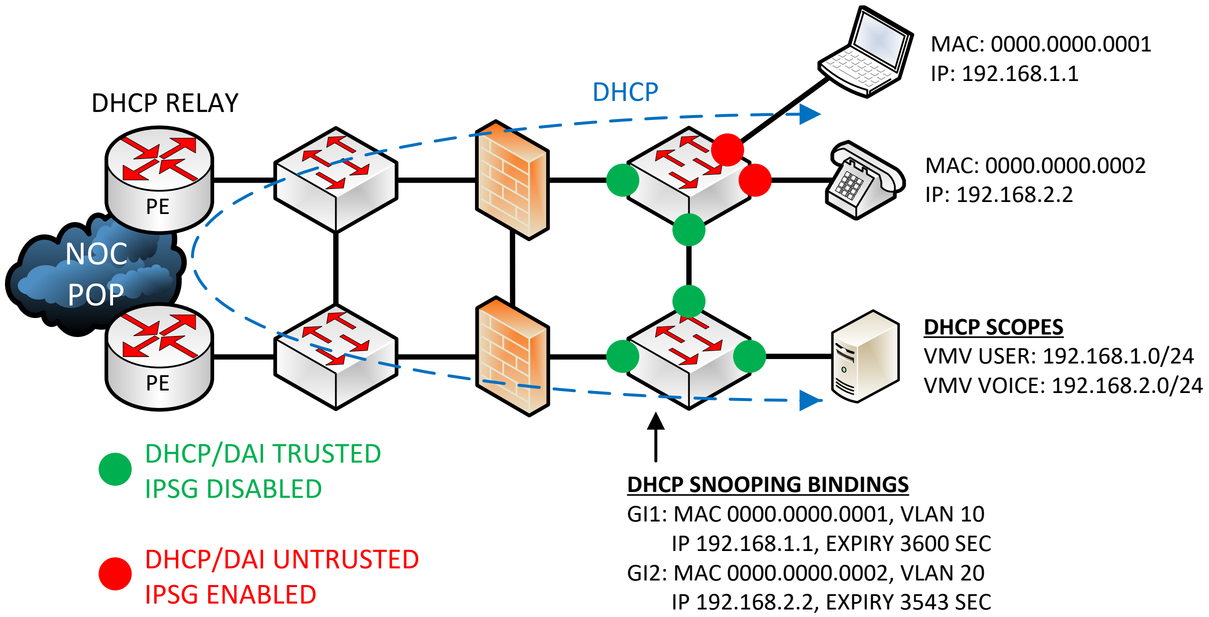
While firewalls are effective, they cannot block all threats. On the internal access switches, we enabled several more security techniques on the VMV VLANs. Three of them all work together: Dynamic Host Configuration Protocol (DHCP) snooping, Dynamic ARP Inspection (DAI) and IP Source Guard (IPSG).

DHCP snooping separates switch ports into two categories: trusted or untrusted. Trusted ports can send any kind of DHCP message while untrusted ports can only send client-related messages, such as Discover, Request, and Release. DHCP snooping was enabled for all VLANs and only ports directly connected to DHCP servers (which were virtual machines hosted on a shared hardware server) were trusted. This prevents rogue DHCP servers from handing out bogus IP addresses, which could happen if one of the LAN devices was compromised.

Additionally, it records the DHCP bindings for each client, including the issued IP address, client MAC, physical interface, and VLAN. Because the DHCP servers are in a different VLAN than the client it serves, the PEs act as DHCP relays to facilitate the DHCP messaging.

Given these DHCP snooping bindings, DAI can validate ARP messages between hosts on the subnet. If a client tries to spoof an ARP message, effectively pretending to be another host, DAI will block and log the offense. IPSG further reinforces security by using the DHCP snooping bindings to ensure clients cannot spoof IP packets. The source IP and source MAC must match the bindings, and if they do not, the packets are discarded and optionally logged. It is commonly believed that DHCP is a security liability (and that static IP addressing is “more secure”). This is simply false; DHCP is a security asset when combined with first-hop security techniques like DHCP snooping, DAI, and IPSG. The diagram below illustrates how these technologies work together to provide LAN security.

Figure 43 - Layer-2 Defense in Depth Security Design



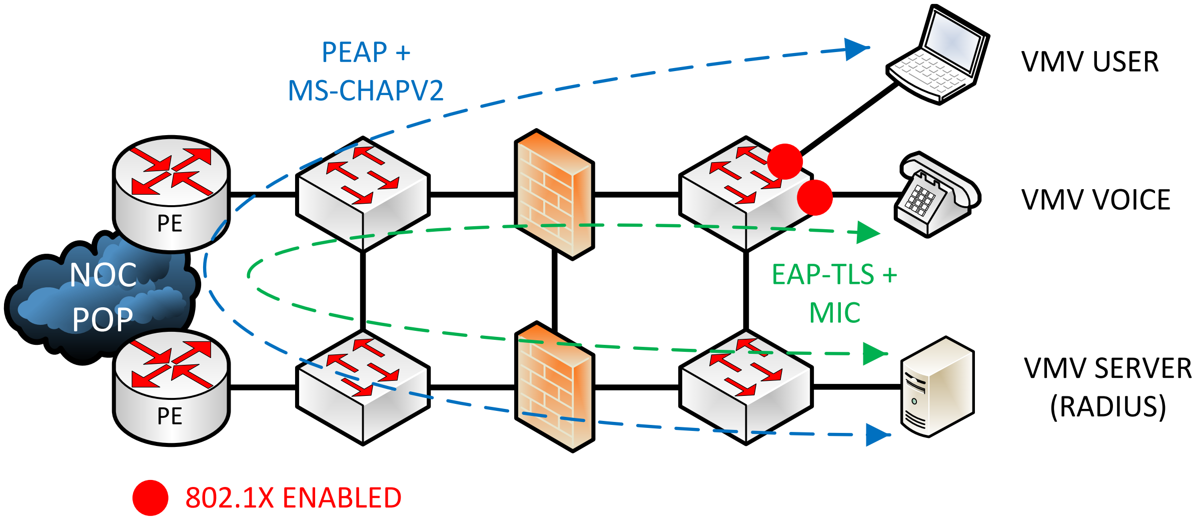
While powerful, these technologies do nothing to validate the authenticity of each client. Attackers who gain physical access to our facility or NOC personnel who connect unauthorized devices should not be able to join the network at all, even if they don’t intend on causing mischief. To solve this problem, we deployed 802.1X for network access control. Because we did not have a Public Key Infrastructure (PKI) nor any degree of PKI operational experience within our team, we opted to use Protected Extensible Authentication Protocol (PEAP) with Microsoft Challenge Handshake Authentication Protocol (MS-CHAPv2) for authentication. The PEAP outer method relies on a one-way certificate trust (client trusts server) to establish a secure TLS connection with the RADIUS authentication server. Once established, the supplicant provides its credentials (in our case, a per-machine username/password) to the authentication server using the MS-CHAPv2 inner method. In our view, this approach provided an operationally sustainable and moderately security posture. With a proper PKI deployed, EAP-TLS would have been a superior option as each client would have its own client certificate for authentication, providing a stronger access control solution in general. Some operators use MAC Authentication Bypass (MAB) instead of 802.1X, which is significantly less secure, but often better than nothing. MAB was only used in our network to support devices lacking an 802.1X supplicant, which was a rare occurrence.

Our VoIP phones were exclusively Cisco and thus supported a variety of 802.1X EAP methods. We opted to use EAP-TLS using the Manufacturer Installed Certificate (MIC). This is hard-coded into each device and is signed by a Cisco Certificate Authority (CA). Assuming the authentication server trusts the Cisco CA, the phones can authenticate using EAP-TLS. Note that the EAP-TLS + MIC technique only guarantees that the phone is a Cisco IP phone. An attacker could plug in a compromised Cisco IP phone, which would pass 802.1X authentication, and launch an attack. We saw this as an unlikely attack vector as physical security was relatively tight in our NOCs. Combined with all the other security defenses described earlier, EAP-TLS + MIC was a good design choice for IP phones as it balanced security and operational simplicity.

Note that 802.1X was not enabled for any servers, physical or virtual, and was limited to VMV users and VMV IP phones. Because GMV users are already quite limited in what they could accomplish on the network, 802.1X was not implemented for them.

The diagram below illustrates the high-level 802.1X design within a NOC. Note that the devices were not able to send any traffic into the network, other than traffic relating to 802.1X, until authentication was complete.

Figure 44 - 802.1X for NOC Users and IP Phones



As was true for all critical services, such as DHCP, RADIUS, TACACS, and RADIUS, each NOC had at least one of each server. This allowed clients to operate correctly at any NOC even if the local servers at that NOC were offline.

2.5.4. TACACS Command Authorization

Our environment was just beginning its automation journey as the network was being designed and built. As modern networks rely on command-line interfaces (CLI) less and less, it was still the primary means of network management for us. To improve security, visibility, and our overall change management process, we deployed TACACS for command authorization. In short, every command issued on every device had to be authorized for each user. We classified users into one of three categories, called tiers.

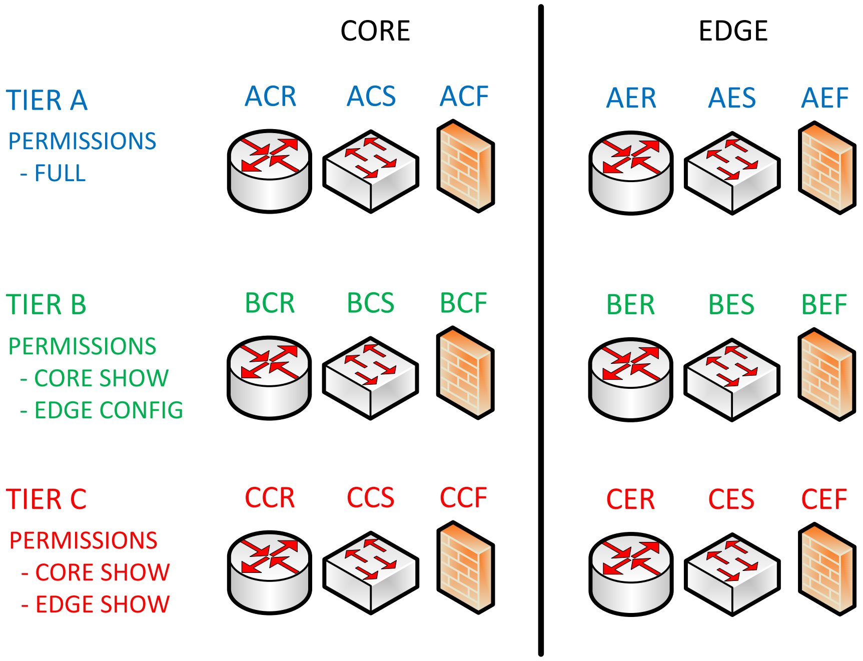
Tier C was designed for NOC operators. These individuals were primarily responsible for answering phones, resolving customer trouble tickets, monitoring visual displays, and running basic, low-risk CLI commands. Some customers demanded Tier C access into the remote POPs deployed within their facilities, which we often granted, as the risk of compromise or a network outage was low.

Tier B was designed for NOC engineers. This group had all the permissions of Tier C, with some limited configuration permissions. These configuration permissions were limited to edge VPN provisioning only, such as managing RD/RT definitions, IP addressing, VRF assignments, PE-CE routing, etc. Tier B was not granted any write access to core configuration items, such as BGP VPN peering, CSC uplinks, BGP-LU, IGP/LDP, or device security features.

Tier A was designed exclusively for senior engineers, most of whom did not work in the NOC on a regular basis. This tier granted full permissions over the entire network. Regardless of the tier in which a user was placed, all commands were still authorized and logged.

In order to more easily identify the precise commands for each tier, devices were classified by type and by location. We defined three types: router, switch, and firewall. We further categorized devices by location using two options: core and edge. Thus, there were 18 total permutations when accounting for user tiers, device types, and device locations. These permutations corresponded to 18 different TACACS command sets, providing granular control over who could make what changes and on which devices. The diagram below illustrates the high-level TACACS design and the 18 command sets that exist. Even though some commands sets were nearly identical (for example, all of the Tier A command sets basically said “permit any”), creating the individual command sets allowed us to more finely tune the command authorization rules later.

Figure 45 - Tiered TACACS Design and Command Sets



2.5.5. Automation Strategy and Use Cases

Managing a global, high-scale network of any kind typically requires automation to some degree. This section summarizes some of the automation use-cases for which Ansible playbooks were developed and used in real life. Many of these use cases are specific to MPLS carrier environments, but with small adaptations, could be used in a variety of business environments. All of these tools have been open-sourced and this document includes links to GitHub for each.

2.5.5.1. Data Collection for Archival and Troubleshooting

For compliance reporting purposes, maintaining configuration backups was an important regulatory requirement. While there are many existing commercial tools that can perform this task, the author developed a custom Ansible playbook instead. Collecting configurations is necessary but not sufficient, as there are many other useful "show" command outputs that could be captured.

From an archival perspective, this might include software version, hardware inventory, active product licenses, current users, and various packet counters relating to access lists, interfaces, and QoS. From an operational perspective, consider rapidly collecting all of the FIB and FIB tables from all PEs in the network to troubleshoot a globally-disruptive routing loop (true story). This can be helpful to quickly resolve customer-impacting problems in production.

Note that this was a Tier C task and users at that tier were allowed to run this playbook to collection information from the production network (very low risk). The tool is open-sourced and the source code is available here: https://github.com/nickrusso42518/racc

2.5.5.2. MPLS Route-target (RT) Management

In any service provider network, customers come and go. Our network had particularly high turnover as customer engagements were very short lived, typically between one week and a few months, as opposed to multi-year commercial contracts. Many customers had complex inter-VPN connectivity requirements with partners, resulting in a web of extranets. Managing route-targets manually was difficult, error prone, and time consuming.

Using the concepts of infrastructure as code, the author developed an Ansible playbook to ensure the proper RTs were present or absent for each VRF on every PE. Each PE is represented by an individual state file, containing a list of VRFs. Each VRF has a list of import and export RTs. Those lists combine to form the intended VRF configuration. Any RTs not present in those lists are removed, and if a PE is missing any of those RTs, the proper RTs are added. The playbook summarizes the changes made on each PE, providing "drift detection" via persistent logs.

Note that this was a Tier B task and users at that tier were allowed to run this playbook to manage RTs in production. The tool is open-sourced and the source code is available here: https://github.com/nickrusso42518/vpnm

2.5.5.3. Inter-POP Performance Measurement

As the network grew, so too did the challenges regarding network performance monitoring. Our network spanned from the Mariana Islands to Hawaii (the long way) and round-trip times varied from 5 ms to 300 ms between any pair of POPs. Monitoring the performance between POPs, including latency, jitter, voice mean opinion score (MOS), MTU/fragmentation, and other metrics become increasingly important as customers demanded better service. The tool also uses MPLS label switch path verification (LSPV) probes to test the end-to-end health of the MPLS network, making it useful as a troubleshooting aid.

Using Ansible, the author solved this problem via Cisco's IP service level agreement (SLA) feature to synthesize VOIP traffic in a full-mesh between all sites. The tool could be used on-demand by generating the traffic at the time of execution and measuring the results, providing immediate, point-in-time feedback. The tool could also configure long-term traffic flows to track the metrics for trends.

Note that this was a Tier C task and users at that tier were allowed to run this playbook to measure performance data between POPs. The tool is open-sourced and the source code is available here: https://github.com/nickrusso42518/perf

2.5.5.4. Extranet IP Address Overlap/Translation Management

In a previous section, this document explained how Inter-AS MPLS Option A is typically used to connect to extranet partners that operate in different regions or provide access to new services. During one particular integration, the peer carrier did not allow RFC1918 IP addressing into their network. Rather than force all of our customers to stop using such addressing or to configure NAT in their own networks, we deployed NAT at the ASBRs. For security traceability, only 1:1 NAT was allowed, creating an enormous management burden as our customers onboarded and offboarded. Even without this regulatory requirement, overlapping IP addresses between carrier services is a common problem solved using 1:1 NAT.

The author developed an Ansible playbook, much like the previous route-target example, to manage 1:1 NAT statements on the ASBRs. Each ASBR had a corresponding state file with a list of NAT entries specifying their inside and outside IP addresses. Individual entries could be switched on (present) or off (absent), and the playbook would ensure the desired state was configured on each device.

Note that this was a Tier B task and users at that tier were allowed to run this playbook to manage static NAT statements in production. The tool is open-sourced and the source code is available here: https://github.com/nickrusso42518/natm

2.5.5.5. Customer Onboarding Assistance

Unlike in the commercial world, where "not my problem" is the standard carrier response to integration issues, the NOC was contractually obligated to assist with CE equipment setup and troubleshooting. This requirement comes in addition to managing the network and the customer services it provided. Because our operators rarely had access to customer equipment, automating the service onboarding and offboarding process was challenging.

The most legally and politically tolerable approach was to focus on two key areas:

1. Producing setup and teardown configuration snippets for customers. The NOC would provide complete, copy/paste capable text files for each device that customer was expected to use when consuming the carrier's services.
2. Producing an automatically-generated technical how-to guide that explained what to do, when to do it, and what the result should be. These guides were customized for each customer, using the customer-specific IP addressing, hostnames, VLANs, and more.

Both the configuration snippets and technical guide are generated automatically using an Ansible playbook which integrates with LaTeX. The result is a ZIP file containing all required snippets (plain text files), plus the technical guide (a PDF file), which can be delivered to the customer via email or through a web portal. While this technique introduces "user error" risk and uncertainty, it greatly reduced the quantity of human errors our NOC observed in real life.

Because this tool does not interact with any network devices, anyone could use it, and there was no risk to the network (TACACS was wholly unaware of its existence). The tool is open-sourced and the source code is available here: https://github.com/nickrusso42518/mkfd

2.6. Example Customer Use Cases

This section details a subset of real-life use cases made available by this network design.

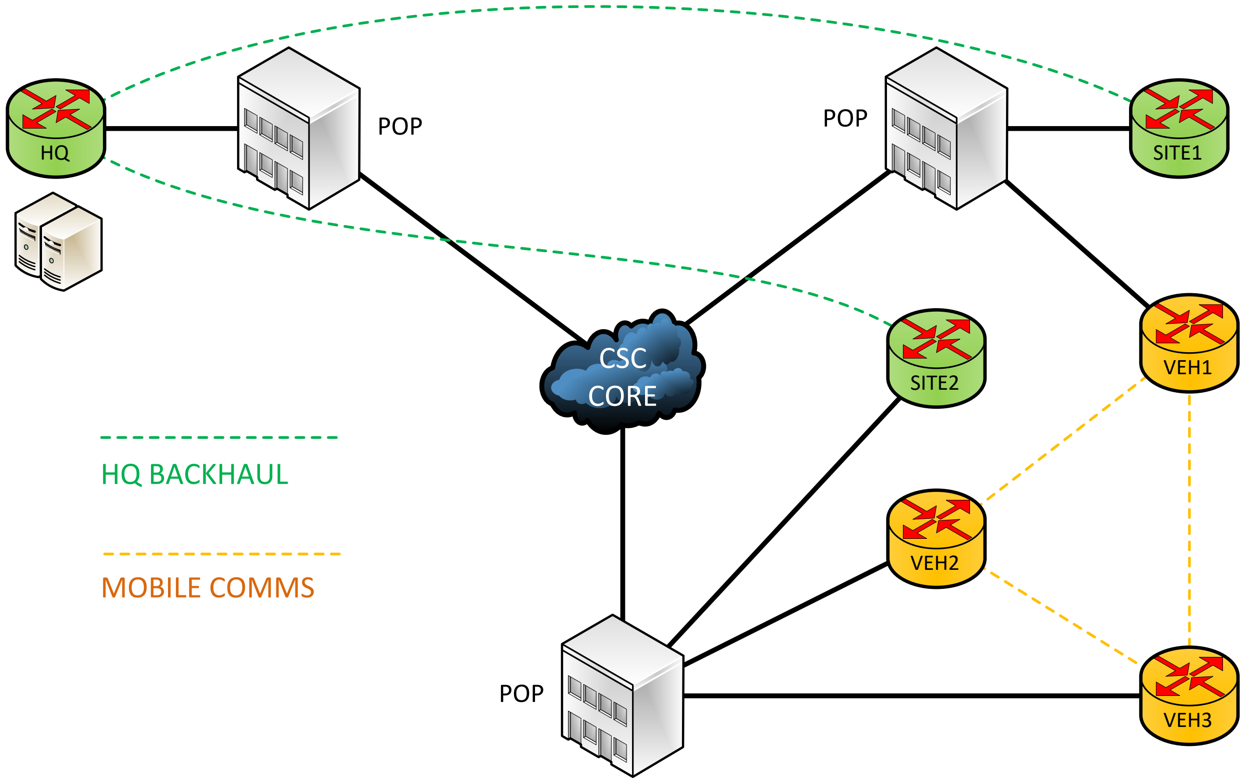
2.6.1. Geographic Extension with Multi-tenancy

A primary use case of MPLS VPNs in general is to create multi-tenant connections across geographic distances. By a wide margin, this was the most popular service we offered to our customers. This service is delivered by simply managing L3VPNs as one would normally do. Our most common applications include:

a. A headquarters location hosting centralized services connecting to remote sites.
b. Disparate mobile elements (such as vehicles or field expedient tents) that are part of the same organization needing to communicate laterally across the world.

The diagram below illustrates some of these examples, which due to the multi-tenant design of MPLS VPNs, can all be supported concurrently. Because this use case has already been extensively explained, this document will not detail it further. Note that mobile customers used IPsec VPNs or other secure transport technologies to connect to PEs whereby the IPsec tunnel is the logical PE-CE link. Other than MTU calculations and some platform-specific QoS limitations, this minor deviation is not significant to the overall service offering.

Figure 46 - Use Case: Connecting Geographically Dispersed Nodes

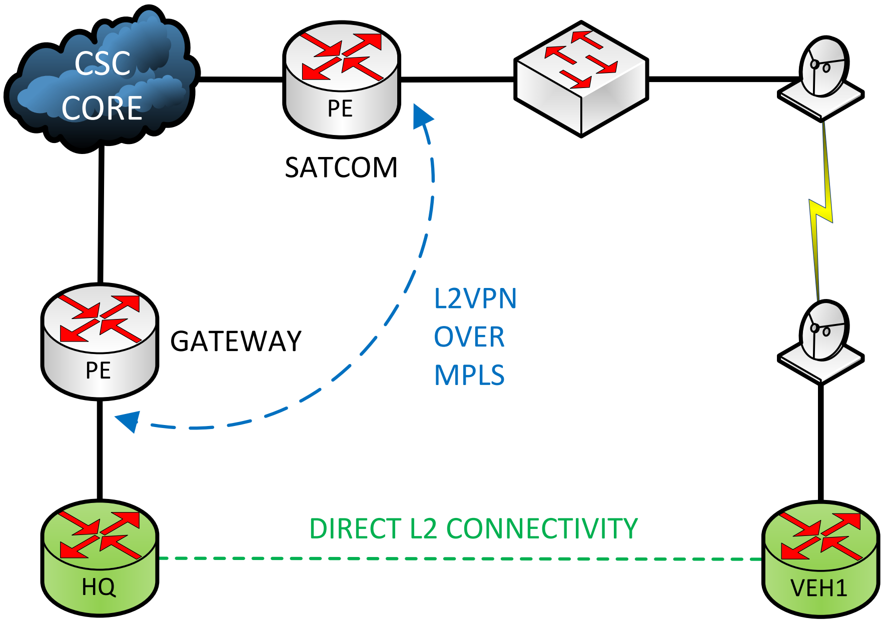


At the time of this writing, the scourge of Coronavirus 2019 (COVID-19) was present in global life. This latter use case regarding mobile elements could be potentially utilized by pop-up hospitals, office spaces, legal courts, classrooms, or any other disaggregated business attempting to socially distance. This also applies to general-purpose disaster relief, emergency communications, and various types of mobile units.

2.6.2. Satellite Communications (SATCOM) Remoting

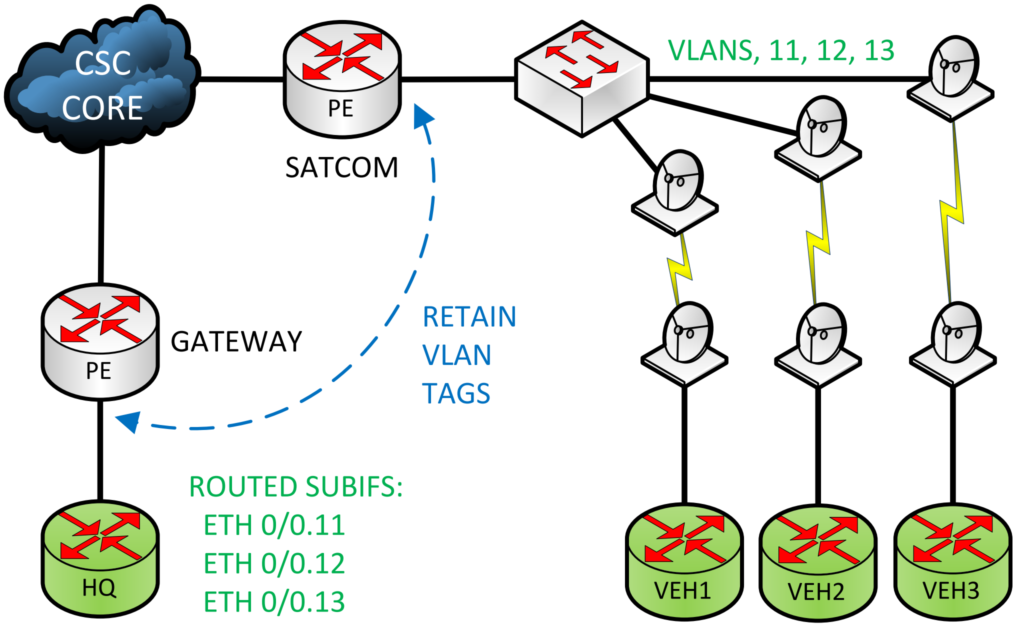
Any kind of VPN service is available across the network, including layer-2 VPNs and multicast VPNs as discussed previously. A common use case for layer-2 VPNs, beyond the standard convenience of simplified customer routing, was to decouple SATCOM transport from the target services. For example, a mobile user could connect to one regional POP using SATCOM and be transparently backhauled to another region in a layer-2 VPN to access services there. This transparency allows the user to leverage the SATCOM equipment at one site while connecting to a networked device (another CE) at a different site, improving their global mobility. It’s somewhat comparable to an inter-site wireless mobility design using concepts of “anchor” and “foreign” to identify the two sites in question. While the same outcome could be achieved using standard layer-3 VPNs, regulatory restrictions and political barriers often mandated a separation of routing domains, making L2VPNs attractive. Declaring one site as the “SATCOM” site and the other as the “gateway” site assuaged the critics. The diagram below illustrates the high-level concept for this service.

Figure 47 - High-level SATCOM Remoting Design



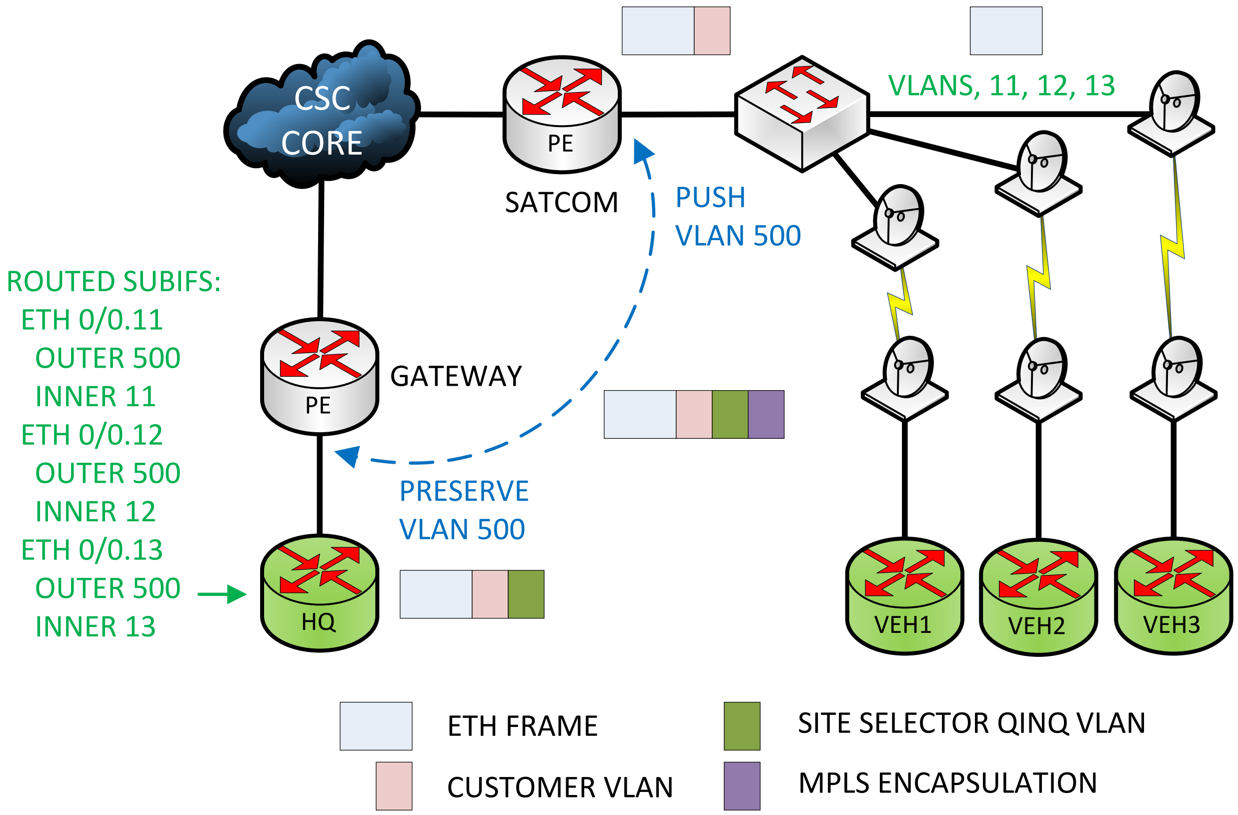
When there are many mobile users connecting to the same SATCOM POP to access services in the gateway POP, creating individual point-to-point layer-2 VPNs for each customer is burdensome and scales poorly, especially without automation. Instead, a single EV-LINE QinQ-based L2VPN can transport multiple connections. For example, suppose there are 3 different mobile users that use connect to a given SATCOM site using a point-to-point wireless technology. Those SATCOM modems are each placed in different access VLANs numbered 11, 12, and 13. Thus, the higher-level IP addressing on each link is a different subnet. If all three SATCOM connections terminate on the same gateway PE, they can all “ride together”, somewhat analogous to carpooling. A single EV-LINE circuit could transport all of them, and from the logical perspective of the headquarters node, it would be a hub/spoke network with three different Ethernet point-to-point links. The VLAN tags would be retained end-to-end allowing the CE router to peel off each VLAN using routed subinterfaces. The diagram below illustrates this scaling technique.

Figure 48 - Using a Single L2VPN to Connect Multiple Sites



In our experience, different POPs did not always have the same VLANs available for transport, and individually remarking VLANs on a per L2VPN basis does not scale. Sometimes, the VLANs overlapped between sites, implying that VLAN transparency cannot work. Furthermore, it was a common occurrence that all 3 SATCOM clients were terminating on the same PE. We used QinQ to add additional encapsulation to represent the “site selector”. The three SATCOM VLANs would be trunked to the ingress PE as discussed before with their VLAN encapsulation retained. The ingress PE adds a new “site selector” VLAN tag to the Ethernet frame before imposing MPLS encapsulation. The egress PE would preserve (i.e., not remove) this extra QinQ tag, ensuring that the QinQ VLAN was provisioned across its last-mile switching fabric between PE and CE. The router terminating all of the SATCOM links could then match both the outer tag (QinQ site selector VLAN) and the inner tag (SATCOM customer VLAN) using routed subinterfaces. In our experience, we found this to be a suitable solution for scaling L2VPNs while also not requiring sites to synchronize their VLAN numbering, allocation, and consumption schemes. It extends the “ride together” carpooling logic to the CE rather than terminating it at the PE. The diagram below illustrates this QinQ “site selector” technique. Although not depicted for brevity, there were often many layer-2 switches between the PE and customer headquarters node, making this technique both scalable and transparent.

Figure 49 - Using QinQ Tunneling to Avoid VLAN Rewrites



2.6.3. Highly-Available Internet Access

Most customers connected to our network using private transport networks, such as SATCOM, dark fiber, or E-LINE services from local carriers. These transport types do not have direct Internet access; many customers required such access, often for humanitarian missions and/or disaster relief. Providing Internet service over MPLS networks is a complex topic with many high-level solutions:

a. Carry the Internet IPv4/v6 tables in global BGP, and form another peering in addition to the VRF-aware peering to provide Internet service
b. Carry the Internet IPv4/v6 tables in global BGP, and leak these routes (or a subset/aggregate of them) into the VRF routing table on the PE facing the customer
c. Carry the Internet IPv4/v6 tables in a VPN with a central services RT that customers can import as necessary

The first option provides the best scale as it guarantees the BGP tables will only be stored once, as the routes never end up in any VRFs. This can cause security problems as it exposes the MPLS customer carrier to the Internet, potentially compromising the entire network. While infrastructure ACLs can deny all traffic destined to infrastructure networks from both Internet peering points and from Internet-subscribed customers, they are only effective if kept up to date and regularly reviewed.

The second option has slightly less scale and still exposes the customer carrier directly to the Internet, but is far more convenient for customers. They'll use their BGP uplink both for VPN routes and for Internet routes as the route leaking between tables happens on the PE. If there are many customers on a single PE that require full Internet tables, the routes will have to be copied in memory with new RDs, potentially taxing the router's memory.

The third option is the most secure, easiest to conceptualize/operate, but the least scalable. Because the Internet connectivity always exists in a VPN, there is no need for infrastructure Access Control Lists (ACLs) to control inbound/outbound core access and very little chance of an Internet attack reaching the customer carrier's network. Scale is poor considering each customer that imports the Internet RT will copy routes from the RD-indexed table into the VRF-specific table on a given PE. Some platforms may implement memory optimization techniques here, but this generally isn't a safe assumption.

This Internet-in-a-VPN approach was our best choice, and we decided to improve the scalability by only allowing our customers to follow a default route towards the Internet. No customers received any longer matches for any Internet prefix. The ASBRs touching the Internet maintained full routing tables, but generated local 0.0.0.0/0 and ::/0 aggregates for advertisement into BGP VPNv4/v6. This allowed us, in emergency situations, to advertise more specific Internet routes into BGP if a customer required it; this was a rare occurrence.

To provide Internet connectivity to a customer, we defined a pair of Internet RTs for import and export that would grant access in a hub/spoke fashion. In a given customer VRF, the operators would import the Internet RT to receive the default route and export the Internet RT so that the Internet VRF would import the customer's routes. This hub/spoke RT exchange has been detailed earlier in the document in other contexts, but the same logic applies here.

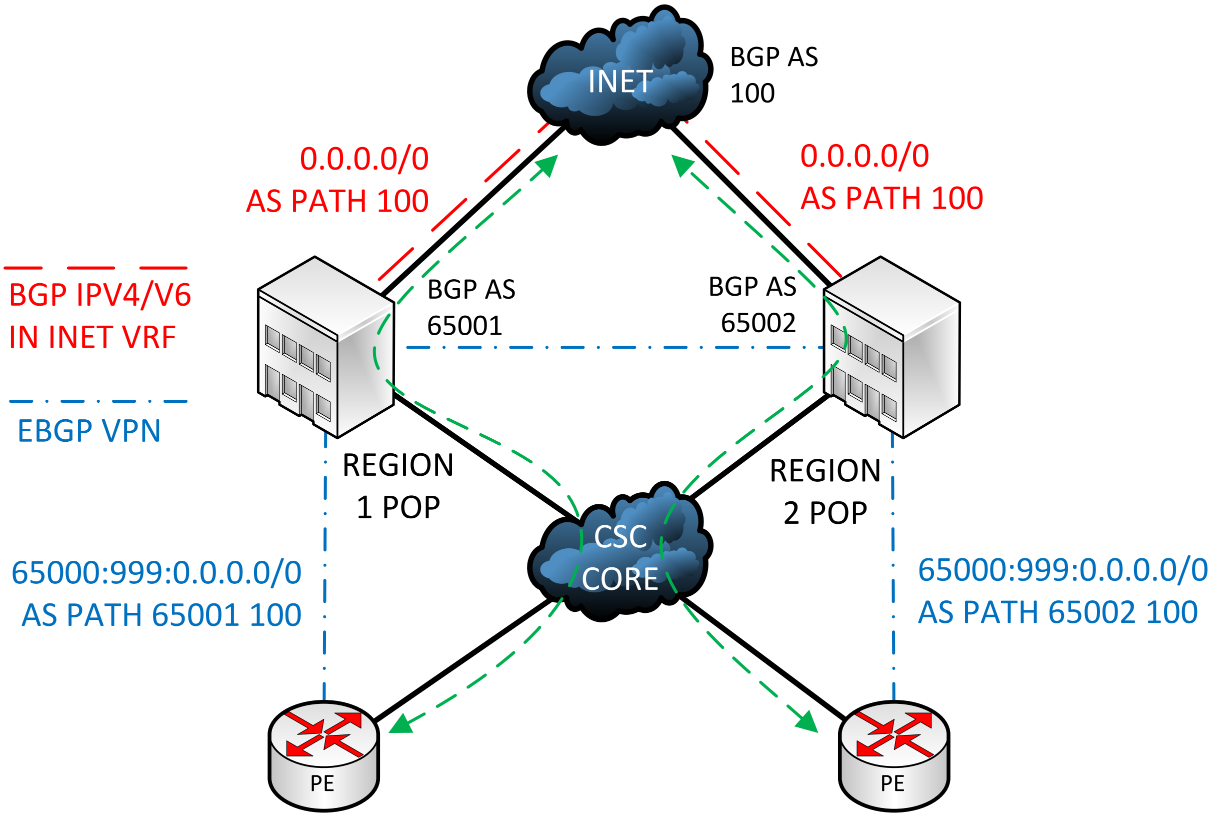
To provide high availability, the carrier should have multiple Internet uplinks with multiple default routes originated. The simplest and most automatic way to handle this is to rely on the BGP AS-path length to select the shortest egress path. If every region has one Internet connection (which was generally true for us), then the following is true:

1. Customers attached to the regional POP hosting the Internet connection will see an AS-path length of 2:
    a. ISP ASN
    b. Locally-connected regional POP ASN
2. Customers attached to satellite POPs within that region will see an AS-path length of 3:
    a. ISP ASN
    b. Parent regional POP ASN
    c. Satellite POP ASN
3. Customers attached to regional POPs that currently have a broken Internet connection will see an AS-path length of 3:
    a. ISP ASN
    b. Internet-connected remote regional POP ASN
    c. Locally-connected regional POP ASN
4. Customers attached to satellite POPs within a region that currently has a broken Internet connection will see an AS-path length of 4
    a. ISP ASN

b. Internet-connected remote regional POP ASN
c. Parent regional POP ASN
d. Satellite POP ASN

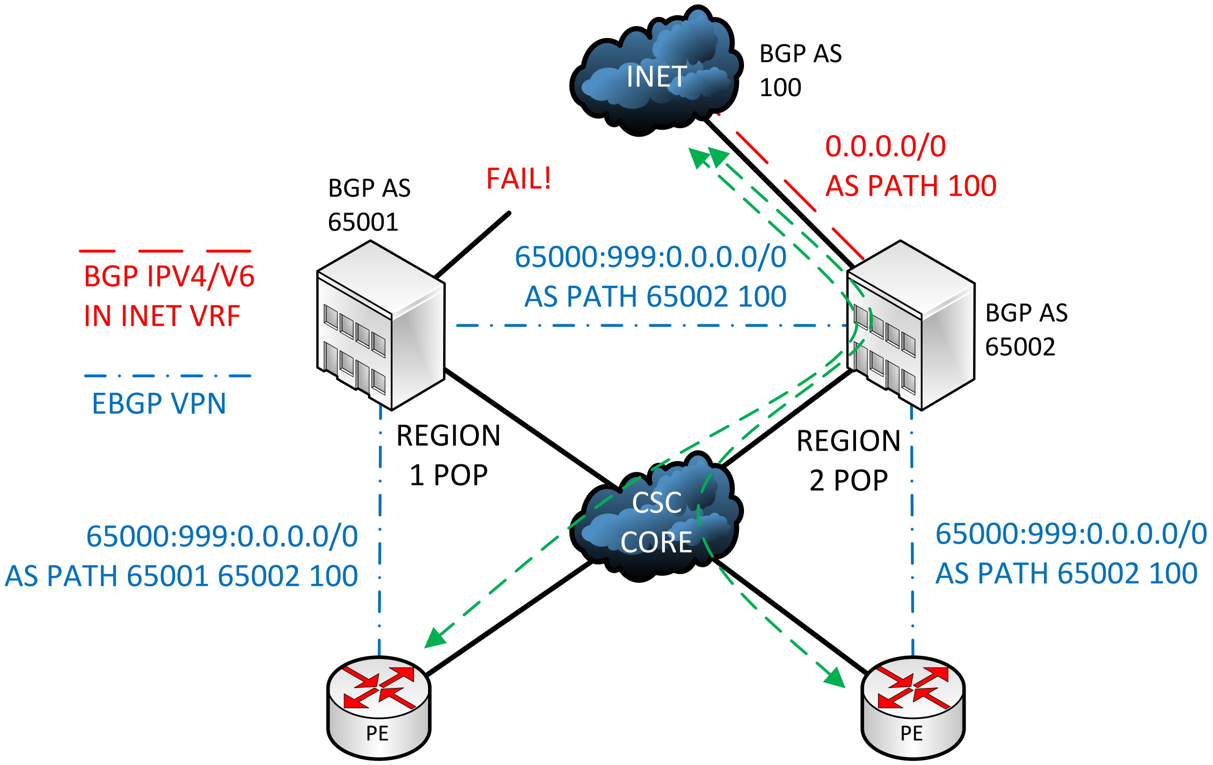
This document will not detail the succession of failovers between regions and the prioritization between them. Such configuration techniques are well-known and can be implemented using a variety of BGP path attributes and/or communities. The diagram below illustrates this design using two regions for simplicity.

Figure 50 - Internet-in-VRF High-level Design



Suppose the Region 1 Internet connection fails. Customers within that region, both within the regional POP and connected satellite POPs, would be able to consume the Internet connection via Region 2. Note that the Option C over CSC designs allows these satellite POPs to route directly to Region 2 (not transiting Region 1) due to the eBGP next-hop preservation feature. The diagram below illustrates this failover.

Figure 51 - Internet-in-VRF Regional Failover



For carriers offering "wires only" service, the expectation is customers have already advertised Internet-routable prefixes to the carrier via BGP. Internet-destined traffic must be sourced from one of these Internet-routable prefixes, whether provider aggregate (PA) or provider independent (PI), in order for return traffic to function correctly. In managed service providers, a perimeter security stack that inspects all Internet traffic and performs NAT for customers is sometimes deployed. Both of these topics are out of scope for this document. Note that NAT is important as stateful firewalls will expect traffic flows to be symmetric. That is to say, if traffic egresses through Region 1, return traffic must ingress through Region 1. This becomes more challenging with IPv6 Internet traffic unless Network Prefix Translation for IPv6 (NPTv6) is used.

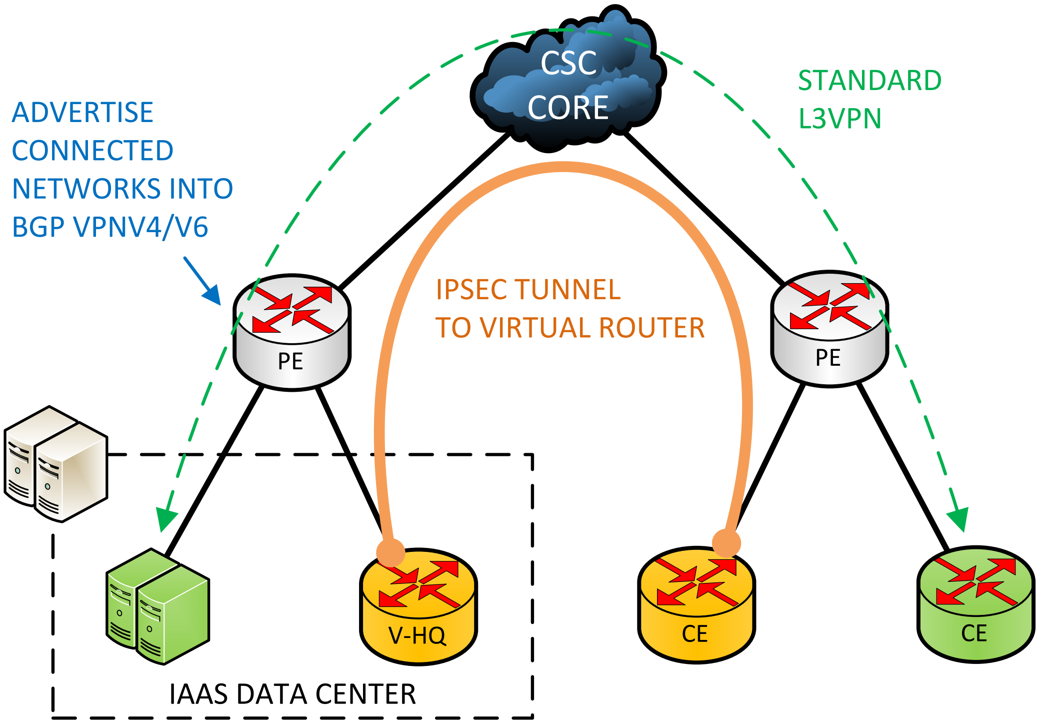
2.6.4. WAN Aggregation and Cloud Data Center

Some customers are less interested in connecting remote sites or accessing central, shared services like the public Internet. Instead, they want to access a hosted compute environment, effectively infrastructure-as-a-service (IaaS), to host their own services. The design and rollout of this new service was being conceptualized as I was leaving the organization, so this section provides only a conceptual overview of the solution.

Regional POPs contain general-purpose data centers whereby individual customers can be given access to a share of the resources. To provide an even more enterprise-like experience, customers can deploy their own virtual routers in this data center to serve as a WAN aggregation point instead of terminating IPsec tunnels on PEs. This gives the customer ultimate control over their security posture while using the carrier for purely transport reasons, which is true in commercial WAN designs.

To support multiple customers, the PEs servicing the data center would typically use a different VLAN+VRF combination per tenant with individual BGP sessions connecting the virtual CE router and the PE. The PE-CE links in the diagram illustrate these logical connections. Although the vast majority of customers prefer to only advertise their IPsec tunnel source (perhaps a loopback or just the connected network), other customers may forego IPsec and just advertise the backend server networks directly. Both designs are supported and given the similarities with standard L3VPN use cases, there aren't many special design considerations with respect to routing. The diagram below illustrates this conceptual design.

Figure 52 - Managed IaaS High-level Design



3. 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 more traditional MPLS deployments, such as those not using CSC or Option C. It is also relevant when analyzing the different options within the aforementioned design regarding routing, QoS, and management design decisions.

3.1. 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.

This solution has relatively low overall state as the different layers of hierarchy contain different sets of information:

1. Option C is highly scalable because it does not require ASBRs (CSC-CEs in our case) to retain all VPN routes. Because our CSC-CEs happened to be RRs, and because the number of CSC-CEs was typically equal to the number of RRs within a POP, our particular environment did not benefit from the scaling advantage with respect to state.
2. CSC is highly scalable because it decouples transport networks, such as PE and RR loopbacks, from customer networks. This is beneficial for both the customer carrier P routers and the entire core carrier's network, including the CSC-PEs.
3. In general, using IGP+LDP is a scalable approach when compared to RSVP-TE as a strategic tool for building and maintaining LSPs. Considering the number of LSPs within each POP was small, and all inter-POP LSPs were governed by BGP-LU, we did not benefit much from this scaling advantage, although it did exist.

Various other decisions contributed positively towards reducing state in the network:

1. Not peering the A/B mesh RRs within each POP.
2. Aggregating the Internet table to default routes on the Internet-facing PEs.
3. Requiring customers to perform their own NAT for Internet-destined traffic.
4. Not allowing regions to "double advertise" inter-regional routes between one another using community-based prefix filtering.
5. Use hub/spoke VPNs versus any-to-any VPNs when appropriate (e.g., VMV).
6. Building a network capable of nearly unlimited growth over CSC, compared to a flat E-LAN/VPLS design where IGP neighbor limits would likely restrict expansion.

3.2. 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.

With respect to traffic forwarding, all LSPs are optimal in that there is no hair-pinning. For example, intermediate POPs are never in the transit path, thanks to the any-to-any CSC transport combined with Option C. Furthermore, IP multicast is efficiently transported without ingress replication (as would be present for any non-EVPN style E-LAN service) across CSC between POPs within the customer carrier.

One case where totally optimal IP forwarding may be jeopardized is the Internet access use case. If a client connected to the western regional POP needs to access a server in the east, traffic still must egress through the western regional POP. Assuming stateful firewalls and NAT do not exist at the Internet edge, return traffic may ingress through the eastern regional POP. Only one of these paths can objectively be "optimal", implying that the other is suboptimal, but probably not by much, assuming both links are operational.

The bigger issue happens when a western customer routes through the eastern regional POP to reach a western server. This would only happen when the western regional POP Internet connection is offline, but since this is a failover case and not the "steady state" of the network, it isn't a major drawback with respect to optimization analysis.

3.3. 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.

The transport and VPN architectures are highly decoupled. The former is built on IGP+LDP and BGP-LU to establish connectivity between MPLS routers and the latter is based on a different BGP VPN topology. This improves scale (discussed earlier) and allows the two topologies to evolve at difference paces and in different ways.

The transport architecture itself is comprised of two tightly integrated components: IGP+LDP within the POP and eBGP-LU from CSC-CE to CSC-PE for inter-POP connectivity. Route redistribution occurs on the CSC-CE to connect these two different label switching environments. This surface interaction is wide as it occurs on every regional POP and is also deep because large quantities of data (routes and corresponding labels, etc) are redistributed/readvertised.

Extending iBGP-LU from CSC-CE to PE (discussed earlier) would eliminate this surface interaction completely, but create new ones. For example, PEs would need to originate PIM proxy vectors in leaf/spine POPs and label stack depths/MTUs would need to be recomputed. This creates newer, and in our professional opinion, more complex surface interactions between components.

Appendix A – Acronyms

Acronym	Definition
AAA	Authentication, Authorization, Accounting
ABR	Area Border Router
AC	Admission Control
ACL	Access Control List
AD	Administrative Distance
AF	Assured Forwarding
AGI	Attached Group Identifier
AH	Authentication Header
AS	Autonomous System
ASBR	AS Boundary Router
ASM	Any Source Multicast
ASN	AS Number
ATM	Asynchronous Transfer Mode
Bc	Burst Committed
Be	Burst Excess
BFD	Bidirectional Forwarding Detection
BGP	Border Gateway Protocol
CE	Customer Edge router
CIR	Committed Information Rate
CLI	Command Line Interface
CS	Class Selector

Acronym	Definition
CSC	Carrier Supporting Carrier
DAI	Dynamic ARP Inspection
DF	Default Forwarding
DHCP	Dynamic Host Configuration Protocol
DNS	Domain Name System
DSCP	DiffServ Code Point
EAP	Extensible Authentication Protocol
ECMP	Equal-cost Multi-path
EF	Expedited Forwarding
EVPN	Ethernet VPN
EXP	MPLS Experimental bits
FHRP	First Hop Redundancy Protocol
FRR	Fast Re-Route
GMV	Global Management View
GRE	Generic Routing Encapsulation
HTTP/S	Hypertext Transfer Protocol (Secure)
IaaS	Infrastructure as a Service
IGP	Interior Gateway Protocol
IP	Internet Protocol
I-PMSI	Inclusive PMSI
IPP	IP Precedence
IPSG	IP Source Guard
IR	Ingress Replication

Acronym	Definition
IS-IS	Intermediate System to Intermediate System
ISP	Internet Service Provider
iSPF	Incremental SPF
IXP	Internet Exchange Point
LAN	Local Area Network
LDP	Label Distribution Protocol
LFA	Loop Free Alternative
LIB	Label Information Base
LLQ	Low Latency Queuing
LSA	Link State Advertisement
LSP	Label-switched Path
LU	Labeled Unicast
MAC	Media Access Control (Ethernet)
MDT	Multicast Delivery Tree
MED	Multi-exit Discriminator (BGP)
MLD	Multicast Listener Discovery (IPv6 ICMP)
mLDP	Multicast LDP
MP2P	Multipoint-to-point
MPLS	Multi-protocol Label Switching
MS-CHAP	Microsoft Challenge Handshake Authentication Protocol
MTU	Maximum Transmission Unit
MVPN	Multicast MVPN
NAT	Network Address Translation

Acronym	Definition
NOC	Network Operations Center
NPTv6	Network Prefix Translation for IPv6
NTP	Network Time Protocol
OSPF	Open Shortest Path First
P2P	Point-to-point
PA	Provider Aggregate
PE	Provider Edge router
PEAP	Protected EAP
PHB	Per-hop Behavior
PI	Provider Independent
PIM	Protocol Independent Multicast
PIR	Peak Information Rate
PMSI	Provider Multicast Service Interface
POP	Point of Presence
QoS	Quality of service
RADIUS	Remote Authentication Dial-In User Service
RD	Route Distinguisher
RP	Rendezvous Point
RPF	Reverse Path Forwarding
RR	Route Reflector
RS	Route Server
RSVP	Resource Reservation Protocol
RT	Route Target

Acronym	Definition
SAN	Storage Area Network
SATCOM	Satellite Communications
SIP	Session Initiation Protocol
SNMP	Simple Network Management Protocol
SOS	State Optimization Surface
S-PMSI	Selective PMSI
SR	Segment Routing
SSH	Secure Shell
TACACS	Terminal Access Controller Access Control System
Tc	Time Committed
TE	Traffic Engineering
TLS	Transport Layer Security
VLAN	Virtual LAN
VMV	VPN Management View
VPLS	Virtual Private LAN Service
VPN	Virtual Private Network
VPWS	Virtual Private WAN Service
VRF	VPN Routing and Forwarding
VRRP	Virtual Router Redundancy Protocol
VXLAN	Virtual eXtensible LAN
WAN	Wide Area Network

Appendix B – References

Basic IP LFA Techniques (IETF RFC 5286)

Border Gateway Protocol (IETF RFC 4271)

BGP Labeled Unicast (IETF RFC 3107)

BGP MVPN Address Family (IETF RFC 6514)

BGP Route Servers in IXPs (IETF RFC 7947)

BGP/MPLS IP VPNs (IETF RFC 4364)

DiffServ QoS (IETF RFC 4594)

Draft Rosen MDT for MVPN (IETF RFC 6037)

Extensible Authentication Protocol (IETF RFC 3748)

Label Distribution Protocol (IETF RFC 5036)

Multicast in MPLS/BGP IP VPNs (IETF RFC 6513)

Navigating Network Complexity (White and Tantsura)

Network Prefix Translation for IPv6 (IETF RFC 6296)

NTP version 4 Specification (IETF RFC 5905)

Open Shortest Path First version 2 (IETF RFC 2328)

RADIUS Specification (IETF RFC 2865)

SNMP version 3 Specification (IETF RFC 3414)

TACACS Specification (IETF RFC 1492)

VPLS with BGP Discovery and LDP Signaling (IETF RFC 4762)