White Paper

Technology Overview
Virtual Private LAN Service Architectures and Operation

Introduction

Many enterprises are using next-generation IT applications to change their business processes in order to reduce costs and improve operational efficiencies. The attributes and the mission-critical nature of these applications are increasing the expectations of the network in that predictable performance with low transactional delay and high availability are becoming essential requirements. Additionally, new applications are emerging that exploit peer-to-peer communications rather than the traditional client-server models.

Although the requirements for predictable, any-to-any communications dictated by peer-to-peer applications are easily supported within Ethernet-switched campus networks, they become problematic to deliver with the required performance using traditional connectivity services, such as Frame Relay or ATM Layer 2 VPNs (L2VPNs). Although Multiprotocol Label Switching (MPLS) Layer 3 VPNs provide multipoint "any to any" connectivity, some enterprises are reluctant to relinquish routing control of their network and desire L2VPN services with multipoint connectivity.

To address the requirements for a high-performance transport that supports point-to-point as well as multipoint L2VPNs, many enterprises and service providers started to look to Ethernet as a next-generation MAN and WAN technology to provide LAN-like services. This has several benefits for the enterprise because Ethernet is the standard for network-capable devices that typically have in-built Ethernet 10/100- or 10/100/1000-Mbps Ethernet interfaces. The benefits of using Ethernet as a service interface for LAN/MAN and WAN for enterprises and service providers are:

- Flexible access to Layer 3 VPN and Layer 2 multipoint and point-to-point VPN services
- Flexible interface and bandwidth provisioning
- Ubiquitous interface type
- Sympathetic extension of enterprise LAN network
- Lower total cost of ownership
- Simplified operational support requirements

These factors are leading an increasing number of enterprises to request Ethernet connectivity for MAN and WAN services from their service providers as an Ethernet UNI offers tremendous business value. Although this document focuses on multipoint Ethernet services, specifically Virtual Private LAN Service (VPLS), it should be remembered that Ethernet may also be used to deliver Layer 3 VPN as well as Layer 2 point-to-point VPN services. This is important to note because, although Multipoint Layer 2 VPN provides a valuable service, other services such as Layer 3 VPN that may also be delivered using Ethernet, provide unique attributes that may be a better fit for the enterprise's requirements.

Although it is possible to build large-scale Ethernet switched networks to provide multipoint Ethernet L2VPN services using techniques such as 802.1q Tunneling (802.1q-in-802.1q), Ethernet switches cannot be used to build Internet-scale, geographically distributed networks because of technical limitations inherent to Ethernet control protocols and the operation of Ethernet bridges. These limitations have led to the development of mechanisms within the IETF whereby transport protocols, such as Ethernet, can be transported over IP or MPLS networks. Although these technologies allow Ethernet to be transported over IP or MPLS, the mechanisms are inherently point-to-point and are unable to offer multipoint bridged operation.

To provide multipoint Ethernet services, or Transparent LAN Service (TLS) the concept of inter-connecting virtual Ethernet bridges using MPLS pseudowires to provide bridged multipoint connectivity was conceived and VPLS was born. This document discusses VPLS from a technology perspective including VPLS and hierarchical VPLS, autodiscovery of VPLS membership, signaling of emulated wires and forwarding of frames within a VPLS.

What Is a Virtual Private LAN Service?

Very simply, VPLS is an architecture that allows MPLS networks to provide multipoint Ethernet LAN services, often referred to as Transparent LAN Service (TLS). A multipoint network service is one that allows a customer edge (CE) endpoint or node to communicate directly with all other CE nodes associated with the multipoint service. By contrast, using a point-to-point network service such as ATM, the end customer typically designates one CE node to be the hub to which all spoke sites are connected. In this scenario, if a spoke site needs to communicate with another spoke site, it must communicate through the hub, and this requirement can introduce transmission delay.

To provide multipoint Ethernet capability, the IETF VPLS drafts describe the concept of linking virtual Ethernet bridges using MPLS Pseudo-Wires (PWs). As a VPLS forwards Ethernet frames at Layer 2, the operation of VPLS is exactly the same as that found within IEEE 802.1 bridges in that VPLS will self learn source MAC address to port associations, and frames are forwarded based upon the destination MAC address. If the destination address is unknown, or is a broadcast or multicast address, the frame is flooded to all ports associated with the virtual bridge. Although the forwarding operation of VPLS is relatively simple, the VPLS architecture needs to be able to perform other operational functions, such as:

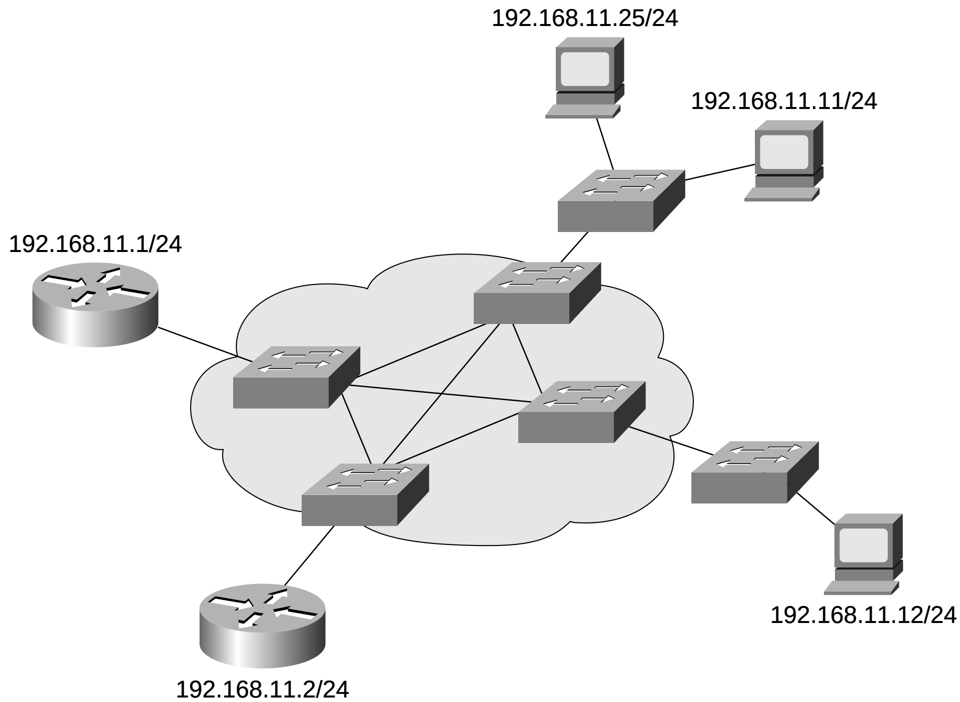
- Autodiscover other provider edges (PEs) associated with a particular VPLS instance
- Signaling of PWs to interconnect VPLS virtual switch instances (VSIs)
- Loop avoidance
- MAC address withdrawal

These features, including frame forwarding, are described below and are also compared to other VPLS implementations and how these differ to those supported by Cisco® VPLS implementations.

VPLS Architecture

The current VPLS working group documents describe two basic architectures; a nonhierarchical, flat architecture and a hierarchical architecture. Cisco Systems® was one of the first vendors to realize the scaling limitations imposed by having a nonhierarchical architecture and developed the concept of a hierarchical VPLS architecture using Ethernet bridging techniques at the edge and MPLS at the core. Please refer to Figure 1 and Figure 2 for a comparison of hierarchical vs. nonhierarchical VPLS architectures.

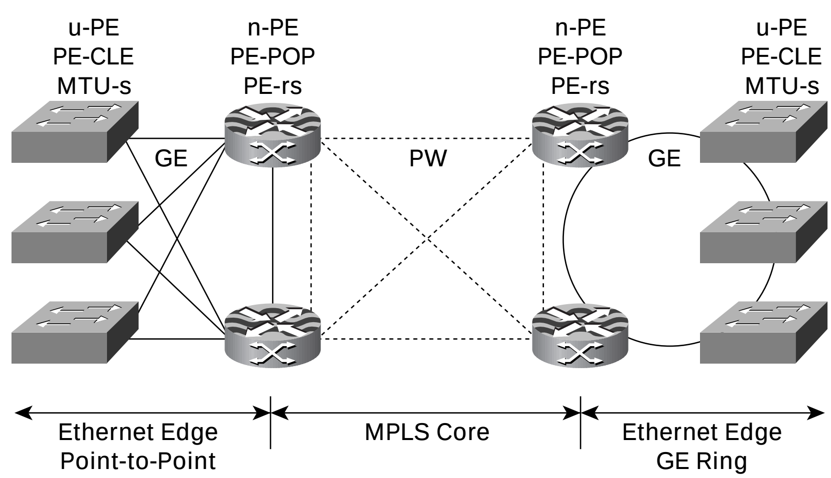
Figure 1
Nonhierarchical VPLS



Although VPLS describes an Ethernet Multipoint Service, the architecture described in Hierarchical VPLS shown in Figure 2 provides an extremely flexible architectural model that also enables multipoint Ethernet services (VPLS), as well as Ethernet Point-to-Point L2VPN services as well as Ethernet access to Layer 3 VPN services.

This document describes the operation of a VPLS from a data forwarding, autodiscovery and signaling perspective and contrasts the implementation of various VPLS architectures existing within the market. This document also describes the IEEE 802.1ad Provider Bridge working group.

Figure 2
Hierarchical VPLS–Ethernet Edge and MPLS Core



VPLS

Metropolitan Ethernet services have garnered considerable interest from service providers and enterprises that wish to exploit the attributes and economics of Ethernet as a transport technology. One emerging technology, the Virtual Private LAN Service (VPLS)[1], provides a mechanism that delivers true Transparent LAN Service (TLS) capabilities across IP/MPLS networks.

The definition of VPLS within the IETF is generating a large amount of interest within service providers because it enables these companies to offer new Layer 2 multipoint connectivity services. The Layer 2 forwarding model is attractive to many service providers because it is a new service model that is complementary to (rather than competitive with) their existing point-to-point ATM and Frame Relay services and also does not require them to interoperate with the customer's routing hierarchy if L3VPNs were deployed.

Several VPLS drafts have been proposed that deliver the same functionality, but, because they use different tunneling mechanisms to automatically discover Layer 2 control protocols, they are mutually exclusive. This document discusses the evolution of IETF *draftietf-l2vpn-vpls-ldp-00* [VPLS-LDP], describes VPLS operations such as architecture, frame forwarding, autodiscovery and recovery, and describes alternative approaches where applicable.

Two important items with respect to VPLS are redundancy and availability, which are important to service providers and service users because network availability has become a critical requirement for enterprise applications. Redundancy within VPLS networks is predicated upon loop avoidance, due to the Layer 2 Ethernet header not having a time-to-live (TTL) field that is decremented at each bridge hop. If a loop exists within an Ethernet bridged network, the lack of a TTL mechanism can cause a single packet to loop endlessly, often referred to as a broadcast storm, which generally makes the bridged network unusable.

It should be noted that although this document focuses on the emerging VPLS technology for multipoint services, Ethernet as an access technology can be used to deliver other services such as point-to-point L2VPN services, multipoint services such as those delivered using VPLS, and more sophisticated IP VPN services to deliver value-added services to enterprise customers.

Cisco IOS® Software supports feature-rich Layer 3 VPNs, as well as Layer 2 point-to-point VPNs using Any Transport over MPLS (AToM), across a broad range of products today. Cisco is committed to delivering a robust and feature-rich VPLS implementation that enhances Cisco's broad portfolio of VPN capabilities. Cisco IOS Software will also support all VPN service types, L3VPN, Layer 2 point-to-point, and Layer 2 multipoint (VPLS) concurrently, thereby allowing service providers the flexibility to deliver any service at any point within the network.

1. Please note that all Internet drafts are works in progress and, as such, should not be referenced because they are subject to deletion from the IETF Website. However, for completeness, the relevant drafts are identified within this document. Should the reader wish to refer to older IETF drafts, http://www.watersprings.org/pub/id/ provides an archive service for 'retired' IETF drafts.

VPLS Terminology

Virtual Private LAN Service (VPLS) drafts, as with any architecture, have many common terms and definitions that are required to enable people to discuss the concepts and architectures with a common language. The terms and concepts most commonly encountered are listed in Table 1 for reference.

As a VPLS emulates the functions of an IEEE 802.1 bridge, this document assumes some familiarity with IEEE bridge concepts, such as loop avoidance using spanning tree protocol.

Table 1 VPLS Terminology

Term.	Description
VPLS	Virtual Private LAN Service. VPLS describes an architecture that delivers Layer 2 service that in all respects emulates an Ethernet LAN across a Wide Area Network (WAN) and inherits the scaling characteristics of a LAN.
u-PE	User facing-Provider Edge. Also referred to as Provider Edge-Customer Located Equipment (PE-CLE). The u-PE is the user facing PE device that is used to connect Customer Edge (CE) devices to the service.
n-PE	Network Provider Edge. Also referred to as Provider Edge Point of Presence (PE-POP). The n-PE is the network facing PE device that acts as a gateway between the MPLS core and edge domain, which may be MPLS or Ethernet.
PE-Agg	Provider Edge—Aggregation Device. The PE-Agg device is an Ethernet switch that aggregates several u-PE connections for onward connection to the n-PE.
PE-s	Provider Edge—Switch. This is a device that is capable of Ethernet bridge operations but has no MPLS capability. This term has been replaced by the u-PE definition.
PE-r	Provider Edge—Router. This is a device that is MPLS capable but has no Ethernet bridging capability.
PE-rs	Provider Edge—Routing Switch. This is a device that is capable of supporting MPLS and Ethernet bridge operation. This term is replaced by the n-PE definition.

Table 1 VPLS Terminology (Continued)

Term.	Description
VSI	Virtual Switch Instance. A VSI describes an Ethernet bridge function within an n-PE that equates to a multipoint L2VPN. A unique attribute of a VSI is that it terminates PW virtual interfaces, which differs from an Ethernet bridge that terminates physical Ethernet interfaces.
PW	Pseudo-Wire. A PW is a virtual connection that in the context of VPLS connects two VSIs. A PW is bi-directional in nature and consists of a pair of uni-directional MPLS Virtual Circuits (VCs). A PW may also be used to connect a point-to-point circuit. Also referred to as an emulated circuit.
AC	Attachment Circuit. An AC is the customer connection to a service provider network. An AC may be a physical port, or virtual port and may be any transport technology, i.e. Frame Relay DLCI, ATM PVC, Ethernet VLAN. In the context of a VPLS, an AC is typically an Ethernet port.
Split Horizon Forwarding	Split Horizon Forwarding within the VPLS drafts is used to obviate the requirement for supporting a spanning tree protocol. Split horizon is a mechanism that simply states that a packet received on an interface should never be forwarded back out of the same interface.
Tag Stacking	Tag Stacking is a Cisco's implementation of Q-in-Q. Tag Stacking in the context of VPLS is used to bundle all customer VLANs into a single L2VPN identifier that identifies which VSI is used to switch the frame. The outer 802.1q label in the Tag Stack is a service delimiting Tag.
QinQ	A mechanism for constructing multipoint Layer 2 VPN using Ethernet switches. Although widely supported within the industry, there is no standard for QinQ encapsulation or features such as tunneling BPDUs. See "Tag Stacking".
Service Delimiting Tag	A Service Delimiting Tag is a frame identifier that a service provider appends to a customer frame to identify a particular VPN or forwarding behavior. A service delimiting tag has local significance only. An example of a service delimiting tag is the outer 802.1q label in a Tag Stacking edge network. If the frame is forwarded to another VSI across an MPLS pseudo-wire, the outer tag may be stripped at the egress n-PE that connects to another Tag stacking area. The ingress VSI may then append a new 802.1q provider tag as required.
IEEE 802.1ad	IEEE Provider Bridges. This project is in response to the IETF VPLS and QinQ implementations that are essentially using IEEE Bridge functions. The project is currently in the draft phase and addresses standardization of QinQ as a mechanism and how BPDUs and other Layer 2 control protocols are handled by the service. IEEE 802.1ad also describes the relationship between an IETF VPLS and how this interacts with an Ethernet bridge. Please refer to the "*IEEE 802.1ad Provider Bridges*" section.
Spanning Tree	The IEEE has defined several spanning tree protocols that determine a loop free forwarding topology within redundant bridge networks. The motivation for spanning tree is that the MAC layer does not have a concept of routing protocols or Time To Live (TTL) field and loops cause packets to be forwarded forever[1]. The IEEE standards that define spanning tree protocols are IEEE 802.1d-1998, IEEE 802.1w Rapid Spanning Tree and IEEE 802.1s Multi-Instance Spanning Tree that is based upon mechanisms described within IEEE 802.1w.
IEEE 802.1s MST	The IEEE 802.1s standard describes a spanning tree protocol that calculates a loop free, bridged topology. The 802.1s standard also describes how multiple spanning tree topologies may be derived such that different VLANs may follow different forwarding topologies on a bridged network.
EE-H-VPLS	Ethernet Edge—Hierarchical VPLS describes a VPLS architecture within [VPLS-LDP] that uses an Ethernet Edge network and MPLS core to provide VPLS services. This architecture is also the basis of the IEEE802.1ad project.
ME-H-VPLS	MPLS Edge—Hierarchical VPLS describes a VPLS architecture within [VPLS-LDP] that uses MPLS end-to-end to provide VPLS services.
FF	Forwarding Function. The forwarding function is an IETF function that is used to forward Ethernet frames to the correct PW within a PW VPLS "bundle". The FF decouples the PW bundles from an IEEE bridge such that a PW bundle is abstracted behind a virtual interface and emulates a LAN segment. The FF function was introduced within [VPLS-LDP] by Cisco to achieve compliance to existing IEEE Bridge standards and IETF Emulated LAN segments.
SI	Service Instance. A service instance as defined within IEEE 802.1ad defines an emulated LAN segment that is used to link two or more services islands that share a common Ethernet service. An SI does not transport provider bridge spanning tree BPDUs between service islands.
SI-MST	Service Instance-Multi-Instance Spanning Tree. The SI-MST is a special case of an SI in that an SI-MST is used to transport provider bridge BPDUs between n-PEs that belong to the same service island. The SI-MST is the only SI that may be used to transport provider bridge BPDUs. This ensures that the service island spanning tree topologies are independent and bounded, and also ensures that loops cannot occur within the network.

1. Forever meaning until the loop is broken, which should be a relatively short time, or until the constituent electrons lose their energy, which will take a very long time to occur.

VPLS Discovery and Signaling

An important aspect of VPN technologies, including VPLS, is the ability of network devices to discover other VPN members and signal PWs to interconnect a particular VPN, often referred to as autodiscovery and signaling mechanisms. Although a lot of attention is focused on these mechanisms, bear in mind that robust network management and operational support systems (OSS) are critical elements in the deployment and management of VPN technologies, whether these are Layer 2 or Layer 3 VPNs.

Discovery mechanisms can be broadly characterized as distributed mechanisms that reside within the network devices, or centralized services that the network devices query to learn VPN associations. Distributed mechanisms such as BGP and LDP require each network device to be configured with VPN associations that the autodiscovery mechanism then advertises to other network devices. Although distributed mechanisms are preferable, they can be prone to configuration errors and security issues such as injection of false information or Denial of Service (DOS) attacks.

Centralized mechanisms such as Dynamic Name Service (DNS) and Radius and Directory Services (AD) require the network devices to poll the centralized server(s) to learn VPN associations. These mechanisms provide a single point of configuration and typically offer robust security, but add additional management elements to the overall network. Another discovery mechanism that is often overlooked is the use of an NMS/OSS that distributes VPN membership to PE devices as the VPN is created by the service management software. This provides desirable features such as service integrity and syntax checking as well as system security as only the network devices that need to be associated with VPN are configured.

Once the PEs associated with a particular VPLS have been identified, the PEs set up unidirectional VCs between each of the PEs identified. As the VCs are set up between the individual pairs of PEs, these VC pairs are then bound together to form bidirectional PWs that are then bound to a particular VSI. Once the unidirectional VCs are signaled as up, and the VCs are bound together to form a bidirectional PW, the PW is considered to be operational and traffic may be forwarded across the PW.

The Cisco IOS® Software VPLS implementation roadmap—which is based upon *draft-ietf-l2vpn-vpls-ldp-00.txt* [VPLS-LDP] differs from other VPLS drafts in that [VPLS-LDP] does not specify a mechanism for the discovery of other n-PEs associated with a particular VPLS. As different service providers may wish to utilize different mechanisms for distributing information to n-PEs, [VPLS-LDP] has simply identified three possible solutions that may be used, BGP[2], DNS[3] and LDP[4] for autodiscovery. [VPLS-LDP] however does not rule out the possibility of using other mechanisms such as directory services or Radius attributes to distribute information regarding VPLS membership to n-PEs within a given network. Once the PEs have received the PE-to-VPLS mapping information, [VPLS-LDP] recommends that LDP is used to set up the PWs to other PEs associated with a particular VPLS instance.

Other VPLS drafts utilize different schemes for autodiscovery of VPLS membership. The original [lasserre] draft utilized LDP signaling to perform both autodiscovery and label negotiation services. Another approach defined within [kompella] is to use BGP to perform autodiscovery and label negotiation services. Although both mechanisms work, they impose a particular mechanism upon the service provider and that have limitations with respect to the scaling and efficient distribution of information to specific end points.

The advantages of a decoupled autodiscovery mechanism, such as BGP, and signaling mechanism, such as LDP, is important to understand as it has ramifications for the scaling and operational aspects of the network. There is little argument that BGP is an excellent mechanism for distributing information regarding VPN membership within a given network, or between networks. However, it is questionable whether BGP is a good mechanism for distributing MPLS label information that is inherently peer-to-peer in nature.

The decision taken in [VPLS-LDP] to not dictate a particular autodiscovery mechanism gives a service provider the flexibility to deploy the solution that is most applicable for their particular requirements. The decoupling of VPLS association from label negotiation also provides an elegant solution for signaling PW parameters between PEs associated with a particular VPLS.

The authors of [VPLS-LDP] and Cisco Systems believe that autodiscovery and label distribution are two distinct services that provide different functions within a given network. As LDP is used to distribute label information within an MPLS network, it makes sense to utilize the point-to-point signaling and information exchange capabilities of LDP for PW signaling. Autodiscovery by contrast is a “broadcast function” in that multiple PEs require to be informed of VPLS membership. This functionality can be provided using a number of different mechanisms including BGP. This flexibility within the [VPLS-LDP] proposal allows a service provider to deploy the right solution for their particular requirements.

2. draft-ietf-ppvpn-bgpvpn-auto-03.txt
3. draft-luciani-ppvpn-vpn-discovery-03.txt
4. draft-stokes-ppvpn-vpls-discover-00.txt

VPLS Discovery

One of the primary considerations for any service provider is the amount of work that is required to enable a particular service and the number of devices that need to be added or configured to enable the service. In most cases, sophisticated network management systems (NMS) and operation support systems (OSS) are used to automate much of the configuration that augments mechanisms inherent within the network. In simple terms, autodiscovery allows PE devices to automatically discover other PE devices that have an association with a particular VPLS instance. Once the PEs have discovered other PEs that have an association with a particular VPLS instance, the PEs can then signal connections to interconnect the PEs associated with a particular VPLS instance.

There are many mechanisms that can be used to distribute VPLS associations between PE devices. One of the most discussed mechanisms is the use of extensions to Border Gateway protocol Version 4 (BGP). However, other mechanisms have been described within the IETF that utilize Label Distribution Protocol (LDP), Dynamic Name Service (DNS) and RADIUS to provide autodiscovery. Although not automatic, static configuration[5] can also be used to define VPLS associations.

Each mechanism has a unique set of scaling and security attributes with respect to autodiscovery and service providers will select the appropriate mechanism to support their particular requirements. A brief discussion of the attributes of the different discovery methods follows.

1. Static configuration requires that each PE associated with a particular VPLS is configured as a peer and it can be seen that the scalability of that solution is low as manual, and often error prone, configuration is required every time a VPLS is added, changed, or deleted. However, as the peers are specifically configured, the security and flexibility to signal additional attributes—such as bandwidth profiles and the like—of the solution is fairly robust as there is no automatic acceptance of received information.
2. NMS/OSS configuration uses a central management point that distributes VPLS membership to each PE associated with a particular VPLS. This provides for syntax checking based upon the type of device being configured and allows other service specific attributes such as bandwidth to be provisioned at the same time. As the peers are specifically configured, the security of the solution is robust as there is no automatic acceptance of received information. Additionally, the NMS/OSS system can have built-in security such as one-time password control and the PE devices can be configured to accept configurations from the NMS/OSS system only.
3. DNS configuration uses a DNS to distribute VPLS membership information. This mechanism provides centralized management and also uses a common syntax. The security attributes of using DNS are good in that the mechanism described in *draft-luciani-ppvpn-vpn-discovery-03.txt* requires that the requesting PE must belong to the DNS entries for the VPLS instance. However, DNS cannot signal additional attributes and require an additional mechanism to provide this information.
4. Radius configuration uses Radius attributes to distribute VPLS membership information. This mechanism provides centralized management and uses a common syntax. The security attributes of using Radius are good in that the mechanism requires that the requesting PE must belong to the have a Radius attribute associating the PE with the requested VPLS. Additionally, Radius A/V pairs can be used to signal additional attributes and provide a flexible discovery mechanism.
5. LDP signaling requires that each PE is identified and a targeted LDP session is active for autodiscovery to take place. Although the configuration can be automated using NMS/OSS the overall scalability of the solution is poor as a PE must be associated with all other PEs for LDP discovery to work, which can lead to a large number of targeted LDP sessions (n2), which may be largely unused as not all VPLS will be associated with every PE. The security attributes of LDP are reasonably good, although additional configuration is required to prevent unauthorized sessions being set up. Although LDP can signal additional attributes, it requires additional configuration either from an NMS/OSS or static configuration.
6. BGP requires that a PE is associated with a particular VPLS is configured under a BGP process. BGP then advertises VPLS membership using NRLIs described within *draft-ietf-ppvpn-bgpvpn-auto-03.txt* that provides a scalable mechanism for distributing VPLS membership. However, as BGP is essentially a broadcast mechanism, by default, the security of the solution can be quite low unless specific mechanisms such as filters are implemented. Additionally, as BGP is a distributed mechanism, it cannot easily distribute attributes such as bandwidth profiles without introducing additional overhead.

The relative scalability, security, and the ability to signal other attributes—such as bandwidth profiles—of each autodiscovery mechanism is tabulated in Table 2.

5. It should be noted that although not explicitly described Directory Services may also be used to distribute VPLS associations.

Table 2 Comparison of VPLS Autodiscovery Mechanisms Summary

Autodiscovery Mechanism	Centralized/Distributed	Scalability	Security	Attributes
Static Configuration	Distributed	Poor	Good	Good
NMS/OSS Configuration	Centralized	Good	Good	Good
DNS	Centralized	Good	Good	Poor
RADIUS	Centralized	Good	Good	Good
LDP	Distributed	Poor	Good	Poor
BGP	Distributed	Good	Fair	Fair

As each mechanism has specific benefits that are of interest to different service providers, Cisco will support a broad range of autodiscovery mechanisms that provide a flexible set of options from which service providers may select the most appropriate autodiscovery mechanism for their network. The flexibility in choosing a discovery mechanism is inherent to the VPLS working group document, draft-ietf-l2vpn-vpls-ldp-01.txt, as it does not describe any particular autodiscovery mechanism as many service providers wish to utilize different mechanisms for autodiscovery purposes.

Signaling

Once the PEs have ascertained that other PEs have an association with the same VPLS instance, each PE needs to set up a PW between the PEs and bind the PWs to the particular VSI. Within the IETF there are two solutions that have been described for signaling of PWs between PEs; one that describes using BGP as the signaling mechanisms and the other that describes using LDP. The difference between these solutions is important to understand as, although one appears to be more efficient and scalable, once the mechanics of the solution are understood the opposite is true.

Both VPLS and H-VPLS utilizes a full mesh of PWs between all n-PEs associated with a particular VPLS, and it has been argued that a "broadcast" mechanism such as BGP, described within [VPLS-BGP], may be used to signal LDP labels. Although this may appear to be desirable due to the familiarity of BGP to the service providers, the actual mechanics of the implementation does not gracefully address the requirements for label distribution of PW signaling between n-PE devices. The implementation described within [VPLS-BGP] advertises an n-PE's association with a particular VPLS as well as a label block from which labels may be assigned to communicate with that n-PE[6]. This approach, although attractive at first glance has the following disadvantages when compared to [VPLS-LDP].

1. All label information is broadcast to all n-PEs associated with a particular VPLS. Although this is acceptable for initial VPLS autodiscovery, subsequent PW signaling is inefficient.
2. PW signaling of essentially peer-to-peer parameters are broadcast to all n-PEs which wastes bandwidth and reduces the efficiency of BGP scaling features such as BGP Route Reflectors.
3. Not all PWs will have the same characteristics between each n-PE associated with a VPLS. Each individual PW may have differing bandwidth profiles depending upon the destination of the PW. As an example a PW within a MAN may have relatively large bandwidth assigned, whereas a transcontinental PW may need to be constrained due to the "cost" in terms of transporting that PW i.e. each PW may have differing bandwidth characteristics per PW or different sequencing characteristics[7].
4. OAM—OAM functions associated with a PW are specific to that PW only and other PWs have no interest in OAM frames or functions of other PWs.
5. Sequencing—One often overlooked attribute of Layer 2 bridged networks is that in-sequence packet delivery is a desirable attribute as some applications do not gracefully recover from out-of-sequence packet delivery and suffer poor performance as a result. To ensure sequenced packet delivery PWs contain a sequence number field that may be used to provide sequenced packet delivery. Again it can be seen that sequenced packet delivery is point-to-point attribute and negotiation or renegotiation of sequence numbers is only required between the PEs associated with a particular PW.

6. The mechanism for assigning labels is described within draft-kompella-mpls-l2vpn-02.txt

7. An attribute of IEEE bridge definitions is that out-of-sequence frames are to be avoided as they may cause some Layer 2 protocols to break. A service provider may therefore differentiate his network or charge more for packet sequence protection.

6. [VPLS-BGP] utilizes BGP signaling for MAC address flush instead of IEEE Spanning Tree TCNs that makes this approach incompatible with IEEE bridges.
7. Label resources are recommended to be over-provisioned to cater for future growth requirements. The allocation of label blocks within the n-PE device causes the label space to become fragmented over time and resources may require re-allocation of later date.
8. BGP scaling mechanisms such as Route Reflectors have to cater for increased signaling overhead that further complicates the BGP implementation and scaling.
9. Pseudo-Wire Edge-to-Edge Emulation (PWE^3) working group utilizes LDP signaling to negotiate VC label assignment and PW parameters[8]. BGP has not been accepted as a mechanism for label negotiation within the PWE^3 working group.

It can be seen that although a full mesh of PWs is formed between VSIs within a VPLS, each individual PW has a set of unique attributes that are specific to a PW and have significance to that PW only. As the attributes are inherently point-to-point in nature, then signaling of these attributes is best performed using a peer-to-peer protocol such as Targeted Label Distribution Protocol (LDP). As most, if not all implementations of VPLS are based upon MPLS, the use of LDP does not introduce new protocols into the network and does not incur any significant overhead.

Table 3 Comparison of VPLS Signaling Mechanisms

Signaling Mechanism	Scalability	Efficiency
LDP	Good	Good
BGP	Fair	Poor

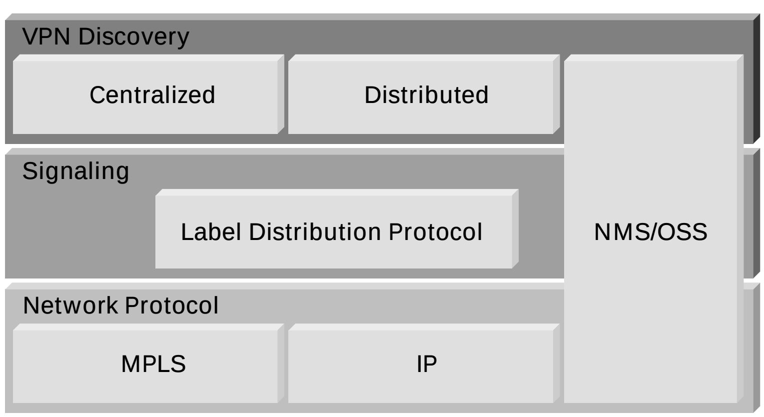
Although some argue that BGP is more efficient for VC label distribution, what can be seen is that the label distribution requires redundant information to be distributed, whereby much of the information received by a PE is simply ignored, and also requires some overprovisioning of VC label resources to make future expansion of the VPLS easier. This inefficiency is further compounded when Route Reflectors and/or BGP Confederations are used, as the information is essentially broadcast to all members of the VPLS, even though the information is specific to a single PW.

Cisco IOS Software's VPLS implementation uses LDP to signal the set up, maintenance and tear down of a PW between two PE devices. Once a PE has discovered that other PEs have an association with a particular VPLS instance, the PEs will signal using targeted LDP to the other PE that a PW is required to be set up between the VPLS VSIs. Cisco supports the signaling and encapsulation described within *draft-martini-l2circuit- trans-mpls-11.txt* to signal PW setup and attribute exchange.

Cisco is also a key contributor to the PWE^3 working group and has made several contributions to describe signaling mechanisms that require minimal configuration that have been adopted by the PWE^3 working group for IP and MPLS signaling.

8. draft-rosen-ppvpn-l2-signaling-02.txt

Figure 3
Layer 2 Virtual Private Network Autodiscovery and Signaling Hierarchy



VPLS Architectures

The current [VPLS-LDP] draft describes three architectures that deliver multipoint Ethernet based services. The following section describes the three architectures that are supported within [VPLS-LDP] with respect to the characteristics, advantages and disadvantages of the different architectures. It is important to understand the background to these drafts as they have influenced both current and future product implementations from Cisco and others.

Lasserre-VPLS

Draft-lasserre-mpls-tls [lasserre] was one of the first drafts that described how a VPLS could be built using Pseudo-Wires (PW) as virtual Ethernet wires to inter-connect virtual Ethernet switches. The key enabler for VPLS technology was the definition of Ethernet over MPLS (EoMPLS) using draft-martini encapsulation[9] that describes a mechanism whereby Ethernet, as well as other transport protocols, can be encapsulated within an MPLS virtual circuit for transport across MPLS networks. Conceptually, VPLS can be thought of as an emulated Ethernet switch with a VSI being analogous to a Virtual LAN (VLAN).

[lasserre] describes an architecture that uses autodiscovery to learn which PEs are associated with a particular VPLS instance, and then automatically provisions PWs (MPLS VCs) to provide connectivity between VPLS VSIs. The [lasserre] draft recommends that LDP is used to automatically discover and provision the VPLS architecture. Although other mechanisms such as BGP are referenced as alternative solutions for autodiscovery these are not discussed. This is important to note as [lasserre] assumes that targeted LDP sessions exist between all PE devices, which limits the overall scalability and flexibility of the solution.

9. draft-martini-l2circuit-trans-mpls-11.txt

When a PE device is associated with a particular VSI, LDP will transmit an LDP label-mapping message (downstream unsolicited) with a VC Type 0x0005 and a 4-byte VC-ID[10] value. If a PE has an association with that particular VC-ID, the PE will accept the LDP mapping message and respond with a label mapping message[11] of its own. Once the two unidirectional VCs are signaled as operational, these are combined to form a single bidirectional PW that terminates as a virtual Ethernet port on the VSI. It should be noted that if one of the unidirectional VCs that forms the PW fails, the terminating device should be able to detect the failure and signal to the peer PE device that the PW is considered as down. The VC is signaled as a VC type 0x0005 Ethernet encapsulation as any VLAN information contained within the frame is not considered to be a service-delimiting field[12].

To reduce the complexity of the VPLS architecture, [lasserre] describes a flat architecture whereby all VSIs that are associated with a single multipoint L2VPN are interconnected using a full mesh of MPLS VCs as shown in Figure 4. As all VSIs are interconnected in a full mesh, [lasserre] avoided implementing a spanning tree by using a technique known as split horizon forwarding. Split horizon forwarding is a frame forwarding technique that prevents packets looping by simply not transmitting a frame back out of the interface the frame was received upon. In the case of [lasserre] if a frame is received on a PW, the frame cannot be forwarded on any other PW associated with a particular VSI as shown in Figure 5. The concept of split horizon forwarding is well-known with routing protocols such as RIP and IGRP, and is important to understand with respect to VPLS as it is reused within other drafts, including [VPLS-LDP].

10. The VC-ID field will be substituted with a VPNID TLV at a later stage.

11. Assuming correct LDP procedure.

12. It should be noted that the Label mapping procedure described within this document adheres to that described within [VPLS-LDP] and not the original [lasserre] specification. This is due to the [lasserre] topology being subsumed within [VPLS-LDP] and it is feasible that some providers may build VPLS architectures with a single hierarchy as described within [lasserre].

Figure 4
Draft-Lasserre Architecture

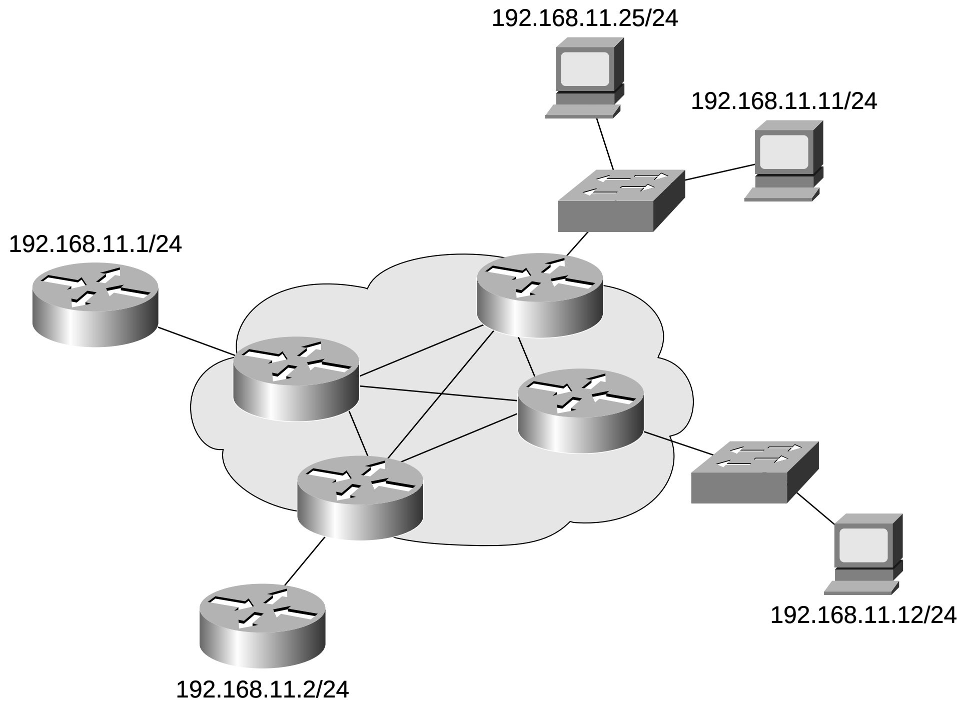
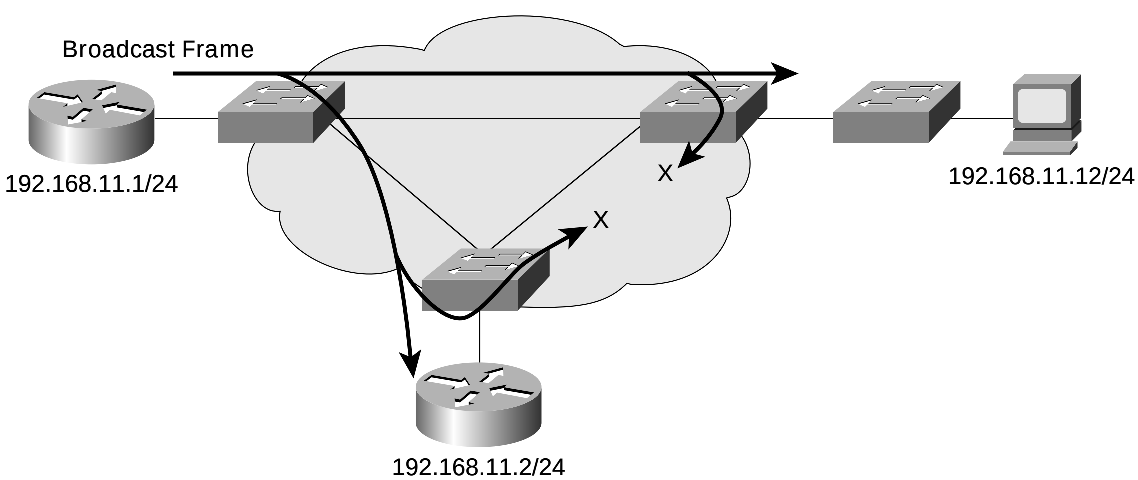


Figure 5
Split Horizon Frame Forwarding



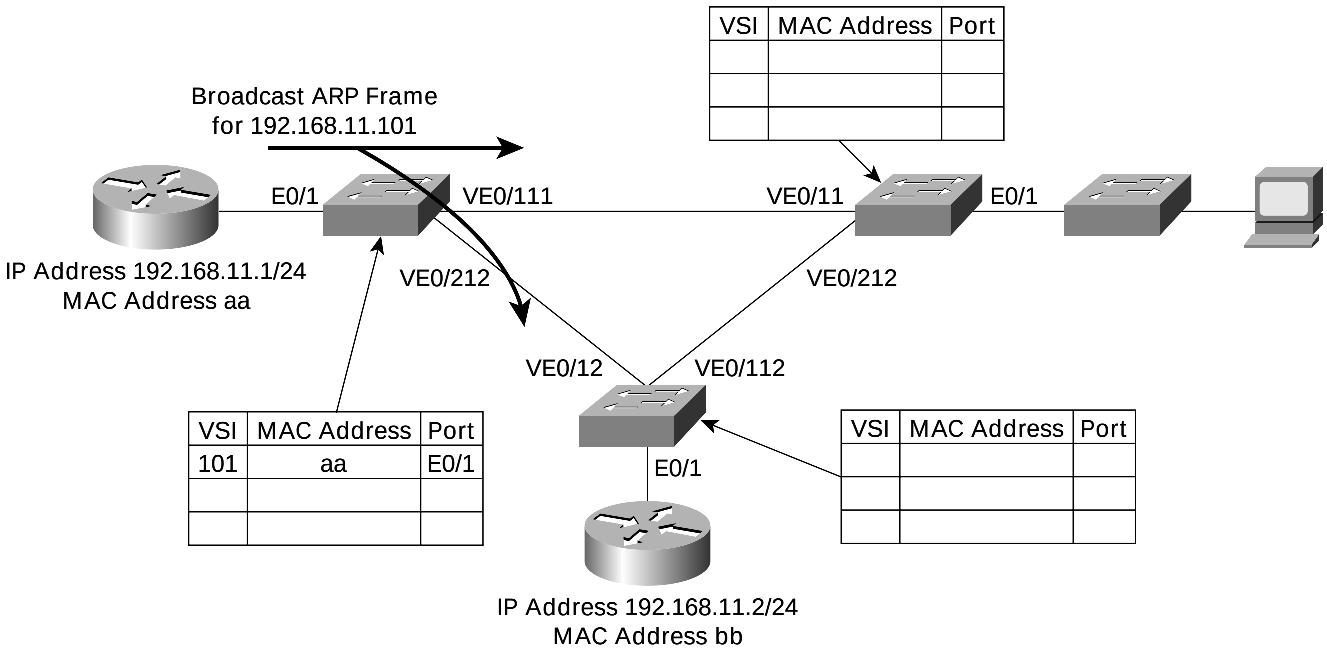
One important consideration for the [lasserre] specification is that although it states that a spanning tree protocol is not desirable due to the scalability of spanning tree protocols[13], it does allow customer BPDUs to be transported across the network. However, [lasserre] simply states that customer BPDUs should be tunneled without describing a solution. This lack of direction leaves BPDU and other Layer 2 control protocol tunneling open to interpretation regarding what is a suitable tunneling mechanism and may preclude interoperability between VPLS implementations.

Additionally, spanning tree BPDUs also carry information that tells other bridges that a topology change has occurred and should flush their bridge forwarding tables. The issues surrounding Layer 2 protocol tunneling such as BPDUs is however common to all VPLS implementations and is currently not addressed.

13. This statement does not consider the scaling attributes of the IEEE802.1s protocols or developments in the architectures of Ethernet switches.

The operation of [lasserre] with respect to frame forwarding follows the same rules as those for Ethernet bridge operation in that it forwards frames using the destination MAC address of an Ethernet frame, self learns source MAC address to port associations, and floods destination-unknown, broadcast, and multicast frames. VPLS address learning can be considered to be a data plane function that uses the same forwarding and filtering rule of Ethernet bridges. Contrast this to Layer 3 VPNs, where each customer advertises subnet reachability information to the service provider, which is then distributed to other members within a particular VPN using iBGP that is considered to be a control plane mechanism.

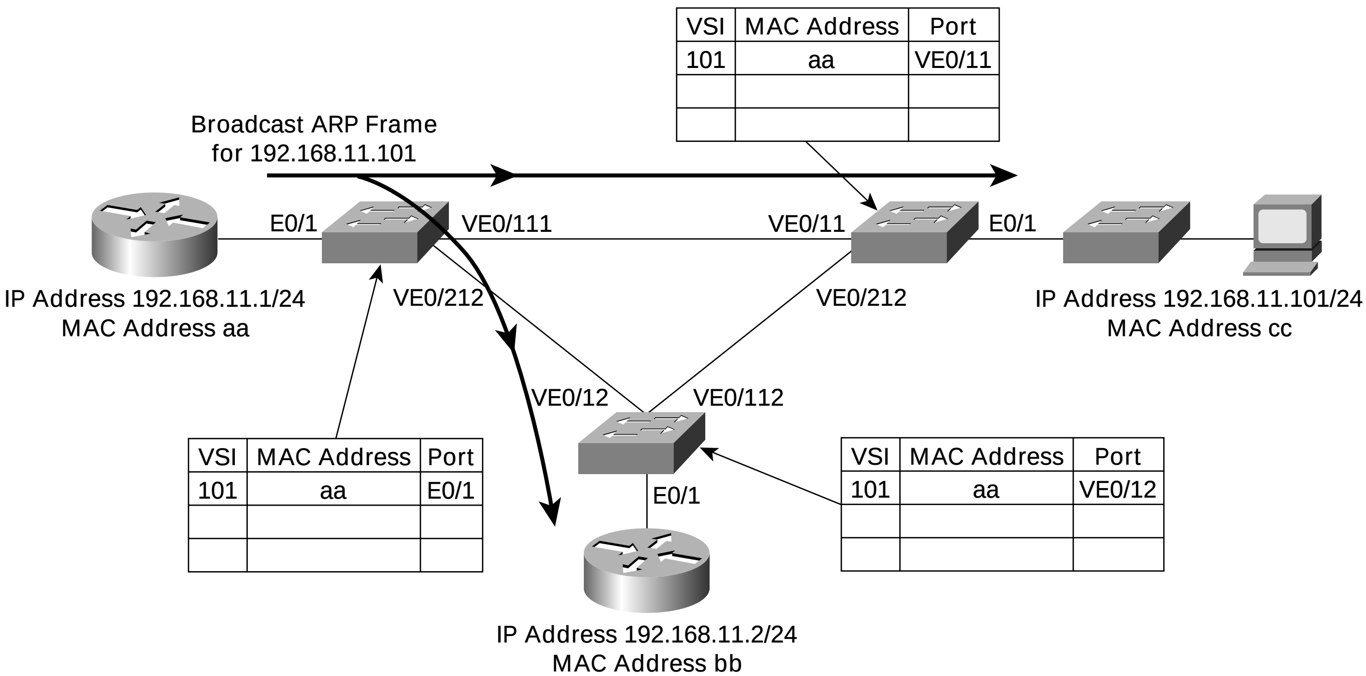
Figure 6
PE#1 Flooding and Address Learning



Using Figure 6 as a reference, client workstations ‘A’, ‘B’ and ‘C’ are attached to an EMS. Client ‘A’ transmits an ARP frame to resolve the hardware address of client ‘B’. When the frame enters the VPLS network, the VSI associated with the EMS in PE #1 perform a table lookup to determine which port the destination address is associated with. As the ARP request has a broadcast MAC address (0xff-ff-ff-ff-ff-ff), the frame is simply flooded to all ports attached to the VSI. If a frame is received that does not have a broadcast or multicast destination address, and no MAC-to-port association exists, the frame is simply flooded to all ports of the VSI as if it were a broadcast. At this time, PE #1 also learns or associates Client ‘A’s MAC address with the port the frame was received upon and stores this information within a forwarding table for future forwarding decisions, as shown in Figure 6.

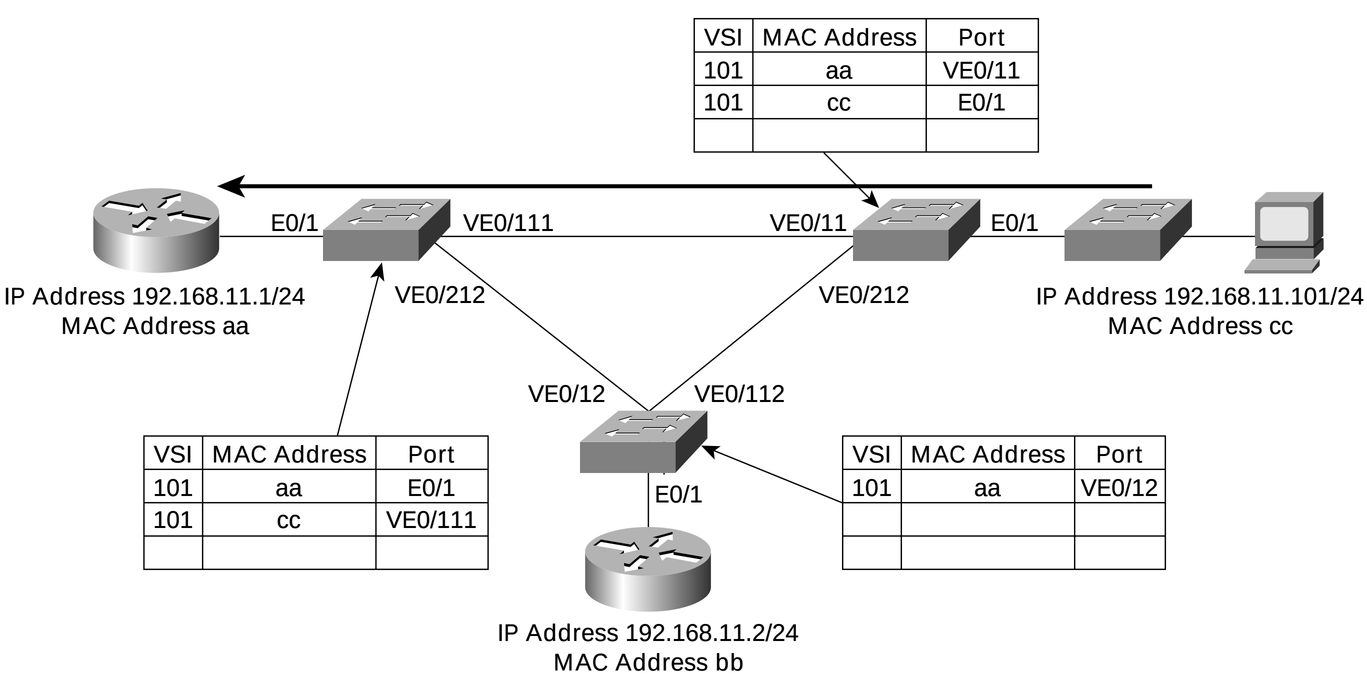
As the VSI does not have a port association for Client ‘B’s MAC address, the frame is flooded to all other VSIs associated with the EMS whereupon each VSI performs a destination lookup. As the frame has a broadcast destination address, the frame is flooded to all ports of the VSI except the PW interfaces following split horizon forwarding procedures. The VSI will also associate the received frame’s source MAC address with the ingress PW within its forwarding table for future forwarding decisions. Please refer to Figure 7 for operation details.

Figure 7
PE#2 and PE#3 Flooding and Address Learning



When the ARP request frame is received at Client ‘B’, Client ‘B’ will respond with a unicast response with Client ‘A’s MAC address as the destination address and Client ‘B’s as the source. When the VSI within PE#2 receives the frame it will perform a destination address lookup for client ‘A’s MAC address. As the VSI will have a MAC-to-port association for Client ‘A’s MAC address, the frame will be forwarded only on the associated port. The VSI will also associate Client ‘B’s address with the port the frame was received upon and store this in its forwarding table. The VSIs in PE#1 and PE#2 now have MAC-to-port associations for Client ‘A’ and Client ‘B’ and frames are switched using the destination MAC address directly to the port associated with that address. Please refer to Figure 8 for operation details.

Figure 8
Constrained Frame Forwarding



The forwarding table associations in each VSI will be refreshed each time a frame is received that matches the source address and port association. If the source MAC address changes location, the VSI will automatically update its forwarding table to reflect the new port association. If a switch does not refresh the MAC-to-port associations within a default time period (IEEE 802.1s/w specifies a default of 300 seconds), the MAC-to-port association is flushed from the forwarding table.

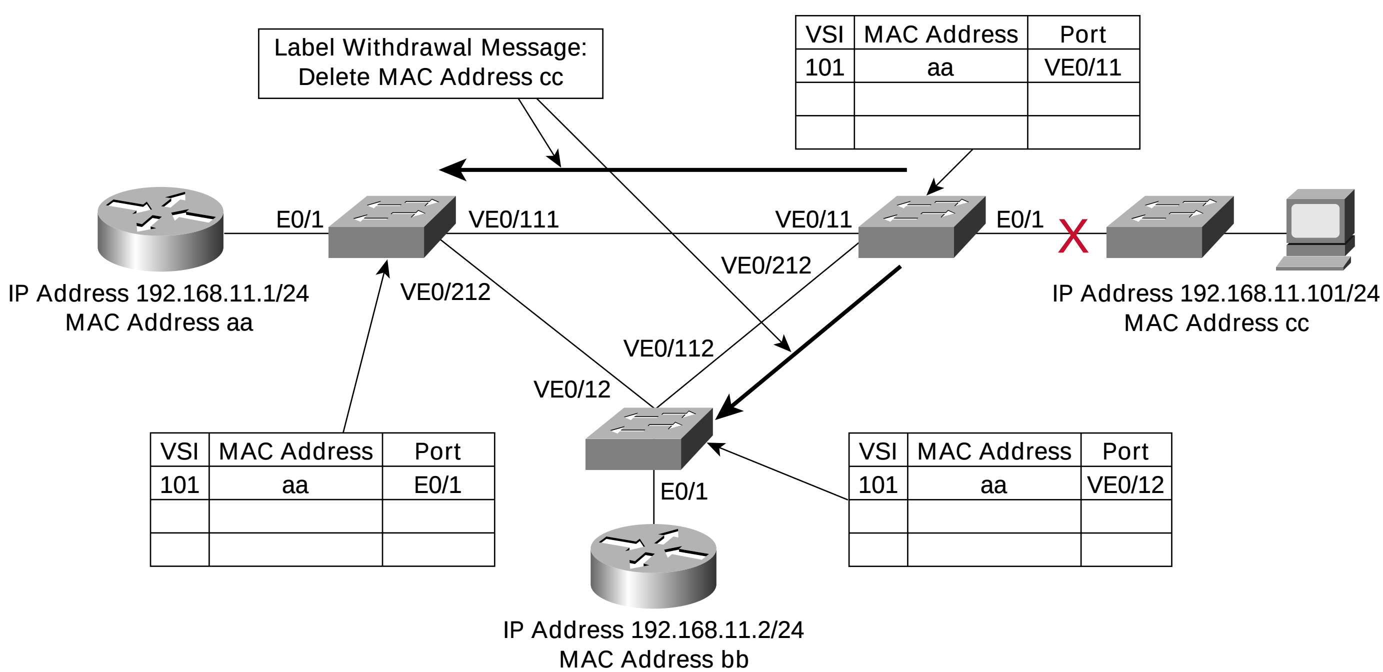
[lasserre] describes a MAC flush mechanism that allows PE devices to inform other PEs that the topology has failed and forwarding entries need to be flushed. The MAC flush mechanism described for [lasserre] that also applies to the H-VPLS variant within [VPLS-LDP] uses an LDP MAC withdraw mechanism whereby an LDP message is transmitted that contains the MAC addresses to be withdrawn. When the MAC flush message is received, the receiving device simply deletes the MAC entries identified within the MAC flush message. The VSI will then flood and relearn MAC address-to-port associations as described previously.

It should be noted that the flush mechanism described within [VPLS-LDP] is incompatible with the IEEE 802.1d/w/s specification with respect to flushing MAC addresses as the IEEE Spanning Tree protocols utilize an in-band Topology Change Notification (TCN) to inform the network that a topology change has occurred.

Although the LDP MAC withdrawal mechanism works for [lasserre] and [H-VPLS], this mechanism cannot support an Ethernet switched edge domain, as the edge switch does not participate in LDP signaling and stale MAC addresses may result. This forces the service provider to deploy an MPLS edge device at the edge of the network that is more expensive than that of a "standard" Ethernet switch. This has significant ramifications with respect to the economics and operational aspects of deploying and maintaining the network.

Within [lasserre] certain aspects such as IP Multicast and security of the service are not addressed. From the perspective of IP Multicast, [lasserre] simply elects to broadcast IP Multicast frames, although the draft does mention that IGMP snooping may be used to constrain IP Multicast traffic. This reference misses an important point about constrained IP Multicast traffic because if routers are attached to the VPLS service, IGMP report messages will not be seen within the service as the routers will send either PIM join messages or DVMRP graft messages. For simplicity, the default approach of broadcasting multicast frames is used. This does not preclude future extensions to VPLS that would allow interaction with 802.1 GMRP protocol, IGMP snooping or PIM snooping to provide constrained IP Multicast forwarding.

Figure 9
MAC Withdrawal and Topology Change



[lasserre] also, quite unintentionally, trivializes the security aspects of the architecture by stating that [lasserre] does not introduce any new security considerations. Although this is true from an MPLS perspective, the security implications for the service and platforms are considerable. Network hackers have become increasingly sophisticated in how they can launch denial of service (DoS) and theft of service attacks against networks, and Layer 2-switched networks are particularly vulnerable. Although it may be argued that security is a vendor “value add” differentiator, it is important to consider the effects of attacks such as MAC flooding, ARP spoofing, DNS spoofing, etc.[14] and the affect these have on the overall security of the service.

Lasserre-VPLS Summary

The original [lasserre] specification provides an architecture that enables a multipoint emulated Ethernet switch service to be delivered over MPLS infrastructure. However, the [lasserre] specification has several architectural limitations that limit the extent to which VPLS may expand. An example of the scaling limitations is the requirement for a full mesh of targeted LDP sessions for VPLS discovery and PW signaling and for a full mesh of PWs for split horizon forwarding within the core.

The requirement for a full mesh of targeted LDP sessions is problematic as the number of provider edge (PE) devices scales, as each edge devices must form an adjacency with all other PE devices. This requires that the edge devices must have the IP address of all remote PEs in it’s routing table, but also requires the PE to exchange label information with all remote PE devices that introduces an n-1 control plane scaling issue, for example, if there are 100 n-PE devices, a total of 99 LDP sessions are required.

14. For more information relating to these attacks please refer to http://naughty.monkey.org/~dugsong/dsniff/.

Other scaling attributes can be calculated by applying relevant network parameters. If a network of 100 PEs supports 1000 VPLS instances that have 100 MAC addresses each, the following scaling figures can be calculated

- Total number of MAC addresses = 100*1,000 = 100,000
- Number of LDP sessions per PE = 99 (n-1)
- Total Number of LSP tunnels = (100*(100-1))/2 = 4950 [n*(n-1)/2]

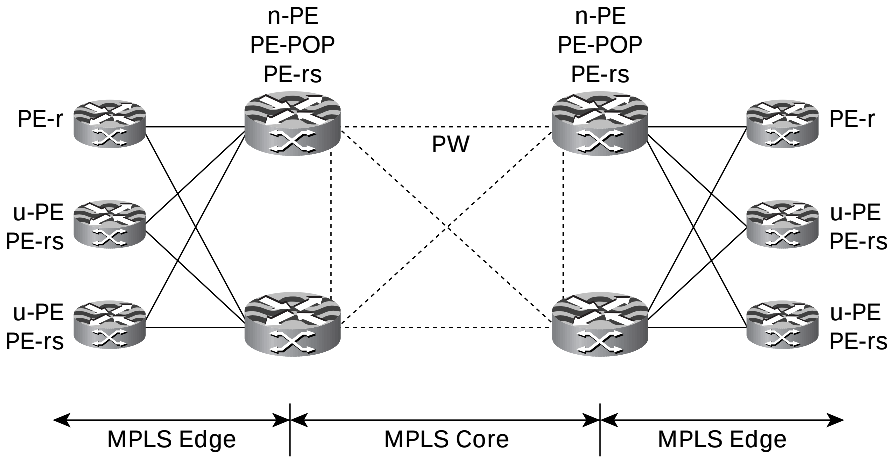
Additionally, the full mesh of PWs between PE devices associated with a particular VPLS instance to meet the requirements for split horizon forwarding is also problematic from a data forwarding perspective. As each remote PE has an associated pair of PWs, the replication of broadcast and multicast traffic needs to be performed at the ingress PE device. If a VPLS instance has a large number of remote PE associations, there is an inefficient use of network bandwidth and system resources as the ingress PE must replicate each frame and append MPLS labels for each remote PE.

Hierarchical VPLS

To address the scaling limitations inherent within [lasserre], two subsequent drafts, draft-khandekar-ppvpn-hvpls-mpls-00 [khandekar] and draft-sajassi-vpls-architectures [sajassi], described the concept of a Hierarchical VPLS (H-VPLS) architecture. H-VPLS describes an architecture that employs a distributed switch architecture that consists of edge domains inter-connected using an MPLS core that addresses some of the limitations of the [lasserre] proposal. The two architectures described within [kandekhar] and [sajassi] have been subsumed into draft-ietf-l2vpn-vpls-ldp-00.txt [VPLS-LDP] and describe a distributed VPLS architecture that may consist of an end-to-end MPLS network, an Ethernet switched edge and MPLS core or a combination of the two. This hierarchical architecture allows the most flexibility in terms of deployment options and the economics of the network.

Figure 10

Hierarchical ME H-VPLS Architecture—MPLS Edge to Edge



The operation of Hierarchical VPLS is similar to that of VPLS in that the operation of the VSI elements at the u-PE and n-PE is that of a standard Ethernet bridge. However, there are subtle differences in how packets are switched due to the architecture of H-VPLS and the differences in how loop avoidance is achieved. The following sections describe the different architectural models, how redundancy may be achieved and how packets flow within each of the models. As [VPLS-LDP] does not utilize a spanning tree protocol for loop avoidance, redundancy in hierarchical networks is critical to the stability and economics of the service. The redundancy mechanisms that may be deployed within a VPLS will be explored in this section.

The physical topology of Ethernet Edge H-VPLS (EE H-VPLS) can be formed of point-to-point Ethernet connections, or Ethernet rings using a spanning tree protocol to provide redundancy and loop avoidance. Other edge architectures that are shown Figure 12 utilize an aggregation layer between the u-PE and n-PE, or indeed utilize Ethernet over SONET/SDH (EoS), or Resilient Packet Ring (RPR) as a transport between the u-PE and n-PE.

Figure 11

Hierarchical EE H-VPLS Architecture—Ethernet Edge, MPLS Core

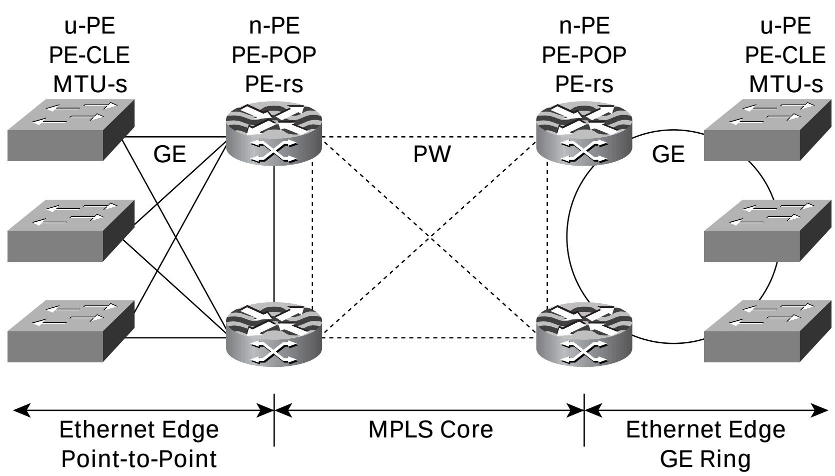
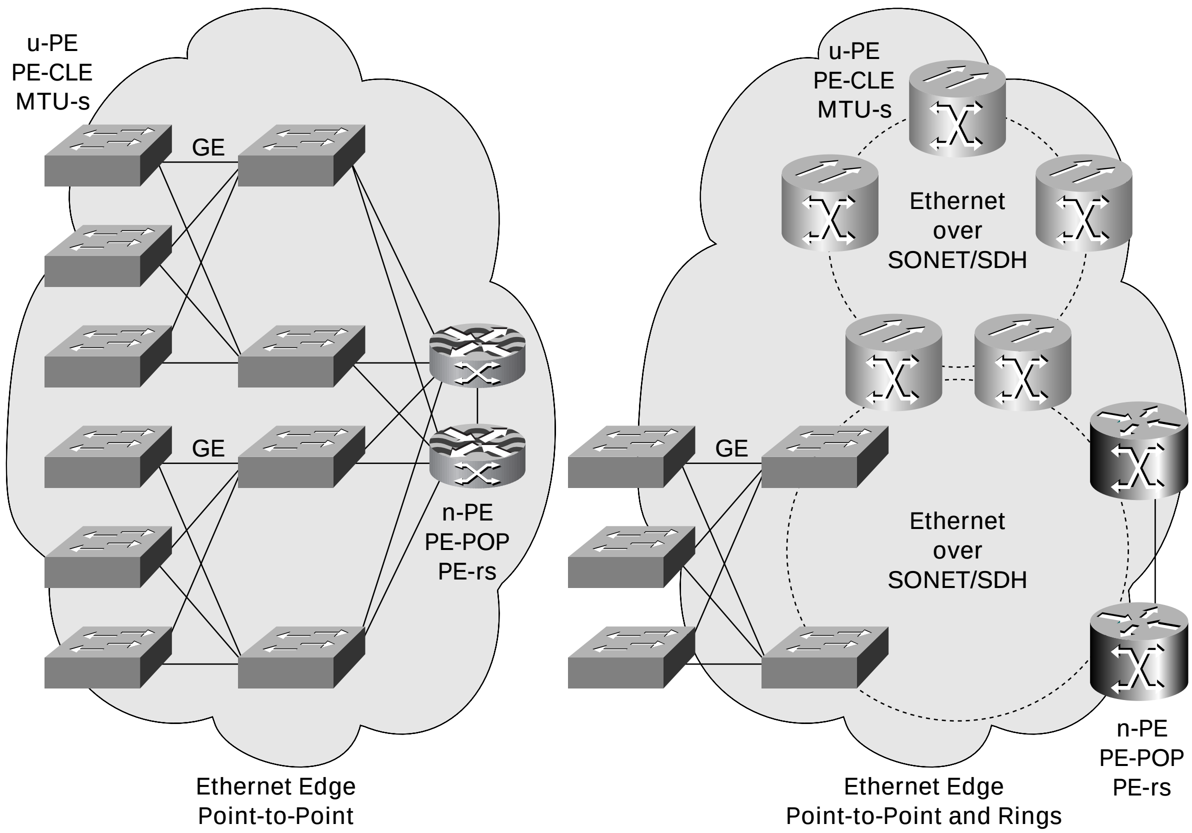


Figure 12

Alternative Ethernet Edge EH-MPLS Architectures



The architecture discussed within [VPLS-LDP] describes a VPLS architecture that consists of PWs to inter-connect VSIs within an MPLS core and either an MPLS edge or Ethernet edge. The current [VPLS-LDP] draft does not specifically discuss autodiscovery, but allows an n-PE to use either DNS, BGP Radius attributes or Directory services[15] to discover other members of a particular VPLS instance. The signaling and binding of labels is also not discussed as semantics of PW signaling are discussed within draft-ietf-pwe3-control-protocol-02.txt, which specifies LDP. This decision to decouple autodiscovery and PW signaling allows a service provider to implement the most applicable autodiscovery solution to meet their requirements.

It should be noted that solutions addressing IP Multicast and network security are not discussed within the [VPLS-LDP] and the considerations for these features are the same as for [lasserre].

15. Manual provisioning of n-PE peers is also supported.

H-VPLS Economics

Within [VPLS-LDP] there are two distinct architectural solutions that are dictated by economic considerations. If we consider an edge-to-edge MPLS VPLS architecture, the edge device must by definition support MPLS label imposition and disposition, LDP signaling and an IP routing protocol. Additionally, the edge device must also support sufficient system resources such as memory and CPU to ensure consistent operation. Additionally, the n-PE devices have to support additional VPLS switching logic that requires label imposition and disposition that is not required by an MPLS P router.

By contrast, if we consider an Ethernet edge domain and an MPLS core, the edge device can be a standard[16] Ethernet switch that uses IEEE 802.1q VLAN tags to provide VPN separation. This allows the Ethernet switch to handle transport characteristics, such as redundancy, as well as quality of service and security.

16. The standard switch nomenclature is slightly misleading as sophisticated features are required to provide robust transport architectures. These features include, Layer 2 protocol Tunneling L2PT, BPDU Guard, ROOT Guard, Port Security 802.1x authentication, Traffic Policing, Sophisticated buffer management, etc. that are all features supported by Cisco Catalyst switches.

If the cost dynamics of the two edge devices are compared, it can be seen that the MPLS edge device will be more expensive than the Ethernet device due to production costs, and the cost to support MPLS and associated MPLS signaling. Additionally, MPLS to the edge offers no real technical benefit when compared to Ethernet mechanisms, and may, arguably, complicate the design and operational aspects of the overall network.

Redundancy in H-VPLS Networks

The redundancy requirements of the two architectures described in [VPLS-LDP] are very different. The architecture as described within [kandekhar] utilizes MPLS edge to edge and redundancy can be achieved at the routing level for intermediate hops between the u-PE and n-PE. However, some form of loop avoidance is required to prevent packets looping between VSIs between the n-PE and u-PE devices.

However, the dynamics of the Ethernet edge model as described within [sajassi] requires some additional configuration. As the edge domain is predicated upon Ethernet switch devices, these devices use spanning tree to form loop free forwarding paths. However, within the core, loop avoidance is predicated around the use of split horizon forwarding which is a different loop avoidance mechanism. Additionally, to limit the extent to which spanning tree may propagate across the network, it is not desirable to simply forward service provider BPDUs across the core, as this will create a large spanning tree domain.

In EE-H-VPLS networks, if spanning tree BPDUs are simply transmitted across the MPLS core, the network will have a single spanning tree topology rooted at a common ROOT bridge. This is generally not desirable as he network fault domain is now extended. Additionally, spanning tree explicitly requires a single ROOT bridge that results in a single administrative domain. This is very undesirable to service providers that wish to extend their service across another provider's network as they will be required to agree on a single authoritative spanning tree root bridge.

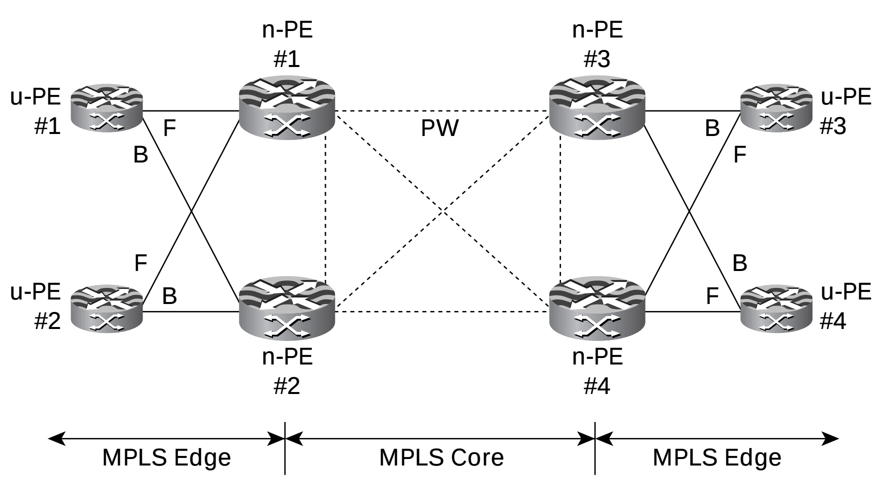
By limiting the extent of a spanning tree to a particular edge domain, the topology of the spanning tree can be greatly simplified and will be faster to react to and recover from failures. Additionally, as the spanning tree is not extended between domains, interprovider services are easier to negotiate as each provider has responsibility for his administrative domain only and does not need to negotiate root bridge placement or priorities.

Packet Forwarding and Redundancy in [VPLS-LDP] MPLS Edge ME-H-VPLS Architectures

The redundancy and loop avoidance attributes of MPLS Edge H-VPLS (ME-H-VPLS), as detailed within [VPLS-LDP], can be broken into two distinct redundancy and loop avoidance problems; u-PE to n-PE and n-PE to n-PE. The [VPLS-LDP] specification uses a full mesh of PWs and split horizon forwarding between the core n-PEs to provide loop free connectivity. The creation of the full mesh of PWs is the same as that described within the [lasserre] section detailed above and will not be discussed further.

The assumption that a full mesh of PWs has considerations for edge connectivity as forwarding loops will occur between the u-PE and u-PE devices as shown in Figure 13. As [VPLS-LDP] does not discuss the use of a spanning tree protocol, other mechanisms are required to prevent loops at the edge domains.

Figure 13
Loop Avoidance in MPLS Edge H-VPLS



At the edge domains, as the u-PE devices are typically smaller devices, the [VPLS-LDP] specification reduces the signaling overhead of the u-PE by reducing the connectivity requirements within the edge domain. Within the MPLS core a full mesh of PWs is formed and split horizon is used to prevent loops. This mechanism could be used within the edge domain, but would increase the cost and functionality of the u-PE devices. Additionally, as nonswitching devices may also be required to participate within a VPLS, the full mesh requirements would necessitate another mechanism to integrate these types of device. To simplify the u-PE device and also allow nonbridging capable devices to participate within an ME-H-VPLS, [VPLS-LDP] states that a u-PE or nonbridging PE device simply forms PW adjacencies with its associated n-PEs.

From the edge u-PEs perspective, it can be seen in Figure 13 that a loop exists between n-PE # 1 and n-PE #2 across the MPLS network if the u-PE ports were not blocked and therefore loop avoidance is required. Once a particular VPLS instance (VSI) is configured on the u-PE #1, u-PE #1 will exchange label mapping messages with n-PE #1 and n-PE #2 in the same fashion described for [lasserre] to establish two LSPs that are associated with the VSIs within u-PE and n-PE #1 and n-PE #2. It should be noted that the u-PE will only form LSP adjacencies with the n-PEs associated with the edge area, not to other u-PEs within the edge domain. Once the PWs have been established between the u-PE and n-PEs and bound to the particular VSI, the u-PE must block the transmission and reception of packets on one of the PWs, or forwarding loops will occur. This is reasonably simple to achieve by blocking the PW to the n-PE with the lowest[17] IP address that, in Figure 13, are n-PE #2 and n-PE #3.

Once the u-PE and n-PE devices have negotiated all of the PWs and calculated their forwarding topology, frames can be forwarded within the VPLS. The forwarding logic is much the same as described within the [lasserre] section, although there are some differences that, in the interest of completeness, will be described in the following paragraphs.

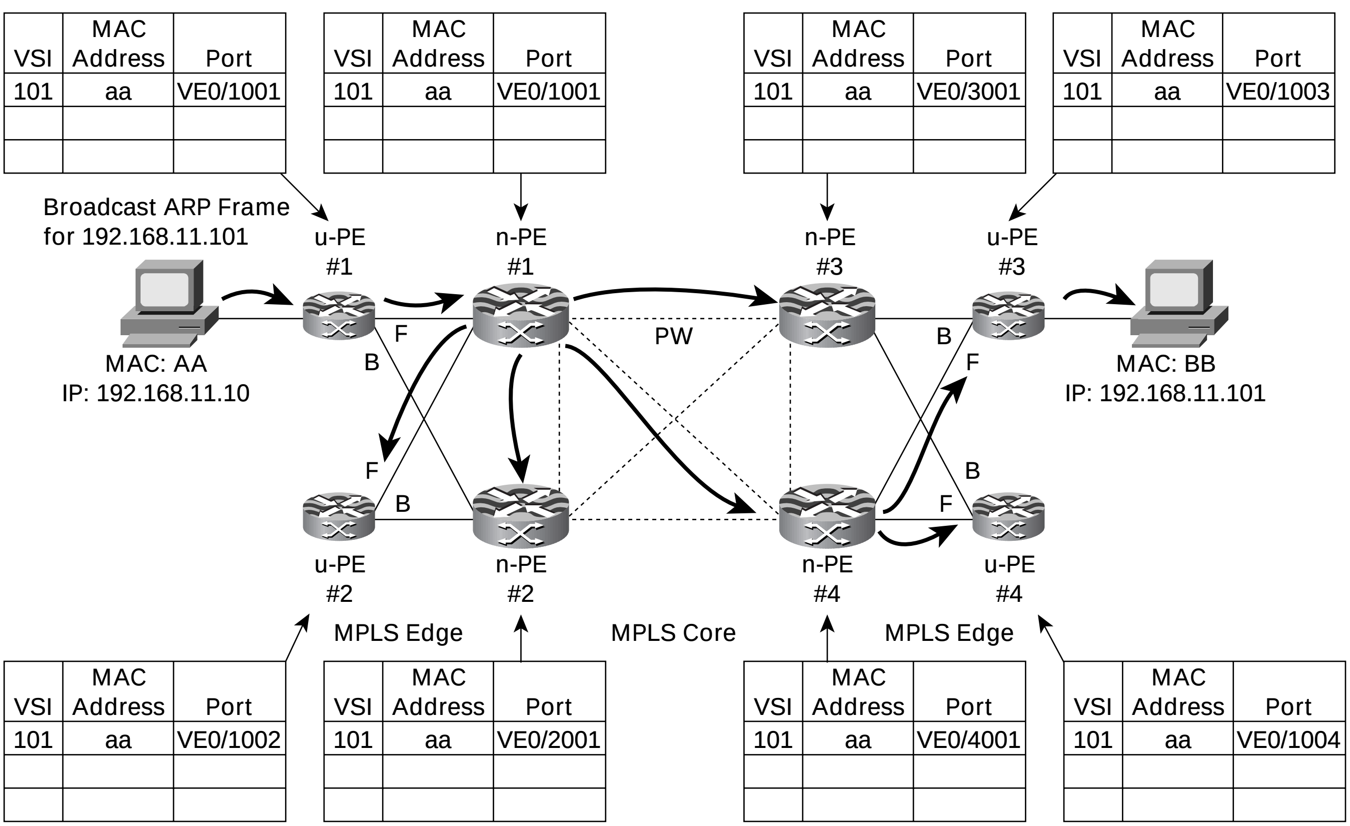
Using Figure 14 as a reference, workstation ‘AA’ connected to u-PE #1 transmits an ARP frame for workstation ‘BB’ connected to u-PE #4. Upon receiving the frame, u-PE #1 will determine that the frame is a broadcast frame and flood the frame over all PWs that are in a forwarding state for that VSI. At this time, u-PE #1 will associate workstation ‘AA’s MAC address with the ingress port the frame was received upon in the VSI forwarding table.

When the frame is received at n-PE #1, the VSI associated with the PW performs a destination address lookup. As the frame is an ARP request and has a broadcast MAC address, the frame is flooded to all PWs associated within the VSI, except the PW the frame was

17. Or highest, it doesn’t matter from an implementation perspective as the decision is local to the u-PE.

Figure 14

H-VPLS MAC Tables Workstation “A” to “B”



received on. This is slightly different to [lasserre] in that [lasserre], having a single hierarchy, does not forward frames received on a PW to other PWs associated with a VSI. This rule is broken within [VPLS-LDP] ME VPLS networks and communications would not be possible. The operation of ME VPLS networks requires additional logic to prevent frames looping within a service. This logic is vendor specific currently as there is no mechanism described within [VPLS-LDP], although a possible solution is discussed below.

Within H-VPLS, the configuration of the VSI dictates whether frames will be subject to split horizon or not. This behavior can be controlled by a neighbor configuration option that determines whether the neighbor needs to have split horizon enabled or disabled. If the neighbor is an n-PE split horizon will be enabled and disabled if the neighbor is a u-PE. Therefore if a broadcast, multicast or destination unknown frame is received from a PW associated with a u-PE, the frame will be copied to all PWs associated with the VSI as split horizon is disabled. If a broadcast, multicast or destination unknown frame is received on a PW associated with an n-PE, the frame will be flooded to all ports associated with the VSI, except PWs that are associated with other n-PEs as split horizon is enabled for those adjacencies.

When the frame is received on the PW at n-PE #2, n-PE #3 and n-PE #4, the frames destination address is determined to be a broadcast and will be flooded on all PWs except those associated with other n-PEs. At this stage, workstation 'AA's MAC address will be associated with the ingress PW in the MAC forwarding table.

When the frame is received at u-PE #3 and u-PE #4, the frame will be recognized as a broadcast frame and will be flooded to all forwarding ports associated with the VSI, and also associated workstation 'AA's MAC address with the PW the frame was received on. All VSIs associated with the L2VPN within the VPLS network will now contain MAC forwarding entry for workstation 'AA' as depicted in Figure 14.

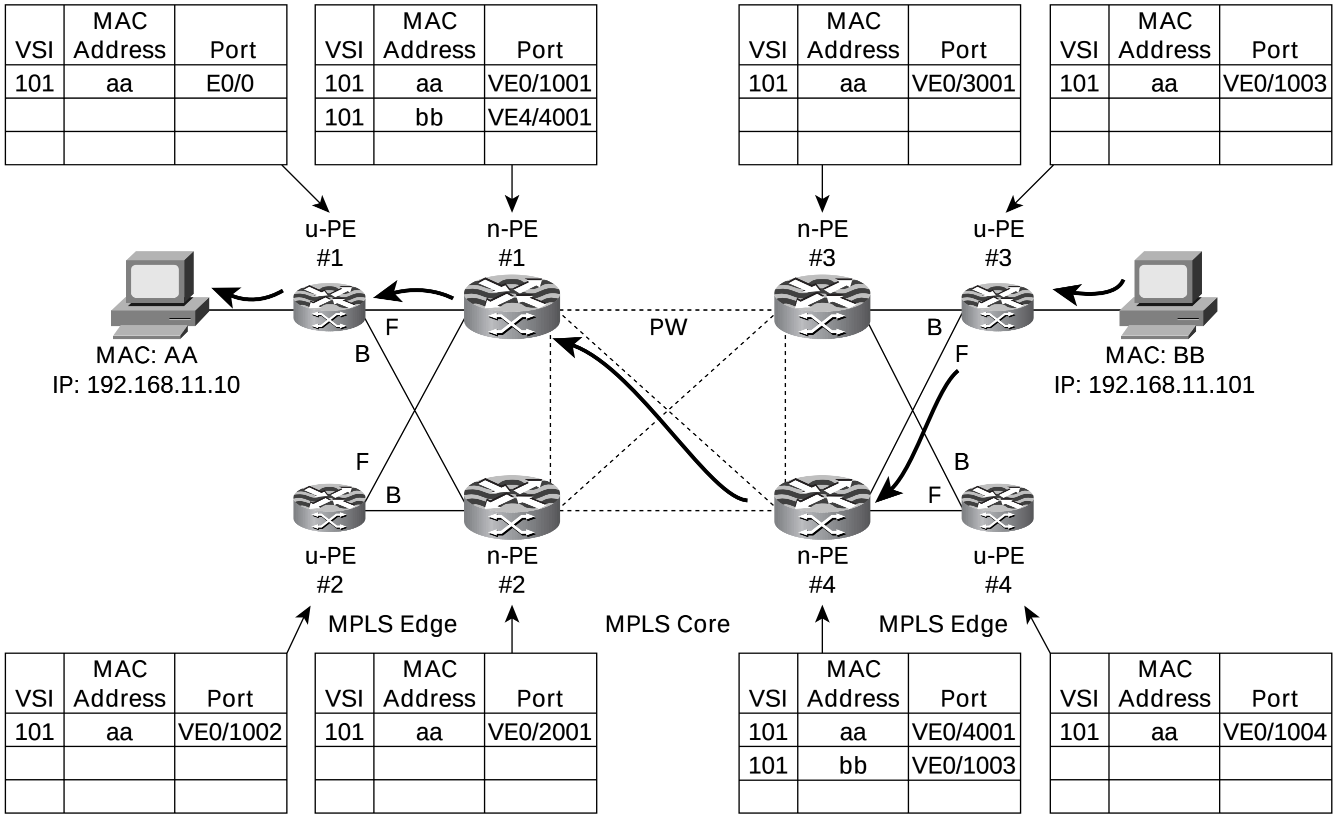
Workstation 'BB', along with any other workstations associated with the VPLS, will have received workstation 'AA's ARP request and processed the frame. Workstation 'BB' will unicast a response to workstation 'AA's ARP request using a destination MAC address 'AA', source MAC address 'BB'.

When the frame is received at u-PE #3, the VSI will perform a destination look up and determine the frame is a unicast and perform a look up within the VSI' MAC forwarding table. The destination look up will find the forwarding entry for MAC address 'AA', and the frame will be forwarded on the PW to n-PE #3. At this stage, the source MAC address 'BB' will be learned and associated with the frames ingress port.

When the frame is received at n-PE #3, the VSI will perform a destination look up and determine the frame is a unicast and perform a look up within the VSI's MAC forwarding table. The destination look up will find the forwarding entry for MAC address 'AA', and the frame will be forwarded on the PW to n-PE #2. Additionally, source MAC address 'BB' will be learned and associated with the frames ingress port.

This process is repeated at n-PE #1 and u-PE #1 until the ARP response frame is finally delivered back to workstation 'AA'. All subsequent frames that are forwarded between workstation 'AA' and 'BB' will be forwarded on the PWs associated with the destination MAC address of each device as depicted in Figure 15.

Figure 15
H-VPLS MAC Tables Workstation “B” to “A”

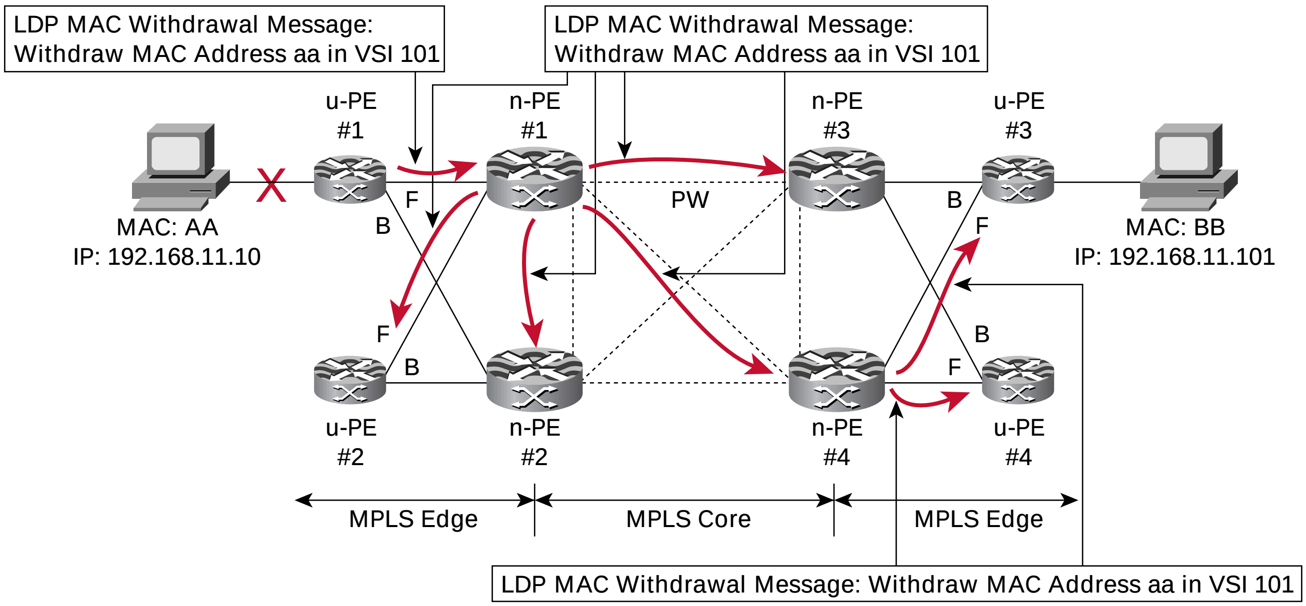


In the event that frames from a particular workstation are not received by a VSI for 300 seconds, the MAC address entry for that particular workstation is flushed from the forwarding table. If a VSI receives a frame for which it does not have a forwarding entry, it will simply flood the frame to all ports associated with the VSI following the rules for split horizon forwarding discussed previously.

If the PW between the u-PE and n-PE, or the forwarding n-PE were to fail, the u-PE simply activates the blocked PW to the standby n-PE. This is a relatively simple mechanism that distributes the redundancy logic to the u-PE devices that reduces signaling overhead on the core n-PE devices. In this instance, the n-PE must send a MAC flush message to other n-PEs to inform them that the topology has changed and the MAC forwarding tables may be invalid.

The MAC flush mechanism described for H-VPLS in [VPLS-LDP] uses an LDP MAC withdraw mechanism, as described within [lasserre], whereby an LDP message is sent that contains the MAC addresses to be withdrawn. When the MAC flush message is received, the receiving device simply deletes the MAC entries identified within the MAC flush message as depicted in Figure 16. The VSI will subsequently flood and relearn MAC-to-port associations as previously described.

Figure 16
MAC Withdrawal and Topology Change



It should be noted that the flush mechanism described within [VPLS-LDP] is incompatible with the IEEE 802.1d/w/s specification with respect to flushing MAC addresses as the IEEE Spanning Tree Protocol utilizes an in-band mechanism to inform the network a topology change has occurred. Although this mechanism works for [lasserre] and ME-H-VPLS, this mechanism cannot support an Ethernet switched edge domain as the edge switch does not participate in LDP signaling and stale MAC addresses in the edge switch forwarding tables may result.

[VPLS-LDP] also describes how nonswitching PE-r capable devices may also participate within a VPLS network. As the PE-r cannot switch frames using the Ethernet destination MAC address, the PE-r forwards the traffic to be switched to the n-PE where the VSI provides the necessary switching capability.

When a port or a VLAN on a PE-r is configured as being associated with a particular VPLS instance, the PE-r signals to the n-PE that a PW needs to be established and associated with a particular VSI. Once the PW is established, the PE-r will simply map traffic that is received on a particular port or VLAN onto a PW and the frame is forwarded to the n-PE device for forwarding.

If a PE-r has several ports or VLANs associated with a particular VPLS, each port or VLAN is mapped to an individual PW that terminates on the same VSI at an n-PE. If traffic needs to be switched between ports on a PE-r, when a frame is received on an ingress port the frame is forwarded on the PW associated with that port to the VSI within an n-PE. The VSI then switches the frame to the egress PW and is forwarded back to the PE-r. The PE-r then simply forwards the frame onto the port or VLAN associated with the PW that the frame is received on. It should be noted that both ME-H-VPLS and EE-H-VPLS can accommodate nonswitching PE-r devices using this relatively simple mechanism.

Packet Forwarding and Redundancy in [VPLS-LDP] Ethernet Edge MPLS Core (EE-H-VPLS) Architectures

The second architecture described in [VPLS-LDP] utilizes an Ethernet edge and MPLS core (EE-H-VPLS) to deliver a VPLS service, which was first described within [sajassi]. Although, this model utilizes Ethernet mechanisms such as 802.1 spanning tree to provide loop avoidance at the edge, [sajassi] retains split horizon forwarding for the core to simplify the operation of the VPLS architecture and also allows the network to scale by limiting the extent to which spanning tree is extended across the network.

The extent to which spanning tree is extended is limited by not propagating spanning tree BPDUs across the MPLS connections. This filtering does however cause some issues as loops can occur due to the noncongruent nature of the loop avoidance mechanism (802.1d/s and split horizon). This section details the solutions that are available to address redundancy within EE-H-VPLS.

Although [H-VPLS] and [sajassi] describe similar architectures, there are differences that need to be taken into consideration with respect to redundancy. In the respect of packet forwarding both architectures are similar to that described in the ME H-VPLS section and will not be repeated here. The redundancy attributes are however very different and will be discussed in detail.

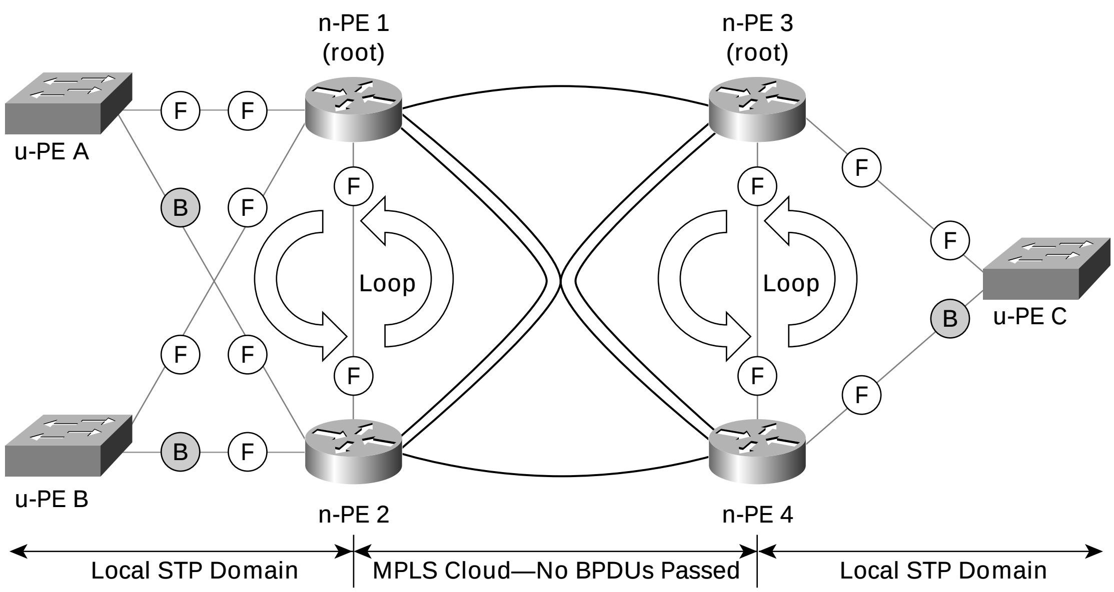
The diagram shown in Figure 17 depicts a simplified H-VPLS architecture based upon an Ethernet edge[18] and MPLS core that depicts the possible loops that are inherent within the design. This is due to the design choice that was made to block spanning tree BPDUs to limit the extent to which a spanning tree topology may extend within the network. As BPDUs are prevented from crossing the MPLS PWs, the n-PEs will not receive BPDUs from other n-PEs.

In practice, if an Ethernet bridge does not receive BPDUs on a particular port, it assumes that there is no neighboring bridge and therefore a loop cannot exist. The bridge will subsequently transition the port to forwarding status and will start to forward frames, thereby causing a loop. It can be seen in Figure 17 that a loop exists between n-PE #1 and n-PE#2, as a PW exists between them that provides redundant connectivity for VSI (VLAN) 101. In fact there are several loops that can be identified, although only two are shown for clarity.

The loop avoidance characteristics of EE-H-VPLS are complicated by the disjointed approach taken for edge and core redundancy and loop avoidance mechanisms and the interaction of the two.

18. Whether the Ethernet edge domain is IEEE 802.1q or a Tag Stacking network is largely irrelevant.

Figure 17
Loops Inherent to EE H-VPLS



To break the loops within the network there are three solutions that may be considered:

1. Disable spanning tree at the n-PEs.
2. Utilize spanning tree and packet filtering to constrain data traffic to provide a loop free topology
3. Utilize IP addressing and routing protocol policy to constrain the forwarding topology.

The first option has the disadvantage that it allows spanning tree BPDUs to be propagated across the MPLS core, thereby extending the spanning tree topologies and reducing the scalability of the VPLS network, and will not be considered further. The last two options will be discussed in detail in the following section.

Constrained Data Forwarding

This solution to redundancy in EE H-VPLS relies upon a combination of spanning tree and access control lists (ACLs) to control how traffic is forwarded within the network and thereby provide loop-free forwarding and redundancy. In Figure 17, the Ethernet forwarding path has been correctly calculated and the u-PE ports are blocked to prevent loops. However, a forwarding loop exists between n-PE #1 and n-PE #2[19] due to the forwarding behavior of the MPLS core network and the Ethernet edge.

If a broadcast[20] frame is forwarded from u-PE #1, n-PE #1 will forward the frame to all ports associated with the VSI. A copy of the broadcast frame will then be received by n-PE #2 on the Gigabit

19. It should be noted that loops will exist between any redundant n-PE pair.
20. Broadcast, multicast or destination unknown for completeness.

trunk 1/1 and the PW linking n-PE #1 and n-PE #2. Upon receipt of the broadcast frames, n-PE #2 will forward the broadcast received on the PW to all ports associated with the VSI and also forward a copy of the broadcast received on the Gigabit Ethernet trunk to all ports associated with the VSI. Although this sounds like a violation of split horizon, the VSI is operating correctly as the broadcast is received on two different interfaces.

When copies of the broadcast are received on the Gigabit Ethernet and PW ports of n-PE #1, n-PE # 1 simply copies frames to all ports associated with the VSI again. This packet looping is self-perpetuating and the frames will now circulate until the loop is broken. This behavior is obviously suboptimal and a mechanism to suppress these loops is required.

One simple way to prevent a loop is to artificially filter traffic that crosses the Ethernet connection between the n-PEs. Figure 18 depicts EE-H-VPLS loop avoidance that shows a single u-PE that is dual-homed to the two n-PEs. As n-PE #1 is the root of the spanning tree, one port on the edge u-PEs is blocked anyway. In this configuration, the only links that need to be filtered are the inter-n-PE link. However, as spanning tree will still be required to block u-PE interfaces (and so as not to introduce another loop) spanning tree BPDUs must be allowed across the inter-n-PE link. This makes the filter very simple as only BPDUs are allowed across the link between the n-PEs

Now when a broadcast frame is forwarded, the broadcast frame cannot be forwarded across the Gigabit Ethernet link between the n-PEs as these frames are filtered, and the spanning tree topology causes the u-PEs to block ports correctly, so the loop is broken and correct forwarding behavior is maintained. Please refer to Figure 18 for a representation of the principles and blocking behavior.

In the event that a u-PE loses its primary link to n-PE #1, the u-PE will unblock it's previously blocked port and normal forwarding will occur. It should be noted that if a frame requires to be forwarded between u-PE # 1 and u-PE #2, the forwarding path would be via the MPLS core and not via the Gigabit Ethernet trunk. This is due to the filtering of frames that prevents addresses from being learned as available via this link.

Figure 18
Loop Avoidance in EE H-VPLS Network Using ACL

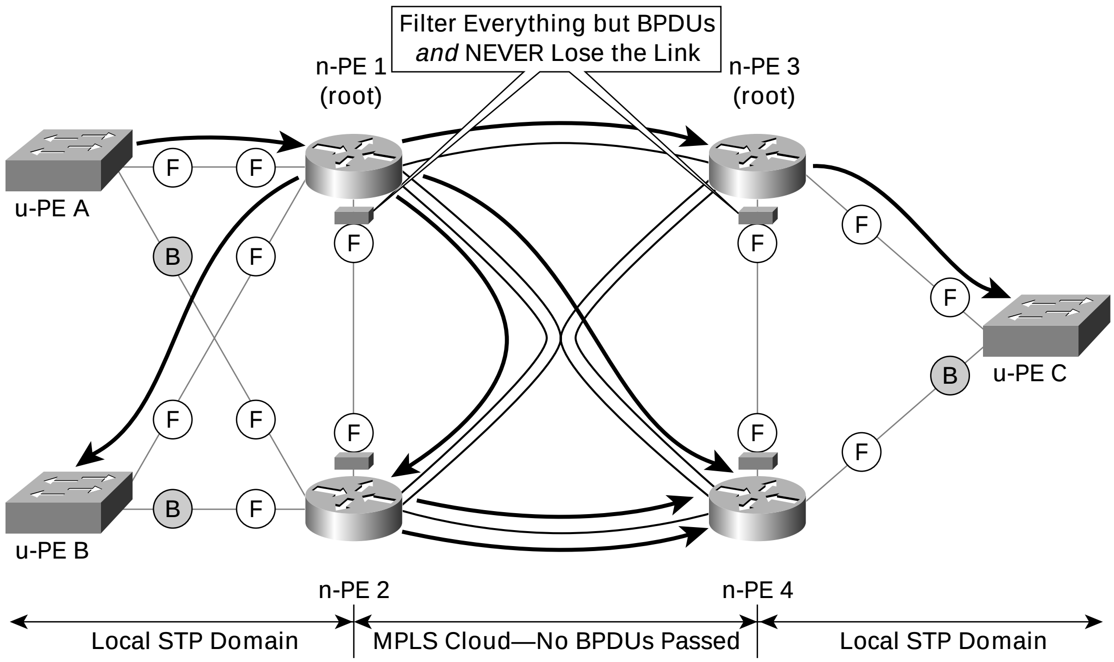
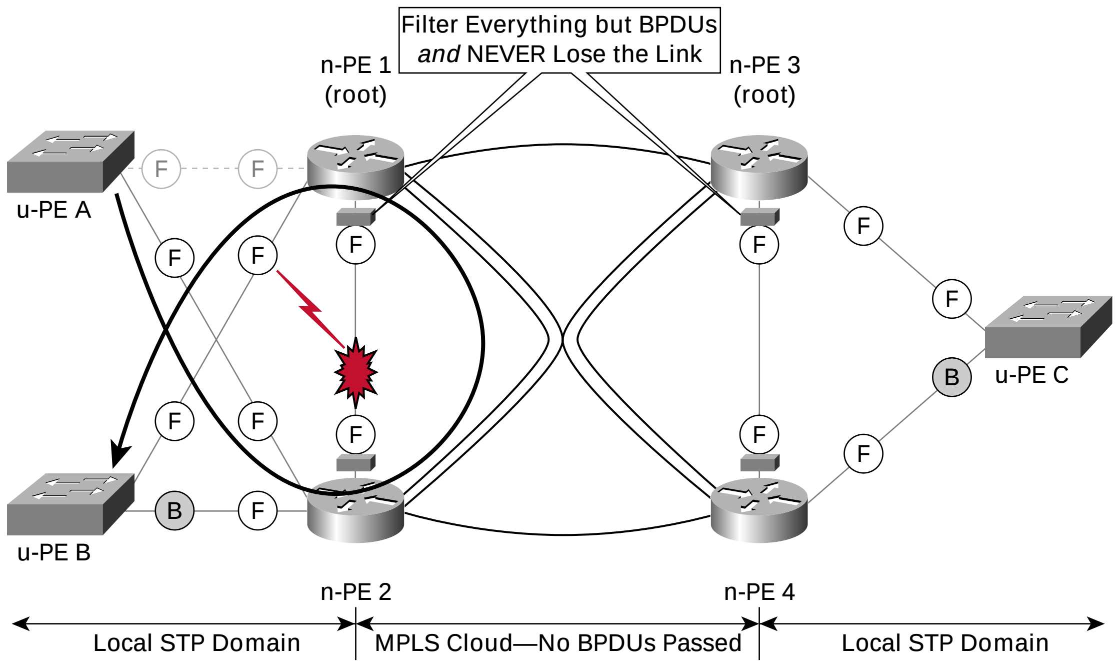


Figure 19

EE H-VPLS u-PE Failure Scenario



Note: Traffic associated with a particular u-PE will always be forwarded via the MPLS core following the failure of its primary n-PE connection.

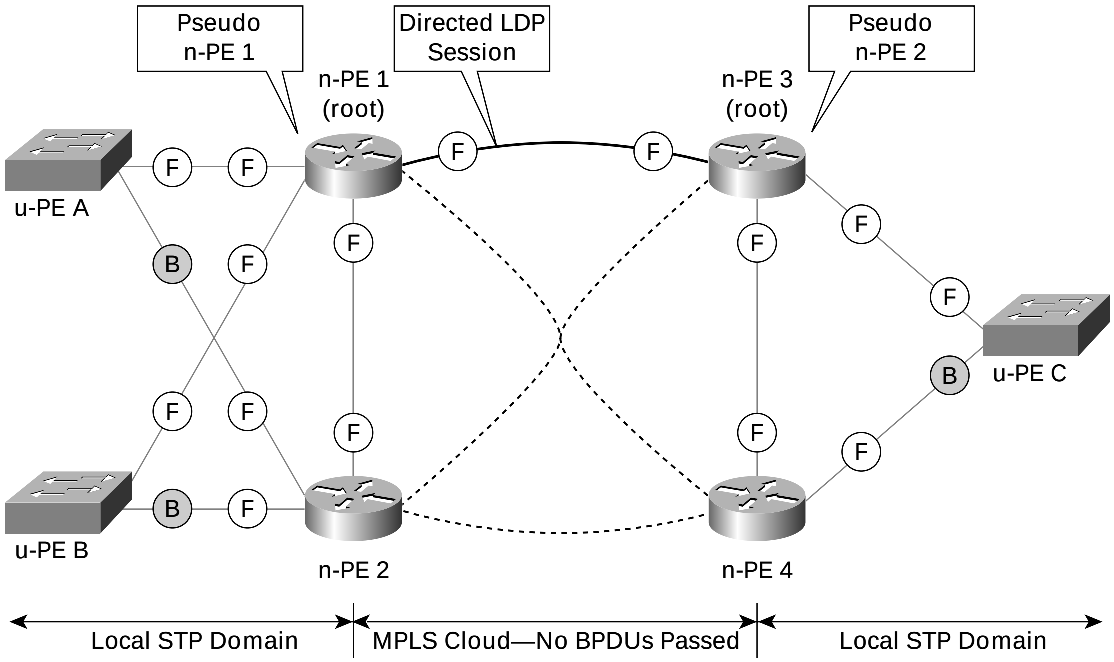
The drawback of this solution is that the inter-n-PE switch links must always be active and forwarding or a loop will occur via one of the u-PE edge devices. Although it is possible to reduce the probability that this may occur using EtherChannel that is split across different modules between the two n-PEs, the possibility that the inter-n-PE fiber fails needs to be ascertained as to the suitability of the solution.

A second solution that provides loop avoidance and redundancy, it does rely upon a specific configuration that may be compromised. A second solution that provides loop avoidance and redundancy is to subvert the MPLS full mesh of VCs by using routing policy and the concept of a pseudo-n-PE as depicted in Figure 20.

The pseudo-n-PE concept changes the way in which the MPLS network is viewed from an LDP perspective that dramatically simplifies the MPLS full-mesh requirements. This is achieved by configuring each redundant pair of n-PEs with the same IP address that is used for LDP peering purposes. In effect, each n-PE will be 'spoofed' into thinking that although a pair of remote n-PEs exist, from an LDP perspective only a single pseudo-n-PE exists.

Figure 20

EE H-VPLS Pseudo-n-PE Redundancy



A pseudo-n-PE configuration is achieved by configuring a loopback interface in each n-PE that is then configured with the same IP address for the pair of redundant n-PEs. To help ensure that the forwarding topology of the network is known, the choice of active and standby n-PE can be influenced by configuring a routing protocol metric against the loopback interface. The example depicted in Figure 21 details the specific configuration commands that may be used to influence the active LDP session partners and forwarding topology.

One particular issue that needs to be avoided is that routing protocols such as OSPF and internal BGP typically use the highest IP address of any interface within a chassis or the loopback's IP address as the protocol router ID. To avoid any routing protocol operational issues, it is good practice to configure a second loopback interface for pseudo-n-PE purposes and explicitly configure the OSPF and BGP router IDs.

Figure 21

EE H-VPLS Sample Pseudo-n-PE Configurations

```
n-PE-1#
!
mpls ldp router-id loopback 666
mpls label protocol ldp

interface loopback 0
 description Unique Router ID and management
Loopback
 ip address 192.168.255.1 255.255.255.255
!
interface loopback 666
 description pseudo-n-PE ID – ACTIVE PSEUDO-NODE
 ip address 192.168.255.254 255.255.255.255
 ip ospf cost 5000
!
router ospf 666
 router-id 192.168.255.1
 [text suppressed]
!
router bgp 666
 neighbor 192.168.255.2 remote-as 666
 neighbor 192.168.255.2 update-source loopback 0
 [text suppressed]
!

n-PE-2#
!
mpls ldp router-id loopback 666
mpls label protocol ldp

interface loopback 0
 description Unique Router ID and management
Loopback
 ip address 192.168.255.2 255.255.255.255
!
interface loopback 666
 description pseudo-n-PE ID – STANDBY PSEUDO-NODE
 ip address 192.168.255.254 255.255.255.255
 ip ospf cost 10000
!
router ospf 666
 router-id 192.168.255.2
 [text suppressed]
!
router bgp 666
 neighbor 192.168.255.1 remote-as 666
 neighbor 192.168.255.1 update-source loopback 0
 [text suppressed]
```

In operation, the routers that have been designated as the active pseudo-n-PEs will form an LDP session and exchange labels over this session. Consequently, a single PW will be established between the VSI in n-PE #1 and n-PE #3. As the edge Ethernet domain has redundant connectivity to each redundant n-PE, by establishing a single PW there are no loops introduced by the full mesh requirement of [VPLS-LDP] without compromising connectivity or redundancy between n-PEs.

Using Figure 20 as a reference, if a broadcast frame is transmitted from u-PE #1, as the link to n-PE #2 is blocked from a spanning tree perspective, n-PE #1 receives the broadcast frame and forwards to all ports associated with the VSI[21]. n-PE #2 will receive a copy of the frame on its Gigabit Ethernet trunk interface and will forward a copy to all ports associated with the VSI the frame was received upon. The broadcast frame will be received by u-PE #2 on the uplink to n-PE #1 only, as the link to n-PE #2 is blocked from a spanning tree perspective. It should be noted that n-PE#2 has no active PW associations as it does not have an LDP session and frames will be not be transmitted to, or received from, the MPLS network.

The broadcast frame will be transmitted using the negotiated PW to n-PE #3 whereupon the frame will be flooded to all ports associated with the VSI. n-PE #4 will receive a copy of the frame and flood a copy of the frame to all ports associated with the VSI the frame was received on. Again it should be noted that as n-PE #3 is the standby n-PE for a pseudo-n-PE pair, n-PE #3 would not have any PWs associated with the VSI. u-PE # 3 and u-PE #4 will receive a single copy of the original broadcast frame on the forwarding uplink attached to n-PE #3. It should be noted that the two spanning tree domains are entirely autonomous of each other and may have a different forwarding topology to that of other pseudo-n-PEs.

In the event of an uplink failure between a u-PE and an n-PE, the forwarding changes are purely local and do not affect the pseudo-n-PE topology. This is attractive to service providers as packets need not traverse expensive[22] MPLS modules for purely local connectivity requirements.

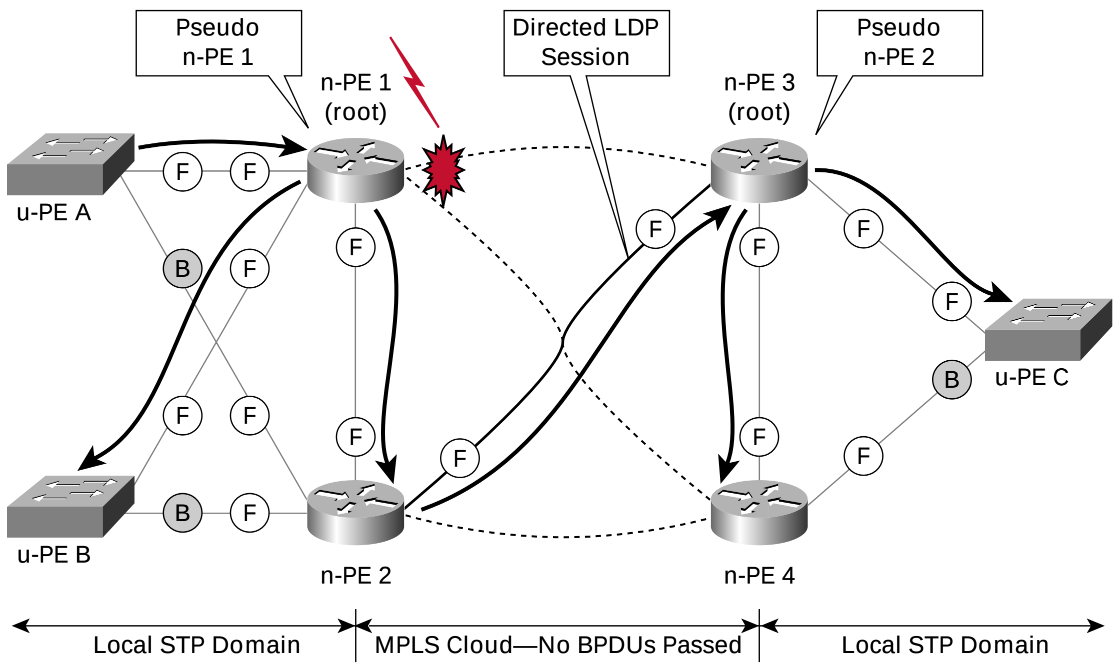
Using Figure 22 and Figure 23 as a reference, there are two distinct failures that may be envisaged: an MPLS interface failure and the failure of an n-PE. In the event that an MPLS capable module were to fail, the remote active n-PE within a pseudo-n-PE will simply think that the cost to reach the remote pseudo-n-PE has increased and the sessions will be reestablished with the standby n-PE.

21. It goes without saying that the source MAC address will be associated with ingress port within the MAC forwarding tables as per standard bridge self learning and this operation is omitted for clarity and brevity of explanation.

22. In terms of monetary cost as well as system resources.

Figure 22

EE H-VPLS Pseudo-n-PE—Interface Failure

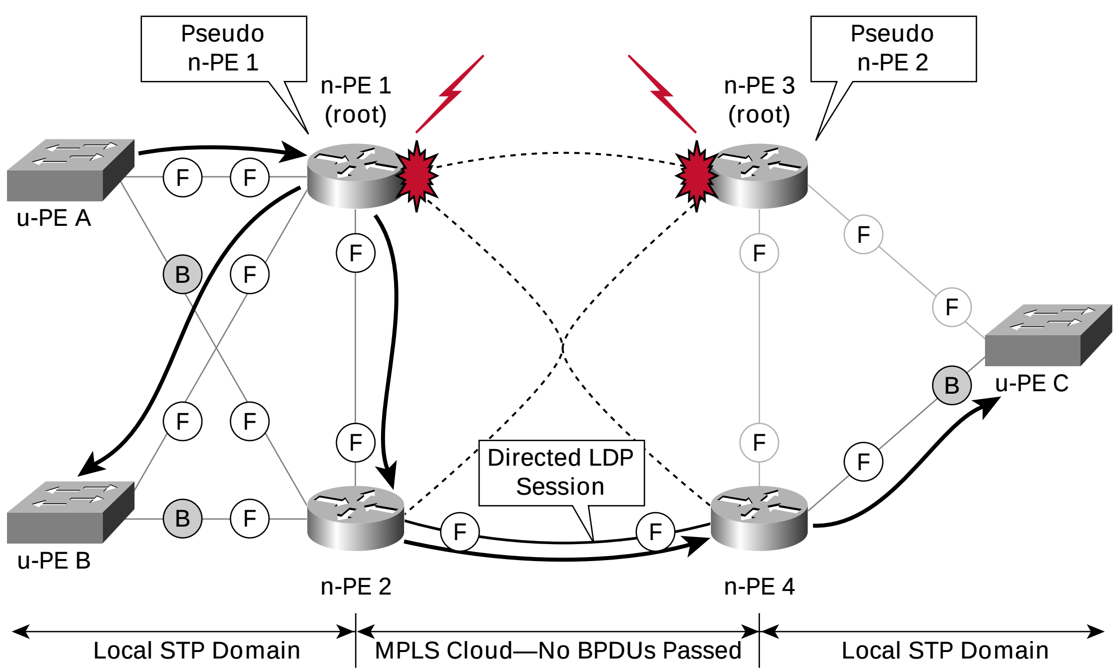


In the event of an entire n-PE failing, the remote pseudo-n-PE active n-PE will again think that the cost to the neighboring pseudo-n-PE has increased and the session will be reestablished to the standby pseudo-n-PE.

Figure 23

EE H-VPLS Pseudo-n-PE—Node Failure

.



It can be seen that the concept of a pseudo-n-PE very neatly, simply and logically addresses the redundancy and loop avoidance required within EE H-VPLS networks without extending spanning tree across the entire network. The drawback with pseudo-n-PE redundancy is that all LDP-based VPN connectivity will be forwarded through the active n-PE as the standby n-PE cannot be active at the same time. This is due to the operation of LDP and the use of a single IP address to identify a single pair of n-PEs. However, the pseudo-n-PE solution offers a simple and easy-to-understand solution to loop avoidance without the requirement to extend IEEE Spanning Tree across an Ethernet edge VPLS network.

[VPLS-LDP] Summary

The [VPLS-LDP] draft describes three distinct solutions within a single architectural model: one predicated upon MPLS end-to-end and the other on Ethernet as an edge domain with an MPLS core. Although the Ethernet edge model has some drawbacks in terms of redundancy, the economics of an Ethernet vs. MPLS solution are very different. The price point of an Ethernet switch is significantly less than a similar MPLS capable device and has a dramatic impact on the overall solutions costs. Additionally, any technical benefits of MPLS to the edge are largely outweighed by the additional complexity required to support MPLS to the edge.

Cisco engineering leaders are actively driving the [VPLS-LDP] draft within the IETF and were the first to describe a hierarchical model with an Ethernet edge and MPLS core. Cisco has also been instrumental in clarifying and correcting significant portions of [VPLS-LDP] with respect to the operation of IEEE 802.1 bridges. These areas include corrections to qualified and unqualified learning modes to reflect how IEEE 802.1Q bridges associate forwarding entries with filtering rules, referred to as Shared & Independent VLAN Learning within the IEEE 802.1Q standard.

Cisco also drove the definition of the PW and forwarding functions that emulate an Ethernet LAN from the perspective of an IEEE bridge, and were instrumental in ensuring that EE-H-VPLS described within [VPLS-LDP] is compatible with the proposed architecture and operation of IEEE 802.1ad bridges.

The IEEE 802.1ad Provider Bridges working group is a significant development in that it is addressing the requirements for Ethernet bridges to support service provider Metro Ethernet requirements. However, as the 802.1ad is still a work in progress, the details of Provider Bridge operation are outside of the scope of this document.

VPLS Summary

The IETF VPLS drafts have attracted considerable attention from service providers and enterprises alike, due to Ethernet's sympathetic operation with enterprise applications and the new services capabilities that Ethernet allows service providers to offer. Several drafts have been proposed to the IETF that deliver VPLS capabilities, although not all VPLS drafts are compatible with other VPLS drafts.

The VPLS draft that has most vendor support is *draft-ietf-l2vpn-vpls-ldp-00.txt* which has evolved from several earlier drafts that describe a hierarchical architecture for delivering VPLS services. The IEEE has also reacted to the considerable interest in VPLS, primarily due to the inconsistencies between VPLS bridge operation and IEEE bridge operation that in many cases make their operation mutually exclusive, by initiating the IEEE 802.1ad Provider Bridges working group.

Many other VPLS drafts have been proposed that rely upon proprietary mechanisms, or use mechanisms that are incompatible with IEEE 802.1 bridges. Although some of these mechanisms appear to be attractive at first glance, the actual operation needs careful evaluation, because the mechanisms used may not be as technically elegant or desirable as it first appears.

Cisco is also developing the features and functions necessary for service providers to deliver end-to-end Ethernet-based services. These features include the definition of an Ethernet User-to-Network Interface (E-UNI) and Ethernet LMI capabilities that allow the service provider to signal to the CE device service parameters such as up/down and bandwidth profiles. Cisco is actively driving the definition and mechanisms to provide end-to-end Ethernet OAM functions, such as Ethernet ping and Ethernet trace capabilities within ITU-T Study Group 13.

One area of considerable importance is Service Inter-Working between Ethernet and other transport technologies such as ATM and Frame Relay. This is critical to the success of Ethernet as a WAN protocol due to the large installed base of Frame Relay and ATM services and relatively limited footprint of Ethernet connectivity currently. Cisco has also published the IETF draft, *draft-sajassi-l2vpn-interworking-02.txt*, that describes what address resolution and frame conversions are necessary to provide Layer 2 Service Inter-Working.

Cisco is promoting changes within [VPLS-LDP] to make VPLS compatible with IEEE bridge operations and also driving changes to IEEE bridge operation to provide scalable VPLS services and interprovider operations. Cisco Systems, while actively driving standards-based VPLS solutions within the IETF and IEEE, is also actively driving standards-based Ethernet solutions for OAM within the ITU-T and Metro Ethernet Forum (MEF). Within the MEF Cisco is driving the adoption of Ethernet service definitions such as E-line and E-LAN services and defining key technologies such as Ethernet LMI, NNI, and OAM that will drive the adoption of Ethernet as a key WAN technology.

Cisco Systems is committed to delivering standards-based VPLS implementation based upon [VPLS-LDP] and the IEEE 802.1ad drafts. Cisco Systems is committed to delivering a broad range of solutions for Metro Ethernet Service delivery across a broad range of products including the industry-leading Cisco Catalyst® switch, Cisco router, and Cisco optical platforms that deliver Ethernet Layer 2 point-to-point and multipoint VPN capabilities as well as Ethernet access to Layer 3 VPN and Internet services on a single, consistent architectural construct. The Cisco commitment to standards is critical in the adoption of Ethernet as a WAN protocol and in delivering the capabilities that support next-generation enterprise applications.

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
Capital Tower
168 Robinson Road
#22-01 to #29-01
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

N2/MH/LW5340 1//04