Next Generation Enterprise MPLS VPN-Based WAN Design and Implementation Guide

Cisco Validated Design I

October 1, 2007

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Cisco Validated Design

The Cisco Validated Design Program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit www.cisco.com/go/validateddesigns.

Next Generation Enterprise MPLS VPN-Based WAN Design and Implementation Guide

CONTENTS

CHAPTER 1

Solution Description

Enterprise customers have in the past relied heavily upon traditional WAN/MAN services for their connectivity requirements. Layer 2 circuits based on TDM, Frame Relay, ATM, and SONET have formed the mainstay of most low-speed WAN services. More recently, high-speed MAN solutions have been delivered directly over Layer 1 optical circuits, SONET, or through the implementation of point-to-point or point-to-multipoint Ethernet services delivered over one of these two technologies.

Today, many enterprise customers are turning to Multiprotocol Label Switching (MPLS)-based VPN solutions because they offer numerous secure alternatives to the traditional WAN/MAN connectivity offerings. The significant advantages of MPLS-based VPNs over traditional WAN/MAN services include the following:

- Provisioning flexibility
- Wide geographical availability
- Little or no distance sensitivity in pricing
- The ability to mix and match access speeds and technologies
- Perhaps most importantly, the ability to securely segment multiple organizations, services, and applications while operating a single MPLS-based network

Although service providers have been offering managed MPLS-based VPN solutions for years, the larger enterprises, universities, and federal and state governments are now beginning to investigate and deploy MPLS in their own networks to implement self-managed MPLS-based VPN services. The concept of self-managed enterprise networks is not new; many enterprise customers purchase Layer 2 TDM, Frame Relay, or ATM circuits and deploy their own routed network for these circuits. The largest of enterprise customers even manage their own core networks by implementing Frame Relay or ATM-based switching infrastructures and “selling” connectivity services to other organizations within their companies.

Both of these solutions have had disadvantages; deploying an IP-based infrastructure over leased lines offers little flexibility and segmentation capabilities that are cumbersome at best. Deploying a switched Frame Relay or ATM infrastructure to allow for resiliency and segmentation is a solution within reach of only the largest and most technically savvy enterprises.

As noted, the self-managed MPLS-based network is typically reserved for larger enterprises willing to make an investment in network equipment and training, with an IT staff that is comfortable with a high degree of technical complexity. A self-managed MPLS VPN can be an attractive option if a business meets these requirements and wants to fully control its own WAN or MAN and to increase virtualization across multiple sites to guarantee delivery of specific applications. There are alternate approaches to full-fledged MPLS implementations such as Multi-VRF or a combination of both MPLS and Multi-VRF that allow existing networks to be easily transitioned to virtualized ones. The level of security between separated networks is comparable to private connectivity without needing service provider intervention, allowing for consistent network segmentation of departments, business functions, and user groups.

Corporations with a propensity for mergers and acquisitions benefit from the inherent any-to-any functions of MPLS that, when the initial configuration is completed, allow even new sites with existing networks to be merged with the greater enterprise network with minimal overhead. Secure partner networks can also be established to share data and applications as needed, on a limited basis. Theself-managed MPLS is also earning greater adoption as an important and viable method for meeting and maintaining compliance with regulatory privacy standards such as HIPAA and the Sarbanes-Oxley Act.

While the technology enables you to create the logical separation across networks, it is important to understand the reasons for creating these logical networks. Enterprise customers increasingly require segmentation for a number of different reasons:

- Closed User Groups (CUG)—The CUGs could be created based on a number of different business criteria, with guest Internet access for onsite personnel being the simplest example. Providing NAC/isolation services also creates a need to separate the non-conforming clients. While this can be done using VLANs within a Layer 2 campus network, it requires Layer 3 VPN functionality to extend it across Layer 3 boundaries. CUGs could be created with partners, either individually or as a sub-group, where the segmentation criteria are resources that are to be shared/accessed. This simplifies the information sharing with partners while still providing security and traffic separation.
- Virtualization—Segmentation to the desktop is driving virtualization in the application server space. This means that even existing employees can be segmented into different CUGs where they are provided access to internal services based on their group membership.
- Enterprise as a Service Provider—With some of the enterprise networks expanding as their organization expands, IT departments at some of the large enterprises have become internal Service Providers. They leverage a shared network infrastructure to provide network services to individual Business Units within the enterprise. This not only requires creating VPNs, but also requires the ability of each of the BUs to access shared corporate applications. Such a model can be expanded to include scenarios in which a company acquires another company (possibly with an overlapping IP addressing scheme) and needs to eventually consolidate the networks, the applications, and the backoffice operations.
- Protecting critical applications—Another segmentation criteria could be based off the applications themselves rather than the users. An organization that feels that its critical applications need to be separated from everyday network users can create VPNs for each or a group of applications. This not only allows it to protect them from any malicious traffic, but also more easily control user access to the applications. An example of this is creating separate VPNs for voice and data.

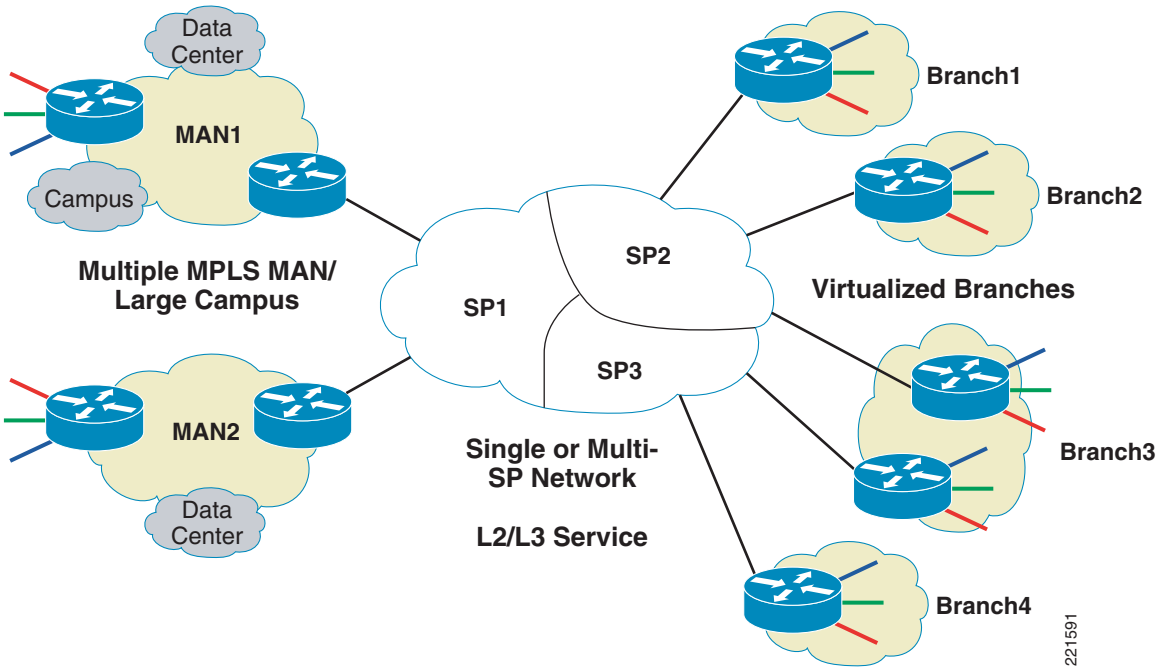
Beyond the segmentation criteria, the overarching consideration should be based on the need to share. The VPNs create a closed user group that can easily share information, but there will always be the scenario that requires sharing across the VPNs. For example, a company-wide multicast stream would need to be accessible by all the employees irrespective of their group association. Thus the VPNs should be created based on practical considerations that conform to the business needs of the organization.

The first phase of the solution provided design guidelines for creating a self deployed MPLS MAN for the enterprise. It focused on Layer 3 VPNs and Layer 2 VPNs deployments, multicast and voice services supported by a end-to-end QoS model. It provided models for shared services deployments such as Internet access.

The next logical step for expanding virtualization across enterprise networks is to extend it into two other areas (Figure 1-1):

- Connecting large MAN/campus MPLS networks
- Remote branches

Figure 1-1 NGMAN/WAN 2.0—Solution Scope



While the MAN deployment in phase 1.0 was characterized by higher bandwidth requirements, WAN deployment (especially branch aggregation) is characterized by higher scale requirements to support 100s to 1000s of sites. A WAN may be a Layer 2 or Layer 3 service and may be spread across providers. A Layer 3 WAN may require some form of overlay network to support enterprise virtualization. The virtualization deployment in the WAN also has a important bearing on how the sites are integrated with the core MPLS network.

The remaining chapters explore the different deployment models for inter-MAN MPLS connectivity and virtualized branch aggregation. It focuses on data, voice, and multicast services along with QoS. Each of the models is discussed with lab tested and deployable examples.

For a technology refresher and MPLS MAN deployment, refer to the phase 1.0 DIG:

http://www.cisco.com/application/pdf/en/us/guest/netsol/ns241/c649/ccmigration_09186a008055edcf.pdf

CHAPTER 2

Deployment Architectures

As mentioned earlier, network virtualization extension across the WAN can be broadly classified into two categories:

- Inter-MAN/ large campus connectivity or WAN core
- Virtualized branch aggregations or WAN edge

WAN Core

After creating MPLS MAN islands, this is the next logical step when migrating to MPLS-based enterprise networks. The different options for interconnecting MPLS MANs are:

- MPLSoL2 service—If it is a legacy Layer 2 or Layer 2 VPN service from SP
- MPLSoGRE—If it is a Layer 3 VPN service from SP
- Carrier Supporting Carrier (CSC)

Typically, in a large enterprise, the WAN core consists of dedicated point-to-point, high-bandwidth links. We do not expect these to move to Layer 3-based (such as Layer 3 VPNs) connections because these are deemed critical links that require the fastest possible round-trip times and higher bandwidths —hence Layer 2circuits are preferred. Additionally, these are few in numbers and hence cost advantages of Layer 3 services are not necessarily applicable.

While all three options are discussed in this chapter, only MPLSoL2 Service is discussed in depth in WAN Core—MPLSoL2 Service since it is expected to be the most-widely deployed.

MPLSoL2 Service

This is the simplest deployment model for connecting MANs if the enterprise already has Layer 2 connectivity between them either via legacy WAN (FR/ATM/POS) or via Layer 2 VPN service (AToM) from a provider. The migration involves converting the edge devices into a P/PE and making it part of the MPLS MAN network.

The MANs are assumed to be already MPLS enabled and configured for enterprise-deployed VPNs. As shown in Figure 2-1, the WAN edge router used for interconnecting MANs plays the role of a P device. It is expected to label switch packets between the MANs across the SP network.

Figure 2-1 MPLS Over Legacy Layer 2 Service

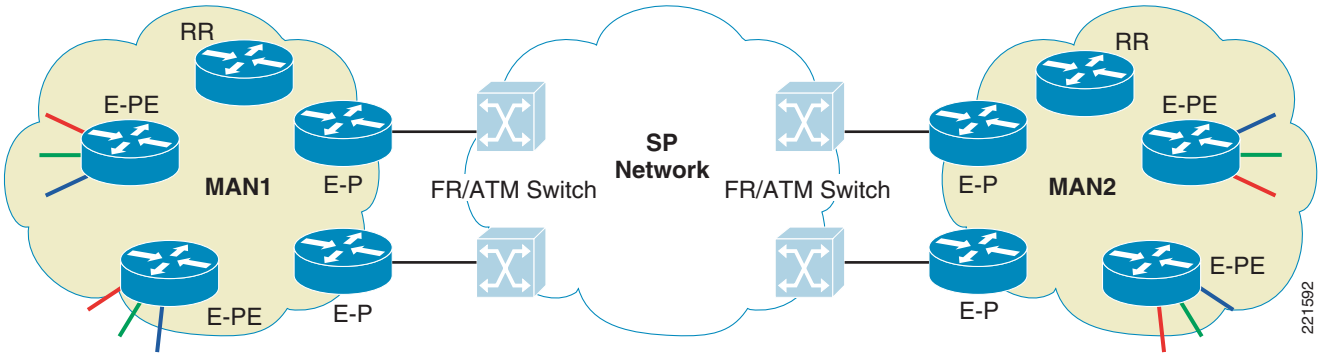
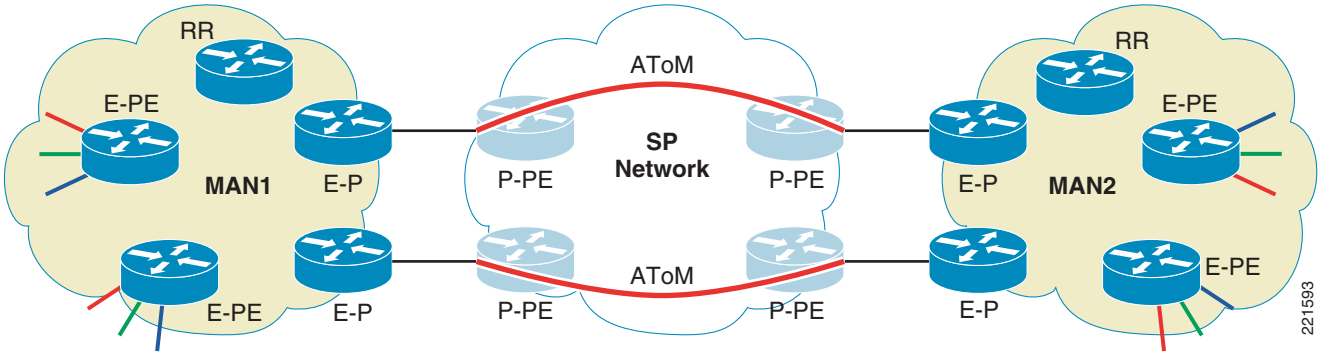


Figure 2-2 MPLS Over Layer 2 VPN Service



From a control plane perspective, the following are expected to be run over the Layer 2 link:

- IGP such as EIGRP or OSPF for MPLS device reachability (P/PE/RR)
- LDP for label distribution
- MP-iBGP for VPN route/label distribution

If these MAN islands/campuses are under different administrative control, then Inter-AS can be implemented. Typical Inter-AS models are:

- Back-to-back VRFs
- ASBR-to-ASBR with MP-eBGP
- ASBR-to-ASBR with multihop EBGP using Route Reflectors

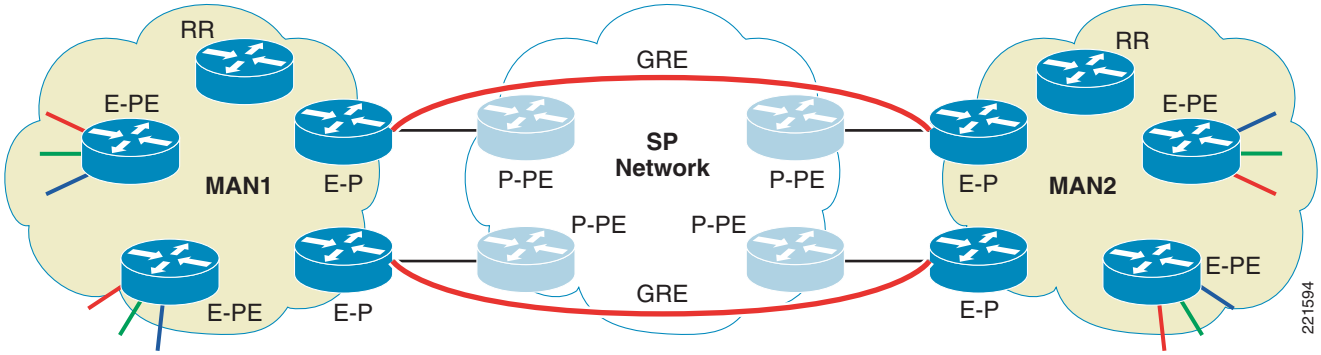
Apart from being a simple solution to deploy, it also offers wider platform options. All the platforms that support P roles should be deployable. All the features that would be deployed within a MPLS network (such as TE) can also be deployed across the WAN core.

MPLSoGRE

The implementation assumes that the enterprise has a Layer 3-based service such a Layer 3 VPNs from a provider interconnecting the MPLS MANs. The MANs may have multiple connections between them to provide load balancing and/or redundancy. It might be desirable to obtain the redundant connectivity services from multiple providers.

As shown in Figure 2-3, the WAN edge router used for interconnecting MANs plays the role of a P device even though it is a CE for the SP VPN service. It is expected to label switch packets between the MANs across the SP network.

Figure 2-3 MPLS over GRE

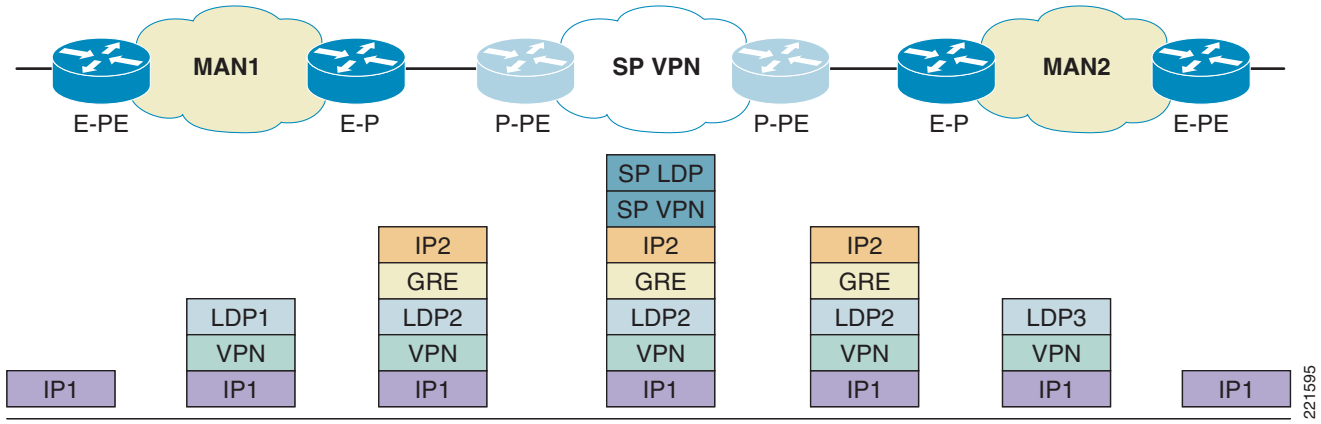


A point-to-point GRE tunnel is set up between each edge router pair (if full mesh is desired). From a control plane perspective, the following are expected to be run within the GRE tunnels:

- IGP such as EIGRP or OSPF for MPLS device reachability (P/PE/RR)
- LDP for label distribution
- MP-iBGP for VPN route/label distribution

Once the route/label distribution is done in the control plane, the enterprise edge device acts like a label switching router (LSR/P) where it treats the GRE interfaces as normal access interfaces. Figure 2-4 shows end-to-end packet flow between campuses across different MANs.

Figure 2-4 MPLS over GRE—Forwarding Plane



As can be seen from the headers, this adds a large amount of overhead to the MTU and hence is not the most desired option. In addition, platform support is also limited to the 7200 and ISRs. Thus none of the high-end platforms (7600, 12000) support it. 7600 supports MPLSoGRE in PE-PE mode only, but ideally we would prefer P-P setup for inter-MAN connectivity. Thus MPLSoGRE is not a preferred option for Inter-MAN connectivity.

Carrier Supporting Carrier (CSC)

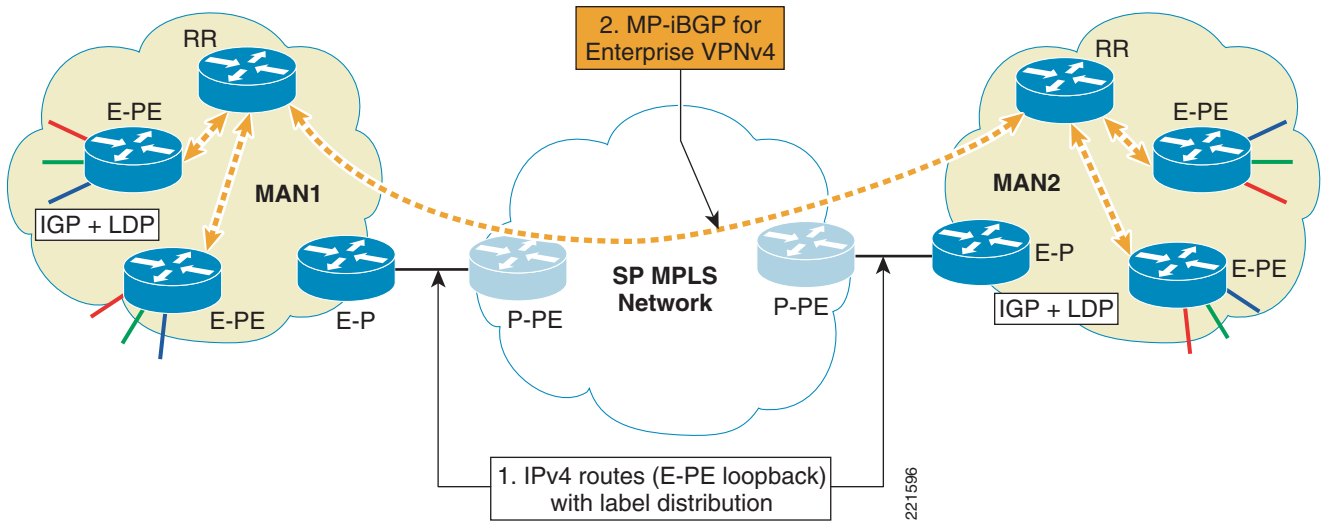
The Carrier Supporting Carrier was developed for MPLS enabled SPs to support other MPLS/VPN SPs. With the growth in enterprise network growth, CSC is a service that SPs can provide to large enterprises as well. For an enterprise, this involves getting a label transport service from a provider. The enterprise edge devices perform the P role in the enterprise MPLS network.

The advantages of a label transport service include the fact that there is no overlay such as GRE required. The SP provides any-to-any connectivity as well, so multiple dedicated links are not required between the MANs. Based on the incoming label, the SP network ensures that the packet is forwarded to the correct MAN location.

From a control plane perspective there are two major elements (Figure 2-5):

1. A IPv4 route and label exchange between the enterprise edge P and the Provider PE—the only routes that needed to be exchanged are for the PE and RR reachability (loopbacks). This can be achieved in two ways:
   a. Running a IGP with the Provider PE to advertise the loopback addresses and running LDP to advertise the labels associated with those loopback addresses.
   b. Alternatively, the SP may prefer to run EBGP+label option to advertise the loopback addresses along with their associated labels.
2. Enterprise running MP-iBGP between its RRs to exchange VPNv4+lable information for its own VPNs.

Figure 2-5 Carrier Supporting Carrier (CSC)

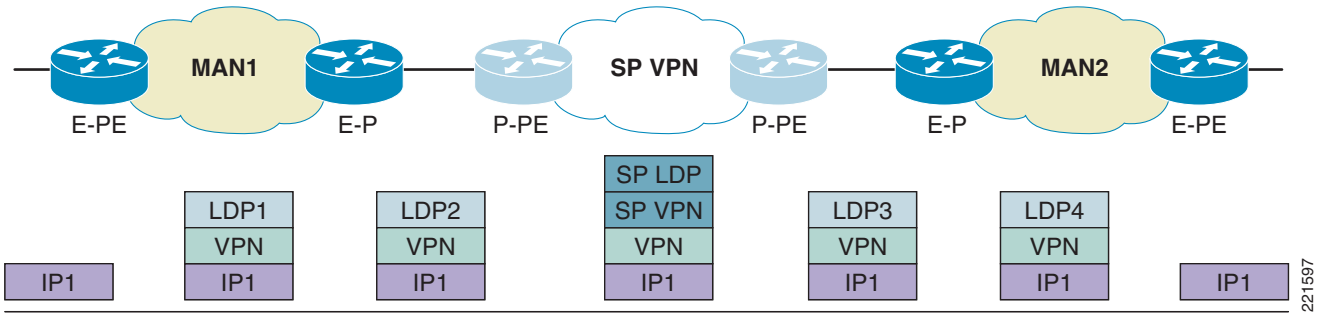


In the forwarding plane, as shown in Figure 2-6, the following occurs:

- The P router at the enterprise edge (E-P1) receives a labeled packet from the MAN and label switches it based on the label learned via the provider for E-P2.
- The provider PE (P-PE1) label switches the incoming top label with the label advertised via P-PE2. This is treated as the "VPN" label within the provider network.
- P-PE1 prepends the top label (SP LDP) for P-PE1 to P-PE2 reachability.
- P-PE2 receives the packet with SP LDP label popped (if PHP is enabled). It label switches the exposed label corresponding to the one advertised by E-P2.

- E-P2 label switches the top label and sends it across the MAN2 network to the appropriate PE.

Figure 2-6 Carrier Supporting Carrier—Forwarding Plane



Depending on the size of the deployments, the platforms can range from 12000, 7600, 7304, 7200, or 3800.

CSC does have two major draw backs:

- There is a very limited offering from the SPs to the enterprises—their CSC service is designed primarily for other SPs. Additionally, since the VPNs are now maintained by the enterprises, the SP essentially sells a single VPN service.
- Since it is essentially a Layer 3 service from a provider, the enterprise may wish to encrypt the traffic, which is not feasible as currently there are no mechanisms that allow for encryption of labelled packets.

WAN Edge

If enterprises have requirements for virtualization of the branches, then the following deployment models are technically available:

- Multi-VPN service from SP
- Carrier Supporting Carrier
- MPLSoL2 infrastructure
- Self-Deployed Multi-VRF with mGRE/DMVPN (DMVPN per VRF)
- MPLS VPN over DMVPN— 2547oDMVPN (Hub and Spoke only)
- MPLS VPN over IP using L2TPv3 (2547oL2TPv3)

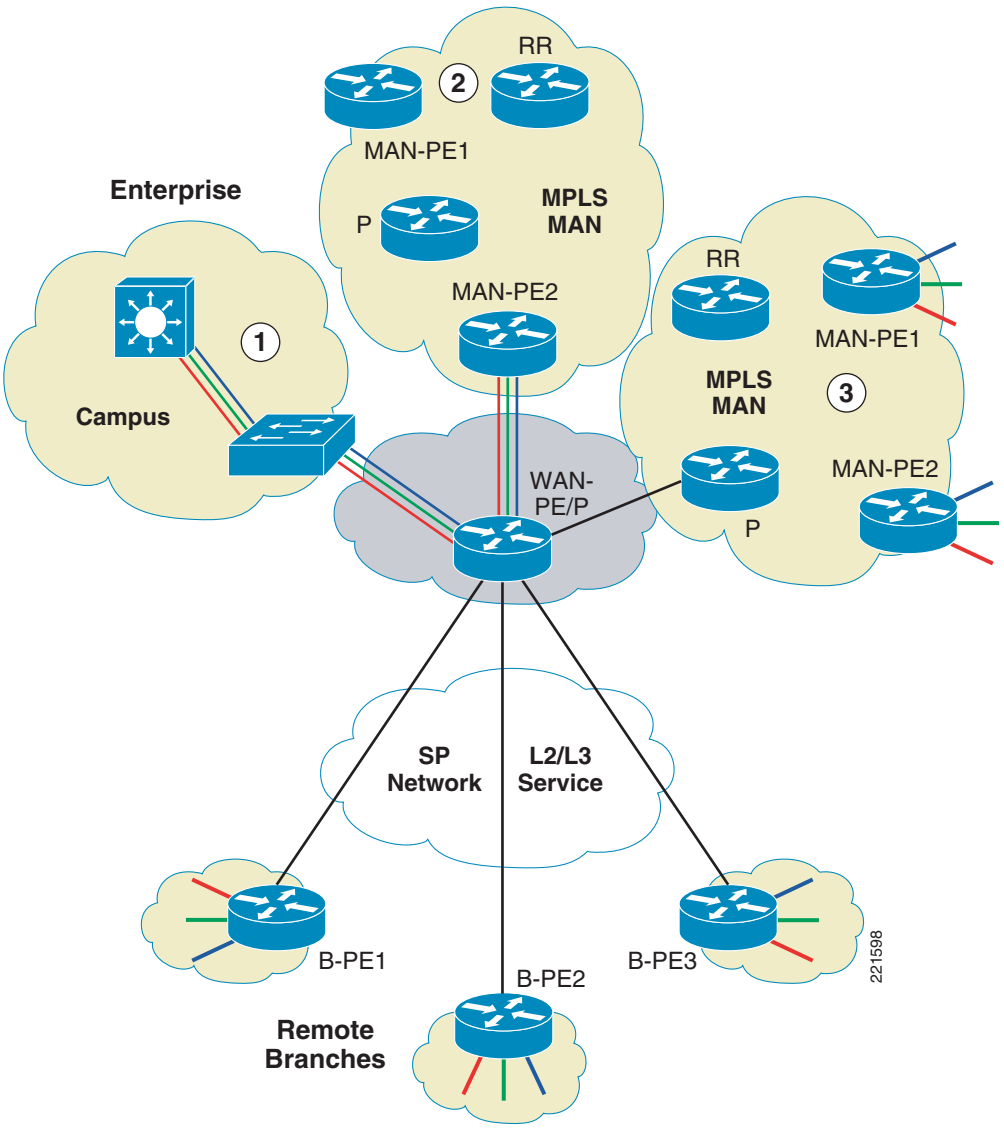
There are scenarios where virtualization may be required at the large branches only. The rest of the branches may not have any VRFs, but may have to be placed in there own VRF, default common VRF, or global table at the headend depending on enterprise requirements.

The deployment models should support integration of branches ranging from 10s to 1000s. The interface types on the headend are expected to range from the DS3 to GE. The speeds on the branches would typically be T1 or below.

WAN Edge Integration

The deployment models listed above allow connectivity between the segmented branches and the headend. The headend itself (as shown inFigure 2-7) may be connected to the rest of the core network in at least three different ways (discussed below). Not every model may support these integration options.

Figure 2-7 Integrating WAN Edge with MPLS MAN



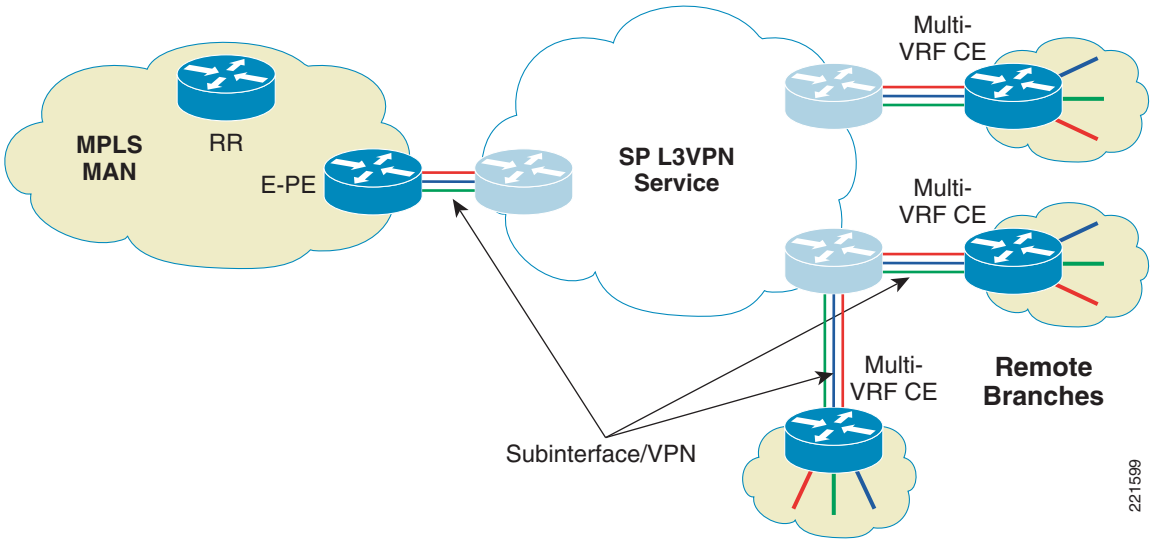
1. Direct connectivity with campus—In this scenario the WAN edge router is running in VRF-lite mode towards the campus where the VRFs from the branches are extended into the campus by using VLANs. The assumption here is that campus is virtualized using a combination of VLANs and VRF-lite as well. A separate routing instance is running per VRF within the campus that is extended to the branches via the WAN edge device. This can be deployed with any of the deployment models above.

2. Back-to-back PEs with MPLS MAN—The WAN edge router is running VRF-lite mode towards the core network in this scenario as well, but instead of connecting into a VRF-lite campus, it connects back-to-back with a MPLS MAN PE. This option is ideal for scenarios where the WAN Edge router may have P capabilities to extend the core MPLS to the branches or if it cannot be directly made into a core PE. This option can be deployed with any of the models above.
3. Direct Connectivity with MPLS MAN—This is the easiest integration option since the WAN edge router performs the P functionality and the branch PEs are integrated directly with the core MPLS network. This is ideal for MPLSoL2 and 2547oDMVPN deployments although its currently not supported fully in the later scenario.

Multi-VPN Service from Provider

A simple solution for enterprises to extend virtualization to the branches is to obtain multiple Layer 3 VPN services from a provider. As shown in Figure 2-8, the branch routers become Multi-VRF CEs and the headend may be a Multi-VRF CE or a PE (if it is directly connected to the MPLS MAN). This may be a desirable solution from a cost perspective if there are a small number of branches that have a requirement for virtualization and the number of VRFs is low.

Figure 2-8 Multi-VPN service from SP



In the control plane each of the CEs run a routing protocol such as OSPF, EIGRP or BGP with the SP PE on per VRF basis. Thus any design recommendations implemented while getting a single VPN service (in terms of routing, QoS, Multicast, etc.) would have to be followed for each of these VPN instances as well.

Carrier Supporting Carrier

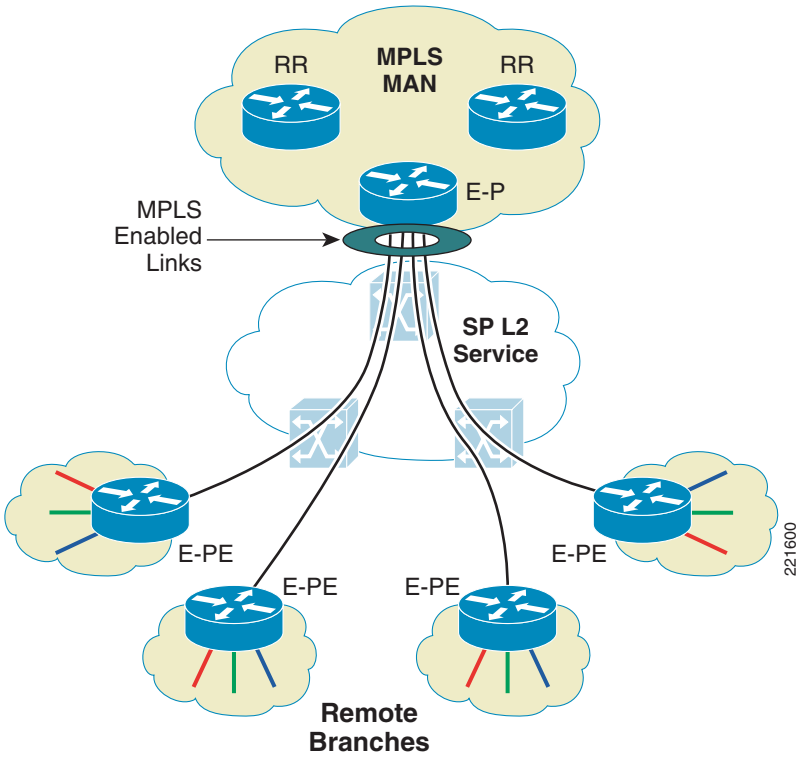
This is same labeled transport service that was discussed in Carrier Supporting Carrier (CSC). The requirement here is limited to branch routers such as ISRs.

MPLSoL2 Service

This model assumes that the enterprise has existing Layer 2 services for connecting branches and wants to enable MPLS over them. Since such Layer 2 connectivity is typically hub and spoke or partial mesh, the MPLS overlay also inherits the same connectivity characteristics. If spoke-to-spoke communication is required, it has to be handled via the hub.

The branch aggregation router is converted into a P role for the MPLS network and is expected to label switch packets as shown in Figure 2-9. The branch routers become PE routers with VRF interfaces facing the branch and MPLS-enabled interface facing the headend.

Figure 2-9 MPLSoL2 Service



In the control plane, each of the remote branch PEs would have a LDP session and a IGP session with the headend aggregator. They would also have MP-iBGP sessions with the route reflectors that would typically reside behind the headend aggregating device.

In cases where virtualization is not required at certain branches, then those branch routers do not need to have their WAN connection MPLS enabled. On the headend, depending on the connectivity model (point-to-point vs. multipoint) and interface flexibility, the enterprise has a few options:

- If using multipoint interface(s) at headend, then separate the MPLS-enabled and the non-MPLS connections into separate multipoint groups. Within the non-MPLS group, they may need to be further separated based on the VRF(s) into which they need to be placed.
- If using point-to-point interfaces, then each individual connection can be MPLS enabled or placed in a VRF.
- A combination of point-to-point and multipoint interfaces can be supported as well.

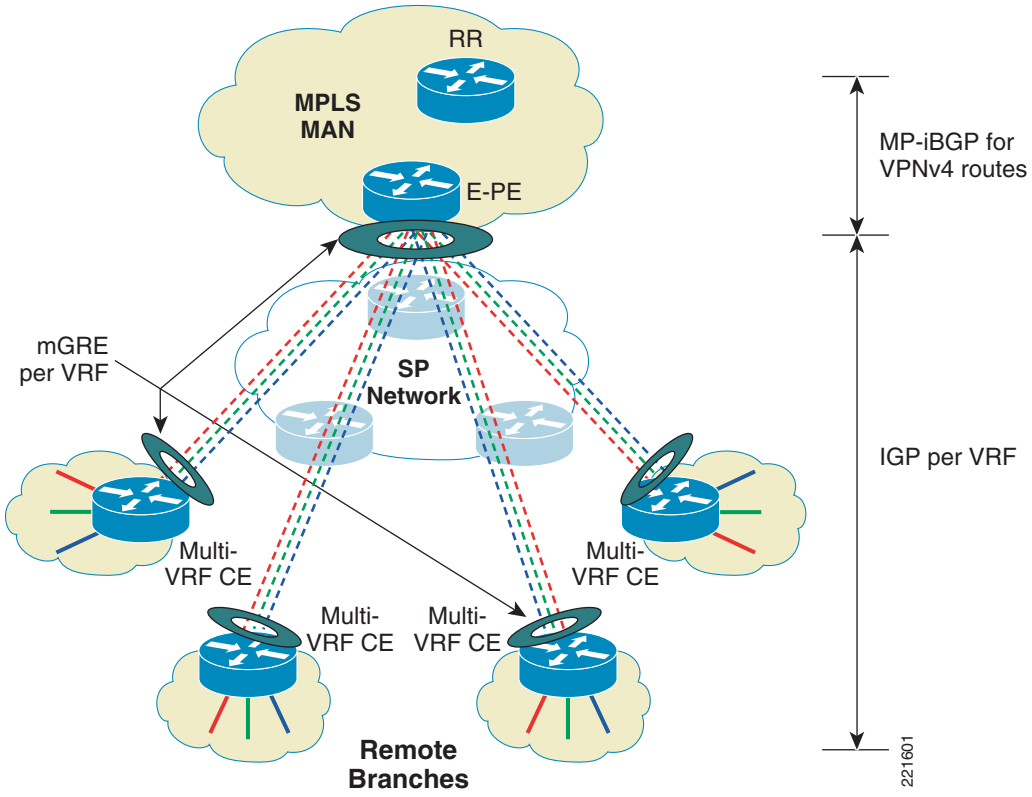
Note In each of the cases, the aggregation device have to perform both the roles—P as well as PE.

DMVPN per VRF

This model can be used over a Layer 2 or Layer 3 service from a provider. If it is a Layer 3 VPN service, then the enterprise purchases only a single VPN from the provider but overlays its own VPNs by using a combination of Multi-VRF and GRE. The headend has a mGRE tunnel per VRF, the branches have either GRE (if no spoke-to-spoke communication is required) or mGRE (if spoke-to-spoke communication is required) tunnels per VRF. DMVPN provides for spoke-to-spoke communication as well as encryption (although encryption is optional for our deployment model). By configuring mGRE on certain spokes, it provides them with the ability to create dynamic tunnels to other spokes (which should be configured with mGRE as well) on a per-VRF basis.

Most enterprises only have a partial mesh requirement—large sites need to be meshed together but the smaller sites are typically only hub and spoke. Thus the deployment is expected to be a combination of GRE and mGRE at the spokes (Figure 2-10).

Figure 2-10 DMVPN per VRF



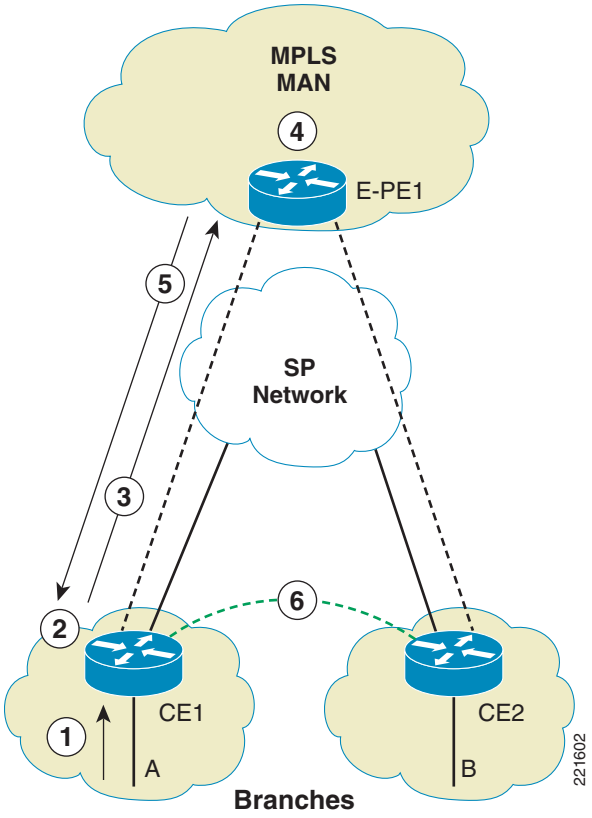
The hub device, while aggregating the branches, is also a PE for the MPLS MAN network. It has a IGP instance running within each VPN with each of the spokes. The IPv4 addresses learned from the spokes are converted to VPNv4 addresses before being advertised to the RRs using MP-iBGP. IGP might limit the scale of the deployment requiring the use of multiple hub routers. Scaling options include:

- Use of BGP as the hub and spoke protocol with dedicated RRs for this purpose

- Having multiple termination devices at the headend based—for example, one for each VRF if there are a low number of VRFs

DMVPN uses NHRP to keep track of the next-hop to physical address mapping. The hub is the NHRP server that maintains the mapping table on per VRF basis. In the example shown in Figure 2-11, once the GRE tunnel is established with the hub, both Branch1 (CE1) and Branch2 (CE2) register with the hub (E-PE1) using NHRP on per VRF basis. The hub learns about each of the branch VPN routes and advertises them back out to the other branches.

Figure 2-11 DMVPN per VRF—Spoke-to-Spoke Communication

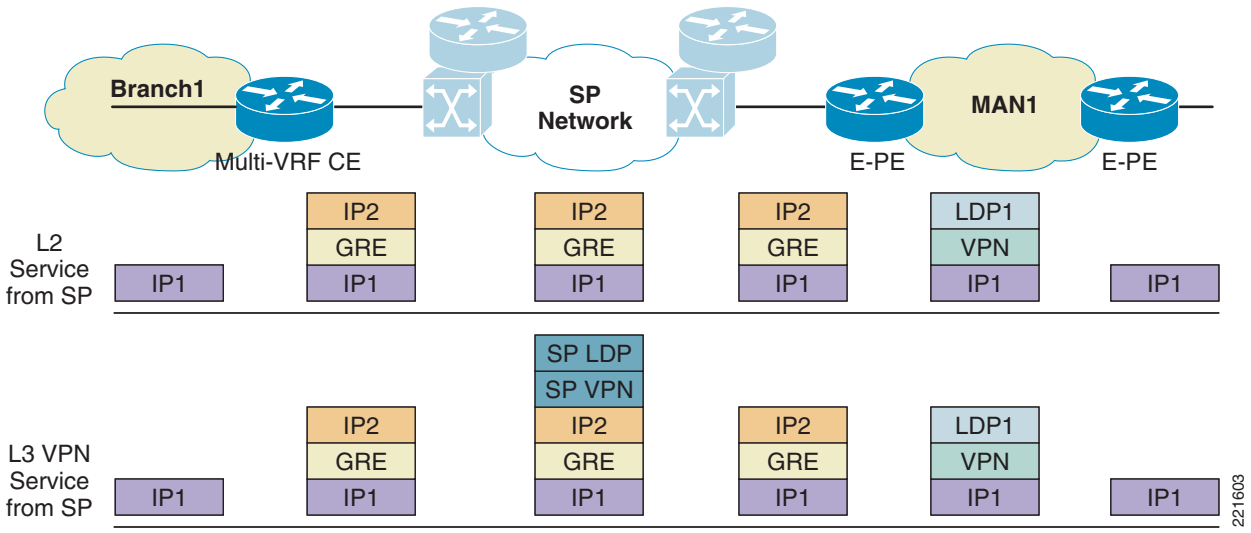


As shown in Figure 2-11, in the forwarding plane, the following sequence occurs on a per-VRF basis:

1. A sends packets to CE1 destined for B.
2. CE1 looks up its VRF table and finds B with next hop of CE2's GRE tunnel address.
3. CE1 sends a NHRP query to E-PE1 to resolve the next-hop address.
4. E-PE1 looks up the per-VRF NHRP database and associates CE2's tunnel address with its physical interface address.
5. E-PE1 sends the NHRP response back to CE1 with CE2's physical interface address.
6. CE1 sets up the dynamic GRE tunnel with CE2.

Figure 2-12 shows the end-to-end packet encapsulation when using this model. The packets from the remote branches are encapsulated in GRE on a per-VRF basis and forwarded across the SP network. The enterprise PE at the hub site decapsulates the GRE headers and performs label pushing based on the VRF in which the GRE interface is configured. The packets are then forwarded as normal MPLS VPN packets across the MAN (with the VPN and the LDP label).

Figure 2-12 DMVPN per VRF—Forwarding Plane



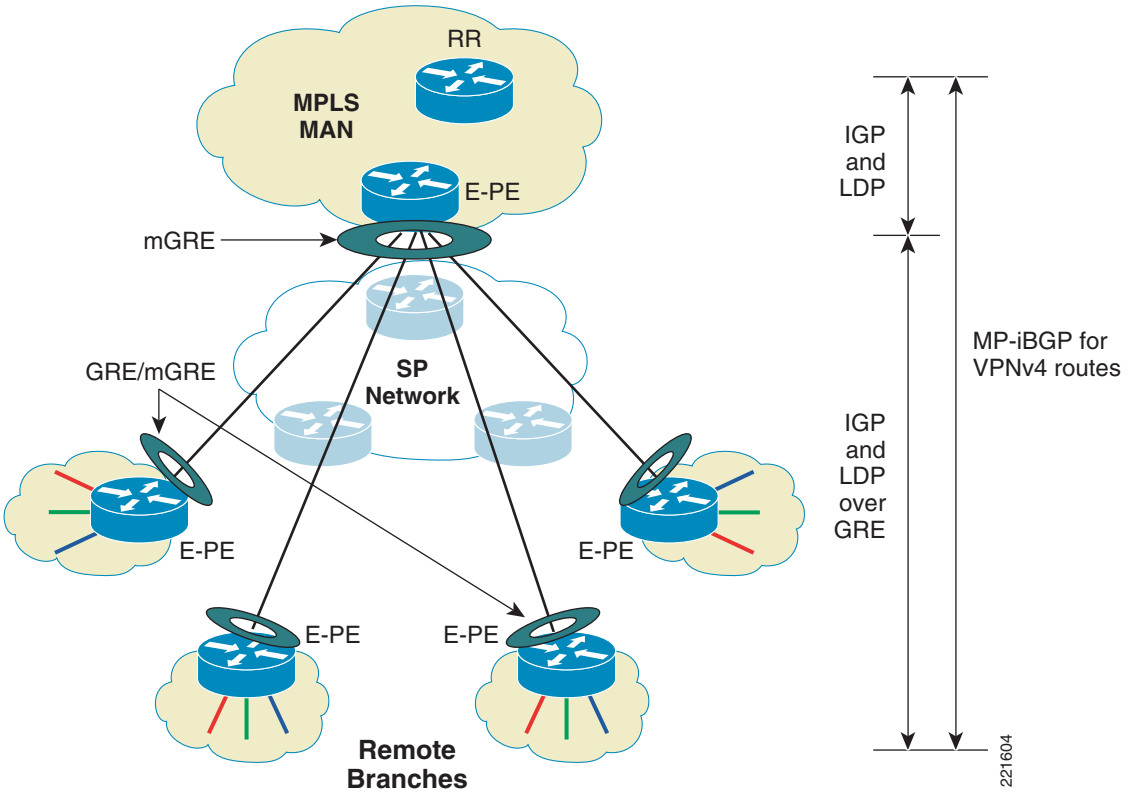
The hub site PE can be a 7600 or a 7200 depending on the scale or encryption requirements. The spokes can be aa ISR such as 2800 and 3800 depending on the branch requirements in terms of number of VRFs, throughput, and other non-transport related features.

MPLS VPN over DMVPN—2547oDMVPN (Hub & Spoke Only)

This model does not have some of the scale limitations of the Multi-VRF based solutions because the GRE tunnels are created outside the VRFs and hence a single tunnel can be shared for transporting many VRFs. The hub is configured with a single mGRE tunnel while spokes have a single GRE tunnel.

Note This is designed to be used for hub and spoke communication only and currently the dynamically created spoke-to-spoke tunnels are not supported.

Figure 2-13 2547oDMVPN (Hub & Spoke Only)



As shown in Figure 2-13, in the control plane the following protocols exist:

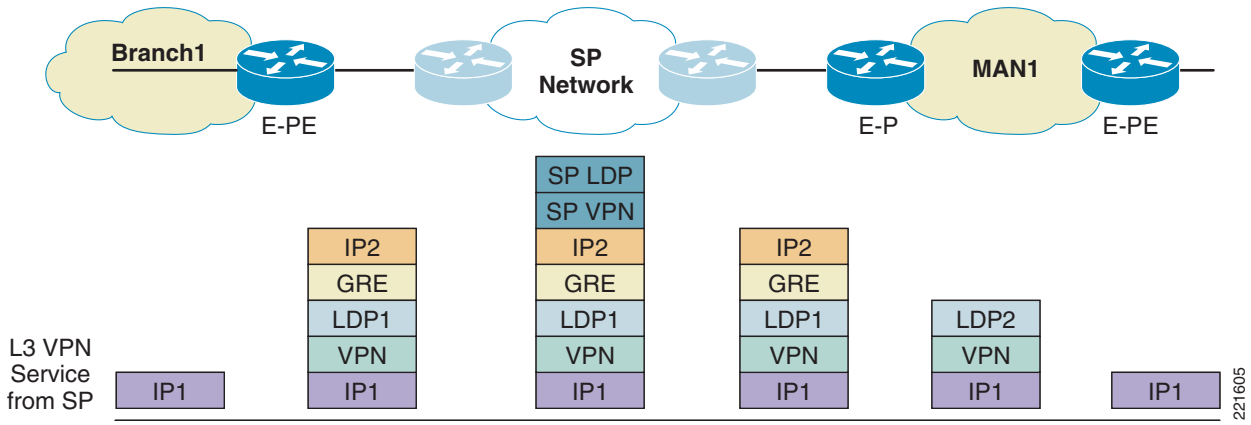
- Routing protocol of the provider to learn the Branch and headend's physical interface addresses (tunnel source address). Statics could be used as well if these are easily summarizable.
- GRE tunnel between the branch PE and the headend P.
- IGP running in the enterprise global space over the GRE tunnel to learn remote PE's and RR's loopback address.
- LDP session over the GRE tunnel with label allocation/advertisement for the GRE tunnel address by the branch router.
- MP-iBGP session with RR, where the branch router's BGP source address is the tunnel interface address—this forces the BGP next-hop lookup for the VPN route to be associated with the tunnel interface.

Additionally, IPsec can be used to encrypt the GRE tunnels; encryption happens after the GRE encapsulation.

Hub as a P Router

As shown in Figure 2-14, the branch router attaches the appropriate VPN label for the destination along with the LDP label advertised by the hub P for the destination next-hop address. It then encapsulates the labeled packet in a GRE tunnel with the hub P as the destination before sending it to the provider. Since in this example SP is providing Layer 3 VPN service, it further prepends its own VPN and LDP labels for transport within its network. The hub P receives a GRE encapsulated labeled packet. It decapsulated the tunnel headers before label switching it out to the appropriate outgoing interface in the MPLS MAN for the packet to reach the eventual PE destination.

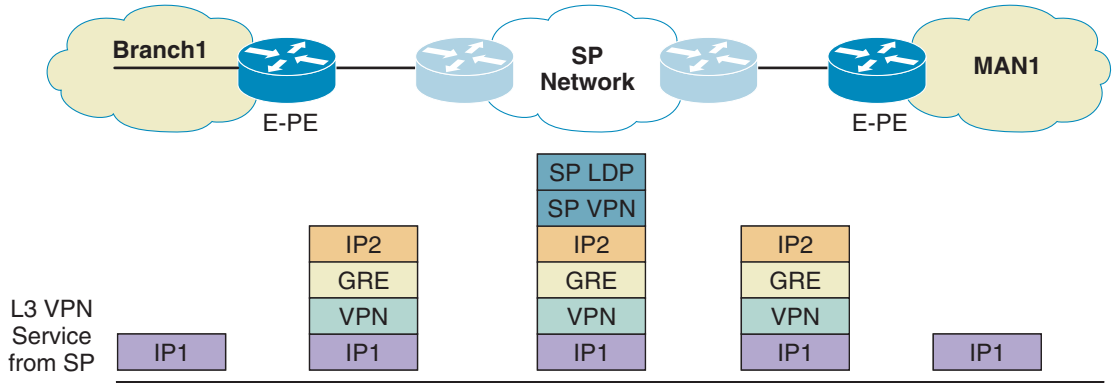
Figure 2-14 2547oDMVPN—Forwarding Plane with Hub as a P



Hub as a PE Router

As shown in Figure 2-15, the branch router attaches the appropriate VPN label for the destination advertised by the hub PE router. It then encapsulates the labeled packet in a GRE tunnel with the hub PE as the destination before sending it to the provider. Since in this example SP is providing Layer 3 VPN service, it further prepends its own VPN and LDP labels for transport within its network. The hub PE receives a GRE encapsulated labeled packet. It decapsulated the tunnel headers before forwarding it out to the appropriate outgoing interface based on the VPN label information and the VRF routing table.

Figure 2-15 2547oDMVPN—Forwarding Plane with Hub as a PE



MPLS VPN Over IP Using L2TPv3—2547oL2TPv3

This is currently not supported on the relevant platforms (7600, ISRs) and hence is not discussed, but is listed for the sake of completeness.

The rest of the guide focuses on providing design and implementation guidelines for the following deployment model for WAN Core:

- MPLSoL2 Service

The rest of the guide also focuses on providing design and implementation guidelines for the following deployment model for WAN Edge:

- MPLSoL2 Infrastructure
- DMVPN per VRF
- 2547oDMVPN

CHAPTER 3

WAN Core—MPLSoL2 Service

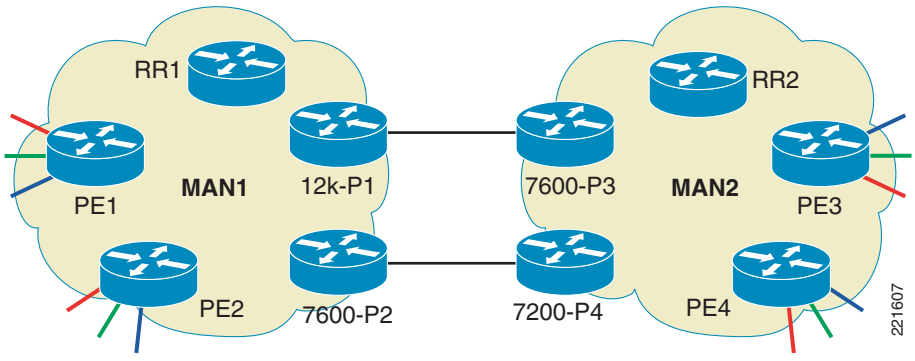
As stated earlier WAN Core typically consists of connecting large corporate islands such as MANs or large campuses. These are typically dedicated high bandwidth point-to-point connections. If these large islands are already MPLS enabled. then it is recommended to convert the edge devices into Provider (P) routers. This creates an integrated but flat MPLS network.

Platforms

The edge platforms should be capable of performing a P role, beyond which the selection is based on the interface/throughput requirements. Thus any Cisco platform above the 7200VXR series can be used as the edge router.

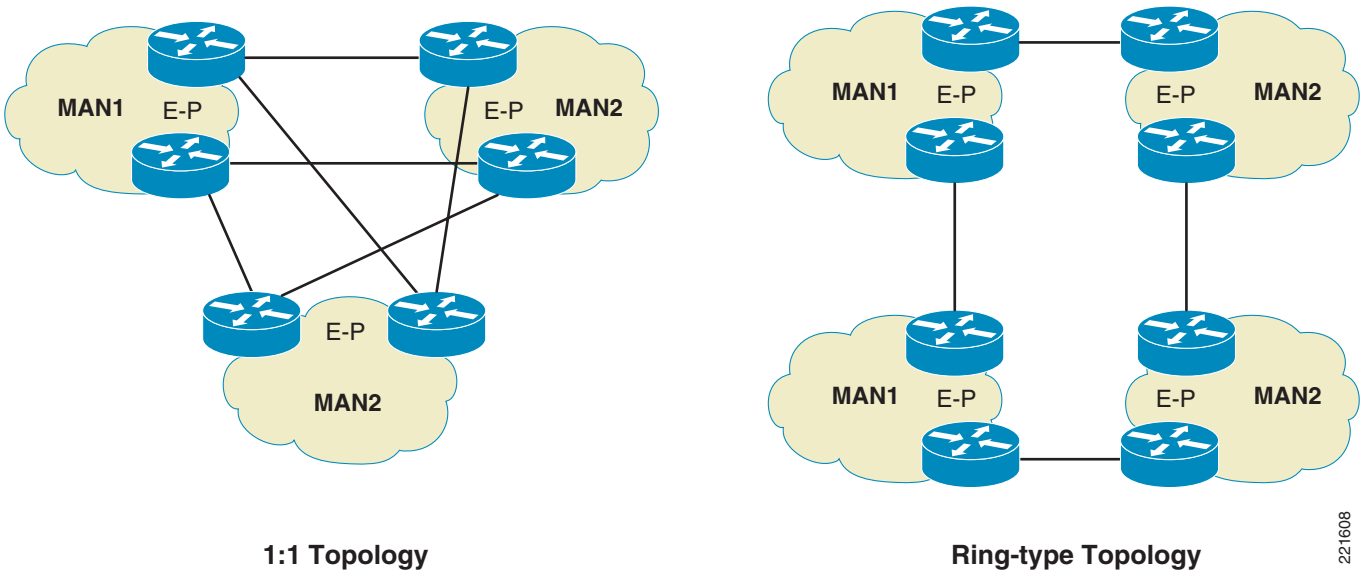
In Figure 3-1, we have a combination of GSR, 7600, and 7200 connecting to two large MPLS networks. Each network has its own sets of P, PE, and RR routers. But by connecting the P routers together as shown, we can create a single integrated network where the VPN traffic can access both networks seamlessly. No major changes are required in the existing networks. Functionally, these edge routers act like the other P routers in their respective locations and label switch traffic between locations.

Figure 3-1 MPLSoL2



While our test network only had two locations connected together, as shown in Figure 3-2, any number of networks could be connected by connecting the edge P routers. In certain cases, the edge P router could become transit for other sites. In such case care should be taken to ensure that the platform and the links are properly scoped for such scenarios. Additionally, enough redundancy should be built in to account for any single point of failure. Having multiple links between two sites allows the traffic to be load balanced. The links can be chosen based on the IGP metric to the PE next-hop address. In case of a single link failure, the traffic should reconverge and all traffic should transit the remaining up links.

Figure 3-2 MPLSoL2—Multi-Site Topologies



1:1 Topology **Ring-type Topology**

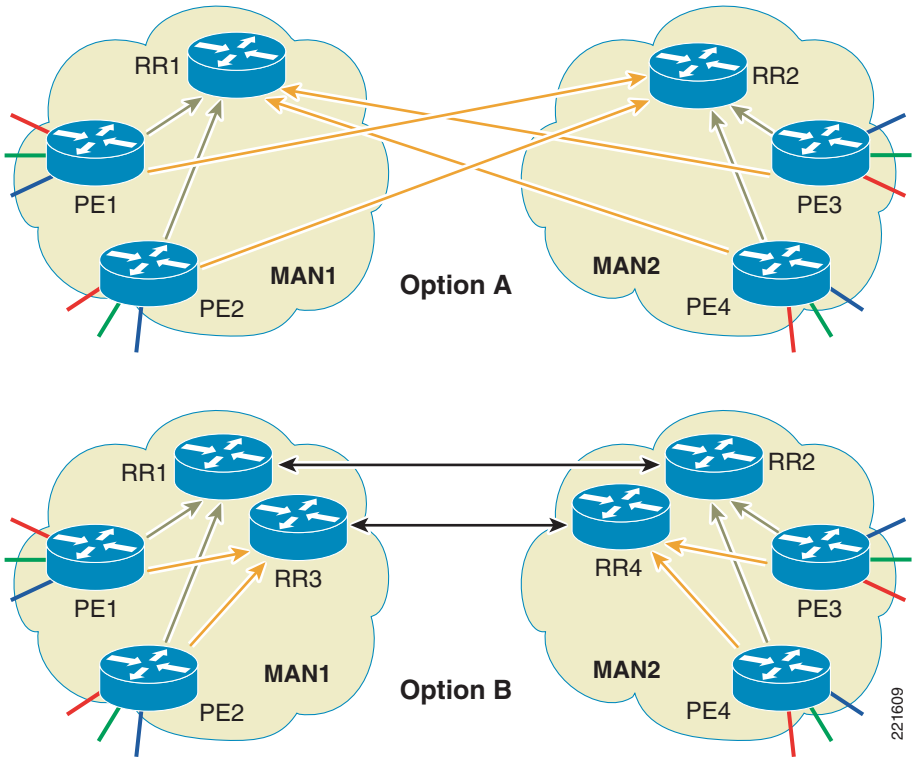
The only change that may be required is extending the IGP across the two networks. Since the assumption is that the two networks are under the same administrative control, the IGPs should be the same but may require some redesign. The ultimate goal is to provide reachability between the PE next-hop addresses. Typical recommendations are to allocate a subnet address with netmask of /32 from the same /24 subnet and not summarize it anywhere in the network. This ensures that a label is assigned for each /32.

Another change that may be required is in the Route Reflector placement or configuration. Figure 3-3 shows two of the options:

- Option A has a RR per MAN but every PE is peered to both the RRs. This provides RR redundancy where a loss of RR does not affect the VPNs. The RRs do not need to be peered to each other.
- Option B has two RR per MAN. The PEs in each MAN peer with their local RRs. To be able to learn the VPN routes from the other MAN, the RRs are peered with each other in a full mesh to provide additional redundancy. This option creates an additional level of hierarchy that can have an impact on convergence (more RR hops thorough which the VPN route needs to be propagated).

Option A provides the simplest implementation while option B provides better scale and partitioning capabilities that may be required when servicing large number of PEs and VPN routes.

Figure 3-3 Route Reflector Peering



Note While in large networks (SP-type) it is recommended to have dedicated RRs which are placed out of the forwarding path, in smaller enterprise networks it is possible to use an existing PE as a RR. Care should be taken to scope the additional processing overhead on the PE.

The basic P configurations have been discussed in phase 1 of the design guide and are applicable here as well.

For multicast, the recommendation is to use MVPN as discussed in the phase 1 guide. If the VRFs are using anycast RP, then they should be configured in each of the sites to provide local accessibility.

QoS recommendations from phase 1 hold true here as well. The only additional capability that may be required is the ability to shape the outgoing traffic. For example, in Figure 3-1 P1 and P3 are connected via GE ports, but the underlying service may only be a sub-rate GE. Thus outgoing traffic on both sides needs to be shaped to match the sub-rate.

There could be scenarios that may require the use of Inter-AS to connect the two MPLS networks. This could be when the two networks are part of two different administrative domains or are large enough that a flat network is not recommended. Since we have not seen such requirements yet from enterprise networks, this is not discussed here but future versions may be updated to include it if required.

CHAPTER 4

WAN Edge—MPLSoL2 Service

While Layer 3 VPN services are becoming increasing popular as a primary connection for the WAN, there are a much larger percentage of customers still using Layer 2 services such Frame-Relay (FR). A big factor in migrating to Layer 3 services is cost and bandwidth scalability and flexibility. There will be customers who may not want to migrate to Layer 3 services, but rather maintain their Layer 2 infrastructure (such as Financials) or are at least slow in moving towards it. For such customers, extending virtualization to the branches involves converting the branch routers into MPLS edge devices and enabling MPLS on the Layer 2 links. The WAN aggregation device is converted into a P router and connected directly to the MPLS network. Thus the branch routers (PE) are now part of the hub MPLS network.

The existing IGP can be used to distribute the PE and Router Reflector (RR) reachability information. The WAN aggregation router maintains LDP sessions with ever branch router to advertise label information for the PEs. The branch routers establish the MP-ibgp session with the core MPLS network RRs for VPN information. Since they are now part of the MPLS network, services such as MVPN can be extended to them as well.

To extend MPLS to the branches:

- Create loopback interfaces on the branch routers for MP-iBGP peering.
- Enable LDP on the Layer 2 links connecting the branches with the WAN aggregation hub.
- Ensure that the loopback addresses are advertised via IGP and that LDP labels are allocated for them.
- Configure the VRFs and place the user LAN or VLANs into appropriate VRF at each of the branches.
- Make the branch routers clients of the core RR for MP-BGP. This allows VPN information to be exchanged between the branch PEs and core PEs.

For network redundancy, the spoke could be dual homed with two PVCs to two aggregators. MPLS could be enabled on both the links and the spoke can load balance traffic destined to other PEs.

Note If the existing deployment uses IPSec encryption on the routers, then this model may present some challenges. Labeled packets cannot be encrypted by the routers.

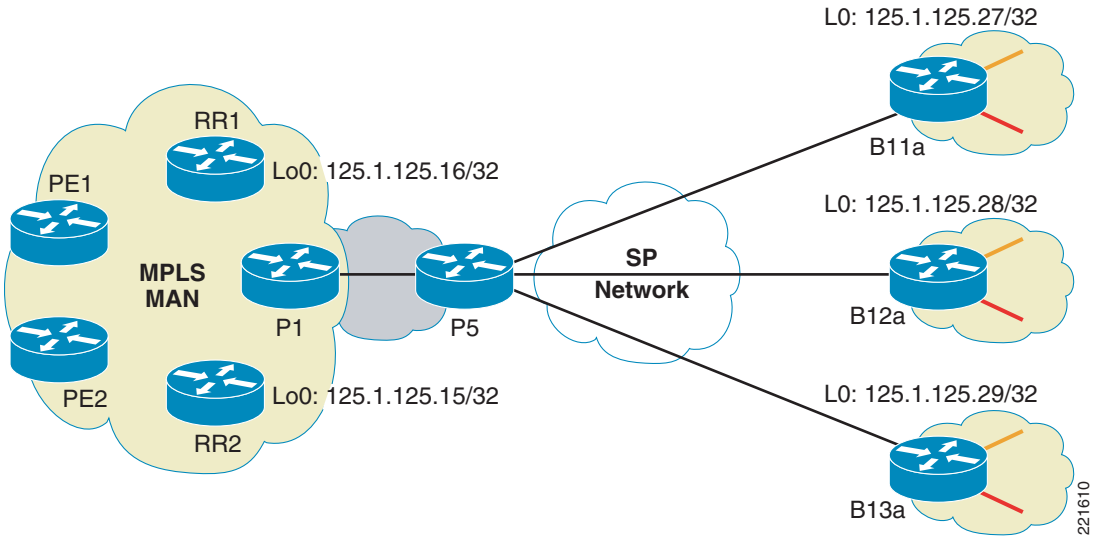
Platforms

The WAN aggregation hub could be any router that supports P functionality and meets the performance requirements, such as 12000, 7600, or 7200s. ISRs are typically recommended as spoke routers. The latest 12.4T images are recommended for the ISRs and 7200s used as branch routers. The image selection for GSR, 7600, and 7200 WAN aggregation routers need to be selected based on the feature, hardware requirement, and compatibilities.

Example:

As shown in Figure 4-1, sites B11, B12, and B13 have existing FR connections to WAN aggregation device (P5). The aggregation device is connected to the core P (P1). The three branch routers have loopbacks created that are advertised in the IGP—B11a (125.1.125.27/32), B12a (125.1.125.28/32), and B13a (125.1.125.29/32). The core has two RRs (125.1.125.15 and 16) to which each of the branch PEs is peering.

Figure 4-1 MPLSoL2 Deployment



P5:

```
mpls label protocol ldp
tag-switching tdp router-id Loopback0 force
!
interface Loopback0
 ip address 125.1.125.11 255.255.255.255
!
interface GigabitEthernet1/3
 description To P1 - intf G2/0/1
  ip address 125.1.100.102 255.255.255.252
  tag-switching ip
  mls qos trust dscp
!
interface Serial2/0/0
 mtu 1500
 no ip address
 encapsulation frame-relay
 dsu bandwidth 44210
 framing c-bit
 cablelength 10
 clock source internal
```

```
!
interface Serial2/0/0.1 point-to-point
 ip address 125.1.201.2 255.255.255.252
 ip pim sparse-mode
 tag-switching ip
 frame-relay interface-dlci 17
!
interface Serial2/0/0.2 point-to-point
 ip address 125.1.202.2 255.255.255.252
 ip pim sparse-mode
 tag-switching ip
 frame-relay interface-dlci 18
!
interface Serial2/0/0.3 point-to-point
 ip address 125.1.203.2 255.255.255.252
 ip pim sparse-mode
 tag-switching ip
 frame-relay interface-dlci 20
!
router ospf 10
 log-adjacency-changes
 network 125.1.201.0 0.0.0.3 area 0
 network 125.1.202.0 0.0.0.3 area 0
 network 125.1.203.0 0.0.0.3 area 0
 network 125.0.0.0 0.255.255.255 area 0
 maximum-paths 8
```

Note While the IGP configuration here shows all the spokes in Area 0 for simplicity, in practice any existing hierarchical design can be maintained as long as PE loopback addresses are never summarized and a label is allocated for each of the /32 addresses.

```
ip cef
mpls label protocol ldp
!
ip vrf red-data
 rd 10:1033
 route-target export 10:103
 route-target import 10:103
!
ip vrf red-voice
 rd 10:1043
 route-target export 10:104
 route-target import 10:104
!
interface Loopback0
 ip address 125.1.125.27 255.255.255.255
!
interface GigabitEthernet0/1.1
 encapsulation dot1Q 261
 ip vrf forwarding red-data
 ip address 125.1.20.1 255.255.255.0
!
interface GigabitEthernet0/1.2
 encapsulation dot1Q 262
 ip vrf forwarding red-voice
 ip address 125.1.20.1 255.255.255.0
!
interface Serial1/0/0
 no ip address
 encapsulation frame-relay
 load-interval 30
```

```
 clock rate 2000000
!
interface Serial1/0/0.1 point-to-point
 ip address 125.1.201.1 255.255.255.252
 mpls ip
 frame-relay interface-dlci 16
!
router ospf 10
 log-adjacency-changes
 network 125.1.125.27 0.0.0.0 area 0
 network 125.1.201.0 0.0.0.3 area 0
!
router bgp 1
 no bgp default ipv4-unicast
 bgp log-neighbor-changes
 neighbor 125.1.125.15 remote-as 1
 neighbor 125.1.125.15 update-source Loopback0
 neighbor 125.1.125.16 remote-as 1
 neighbor 125.1.125.16 update-source Loopback0
 !
 address-family vpnv4
 neighbor 125.1.125.15 activate
 neighbor 125.1.125.15 send-community extended
 neighbor 125.1.125.16 activate
 neighbor 125.1.125.16 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf red-voice
 redistribute connected
 no synchronization
 exit-address-family
 !
 address-family ipv4 vrf red-data
 redistribute connected
 no synchronization
 exit-address-family
```

Multicast

As in a MPLS/Layer 3VPN network, mVPN is the technique to bring the multicast traffic of individual customer (or segmented user groups) across the core network MVRFs are configured for every VRF where multicast traffic is expected on the branch PEs. Default and Data MDTs are also configured if used within the core MPLS network.

Since the Layer 2 service is typically hub and spoke or partially meshed for larger branches, it is recommended to keep the multicast sources at or behind the hub as much as possible. In either case, the WAN aggregator at the hub would end up doing most of the multicast replication as it would be the last hop P for all the branch PEs that are receivers. Thus the multicast replication performance of the aggregator becomes critical to solution scaleability.

Use of Data MDTs with a very low threshold is highly recommended. This would limit the replication at the aggregator to only those branches that have sent a explicit join to the Data MDT. Keeping the threshold low ensures that the Data MDT is spawned for the more specific (S,G) as soon as possible to reduce the overhead at the aggregator.

Most deployments use either PIM SSM or PIM SM with RP in the MPLS core. PIM-SM tries to constrain data distribution so that a minimal number of routers in the network receive it. Packets are sent only if they are explicitly requested at the RP. By default, members of a multicast group receive data from senders to the group across a single data distribution tree rooted at the RP. PIM-SSM is similar to

PIM-SM with the additional ability to report interest in receiving packets from specific source addresses to an IP multicast address. It does not use RPs but uses source-based forwarding trees only. While PIM SSM provides the simplest implementation, it does create additional memory overhead since the number of mroutes now increases. This is not expected to be an issue in most enterprise deployments.

Example:

Continuing with our earlier example, we add MVPN to sites B11, B12, and B13. We have Data MDT setup with very low threshold (1kbps) to ensure that it gets initiated almost instantly for any stream. We will use PIM SSM for the Data MDTs.

B11a:

```
ip vrf red-data
 rd 10:1033
 route-target export 10:103
 route-target import 10:103
 mdt default 239.232.10.3
 mdt data 239.232.20.32 0.0.0.15 threshold 1
!
ip multicast-routing
ip multicast-routing vrf red-data
!
interface Loopback0
 ip address 125.1.125.27 255.255.255.255
 ip pim sparse-mode
!
interface GigabitEthernet0/1.1
 encapsulation dot1Q 261
 ip vrf forwarding red-data
 ip address 125.1.20.1 255.255.255.0
 ip pim sparse-mode
!
interface Serial1/0/0.1 point-to-point
 ip address 125.1.201.1 255.255.255.252
 ip pim sparse-mode
 mpls ip
!
ip pim ssm range 1
ip pim vrf red-data rp-address 3.3.3.11
!
access-list 1 permit 239.232.0.0 0.0.255.255
```

Note When the MPLS PE function is extended to the branch routers and the enterprise large campus is connected via an Inter-AS solution, multicast would be a problem if Inter-AS mVPN is not supported on the branch platforms. Currently the ISR platform does not support the Inter-AS mVPN feature, while c7200 does, including both NPE-G1 and NPE-G2. Consult the Cisco IOS feature navigator for up-to-date information.

QoS

Existing WAN Edge QoS models can still be implemented with MPLS WAN setup. At the headend the expectation is that a interface with high link speed is used (DS3 and up). At these speeds, link-efficiency policies such as LFI and cRTP are not required. The Enterprise QoS SRND recommends 5-11 classes at the WAN edge. At the branches, the PE could be configured to map the COS to DSCP, but in our example we assume that the packets are already marked with the appropriate DSCP. If the branches have slow/medium speed links (<T1), then a 3-5 class model is recommended.

The traffic classification has to be done based on EXP (3 bits). This restricts us to at the most 8 classes at the edge. The original IP packet DSCP are preserved and only the 3 bits of IP precedence are copied on to the outgoing EXP at the branches. There are three different variations of core QOS behaviors within a MPLS network:

- Uniform Mode—This is typically deployed when the core MPLS network and the VPNs are part of the same DiffServ domain as would be the case in enterprises. If policers or any other mechanisms re-mark the MPLS EXP values within the MPLS core, these marking changes are propagated to lower-level labels and eventually are propagated to the IP ToS field (MPLS EXP bits are mapped to IP Precedence values on the egress PE).
- Short Pipe Mode—This is typically deployed if the core MPLS network and the VPNs are part of different DiffServ domains. In the case of any re-marking occurrence within the core MPLS network, changes are limited to MPLS EXP re-marking only and are not propagated down to the underlying IP packet's ToS byte.
- Pipe Mode—The main difference between Short Pipe Mode and Pipe Mode is that the PE egress policies (toward the CEs) are provisioned according to the core network's explicit markings and re-markings, not the IP DiffServ markings used within the VPN (although these are preserved). As with Short Pipe Mode, any changes to label markings that occur within the core MPLS cloud do not get propagated to the IP ToS byte when the packet leaves the MPLS network.

Example:

We use a modified version of the 8 class model from the Enterprise QoS SRND (scavenger combined with bulk data) with dual LLQ for voice and video. Recall that MVPN encapsulates the multicast packets into GRE and forwards it as IP packets and not MPLS. So we must ensure that the MVPN packets are accounted for in a specific class, otherwise they would be dropped into the default class. We use the video class for interactive video, streaming video, and any other multicast traffic. A hierarchical policy is applied to shape all the traffic leaving the branch PE.

The sample below shows branch PE configuration for QoS. Similar configuration can be applied at the aggregator adjusted for the link speed.

B11a:

```
class-map match-any Bulk-Data
 match mpls experimental topmost 1
class-map match-any Video
 match mpls experimental topmost 4
 match ip precedence 4
class-map match-any Network-Control
 match mpls experimental topmost 6
 match mpls experimental topmost 7
class-map match-any Critical-Data
 match mpls experimental topmost 2
class-map match-any Call-Signaling
 match mpls experimental topmost 3
 match ip dscp af31
class-map match-any Voice
 match mpls experimental topmost 5
!
policy-map WAN-EDGE
 class Voice
  priority percent 18
 class Video
  priority percent 15
 class Call-Signaling
  bandwidth percent 5
 class Network-Control
  bandwidth percent 5
 class Critical-Data
```

```
  bandwidth percent 27
  random-detect dscp-based
 class Bulk-Data
  bandwidth percent 5
  random-detect dscp-based
 class class-default
  bandwidth percent 25
  random-detect
policy-map MQC-FRTS-1536
 class class-default
  shape average 1460000 14600 0
  service-policy WAN-EDGE
!
interface Serial1/0/0.1 point-to-point
 ip address 125.1.201.1 255.255.255.252
 mpls ip
 frame-relay interface-dlci 16
  class FR-MAP-CLASS-1536
!
map-class frame-relay FR-MAP-CLASS-1536
 service-policy output MQC-FRTS-1536
```

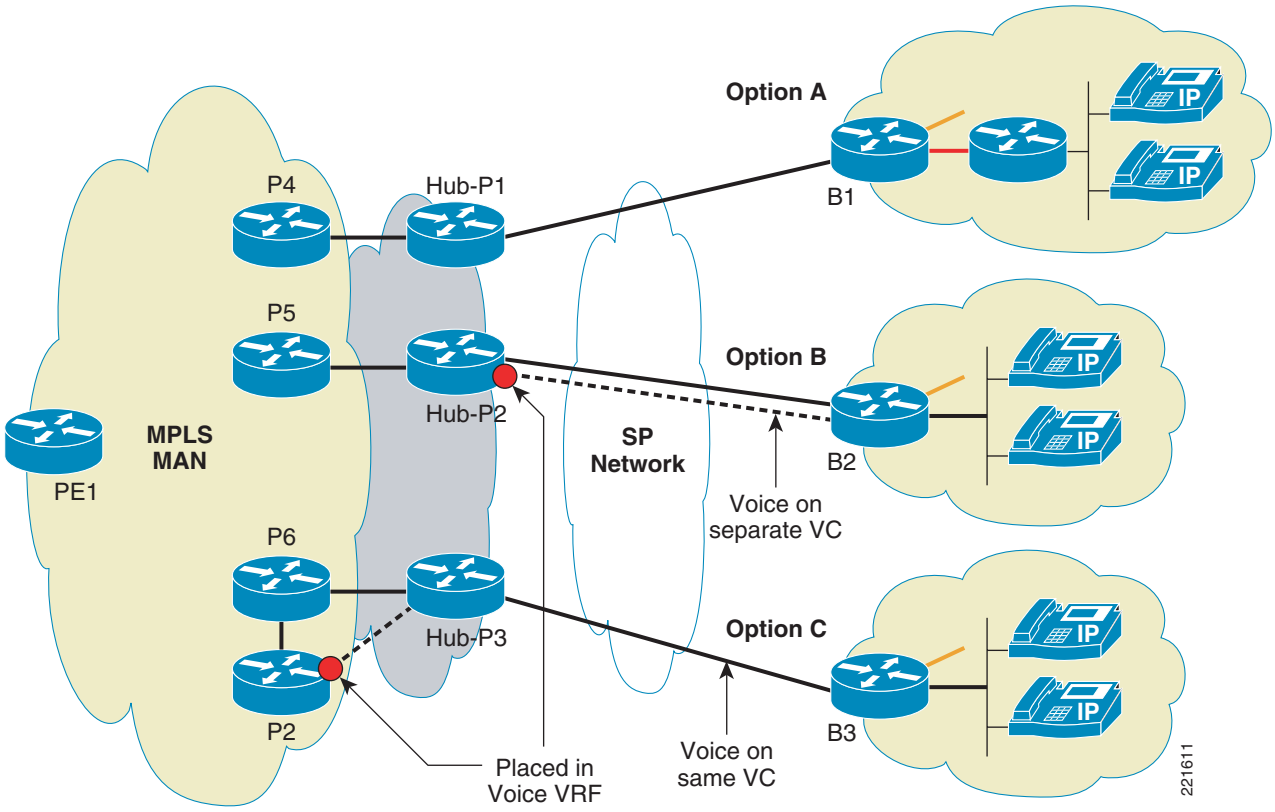
Voice and VRFs

Typically voice traffic has no dependency on the network type since they are just transported as IP packets and require correct QoS behavior applied to them. An exception is when routers are used as gateways for voice services because a lot of voice features and protocols deployed at the branches are not VRF aware (for example, SRST, CME, etc.). Thus just getting the voice traffic in a VRF could be a challenge. This is apart from larger issues of having the voice in a VRF—while you can have the IP phones within a VRF, other services such as softphones VT advantage may be in a different VRF. There are challenges in implementing Inter-VRF IP communications, but they are not discussed here as it is part of the larger virtualization architecture issue. The current recommendation is to keep voice within the global space especially at the branches. At the hub they could remain in the global space or would have to be placed within its own VRF. We look at both options, getting the voice in the VRF at the branch as well keeping it in the global table at the branch.

Voice in a VRF at the Branch

If we need to put the voice in the VRF and still want to use voice features such as CME, then the only way to currently do this is by having two separate routers at the branch. The branch edge router still has a voice VRF configured but treats it like any other VRF. It has a second router (such as a low end ISR) connected to its voice VRF VLAN. The CCME is implemented in the second router, as shown in Figure 4-2 (option A), has all the phones attached to it. Cost might be an issue with this approach as it requires two routers at every such branch site.

Figure 4-2 MPLSoL2—Voice and VRFs



Voice Global at the Branch

If we choose to keep the voice in the global space at the branch, then a single router would be sufficient. The voice VLAN is connected to the branch router but remains in the global space. If the voice is gong to be kept in the global space within the hub network as well, then it can be transported over the existing connection to the hub (MPLS-switched traffic and IP-forwarded traffic share the same link). But at the hub if this traffic needs to be placed within its own VRF, then we would need a separate logical link between the hub and the spoke. This link would be in the global space at the spoke, but be placed within the voice VRF at the hub as shown in Figure 4-2 (option B). The reason we need a separate logical link is that the MPLS link at the hub cannot be placed in a VRF since its configured with “mpls ip” for tag switching. This can potentially increase the circuit cost for the Layer 2 service.

A third option as shown in Figure 4-2 (option C) is to have a separate link at the headend to a PE device which puts the traffic into a VRF. We would need proper routing mechanisms at the hub including route filtering to control the route advertisement within the core network as well as voice VRF.

System Scale and Performance Considerations

Some of the considerations that need to be accounted for from a system scale and performance perspective:

- The WAN aggregator now has IGP and LDP sessions to all the branch routers; the number of peers that it can support can affect the system

- Typically there are large number of branches (into the thousands) and with each one peering directly to the core RR, a more distributed RR design may need to be adopted depending on number of peers supported on a platform.
- In case of MVPN the headend is expected to replicate multicast packets for every spoke receiver and this performance bottleneck can affect the scale (number of branches terminated on each WAN aggregator).
- Converting all the branch routers to PEs increases the footprint of the MPLS network exponentially, from a few 10s within the core network to potentially thousands. This can present a management challenge if the right tools are not used.

Note Inter-AS is another option mentioned in the architecture chapter that will be addressed in the future phases of the solution.

CHAPTER 5

WAN Edge—DMVPN Per VRF

DMVPN is used widely by enterprises to securely extend their private networks across public networks such as Internet. In a number of scenarios it provides backup to primary Layer 2 WAN connection. One of the ways that the existing DMVPN setup can be leveraged and expanded is by using it to extend virtualization to the branches. All the DMVPN functionality remains intact including bulk encryption and dynamic tunnel building.

Instead of the tunnel residing in the global space, it resides within the VRF. Thus for every VRF, you have to create a separate DMVPN cloud. DMVPN per VRF can create challenges, especially in terms of scale, management, and troubleshooting. So the overall recommendation is to implement this model only if the expectation is that total number of VRFs will remain three or less.

Since this is not new functionality, we focus on the implementation aspects of the solution, such as basic configuration, Multicast, QoS, and redundancy.

Platforms

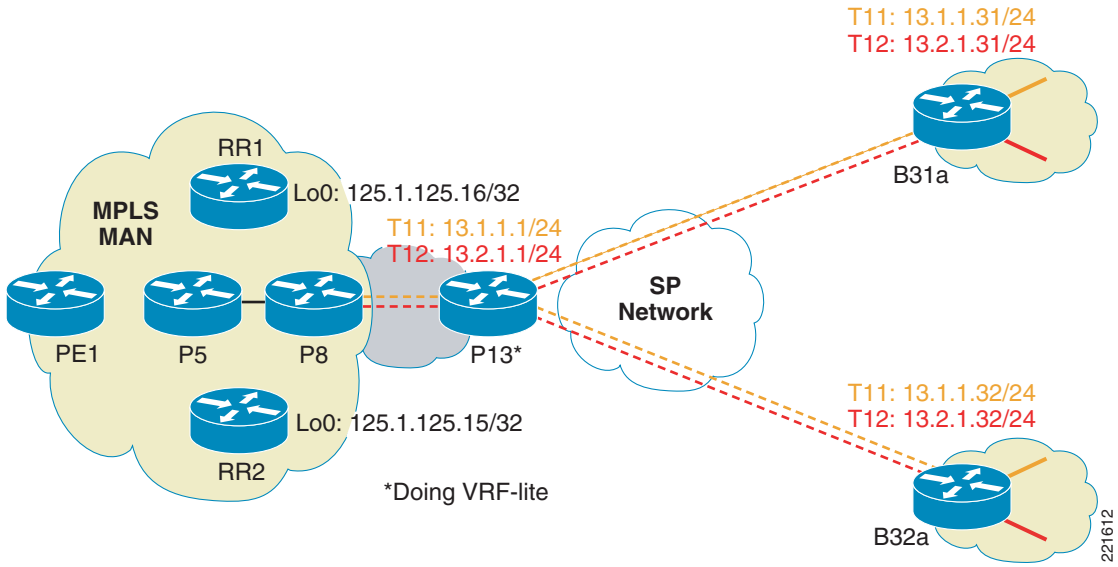
The hub can be a 7200VXR (NPE-G1/G2) with encryption modules (VAM2/VAM2+/VSA) or a 7600 (Sup720-3BXL recommended) with encryption modules (VPNSM/VPN SPA). ISRs with hardware encryption accelerators (AIM II) are recommended as spoke routers. The lab tests were done with the following images:

- 7200VXR with NPE-G1/G2—12.4(11)T1
- 7600 with Sup720-3BXL—12.2(18)SXF
- ISRs (3825/2851)—12.4(11)T1

Example:

We discuss the basic implementation with an example. As shown in Figure 5-1, we have two branches 31 (B31a) and 32 (B32a) connecting to hub PE13. The branch routers and the hub are running VRF-lite in this example. The hub is connected to a PE in the MPLS network. It can be a PE too, an example of which we will see later. The hub is not doing encryption in this example.

Figure 5-1 DMVPN per VRF Deployment



Hub PE13:

```
ip vrf red-data
 rd 10:103
!
ip vrf red-voice
 rd 10:104
!
interface Tunnel11
 ip vrf forwarding red-data
 ip address 13.1.1.1 255.255.255.0
 ip nhrp authentication spe
 ip nhrp map multicast dynamic
 ip nhrp network-id 11
 ip ospf network broadcast
 ip ospf priority 100
 tunnel source GigabitEthernet0/2
 tunnel mode gre multipoint
 tunnel key 1111
!
interface Tunnel12
 ip vrf forwarding red-voice
 ip address 13.2.1.1 255.255.255.0
 ip nhrp authentication spe
 ip nhrp map multicast dynamic
 ip nhrp network-id 12
 ip ospf network broadcast
 ip ospf priority 100
 tunnel source GigabitEthernet0/2
 tunnel mode gre multipoint
 tunnel key 2222
!
interface GigabitEthernet0/2
 ip address 135.0.16.2 255.255.255.252
!
interface GigabitEthernet0/3.1
 encapsulation dot1Q 301
 ip vrf forwarding red-data
 ip address 125.1.108.2 255.255.255.252
!
```

```
interface GigabitEthernet0/3.2
 encapsulation dot1Q 302
 ip vrf forwarding red-voice
 ip address 125.1.108.2 255.255.255.252
!
router ospf 1 vrf red-data
 log-adjacency-changes
 capability vrf-lite
 network 13.1.1.0 0.0.0.255 area 0
 network 125.1.108.0 0.0.0.3 area 0
!
router ospf 2 vrf red-voice
 router-id 125.1.125.31
 log-adjacency-changes
 capability vrf-lite
 network 13.2.1.0 0.0.0.255 area 0
 network 125.1.108.0 0.0.0.3 area 0
!
router bgp 1
 bgp log-neighbor-changes
 neighbor 135.0.16.1 remote-as 2
 !
 address-family ipv4
 neighbor 135.0.16.1 activate
 neighbor 135.0.16.1 allowas-in
 no auto-summary
 no synchronization
 exit-address-family
```

Spoke B31a:

```
ip vrf red-data
 rd 10:103
!
ip vrf red-voice
 rd 10:104
!
interface Tunnel11
 ip vrf forwarding red-data
 ip address 13.1.1.31 255.255.255.0
ip nhrp authentication spe
 ip nhrp network-id 11
 ip nhrp nhs 13.1.1.1
 ip ospf network broadcast
 ip ospf priority 0
 tunnel source FastEthernet1/1
 tunnel destination 135.0.16.2
 tunnel key 1111
!
interface Tunnel12
 ip vrf forwarding red-voice
 ip address 13.2.1.31 255.255.255.0
 ip nhrp authentication spe
 ip nhrp network-id 12
 ip nhrp nhs 13.2.1.1
 ip ospf network broadcast
 ip ospf priority 0
 tunnel source FastEthernet1/1
 tunnel destination 135.0.16.2
 tunnel key 2222
!
interface GigabitEthernet0/1.1
 encapsulation dot1Q 241
```

```
 ip vrf forwarding red-data
 ip address 125.1.18.1 255.255.255.0
!
interface GigabitEthernet0/1.2
 encapsulation dot1Q 242
 ip vrf forwarding red-voice
 ip address 125.1.18.1 255.255.255.0
!
interface FastEthernet1/1
 ip address 135.0.6.2 255.255.255.252
!
router ospf 1 vrf red-data
 log-adjacency-changes
 capability vrf-lite
 passive-interface GigabitEthernet0/1.1
 network 13.1.1.0 0.0.0.255 area 0
 network 125.1.18.0 0.0.0.255 area 0
!
router ospf 2 vrf red-voice
 log-adjacency-changes
 capability vrf-lite
 passive-interface GigabitEthernet0/1.2
 network 13.2.1.0 0.0.0.255 area 0
 network 125.1.18.0 0.0.0.255 area 0
!
router bgp 1
 bgp log-neighbor-changes
 neighbor 135.0.6.1 remote-as 2
 !
 address-family ipv4
 neighbor 135.0.6.1 activate
 neighbor 135.0.6.1 allowas-in
 no auto-summary
 no synchronization
 exit-address-family
```

Configuration Notes:

- Every multipoint tunnel corresponds to a VRF. Each tunnel is placed in its own VRF.
- We are running OSPF within each VRF configured with “capability vrf-lite”.
- On the hub, the tunnel interfaces and the VLAN to the core MPLS PE are part of the corresponding OSPF process. On the spokes, the tunnel interface and optionally the LAN-facing VLAN (if there are other OSPF speakers on the LAN) are part of the OSPF process.

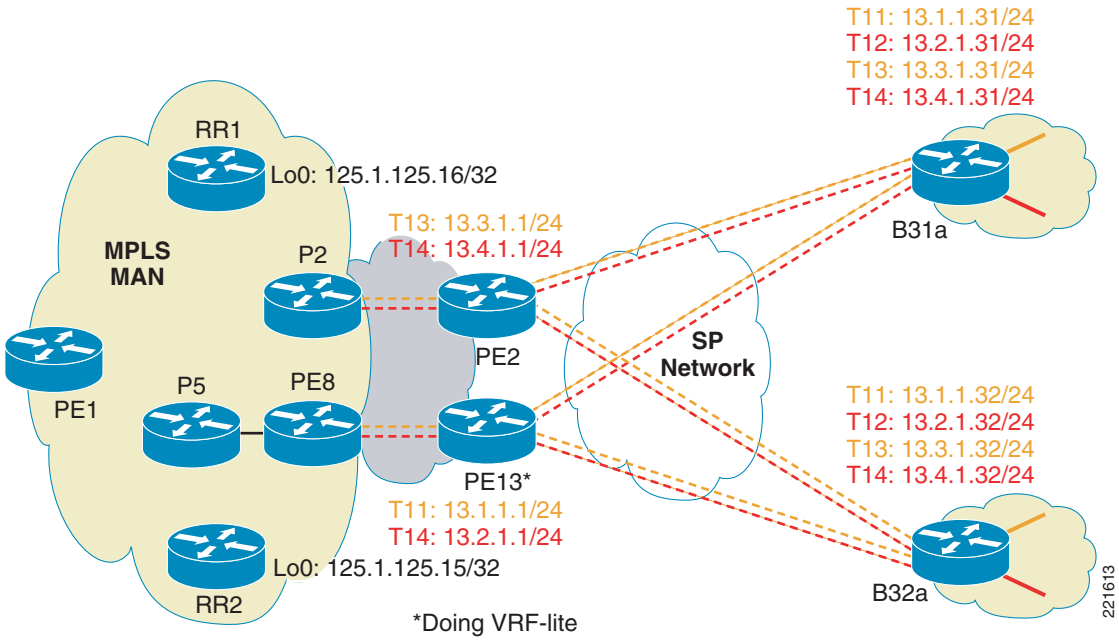
Building Redundancy

As in a normal DMVPN network, it is recommended to have multiple hubs. From the spoke perspective, it keeps connection to both the hubs but prefers one over the other. This can be done by changing the tunnel metric depending on the IGP—change delay for EIGRP and interface cost for OSPF. The return traffic from the headend network just picks the best path. The advantage of keeping such a arrangement is that it allows the hubs to be engineered to maintain a certain number of tunnels and level of traffic. With fast convergence mechanisms configured a tunnel failure would quickly switch the traffic to the backup tunnel.

Example:

As shown in Figure 5-2, we introduce a second hub to our earlier setup. PE2 supports the same VRFs but is a also a PE in the core MPLS network (connected to a P). We show the configuration from a 7600 hub PE with encryption enabled.

Figure 5-2 DMVPN per VRF Redundancy



Hub PE2:

```
ip vrf red-data
 rd 10:1032
 route-target export 10:103
 route-target import 10:103
!
ip vrf red-voice
 rd 10:1042
 route-target export 10:104
 route-target import 10:104
!
mls mpls tunnel-recir
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco12345 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
crypto ipsec transform-set T1 esp-3des esp-sha-hmac
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
crypto engine mode vrf
!
interface Loopback0
 ip address 125.1.125.6 255.255.255.255
!
interface Loopback1
```

```
 ip address 135.10.1.1 255.255.255.255
 crypto engine slot 6
!
interface Loopback2
 ip address 135.10.1.2 255.255.255.255
 crypto engine slot 6
!
interface Tunnel13
 ip vrf forwarding red-data
 ip address 13.3.1.1 255.255.255.0
 no ip redirects
 ip nhrp authentication spe
 ip nhrp map multicast dynamic
 ip nhrp network-id 13
 ip ospf network broadcast
 ip ospf priority 100
 tunnel source Loopback1
 tunnel mode gre multipoint
 tunnel protection ipsec profile P1
 crypto engine slot 6
!
interface Tunnel14
 ip vrf forwarding red-voice
 ip address 13.4.1.1 255.255.255.0
 no ip redirects
 ip nhrp authentication spe
 ip nhrp map multicast dynamic
 ip nhrp network-id 14
 ip ospf network broadcast
 ip ospf priority 100
 tunnel source Loopback2
 tunnel mode gre multipoint
 tunnel protection ipsec profile P1
 crypto engine slot 6
!
interface GigabitEthernet2/9
 description To P2
 ip address 125.1.100.78 255.255.255.252
 tag-switching ip
 mls qos trust dscp
!
interface GigabitEthernet2/10
 description To SP
 ip address 135.0.8.2 255.255.255.252
 mls qos trust dscp
 crypto engine slot 6
!
interface GigabitEthernet6/1
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 1,1002-1005
 switchport mode trunk
 mtu 4500
 no ip address
 load-interval 30
 flowcontrol receive on
 flowcontrol send off
 spanning-tree portfast trunk
!
interface GigabitEthernet6/2
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 1,1002-1005
 switchport mode trunk
```

```
 mtu 4500
 no ip address
 load-interval 30
 flowcontrol receive on
 flowcontrol send off
 spanning-tree portfast trunk
!
router ospf 2 vrf red-voice
 log-adjacency-changes
 redistribute bgp 1 subnets
 network 13.4.1.0 0.0.0.255 area 0
 network 125.1.101.12 0.0.0.3 area 0
!
router ospf 1 vrf red-data
 log-adjacency-changes
 redistribute connected subnets
 redistribute bgp 1 subnets
 network 13.3.1.0 0.0.0.255 area 0
 network 125.1.101.8 0.0.0.3 area 0
!
router ospf 10
 log-adjacency-changes
 network 125.0.0.0 0.255.255.255 area 0
 maximum-paths 8
!
router bgp 1
 no synchronization
 bgp log-neighbor-changes
 network 135.10.1.1 mask 255.255.255.255
 network 135.10.1.2 mask 255.255.255.255
 neighbor 125.1.125.15 remote-as 1
 neighbor 125.1.125.15 update-source Loopback0
 neighbor 125.1.125.16 remote-as 1
 neighbor 125.1.125.16 update-source Loopback0
 neighbor 135.0.8.1 remote-as 2
 no auto-summary
 !
 address-family vpnv4
 neighbor 125.1.125.15 activate
 neighbor 125.1.125.15 send-community extended
 neighbor 125.1.125.16 activate
 neighbor 125.1.125.16 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf red-voice
 redistribute ospf 2 vrf red-voice match internal external 1 external 2
 maximum-paths ibgp unequal-cost 8
 no auto-summary
 no synchronization
 exit-address-family
 !
 address-family ipv4 vrf red-data
 redistribute ospf 1 vrf red-data match internal external 1 external 2
 maximum-paths ibgp unequal-cost 8
 no auto-summary
 no synchronization
 exit-address-family
```

DMVPN per VRF Configuration Notes:

- On a 7600, when using tunnel keys, the packets get process switched. Thus use of tunnel keys should be disabled on the 7600 as well as the connecting branch routers.

- On a 7600, GRE tunnels in different VRFs cannot share the same tunnel source. Hence in our example we are using two different loopback interfaces as sources for the VRF red-data (loopback 1) and red-voice (loopback 2) which are advertised to the SP via BGP.
- On a 7600, the "crypto engine <slot>" commands needs to be configured on the tunnel source interface (loopbacks in this case) and on the tunnels themselves.
- On a 7600, "mls mpls tunnel-recir" needs to be configured when implementing VRF-aware DMVPN.
- On a 7200, "tunnel protection … shared" must be configured on the tunnel interface if the tunnel is sourced from a single address and tunnel key is used to distinguish different tunnels.
- Since this hub is a full PE, it is configured with route import/export configuration under the VRF. It is also peered with couple of Route Reflectors (125.1.125.15 and 125.1.125.16) in the core MPLS network.
- OSPF is used over the tunnel interface. So a process is created for each VRF and the hub peers with every branch router to which it is connected. The routes are mutually distributed with MP-BGP to establish the connectivity between the core MPLS network and the virtualized branches.
- The spoke configuration is similar to the ones shown in the earlier section, with additional tunnels configured to connect to the 7600. On spokes, for two hub connections, we need two tunnels configured per VRF. Thus, the number of tunnels start multiplying as the number of VRFs increases (four tunnels for two VRFs and two hubs in our example).

Implementing Multicast

Multicast is implemented as a combination of MVPN and non-VRF configuration. The RPs would still reside within the core MPLS network, but need to be reachable by the branches from within the VRF. The branches have multicast capabilities enabled for each VRF. The hub if it is configured as VRF-lite then would only require the appropriate PIM mode to be enabled on the interfaces, RP reachability configured, and multicast enabled for each VRF, similar to the spokes. If the hub is a PE for the core MPLS network, then it will have a full MVPN configuration which includes default and data MDTs for the core network.

Example:

Continuing with our dual hub example, hub PE13 is configured for VRF-lite and hub PE2 is configured for MVPN since its a full fledged PE. Only the additions to earlier configuration are shown here.

Hub PE13:

```
ip multicast-routing
ip multicast-routing vrf red-data
ip multicast-routing vrf red-voice
!
interface Tunnel11
 ip pim nbma-mode
 ip pim sparse-mode
!
interface Tunnel12
 ip pim nbma-mode
 ip pim sparse-mode
!
ip pim vrf red-data rp-address 3.3.3.11
ip pim vrf red-voice rp-address 4.4.4.11
```

Hub PE2:

```
ip vrf red-data
 mdt default 239.232.10.3
 mdt data 239.232.20.32 0.0.0.15 threshold 1
!
ip vrf red-voice
 mdt default 239.232.10.4
 mdt data 239.232.20.48 0.0.0.15 threshold 1
!
ip multicast-routing
ip multicast-routing vrf red-data
ip multicast-routing vrf red-voice
!
interface Tunnel13
 ip pim nbma-mode
 ip pim sparse-mode
!
interface Tunnel14
 ip pim nbma-mode
 ip pim sparse-mode
!
ip pim ssm range 1
ip pim vrf red-data rp-address 3.3.3.11
ip pim vrf red-voice rp-address 4.4.4.11
!
access-list 1 permit 239.232.20.0 0.0.0.255
```

The spoke configurations are similar to the VRF-lite case (hub PE13).

Configuration Notes:

- PIM sparse mode is configured on all the interfaces including the core facing (not shown).
- PIM NBMA mode is configured on the multipoint GRE tunnels. This creates the spoke specific entries in the Multicast Output Interface List (OIL).
- In the MPLS network, PIM SSM is used for the data MDTs in the core.

Note Multicast is not supported in DMVPN per VRF on Cat6500 and c7600 routers. It is not recommended to use either of these platforms as a DMVPN mGRE hub if multicast must be implemented.

Implementing QoS

QoS configurations and recommendations do not need to change with DMVPN per VRF. Policies used for existing DMVPN setup are applicable as well. One exception is the DMVPN Hub Support by QoS (http://www.cisco.com/en/US/products/ps6558/prod_bulletin0900aecd803f93d6.html) is not supported).

Example:

The following example shows the configuration on the 7200 hub (PE13) with a sub-rate GE connection to the provider. Hence we will apply a Hierarchical QoS policy to the outgoing interface. It shows a 8-class model with dual LLQ for voice and interactive video traffic.

```
class-map match-all Bulk-Data
 match ip dscp af11  af12
class-map match-any Network-Control
 match ip dscp cs6
 match ip dscp cs2
class-map match-all Critical-Data
 match ip dscp af21  af22
```

```
class-map match-any Call-Signaling
 match ip dscp cs3
 match ip dscp af31
class-map match-any Video
 match ip dscp af41
 match ip dscp af42
class-map match-all Voice
 match ip dscp ef
class-map match-all Scavenger
 match ip dscp cs1
!
policy-map WAN-EDGE-child
 class Voice
  priority percent 18
 class Call-Signaling
  bandwidth percent 5
 class Network-Control
  bandwidth percent 5
 class Critical-Data
  bandwidth percent 27
  random-detect dscp-based
 class Bulk-Data
  bandwidth percent 4
  random-detect dscp-based
 class Scavenger
  bandwidth percent 1
 class Video
  priority percent 15
 class class-default
  bandwidth percent 25
  random-detect
!
policy-map WAN-EDGE-parent
 class class-default
  shape peak 500000000
  service-policy WAN-EDGE-child
!
interface GigabitEthernet0/2
 ip address 135.0.16.2 255.255.255.252
 service-policy output WAN-EDGE-parent
```

Scale Considerations

Traditional non-VRF DMVPN deployment scales have been limited by the number of IGP peers that can be supported per DMVPN cloud by the headend. For example, tests conducted by ESE/NSITE suggest that a single DMVPN domain can typically support 350-500 EIGRP peers on a 7200 with NPE-G1/VAM2. A platform itself can support two such domains before CPU becomes the limiting factor. Better performance can be expected with NPE-G2. One of the workarounds for that includes splitting the cloud into multiple headends. DMVPN per VRF has the same scale limitations, but now they are multiplied by the number of VRFs.

One approach is to use a similar concept as non-VRF DMVPN deployments and split the DMVPN clouds to multiple headends as the headend reaches the peer scale limits.

Another approach may be to terminate the different VRFs at different hubs. Depending on the requirements, this could mean a subset of VRFs are terminated on a particular hub. While this does not help with the peer scale limits, it can be useful in cases where DMVPN per VRF is used for large number of VRFs but small number of sites per VRF.

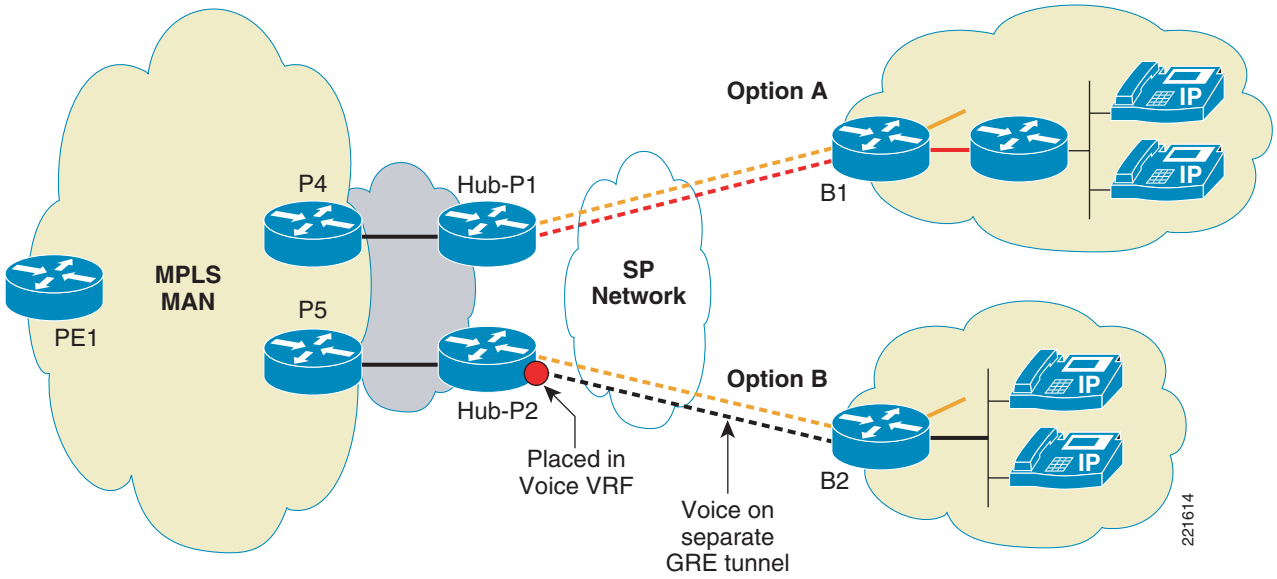
Voice and VRFs

Typically voice traffic has no dependency on the network type since they are just transported as IP packets and require correct QoS behavior applied to them. An exception is when routers are used as gateways for voice services because a lot of voice features and protocols deployed at the branches are not VRF aware (for example, SRST, CME, etc.). Thus just getting the voice traffic in a VRF could be a challenge. This is apart from larger issues of having the voice in a VRF; while you can have the IP phones within a VRF, other services such as softphones VT advantage may be in a different VRF. There are challenges in implementing Inter-VRF IP communications. These are not discussed here as it is part of the larger virtualization architecture issue. The current recommendation is to keep voice within the global space especially at the branches. At the hub they could remain in the global space or would have to be placed within its own VRF. We look at both the options, getting the voice in the VRF at the branch as well keeping it in the global table at the branch.

Voice in a VRF at the Branch

If we need to put the voice in the VRF and still want to use voice features such as CME, then the only way to currently do this is by having two separate routers at the branch. The branch edge router still has a voice VRF configured but treats it like any other VRF. It has a second router (such as a low end ISR) connected to its voice VRF VLAN. The second router, as shown in Figure 5-3 (option A) has all the phones attached to it. Since it requires two routers at every such branch, this can be a expensive proposition

Figure 5-3 DMVPN per VRF—Voice and VRFs



Voice Global at the Branch

If we choose to keep the voice global at the branch then a single router would suffice. The voice VLAN is connected to the branch router but remains in the global space. It is carried across the a global GRE tunnel as normal IPv4 traffic. The DMVPN tunnels per VRF would co-exist with the global tunnel. At

the hub, the DMVPN tunnel carrying voice traffic is placed in the voice VRF from where on it is treated like other VRF traffic (Figure 5-3 option B). Another option at the hub could be to keep the tunnel in the global table for scenarios that keep the voice traffic in the global space even within the hub network.

CHAPTER 6

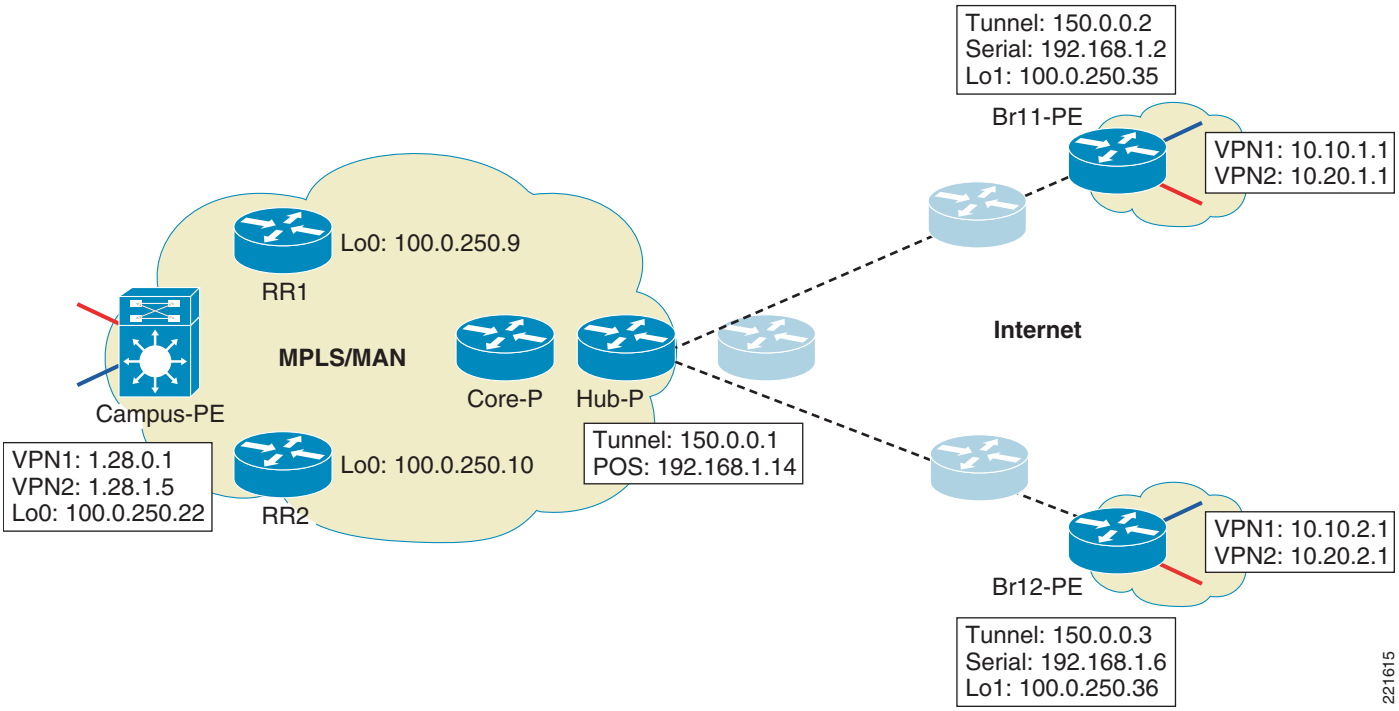
WAN Edge—MPLS VPN over DMVPN—2547oDMVPN (Hub and Spoke Only)

DMVPN provides two key advantages for extending MPLS VPNs to the branches, bulk encryption and, more importantly, a scalable overlay model. Since the assumption here is that the branches in this deployment are connected to the hub through a Layer 3 SP service, a tunneled model using GRE is needed to extend MPLS to the branches. Coupled with the fact that there is large number of existing DMVPN deployments, this solution becomes an attractive deployment option.

DMVPN allows the hub to have a single multipoint GRE tunnel interface to support large numbers of spokes. The spokes can be point-to-point or multipoint GRE tunnels depending on the requirement of direct spoke-to-spoke communication. Spoke-to-Spoke Communication (via Hub) discusses the advantages of point-to-point GRE tunnels at the spokes in the context of the current implementation of MPLS VPN over DMVPN.

To seamlessly extend the enterprise MPLS/Layer 3VPN MAN network to the remote branches, the WAN edge router (also the DMVPN hub in this case) should be a P device to label switching packets between the hub and the branches. As shown in Figure 6-1, the WAN hub router acts as a MPLS/Layer 3VPN P router to establish the LDP neighbor relationship and label switch packet with branch routers which act as a MPL3/Layer 3VPN PE router. The single IGP process is running on the entire enterprise MAN/WAN network to enable the branch routers to establish the MP-iBGP session with RRs in the enterprise MPLS MAN network.

Figure 6-1 2547oDMVPN Deployment



Platforms

Only 7200VXR is supported as the hub router. NPE-G1/G2 is recommended, along with VAM2/VAM2+/VSA modules for encryption. ISRs are recommended as spoke devices. The following images were tested in the lab:

- 7200VXR with NPE-G1/G2—12.4(11)T1
- ISRs—12.4(11)T1

Hub and Spoke Communication

The hub and spoke communication is straightforward as it follows the normal P-PE forwarding mechanism. The example below gives the typical configurations for a DMVPN hub router used as a MPLS P device and a DMVPN spoke router used as a MPLS PE device.

Hub11:

```
!
hostname ngwan-hub11
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco123 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
!
```

```
crypto ipsec transform-set T1 esp-3des
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
interface Loopback1
 ip address 100.0.250.33 255.255.255.255
 ip pim sparse-mode
!
interface Tunnel1
 bandwidth 1500
 ip address 150.0.0.1 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp map multicast dynamic
 ip nhrp network-id 1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
 ip ospf priority 100
 load-interval 30
 mpls ip
 tunnel source 192.168.1.14
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1
!
router ospf 1
 mpls traffic-eng router-id Loopback1
 mpls traffic-eng area 0
 router-id 100.0.250.33
 log-adjacency-changes
 network 100.0.250.33 0.0.0.0 area 0
 network 100.0.0.0 0.0.255.255 area 0
 network 150.0.0.0 0.0.0.255 area 0
!
router ospf 100
 log-adjacency-changes
network 192.168.1.12 0.0.0.3 area 3
!
ip pim ssm range 1
access-list 1 permit 239.232.0.0 0.0.255.255
```

Spoke br11:

```
hostname ngwan-br11
!
ip vrf vpn1
 rd 100:10
 route-target export 100:110
 route-target import 100:110
 mdt default 239.232.1.1
 mdt data 239.232.1.128 0.0.0.127 threshold 10
!
ip vrf vpn2
 rd 100:20
 route-target export 100:120
 route-target import 100:120
 mdt default 239.232.2.1
 mdt data 239.232.2.128 0.0.0.127 threshold 10
!
ip multicast-routing
ip multicast-routing vrf vpn1
```

```
ip multicast-routing vrf vpn2
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco123 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
!
crypto ipsec transform-set T1 esp-3des
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
interface Loopback1
 ip address 100.0.250.35 255.255.255.255
 ip pim sparse-mode
!
interface Tunnel1
 bandwidth 1500
 ip address 150.0.0.2 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp map 150.0.0.1 192.168.1.14
 ip nhrp map multicast 192.168.1.14
 ip nhrp network-id 1
 ip nhrp nhs 150.0.0.1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
 ip ospf priority 0
 load-interval 30
 mpls ip
 qos pre-classify
 tunnel source 192.168.1.2
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1
!
interface POS5/0
 ip address 192.168.1.2 255.255.255.252
 load-interval 30
 crc 32
 clock source internal
 service-policy output wan-edge
!
router ospf 100
 log-adjacency-changes
 network 192.168.1.0 0.0.0.3 area 3
!
router ospf 1
 mpls traffic-eng router-id Loopback1
 mpls traffic-eng area 0
 router-id 100.0.250.35
 log-adjacency-changes
 network 100.0.250.35 0.0.0.0 area 0
 network 150.0.0.0 0.0.0.255 area 0
 network 150.0.1.0 0.0.0.255 area 0
!
router bgp 64512
 bgp log-neighbor-changes
 neighbor RRs peer-group
 neighbor RRs remote-as 64512
```

```
 neighbor RRs update-source Loopback1
 neighbor 100.0.250.9 peer-group RRs
 neighbor 100.0.250.10 peer-group RRs
 !
 address-family ipv4
 neighbor 100.0.250.9 activate
 neighbor 100.0.250.10 activate
 no auto-summary
 no synchronization
 exit-address-family
 !
address-family ipv4 vrf vpn2
 redistribute connected
 no synchronization
 exit-address-family
 !
 address-family ipv4 vrf vpn1
 redistribute connected
 no synchronization
 exit-address-family
 !
ip pim ssm range 1
access-list 1 permit 239.232.0.0 0.0.255.255

ngwan-br11#sh ip bgp vpn vrf vpn1 1.28.0.1
ngwan-br11#sh ip bgp vpnv4 vrf vpn1 1.28.0.1
BGP routing table entry for 100:10:1.28.0.0/30, version 269
Paths: (1 available, best #1, table vpn1)
  Not advertised to any peer
  Local, imported path from 100:180:1.28.0.0/30
    100.0.250.22 (metric 69) from 100.0.250.9 (100.0.250.9)
      Origin incomplete, metric 0, localpref 100, valid, internal, best
      Extended Community: SoO:100:110 RT:100:110
      Originator: 100.0.250.22, Cluster list: 0.0.0.1
      mpls labels in/out nolabel/21
ngwan-br11#sh ip cef vrf vpn1 1.28.0.1
1.28.0.0/30, version 12, epoch 0
0 packets, 0 bytes
  tag information set
    local tag: VPN-route-head
    fast tag rewrite with Tu1, 150.0.0.1, tags imposed: {63 21}
  via 100.0.250.22, 0 dependencies, recursive
    next hop 150.0.0.1, Tunnel1 via 100.0.250.22/32
    valid adjacency
    tag rewrite with Tu1, 150.0.0.1, tags imposed: {63 21}
ngwan-br11#

ngwan-br11#ping vrf vpn1 1.28.0.1

Type escape sequence to abort.
Sending 5, 100-byte ICMP Echos to 1.28.0.1, timeout is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/4 ms
ngwan-br11#
```

Notice that there are two labels assigned to the destination address in one of the VRFs within the enterprise MPLS MAN. This is exactly the expected behavior for using MPLS to extend the network segmentation to remote branches. Further test with ping shows the LSP and successfully established.

Spoke-to-Spoke Communication (via Hub)

In branch aggregation scenarios, while most traffic is typically between the hub and the spokes, there is always a requirement for spoke-to-spoke communication. VOIP traffic is a good example of peer-to-peer traffic. In a normal DMVPN scenario this would have been achieved by dynamically creating direct spoke-to-spoke tunnels. While there are obvious advantages to this approach, there are issues as well:

- Depending on the underlying physical connectivity, in certain cases the spoke-to-spoke path may not necessarily be better then the spoke-hub-spoke path which could be a problem for latency sensitive traffic such as VOIP.
- There is a possibility of receiving out-of-order packets as during the initial tunnel setup time the traffic traverses the hub, but once the spoke-to-spoke tunnel is setup it switches it over.
- Depending on the number of spoke-to-spoke tunnels that need to created/maintained simultaneously, this can put scale pressures on the spoke router especially if it is a low-end CPE.

Additionally, the MPLS network requires packets to be label switched all the way between source PEs and destination PEs. Running MPLS over DMVPN tunnels makes the remote branch router a full function PE router, which means label imposition is done in the branch router and label switching must be performed all the way between spokes. This requirement make the direct spoke-spoke communication impossible due to the lack of a label allocation mechanism on the dynamically created spoke-spoke tunnels. However, label switching between spoke PE routers can easily be done if spoke-hub-spoke switching path is implemented. With this approach, the hub router act as a MPLS P router, maintains the LDP neighbor relationship, and exchanges label allocation information with all spoke routers. The hub router label switches the packets in-and-out the mGRE interface between the spokes. Since it is done in the fast path (whether encrypted or not), there should be minimal performance implications other then the increase in the hub traffic.

While this solution breaks the benefit of dynamically building spoke-to-spoke tunnels, it provides a acceptable and often more deterministic path for spoke-to-spoke communications and meets the segmentation requirement. It is a very attractive solution when the large enterprise needs to extend their MPLS-segmented data center or large campus to remote branches.

Configuration Example:

The following example shows the two VPNs in the two remote branches (br11 and br12) communicate to each other via the DMVPN hub router (hub11). The router hub11's configuration is the same as shown in Hub and Spoke Communication. The VPN naming and address scheme, along with the address of hub router and branch routers, are illustrated in Figure 6-1.

Br11:

```
hostname ngwan-br11
!
ip vrf vpn1
 rd 100:10
 route-target export 100:110
 route-target import 100:110
 mdt default 239.232.1.1
 mdt data 239.232.1.128 0.0.0.127 threshold 10
!
ip vrf vpn2
 rd 100:20
 route-target export 100:120
 route-target import 100:120
 mdt default 239.232.2.1
 mdt data 239.232.2.128 0.0.0.127 threshold 10
!
ip multicast-routing
```

```
ip multicast-routing vrf vpn1
ip multicast-routing vrf vpn2
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco123 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
!
crypto ipsec transform-set T1 esp-3des
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
interface Loopback1
 ip address 100.0.250.35 255.255.255.255
 ip pim sparse-mode
!
interface Tunnel1
 bandwidth 1500
 ip address 150.0.0.2 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp map 150.0.0.1 192.168.1.14
 ip nhrp map multicast 192.168.1.14
 ip nhrp network-id 1
 ip nhrp nhs 150.0.0.1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
 ip ospf priority 0
 load-interval 30
 mpls ip
 qos pre-classify
 tunnel source 192.168.1.2
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1
!
interface POS5/0
 ip address 192.168.1.2 255.255.255.252
 load-interval 30
 crc 32
 clock source internal
 service-policy output wan-edge
!
router ospf 100
 log-adjacency-changes
 network 192.168.1.0 0.0.0.3 area 3
!
router ospf 1
 mpls traffic-eng router-id Loopback1
 mpls traffic-eng area 0
 router-id 100.0.250.35
 log-adjacency-changes
 network 100.0.250.35 0.0.0.0 area 0
 network 150.0.0.0 0.0.0.255 area 0
 network 150.0.1.0 0.0.0.255 area 0
!
router bgp 64512
 bgp log-neighbor-changes
 neighbor RRs peer-group
```

```
 neighbor RRs remote-as 64512
 neighbor RRs update-source Loopback1
 neighbor 100.0.250.9 peer-group RRs
 neighbor 100.0.250.10 peer-group RRs
 !
 address-family ipv4
 neighbor 100.0.250.9 activate
 neighbor 100.0.250.10 activate
 no auto-summary
 no synchronization
 exit-address-family
 !
address-family ipv4 vrf vpn2
 redistribute connected
 no synchronization
 exit-address-family
 !
 address-family ipv4 vrf vpn1
 redistribute connected
 no synchronization
 exit-address-family
 !
ip pim ssm range 1
access-list 1 permit 239.232.0.0 0.0.255.255
```

Br12:

```
!
hostname ngwan-br12


ip vrf vpn1
 rd 100:10
 route-target export 100:110
 route-target import 100:110
 mdt default 239.232.1.1
 mdt data 239.232.1.128 0.0.0.127 threshold 10
!
ip vrf vpn2
 rd 100:20
 route-target export 100:120
 route-target import 100:120
 mdt default 239.232.2.1
 mdt data 239.232.2.128 0.0.0.127 threshold 10
!
mpls label protocol ldp
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco123 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
!
crypto ipsec transform-set T1 esp-3des
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
interface Loopback1
 ip address 100.0.250.36 255.255.255.255
```

```
 ip pim sparse-mode
!
interface Tunnel1
 ip address 150.0.0.3 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp authentication spe
 ip nhrp map 150.0.0.1 192.168.1.14
 ip nhrp map multicast 192.168.1.14
 ip nhrp network-id 1
 ip nhrp holdtime 360
 ip nhrp nhs 150.0.0.1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
 ip ospf priority 0
 load-interval 30
 mpls label protocol ldp
 mpls ip
 qos pre-classify
 tunnel source Serial2/0
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1

interface Serial2/0
 ip address 192.168.1.6 255.255.255.252
 load-interval 30
 dsu bandwidth 44210
 service-policy output wan-edge
!
router ospf 100
 log-adjacency-changes
 network 192.168.1.4 0.0.0.3 area 3
!
router ospf 1
 mpls traffic-eng router-id Loopback1
 mpls traffic-eng area 0
 log-adjacency-changes
 network 100.0.250.36 0.0.0.0 area 0
 network 150.0.0.0 0.0.0.255 area 0
 network 150.0.1.0 0.0.0.255 area 0
!
router bgp 64512
 bgp log-neighbor-changes
 neighbor RRs peer-group
 neighbor RRs remote-as 64512
 neighbor RRs update-source Loopback1
 neighbor 100.0.250.9 peer-group RRs
 neighbor 100.0.250.10 peer-group RRs
 !
 address-family ipv4
 neighbor 100.0.250.9 activate
 neighbor 100.0.250.10 activate
 no auto-summary
 no synchronization
 exit-address-family
 !
 address-family vpnv4
 neighbor RRs send-community extended
 neighbor 100.0.250.9 activate
 neighbor 100.0.250.10 activate
 exit-address-family
```

```
 !
 address-family ipv4 vrf vpn2
 redistribute connected
 no synchronization
 exit-address-family
 !
 address-family ipv4 vrf vpn1
 redistribute connected
 no synchronization
 exit-address-family
!
ip pim ssm range 1!
access-list 1 permit 239.232.0.0 0.0.255.255
!
```

The following command shows the seamless integrating of remote spokes with the Enterprise MPLS network. VPN routes are distributed by RRs in the enterprise MPLS network via MP-iBGP:

```
ngwan-br11#sh ip bgp vpnv4 vrf vpn1 10.10.2.1
BGP routing table entry for 100:10:10.10.2.0/24, version 3267
Paths: (2 available, best #2, table vpn1)
  Not advertised to any peer
  Local
    100.0.250.36 (metric 133) from 100.0.250.10 (100.0.250.10)
      Origin incomplete, metric 0, localpref 100, valid, internal
      Extended Community: RT:100:110
      Originator: 100.0.250.36, Cluster list: 0.0.0.1
      mpls labels in/out nolabel/76
  Local
    100.0.250.36 (metric 133) from 100.0.250.9 (100.0.250.9)
      Origin incomplete, metric 0, localpref 100, valid, internal, best
      Extended Community: RT:100:110
      Originator: 100.0.250.36, Cluster list: 0.0.0.1
      mpls labels in/out nolabel/76
ngwan-br11#
```

The following command shows two labels are allocated for the VPN routes in spoke routers and hub is in the middle of the LSP:

```
ngwan-br11#sh ip cef vrf vpn1 10.10.2.1 detail
10.10.2.0/24, version 423, epoch 0
0 packets, 0 bytes
  tag information set
    local tag: VPN-route-head
    fast tag rewrite with Tu1, 150.0.0.1, tags imposed: {78 76}
  via 100.0.250.36, 0 dependencies, recursive
    next hop 150.0.0.1, Tunnel1 via 100.0.250.36/32
    valid adjacency
    tag rewrite with Tu1, 150.0.0.1, tags imposed: {78 76}
ngwan-br11#
```

The following command shows the VPN traffic between spokes are label switched via hub router:

```
ngwan-br11#traceroute vrf vpn1 10.10.2.1

Type escape sequence to abort.
Tracing the route to 10.10.2.1

  1 150.0.0.1 [MPLS: Labels 78/76 Exp 0] 4 msec 0 msec 0 msec
  2 10.10.2.1 4 msec *  8 msec
ngwan-br11#
```

Connecting to the Core MPLS Network

The core MPLS network and the DMVPN-based MPLS network are fully integrated together when the DMVPN hub router act as a MPLS P router. The normal MPLS/LDP configuration applies here when it connects to the enterprise MPLS core networks.

Building Redundancy

Redundancy can be built at various points within the networks:

- Use of multiple routers and HSRP/GLBP with the Enhanced Object Tracking at the branch
- Multiple hub routers at the headend
- Hub connects to multiple core routers

Ideally all three should be used to provide a robust end-to-end connectivity solution. For cost reasons, it may not be feasible to have two routers at every branch; however it should be implemented at least at the large branches when the high-available requirement is a must. While the loss of a spoke router may not be critical to the network, loss of the hub may mean loss of multiple sites and hence more critical. Thus each spoke should be connected to multiple hubs via GRE tunnels which are maintained in active/standby state by controlling the route metrics.

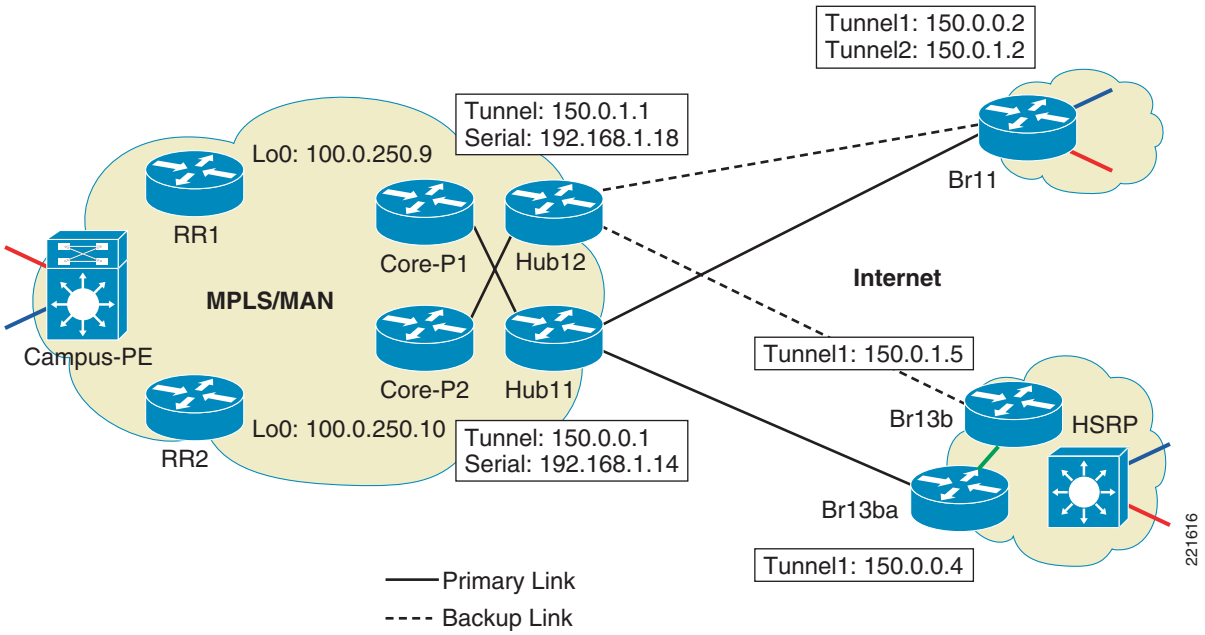
While it may seem desirable to keep both (or all) the hubs as active/active and allow the traffic to be load balanced, we do not see any true advantage in doing so. Keeping the tunnels as active/standby allows the hubs to be better engineered for steady state performance. It also reduces the load on the spoke routers while allowing more deterministic traffic path characteristics.

The various options discussed here will are illustrated using an example.

Example:

In the following example (Figure 6-2), two remote spoke sites br11 and br13 (br13a, br13b) are connected to two hubs (hub11 and hub12). Br11 is considered as a single-tier branch, where it has only one WAN router with two DMVPN tunnels terminated at hub11 and hub12. Hub11 is the primary hub and hub12 is the backup. Br13 is considered as a Dual-tier branch, representing a large branch with two WAN routers with the dual DMVPN tunnel which provides the WAN link redundancy. HSRP with enhanced object tracking is used to provide the network resiliency for the clients on the branch.

Figure 6-2 2547oDMVPN Redundancy



Note The encryption configuration is shown here as an example and standard best practices for IPSec should be followed in an actual deployment.

Below is the configuration example for br11, which use dual tunnels on the same router for the WAN redundancy. OSPF cost is used to tune the routing matrix to select which hub is the primary one.

Spoke B21a:

```
hostname ngwan-br11
!
crypto isakmp policy 1
 encr 3des
 authentication pre-share
crypto isakmp key Cisco123 address 0.0.0.0 0.0.0.0
crypto isakmp keepalive 10 5
!
crypto ipsec transform-set T1 esp-3des
 mode transport
!
crypto ipsec profile P1
 set transform-set T1
!
interface Tunnel1
 bandwidth 1500
 ip address 150.0.0.2 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp authentication ngwan
 ip nhrp map 150.0.0.1 192.168.1.14
 ip nhrp map multicast 192.168.1.14
 ip nhrp network-id 1
 ip nhrp nhs 150.0.0.1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
```

```
 ip ospf priority 0
 load-interval 30
 mpls ip
 qos pre-classify
 tunnel source 192.168.1.2
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1 shared
!!
interface Tunnel2
 ip address 150.0.1.2 255.255.255.0
 no ip redirects
 ip mtu 1368
 ip pim sparse-mode
 ip nhrp authentication ngwan
 ip nhrp map 150.0.0.1 192.168.1.18
 ip nhrp map multicast 192.168.1.18
 ip nhrp network-id 2
 ip nhrp nhs 150.0.1.1
 ip nhrp cache non-authoritative
 ip ospf network point-to-multipoint
 ip ospf cost 100
 ip ospf priority 0
 load-interval 30
 qos pre-classify
 mpls ip
 tunnel source POS5/0
 tunnel mode gre multipoint
 tunnel key 888
 tunnel protection ipsec profile P1 shared

!
```

Spokes Br13a and Br13b:

Spoke Br13a has the DMVPN tunnel connected to Hub11 while br13b's DMVPN tunnel connects to hub12. Br13a is selected as the HSRP active router (higher priority and preemption enabled) and br13b is configured as the standby, This also makes the hub11 as the primary hub and hub12 as the backup hub for this remote site. Working together with the HSRP, Enhanced object tracking is also configured to track two objects, the line protocol on the tunnel interface and the reachability to the hub itself (tunnel destination address). The latter is a more reliable object to track since there are scenarios where the line protocol on the tunnel interface may not go down.

Since the rest of the configuration is similar to other spokes, only the HSRP relevant configuration is shown here:

```
hostname ngwan-br13a
!
track 1 interface Tunnel1 line-protocol
track 2 ip route 192.168.1.14.255.255.255.252 reachability
!
hostname ngwan-br13a
!
track 1 interface Tunnel1 line-protocol
 delay up 50
track 2 ip route 192.168.1.14.255.255.255.252 reachability
 delay up 50
!
interface Vlan110
 ip vrf forwarding vpn1
 ip address 10.10.4.2 255.255.255.0
 standby 1 ip 10.10.4.3
```

```
 standby 1 timers 1 3
 standby 1 priority 105
 standby 1 preempt
 standby 1 track 1 decrement 10
 standby 1 track 2 decrement 10
!
```

```
hostname ngwan-br13b
!
track 1 interface Tunnel1 line-protocol
 delay up 50
track 2 ip route 192.168.1.18.255.255.255.252 reachability
 delay up 50
!
interface Vlan110
 ip vrf forwarding vpn1
 ip address 10.10.4.2 255.255.255.0
 standby 1 ip 10.10.4.3
 standby 1 timers 1 3
 standby 1 preempt
 standby 1 track 1 decrement 10
 standby 1 track 2 decrement 10
!
```

Understanding Convergence

We focus on traffic convergence for the hub <-> spoke traffic. As seen in the redundancy section, there are two major backup options available for two types of branch architecture—single-tier branch which having multiple tunnels originating on the same router and dual-tier branch which use two separate routers with HSRP at the branches.

Single-Tier Branches—Backup Tunnel on the Same Router

When the backup tunnel is on the same router, the traffic convergence is primarily dependent on the IGP. By keeping the default timers, following test has conducted to know the network convergence time.

The failure/recovery is simulated by shut/no shut of the link connecting the hub to the SP (doing it on the SP router).

Table 6-1 Convergence When the Primary Tunnel is Down

Iteration	1		2		3	
	Down	Up	Down	Up	Down	Up
Spoke-to-hub traffic	5.5s	0s	5s	0s	4s	0s
Hub-to-spoke traffic	6s	0s	5s	0s	5s	1s

Table 6-2 Convergence When the Primary Router is Reloaded

Iteration	1		2		3	
	Down	Up	Down	Up	Down	Up
Spoke-to-hub traffic	1.5s	0s	2s	4s	2s	1s
Hub-to-spoke traffic	15s	0s	1.5s	3s	2s	0.6s

As can be seen tuning down the BGP timers provides a much faster convergence. While tuning down to (1s, 3s) provides the best end-to-end convergence performance, it is not recommended in a scaled environment due to the additional overhead on the hub router. At the very least the performance impact needs to be studied in a scaled environment before tuning it down to such levels.

Dual-Tier Branches—Backup Tunnel on Different Routers

With two WAN routers used in the dual-tier branch, the primary and backup tunnels are configured on two different WAN routers (HSRP enabled). Since the branch routers is actually a MPLS PE device maintaining the MP-iBGP session with RRs, the MP-iBGP convergence time is a big factor when the failover happens. In addition, two other factor need to be considered as well—object tracking detection of absence of the DMVPN tunnel availability and HSRP switchover.

We test the failure as in the previous case by shutting the SP link to the hub1 as well as reloading the primary hub. Two sets of BGP timer are tested: BGP default timer and BGP keepalive timers set to (2s, 6s). Shorter BGP timers like (2s, 6s) are not recommended for large-scale deployments without additional testing and is provided here for comparison only.

Convergence Time When BGP Default Timer is Used

Two scenarios, primary link down/up and primary router reload, have been tested.

Table 6-3 Convergence When the Primary Tunnel is Down

Iteration	1		2		3	
	Down	Up	Down	Up	Down	Up
Spoke-to-hub traffic	4s	0s	6s	2s	5s	1s
Hub-to-spoke traffic	3s	0s	5s	0s	5s	1s

Table 6-4 Convergence When the Primary Router is Reload

Iteration	1		2		3	
	Down	Up	Down	Up	Down	Up
Spoke-to-hub traffic	8s	1.5s	10s	1.2s	8s	1.5s
Hub-to-spoke traffic	8s	1s	10s	1.5s	8s	1s

While HSRP provides better redundancy and fast LAN convergence, the convergence on the WAN side is affected by other factors. With object tracking we are watching for the hub's tunnel source interface, which comes up first (in the case of up convergence), but BGP itself takes much longer since it has to wait for the Tunnel itself to come up. Thus even though HSRP has switched over to the original active router (because of preempt), BGP convergence takes longer. This issue can be addressed by delaying the HSRP switchover time when BGP is converging, which can be done by adding the delay statement for the objects that have been tracked. The delay timer implemented in the network needs to adjust based on the BGP convergence time, which needs to be tested in a scenario that's very close to the production network.

From the network design perspective, how the network redundancy is implemented in the remote branch is an integrated part of the overall branch architecture. As shown above, different redundancy approaches yield to different convergence time and which one should be used needs to be evaluated from the overall branch architecture point of view.

Note HSRP timers were kept at 1s for hello and 3s for holdtime in both the cases.

Implementing Multicast

Multicast VPN (MVPN) is the technique used to delivery multicast traffic across the MPLS network for different VPNs (user groups). From a multicast perspective, DMVPN is treated as any other transport media although we do have to account for its multipoint nature in the design.

Assuming that the enterprise MAN MPLS network at the headend is already MVPN enabled, then it is a matter of extending the functionality to the branches. In our example, each VRF is set up with static anycast RP with MSDP enabled, which provides simplicity and redundancy. The RPs typically reside closer to the source and this case the RP is configured in campus CE device at the enterprise MAN data center for each VPN (user group), where the source is connected. All the VPNs in each of the spokes has reachability to the RP and source, so from a VRF perspective the setup looks similar to a normal multicast network.

In the global space mVPN need to be implemented across the entire MPLS network where multicast is required. PIM-SSM or PIM-Bidir are the recommended protocols for the core. Default MDT is used to maintain the control plan traffic and low rate data traffic as well. Data MDT is created automatically when the traffic exceeds the configured threshold. Data MDTs are even more important in the DMVPN network because without them the hub would end up replicating the multicast traffic for all the spokes that are attached to it irrespective of whether they have receivers or not. With Data MDTs the spokes would only join the specific (S,G) if they had receivers for it. This saves CPU resources on the headend device and bandwidth at the hub and the spokes. Two conditions need to be met for the Data MDTs to be initiated:

- The traffic threshold needs to be low enough to enable (set it to 1kbps for almost instantaneous initiation).
- (S,G) entries need to exist within the VRF.

Additionally, PIM NBMA mode needs to be configured on the mGRE interface. This creates the spoke specific entry in the Multicast Output Interface List (OIL).

Caveats:

- CSCse05807 identifies problems with multicast forwarding—received mvpn traffic is process switched on ISRs. This problem is observed when the ISRs are used as a PE (irrespective of using 2547oDMVPN).

Configuration Example:

The following is built on our earlier example (Figure 6-2) and only the multicast-relevant configurations are shown here. The VPN and address scheme is illustrated in Figure 6-1. The multicast traffic is delivered from the campus to remote branches.

Hub11:

```
Hostname ngwan-hub11
!
ip multicast-routing
!
interface Tunnel1
 ip address 150.0.0.1 255.255.255.0
 ip pim nbma-mode
 ip pim sparse-mode
!
ip pim ssm range 1
access-list 1 permit 239.232.0.0 0.0.255.255
!
```

Br11:

```
ip vrf vpn1
 !
 mdt default 239.232.2.1
 mdt data 239.232.2.128 0.0.0.127 threshold 10
!
ip multicast-routing
ip multicast-routing vrf vpn1
!
interface Tunnel1
 ip address 150.0.0.2 255.255.255.0
 no ip redirects
 ip pim sparse-mode
 !

interface Loopback1
 ip address 100.0.250.35 255.255.255.255
 ip pim sparse-mode
!
!
router bgp 64512

 !
 address-family vpnv4
 neighbor RRs send-community extended
 neighbor 100.0.250.9 activate
 neighbor 100.0.250.10 activate
 exit-address-family
!
ip pim spt-threshold infinity
ip pim ssm range 1
ip pim vrf vpn1 rp-address 1.28.103.1
access-list 1 permit 239.232.0.0 0.0.255.255
!
```

Below are the show commands that illustrate the packet flow from the source PE in enterprise MPLS campus to the receiving PEs in the remote branches.

Let's first check the mroute table in both global and VPN space on the campus PE connecting to the source:

```
campus-pe1#sh ip mrou 100.0.250.22 239.232.2.128
IP Multicast Routing Table
```

```
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group
       V - RD & Vector, v - Vector
Outgoing interface flags: H - Hardware switched, A - Assert winner
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(100.0.250.22, 239.232.2.128), 02:19:15/00:03:20, flags: sTz
  Incoming interface: Loopback0, RPF nbr 0.0.0.0, RPF-MFD
  Outgoing interface list:
    GigabitEthernet4/15, Forward/Sparse, 02:19:15/00:02:57, H

campus-pe1#
ngden-7606-pe1#sh ip mrou 239.232.2.128 ac
Active IP Multicast Sources - sending >= 4 kbps

Group: 239.232.2.128, (?)
   Source: 100.0.250.22 (?)
     Rate: 2840 pps/1681 kbps(1sec), 1681 kbps(last 10 secs), 1671 kbps(life avg)
campus-pe1#

campus-pe1#sh ip mrou vrf vpn2 224.2.253.249
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group
       V - RD & Vector, v - Vector
Outgoing interface flags: H - Hardware switched, A - Assert winner
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 224.2.253.249), 02:07:15/00:03:27, RP 1.28.103.1, flags: S
  Incoming interface: GigabitEthernet4/3.1121, RPF nbr 1.28.1.1, RPF-MFD
  Outgoing interface list:
    Tunnel2, Forward/Sparse, 02:07:15/00:03:27, H

(1.28.101.2, 224.2.253.249), 02:07:03/00:03:25, flags: Ty
  Incoming interface: GigabitEthernet4/3.1121, RPF nbr 1.28.1.1, RPF-MFD
  Outgoing interface list:
    Tunnel2, Forward/Sparse, 02:07:15/00:03:27, H

campus-pe1#sh ip mrou vrf vpn2 224.2.253.249 ac
Active IP Multicast Sources - sending >= 4 kbps

Group: 224.2.253.249, (?)
   Source: 1.28.101.2 (?)
     Rate: 2840 pps/1045 kbps(1sec), 1045 kbps(last 0 secs), 1038 kbps(life avg)
campus-pe1#
```

Now let's take a look the mroute table in the DMVPN hub router, which is a P device in the MPLS network:

```
ngwan-hub11#sh ip mrou 100.0.250.22 239.232.2.128
IP Multicast Routing Table
```

```
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group
Outgoing interface flags: H - Hardware switched, A - Assert winner
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(100.0.250.22, 239.232.2.128), 02:22:34/00:03:20, flags: sT
  Incoming interface: GigabitEthernet0/2, RPF nbr 100.0.33.1
  Outgoing interface list:
    Tunnel1, Forward/Sparse, 02:22:34/00:03:05

ngwan-hub11#sh ip mrou 239.232.2.128 ac
Active IP Multicast Sources - sending >= 4 kbps
 a negative (-) Rate counts pps being fast-dropped

Group: 239.232.2.128, (?)
   Source: 100.0.250.22 (?)
     Rate: 2394 pps/1340 kbps(1sec), 1340 kbps(last 0 secs), 357 kbps(life avg)
ngwan-hub11#
```

Finally let's examine the mroute and active multicast traffic in both global and VPN space in the remote spokes (receiving PEs) attached with multicast receiver.

```
ngwan-br12#sh ip mrou 100.0.250.22 239.232.2.128
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group
Outgoing interface flags: H - Hardware switched, A - Assert winner
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(100.0.250.22, 239.232.2.128), 02:55:26/00:02:57, flags: sTIZ
  Incoming interface: Tunnel1, RPF nbr 150.0.0.1
  Outgoing interface list:
    MVRF vpn2, Forward/Sparse, 00:10:10/00:01:50

ngwan-br12#sh ip mrou 239.232.2.128 ac
Active IP Multicast Sources - sending >= 4 kbps
 a negative (-) Rate counts pps being fast-dropped

Group: 239.232.2.128, (?)
   Source: 100.0.250.22 (?)
     Rate: 4787 pps/2680 kbps(1sec), 2654 kbps(last 30 secs), 270 kbps(life avg)
ngwan-br12#


ngwan-br12#sh ip mrou vrf vpn2 224.2.253.249
IP Multicast Routing Table
Flags: D - Dense, S - Sparse, B - Bidir Group, s - SSM Group, C - Connected,
       L - Local, P - Pruned, R - RP-bit set, F - Register flag,
       T - SPT-bit set, J - Join SPT, M - MSDP created entry,
       X - Proxy Join Timer Running, A - Candidate for MSDP Advertisement,
       U - URD, I - Received Source Specific Host Report,
```

```
       Z - Multicast Tunnel, z - MDT-data group sender,
       Y - Joined MDT-data group, y - Sending to MDT-data group
Outgoing interface flags: H - Hardware switched, A - Assert winner
 Timers: Uptime/Expires
 Interface state: Interface, Next-Hop or VCD, State/Mode

(*, 224.2.253.249), 02:51:45/stopped, RP 1.28.103.1, flags: SJC
  Incoming interface: Tunnel3, RPF nbr 100.0.250.22
  Outgoing interface list:
    GigabitEthernet0/1.112, Forward/Sparse, 02:51:42/00:02:01

(1.28.101.2, 224.2.253.249), 02:51:42/00:02:55, flags: JTY
  Incoming interface: Tunnel3, RPF nbr 100.0.250.22,
MDT:[100.0.250.22,239.232.2.128]/00:02:42
  Outgoing interface list:
    GigabitEthernet0/1.112, Forward/Sparse, 02:51:42/00:02:01

ngwan-br12#
ngwan-br12#sh ip mrou vrf vpn2 224.2.253.249 ac
Active IP Multicast Sources - sending >= 4 kbps
 a negative (-) Rate counts pps being fast-dropped

Group: 224.2.253.249, (?)
   Source: 1.28.101.2 (?)
     Rate: 2357 pps/867 kbps(1sec), 879 kbps(last 40 secs), 53 kbps(life avg)
ngwan-br12#
```

Configuration Checklist:

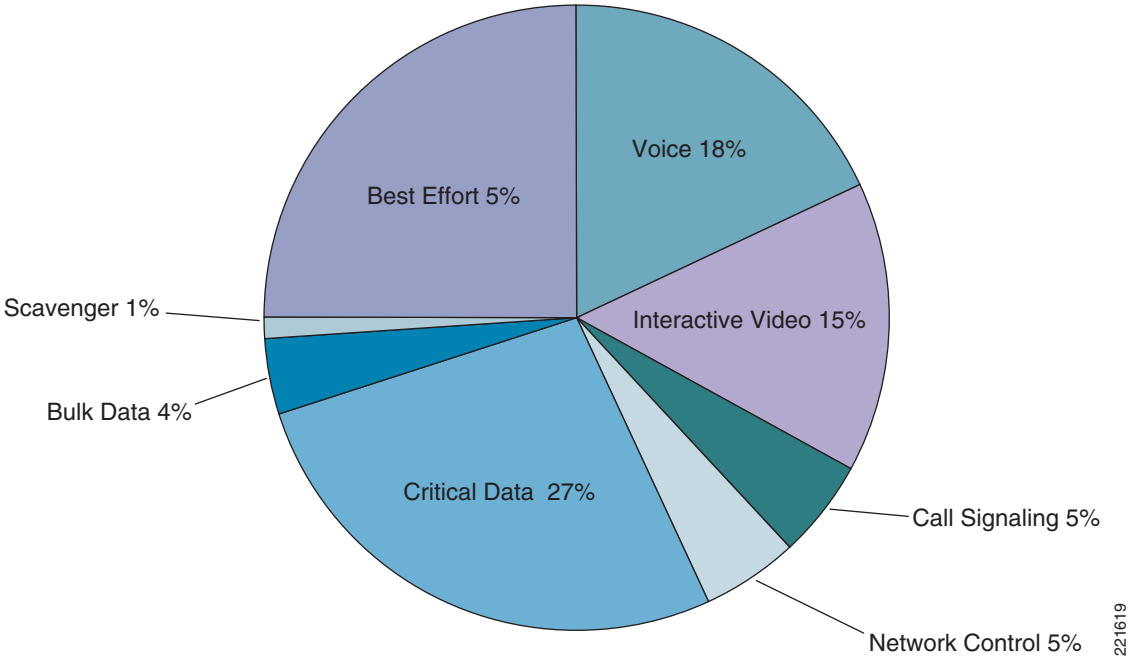
- Configure PIM nbma-mode on the mGRE interface at the headend.
- Ensure that the MDT switchover threshold is set to the lowest value to enable the data MDTs.

Implementing QoS

A basic assumption of this implementation is that the enterprise is getting a Layer 3 VPN service from a provider. Thus the level of QoS service from a provider becomes important. Typically, a service with 3-5 classes of service can be expected. We do not focus on SP service in this design guide as this has been discussed extensively in the Consumer Guidance WP (http://www.cisco.com/application/pdf/en/us/guest/netsol/ns465/c654/cdccont_0900aecd80375d78.pdf). We focus on the aspects of QoS that are within enterprise control, primarily on WAN Edge QoS at the DMVPN headend and the branches.

QoS policies required on the headend include queuing, shaping, selective dropping, and link-efficiency policies in the outbound direction of the WAN link. Traffic is assumed to be correctly classified and marked (at Layer 3) before it ingresses the headend router. At the headend the expectation is that a interface with high link speed is used (DS3/OC3/GE range). At these speeds, link-efficiency policies such as LFI and cRTP are not required. The Enterprise QoS SRND recommends 5-11 classes at the WAN edge. The choice would be dependent on the existing core QoS deployment. We use a 8-class model in our example. The typical bandwidth allocation for a 8-class model is shown in Figure 6-3 (from the SRND).

Figure 6-3 8 Class QoS Model

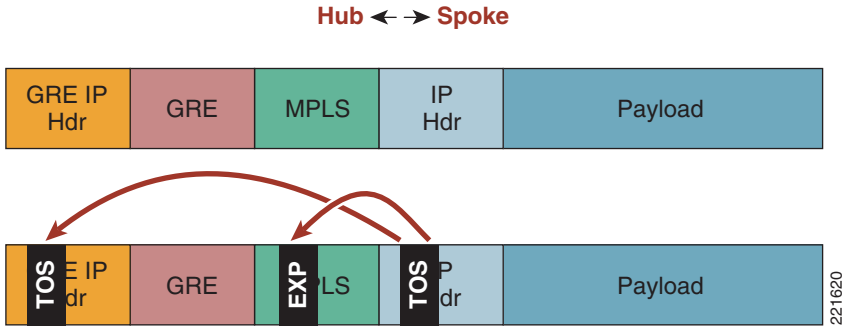


At the branches, the PE could be configured to map the COS to DSCP, but in our example we assume that the packets are already marked with the appropriate DSCP. If the branches have slow/medium speed links (<T1) then a 3-5 class model is recommended. One option could be to match the model used by the SP providing the Layer 3 VPN service. We assume that the branches have higher speed links (>T1) and implement a 8-class model as well (similar to the hub).

Overall the WAN QoS recommendation made in the Enterprise QoS SRND remain for 2547oDMVPN as well. This is because the labeled packets are encapsulated in the GRE header and at the outgoing interface the packet looks like a normal IP packet which can be treated under existing guidelines.

For packets going from hub to spoke and vice-versa, as shown in Figure 6-4, the DSCP/TOS from original IP packet is copied to the MPLS EXP (automatic IP Precedence to EXP mapping) as well as to outer headers DSCP/TOS field.

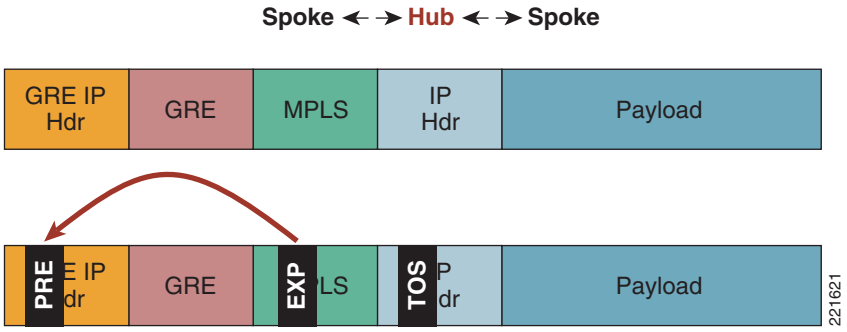
Figure 6-4 Headers for QoS—Hub and Spoke Traffic



Packets traverse the hub in the case of spoke-to-spoke communication. In this case the source branch behavior remains same as above. At the hub, the GRE header is stripped off before a forwarding decision is taken, which in this case requires adding another GRE header before forwarding it back out to the

destination branch. As shown in Figure 6-5, the EXP gets copied to the outgoing GRE IP header as Precedence (3 bits only). Thus the outgoing policy on the hub should account for DSCP as well as Precedence.

Figure 6-5 Headers for QoS—Spoke to Spoke Traffic



The original IP headers marking are always preserved in either case.

Note At the hub, since all the traffic is placed inside the same GRE tunnel, per spoke QoS is not supported.

Example:

Hub1 has a OC3 ATM connection to the SP and spoke B21a has a high speed Ethernet connection. LLQ is used for Voice and Interactive Video traffic. The rest of the classes are provided a bandwidth percentage. DSCP-based WRED is enabled for both Critical Data and Bulk Data.

Hub1:

```
class-map match-any Bulk-Data
 match ip dscp af11
 match ip dscp af12
 match ip precedence 1
class-map match-any Interactive-Video
 match ip dscp af41
 match ip dscp af41
 match ip precedence 4
class-map match-any Network-Control
 match ip dscp cs6
 match ip dscp cs2
 match ip precedence 6
class-map match-any Critical-Data
 match ip dscp af21
 match ip dscp af22
 match ip precedence 2
class-map match-any Call-Signaling
 match ip dscp cs3
 match ip dscp af31
 match ip precedence 3
class-map match-any Voice
 match ip dscp ef
 match ip precedence 5
class-map match-any Scavenger
 match ip dscp cs1
!
policy-map WAN-EDGE
 class Interactive-Video
  priority percent 15
 class Call-Signaling
```

```
  bandwidth percent 5
 class Network-Control
  bandwidth percent 5
 class Critical-Data
  bandwidth percent 27
  random-detect dscp-based
 class Bulk-Data
  bandwidth percent 4
  random-detect dscp-based
 class Scavenger
  bandwidth percent 1
 class Voice
  priority percent 18
 class class-default
  bandwidth percent 25
  random-detect
!
interface ATM5/0
 ip address 135.0.13.2 255.255.255.252
 pvc 1/1
  vbr-nrt 44209 44209
  service-policy output WAN-EDGE
  max-reserved-bandwidth 100
```

Note Scavenger traffic is mapped to Bulk Data at the hub for spoke-to-spoke communication.

Spoke B21a:

```
class-map match-all Bulk-Data
 match ip dscp af11  af12
class-map match-all Interactive-Video
 match ip dscp af41  af42
class-map match-any Network-Control
 match ip dscp cs6
 match ip dscp cs2
class-map match-all Critical-Data
 match ip dscp af21  af22
class-map match-any Call-Signaling
 match ip dscp cs3
 match ip dscp af31
class-map match-all Voice
 match ip dscp ef
class-map match-all Scavenger
 match ip dscp cs1
!
policy-map WAN-EDGE
 class Voice
  priority percent 18
 class Interactive-Video
  priority percent 15
 class Call-Signaling
  bandwidth percent 5
 class Network-Control
  bandwidth percent 5
 class Critical-Data
  bandwidth percent 27
  random-detect dscp-based
 class Bulk-Data
  bandwidth percent 4
  random-detect dscp-based
 class Scavenger
  bandwidth percent 1
```

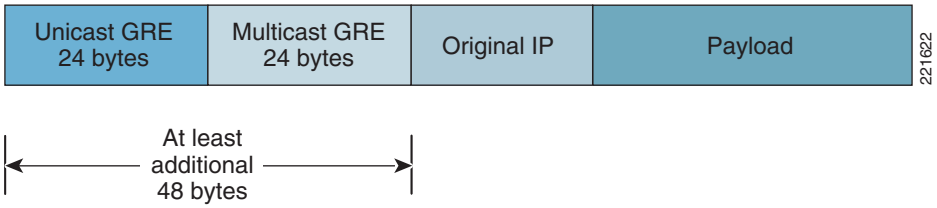
```
 class class-default
  bandwidth percent 25
  random-detect
!
interface FastEthernet1/1
 description To SP
 ip address 135.0.5.1 255.255.255.252
 max-reserved-bandwidth 100
 service-policy output WAN-EDGE
```

Note QoS policy is applied to the outgoing physical interface only. No policy is required on the mGRE interface.

MTU Issues

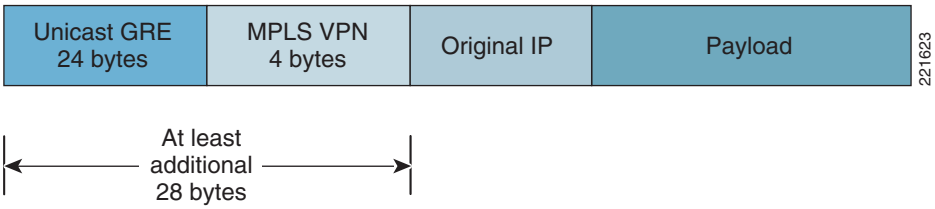
As with any tunneled implementation, MTU size can become an issue. This becomes particular acute with MVPN over 2547oDMVPN when the underlying service is a Layer 3 VPN service from a provider and mVPN is used to delivery the multicast packets. As can be seen in Figure 6-6, at the hub the original IP multicast packet is encapsulated in a multicast GRE header (MTI) which is encapsulated in the unicast GRE header (DMVPN) before being sent to the SP (with appropriate Layer 2 header added). This means that the original IP packet ends up with an additional overhead of at least 48 bytes (without encryption).

Figure 6-6 2547oDMVPN—Multicast Packet Overhead



In the case of unicast, things are a little better. The original IP packet has a MPLS label attached to it before being encapsulated into the unicast GRE (DMVPN). As shown in Figure 6-7, the additional overhead can be expected to be at least 28 bytes (without encryption).

Figure 6-7 2547oDMVPN—Unicast Packet Overhead



The safest MTU (the worst case MTU) for tunnel interface to avoid fragmentation is 1400 bytes. Taking into consideration that each MPLS label is 4 bytes, the safest MTU on 2547oDVMPN tunnel is 1392 Bytes for unicast traffic and 1368 for multicast traffic.

"ip tcp adjust-mss <value>" can also be used to inform the end device to use the correct MSS for TCP transmissions. The MSS must be set to a value that equals the interface MTU minus the size of IP, TCP, GRE, and MPLS headers.

The GRE interface can also be configured with "mpls mtu" which sets the MTU for the labeled packets. MPLS MTU is derived by adding (label stack x 4bytes) to the interface MTU.

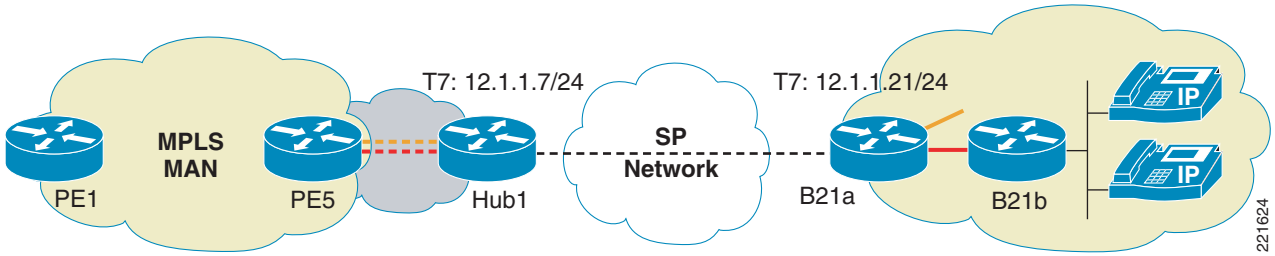
Voice and VRFs

Typically voice traffic has no dependency on the network type since they are just transported as IP packets and require correct QoS behavior applied to them. An exception is when routers are used as gateways for voice services because a lot of voice features and protocols deployed at the branches are not VRF aware (for example, SRST, CME, etc. Thus just getting the voice traffic in a VRF could be a challenge. This is apart from larger issues of having the voice in a VRF; while you can have the IP phones within a VRF, other services such as softphones VT advantage may be in a different VRF. There are challenges in implementing Inter-VRF IP communications. These are not discussed here as its part of the larger virtualization architecture issue. The current recommendation is to keep voice within the global space especially at the branches. At the hub they could remain in the global space or would have to be placed within its own VRF. We look at both options, getting the voice in the VRF at the branch as well keeping it in the global table at the branch.

Voice in a VRF at the Branch

If we need to put the voice in the VRF and still want to use voice features such as CME, then the only way to currently do this is by having two separate routers at the branch. The branch edge router still has a voice VRF configured but treats it like any other VRF. It has a second router (such as a low end ISR) connected to its voice VRF VLAN. The second router, as shown in Figure 6-8, has all the phones attached to it. Since it requires two routers at every such branch, this can be a expensive proposition.

Figure 6-8 2547oDMVPN—Voice in a VRF at the Branch



Example:

Branch 21 has the voice VRF configured on B21a. B21a has 21b connected to it within VRF red-voice. B21b provides the connection support for IP/Pots phones within the branch.

B21a:

```
ip vrf red-voice
 rd 10:104
 route-target export 10:104
 route-target import 10:104
!
interface GigabitEthernet0/0
 description to voice-B21b
```

```
 ip vrf forwarding red-voice
 ip address 125.1.14.129 255.255.255.128
!
router ospf 2 vrf red-voice
 log-adjacency-changes
 redistribute bgp 1 subnets
 network 125.1.14.0 0.0.0.255 area 0
!
router bgp 1
<snip>
!
address-family ipv4 vrf red-voice
redistribute connected
redistribute ospf 2 vrf red-voice match internal external 1 external 2
no synchronization
exit-address-family
```

Voice Global at the Branch

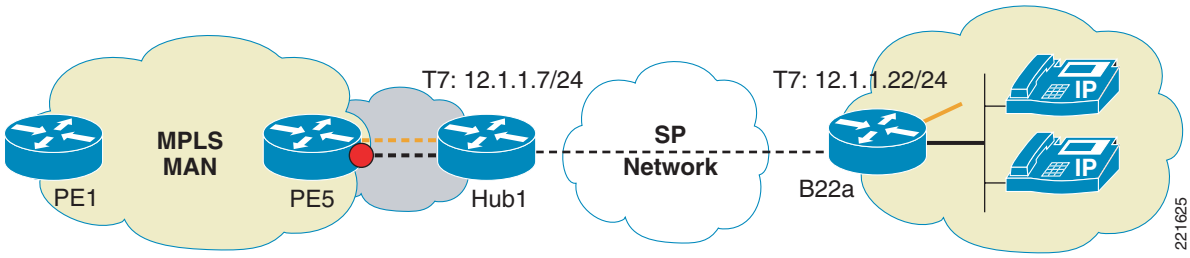
If we choose to keep the voice global at the branch then a single router would suffice. The voice VLAN is connected to the branch router but remains in the global space. It is carried across the same GRE that is used to carry labeled packets but as normal IPv4 traffic. At the hub router, it remains in the global space. It can be forwarded to the next hop router on a global VLAN or can be forwarded to core MPLS PE where it can be placed in its own VRF.

Note To get the voice traffic routed, a routing protocol needs to be used over the GRE tunnels. This potentially has a scaleability impact on the DMVPN setup (relying solely on MP-BGP vs. a combination of MP-BGP and IGP for route distribution).

Example:

B22a in our example is a PE which uses MP-BGP for all the VRFs except voice. Voice traffic exists in global space and hence voice VRF is not defined on it. It runs OSPF with the hub. At the hub the traffic remains in the global space but it has a OSPF peering with PE5. On PE5 the traffic is placed within voice VRF as shown in Figure 6-9.

Figure 6-9 2547oDMVPN—Voice Global at the Branch



B22a:

```
interface Tunnel7
 ip address 12.1.1.22 255.255.255.0
 ip pim sparse-mode
 ip nhrp authentication spe
 ip nhrp network-id 7
 ip nhrp nhs 12.1.1.7
 ip ospf network broadcast
```

```
 ip ospf priority 0
 load-interval 30
 mpls ip
 tunnel source 135.0.3.1
 tunnel destination 135.0.13.2
 tunnel key 777
 tunnel protection ipsec profile P1
!
interface GigabitEthernet0/1.2
 description voice in global
 encapsulation dot1Q 222
 ip address 125.1.15.1 255.255.255.0
 ip pim sparse-mode
 no snmp trap link-status
!
router ospf 3
 log-adjacency-changes
 network 12.1.1.0 0.0.0.255 area 0
 network 121.1.1.0 0.0.0.255 area 0
 network 125.1.15.0 0.0.0.255 area 0
```

Hub1:

```
interface Tunnel7
 ip address 12.1.1.7 255.255.255.0
 no ip redirects
 ip pim nbma-mode
 ip pim sparse-mode
 ip nhrp authentication spe
 ip nhrp map multicast dynamic
 ip nhrp network-id 7
 ip ospf network broadcast
 ip ospf priority 100
 load-interval 30
 mpls ip
 qos pre-classify
 tunnel source 135.0.13.2
 tunnel mode gre multipoint
 tunnel key 777
 tunnel protection ipsec profile P1
!
interface GigabitEthernet0/2.4
 description To PE5
 encapsulation dot1Q 104
 ip address 125.1.103.110 255.255.255.252
!
router ospf 3
 log-adjacency-changes
 network 12.1.1.0 0.0.0.255 area 0
 network 125.1.103.108 0.0.0.3 area 0
 network 125.1.125.22 0.0.0.0 area 0
```

PE5:

```
interface GigabitEthernet1/6.4
 description To 7200-hub1
 encapsulation dot1Q 104
 ip vrf forwarding red-voice
 ip address 125.1.103.109 255.255.255.252
!
```

```
router ospf 2 vrf red-voice
 log-adjacency-changes
 redistribute bgp 1 subnets
 network 125.1.103.108 0.0.0.3 area 0
!
router bgp 1
<snip>
address-family ipv4 vrf red-voice
redistribute ospf 2 vrf red-voice match internal external 1 external 2
no auto-summary
no synchronization
exit-address-family
```

Scale Considerations

The 2547oDMVPN hub can support at least 500 remote spokes as we have tested this scenario in the lab with a simulated customer-representative environment. OSPF is used as the routing protocol between DMVPN hub and spoke. Hence the following scalability numbers apply to this setup; these numbers will satisfy many of the customer scaling requirements:

- 500 OSPF neighbors on hub router
- 500 LDP sessions on hub router
- 500 NHRP entries on hub router
- 500 IPSec sessions on hub router
- 500 MP-iBGP sessions on RRs[1]

Solution Caveats Summary

2547oDMVPN provides a scalable solution for both greenfield virtualization deployments and established DMVPN deployments moving towards virtualization. The following list summarizes the caveats for the deployment model discussed here:

- No direct spoke-to-spoke tunnels can be established. Spoke-to-spoke communication has to happen through the hub.
- For multicast ensure that PIM-NBMA mode is configured on the tunnel interface.
- Use Data MDTs where possible to avoid unnecessary flows to the bandwidth-sensitive remote spokes.
- MTU overhead due to MPLS labels and GRE headers need to be considered.

1. The RRs also handle the other MP-iBGP session for the rest of the network.

CHAPTER 7

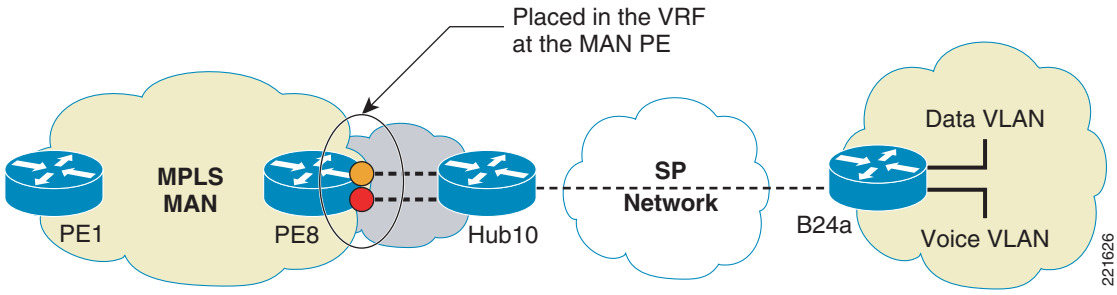
Migration Strategy and Integrating Non-VRF Sites

Either during migration or otherwise, there will be sites that do not have any VRFs configured since the branches are not virtualized. Thus while migrating from a totally non-virtualized WAN to a virtualized WAN it is recommended to have a parallel headend with VRFs enabled. Thus individual branches can be moved from hub1 (non virtualized) to hub2 as each of them is virtualized.

A similar setup is needed to support non-VRF sites as well. There could be scenarios where the enterprise may choose to implement virtualization in only some of the branches (larger ones) and choose to keep the existing setup in the smaller sites. But at the hub, the network is already virtualized with each segment in its own VRF. In such a case, while there is no separation of routes/traffic at the smaller branches, the routes/traffic need to be separated at the hub and placed in their own VRF.

Figure 7-1 demonstrates one such method. This uses DMVPN as a example but would have relevance in the other cases as well. The DMVPN hub terminates non-virtualized spokes. This would mean that any existing WAN setup can remain un changed. Physically, the only change required is the creation of subinterfaces that connect the hub to the MPLS PE upstream. The PE has the VRFs configured while the hub has everything in the global routing table.

Figure 7-1 Integrating non-VRF Sites



Successful implementation requires that the routes advertisement is controlled between the MPLS PE and the DMVPN hub, so that the traffic flows appropriately as well. Traffic from spoke to hub do not have any issues as the MPLS PE advertises the VRF routes over the matching VLAN to the DMVPN hub, thus creating separate paths for the segmented traffic. The traffic from hub to spoke can get a little tricky. It requires the DMVPN hub to control the route advertisement into the MPLS PE. Spoke routes that need to be visible in VRF A need to be advertised over VLAN A and routes that need to be visible in VRF B need to be advertised over VLAN B. This requires two key abilities in the way the networks have been addressed:

- The spoke addresses need to be easily identifiable such that they can be classified into VRFs at the hub.
- These identifiable addresses need to be summarizable for ease of configuration.

Note It is recommended to use BGP as the routing protocol between MPLS PE and DMVPN hub. It allows for most flexibility in filtering routes.

Note Overlapping addresses cannot exist between the VRFs (unless they are NATed) since the spokes and the DMVPN hub would have no way of differentiating between them since they would be in the global table.

Example:

In the following example the DMVPN hub (hub10) supporting spokes (B24a shown here), connected to MPLS PE (PE8). PE8 has two VRFs (red-data and red-voice) configured on the two VLANs connecting to hub10. The spoke has identifiable subnets that correspond to the two VRFs although they co-exist in the global table till they reach PE8. OSPF is running over DMVPN between the hub and the spoke. BGP is running between PE8 and the hub—1 session per VRF to be supported. Routes are mutually redistributed between OSPF and BGP but filtered in BGP when advertised out from the hub to PE8. B24a and PE8 are configured normally. Relevant portions of hub10 configuration are shown below.

Hub10:

```
interface Tunnel11
 ip address 12.2.1.1 255.255.255.0
 no ip redirects
 ip nhrp authentication spoke
 ip nhrp map multicast dynamic
 ip nhrp network-id 1
 ip ospf network point-to-multipoint
 ip ospf priority 100
 tunnel source POS1/0
 tunnel mode gre multipoint
 tunnel key 111
!
interface GigabitEthernet0/3
 description To PE8
!
interface GigabitEthernet0/3.1
 description VLAN for red-data
 encapsulation dot1Q 201
 ip address 125.1.141.2 255.255.255.252
 no snmp trap link-status
!
interface GigabitEthernet0/3.2
 description VLAN for red-voice
 encapsulation dot1Q 202
 ip address 125.1.141.6 255.255.255.252
 no snmp trap link-status
!
router ospf 10
 log-adjacency-changes
 redistribute bgp 3 subnets
 network 12.2.1.0 0.0.0.255 area 0
!
router bgp 3
 no synchronization
 bgp log-neighbor-changes
 redistribute ospf 10 match internal external 1 external 2
```

```
 neighbor 125.1.141.1 remote-as 1
 neighbor 125.1.141.1 update-source GigabitEthernet0/3.1
 neighbor 125.1.141.1 distribute-list data out
 neighbor 125.1.141.5 remote-as 1
 neighbor 125.1.141.5 update-source GigabitEthernet0/3.2
 neighbor 125.1.141.5 distribute-list voice out
 no auto-summary
!
ip access-list standard data
 permit 125.1.17.0 0.0.0.63
ip access-list standard voice
 permit 125.1.17.64 0.0.0.63
```