High Scale Data Center Interconnect

LAN Extension Using MC-LAG to VPLS on the Cisco ASR-9000

Last Updated: July 11, 2011

Cisco Validated Design

Building Architectures to Solve Business Problems

About the Authors

Max Ardica, Technical Marketing Engineer, Systems Architecture and Strategy Unit (SASU), Cisco Systems

Max Ardica, CCIE #13808 and a 13 year Cisco veteran, is a Solution Architect for data center technologies in the Cisco System Architecture and Strategy Unit. His primary focus is driving architectural definition and systems validation for data center solutions, mainly around the topic of data center virtualization and interconnection. Prior to this work, Max was a Technical Marketing Engineer in the Enterprise Systems Engineering (ESE) group at Cisco, where he focused on the development and verification of the best practices for Network Virtualization and Identity related technologies for Cisco Campus Networks.

Nash Darukhanawalla, Development Engineering Manager, Enhanced Customer Aligned Testing Services (ECATS), Cisco Systems

Nash Darukhanawalla, CCIE No. 10332, has more than 27 years of internetworking experience. He has held a wide variety of consulting, technical, product development, customer support, and management positions. Nash's technical expertise includes extensive experience in designing and supporting complex networks with a strong background in configuring, troubleshooting, and analyzing network systems.

Patrice Bellagamba, Distinguished Systems Engineer, Innovation Consulting Engineering, Cisco Systems

Patrice Bellagamba is a recognized expert in innovative IP and MPLS technologies and networks and has been honored as a Distinguished Systems Engineer at Cisco Systems. Patrice has been in the networking industry for more than 27 years and has spent more than 10 years in development engineering. Patrice has been a significant contributor to Data Center Interconnect (DCI), with a special focus on the usage of the VPLS technology.

About Cisco Validated Design (CVD) Program

The CVD program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit http://www.cisco.com/go/designzone.

High Scale Data Center Interconnect LAN Extension Using MC-LAG to VPLS on the Cisco ASR-9000

CONTENTS

Preface

This document provides design guidance, configuration examples, and Cisco recommended best practices for interconnecting geographically dispersed data centers and implementing Layer 2 connectivity across Layer 3 network infrastructure using VPLS. The specific solution presented in this paper is named “MC-LAG to VPLS” and leverages the VPLS and MC-LAG functions available on Cisco ASR 9000 platforms. MC-LAG allows interconnecting two physically independent ASR 9000 PE devices to a single (physical or logical) device with a port-channel connection. The paper highlights the design considerations when ASR 9000 PE devices are connected to pair of Nexus 7000 using the virtual Port-Channels (vPC) technology. This solution is specifically targeted to high scale DCI deployments, usually found in large enterprise or Service Provider market segments. Two different deployment scenarios are the focus of this paper; the first one aiming to provide LAN extension services between data center sites for 500 VLANs. The second increasing this figure to 1200 VLANs. Also, despite the fact that the main focus of the paper is on the integration of ASR 9000 and Nexus 7000 platforms, interoperability with other Cisco platforms (as 7600 routers and Catalyst 6500 in VSS mode) was also considered during the validation effort.

Audience

This document is intended for customers and system engineers who are designing solutions or looking for design guidance with interconnecting data centers ensuring high availability Layer 2 connectivity and STP isolation. In addition, the solution presented in this paper applies to large-scale Layer 2 extension.

Organization

This document is organized as follows:

- MC-LAG to VPLS Introduction, page 1-1 Provides an overview of design considerations and the Cisco Validated Design (CVD) program.
- Data Center Interconnect Solution Overview, page 2-1 Provides an overview of Cisco Data Center Interconnect solutions, highlighting their main functional components and functional requirements, specifically focusing on the LAN Extension aspects of such solutions.

- MC-LAG to VPLS Technology and Solution Overview, page 3-1 Provides an overview of the technologies required to deploy the MC-LAG to VPLS solution, including MC-LAG, vPC and VPLS. An high level architecture overview is also provided in this chapter, highlighting alternative solution deployment models.
- Data Center Multitier Model and Testbed Topology, page 4-1 Describes the Cisco-recommended data center multitier model and the testbed that was used to validate the specific MC-LAG to VPLS solution.
- Deploying MC-LAG to VPLS Solution, page 5-1 Discusses in detail the deployment of this specific LAN Extension solution, presenting also specific results achieved under various failure scenarios.

Obtaining Documentation, Support, and Security Guidelines

For information about obtaining documentation, submitting a service request, and gathering additional information, see the monthly What's New in Cisco Product Documentation, which also lists all new and revised Cisco technical documentation, at:

http://www.cisco.com/en/US/docs/general/whatsnew/whatsnew.html

Subscribe to the What's New in Cisco Product Documentation as a Really Simple Syndication (RSS) feed and set content to be delivered directly to your desktop using a reader application. The RSS feeds are a free service and Cisco currently supports RSS version 2.0.

CHAPTER 1

MC-LAG to VPLS Introduction

Various data center requirements have resulted in an expansion of Layer 2 domains, thus increasing the extension of Spanning Tree domain at the network level. Considering the fact that the Spanning Tree Protocol was developed to handle a small network diameter, the enterprise/SP network needs to meet the required Layer 2 connectivity challenges to ensure high availability between geographically dispersed data centers.

Exponential growth in data center resources and security requirements are driving the need to connect multiple data centers over larger distances. As a result, customers are facing additional challenges such as maintaining the high availability of applications and dealing with complex multi-sites interconnections.

This document covers one specific VPLS-based solution (referred to as “MC-LAG to VPLS”) that provides a high-scale, low-latency network and Spanning Tree Protocol isolation between data centers. This document encompasses issues related with large Layer 2 bridging domains and provides guidance for extending VLANs over Layer 3 network using VPLS technology.

Extensive manual testing was conducted in a large-scale customer representative network. The MC-LAG to VPLS solution was validated with a wide range of system test types, including system integration, fault and error handling, and redundancy to ensure a successful customer deployment. An important part of the testing was end-to-end verification of unicast and multicast traffic.

Cisco Validated Design Program

The Cisco Validated Design (CVD) Program consists of systems and solutions that are designed, tested, and documented to facilitate faster, more reliable and more predictable customer deployments. These designs incorporate a wide range of technologies and products into a broad portfolio of solutions that meet the needs of our customers. There are two levels of designs in the program: Cisco Validated Design and CVD System Assurance.

Cisco Validated Design

Cisco Validated Designs are systems or solutions that have been validated through architectural review and proof-of concept testing in a Cisco lab. Cisco Validated Design provides guidance for deploying new technologies or in applying enhancements to existing infrastructure.

CVD System Assurance

Cisco Validated Design System Assurance is a program that identifies systems that have undergone architectural and customer relevant testing. Designs at this level have met the requirements of a CVD design as well as being certified to a baseline level of quality that is maintained through ongoing testing and automated regression for a common design and configuration.

Certified designs are architectural best practices that have been reviewed and updated with appropriate customer feedback and can be used in pre and post-sales opportunities. Certified designs are supported with forward looking CVD roadmaps and system test programs that provide a mechanism to promote new technology and design adoption. CVD Certified Designs advance Cisco System's competitive edge and maximize our customers' return on investment while ensuring operational impact is minimized.

A CVD certified design is a highly validated and customized solution that meets the following criteria:

- Reviewed and updated for general deployment
- Achieves the highest levels of consistency and coverage within the Cisco Validated Design program
- Solution requirements successfully tested and documented with evidence to function as detailed within a specific design in a scaled, customer representative environment
- Zero observable operation impacting defects within the given test parameters, that is, no defects that have not been resolved either outright or through software change, redesign, or workaround (refer to test plan for specific details)
- A detailed record of the testing conducted is generally available to customers and field teams, which provides:
  - Design baseline that provides a foundational list of test coverage to accelerate a customer deployment
  - Software baseline recommendations that are supported by successful testing completion and product roadmap alignment
  - Detailed record of the associated test activity that includes configurations, traffic profiles, and expected results as compared to actual testing results

For more information about the Cisco CVD program, refer to:

http://cisco.com/en/US/netsol/ns741/networking_solutions_program_home.html

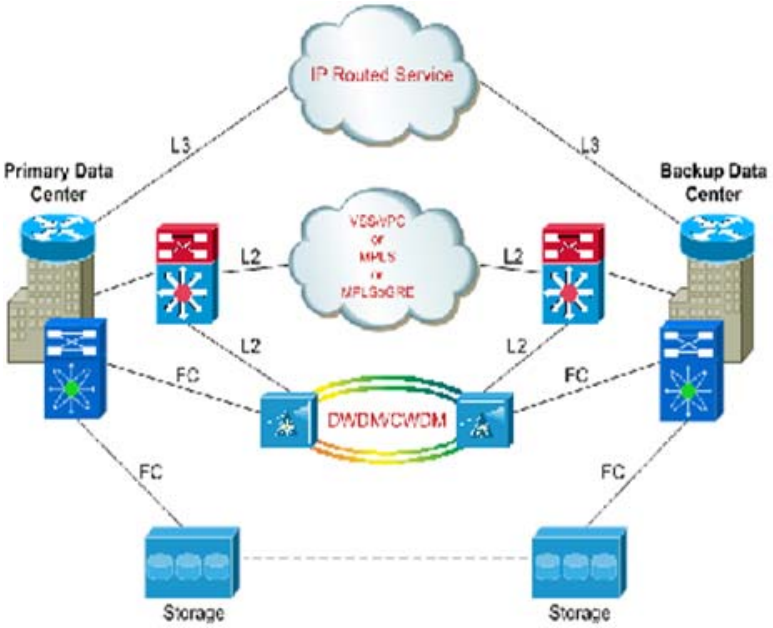
CHAPTER 2

Data Center Interconnect Solution Overview

The term DCI (Data Center Interconnect) is relevant in all scenarios where different levels of connectivity are required between two or more data center locations in order to provide flexibility for deploying applications and resiliency schemes.

Figure 2-1 summarizes the three general types of connectivity required for a DCI solution.

Figure 2-1 DCI Connectivity Overview



- **LAN Extension:** Provides a single Layer 2 domain across data centers. The data center applications are often legacy or use embedded IP addressing that drives Layer 2 expansion across data centers. Layer 2 Extension provides a transparent mechanism to distribute the physical resources required by some application frameworks such as the mobility of the active machine (virtual or physical).
- **Layer 3 Extension:** Provides routed connectivity between data centers used for segmentation/virtualization and file server backup applications. This may be Layer 3 VPN-based connectivity, and may require bandwidth and QoS considerations.

- **SAN Extension:** This presents different types of challenges and considerations because of the requirements in terms of distance and latency and the fact that Fibre Channel cannot natively be transported over an IP network.

In addition to the 3 functional component listed above, a holistic DCI solution usually leverages an additional building block. This is usually referred to as Path Optimization and deals with the fact that every time a specific VLAN (subnet) is stretched between two (or more) locations that are geographically remote, specific considerations need to be made regarding the routing path between client devices that need to access application servers located on that subnet. Same challenges and considerations also apply to server-to-server communication, especially for multi-tier application deployments. Path Optimization includes various technologies that allow optimizing the communication path in these different scenarios. Integration of network services (as FW, load-balancers) also represents an important design aspect of a DCI solution, given the challenges brought up by the usual requirement of maintaining stateful services access while moving workloads betwee data center sites.

LAN Extension represents a very important component of DCI and is the main focus of this document. The following sections of this chapter present the most common business requirements for LAN Extension, listing also its main technical requirements. The last section provides an overview of Cisco LAN Extension solution offering.

LAN Extension Business Drivers

There are various business reasons driving the deployment of DCI solutions. Traditionally, Cisco recommends isolating and reducing Layer 2 networks to their smallest scope, usually limiting them to the access layer.

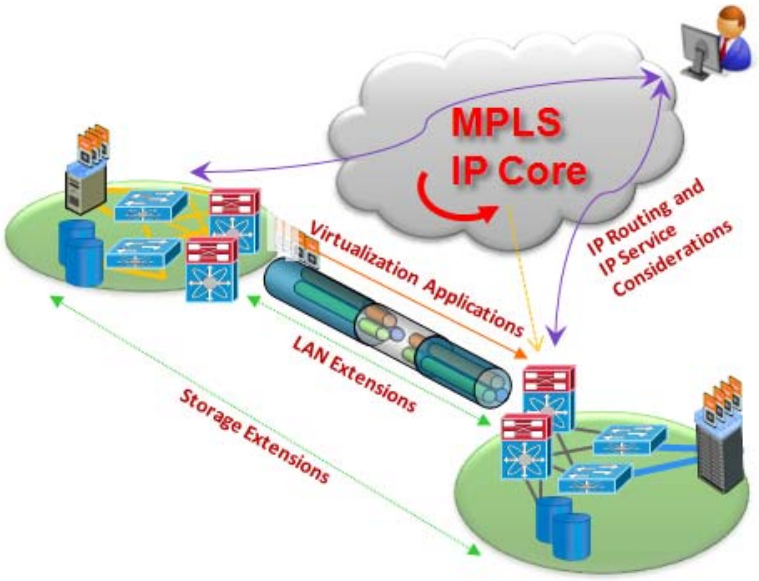
However, in some situations Layer 2 must be extended beyond the single data center, specifically when the framework or scenario developed for a campus has been extended beyond its original geographic area and over multiple data centers across long distances. Such scenarios are becoming more prevalent as high-speed service provider connectivity becomes more available and cost effective.

High-availability clusters, server migration, and application mobility are some important use cases that require Layer 2 extension.

Workload Mobility (Active/Active Data Centers)

The deployment of LAN Extension technologies can also facilitate and maximize a company's server virtualization strategy, adding flexibility in terms of where compute resources (workload) reside physically and being able to shift them around geographically as needs dictate.

Figure 2-2 Workload Mobility across DC Sites



Some applications that offer virtualization of operating systems allow the move of virtual machines between physical servers separated by long distances. To synchronize the software modules of the virtual machines during a software move and to keep the active sessions up and running, the same extended VLANs between the physical servers must be maintained.

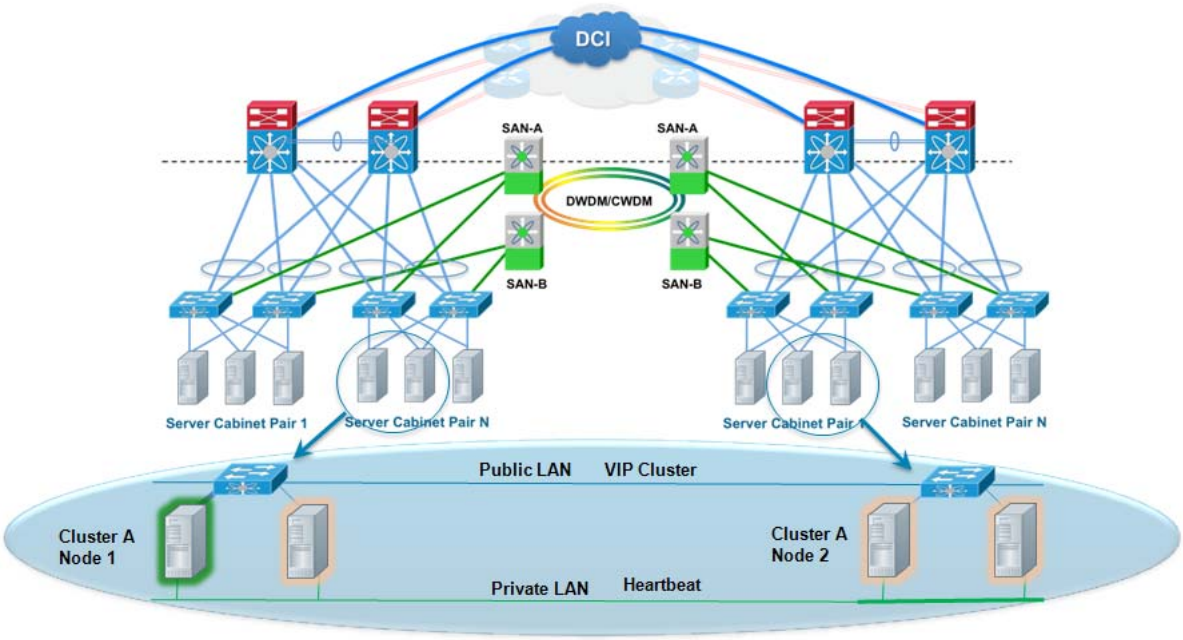
Migrating virtual machines between data centers provides compute power from data centers closer to the clients ("follow the sun") or to load-balance across multiple sites. Enterprises with multiple sites can also conserve power and reduce cooling costs by dynamically consolidating virtual machines in fewer data centers.

Business Continuance: High-Availability Clusters

Despite the fact that application clustering is evolving and has started supporting deployments across L3 boundaries, there are still a long list of applications that require L2 adjacency between the cluster nodes. These applications include:

- Private inter-process communication (such as heartbeat and database replication) used to maintain and control the state of the active node.
- Public communication (from the client to the virtual IP of the cluster).

Figure 2-3 Multi-Site HA Cluster



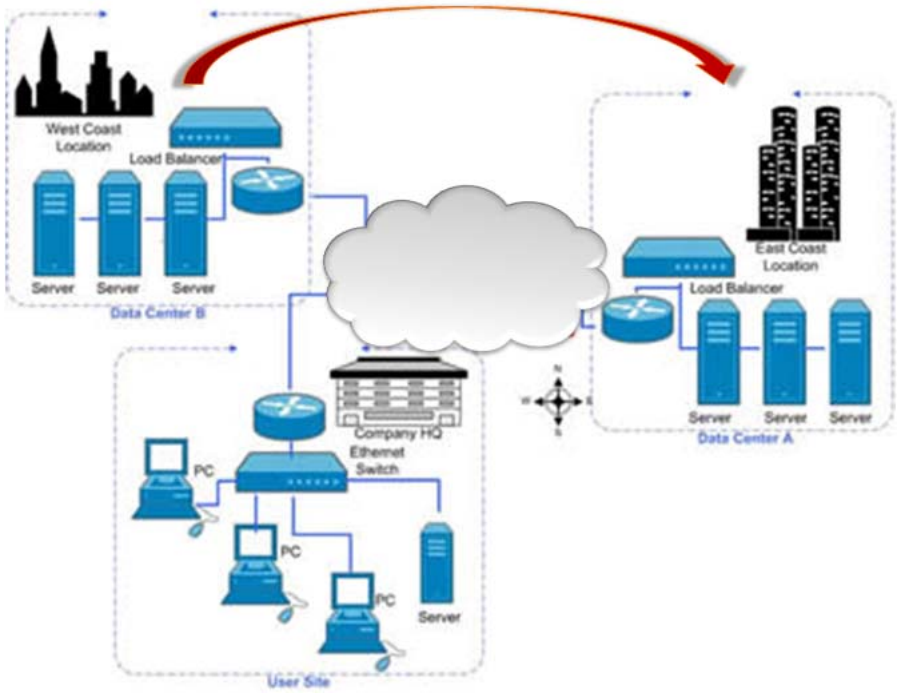
Despite the fact that application clustering is evolving and has started supporting deployments across L3 boundaries, there are still a long list of applications that require L2 adjacency between the cluster nodes. These applications include:

- Microsoft MSCS
- Veritas Cluster Server (Local)
- Solaris Sun Cluster Enterprise
- VMware Cluster (Local)
- Oracle Real Application Cluster (RAC)
- IBM HACMP
- EMS/Legato Automated Availability Manager
- NetApp Metro Cluster
- HP Metrocluster

Data Center Migration and Consolidation

Migration of servers (physical or virtual) between data centers is becoming more popular. This is often driven by different and sometimes opposite requirements, like the consolidation of a large number of DC sites into fewer ones, or the expansion of services from a few sites to multiple locations.

Figure 2-4 Data Center Migration



Providing a L2 path between sites is often desirable in these cases to ensure a smooth migration and minimize the experienced application down time. It is also important to keep in mind additional considerations:

- IP renumbering of servers to be moved is complex and costly. Avoiding IP address renumbering makes physical migration projects easier and reduces cost substantially.
- Some applications may be difficult to readdress at Layer 3 (mainframe applications, for example). In this case, it is easier to extend the Layer 2 VLAN outside the access layer, to keep the original configuration of the systems after the move.
- During phased migration, when only part of the server farm is moving at any given time, Layer 2 adjacency is often required across the whole server farm for business-continuity purposes.

LAN Extension Considerations

As mentioned above, LAN extension solutions are commonly used to extend subnets beyond the traditional Layer 3 boundaries of a single data center. Stretching the network space across two or more data centers can accomplish many things. Doing so also presents a challenge, since providing these LAN extension capabilities may have an impact on the overall network design. Simply allowing Layer 2 connectivity between sites that were originally connected only at Layer 3 would have the consequence of creating new traffic patterns between the sites: STP BPDUs, unicast floods, broadcasts, ARP requests, and so on. This can create issues, some of them related to attacks (ARP or flood storms), others related to stability issues (size of STP domain) or scale (ARP caches or MAC address table sizes). How does an extended spanning-tree environment avoid loops and broadcast storms? How does a core router know where an active IP address or subnet exists at any given time?

LAN Extension Technical Requirements

For deploying a solid LAN extension solution, it is important to keep into considerations two main following requirements:

- Spanning-Tree (STP) Isolation: the first basic requirement is to isolate the Spanning Tree domains between the data center sites belonging to the extended Layer 2 network. This is important to protect against any type of global disruptions that could be generated by a remote failure, and to mitigate the risk of propagating unwanted behavior such as topology change or root bridge movement from one data center to another. These packets could be flooded throughout the Layer 2 network, making all remote data centers and resources unstable, or even inaccessible.
- End-to-End loop prevention: In each data center site, the deployment of redundant physical devices providing LAN extension services is recommended to improve the overall resiliency of the LAN Extension solution. Therefore, a solution must eliminate any risk of creating an end-to-end Layer 2 loop; STP cannot be used for this purpose, given the previous requirement of isolating the STP domains between remote DC sites.

In addition to these, other requirements to be considered are:

- **WAN Load Balancing:** Typically, WAN links are expensive, so the uplinks need to be fully utilized, with traffic load-balanced across all available uplinks.
- **Core Transparency:** The LAN extension solution should ideally be transparent to the existing enterprise core, to minimize the operational impact.
- **Data Center Site Transparency:** The LAN extension solution should not affect the existing data center network deployment.
- **VLAN Scalability:** The solution must be able to scale to extend up to hundreds (sometimes few thousands) of VLANs.
- **Multisite Scalability:** The LAN extension solution should be able to scale to connect multiple data centers.
- **Hierarchical Quality of Service (HQoS):** HQoS is typically needed at the WAN edge to shape traffic when an enterprise subscribes to a subrate service provider service or a multipoint Ethernet virtual private line (EVPL) service.
- **Encryption:** The requirement for LAN extension cryptography is increasingly prevalent, to meet federal and regulatory requirements.

Cisco LAN Extension Solutions

Cisco LAN Extension solutions can been divided in three categories:

Ethernet Based Solutions

This category includes technologies like virtual Port-Channel (vPC) or Virtual Switching Systems (VSS), originally deployed for intra Data Center designs, but readapted for DCI deployments. The idea is to extend VLANs between remote sites by leveraging Multi Chassis Etherchannels (MCECs) established between devices deployed in different sites. As such, this solution mostly applies to point-to-point deployments, where the sites are connected via dedicated dark fiber links or protected DWDM optical circuits.

Note More information around Ethernet Based LAN Extension solutions (VSS/vPC) can be found in the following paper:
http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns949/ns304/ns975/data_center_interconnect_design_guide.pdf

It is worth noticing that an emerging technology, Cisco FabricPath, originally positioned to deploy large L2 domains inside a data Center network, could also be considered for LAN Extension purposes.

MPLS Based Solutions

This category includes MPLS technologies providing L2 connectivity services over a L3 network service. Depending on the nature of the transport infrastructure between data center sites and the number of data center sites to be interconnected, different technologies can address the connectivity requirements. EoMPLS and VPLS are usually positioned for point-to-point and multipoint deployments respectively over native MPLS based infrastructure. This is often the case with large enterprise or SP deployments. When only a generic IP service is available to interconnect different DC sites which is usually the case for small and medium enterprises acquiring connectivity services from one or more SPs, the same EoMPLS/VPLS technologies can be deployed over a logical overlay connection built leveraging GRE tunnels also known as EoMPLSoGRE or VPLSoGRE deployments.

Note More information on MPLS Based LAN Extension solutions can be found in the following paper:
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/DCI2_External.pdf

IP Based Solutions

Part of this category is an emerging Cisco technology, called Overlay Transport Virtualization (OTV). OTV is an IP based functionality that has been designed from the ground up to provide Layer 2 extension capabilities over any transport infrastructure: Layer 2 based, Layer 3 based, IP switched, label switched, and so on. The only requirement from the transport infrastructure is providing IP connectivity between remote data center sites. In addition, OTV provides an overlay that enables Layer 2 connectivity between separate Layer 2 domains while keeping these domains independent and preserving the fault-isolation, resiliency, and load-balancing benefits of an IP-based interconnection.

As of this writing, OTV is supported only on the Nexus 7000. However, there are plan to extend OTV support to other Cisco platforms in the future.

Note More information on OTV can be found in the following paper:
http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns949/ns304/ns975/OTV_intro_wp.pdf

CHAPTER 3

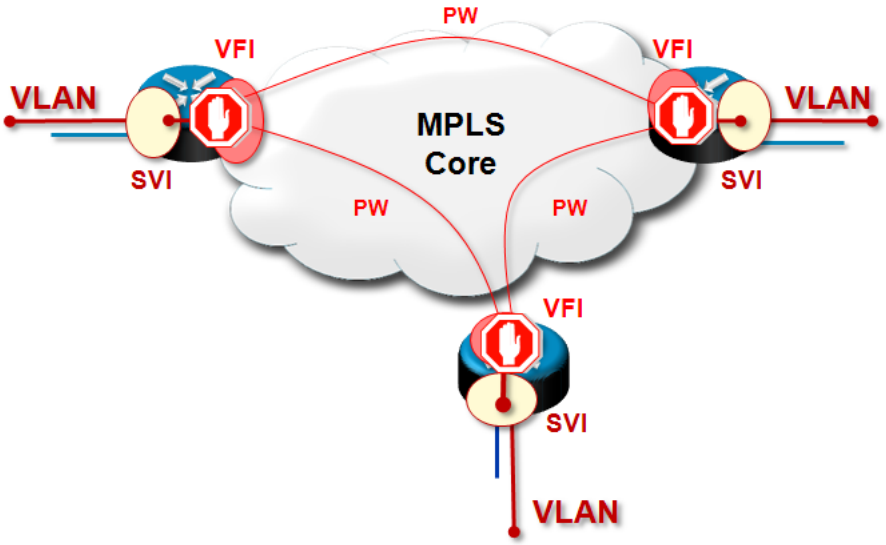
MC-LAG to VPLS Technology and Solution Overview

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.

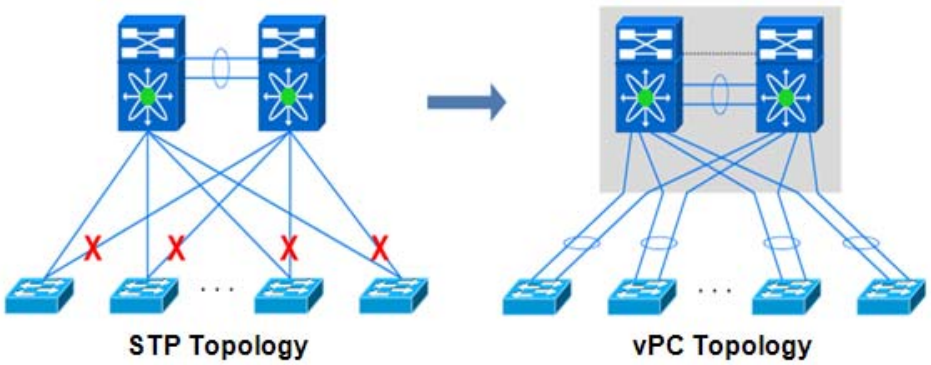
Compared to traditional LAN switching technologies, VPLS is also more flexible in its geographic scaling, so that Customer Edge (CE) sites may be within the same metropolitan domain, or may be geographically dispersed on a regional or national basis. The increasing availability of Ethernet-based multipoint 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

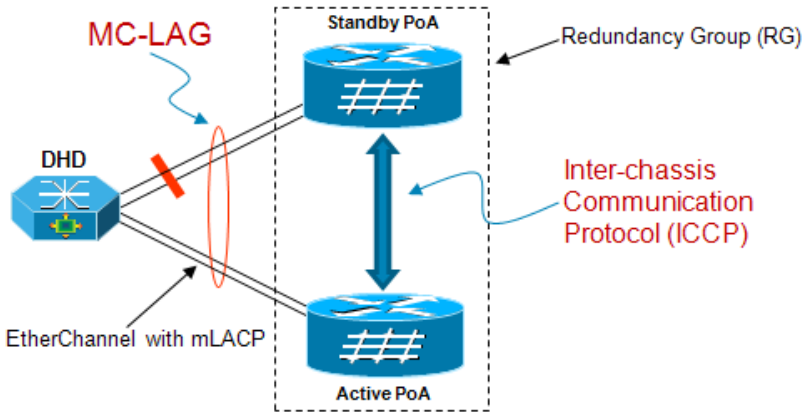
Similarly to what said in the previous section for vPC, the Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of a logical port-channel connection with a redundant pair of independent nodes. The two nodes run an independent control plane. The purpose of multi-chassis 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.

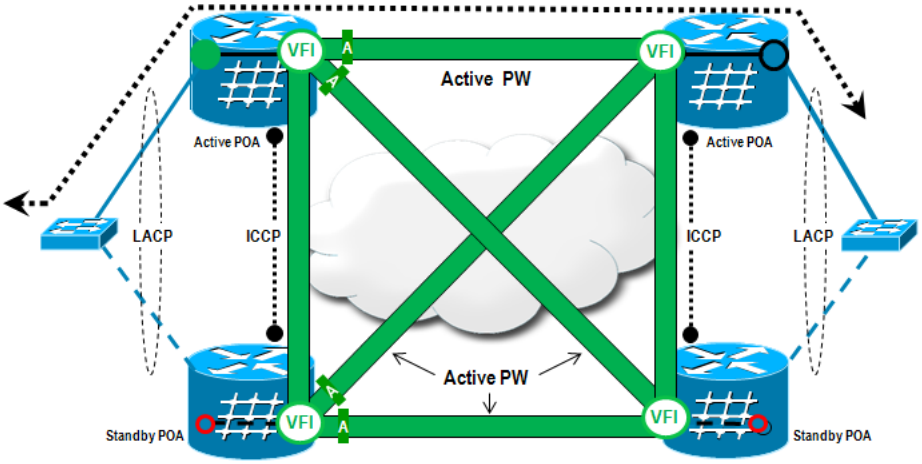
MC-LAG uses dynamic port priority management to ensure link state between DHD and POA. This mechanism involves dynamic allocation of port priority of links in the LAG such that the priority of the standby link is lower than the primary one, while trigger switchover would just be accomplished thru decrease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

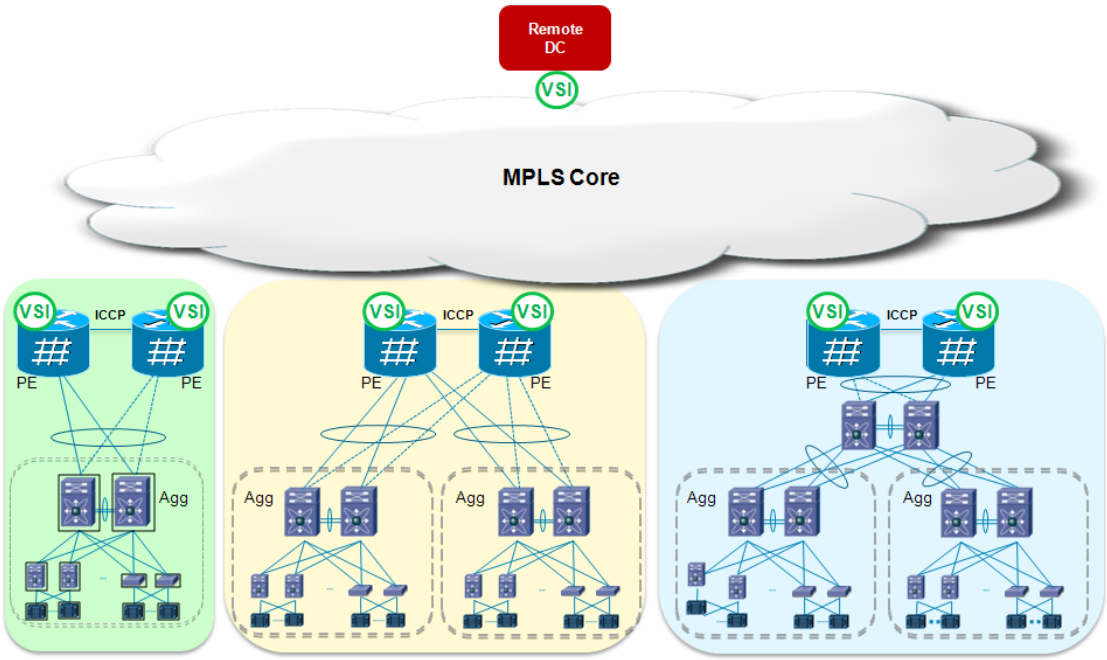
When there are n PEs in a full mesh topology, and a PE is required to flood a frame, it will have to replicate and forward them on (n-1) pseudowires. In theory, when a PE receives for example 1 Gbps of traffic from the access side and a MAC table flush results in the removal of all MAC addresses, then 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



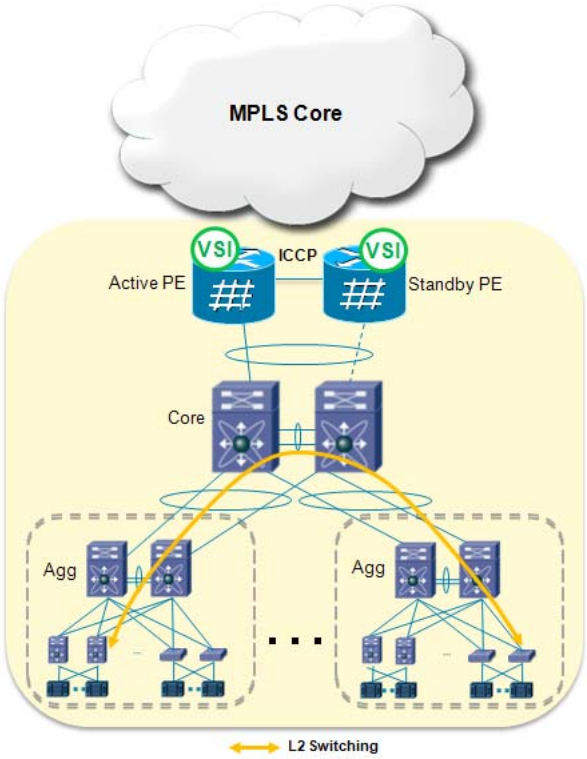
The key functional blocks of the solution are the following:

- 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.

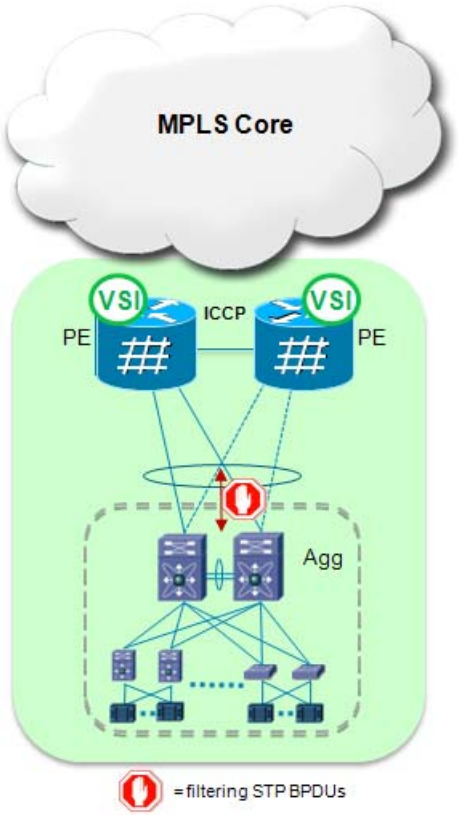
- Depending on the number of aggregation blocks deployed in a single site, the deployment of a dedicated DC core layer is possible, as shown on the right side of Figure 3-5. In that scenario, each aggregation block connects to the core leveraging vPC connections, and the core devices are now connecting 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.

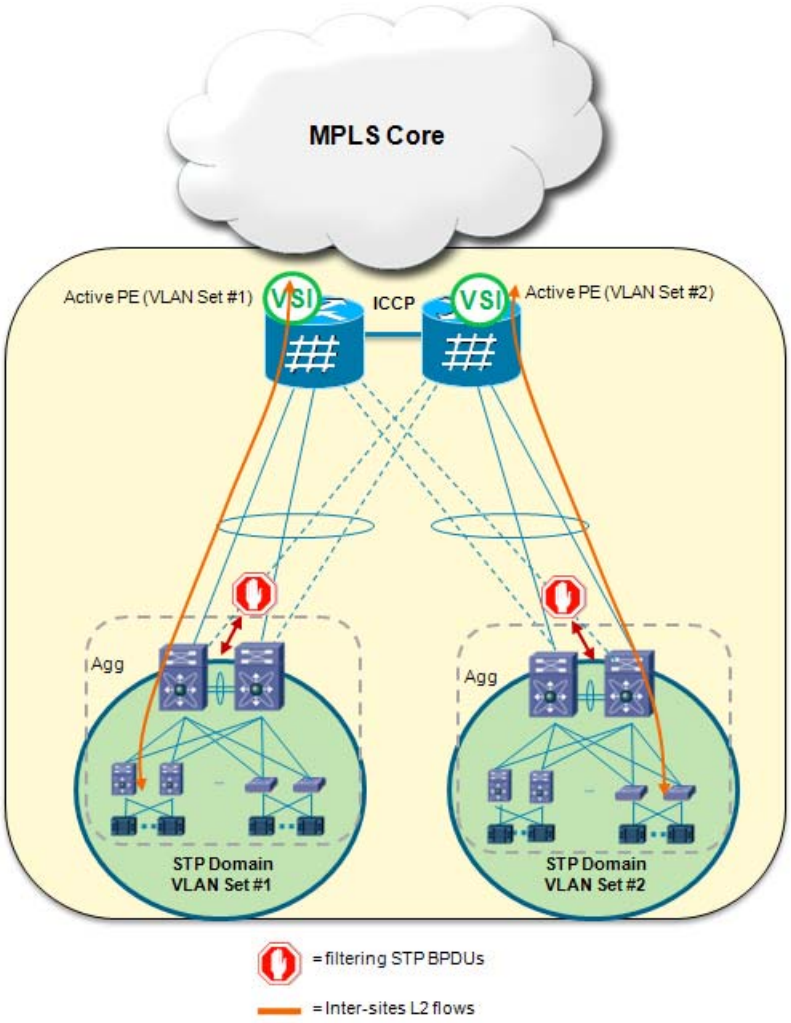
Figure 3-7 STP Isolation from each POD



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).

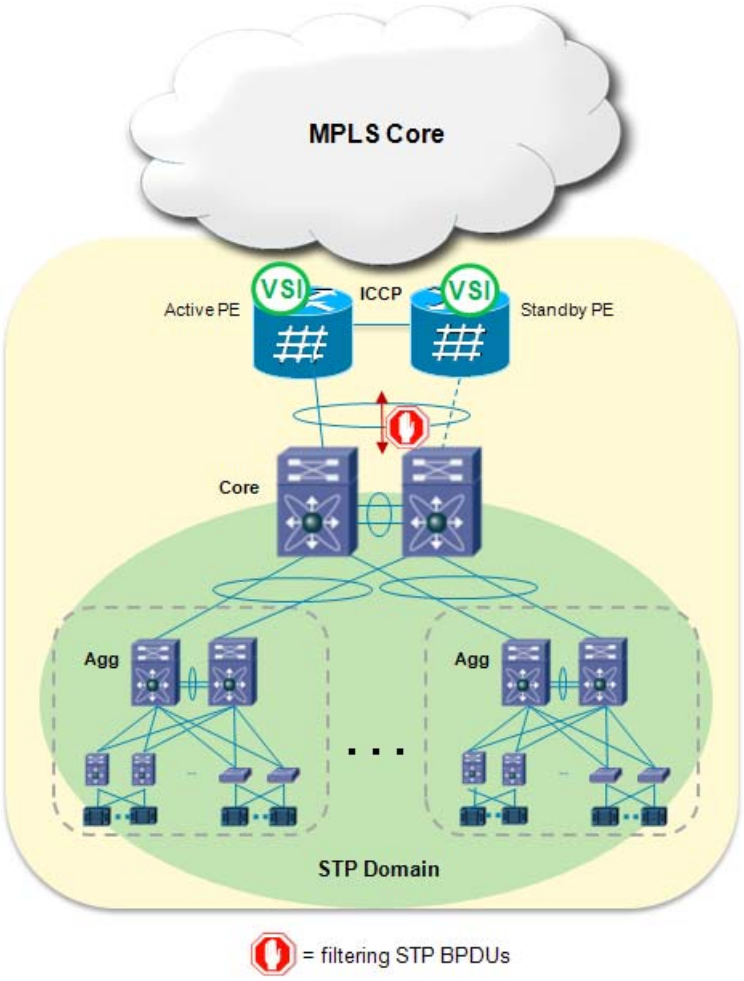
Figure 3-8 Independent VLAN Sets per POD



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.

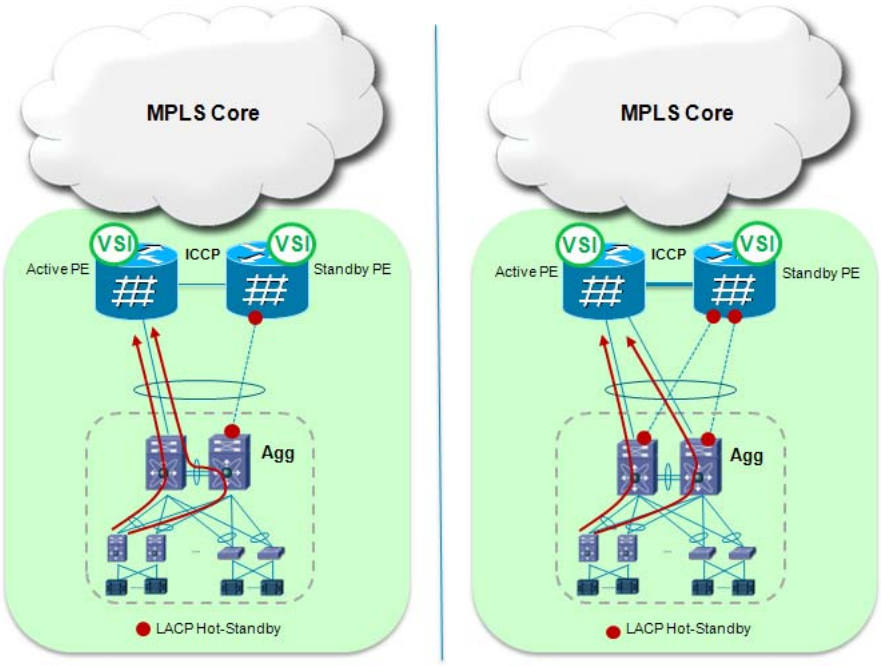
Active/Standby Deployment

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

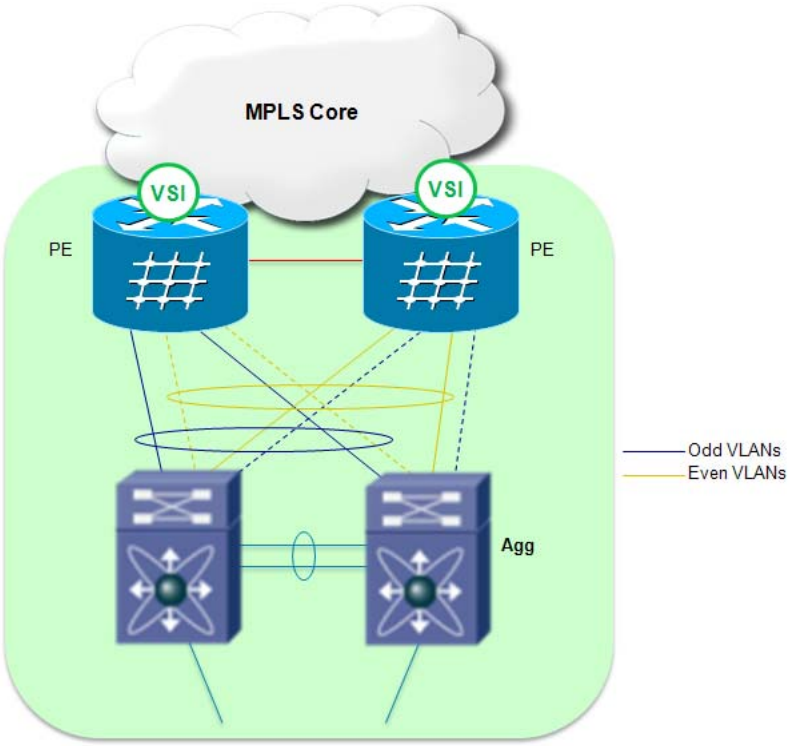


As shown in Figure 3-10, when the vPC is created with only two physical links bundled together, the outbound L2 traffic flows may follow a sub-optimal path across the peer-link connecting the aggregation layer device. This behavior can be improved by bundling together 4 links in the vPC, so that a direct 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



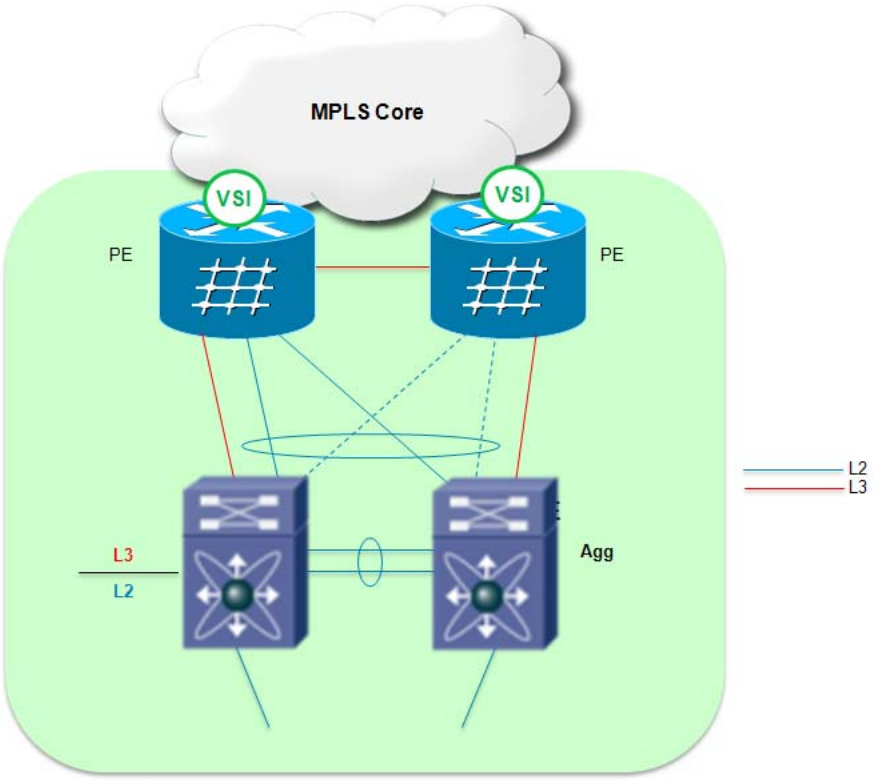
The idea is to create two logical connections between these devices, leveraging two separate vPCs and MC-LAG bundles and divide the VLANs to be extended between these. In this way, each PE router can become active for half of the VLANs to be extended and traffic can better be load-balanced both in outbound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

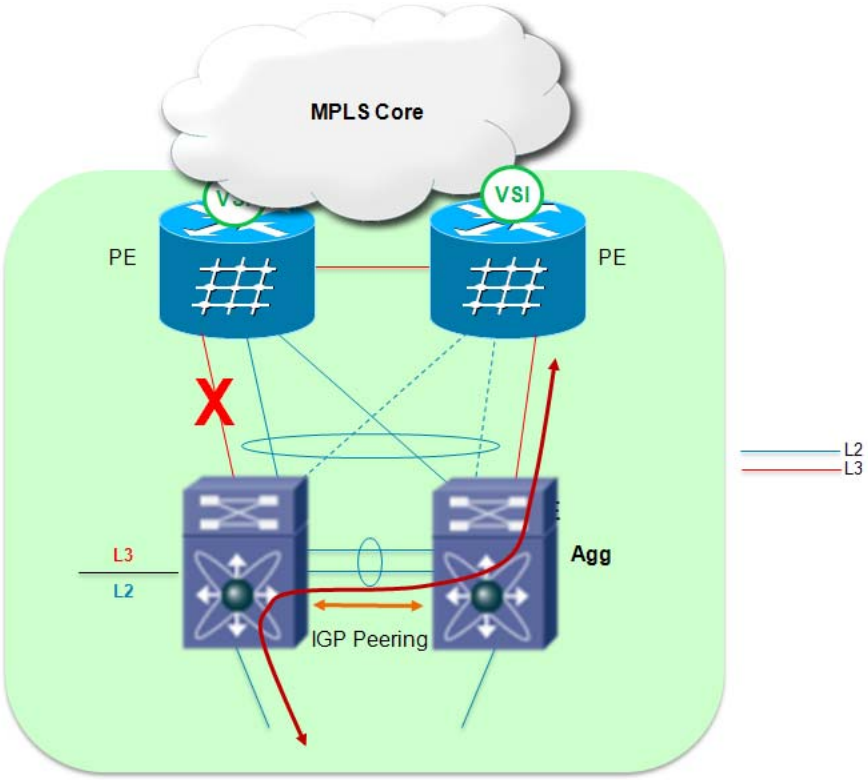


Figure 3-13 shows L3 traffic path during failure of the L3 link connecting the left aggregation device to the PE router.

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.

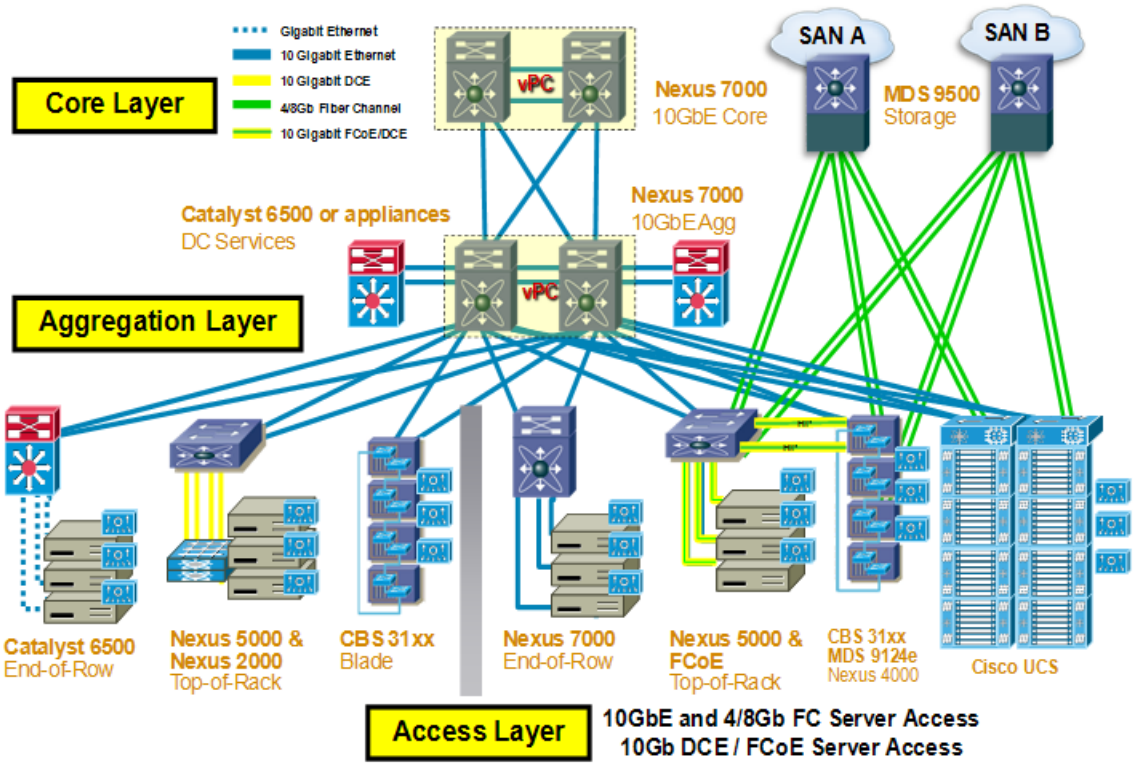
The use of QinQ is not considers as part of the scope for this paper. This implies that every extended VLAN is associated to a separate bridge domain (VSI) and corresponding VPLS PseudoWire (one VSI per VLAN deployment model).

CHAPTER 4

Data Center Multitier Model and Testbed Topology

The modern data center is evolving to meet new business requirements emanating from Web 2.0 and social network changes to the Internet and corporate networks. Figure 4-1 depicts the diversity of compute, storage, and networking elements that compose the modern data center.

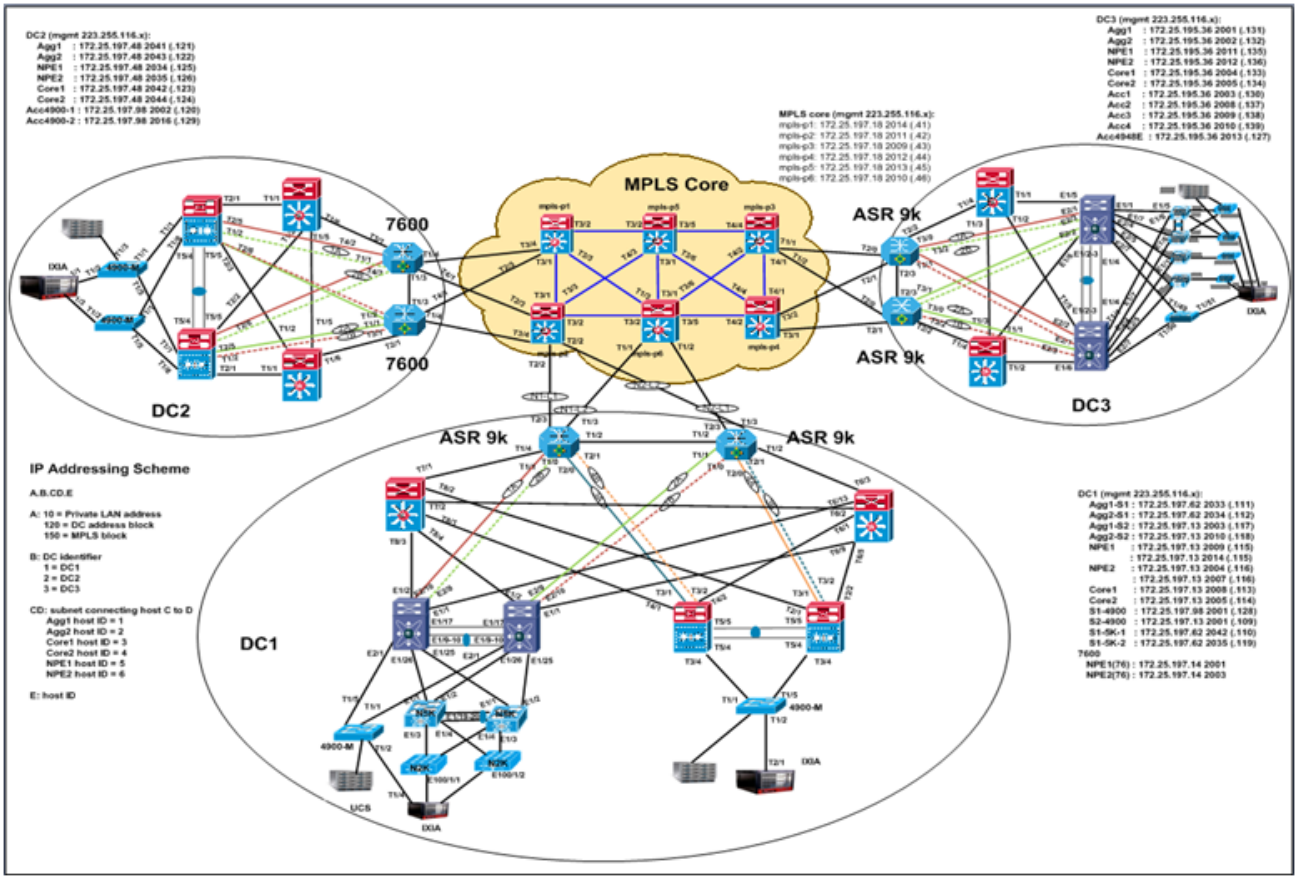
Figure 4-1 Data Center Multitier Model



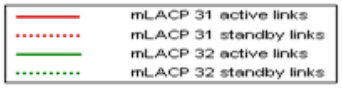
A data center is home to computational power, storage, and applications that support an enterprise business. The data center design is based on a proven layered approach, which has been tested and improved over several years in some of the largest data center implementations. The layered approach provides the basic foundation of a data center design that improves scalability, performance, flexibility, resiliency and maintenance.

The validated network design consists of data centers that interconnect via the enterprise core network. Each data center is built on the multilayer design and includes N-PE routers in addition to core, aggregation and access layer devices.

Figure 4-2 VPLS-based Data Center Interconnection - Lab Topology



Legend:



End-to-end service validation was performed by using traffic tools to generate IP unicast and multicast traffic on the network. A variety of health checks were performed, including memory and CPU utilization, trace-backs, memory alignment errors, interface errors, line card status and syslog messages. These checks were performed during every stage of network provisioning and before and after including errors and failures to determine the convergence of traffic.

Traffic Profile

This section outlines traffic flows used during extension of 500 and 1200 VLANs between data centers and the layer 2 and layer 3 test parameters configured to determine end-to-end convergence for unicast and multicast traffic.

Traffic flows

Bidirectional traffic was provisioned between DC2 and DC3, however all negative tests, i.e. link and node failure test cases were executed only in DC3 (where Cisco ASR9000 were deployed as PE routers) and network convergence was measured.

Distance between the DC2 and DC3 was 400 kms simulated using Anue and all interfaces end-to-end were 10gig Ethernet.

500 VLANs extension: Total 6.1 Gbps of aggregate traffic; 128-byte packet size

- **Unicast (1200 Flows):** 4.9 Gbps
  - Bidirectional intra-VLAN traffic: 1000 flows (L2)
  - Bidirectional inter-VLAN: 100 flows
  - Bidirectional inter-VLAN: 100 flows. These flows were provisioned for VLANs that were not extended between data centers
  - In addition, bidirectional Layer 2 traffic to generate 20,000 MAC addresses. Total 40 MAC addresses per vlan: 20 MAC addresses from DC2 and 20 MAC addresses from DC3

    500 VLANS * 40 MAC addresses = 20,000 MAC addresses
- **Multicast (300 Flows and 300 Groups):** 1.2 Gbps
  - Intra-VLAN traffic: 100 flows, 100 groups
  - Inter-VLAN: 100 flows, 100 groups
  - Inter-VLAN: 100 flows, 100 groups. These flows were provisioned for VLANs that were not extended between data centers
- **VLAN Distribution in DC3**
  - mLACP 31: Po31 – VLANs 1,76-80,100-349 = 256 VLANs
  - mLACP 32: Po32 – VLANs 1200-1499 = 250 VLANs
  - vPC to N5k : Po10 – VLANs 1,76-80,100-349 = 256 VLANs
  - 4948: Po13 – VLANs 1,1200-1499,3051-3100 (L3) = 301 VLANs

1200 VLANs extension: Total of 10.6 Gbps aggregate traffic; 128-byte packet size

- **Unicast (2590 Flows):** 10.6 Gpbs
  - Bidirectional intra-VLAN traffic: 2400 flows (L2)
  - Bidirectional inter-VLAN: 90 flows
  - Bidirectional inter-VLAN: 100 flows. These flows were provisioned for VLANs that were not extended between data centers
  - In addition, bidirectional Layer 2 traffic to generate 20,000 MAC addresses. Total 40 MAC addresses per vlan: 20 MAC addresses from DC2 and 20 MAC addresses from DC3

    500 VLANS * 40 MAC addresses = 20,000 MAC addresses

- **VLAN Distribution in DC3**
  - mLACP 31: Po31 – VLANs 1,76-80,100-999,1100-1199 = 1006 VLANs
  - mLACP 32: Po32 – VLANs 1200-1399 = 200 VLANs
  - vPC to N5k : Po10 – VLANs 1,76-80,100-349 = 256 VLANs
  - 4948: Po13 – VLANs 1,350-999,1100-1399,3051-3100 (L3) = 1001 VLANs

Test setup for Layer 2 and Layer 3 parameters

- **Layer 2**
  - Rapid PVST between the access and aggregation switches
  - vPC peer-switch on both N7K aggregation switches for single STP root in L2 topology
  - dot1q trunks and port-channels between aggregation and PEs. One active link from each N7k aggregation to ASR9k-1 and one backup link from each N7k to ASR9k-2 for one mLACP group. Two such mLACP groups configured between N7k's and ASR9k's for VLANs load-sharing
  - Storm-control on N7k port-channels facing PEs to prevent storm propagating from remote datacenters
- **Layer 3**
  - HSRP configured on agg switches, with N7k as vPC primary being HSRP active and N7k as vPC secondary being HSRP standby
  - 500 SVIs across 16 VRFs; 33 SVIs per each VRF
  - OSPF between aggregation and core switches and between core and PE devices

Table 4-1 provides L2 and L3 test parameters summary.

Table 4-1 L2 and L3 Test Parameters

Datacenter 2 (DC2)	Datacenter 3 (DC3)
• N-PEs – 2 x Cisco 7600 – 160 IGP routes – 2 mLACP groups for 500 VLANs (active/standby) – 500 VFIs (1 VFI / VLAN) • AGGREGATION – 2 x Catalyst 6500 (VSS) – 500 (1200) VLANs + OSPF adjacency (L3 traffic) – 500 SVI 33 SVIs in 1 VRF 16 VRFs for 500 VLANs – 4 MEC po10, po30 towards access po21, po22 towards PEs – Storm control	• N-PEs – 2 x ASR9000 – 160 IGP routes – 2 mLACP groups for 500 VLANs (active/standby) – 500 VFIs (1 VFI / VLAN) • AGGREGATION – 2 x Nexus 7000 (vPC) – 500 (1200) VLANs + OSPF adjacency (L3 traffic) – 500 SVI 33 SVIs in 1 VRF 16 VRFs for 500 VLANs – 4 vPCs po10, po13 towards access po31, po32 towards PEs Storm control

Note DC3 was the portion of the network under test. All the considerations and test results contained in this paper apply to that specific part of the overall topology.

CHAPTER 5

Deploying MC-LAG to VPLS Solution

This solution requires that the two Nexus 7000 aggregation switches be converted to vPC and connected to both the PE routers via mLACP. The PWs on the both the N-PEs are in active state because of the decoupled mode of operation where the state of the attachment circuits controlled by mLACP is independent from the state of the PWs.

This section provides information about the test scope, hardware and software version used during design validation and includes key configuration details to implement this solution.

Scope

This engagement validated the MC-LAG to VPLS LAN extension solution with Cisco ASR 9000 series routers as PE. The testing was performed for the following features:

Table 5-1 **Test Scope**

Features/Tests	Description
L2 Interconnect	Verify L2 Interconnect between data centers and VLAN extension using VPLS
VPLS	Verify VPLS using ASR 9000 (one VFI/VLAN) as N-PE with Nexus 7000 as aggregation devices
mLACP	Verify mLACP on ASR 9000 to achieve N-PE redundancy
RSTP and HSRP	Verify RSTP and HSRP functionality
Storm Control	Verify Storm Control functionality
VLAN extension	Extend 500 and 1200 VLANs between data centers
Negative Tests	
Reload	Reload Nexus 7000and ASR 9000
Link Failure	Shut/No Shut links between various nodes
SSO	SSO on Nexus 7000
ISSU	ISSU on Nexus 7000

Hardware and Software

Table 5-2 lists all the hardware and software used in validating MC-LAG based VPLS solution to interconnect data centers.

Table 5-2 **Hardware and Software Information**

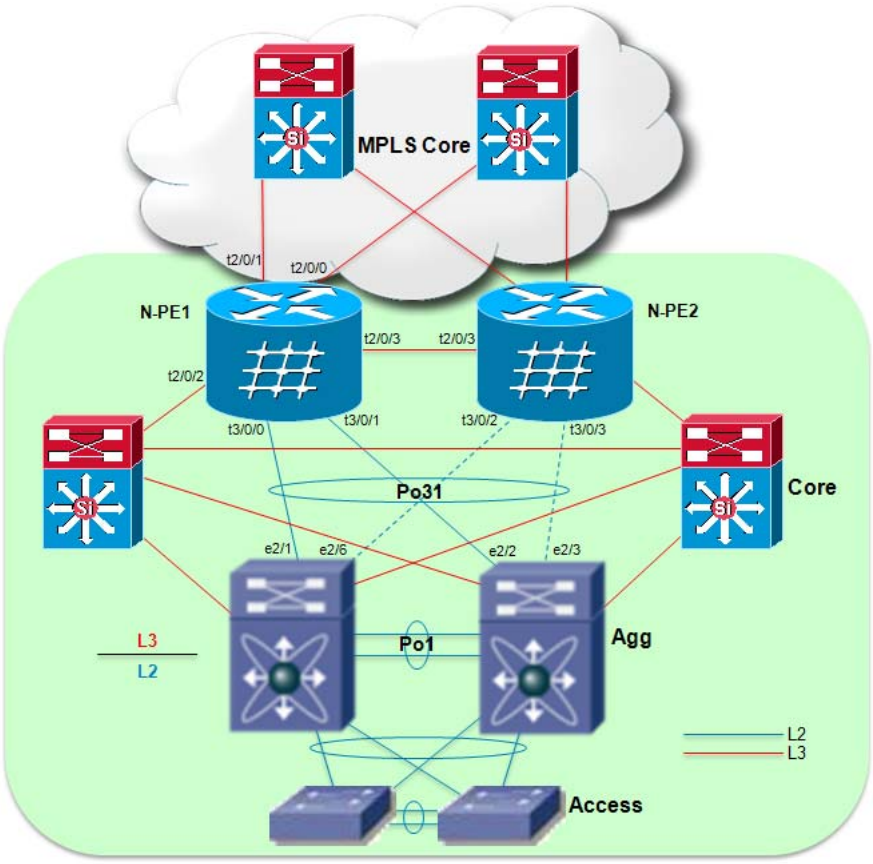
Qty	Hardware Platform	Software Version	Role	Line cards / Interfaces
2	Nexus 7010	Engineering image based on 5.1.2 [1]	DC3 Aggregation	N7K-M108X2-12L N7K-M148GT-11
4	Catalyst 6500	12.2(33)SXI3	DC2 and DC3 core	WS-6704-10GE WS-6708-10GE
2	Cisco 7600	12.2(33)SRE2	DC2 PE	7600-ES+4TG3CXL 7600-ES+XT-2TG3CXL 7600-ES+2TG3CXL
2	ASR 9000	4.0.1	DC3 PE	
2	Nexus 5000	4.2(1)N2(1)	DC3 Access	
1	Cisco WS-C4948E	12.2 (54)SG	DC3 Access	
1	Nexus 1000V	4.0(4)SV1(3a)	Virtual Switch	
2	Fabric Interconnect	4.1(3)N2(1.3i)	Fabric Interconnect for UCS	
2	UCS chassis	1.3(1i)	Blade server	
2	Cisco C4900M	12.2(50)SG	DC2 Access	

1. Engineering image based on NX-OS version 5.1(2) was used during testing on Nexus 7000 switches since issues attributing to high convergence times under failure conditions were uncovered. Thus, it is strongly recommended to use this NX-OS software release 5.2 (and higher) to deploy the solution object of this paper.

Configuration Details

The following list provides snippets of configuration from the PEs and aggregation devices and output from various show commands for verification. All the configuration samples refer to the DC3 site, which is represented again for reference in Figure 5-1.

Figure 5-1 DC3 Site Under Test



Note In the actual testbed, an active/active MC-LAG deployment was validated, leveraging a redundant set of connections between the aggregation devices and the PE routers (as previously shown in Figure 15). For the sake of simplicity, only one set of cables is shown in Figure 5-1. Also, multiple access layer switches were actually connected to the two aggregation devices (to allow carrying up to 1200 VLANs toward the PE routers), even if the network diagram above shows only a pair of Nexus 5000 for simplicity sake.

Step 1 Configure IGP, MPLS and targeted-LDP on PE Routers.

a. OSPF (IGP) configuration on ASR9000 routers: Cisco recommends enabling BFD to detect failures in the path between adjacent L3 enabled interfaces. Also, it is recommended to tune the OSPF timers (throttle lsa, spf and lsa min-arrival) to ensure to run a faster SPF calculation after notification of a topology change event (arrival of an LSA).

```
router ospf ospf300
 bfd minimum-interval 200
 bfd multiplier 3
 timers throttle lsa all 100 100 5000
 timers throttle spf 100 100 5000
 timers lsa min-arrival 80
 area 0
  interface Loopback100
  !
  interface TenGigE0/2/0/0 << MPLS core facing interface
```

```
   bfd fast-detect
   network point-to-point
  !
  interface TenGigE0/2/0/1 << MPLS core facing interface
   bfd fast-detect
   network point-to-point
  !
  interface TenGigE0/2/0/2 << Local datacenter core facing interface
   bfd fast-detect
   network point-to-point
  !
  interface TenGigE0/2/0/3 << Connected to N-PE2 (Redundant PE in the same datacenter)
   bfd fast-detect
   network point-to-point
  !
 !
!
RP/0/RSP1/CPU0:DC3-ASR9K-NPE1#show ospf neighbor

* Indicates MADJ interface

Neighbors for OSPF ospf300

Neighbor ID     Pri   State           Dead Time   Address         Interface
150.1.1.103     1     FULL/  -        00:00:38    150.3.36.3      TenGigE0/2/0/0
    Neighbor is up for 12:55:08
150.1.1.104     1     FULL/  -        00:00:32    150.3.46.4      TenGigE0/2/0/1
    Neighbor is up for 12:55:07
120.3.1.4       1     FULL/  -        00:00:38    120.3.46.4      TenGigE0/2/0/2
    Neighbor is up for 12:55:08
150.3.3.5       1     FULL/  -        00:00:39    150.3.33.5      TenGigE0/2/0/3
    Neighbor is up for 12:55:07

Total neighbor count: 4
```

b. Configure MPLS and LDP.

```
mpls ldp
 router-id 150.3.3.5
 session protection << MPLS LDP session protection
 interface TenGigE0/2/0/0 << MPLS enabled interfaces
 !
 interface TenGigE0/2/0/1
 !
 interface TenGigE0/2/0/3
 !
!

RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#show mpls interfaces
Interface                  LDP      Tunnel   Enabled
-------------------------- -------- -------- --------
TenGigE0/2/0/0             Yes      No       Yes
TenGigE0/2/0/1             Yes      No       Yes
TenGigE0/2/0/3             Yes      No       Yes
```

Step 2 Configure Inter-chassis Communication Protocol (ICCP) on PEs.

Table 5-3 shows two mLACP groups, group 31 and 32, are configured for load-sharing VLANs between the PEs (as previously discussed in the “Active/Active Deployment with Redundant Physical Connections” section in Chapter 3). Notice that the required configuration on the two PE devices is

pretty much identical, with the exception of few commands. Particularly important is to configure the same mLACP system MAC value, because the key principle of MC-LAG is to make the two PE devices appearing as a single entity to the Nexus 7000 switches deployed at the aggregation layer.

Table 5-3 LACP groups 31 and 32 configured for load-sharing VLANs between the N-PEs

On N-PE1:

```
redundancy
 iccp
  group 31
   mlacp node 1 << Unique on each POA
   mlacp system mac 0000.0000.0031 << LACP
system ID. Recommended to be same on both
N-PEs
   mlacp system priority 1  <<  Recommended
be lower than DHD (Aggregation)
   mlacp connect timeout 0
   member
    neighbor 150.3.3.6 << N-PE2 loopback
address
   !
   backbone
    interface TenGigE0/2/0/0 << Facing MPLS
core
    interface TenGigE0/2/0/1 << Facing MPLS
core
   !
   isolation recovery-delay 100 << Used
when core isolation condition is cleared
  !
  group 32
   mlacp node 2
   mlacp system mac 0000.0000.0032
   mlacp system priority 2
   mlacp connect timeout 0
   member
    neighbor 150.3.3.6
   !
   backbone
    interface TenGigE0/2/0/0
    interface TenGigE0/2/0/1
   !
   isolation recovery-delay 100
  !
 !
!
```

On N-PE2:

```
redundancy
 iccp
  group 31
   mlacp node 2 << Unique on each POA
   mlacp system mac 0000.0000.0031 << LACP
system ID. Recommended to be same on both
N-PEs
   mlacp system priority 1 << Recommended
be lower than DHD (Aggregation)
   mlacp connect timeout 0
   member
    neighbor 150.3.3.5 << N-PE1 loopback
address
   !
   backbone
    interface TenGigE0/2/0/0 << Facing MPLS
core
    interface TenGigE0/2/0/1 << Facing MPLS
core
   !
   isolation recovery-delay 100 << Used
when core isolation condition is cleared
  !
  group 32
   mlacp node 1
   mlacp system mac 0000.0000.0032
   mlacp system priority 1
   mlacp connect timeout 0
   member
    neighbor 150.3.3.5
   !
   backbone
    interface TenGigE0/2/0/0
    interface TenGigE0/2/0/1
   !
   isolation recovery-delay 100
  !
 !
!
```

On N-PE1:

```
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#show iccp group 31
Redundancy Group 31
  member ip:150.3.3.6 (DC3-ASR9K-NPE2), up (connected)
    monitor: route-watch (up)
  backbone interface Te0/2/0/0: up
  backbone interface Te0/2/0/1: up
  enabled applications: mLACP, IGMP Snooping
  isolation recovery delay timer: 100 s, not running
```

On N-PE2:

```
RP/0/RSP1/CPU0:DC3-ASR9K-NPE2#show iccp group 31
Redundancy Group 31
```

```
member ip:150.3.3.5 (DC3-ASR9K-NPE1), up (connected)
  monitor: route-watch (up)
backbone interface Te0/2/0/0: up
backbone interface Te0/2/0/1: up
enabled applications: mLACP, IGMP Snooping
isolation recovery delay timer: 100 s, not running
```

Step 3 Configure Attachment Circuits on each N-PE (interface towards aggregation switch)

a. Configure interfaces that are members of port-channel connecting to aggregation switches. On each PE the 4 physical interfaces were assigned to the two separate MC-LAG groups previously defined to achieve active/active traffic load-balancing.

```
interface TenGigE0/3/0/0
 bundle id 31 mode active << Bundle 31
 cdp
 lacp period short << lacp fast hellos
 lacp period short transmit 100
 carrier-delay up 0 down 0
!
interface TenGigE0/3/0/1
 bundle id 31 mode active
 cdp
 lacp period short
 lacp period short transmit 100
 carrier-delay up 0 down 0
!
interface TenGigE0/3/0/2
 bundle id 32 mode active
 cdp
 lacp period short
 lacp period short transmit 100
 carrier-delay up 0 down 0
!
interface TenGigE0/3/0/3
 bundle id 32 mode active
 cdp
 lacp period short
 lacp period short transmit 100
 carrier-delay up 0 down 0
```

b. Configure Bundle interfaces: these are the logical interfaces equivalent to a port-channel on ASR 9000 devices. Each defined bundle interface is assigned to a specific ICCP group (previously created).

```
interface Bundle-Ether31
 lacp switchover suppress-flaps 100
 mlacp iccp-group 31 << Attach this bundle to previously defined ICCP group
 mlacp switchover type revertive << Automatic switchback to Primary role upon failure
recovery
 mlacp switchover recovery-delay 40 << Timer to wait before switching back
 bundle wait-while 0 << Recommended to be set to 0 to improve convergence
 bundle maximum-active links 2 << Defines maximum active links in port-channel.
 !
interface Bundle-Ether32
 lacp switchover suppress-flaps 100
 mlacp iccp-group 32 << Attach this bundle to previously defined ICCP group
 mlacp switchover type revertive << Automatic switchback to Primary role upon failure
recovery
 mlacp switchover recovery-delay 40 << Timer to wait before switching back
 bundle wait-while 0 << Recommended to be set to 0 to improve convergence
 bundle maximum-active links 2 << Defines maximum active links in port-channel.
```

c. Configure sub-interfaces for all the VLANs to be extended on both N-PEs: a separate sub-interface is created for each VLAN that needs to be extended across the MPLS core.

```
interface Bundle-Ether31.100 l2transport
 encapsulation dot1q 100
 rewrite ingress tag pop 1 symmetric
!
interface Bundle-Ether31.101 l2transport
 encapsulation dot1q 101
 rewrite ingress tag pop 1 symmetric
!
interface Bundle-Ether31.102 l2transport
 encapsulation dot1q 102
 rewrite ingress tag pop 1 symmetric
!
interface Bundle-Ether31.103 l2transport
 encapsulation dot1q 103
 rewrite ingress tag pop 1 symmetric
!
…
!
```

Step 4 Pseudowire configuration on PEs (one VFI per VLAN).

The VPLS configuration that was validated leverages a separate VFI and bridge domain for each VLAN that needed to be extended via VPLS. In IOS-XR this can be achieved by defining a “bridge group” container, under which all the different bridge domains and VFIs are created. ASR9000 supports the use of BGP for auto-discovery of the neighbor PE devices, but in this specific validation effort leveraged static neighbor configuration.

```
l2vpn << l2vpn configuration mode
pw-status
 logging
  pseudowire << Enable Pseudowire status logging
 !
 pw-class vpls-pw-class << PW class to enable mpls encapsulation
  encapsulation mpls
  !
 !
 bridge group group1
    bridge-domain vlan100  << Define bridge domain
     interface Bundle-Ether31.100 << Aggregation facing sub-interface for VFI
     !
     vfi vfi100  << Define VFI
      neighbor 150.2.2.5 pw-id 100 << Pseudowire peer with VC identifier
       pw-class vpls-pw-class
      !
      neighbor 150.2.2.6 pw-id 100  << Pseudowire peer
       pw-class vpls-pw-class
      !
      neighbor 150.3.3.6 pw-id 100  << Pseudowire peer
       pw-class vpls-pw-class
      !
      neighbor 150.11.11.5 pw-id 100 << Pseudowire peer
       pw-class vpls-pw-class
      !
      neighbor 150.11.11.6 pw-id 100 << Pseudowire peer
       pw-class vpls-pw-class
      !
     !
    !
    bridge-domain vlan101
     interface Bundle-Ether31.101
     !
     vfi vfi101
```

```
      neighbor 150.2.2.5 pw-id 101
       pw-class vpls-pw-class
      !
      neighbor 150.2.2.6 pw-id 101
       pw-class vpls-pw-class
      !
      neighbor 150.3.3.6 pw-id 101
       pw-class vpls-pw-class
      !
      neighbor 150.11.11.5 pw-id 101
       pw-class vpls-pw-class
      !
      neighbor 150.11.11.6 pw-id 101
       pw-class vpls-pw-class
      !
     !
    !
```

Similarly configure PWs for all VLANs to be extended between data centers

Verify that Pseudowire is in UP state

```
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#sh l2vpn bridge-domain group group1 brief | inc
group1:vlan100
group1:vlan100              5     up         1/1            5/5
```

Step 5 Configure virtual port-channel (vPC) on Nexus 7000 aggregation switches and filter HSRP hellos from remote data centers.

a. Configure member interface to be bundled in virtual port-channel

On Agg1:

```
interface Ethernet2/1
  lacp rate fast  << LACP fast hellos
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1,76-80,100-349  << VLANs to be extended to remote
data centers
  channel-group 31 mode active
  no shutdown

interface Ethernet2/2
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1200-1449
  channel-group 32 mode active
  no shutdown

interface Ethernet2/3
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1200-1449
  channel-group 32 mode active
  no shutdown

interface Ethernet2/6
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1,76-80,100-349
  channel-group 31 mode active
  no shutdown
```

On Agg2:

```
interface Ethernet2/1
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1200-1449
  channel-group 32 mode active
  no shutdown

interface Ethernet2/2
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1,76-80,100-349
  channel-group 31 mode active
  no shutdown

interface Ethernet2/3
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1,76-80,100-349
  channel-group 31 mode active
  no shutdown

interface Ethernet2/6
  lacp rate fast
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1200-1449
  channel-group 32 mode active
  no shutdown
```

b. Configure port-channel interface on both Nexus 7000 aggregation switches

On Agg1:

```
interface port-channel31
  switchport
  switchport mode trunk
  ip port access-group HSRP_Deny in   << Filter HSRP hellos from remote datacenters
  switchport trunk allowed vlan 1,76-80,100-349
  spanning-tree port type edge trunk   << Define edge port to improve convergence
  spanning-tree bpdufilter enable   << Filter BPDU for STP isolation
  lacp max-bundle 1   << Maximum links to be active. Configure 1 on both N7k
  vpc 31   << vPC domain for virtual port-channel

interface port-channel32
  switchport
  switchport mode trunk
  ip port access-group HSRP_Deny in
  switchport trunk allowed vlan 1200-1449
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  lacp max-bundle 1
  vpc 32
```

On Agg2:

```
interface port-channel31
  switchport
  switchport mode trunk
  ip port access-group HSRP_Deny in
```

```
  switchport trunk allowed vlan 1,76-80,100-349
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  lacp max-bundle 1
  vpc 31

interface port-channel32
  switchport
  switchport mode trunk
  ip port access-group HSRP_Deny in
  switchport trunk allowed vlan 1200-1449
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  lacp max-bundle 1
  vpc 32
```

c. Access list to filter HSRP hellos configured on both aggregation switches

```
ip access-list HSRP_Deny
  statistics per-entry
  10 deny udp any 224.0.0.102/32 eq 1985 << Filter specific to HSRP v2 hellos (for
HSRP v1 that are the default version used on Nexus 7000the address to use would be
224.0.0.2/32)
  20 permit ip any any
```

d. Verify vPC and port-channel status

```
DC3-N7K-AGG1# show vpc brief
Legend:
            (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                   : 1
Peer status                     : peer adjacency formed ok
vPC keep-alive status           : peer is alive
Configuration consistency status: success
Type-2 consistency status       : success
vPC role                        : primary
Number of vPCs configured       : 5
Peer Gateway                    : Disabled
Dual-active excluded VLANs      : -

vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po1    up     1,76-80,100-999,1100-2199,3051-3100

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason              Active vlans
--   ----   ------ ----------- ------              ------------
..
31   Po31   up     success     success             1,76-80,100-599
..
```

Step 6 Verify that MC-LAG operation on ASR9000 and Nexus 7000 Configure Nexus 7000 aggregation switches for virtual port-channel (vPC).

Step 7 The final results of the LACP negotiation between the PE routers and the aggregation layer switches is to activate, on a per redundancy group basis, all the links connected to the Primary PE device. The connections to the Secondary PE must end up in Hot-Standby state, and this needs to happen both on the PE and aggregation devices, to avoid traffic black holing. The following CLI commands allow verifying this behavior.

On N-PE1:

```
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#show bundle bundle-ether 31

Bundle-Ether31
  Status:                                  Up
  Local links <active/standby/configured>: 2 / 0 / 2
  Local bandwidth <effective/available>:   20000000 (20000000) kbps
  MAC address (source):                   18ef.63e4.249b (Chassis pool)
  Minimum active links / bandwidth:        1 / 1 kbps
  Maximum active links:                    2
  Wait while timer:                        Off
  LACP:                                    Operational
    Flap suppression timer:                100 ms
  mLACP:                                   Operational
    ICCP Group:                            31
    Role:                                  Active
    Foreign links <active/configured>:     0 / 2
    Switchover type:                       Revertive
    Recovery delay:                        40 s
    Maximize threshold:                    1 link
  IPv4 BFD:                                Not configured

  Port           Device         State      Port ID           B/W, kbps
  -------------  -------------  ---------  ------------      ---------
  Te0/3/0/0      Local          Active     0x82d9, 0x9001    10000000
      Link is Active
  Te0/3/0/1      Local          Active     0x82d9, 0x9002    10000000
      Link is Active
  Te0/3/0/2      150.3.3.6      Standby    0x82da, 0xa003    10000000
      Link is marked as Standby by mLACP peer
  Te0/3/0/3      150.3.3.6      Standby    0x82da, 0xa004    10000000
      Link is marked as Standby by mLACP peer
```

On N-PE2:

```
RP/0/RSP1/CPU0:DC3-ASR9K-NPE2#show bundle bundle-ether 31

Bundle-Ether31
  Status:                                  mLACP hot standby
  Local links <active/standby/configured>:   0 / 2 / 2
  Local bandwidth <effective/available>:     0 (0) kbps
  MAC address (source):                    18ef.63e4.249b (Peer)
  Minimum active links / bandwidth:          1 / 1 kbps
  Maximum active links:                      2
  Wait while timer:                          Off
  LACP:                                      Operational
    Flap suppression timer:                  100 ms
  mLACP:                                     Operational
    ICCP Group:                              31
    Role:                                    Standby
    Foreign links <active/configured>:       2 / 2
    Switchover type:                         Revertive
    Recovery delay:                          40 s
    Maximize threshold:                      1 link
  IPv4 BFD:                                  Not configured

  Port            Device        State        Port ID          B/W, kbps
  --------------  ------------  -----------  --------------   ----------
  Te0/3/0/2       Local         Standby      0x82da, 0xa003   10000000
      mLACP peer is active
  Te0/3/0/3       Local         Standby      0x82da, 0xa004   10000000
      mLACP peer is active
  Te0/3/0/0       150.3.3.5     Active       0x82d9, 0x9001   10000000
```

```
      Link is Active
  Te0/3/0/1       150.3.3.5     Active       0x82d9, 0x9002  10000000
      Link is Active
```

On Agg1:

```
DC3-N7K-AGG1#show port-channel summary
Flags:  D - Down        P - Up in port-channel (members)
        I - Individual  H - Hot-standby (LACP only)
        s - Suspended   r - Module-removed
        S - Switched    R - Routed
        U - Up (port-channel)
        M - Not in use. Min-links not met
--------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports      Channel
--------------------------------------------------------------------------
1     Po1(SU)     Eth      LACP      Eth1/2(P)    Eth1/3(P)
..
31    Po31(SU)    Eth      LACP      Eth2/1(P)    Eth2/6(H)
..
DC3-N7K-AGG1#..
```

On Agg2:

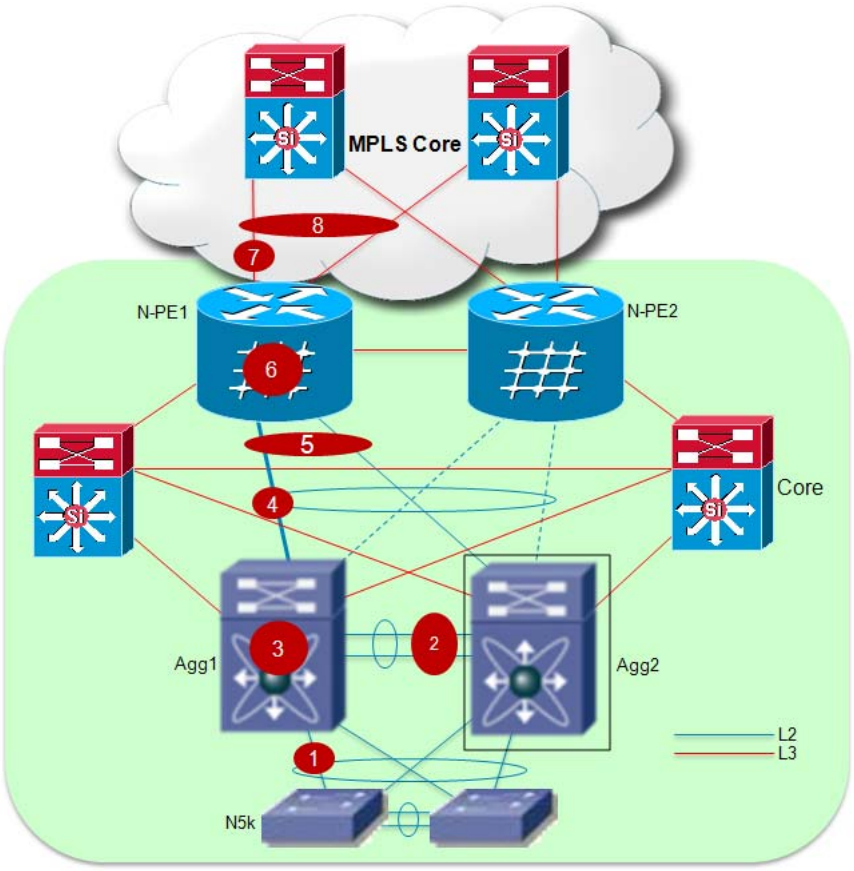
```
DC3-N7K-AGG2#show port-channel summary
Flags:  D - Down        P - Up in port-channel (members)
        I - Individual  H - Hot-standby (LACP only)
        s - Suspended   r - Module-removed
        S - Switched    R - Routed
        U - Up (port-channel)
        M - Not in use. Min-links not met
--------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports      Channel
--------------------------------------------------------------------------
1     Po1(SU)     Eth      LACP      Eth1/2(P)    Eth1/3(P)
..
31    Po31(SU)    Eth      LACP      Eth2/2(P)    Eth2/3(H)
....
```

Convergence Tests

Convergence testing was performed to measure convergence times for unicast and multicast traffic during various link and node failures. Convergence was measured from the data source to the receiver (end-to-end network convergence) by determining packet loss for each flow. For example, a packet rate of 1000 packets per second (pps) corresponds to 1-millisecond (ms) convergence time for each packet dropped.

Figure 5-2 provides a view of the different failure and recovery scenarios that were validated.

Figure 5-2 Failure/Recovery Test Cases



Each failure and recovery scenario will be analized, discussing the mechanisms leveraged to recover traffic flows and presenting the specific test results achieved during the validation effort. Test results will be presented in the two specific scenarios where 500 and 1200 VLANs were extended by leveraging two separate MC-LAG groups, as shown in Figure 2-1. For the 500 VLANs extension case, each MC-LAG group was carrying 250 VLANs; for the 1200 VLANs extension case, one MC-LAG group (the one to which failure/recovery scenarios were applied) was carrying 1000 VLANs, whereas the other one was carrying 200 VLANs.

To better understanding the convergence results achieved, it is important to describe the various traffic flows that were used during testing, distinguishing the cases where 500 VLANs or 1200 VLANs were extended.

500 VLANs

- **L2:Intra-VLAN-100-349:** these are pure L2 traffic flows (250 VLANs) extended between DC2 and DC3 sites leveraging the first MC-LAG group. Most of the failure/recovery scenarios directly affected links and devices relative to this MC-LAG group, so it is expected to notice traffic outage.
- **L2:Intra-VLAN-1200-1449:** these are pure L2 traffic flows (250 VLANs) extended between DC2 and DC3 sites leveraging the second MC-LAG group. In most of the validated failure scenarios, traffic flowing across this connection should remain unaffected. This is because most of the link/node failures are performed to only affect the connections forming the first MC-LAG group.
- **L3L2: Inter-VLAN:** these are traffic flows that are first routed at the aggregation layer between a source and a destination VLAN. The destination VLAN is then extended to the remote site, so the behavior is expected to be very similar to the pure L2 traffic flows mentioned above.

- **L3:Inter-VLAN:** these are pure routed flows that leverage the dedicated L3 connections established between the aggregation and core layre devices. Most of the failire scenarios affecting the MC-LAG connection should not affect these flows.
- **Multicast L2:** Intra-VLAN: these are 100 multicast flows (one in each unique VLAN) characterized by having 1 source and 1 receiver in the same VLAN in separate data center site.

Note L3 multicast was not the main focus so convergence numbers are not specifically referenced. However, it was validated that L3 multicast stream always recovered in all failure scenarios.

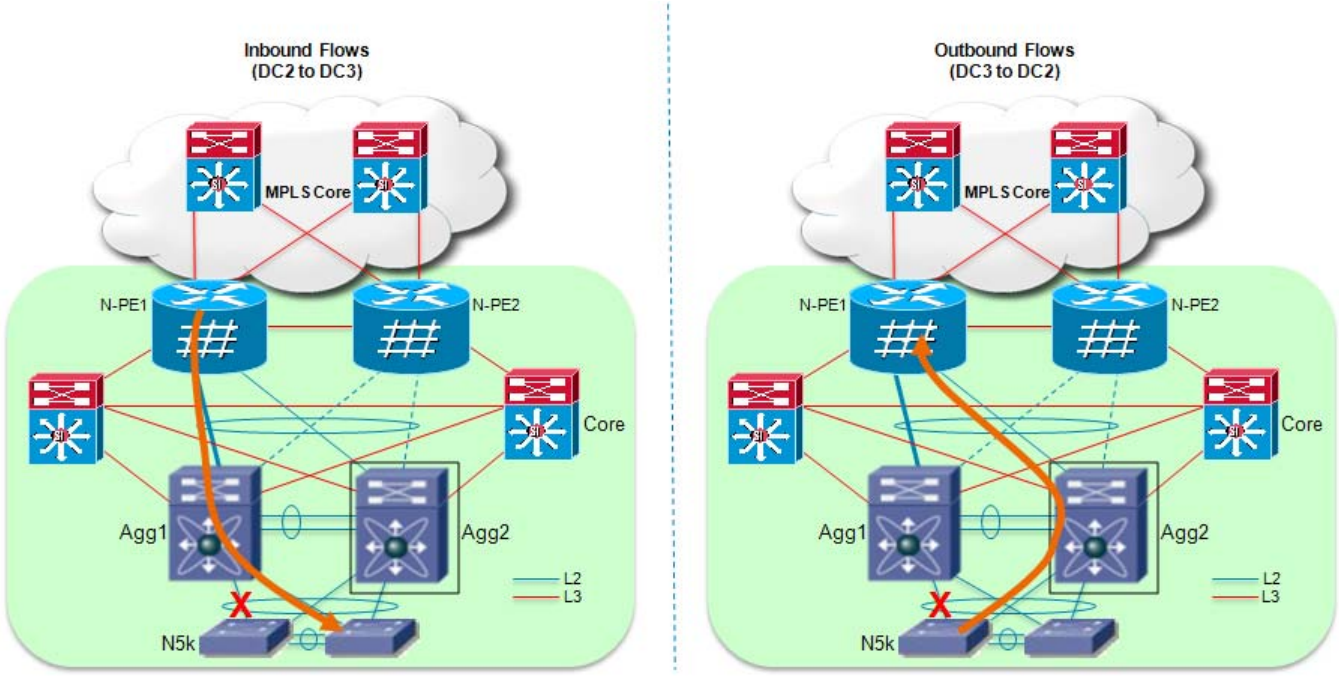
1200 VLANs

- **L2:Intra-VLAN-100-349:** these are pure L2 traffic flows (250 VLANs) carried between a specific access layer device and the Nexus 7000 in aggregation. These flows are then combined with the ones described below on the MC-LAG connection between Nexus 7000 switches and PE routers (in order to achieve a total of 1000 VLANs on that connection).
- **L2:Intra-VLAN-350-999-1100-1199:** these are pure L2 traffic flows (750 VLANs) carried between another access layer device and the aggregation switches. As mentioned above, these flows are then carried on the main MC-Lag connection under test toward the PE routers.
- **L2:Intra-VLAN-1200-1399:** these are pure L2 traffic flows (200 VLANs) extended between DC2 and DC3 sites leveraging the second MC-LAG group. In most of the validated failure scenarios, traffic flowing across this connection should remain unaffected.
- **L3L2: Inter-VLAN:** these are traffic flows that are first routed at the aggregation layer between a source and a destination VLAN. The destination VLAN is then extended to the remote site, so the behavior is expected to be very similar to the pure L2 traffic flows mentioned above.
- **L3:Inter-VLAN:** these are pure routed flows that leverage the dedicated L3 connections established between the aggregation and core layre devices. Most of the failire scenarios affecting the MC-LAG connection should not affect these flows.

Test 1: Access to Aggregation Uplink Failure and Recovery

When the uplink between access and aggregation layer devices fails, the traffic recovering mechanism is the same independently from the direction of the traffic (Figure 5-3).

Figure 5-3 Failure/Recovery Test Cases



Inbound flows (originated in the remote site DC2 and destined to local DC3) received by the left aggregation device are switched via the remaining link connecting to the access layer.

Outbound flows (originated in the local DC3 site and destined to remote DC2) that were originally sent via the failed uplink need also to be shifted to the remaining available uplink connecting to the right aggregation switch. The access layer device is responsible for detecting the uplink failure event and performing etherchannel re-hashing. Traffic flows that were originally hashed to the right uplinks are unaffected by the failure.

Note The opposite behavior is required when the uplink is recovered.

The convergence results achieved when extending 500 VLANs are shown in Table 5-4.

Table 5-4 Test 1 results with 500 VLANs (Unicast and Multicast)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Access to aggregation uplink failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.099	0.124
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.099	0.124
		L3:Inter-Vlan	0	0
	no shut	L2:Intra-Vlan-100-349	1.453	1.381
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	1.453	1.381
		L3:Inter-Vlan	0	0

Table 5-4 Test 1 results with 500 VLANs (Unicast and Multicast)

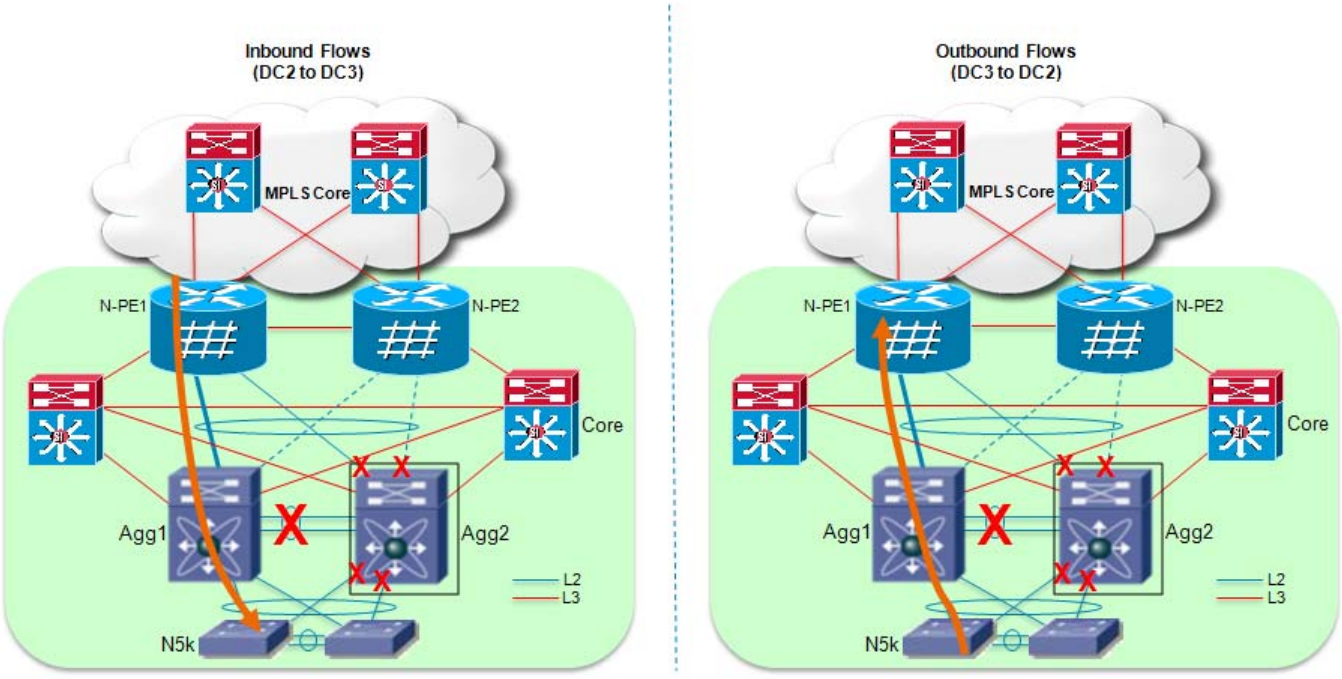
Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Access to aggregation uplink failure & recovery **Multicast Traffic**	shut	Multicast L2: Intra-Vlan	0.099	0.188
	no shut	Multicast L2:Intra-Vlan	1.453	1.423

Note Convergence results for 1200 VLANs are not available for this specific failure scenario because Nexus 5000 can support up to 512 VLANs in the software release that was validated as part of this solution.

Test 2: Complete vPC Peer-Link Failure and Recovery

As a consequence of the vPC peer-link failure, assuming that the two aggregation devices can still communicate via the peer-keepalive link, the device operating in "vPC secondary" role (the right aggregation switch in Figure 5-4 would bring down all the physical interface part of configured vPCs.

Figure 5-4 Complete vPC Peer-Link Failure Scenario



Inbound flows will be re-hashed by the ASR 9000 PE1 device on the remaining links connecting to the left Nexus 7000. Outbound traffic flows will be re-hashed by the access layer devices on the remaining uplinks (similarly to what discussed in Test 1).

When the peer-link connection is re-established, the vPC secondary device will re-enable the vPC physical links and traffic will start flow again on these connections.

The results shown in Table 5-5 and Table 5-6 highlight how also pure L3 traffic is impacted in this scenario. This is expected given that no physical path is available after the peer-link failure and re-rouing of inbound L3 traffic is required at the DC Core layer.

Table 5-5 Test 2 results with 500 VLANs (Unicast and Multicast)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Complete vPC peer link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.011	0.126
		L2:Intra-Vlan-1200-1449	0	0.271
		L3L2: Inter-Vlan	0.011	0.271
		L3:Inter-Vlan	1.767	0.215
	no shut	L2:Intra-Vlan-100-349	1.162	0.529
		L2:Intra-Vlan-1200-1449	0	0.534
		L3L2: Inter-Vlan	1.162	0.534
		L3:Inter-Vlan	2.065	0.535
Complete vPC peer link failure and recovery **Multicast Traffic**	shut	Multicast L2:Intra-Vlan	0.011	0.271
	no shut	Multicast L2:Intra-Vlan	1.160	0.534

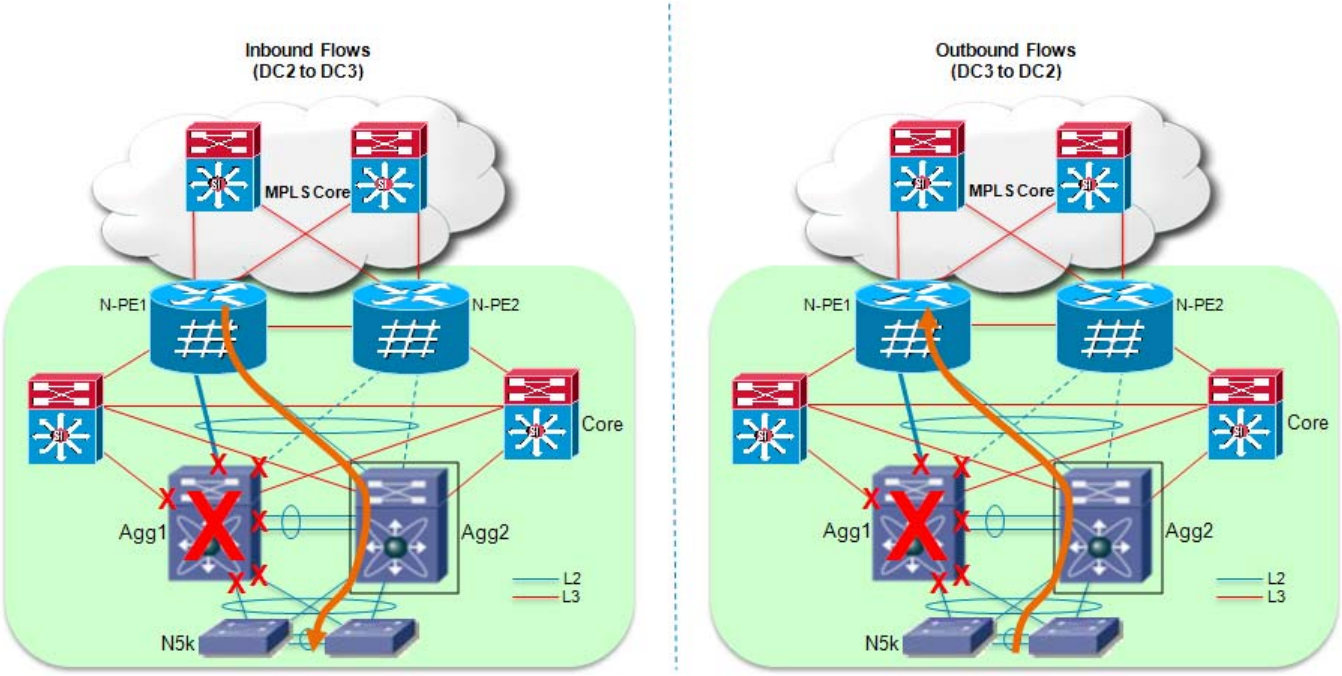
Table 5-6 Test 2 results with 1200 VLANs (Unicast Traffic)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Complete vPC peer link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0	0.141
		L2:Intra-Vlan-350-999-1100-1199	0	0.239
		L2:Intra-Vlan-1200-1399	0.010	0.262
		L3L2:Inter-Vlan	0.010	0.262
		L3:Inter-Vlan	5.636	0.228
	no shut	L2:Intra-Vlan-100-349	0	2.335
		L2:Intra-Vlan-350-999-1100-1199	0	2.063
		L2:Intra-Vlan-1200-1399	3.164	2.065
		L3L2:Inter-Vlan	3.163	2.335
		L3:Inter-Vlan	2.797	2.066

Test 3: Aggregation Device Failure and Recovery

When one of the two aggregation Nexus 7000 device fails, all the physical connections to and from that device are obviously torn down as well. Inbound and outbound traffic paths after the failure are shown in Figure 5-5.

Figure 5-5 Aggregation Device Failure Scenario



Inbound flows are re-hashed by the ASR 9000 PE device on the remaining active links connected to the right aggregation layer device (this recovery mechanism is similar to the one discuss for the vPC peer-link failure scenario). Outbound flows are instead re-hashed by the access layer switches, similarly to how noticed for the access to aggregation uplink failure test case.

The physical failure of the aggregation device would obviously impact also the pure L3 routed flows, since the dedicated L3 connections will also fail.

Table 5-7 **Test 3 results with 500 VLANs (Unicast and Multicast)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Aggregation device failure and recovery **Unicast Traffic**	reload	L2:Intra-Vlan-100-349	0.018	0.123
		L2:Intra-Vlan-1200-1449	0.013	0.228
		L3L2: Inter-Vlan	0.018	0.228
		L3:Inter-Vlan	0.048	0.228
	restore	L2:Intra-Vlan-100-349	0.029	1.837
		L2:Intra-Vlan-1200-1449	1.894	0.751
		L3L2: Inter-Vlan	1.893	1.837
		L3:Inter-Vlan	1.284	0.749
Aggregation device failure and recovery **Multicast Traffic**	reload	Multicast L2: Intra-Vlan	0.018	0.228
	restore	Multicast L2:Intra-Vlan	1.891	1.881

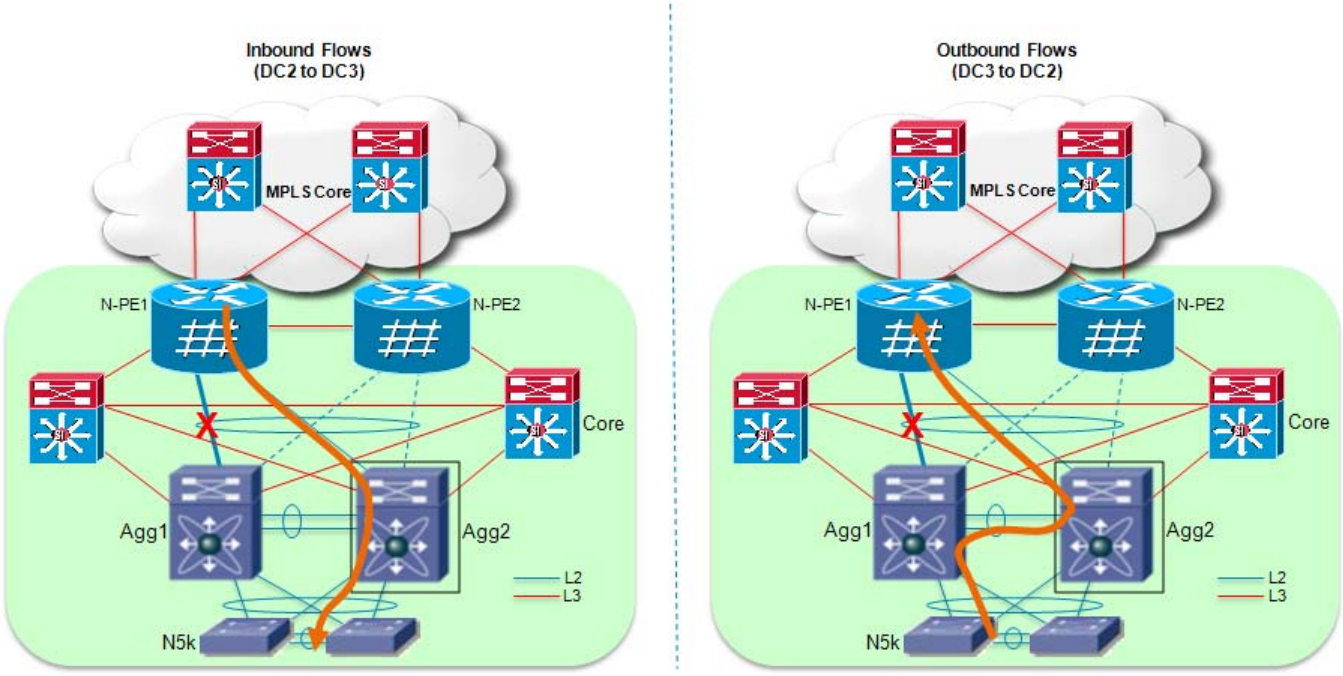
Table 5-8 **Test 3 results with 1200 VLANs (Unicast Traffic)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Aggregation device failure and recovery **Unicast Traffic**	reload	L2:Intra-Vlan-100-349	0.017	0.125
		L2:Intra-Vlan-350-999-1100-1199	0	0.250
		L2:Intra-Vlan-1200-1399	0	0.250
		L3L2:Inter-Vlan	0.017	0.250
		L3:Inter-Vlan	0.059	0.250
	restore	L2:Intra-Vlan-100-349	0.023	3.986
		L2:Intra-Vlan-350-999-1100-1199	0	0
		L2:Intra-Vlan-1200-1399	3.245	3.247
		L3L2:Inter-Vlan	0.023	3.985
		L3:Inter-Vlan	4.625	4.606

Test 4: Aggregation to PE Active Link Failure and Recovery

The failure of an active L2 link between aggregation and PE device would cause the same recovery mechanism for inbound traffic flows already discussed in the previous failure scenario, where the ASR 9000 PE router is responsible to re-hash the flows on the remaining active link connecting to the second aggregation switch (Figure 5-6).

Figure 5-6 *Aggregation Device Failure Scenario*



Outbound flows that were originally sent on the failed link need instead to be switched via the peer-link connecting the aggregation devices. Once again, this should be taken into consideration when designing the bandwidth to be dedicated to this connection.

Table 5-9 **Test 4 results with 500 VLANs (Unicast and Multicast)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Aggregation to PE active link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0	0.658
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0	0.658
		L3:Inter-Vlan	0	0
	no shut	L2:Intra-Vlan-100-349	0	0.553
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0	0.553
		L3:Inter-Vlan	0	0
Aggregation to PE active link failure and recovery **Multicast Traffic**	shut	Multicast L2:Intra-Vlan	0.011	0.657
	no shut	Multicast L2:Intra-Vlan	0.167	0.553

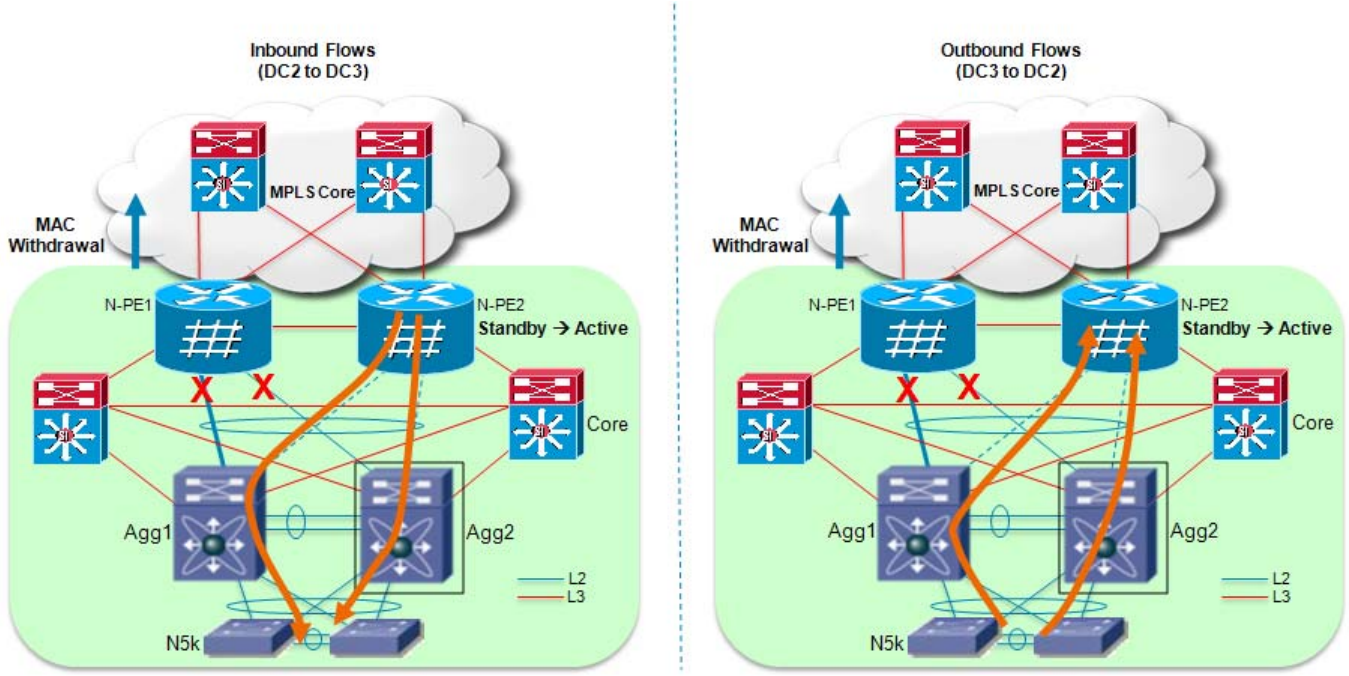
Table 5-10 *Test 4 results with 1200 VLANs (Unicast Traffic)*

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Aggregation to PE active link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.011	0.849
		L2:Intra-Vlan-350-999-1100-1199	0.011	0.849
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	0.010	0.849
		L3:Inter-Vlan	0	0
	no shut	L2:Intra-Vlan-100-349	1.223	2.086
		L2:Intra-Vlan-350-999-1100-1199	1.231	2.094
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	1.223	2.086
		L3:Inter-Vlan	0	0

Test 5: PE Dual Active Links to Aggregation Failure and Recovery

The contemporary failure of both links connecting the active PE to the aggregation layer devices is the first scenario where the standby PE needs to become active to allow for L2 traffic flows recovery. In order for that to happen, the initially active PE communicate via ICCP to its peer that it losts both active links, forcing the standby router to become active (Figure 5-7).

Figure 5-7 PE Dual Active Links to Aggregation Failure Scenario



Inbound traffic recovery is mainly dictated by two mechanisms:

- Time required for the standby PE to transition to an active role. This would cause (through LACP negotiation) moving the interfaces connecting the aggregation devices to this PE from an Hot-Standby to an Active state.
- Time required for the remote PE devices to flush their MAC address tables (as a result of a MAC notification originated by the local PE) to ensure that traffic can now be flooded and can reach the newly activated PE router. Flooding will stop once bidirectional communication is established between sites, and the remote PE routers correcty populate the information in their MAC address tables.

Outbound traffic outage is mostly dictated by the first item discussed above (i.e. the time required to transiction hot-standby interfaces to an active role).

Note To minimize the occurrence of this failure scenario, Cisco recommends spreading the links connecting each PE router to the aggregation switched on different linecards.

Table 5-11 Test 5 results with 500 VLANs (Unicast and Multicast)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
PE dual active links to aggregation failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.378	0.374
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.378	0.374
		L3:Inter-Vlan	0	0
	no shut	L2:Intra-Vlan-100-349	0.592	0.322
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.591	0.320
		L3:Inter-Vlan	0	0
PE dual active links to aggregation failure and recovery **Multicast Traffic**	shut	Multicast L2:Intra-Vlan	0.210	0.443
	no shut	Multicast L2:Intra-Vlan	3.269	0.193

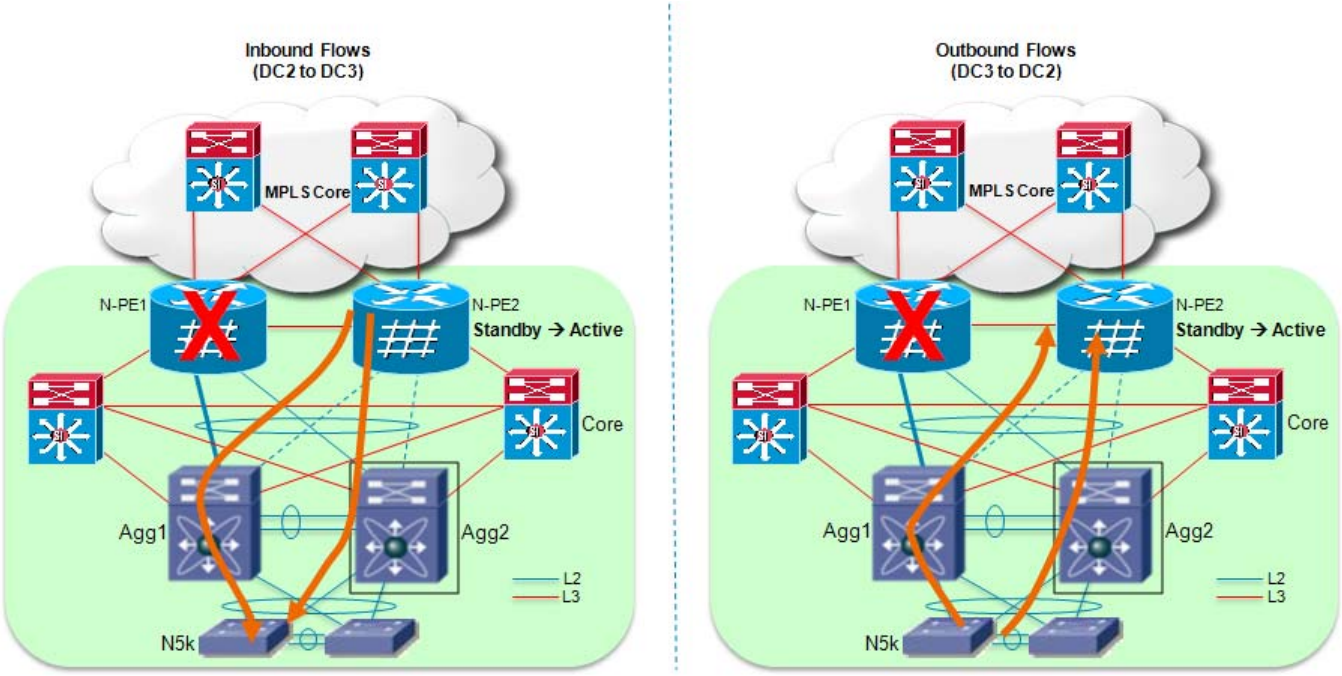
Table 5-12 Test 5 results with 1200 VLANs (Unicast Traffic)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
PE dual active links to aggregation failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.693	0.668
		L2:Intra-Vlan-350-999-1100-1199	0.693	0.668
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	0.693	0.668
		L3:Inter-Vlan	0	0
	no shut	L2:Intra-Vlan-100-349	1.994	1.145
		L2:Intra-Vlan-350-999-1100-1199	2.877	2.303
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	1.992	1.142
		L3:Inter-Vlan	0	0

Test 6: Active PE Router Failure and Recovery

The failure of the active PE router is a second scenario that will force the standby PE to transition to active state.

Figure 5-8 Active PE Router Failure Scenario



The recovery for inbound traffic is similar to what discussed in the previous test case. The main difference is that the remote PE routers will start flooding traffic directed to DC3 not because of the reception of a MAC withdrawal notification, but as a consequence of the fact the PWs connected to the failed router are brought down. The end result is the same, with traffic being received by the newly activated PE router, which will then send it toward the aggregation layer switches as soon as the physical links are transitioned fron Hot-Standby to Active state.

Outbound traffic recovery once again is dependant on the activation of the standby links only.

Table 5-13 **Test 6 results with 500 VLANs (Unicast and Multicast)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE router failure and recovery **Unicast Traffic**	reload	L2:Intra-Vlan-100-349	0.973	0.727
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.972	0.726
		L3:Inter-Vlan	0	0
	restore	L2:Intra-Vlan-100-349	0.598	0.327
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.598	0.324
		L3:Inter-Vlan	0.098	0
Active PE router failure and recovery **Multicast Traffic**	reload	Multicast L2: Intra-Vlan	0	0.075
	restore	Multicast L2:Intra-Vlan	4.986	0.209

Table 5-14 **Test 6 results with 1200 VLANs (Unicast Traffic)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE router failure and recovery **Unicast Traffic**	reload	L2:Intra-Vlan-100-349	2.756	2.068
		L2:Intra-Vlan-350-999-1100-1199	2.985	2.518
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	2.755	2.068
		L3:Inter-Vlan	0	0
	restore	L2:Intra-Vlan-100-349	1.667	1.039
		L2:Intra-Vlan-350-999-1100-1199	2.816	2.334
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	1.666	1.937
		L3:Inter-Vlan	0	0

Test 7: Active PE Router Core Link Failure and Recovery

The failure of one of the routed link connecting a PE router to the MPLS core can be recovered by routing the VPLS traffic on the alternate L3 path available.

For inbound flows, the re-routing happens in the MPLS core (since traffic is destined to the PE loopback interface used to establish the LDP session with the remote PEs).

For outbound flows, the re-routing happens locally on the PE affected by the link failure, since traffic is destined to the loopback interface of a remote PE device.

In both cases, only the flows that were sent via the failed link are affected, whereas the traffic originally routed via the remaining L3 uplink continues to flow undisturbed.

Figure 5-9 Active PE Router Core Link Failure Scenario

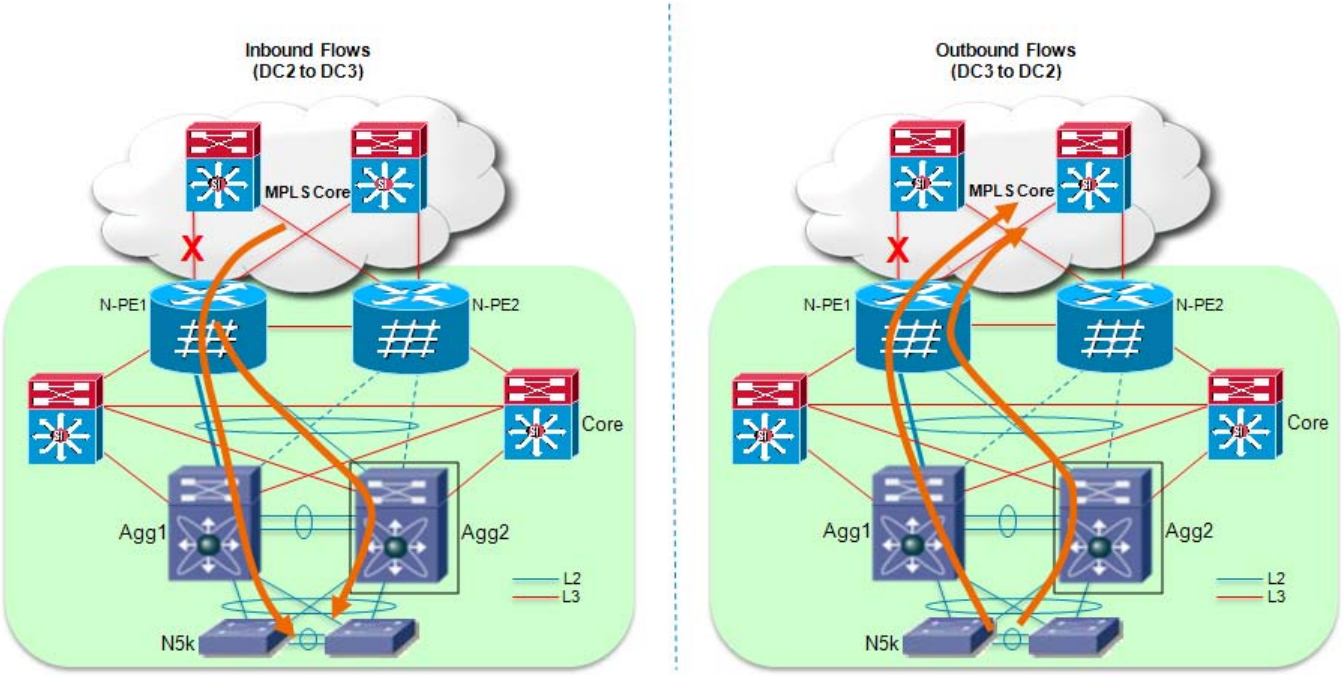


Table 5-15 **Test 7 results with 500 VLANs (Unicast and Multicast)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE router core link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.241	0.246
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.241	0.246
		L3:Inter-Vlan	0.141	0.299
	no shut	L2:Intra-Vlan-100-349	0	0
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0	0
		L3:Inter-Vlan	0	0
Active PE router core link failure and recovery **Multicast Traffic**	shut	Multicast L2:Intra-Vlan	0.241	1.613
	no shut	Multicast L2:Intra-Vlan	0	1.542

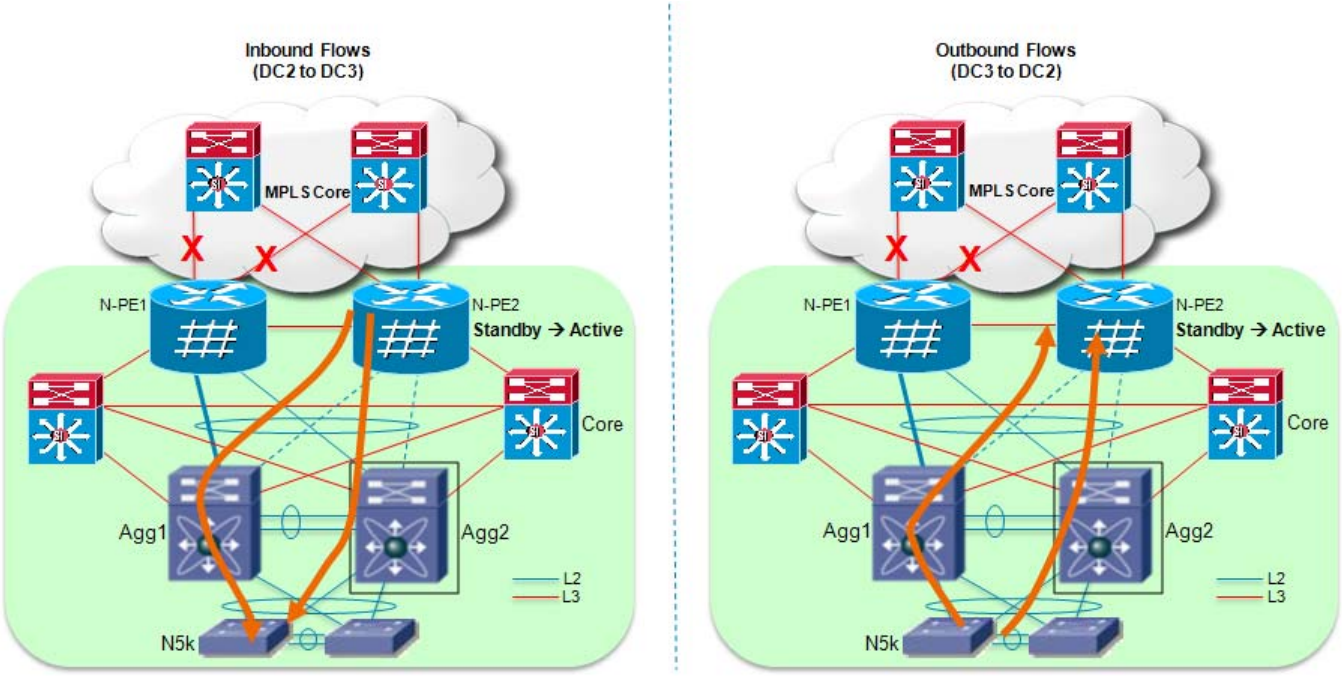
Table 5-16 Test 7 results with 1200 VLANs (Unicast Traffic)

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE router core link failure and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.243	0.238
		L2:Intra-Vlan-350-999-1100-1199	0.243	0.238
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	0.243	0.238
		L3:Inter-Vlan	0.132	0.297
	no shut	L2:Intra-Vlan-100-349	0	0
		L2:Intra-Vlan-350-999-1100-1199	0	0
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	0	0
		L3:Inter-Vlan	0	0

Test 8: Active PE Core Isolation and Recovery

If the active PE loses both direct connections to the MPLS core, a “core isolation” recovery is triggered forcing the other PE router to transition to the Active state. The behavior is the result of a specific configuration applied on the active PE router when defining an MC-LAG group: the interfaces used to connect to the core are explicitely defined as “backbone” interfaces. When they both fail, the PE router leverages ICCP to communicate the event to its peer, which will get activated.

Figure 5-10 Active PE Core Isolation Scenario



Similarly to how discuss Test 6 (PE failure scenarios), traffic recovery in both inbound and outbound directions is mainly dependant on how fast the hot-standby links can be activated.

Table 5-17 **Test 8 results with 500 VLANs (Unicast and Multicast)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE core isolation and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	0.602	0.598
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.602	0.598
		L3:Inter-Vlan	0.132	0.297
	no shut	L2:Intra-Vlan-100-349	0.534	0.349
		L2:Intra-Vlan-1200-1449	0	0
		L3L2: Inter-Vlan	0.534	0.349
		L3:Inter-Vlan	0	0.001
Active PE core isolation and recovery **Multicast Traffic**	shut	Multicast L2: Intra-Vlan	0.210	0.443
	no shut	Multicast L2:Intra-Vlan	3.269	0.193

Table 5-18 **Test 8 results with 1200 VLANs (Unicast Traffic)**

Failure Type	Action	Flows	DC2→DC3	DC3→DC2
Active PE core isolation and recovery **Unicast Traffic**	shut	L2:Intra-Vlan-100-349	1.936	1.602
		L2:Intra-Vlan-350-999-1100-1199	2.800	2.680
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	1.935	1.600
		L3:Inter-Vlan	0.133	0.282
	no shut	L2:Intra-Vlan-100-349	2.461	1.297
		L2:Intra-Vlan-350-999-1100-1199	2.852	2.324
		L2:Intra-Vlan-1200-1399	0	0
		L3L2:Inter-Vlan	2.460	1.284
		L3:Inter-Vlan	0	0

Deployment Recommendations

This chapter addresses issues that you should consider when deploying MC-LAG based VPLS solution to interconnect data centers.

1. **Improve multicast convergence number using "mrouter" command**

By default on ASR 9000, ICCP process running between the two POAs only synchronizes IGMP entries from the access side within a data center to mLACP standby peer. Also, mrouter ports are dynamically learned via multicast protocols. This adds delay under failure conditions during mLACP switchover.

Network convergence time for multicast traffic can be improved by configuring "mrouter" on ASR 9000. This command does the following:

- Statically configures an interface as mrouter port which otherwise has to rely on PIM or IGMP for dynamic learning
- Allows synchronization of IGMP entries from the MPLS core to N-PE in mLACP standby mode

This recommendation requires ASR 9000 as the edge router on all data centers.

Below is the relevant configuration. Configuring mrouter is a two-step process:

a. Create an IGMP profile

b. Configure mrouter under this profile. The second step is to apply this IGMP profile to all neighbors under VFIs terminating in remote datacenters.

Note Do not apply this profile to the pseudowire between the NPE's within the same data center.

mrouter configuration on ASR 9000 router

```
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1#conf t
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1(config)#igmp snooping profile igmp-mrouters
RP/0/RSP0/CPU0:DC3-ASR9(config-igmp-snooping-profile)#mrouter
RP/0/RSP0/CPU0:DC3-ASR9(config-igmp-snooping-profile)#commit
```

```
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1(config)#l2vpn bridge group group1
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1(config-l2vpn-bg)#bridge-domain vlan107
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1(config-l2vpn-bg-bd)#vfi vfi107
RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1(config-l2vpn-bg-bd-vfi)#neighbor 150.3.3.6 pw-id 107
RP/0/RSP0/CPU0:DC3-ASR 9000-N(config-l2vpn-bg-bd-vfi-pw)#igmp snooping profile
igmp-mrouters
RP/0/RSP0/CPU0:DC3-ASR 9000-N(config-l2vpn-bg-bd-vfi-pw)#commit

RP/0/RSP0/CPU0:DC3-ASR 9000-NPE1#sh run l2vpn bridge group group1 bridge-domain vlan107
Thu Mar 10 12:35:21.624 PST
l2vpn
 bridge group group1
  bridge-domain vlan107
   igmp snooping profile igmp-snoop
   interface Bundle-Ether31.107
   !
   vfi vfi107
    neighbor 150.2.2.5 pw-id 107
     pw-class vpls-pw-class
     igmp snooping profile igmp-mrouters
    !
    neighbor 150.2.2.6 pw-id 107
     pw-class vpls-pw-class
     igmp snooping profile igmp-mrouters
    !
# Neighbor 150.3.3.6 is the N-PE (ASR9000) within the same data center. Hence IGMP mrouter
profile is not configured #

    neighbor 150.3.3.6 pw-id 107
     pw-class vpls-pw-class
     !
    neighbor 150.11.11.5 pw-id 107
     pw-class vpls-pw-class
     igmp snooping profile igmp-mrouters
    !
    neighbor 150.11.11.6 pw-id 107
     pw-class vpls-pw-class
     igmp snooping profile igmp-mrouters
    !
```

As shown in Figure 4-2, Cisco 7600 routers deployed as PEs in DC2 do not support syncing of IGMP entries either from access or from core. Due to this, IGMP entries have to be relearned during mLACP switchover. In this scenario, network convergence for multicast traffic depends on the rate at which mrouter ports are dynamically learned which is a factor of PIM and IGMP timers.

Note This issue is going to be fixed in the upcoming 7600 router 15.2(1)S release planned for July.

2. Avoid IGMP packets looping under specific mLACP failure conditions

In case of deploying multiple aggregation blocks connected to the pair of PE devices and extending the same set of VLANs, a local pseudowire must be established between the PEs and this may induce IGMP packets looping. To avoid this problem, the recommendation is to configure the “router-guard” command under IGMP profile and assign that IGMP profile to the pseudowire between the NPE’s within the same datacenter.

As shown below, configuring router-guard is a two-step process. First an IGMP profile has to be created, and the router-guard command should be configured under this profile. The second step is to apply this IGMP profile to all neighbors under VFIs between the NPE’s within the same datacenter.

router-guard configuration on ASR 9000 router

```
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#conf t
Mon Mar 14 15:14:01.865 PST
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1(config)#igmp snooping profile router-guard
RP/0/RSP0/CPU0:DC3-ASR9(config-igmp-snooping-profile)#router-guard
RP/0/RSP0/CPU0:DC3-ASR9(config-igmp-snooping-profile)#commit

RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#conf t
Mon Mar 14 15:14:23.503 PST
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1(config)#l2vpn bridge group group1 bridge-domain vlan111
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1(config-l2vpn-bg-bd)#vfi vfi111
RP/0/RSP0/CPU0:DC3-ASR9K-NPE1(config-l2vpn-bg-bd-vfi)#neighbor 150.3.3.6 pw-id 111
RP/0/RSP0/CPU0:DC3-ASR9K-N(config-l2vpn-bg-bd-vfi-pw)#igmp snooping profile router-guard
RP/0/RSP0/CPU0:DC3-ASR9K-N(config-l2vpn-bg-bd-vfi-pw)#commit

RP/0/RSP0/CPU0:DC3-ASR9K-NPE1#sh run l2vpn bridge group group1 bridge-domain v$
Mon Mar 14 15:22:52.099 PST
l2vpn
 bridge group group1
  bridge-domain vlan111
   igmp snooping profile igmp-snoop
   interface Bundle-Ether31.111
   !
   vfi vfi111
    neighbor 150.2.2.5 pw-id 111
     pw-class vpls-pw-class
    !
    neighbor 150.2.2.6 pw-id 111
     pw-class vpls-pw-class
    !

# Neighbor 150.3.3.6 is the N-PE (ASR9000) within the same data center. Hence IGMP profile
with router-guard is configured #

    neighbor 150.3.3.6 pw-id 111
     pw-class vpls-pw-class
     igmp snooping profile router-guard
    !
    neighbor 150.11.11.5 pw-id 111
     pw-class vpls-pw-class
    !
    neighbor 150.11.11.6 pw-id 111
     pw-class vpls-pw-class
```

Summary

Globalization, security and disaster recovery considerations are driving divergence in business locations across multiple regions. In addition, organizations are looking to distribute workload between computers, share network resources effectively and increase the availability of applications.

As data centers grow in size and complexity, enterprises are adopting server virtualization technologies to achieve increase efficiency and use of resources. Due to the exponential growth, most of these customers are looking at interconnecting more data centers, extending large number of VLANs with high layer 2 traffic capabilities between these data centers.

This design guide describes the deployment of the MC-LAG to VPLS technology on Cisco ASR 9000 routers. While active/standby by nature, an option to provide active/active connectivity was also discussed thus providing greater redundancy and VLAN load sharing between the two POA devices. With the deployment of vPC on Nexus 7000 switches in the aggregation, this solution provides link and chassis level redundancy and faster convergence during link and node failures. In addition, the solution is also fully compatible with the use of VSS technology on Cisco Catalyst 6500s in the aggregation layer.

In summary, the MC-LAG based VPLS solution documented in this design guide provides a high-speed, low latency network with STP isolation between data centers. The solution is extremely flexible and highly scalable and offers key features required for large-scale data center interconnectivity.