CHAPTER 3

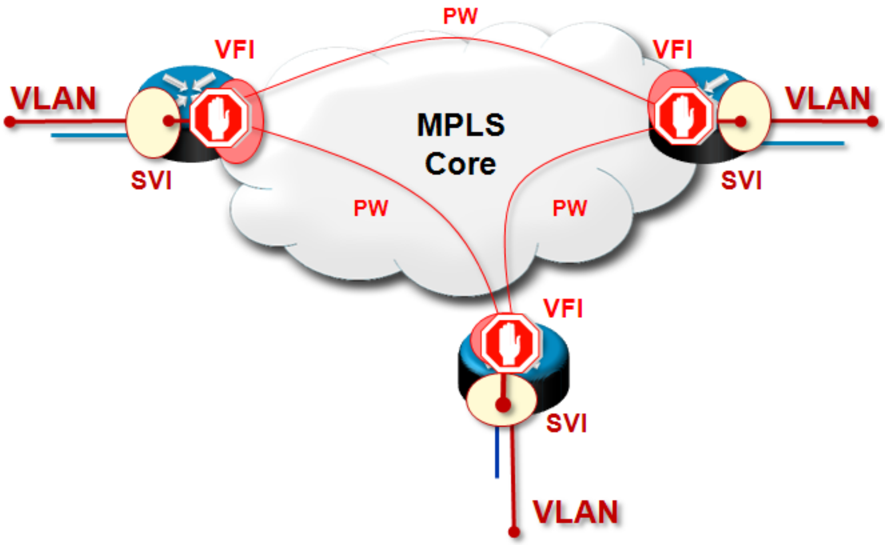
MC-LAG to VPLS Technology and Solution Overview

Virtual Private LAN Service (VPLS) is an architecture that provides multipoint Ethernet LAN services, often referred to as Transparent LAN Services (TLS) across geographically dispersed locations using MPLS as transport.

VPLS is often used by service providers to provide Ethernet Multipoint Services (EMS) and is also being adopted by Enterprises on a self-managed MPLS-based metropolitan area network (MAN) to provide high-speed any-to-any forwarding at Layer 2 without relying on spanning tree to loop free logical topology. The MPLS core uses a full mesh of pseudowires and split-horizon to avoid loops.

To provide multipoint Ethernet capability, IETF VPLS drafts describe the concept of linking virtual Ethernet bridges using MPLS pseudowires. At a basic level, VPLS can be defined as a group of Virtual Switch Instances (VSIs or VFIs) that are interconnected using EoMPLS circuits in a full mesh topology to form a single, logical bridge, as shown in Figure 3-1.

Figure 3-1 VPLS



In concept, a VSI is similar to the bridging function found in IEEE 802.1q bridges in that a frame is switched based upon the destination MAC and membership in a Layer 2 VPN (a virtual LAN or VLAN). VPLS forwards Ethernet frames at Layer 2, dynamically learns source MAC address to port associations, and forwards frames based upon the destination MAC address. If the destination address is unknown, or is a broadcast or multicast address, the frame is flooded to all ports associated with the virtual bridge. Therefore in operation, VPLS offers the same connectivity experienced if a device were attached to an Ethernet switch by linking virtual switch instances (VSIs) using MPLS pseudowires to form an “emulated” Ethernet switch.

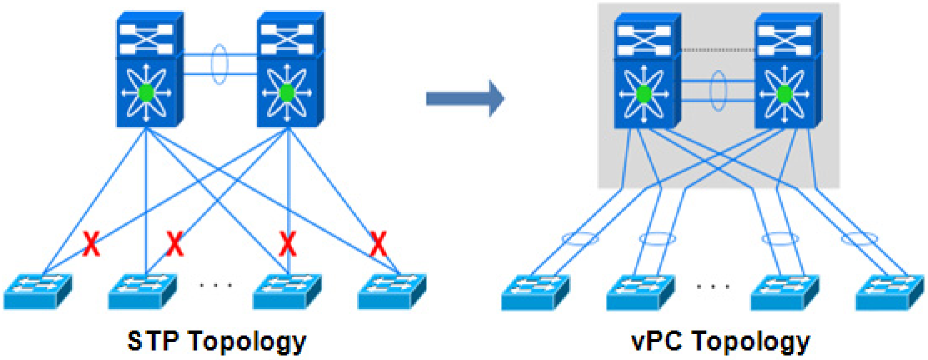
Compared to traditional LAN switching technologies, VPLS is also more flexible in its geographic scaling, so that Customer Edge (CE) sites may be within the same metropolitan domain, or may be geographically dispersed on a regional or national basis. The increasing availability of Ethernet-based multipoint service architectures from service providers, for both L2 VPN and L3 VPN services, is resulting in a growing number of enterprises transitioning their WANs to these multipoint services and VPLS is playing an increasingly important role in this transition. As highlighted in Figure 3-1, a VFI is linked (with a 1:1 mapping) to a Switch Virtual Interface (SVI). This is done for all the VLANs that need to be extended across the VPLS domain.

vPC Overview

The virtual Port Channel (vPC) functionality allows establishing port channel distributed across two devices, allowing redundant yet loop-free topology. Currently, vPC technology is offered on the Nexus 7000 and Nexus 5000 platforms.

Compared to traditional STP-based environments, vPC allows redundant paths between a downstream device and its two upstream neighbors. With STP, the port channel is a single logical link that allows for building Layer 2 topologies that offer redundant paths without STP blocking redundant links.

Figure 3-2 vPC Physical Topology



The deployment of these Multi-Chassis EtherChannel (MCEC) connections between the vPC peers and the downstream devices provides the following benefits:

- Removes dependence on STP for link recovery
- Doubles effective bandwidth by utilizing all MEC links

The use of vPC is usually positioned in the L2 domain of the network. This is a consequence of two current restrictions in the interaction of vPC with L3:

1. Only L2 links (access or trunk interfaces) can be bundled together using vPC. In other words, it is not possible to create a L3 virtual Port-Channel resulting in the bundle of L3 interfaces.
2. Establishment of dynamic routing adjacencies is currently not supported across a vPC (static routing is supported instead).

From a control plane perspective, each vPC peer is configured separately and runs its own independent instance of the operating system (control plane independency). The only interaction between the two chassis is facilitated using the Cisco Fabric Service (CFS) protocol, which assures that relevant configuration and MAC address tables of the two peers are in synch.

A downstream device sees the vPC domain as a single LACP peer since it uses a single LACP ID. Therefore the downstream device does not need to support anything beyond IEEE 802.3ad LACP. In the specific case where downstream device doesn't support 802.3ad LACP, a port channel can be statically configured ("channel-group group mode on"). Currently, NX-OS does not support PAgP that typically does not pose a problem given LACP standardization acceptability and longevity.

MC-LAG Overview

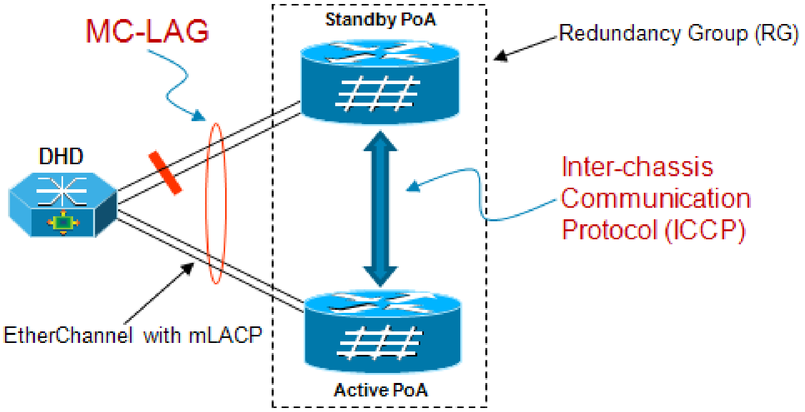
Similarly to what said in the previous section for vPC, the Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of a logical port-channel connection with a redundant pair of independent nodes. The two nodes run an independent control plane. The purpose of multi-chassis LAG is to provide a solution to achieve higher network availability. In order to improve the network reliability, network access elements are typically dual-homed to Provider Edge (PE) network devices. Network access element can be any type of device implementing LACP link bundling like Customer Edge (CE) router, DSLAM or in the case of Data-Center site connection, aggregation boxes running feature that support multi-chassis link bundling.

MC-LAG defines a solution for MPLS PE device and relies on the Inter Chassis Communication Protocol (ICCP) to form a redundancy Group (RG) that allows redundant attachment-circuit using multi-chassis Link Aggregation Control Protocol (mLACP). The PEs of a given redundancy group (RG) may be either physically co-located (e.g., in the same physical site or in the same central office) or geo-redundant (e.g., in sites or central central offices that are far apart). Connectivity for ICCP between PEs can be provided either via dedicated back-to-back links or shared link through the core network.

MC-LAG Components

Figure 3-3 highlights the main functional components building up the MC-LAG solution.

Figure 3-3 MC-LAG Components



- **Dual-Homed Device (DHD):** a DHD can be either a regular access device in the form of an access switch, or virtual switch implementing its own multi-chassis bundling solution. Considered only in the context of Layer 2 connectivity, it is assumed that the DHD is a relatively simple device and is not running any loop prevention algorithms, such as MSTP. The only baseline assumption is support of Link Aggregation Groups with LACP as defined in IEEE 802.3ad.
- **Point of Attachment (POA):** the POA is the point at which one of the DHD's uplinks is connected to an upstream system. In normal LACP operation a device would have 2 or more link's point of attachment to a common (single) chassis, or system.
- **Multi-chassis Link Aggregation Control Protocol (mLACP):** mLACP is an extension usage of standard based "Link Aggregation Control Protocol" (LACP) defined in IEEE 802.3ad to convey to the DHD that it is connected to a single virtual LACP peer as opposed to two independent devices. Note that the MC-LAG solution relies exclusively on LACP and does not work when using Port Aggregation Protocol (PAgP) or static bundles.
- **Inter-Chassis Communication Protocol (ICCP):** The POA nodes forming a virtual LACP peer, from the perspective of the DHD, are said to be members of a Redundancy Group (RG). State synchronization between the POA nodes in a RG is required in order for them to appear as a single device to the DHD. This is achieved through an Inter-Chassis Communication Protocol (ICCP), which provides a control-only Inter-chassis Communication Channel (ICC). ICCP runs over an LDP session established between two POA nodes. L3 IP connectivity is only required between these devices, so they don't need to run MPLS LDP between them.

Note More information on ICCP can be found in the latest version of the following IETF Draft: http://tools.ietf.org/html/draft-ietf-pwe3-iccp-05

Inter-chassis Coordination / Communication Considerations

A method to coordinate states and handle failover conditions needs to be implemented between POAs. This requires a reliable communication protocol that is flexible and allows intelligent policy control. This communication protocol is referred to as the Redundancy Manager (RM) whose generic functions can be characterized by the following:

- Manage state between Active & Standby POAs
- Manage communication sessions between POAs
- Interpret access-circuit driven events and drive action on network-side
- Allow other resiliency protocols to take advantage of RM's function.
- Allow for operator originated policy to provide deterministic failover behavior
- Provide a means to monitor and take action on POA peer failure events (i.e. IP Route Watch, BFD, etc.)
- Trigger remote system notification via other protocols & redundancy mechanisms (i.e. 2-way status bit signaling, MIRP MAC withdrawal, etc.)

To address these requirements, ICCP is modeled comprising three layers:

1. **Application Layer:** This provides the interface to the various redundancy applications that make use of the services of ICCP.

2. **Inter Chassis Communication (ICC) Layer:** This layer implements the common set of services that ICCP offers to the client applications. It handles protocol versioning, Redundancy Group membership, Redundant Object identification, PE node identification and ICCP connection management.
3. **Transport Layer:** This layer provides the actual ICCP message transport. It is responsible for addressing, route resolution, flow-control, reliable and in-order message delivery, connectivity resiliency/redundancy and finally PE node failure detection. This Transport layer may differ depending on the Physical Layer of the interconnection, but current implementation relies on the targetted Label Distribution Protocol (LDP) that is the Pseudo-Wires establishment control–plane. When an RG is enabled on a particular PE, the capability of supporting ICCP must be advertised to all LDP peers in that RG. This is achieved by using the methods in [RFC5561] and advertising the ICCP LDP capability TLV.

Multi-chassis Link Aggregation Control Protocol (mLACP) Considerations

Link Aggregation Control Protocol (LACP) defined in IEEE 802.3ad is a link-level control protocol that allows the dynamic negotiation and establishment of link aggregation groups (LAGs). It was designed to form link aggregation between two devices and the challenge is that it was never designed to form link aggregation using multiple nodes. mLACP circumvents this by creating a virtual LACP peer in such a way that the connected device do not notice that its bundle is connected to two or more PEs.

LACP is a link layer protocol and operates such that all messages exchanged over a given link contain information that is specific and localized to the link itself. The exchanged information includes:

- System Attributes: Priority and MAC Address
- Link Attributes: Key, Priority, Port Number and State

When extending LACP to operate over a multi-chassis setup, it is required to synchronize the protocol attributes and states between the two chassis.

LACP relies on a System MAC Address to determine the identity of the remote device connected over a particular link. Therefore, in order to mask the fact that the attached device is connected to two separate devices, it is essential to coordinate the System MAC address between the two PE.

In general, the LACP System MAC Address defaults to the ROM MAC address on the backplane and cannot be changed by configuration. For purpose of multi-chassis operation, the following two requirements should be addressed:

- System MAC Address for each POA should be communicated to its peer. The POAs would, for example, elect the MAC Address with the lower numeric value to be the System MAC. The arbitration scheme should be deterministic, i.e. always resolve to the same value. Selecting the lower numeric MAC address value has the advantage since it provides higher System Priority.
- System MAC Address should be configurable. This is required because the System Priority depends, in part, on the MAC Address, and there is the need to guarantee that the PoAs have higher priority than the DHD (for example: if both DHD and PoA are configured with the same System Priority and SP has no control over DHD). This guarantees that the PoA Port Priorities take precedence over the DHD's Port Priority configuration. In the scenario where the user configures the System MAC address, it is preferred that the user guarantees that the addresses are uniform on both PoAs; otherwise, the system will automatically arbitrate the discrepancy as in the case of the default MAC above (ie pick the lowest configured value).

The DHD should not be aware of the existence of separate PoAs. This will require, initially, to be implemented with only one active Point of Attachment (POA), but not one active link. In other words, you could have two links going to POA1 and three links going to POA2. The key is that in an exclusive fashion either all the links to POA1 are active or all the links to POA2 are active in this phase. This mode of operation is referred to as ACTIVE / STANDBY.

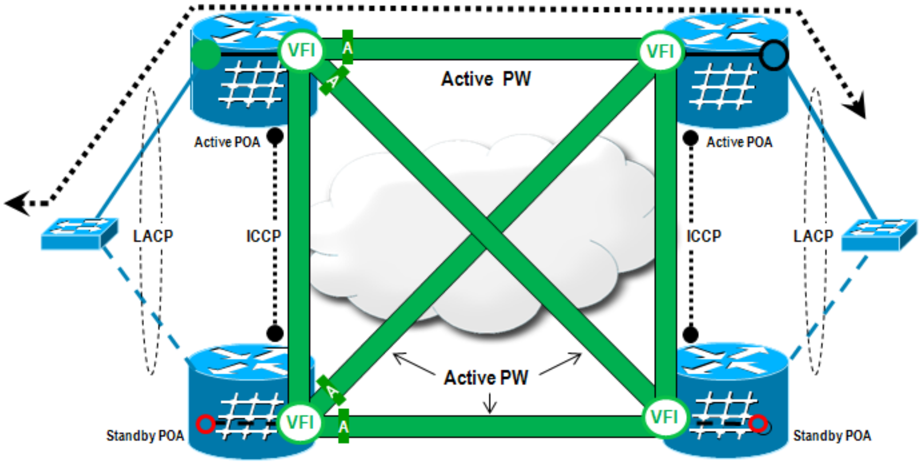
MC-LAG uses dynamic port priority management to ensure link state between DHD and POA. This mechanism involves dynamic allocation of port priority of links in the LAG such that the priority of the standby link is lower than the primary one, while trigger switchover would just be accomplished thru decrease of link priority.

It is recommended to allocate higher system priority to POA to have it manages which link is active or standby based on ICCP exchange.

MC-LAG and VPLS Integration Considerations

The MC-LAG solution discussed in this document is combined with VPLS, in order to extend L2 communication to remote sites across an MPLS core. The way these two functionalities are integrated is shown in Figure 8 - MC-LAG to VPLS (Decoupled Mode).

Figure 3-4 MC-LAG to VPLS (Decoupled Mode)



As highlighted in Figure 3-4, the status of the VPLS PseudoWires (PW) originated from each POA is independent of the link bundle or POA status. This is called decoupled mode, where all the PWs are in active state unless the remote PE router signals a standby state. The use of decouple mode brings an advantage during link failure/recovery scenarios, since traffic outage is only affected by the time taken by the Standby POA to gain the active role.

Note Coupled mode is also available when integrating MC-LAG and VPLS. ASR 9000 platforms only support decoupled mode and since they are the only POA devices discussed in this paper, the discussion is limited to this mode of operation.

VPLS MAC Withdrawal Considerations

When a failure at the customer site results in a topology change such that a particular host becomes unreachable via the original path, a MAC flush on all PE routers will result in a new flooding of traffic to that host until new path is learned. This will allow frames to reach that host without the need to first wait for the MAC addresses to age out on all devices, or – in specific situations – to wait for traffic from the host to reprogram the MAC tables.

In IOS XR, MAC withdraw is triggered on a DOWN event that are associated to topology change and LDP MAC withdraw messages are sent out on the following events:

1. An attachment Circuit (AC) is brought down
2. An access PW (H-VPLS topology) is brought down
3. An attachment Circuit (AC) is removed from a bridge domain (unconfiguration)
4. An access PW is removed from a bridge domain (unconfiguration)
5. Bridge domain MAC addresses are cleared via CLI using “clear l2vpn bridge-domain”
6. MSTP, RSTP or REP MAC flush received over attachment circuit

When a MAC flush is triggered, a second MAC flush action is performed 7 seconds later. This produces the effect of sending a new MAC withdraw. This is to reduce the possibility of MAC withdraw messages reaching the receiving PE before the data in transit. Should this happen, the receiving PE flushes the MAC addresses are relearned the MAC addresses on the sending PE immediately due to transiting data.

In VPLS, such a MAC flush is performed using a MAC Address withdrawn which is a LDP message that is sent by a PE to remote peers in order to trigger a MAC flush on those peers. LDP has the option to specify a list of MAC addresses to flush (use of a list) or to use an empty list. RFC4762 (Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling), Section 6.2 states the following:

For a MAC Address Withdraw message with empty list: - Remove all the MAC addresses associated with the VPLS instance (specified by the FEC TLV) except the MAC addresses learned over the PW associated with this signaling session over which the message was received.”

In IOS-XR, receiving an LDP MAC withdraw message will flush all MAC addresses associated with the VPLS instance thru receiving a MAC Withdraw (MW) message with empty list, including the MAC addresses learned over the pseudowire associated with this signaling session over which the message was received.

Asr9k learns MAC addresses in hardware, hence is able to learn MACs at linerate. This has been measured at a learning rate of 512000 MACs per second, which was the highest possible rate that we could use.

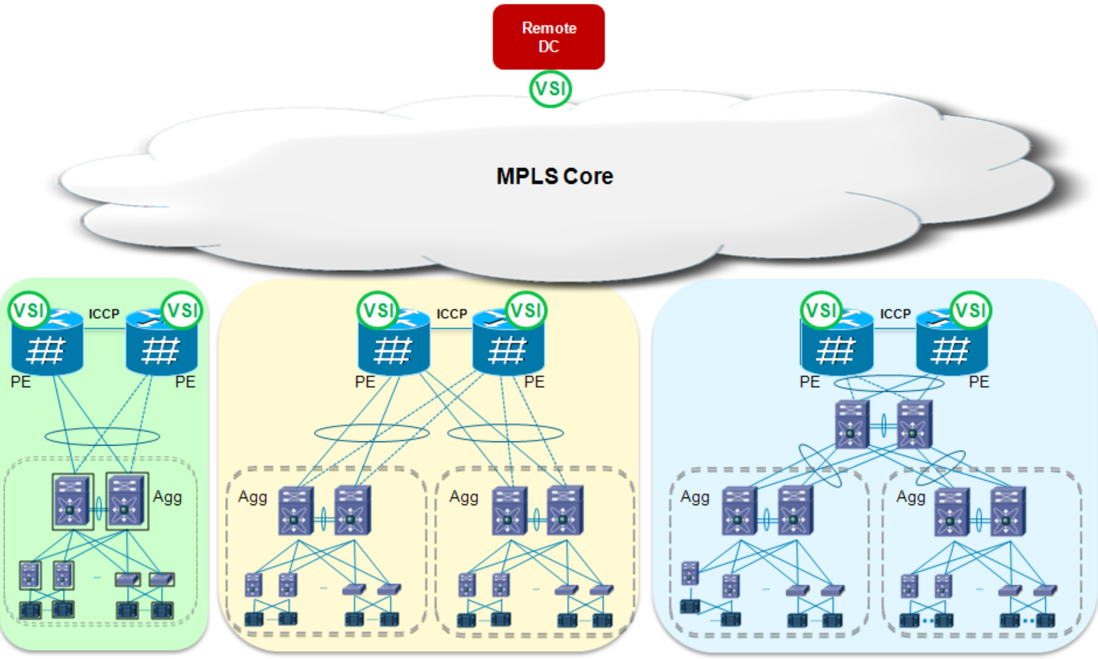
When there are n PEs in a full mesh topology, and a PE is required to flood a frame, it will have to replicate and forward them on (n-1) pseudowires. In theory, when a PE receives for example 1 Gbps of traffic from the access side and a MAC table flush results in the removal of all MAC addresses, then this PE would initially forward 3 Gbps of traffic into the core Service Provider network. In practice, whether there will be a sudden burst of 3 Gbps will depend mainly on what type of applications were running and how fast the new MAC addresses will again be learned.

The number of transmitted and received MAC withdraw messages over a given PW is provided in the output of show l2vpn bridge-domain detail.

Architecture Overview

The architecture that can be deployed to provide LAN extension services between data center sites leveraging MC-LAG and VPLS is shown in Figure 3-5.

Figure 3-5 MC-LAG to VPLS Architecture



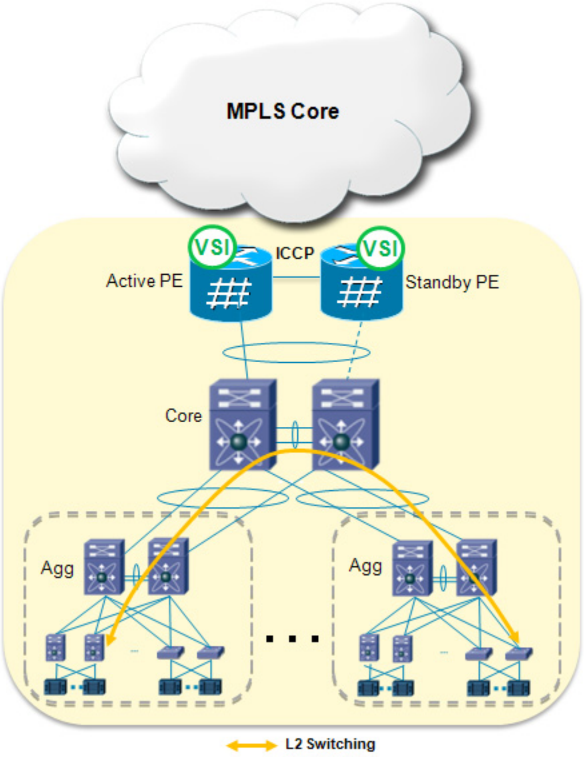
The key functional blocks of the solution are the following:

- Pair of PE devices performing the VPLS traffic forwarding and running the ICCP control protocol in order to support the MC-LAG functionality. This document specifically positions the use of Cisco ASR 9000 in this role.
- One of more aggregation blocks connecting to the PE devices. A pair of aggregation layer devices connecting multiple access layer switches usually represents an aggregation block. The number of aggregation blocks deployed depends on the size of the DC. Nexus 7000 platforms are presented in this role because of their support of vPC providing Multi-Chassis EtherChannel functionality.

Note In the context of this paper, a given aggregation block can also be referred to as POD.

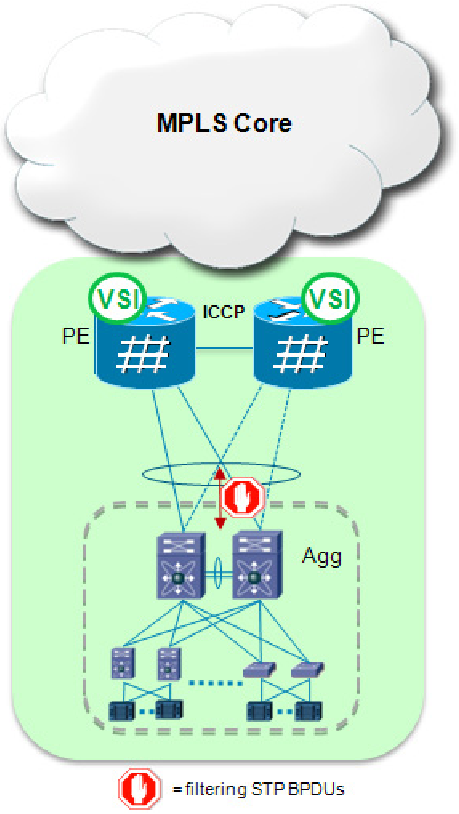
- Depending on the number of aggregation blocks deployed in a single site, the deployment of a dedicated DC core layer is possible, as shown on the right side of Figure 3-5. In that scenario, each aggregation block connects to the core leveraging vPC connections, and the core devices are now connecting to the PE routers via MC-LAG. The introduction of the core layer is usually positioned to simplify the deployment (and minimize the number of the required interfaces on the PE routers) when a large number of PODs are deployed inside the data center. The use of a core layer is definitely recommended when LAN extension is also required between PODs deployed in the same DC site. In that case, Nexus 7000 switches deployed in the core can perform the L2 switching functionality, leveraging full active/active vPC connections with the various PODs (Figure 3-6).

Figure 3-6 LAN Extension between PODs



As discussed in the "LAN Extension Technical Requirements" section, it is desirable to limit as much as possible the extension of the STP domain inside a given aggregation block. This can be achieved by filtering STP BPDUs on the MCEC connection between a specific POD and the pair of upstream devices, as shown in Figure 3-7.

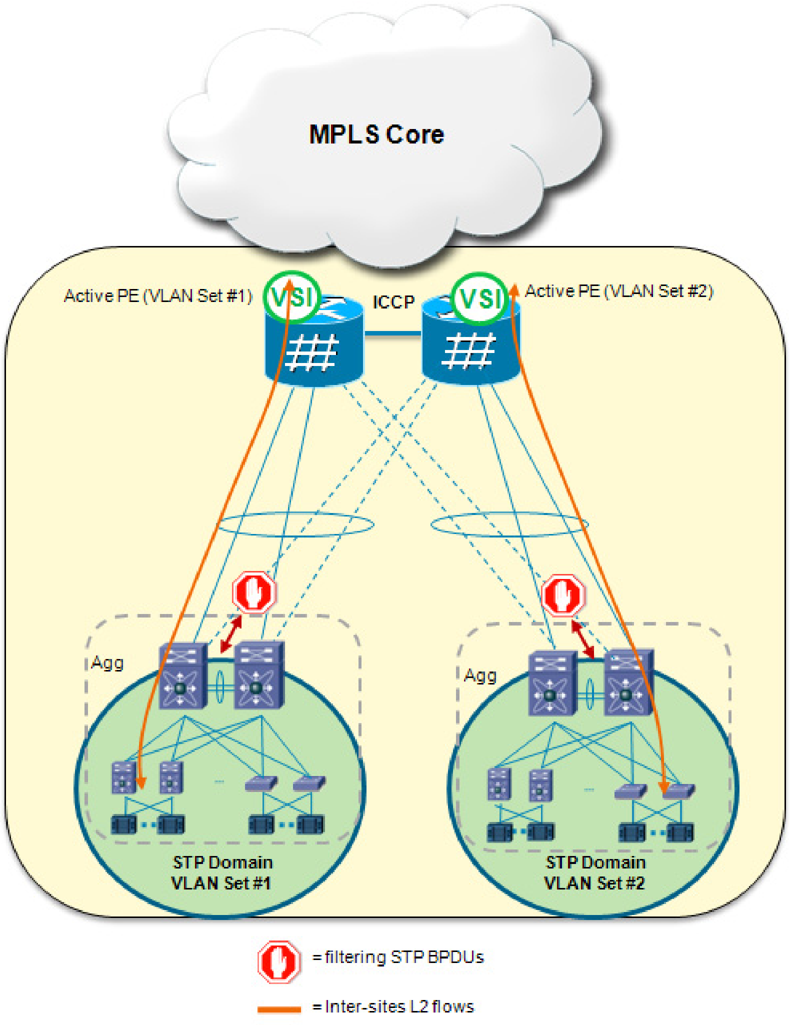
Figure 3-7 STP Isolation from each POD



The configuration of BPDU filtering on the individual vPC connection ensures that the box does not send any BPDUs and drops all BPDUs that it receives. This could be important in scenarios where the PE routers are for example managed by a different entity from the team managing the DC switches (could be a separate team inside the enterprise or even a SP), because it ensure complete isolation between each POD and the rest of the network.

However, in scenarios where multiple PODs are deployed, it is important to distinguish a couple of different cases. When the VLANs defined inside each POD are different (i.e. no inter-POD LAN extension is required), the STP BPDU filtering can still be applied on the vPC connection out of each POD, allowing for the creation of independent STP domains in each POD (Figure 3-8).

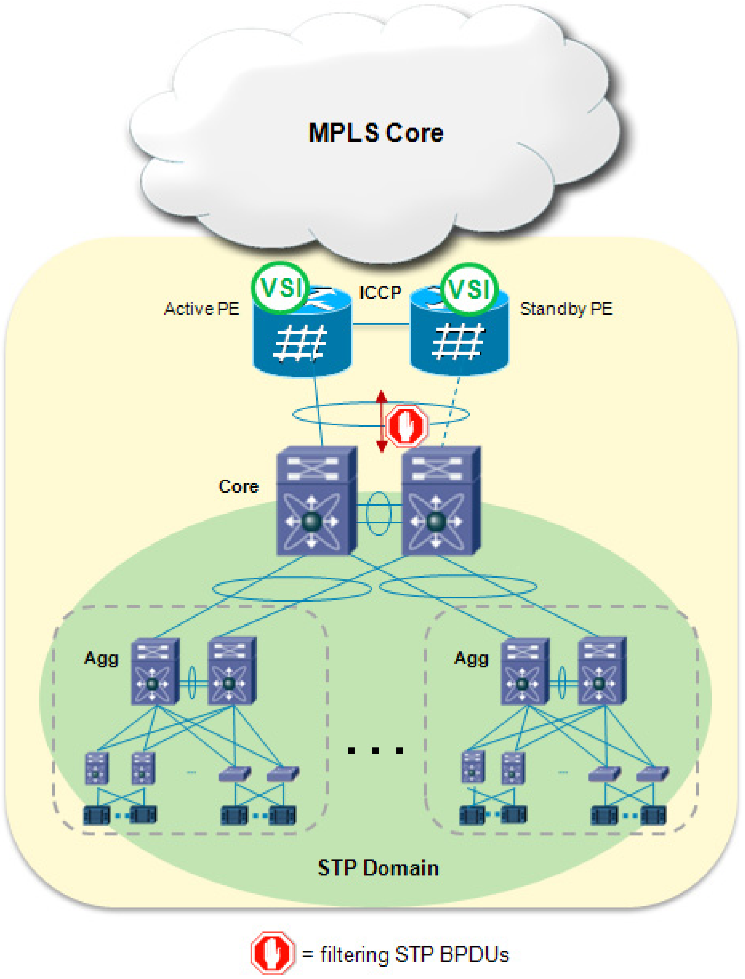
Figure 3-8 Independent VLAN Sets per POD



As shown above, in this specific deployment scenario, it is also recommended to tune the configuration in order to ensure each deployed PE router is active for a given VLAN set. In this way, a more optimal traffic load-balnacing can be achieved for inbound and outbound inter-sites L2 flows.

In deployments where VLANs span between PODs, filtering STP BPDUs is not recommended not to expose the design to the risk of creating an STP loop as a consequence of a cabling mistake that creates a L2 backdoor between aggregation blocks. As shown in Figure 3-9, this leads to the creation of a large STP domain usually rooted on the core devices (at least for the VLANs that require inter-PODs extension). Notice also how BPDU filtering is still applied between the DC core switches and the PE routers, for the same reasons previously discussed.

Figure 3-9 Extension of the STP domain



In the example above, STP BPDUs are exchanged between the PODs via the core routers. From a data plane perspective, local switching on the core devices provides intra-site L2 connectivity, whereas VPLS is leveraged to extend L2 to remote sites across the MPLS enabled core.

Note The design validate for the creation of this paper was leveraging the STP BPDU filtering configuration. As a consequence, also the convergence results discussed in the rest of the paper apply to this specific deployment model.

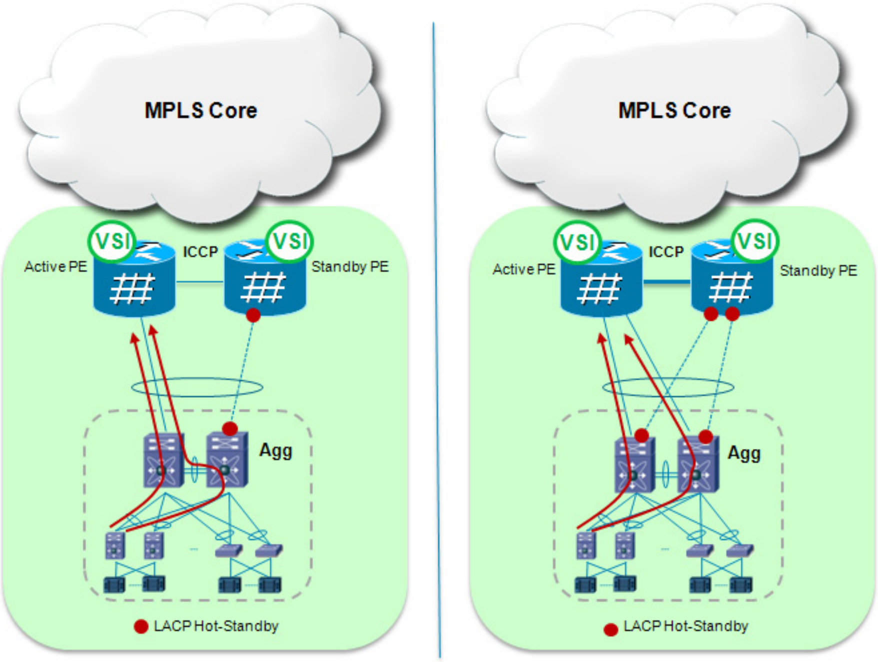
Active/Standby Deployment

As previously mentioned, MC-LAG is by definition an Active/Standby solution: one POA (PE device) is active (i.e. forwarding traffic in both inbound and outbound directions), whereas the other POA is standby. All the member ports on the Standby POA are in standby mode to prevent forwarding of traffic.

When interconnecting an aggregation block with Nexus 7000 devices in aggregation, the use of vPC facing the PE devices results in interfacing an Active/Active port-channel technology (vPC) with an Active/Standby one (MC-LAG). For this reason, the use of LACP negotiation becomes key to ensure that the Nexus 7000 interfaces facing the Standby POA are in Standby state as well, in order to prevent black-holing of traffic.

This may also have implications on the traffic path for traffic flows originated from inside the DC site and directed toward the PE devices.

Figure 3-10 Traffic Flows with 2 and 4 vPC Links

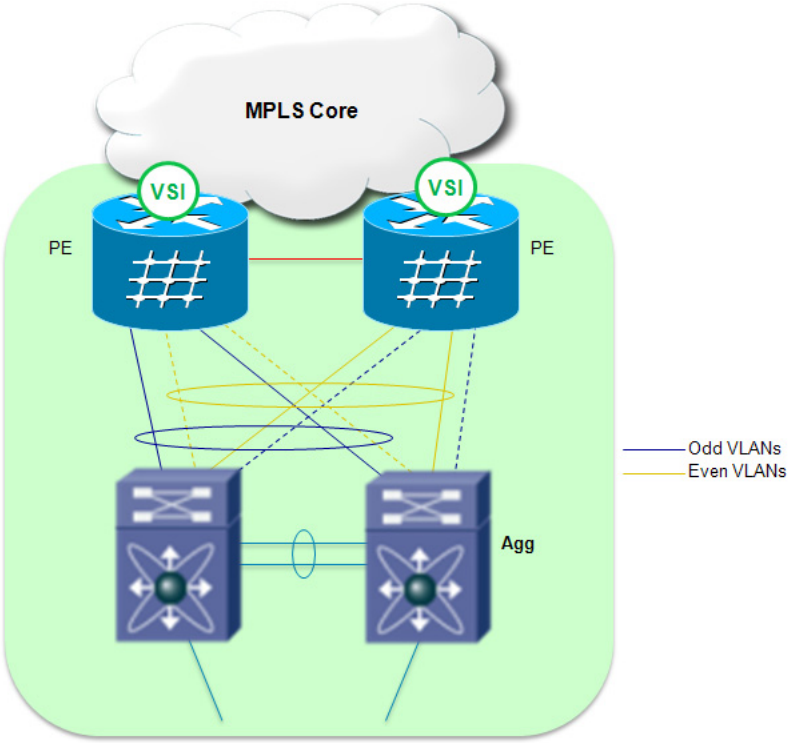


As shown in Figure 3-10, when the vPC is created with only two physical links bundled together, the outbound L2 traffic flows may follow a sub-optimal path across the peer-link connecting the aggregation layer device. This behavior can be improved by bundling together 4 links in the vPC, so that a direct path exists all the time between each Nexus 7000 device and the active PE router (right picture above).

Active/Active Deployment with Redundant Physical Connections

The Active/Standby behavior previously described can be improved by doubling the number of physical connection between the aggregation devices and the PE router, as highlighted in Figure 3-11.

Figure 3-11 Active/Active MC-LAG Deployment



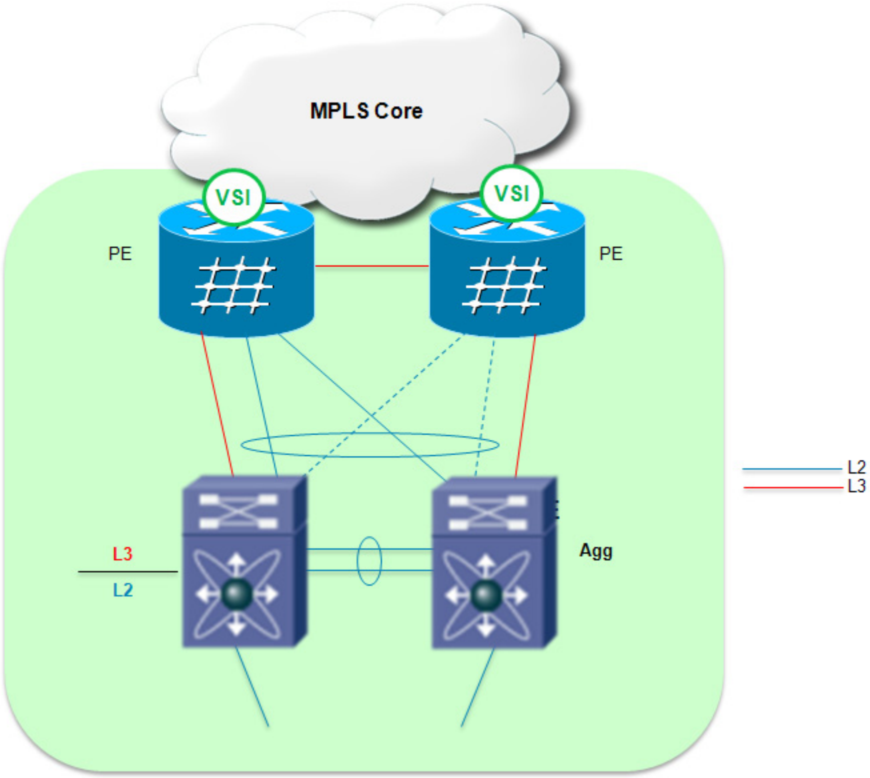
The idea is to create two logical connections between these devices, leveraging two separate vPCs and MC-LAG bundles and divide the VLANs to be extended between these. In this way, each PE router can become active for half of the VLANs to be extended and traffic can better be load-balanced both in outbound and inbound directions.

Note The reference to “Odd” and “Even” VLANs separation is just to provide an intuitive example. From an operational perspective it would probably be easier to group the VLANs to be activated via the two separate paths in two contiguous sets.

Routing Considerations

All the design considerations included in the previous sections were referring to L2 traffic flows established within and across data centers. Very often the same PE routers providing LAN extension services are also leveraged for L3 connectivity, and this raises the question on how the aggregation layer devices should route traffic toward the PEs.

Figure 3-12 Deploying Dedicated Links for L3 Communication



Routing of traffic is usually required because the L2/L3 boundary for all the VLANs defined at the access layer of each aggregation block often remains on the aggregation layer devices, as shown in Figure 3-12. The blue links represents L2 trunks carrying the VLAN traffic to the PE devices, in order to be extended to other aggregation blocks via local switching or to remote sites via VPLS. While MC-LAG supports L3 peering over bundle, however, it is not possible to use these same L2 trunk connections to establish a routing peering between the aggregation devices and the PE router, because L3 virtual Port-Channel is not supported when deploying vPC (as previously mentioned in the "vPC Overview" section).

As a consequence, dedicated L3 links are required to provide routing services. A single link between the aggregation device and the PE router can be used, but it is important to also establish L3 peering on a dedicated VLAN across the peer-link between aggregation devices.

Figure 3-13 Establishing IGP Peering via the Peer-Link

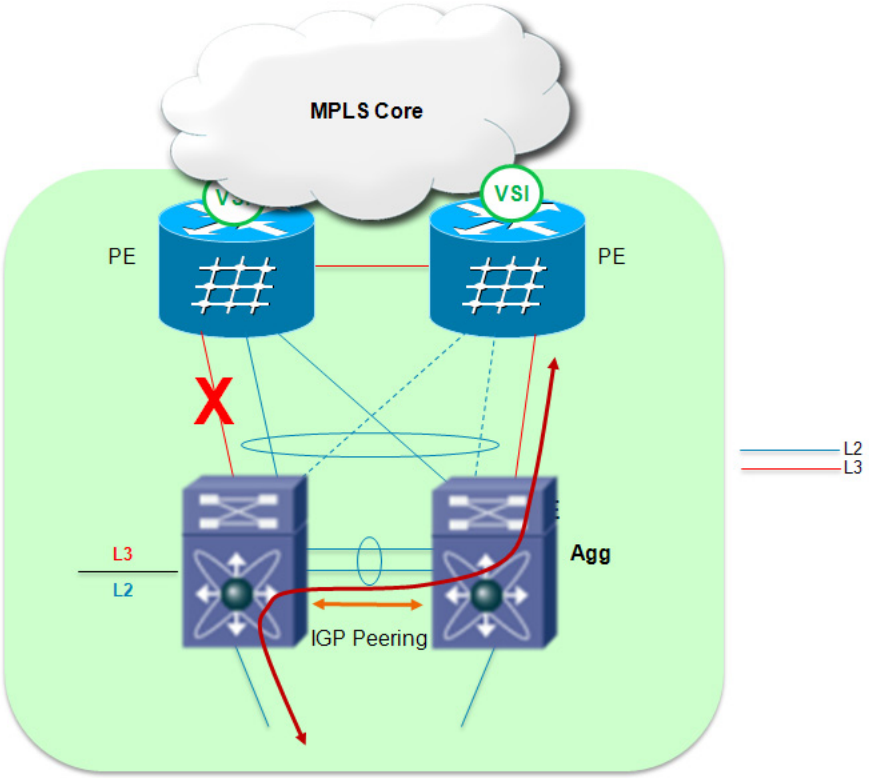


Figure 3-13 shows L3 traffic path during failure of the L3 link connecting the left aggregation device to the PE router.

Note The dedicated L3 connections may be established with separate DC core devices (if available). This specific deployment will be highlighted in the "Deploying MC-LAG to VPLS Solution" section.

Selective Q-in-Q Considerations

Cisco ASR 9000 platforms offer support for an additional functionality named "selective Q-in-Q". When deploying such functionality, a set of VLANs can be grouped together in the same bridge domain and associated to a single VPLS Pseudowire connecting to remote data center sites. This is usually positioned to increase the scalability characteristics of the solution while reducing the overall OPEX associated to it. Also, the deployment of QinQ is typical in multi-tenant deployments, where separate sets of "customer VLANs" can be grouped and associated to unique "core VLANs" transported on dedicated PWs across the MPLS cloud.

The use of QinQ is not considers as part of the scope for this paper. This implies that every extended VLAN is associated to a separate bridge domain (VSI) and corresponding VPLS PseudoWire (one VSI per VLAN deployment model).