

L2VPN Working Group
Internet Draft
Intended status: Informational
Expires: April 2012

Nabil Bitar
Verizon

Florin Balus
Marc Lasserre
Wim Henderickx
Alcatel-Lucent

Ali Sajassi
Luyuan Fang
Cisco

Yuichi Ikejiri
NTT Communications

Mircea Pisica
BT

October 31, 2011

Cloud Networking: Framework and VPN Applicability
draft-bitar-datacenter-vpn-applicability-01.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 31, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

Cloud Computing has been attracting a lot of attention from the networking industry. Some of the most publicized requirements are related to the evolution of the Cloud Networking Infrastructure to accommodate a large number of tenants, efficient network utilization, scalable loop avoidance, and Virtual Machine Mobility.

This draft describes a framework for cloud networking, highlighting the applicability of existing work in various IETF Working Groups (e.g., RFCs and drafts developed in IETF L2VPN and L3VPN Working Groups) to cloud networking, and the gaps and problems that need to be further addressed. That is, the goal is to understand what may be re-used from the current protocols and call out requirements specific to the Cloud space that need to be addressed by new standardization work with proposed solutions in certain cases.

Table of Contents

1. Introduction.....	3
2. General terminology.....	4
2.1. Conventions used in this document.....	5
3. Brief overview of Ethernet, L2VPN and L3VPN deployments.....	5
4. Cloud Networking Framework.....	6
5. DC problem statement.....	9
5.1. VLAN Space.....	9
5.2. MAC, IP, ARP Explosion.....	10
5.3. Per VLAN flood containment.....	11
5.4. Convergence and multipath support.....	12
5.5. Optimal traffic forwarding.....	12
5.6. Efficient multicast support.....	14
5.7. Connectivity to existing VPN sites.....	14
5.8. DC Inter-connect requirements.....	15
5.9. L3 virtualization considerations.....	15

5.10. VM Mobility requirements.....	15
6. L2VPN Applicability to Cloud Networking.....	16
6.1. VLANs and L2VPN toolset.....	16
6.2. PBB and L2VPN toolset.....	17
6.2.1. Addressing VLAN space exhaustion and MAC explosion..	18
6.2.2. Fast convergence, L2 multi-pathing.....	19
6.2.3. Per ISID flood containment.....	20
6.2.4. Efficient multicast support.....	20
6.2.5. Tunneling options for PBB ELAN: Ethernet, IP, MPLS..	20
6.2.6. Use Case examples.....	20
6.2.6.1. PBBN in DC, L2 VPN in DC GW.....	20
6.2.6.2. PBBN in VSw, L2VPN in the ToR.....	22
6.2.7. Connectivity to existing VPN sites and Internet.....	23
6.2.8. DC Interconnect.....	25
6.2.9. Interoperating with existing DC VLANs.....	25
6.3. TRILL and L2VPN toolset.....	27
7. L3VPN applicability to Cloud Networking.....	28
8. Solutions for other DC challenges.....	29
8.1. Addressing IP/ARP explosion.....	29
8.2. Optimal traffic forwarding.....	29
8.3. VM Mobility.....	29
9. Security Considerations.....	30
10. IANA Considerations.....	30
11. References.....	30
11.1. Normative References.....	30
11.2. Informative References.....	31
12. Acknowledgments.....	32

1. Introduction

The initial Data Center (DC) networks were built to address the needs of individual enterprises and/or individual applications. Ethernet VLANs and regular IP routing are used to provide connectivity between compute, storage resources and the related customer sites.

The virtualization of compute resources in a DC environment provides the foundation for selling compute and storage resources to multiple customers, or selling application services to multiple customers. For example, a customer may buy a group of Virtual Machines (VMs) that may reside on server blades distributed throughout a DC or across DCs. In this latter case, the DCs may be owned and operated by a cloud service provider connected to one or more network service providers, two or more cloud service providers each connected to one or more network service providers, or a hybrid of DCs operated by the customer and the cloud service provider(s). In addition, multiple

customers may be assigned resources on the same compute and storage hardware.

In order to provide access for multiple customers to the virtualized compute and storage resources, the DC network and DC interconnect have to evolve from the basic VLAN and IP routing architecture to provide equivalent connectivity virtualization at a large scale.

This document describes in separate sections existing DC networking architecture, challenges faced by existing DC network models, and the applicability of VPN technologies to address such challenges. In addition, challenges not addressed by existing solutions are called out to describe the problem or to suggest solutions.

2. General terminology

Some general terminology is defined here; most of the terminology used is from [802.1ah] and [RFC4026]. Terminology specific to this memo is introduced as needed in later sections.

DC: Data Center

ELAN: MEF ELAN, multipoint to multipoint Ethernet service

EVPN: Ethernet VPN as defined in [EVPN]

PBB: Provider Backbone Bridging, new Ethernet encapsulation designed to address VLAN exhaustion and MAC explosion issues; specified in IEEE 802.1ah [802.1ah]

PBB-EVPN: defines how EVPN can be used to transport PBB frames

BMAC: Backbone MACs, the backbone source or destination MAC address fields defined in the 802.1ah provider MAC encapsulation header.

CMAC: Customer MACs, the customer source or destination MAC address fields defined in the 802.1ah customer MAC encapsulation header.

BEB: A backbone edge bridge positioned at the edge of a provider backbone bridged network. It is usually the point in the network where PBB encapsulation is added or removed from the frame.

BCB: A backbone core bridge positioned in the core of a provider backbone bridged network. It performs regular Ethernet switching using the outer Ethernet header.

2.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Brief overview of Ethernet, L2VPN and L3VPN deployments

Initial Ethernet networks have been deployed in LAN environments, where the total number of hosts (hence MAC addresses) to manage was limited. Physical Ethernet topologies in LANs were pretty simple. Hence, a simple loop resolution protocol such as the Spanning Tree Protocol was sufficient in the early days. Efficient utilisation of physical links was not a major concern in LANs, while at the same time leveraging existing and mature technologies.

As more hosts got connected to a LAN, or the need arose to create multiple LANs on the same physical infrastructure, it became necessary to partition the physical topology into multiple Virtual LANs (VLANs). STP evolved to cope with multiple VLANs with Multiple-STP (MSTP). Bridges/Switches evolved to learn behind which VLAN specific MACs resided, a process known as qualified learning. As Ethernet LANs moved into the provider space, the 12-bit VLAN space limitation (i.e. a total of 4k VLANs) led to Q-in-Q and later to Provider backbone Bridging (PBB).

With PBB, not only can over 16M virtual LAN instances (24-bit Service I-SID) be supported, but a clean separation between customer and provider domains has been defined with separate MAC address spaces (Customer-MACs (CMACs) versus Provider Backbone-MACs (BMACs)). CMACs are only learned at the edge of the PBB network on PBB Backbone Edge Bridges (BEBs) in the context of an I-component while only B-MACs are learnt by PBB Backbone Core Bridges (BCBs). This results in BEB switches creating MAC-in-MAC tunnels to carry customer traffic, thereby hiding C-MACs in the core.

In the meantime, interconnecting L2 domains across geographical areas has become a necessity. VPN technologies have been defined to carry both L2 and L3 traffic across IP/MPLS core networks. The same technologies could also be used within the same data center to provide for scale or for interconnecting services across L3 domains, as needed. Virtual Private LAN Service (VPLS) has been playing a key

role to provide transparent LAN services over IP/MPLS WANs while IP VPNs, including BGP/MPLS IP VPNs and IPsec VPNs, have been used to provide virtual IP routing instances over a common IP/MPLS core network.

All these technologies have been combined to maximize their respective benefits. At the edge of the network, such as in access networks, VLAN and PBB are commonly used technologies. Aggregation networks typically use VPLS or BGP/MPLS IP VPNs to groom traffic on a common IP/MPLS core.

It should be noted that Ethernet has kept evolving because of its attractive features, specifically its auto-discovery capabilities and the ability of hosts to physically relocate on the same LAN without requiring renumbering. In addition, Ethernet switches have become commodity, creating a financial incentive for interconnecting hosts in the same community with Ethernet switches. The network layer (layer3), on the other hand, has become pre-dominantly IP. Thus, communication across LANs uses IP routing.

4. Cloud Networking Framework

A generic architecture for Cloud Networking is depicted in Figure 1:

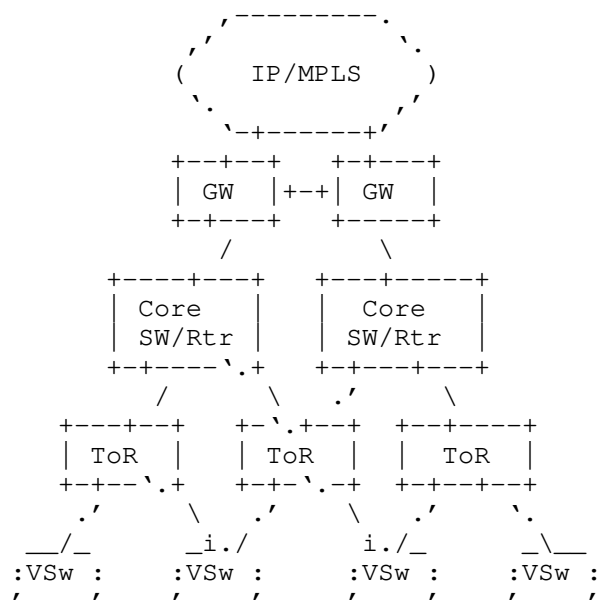


Figure 1 : A Generic Architecture for Cloud Networking

A cloud network is composed of intra-Data Center (DC) networks and network services, and inter-DC network connectivity. DCs may belong to a cloud service provider connected to one or more network service providers, different cloud service providers each connected to one or more network service providers, or a hybrid of DCs operated by the enterprise customers and the cloud service provider(s). It may also provide access to the public and/or enterprise customers.

The following network components are present in a DC:

- VSw or virtual switch - software based Ethernet switch running inside the server blades. VSw may be single or dual-homed to the Top of Rack switches (ToRs). The individual VMs appear to a VSw as IP hosts connected via logical interfaces. The VSw may evolve to support IP routing functionality.
- ToR or Top of Rack - hardware-based Ethernet switch aggregating all Ethernet links from the server blades in a rack representing the entry point in the physical DC network for the hosts. ToRs may also perform routing functionality. ToRs are usually dual-homed to the Core SW. Other deployment scenarios

may use an EoR (End of Row) switch to provide similar function as a ToR.

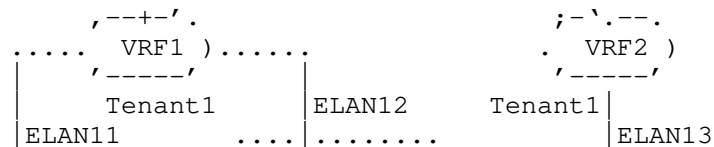
- Core SW (switch) - high capacity core node aggregating multiple ToRs. This is usually a cost effective Ethernet switch. Core switches can also support routing capabilities.
- DC GW - gateway to the outside world providing DC Interconnect and connectivity to Internet and VPN customers. In the current DC network model, this may be a Router with Virtual Routing capabilities and/or an IPVPN/L2VPN PE.

A DC network also contains other network services, such as firewalls, load-balancers, IPsec gateways, and SSL acceleration gateways. These network services are not currently discussed in this draft as the focus is on the routing and switching services. The usual DC deployment employs VLANs to isolate different VM groups throughout the Ethernet switching network within a DC. The VM Groups are mapped to VLANs in the VSws. The ToRs and Core SWs may employ VLAN trunking to eliminate provisioning touches in the DC network. In some scenarios, IP routing is extended down to the ToRs, and may be further extended to the hypervisor.

Any new DC and cloud networking technology needs to be able to fit as seamlessly as possible with this existing DC model, at least in a non-greenfield environment. In particular, it should be possible to introduce enhancements to various tiers in this model in a phased approach without disrupting the other elements.

Depending upon the scale, DC distribution, operations model, Capex and Opex aspects, DC switching elements can act as strict L2 switches and/or provide IP routing capabilities, including VPN routing and/or MPLS support. In smaller DCs, it is likely that some tier layers will be collapsed, and that Internet connectivity, inter-DC connectivity and VPN support will be handled by Core Nodes which perform the DC GW role.

The DC network architecture described in this section can be used to provide generic L2-L3 service connectivity to each tenant as depicted in Figure 2:



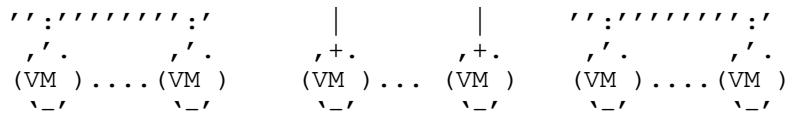


Figure 2 : Logical Service connectivity for one tenant

In this example one or more virtual routing contexts distributed on multiple DC GWs and one or more ELANs (e.g., one per Application) running on DC switches are assigned for DC tenant 1. ELAN is a generic term for Ethernet multipoint service, which in the current DC environment is implemented using 12-bit VLAN tags. Other possible ELAN technologies are discussed in section 6.

For a multi-tenant DC, this type of service connectivity or a variation could be used for each tenant. In some cases only L2 connectivity is required, i.e., only an ELAN may be used to interconnect VMs and customer sites.

5. DC problem statement

This section summarizes the challenges faced with the present mode of operation described in the previous section and implicitly describes the requirements for next generation DC network.

With the introduction of Compute virtualization, the DC network must support multiple customers or tenants that need access to their respective computing and storage resources in addition to making some aspect of the service available to other businesses in a B-to-B model or to the public. Every tenant requires service connectivity to its own resources with secure separation from other tenant domains. Connectivity needs to support various deployment models, including interconnecting customer-hosted data center resources to cloud service provider hosted resources (Virtualized DC for the customer). This connectivity may be at layer2 or layer3.

Currently, large DCs are often built on a service architecture where VLANs configured in Ethernet edge and core switches are interconnected by IP routing running in a few centralized routers. There may be some cases though where IP routing might be used in the core nodes or even in the TORs inside a DC.

5.1. VLAN Space

Existing DC deployments provide customer separation and flood containment, including support for DC infrastructure

interconnectivity, using Ethernet VLANs. A 12-bit VLAN tag provides support for a maximum of 4K VLANs.

4K VLANs are inadequate for a Cloud Provider looking to expand its customer base. For example, there are a number of VPN deployments (VPLS and IP VPN) which serve more than 20K customers, each requiring multiple VLANs. Thus, 4K VLANs will likely support less than 4K customers.

The cloud networking infrastructure needs to provide support for a much bigger number of virtual L2 domains.

5.2. MAC, IP, ARP Explosion

Virtual Machines are the basic compute blocks being sold to Cloud customers. Every server blade supports today 16-40 VMs with 100 or more VMs per server blade coming in the near future. Every VM may have multiple interfaces for provider and enterprise management, VM mobility and tenant access, each with its own MAC and IP addresses. For a sizable DC, this may translate into millions of VM IP and MAC addresses. From a cloud network viewpoint, this scale number will be an order of magnitude higher.

Supporting this amount of IP and MAC addresses, including the associated dynamic behavior (e.g., ARP), throughout the DC Ethernet switches and routers is very challenging in an Ethernet VLAN and regular routing environment. Core Ethernet switches running Ethernet VLANs learn the MAC addresses for every single VM interface that sends traffic through that switch. Throwing memory to increase the MAC Forwarding DataBase (FDB) size affects the cost of these switches. In addition, as the number of MACs that switches need to learn increases, convergence time could increase, and flooding activity will increase upon a topology change as the core switches flush and re-learn the MAC addresses. Simple operational mistakes may lead to duplicate MAC entries within the same VLAN domain and security issues due to administrative MAC assignment used today for VM interfaces. Similar concerns about memory requirements and related cost apply to DC Edge switches (ToRs/EoRs) and DC GWs.

From a router perspective, it is important to maximize the utilization of available resources in both control and data planes through flexible mapping of VMs and related VLANs to routing interfaces. This is not easily done in the current VLAN based deployment environment where the use of VLAN trunking limits the allocation of VMs to only local routers.

The amount of ARP traffic grows linearly with the number of hosts on a LAN. For 1 million VM hosts, it can be expected that the amount of ARP traffic will be in the range of half million ARPs per second at the peak, which corresponds to over 200 Mbps of ARP traffic [MYERS]. Similarly, on a server, the amount of ARP traffic, grows linearly with the number of virtual L2 domains/ELANs instantiated on that server and the number of VMs in that domain. Besides the link capacity wasted, which may be small compared to the link capacities deployed in DCs, the computational burden may be prohibitive. In a large-DC environment, the large number of hosts and the distribution of ARP traffic may lead to a number of challenges:

- . Processing overload and overload of ARP entries on the Server/Hypervisor. This is caused by the increased number of VMs per server blade and the size of related ELAN domains. For example, a server blade with 100 VMs, each in a separate L2 domain with 100 VMs each would need to support 10K ARP entries and the associated ARP processing while performing the other compute tasks.
- . Processing overload and exhaustion of ARP entries on the Routers/PEs and any other L3 Service Appliances (Firewall (FW), Load-Balancer (LB) etc). This issue is magnified by the L3 virtualization at the service gateways. For example, a gateway PE handling 10K ELANs each with 10 VMs will result in 100K hosts sending/receiving traffic to/from the PE, thus requiring the PE to learn 100K ARP entries. It should be noted that if the PE supports Integrated Routing and Bridging (IRB), it must support the associated virtual IP RIBs/FIBs and MAC FDBs for these hosts in addition to the ARP entries.
- . Flood explosion throughout Ethernet switching network. This is caused by the use of VLAN trunking and implicitly by the lack of per VPN flood containment.

DC and DC-interconnect technologies that minimize the negative impact of ARP, MAC and IP entry explosion on individual network elements in a DC or cloud network hierarchy are needed.

5.3. Per VLAN flood containment

From an operational perspective, DC operators try to minimize the provisioning touches required for configuring a VLAN domain by employing VLAN trunks on the L2 switches. This comes at the cost of flooding broadcast, multicast and unknown unicast frames outside of the boundaries of the actual VLAN domain.

The cloud networking infrastructure needs to prevent unnecessary traffic from being sent/leaked to undesired locations.

5.4. Convergence and multipath support

Spanning Tree is used in the current DC environment for loop avoidance in the Ethernet switching domain.

STP can take 30 to 50 seconds to repair a topology. Practical experience shows that Rapid STP (RSTP) can also take multiple seconds to converge, such as when the root bridge fails.

STP eliminates loops by disabling ports. The result is that only one path is used to carry traffic. The capacity of disabled links cannot be utilized, leading to inefficient use of resources.

In a small DC deployment, multi-chassis LAG (MC-LAG) support may be sufficient initially to provide for loop-free redundancy as an STP alternative. However, in medium or large DCs it is challenging to use MC-LAGs solely across the network to provide for resiliency and loop-free paths without introducing a layer2 routing protocol: i.e. for multi-homing of server blades to ToRs, ToRs to Core SWs, Core SWs to DC GWs. MC-LAG may work as a local mechanism but it has no knowledge of the end-to-end paths so it does not provide any degree of traffic steering across the network.

Efficient and mature link-state protocols, such as IS-IS, provide rapid failover times, can compute optimal paths and can fully utilize multiple parallel paths to forward traffic between 2 nodes in the network.

Unlike OSPF, IS-IS runs directly at L2 (i.e. no reliance on IP) and does not require any configuration. Therefore, IS-IS based DC networks are to be favored over STP-based networks. IEEE Shortest Path Bridging (SPB) based on IEEE 802.1aq and IEEE 802.1Qbp, and IETF TRILL [RFC6325] are technologies that enable Layer2 networks using IS-IS for Layer2 routing.

5.5. Optimal traffic forwarding

Optimal traffic forwarding requires (1) efficient utilization of all available link capacity in a DC and DC-interconnect, and (2) traffic forwarding on the shortest path between any two communicating VMs within the DC or across DCs.

Optimizing traffic forwarding between any VM pair in the same virtual domain is dependent on (1) the placement of these VMs and their relative proximity from a network viewpoint, and (2) the technology used for computing the routing/switching path between these VMs. The latter is especially important in the context of VMotion, moving a VM from one network location to another, while maintaining its layer2 and Layer3 addresses.

Ethernet-based forwarding between two VMs relies on the MAC-destination Address that is unique per VM interface in the context of a virtual domain. In traditional IEEE technologies (e.g., 802.1ad, 802.1ah) and IETF L2VPN (i.e., VPLS), Ethernet MAC reachability is always learnt in the data plane. That applies to both B-MACs and C-MACs. IETF EVPN [EVPN] supports C-MAC learning in the control plane via BGP. In addition, with newer IEEE technologies (802.1aq and 802.1Qbp) and IETF PBB-EVPN [PBB-EVPN], B-MAC reachability is learnt in the control plane while C-MACs are learnt in the data plane at BEBs, and tunneled in PBB frames. In all these cases, it is important that as a VM is moved from one location to another: (1) VM MAC reachability convergence happens fast to minimize traffic black-holing, and (2) forwarding takes the shortest path.

IP-based forwarding relies on the destination IP address. ECMP load balancing relies on flow-based criteria. An IP host address is unique per VM interface. However, hosts on a LAN share a subnet mask, and IP routing entries are based on that subnet address. Thus, when VMs are on the same LAN and traditional forwarding takes place, these VMs forward traffic to each other by relying on ARP or IPv6 Neighbor discovery to identify the MAC address of the destination and on the underlying layer2 network to deliver the resulting MAC frame to its destination. However, when VMs, as IP hosts across layer2 virtual domains, need to communicate they rely on the underlying IP routing infrastructure.

In addition, when a DC is an all-IP DC, VMs are assigned a host address with /32 subnet in the IPv4 case, or /64 or /128 host address in the IPv6 case, and rely on the IP routing infrastructure to route the IP packets among VMs. In this latter case, there is really no need for layer2 awareness potentially beyond the hypervisor switch at the server hosting the VM. In either case, when a VM moves location from one physical router to another while maintaining its IP identity (address), the underlying IP network must be able to route the traffic to the destination and must be able to do that on the shortest path.

Thus, in the case of IP address aggregation as in a subnet, optimality in traffic forwarding to a VM will require reachability to the VM host address rather than only the subnet. That is what is often referred to as punching a hole in the aggregate at the expense of routing and forwarding table size increase.

As in layer2, layer3 may capitalize on hierarchical tunneling to optimize the routing/FIB resource utilization at different places in the network. If a hybrid of subnet-based routing and host-based routing (host-based routing here is used to refer to hole-punching in the aggregate) is used, then during VMotion, routing transition can take place, and traffic may be routed to a location based on subnet reachability or to a location where the VM used to be attached. In either of these cases, traffic must not be black-holed. It must be directed potentially via tunneling to the location where the VM is. This requires that the old routing gateway knows where the VM is currently attached. How to obtain that information can be based on different techniques with tradeoffs. However, this traffic triangulation is not optimal and must only exist in the transition until the network converges to a shortest path to the destination.

5.6. Efficient multicast support

STP bridges typically perform IGMP and/or PIM snooping in order to optimize multicast data delivery. However, this snooping is performed locally by each bridge following the STP topology where all the traffic goes through the root bridge. This may result in sub-optimal multicast traffic delivery. In addition, each customer multicast group is associated with a forwarding tree throughout the Ethernet switching network. Solutions must provide for efficient Layer2 multicast. In an all-IP network, explicit multicast trees in the DC network can be built via multicast signaling protocols (e.g., PIM-SSM) that follows the shortest path between the destinations and source(s). In an IPVPN context, Multicast IPVPN based on [MVPN] can be used to build multicast trees shared among IPVPNs, specific to VPNS, and/or shared among multicast groups across IPVPNs.

5.7. Connectivity to existing VPN sites

It is expected that cloud services will have to span larger geographical areas in the near future and that existing VPN customers will require access to VM and storage facilities for virtualized data center applications. Hence, the DC network virtualization must interoperate with deployed and evolving VPN solutions - e.g. IP VPN, VPLS, VPWS, PBB-VPLS, E-VPN and PBB-EVPN.

5.8. DC Inter-connect requirements

Cloud computing requirements such as VM Mobility across DCs, Management connectivity, and support for East-West traffic between customer applications located in different DCs imply that inter-DC connectivity must be supported. These DCs can be part of a hybrid cloud operated by the cloud service provider(s) and/or the end-customers.

Mature VPN technologies can be used to provide L2/L3 DC interconnect among VLANs/virtual domains located in different DCs.

5.9. L3 virtualization considerations

In order to provide customer L3 separation while supporting overlapping IP addressing and privacy, a number of schemes were implemented in the DC environment. Some of these schemes, such as double NATing are operationally complex and prone to operator errors. Virtual Routing contexts (or Virtual Device contexts) or dedicated hardware-routers are positioned in the DC environment as an alternative to these mechanisms. Every customer is assigned a dedicated routing context with associated control plane protocols. For instance, every customer gets an IP Forwarding instance controlled by its own BGP and/or IGP routing. Assigning virtual or hardware routers to each customer while supporting thousands of customers in a DC is neither scalable nor cost-efficient.

5.10. VM Mobility requirements

The ability to move VMs within a resource pool, whether it is a local move within the same DC to another server or to a distant DC, offers multiple advantages for a number of scenarios, for example:

- In the event of a possible natural disaster, moving VMs to a safe DC location decreases downtime and allows for meeting SLA requirements.
- Optimized resource location: VMs can be moved to locations that offer significant cost reduction (e.g. power savings), or locations close to the application users. They can also be moved to simply load-balance across different locations.

When VMs change location, it is often important to maintain the existing client sessions. The VM MAC and IP addresses must be preserved, and the state of the VM sessions must be copied to the new location.

Current VM mobility tools like VMware VMotion require L2 connectivity among the hypervisors on the servers participating in a VMotion pool. This is in addition to "tenant ELAN" connectivity which provides for communication between the VM and the client(s).

A VMotion ELAN might need to cross multiple DC networks to provide the required protection or load-balancing. In addition, in the current VMotion procedure, the new VM location must be part of the tenant ELAN domain. When the new VM is activated, a Gratuitous ARP is sent so that the MAC FIB entries in the "tenant ELAN" are updated to direct traffic destined to that VM to the new VM location. In addition, if a portion of the path requires IP forwarding, the VM reachability information must be updated to direct the traffic on the shortest path to the VM.

VM mobility requirements may be addressed through the use of Inter-DC VLANs to address VMotion and tenant ELANs. However expanding "tenant VLANs" across two or more DCs will accelerate VLAN exhaustion and MAC explosion issues. In addition, STP needs to run across DCs leading to increased convergence times and the blocking of expensive WAN bandwidth. VLAN trunking used throughout the network creates indiscriminate flooding across DCs.

L2 VPN solutions over IP/MPLS are designed to interconnect sites located across the WAN.

6. L2VPN Applicability to Cloud Networking

The following sections will discuss different solution alternatives, re-using IEEE and IETF technologies to provide a gradual migration path from the current Ethernet switching VLAN-based model to more advanced Ethernet switching and IP/MPLS based models. This evolution is targeted to address inter-DC requirements, cost considerations and the efficient use of processing/memory resources on DC networking components.

6.1. VLANs and L2VPN toolset

One approach to address some of the DC challenges discussed in the previous section is to gradually deploy additional technologies within existing DC networks. For example, an operator may start by breaking its DC VLAN domains into different VLAN islands so that each island can support up to 4K VLANs. VLAN Domains can then be interconnected via VPLS using the DC GW as a VPLS PE. An ELAN service can be identified with one VLAN ID in one island and another VLAN ID in another island with the appropriate VLAN ID processed at the GW.

As the number of tenants in individual VLAN islands surpasses 4K, the operator could push VPLS deployment deeper in the DC network. It is possible in the end to retain existing VLAN-based solution only in VSw and to provide L2VPN support starting at the ToRs. The ToR and DC core elements need to be MPLS enabled with existing VPLS solutions.

However, this model does not solve the MAC explosion issue as ToRs still need to learn VM MAC addresses. In addition, it requires management of both VLAN and L2VPN addressing and mapping of service profiles. Per VLAN, per port and per VPLS configurations are required at the ToR, increasing the time it takes to bring up service connectivity and complicating the operational model.

6.2. PBB and L2VPN toolset

As highlighted in the problem statement section, the expected large number of VM MAC addresses in the DC calls out for a VM MAC hiding solution so that the ToRs and the Core Switches only need to handle a limited number of MAC addresses.

PBB IEEE 802.1ah encapsulation is a standard L2 technique developed by IEEE to achieve this goal. It was designed also to address other limitations of VLAN-based encapsulations while maintaining the native Ethernet operational model deployed in the DC network.

A conceptual PBB encapsulation is described in Figure 3 (for detailed encapsulation see [802.1ah]):

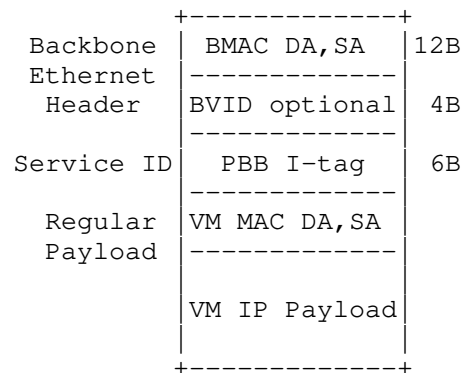


Figure 3 PBB encapsulation

The original Ethernet packet used in this example for Inter-VM communication is encapsulated in the following PBB header:

- I-tag field - organized similarly with the 802.1q VLAN tag; it includes the Ethertype, PCP and DEI bits and a 24 bit ISID tag

which replaces the 12 bit VLAN tag, extending the number of virtual L2 domain support to 16 Million. It should be noted that the PBB I-Tag includes also some reserved bits, and most importantly the C-MAC DA and SA. What is designated as 6 bytes in the figure is the I-tag information excluding the C-MAC DA and SA.

- An optional Backbone VLAN field (BVLAN) may be used if grouping of tenant domains is desired.
- An outer Backbone MAC header contains the source and destination MAC addresses for the related server blades, assuming the PBB encapsulation is done at the hypervisor virtual switch on the server blade.
- The total resulting PBB overhead added to the VM-originated Ethernet frame is 18 or 22 Bytes (depending on whether the BVID is excluded or not)
- Note that the original PBB encapsulation allows the use of CVLAN and SVLAN in between the VM MACs and IP Payload. These fields were removed from Figure 3 since in a VM environment these fields do not need to be used on the VSw, their function is relegated to the I-SID tag.

6.2.1. Addressing VLAN space exhaustion and MAC explosion

In a DC environment, PBB maintains traditional Ethernet forwarding plane and operational model. For example, a VSw implementation of PBB can make use of the 24 bit ISID tag instead of the 12 bit VLAN tag to identify the virtual bridging domains associated with different VM groups. The VSw uplink towards the ToR in Figure 1 can still be treated as an Ethernet backbone interface. A frame originated by a VM can be encapsulated with the ISID assigned to the VM VSw interface and with the outer DA and SA MACs associated with the respective destination and source server blades, and then sent to the ToR switch. Performing this encapsulation at the VSw distributes the VM MAC learning to server blades with instances in the corresponding layer2 domain, and therefore alleviates this load from ToRs that aggregate multiple server blades. Alternatively, the PBB encapsulation can be done at the ToR.

With PBB encapsulation, ToRs and Core SWs do not have to handle VM MAC addresses so the size of their MAC FIB tables may decrease by two or more orders of magnitude, depending on the number of VMs

configured in each server blade and the number of VM virtual interfaces and associated MACs.

The original PBB specification [802.1ah] did not introduce any new control plane or new forwarding concepts for the PBB core. Spanning Tree and regular Ethernet switching based on MAC Learning and Flooding were maintained to provide a smooth technology introduction in existing Ethernet networks.

6.2.2. Fast convergence and L2 multi-pathing

Additional specification work for PBB control plane has been done since then in both IEEE and IETF L2VPN.

As stated earlier, STP-based layer2 networks underutilize the available network capacity as links are put in an idle state to prevent loops. Similarly, existing VPLS technology for interconnecting Layer2 network-islands over an IP/MPLS core does not support active-active dual homing scenarios.

IS-IS controlled layer2 networks allow traffic to flow on multiple parallel paths between any two servers, spreading traffic among available links on the path. IEEE 802.1aq Shortest Path Bridging (SPB) [802.1aq] and emerging IEEE 802.1Qbp [802.1Qbp] are PBB control plane technologies that utilize different methods to compute parallel paths and forward traffic in order to maximize the utilization of available links in a DC. In addition, a BGP based solution [PBB-EVPN] was submitted and discussed in IETF L2VPN WG.

One or both mechanisms may be employed as required. IS-IS could be used inside the same administrative domain (e.g., a DC), while BGP may be employed to provide reachability among interconnected Autonomous Systems. Similar architectural models have been widely deployed in the Internet and for large VPN deployments.

IS-IS and/or BGP are also used to advertise Backbone MAC addresses and to eliminate B-MAC learning and unknown unicast flooding in the forwarding plane, albeit with tradeoffs. The BMAC FIB entries are populated as required from the resulting IS-IS or BGP RIBs.

Legacy loop avoidance schemes using Spanning Tree and local Active/Active MC-LAG are no longer required as their function (layer2 routing) is replaced by the indicated routing protocols (IS-IS and BGP).

6.2.3. Per ISID flood containment

Service auto-discovery provided by 802.1aq SPB [802.1aq] and BGP [PBB-EVPN] is used to distribute ISID related information among DC nodes, eliminating any provisioning touches throughout the PBB infrastructure. This implicitly creates backbone distribution trees that provide per ISID automatic flood and multicast containment.

6.2.4. Efficient multicast support

IS-IS [802.1aq] and BGP [PBB-EVPN] could be used to build optimal multicast distribution trees. In addition, PBB and IP/MPLS tunnel hierarchy may be used to aggregate multiple customer multicast trees sharing the same nodes by associating them with the same backbone forwarding tree that may be represented by a common Group BMAC and optionally a P2MP LSP. More details will be discussed in a further version of the draft.

6.2.5. Tunneling options for PBB ELAN: Ethernet, IP and MPLS

The previous section introduces a solution for DC ELAN domains based on PBB ISIDs, PBB encapsulation and IS-IS and/or BGP control plane.

IETF L2 VPN specifications [PBB-VPLS] or [PBB-EVPN] enable the transport of PBB frames using PW/MPLS or simply MPLS, and implicitly allow the use of MPLS Traffic Engineering and Resiliency toolset to provide for advanced traffic steering and faster convergence.

Transport over IP/L2TPv3 [RFC 4719] or IP/GRE is also possible as an alternative to MPLS tunneling. Additional header optimization for PBB over IP/GRE encapsulated packets may also be feasible. These specifications would allow for ISID based L2 overlay using a regular IP backbone.

6.2.6. Use Case examples

6.2.6.1. PBBN in DC, L2VPN in DC GW

DC environments based on VLANs and native Ethernet operational model may want to consider using the native PBB option to provide L2 multi-tenancy, in effect the DC ELAN from Figure 2. An example of a network architecture that addresses this scenario is depicted in Figure 4:

, '-----'.
, ' Inter-DC '.
(L2VPN (PBB-VPLS))

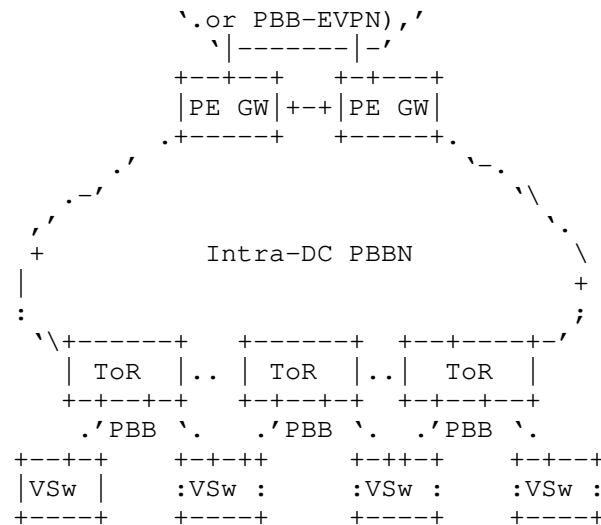


Figure 4 PBB in DC, PBB-VPLS or PBB-EVPN for DC Interconnect

PBB inside the DC core interoperates seamlessly with VPLS used for L2 DC-Interconnect to extend ELAN domains across DCs. This expansion may be required to address VM Mobility requirements or to balance the load on DC PE gateways. Note that in PBB-VPLS case, just one or a handful of infrastructure B-VPLS instances are required, providing Backbone VLAN equivalent function.

PBB encapsulation addresses the expansion of the ELAN service identification space with 16M ISIDs and solves MAC explosion through VM MAC hiding from the Ethernet core.

PBB SPB [802.1aq] is used for core routing in the ToRs, Core SWs and PEs. If the DCs that need to be interconnected at L2 are part of the same administrative domain, and scaling is not an issue, SPB/IS-IS may be extended across the VPLS infrastructure. If different AS domains are present, better load balancing is required between the DCs and the WAN, or IS-IS extension across DCs causes scaling issues, then BGP extensions described in [PBB-EVPN] must be employed.

The forwarding plane, MAC FIB requirements and the Layer2 operational model in the ToR and Core SW are maintained. The VSw sends PBB encapsulated frames to the ToR as described in the previous section. ToRs and Core SWs still perform standard Ethernet switching using the outer Ethernet header.

From a control plane perspective, VSw uses a default gateway configuration to send traffic to the ToR, as in regular IP routing case. VSw BMAC learning on the ToR is done through either LLDP or VM Discovery Protocol (VDP) described in [802.1Qbg]. Identical mechanisms may be used for the ISID. Once this information is learned on the ToR it is automatically advertised through SPB. If PBB-EVPN is used in the DC GWs, MultiProtocol (MP)-BGP will be used to advertise the ISID and BMAC over the WAN as described in [PBB-EVPN].

6.2.6.2. PBBN in VSw, L2VPN in the ToR

A variation of the use case example from the previous section is depicted in Figure 5:

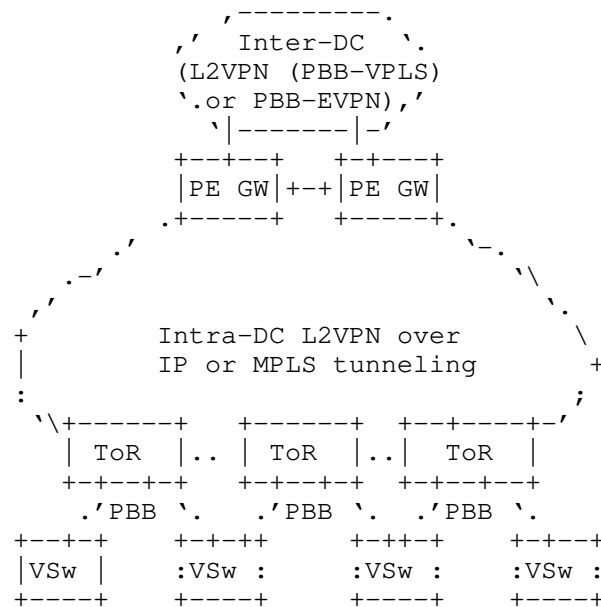


Figure 5 PBB in VSw, L2VPN at the ToR

The procedures from the previous section are used at the VSw: PBB encapsulation and Ethernet BVLANS can be used on the VSw uplink. L2VPN infrastructure is replacing the BVLAN at the ToR enabling the use of IP (GRE or L2TP) or MPLS tunneling.

L2 networking still has the same control plane choices: IS-IS [802.1aq] and/or BGP [PBB-EVPN], independently from the tunneling choice.

6.2.7. Connectivity to existing VPN sites and Internet

The main reason for extending the ELAN space beyond the 4K VLANs is to be able to serve multiple DC tenants whereby the total number of service domains needed exceeds 4K. Figure 6 represents the logical service view where PBB ELANs are used inside one or multiple DCs to connect to existing IP VPN sites. It should be noted that the PE GW should be able to perform integrated routing in a VPN context and bridging in VSI context:

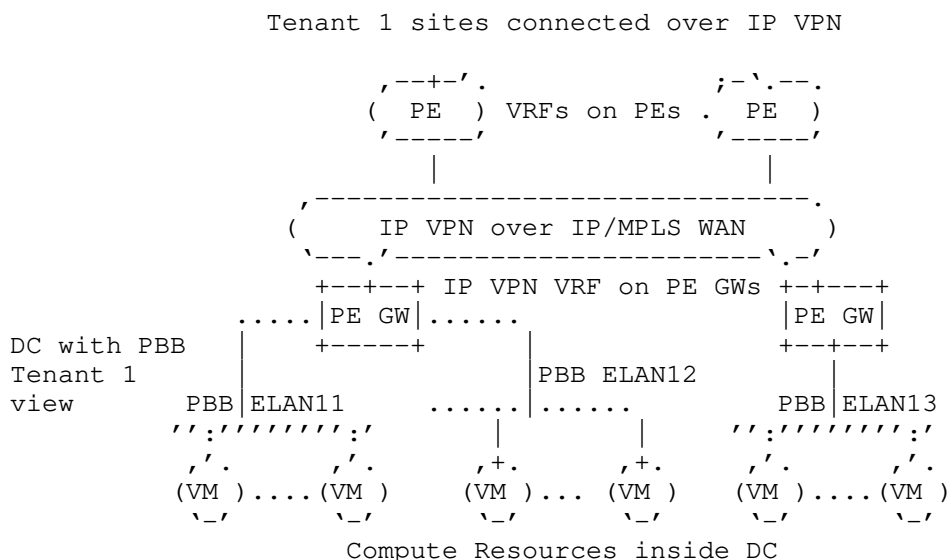


Figure 6 Logical Service View with IP VPN

DC ELANs are identified with 24-bit ISIDs instead of VLANs. At the PE GWs, an IP VPN VRF is configured for every DC tenant. Each "ISID ELAN" for Tenant 1 is seen as a logical Ethernet endpoint and is assigned an IP interface on the Tenant 1 VRF. Tenant 1 enterprise sites are connected to IP VPN PEs distributed across the WAN. IP VPN instances on PE GWs can be automatically discovered and connected to the WAN IP VPN using standard procedures - see [RFC4364].

In certain cases, the DC GW PEs are part of the IPVPN service provider network providing IPVPN services to the enterprise customers. In other cases, DC PEs are operated and managed by the DC/cloud provider and interconnect to multiple IPVPN service providers using inter-AS BGP/MPLS models A, B, or C [RFC4364]. The

same discussion applies to the case of IPSec VPNs from a PBB ELAN termination perspective.

If tenant sites are connected to the DC using WAN VPLS, the PE GWs need to implement the BEB function described in the PBB-VPLS PE model [PBB-VPLS] and the procedures from [PBB-Interop] to perform the required translation. Figure 7 describes the VPLS WAN scenario:

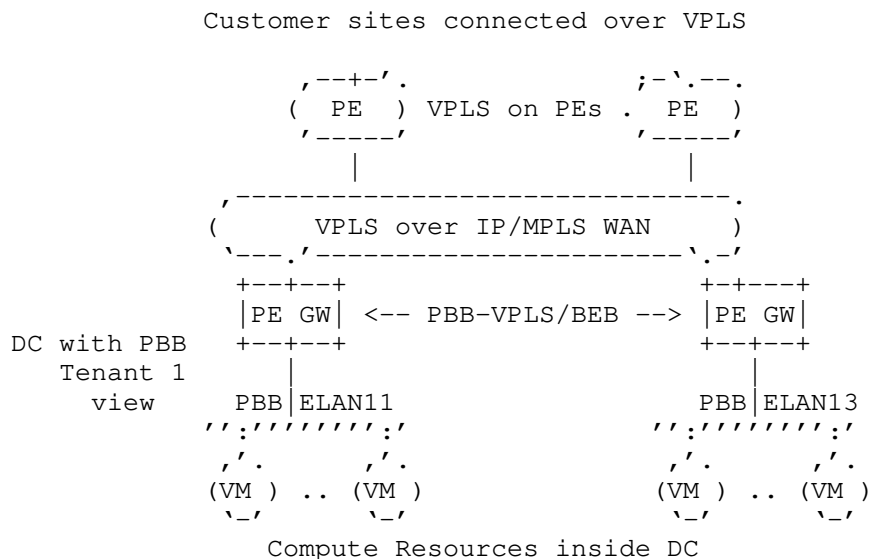


Figure 7 Logical Service View with VPLS WAN

One VSI is required at the PE GW for every DC ELAN domain. Same as in the IP VPN case, DC PE GWs may be fully integrated as part of the WAN provider network or using Inter-AS/Inter-provider models A,B or C.

The VPN connectivity may be provided by one or multiple PE GWs, depending on capacity need and/or the operational model used by the DC/cloud operator.

If a VM group is serving Internet connected customers, the related ISID ELAN will be terminated into a routing context (global public instance or another VRF) connected to the Internet. Same as in the IP VPN case, the 24bit ISID will be represented as a logical Ethernet endpoint on the Internet routing context and an IP interface will be allocated to it. Same PE GW may be used to provide both VPN and Internet connectivity with the routing contexts separated internally using the IP VPN models.

6.2.8. DC Interconnect

L2 DC interconnect may be required to expand the ELAN domains for Management, VM Mobility or when a VM Group needs to be distributed across DCs.

PBB may be used to provide ELAN extension across multiple DCs as depicted in Figure 8:

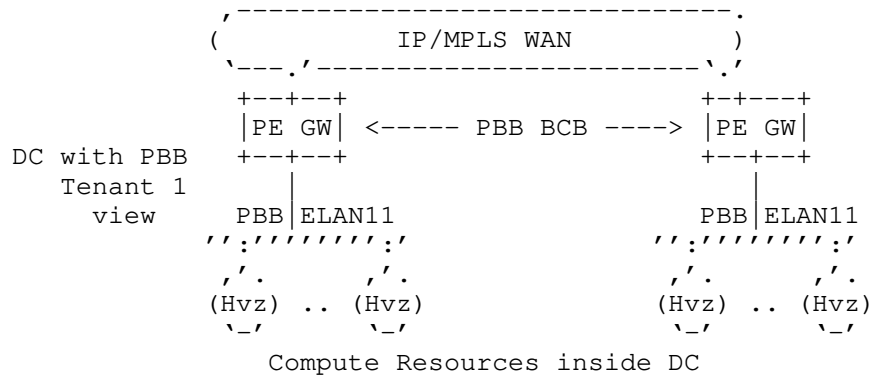
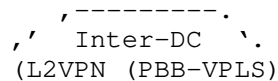


Figure 8 PBB BCB providing VMotion ELAN

ELAN11 is expanded across DC to provide interconnect for the pool of server blades assigned to the same VMotion domain. This time Hypervisors are connected directly to ELAN11. The PE GW operates in this case as a PBB Backbone Core Bridge (BCB) [PBB-VPLS] combined with PBB-EVPN capabilities [PBB-EVPN]. The ISID ELANs do not require any additional provisioning touches and do not consume additional MPLS resources on the PE GWs. Per ISID auto-discovery and flood containment is provided by IS-IS/SPB [802.1aq] and BGP [PBB-EVPN].

6.2.9. Interoperating with existing DC VLANs

While green field deployments will definitely benefit from all the advantages described in the previous sections, in many other scenarios, existing DC VLAN environments will have to be gradually migrated to the new architecture. Figure 9 depicts an example of a possible migration scenario where both PBB and VLAN technologies are present:



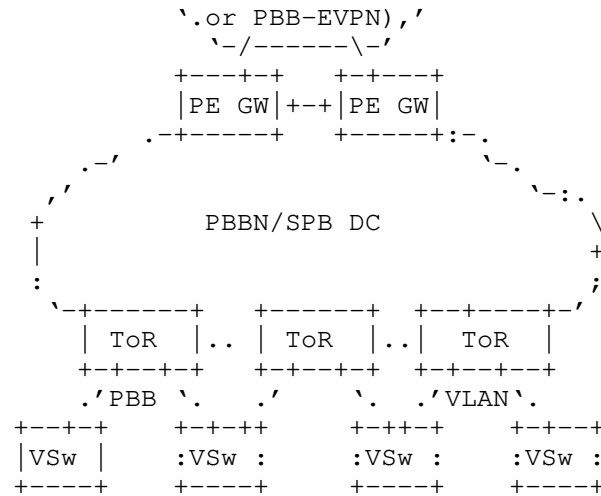


Figure 9 DC with PBB and VLANs

This example assumes that the two VSWs on the right do not support PBB but the ToRs do. The VSw on the left side are running PBB while the ones on the right side are still using VLANs. The left ToR is performing only Ethernet switching whereas the one on the right is translating from VLANs to ISIDs and performing PBB encapsulation using the BEB function [802.1ah] and [PBB-VPLS]. The ToR in the middle is performing both functions: core Ethernet tunneling for the PBB VSw and BEB function for the VLAN VSw.

The SPB control plane is still used between the ToRs, providing the benefits described in the previous section. The VLAN VSw must use regular multi-homing functions to the ToRs: for example STP or Multi-chassis-LAG.

DC VLANs may be also present initially on some of the legacy ToRs or Core SWs. PBB interoperability will be performed as follows:

- . If VLANs are used in the ToRs, PBB BEB function may be performed by the Core SW(s) where the ToR uplink is connected
- . If VLANs are used in the Core SW, PBB BEB function may be performed by the PE GWs where the Core SW uplink is connected

It is possible that some DCs may run PBB or PBB-VLAN combination while others may still be running VLANs. An example of this interoperability scenario is described in Figure 10:

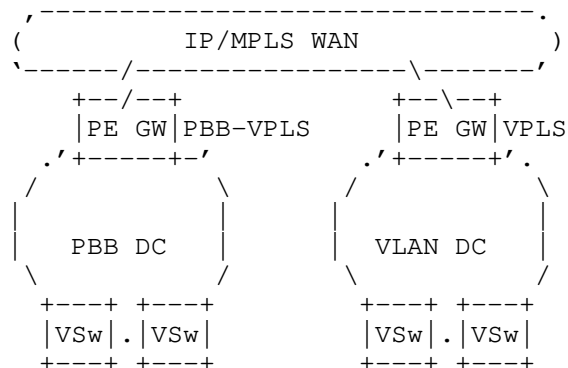


Figure 10 Interoperability to a VLAN-based DC

Interoperability with existing VLAN DC is required for DC interconnect. The PE-GW in the PBB DC or the PE GW in the VLAN DC must implement PBB-VPLS PE model described in [PBB-VPLS]. This interoperability scenario is addressed in detail in [PBB-Interop].

Connectivity to existing VPN customer sites (IP VPN, VPLS, IPSec) or Internet does not require any additional procedures beyond the ones described in the VPN connectivity section. The PE GW in the DC VLAN will aggregate DC ELANs through IP interfaces assigned to VLAN logical endpoints whereas the PE GW in the PBB DC will assign IP interfaces to ISID logical endpoints.

If EVPN is used to interconnect the two DCs, PBB-EVPN functions described in [PBB-EVPN] must be implemented in one of the PE-GWs.

6.3. TRILL and L2VPN toolset

TRILL and SPB control planes provide similar functions. IS-IS is the base protocol used in both specifications to provide multi-pathing and fast convergence for core networking. [PBB-EVPN] describes how seamless Inter-DC connectivity can be provided over an MPLS/IP network for both TRILL [RFC6325] and SPB [802.1aq]/[802.1Qbp] networks.

The main differences exist in the encapsulation and data plane forwarding. TRILL encapsulation [RFC6325] was designed initially for large enterprise and campus networks where 4k VLANs are sufficient. As a consequence the ELAN space in [RFC6325] is limited to 4K VLANs; however, this VLAN scale issue is being addressed in [Fine-Grained].

7. L3VPN applicability to Cloud Networking

This section discusses the role of IP VPN technology in addressing the L3 Virtualization challenges described in section 5.

IP VPN technology defined in L3VPN working group may be used to provide L3 virtualization in support of multi-tenancy in the DC network as depicted in Figure 11.

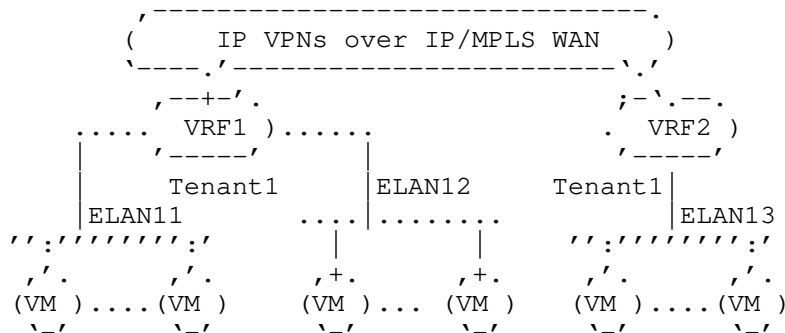


Figure 11 Logical Service View with IP VPN

Tenant 1 might buy Cloud Services in different DC locations and choose to associate the VMs in 3 different groups, each mapped to a different ELAN: ELAN11, ELAN12 and ELAN13. L3 interconnect between the ELANs belonging to tenant1 is provided using an IP/MPLS VPN and associated VRF1 and VRF2, possibly located in different DCs. Each tenant that requires L3 virtualization will be allocated a different IP VPN instance. Using full fledged IP VPN for L3 Virtualization inside a DC presents the following advantages compared with existing DC technologies like Virtual Routing:

- Interoperates with existing WAN VPN technology
- Deployment tested, provides a full networking toolset
- Scalable core routing - only one BGP-MP routing instance is required compared with one per customer/tenant in the Virtual Routing case
- Service Auto-discovery - automatic discovery and route distribution between related service instances
- Well defined and deployed Inter-Provider/Inter-AS models

- Supports a variety of VRF-to-VRF tunneling options accommodating different operational models: MPLS [RFC4364], IP or GRE [RFC4797]

To provide Cloud services to related customer IP VPN instances located in the WAN the following connectivity models may be employed:

- DC IP VPN instance may participate directly in the WAN IP VPN
- Inter-AS Options A, B or C models may be employed with applicability to both Intra and Inter-Provider use cases [RFC4364]

8. Solutions for other DC challenges

This section touches on some of the DC challenges that may be addressed by a combination of IP VPN, L2VPN and IP toolset. Additional details will be provided in a future revision.

8.1. Addressing IP/ARP explosion

Possible solutions for IP/ARP explosion are discussed in [EVPN], [PBB-EVPN], [ARPproxy] and in ARMD WG that address certain aspects. More discussion is required to clarify the requirements in this space, taking into account the different network elements potentially impacted by ARP.

8.2. Optimal traffic forwarding

IP networks, built using links-state protocols such as OSPF or ISIS and BGP provide optimal traffic forwarding through the use of equal cost multiple path (ECMP) and ECMP traffic load-balancing, and the use of traffic engineering tools based on BGP and/or MPLS-TE as applicable. In the Layer2 case, SPB or TRILL based protocols provide for load-balancing across parallel paths or equal cost paths between two nodes. Traffic follows the shortest path. For multicast, data plane replication at layer2 or layer3 happens in the data plane albeit with different attributes after multicast trees are built via a control plane and/or snooping. In the presence of VM mobility, optimal forwarding relates to avoiding triangulation and providing for optimum forwarding between any two VMs.

8.3. VM Mobility

IP VPN technology may be used to support DC Interconnect for different functions like VM Mobility and Cloud Management. A

description of VM Mobility between server blades located in different IP subnets using extensions to existing BGP-MP and IP VPN procedure is described in [VM-Mobility]. Other solutions can exist as well. What is needed is a solution that provides for fast convergence toward the steady state whereby communication among any two VMs can take place on the shortest path or most optimum path, transit triangulation time is minimized, traffic black-holing is avoided, and impact on routing scale for both IPv4 and IPv6 is controllable or minimized.

9. Security Considerations

No new security issues are introduced beyond those described already in the related L2VPN drafts.

10. IANA Considerations

IANA does not need to take any action for this draft.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4761] Kompella, K. and Rekhter, Y. (Editors), "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and Kompella, V. (Editors), "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [PBB-VPLS] Balus, F. et al. "Extensions to VPLS PE model for Provider Backbone Bridging", draft-ietf-l2vpn-pbb-vpls-pe-model-04.txt (work in progress), October 2011.
- [PBB-Interop] Sajassi, A. et al. "VPLS Interoperability with Provider Backbone Bridging", draft-ietf-l2vpn-pbb-vpls-interop-02.txt (work in progress), July 2011.
- [802.1ah] IEEE 802.1ah "Virtual Bridged Local Area Networks, Amendment 6: Provider Backbone Bridges", Approved Standard June 12th, 2008

- [802.1aq] IEEE Draft P802.1aq/D4.3 "Virtual Bridged Local Area Networks, Amendment: Shortest Path Bridging", Work in Progress, September 21, 2011
- [RFC6325] Perlman, et al., "Routing Bridges (Rbridges): Base Protocol Specification", RFC 6325, July 2011.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4797] Rosen, E. and Y. Rekhter, " Use of Provider Edge to Provider Edge (PE-PE) Generic Routing encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks ", RFC 4797, January 2007.

11.2. Informative References

- [RFC4026] Andersson, L. et Al., "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, May 2005.
- [802.1Qbp] IEEE Draft P802.1Qbp/D0.1 "Virtual Bridged Local Area Networks, Amendment: Equal Cost Multiple Paths (ECMP)", Work in Progress, October 13, 2011
- [802.1Qbg] IEEE Draft P802.1Qbg/D1.8 "Virtual Bridged Local Area Networks, Amendment: Edge Virtual Bridging", Work in Progress, October 17, 2011
- [EVPN] Raggarwa, R. et al. "BGP MPLS based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-04.txt (work in progress), September 2011.
- [PBB-EVPN] Sajassi, A. et al. "PBB-EVPN", draft-sajassi-l2vpn-pbb-evpn-02.txt (work in progress), July 2011.
- [VM-Mobility] Raggarwa, R. et al. "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-01.txt (work in progress), September 2011.
- [RFC4719] Aggarwal, R. et al., "Transport of Ethernet over Layer 2 Tunneling Protocol Version 3 (L2TPv3)", RFC 4719, November 2006.

- [MVPN] Rosen, E. and Raggarwa, R. "Multicast in MPLS/BGP IP VPN", draft-ietf-l3vpn-2547bis-mcast-10.txt (work in progress), January 2010.
- [ARProxy] Carl-Mitchell, S. and Quarterman, S., "Using ARP to implement transparent subnet gateways", RFC 1027, October 1987.
- [MYERS] Myers, A., Ng, E. and Zhang, H., "Rethinking the Service Model: Scaling Ethernet to a Million Nodes" <http://www.cs.cmu.edu/~acm/papers/myers-hotnetsIII.pdf>
- [Fine-Grained] Eastlake, D. et Al., "RBridges: Fine-Grained Labeling", draft-eastlake-trill-rbridge-fine-labeling-01.txt (work in progress), October 2011.

12. Acknowledgments

In addition to the authors the following people have contributed to this document:

Javier Benitez, Colt

Dimitrios Stiliadis, Alcatel-Lucent

Samer Salam, Cisco

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
Email: nabil.bitar@verizon.com

Florin Balus
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA, USA 94043
Email: florin.balus@alcatel-lucent.com

Marc Lasserre
Alcatel-Lucent
Email: marc.lasserre@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, USA
Email: sajassi@cisco.com

Luyuan Fang
Cisco
111 Wood Avenue South
Iselin, NJ 08830
Email: lufang@cisco.com

Yuichi Ikejiri
NTT Communications
1-1-6, Uchisaiwai-cho, Chiyoda-ku
Tokyo, 100-8019 Japan
Email: y.ikejiri@ntt.com

Mircea Pisica
BT
Telecomlaan 9
Brussels 1831, Belgium
Email: mircea.pisica@bt.com

Network Working Group
Internet-Draft
Updates: 4761 (if approved)
Intended status: Standards Track
Expires: February 6, 2012

B. Kothari
Cisco Systems
K. Kompella
Juniper Networks
W. Henderickx
F. Balus
Alcatel-Lucent
J. Uttaro
AT&T
July 6, 2011

BGP based Multi-homing in Virtual Private LAN Service
draft-ietf-l2vpn-vpls-multihoming-03.txt

Abstract

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for the Service Provider (SP) to give the customer redundant connectivity to some sites, often called "multi-homing". This memo shows how BGP-based multi-homing can be offered in the context of LDP and BGP VPLS solutions.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 6, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
1.1. General Terminology	4
1.2. Conventions	4
2. Background	6
2.1. Scenarios	6
2.2. VPLS Multi-homing Considerations	7
3. Multi-homing Operation	8
3.1. Provisioning Model	8
3.2. Multi-homing NLRI	8
3.3. Designated Forwarder Election	9
3.3.1. Attributes	9
3.3.2. Variables Used	9
3.3.3. Election Procedures	11
3.4. DF Election on PEs	13
4. Multi-AS VPLS	14
4.1. Route Origin Extended Community	14
4.2. VPLS Preference	14
4.3. Use of BGP-MH attributes in Inter-AS Methods	15
4.3.1. Inter-AS Method (b): EBGW Redistribution of VPLS Information between ASBRs	15
4.3.2. Inter-AS Method (c): Multi-Hop EBGW Redistribution of VPLS Information between ASes	16
5. MAC Flush Operations	18
5.1. MAC List Flush	18
5.2. Implicit MAC Flush	18
5.3. Minimizing the effects of fast link transitions	20
6. Backwards Compatibility	21
6.1. BGP based VPLS	21
6.2. LDP VPLS with BGP Auto-discovery	21
7. Security Considerations	22
8. IANA Considerations	23
9. Acknowledgments	24
10. References	25
10.1. Normative References	25
10.2. Informative References	25
Authors' Addresses	27

1. Introduction

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for a Service Provider (SP) to give the customer redundant connectivity to one or more sites, often called "multi-homing". [RFC4761] explains how VPLS can be offered using BGP for auto-discovery and signaling; section 3.5 of that document describes how multi-homing can be achieved in this context. [RFC6074] explains how VPLS can be offered using BGP for auto-discovery, (BGP-AD) and [RFC4762] explains how VPLS can be offered using LDP for signaling. This document provides a BGP-based multi-homing solution applicable to both BGP and LDP VPLS technologies. Note that BGP MH can be used for LDP VPLS without the use of the BGP-AD solution.

Section 2 lays out some of the scenarios for multi-homing, other ways that this can be achieved, and some of the expectations of BGP-based multi-homing. Section 3 defines the components of BGP-based multi-homing, and the procedures required to achieve this. Section 7 may someday discuss security considerations.

1.1. General Terminology

Some general terminology is defined here; most is from [RFC4761], [RFC4762] or [RFC4364]. Terminology specific to this memo is introduced as needed in later sections.

A "Customer Edge" (CE) device, typically located on customer premises, connects to a "Provider Edge" (PE) device, which is owned and operated by the SP. A "Provider" (P) device is also owned and operated by the SP, but has no direct customer connections. A "VPLS Edge" (VE) device is a PE that offers VPLS services.

A VPLS domain represents a bridging domain per customer. A Route Target community as described in [RFC4360] is typically used to identify all the PE routers participating in a particular VPLS domain. A VPLS site is a grouping of ports on a PE that belong to the same VPLS domain. A Multi-homed (MH) site is uniquely identified by a MH site ID (MH-ID). Sites are referred to as local or remote depending on whether they are configured on the PE router in context or on one of the remote PE routers (network peers).

1.2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this

document are to be interpreted as described in [RFC2119].

2. Background

This section describes various scenarios where multi-homing may be required, and the implications thereof. It also describes some of the singular properties of VPLS multi-homing, and what that means from both an operational point of view and an implementation point of view. There are other approaches for providing multi-homing such as Spanning Tree Protocol, and this document specifies use of BGP for multi-homing. Comprehensive comparison among the approaches is outside the scope of this document.

2.1. Scenarios

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity.

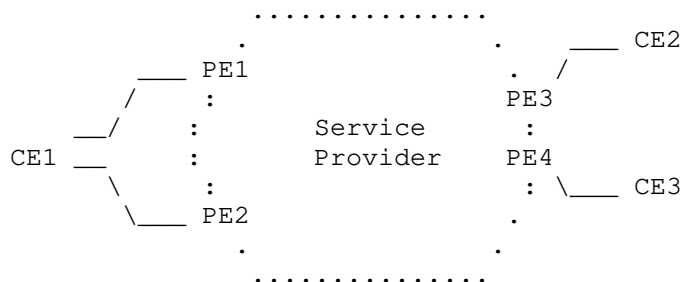


Figure 1: Scenario 1

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity. However, CE4, which is also in the same VPLS domain, is single-homed to just PE1.

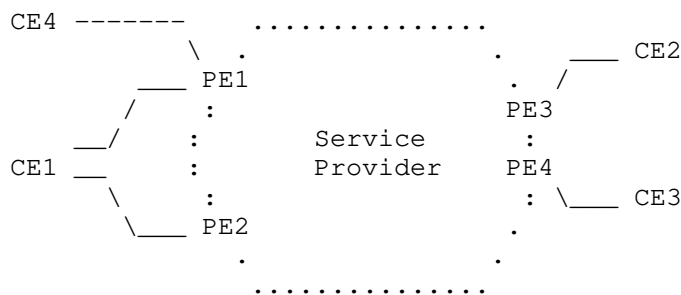


Figure 2: Scenario 2

2.2. VPLS Multi-homing Considerations

The first (perhaps obvious) fact about a multi-homed VPLS CE, such as CE1 in Figure 1 is that if CE1 is an Ethernet switch or bridge, a loop has been created in the customer VPLS. This is a dangerous situation for an Ethernet network, and the loop must be broken. Even if CE1 is a router, it will get duplicates every time a packet is flooded, which is clearly undesirable.

The next is that (unlike the case of IP-based multi-homing) only one of PE1 and PE2 can be actively sending traffic, either towards CE1 or into the SP cloud. That is to say, load balancing techniques will not work. All other PEs MUST choose the same designated forwarder for a multi-homed site. Call the PE that is chosen to send traffic to/from CE1 the "designated forwarder".

In Figure 2, CE1 and CE4 must be dealt with independently, since CE1 is dual-homed, but CE4 is not.

3. Multi-homing Operation

This section describes procedures for electing a designated forwarder among the set of PEs that are multi-homed to a customer site. The procedures described in this section are applicable to BGP based VPLS, LDP based VPLS with BGP-AD or a VPLS that contains a mix of both BGP and LDP signaled PWs.

3.1. Provisioning Model

Figure 1 shows a customer site, CE1, multi-homed to two VPLS PEs, PE1 and PE2. In order for all VPLS PEs within the same VPLS domain to elect one of the multi-homed PEs as the designated forwarder, an indicator that the PEs are multi-homed to the same customer site is required. This is achieved by assigning the same multi-homed site ID (MH-ID) on PE1 and PE2 for CE1. When remote VPLS PEs receive NLRI advertisement from PE1 and PE2 for CE1, the two NLRI advertisements for CE1 are identified as candidates for designated forwarder selection due to the same MH-ID. Thus, same MH-ID SHOULD be assigned on all VPLS PEs that are multi-homed to the same customer site.

Note that a MH-ID=0 is invalid and a PE should discard such an advertisement.

3.2. Multi-homing NLRI

Section 3.2.2 in [RFC4761] describes the encoding of the BGP VPLS NLRI. This NLRI contains fields VE-ID, VE block offset, VE block size and label base. For multi-homing operation, the same NLRI is used for identifying the multi-homed customers sites. The VE-ID field in the NLRI is set to MH-ID; the VE block offset, VE block size and label base are set to zero. Thus, the NLRI contains 2 octets indicating the length, 8 octets for Route Distinguisher, 2 octets for MH-ID and 7 octets with value zero.

Figure 2 shows two customer sites, CE1 and CE4, connected to PE1 with CE1 multi-homed to PE1 and PE2. CE4 does not require special addressing, being associated with the base VPLS instance identified by the VSI-ID for LDP VPLS and VE-ID for BGP VPLS. However, CE1 which is multi-homed to PE1 and PE2 requires configuration of MH-ID and both PE1 and PE2 MUST be provisioned with the same MH-ID for CE1.

It is valid to have non-zero VE block offset, VE block size and label base in the VPLS NLRI for a multi-homed site. However, multi-homing operations in such a case are outside the scope of this document.

3.3. Designated Forwarder Election

BGP-based multi-homing for VPLS relies on BGP DF election and VPLS DF election. The net result of doing both BGP and VPLS DF election is that of electing a single designated forwarder (DF) among the set of PEs to which a customer site is multi-homed. All the PEs that are elected as non-designated forwarders MUST keep their attachment circuit to the multi-homed CE in blocked status (no forwarding).

These election algorithms operate on VPLS advertisements, which include both the NLRI and attached BGP attributes. In order to simplify the explanation of these algorithms, we will use a number of variables derived from fields in the VPLS advertisement. These variables are: RD, MH-ID, VBO, DOM, ACS, PREF and PE-ID. The notation ADV -> <RD, MH-ID, VBO, DOM, ACS, PREF, PE-ID> means that from a received VPLS advertisement ADV, the respective variables were derived. The following sections describe two attributes needed for DF election, then describe the variables and how they are derived from fields in VPLS advertisement ADV, and finally describe how DF election is done.

3.3.1. Attributes

The procedures below refer to two attributes: the Route Origin community (see Section 4.1) and the L2-info community (see Section 4.2). These attributes are required for inter-AS operation; for generality, the procedures below show how they are to be used. The procedures also say how to handle the case that either or both are not present.

3.3.2. Variables Used

3.3.2.1. RD

RD is simply set to the Route Distinguisher field in the NLRI part of ADV.

3.3.2.2. MH-ID

MH-ID is simply set to the VE-ID field in the NLRI part of ADV.

3.3.2.3. VBO

VBO is simply set to the VE Block Offset field in the NLRI part of ADV. This field will typically be zero.

3.3.2.4. DOM

This variable, indicating the VPLS domain to which ADV belongs, is derived by applying BGP policy to the Route Target extended communities in ADV. The details of how this is done are outside the scope of this document.

3.3.2.5. ACS

ACS is the status of the attachment circuits for a given site of a VPLS. ACS = 1 if all attachment circuits for the site are down, and 0 otherwise.

For BGP-based Multi-homing, ADV MUST contain an L2-info extended community; within this community are control flags. One of these flags is the 'D' bit, described in [I-D.kothari-l2vpn-auto-site-id]. ACS is set to the value of the 'D' bit in ADV.

3.3.2.6. PREF

PREF is derived from the Local Preference (LP) attribute in ADV as well as the VPLS Preference field (VP) in the L2-info extended community. If the Local Preference attribute is missing, LP is set to 0; if the L2-info community is missing, VP is set to 0. The following table shows how PREF is computed from LP and VP.

VP Value	LP Value	PREF Value	Comment
0	0	0	malformed advertisement, unless ACS=1
0	1 to $(2^{16}-1)$	LP	backwards compatibility
0	2^{16} to $(2^{32}-1)$	$(2^{16}-1)$	backwards compatibility
>0	LP same as VP	VP	Implementation supports VP
>0	LP != VP	0	malformed advertisement

Table 1

3.3.2.7. PE-ID

If ADV contains a Route Origin (RO) community (see Section 4.1) with type 0x01, then PE-ID is set to the Global Administrator sub-field of the RO. Otherwise, if ADV has an ORIGINATOR_ID attribute, then PE-ID is set to the ORIGINATOR_ID. Otherwise, PE-ID is set to the BGP Identifier.

3.3.3. Election Procedures

The election procedures described in this section apply equally to BGP VPLS and LDP VPLS.

Election occurs in two stages. The first stage divides all received VPLS advertisements into buckets of relevant and comparable advertisements. Distinction MUST NOT be made on whether the NLRI is a multi-homing NLRI or not. In this stage, advertisements may be discarded as not being relevant to DF election. The second stage picks a single "winner" from each bucket by repeatedly applying a tie-breaking algorithm on a pair of advertisements from that bucket. The tie-breaking rules are such that the order in which advertisements are picked from the bucket does not affect the final result. Note that this is a conceptual description of the process; an implementation MAY choose to realize this differently as long as the semantics are preserved.

Note: these procedures supersede the tie breaking rules described in (Section 9.1.2.2) [RFC4271]

3.3.3.1. Bucketization for BGP DF Election

An advertisement

ADV -> <RD, MH-ID, VBO, ACS, PREF, PE-ID>

is put into the bucket for <RD, MH-ID, VBO>. In other words, the information in BGP DF election consists of <RD, MH-ID, VBO> and only advertisements with exact same <RD, MH-ID, VBO> are candidates for DF election.

3.3.3.2. Bucketization for VPLS DF Election

An advertisement

ADV -> <RD, MH-ID, VBO, DOM, ACS, PREF, PE-ID>

is discarded if DOM is not of interest to the VPLS PE. Otherwise, ADV is put into the bucket for <DOM, MH-ID>. In other words, all

advertisements for a particular VPLS domain that have the same MH-ID are candidates for VPLS DF election.

3.3.3.3. Tie-breaking Rules

This section describes the tie-breaking rules for both BGP and VPLS DF election. Tie-breaking rules for BGP DF election are applied to candidate advertisements by any BGP speaker. Since RD must be same for advertisements to be candidates for BGP DF election, use of unique RDs will result in no candidate advertisements for BGP tie-breaking rules and thus, a BGP speaker in such a case will simply not do BGP DF election. Tie-breaking rules for VPLS DF election are applied to candidate advertisements by all VPLS PEs and the actions taken by VPLS PEs based on the VPLS DF election result are described in Section 3.4.

Given two advertisements ADV1 and ADV2 from a given bucket, first compute the variables needed for DF election:

```
ADV1 -> <RD1, MH-ID1, VBO1, DOM1, ACS1, PREF1, PE-ID1>
ADV2 -> <RD2, MH-ID2, VBO2, DOM2, ACS2, PREF2, PE-ID2>
```

Note that MH-ID1 = MH-ID2 and DOM1 = DOM2, since ADV1 and ADV2 came from the same bucket. If this is for BGP DF election, RD1 = RD2 and VBO1 = VBO2 as well. Then the following tie-breaking rules MUST be applied in the given order.

1. if (ACS1 != 1) AND (ACS2 == 1) ADV1 wins; stop;
 if (ACS1 == 1) AND (ACS2 != 1) ADV2 wins; stop;
 else continue
2. if (PREF1 > PREF2) ADV1 wins; stop;
 else if (PREF1 < PREF2) ADV2 wins; stop;
 else continue
3. if (PE-ID1 < PE-ID2) ADV1 wins; stop;
 else if (PE-ID1 > PE-ID2) ADV2 wins; stop;
 else ADV1 and ADV2 are from the same VPLS PE

For BGP DF election, if there is no winner and ADV1 and ADV2 are from the same PE, BGP DF election should simply consider this as an update.

For VPLS DF election, if there is no winner and ADV1 and ADV2 are from the same PE, a VPLS PE MUST retain both ADV1 and ADV2.

3.4. DF Election on PEs

DF election algorithm MUST be run by all multi-homed VPLS PEs. In addition, all other PEs SHOULD also run the DF election algorithm. As a result of the DF election, multi-homed PEs that lose the DF election for a MH-ID MUST put the ACs associated with the MH-ID in non-forwarding state.

DF election result on the egress PEs can be used in traffic forwarding decision. Figure 2 shows two customer sites, CE1 and CE4, connected to PE1 with CE1 multi-homed to PE1 and PE2. If PE1 is the designated forwarder for CE1, based on the DF election result, PE3 can chose to not send unknown unicast and multicast traffic to PE2 as PE2 is not the designated forwarder for any customer site and it has no other single homed sites connected to it.

4. Multi-AS VPLS

This section describes multi-homing in an inter-AS context.

4.1. Route Origin Extended Community

Due to lack of information about the PEs that originate the VPLS NLRI in inter-AS operations, Route Origin Extended Community [RFC4360] is used to carry the source PE's IP address.

To use Route Origin Extended Community for carrying the originator VPLS PE's loopback address, the type field of the community MUST be set to 0x01 and the Global Administrator sub-field MUST be set to the PE's loopback IP address.

4.2. VPLS Preference

When multiple PEs are assigned the same site ID for multi-homing, it is often desired to be able to control the selection of a particular PE as the designated forwarder. Section 3.5 in [RFC4761] describes the use of BGP Local Preference in path selection to choose a particular NLRI, where Local Preference indicates the degree of preference for a particular VE. The use of Local Preference is inadequate when VPLS PEs are spread across multiple ASes as Local Preference is not carried across AS boundary. A new field, VPLS preference (VP), is introduced in this document that can be used to accomplish this. VPLS preference indicates a degree of preference for a particular customer site. VPLS preference is not mandatory for intra-AS operation; the algorithm explained in Section 3.3 will work with or without the presence of VPLS preference.

Section 3.2.4 in [RFC4761] describes the Layer2 Info Extended Community that carries control information about the pseudowires. The last two octets that were reserved now carries VPLS preference as shown in Figure 3.

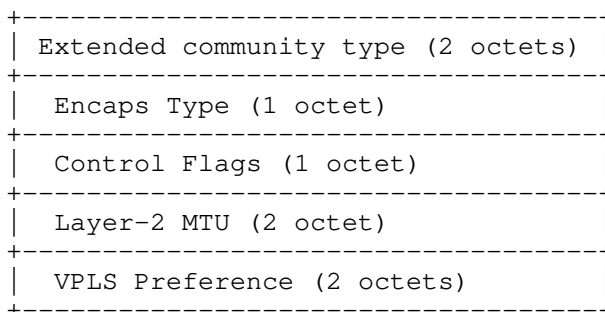


Figure 3: Layer2 Info Extended Community

A VPLS preference is a 2-octets unsigned integer. A value of zero indicates absence of a VP and is not a valid preference value. This interpretation is required for backwards compatibility. Implementations using Layer2 Info Extended Community as described in (Section 3.2.4) [RFC4761] MUST set the last two octets as zero since it was a reserved field.

For backwards compatibility, if VPLS preference is used, then BGP Local Preference MUST be set to the value of VPLS preference. Note that a Local Preference value of zero for a MH-ID is not valid unless 'D' bit in the control flags is set (see [I-D.kothari-l2vpn-auto-site-id]). In addition, Local Preference value greater than or equal to 2^{16} for VPLS advertisements is not valid.

4.3. Use of BGP-MH attributes in Inter-AS Methods

Section 3.4 in [RFC4761] and section 4 in [RFC6074] describe three methods (a, b and c) to connect sites in a VPLS to PEs that are across multiple AS. Since VPLS advertisements in method (a) do not cross AS boundaries, multi-homing operations for method (a) remain exactly the same as they are within an AS. However, for method (b) and (c), VPLS advertisements do cross AS boundary. This section describes the VPLS operations for method (b) and method (c). Consider Figure 4 for inter-AS VPLS with multi-homed customer sites.

4.3.1. Inter-AS Method (b): EBGp Redistribution of VPLS Information between ASBRs

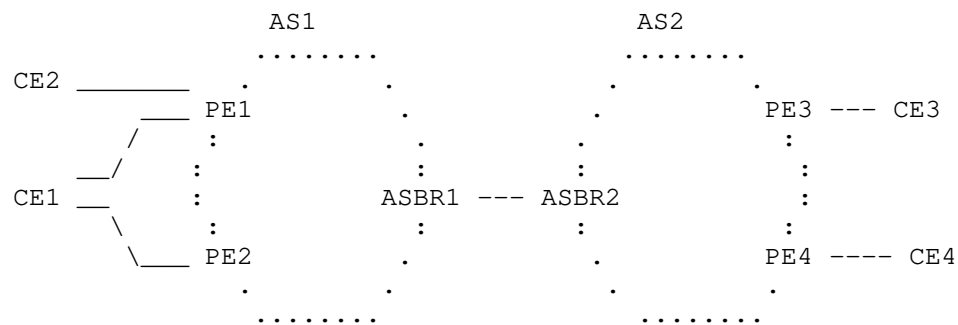


Figure 4: Inter-AS VPLS

A customer has four sites, CE1, CE2, CE3 and CE4. CE1 is multi-homed to PE1 and PE2 in AS1. CE2 is single-homed to PE1. CE3 and CE4 are also single homed to PE3 and PE4 respectively in AS2. Assume that in addition to the base LDP/BGP VPLS addressing (VSI-IDs/VE-IDs), MH ID 1 is assigned for CE1. After running DF election algorithm, all four VPLS PEs must elect the same designated forwarder for CE1 site. Since BGP Local Preference is not carried across AS boundary, VPLS preference as described in Section 4.2 MUST be used for carrying site preference in inter-AS VPLS operations.

For Inter-AS method (b) ASBR1 will send a VPLS NLRI received from PE1 to ASBR2 with itself as the BGP nexthop. ASBR2 will send the received NLRI from ASBR1 to PE3 and PE4 with itself as the BGP nexthop. Since VPLS PEs use BGP Local Preference in DF election, for backwards compatibility, ASBR2 MUST set the Local Preference value in the VPLS advertisements it sends to PE3 and PE4 to the VPLS preference value contained in the VPLS advertisement it receives from ASBR1. ASBR1 MUST do the same for the NLRIs it sends to PE1 and PE2. If ASBR1 receives a VPLS advertisement without a valid VPLS preference from a PE within its AS, then ASBR1 MUST set the VPLS preference in the advertisements to the Local Preference value before sending it to ASBR2. Similarly, ASBR2 must do the same for advertisements without VPLS Preference it receives from PEs within its AS. Thus, in method (b), ASBRs MUST update the VPLS and Local Preference based on the advertisements they receive either from an ASBR or a PE within their AS.

In Figure 4, PE1 will send the VPLS advertisements with Route Origin Extended Community containing its loopback address. PE2 will do the same. Even though PE3 receives the VPLS advertisements for VE-ID 1 and 2 from the same BGP nexthop, ASBR2, the source PE address contained in the Route Origin Extended Community is different for the CE1 and CE2 advertisements, and thus, PE3 creates two PWs, one for CE1 (for VE-ID 1) and another one for CE2 (for VE-ID 2).

4.3.2. Inter-AS Method (c): Multi-Hop EBGp Redistribution of VPLS Information between ASes

In this method, there is a multi-hop E-BGP peering between the PEs or Route Reflectors in AS1 and the PEs or Route Reflectors in AS2. There is no VPLS state in either control or data plane on the ASBRs. The multi-homing operations on the PEs in this method are exactly the same as they are in intra-AS scenario. However, since Local Preference is not carried across AS boundary, the translation of LP to VP and vice versa MUST be done by RR, if RR is used to reflect VPLS advertisements to other ASes. This is exactly the same as what

a ASBR does in case of method (b). A RR must set the VP to the LP value in an advertisement before sending it to other ASes and must set the LP to the VP value in an advertisement that it receives from other ASes before sending to the PEs within the AS.

5. MAC Flush Operations

In a service provider VPLS network, customer MAC learning is confined to PE devices and any intermediate nodes, such as a Route Reflector, do not have any state for MAC addresses.

Topology changes either in the service provider's network or in customer's network can result in the movement of MAC addresses from one PE device to another. Such events can result into traffic being dropped due to stale state of MAC addresses on the PE devices. Age out timers that clear the stale state will resume the traffic forwarding, but age out timers are typically in minutes, and convergence of the order of minutes can severely impact customer's service. To handle such events and expedite convergence of traffic, flushing of affected MAC addresses is highly desirable.

This section describes the scenarios where VPLS flush is desirable and the specific VPLS Flush TLVs that provide capability to flush the affected MAC addresses on the PE devices. All operations described in this section are in context of a particular VPLS domain and not across multiple VPLS domains. Mechanisms for MAC flush are described in [I-D.kothari-l2vpn-vpls-flush] for BGP based VPLS and in [RFC4762] for LDP based VPLS.

5.1. MAC List Flush

If multiple customer sites are connected to the same PE, PE1 as shown in Figure 2, and redundancy per site is desired when multi-homing procedures described in this document are in effect, then it is desirable to flush just the relevant MAC addresses from a particular site when the site connectivity is lost.

To flush particular set of MAC addresses, a PE SHOULD originate a flush message with MAC list that contains a list of MAC addresses that needs to be flushed. In Figure 2, if connectivity between CE1 and PE1 goes down and if PE1 was the designated forwarder for CE1, PE1 MAY send a list of MAC addresses that belong to CE1 to all its BGP peers.

It is RECOMMENDED that in case of excessive link flap of customer attachment circuit in a short duration, a PE should have a means to throttle advertisements of flush messages so that excessive flooding of such advertisements do not occur.

5.2. Implicit MAC Flush

Implicit MAC Flush refers to the use of BGP MH advertisements by the PEs to flush the MAC addresses learned from the previous designated

forwarder.

In case of a failure, when connectivity to a customer site is lost, remote PEs learn that a particular site is no longer reachable. The local PE either withdraws the VPLS NLRI that it previously advertised for the site or it sends a BGP update message for the site's VPLS NLRI with the 'D' bit set. In such cases, the remote PEs can flush all the MACs that were learned from the PE which reported the failure.

However, in cases when a designated forwarder change occurs in absence of failures, such as when an attachment circuit comes up, the BGP MH advertisement from the PE reporting the change is not sufficient for MAC flush procedures. Consider the case in Figure 2 where PE1-CE1 link is non-operational and PE2 is the designated forwarder for CE1. Also assume that Local Preference of PE1 is higher than PE2. When PE1-CE1 link becomes operational, PE1 will send a BGP MH advertisement to all its peers. If PE3 elects PE1 as the new designated forwarder for CE1 and as a result flushes all the MACs learned from PE1 before PE2 elects itself as the non-designated forwarder, there is a chance that PE3 might learn MAC addresses from PE2 and as a result may black-hole traffic until those MAC addresses are deleted due to age out timers.

A new flag 'F' is introduced in the Control Flags Bit Vector as a deterministic way to indicate when to flush.

Control Flags Bit Vector

```

0 1 2 3 4 5 6 7
+---+---+---+---+
|D|A|F|Z|Z|Z|C|S| (Z = MUST Be Zero)
+---+---+---+---+

```

Figure 5

A designated forwarder must set the F bit and a non-designated forwarder must clear the F bit when sending BGP MH advertisements. A state transition from one to zero for the F bit can be used by a remote PE to flush all the MACs learned from the PE that is transitioning from designated forwarder to non-designated forwarder.

5.3. Minimizing the effects of fast link transitions

Certain failure scenarios may result in fast transitions of the link towards the multi-homing CE which in turn will generate fast status transitions of one or multiple multi-homed sites reflected through multiple BGP MH advertisements and LDP MAC Flush messages.

It is recommended that a timer to damp the link flaps be used for the port towards the multi-homed CE to minimize the number of MAC Flush events in the remote PEs and the occurrences of BGP state compressions for F bit transitions. A timer value more than the time it takes BGP to converge in the network is recommended.

6. Backwards Compatibility

No forwarding loops are formed when PEs or Route Reflectors that do not support procedures defined in this section co exist in the network with PEs or Route Reflectors that do support.

6.1. BGP based VPLS

As explained in this section, multi-homed PEs to the same customer site MUST assign the same MH-ID and related NLRI SHOULD contain the block offset, block size and label base as zero. Remote PEs that lack support of multi-homing operations specified in this document will fail to create any PWs for the multi-homed MH-IDs due to the label value of zero and thus, the multi-homing NLRI should have no impact on the operation of Remote PEs that lack support of multi-homing operations specified in this document.

6.2. LDP VPLS with BGP Auto-discovery

The BGP-AD NLRI has a prefix length of 12 containing only a 8 bytes RD and a 4 bytes VSI-ID. If a LDP VPLS PEs running BGP AD lacks support of multi-homing operations specified in this document, it SHOULD ignore a MH NLRI with the length field of 17. As a result it will not ask LDP to create any PWs for the multi-homed Site-ID and thus, the multi-homing NLRI should have no impact on LDP VPLS operation. MH PEs may use existing LDP MAC Flush to flush the remote LDP VPLS PEs or may use the implicit MAC Flush procedure.

7. Security Considerations

No new security issues are introduced beyond those that are described in [RFC4761] and [RFC4762].

8. IANA Considerations

At this time, this memo includes no request to IANA.

9. Acknowledgments

The authors would like to thank Yakov Rekhter, Nischal Sheth, and Mitali Singh for their insightful comments and probing questions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC6074] Rosen, E., "Provisioning, Autodiscovery, and Signaling in L2VPNs", RFC 6074, January 2011.

10.2. Informative References

- [I-D.kothari-l2vpn-vpls-flush]
Kothari, B. and R. Fernando, "VPLS Flush in BGP-based Virtual Private LAN Service",
draft-kothari-l2vpn-vpls-flush-00 (work in progress),
October 2008.
- [I-D.kothari-l2vpn-auto-site-id]
Kothari, B., Kompella, K., and T. IV, "Automatic Generation of Site IDs for Virtual Private LAN Service",
draft-kothari-l2vpn-auto-site-id-01 (work in progress),
October 2008.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Authors' Addresses

Bhupesh Kothari
Cisco Systems
3750 Cisco Way
San Jose, CA 95134, US
Email: bhupesh@cisco.com

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US
Email: kireeti@juniper.net

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.be

Florin Balus
Alcatel-Lucent
Email: florin.balus@alcatel-lucent.com

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748, US
Email: uttaro@att.com

Network Working Group
Internet Draft
Intended status: Informational

B. Mack-Crane
L. Yong
Huawei
October 17, 2011

Expires: April 2012

Shortest Path Bridging (SPB) over an MPLS Packet Switched Network
draft-mack-crane-l2vpn-spb-o-mpls-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the DNSEXT working group mailing list: <rbridge@postel.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this

document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

Abstract

This informational document describes ways to interconnect a Shortest Path Tree (SPT) Region over WAN connections using MPLS Pseudo Wires (PWs) with existing SPB and MPLS standards. It also describes how a combination of SPB and MPLS can provide a hierarchical scalable L2VPN.

Table of Contents

1. Introduction.....	2
2. Use Cases.....	3
2.1. Point-To-Point Interconnection.....	4
2.2. Multiple Interconnections.....	5
2.3. Hierarchical L2VPN with SPB and MPLS.....	7
3. Security Considerations.....	9
4. IANA Considerations.....	9
5. Acknowledgements.....	9
6. References.....	9
6.1. Normative References.....	9
6.2. Informative References.....	10

1. Introduction

The IEEE Shortest Path Bridging (SPB) standard [802.1aq] provides optimal pair-wise data frame forwarding with little or no configuration in multi-hop networks of arbitrary topology. This network behavior is implemented by Shortest Path Tree (SPT) Bridges that automatically confederate (i.e., recognize compatibly configured neighbors) to form SPT Regions within which shortest path bridging is provided. The data plane controlled by SPT Bridges is unchanged from earlier bridging standards except for the addition of a reverse path forwarding check option. The ECMP project [802.1Qbp] will add support for multipath load spreading for both unicast and multicast traffic. SPB enables a new method to construct enterprise and cloud data center networks.

This document describes use cases for SPB over an MPLS Packet Switched Network (PSN) and introduces a new hierarchical L2VPN architecture that uses SPB and IP/MPLS and documents the related

configurations and references for proper interworking. In the use cases described the SPBM mode (MAC address based) is used, implying the existence of a Provider Backbone Edge Bridge function (MAC-in-MAC encapsulation) [802.1Q] at the boundary of the SPT Region.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Acronyms used in this document include the following:

AC - Attachment Circuit

CE - Customer Edge

IS-IS - Intermediate System to Intermediate System

MPLS - Multi-Protocol Label Switching

PE - Provider Edge

PPP - Point to Point Protocol

PW - Pseudo Wire

SPB - Shortest Path Bridging

SPT - Shortest Path Tree

VSI - Virtual Switching Instance

2. Use Cases

SPT Regions at different locations may be interconnected by networks that are implemented with different technologies to form one larger SPT Region. This section describes use cases assuming that IP/MPLS technology is available. From the MPLS network view, SPT Bridges act as Customer Edge (CE) devices and connect to PEs via an attachment circuit (AC). SPT Bridges [802.1aq] support deterministic forwarding behavior over point-to-point links. Section 2.1 describes SPT Region interconnection over a single point-to-point link provided by an MPLS network. Section 2.2 describes interconnecting multiple SPT Regions using multiple PWs. Section 2.3 introduces a hierarchical L2VPN solution that uses SPT Bridges and MPLS in a tiered architecture.

2.1. Point-To-Point Interconnection

Two SPT Bridges are interconnected by either an Ethernet or PPP PW over a MPLS network. The PW is configured between a pair of PEs to provide part of the point-to-point link between two SPT Bridges. Figure 1 illustrates this architecture. Each SPT Bridge connects to a PE via an AC and acts as a CE device. The MPLS PSN is bounded by the PEs. The link across the IP/MPLS PSN enables the site A and site B SPT Bridges to form one SPT Region.

MPLS supports many pseudo wire transport encapsulations [RFC4446]. Two types of links between Bridges have been standardized: Ethernet [RFC4448] and PPP [RFC3518, RFC4618]. A Bridge port connected to an AC may be mapped to a PW with Ethernet encapsulation [RFC4448]. The PW between two PEs can be auto-configured [RFC4447] or manually configured; the two Bridges then appear directly interconnected with an Ethernet link.

When the Bridge ports connected to the ACs are configured with PPP, the PEs may be configured as a PW with PPP encapsulation [RFC4618]. After the PW is established between two PEs, the two R Bridges then appear directly interconnected with a PPP link. Because the frames between the bridges are encapsulated within PPP, if the PEs have the capability to add or remove PPP encapsulation, it is an independent decision for each AC and for the PW whether each is PPP or Ethernet.

An SPB adjacency is automatically established over an Ethernet link or PPP link. The PW provides transparent transport between ACs.

Note: For Ethernet PW configuration, PE SHOULD use the raw mode and non-service-delimiting options.

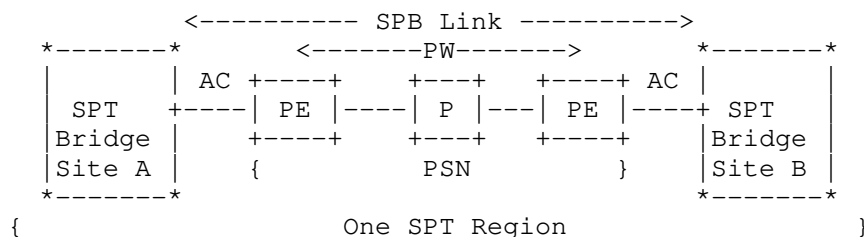


Figure 1 P2P SPB Link over IP/MPLS PSN Use Case I

As networks converge, it is possible that one operator controls both the SPT Region as well as the core MPLS network. Figure 2 illustrates this use case, in which SPT Bridges are also MPLS PE enabled. The interworking between the SPT network and the MPLS PSN is within one device. In this case, a virtual Ethernet interface is configured between the SPT Bridge component and PE component on the SPT/PE device and a Packet-PW is configured between two PE components on two devices to emulate the virtual Ethernet link. An SPB adjacency is established between two RB/PE devices after the PW is established. In this case, SPB runs in the client layer and MPLS runs in the Server Layer; SPB/PE devices support both client and server layer control plane and data plane functions.

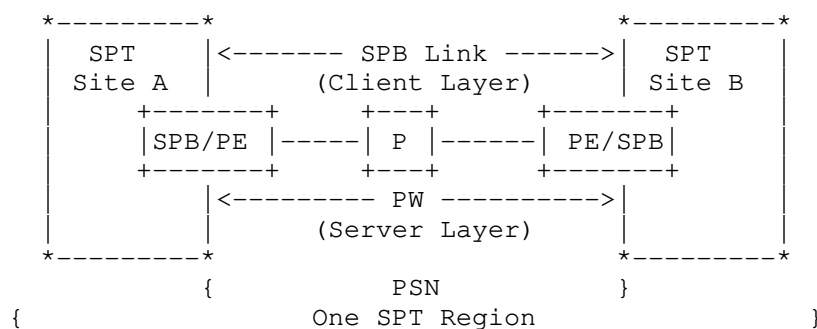


Figure 2 P2P SPB-Link over IP/MPLS PSN Use Case II

In both case I and II, the PE treats an SPT Bridge as a generic CE and has no awareness of SPB capability on the CE. Use case I enables the business models when the SPT Region and Core MPLS may be operated by different operators or the same operator. In the case of different operators, the core MPLS operator can sell a VPWS service to the SPB operator. Use case II provides the model where the SPT Region and the core network are operated by the same operator but use different technologies in edge and core domains of the network.

A PW may cross multiple MPLS domains [RFC5659]. In this case, SPT Bridges connect to T-PEs and it works in the same way as single domain.

2.2. Multiple Interconnections

More than two SPT sites may be interconnected by a full or partial mesh of PWs. The PWs provide a set of links interconnecting the SPT sites and enable the formation of one SPT Region. Interconnecting

multiple sites using PWs is preferable to using a VPLS (VLAN) service because it allows deterministic control of traffic placement and traffic engineering (assuming the PWs provide a bandwidth SLA).

PWs can provide multiple connections to a single physical interface if VLAN tags are used for service selection (Ethernet VLAN ACs). Virtual ports can be provisioned on the SPT Bridge by using a port-mapping S-VLAN component [802.1Qbc]. The S-VID is then used for service selection to map traffic to each PW connection. Figure 3 shows the use of PWs to interconnect three SPT Bridges. One SPT Region is formed across three different sites. Three PWs are configured, providing a full mesh between the three sites. Each SPT site connects to the others via PWs selected by the service-delimiting S-VID on the AC. So in this use case the PEs should use raw mode with service-delimiting.

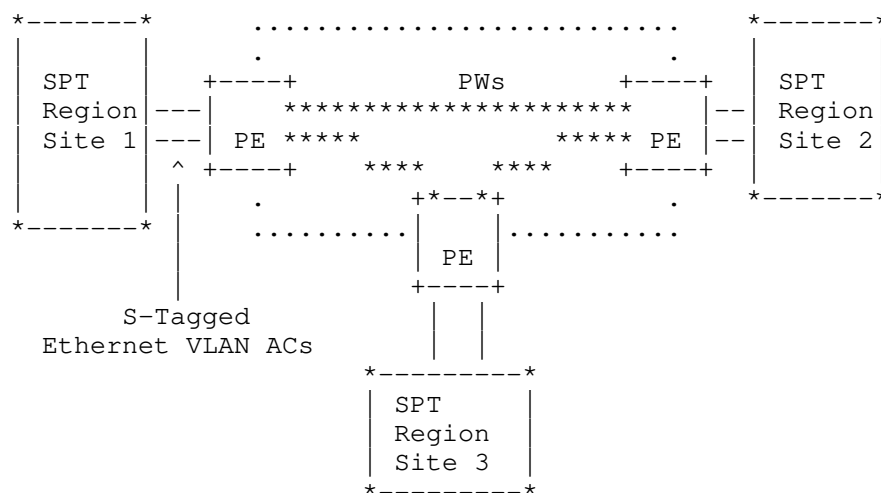


Figure 3 Multiple SPT sites interconnected by PWs

The scenario in Figure 3 can also be applied to interconnect multiple SPT Bridges when a device serves both SPT Bridge and PE functions. This use case is addressed in the following section.

Note: If CEs at a site happen to be regular C-VLAN bridges, the site may be connected to a SPT Bridge via a virtual port bound to an I-Component. This enables MAC-in-MAC encapsulation to be performed

before the traffic enters the SPT Region without requiring upgrade at the C-VLAN bridging site. In this case the PW at the PE connected to the C-VLAN bridging site could be configured as raw mode, non service-delimiting.

2.3. Hierarchical L2VPN with SPB and MPLS

H-VPLS in [RFC4762] describes a two-tier hierarchical solution for the purpose of pseudo wire (PW) scalability improvement. This improvement is achieved by reducing the number of PE devices connected in a full-mesh topology through connecting CE devices via the lower-tier access network, which in turn is connected to the top-tier core network. However, H-VPLS solutions in [RFC4762] require learning and forwarding based on customer MAC addresses, which poses scalability issues as the number of VPLS instances and customer MAC addresses increase. [PBB-VPLS] describes how to use PBB (Provider Backbone Bridges) at the lower-tier access network to solve the scalability issue, in which the transit network nodes only learn and forward on PBB port MAC addresses instead of customer MAC addresses.

Figure 4 depicts the hierarchical L2VPN architecture with SPT Bridge/MPLS technologies. An IP/MPLS network serves the top-tier core network function while an SPT Region serves as the low-tier access network function. A SPB/PE enabled device is placed at the border of the two-tier networks. Ethernet PWs, as described in Section 2.1, are configured between pairs of PE components in the top-tier IP/MPLS network, which construct a full mesh of links among the SPB/PE devices. The SPT Bridge component on a SPB/PE device and other SPT Bridges at the same site serve as the low-tier access network. Customer CEs connect to SPT Bridges at each site directly.

This architecture provides E-LAN or E-VLAN connectivity among customer CEs connecting to the SPT Region sites. The transit SPT Bridge node only forwards and learns other SPT Bridge addresses and the number of PWs in the top-tier core network is not related to the number of devices connecting to Customer CEs. This makes the solution scale very well. In addition, SPB technology supports multiple links from one SPT Bridge to multiple other SPT Bridges and prevents loops, which provides the flexibility to construct the networks based on traffic demands and dynamically reroute traffic when necessary. Figure 4 shows that one SPT Bridge in campus site 1 connects to two SPB/PE devices and one SPB/PE device connects two SPT Bridges at Site 3.

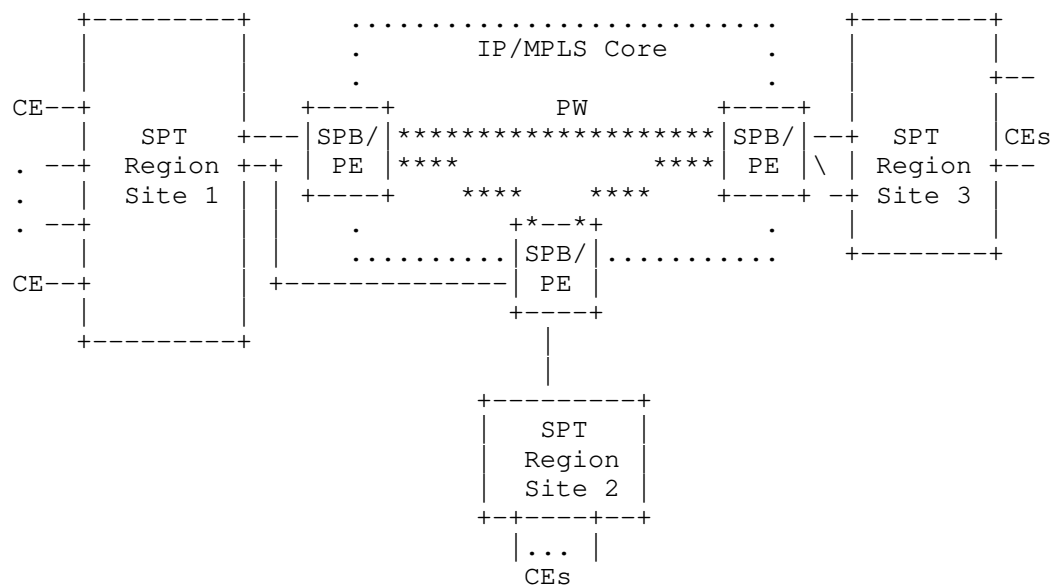


Figure 4 Hierarchical L2VPN with SPB and MPLS

There are several advantages to using SPT Bridge/MPLS based L2VPNs:

1) Scalability improvement; 2) Auto-configuration; 3) Good efficiency and loop prevention; 4) Multipath support (based on 802.1Qbp).

The solution also has advantages over some alternative solutions:

1. SPT Bridges provide deterministic forwarding behavior, allowing network tuning and traffic engineering;
2. SPB supports shortest path for both unicast and multicast traffic;
3. SPT Bridge core interfaces do not have to be upgraded to support a new encapsulation;
4. I-SID supports over 16M tenants; 5) Mature OAM functionality, Ethernet OAM (802.1ag and Y.1731) can be applied to SPB VLANs.

Note: It is possible to construct a Tiered L2VPN in the combination of Figure 4 and 3, i.e. some locations use SPB/PE enabled device and some location use separated SPT Bridge and PE devices in a Hierarchical L2VPN.

3. Security Considerations

The IS-IS authentication mechanism [RFC5304] [RFC5310] can be used to prevent fabrication of link-state control messages including those discussed in this document.

The use cases do not introduce any new security considerations for MPLS networks.

4. IANA Considerations

This document requires no IANA actions.

5. Acknowledgements

The authors would like to acknowledge the contributions of Donald E. rd Eastlake, 3 , Sue Hares, and Sam Aldrin.

6. References

6.1. Normative References

- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels," BCP 14 and RFC 2119, March 1997
- [RFC3518] Higashiyama, M., etc, "Point-to-Point Protocol (PPP) Bridging Control Protocol (BCP)", RFC 3518, April 2003.
- [RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.
- [RFC4447] Martini, L., etc, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC4447, April 2006.
- [RFC4448] Martini, L., "Encapsulation Methods for Transport of Ethernet over MPLS Networks", BCP 116, RFC 4446, April 2006.
- [RFC4618] Martini, L., "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", BCP 116, RFC 4618, September 2006.
- [RFC4762] Lasserre, M., and Kompella, V., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC4762, January 2007

- [RFC5304] Li, T. and Atkinson, R, "IS-IS Cryptographic Authentication," RFC 5304, October 2008
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009
- [RFC5659] Bocci, M and Bryant, S, "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [802.1Q] IEEE Std 802.1Q 2011, Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks, August 2011.
- [802.1Qbc] IEEE Std 802.1Qbc 2011, Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks-Amendment 16: Provider Bridging-Remote Customer Service Interfaces, September 2011.

6.2. Informative References

- [PBB-VPLS] Sajassi, A, etc, "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop, work in progress, 2011

Authors' Addresses

Ben Mack-Crane
Huawei Technologies
5340 Legacy Drive
Plano, TX 75025

Phone: 630-810-1132
Email: ben.mackcrane@huawei.com

Lucy Yong
Huawei Technologies
5340 Legacy Drive
Plano, TX 75025

Phone: 469-227-5837
Email: lucy.yong@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: May 3, 2012

T. Narten, Ed.
IBM
M. Sridharan
Microsoft
D. Dutt
Cisco
D. Black
EMC
L. Kreeger
Cisco
October 31, 2011

Problem Statement: Overlays for Network Virtualization
draft-narten-nvo3-overlay-problem-statement-01

Abstract

This document describes issues associated with providing multi-tenancy in large data center networks and an overlay-based network virtualization approach to addressing them. A key multi-tenancy requirement is traffic isolation, so that a tenant's traffic is not visible to any other tenant. This isolation can be achieved by assigning one or more virtual networks to each tenant such that traffic within a virtual network is isolated from traffic in other virtual networks. The primary functionality required is provisioning virtual networks, associating a virtual machine's NIC with the appropriate virtual network, and maintaining that association as the virtual machine is activated, migrated and/or deactivated. Use of an overlay-based approach enables scalable deployment on large network infrastructures.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. Problem Details	5
2.1. Multi-tenant Environment Scale	5
2.2. Virtual Machine Mobility Requirements	5
2.3. Span of Virtual Networks	5
2.4. Inadequate Forwarding Table Sizes in Switches	6
2.5. Decoupling Logical and Physical Configuration	6
2.6. Support Communication Between VMs and Non-virtualized Devices	6
2.7. Overlay Design Characteristics	6
3. Defining Virtual Networks and Tenants	7
3.1. Limitations of Existing Virtual Network Models	8
3.2. Virtual Network Instance	8
3.3. Tenant	9
4. Network Overlays	9
4.1. Benefits of an Overlay Approach	10
4.2. Standardization Issues for Overlay Networks	10
4.2.1. Overlay Header Format	10
4.2.2. Fragmentation	11
4.2.3. Checksums and FCS	11
4.2.4. Middlebox Traversal	12
4.2.5. OAM	12
5. Control Plane	12
5.1. Populating the Forwarding Table of a Virtual Network Instance	12
5.2. Handling Multi-destination Frames	13
5.3. Associating a VNID With An Endpoint	13
5.4. Disassociating a VNID on Termination or Move	13
6. Related Work	13
6.1. ARMD	13
6.2. TRILL	14
6.3. L2VPNs	14
6.4. Proxy Mobile IP	14
6.5. LISP	14
6.6. Individual Submissions	15
7. Further Work	15
8. Summary	15
9. Acknowledgments	15
10. IANA Considerations	15
11. Security Considerations	15
12. Informative References	16
Authors' Addresses	17

1. Introduction

Server virtualization is increasingly becoming the norm in data centers. With server virtualization, each physical server supports multiple virtual machines (VMs), each running its own operating system, middleware and applications. Virtualization is a key enabler of workload agility, i.e., allowing any server to host any application and providing the flexibility of adding, shrinking, or moving services within the physical infrastructure. Server virtualization provides numerous benefits, including higher utilization, increased data security, reduced user downtime, reduced power usage, etc.

Large scale multi-tenant data centers are taking advantage of the benefits of server virtualization to provide a new kind of hosting, a virtual hosted data center. Multi-tenant data centers are ones in which each tenant could belong to a different company (in the case of a public provider) or a different department (in the case of a internal company data center). Each tenant has the expectation of a level of security and privacy separating their resources from those of other tenants. Each virtual data center looks similar to its physical counterpart, consisting of end stations connected by a network, complete with services such as load balancers and firewalls. The network within each virtual data center can be a pure routed network, a pure bridged network or a combination of bridged and routed network. The key requirement is that each such virtual network is isolated from the others, whether the networks belong to the same tenant or different tenants.

This document outlines the problems encountered in scaling the number of isolated networks in a data center, as well as the problems of managing the creation/deletion, membership and span of these networks and makes the case that an overlay based approach, where individual networks are implemented within individual virtual networks that are dynamically controlled by a standardized control plane provides a number of advantages over current approaches. The purpose of this document is to identify the set of problems that any solution has to address in building multi-tenant data centers. With this approach, the goal is to allow the construction of standardized, interoperable implementations to allow the construction of multi-tenant data centers.

Section 2 describes the problem space details. Section 3 defines virtual networks. Section 4 provides a general discussion of overlays and standardization issues. Section 5 discusses the control plane issues that require addressing for virtual networks. Section 6 and 7 discuss related work and further work.

2. Problem Details

The following subsections describe aspects of multi-tenant networking that pose problems for large scale network infrastructure. Different problem aspects may arise based on the network architecture and scale.

2.1. Multi-tenant Environment Scale

Cloud computing involves on-demand elastic provisioning of resources for multi-tenant environments. A common example of cloud computing is the public cloud, where a cloud service provider offers these elastic services to multiple customers over the same infrastructure. This elastic on-demand nature in conjunction with trusted hypervisors to control network access by VMs calls for resilient distributed network control mechanisms.

2.2. Virtual Machine Mobility Requirements

A key benefit of server virtualization is virtual machine (VM) mobility. A VM can be migrated from one server to another, live i.e. as it continues to run and without shutting down the VM and restarting it at a new location. A key requirement for live migration is that a VM retain its IP address(es) and MAC address(es) in its new location (to avoid tearing down existing communication). Today, servers are assigned IP addresses based on their physical location, typically based on the ToR (Top of Rack) switch for the server rack or the VLAN configured to the server. This works well for physical servers, which cannot move, but it restricts the placement and movement of the more mobile VMs within the data center (DC). Any solution for a scalable multi-tenant DC must allow a VM to be placed (or moved to) anywhere within the data center, without being constrained by the subnet boundary concerns of the host servers.

2.3. Span of Virtual Networks

Another use case is cross pod expansion. A pod typically consists of one or more racks of servers with its associated network and storage connectivity. Tenants may start off on a pod and, due to expansion, require servers/VMs on other pods, especially the case when tenants on the other pods are not fully utilizing all their resources. This use case requires that virtual networks span multiple pods in order to provide connectivity to all of the tenant's servers/VMs.

2.4. Inadequate Forwarding Table Sizes in Switches

Today's virtualized environments place additional demands on the forwarding tables of switches. Instead of just one link-layer address per server, the switching infrastructure has to learn addresses of the individual VMs (which could range in the 100s per server). This is a requirement since traffic from/to the VMs to the rest of the physical network will traverse the physical network infrastructure. This places a much larger demand on the switches' forwarding table capacity compared to non-virtualized environments, causing more traffic to be flooded or dropped when the addresses in use exceeds the forwarding table capacity.

2.5. Decoupling Logical and Physical Configuration

Data center operators must be able to achieve high utilization of server and network capacity. For efficient and flexible allocation, operators should be able to spread a virtual network instance across servers in any rack in the data center. It should also be possible to migrate compute workloads to any server anywhere in the network while retaining the workload's addresses. This can be achieved today by stretching VLANs (e.g., by using TRILL or OTV).

However, in order to limit the broadcast domain of each VLAN, multi-destination frames within a VLAN should optimally flow only to those devices that have that VLAN configured. When workloads migrate, the physical network (e.g., access lists) may need to be reconfigured which is typically time consuming and error prone.

2.6. Support Communication Between VMs and Non-virtualized Devices

Within data centers, not all communication will be between VMs. Network operators will continue to use non-virtualized servers for various reasons, traditional routers to provide L2VPN and L3VPN services, traditional load balancers, firewalls, intrusion detection engines and so on. Any virtual network solution should be capable of working with these existing systems.

2.7. Overlay Design Characteristics

There are existing layer 2 overlay protocols in existence, but they were not necessarily designed to solve the problem in the environment of a highly virtualized data center. Below are some of the characteristics of environments that must be taken into account by the overlay technology:

1. Highly distributed systems. The overlay should work in an environment where there could be many thousands of access switches (e.g. residing within the hypervisors) and many more end systems (e.g. VMs) connected to them. This leads to a distributed mapping system that puts a low overhead on the overlay tunnel endpoints.
2. Many highly distributed virtual networks with sparse connectivity. Each virtual network could be highly dispersed inside the data center. Also, along with expectation of many virtual networks, the number of end systems connected to any one virtual network is expected to be relatively low; Therefore, the percentage of access switches participating in any given virtual network would also be expected to be low. For this reason, efficient pruning of multi-destination traffic should be taken into consideration.
3. Highly dynamic end systems. End systems connected to virtual networks can be very dynamic, both in terms of creation/deletion/power-on/off and in terms of mobility across the access switches.
4. Work with existing, widely deployed network Ethernet switches and IP routers without requiring wholesale replacement. The first hop switch that adds and removes the overlay header will require new equipment and/or new software.
5. Network infrastructure administered by a single administrative domain. This is consistent with operation within a data center, and not across the Internet.

3. Defining Virtual Networks and Tenants

Virtual Networks are used to isolate a tenant's traffic from other tenants (or even traffic within the same tenant that requires isolation). There are two main characteristics of virtual networks:

1. Providing network address space that is isolated from other virtual networks. The same network addresses may be used in different virtual networks on the same underlying network infrastructure.
2. Limiting the scope of frames to not exit a virtual network except through controlled exit points or "gateways".

3.1. Limitations of Existing Virtual Network Models

Virtual networks are not new to networking. VLANs are a well known construct in the networking industry. VLAN is a bridging construct which provides the semantics of virtual networks mentioned above: a MAC address is unique within a VLAN, but not necessarily across VLANs and broadcast traffic is limited to the VLAN it originates from. In the case of IP networks, routers have the concept of a Virtual Routing and Forwarding (VRF). The same router can run multiple instances of routing protocols, each with their own forwarding table. Each instance is referred to as a VRF, which is a mechanism that provides address isolation. Since broadcasts are never forwarded across IP subnets, limiting broadcasts are not applicable to VRFs. In the case of both VLAN and VRF, the forwarding table is looked up using the tuple {VLAN, MAC address} or {VRF, IP address}.

But there are two problems with these constructs. VLANs are a pure bridging construct while VRF is a pure routing construct. VLANs are carried along with a frame to allow each forwarding point to know what VLAN the frame belongs to. VLAN today is defined as a 12 bit number, limiting the total number of VLANs to 4096 (though typically, this number is 4094 since 0 and 4095 are reserved). Due to the large number of tenants that a cloud provider might service, the 4094 VLAN limit is often inadequate. In addition, there is often a need for multiple VLANs per tenant, which exacerbates the issue.

There is no VRF indicator carried in frames. The VRF is derived at each hop using a combination of incoming interface and some information in the frame. Furthermore, the VRF model has typically assumed that a separate control plane governs the population of the forwarding table within that VRF. Thus, a traditional VRF model assumes multiple, independent control planes and has no specific tag within a frame to identify the VRF of the frame.

3.2. Virtual Network Instance

To overcome the limitations of a traditional VLAN or VRF model, we define a new mechanism for virtual networks called a virtual network instance. Each virtual network is assigned a virtual network instance ID, shortened to VNID for convenience. A virtual network instance provides the semantics of a virtual network: address disambiguation and multi-destination frame scoping. A virtual network can be either routed or bridged. So, a VNID can be used for both bridged networks and routed networks and so is unlike a VLAN or a VRF. To build large multi-tenant data centers, a larger number space than the 12b VLAN is required. 24 bits is the most common value identified by multiple solutions that attempt to address this problem space (or similar problem spaces). To simplify the building and

administration of these large data centers, we require that the VNID be carried with each frame (similar to a VLAN, but unlike a VRF). Finally, because of the nature of a virtual data center and to allow scaling virtual networks to massive scales, we don't require a separate control plane to run for each virtual network. We'll identify other possible mechanisms to populate the forwarding tables for virtual networks in section 5.1.

3.3. Tenant

Tenant is the administrative entity that that is responsible for and manages a specific virtual network and its associated services (whether virtual or physical). In a cloud environment, a tenant would correspond to the customer that has defined and is using a particular virtual network. However, there is a one-to-many mapping between tenants and virtual network instances. A single tenant may operate multiple individual virtual networks, each associated with a different service.

4. Network Overlays

To address the problems of decoupling physical and logical configuration and allowing VM mobility without exploding the forwarding table sizes in the switches and routers, a network overlay model can be used.

The idea behind an overlay is quite straightforward. The original frame is encapsulated by the first hop network device. The encapsulation identifies the destination as the device that will perform the decapsulation before delivering the frame to the endpoint. The rest of the network forwards the frame based on the encapsulation header and can be oblivious to the payload that is carried inside. To avoid belaboring the point each time, the first hop network device can be a traditional switch or router or the virtual switch residing inside a hypervisor. Furthermore, the endpoint can be a VM or it can be a physical server. Some examples of network overlays are tunnels such as IP GRE [RFC2784], LISP[I-D.ietf-lisp] or TRILL [RFC6325].

With an overlay, the VNID can be carried within the overlay header so that every frame has its VNID explicitly identified in the frame. Since both routed and bridged semantics can be supported by a virtual data center, the original frame carried within the overlay header can be an Ethernet frame complete with MAC addresses or just the IP packet.

4.1. Benefits of an Overlay Approach

The use of a large (e.g., 24-bit) VNID would allow 16 million distinct virtual networks within a single data center, eliminating current VLAN size limitations. This VNID needs to be carried in the data plane along with the packet. Adding an overlay header provides a place to carry this VNID.

A key aspect of overlays is the decoupling of the "virtual" MAC and IP addresses used by VMs from the physical network infrastructure and the infrastructure IP addresses used by the data center. If a VM changes location, the switches at the edge of the overlay simply update their mapping tables to reflect the new location of the VM within the data center's infrastructure space. Because an overlay network is used, a VM can now be located anywhere in the data center that the overlay reaches without regards to traditional constraints implied by L2 properties such as VLAN numbering, or the span of an L2 broadcast domain scoped to a single pod or access switch.

Multi-tenancy is supported by isolating the traffic of one virtual network instance from traffic of another. Traffic from one virtual network instance cannot be delivered to another instance without (conceptually) exiting the instance and entering the other instance via an entity that has connectivity to both virtual network instances. Without the existence of this entity, tenant traffic remains isolated within each individual virtual network instance. External communications (from a VM within a virtual network instance to a machine outside of any virtual network instance, e.g. on the Internet) is handled by having an ingress switch forward traffic to an external router, where an egress switch decapsulates a tunneled packet and delivers it to the router for normal processing. This router is external to the overlay, and behaves much like existing external facing routers in data centers today.

Overlays are designed to allow a set of VMs to be placed within a single virtual network instance, whether that virtual network provides the bridged network or a routed network.

4.2. Standardization Issues for Overlay Networks

4.2.1. Overlay Header Format

Different overlay header formats are possible as are different possible encodings of the VNID. Existing overlay headers maybe extended or new ones defined. This document does not address the exact header format or VNID encoding except to state that any solution MUST:

1. Carry the VNID in each frame
2. Allow the payload to be either a complete Ethernet frame or only an IP packet

4.2.2. Fragmentation

Whenever tunneling is used, one faces the potential problem that the packet plus the encapsulation overhead will exceed the MTU of the path to the egress router. If the outer encapsulation is IP, fragmentation could be left to the IP layer, or it could be done at the overlay level in a more optimized fashion that is independent of the overlay encapsulation header, or it could be left out altogether, if it is believed that data center networks can be engineered to prevent MTU issues from arising.

Related to fragmentation is the question of how best to handle Path MTU issues, should they occur. Ideally, the original source of any packet (i.e., the sending VM) would be notified of the optimal MTU to use. Path MTU problems occurring within an overlay network would result in ICMP MTU exceeded messages being sent back to the egress tunnel switch at the entry point of the overlay. If the switch is embedded within a hypervisor, the hypervisor could notify the VM of a more appropriate MTU to use. It may be appropriate to specify a set of best practices for implementers related to the handling of Path MTU issues.

4.2.3. Checksums and FCS

When tunneling packets, both the inner and outer headers could have their own checksum, duplicating effort and impacting performance. Therefore, we strongly recommend that any solution carry only one set of checksum or frame FCS.

When the inner packet is TCP or UDP, they already include their own checksum, and adding a second outer checksum (using the same 1's complement algorithm) provides little value. Similarly, if the inner packet is an Ethernet frame, the frame FCS protects the original frame and a new frame FCS over both the original frame and the overlay header protects the new encapsulated frame.

In IPv4, UDP checksums can be disabled on a per-packet basis simply by setting the checksum field to zero. IPv6, however, specifies that UDP checksums must always be included. But even for IPv6, the LISP protocol [I-D.ietf-lisp] already allows a zero checksum field. The 6man working group is also currently considering relaxing the IPv6 UDP checksum requirement [I-D.ietf-6man-udpzero].

For Ethernet frames, L2 overlays such as TRILL already mandate only a single frame FCS.

4.2.4. Middlebox Traversal

One issue to consider is to whether the overlay will need to run over networks that include middleboxes such as NAT. Middleboxes may have difficulty properly supporting multicast or other aspects of an overlay header. Inside a data center, it may well be the case that middlebox traversal is a non-issue. But if overlays are extended across the broader Internet, the presence of middleboxes may be of concern.

4.2.5. OAM

Successful deployment of an overlay approach will likely require appropriate Operations, Administration and Maintenance (OAM) facilities.

5. Control Plane

The control plane needs to address the following pieces, at least:

1. A mechanism to populate the forwarding table of a virtual network instance.
2. A mechanism to handle multi-destination frames within a virtual network instance.
3. A mechanism to allow an endpoint to inform the access switch which virtual network instance it wishes to join on a virtual network interface.
4. A mechanism to allow an endpoint to inform the access switch about its leaving the network so that the access switch can clean up state.

5.1. Populating the Forwarding Table of a Virtual Network Instance

When an access switch has to forward a frame from one endpoint to another, across the network, it has to consult some form of a forwarding table. When we use network overlays, the problem boils down to deriving the mapping between the inner and outer addresses i.e. deriving the destination address in the overlay header based on the destination address sent by the endpoint. Two well known mechanisms for populating the forwarding table (or deriving the mapping table) of a switch are (i) via a routing control protocol and

(ii) learning from the data plane as Ethernet bridges do. Another mechanism is through a centralized mapping database. Any solution must avoid problems associated with scaling a virtual network instance across a large data center.

5.2. Handling Multi-destination Frames

Another aspect of address mapping concerns the handling of multi-destination frames, i.e. broadcast and multicast frames, or the delivery of unicast packets when no mapping exists. Associating a infrastructure multicast address is one possible way of connecting together all the machines belonging to the same VNID. However, existing multicast implementations do not scale to efficiently handle hundreds of thousands of multicast groups, as would be required if one multicast group were assigned to each VNID.

5.3. Associating a VNID With An Endpoint

When an endpoint, such as VM or physical server, connects to the infrastructure, we must define a mechanism to allow the endpoint to identify to the access switch the network instance that it wishes to join. Typically, it is a virtual NIC (the one connected to the VM) coming up that triggers this association. The access switch can then determine the VNID to be associated with this virtual NIC. A standard protocol that all types of overlay encapsulation points can use to identify the VNID associated with an endpoint will be beneficial for supporting multi-vendor implementations. This protocol could also be used to distribute any per virtual network information (e.g. a multicast group address). This signaling can provide the stimulus to trigger the overlay termination points to perform any actions needed within the infrastructure network (e.g. use IGMP to join a multicast group).

5.4. Disassociating a VNID on Termination or Move

To enable cleaning up state in the access switch, we must define a mechanism to allow an endpoint to signal its disconnection from the network.

6. Related Work

6.1. ARMD

ARMD is chartered to look at data center scaling issues with a focus on address resolution. ARMD is currently chartered to develop a problem statement and is not currently developing solutions. While an overlay-based approach may address some of the "pain points" that

have been raised in ARMD (e.g., better support for multi-tenancy), an overlay approach may also push some of the L2 scaling concerns (e.g., excessive flooding) to the IP level (flooding via IP multicast). Analysis will be needed to understand the scaling trade offs of an overlay based approach compared with existing approaches. On the other hand, existing IP-based approaches such as proxy ARP may help mitigate some concerns.

6.2. TRILL

TRILL is an L2 based approach aimed at improving deficiencies and limitations with current Ethernet networks. Approaches to extend TRILL to support more than 4094 VLANs are currently under investigation [I-D.eastlake-trill-rbridge-fine-labeling]

6.3. L2VPNs

The IETF has specified a number of approaches for connecting L2 domains together as part of the L2VPN Working Group. That group, however has historically been focused on Provider-provisioned L2 VPNs, where the service provider participates in management and provisioning of the VPN. In addition, much of the target environment for such deployments involves carrying L2 traffic over WANs. Overlay approaches are intended be used within data centers where the overlay network is managed by the data center operator, rather than by an outside party. While overlays can run across the Internet as well, they will extend well into the data center itself (e.g., up to and including hypervisors) and include large numbers of machines within the data center itself.

Other L2VPN approaches, such as L2TP [RFC2661] require significant tunnel state at the encapsulating and decapsulating end points. Overlays require less tunnel state than other approaches, which is important to allow overlays to scale to hundreds of thousands of end points. It is assumed that smaller switches (i.e., virtual switches in hypervisors or the physical switches to which VMs connect) will be part of the overlay network and be responsible for encapsulating and decapsulating packets.

6.4. Proxy Mobile IP

Proxy Mobile IP [RFC5213] [RFC5844] makes use of the GRE Key Field [RFC5845] [RFC6245], but not in a way that supports multi-tenancy.

6.5. LISP

LISP[I-D.ietf-lisp] essentially provides an IP over IP overlay where the internal addresses are end station Identifiers and the outer IP

addresses represent the location of the end station within the core IP network topology. The LISP overlay header uses a 24 bit Instance ID used to support overlapping inner IP addresses.

6.6. Individual Submissions

Many individual submissions also look to addressing some or all of the issues addressed in this draft. Examples of such drafts are VXLAN [I-D.mahalingam-dutt-dcops-vxlan], NVGRE [I-D.sridharan-virtualization-nvgre] and Virtual Machine Mobility in L3 networks [I-D.wkumari-dcops-l3-vmmobility].

7. Further Work

It is believed that overlay-based approaches may be able to reduce the overall amount of flooding and other multicast and broadcast related traffic (e.g, ARP and ND) currently experienced within current data centers with a large flat L2 network. Further analysis is needed to characterize expected improvements.

8. Summary

This document has argued that network virtualization using L3 overlays addresses a number of issues being faced as data centers scale in size. In addition, careful consideration of a number of issues would lead to the development of interoperable implementation of virtualization overlays.

9. Acknowledgments

Helpful comments and improvements to this document have come from Ariel Hendel, Vinit Jain, and Benson Schliesser.

10. IANA Considerations

This memo includes no request to IANA.

11. Security Considerations

TBD

12. Informative References

- [I-D.eastlake-trill-rbridge-fine-labeling]
Eastlake, D., Zhang, M., Agarwal, P., Dutt, D., and R. Perlman, "RBrigdes: Fine-Grained Labeling", draft-eastlake-trill-rbridge-fine-labeling-02 (work in progress), October 2011.
- [I-D.hasmit-otv]
Grover, H., Rao, D., Farinacci, D., and V. Moreno, "Overlay Transport Virtualization", draft-hasmit-otv-03 (work in progress), July 2011.
- [I-D.ietf-6man-udpzero]
Fairhurst, G. and M. Westerlund, "IPv6 UDP Checksum Considerations", draft-ietf-6man-udpzero-04 (work in progress), October 2011.
- [I-D.ietf-lisp]
Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol (LISP)", draft-ietf-lisp-15 (work in progress), July 2011.
- [I-D.mahalingam-dutt-dcops-vxlan]
Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-00 (work in progress), August 2011.
- [I-D.sridharan-virtualization-nvgre]
Sridharan, M., Duda, K., Ganga, I., Greenberg, A., Lin, G., Pearson, M., Thaler, P., Tumuluri, C., Venkataramaiah, N., and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-00 (work in progress), September 2011.
- [I-D.wkumari-dcops-l3-vm-mobility]
Kumari, W. and J. Halpern, "Virtual Machine mobility in L3 Networks.", draft-wkumari-dcops-l3-vm-mobility-00 (work in progress), August 2011.
- [RFC2661] Townsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., and B. Palter, "Layer Two Tunneling Protocol "L2TP"", RFC 2661, August 1999.
- [RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P.

- Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.
- [RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.
- [RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.
- [RFC5844] Wakikawa, R. and S. Gundavelli, "IPv4 Support for Proxy Mobile IPv6", RFC 5844, May 2010.
- [RFC5845] Muhanna, A., Khalil, M., Gundavelli, S., and K. Leung, "Generic Routing Encapsulation (GRE) Key Option for Proxy Mobile IPv6", RFC 5845, June 2010.
- [RFC6245] Yegani, P., Leung, K., Lior, A., Chowdhury, K., and J. Navali, "Generic Routing Encapsulation (GRE) Key Extension for Mobile IPv4", RFC 6245, May 2011.
- [RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

Authors' Addresses

Thomas Narten (editor)
IBM

Email: narten@us.ibm.com

Murari Sridharan
Microsoft

Email: muraris@microsoft.com

Dinesh Dutt
Cisco

Email: ddutt@cisco.com

David Black
EMC

Email: david.black@emc.com

Lawrence Kreeger
Cisco

Email: kreeger@cisco.com

Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Cisco

Florin Balus
Wim Henderickx
Alcatel-Lucent

Nabil Bitar
Verizon

Clarence Filsfils
Dennis Cai
Cisco

Aldrin Issac
Bloomberg

Lizhong Jin
ZTE

Expires: April 28, 2012

October 28, 2011

PBB E-VPN
draft-sajassi-l2vpn-pbb-evpn-03.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 28, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

This document discusses how Ethernet Provider Backbone Bridging [802.1ah] can be combined with E-VPN in order to reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119

Table of Contents

1. Introduction.....	3
2. Contributors.....	4
3. Terminology.....	4
4. Requirements.....	4
4.1. MAC Advertisement Route Scalability.....	4
4.2. C-MAC Mobility with MAC Sub-netting.....	5
4.3. C-MAC Address Learning and Confinement.....	5
4.4. Interworking with TRILL and 802.1aq Access Networks with C-MAC Address Transparency.....	5
4.5. Per Site Policy Support.....	6
4.6. Avoiding C-MAC Address Flushing.....	6
5. Solution Overview.....	6
6. BGP Encoding.....	7
6.1. BGP MAC Advertisement Route.....	7
6.2. Ethernet Auto-Discovery Route.....	8
6.3. Per VPN Route Targets.....	8
6.4. MAC Mobility Extended Community.....	8

7. Operation.....	8
7.1. MAC Address Distribution over Core.....	8
7.2. Device Multi-homing.....	8
7.2.1. MES MAC Layer Addressing & Multi-homing.....	8
7.2.2. Split Horizon and Designated Forwarder Election.....	11
7.3. Network Multi-homing.....	11
7.3.1. B-MAC Address Advertisement.....	12
7.3.2. Failure Handling.....	12
7.4. Frame Forwarding.....	13
7.4.1. Unicast.....	13
7.4.2. Multicast/Broadcast.....	14
8. Minimizing ARP Broadcast.....	14
9. Seamless Interworking with TRILL and IEEE 802.1aq/802.1Qbp.....	14
9.1. TRILL Nickname Advertisement Route.....	15
9.2. IEEE 802.1aq / 802.1Qbp B-MAC Advertisement Route.....	16
9.3. Operation.....	16
10. Solution Advantages.....	17
10.1. MAC Advertisement Route Scalability.....	18
10.2. C-MAC Mobility with MAC Sub-netting.....	18
10.3. C-MAC Address Learning and Confinement.....	18
10.4. Interworking with TRILL and 802.1aq Access Networks with C-MAC Address Transparency.....	18
10.5. Per Site Policy Support.....	19
10.6. Avoiding C-MAC Address Flushing.....	19
11. Acknowledgements.....	20
12. Security Considerations.....	20
13. IANA Considerations.....	20
14. Intellectual Property Considerations.....	20
15. Normative References.....	20
16. Informative References.....	20
17. Authors' Addresses.....	20

1. Introduction

[E-VPN] introduces a solution for multipoint L2VPN services with advanced multi-homing capabilities using BGP for distributing customer/client MAC address reach-ability information over the core MPLS/IP network. [802.1ah] defines an architecture for Ethernet Provider Backbone Bridging (PBB), where MAC tunneling is employed to improve service instance and MAC address scalability in Ethernet networks and in VPLS networks [PBB-VPLS].

In this document, we discuss how PBB can be combined with E-VPN in order to reduce the number of BGP MAC advertisement routes by aggregating Customer/Client MAC (C-MAC) addresses via Provider Backbone MAC address (B-MAC), provide client MAC address mobility using C-MAC aggregation and B-MAC sub-netting, confine the scope of C-MAC learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes. The combined solution is referred to as PBB-EVPN.

2. Contributors

In addition to the authors listed above, the following individuals also contributed to this document.

Keyur Patel
Cisco

3. Terminology

BEB: Backbone Edge Bridge
B-MAC: Backbone MAC Address
CE: Customer Edge
C-MAC: Customer/Client MAC Address
DHD: Dual-homed Device
DHN: Dual-homed Network
LACP: Link Aggregation Control Protocol
LSM: Label Switched Multicast
MDT: Multicast Delivery Tree
MES: MPLS Edge Switch
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point
PoA: Point of Attachment
PW: Pseudowire
E-VPN: Ethernet VPN

4. Requirements

The requirements for PBB-EVPN include all the requirements for E-VPN that were described in [EVPN-REQ], in addition to the following:

4.1. MAC Advertisement Route Scalability

In typical operation, an [E-VPN] MES sends a BGP MAC Advertisement Route per customer/client MAC (C-MAC) address. In certain applications, this poses scalability challenges, as is the case in virtualized data center environments where the number of virtual machines (VMs), and hence the number of C-MAC addresses, can be in the millions. In such scenarios, it is required to reduce the number of BGP MAC Advertisement routes by relying on a MAC 'summarization'

scheme, as is provided by PBB. Note that the MAC sub-netting capability already built into E-VPN is not sufficient in those environments, as will be discussed next.

4.2. C-MAC Mobility with MAC Sub-netting

Certain applications, such as virtual machine mobility, require support for fast C-MAC address mobility. For these applications, it is not possible to use MAC address sub-netting in E-VPN, i.e. advertise reach-ability to a MAC address prefix. Rather, the exact virtual machine MAC address needs to be transmitted in BGP MAC Advertisement route. Otherwise, traffic would be forwarded to the wrong segment when a virtual machine moves from one Ethernet segment to another. This hinders the scalability benefits of sub-netting.

It is required to support C-MAC address mobility, while retaining the scalability benefits of MAC sub-netting. This can be achieved by leveraging PBB technology, which defines a Backbone MAC (B-MAC) address space that is independent of the C-MAC address space, and aggregate C-MAC addresses via a B-MAC address and then apply sub-netting to B-MAC addresses.

4.3. C-MAC Address Learning and Confinement

In E-VPN, all the MES nodes participating in the same E-VPN instance are exposed to all the C-MAC addresses learnt by any one of these MES nodes because a C-MAC learned by one of the MES nodes is advertised in BGP to other MES nodes in that E-VPN instance. This is the case even if some of the MES nodes for that E-VPN instance are not involved in forwarding traffic to, or from, these C-MAC addresses. Even if an implementation does not install hardware forwarding entries for C-MAC addresses that are not part of active traffic flows on that MES, the device memory is still consumed by keeping record of the C-MAC addresses in the routing table (RIB). In network applications with millions of C-MAC addresses, this introduces a non-trivial waste of MES resources. As such, it is required to confine the scope of visibility of C-MAC addresses only to those MES nodes that are actively involved in forwarding traffic to, or from, these addresses.

4.4. Interworking with TRILL and 802.1aq Access Networks with C-MAC Address Transparency

[TRILL] and [802.1aq] define next generation Ethernet bridging technologies that offer optimal forwarding using IS-IS control plane, and C-MAC address transparency via Ethernet tunneling technologies. When access networks based on TRILL or 802.1aq are interconnected over an MPLS/IP network, it is required to guarantee C-MAC address transparency on the hand-off point and the edge (i.e. MES) of the MPLS network. As such, solutions that require termination of the access data-plane encapsulation (i.e. TRILL or

802.1aq) at the hand-off to the MPLS network do not meet this transparency requirement, and expose the MPLS edge devices to the MAC address scalability problem.

PBB-EVPN supports seamless interconnect with these next generation Ethernet solutions while guaranteeing C-MAC address transparency on the MES nodes.

4.5. Per Site Policy Support

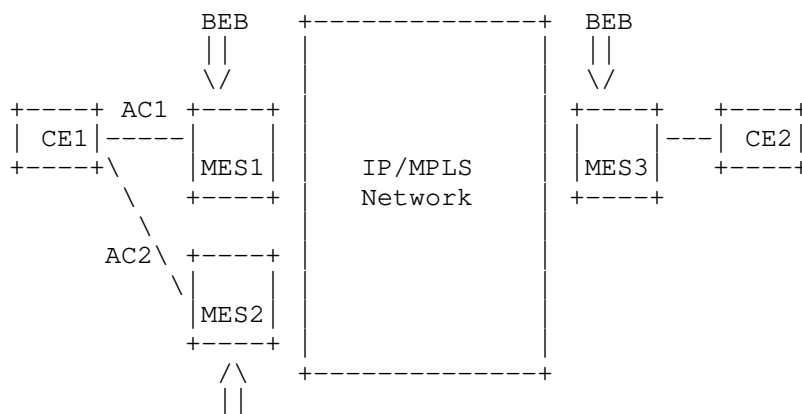
In many applications, it is required to be able to enforce connectivity policy rules at the granularity of a site (or segment). This includes the ability to control which MES nodes in the network can forward traffic to, or from, a given site. PBB-EVPN is capable of providing this granularity of policy control. In the case where per C-MAC address granularity is required, the EVI can always continue to operate in E-VPN mode.

4.6. Avoiding C-MAC Address Flushing

It is required to avoid C-MAC address flushing upon link, port or node failure for multi-homed devices and networks. This is in order to speed up re-convergence upon failure.

5. Solution Overview

The solution involves incorporating IEEE 802.1ah Backbone Edge Bridge (BEB) functionality on the E-VPN MES nodes similar to PBB-VPLS PEs (PBB-VPLS) where BEB functionality is incorporated in PE nodes. The MES devices would then receive 802.1Q Ethernet frames from their attachment circuits, encapsulate them in the PBB header and forward the frames over the IP/MPLS core. On the egress E-VPN MES, the PBB header is removed following the MPLS disposition, and the original 802.1Q Ethernet frame is delivered to the customer equipment.



draft-sajassi-l2vpn-pbb-evpn-03.txt

BEB
<-802.1Q-> <-----PBB over MPLS-----> <-802.1Q->

Figure 1: PBB-EVPN Network

The MES nodes perform the following functions:

- Learn customer/client MAC addresses (C-MACs) over the attachment circuits in the data-plane, per normal bridge operation.
- Learn remote C-MAC to B-MAC bindings in the data-plane from traffic ingress from the core per [802.1ah] bridging operation.
- Advertise local B-MAC address reach-ability information in BGP to all other MES nodes in the same set of service instances. Note that every MES has a set of local B-MAC addresses that uniquely identify the device. More on the MES addressing in section 5.
- Build a forwarding table from remote BGP advertisements received associating remote B-MAC addresses with remote MES IP addresses and the associated MPLS label(s).

6. BGP Encoding

PBB-EVPN leverages the same BGP Routes and Attributes defined in [E-VPN], adapted as follows:

6.1. BGP MAC Advertisement Route

The E-VPN MAC Advertisement Route is used to distribute B-MAC addresses of the MES nodes instead of the C-MAC addresses of end-stations/hosts. This is because the C-MAC addresses are learnt in the data-plane for traffic arriving from the core. The MAC Advertisement Route is encoded as follows:

- The RD is set to a Type 1 RD RD [RFC4364]. The value field encodes the IP address of the MES (typically, the loopback address) followed by 0. The reason for such encoding is that the RD cannot be that of a single EVI since the same B-MAC address can span across multiple EVIs.
- The MAC address field contains the B-MAC address.
- The Ethernet Tag field is set to 0.

The route is tagged with the set of RTs corresponding to all EVIs associated with the B-MAC address.

All other fields are set as defined in [E-VPN].

6.2. Ethernet Auto-Discovery Route

This route and any of its associated modes is not needed in PBB-EVPN.

6.3. Per VPN Route Targets

PBB-EVPN uses the same set of route targets defined in [E-VPN]. More specifically, the RT associated with a VPN is set to the value of the I-SID associated with the service instance. This eliminates the need for manually configuring the VPN-RT.

6.4. MAC Mobility Extended Community

This extended community is a new transitive extended community. It may be advertised along with MAC Advertisement routes. When used in PBB-EVPN, it indicates that the C-MAC forwarding tables for the I-SIDs associated with the RTs tagging the MAC Advertisement routes must be flushed. This extended community is encoded in 8-bytes as follows:

- Type (1 byte) = Pending IANA assignment.
- Sub-Type (1 byte) = Pending IANA assignment.
- Reserved (2 bytes)
- Counter (4 bytes)

Note that all other BGP messages and/or attributes are used as defined in [E-VPN].

7. Operation

This section discusses the operation of PBB-EVPN, specifically in areas where it differs from [E-VPN].

7.1. MAC Address Distribution over Core

In PBB-EVPN, host MAC addresses (i.e. C-MAC addresses) need not be distributed in BGP. Rather, every MES independently learns the C-MAC addresses in the data-plane via normal bridging operation. Every MES has a set of one or more unicast B-MAC addresses associated with it, and those are the addresses distributed over the core in MAC Advertisement routes. Given that these B-MAC addresses are global within the provider's network, there's no need to advertise them on a per service instance basis.

7.2. Device Multi-homing

7.2.1. MES MAC Layer Addressing & Multi-homing

In [802.1ah] every BEB is uniquely identified by one or more B-MAC addresses. These addresses are usually locally administered by the

Service Provider. For PBB-EVPN, the choice of B-MAC address(es) for the MES nodes must be examined carefully as it has implications on the proper operation of multi-homing. In particular, for the scenario where a CE is multi-homed to a number of MES nodes with all-active redundancy and flow-based load-balancing, a given C-MAC address would be reachable via multiple MES nodes concurrently. Given that any given remote MES will bind the C-MAC address to a single B-MAC address, then the various MES nodes connected to the same CE must share the same B-MAC address. Otherwise, the MAC address table of the remote MES nodes will keep flip-flopping between the B-MAC addresses of the various MES devices. For example, consider the network of Figure 1, and assume that MES1 has B-MAC BM1 and MES2 has B-MAC BM2. Also, assume that both links from CE1 to the MES nodes are part of an all-active multi-chassis Ethernet link aggregation group. If BM1 is not equal to BM2, the consequence is that the MAC address table on MES3 will keep oscillating such that the C-MAC address CM of CE1 would flip-flop between BM1 or BM2, depending on the load-balancing decision on CE1 for traffic destined to the core.

Considering that there could be multiple sites (e.g. CEs) that are multi-homed to the same set of MES nodes, then it is required for all the MES devices in a Redundancy Group to have a unique B-MAC address per site. This way, it is possible to achieve fast convergence in the case where a link or port failure impacts the attachment circuit connecting a single site to a given MES.

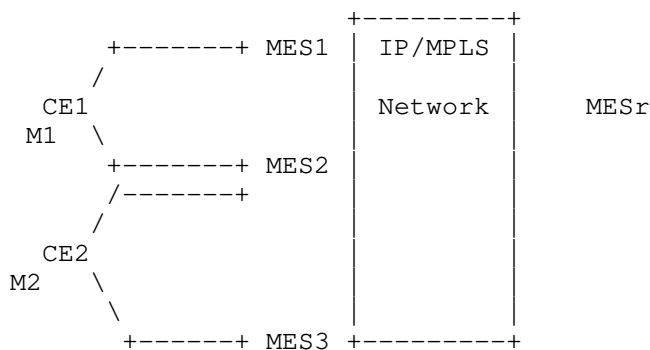


Figure 2: B-MAC Address Assignment

In the example network shown in Figure 2 above, two sites corresponding to CE1 and CE2 are dual-homed to MES1/MES2 and MES2/MES3, respectively. Assume that BM1 is the B-MAC used for the site corresponding to CE1. Similarly, BM2 is the B-MAC used for the site corresponding to CE2. On MES1, a single B-MAC address (BM1) is required for the site corresponding to CE1. On MES2, two B-MAC addresses (BM1 and BM2) are required, one per site. Whereas on MES3,

a single B-MAC address (BM2) is required for the site corresponding to CE2. All three MES nodes would advertise their respective B-MAC addresses in BGP using the MAC Advertisement routes defined in [E-VPN]. The remote MES, MESr, would learn via BGP that BM1 is reachable via MES1 and MES2, whereas BM2 is reachable via both MES2 and MES3. Furthermore, MESr establishes via the normal bridge learning that C-MAC M1 is reachable via BM1, and C-MAC M2 is reachable via BM2. As a result, MESr can load-balance traffic destined to M1 between MES1 and MES2, as well as traffic destined to M2 between both MES2 and MES3. In the case of a failure that causes, for example, CE1 to be isolated from MES1, the latter can withdraw the route it has advertised for BM1. This way, MESr would update its path list for BM1, and will send all traffic destined to M1 over to MES2 only.

For single-homed sites, it is possible to assign a unique B-MAC address per site, or have all the single-homed sites connected to a given MES share a single B-MAC address. The advantage of the first model over the second model is the ability to avoid C-MAC destination address lookup on the disposition PE (even though source C-MAC learning is still required in the data-plane). Also, by assigning the B-MAC addresses from a contiguous range, it is possible to advertise a single B-MAC subnet for all single-homed sites, thereby rendering the number of MAC advertisement routes required at par with the second model.

In summary, every MES may use a unicast B-MAC address shared by all single-homed CEs or a unicast B-MAC address per single-homed CE, and in addition a unicast B-MAC address per dual-homed CE. In the latter case, the B-MAC address MUST be the same for all MES nodes in a Redundancy Group connected to the same CE.

7.2.1.1. Automating B-MAC Address Assignment

The MES B-MAC address used for single-homed sites can be automatically derived from the hardware (using for e.g. the backplane's address). However, the B-MAC address used for multi-homed sites must be coordinated among the RG members. To automate the assignment of this latter address, the MES can derive this B-MAC address from the MAC Address portion of the CE's LACP System Identifier by flipping the 'Locally Administered' bit of the CE's address. This guarantees the uniqueness of the B-MAC address within the network, and ensures that all MES nodes connected to the same multi-homed CE use the same value for the B-MAC address.

Note that with this automatic provisioning of the B-MAC address associated with multi-homed CEs, it is not possible to support the uncommon scenario where a CE has multiple bundles towards the MES nodes, and the service involves hair-pinning traffic from one bundle to another. This is because the split-horizon filtering relies on B-MAC addresses rather than Site-ID Labels (as will be described in

the next section). The operator must explicitly configure the B-MAC address for this fairly uncommon service scenario.

Whenever a B-MAC address is provisioned on the MES, either manually or automatically (as an outcome of CE auto-discovery), the MES MUST transmit an MAC Advertisement Route for the B-MAC address with a downstream assigned MPLS label that uniquely identifies that address on the advertising MES. The route is tagged with the RTs of the associated EVIs as described above.

7.2.2. Split Horizon and Designated Forwarder Election

[E-VPN] relies on access split horizon, where the Ethernet Segment Label is used for egress filtering on the attachment circuit in order to prevent forwarding loops. In PBB-EVPN, the B-MAC source address can be used for the same purpose, as it uniquely identifies the originating site of a given frame. As such, Segment Labels are not used in PBB-EVPN, and the egress filtering is done based on the B-MAC source address. It is worth noting here that [802.1ah] defines this B-MAC address based filtering function as part of the I-Component options, hence no new functions are required to support split-horizon beyond what is already defined in [802.1ah]. Given that the Segment label is not used in PBB-EVPN, the MES sets the Label field in the Ethernet Segment Route to 0.

The Designated Forwarder election procedures remain unchanged from [E-VPN].

7.3. Network Multi-homing

When an Ethernet network is multi-homed to a set of MES nodes running PBB-EVPN, an all-active redundancy model can be supported with per service instance (i.e. I-SID) load-balancing. In this model, DF election is performed to ensure that a single MES node in the redundancy group is responsible for forwarding traffic associated with a given I-SID. This guarantees that no forwarding loops are created. Filtering based on DF state applies to both unicast and multicast traffic, and in both access-to-core as well as core-to-access directions (unlike the multi-homed device scenario where DF filtering is limited to multi-destination frames in the core-to-access direction).

Similar to the multi-homed device scenario, a unique B-MAC address is used on the MES per multi-homed network (Segment). This helps eliminate the need for C-MAC address flushing in all but one failure scenario (more details on this in the Failure Handling section below). The B-MAC address may be auto-provisioned by snooping on the BPDUs of the multi-homed network: the B-MAC address is set to the root bridge ID of the CIST albeit with the 'Locally Administered' bit set.

7.3.1. B-MAC Address Advertisement

For every multi-homed network, the MES advertises two MAC Advertisement routes with different RDs and identical MAC addresses and ESIs. One of these routes will be tagged with a lower Local Pref attribute than the other. The route with the higher Local Pref will be tagged with the RTs corresponding to the I-SIDs for which the advertising MES is the DF. Whereas, the route with the lower Local Pref will be tagged with the RTs corresponding to the I-SIDs for which the advertising MES is the backup DF. Consider the example network of the figure below, where a multi-homed network (MHN1) is connected to two MES nodes (MES1 and MES2).

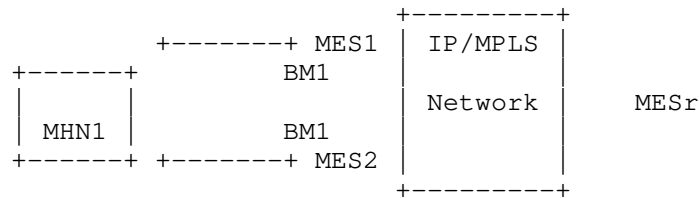


Figure 3: Multi-homed Network

Both MES nodes use the same B-MAC address (BM1) for the Ethernet Segment (ESI1) associated with MHN1. Assume, for instance, that MES1 is the DF for the even I-SIDs whereas MES2 is the DF for the odd I-SIDs. In this example, the routes advertised by MES1 and MES2 would be as follows:

MES1:

Route 1: RD11, BM1, ESI1, Local Pref = 120, RT2, RT4, RT6...
 Route 2: RD12, BM1, ESI1, Local Pref = 80, RT1, RT3, RT5...

MES2:

Route 1: RD21, BM1, ESI1, Local Pref = 120, RT1, RT3, RT5...
 Route 2: RD22, BM1, ESI1, Local Pref = 80, RT2, RT4, RT6

Upon receiving the above MAC Advertisement routes, the remote MES nodes (e.g. MESr) would install forwarding entries for BM1 towards MES1 for the even I-SIDs, and towards MES2 for the odd I-SIDs.

It is worth noting that the procedures of this section can also be used for a multi-homed device in order to support all-active redundancy with per I-SID load-balancing.

7.3.2. Failure Handling

In the case of an MES node failure, or when the MES is isolated from the multi-homed network due to a port or link failure, the affected

MES withdraws its MAC Advertisement routes for the associated B-MAC. This serves as a trigger for the remote MES nodes to adjust their forwarding entries to point to the backup DF. Because the same B-MAC address is used on both the DF and backup DF nodes, then there is no need to flush the C-MAC address table upon the occurrence of these failures.

In the case where the multi-homed network is partitioned, the MES nodes can detect this condition by snooping on the network's BPDUs. When a MES detects that the root bridge ID has changed, it must change the value of the B-MAC address associated with the Ethernet Segment. This is done by the MES withdrawing the previous MAC Advertisement route, and advertising a new route for the updated B-MAC. The MES, which detects the failure, must inform the remote MES nodes to flush their C-MAC address tables for the affected I-SIDs. This is required because when the multi-homed network is partitioned, certain C-MAC addresses will move from being associated with the old B-MAC address to the new B-MAC addresses. Other C-MAC addresses will have their reachability remaining intact. Given that the MES node has no means of identifying which C-MACs have moved and which have not, the entire C-MAC forwarding table for the affected I-SIDs must be flushed. The affected MES signals the need for the C-MAC flushing by sending the MAC Mobility Extended Community in the MP_UNREACH_NLRI attribute containing the E-VPN NLRI for the withdrawn MAC Advertisement route.

7.4. Frame Forwarding

The frame forwarding functions are divided in between the Bridge Module, which hosts the [802.1ah] Backbone Edge Bridge (BEB) functionality, and the MPLS Forwarder which handles the MPLS imposition/disposition. The details of frame forwarding for unicast and multi-destination frames are discussed next.

7.4.1. Unicast

Known unicast traffic received from the AC will be PBB-encapsulated by the MES using the B-MAC source address corresponding to the originating site. The unicast B-MAC destination address is determined based on a lookup of the C-MAC destination address (the binding of the two is done via transparent learning of reverse traffic). The resulting frame is then encapsulated with an LSP tunnel label and the MPLS label which uniquely identifies the B-MAC destination address on the egress MES. If per flow load-balancing over ECMPs in the MPLS core is required, then a flow label is added as the end of stack label.

For unknown unicast traffic, the MES forwards these frames over MPLS core. When these frames are to be forwarded, then the same set of

options used for forwarding multicast/broadcast frames (as described in next section) are used.

7.4.2. Multicast/Broadcast

Multi-destination frames received from the AC will be PBB-encapsulated by the MES using the B-MAC source address corresponding to the originating site. The multicast B-MAC destination address is selected based on the value of the I-SID as defined in [802.1ah]. The resulting frame is then forwarded over the MPLS core using one out of the following two options:

Option 1: the MPLS Forwarder can perform ingress replication over a set of MP2P tunnel LSPs. The frame is encapsulated with a tunnel LSP label and the E-VPN ingress replication label advertised in the Inclusive Multicast Route.

Option 2: the MPLS Forwarder can use P2MP tunnel LSP per the procedures defined in [E-VPN]. This includes either the use of Inclusive or Aggregate Inclusive trees.

Note that the same procedures for advertising and handling the Inclusive Multicast Route defined in [E-VPN] apply here.

8. Minimizing ARP Broadcast

The MES nodes implement an ARP-proxy function in order to minimize the volume of ARP traffic that is broadcasted over the MPLS network. This is achieved by having each MES node snoop on ARP request and response messages received over the access interfaces or the MPLS core. The MES builds a cache of IP / MAC address bindings from these snooped messages. The MES then uses this cache to respond to ARP requests ingress on access ports and targeting hosts that are in remote sites. If the MES finds a match for the IP address in its ARP cache, it responds back to the requesting host and drops the request. Otherwise, if it does not find a match, then the request is flooded over the MPLS network using either ingress replication or LSM.

9. Seamless Interworking with TRILL and IEEE 802.1aq/802.1Qbp

PBB-EVPN enables seamless connectivity of TRILL or 802.1aq/802.1Qbp networks over an MPLS/IP core while maintaining control-plane separation among these networks. We will refer to one or any of TRILL, 802.1aq or 802.1Qbp networks collectively as 'NG-Ethernet networks' thereafter.

Every NG-Ethernet network that is connected to the MPLS core runs an independent instance of the corresponding IS-IS control-plane. Each MES participates in the NG-Ethernet network control plane of its local site. The MES peers, in IS-IS protocol, with the switches internal to the site, but does not terminate the TRILL / PBB data-

plane encapsulation. So, from a control-plane viewpoint, the MES appears as an edge switch; whereas, from a data-plane viewpoint, the MES appears as a core switch to the NG-Ethernet network. The MES nodes encapsulate TRILL / PBB frames with MPLS in the imposition path, and de-capsulate them in the disposition path.

9.1. TRILL Nickname Advertisement Route

A new BGP route is defined to support the interconnection of TRILL networks over PBB-EVPN: the TRILL Nickname Advertisement' route, encoded as follows:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
Nickname Length (1 octet)
RBridge Nickname (2 octets)
MPLS Label ($n * 3$ octets)

Figure 4: TRILL Nickname Advertisement Route

The MES uses this route to advertise the reachability of TRILL RBridge nicknames to other MES nodes in the VPN instance. The MPLS label advertised in this route can be allocated on a per VPN basis and serves the purpose of identifying to the disposition MES that the MPLS-encapsulated packet holds an MPLS encapsulated TRILL frame.

The encapsulation for the transport of TRILL frames over MPLS is encoded as shown in the figure below:

IP/MPLS Header
TRILL Header
Ethernet Header
Ethernet Payload
Ethernet FCS

Figure 5: TRILL over MPLS Encapsulation

It is worth noting here that while it is possible to transport Ethernet encapsulated TRILL frames over MPLS, that approach unnecessarily wastes 16 bytes per packet. That approach further requires either the use of well-known MAC addresses or having the MES nodes advertise in BGP their device MAC addresses, in order to resolve the TRILL next-hop L2 adjacency. To that end, it is simpler and more efficient to transport TRILL natively over MPLS and that is why we are defining the above BGP route for TRILL Nickname advertisement.

9.2. IEEE 802.1aq / 802.1Qbp B-MAC Advertisement Route

B-MAC addresses associated with 802.1aq / 802.1Qbp switches are advertised using the BGP MAC Advertisement route already defined in [E-VPN].

The encapsulation for the transport of PBB frames over MPLS is similar to that of classical Ethernet, albeit with the additional PBB header, as shown in the figure below:

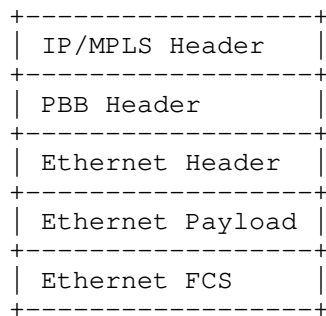


Figure 6: PBB over MPLS Encapsulation

9.3. Operation

For correct connectivity, the TRILL Nicknames or 802.1aq/802.1Qbp B-MACs must be globally unique in the network. This can be achieved, for instance, by using a hierarchical Nickname (or B-MAC) assignment paradigm, and encoding a Site ID in the high-order bits of the Nickname (or B-MAC):

Nickname (or B-MAC) = [Site ID : Rbridge ID (or MAC)]

The only practical difference between TRILL Nicknames and B-MACs, in this regards, is with respect to the size of the address space: Nicknames are 16-bits wide whereas B-MACs are 48-bits wide.

Every MES then advertises (in BGP) the Nicknames (or B-MACs) of all switches local to its site in the TRILL Nickname Advertisement routes (or MAC Advertisement routes).

Furthermore, the MES advertises in IS-IS (to the local island) the Rbridge nicknames (or B-MACs) of all remote switches in all the other TRILL (or IEEE 802.1aq/802.1Qbp) islands that the MES has learned via BGP.

Note that by having multiple MES nodes (connected to the same TRILL or 802.1aq /802.1Qbp island) advertise routes to the same RBridge nickname (or B-MAC), with equal BGP Local_Pref attribute, it is possible to perform active/active load-balancing to/from the MPLS core.

When a MES receives an Ethernet-encapsulated TRILL frame from the access side, it removes the Ethernet encapsulation (i.e. outer MAC header), and performs a lookup on the egress RBridge nickname in the TRILL header to identify the next-hop. If the lookup yields that the next hop is a remote MES, the local MES would then encapsulate the TRILL frame in MPLS. The label stack comprises of the VPN label (advertised by the remote MES), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition MES, assuming penultimate-hop-popping is employed, the MES receives the MPLS-encapsulated TRILL frame with a single label: the VPN label. The value of the label indicates to the disposition MES that this is a TRILL packet, so the label is popped, the TTL field (in the TRILL header) is reinitialized and normal TRILL processing is employed from this point onwards.

By the same token, when a MES receives a PBB-encapsulated Ethernet frame from the access side, it performs a lookup on the B-MAC destination address to identify the next hop. If the lookup yields that the next hop is a remote MES, the local MES would then encapsulate the PBB frame in MPLS. The label stack comprises of the VPN label (advertised by the remote PE), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition MES, assuming penultimate-hop-popping is employed, the MES receives the MPLS-encapsulated PBB frame with a single label: the VPN label. The value of the label indicates to the disposition MES that this is a PBB frame, so the label is popped, the TTL field (in the 802.1Qbp F-Tag) is reinitialized and normal PBB processing is employed from this point onwards.

10. Solution Advantages

In this section, we discuss the advantages of the PBB-EVPN solution in the context of the requirements set forth in section 3 above.

10.1. MAC Advertisement Route Scalability

In PBB-EVPN the number of MAC Advertisement Routes is a function of the number of segments (sites), rather than the number of hosts/servers. This is because the B-MAC addresses of the MESes, rather than C-MAC addresses (of hosts/servers) are being advertised in BGP. And, as discussed above, there's a one-to-one mapping between multi-homed segments and B-MAC addresses, whereas there's a one-to-one or many-to-one mapping between single-homed segments and B-MAC addresses for a given MES. As a result, the volume of MAC Advertisement Routes in PBB-EVPN is multiple orders of magnitude less than E-VPN.

10.2. C-MAC Mobility with MAC Sub-netting

In PBB-EVPN, if a MES allocates its B-MAC addresses from a contiguous range, then it can advertise a MAC prefix rather than individual 48-bit addresses. It should be noted that B-MAC addresses can easily be assigned from a contiguous range because MES nodes are within the provider administrative domain; however, CE devices and hosts are typically not within the provider administrative domain. The advantage of such MAC address sub-netting can be maintained even as C-MAC addresses move from one Ethernet segment to another. This is because the C-MAC address to B-MAC address association is learnt in the data-plane and C-MAC addresses are not advertised in BGP. To illustrate how this compares to E-VPN, consider the following example:

If a MES running E-VPN advertises reachability for a MAC subnet that spans N addresses via a particular segment, and then 50% of the MAC addresses in that subnet move to other segments (e.g. due to virtual machine mobility), then in the worst case, $N/2$ additional MAC Advertisement routes need to be sent for the MAC addresses that have moved. This defeats the purpose of the sub-netting. With PBB-EVPN, on the other hand, the sub-netting applies to the B-MAC addresses which are statically associated with MES nodes and are not subject to mobility. As C-MAC addresses move from one segment to another, the binding of C-MAC to B-MAC addresses is updated via data-plane learning.

10.3. C-MAC Address Learning and Confinement

In PBB-EVPN, C-MAC address reachability information is built via data-plane learning. As such, MES nodes not participating in active conversations involving a particular C-MAC address will purge that address from their forwarding tables. Furthermore, since C-MAC addresses are not distributed in BGP, MES nodes will not maintain any record of them in control-plane routing table.

10.4. Seamless Interworking with TRILL and 802.1aq Access Networks

Consider the scenario where two access networks, one running MPLS and the other running 802.1aq, are interconnected via an MPLS backbone network. The figure below shows such an example network.

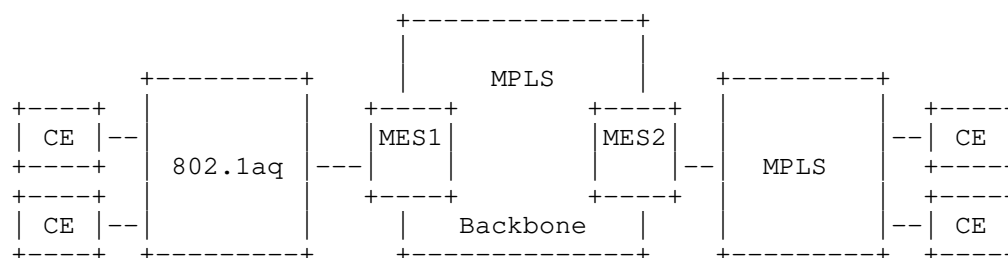


Figure 7: Interoperability with 802.1aq

If the MPLS backbone network employs E-VPN, then the 802.1aq data-plane encapsulation must be terminated on MES1 or the edge device connecting to MES1. Either way, all the MES nodes that are part of the associated service instances will be exposed to all the C-MAC addresses of all hosts/servers connected to the access networks. However, if the MPLS backbone network employs PBB-EVPN, then the 802.1aq encapsulation can be extended over the MPLS backbone, thereby maintaining C-MAC address transparency on MES1. If PBB-EVPN is also extended over the MPLS access network on the right, then C-MAC addresses would be transparent to MES2 as well.

Interoperability with TRILL access network will be described in future revision of this draft.

10.5. Per Site Policy Support

In PBB-EVPN, a unique B-MAC address can be associated with every site (single-homed or multi-homed). Given that the B-MAC addresses are sent in BGP MAC Advertisement routes, it is possible to define per site (i.e. B-MAC) forwarding policies including policies for E-TREE service.

10.6. Avoiding C-MAC Address Flushing

With PBB-EVPN, it is possible to avoid C-MAC address flushing upon topology change affecting a multi-homed device. To illustrate this, consider the example network of Figure 1. Both MES1 and MES2 advertize the same B-MAC address (BM1) to MES2. MES2 then learns the C-MAC addresses of the servers/hosts behind CE1 via data-plane learning. If AC1 fails, then MES3 does not need to flush any of the C-MAC addresses learnt and associated with BM1. This is because MES1 will withdraw the MAC Advertisement routes associated with BM1,

thereby leading MES3 to have a single adjacency (to MES2) for this B-MAC address. Therefore, the topology change is communicated to MES3 and no C-MAC address flushing is required.

11. Acknowledgements
TBD.

12. Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be discussed here.

13. IANA Considerations

This document requires IANA to assign a new SAFI value for L2VPN_MAC SAFI.

14. Intellectual Property Considerations

This document is being submitted for use in IETF standards discussions.

15. Normative References

[802.1ah] "Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges", IEEE Std. 802.1ah-2008, August 2008.

16. Informative References

[PBB-VPLS] Sajassi et al., "VPLS Interoperability with Provider Backbone Bridges", draft-ietf-l2vpn-vpls-pbb-interop-00.txt, work in progress, September, 2011.

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt, work in progress, October, 2010.

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-01.txt, November, 2010., work in progress, June, 2010.

17. Authors' Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Samer Salam

Cisco
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Florin Balus
Alcatel-Lucent
701 E. Middlefield Road
Mountain View, CA, USA 94043
Email: florin.balus@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.be

Clarence Filsfils
Cisco
Email: cfilsfil@cisco.com

Dennis Cai
Cisco
Email: dcai@cisco.com

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn

Network Working Group
Internet Draft
Intended status: Informational

L. Yong
D. Eastlake
S. Aldrin
Huawei
J. Hudson
Brocade

Expires: April 2012

October 23, 2011

Transparent Interconnection of Lots of Links (TRILL) over an MPLS
PSN (Packet Switched Network)
draft-yong-trill-trill-o-mpls-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the DNSEXT working group mailing list: <rbridge@postel.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

Abstract

This informational document describes ways to interconnect TRILL R Bridges over WAN connections by using MPLS Pseudo Wire (PW) or Virtual Private LAN Service (VPLS) with existing TRILL and MPLS standards. It also describes the combination of R Bridge and MPLS to provide a hierarchical scalable L2VPN.

Table of Contents

1. Introduction.....	2
2. Use Cases.....	3
2.1. Point-To-Point Interconnection.....	3
2.2. Multi-Access Link Interconnection.....	6
2.3. Hierarchical L2VPN with R Bridges and MPLS.....	8
3. R Bridge Behavior for MPLS Pseudo Wire.....	10
4. Security Considerations.....	11
5. IANA Considerations.....	11
6. Acknowledgements.....	11
7. References.....	11
7.1. Normative References.....	11
7.2. Informative References.....	13

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) standard [RFC6325] [RFC6326] provides optimal pair-wise data frame forwarding without configuration in multi-hop networks with arbitrary topology, and support for multipathing of both unicast and multicast traffic. TRILL enables a new method to construct a campus or data center network. Devices that implement TRILL are called R Bridges.

This document describes the use cases of TRILL over an MPLS PW or VPLS, and introduces a new hierarchical L2VPN architecture that uses R Bridges and IP/MPLS and documents the related configurations and references for the proper interworking.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Acronyms used in this document include the following:

AC - Attachment Circuit

CE - Customer Edge

IS-IS - Intermediate System to Intermediate System

MPLS - Multi-Protocol Label Switching

PE - Provider Edge

PPP - Point to Point Protocol

PW - Pseudo Wire

RBridge - Routing Bridge

TRILL - Transparent Interconnection of Lots of Links

VPLS - Virtual Private LAN Service

VSI - Virtual Service Instance

2. Use Cases

RBridge campuses at different locations may interconnect by networks that are implemented with different technologies to form one RBridge campus. This section describes use cases assuming that IP/MPLS technology is available. From the MPLS network view, an RBridge device acts as a Customer Edge (CE) device and connects to PE via an attachment circuit (AC). RBridges [RFC6325] support both point-to-point links and multi-access links. Section 2.1 describes point-to-point link, i.e. TRILL over either Ethernet or PPP point-to-point link that is over an MPLS network. Section 2.2 describes TRILL over a bridged LAN that is implemented by MPLS/VPLS. Section 2.3 introduces a new hierarchical L2VPN solution that uses the RBridges and MPLS tiered architecture.

2.1. Point-To-Point Interconnection

Two RBridges are interconnected by either Ethernet or PPP link that is over a MPLS network. A Pseudo wire (PW) is configured between a

illustrates this use case, in which R Bridges are also MPLS PE enabled devices. The interworking between the R Bridge network and the MPLS PSN is within the device. This has a similar architecture to MPLS/VPLS [RFC4762]. In this case, a virtual Ethernet interface is configured at the R Bridge component; an Ethernet encapsulated PW is configured between two interfaces, which brings up an TRILL link between two R Bridge components. The outer MAC address can be a local Ethernet address. In this case, the Campus R Bridges run in the client layer and MPLS runs in the Server Layer; RB/PE devices support both client and server layer control plane and data plane functions.

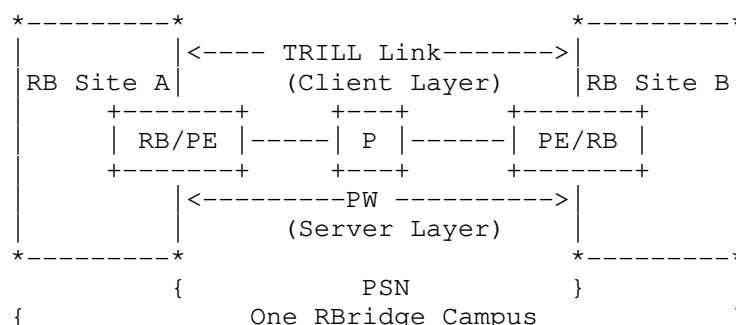


Figure 2 P2P TRILL-Link over IP/MPLS PSN Use Case II

In both case I and II, the PE treats an R Bridge as a generic CE and has no awareness of TRILL capability on the CE. Use case I enables the business models when the R Bridge campus and Core MPLS may be operated by different operators or the same operator. In the case of different operators, the core MPLS operator can sell a VPWS service to R Bridge operator. Use case II provides the model when the R Bridge campus and the core network are operated by the same operator but use different technologies in each network.

Technically speaking, it is possible to create a specially designated TRILL encapsulated pseudo wire for point-to-point TRILL over MPLS. However, the authors think that this is not worth the effort because of available technologies as mentioned above, particularly the highly-efficient PPP link technology.

A PW may cross multiple MPLS domains.[RFC5659] In this case, R Bridges connect to T-PEs and it works in the same way as a single domain. The PSN can provide transport resiliency for a PW. The dual homing (two ACs) can be used for AC protection. In this case, two

TRILL links are established; RBridge device perform load balance over two links.

2.2. Multi-Access Link Interconnection

Multiple RBridges may interconnect via an 802.1Q Bridged LAN that acts as a hub. The bridged LAN simply forwards on the outer Ethernet header of the TRILL frames. This configuration creates what appears to each connected RBridge as a multi-access link. In other words, each RBridge connecting to a bridged LAN has connectivity to every other RBridges connecting to the same LAN.

MPLS/VPLS can provide the same capability when multiple parts of an RBridge campus are interconnected over an IP/MPLS PSN and make each RBridge attaching to the VPLS to appear as having a multi-access TRILL link. Figure 3 shows the use of MPLS/VPLS for RBridge interconnection. One RBridge campus is split between three different sites. One VPLS instance is configured on three PEs and the PWs are configured for the VPLS instance. Each RBridge Site connects to the VSI on a PE via an AC (Ethernet Type). The VSI on a PE forwards TRILL frames based on the outer Ethernet header of the frames. [RFC6325] Either BGP [RFC4761] or LDP [RFC4762] protocol can be used to automatically construct the VPLS instance on the PEs. A PE may connect to several different RBridge campuses that belong to different customers. Separated VPLS instances are configured for individual customers and customer traffic is completely isolated by VPLS instance. The PE treats an RBridge as a generic CE and has no awareness of TRILL.

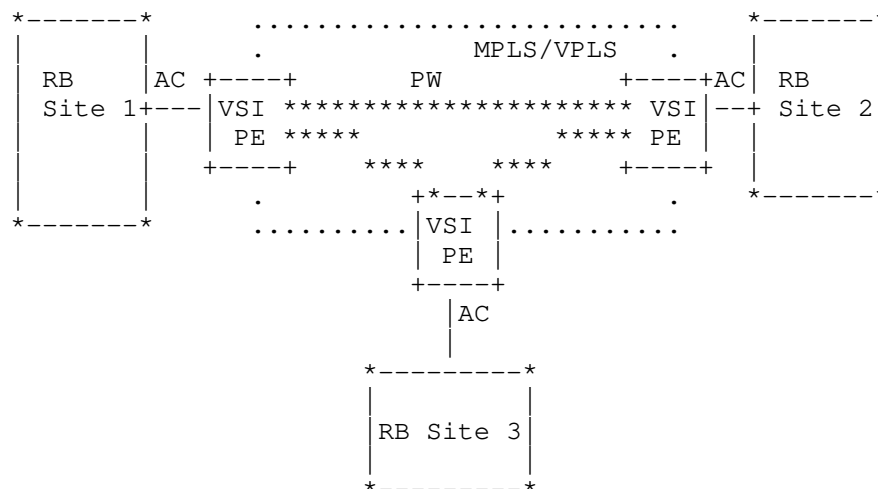


Figure 3 Multi-Access TRILL Link over MPLS/VPLS

The outer Ethernet MAC of TRILL frames may be either a next-hop RBridge MAC address (for unicast frames) or one of TRILL defined multicast addresses (ALL-IS-IS-RBridges and All-RBridges). [RFC6325] The VSI at each PE learns the source MAC addresses on each VSI interface and forward the frame based on the destination MAC. For the multicast frames, the VSI replicates the frames to all PWs it associates. If a VPLS is configured with some optimization capability [VPLS-BCAST], the multicast frames can be delivered over a point-to-multipoint PW while unicast frames are carried over a point-to-point PW.

The scenario in Figure 3 can also be extended to multiple RBridges interconnections when a device serves both RBridge and PE functions. This use case is discussed in the following section.

Note: If the CEs associated with one VPLS instances happen to include some RBridges and some end stations or IEEE 802.1Q bridges to end stations, TRILL will, by default, be able to handle this by providing both through service and end station service. However, the end station addresses will be visible to the VPLS instance. If, in such a case, all the RBridge ports connected to the VPLS are configured as trunk ports (see Section 4.9.2 of [RFC6325]), then they will not provide any end station service.

2.3. Hierarchical L2VPN with RBridges and MPLS

H-VPLS in [RFC4762] describes a two-tier hierarchical solution for the purpose of pseudo wire (PW) scalability improvement. This improvement is achieved by reducing the number of PE devices connected in a full-mesh topology through connecting CE devices via the lower-tier access network, which in turn is connected to the top-tier core network. However, H-VPLS solutions in [RFC4762] require learning and forwarding based on customer MAC addresses, which poses scalability issues as the number of VPLS instances and customer MAC addresses increase. [PBB-VPLS] describes how to use PBB (Provider Backbone Bridges) at the lower-tier access network to solve the scalability issue, in which the transit network nodes only learn and forward on PBB port MAC addresses instead of customer MAC addresses.

RBridges over IP/MPLS provide an alternative solution for a scalable L2VPN over WAN networks. Figure 4 depicts the hierarchical L2VPN architecture with RBridge/MPLS technologies. An IP/MPLS network serves as the top-tier core network function while an RBridge campus serves as the low-tier access network function. A RB/PE enabled device is placed at the boundary between the two-tier networks. A PW is configured between each pair of PE components in the top-tier IP/MPLS network, which constructs a full mesh TRILL links among the RB/PE devices. The RBridge component on a RB/PE device and other RBridges at the same site serves as the low-tier access network. Customer CEs connect to RBridges at each site directly. This architecture provides E-LAN or E-VLAN connectivity among customer CEs connecting to the RBridge campus sites. The transit RBridge node only forwards and learns other RBridge addresses and the number of PWs in the top-tier core network is not relate to the number of devices connecting to Customer CEs. This makes the solution scale very well. In addition, TRILL technology already supports multi-TRILL links from one RBridge to one or multiple RBridges and prevents loops, which provides the flexibility to construct the networks based on their network condition. Figure 4 shows that one RBridge in site 1 connects two RB/PE devices and one RB/PE device connects two RBridges at Site 2 via Ethernet links.

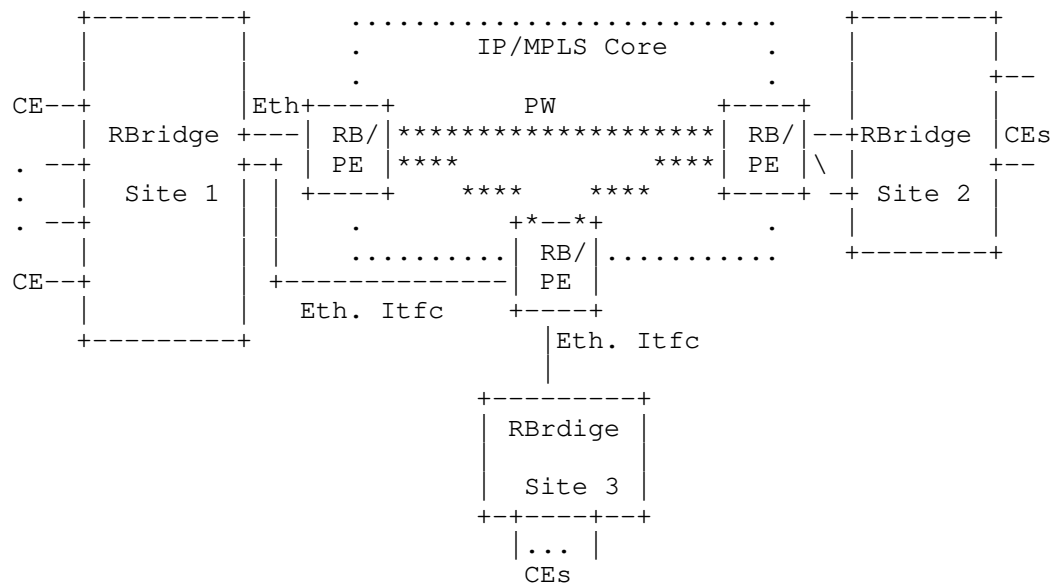


Figure 4 Hierarchical L2VPN with RBridge and MPLS

The following advantages of using RBridge/MPLS based L2VPN: 1) Scalability improvement; 2) Auto-configuration; 3) Good efficiency and loop prevention; and 4) Multipath support.

The solution also has the following advantages: 1) Since RBridge terminates customer spanning tree protocol (STP), individual STPs in attached customer bridged LANs will be separated and will converge faster. 2) low over head per frame in number of added bytes and scalable routing computations; 3) MPLS just provides P2P PWs, MAC forwarding and learning does not exist within the MPLS network, thus multi-homing issue does not exist.

Note: it is good to mention another scenario when the device has both RB/PE capabilities, i.e. configure a VPLS instance among PE components in the top-tier network to provide a multi-access link to RBridge component on the RB/PE devices. Although this solution can also provide scalability, it requires both the RBridge component and VSI/PE component on a device to perform the same MAC forwarding and learning functions, which is redundant. The number of PWs configured in this case is the same as of the number of PWs in Figure 4. Thus, authors do not recommend this configuration. For the same reason, the use case in Section 2.2 is not viewed as the recommended L2VPN solution for the WAN networks. Instead, it is useful when a Core

Service Provider provides a VPLS service to the customer who needs to interconnect the RBridge campus sites over IP/MPLS PSN.

It is possible to construct a Tiered L2VPN in the combination of Figure 4 and 3, i.e. some locations use RB/PE enabled device and some location use separated RBridge and PE devices in a Hierarchical L2VPN. When using separated RBridge and PE devices at some locations, the MPLS network has to run a VPLS instance, which makes RB/PE devices perform MAC forwarding and learning function two times. In addition, it becomes operator responsibility to ensure that the top tiered MPLS core is fully surrounded by an RBridge campus. Missing configuration may increase the scalability problem in the core network.

Auto configuration for the Hierarchical L2VPN will be addressed in another draft.

3. RBridge Behavior for MPLS Pseudo Wire

This section describes RBridge behaviors for TRILL Ethernet or TRILL PPP links over MPLS pseudo wire (PW) as described in Sections 2.1 .

1. For two RBridge ports connecting via a PPP PW, the ports MUST be configured as IS-IS point-to-point. Thus TRILL will use IS-IS P2P Hellos that, per "Point-to-Point IS to IS Hello PDU" (section 9.7 of [IS-IS]), do not use Neighbor TLVs in the same manner as on a multi-access link. However, per section 4.2.4.1 of [RFC6325], three-way IS-IS handshake using extended circuit IDs is required.
2. For two RBridge ports connecting via an Ethernet PW, it is RECOMMENDED that the ports be configured as IS-IS point-to-point for the same reason able. Note: an RBridge port by default supports multi-access links.
3. Any MPLS forwarder within an MPLS PSN does not change the TRILL Header Hop Count. RBridges is never aware of the packet forwarders in MPLS PSN.
4. If it is desired for MPLS PSN to perform QoS in the same way as in the RBridge campus, RBridges MUST be configured to send an Outer.VLAN tag on the RBridge port. The PE can then copy the priority value from the Outer.VLAN tag to the COS filed of the PW label prior to the forwarding. [RFC5462]

5. TRILL MTU-probe and TRILL MTU-ack messages (section 4.3.2 of [RFC6325]) are not needed on a pseudo wire link. Implementations MUST NOT send MTU-probe and SHOULD NOT reply to these messages. The MTU pseudo wire interface parameter SHOULD be used instead. PE Must configure the MTU size as the originating RBridges Size specified in Section 4.3.1 of [RFC6325].

4. Security Considerations

The IS-IS authentication mechanism [RFC5304] [RFC5310], at the TRILL IS-IS layer, can be used to prevent fabrication of link-state control messages over TRILL links including those discussed in this document.

For general TRILL protocol security considerations, see [RFC6325].

The use case does not introduce any security considerations for MPLS network.

5. IANA Considerations

No IANA action is required by this document.

6. Acknowledgements

The authors sincerely acknowledge the contributions of Ben Mack-Crane and Sue Hares.

7. References

7.1. Normative References

- [RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels," BCP 14 and RFC 2119, March 1997
- [RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.
- [RFC4447] Martini, L., etc, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC4447, April 2006.
- [RFC4448] Martini, L., "Encapsulation Methods for Transport of Ethernet over MPLS Networks", BCP 116, RFC 4446, April 2006.

- [RFC4618] Martini, L., "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", BCP 116, RFC 4618, September 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and Kompella, V., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC4762, January 2007
- [RFC5304] Li, T. and Atkinson, R., "IS-IS Cryptographic Authentication," RFC 5304, October 2008
- [RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009
- [RFC5462] Andersson, L. and Asati, R., "Multiprotocol Label Switching (MPLS) Label Stack entry: "Exp" Field Rename to "Traffic Class" Field", RFC5462, February 2009
- [RFC5659] Bocci, M and Bryant, S., "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A.Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC6325, July 2011.
- [RFC6326] Eastlake 3rd, D., Banerjee, A., Dutt, D., Perlman, R., and Ghanwani, A. "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC6326, July 2011.
- [RFC6361] Carlson, J., and D. Eastlake, "PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol", RFC6361, August 2011.

7.2. Informative References

- [IS-IS] International Organization for Standardization,
"Intermediate system to Intermediate system intra-domain
routing information exchange protocol for use in
conjunction with the protocol for providing the
connectionless-mode Network Service (ISO 8473)", ISO/IEC
10589:2002, Second Edition, Nov 2002
- [VPLS-BCAST] Delord, S, and Key, R., "Extension to LDP-VPLS for
Ethernet Broadcast and Multicast", draft-ietf-l2vpn-ldp-
vpls-broadcast-exten-02, work in progress, 2011.
- [PBB-VPLS] Sajarssi, A, etc, "VPLS Interoperability with Provider
Backbone Bridges", draft-ietf-l2vpn-pbb-vpls-interop, work
in progress, 2011

Authors' Addresses

Lucy Yong
Huawei Technologies (USA)
5340 Legacy Drive
Plano, TX 75025

Phone: +1-469-277-5837
Email: lucy.yong@huawei.com

Donald E. Eastlake, 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Sam Aldrin
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050

Phone: +1-408-330-4517
Email: sam.aldrin@huawei.com

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134

Phone: +1-408-333-4062
jon.hudson@brocade.com

