E-VPN—Contemporary Layer 2 Interconnect

Table of Contents

Executive Summary

This white paper describes the Layer 2 interconnect options available for today's network designer. We highlight MPLS E-VPN architecture as the newest option for high-scale cloud, data center, and private networks.

The Beginnnings

In September 1980, Digital Equipment Corporation, Intel, and Xerox published a first specification for a LAN standard with 48-bit destination and source station addresses. Colloquially known as "Ethernet", this new link-layer standard quickly went on to dominate LAN environments as open and inexpensive technology designed to work over diverse media. While the first Ethernet networks were based on the coaxial cable and collision domain concept, newer and faster versions emerged to run over twisted pair and optical fiber in the full-duplex mode, with IEEE 802.3 committee constantly working on standard improvements. Thirty-five years after its conception, Ethernet is still the most ubiquitous wire technology, joining a massive amount of consumer, business, and service provider equipment over residential, local and wide area networks.

This situation have inevitably led to the need to connect large islands of Ethernet connectivity (datacenter) over the long-haul links. Such links together with infrastructure that permits link-layer data exchange (routers and switches) are widely known as L2 interconnect.

As often happens in the technology world, the problem of L2 wide area interconnect can be solved in a multitude of ways, which evolved over time as customer requirements grew more sophisticated. This set of competing interconnect options can be intimidating for network engineers and architects in the need of robust and agile Ethernet services, especially in those cases where the cost of design errors can be unacceptably high. The focus of this paper is to explore scenarios where the architecture can make a critical difference to performance of L2 interconnect.

If you are a reader with a basic level of L2 interconnect knowledge, we advise reading this text from the beginning. Network engineers familiar with legacy data plane-based interconnect options such as VPLS and provider backbone bridging (PBB), who are interested in knowing more about control plane-based deployment scenarios, should proceed to the basics of BGP MPLS Ethernet VPNs in the section titled "Legacy L2 Interconnect Options". Engineers already well-versed in E-VPNs may choose to move directly to the section addressing the common scenarios of media access control (MAC mobility) and link failover, and to the review of our "Conclusion" section.

L2 Interconnect Introduction and Use Cases

Up until the mid-1990s, the use of Ethernet connectivity was limited to LAN connections, with a clean separation between layers of Ethernet switching (L2) and routing (L3) in place. This worked well because the traffic destined for the expensive WAN links was meant to be tightly controlled, while Ethernet segments were considered inexpensive shared media (CSMA/CD) domains with ancillary loop avoidance services (spanning tree). In the early days of L2 interconnect, the main business case for sending Ethernet frames over WAN was in support of legacy multiprotocol environments, which were too complex for routers to handle natively. This use case evolved via LAN emulation built into WAN L2.5 protocols (such as cell-based ATM LANE) and later inherited by packet-level tag technologies in the form of MPLS. However, by 1999, most formerly multiprotocol enterprise networks had moved to IP, and the need for point-to-multipoint (P2MP) LAN emulation was waning quickly. The duopoly of IP and MPLS has emerged to control end-to-end packet propagation, and bridged Ethernet segments now seemed to be a thing of the past.

With an IP stack running on every machine in a data center, it was only natural to move to L3 and forget the intricacies of spanning tree operations, topology loop avoidance, and broadcast storms seen in the bridged environment. While point-to-point Ethernet (PWE) were widely sold as a commercial service, deployment of P2MP Ethernet (bridged segments) was in decline.

However, in an example of spiral technology development, it was not long until Ethernet segment interconnect over wide area links came back. The first serious wave of interest to the topic sparked around mid-2000, when a consortium of Ethernet switch vendors came up with the idea of MAC header stacking for connectionless packet routing in carrier networks. While the Ethernet frame format has long supported a 12-bit VLAN ID field, the technology called "double-tagging" was proposed by Nortel (and later implemented as an IEEE 802.1ad standard), allowing the use of a header stack for more than 4,096 multiplexing channels (see Figure 1).

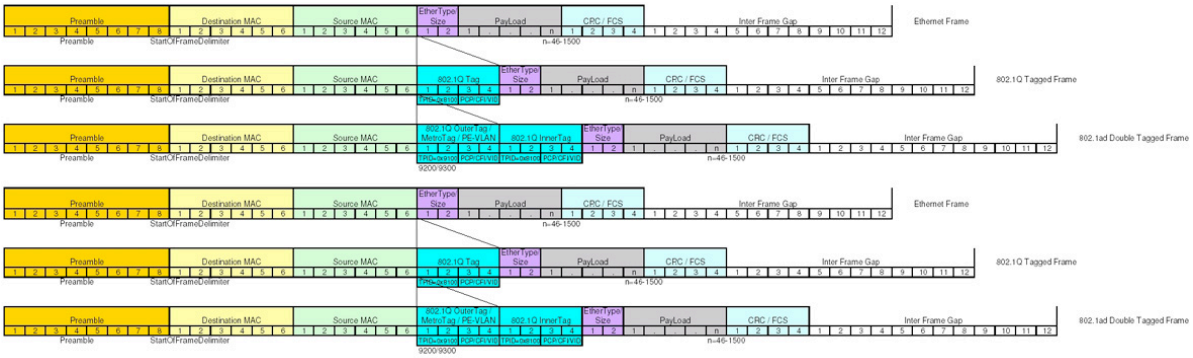


Figure 1: Plain 802.1q and double-stacked 802.1ad frame formats (source: Wikipedia)

Later work in this direction led to a stack generalization in the 802.1ah 2008 standard, where the entire backbone-specific Ethernet headers could be added to customer-level packets.

In theory, this technology (later known as PBB) could enable wide area transport to run entirely on Ethernet switches. MAC-in-MAC encapsulation was going to be used in place of MPLS tags, with bridge hierarchy supporting point-to-point and multipoint interconnects. As such, PBB directly competed with both MPLS pseudowire and virtual private LAN service (VPLS). The main rationale behind PBB was the ability to use inexpensive Ethernet switches in place of the carrier routers, since operations were limited to MAC table lookups and push/pop Ethernet header operations.

This development, however, proved to be short-lived.

One reason behind its demise was that PBB tagging was less efficient than MPLS in terms of the overhead (132 bits in B-DA, B-SA, and B/I-tags in 802.3ah header versus 32-bit MPLS label). Another reason was that off-the-shelf Ethernet switches were not ready for deployment as carrier platforms, with many critical control and data plane features missing. Finally, but perhaps most important, the control plane of PBB dubbed “provider backbone transport” quickly found itself needing to duplicate well-proven resiliency, traffic engineering, and troubleshooting aspects of MPLS. This raised questions about the viability of a new technology and the cost of MAC-level push/pop versus equivalent tag manipulations. As it turned out, swapping and popping MAC addresses was neither faster nor cheaper than doing the same with fixed-length MPLS tags.

Once the shortcomings of the new technology became evident, a business case for wide area bridging largely dissolved, leaving behind only the MAC-in-MAC encapsulation, which survives in hierarchically bridged networks until this day.

This brief detour into the history of L2 interconnect in wide area environments should not distract us from the main question, which is why we are still talking about L2 Ethernet segment transport in carrier networks. As of 2013, the primary reason for that, rather interestingly, is endpoint virtualization. Pioneered by VMware with ESX series products in the mid-2000s, host virtualization propelled data centers into a new era by breaking the link between physical machines and their service-oriented workloads. One implication of that was the widespread use of Ethernet MACs as virtual machine anchors and reference points for station mobility, replication, and failover. This “step back in time” happened precisely because VM supervisors were intended to be transparent to guest operating systems and initially could not offer any abstraction layers other than virtual network interface card services, which pushed the complexity of L2 routing onto the datacenter communication devices.

This development had three major implications for data centers: (1) Ethernet segments suddenly got L2 traffic that no longer had a 1:1 correlation with IP addresses; (2) once again, it became necessary to use VLANs to separate host clusters; and (3) in the event of a virtual machine migration (or failover), the service could move over wide area links and into remote locations.

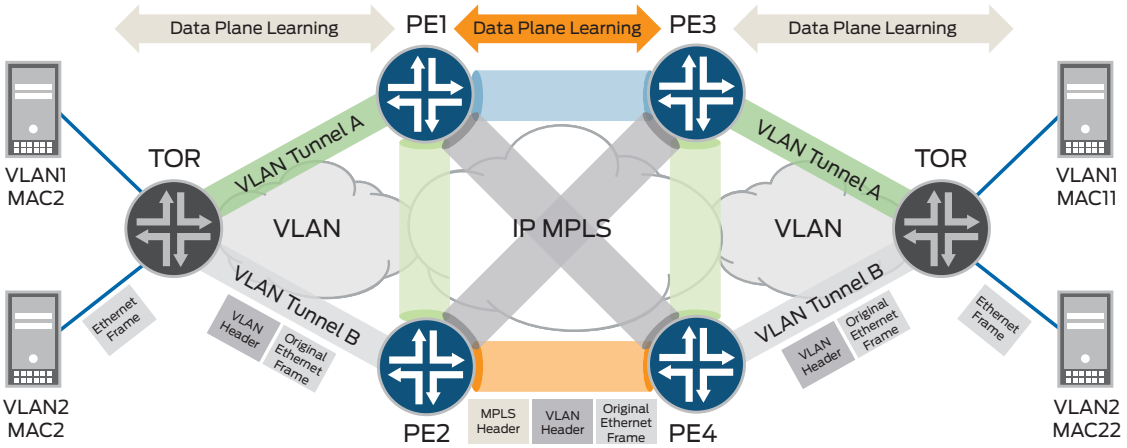
Although the contemporary data center architectures are moving toward supervisor-based routing architectures (such as those supported by Juniper Networks Contrail or VMware Nicira), the ability to support L2 Ethernet segments over WAN natively still remains the lowest common denominator of distributed data centers. To this end, Juniper supports the development and implementation of modern L2 interconnects and offers this functionality in our edge and DC product lineups.

In this paper, we will further assume that the point-to-multipoint L2 connectivity to remote data centers is a requirement, with resiliency and fault tolerance expectations similar to those found in local area broadcast domains.

Legacy L2 Interconnect Options

As we briefly touched on in the Introduction section, the first Ethernet L2 segment support architectures over wide area links were built around LAN emulation. In effect, both ATM LAN emulation (LANE) and MPLS VPLS are quite similar and designed to faithfully replicate the behavior of transparent bridges over non-broadcast media.

The operation of transparent bridge is very simplistic and has not significantly changed since the earliest days of Ethernet. In the most elementary implementation, the pseudo-bridge in a provider edge (PE) router forms a full mesh with all PE endpoints over generic routing encapsulation (GRE) or MPLS tunnels, and listens to incoming packets (typically VLAN-tagged) on its LAN-facing ports (green and gray "VLAN tunnel" links as shown in Figure 2). When a broadcast or multicast packet is seen, it is replicated and forwarded over all outgoing IP/MPLS tunnels. When a unicast packet is seen, it is matched to a forwarding destination in a MAC forwarding table (FIB) of an ingress PE router. This table is the main source of all forwarding decisions and is reactively built based on transit packets. If a match is found, this packet is forwarded to a remote (egress) PE; otherwise it is flooded to all destinations and added to MAC FIB. As is the case with conventional bridges, MAC addresses in emulated bridge FIB tables are subject to aging, flushing, and learning; any L2 connectivity information is, therefore, limited to naive learning mechanisms of the data plane and is not necessarily consistent across all PE devices.



- Multipoint Service Connectivity over MPLS Transport.
- Customer MACs are learnt in Data Plane from Access as well as from core side.
- Active-Standby Multihoming – only one PE is active, while others in dormant state.
- Requires flush of all learnings and flooding and relearning after switchover.
- Switchover delays may cause traffic loss.

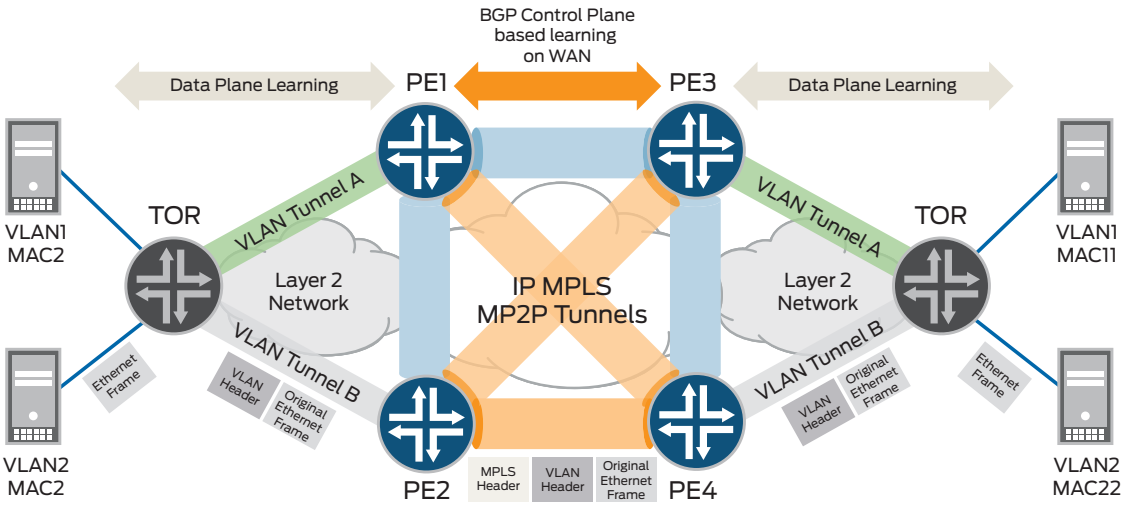
Figure 2. VPLS operation example in the data center environment

Over time, VPLS has received many extensions for hierarchical forwarding decisions, endpoint autodiscovery (AD), PBB encapsulation support, multipoint label-switched paths (LSPs), and other features. Some vendors have also invested in hybrid VPLS/PBB architectures to address the scaling requirements of data centers with switch hierarchy. However, at its core, VPLS has remained the "ships in the night" technology with no interaction between the WAN and emulated Ethernet segments. As such, VPLS inherits restrictions similar to those of legacy bridges—reliance on Spanning Tree Protocol (STP) for path restoration; possibility of L2 forwarding loops; and no native ability to work in the multipath/load-balancing environment.

Therefore, it is best to consider VPLS as a lowest common denominator choice when interoperating with low-end Ethernet access and legacy bridged networks; in this capacity, VPLS and PBB-VPLS will likely continue to to co-exist with more advanced L2 interconnect technologies.

The Rise of the E-VPN

The need for improvement over VPLS became evident shortly after L2 interconnect became popular in data centers. The main motivation for change has come from the fact that a local PE can be acutely aware of WAN state and does not have to pretend to be a dumb bridge.



- Multipoint Service Connectivity over MPLS Transport.
- Remote Customer MACs are learnt in Control Plane.
- Active-Active Multihoming.

Figure 3. Generic E-VPN topology layout

More specifically, the path to a remote Ethernet segment that belongs to the common L2 domain is well-known to L3 routing protocols; moreover, the target segment can be discovered behind multiple PE routers (multihomed) or have multiple path reachability options (multipathed). With this information exchanged in the control plane, local PE routers can make very intelligent reachability, failover, and load-balancing decisions via policy-driven routing protocol extensions such as BGP with new network layer reachability information (NLRI). This migration from pseudowires connecting emulated bridges to explicit reachability information exchanges between PE routers allows for steering of unicast L2 interconnect traffic directly to multipoint-to-point (MP2P) tunnels (broadcast, unknown unicast, and multicast frames still require flooding in manner similar to VPLS).

In all E-VPN architectures, the control plane connections between peering PE routers invariably carries the following reachability information: (1) 10-octet Ethernet Segment Identifier (ESI); (2) 6-octet remote MAC address; and (3) 3-octet MP2P label that corresponds to pairs of (1) and (2). The treatment of this information is, however, different between E-VPN flavors.

In IETF publication draft-ietf-l2vpn-evpn[1] describing E-VPN framework and architecture, every client MAC address learned by the data plane of a PE router on a given Ethernet segment is assumed to be advertised to every remote PE router. This means that the control plane carries the full MAC connectivity table across the virtualized L2 segment. The same framework, however, can be used with a bridge hierarchy to conceal the total number of stations in a segment, as described in another IETF proposal, draft-ietf-l2vpn-pbb-evpn[2]. In this draft, only backbone-bridge addresses (B-MACs) are meant to be carried as control plane routes, while client-station MACs (C-MACs) are learned reactively, when seen on the local or remote (wide area) side of the connection. For brevity, we will refer to the former draft as simply E-VPN and to the latter as PBB-EVPN.

At this point, a careful observer should have already noted that both E-VPN and PBB-EVPN proposals, in addition to a route convergence much faster than in STP, should be able to support active/active configuration of remote PE routers. This is, indeed, the case because the control plane operation allows to ignore the erroneous learning of remote MAC addresses on a local interface when forwarded by an active peer (as it would be done by VPLS); the only special case is the delivery of broadcast/unknown/multicast (BUM) traffic, which requires that at least one of the PE routers is connected to a segment to act as a designated forwarder (DF). In the course of the normal forwarding, E-VPN traffic may utilize load balancing in the backbone network as well as across multiple active/active PE routers. This is a major improvement over VPLS in the data center, as it allows for higher cross-sectional network bandwidth and reduced traffic impact should node failure be detected.

[1] IETF draft on E-VPN: http://http://datatracker.ietf.org/doc/draft-ietf-l2vpn-evpn/
[2] IETF draft on PBB-EVPN: http://https://datatracker.ietf.org/doc/draft-ietf-l2vpn-pbb-evpn/

On the software implementation side, while PBB-EVPN uses the subset of E-VPN BGP messages and attributes, it also requires that the PE router support the PBB edge bridge (BEB) functionality, so not every E-VPN device is necessarily capable of PBB-EVPN operation. As of 2013, the two drafts are being developed and implemented in service provider equipment in parallel.

Bridge Hierarchy

A question that frequently comes up in greenfield deployments is whether bridge hierarchy is a good scaling tool in the first place. The proponents of PBB[3] and related technologies (PBB-VPLS[4] and PBB-EVPN) sometimes make a point that introduction of bridge stacks improves scaling limits, as the control plane can now be freed from client MAC information (see Table 1). This assertion, however, does not always prove to be true.

As we have explained in the Introduction section, the use of PBB does not affect the data plane's scaling limits, because each remote station still needs to be learned and installed into a MAC table for forwarding to work properly. Moreover, PBB can also be less efficient, as bridge hierarchy is achieved by pushing an extra 802.1ad header onto the Ethernet frame (which still has an outer tunnel header to cross the WAN). An extra header can be a notable overhead in small packets and also requires additional lookup cycles in PE routers. This leaves us with the question of why bother with bridge hierarchy in the first place, if it is entirely ephemeral (that is, exists only as an abstraction within PE routers).

Table 1: Scaling Limits of L3 Interconnects

Technology	Control Plane Scaling	Control Plane Limit	Dataplane Limit
PBB-VPLS	Hierarchy of bridges	#PWs = (PE-1)	C-MAC routes
VPLS	None	#PWs = (PE-1) x bridges	C-MAC routes
PBB-EVPN	Hierarchy of B-MACs	B-MAC number is CP	C-MAC routes
E-VPN	Route reflectors	C-MAC number in CP	C-MAC routes

One possible answer to our question is that a stack of virtual backbone bridges conserves routing information in the same fashion as the interior gateway routing protocols (such as OSPF or IS-IS) restrict propagation of reachability behind summary prefixes of connected networks (advertise only B-MACs and not client C-MACs). This analogy, however, is incomplete because L2 networks are flat and have no inherent prefix summarization. Even if virtual MAC addresses can be allocated from segment-specific pools, L2 interconnect is still required to support host routing—that is, an option to move any MAC address to any segment. From this perspective, it can be prudent to carry full L2/L3 reachability information in the control plane, which is exactly what E-VPN does.

Yet another challenge with bridge hierarchy arises with load balancing across multiple uplinks. Since traffic to multiple C-MACs behind the same PE router has an outer Ethernet header with an identical B-MAC address, most MPLS switches will not be able to execute a hash function correctly (it will require parsing an outer Ethernet header, at least two MPLS headers, B-MAC header, and C-MAC). This is not a problem for E-VPN, since it encodes segment reachability within the MPLS tag and thus is not dependent on the 6-level parsing capabilities of a forwarding chipset.

L2/L3 Mobility and Resiliency

What also makes E-VPN technology stand out over PBB-EVPN is its ability to maintain consistent forwarding information in L2 and L3 domains, thus avoiding unnecessary floods, triangular routing, and L3 connection timeouts. This is achieved by the following mechanisms:

- C-MAC reachability propagation in BGP updates
- MAC flush messages
- Ability to maintain a relation between a station'tation to maintain a

To understand why this can be important, let's consider an example (see Figure 4).

[3] 802.1ah Provider Backbone Bridges: http://www.ieee802.org/1/pages/802.1ah.html

[4] IETF draft on PBB-VPLS: https://datatracker.ietf.org/doc/draft-ietf-l2vpn-pbb-vpls-pe-model/

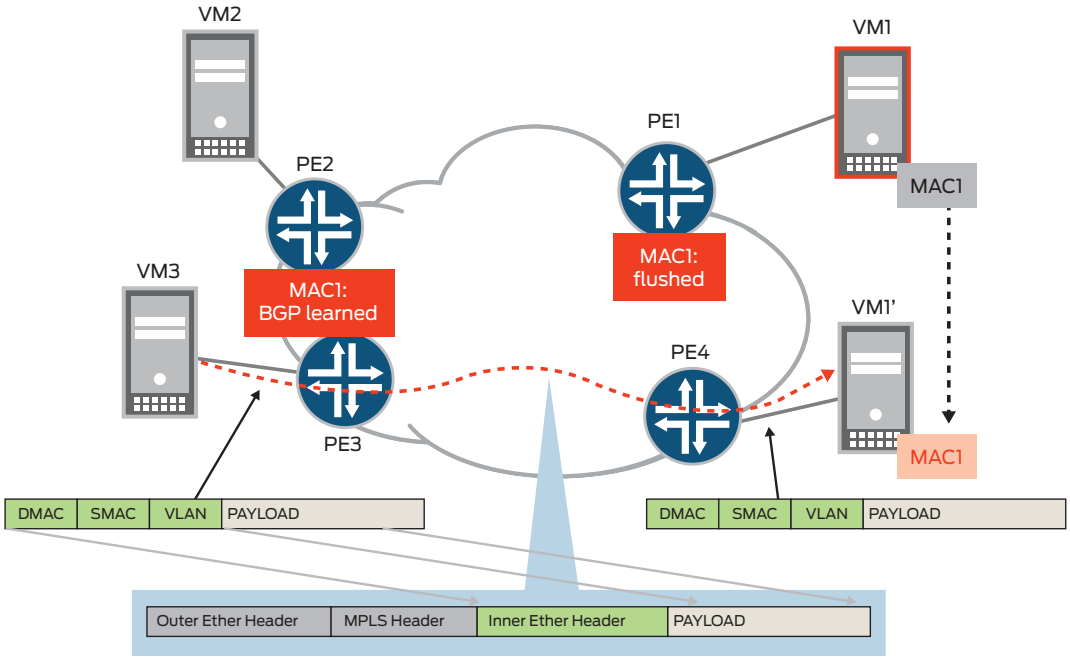


Figure 4: MAC mobility between PE2 and PE3 in E-VPN

Let's suppose we have station VM1 with IP address IP1 and virtual address MAC1 that is migrating across the wide-area link (PE1 to PE4). Such an event may happen in the case of a data center server replication, or during a failure that requires moving a service to another physical location.

After migration, station VM1 must be reachable via at least two different tags: L2 address MAC1 (for intra-cluster transactions) and L3 address IP1 (for connections from outside of its Ethernet segment) as a new VM1 instance. In E-VPN protocol, upon detecting the loss of MAC1 on a locally connected segment, PE1 sends either a MAC flush message to all other PE routers, or withdraws its former MAC1 announcement the first time it sees MAC1 advertised by PE4. In both cases, PE2 and PE3 immediately learn about the change and the network converges. When VM3 sends traffic to MAC1, it no longer goes to PE1 and uses the shortest route to PE4. To prevent loops in the case that VM1 moves back and forth between PE1 and PE4, E-VPN employs a non-decreasing extended community attribute to keep track of the "latest" appearances of a mobile virtual machine (VM) on the network.

Moreover, E-VPN policy can be configured to advertise a host IP route in conjunction with MAC migration. This ensures that VM1 is immediately reachable by clients within its own Ethernet segment (LAN) and from the routed segment. It is an example of how the granularity of E-VPN helps keep the network state coherent.

In the same situation, PBB-based L2 interconnect will exhibit the following problem:

- Let's say that VM1 migrated from PE1 to PE4, but did not send any broadcast or WAN-bound traffic that PE4 could see. In this case, routers PE2-PE3 all have the wrong forwarding information about MAC1 (B-MAC PE1, C-MAC M1) and attempts to reach it via PE1 will fail. At this moment, MAC1 can only be reached directly via the Ethernet segment behind PE4.
- Then let's assume that PE4 learns about MAC1 (when, for instance, VM1 accesses the Internet). This causes PE4 to update the local MAC forwarding table and remove B-MAC1, considering VM1 to be local. However, PE4 still has no means to inform PE2 and PE3 about this important discovery and the latter will continue sending traffic destined for VM1 to PE1.

In this scenario, network state will remain inconsistent until VM1 starts sending a broadcast or targeted unicast to all remote sections of its Ethernet segment, which will allow other PE routers to update FIB. At the same time, even if all PE routers now know that MAC1 is behind C-MAC PE4, they will not be able to install a Layer 3 host route to VM1. This makes it challenging to run IP-based services in parallel with VM mobility, because now an overlay function is needed to inject /32 routes based on MAC address movements.

Propagation of C-MAC addresses across the intervening BGP infrastructure allows the designer to use it as an anchor for connection resiliency. System behavior at node and segment failure events is straightforward and includes the removal of forwarding next-hop information based on the loss of a BGP session with an affected node, or based on a BGP segment withdrawal message. This is by far the easiest way to achieve resiliency in an L2 interconnect environment.

Scaling of L2 Interconnect and SDN

One argument that was frequently brought in the past against the "evolved control plane" architectures (such as E-VPN) is the perceived complexity of large-scale message processing on the routing nodes.

Indeed, one can maintain that an L2 segment with tens of thousands MAC addresses may cause E-VPN headends to expend valuable CPU cycles to maintain MAC FIB tables; such activity can be seen superfluous on a routing engines normally busy with WAN topology maintenance, fast recovery and state replication.

It is, however, interesting to observe that over the last few years, the computing resources available to network nodes are becoming much cheaper as networks are moving towards software-defined (SDN) architectures. As a result, high-end computing nodes (such as x86 machines employed as dedicated route reflectors or NFV processors) are now commonly available in carrier facilities at low cost. For one example, Juniper Networks CSE2000 carrier-grade network appliance provides an Intel Xeon 8-core 1.8Ghz CPU in a hardened package at a small fraction of a cost of a dedicated route-reflection router. For another example, computational elements are now routinely accessible to the carrier network in datacenters, where SDN controllers (such as Juniper Contrail) can create and manage services at a minimal cost. In any case, fault tolerance and resiliency of BGP architecture allows for inexpensive message processing and effectively removes the requirement for high-end control plane hardware to be physically embedded into routers and switches. Thus, virtualization of control plane processing allows, among many other things, for efficient scaling of E-VPN MAC FIBs.

Moving forward, we can definitely project deeper synergy between software-defined networks and L2 interconnect needs as the former development is a step towards more intelligent, centralized forwarding planes that extracts complexity from constituent network nodes. This trend clearly works against PBB-like bridge hierarchies as compression of control plane state (in the form of bridge stacking) is neither necessary nor conducive to fast convergence; on the contrary, an argument can be made that an evolved control plane would be preferred in the environment when the computing resources are free.

Placement of route reflectors in the E-VPN architecture is relatively straightforward and does not call for the dedicated BGP nodes; the same computational facilities can be used to replicate information in different BGP NLRI families. With that, connected datacenters are the most convenient locations for route reflectors stacks, followed by router-tethered computing appliances. Note, that a typical dual- or triple- star BGP reflection architecture does not make any assumptions about reliability of reflector hardware; therefore, non-redundant x86 systems running a virtualized instance of Junos® can be completely adequate for this job.

Conclusion

While the variety of solutions for L2 interconnect can be intimidating, the choice of the right technology can be simplified by looking at strong and weak points of available standards.

If support for legacy/widely heterogeneous sites is a requirement, VPLS stands out as the lowest common denominator. The PBB-VPLS variety, while being useful in certain scaling scenarios, is generally not recommended because it does not significantly add to the technology and introduces a whole new layer of complexity. By way of contrast, the control plane-based solutions (E-VPN and PBB-EVPN) are more sensible, as they can both fit into the modern standards of scaling, availability, and resiliency. Thus, for greenfield deployments the choice comes down to control plane-based protocols.

An attempt to choose one BGP-based technology over another is more complex, however, as it needs to include operational considerations. For one example, E-VPN can be a better choice when VM mobility is important, especially when using data centers for providing services over wide area links. The use of E-VPN removes the need for a separate L3 mobility protocol and does not require concerted B-MAC allocation across PE routers. The fact that every C-MAC is carried in the control plane also makes it straightforward to troubleshoot, as only one BGP route must be checked to locate a station (as opposed to multiple FIB stages of different bridges).

For another example, the use of route reflectors (physical or virtual) may become a requirement in the network with many PE routers. Although using route reflectors is a good and long-standing practice of carrier networks, it is our recommendation to extend this practice over to all BGP-based networks running at scale. Virtual route reflectors (control plane only routers) are especially suitable for L2 interconnect environments, where computation facilities are easily available.

For further help deploying E-VPN solutions, please contact your local Juniper sales office:
http://www.juniper.net/us/en/contact-us/

If you are interested in general information on L2 interconnects, please consider the following resources:

Juniper Networks Junosi operating system technical documentation: **www.juniper.net/techpubs/**

Inter-Data Center Mobility with VMWare (**VPLS solution brief**)

About Juniper Networks

Juniper Networks is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at www.juniper.net.

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.0.207.125.700
Fax: +31.0.207.125.701

2000553-001-EN Sept 2015