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

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

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:

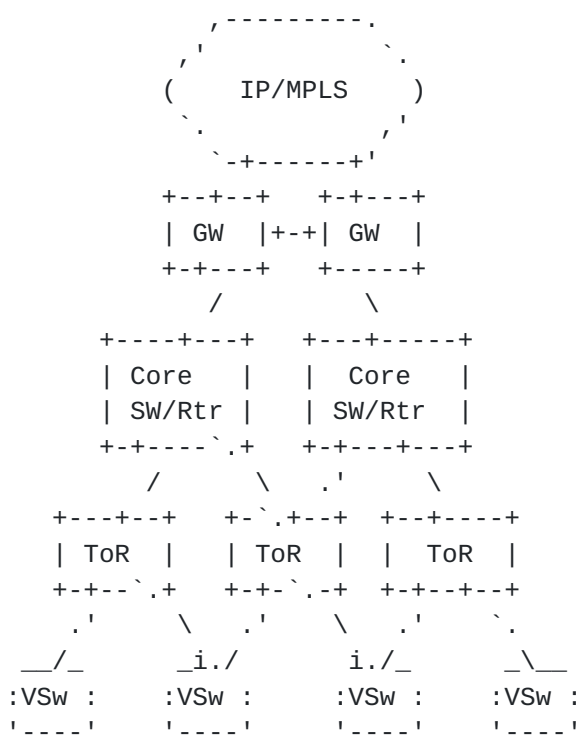

Figure 1 : A Generic Architecture for Cloud Networking

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:

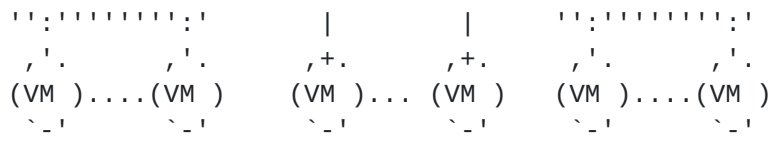

Figure 2 : Logical Service connectivity for one tenant

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 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 is 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]):

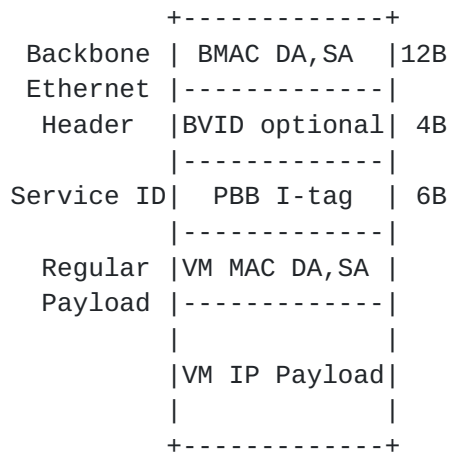

Figure 3 PBB encapsulation

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:

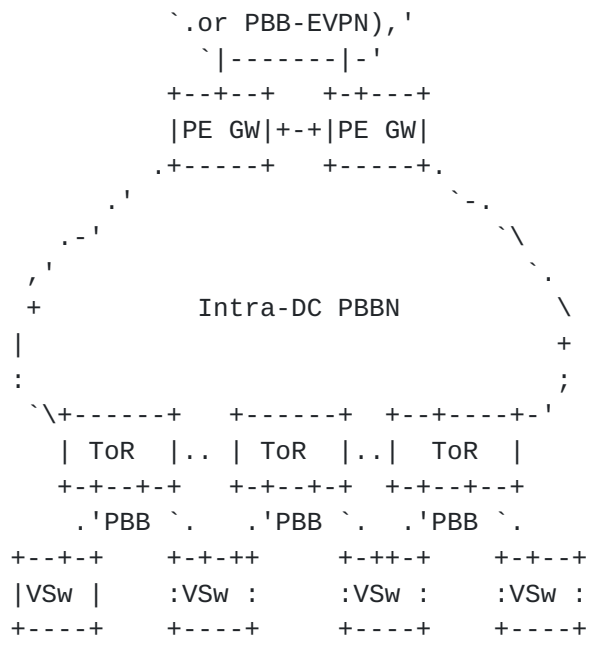

Figure 4 PBB in DC, PBB-VPLS or PBB-EVPN for DC Interconnect

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 than 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, MultiProtcol (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:

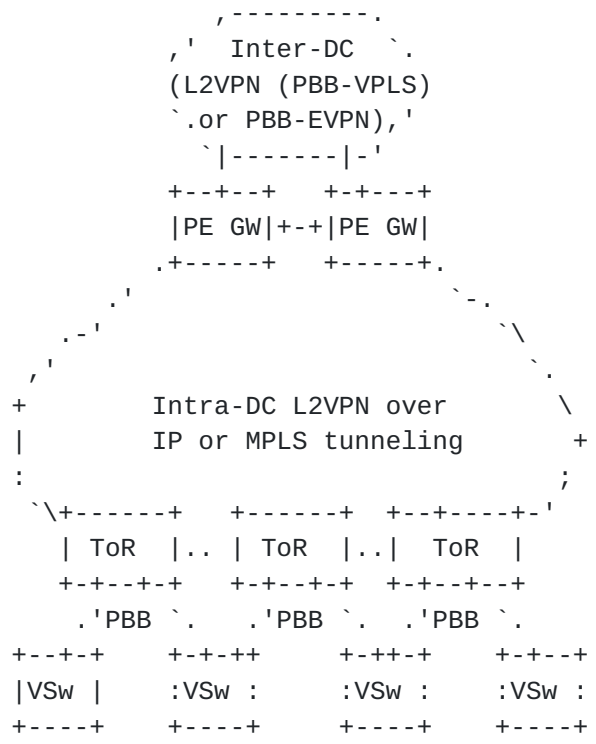

Figure 5 PBB in VSw, L2VPN at the ToR

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:

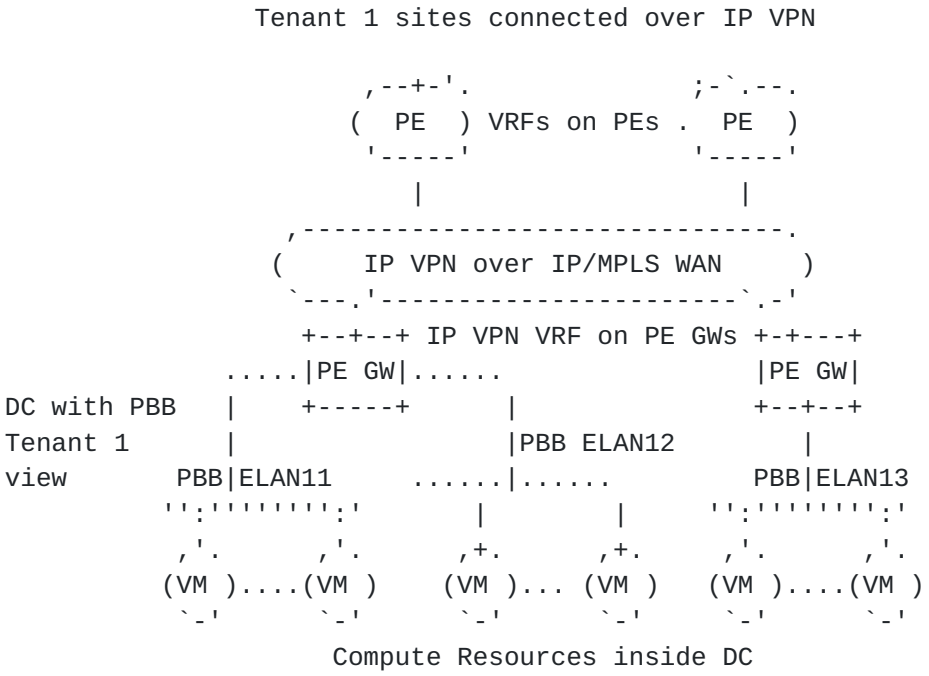

Figure 6 Logical Service View with IP VPN

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:

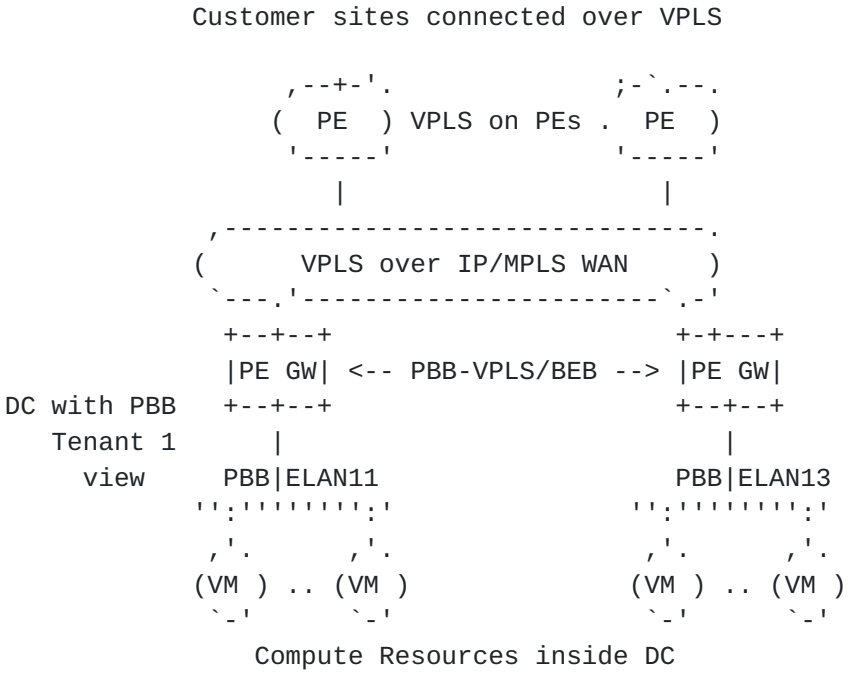

Figure 7 Logical Service View with VPLS WAN

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:

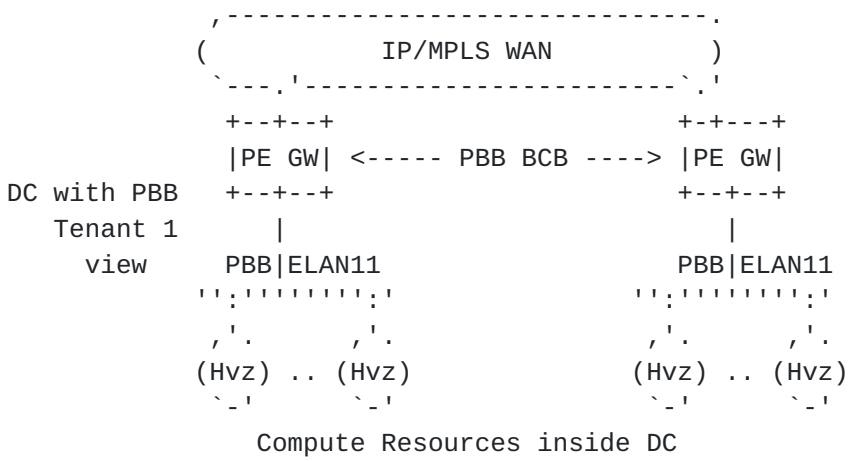

Figure 8 PBB BCB providing VMotion ELAN

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:

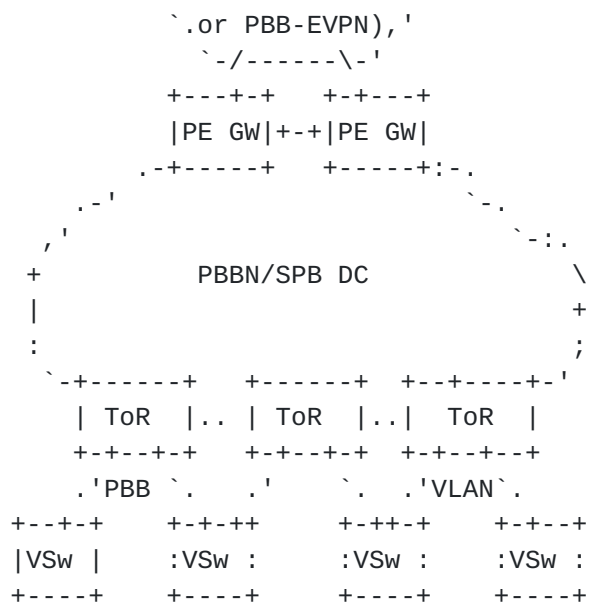

Figure 9 DC with PBB and VLANs

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:

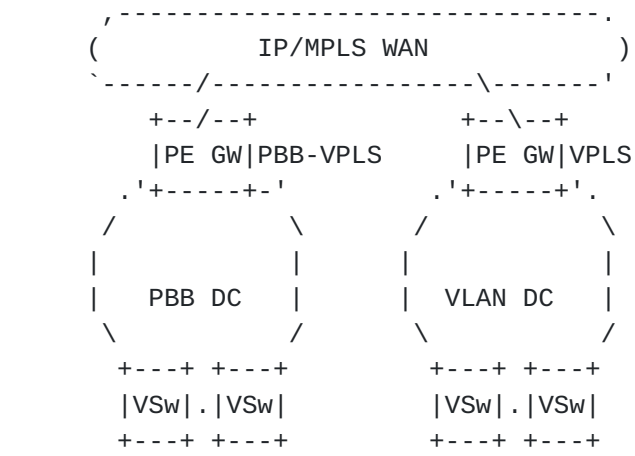

Figure 10 Interoperability to a VLAN-based DC

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.

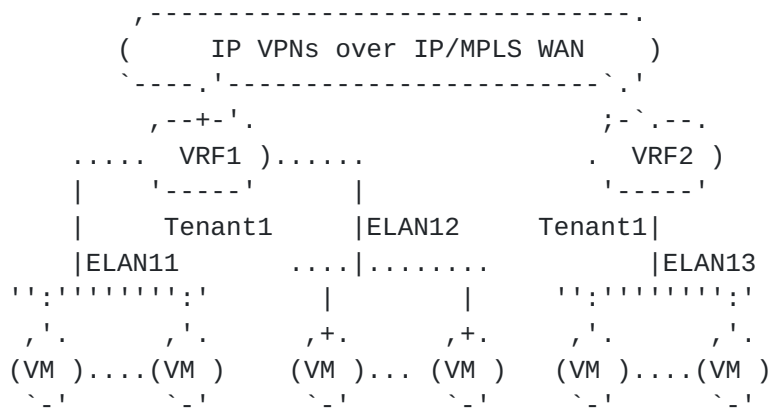

Figure 11 Logical Service View with IP VPN

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 fledge 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.

[ARPproxy] 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