Cisco Data Center Interconnect Design and Deployment Guide, System Release 2.0

Last Updated: November 19, 2010

Cisco Validated

Cisco Data Center Interconnect Design and Deployment Guide, System Release 2.0

CONTENTS

Preface iii

Audience and Doc Suite iv

Motivation v

Document Version v

CHAPTER 1 **Cisco DCI Design Architecture** 1-1

Data Center Network Best Practices 1-3

DC Interconnect Layer 1-5

LAN Extension Solution Requirements 1-8

Spanning Tree Protocol Isolation 1-8

End-to-End Loop Prevention 1-10

High Availability (HA) 1-11

Multi-Path Load Balancing 1-12

Routing Optimization 1-12

Storage Access 1-14

Shared Storage 1-15

Active-Active Storage 1-16

Security and Encryption 1-18

Securing and Hardening Network Infrastructure 1-19

Storm Control and Flooding 1-19

Protecting the Control Plane 1-20

Encryption 1-20

Hierarchical Quality of Service (HQoS) 1-22

DCI Phase 2.0 Networking Technology 1-23

Site-to-Site Deployments 1-23

EoMPLS Port Mode 1-23

EoMPLSoGRE Port Mode 1-29

Multi-Site Deployments 1-32

VPLS 1-32

CHAPTER 2 **Cisco DCI System Solution Deployment** 2-1

Point-to-Point Topologies 2-2

Deploying Port Based EoMPLS 2-3

EoMPLS Configuration 2-5

End-to-End Loop Prevention and STP Isolation 2-8
First Hop Redundancy Protocol (FHRP) Deployment 2-12
Encryption 2-21
Inter Data Centers Routing Considerations 2-24
EoMPLS Failure/Recovery Analysis 2-25
Deploying Port Based EoMPLS over an IP Core 2-38
EoMPLSoGRE Configuration 2-40
IPSec-Based Encryption 2-41
MTU Considerations 2-42
EoMPLSoGRE Failure/Recovery Analysis 2-43
H-QoS Considerations 2-46
Deploying H-QoS with ASR1000 2-47
IPSec and H-QoS Specific Considerations 2-49
Multipoint Topologies 2-51
Deploying VPLS 2-52
VPLS Basic Configuration 2-54
STP Isolation and End-to-End Loop Prevention 2-57
EEM Related Tracking Objects 2-67
First Hop Redundancy Protocol (FHRP) Deployment 2-76
Inter Data Centers Routing Considerations 2-77
VPLS Failure/Recovery Scenarios 2-80
Summary of Design Recommendations 2-96
Point-to-Point Deployment Recommendations 2-96
Multipoint Deployment Recommendations 2-96
Summary 2-98

Preface

This Data Center Interconnect (DCI) Design and Deployment Guide describes a portion of Cisco's system for interconnecting multiple data centers for Layer 2-based business applications.

Given the increase in data center expansion, complexity, and business needs, DCI has evolved to support the following requirements:

- Business Continuity
- Clustering
- Virtualization
- Load Balancing
- Disaster Recovery

There is a strong need to expand the application domain beyond a single data center. DCI is driven by the business requirements shown in Table i-1.

Table i-1 DCI Business Drivers

Business	IT Solutions
Disaster Prevention	Active/Standby Migration
Business Continuance	Server HA clusters, "Geoclustering"
Workload Mobility	Move, consolidate servers, "VMotion"

Several Applications are available for IT solutions to solve these business requirements.

HA Clusters/Geoclusters

- Microsoft MSCS
- Veritas Cluster Server (Local)
- Solaris Sun Cluster Enterprise
- VMware Cluster (Local)
- VMware VMotion
- Oracle Real Application Cluster (RAC)

- IBM HACMP
- EMS/Legato Automated Availability Manager
- NetApp Metro Cluster
- HP Metrocluster

Active/Standby Migration, Move/Consolidate Servers

- VMware Site Recovery Manager (SRM)
- Microsoft Server 2008 Layer 3 Clustering
- VMware VMotion

The applications above drive the business and operation requirement for extending the Layer 2 domain across geographically dispersed data centers. Extending Layer 2 domains across data centers present challenges including, but not limited to:

- Spanning tree isolation across data centers
- Achieving high availability
- Full utilization of cross sectional bandwidth across the Layer 2 domain
- Network loop avoidance, given redundant links and devices without spanning tree

Additional customer demands, such as Quality-of-Service (QoS) and encryption, may be required on an as needed basis.

Cisco's DCI solution satisfies business demands, while avoiding challenges, by providing a baseline for the additional enhancements cited above.

Cisco's DCI solution ensures Cisco's leadership role in the Enterprise/Data Center marketplace, securing their position as the primary competitive innovator.

Audience and Doc Suite

This documentation suite provides detailed data center design architecture, deployment scenarios, and interconnect convergence test descriptions, procedures, results, anomalies and configurations to support the DCI system solution design and implementation recommendations. Verified test results and configurations assist Network Architects, Network Engineers, and Systems Engineers in understanding various Cisco solution recommendations extending geographically Layer 2 networks over multiple distant data centers, while addressing high performance and fast convergence requirements across long distances.

The Data Center Interconnect documentation suite consists of the following:

- Design and Deployment Guide
- Test Verification
- Test Configurations

Motivation

Cisco recommends isolating and reducing Layer 2 networks to their smallest diameter, limited to the access layer. Server-to-server communication, High Availability clusters, networking, and security all require Layer 2 connectivity. In many instances, Layer 2 functionality must extend beyond a single data center, particularly when a campus framework extends beyond its original geography, spanning multiple long distance data centers. This is more pervasive as high-speed service provider connectivity becomes more available and cost effective.

High-Availability clusters, server migration, and application mobility warrant Layer 2 extensions across data centers. To simplify data center deployment, this system level design and configuration guide provides the appropriate governance in configuring, provisioning and scaling DCI by:

- Enabling data center expansion with a Cisco approved solution
- Building new capabilities on top of an existing deployment base, extending the Catalyst 6500 footprint, and positioning high density platforms such as Nexus 7000 series switch
- Extending existing operational capabilities

Document Version

Document Version 2.0 published March 7, 2011.

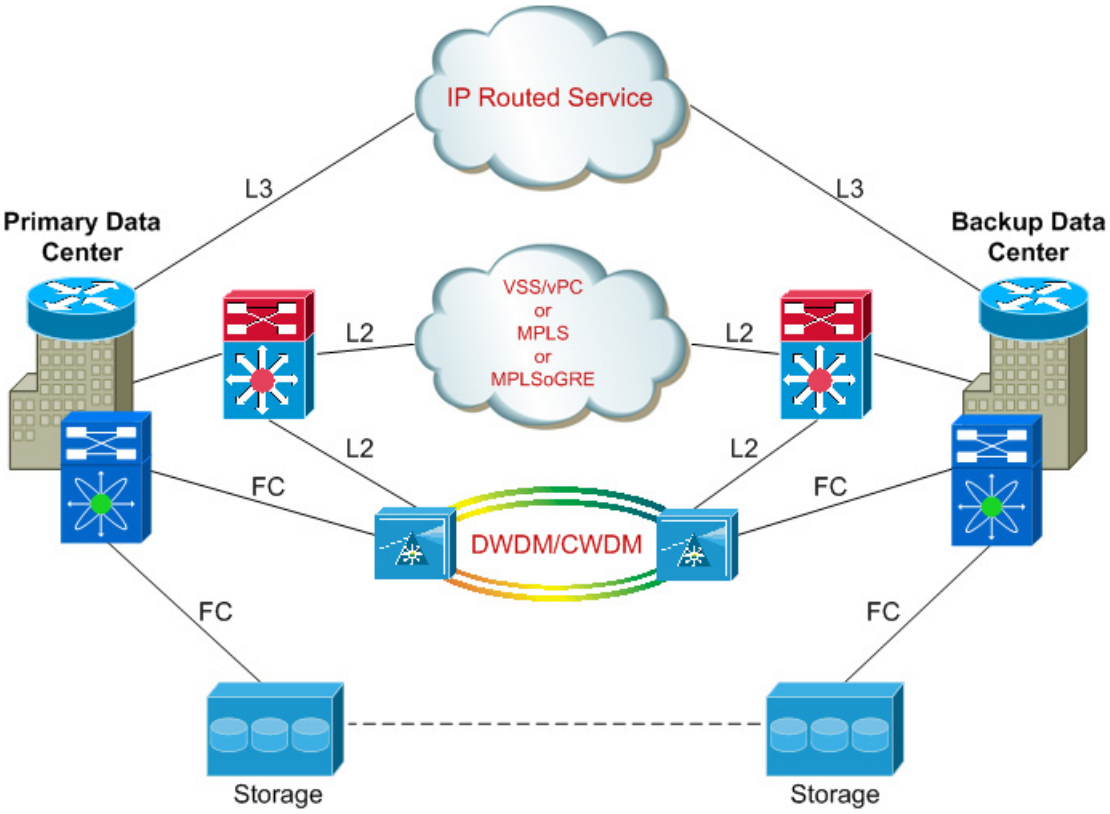
CHAPTER 1

Cisco DCI Design Architecture

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 1-1 summarizes the three general types of connectivity required for a DCI solution:

Figure 1-1 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:** Provide 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.

The focus of this document is on LAN extension solutions (the term DCI is often improperly used as a synonym of this type of connectivity only) but it is important to keep in mind that there are companion solutions in the routing and storage spaces. Layer 3 and SAN extension considerations are integrated wherever they are relevant to provide a complete DCI architecture.

LAN extension solutions (from now on referred to with the generic term "DCI") are commonly used to extend subnets beyond the traditional Layer 3 boundaries of a single site data center. Stretching the network space across two or more data centers can accomplish many things. Many clustering applications, both commercial and those developed in-house, require Layer 2 connectivity between cluster nodes. Putting these nodes in separate data centers can help build resilience into a system. If one data center is lost, the backup resources in the second data center can take over services with a much smaller impact, or a shorter downtime. DCI 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.

DCI also presents a challenge: providing these LAN extension capabilities is going to 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: protocol 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 provider router know where an active IP address or subnet exists at any given time? You should also know which products in the Cisco portfolio strategically work together in support of the DCI solution.

When designing a DCI architecture, you should consider two additional factors:

- Data center sites that a customer maintains and interconnects are categorized in a DCI solution as point-to-point (P2P) or multi-point:
  - **P2P:** Refers to two data center sites connected to each other, deployed in an active/backup or an active/active state.
  - **Multi-point:** Refers to data centers with more than two sites, which may be deployed with all sites active, two sites active and one backup, or in a hierarchical data center design covering different geographical regions.
- Different transport alternatives that can be leveraged to interconnect various data center sites typically include the following:
  - **Dark Fiber (Optical):** This can be considered a Layer 1 type of service. It is popular among many customers today as it allows the transport of various types of traffic, including SAN traffic. It tends to be expensive, especially as the number of sites increases. Dark fiber offerings are also limited in the distance they can span.
  - **MPLS:** This option can usually be deployed when the enterprise owns the infrastructure interconnecting the various data center sites and enables MPLS services on it. This is often the case with large enterprise deployments. For small and medium enterprises, connectivity services are usually offered by one (or more) service providers (SPs): assuming these services are Layer 1 or Layer 2, the enterprise can overlay an MPLS based Layer 2 VPN solution on the SP service, giving the enterprise additional operational flexibility.

  - **IP:** Some SPs offer Layer 3 connectivity services to the enterprise. Independently from how the connectivity is provided inside the SP cloud (through the deployment of IP based or MPLS based technologies) this essentially means that the enterprise edge devices establish a L3 peering with the SP devices. Therefore, the enterprise should deploy an overlay technology to perform the LAN extension between the various sites. The enterprise's choice of overlay solutions tends to be limited to those based on IP, except for the extremely rare instance in which the SP is willing to transport and relay MPLS labels on behalf of the enterprise.

The focus of this DCI System Release 2.0 design guide is on the following specific technical alternatives to provide LAN extension functionality:

- Point-to-point interconnection using Ethernet over Multiprotocol Label Switching (EoMPLS) natively (over an optical or MPLS enabled core) and over a Layer 3 IP core (EoMPLSoGRE)
- Point-to-multipoint interconnections using virtual private LAN services (VPLS) natively (over an optical or MPLS enabled core)

Note The DCI System Release 1.0 design guide focused on Dark Fiber/DWDM-based VSS and vPC deployment options. More information can be found at the following link: http://wwwin.cisco.com/marketing/datacenter/solutions/launches/dci/index.shtml

This chapter contains the following sections:

- Data Center Network Best Practices, page 1-3
- LAN Extension Solution Requirements, page 1-8
- DCI Phase 2.0 Networking Technology, page 1-23

Data Center Network Best Practices

The DCI architecture suggested in this document is deployed following the current best practices for building data center networks.

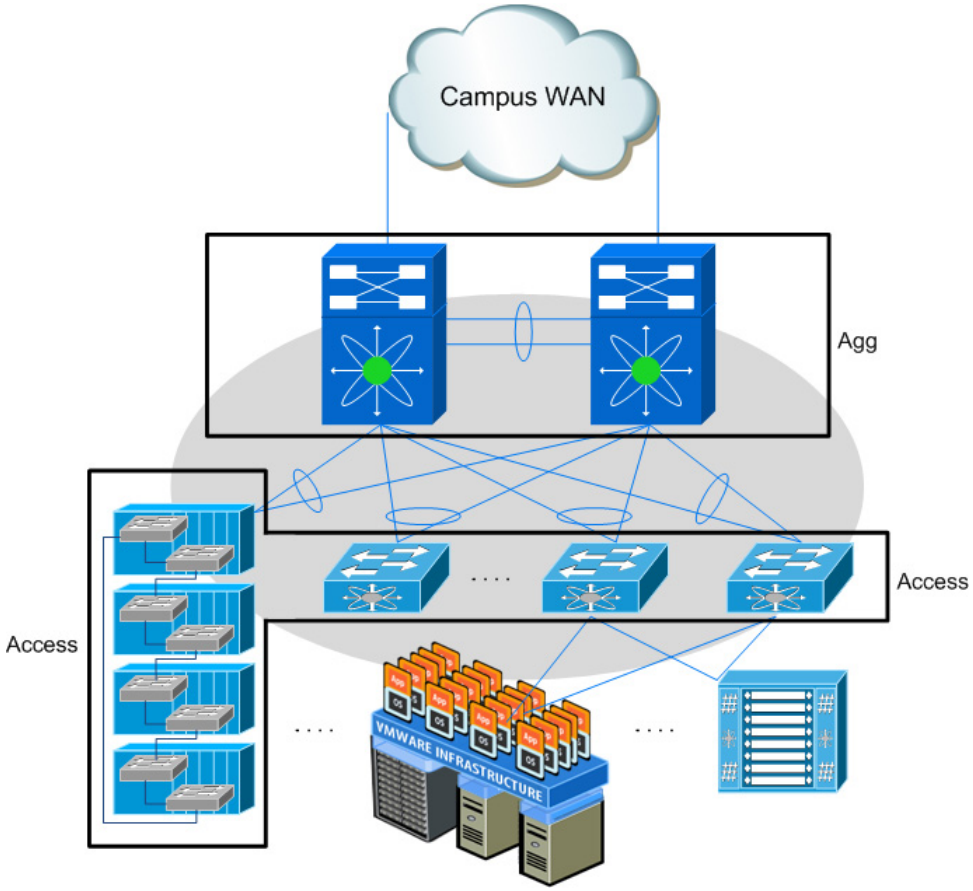
Hierarchical network design has been commonly used in enterprise networking for many years. This model uses redundant switches at each layer of the network topology for device-level failover that creates a highly available transport between end nodes using the network. Data center networks often require additional services beyond basic packet forwarding, such as server load balancing, firewall, or intrusion prevention. These services might be introduced as modules populating a slot of switching nodes in the network, or as standalone appliance devices. Each service approach supports the deployment of redundant hardware to preserve the high availability standards set by the network topology.

A structured data center environment uses a physical layout that corresponds closely with the network topology hierarchy. Decisions on cabling types, and the placement of patch panels and physical aggregation points must match interface types and densities of physical switches being deployed. In a new data center buildout, both can be designed simultaneously, taking into consideration power and cooling resource constraints. Investment concerns regarding the selection of switching platforms in existing data center facilities are strongly influenced by physical environment costs related to cabling, power, and cooling. A flexible networking requirement planning is vital when designing the physical data center environment. A modular approach to data center design ultimately provides flexibility, and scalability, in both network topology design and utilization of physical resources.

Note Data Center best practices are discussed in greater detail in the Cisco Validated Design documents available at the following link: http://www.cisco.com/en/US/netsol/ns748/networking_solutions_design_guidances_list.html.

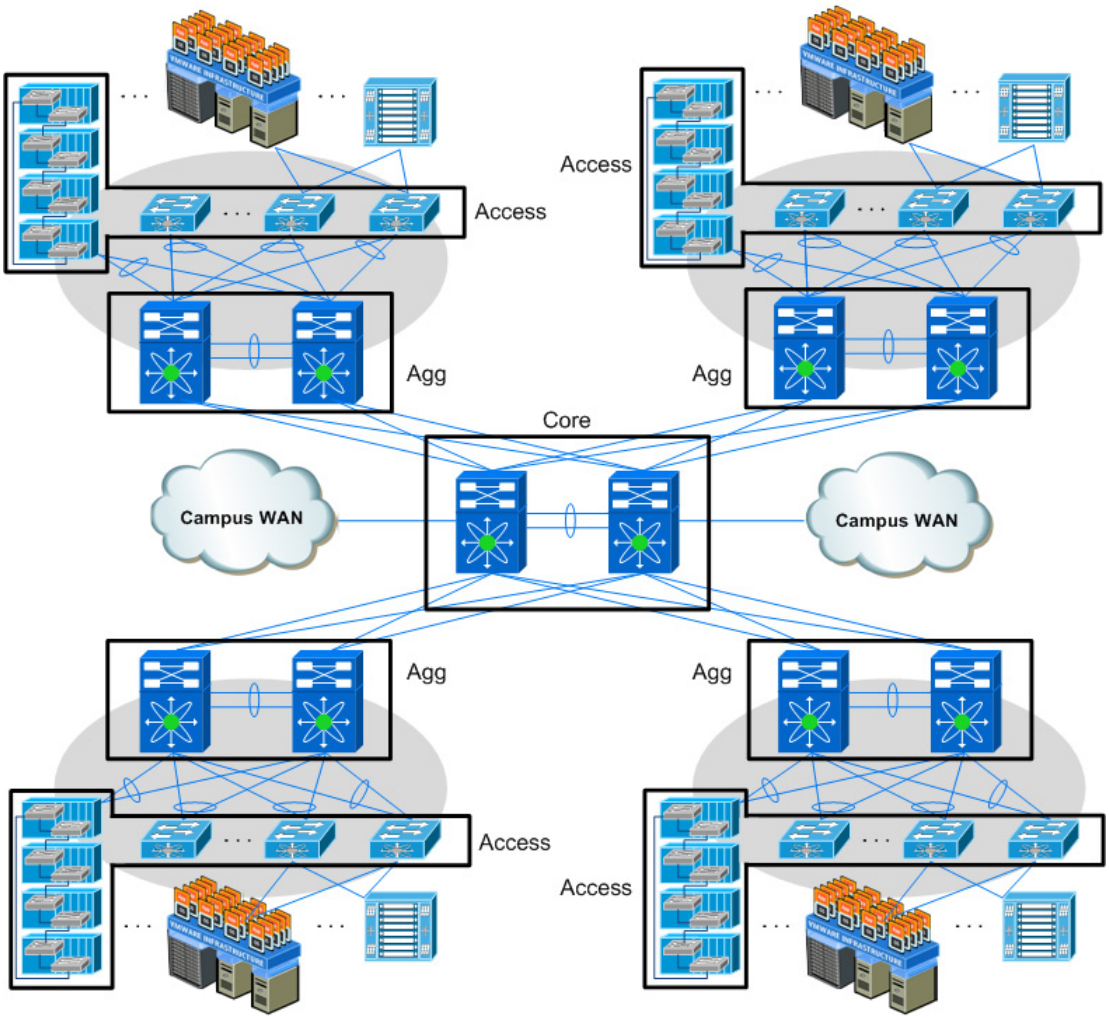
The building block of a data center network is usually named (POD) and consists of the server, and in the context of this document consists of the server, access and aggregation layers, as shown in Figure 1-2:

Figure 1-2 Data Center POD



In large data centers, multiple PODs may be deployed inside the same data center. These PODs are usually interconnected via a Layer 3 core layer, as shown in Figure 1-3.

Figure 1-3 Large Data Center Leveraging a Core Layer



The aggregation layer devices inside a POD traditionally represent the boundary between Layer 2 and Layer 3 network domains. Consequently, when discussing LAN extension requirements, we always talk about inter-POD LAN extension. This happens independently from the fact that interconnected PODs belong to remote data center sites (inter-DC LAN extension) or if they belong to the same location (intra-DC LAN extension).

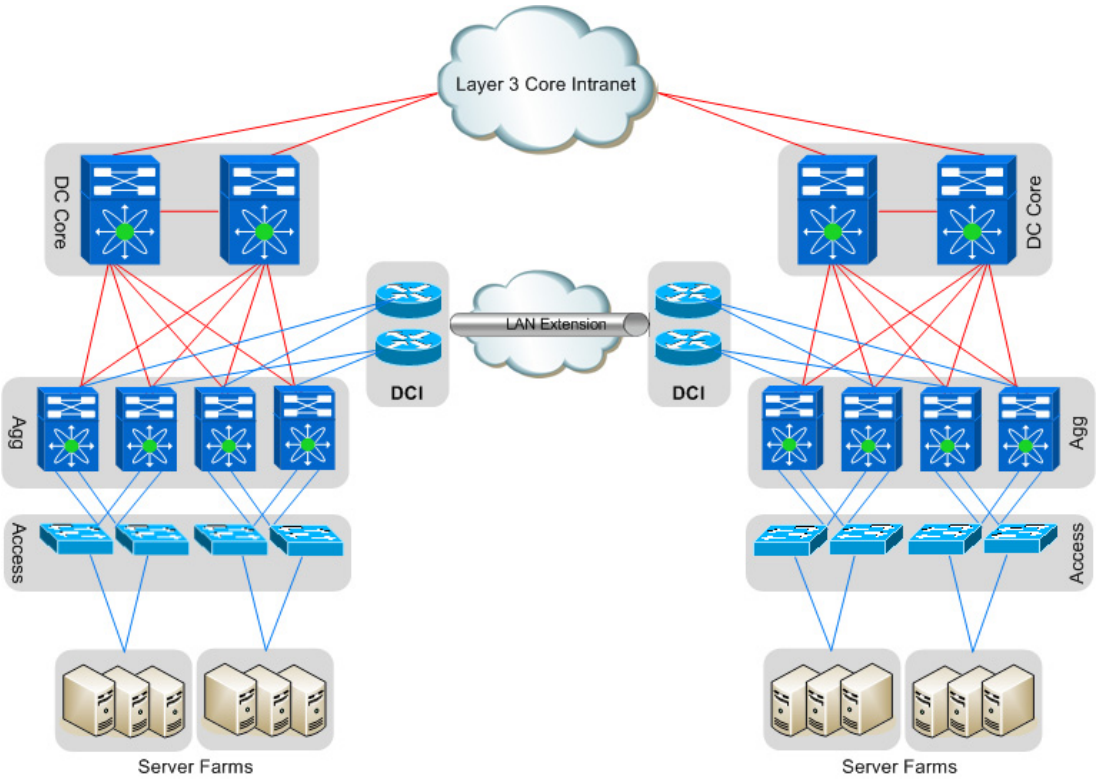
In addition to the traditional access, aggregation and core layers, an additional layer (DCI layer) has been added to the data center network in this DCI solution.

DC Interconnect Layer

The DCI layer is introduced to provide LAN extension functionality core to the DCI solution. This additional layer can be implemented on a pair of dedicated physical network devices or may just represent an additional logical function performed by already existing devices (core layer, aggregation layer, or WAN layer). The two main factors that influence this choice are the size of the data center deployment and the specific technology implemented to provide the LAN extension functionality.

In large-scale data centers where multiple aggregation blocks are built out, a set of redundantly configured devices form a separate DCI layer, which can aggregate L2 links from multiple pairs of Aggregation Layer switches, as shown in Figure 1-4:

Figure 1-4 Large Scale Data Center



When deploying this option, the aggregation layer still the place in the network where routing happens for all traffic originated by the server farms. The VLANs that need to be extended between remote locations (and only these VLANs) are carried on the Layer 2 trunk connections established between the aggregation layer switches and the DCI devices. Each pair of aggregation devices with all the access layer switches connected is referred to as the "Data Center POD."

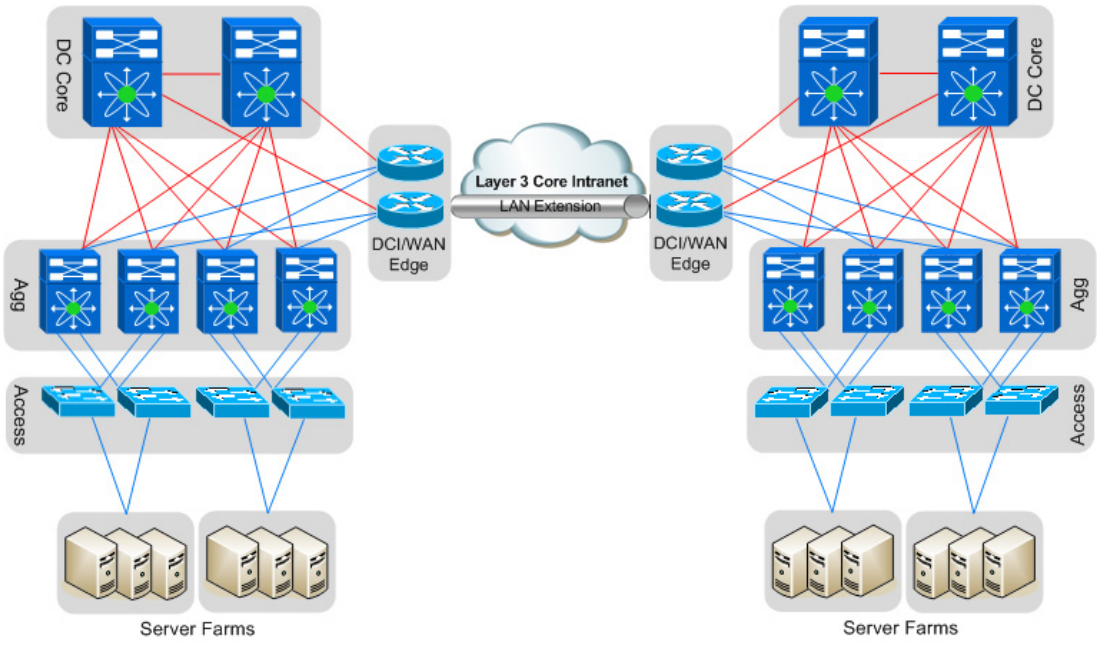
This poses a question: what if a requirement arises to provide LAN extension services between separate PODs inside the same data center? One possible solution to this problem would be to leverage the DCI devices to provide this inter-POD bridging functionality, because a L2 path can always be established between each POD and the DCI devices for all VLANs requiring extension services.

In this scenario, DCI devices would then be dedicated to provide two main functions:

- LAN extension between remote locations.
- Inter-POD bridging if LAN extension is also required inside the same data center location. The capability of performing this local switching functionality usually depends on the specific technology deployed to perform the LAN extension between geographic sites (as will be discussed further in Chapter 2). Also, this approach may lead to the creation of STP loops between each POD and the DCI devices, which may not be desirable. Therefore, we recommend that you limit when possible the LAN extension requirements to a single POD inside a specific data center location or between PODs located in separate geographic sites.

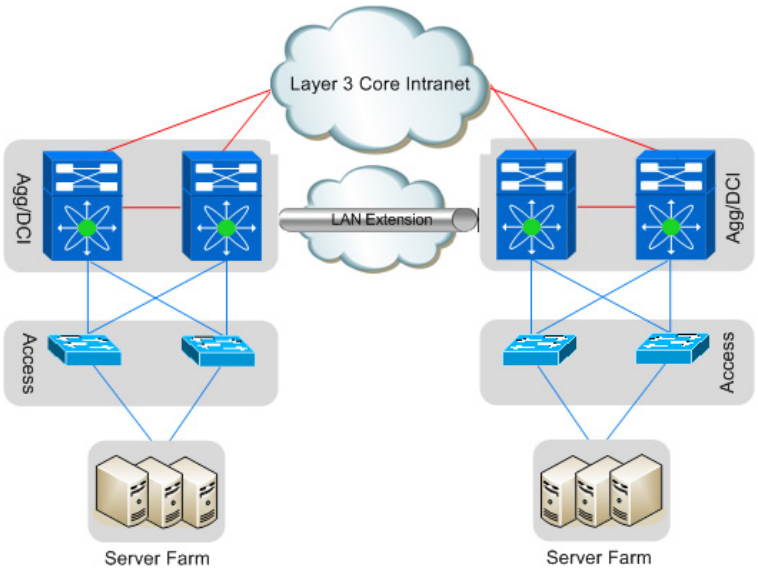
In Figure 1-4, all routing functionality in and out of each data center location is performed by the DC core switches (or by a dedicated pair of WAN edge devices, depending on whether the data center is connected to the Campus core or to a WAN cloud). However, deployments may be possible where a common WAN cloud is used to establish both LAN extension and routing services, as shown in Figure 1-5.

Figure 1-5 Collapsed DCI and WAN Edge layers



For smaller data centers that consist of a single set of aggregation switches, as shown in Figure 1-6, it may be difficult to justify adding layers to extend the Layer 2 domain. In this case, aggregation switches could be used to extend the Layer 2 domain and act as a DCI Layer switch (there is no requirement for the DC core layer in this case).

Figure 1-6 Collapsed Aggregation/DCI Layer



Aggregation layer devices need to provide concurrent routing and bridging functionality under this model. Routing is required to provide connectivity from the DC edge VLANs into the Layer 3 Intranet core, whereas bridging is required to extend the VLANs between remote locations.

LAN Extension Solution Requirements

When deploying a solid DCI solution, you should remember the following solution requirements:

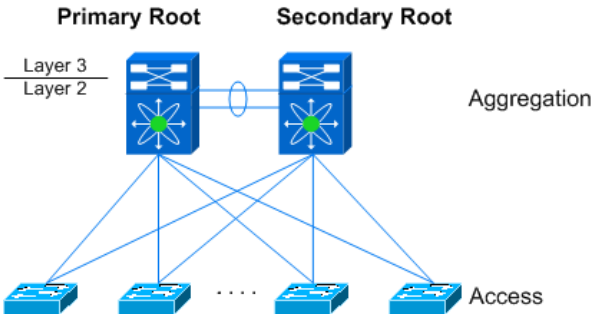
- Spanning Tree Protocol Isolation
- End-to-End loop prevention
- High Availability (HA)
- VLAN scalability
- Multi Path load balancing
- Routing optimization
- Security and encryption
- Hierarchical quality of service (HQoS)

Spanning Tree Protocol Isolation

Typically, an STP domain spans as far as a VLAN reaches. So if multiple data centers share a common VLAN, the STP domain extends to all those data centers.

Figure 1-7 shows that from a STP design standpoint, best practices place the primary and secondary root in an Aggregation Layer. The assumption is that the Aggregation Layer builds the top most level of a Layer 2 hierarchy (the VLAN spans from Access Layer switches to Aggregation Layer switches).

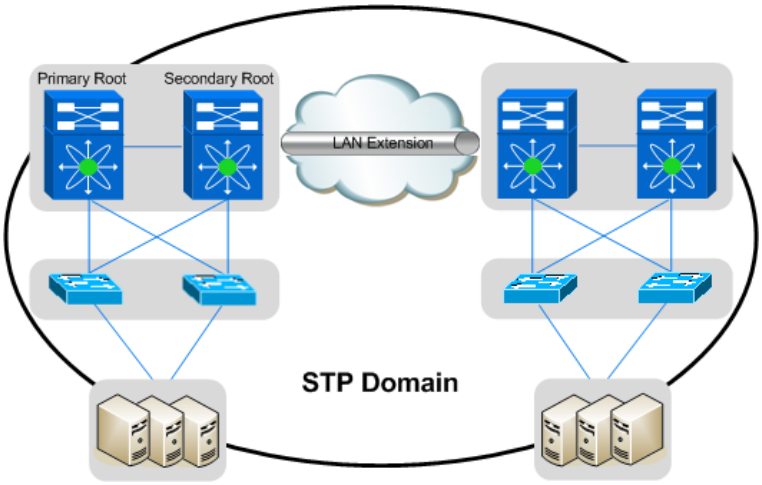
Figure 1-7 Aggregation Layer, Layer 2 Hierarchy



When a VLAN gets extended to another Aggregation Layer, the question arises where the primary and secondary STP root should be placed.

In Figure 1-8, the oval indicates the extended STP domain. It becomes apparent in this illustration that there is no optimal placement of the STP root for data center 2. Each data center should have its own primary and secondary STP root.

Figure 1-8 VLAN Expansion to Another Aggregation Layer



Another STP design consideration regarding multiple data centers is the type of STP used. On Cisco Catalyst and Nexus switches the following STP modes are supported natively or compatibly:

- IEEE 802.1D
- PVST+ (Per VLAN Spanning Tree)
- Rapid-PVST
- IEEE 802.1s MST (Multiple Spanning Tree)

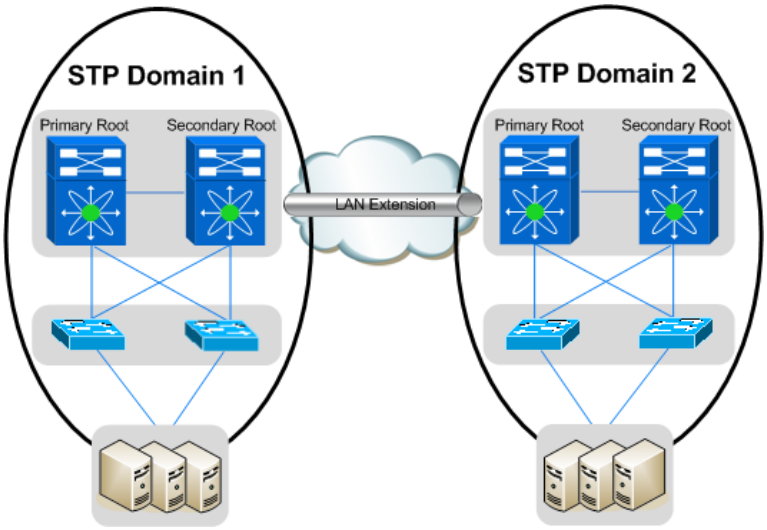
As new data centers are built, the opportunity to move to a more scalable STP mode is given. If VLANs are extended between data centers, it is the existing data center STP mode that typically dictates the mode being used. You should have the freedom of moving to the STP mode that makes the most sense from a device support and scalability standpoint.

Ideally, STP events should be isolated to a single data center. Typically, any link state change of a port that participates in a VLAN, triggers a BPDU with the topology change notification (TCN) bit set. Ports that are configured as edge ports, or configured with Spanning Tree PortFast, do not trigger TCN

BPDUs. When the STP root receives a BPDU with the TCN bit set, it sends out a BPDU with the topology change (TC) bit set resulting in all bridges initiating a fast age out of MAC address table entries (802.1D or PVST+), or immediately flushing those MAC entries (Rapid-PVST or 802.1w). This may result in temporary flooding events in the respective VLANs, which is an undesired but essential side effect to ensure MAC address re-learning.

To address the three design and implementation challenges (root placement, STP mode, BPDU isolation) previously discussed, we recommendation that you isolate the STP domains in the different data center sites, as shown in Figure 1-9.

Figure 1-9 Isolation of STP Domains



In Figure 1-9, the STP domain for a VLAN, or a set of VLANs, is divided (indicated by the two ovals). Consequently, the STP root is localized to each data center (the STP root design is now optimal). Given that there is no interaction between data centers, from an STP standpoint, each data center can make use of an independent STP mode. An existing data center might run Rapid-PVST while the new data center can run MST. Also, BPDU TCN messages are local to the data center and do not initiate any fast aging of MAC address table entries in other connected data centers.

Different techniques can be deployed to ensure the creation of separate STP domains between remote data center sites; the right deployment choice usually depends on the specific DCI solution that is deployed to provide inter-site LAN extension functionality. Refer to DCI Phase 2.0 Networking Technology, page 1-23, for details.

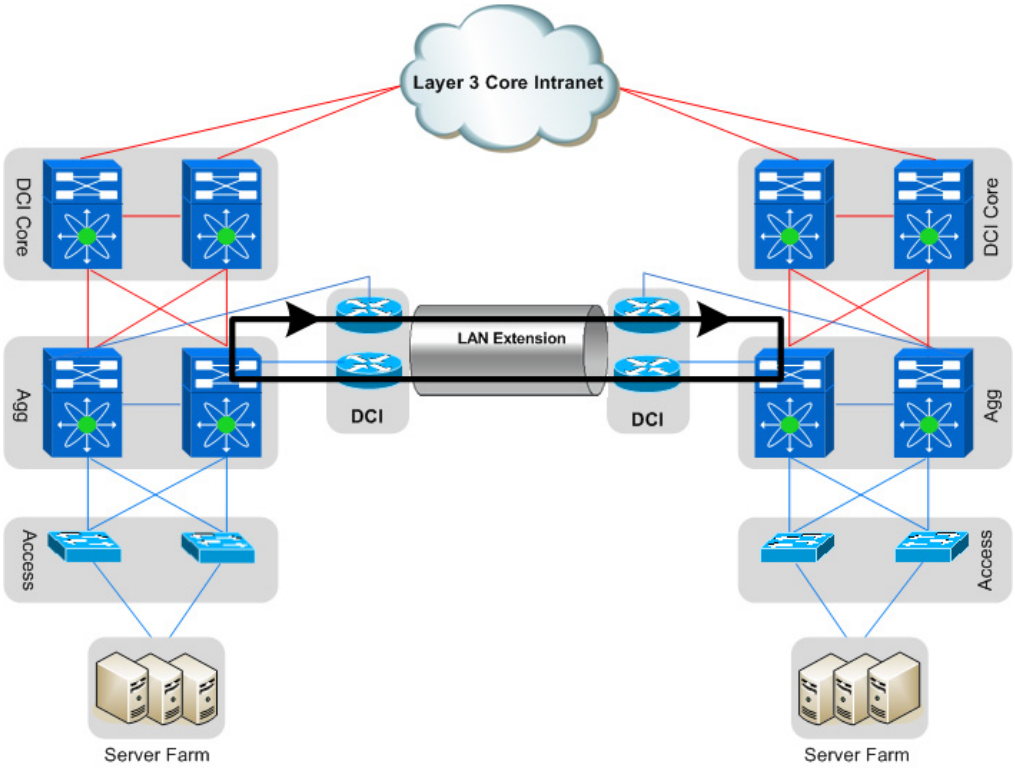
End-to-End Loop Prevention

In addition to STP domain isolation between data center sites, another solution requirement is the ability to prevent end-to-end loops. The potential problem is that while STP ensures the creation of a loop-free topology in each site (or at least represents a safety belt mechanism in local site designs that are by definition loop-free because leveraging technologies like VSS/vPC), the interconnection of two data centers together at Layer 2 may lead to the creation of an end-to-end STP loop. This happens despite the fact that most of the technologies available today to create the Layer 2 connections are actually loop-free in and of themselves.

As shown in Figure 1-10, an end-to-end STP loop is created mainly for two reasons:

- Each data center site is dual attached to the cloud across which the LAN extension services are offered. This is a typical requirement to ensure overall solution resiliency and high availability.
- STP BPDUs are not sent across the DCI connection, because of the previously discussed requirement of isolating the STP domains in separate data center sites.

Figure 1-10 Creation of am End-to-End STP Loop



As will be discussed, different techniques can be utilized to prevent an end-to-end loop. These techniques depend on the solution deployed to provide LAN extension services.

High Availability (HA)

High Availability is one of the most important requirements of any network design, given the 24x7 nature of today's businesses. A couple of relevant HA aspects for DCI solutions are listed below:

- **Network Resiliency:** Building a resilient network infrastructure is paramount to every data center design. This implies the deployment of fully redundant data center sites, characterized by redundant physical devices in each layer (access, distribution, core) of the data center network, and redundant physical connections between these network elements. When deploying a DCI solution to interconnect remote data center sites, it is also required to ensure redundancy in the DCI layer performing LAN extension functions.
- **Disaster Prevention and Recovery:** Disasters can and do strike anytime, anywhere, often without warning. No business can afford to shut down for an extended period of time. An effective business continuity program is imperative to keep a company up, running, and functional.

Before disaster strikes, it is essential for businesses to ensure that the right failover mechanisms are in place, most often in geographically dispersed locations, so that data access can continue, uninterrupted, if one location is disabled.

By integrating virtualization in all aspects of a data center design, with capabilities of seamlessly moving and migrating services across geographically distributed sites, data centers can provide more flexibility in disaster prevention and offer recovery rapidly after a failure occurs.

Additional best practices for disaster prevention and recovery are available at the following link:

http://www.cisco.com/en/US/netsol/ns749/networking_solutions_sub_program_home.html

Multi-Path Load Balancing

When building an interconnect between data centers, availability and redundancy are of utmost concern for the solution. Since each data center site is usually fully redundant, often implying redundant connections also between sites, no single link failure can disrupt communication between them. Unless sites are geographically adjacent (within same campus), connections between those sites are typically costly. Typical requirements are to use as much bandwidth as possible, balance the load across the multiple available connections, and ensure fast failovers under any failure condition. Ideally, a multi-path solution should use all available links.

With Layer 2 extended, STP assures only one active path to prevent loops, which conflicts with the goal of using all links. One way to resolve this conflict is to configure spanning tree parameters so that even VLANs are forwarded over one link and odd VLANs are forwarded over another. This option is discussed in VPLS multipoint deployment. Another alternative is to present a single logical link to spanning tree by bundling links using EtherChannel. The DCI solution validated in phase 1.0 focused on the latter concept, and the same concept will be reused in the point-to-point EoMPLS scenario.

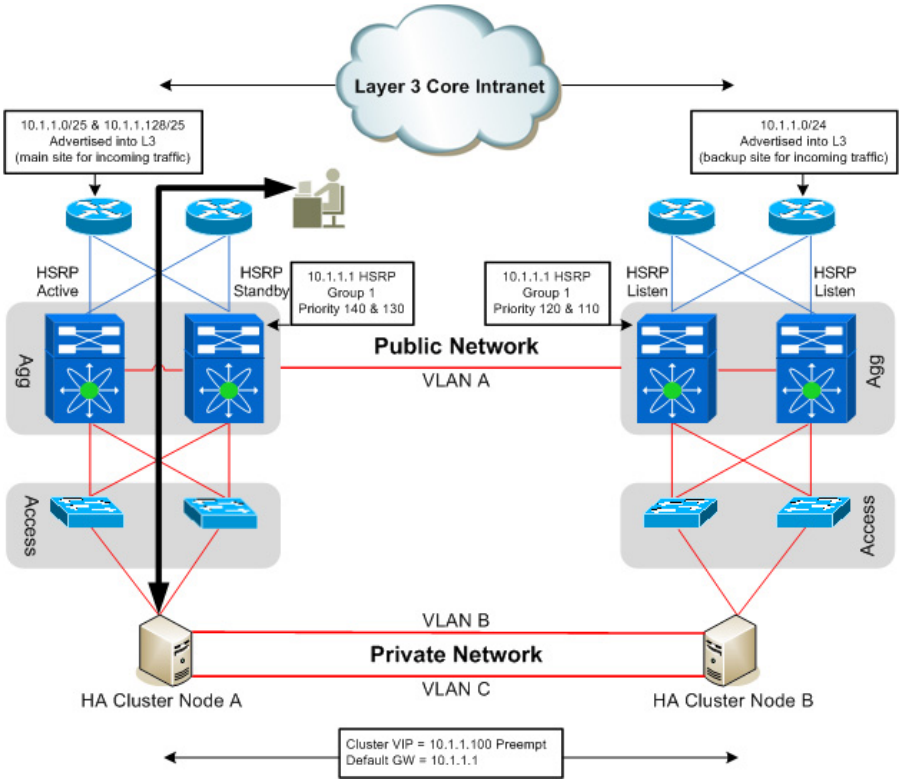
Note It is a best practice to extend only those VLANs that are required to offer virtualization services.

Routing Optimization

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.

It is common practice to elect a primary DC for specific or multiple applications. For example, Figure 1-11 shows a geo-cluster represented by two physical nodes located in two separate data center sites.

Figure 1-11 Client-Server Communication Under Normal Circumstances

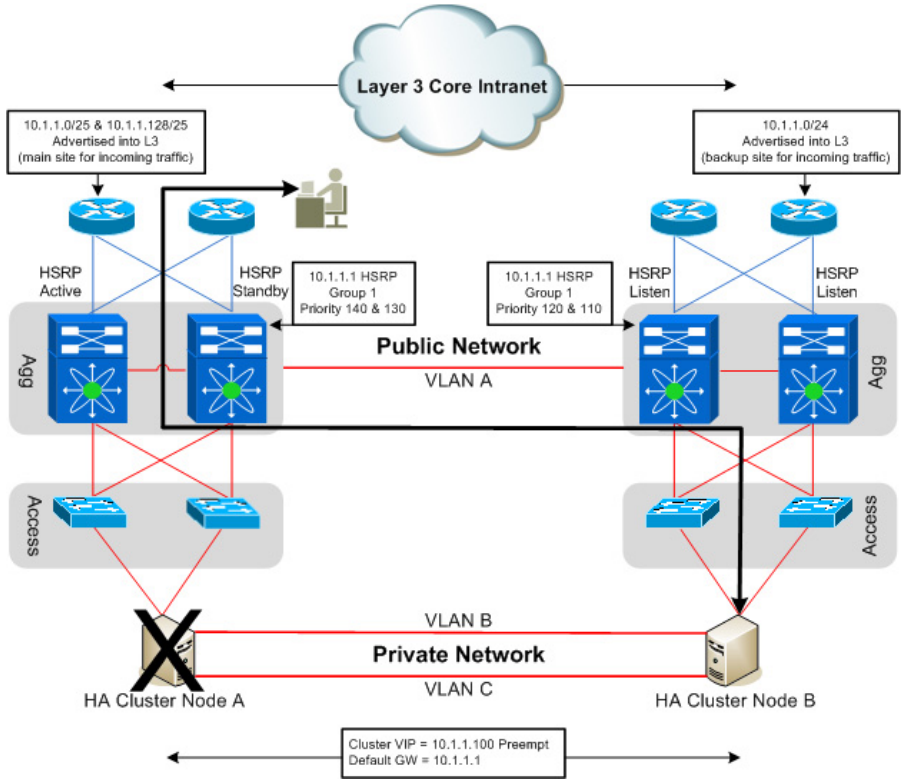


Under normal conditions, the active node is located on the left side and reachable via a specific VIP address (10.1.1.100). To ensure traffic flowing between a client and this active cluster node follows an optimal routing path through the left data center, two traffic flows are required:

- **Inbound traffic (client to server):** DC core devices in the left site can advertise two /25 prefixes (10.1.1.0/25 and 10.1.1.128/25) for the subnet where the cluster VIP address resides. At the same time, the core devices in the right site advertise a less specific /24 prefix (10.1.10/24). This ensures that client-to-server traffic always follows the preferred path through the left data center.
- **Outbound traffic (server to client):** Proper HSRP configuration between the four aggregation layer devices deployed in the two data center ensures, that by design, the active default gateway is located in the same site of the active cluster node. This maintains symmetry between incoming and outgoing traffic flows.

Given the physical redundancy (links and nodes) built in each data center site, any single network device failure can be tolerated without modifying the traffic direction. But what if the cluster node itself fails, as shown in Figure 1-12?

Figure 1-12 Failure of the Active Cluster Node



In this failure scenario, the cluster node in the right data center becomes active (reachable through the same VIP address 10.1.1.100). This event alone does not cause any change to the routing advertisements or HSRP settings previously discussed. Consequently, traffic originated by a client and directed to the VIP address is still attracted by the left data center, whereas return traffic originated by the cluster node can only be routed toward the client after reaching the active default gateway also located in the left site. The net effect is the suboptimal (even if still symmetric) traffic path highlighted in Figure 1-12.

Depending on the specific characteristics of the DCI deployment (distance between sites, available bandwidth, and so on), this behavior may be acceptable. In other cases, however, some enterprises may want to dynamically modify the overall traffic flow to always ensure an optimal path between the client and the application server.

This DCI Phase 2.0 solution emphasizes the first approach in Figure 1-12. Detailed information on how to deploy routing optimization techniques will be covered in future releases.

Note The discussion above refers to a geo-cluster deployment, but the same considerations are valid independently from the specific application driving the LAN extension requirement (another typical example is the long distance VMotion of a virtual machine between two remote sites).

Storage Access

Routing considerations made in the previous section are important to ensure the establishment of optimal and symmetric traffic paths between client and server. Another critical consideration every time a device is moved between remote sites is the access to the storage disk. Possible approaches are Layer 2 mobility

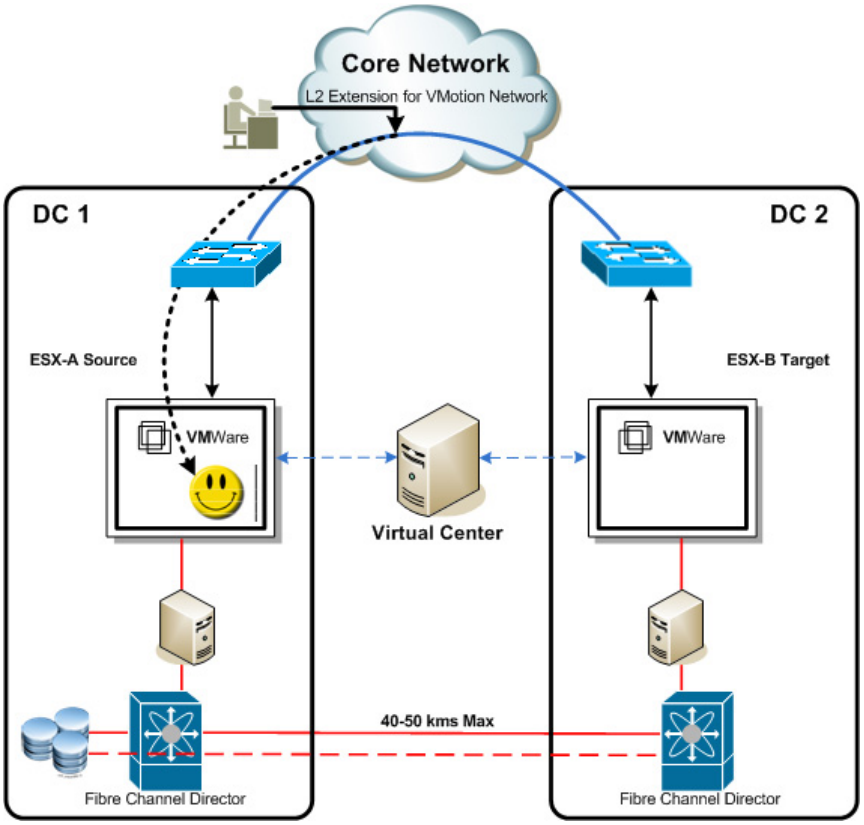
with shared storage, and Layer 2 mobility with active-active storage. The focus in this phase 2.0 is on the validation of the shared storage approach. Alternative solutions (like the active-active one discussed here) may be considered for future phases.

Shared Storage

Figure 1-13 highlights a VMotion scenario where the LAN extension connection between the two data centers is leveraged to move a virtual machine from site 1 to site 2. This process could be initiated manually by a system manager, but VMware also allows a more dynamic process for workload mobility.

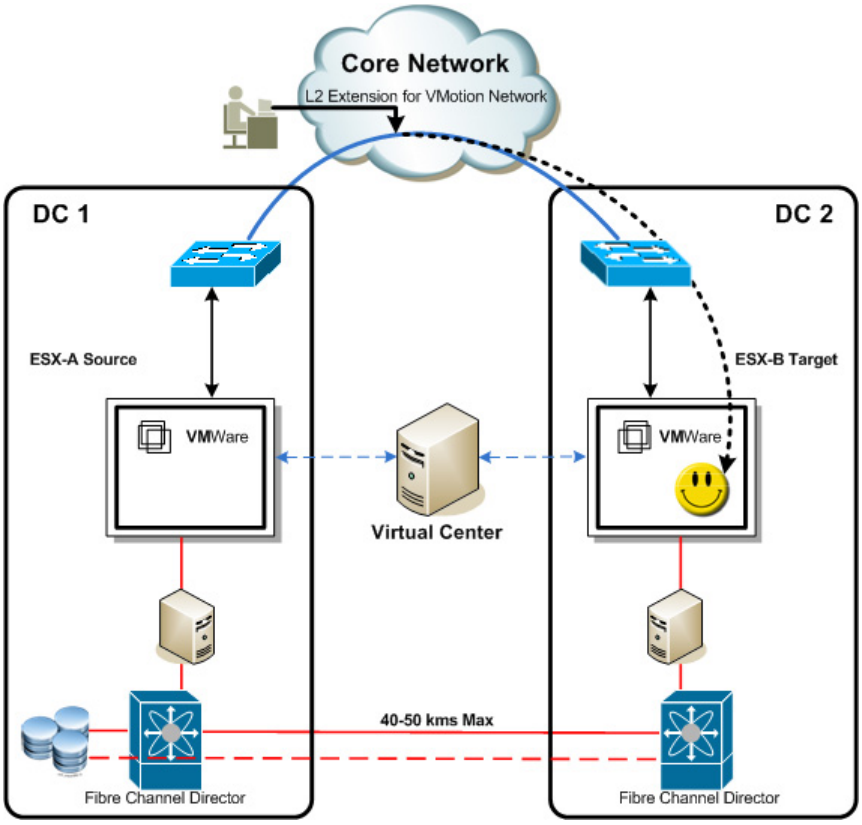
Note Considerations around storage access can also be applied to other applications requiring Layer 2 mobility between geographic sites.

Figure 1-13 Extension for VMotion



The virtual machine is initially located in data center 1 and accesses the storage disk located in the same site. After the virtual machine moves to data center 2, it keeps using the same physical disk sitting in the original location (Figure 1-14).

Figure 1-14 Shared Storage Access with VMotion



To avoid impacting I/O disk performance, you must limit the maximum distance between the two sites. Assuming the interconnection between the FC switches (SAN extension) is achieved through a dedicated connection, the maximum distance should not exceed 40-50kms.

The use of Cisco MDS 9000 FC switches provides an added value for these extended SAN topology challenges. In addition to the performance guaranteed by the high buffer-to-buffer credit capacity, the MDS integrates, in fact, a specific functionality (called I/O Acceleration) to accelerate I/O operations. This feature helps the overall application performance to remain almost the same when the storage is separated by distances longer than the 40-50kms.

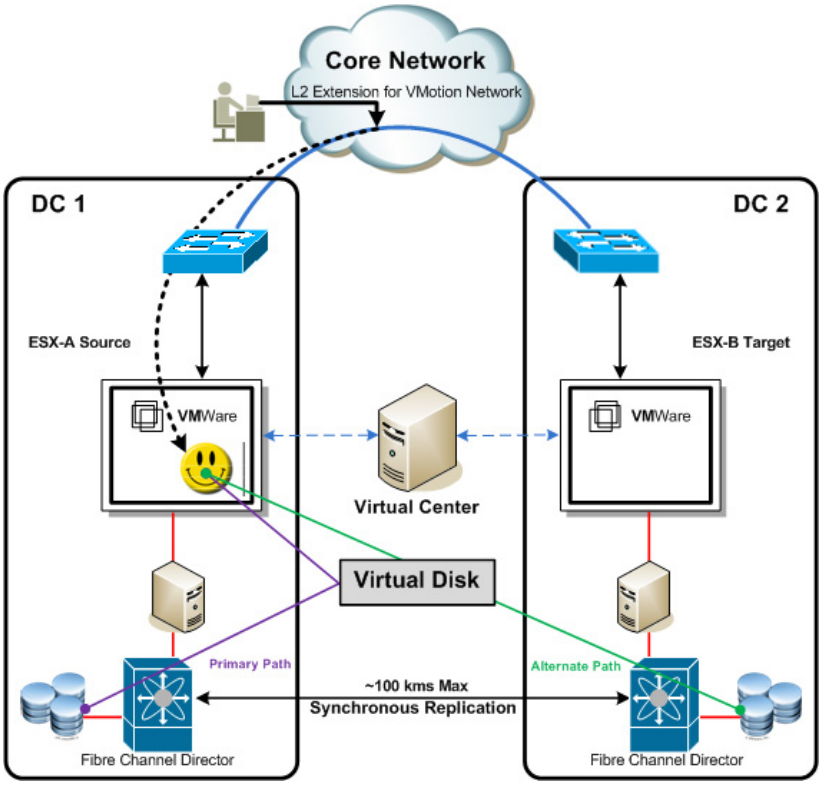
Note The following link provides a white paper discussing this type of deployment, where the VMotion of a virtual machine was achieved between locations distant up to 200kms: http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns836/white_paper_c11-557822.pdf

Active-Active Storage

A different approach to solving the problem of storage access after the movement of a device between remote sites suggests the use of Active-Active storage. An implementation example for this approach leverages the concept of a virtual disk that is offered to the virtual machine (instead of a real storage disk). This virtual disk is then linked to the physical storage via a primary path (local storage disk) and a secondary path (remote storage disk).

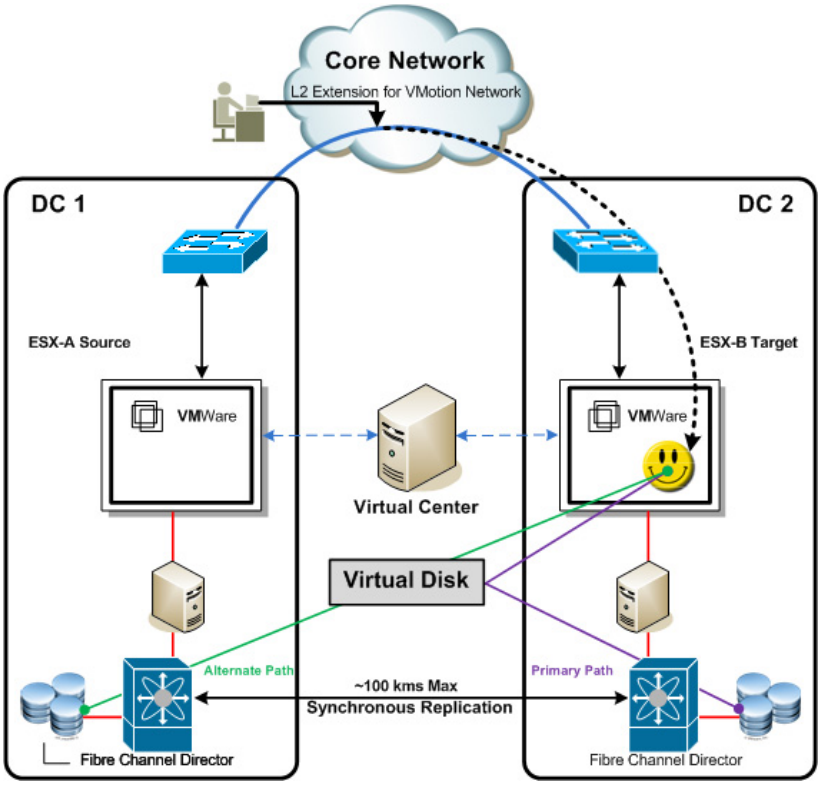
As shown in Figure 1-15, while the virtual machine resides in data center 1, the primary path is leveraged to access the local physical disk (the remote storage would be used only in case of the failure of the primary path).

Figure 1-15 Access to Local Storage Before VMotion



Once the virtual machine moves to data center 2, it still accesses the virtual disk. The difference is that now the primary path provides access to the storage disk located in data center 2, optimizing I/O operations (Figure 1-16).

Figure 1-16 Storage Access After VMotion



Key functionality empowering this Active-Active storage approach is the synchronous mirroring between physical storage disks located in the remote sites. Therefore, this solution is usually deployable between locations with a maximum distance of 100kms.

Note Further consideration regarding active-active storage deployment will be included in future DCI system releases.

Security and Encryption

Security is a primary concern for data center architects. Security considerations range from the physical location over physical access control to the data center, all the way to traffic flow to and from, as well as within and between data centers. Encryption of critical or confidential data is also often vital.

Note This section focuses on security and encryption as it relates to Ethernet and IP networking. Technologies like storage media encryption provided by the Cisco MDS 9500 Fibre Channel director switches is outside the scope of this document, as are physical security aspects of a data center.

Securing and Hardening Network Infrastructure

Running a secure and hardened infrastructure is the foundation for continued operations of a data center. Access via SSH, or console/terminal servers to the infrastructure has to be secured and controlled. Requirements here are no different from what has been known in the past.

Access control via Authentication, Authorization and Accounting (AAA) using protocols such as TACACS+ or RADIUS must be used to prevent unauthorized access to data center devices. Also, access to the management interface should be allowed from specific subnets only. Client/Server subnets should not be able to access the management interface/network. The AAA infrastructure (Cisco ACS) should be built as fault tolerant since it plays a critical role in access and management devices. The infrastructure is often built using an out-of-band (OOB) management network, providing complete separation of management and client/server traffic.

Further security mechanisms include the following:

- **NetFlow:** Allows monitoring the overall amount of traffic and number of flows and that can trigger an alarm in case levels go beyond predefined thresholds.
- **Address Spoofing Filters:** Incorrectly configured end stations might inject traffic which could impact overall network performance.
- **Native VLAN:** Generally it is a best practice not to allow user traffic on the native VLAN. If there is still a requirement to have user traffic on the native VLAN, we recommended that you configure the switch to tag traffic for the native VLAN.
- **Port Security:** Port Security allows for specifying how many MAC addresses should be allowed on a physical port. Faulty network interface cards (NICs) have shown in the past that random traffic may be sent to the network. If the source MAC address is “randomized,” this can result in either creating unnecessary MAC moves and/or CAM tables filling up. The latter can result in unnecessary flooding of traffic. Caution should be used for servers hosting multiple virtual machines since those could be hosting tens of MAC addresses.

For more information on security baseline techniques to harden the network infrastructure, please refer to Cisco SAFE reference guide available at the following location:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Security/SAFE_RG/SAFE_rg.html

Storm Control and Flooding

For large Layer 2 domains, the following traffic types are flooded to all ports that participate in a VLAN:

- Broadcast
- Layer 2 multicast (non-IP-can not be pruned using IGMP)
- Traffic to unknown destinations (unknown unicast)

Some protocols, essential in IP networking, rely on broadcast. The most common is the Address Resolution Protocol (ARP), but depending on the network environment, there may be others. A hardware fault or application level error can cause massive levels of broadcast or unknown unicast traffic. In extreme cases it can happen that a whole data center is affected and effectively no “good” traffic can be passed anymore. To stop a spread of such traffic across multiple data centers, the Catalyst and Nexus families of switches allow suppression of these kinds of traffic types.

It is important that you baseline broadcast, Layer 2 non-IP multicasts, and unknown unicast traffic levels before applying suppression configurations which should be applied with headroom and constantly monitored. The introduction of new, or deprecation of legacy, applications could change traffic levels.

When applied properly, even a STP loop in one data center may only cause a limited amount of traffic to be sent to other data center(s).

For Storm Control details on specific platforms refer to the following:

Catalyst 6500

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/storm.html

Nexus 7000

http://www.cisco.com/en/US/docs/switches/datacenter/sw/4_1/nx-os/security/configuration/guide/sec_storm.html

Protecting the Control Plane

State of the art switching platforms perform forwarding in hardware using Application Specific Integrated Circuits (ASICs). Only traffic destined to the switch, typically management traffic or protocols that run among network devices, uses the switch's CPU. Given that switches aggregate large amounts of bandwidth, the CPU can potentially be stressed beyond normal levels through incorrect configurations and/or malicious traffic. To protect the CPU from such attacks, rate limiters and QoS policies can be applied. This is called Control Plane Policing (CoPP). Various Catalyst as well as Nexus platforms offer hardware-based CoPP capabilities.

Each network environment is different. Protocol scaling, number of servers, and traffic levels dictate the correct CoPP values. The Nexus 7000 has CoPP configured by default. The network administrator can choose from multiple predefined profiles (strict, moderate, lenient, none) and adjust these profiles to meet specific needs.

To find the optimal CoPP configuration, perform a baseline that includes low and high network activity conditions, whereupon an optimal configuration can be found (that is baseline plus margin). Continuous monitoring should be in place to observe CPU trends that may trigger an adjustment of CoPP policies.

For CoPP detailed support considerations on specific platforms, refer to the following:

Catalyst 6500

http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/prod_white_paper0900aecd802ca5d6.html

Nexus 7000

http://www.cisco.com/en/US/docs/switches/datacenter/sw/4_1/nx-os/security/configuration/guide/sec_cppolicing.html

Encryption

Enterprises and service providers store data in data centers that contain information that needs to be protected, such as data representing intellectual property, sensitive information like patient and financial data, and regulatory information. Depending on the type of corporation, there might be regulatory requirements that have to be followed by the respective organization. Examples of such data protection requirements are HIPAA, PCI, Sarbanes-Oxley (SOX), Basel II, and FDA.

Due to these constraints, data passed between locations may have to be encrypted. This requirement holds true for data base synchronization traffic, applications exchanging state information as well as users accessing applications. Generally speaking as soon as traffic leaves the physical data center premises, data needs to be encrypted. This not only assures data privacy but also data integrity. No one should be able to modify data while it is in transit.

Note Encryption may become optional in deployment scenarios where the enterprise owns the core of the network interconnecting the various data center locations.

While some applications encrypt data at the presentation or application layer, it is typical that hundreds if not thousands of applications are active within an enterprise corporation. Changing or enhancing all corporate applications to communicate over a secure channel can be time consuming and costly. Network-based encryption, however, can encrypt all traffic between two or more locations since it acts at the lower layers in the 7 layer OSI model and therefore does not require any changes to the application.

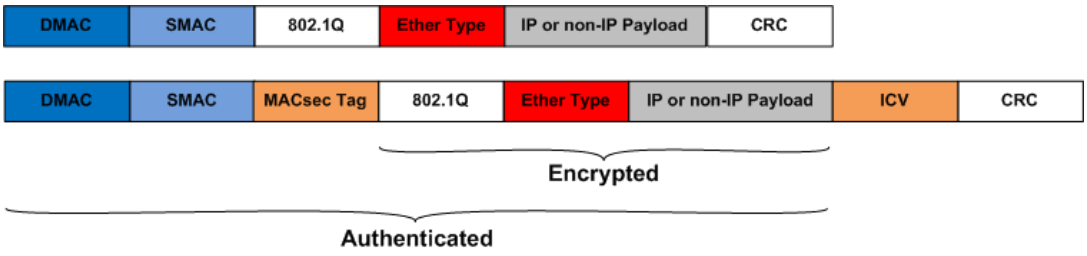
In the context of a DCI solution, the following two network encryption protocols are considered:

- **IPSec:** This protocol is commonly used to encrypt and authenticate IP packets and sends them across an encrypted tunnel. Encryption performance and the operational aspects of setting up IPSec tunnels has been greatly simplified over the years with automated tunnel set up mechanisms (like DMVPN and GETVPN to name a couple of technical alternatives).

  The main issue to consider for applying IPSec-related technologies to a DCI solution is that based on the way the IPSec tunnel technology is defined today, only traffic that is routed (and not bridged) into a tunnel can be encrypted. In other words, IPSec natively does not allow bridged traffic to be encrypted. Because providing LAN extension services is one of the main requirements of a DCI solution, this implies that IPSec can be considered for this type of traffic only when an additional Layer 2-in-Layer 3 tunneling protocol (such as GRE for example) is deployed, but could not be applied, for example, to native MPLS traffic.

- **802.1AE:** In 2006 the IEEE ratified the 802.1AE standard, also known as MAC security standard (MACsec). MACsec encrypts all Ethernet frames, irrespective of the upper layer protocol. With MACsec, not only routed IP packets but also IP packets where the source and destination is in the same subnet or even non-IP traffic are encrypted; this make the technology very appealing for encrypting LAN traffic carried between remote locations. Figure 1-17 shows how a regular Layer 2 frame is encrypted leveraging the 802.1AE functionality.

Figure 1-17 802.1AE Encryption and Authentication



As shown, 802.1AE not only protects data from being read by others sniffing the link, but it also assures message integrity. Data tampering is prevented by authenticating relevant portions of the frame. The MACsec Tag plus the Integrity Check Value (ICV) make up 32 Bytes (no 802.1AE metadata is considered). While the 802.1Q header, the ether type and payload are encrypted, destination and source MAC are not. As for the integrity check, the whole frame, with the exception of the ICV and the CRC, is considered. This assures that not even a source or destination address of the frame could be manipulated.

From a deployment standpoint, two important points need to be highlighted (both valid at the time of writing of this document):

- While IPSec is fundamentally building a tunnel that goes from one IP address to another over potentially many hops, 802.1AE encryption is performed at the link layer and is hence supported between devices that are directly connected (meaning on a hop-by-hop basis).
- The Nexus 7000 is the only platform supporting IEEE 802.1AE standards-based encryption in hardware on Ethernet interfaces. This hardware level encryption is where each port has its dedicated crypto engine that can be turned on in NX-OS configuration. Since each port has dedicated resources, there is no tax on the supervisor or line card CPU. All data encryption and decryption is performed at the port level. Apart from the additional MACsec header, there is no overhead or performance impact then turning on port level encryption.

Hierarchical Quality of Service (HQoS)

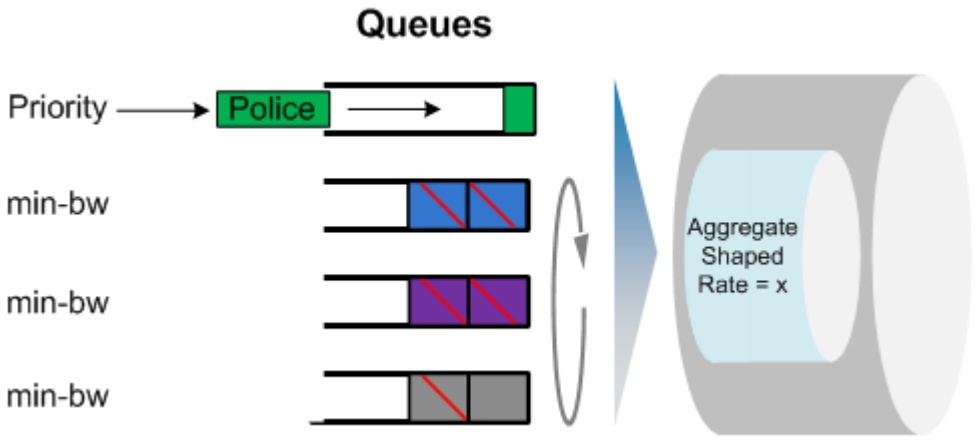
Some DCI deployments may have the luxury of leveraging high bandwidth network connections dedicated to provide the LAN extension functionality. This is the case when deploying dark fiber between the disperse data center sites. In these scenarios the available pipes may be large enough not to require any kind of QoS configuration for the L2 traffic flowing between sites.

In other scenarios, it may be possible that the network used to provide LAN extension services is actually shared with other functions (for example, Layer 3 connectivity or storage synchronization). In that case, the configuration of specific QoS policies on the DCI layer devices may become a mandatory requirement, to ensure a proper SLA for the application traffic flowing between the remote data center sites. This is even more critical when the deployed applications move from faster LAN links, such as 10GE, to slower WAN links, such as subrate GE, on the DCI WAN edge router.

An additional level of complexity is then requested by the customers who are offered (usually by a SP) sub-rate services to interconnect their data centers. In this case, the definition of a QoS policy is not enough, and there is going to be the requirement for certain levels of shaping. This functionality is commonly known with the name Hierarchical QoS (HQoS).

In Figure 1-18, deploying an HQoS policy essentially means first shaping the available bandwidth of the interface to a value below the physical interface pipe, and then applying a more traditional QoS policy to that reduced pipe.

Figure 1-18 Hierarchical QoS



Needless to say, different platforms and HW components may offer different HQoS capabilities (or not offering them at all), so this may be an important factor to keep into consideration when defining a specific DCI solution.

DCI Phase 2.0 Networking Technology

Phase 2.0 testing is focused on the use of MPLS functionality to provide the LAN extension services required for DCI deployments, specifically EoMPLS and VPLS. Given the nature of these technologies, they can be natively deployed only in scenarios where different DC sites are connected via a L1 or L2 types of service offered by a SP, or when the connectivity is achieved through an MPLS enabled core owned by the enterprise.

In cases where the enterprise leverages a L3 service (IP service) offered by the SP (meaning, the customer edge device is peering at L3 with the SP network), it is usually not possible to use these technologies in their native form, since the SP does not usually accept MPLS tagged traffic. To cover these scenarios, GRE tunnels can be established between the enterprise edge devices, and EoMPLS or VPLS, can then be run on top of these logical overlay connections (EoMPLSoGRE and VPLSoGRE).

Note VPLSoGRE deployment discussions are out of the scope of this document.

Note Carrier-supporting-carrier deployments are an exception, but given they are not commonly offered they are not discussed in this document.

By combining the type of connectivity between remote sites with a number of these locations, you can then obtain the different deployment options that were validated during DCI phase 2.0.

Site-to-Site Deployments

- EoMPLS port mode
- EoMPLSoGRE port mode (with or without IPSec)

Multi-Site Deployments

- VPLS

Site-to-Site Deployments

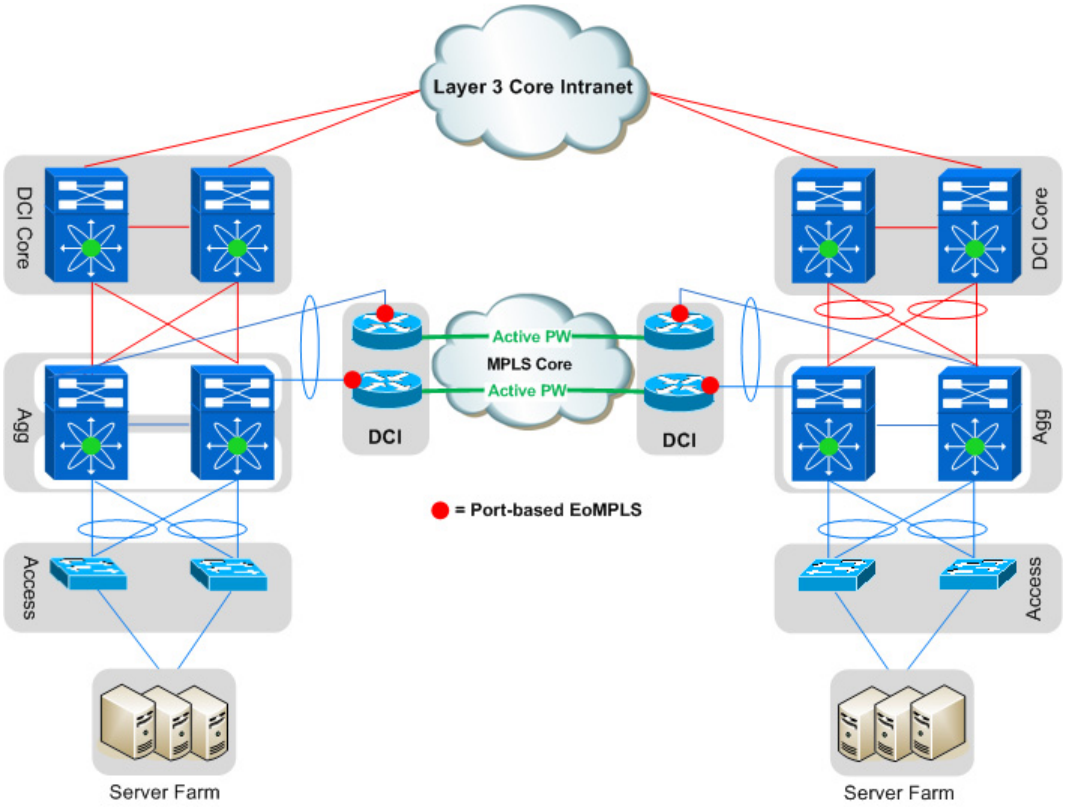
Site-to-site technology deployment recommendations include the following:

- EoMPLS Port Mode, page 1-23
- EoMPLSoGRE Port Mode, page 1-29

EoMPLS Port Mode

The deployment of port-mode EoMPLS, shown in Figure 1-19, represents a simple way to provide LAN extension services between two remote data center locations.

Figure 1-19 EoMPLS Port Mode Deployment

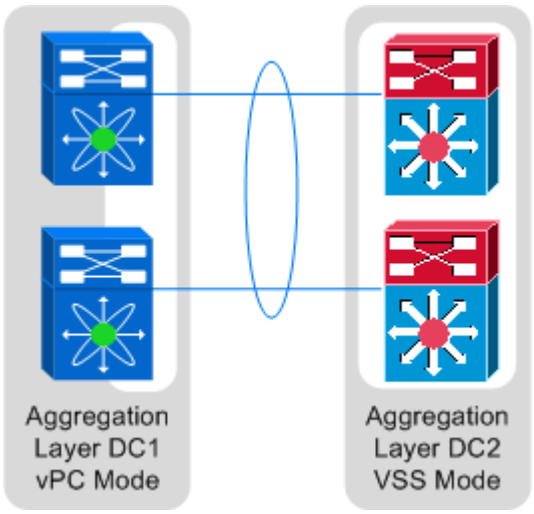


When positioning this solution, you should know the following:

- Given the nature of the EoMPLS technology, this solution is usually suited for point-to-point deployments. Multi-point deployments (where the LAN extension functionality needs to be provided across more than 2 remote locations) are covered leveraging VPLS services.
- The EoMPLS session between devices deployed in the DCI layers can be established under the assumption that the MAN/WAN cloud interconnecting the Data Center is MPLS enabled. This is usually the case if the enterprise owns the cloud. In scenarios where the WAN connectivity services are instead offered by a service provider, the EoMPLS session can actually be established only if the provider offers L1 or L2 connectivity services to the enterprise. In scenarios where the SP offers a L3 service (that is the enterprise edge device peers at L3 with the SP edge device) the use of EoMPLS between the DCI devices is usually not an option, unless the functionality is enabled on a GRE overlay. This specific scenario is discussed later.
- A single EoMPLS Pseudo Wire (PW) is established from each DCI edge device and interconnects with the peer device in the remote site. This allows for the creation of two independent PWs between the remote locations that are always active and forwarding traffic. This is valid independently from the number of VLANs that are extended across the two locations and the result of the EoMPLS port mode functionality. Extension for up to 2000 VLANs will be discussed further in Chapter 2.

From a functional perspective, the use of port-mode EoMPLS allows transparent interconnects of aggregation layer devices deployed in the two remote sites, establishing an end-to-end port-channel between these devices. Logically, the aggregation layer switches are connected back-to-back, as shown in Figure 1-20.

Figure 1-20 Logical Back-to-Back Port-Channel Between Aggregation Layer Switches



This behavior is possible because all the traffic received at the DCI layer on the physical link coming from the aggregation device is blindly tunneled across the EoMPLS PW established with the remote location, effectively making the DCI layer and the EoMPLS cloud in the middle functionally transparent.

The EtherChannel between the aggregation layer devices deployed in different data centers can be established when the same type of device is available on each site, like the Nexus7000 leveraging vPC functionality, or a pair of Catalyst 6500 switches configured in VSS mode. However, as shown in Figure 1-20, the EtherChannel can also be established in a hybrid scenario where a pair of Nexus 7000 are deployed in a data center (usually a green field deployment), whereas a VSS is used on the other data center (representing an installed base deployment).

The following sections discuss how EoMPLS port mode technical option can meet all requirements of a solid DCI solution.

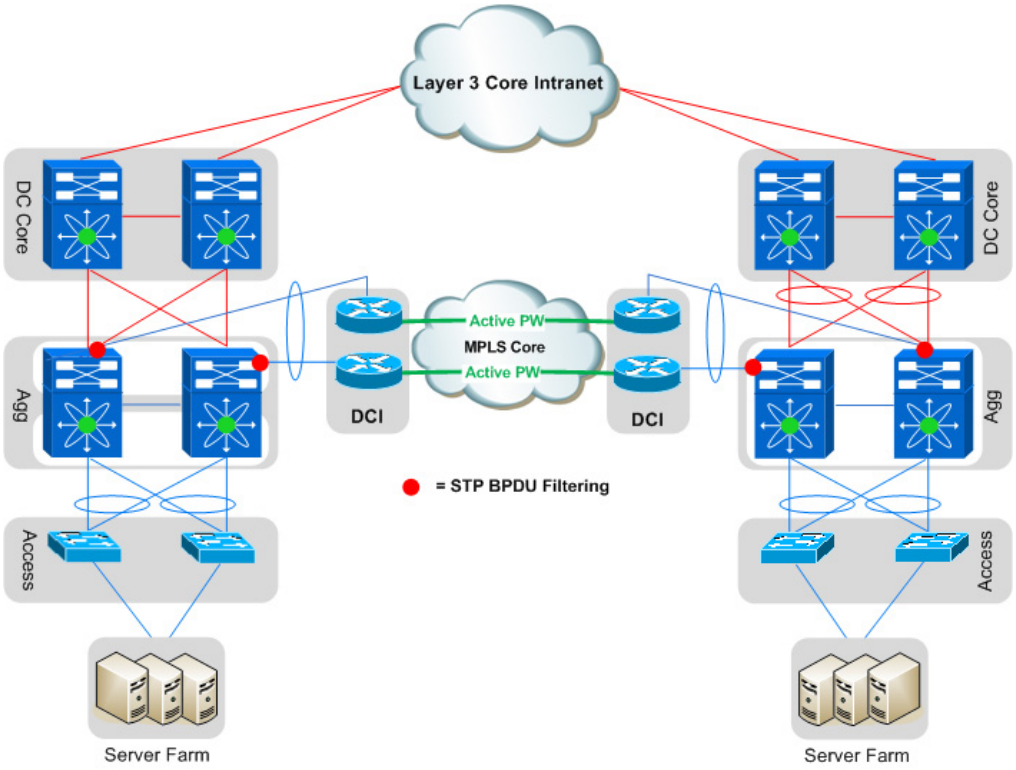
EoMPLS Platforms positioning

Two different Cisco platforms were validated for the role of DCI devices: the first one is the ASR1000, the second is the Catalyst 6500. Notice that both platforms support the EoMPLS port mode functionality natively (Catalyst 6500 requires the use of Sup720-3B supervisor and higher). However, while H-QoS and encryption capabilities are natively supported on ASR1000 platforms, they require the deployment of SIP linecards with the Catalyst 6500 platform.

EoMPLS STP Isolation

The isolation between the STP domains defined in the two remote locations is achieved by preventing STP BPDUs from being sent across the MPLS core over the EoMPLS PWs. As shown in Figure 1-21, this can be accomplished by enabling STP BPDU filtering on the L2 trunk interfaces connecting the aggregation layer switches to the DCI devices.

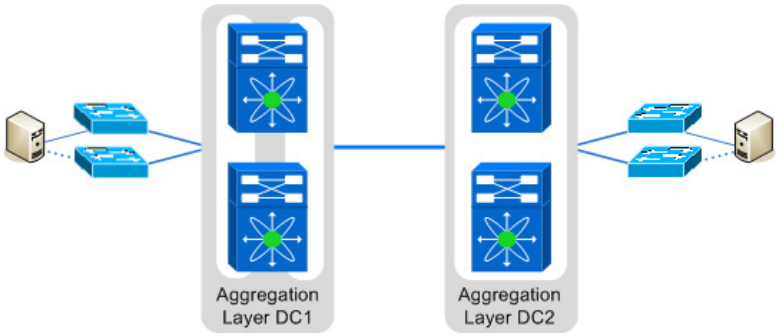
Figure 1-21 STP Isolation Using BPDU Filtering



EoMPLS End-to-End Loop Prevention

Despite the fact that STP BPDUs are not sent across the MPLS core regarding the filtering functionality discussed, the establishment of an EtherChannel between the remote aggregation layer devices ensure the protection against STP loops. As shown in Figure 1-22, the bridged topology from a STP perspective is loopless, given the use of multi-chassis EtherChannel functionality (vPC and VSS).

Figure 1-22 End-to-End Loop Prevention



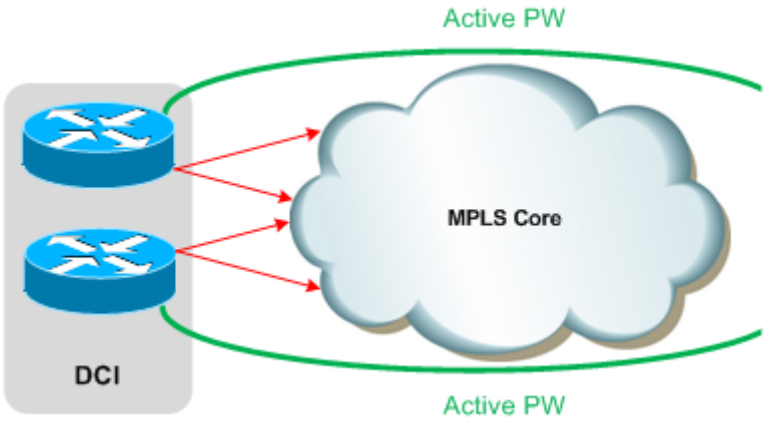
It is recommended to keep STP enabled even in a loop-less topology, as a safety mechanism in case a loop is created by mis-cabling or misconfiguration.

EoMPLS Load Balancing

When discussing traffic load-balancing, it is required to distinguish what happens inside each data center site and on the MAN/WAN cloud.

- Inside each site, from the aggregation layer devices perspective, a single logical L2 trunk (port-channel) is established with the aggregation layer in the second data center, so all the VLANs are allowed on that logical path. This holds true for the aggregation to access layer connections. This implies that traffic can be load-balanced on the various physical links bundled in these logical multi-chassis EtherChannels based on the hashing performed on L2/L3/L4 packet information. Under the assumptions that there are a large number of flows established between devices belonging to the same extended VLAN, it is expected that traffic will be evenly balanced across all physically available paths.
- Once the traffic leaves each data center, it is MPLS tagged (actually double tagged given the way EoMPLS works). Since all the VLANs are carried over the same EoMPLS PW (connecting each DCI device to its peer in the remote site), both Catalyst 6500 and ASR1000 devices would always select only one physical interface connecting to the MPLS core for sending the traffic out. That means that the scenario shown in Figure 1-23, where redundant physical links connect each DCI device to the MPLS core, the second path will remain idle and ready to be activated should the primary link fail.

Figure 1-23 Traffic Load Balancing Toward the MAN/WAN Cloud



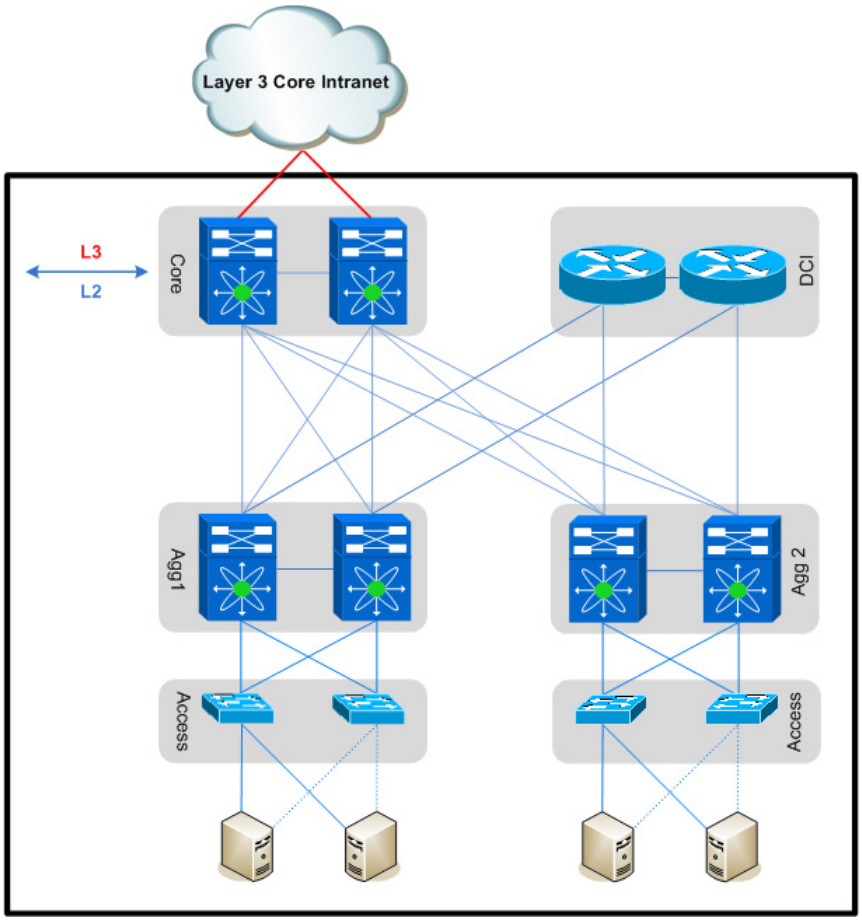
EoMPLS Inter-POD L2 Extension

In some cases there may be a requirement to extend L2 connectivity not only between remote data center sites, but also inside a single data center. This is usually referred to as an inter-POD LAN extension, where a POD is each specific data center building block (represented by the server, access and aggregation layers). Usually the demarcation line between Layer 2 and Layer 3 domains is placed at the aggregation layer of each defined POD.

The nature of the EoMPLS port mode functionality does not allow leveraging the L2 extension between each aggregation layer and the DCI layer to perform inter-POD bridging because all traffic received from the aggregation layer devices is tunneled across the PW toward the remote site.

Consequently, the recommended way to stretch VLANs between PODs in this case is by moving the boundary between L2 and L3 from the aggregation layer to the core, as shown in Figure 1-24.

Figure 1-24 Moving the L2-L3 Demarcation Line to the Core Layer

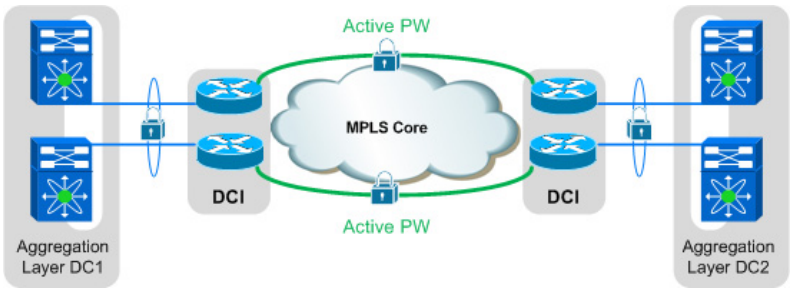


When deploying the approach proposed in Figure 1-24 you should consider the impact of the increase in size of the broadcast domain and the potential of creating STP loops. The second point can be addressed by deploying multi-chassis EtherChannel connections (vPC or VSS based) between the access, aggregation, and core layers.

EoMPLS Encryption

The main issue to consider for encrypting traffic across a DCI connection is that the traditional approach based on IPSec cannot be applied to MPLS traffic. In the context of the EoMPLS solution, this means you cannot apply encryption at the DCI layer to the packets sent toward the MAN/WAN cloud. However, the nature of the EoMPLS port mode functionality allows leveraging the 802.1AE standard (MACSec) because a direct 802.1AE session can be established between devices deployed at the aggregation layer of the two data center sites, as shown in Figure 1-25.

Figure 1-25 802.1AE over EoMPLS



The only platforms supporting 802.1AE in hardware is the Nexus 7000. This implies that the solution depicted above can only be implemented when Nexus 7000s are deployed in the aggregation layer of both data centers.

EoMPLS HQoS

The requirement for Hierarchical Quality of Service (HQoS) arises in cases where enterprise DCI devices connect to a MAN/WAN cloud managed by a service provider offering sub-rate service, typical in small to medium deployments where sub-rate GE connectivity is available to the enterprise.

Using HQoS in port-based EoMPLS deployment depends on two main factors:

- **Specific platform deployed in the DCI layer:** As mentioned, EoMPLS port-mode is supported natively with 6500 platforms. This, however, is not the case for HQoS capability which is limited to SIP modules (SIP-400 and SIP-600). We recommend that SIP support be added every time HQoS becomes a mandatory requirement for this solution.

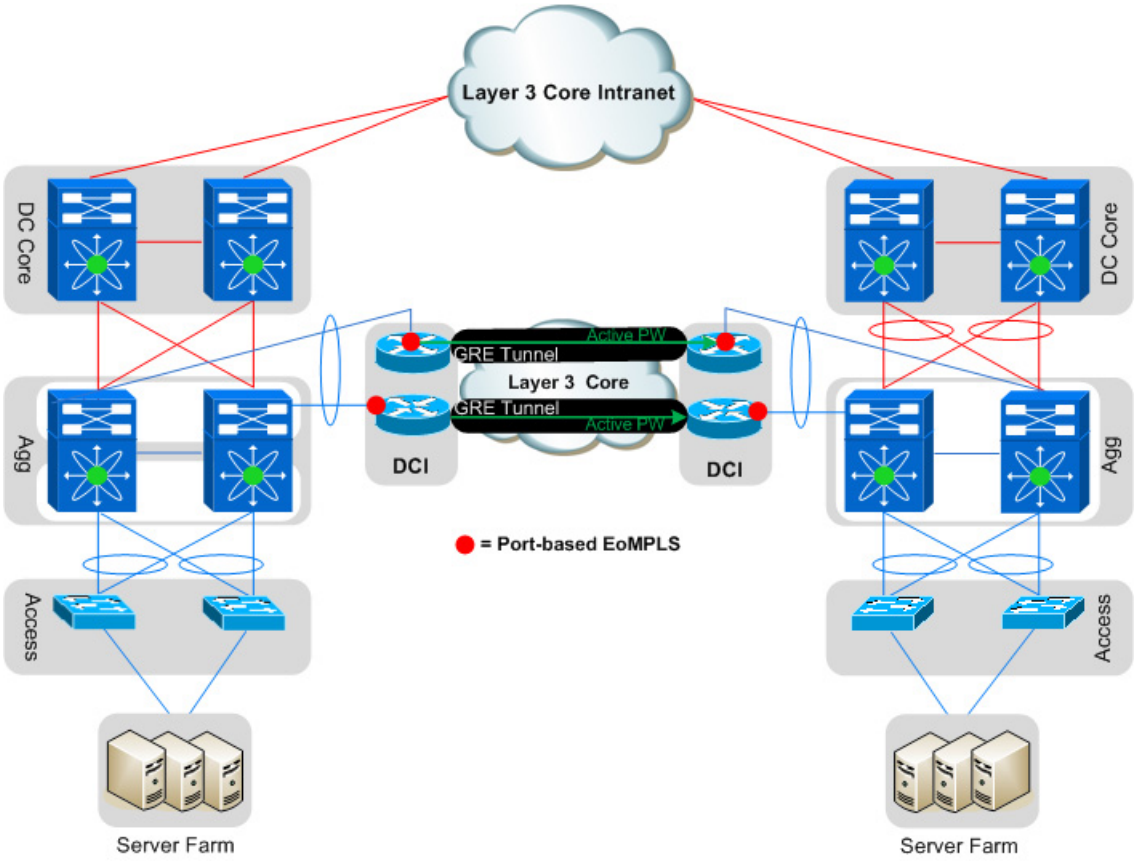
Note HQoS is natively supported on Cisco ASR1000 routers.

- **802.1AE encryption between aggregation boxes:** Figure 1-17 highlights how the 802.1Q header is also encrypted when deploying 802.1AE, which basically does not allow visibility into any QoS marking. Consequently, all traffic is assigned to the default queue and with no prioritization. However, you can sill create a hierarchical policy to ensure a sub-rate amount of traffic is sent to the provider.

EoMPLSoGRE Port Mode

The deployment of port-mode EoMPLSoGRE (Figure 1-26) represents a solution that provides LAN extension services between two remote data center locations in all scenarios where the enterprise is offered MAN/WAN Layer 3 services from a SP. Under these circumstances you can not deploy EoMPLS in native form, and is required to leverage a logical overlay connectivity offered, for example, by GRE tunnels.

Figure 1-26 EoMPLSoGRE Port Mode Deployment



When positioning this solution, remember the following points:

- Given the nature of the EoMPLS technology, this solution is suited for point-to-point deployments. Multi-point deployments (where the LAN extension functionality needs to be provided across more than two remote locations) are covered leveraging VPLS services, still deployed on top of a logical GRE overlay.
- A single EoMPLS Pseudo Wire (PW) is establish from each DCI edge device leveraging a logical point-to-point GRE tunnel established with the peer device in the remote site. This facilitates the creation of two independent PWs between the remote locations that are always active and forwarding traffic. This is valid independent of the number of VLANs extended across the two locations and is the result of the EoMPLS port mode functionality. Extension for up to 1200 VLANs was validated in phase 2 (refer to Chapter 2 convergence results for details).

Functionally, the solution is similar to the native port-mode EoMPLS scenario discussed in the previous section. The aggregation layer devices are transparently interconnected through the EoMPLSoGRE connections establishing end-to-end multi-chassis EtherChannels leveraging vPC or VSS (Figure 1-20).

Given the similarities between the port mode EoMPLSoGRE deployment and native port-mode EoMPLS, refer to EoMPLS Port Mode, page 1-23 regarding the way STP isolation, end-to-end loop prevention, and inter-PODs L2 extension requirements are satisfied. The following sections highlight considerations unique to the EoMPLSoGRE solution.

EoMPLSoGRE Platforms Positioning

Native support for EoMPLSoGRE is not available on the Catalyst 6500 and requires the use of a SIP-400 card facing the MAN/WAN cloud. The ASR1000 is the only validated platform that performs native point-to-point EoMPLSoGRE LAN extension functionality.

EoMPLSoGRE Load Balancing

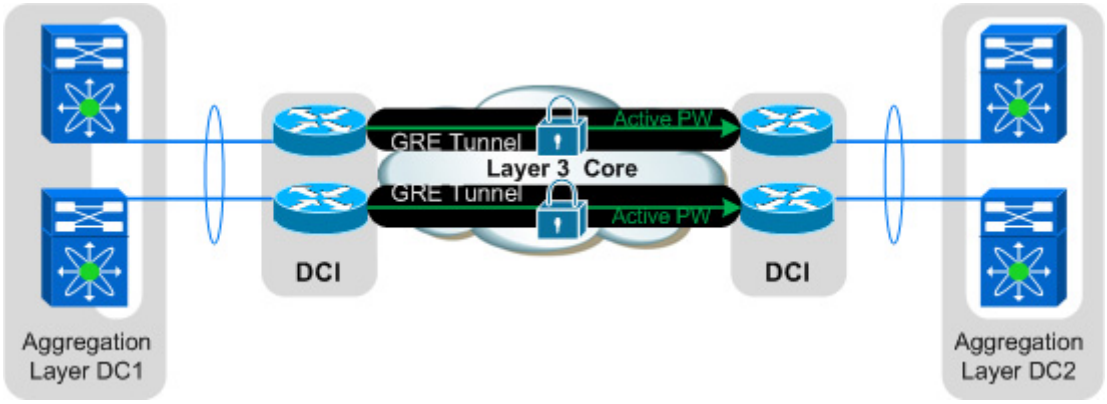
Load balancing inside each data center is achieved in the same manner as native EoMPLS. Different considerations need to be made for traffic leaving each data center, given that now it is not MPLS tagged traffic, but regular IP packets GRE encapsulates. All this traffic is destined through the GRE tunnels (usually a loopback interface defined on the remote DCI device) and sourced from a loopback interface defined on the local DCI device.

The capability of load-balancing outgoing traffic across separate physical links connecting each DCI device to the core (Figure 1-23) depends on the platform specific hashing mechanism. For example, platforms that hash only on the basis of source and destination IP addresses will always pick the same physical link for outgoing traffic.

EoMPLSoGRE Encryption

Establishing EoMPLS sessions over GRE tunnels facilitates traditional IPSec traffic encryption, as shown in Figure 1-27.

Figure 1-27 EoMPLSoGREoIPSec Session



This is an EoMPLSoGREoIPSec solution, given that IPSec encryption is applied to the GRE encapsulated MPLS packets. Figure 1-27 highlights the deployment of IPSec on the DCI layer devices. This is required when 802.1AE capable devices are not deployed in the aggregation layer of the remote interconnected data centers.

Note 802.1AE in HW is currently supported only with Nexus 7000 platforms.

EoMPLSoGRE HQoS

Leveraging IPSec at the DCI layer to encrypt the LAN extension traffic between remote locations implies that DCI devices have visibility into the QoS marking for all packets received from the aggregation layer. This means that both shaping and queuing functionality can be applied to the traffic.

Multi-Site Deployments

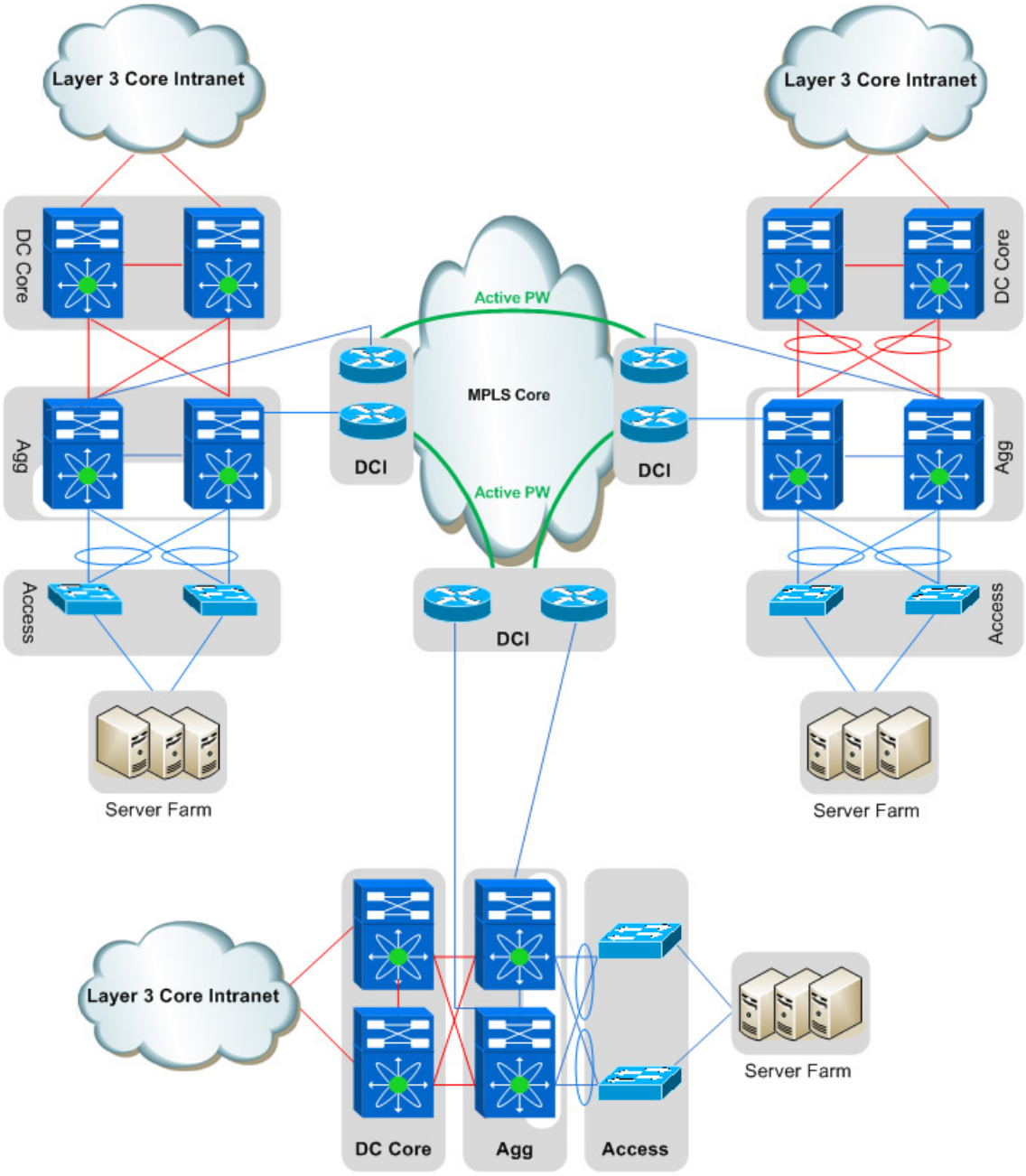
Multi-site technology deployment recommendations include the following:

- VPLS, page 1-32

VPLS

Both the EoMPLS and EoMPLSoGRE scenarios are applicable every time there is a requirement to provide LAN extension services between two point-to-point locations. An enterprise often needs to extend this functionality to three or more sites. In this case, Virtual Private LAN Service (VPLS) provides the technical solution, as shown in Figure 1-28.

Figure 1-28 VPLS Deployment



Some characteristics of a VPLS deployment are highlighted below:

- VPLS defines an architecture that delivers Ethernet multipoint services over an MPLS network. The underlying assumption is that the MAN/WAN cloud interconnecting the data center is MPLS enabled. This is usually the case if the enterprise owns the cloud. In cases where the WAN connectivity services are offered by a SP instead, the VPLS Pseudo Wires can actually be established only if the provider offers L1 or L2 connectivity services to the enterprise. In cases where the SP offers a L3 service (that is, the enterprise edge device peers at L3 with the SP edge device), the use of VPLS between the DCI devices is usually not an option, unless the functionality is enabled on a GRE overlay.

Note In the rest of this section, the terms DCI devices, or N-PE devices, will be used interchangeably. N-PE are network devices deployed in the DCI layer that perform VPLS functions.

- At a basic level, VPLS is a group of virtual switch instances (VSI) that are interconnected by using Ethernet over MPLS (EoMPLS) circuits (Pseudo Wires) in a full-mesh topology to form a single logical bridge. This logical bridge needs to operate like a conventional Layer 2 switch, offering functionality like flooding/forwarding, address learning/aging, and loop prevention.
- A separate Virtual Forwarding Instance (VFI) identifies a group of PWs associated with a VSI. A VFI is created for each VLAN that needs to be extended between remote locations. To minimize the operational costs of such a configuration the maximum number of VLANs validated (and recommended) with VPLS is 300. For cases requiring a higher number of VLANs, the use of QinQ to support the Hierarchical VPLS (H-VPLS) technology may be required.

Note An in-depth discussion of VPLS and H-VPLS is not in the scope of this document. Refer to the following link for more information:
http://www.cisco.com/en/US/products/ps6648/products_ios_protocol_option_home.html

The following sections discuss how this VPLS solution satisfies all requirements of a solid DCI solution.

VPLS Platforms Positioning

The only platform that has been validated for VPLS in this phase is the Catalyst 6500. Since the Catalyst 6500 does not support native VPLS on the Supervisor 720, the SIP line cards (SIP-400 or SIP-600) are introduced as part of the solution. The network processors available on these line cards perform VPLS (and VPLSoGRE) functions.

Note VPLSoGRE deployments are out of the scope of this document.

VPLS STP Isolation

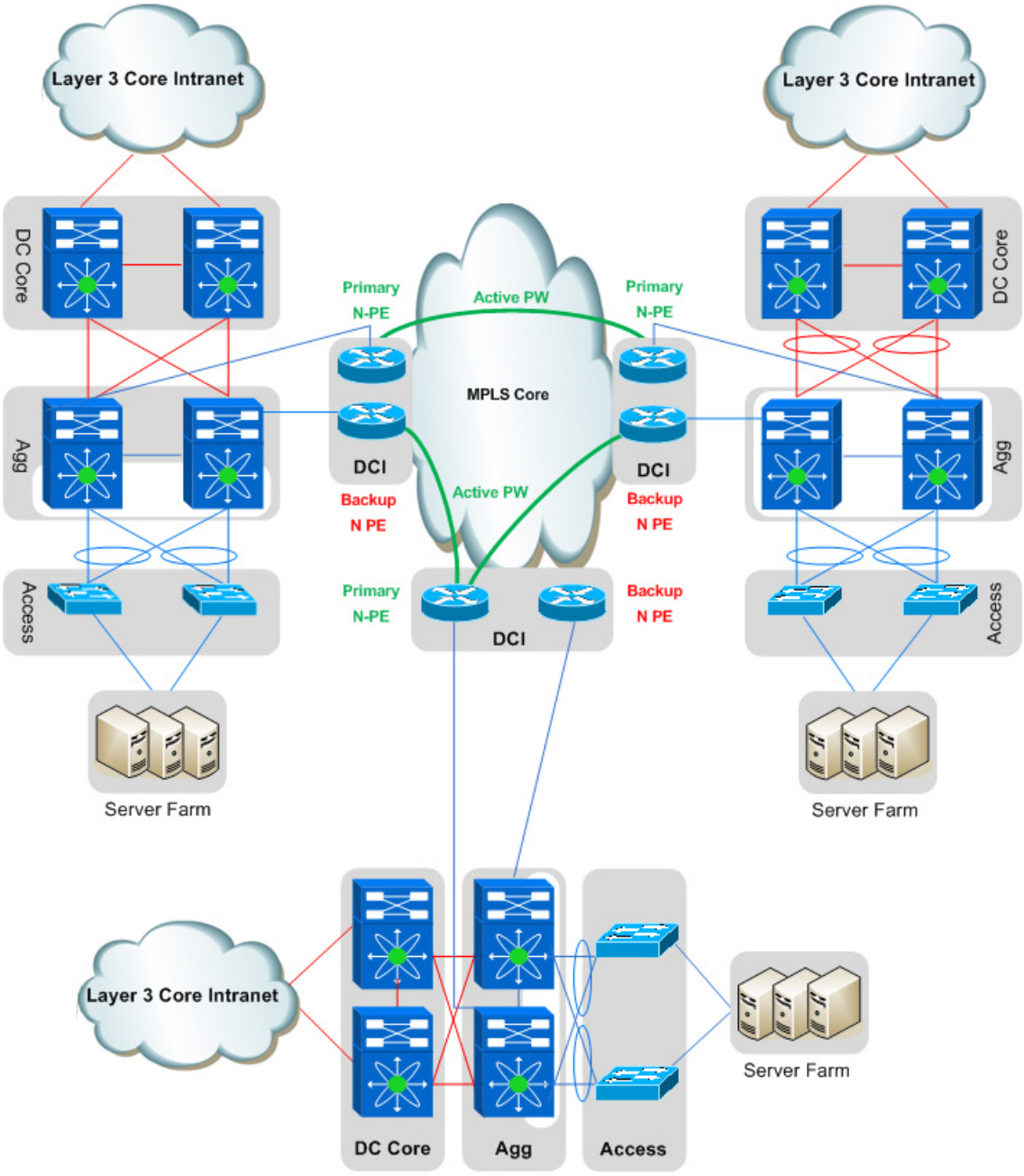
Isolation between the STP domains defined in the two remote locations is achieved given that no BPDUs are allowed across VPLS PW, by default, because VPLS is natively built with an internal mechanism known as split horizon, so the core network does not require STP to prevent L2 loops.

VPLS End-to-End Loop Prevention

The VPLS split horizon functionality is useful to prevent traffic from looping between PWs, but only in single-homed deployments (that is, where a single N-PE is deployed in each site to perform the VPLS functions). To increase the high availability of the solution, it is common to dual home each data center site to the MPLS core deploying two N-PE devices in the DCI layer of each location (Figure 1-28). This may result in an end-to-end STP loop being created if traffic traverses the link between the dual N-PEs, or via the connected aggregation layer devices.

The solution proposed in this design guide for addressing the dual-home N-PE issue leverages the use of Embedded Event Manager (EEM) to ensure, that at any given time, and for each extended VLAN, only one of the N-PEs deployed at each location is active and allowing traffic to flow between the aggregation layer devices and the VPLS enabled cloud (and vice versa), as shown in Figure 1-29.

Figure 1-29 Primary and Backup N-PE Deployment



Traffic flows across the VPLS core only on PWs established between Primary N-PEs. Only in cases where a primary N-PE can no longer perform forwarding functions, because it failed or because its connections to the MPLS core or the aggregation layer failed, will the Standby N-PE be activated to take on forwarding responsibilities.

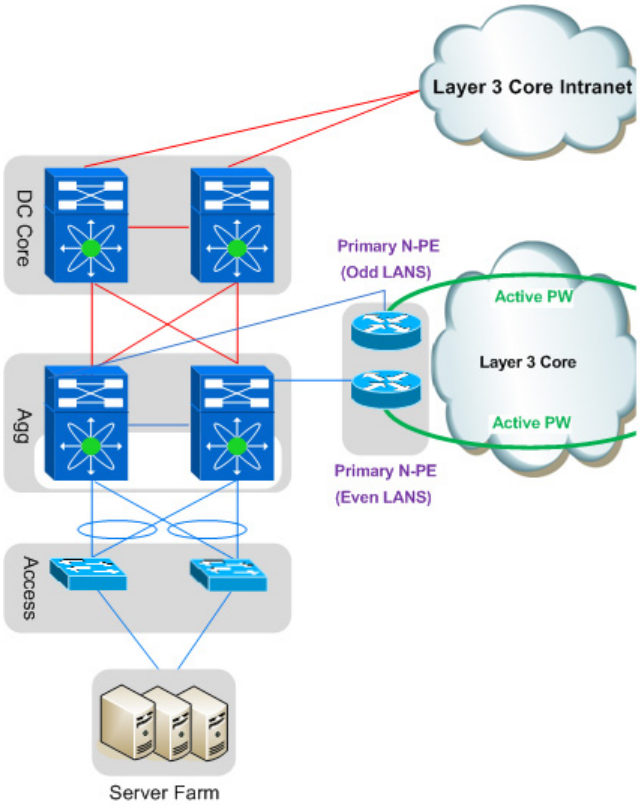
The EEM functionality proposed in this document is deployed at the aggregation layer (and not at the DCI layer) which allows positioning this solution independently from the platform deployed in the DCI layer. We discuss a specific deployment in Chapter 2 leveraging the Catalyst 6500 as PE devices, but the exact same solution may be used with ASR9000 or CRS-1 platforms, a user choice dependent upon the size of their deployment.

VPLS Load Balancing

Like the EoMPLS scenario, traffic load-balancing requires that we distinguish what happens inside each data center site and on the MAN/WAN cloud.

- Inside each site, EEM functionality is deployed to ensure that one of the two N-PEs gains an active role of forwarding traffic to and from the VPLS core. Since this active role can be taken for each VLAN, you can deploy a load-balancing scheme where half the VLANs extended between sites are owned by N-PE1, and the other half are owned by N-PE2 (Figure 1-30).

Figure 1-30 Load Balancing Traffic on a Per VLAN Basis



One N-PE may become the primary device for a set of VLANs (odd VLANs), whereas the other N-PE gains the role of active device for the even set of VLANs. The two N-PEs are also backing each other up in case a box or link fails.

Note The use of odd and even VLAN sets simplifies our understanding. In real world deployments, a VLAN range (for example, VLAN 10 to 100 and from 101 to 200) is used when actually configuring hundreds of VLANs.

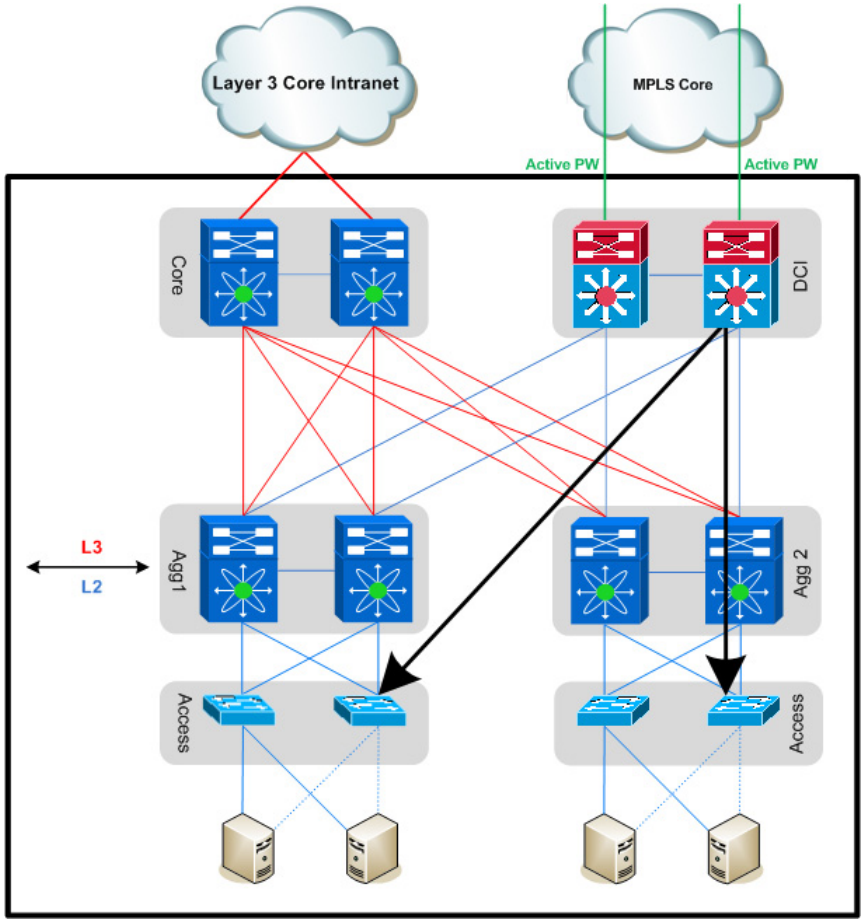
Once traffic leaves each data center, it is MPLS tagged (actually double tagged with VPLS). A separate internal MPLS tag is associated to each extended VLAN. Since the choice of the physical link to use to send the traffic toward the core depends on this MPLS tag, the consequence is that traffic can be load-balanced on a per VLAN basis.

VPLS Inter-POD L2 Extension

In scenarios where we extend L2 connectivity between PODs belonging to the same data center, the advantage of deploying VPLS (compared for example to port-based EoMPLS) is that N-PE devices retain local switching capabilities.

In Figure 1-31 we bridge traffic across the VPLS core (if the destination MAC address of the frame is learned from a remote location) and simultaneously bridge between local interfaces when the destination MAC address is located in the same site of the source. This preserves routing functionality for the VLANs at the aggregation layer where traffic originating from these VLANs is routed toward the data center core. Routed links are used to interconnect the aggregation and core layers if the destination is reachable via the Layer 3 cloud, or bridged toward the DCI layer devices if the destination belongs to the same VLAN (independent of site location).

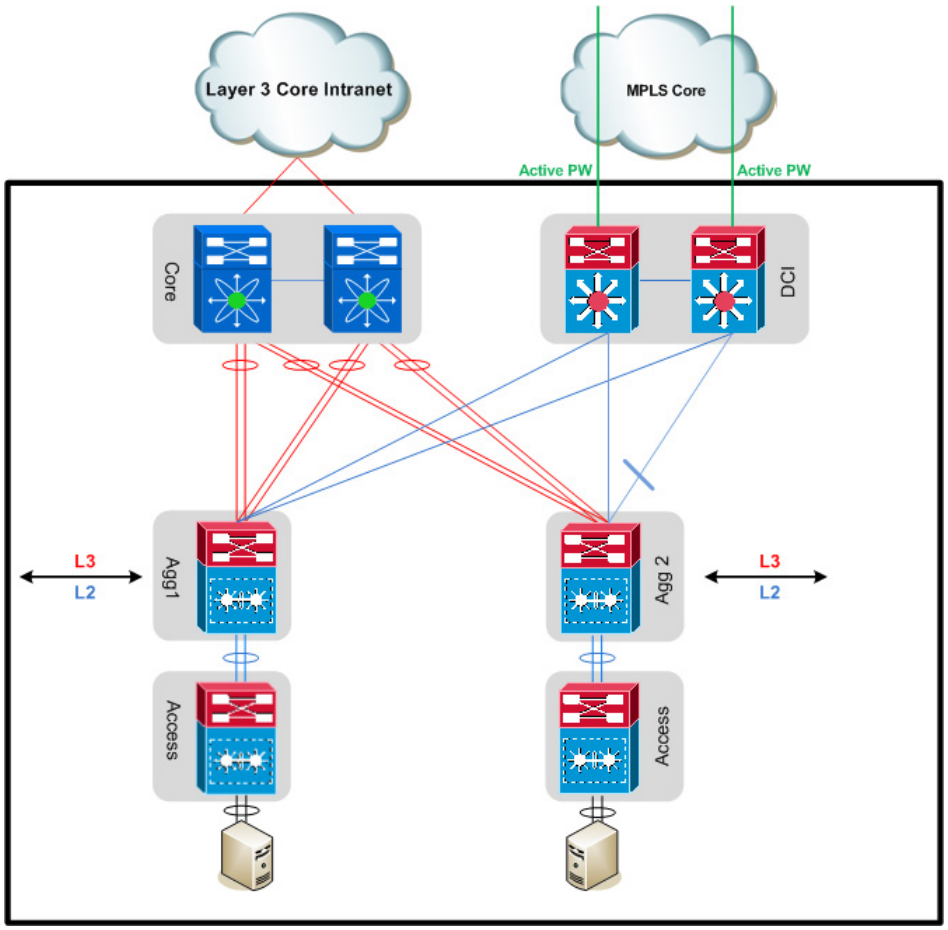
Figure 1-31 Local Switching Performed on DCI Devices



However, a couple of considerations need to be made when deploying this option. First, the local switching functionality performed by the DCI layer devices causes the creation of a larger broadcast domain extending between different PODs of the same DC. Rate-limiting configuration is then recommended on the aggregation devices to minimize the impact of broadcast storms. Second, we already mentioned how the use of Multi-Chassis EtherChannel (MCEC) technologies are recommended inside the data center to build loop-less spanning-tree topologies. VPLS is not supported on platforms

offering MCEC, so be mindful that the inter-POD LAN extension requirement may lead to the creation of a STP loop in each local topology, as shown in Figure 1-32 (in this example VSS is used as an example of MCEC technology).

Figure 1-32 STP Loop Between Aggregation and DCI Layer Devices



We recommended keeping LAN extensions inside the same POD or between PODs belonging to separate data center locations.

VPLS Encryption

A native VPLS deployment does not offer encryption capabilities since IPSec cannot be applied to MPLS-labeled traffic, and 802.1AE encryption can only be performed between devices connected back-to-back (or via a technology that is transparent to Layer 2 traffic, as was the case for port mode EoMPLS). To reintroduce encryption capabilities, you must deploy VPLS over a GRE overlay.

Note VPLSoGRE deployments (with or without IPSec encryption) are out of scope for this document.

VPLS HQoS

HQoS requirements arise in cases where enterprise DCI devices connect to a MAN/WAN cloud managed by a SP offering a sub-rate service. This is typical in small to medium sized deployments where sub-rate GE connectivity is available to the enterprise.

HQoS is supported with SIP card models SIP-400 and SIP-600 required for the Catalyst 6500 platform to deploy VPLS, resulting in traffic and shaping (for sub-rate deployments) support for this solution.

CHAPTER 2

Cisco DCI System Solution Deployment

Cisco Data Center Interconnect (DCI) solutions extend network and SAN connectivity across geographically dispersed data centers. These solutions allow organizations to provide high-performance, non-stop access to business-critical applications and information. They support application clustering, as well as Virtual Machine (VM) mobility between data centers, to optimize computing efficiency and business continuance.

These solutions offer flexible connectivity options to address LAN, Layer 3 and SAN extension needs across optical (Dark Fiber/DWDM using VSS, vPC), Multiprotocol Label Switching (MPLS), or IP infrastructures. They also provide the following:

- Transparency to the WAN network architecture
- Support for multi-site data center architectures
- LAN and SAN extension, with per-data center fault isolation
- A resilient architecture that protects against link, node, and path failure
- Optional payload encryption
- SAN Write Acceleration and compression, to increase replication options

Cisco DCI solutions offer a scalable, flexible platform that satisfies requirements for server and application mobility and business continuance.

Release 2.0 of the solution focuses on the Layer 2 extension of data centers across geographically dispersed data centers using MPLS based technologies. The technology of choice usually depends on the specific topology: for point-to-point deployments, EoMPLS (or EoMPLSoGRE over an IP infrastructure) represents an easy and efficient mechanism to provide LAN extension services. For multipoint deployments, VPLS represents a more suitable choice, given the inherent multipoint characteristics of this technology.

The following sections of this document highlights the deployment considerations for these MPLS based technologies, starting with point-to-point topologies and diving then into details of multipoint designs.

Note A basic explanation of EoMPLS and VPLS features is beyond the scope of this document.

This chapter contains the following sections:

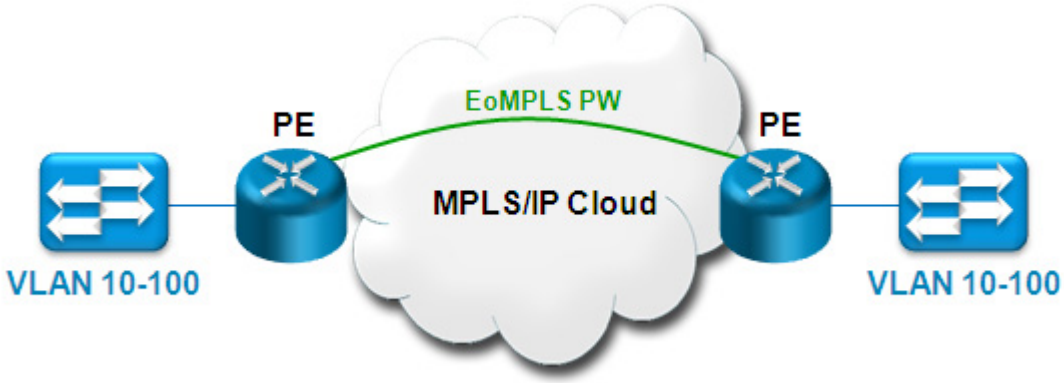
- Point-to-Point Topologies , page 2-2
- Multipoint Topologies, page 2-51
- Summary of Design Recommendations, page 2-96

Point-to-Point Topologies

The deployment of port-mode EoMPLS represents a simple way to provide LAN extension services between two remote data center locations. The basic idea behind this technology is to leverage EoMPLS connections (usually named Pseudowires – PWs) to logically "extend" local physical connections across a Layer 3 cloud.

Figure 2-1 shows the logical PWs are established between network devices performing the role of Provider Edge (PE): this naming convention highlights the service provider origin of these MPLS technologies, and it is usually maintained even when the PE devices are actually owned and managed by an Enterprise IT staff.

Figure 2-1 EoMPLS for LAN Extension



Note The terms "PE" or "N-PE" are used interchangeably in the context of this document, as they reference the same device providing EoMPLS PW services.

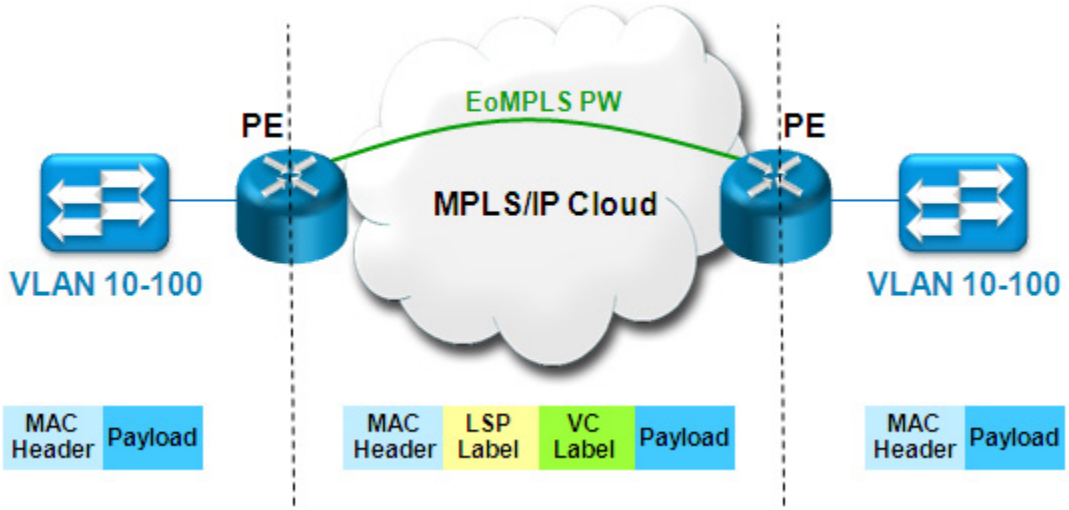
In Figure 2-1shows how the same EoMPLS PW can be used to logically extend a physical Layer 2 trunk connection carrying multiple VLANs. EoMPLS encapsulates Ethernet frames in MPLS packets and forwards them across the MPLS network. Each frame is transported as a single packet, and the PE routers connected to the backbone add and remove labels as appropriate for packet encapsulation. This is the sequence describing the forwarding of EoMPLS frames across the MPLS cloud:

1. The ingress PE router receives an Ethernet frame belonging to VLAN 10 and encapsulates the packet by removing the preamble, the start of frame delimiter (SFD), and the frame check sequence (FCS). The rest of the packet header is not changed.
2. The ingress PE router adds a point-to-point virtual connection (VC) label and a label switched path (LSP) tunnel label for normal MPLS routing through the MPLS backbone. The VC label is negotiated with the remote PE device by leveraging a targeted LDP session, whereas the LSP tunnel label is negotiated with the neighbor devices belonging to the MPLS cloud.
3. The network core routers use the LSP tunnel label to move the packet through the MPLS backbone and do not distinguish Ethernet traffic from any other types of packets in the MPLS backbone.
4. At the other end of the MPLS backbone, the egress PE router receives the packet and de-encapsulates the packet by removing the LSP tunnel label if one is present. The PE router leverages the information in the VC label to determine the interface (attachment circuit) where to switch the frame.

5. The PE removes the VC label from the packet, updates the header, if necessary, and sends the packet out the appropriate interface to the destination switch.

Based on the description above, it is clear how the MPLS devices belonging to the MPLS cloud (P routers) leverage the tunnel labels to transport the packet between the PE routers. The egress PE router uses the VC label to select the outgoing interface for the Ethernet packet.

Figure 2-2 EoMPLS Packet Encapsulation



The same VC label is used for extending all VLANs defined on edge devices above and is maintained, unchanged, across the MPLS cloud. On the other hand, the LSP label is modified at each hop across the MPLS core.

Note EoMPLS PWs are unidirectional by nature. This means that to establish two-way communications across the MPLS cloud, you must create a logical PW in each direction.

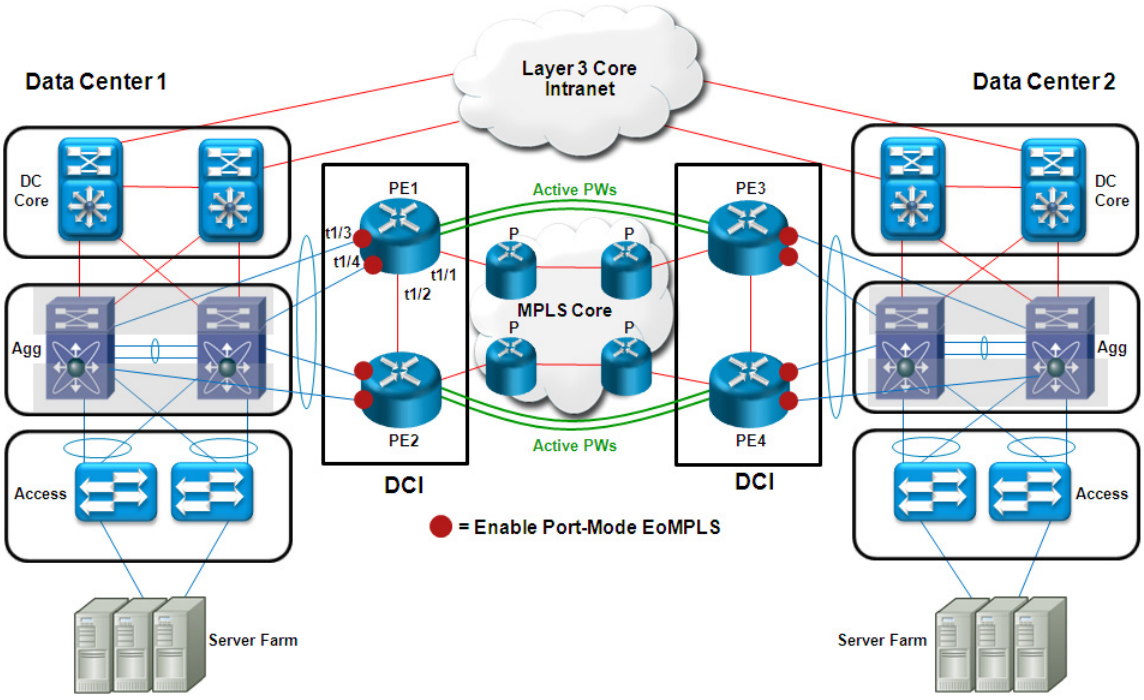
Depending on the nature of the Layer 3 cloud interconnecting the remote sites, two different flavors of EoMPLS can be deployed. If the Layer 3 cloud is MPLS enabled, it is possible to leverage the native EoMPLS functionality. This is usually the case if a given enterprise owns the Layer 3 cloud. On the other side, if the enterprise buys an IP service from a service provider (that is, the enterprise PE device is peering at Layer 3 with the SP device), then the EoMPLS connections can be established across an overlay connection, usually represented by a GRE tunnel. This type of deployment is named EoMPLSoGRE.

The following sections discuss more in detail the deployment aspects of these point-to-point technologies starting with native EoMPLS.

Deploying Port Based EoMPLS

Figure 2-3 shows the network topology that was used to validate EoMPLS as part of DCI phase 2.

Figure 2-3 Validated EoMPLS Topology

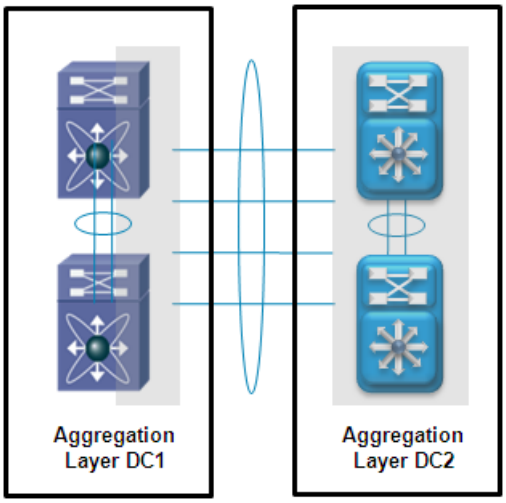


Before we discuss in detail the deployment of EoMPLS, it is important to highlight some characteristics of the proposed design.

- The PE devices performing EoMPLS functions are deployed in each data center in a dedicated DCI layer. The VLANs that need to be extended to a remote data center sites are carried to the DCI layer from the aggregation layer by leveraging Layer 2 trunk connections.
- Each aggregation layer device connects to both the PEs deployed in the DCI layer in a fully meshed fashion. This is recommended to improve the convergence under a couple of node/link recovery scenarios, as it will be discussed more in detail in the "EoMPLS Failure/Recovery Scenarios" section. A specific topology without the full mesh of connections (that is,. having a single Layer 2 trunk connecting each aggregation switch to the corresponding PE in the DCI layer) was also validated to compare the recovery/failure test results.
- The specific type of EoMPLS deployment discussed in the context of this document is called "port-based". In this mode, all the Ethernet frames received on a physical internal PE interface (or sub-interface) are EoMPLS encapsulated and tunneled across the MPLS cloud toward the remote egress PE device. Therefore, the aggregation layer switches appear as if they were connected back-to-back both from a data plane and control plane perspective; this enables for example the exchange of Link Aggregation Control Protocol (LACP) messages, establishing of end-to-end EtherChannels as shown in Figure 2-4.

Note The devices deployed at the aggregation layer of each data center must support some sort of Multi Chassis EtherChannel (MCEC) functionality. The two type of MCEC features that were validated for this design are VSS (supported on Catalyst 6500 switches) and vPC (available on Nexus 7000 devices). Figure 2-4 shows VSS and vPC technologies can inter operate, allowing a logically interconnected data center with different types of network devices deployed in the aggregation layer.

Figure 2-4 Establishment of Back-to-Back EtherChannel Connection



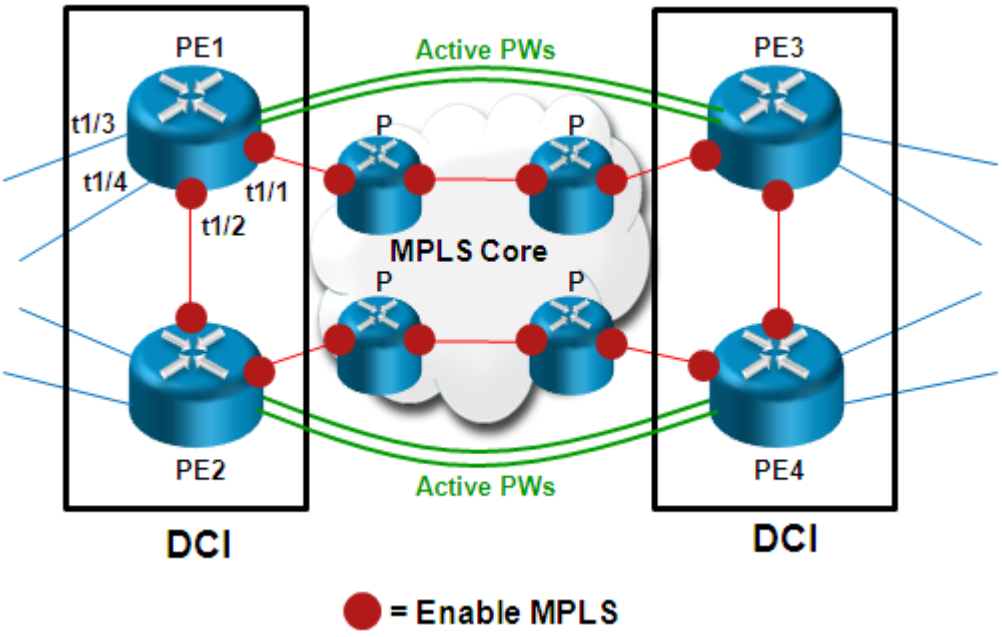
- A logical PW (in each direction, given its unidirectional nature) is associated to each physical link configured in port-mode. This because the PW really represents a logical extension of the physical connection toward the aggregation layer. A pair of logical PWs is then established between PE devices in remote sites when the aggregation layer switches are physically meshed to the DCI layer (as is the case in Figure 2-4).
- The physical paths established between PEs across the MPLS cloud are independent from each other. This was done to simulate a scenario where an enterprise may leverage different physical infrastructures (or SP clouds) to establish the PWs logical connections. A Layer 3 link is recommended between the two PEs deployed in the same data center location, to ensure traffic can be rerouted across it in case all physical paths across the core should fail for one of the PE routers (more considerations for this can be found in the "EoMPLS Failure/Recovery Scenarios" section.
- Two different scenarios were validated for what concerns the Cisco platforms deployed at the DCI layer to perform EoMPLS features:
  - In the first topology, Cisco ASR1000 platforms were used as PEs. ASR1000 supports EoMPLS starting with IOS XE 2.4.0.
  - In the second topology, the PE devices were Catalyst 6500 platforms. Catalyst 6500 is capable of natively supporting the port mode EoMPLS functionality, without requiring the use of any SIP linecard. Port mode EoMPLS is available on Catalyst 6500 on all the IOS releases supporting Sup720-3B supervisors and higher.

EoMPLS Configuration

The MPLS configuration required to enable the EoMPLS services on the PE devices is fairly straightforward and can be summarized in the following steps:

Step 1 Enable MPLS on the Layer 3 links belonging to the PE. This needs to be done both on the physical interface connecting to the MPLS cloud and on the transit link to the peer PE device, as shown in Figure 2-5.

Figure 2-5 Enabling MPLS on Layer 3 Interfaces



The following are required CLI commands:

PE1

```
mpls ldp graceful-restart
mpls ldp session protection
mpls ldp holdtime 15
mpls label protocol ldp
mpls ldp router-id Loopback0 force
!
interface Loopback0
 description OSPF and LDP Router_ID
 ip address 15.0.4.1 255.255.255.255
!
interface TenGigabitEthernet1/1
 description Link to MPLS Core
 mtu 9216
 ip address 50.0.11.1 255.255.255.0
 mpls ip
!
interface TenGigabitEthernet1/2
 description Transit Link to Peer PE
 mtu 9216
 ip address 50.0.54.1 255.255.255.252
 mpls ip
```

Turning on MPLS services on the Layer 3 links also enables LDP (Label Distribution Protocol) with the neighbor devices. LDP is required to exchange the MPLS label information required for packets that need to be sent toward the MPLS cloud. Also, we recommend increasing the MTU on the MPLS enabled interfaces. Every IP packet that needs to be sent across an EoMPLS PW requires two MPLS labels, which means that the configured MTU should be at least 8 bytes bigger than the largest IP packet that need to be sent across the cloud. When possible, configure the maximum MTU size of 9216, as shown in the preceding configuration example.

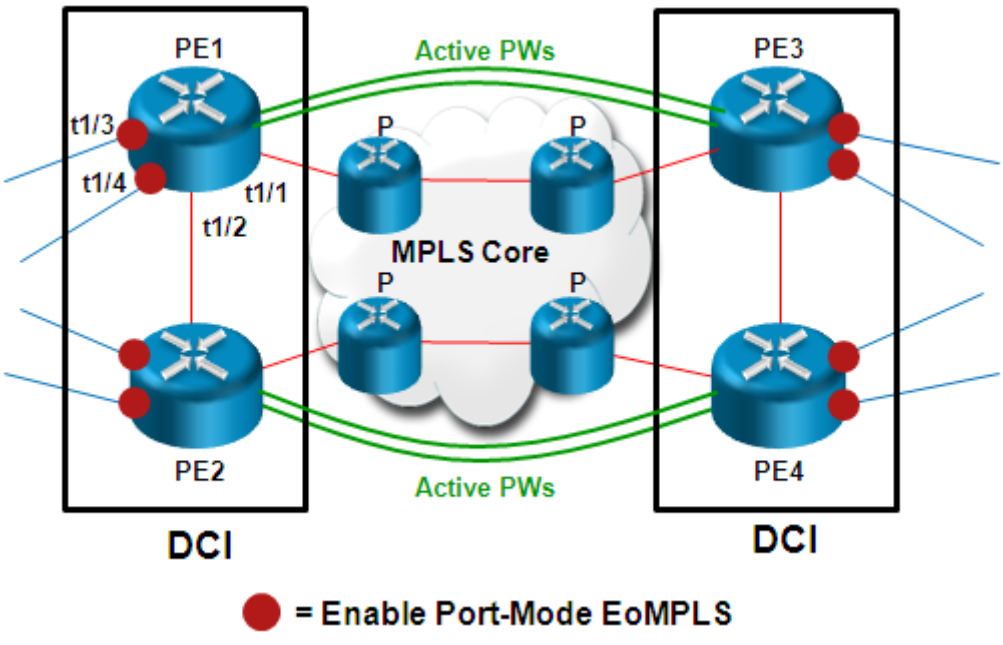
Note The MPLS configuration required on the Provider (P) devices deployed in the MPLS core is similar to the preceding one, since the only basic requirement is to enable MPLS on their Layer 3 interfaces.

As soon as MPLS is enabled on the physical Layer 3 interfaces, LDP neighbors are established, which you can verify with the command below:

```
PE1#sh mpls ldp neighbor
    Peer LDP Ident: 15.0.0.51:0; Local LDP Ident 15.0.4.1:0
        TCP connection: 15.0.0.51.646 - 15.0.4.1.60392
        State: Oper; Msgs sent/rcvd: 2540/2534; Downstream
        Up time: 1d12h
        LDP discovery sources:
          TenGigabitEthernet1/6, Src IP addr: 50.0.11.2
          Targeted Hello 15.0.4.1 -> 15.0.0.51, active, passive
        Addresses bound to peer LDP Ident:
          15.0.0.51       10.0.9.2        50.0.11.2       50.0.12.1
    Peer LDP Ident: 15.0.4.10:0; Local LDP Ident 15.0.4.1:0
        TCP connection: 15.0.4.10.58091 - 15.0.4.1.646
        State: Oper; Msgs sent/rcvd: 2540/2540; Downstream
        Up time: 1d12h
        LDP discovery sources:
          TenGigabitEthernet2/8, Src IP addr: 50.0.54.2
          Targeted Hello 15.0.4.1 -> 15.0.4.10, active, passive
        Addresses bound to peer LDP Ident:
          15.0.4.10       15.0.4.104      15.0.4.11       10.0.9.5
          50.0.30.9       50.0.54.2
```

Step 2 Configure EoMPLS port mode on the PE internal interface. As previously mentioned, the logical PW established with a remote PE allows extending the physical Layer 2 trunk connections originated from the aggregation layer device (Figure 2-6).

Figure 2-6 EoMPLS Port Mode Configuration



PE1

```
interface TenGigabitEthernet1/3
 description Link1 to Aggregation Layer
 mtu 9216
 no ip address
 xconnect 15.0.5.1 2504 encapsulation mpls
!
interface TenGigabitEthernet1/4
 description Link2 to Aggregation Layer
 mtu 9216
 no ip address
 xconnect 15.0.5.1 2515 encapsulation mpls
```

The following are two considerations based on the configuration sample above:

- The configuration required on the PE internal link is pretty minimal and limited to the configuration of the "xconnect" command to implement the port-based EoMPLS functionality. Also, it remains the same independently from the specific interface configuration on the aggregation side (that is, if it is a Layer 2 trunk, a Layer 2 access port or a routed link).
- Two parameters are specified as part of the **xconnect** command: the IP address of the remote PE where EoMPLS PW connection is terminated and the VC ID value identifying the PW. Use a remote loopback address as first parameter to ensure that the PW connection remains active as long as there is a physical path between the two PE devices. The VC ID (identifying the specific PW) must be unique for each internal interface to which the **xconnect** command is applied.

As previously mentioned, a separate PW is associated to each physical interface that connects the PE to the aggregation layer device (usually called "attachment circuit – ac"). Verify this in the following output:

```
PE1#sh xconnect all
Legend: XC ST=Xconnect State, S1=Segment1 State, S2=Segment2 State
UP=Up, DN=Down, AD=Admin Down, IA=Inactive, NH=No Hardware
XC ST  Segment 1                         S1 Segment 2                          S2
------+---------------------------------+--+---------------------------------+--
UP     ac   Te1/3(Ethernet)              UP mpls 15.0.5.1:2504                 UP
UP     ac   Te1/4(Ethernet)              UP mpls 15.0.5.1:2515                 UP
```

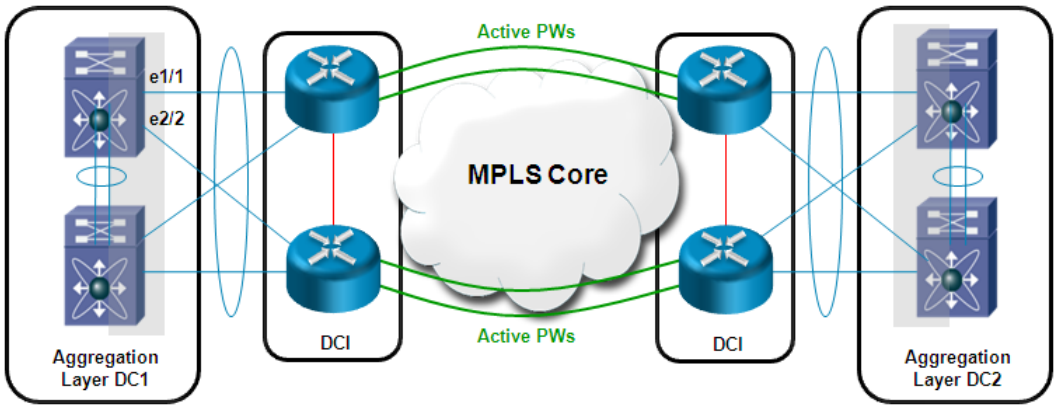
End-to-End Loop Prevention and STP Isolation

The combination of port-mode EoMPLS and Multi Chassis EtherChannel features is key to establishing back-to-back EtherChannels between aggregation layer devices deployed in geographically dispersed data centers. Creating the logical EtherChannel connection in Figure 2-4 is critical for two reasons:

1. Ensures that no end-to-end STP loop can be created. This would be a concern if the physical Layer 2 trunks were carried across the PW logical connections without being bundled together.
2. Isolates the STP domains between the two data center sites. This improves network stability and avoids the possibility that a broadcast storm impacting a given site may affect the other data center.

The basic idea that allows meeting both objectives consists in replacing STP with LACP (802.3ad) as control protocol between the aggregation layer devices belonging to the separate data centers. The required configuration is shown Figure 2-7:

Figure 2-7 Creation of Back-to-Back Port-Channels



DC1-N7K-Agg1

```
interface port-channel70
  description L2 PortChannel to DC 2
  switchport mode trunk
  vpc 70
  switchport trunk allowed vlan 1,1500-1519,1600-2799
  mtu 9216
!
interface Ethernet1/1
  description PortChannel Member
  switchport mode trunk
  switchport trunk allowed vlan 1,1500-1519,1600-2799
  mtu 9216
  channel-group 70 mode active
!
interface Ethernet2/2
  description PortChannel Member
  switchport mode trunk
  switchport trunk allowed vlan 1,1500-1519,1600-2799
  mtu 9216
  channel-group 70 mode active
```

Note The preceding configuration refers to a scenario where Nexus 7000 devices are deployed at the aggregation layer of both data centers, but can easily be extended to designs leveraging Catalyst 6500 deployed in VSS mode.

The following configuration demonstrates how the physical links are in fact bundled together to belong to a PortChannel.

```
DC1-Agg1# sh 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)
--------------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports
      Channel
--------------------------------------------------------------------------------
70    Po70(SU)    Eth      LACP      Eth1/1(P)    Eth2/2(P)
```

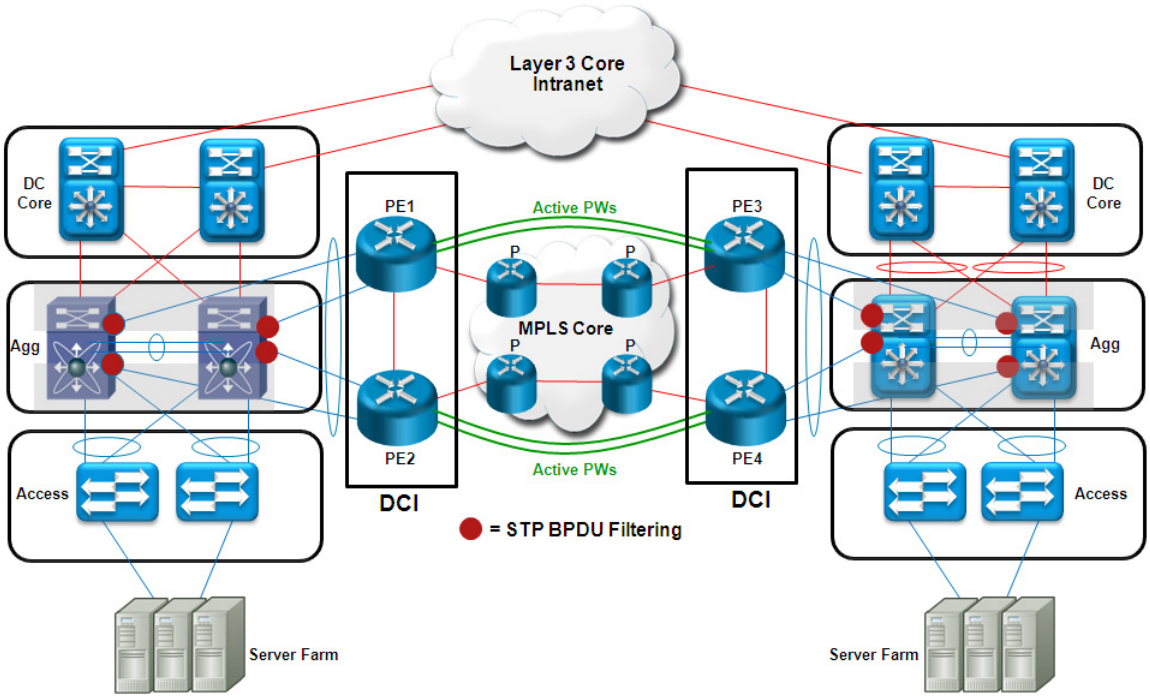
Also, as expected, the devices at the aggregation layer in the two sites appear to be connected back-to-back, and are for example able to become CDP neighbors:

```
DC1-Agg1# sh cdp neighbors
Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater,
                  V - VoIP-Phone, D - Remotely-Managed-Device,
                  s - Supports-STP-Dispute

Device-ID             Local Intrfce Hldtme Capability  Platform      Port ID
<snip>
DC2-Agg1      Eth1/1        176    R S I s     N7K-C7010     Eth1/1
DC2-Agg1      Eth2/2        176    R S I s     N7K-C7010     Eth2/2
```

Once you create the loopless topology of Figure 2-7, you must ensure that no STP messages are exchanged between the two data centers. To do so we recommend you apply BPDU filtering on the EtherChannel connecting the aggregation layer devices with the DCI routers (Figure 2-8).

Figure 2-8 Applying STP BPDU Filtering



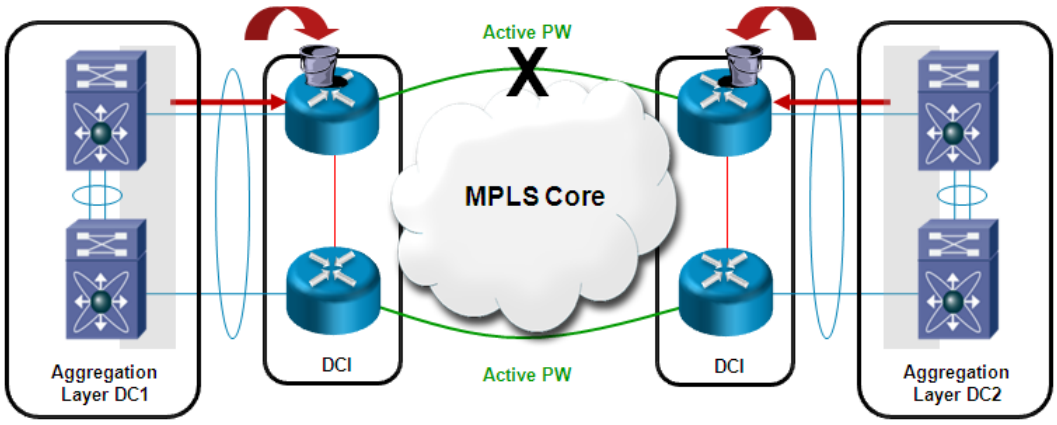
Notice that the BPDU filtering configuration must be applied on the logical port-channel interface, and not on the physical links. Also, together with the filtering, it is also critical to ensure that the PortChannel is deployed as an edge port from a Spanning Tree perspective, to avoid that a local STP convergence event that could cause the PortChannel link to temporarily go in blocking state.

DC1-N7K-Agg1

```
interface port-channel70
  description L2 PortChannel to DC 2
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
```

One of the biggest challenges in deploying the PortChannel across logical PWs (compared with the normal scenario with physical back-to-back connections) lies in dealing with PW failure scenarios. This is because it may happen that communication across the logical PW is compromised even if the physical links connecting the aggregation switches to the DCI layer are still up (from a physical point of view). Under these circumstances, it is hence required to take action in each data center to ensure that the aggregation layer devices stop sending traffic toward the PE device that established the PW logical connection. This is fundamental to avoid black-holing traffic between data centers, (Figure 2-9).

Figure 2-9 Traffic Black-Holing after PW Failure



There are various scenarios in which the connectivity across the logical PW may be jeopardized, as discussed in the "EoMPLS Failure/Recovery Scenarios" section. Depending on the specific platform deployed as PE device in the DCI layer, there are two different ways of handling these PW failure scenarios.

Remote Ethernet Port Shutdown (ASR1000 as PE)

ASR1000 platforms offer a feature, named Remote Ethernet Port Shutdown, that allows the PE router on the local end of an EoMPLS PW to shutdown its transmit signal (for optical link this happens with laser ON/OFF) to the interface configured for EoMPLS port-mode toward the aggregation layer devices once the PW failure is detected.

Therefore, because of this behavior, the state of the interface would be "down" only on the aggregation device, but would remain "up" on the PE.

Note Remote Ethernet Port Shutdown is enabled by default on ASR1000 routers and can be disabled using **no remote link failure notification** command in the xconnect sub-mode. For more information about this feature, please refer to the following link:
http://www.cisco.com/en/US/docs/ios/12_2sr/12_2srb/feature/guide/srbrpsdn.html

Embedded Event Manager – EEM (Catalyst 6500 as PE)

The Catalyst 6500 platform does not offer a feature similar to Remote Ethernet Port Shutdown. Therefore, a workaround is required to be able to achieve the same goal of minimizing the outage in scenarios where the logical PW interconnecting the data center sites should fail.

The solution validated in the context of this document leverages the Embedded Event Manager (EEM) functionality on Catalyst 6500 to monitor the state of the EoMPLS PW and to disable the local link (attachment circuit) toward the aggregation if the PW should fail.

The configuration is pretty simple: a specific event manager applet tracks the state of each defined PW. The recommended approach is to track the syslog message that is produced every time the PW fails, and bounce the corresponding link toward the aggregation layer once it is displayed. In the scenario shown in Figure 2-3 where two logical PWs are defined on each PE device, the required EEM configuration would be the following:

```
event manager applet EOMPLS_T1_1_PW_DOWN
 event syslog pattern "%XCONNECT-5-PW_STATUS: MPLS peer 15.0.5.1 vcid 2504, VC DOWN, VC
state DOWN"
 action 1.0 cli command "enable"
 action 2.0 cli command "conf t"
 action 4.0 cli command "int range ten1/1"
 action 4.1 cli command "shut"
 action 5.0 cli command "no shut"
!
event manager applet EOMPLS_T2_2_PW_DOWN
 event syslog pattern "%XCONNECT-5-PW_STATUS: MPLS peer 15.0.5.1 vcid 2515, VC DOWN, VC
state DOWN"
 action 1.0 cli command "enable"
 action 2.0 cli command "conf t"
 action 4.0 cli command "int range ten2/2"
 action 4.1 cli command "shut"
 action 5.0 cli command "no shut"
```

Notice that bringing the link right back (with the **no shut** command) is important to allow for a dynamic re-establishment of the connectivity as soon as the issue that caused the PW to fail is resolved. At that point LACP messages are again successfully exchanged across the end-to-end logical connection permitting to bundle the link back into the back-to-back PortChannel. Until that happens, that specific link will not be used to exchange traffic between the remote aggregation layer devices (despite that its state is “UP”), avoiding the black-holing of the traffic shown in Figure 2-9.

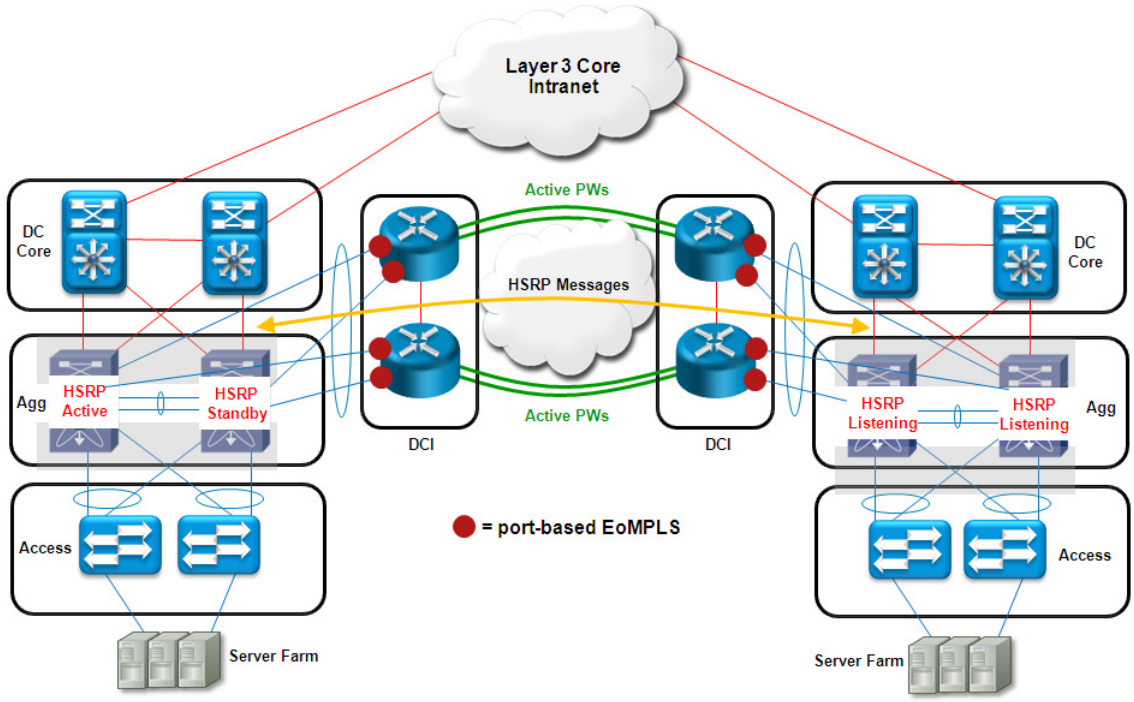
Note The behavior described above, may create a specific issue when deploying EEM and 802.1AE together, as discussed in more details in Encryption, page 2-21.

First Hop Redundancy Protocol (FHRP) Deployment

Some interest design considerations arise when discussing where to deploy the default gateway functionality for devices belonging to a VLAN/subnet that is stretched between remote data center sites. Should the default gateway be positioned in one of the two data centers? Or should each site have its own default gateway?

The solution validated as part of DCI phase 2 is based on the former approach: as shown in Figure 2-10, HSRP runs between the devices deployed in the aggregation layer of the remote data centers. HSRP messages are exchanged for each VLAN leveraging the EoMPLS PW logical connections established between remote sites.

Figure 2-10 HSRP Deployment between DCs



As shown in Figure 2-10, two aggregation devices in DC1 acquire the HSRP "Active" and "Standby" roles, whereas the other two aggregation layer devices in DC2 functions in "Listening" mode. This behavior may obviously be tuned for each VLAN, allowing the "distribution" of the default gateway features on different devices for different VLANs/IP subnets. There are operational implications in doing so, so we recommend that you keep the configuration consistent for all the extended VLANs.

The required configuration to achieve this behavior is shown in Figure 2-11.

Figure 2-11 HSRP Configuration Samples

N7K-Agg1-DC1 (HSRP Active)

```
track 1 ip route <core route>
!
interface Vlan10
 description VLAN Extended
 ip address 10.10.10.2/24
 ip ospf passive-interface
 ip router ospf 200 area 0.0.0.0
 hsrp 10
  authentication md5 key-string test
  preempt delay minimum 60 reload 300
  priority 105
  track 1 decrement 25
  timers 1 3
  ip 10.10.10.1
```

N7K-Agg2-DC1 (HSRP Standby)

```
track 1 ip route <core route>
!
interface Vlan10
 description VLAN Extended
 ip address 10.10.10.3/24
 ip ospf passive-interface
 ip router ospf 200 area 0.0.0.0
 hsrp 10
  authentication md5 key-string test
  preempt delay minimum 60 reload 300
  priority 100
  track 1 decrement 20
  timers 1 3
  ip 10.10.10.1
```

N7K-Agg1-DC2 (HSRP Listening)

```
interface Vlan10
 description VLAN Extended
 ip address 10.10.10.4/24
 ip ospf passive-interface
 ip router ospf 200 area 0.0.0.0
 hsrp 10
  authentication md5 key-string test
  preempt delay reload 300
  priority 90
  timers 1 3
  ip 10.10.10.1
```

N7K-Agg2-DC2 (HSRP Listening)

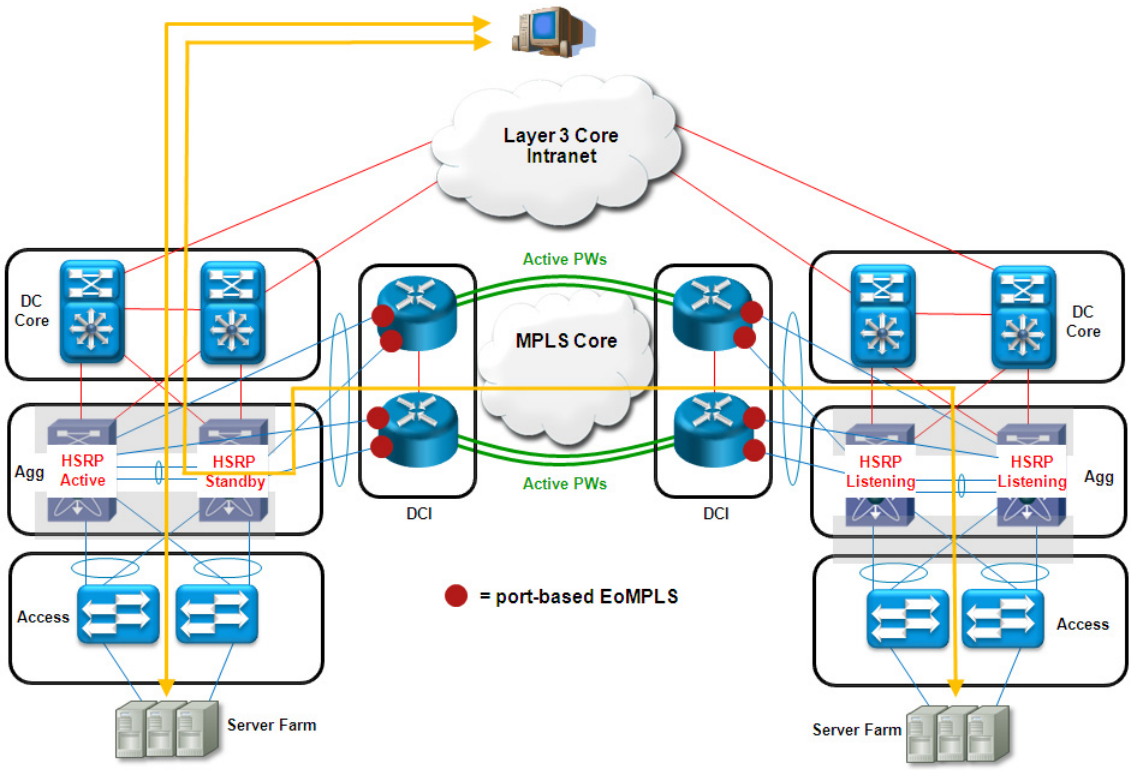
```
interface Vlan10
 description VLAN Extended
 ip address 10.10.10.5/24
 ip ospf passive-interface
 ip router ospf 200 area 0.0.0.0
 hsrp 10
  authentication md5 key-string test
  preempt delay reload 300
  priority 85
  timers 1 3
  ip 10.10.10.1
```

Notice that the preceding configuration should be replicated for each VLAN that is extended between the two data centers. Also, in scenarios where the aggregation layer devices deployed in a data center are Catalyst 6500 in VSS mode, there would only be 3 devices exchanging HSRP messages (two Nexus and one VSS).

Note The considerations made in this section apply only to the VLANs/IP subnets that are extended between the remote data center sites. VLANs that are only defined in a specific site would leverage the gateway defined only on the local aggregation switches pair.

The result of the preceding configuration is that all the outbound traffic (from the data center access layer to the Layer 3 intranet core) would need to be routed by the default gateway in DC1. To ensure symmetric traffic flows, we recommend that you also ensure that inbound traffic destined to the DC access layer devices prefers DC1 as well. This enables you to achieve the behavior shown in Figure 2-12.

Figure 2-12 Inbound and Outbound Routed Traffic Flows



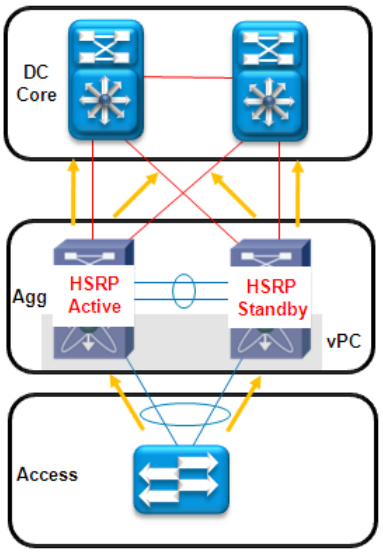
Notice how traffic originated from and destined to devices connected to DC2 access layer must leverage the Layer 2 connections provided by the EoMPLS PWs. This represent sub-optimal traffic behavior that could be acceptable in scenarios where the distance between data centers is limited (up to 100 km). For longer distances, the delay and latency introduced by the requirement of traversing the MPLS core may make this solution inappropriate.

Note Improving this solution may consist in enabling a default gateway for the same VLAN/subnet in each data center and ensuring also that the inbound traffic is directed to the "right" data center (that is, the data center where the destination server actually resides). This functionality are beyond the scope for this document and will be introduced in future releases of the DCI system testing program.

Different convergence mechanisms are enabled for traffic recovery under various link and node failure scenarios, as discussed in detail in the following sections. Understand how traffic flows in a steady state scenario.

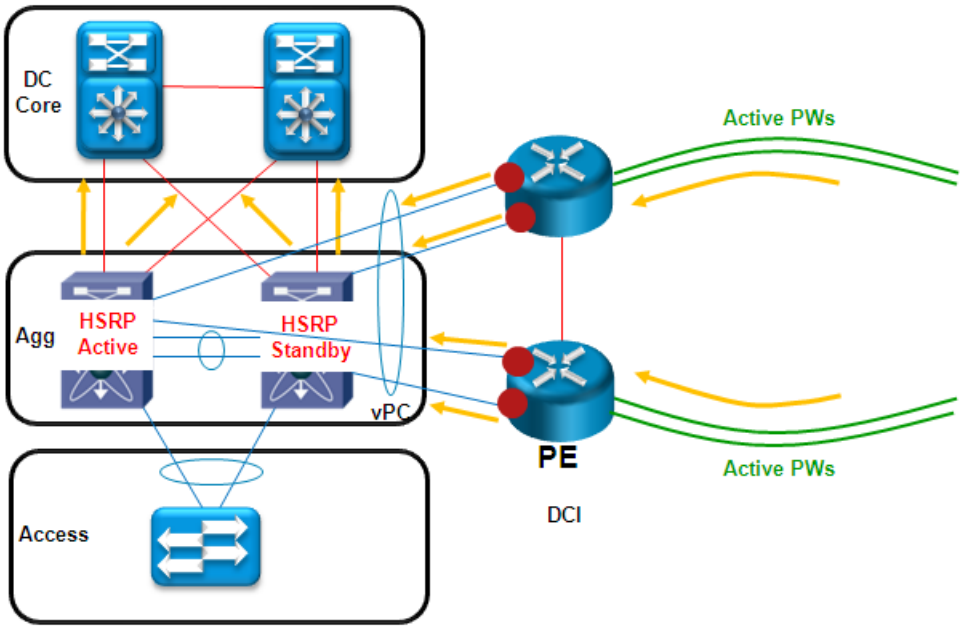
- DC1 Access to Layer 3 Intranet flows: the links connecting each access layer switch to the aggregation layer devices are bundled together. This is possible because of the vPC functionality available with Nexus 7000 switches, which makes the two aggregation switches look like a single logical entity to the external devices. One of the characteristics of HSRP deployments in conjunction with vPC is that the HSRP standby device is capable of routing traffic received on a link member of the vPC. This allow for an optimal utilization of the uplinks between access and aggregation for outbound traffic (Figure 2-13):

Figure 2-13 DC1 Access to Layer 3 Core Traffic Flows



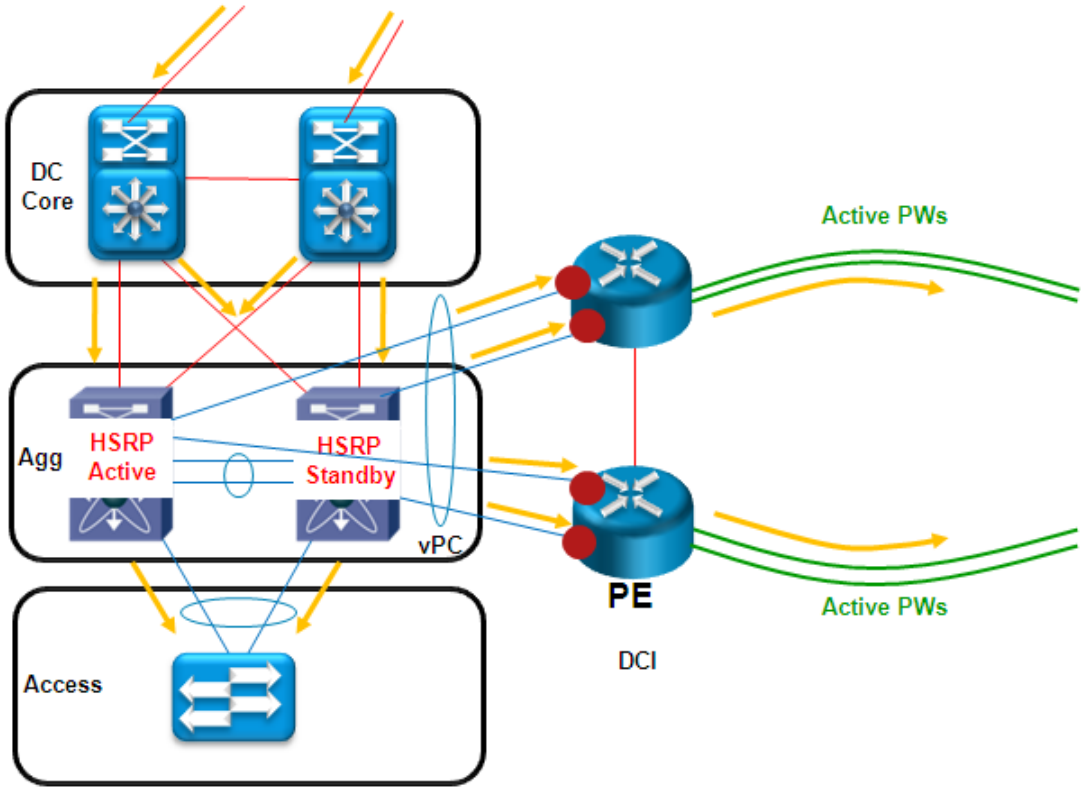
- DC2 Access to Layer 3 Intranet flows: since the two aggregation switches deployed in DC2 are in HSRP "Listening" state, they cannot route the traffic received from the access layer. Therefore, traffic is directed to the aggregation layer devices deployed in DC1, leveraging the Layer 2 extension path established between the remote sites. Once traffic is received by the aggregation layer devices on the links member of the vPC toward the remote aggregation site, the same considerations made above are valid, so all the paths are active as shown in Figure 2-14.

Figure 2-14 DC2 Access to Layer 3 Core Traffic Flows



- Layer 3 Intranet to DC1 and DC2 Access flows: the Layer 3 flows are received by the DC1 core layer devices and are then routed toward the aggregation layer leveraging the four available ECMP paths. At the aggregation layer, traffic is routed to the destination subnets and delivered to the local DC1 access layer or to the remote DC2 access layer leveraging the Layer 2 vPC trunks (Figure 2-15).

Figure 2-15 Layer 3 Core to DC Access Traffic Flows

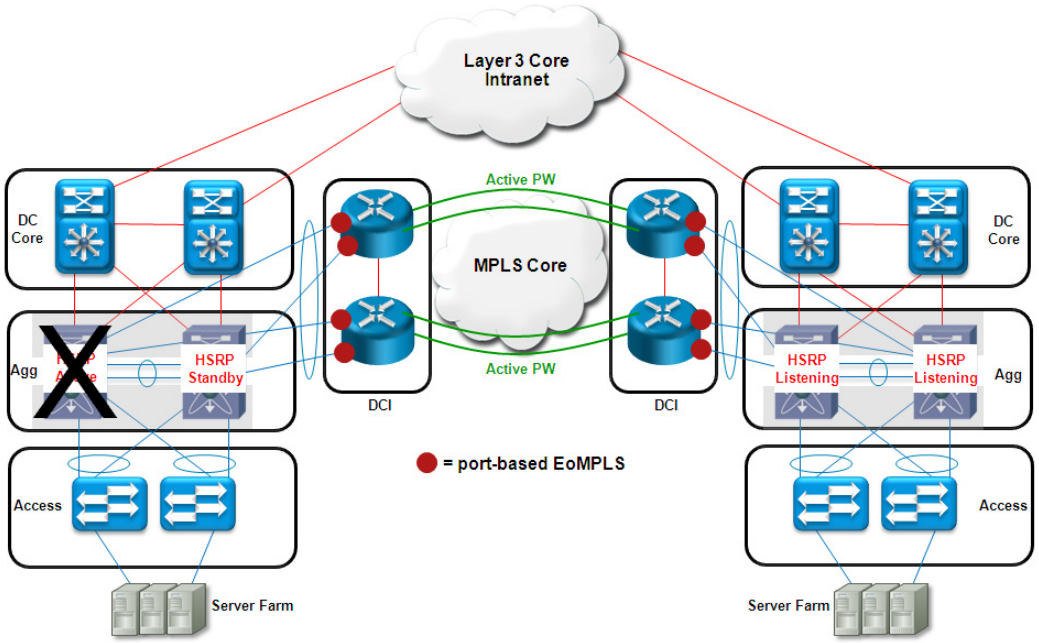


Note The focus here is on the recovery of the Layer 3 traffic flows depicted in Figure 2-12. For more failure/recovery analysis for LAN extension traffic between the two data centers see the "EoMPLS Failure/Recovery Analysis" section.

Test 1: Failure/Recovery of HSRP Active Aggregation Switch

This scenario is shown in Figure 2-16.

Figure 2-16 Failure HSRP Active Switch



Convergence After Failure

- **DC Access to Layer 3 Core flows**: since the HSRP Standby device can route traffic received on links part of the same vPC, only 50% of the traffic flows are actually impacted, and the overall outage is dictated by the recovery of traffic flows on the remaining links member of the same vPC PortChannel. This applies to traffic originated both in DC1 and DC2 access layer.
- **Layer 3 Core to DC Access flows**: Layer 3 traffic flows received at the DC Core from the Intranet are normally routed toward both the aggregation switches because of Equal Cost Multi Pathing (ECMP), so failure of the aggregation switch basically causes a Layer 3 reroute of traffic flows originally sent to this device toward the redundant aggregation layer device.

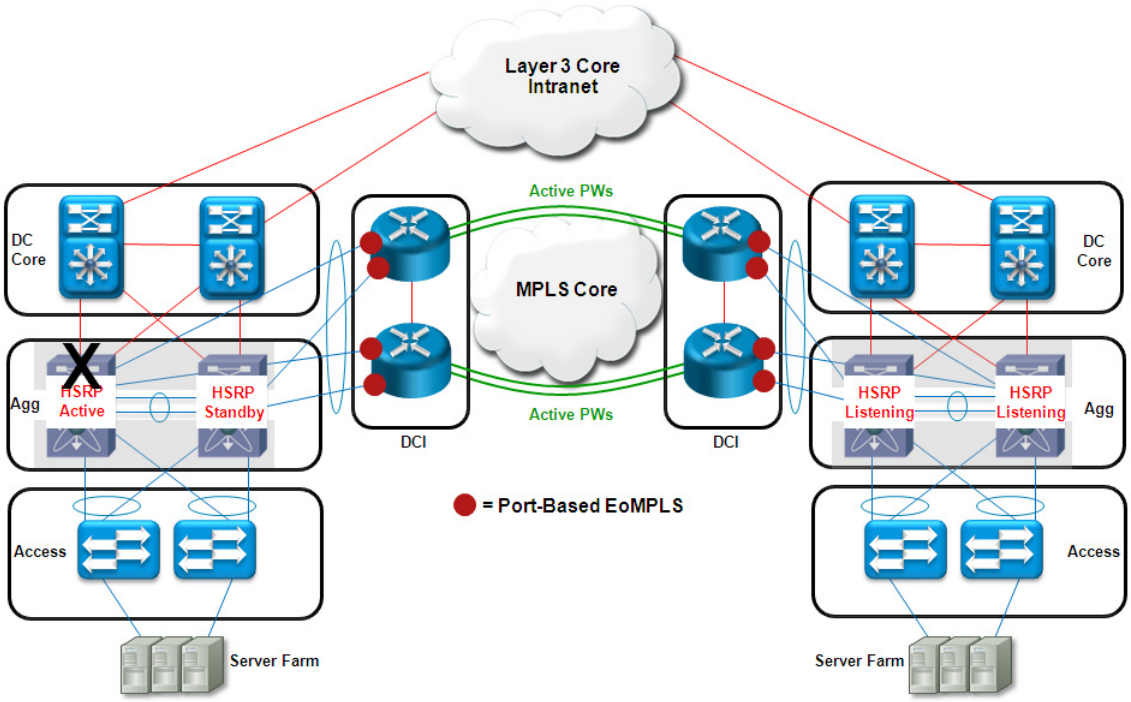
Convergence After Recovery

- When the aggregation switch comes back online, two things need to happen to ensure traffic can flow through it again: first, Layer 3 adjacencies need to be reestablished with the DC core devices. Second, the Layer 2 links connecting to the local access layer and to the DCI layer need to be re-bundled as part of the defined vPCs. Few issues have been discovered for vPC recovery in the 4.2(3) software release validated. Check with your Cisco representative to determine the best software release to deploy.

Test 2: Supervisor Failover in HSRP Active Aggregation Switch

This scenario is shown in Figure 2-17.

Figure 2-17 Supervisor Failover Failure for HSRP Active Switch

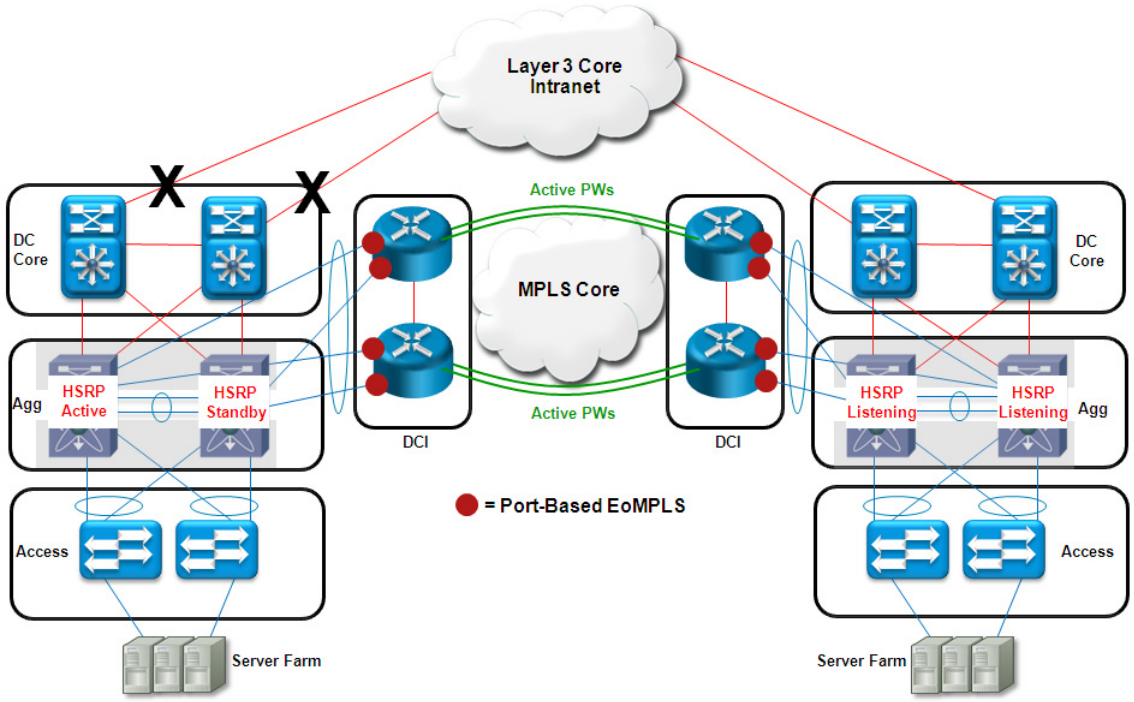


The supervisor failover should not affect traffic in either direction, given the NSF/SSO support on Nexus 7000 platforms,

Test 3: Complete Data Center Isolation Scenario

This failure scenario is shown in Figure 2-18.

Figure 2-18 Complete Data Center Isolation



Convergence After Failure

- **DC Access to Layer 3 Core flows**: the failure of connectivity to the Layer 3 intranet core network by default would not trigger a change in the HSRP state of the aggregation layer devices. The immediate consequence is that the active default gateway is still positioned in DC1 but at that point without having a Layer 3 active path toward the external world, it causes all the traffic to be black-holed. It is hence required to add a specific HSRP tracking configuration (see Figure 2-11) to ensure that the aggregation layer devices in DC1 would lower their HSRP priority under this failure scenario. The consequence is that the active default gateway is moved to DC2 aggregation devices, allowing for the re-establishment of outbound connectivity. The time required for this HSRP role change is the main factor dictating the overall outage experienced in this scenario.

Note HSRP tracking would not be required if the MPLS cloud was also leveraged for establishing Layer 3 connectivity between data centers (as discussed in the "Inter Data Centers Routing Considerations" section). In this section, the assumption is that the MPLS cloud is only used to provide LAN extension services.

- **Layer 3 Core to DC Access flows**: as previously mentioned, all the traffic originated in the Layer 3 internet core is steered toward DC1 to avoid asymmetric inbound/outbound traffic flows. If DC1 gets isolated, a routing update in the core would cause traffic to get steered toward DC2 instead, and the time required to do so is the main factor affecting the inbound outage.

Convergence After Recovery

- **DC Access to Layer 3 Core flows:** once DC1 regains connectivity to the intranet core, the aggregation layer devices in DC1 would raise their HSRP priority to become the active gateways once again. All outbound traffic will start flowing again via DC1 aggregation layer. Use "HSRP delay" to ensure that the aggregation layer devices in DC1 regain their active/standby roles after the Layer 3 portion of the network is converged (to avoid traffic black-holing).
- **Layer 3 Core to DC Access flows:** a routing update in the intranet Layer 3 core would cause the Layer 3 streams to be steered again toward DC1. Because this event may be faster than HSRP re-election, it may be possible to have asymmetric traffic flows for an interim time frame. This is not expected to affect the overall connectivity.

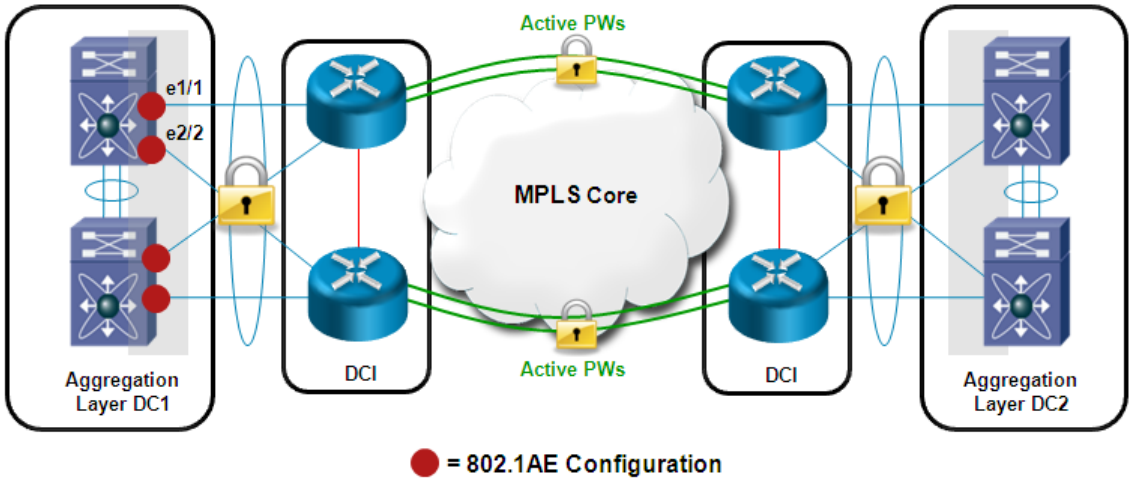
Encryption

The deployment of port-mode EoMPLS as LAN extension technology allows you to leverage 802.1AE (also known as MAC security – MACsec) as encryption mechanism for all the traffic exchanged across the logical PWs interconnecting the remote data center sites. Once again, this is technically possible since the aggregation layer devices deployed in different sites appear as directly connected at Layer 2 (MACsec is usually configured and supported only on back-to-back physical connections).

Note At this time 802.1AE is only supported on Nexus 7000 platforms.

As highlighted in Figure 2-19, the 802.1AE configuration needs to be applied at the physical interface level, despite that these interfaces are then bundled as part of a PortChannel.

Figure 2-19 802.1AE Encryption between DC Sites



The required configuration is quite simple and follows:

```
interface Ethernet1/1
  description PortChannel Member
  cts manual
    no propagate-sgt
    sap pmk 1234000000000000000000000000000000000000000000000000000000000000
!
interface Ethernet2/2
  description PortChannel Member
```

```
cts manual
  no propagate-sgt
  sap pmk 5678000000000000000000000000000000000000000000000000000000000000
```

Manual 802.1AE configuration can be leveraged to establish encrypted sessions between the Nexus 7000 devices. The required SAP pairwise master key (PMK) parameter is an even number of characters with a maximum length of 32 characters. It needs to be configured identically for the two interfaces establishing the 802.1AE encrypted session, but it can vary between different interfaces on the same device (as shown in the preceding example).

Note The PMK values used above are just simple examples used in a testing environment.

The use of the **no propagate-sgt** command is recommended as a workaround for a couple of software defects discovered in the 4.2(3) software release validated. Notice that leveraging this command does not have any specific impact in this deployment scenario, since the use of SGT (Security Group Tags) is not required for establishing static 802.1AE encrypted connections.

You can verify the successful establishment of an 802.1AE session by using the following command:

```
DC1-Agg1# sh cts interface e1/1
```

802.1AE Information for Interface Ethernet1/1:

```
CTS is enabled, mode:   CTS_MODE_MANUAL
IFC state:              CTS_IFC_ST_CTS_OPEN_STATE
Authentication Status:  CTS_AUTHC_SKIPPED_CONFIG
  Peer Identity:
  Peer is:              Unknown in manual mode
  802.1X role:          CTS_ROLE_UNKNOWN
  Last Re-Authentication:
Authorization Status:   CTS_AUTHZ_SKIPPED_CONFIG
  PEER SGT:             0
  Peer SGT assignment:  Not Trusted
SAP Status:             CTS_SAP_SUCCESS
  Configured pairwise ciphers: GCM_ENCRYPT
  Replay protection: Enabled
  Replay protection mode: Strict
  Selected cipher: GCM_ENCRYPT
  Current receive SPI: sci:23ac6554740000 an:0
  Current transmit SPI: sci:225577aac40000 an:0
```

When deploying 802.1AE sessions between the remote aggregation layer devices, realize that any QoS classification or queuing is not be possible for this type of traffic. This is because 802.1AE encrypts the entire frame, with the exception of the source and destination MAC addresses. Therefore, there is no visibility not even into the CoS field and all the traffic will basically fall into the default queue.

Caveats for 802.1AE Deployments

Several specific issues were discovered while validating the use of 802.1AE as encryption mechanism across logical EoMPLS PWs. Some of these caveats apply only when deploying Catalyst 6500 as PE devices, others apply when leveraging ASR1000 routers for that role.

1. As noticed in the preceding output, by default the Cisco NX-OS software enables the data-path reply protection feature. This ensures the integrity of the data-path between the two devices establishing the 802.1AE session by verifying that the frames that originated on one side are received in the same order on the other side. A sequence number is added to each encrypted frame sent out an 802.1AE enabled interface, so that out-of-order frames can be dropped since they may be the result of a man-in-the-middle attack.

This behavior may cause an interaction issue between 802.1AE and LACP when a Catalyst 6500 is deployed as a PE device. This is because any packets with the BPDU bit set (including LACP packets) are punted to the 6500 switch CPU even though they are supposed to be tunneled transparently across the EoMPLS PW. The final result is that LACP frames will almost always be received out of order on the other side and hence dropped, causing the physical link to be unbundled from the PortChannel and compromising network connectivity.

The recommended work around is to change the default 802.1AE behavior and disable the replay protection functionality, as shown in the following configuration sample:

```
interface Ethernet1/1
  description PortChannel Member
  cts manual
    no replay-protection
```

2. A software defect in release 4.2(3) was discovered when 802.1AE was configured on a multiple physical links member of the same vPC when these interfaces were configured in dedicated rate-mode (this is usually done to avoid oversubscription on these interfaces, since on a 32-port 10 GE Ethernet module, each set of four ports can handle 10 gigabits per second of bandwidth). As shown in Figure 2-19, this is the case when using a full mesh of physical connections between the aggregation and the DCI layer devices. This configuration caused a re-key error when sending traffic on the links, which eventually caused the connectivity to be compromised. A workaround for this issue consists of changing the operation mode of the 10GE interfaces from dedicated to shared, as shown in the following configuration:

```
DC1-Agg1(config)# int e1/1
DC1-Agg1(config-if-range)# rate-mode shared
```

3. The final issue is related to the use of EEM to work around the lack of Remote Ethernet Port Shutdown feature on Catalyst 6500 platforms. To make the required EEM applet as simple as possible (as discussed in the previous section), the proposal is to bounce the link (“shut” or “no shut”) connecting the PE device to the aggregation layer switch every time a PW failure is detected.

   The consequence of this configuration is that the link between aggregation and DCI layer may be brought back to a physical “UP” state before connectivity across the EoMPLS PW is re-established. This would cause the 802.1AE process running on the Nexus 7000 interface to actively start negotiating a new encrypted session with the device on the opposite side of the PW; however, after a certain number of attempts (approximately 15 minutes), the 802.1AE process will “give up” on this attempt and bring down the physical interface link connecting to the DCI layer. The only way to recover connectivity across that link (after the PW connection is reactivated) is to manually bounce the Nexus 7000 interface, losing the advantage of leveraging LACP to dynamic bundle back this interface in the Port-Channel.

   To avoid this problem, a more complex EEM logic would need to be implemented to take into account the status of end-to-end IP connectivity across the core between data centers. That would provide the condition for disabling or re-enabling the physical link between DCI and aggregation layer devices. This more complex scenario is out of scope for this release.

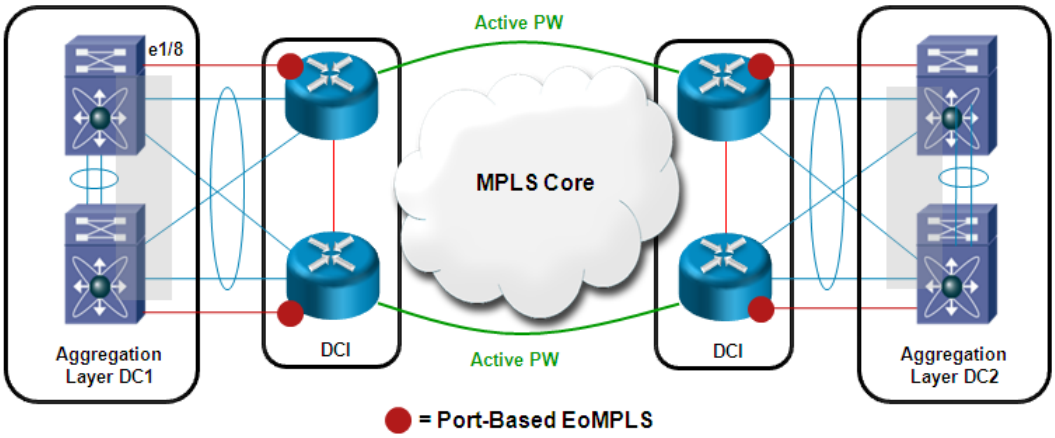
Note Remote Ethernet Port Shutdown is not exposed to this type of issue, since the link between aggregation and DCI layer is kept to a “DOWN” physical state until connectivity through the PW is re-established.

Inter Data Centers Routing Considerations

The EoMPLS PW deployment discussed in the previous sections of this document are generally used to carry the Layer 2 traffic between the remote data centers thus providing the required LAN extension services. When Layer 3 communication is also required between sites there are essentially two different scenarios to consider:

- Routing between remote sites leverages the connectivity across a different network, as shown in Figure 2-3 on page 2-4.
- Routing between data centers happens across the same MPLS cloud used for providing EoMPLS connectivity. This is usually the case when the bandwidth available via the MPLS cloud is higher than the one through the Layer 3 intranet, or if the MPLS cloud is the only connection available between data center sites. In these scenarios, avoid injecting the data center routes into the MPLS core or into the PE routing tables. This can be easily achieved by leveraging a dedicated PW to establish a routing adjacency between the aggregation layer devices, as shown in Figure 2-20.

Figure 2-20 Establishment of EoMPLS PWs for Inter Data Centers Routing



The aggregation layer and the DCI layer device configurations follow, respectively:

Aggregation Layer

```
interface Ethernet1/8
  description L3 peering with remote aggregation device
  cts manual
    no propagate-sgt
    sap pmk 1234000000000000000000000000000000000000000000000000000000000000
  no switchport
  mtu 9216
  ip address 50.0.45.1/30
  ip ospf authentication message-digest
  ip ospf message-digest-key 1 md5 3 9125d59c18a9b015
  ip ospf network point-to-point
  ip router ospf 10 area 0.0.0.0
```

DCI Layer

```
interface TenGigabitEthernet1/8
 description L3 link to aggregation
 mtu 9216
 no ip address
 xconnect 15.0.5.1 2545 encapsulation mpls
```

The configuration on the DCI layer device is similar because the port-mode EoMPLS functionality is independent from the characteristic of the link it is applied to. Notice also how 802.1AE can still be applied on the aggregation layer routed interface to encrypt the Layer 3 traffic flows between data center sites.

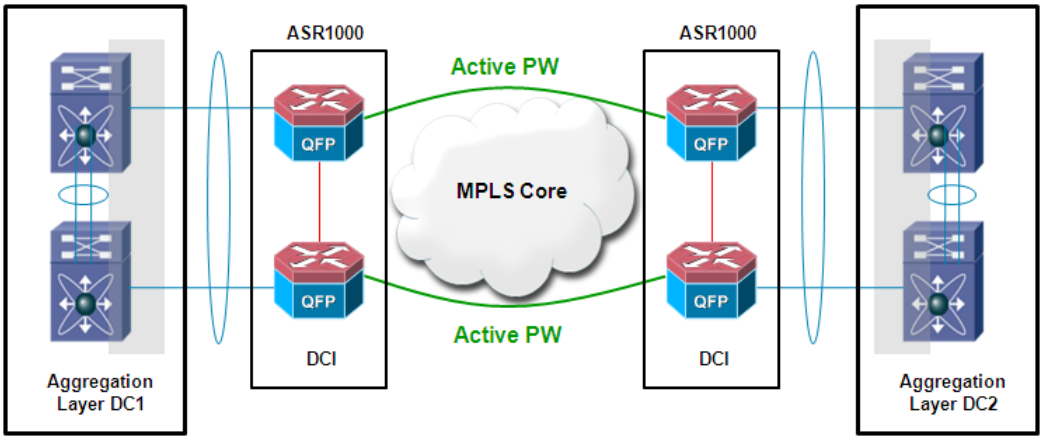
The recommendation for this deployment centers around a need to limit the number of routes injected in the PE routing table, which may become important especially when EEM is used to implement Remote Port Shutdown features. This is because EEM and the routing process may compete for CPU resources, so EEM activation may be delayed during a network convergence event causing a massive routing update.

EoMPLS Failure/Recovery Analysis

Each failure/recovery tests discussed below are executed in two scenarios:

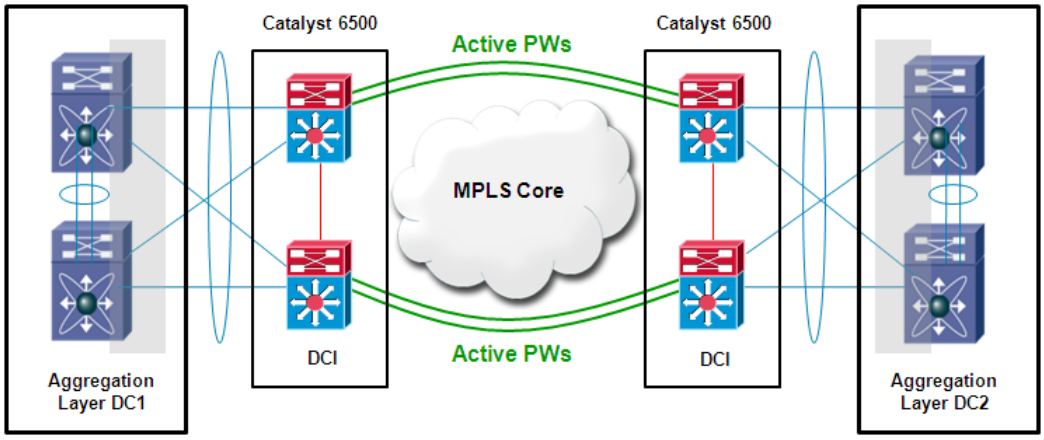
- **ASR1000 deployed as PE**: in this scenario (Figure 2-21) the connectivity between aggregation layer devices and the DCI layer was not leveraging a full mesh of connections.

Figure 2-21 ASR1000 Deployed as PE in the DCI Layer



- **Catalyst 6500 deployed as PEs**: in this scenario (Figure 2-22), a full mesh of connections is established between the aggregation and the DCI layer devices.

Figure 2-22 Catalyst 6500 Deployed as PE in the DCI Layer

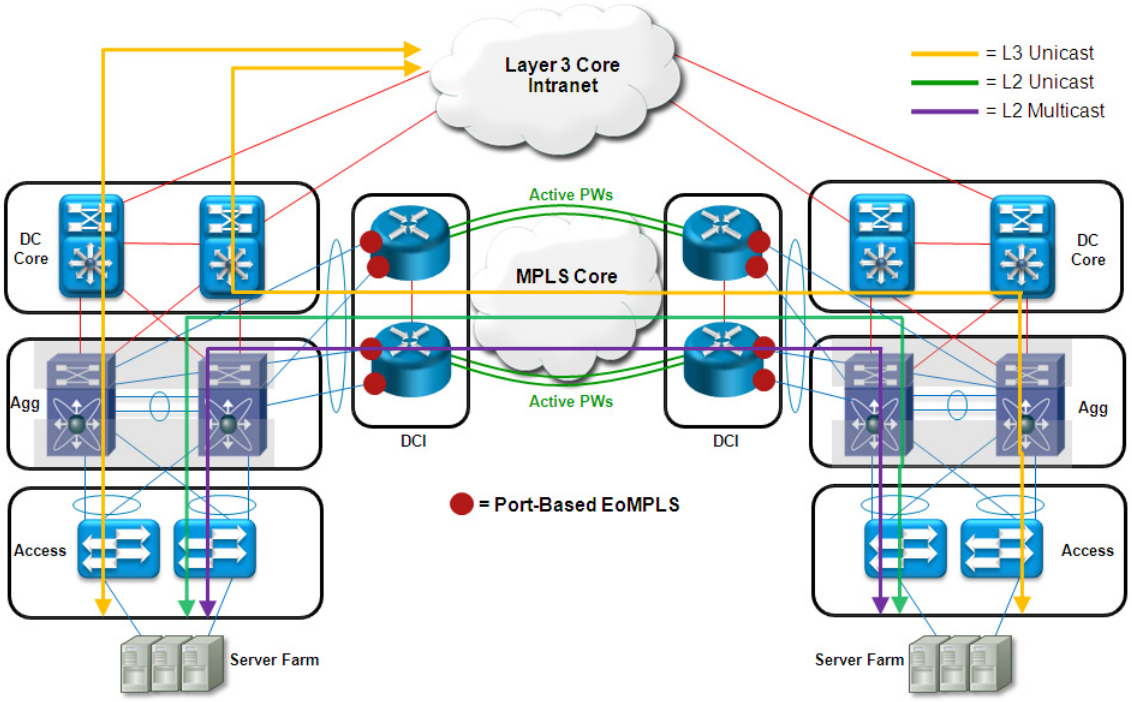


The reason for this choice was to optimize the number of test cases to be executed, while being able to cover all the scenarios and get all the valuable information to provide final design recommendations.

For each test case, results for the non fully meshed and fully meshed topologies will be provided separately in the following sections.

Figure 2-23 shows the traffic flows that were established across the network.

Figure 2-23 Established Traffic Flows



As shown, a mix of Layer 2 and Layer 3 traffic flows were enabled. The goal was to simulate as much as possible the traffic mix expected in a real deployment. For this reason, frames with different IP size were used to build the traffic flows.

From a scalability perspective, the following are the parameters that were validated:

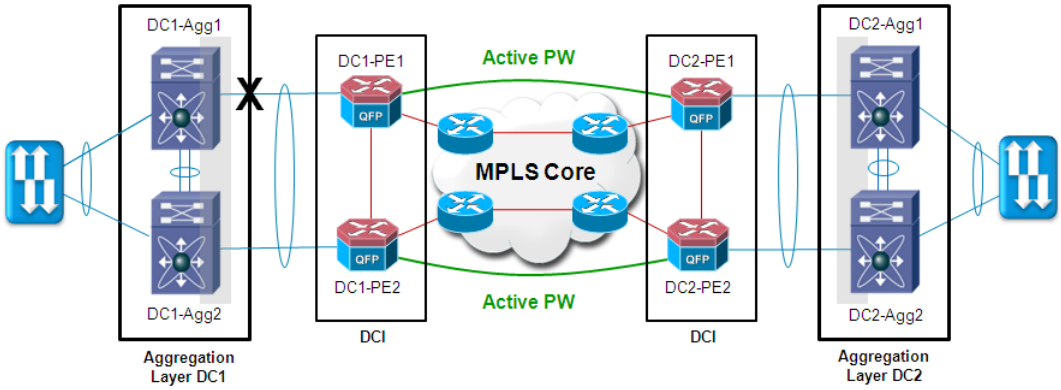
- 1200 VLANs extended between data centers
- 24000 routes injected between the two data centers (12000 in each of them)
- Around 2.5 Gbps of combined traffic was sent between data centers (in each direction)

In the following sections are discussed the various failure and recovery scenarios. For each scenario, the convergence results are captured and explained, both for deployment leveraging ASR1000 or Catalyst 6500 in the DCI layer.

Test 1a : Aggregation to DCI Layer Link Failure/Recovery (AS1000 Topology)

This failure/recovery scenario is shown in Figure 2-24.

Figure 2-24 Aggregation to DCI Layer Link Failure/Recovery (AS1000 Topology)



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the main factor affecting the convergence is how fast the top Nexus7000 aggregation box (DC1-Agg1) is able switch traffic received from the access layer to the vPC peer link connecting to DC1-PE2. This is because the only active path remaining is now via DC1-PE2.
- **DC2 to DC1 Layer 2 Flows**: when the link in DC1 fails, the corresponding PW is brought down (because the attachment circuit has failed). This means that traffic sent from the aggregation layer device in DC2 on the physical link mapped to that specific PW is black-holed until the physical link is also brought down. As previously explained, this is achieved by leveraging the Remote Port Shutdown functionality enabled by default on ASR1000 devices. In addition to the time required for the local link in DC2 to be disabled, you must consider the time required for the aggregation layer device in DC2 2 (DC2-Agg1) start switching the traffic on the vPC peer link connecting to DC2-Agg2 (similarly to what discussed above for the DC1 to DC2 direction).

Convergence After Recovery

- **DC1 to DC2 Layer 2 flows**: convergence after link recovery depends on how fast the link can be bundled back to the vPC. As shown in the results below, it was experienced a longer outage in scenarios where the recovering link is the only one configured on the aggregation box (this is the case in the ASR1000 topology).
- **DC2 to DC1 Layer 2 flows**: convergence depends on re-bundling of both local links in DC1 and DC2 and on the re-establishment of the PW.

Table 2-1 below captures the results achieved with this specific test case.

Table 2-1 Aggregation to DCI Layer Link Failure/Recovery Results (ASR1000 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	1.2 sec	3 sec[1] (1 sec.)	2.8 sec[2]	2.7 sec[3]
Layer 2 Multicast (ASR1000 as PE)	1.2 sec	3 sec[4] (1 sec)	2.7 sec[5]	2.7 sec[6]

1. This result was achieved with the default carrier-delay setting (2 seconds) on the ASR1000 interface. Basically, when the link toward the aggregation layer fails, the ASR1000 does not consider the link down for 2 extra seconds. Therefore, the PW stays up for 2 extra seconds and Remote Ethernet Port Shutdown on DC2-PE1 is delayed as well of the same amount of time causing the flows originated from DC2 to be black-holed (because the remote local link is still up). It was verified that tuning carrier-delay to a lower value (for example 10 msec) removed this extra delay the number in brackets were achieved after this tuning was in place). To accomplish this task, enter the following CLI:

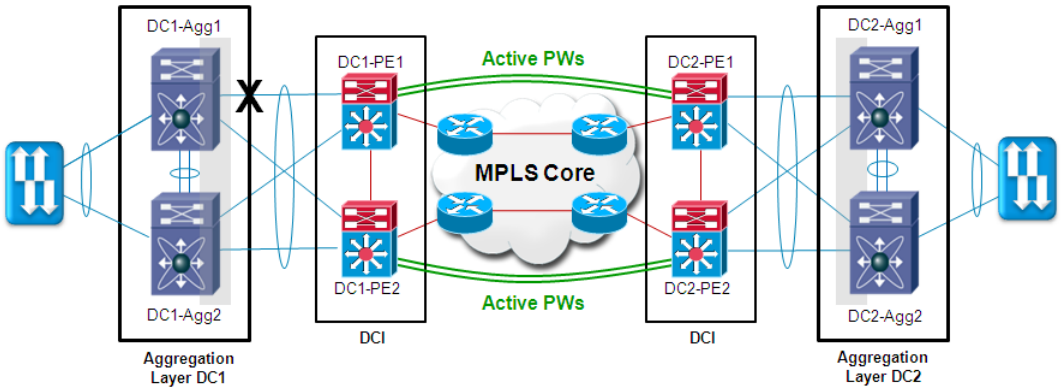
   interface TenGigabitEthernet1/0/0
   mtu 9216
   no ip address
   carrier-delay msec 10
   xconnect 11.0.2.31 100 encapsulation mpls

   Note: A carrier-delay of 0 is not recommended because it may cause interface flapping when enabling 802.1AE on the aggregation layer side (this was not consistently happening and it was only verified during link recovery).
2. This higher recovery value was caused by not having a full mesh topology between the aggregation and the DCI layer switches. Therefore, we recommend deploying full mesh connections so the end-to-end EtherChannel between the remote aggregation switches is built with 4 links (2 on each aggregation switch).
3. Same as 2
4. Same as 1
5. Same as 2
6. Same as 2

Test 1b: Aggregation to DCI Layer Link Failure/Recovery (Catalyst 6500 Topology)

This failure/recovery scenario is shown in Figure 2-25.

Figure 2-25 Aggregation to DCI Layer Link Failure/Recovery (Catalyst 6500 Topology)



Convergence After Failure

- DC1 to DC2 Layer 2 flows: since two links belonging to the same vPC are available on DC1-Agg1, the main factor affecting the convergence is how fast the device is able to reprogram the EtherChannel hashing logic to be able to send the flows via the remaining link available toward the DCI layer.
- DC2 to DC1 Layer 2 flows: when the link in DC1 fails, the corresponding PW is brought down (because the attachment circuit has failed). This means that traffic sent from the Aggregation layer device in DC2 on the physical link mapped to that specific PW is black-holed until the physical link is also brought down. As previously explained, this is achieved by leveraging a simple EEM script running on the Catalyst 6500 PE device. In addition to the time required for the local link in DC2 to be disabled, it must be considered the time required for the aggregation layer device in DC2 to re-hash the flows across the remaining EtherChannel links (similarly to what discussed above for the DC1 to DC2 direction).

Convergence After Recovery

- DC1 to DC2 Layer 2 flows: convergence after link recovery depends on how fast the link can be bundled back to the vPC. Testing results showed that this is happening faster on a Nexus7000 device if there is already a link belonging to the same vPC already up and bundled.
- DC2 to DC1 Layer 2 flows: convergence depends on re-bundling of both local links in DC1 and DC2 and on the re-establishment of the PW.

Table 2-2 captures the results achieved with this specific test case.

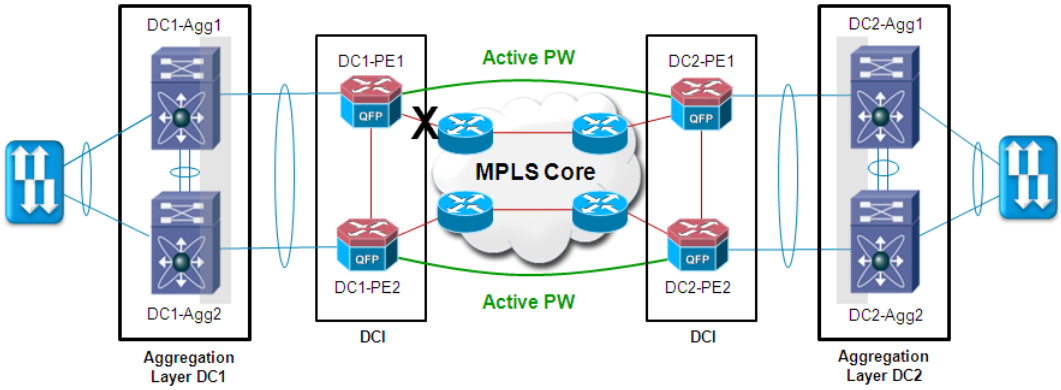
Table 2-2 Aggregation to DCI Layer Link Failure/Recovery Results (Catalyst 6500 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (6500 as PE)	<1 sec	1.1 sec	<1 sec	<1 sec
Layer 2 Multicast (6500 as PE)	1 sec	1.2 sec	<1 sec	<1 sec

Test 2a: DCI Layer to MPLS Core Failure/Recovery (ASR1000 Topology)

This failure/recovery scenario is shown in Figure 2-26.

Figure 2-26 DCI Layer to MPLS Core Failure/Recovery (ASR1000 Topology)



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the PW established between DC1-PE1 and DC2-PE1 remains active in this case, since MPLS traffic is rerouted across the transit link interconnecting the two DCI devices in DC1. Therefore, the convergence is exclusively dictated by how fast this traffic rerouting can happen.
- **DC2 to DC1 Layer 2 Flows**: similar considerations are valid in this direction and traffic rerouting is the main responsible for convergence.

In the testbed used to validate the EoMPLS solution, OSPF was the routing protocol used in the MPLS core. Aggressively tune the OSPF LSA and throttle timers to speed up convergence under core link/box failure scenarios. In addition to that, tuning of OSPF timers under the Layer 3 MPLS enabled interfaces is also recommended. The required configuration follows (it needs to be applied to all the PE and P devices).

```
interface TenGigabitEthernet0/0/0
 mtu 9216
 ip address 40.0.10.1 255.255.255.0
 ip ospf hello-interval 1
!
router ospf 1
 timers throttle spf 10 100 5000
 timers throttle lsa 10 100 5000
 timers lsa arrival 80
```

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: link recovery causes in this case another routing adjustment, so that all MPLS traffic is carried across the optimal path (bypassing the transit link between DCI layer devices).
- **DC2 to DC1 Layer 2 Flows**: same as above.

Table 2-3 captures the results achieved with this specific test case.

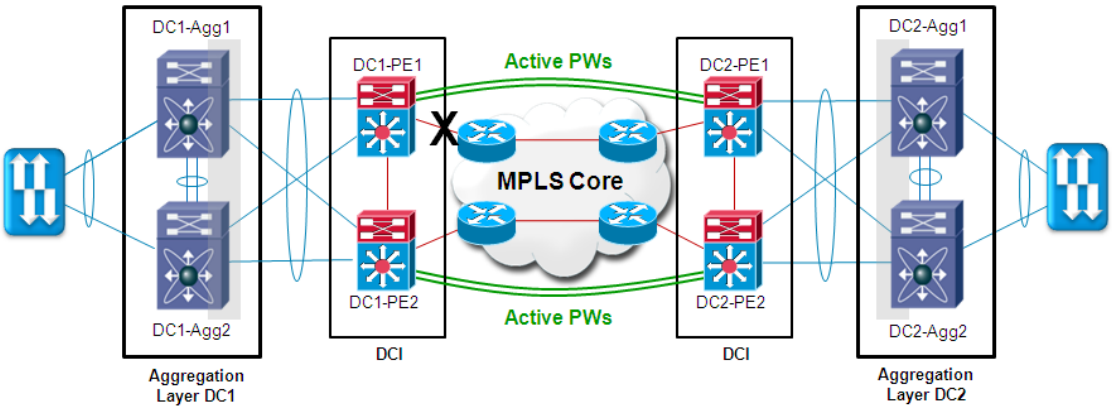
Table 2-3 DCI Layer to MPLS Core Failure/Recovery Results (ASR1000 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	< 1sec	< 1sec	< 1sec	< 1sec
Layer 2 Multicast (ASR1000 as PE)	< 1sec	< 1sec	< 1sec	< 1sec

Test 2b: DCI Layer to MPLS Core Failure/Recovery (Catalyst 6500 Topology)

This failure/recovery scenario is shown in Figure 2-27.

Figure 2-27 DCI Layer to MPLS Core Failure/Recovery (Catalyst 6500 Topology)



Convergence After Failure

- DC1 to DC2 Layer 2 flows: both the PWs established between the DC1-PE1 and DC2-PE1 remain active in this case, since MPLS traffic is re-routed across the transit link interconnecting the two DCI devices in DC1. Therefore, the convergence is exclusively dictated by how fast this traffic re-routing can happen.
- DC2 to DC1 Layer 2 flows: similar considerations are valid in this direction and traffic re-routing is the main responsible for convergence.

Note The same OSPF timers tuning discussed for the ASR1000 topology was also applied in this case.

Convergence After Recovery

- DC1 to DC2 Layer 2 flows: link recovery causes in this case another routing adjustment, so that all the MPLS traffic is now carried across the optimal path (bypassing the transit link between DCI layer devices).
- DC2 to DC1 Layer 2 flows: same as above.

Table 2-4 captures the results achieved with this specific test case.

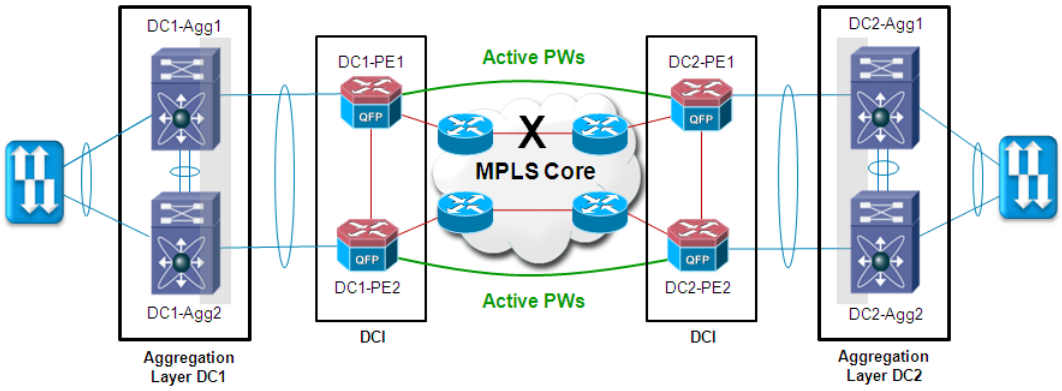
Table 2-4 DCI Layer to MPLS Core Failure/Recovery Results (Catalyst 6500 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (6500 as PE)	0.16 sec	0.11 sec	0 sec	0 sec
Layer 2 Multicast (6500 as PE)	0.1 sec	0.1sec	0 sec	0 sec

Test 3a: MPLS Core "Brown Out" Failure/Recovery (ASR1000 Topology)

This failure/recovery scenario is shown in Figure 2-28.

Figure 2-28 MPLS Core Brown Out Failure/Recovery (ASR1000 Topology)



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the failure of the link between core devices would cause a reroute of MPLS traffic across the transit link connecting the two PE devices in the DCI layer. This is similar to the test 2 scenario, with the only difference is that the DCI layer device triggers an IGP rerouting after receiving an IGP notification from the core router because of a direct link failure.
- **DC2 to DC1 Layer 2 Flows**: the behavior is identical to what is discussed above for the opposite direction.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: link recovery causes in this case another routing adjustment, so that all the MPLS traffic is carried across the optimal path (bypassing the transit link between DCI layer devices).
- **DC2 to DC1 Layer 2 Flows**: same as described in DC! to DC2 Layer 2 flows.

Table 2-5 captures the results achieved with this specific test case.

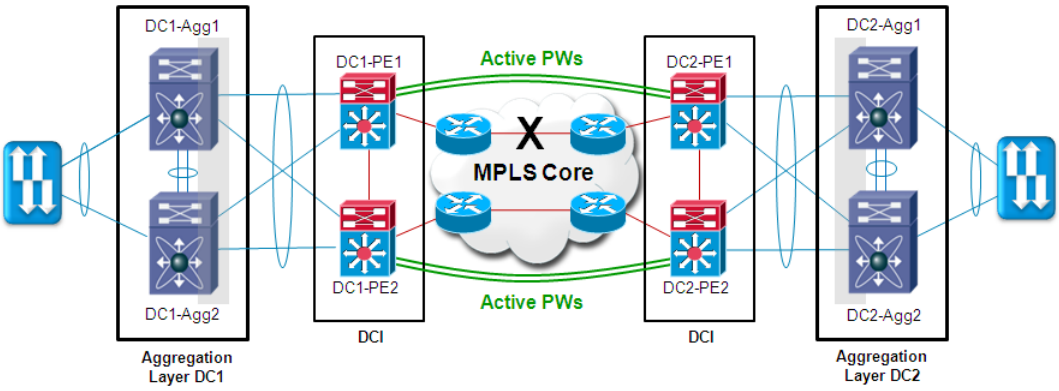
Table 2-5 MPLS Core Brown Out Failure/Recovery Results (ASR1000 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	< 1sec	< 1sec	< 1sec	< 1sec
Layer 2 Multicast (ASR1000 as PE)	< 1sec	< 1sec	< 1sec	< 1sec

Test 3b: MPLS Core "Brown Out" Failure/Recovery (Catalyst 6500 Topology)

This failure/recovery scenario is shown in Figure 2-29.

Figure 2-29 MPLS Core Brown Out Failure/Recovery (Catalyst 6500Topology)



Convergence After Failure

- DC1 to DC2 Layer 2 flows: the failure of the link between core devices would cause a re-route of MPLS traffic across the transit link connecting the two PE devices in the DCI layer. This is similar to the test 2 scenario, with the only difference that the DCI layer device will trigger an IGP re-routing after receiving an IGP notification from the core router (instead than because of a direct link failure).
- DC2 to DC1 Layer 2 flows: the behavior is identical to what discussed above for the opposite direction.

Convergence After Recovery

- DC1 to DC2 Layer 2 flows: link recovery causes in this case another routing adjustment, so that all the MPLS traffic is now carried across the optimal path (bypassing the transit link between DCI layer devices). This is not expected to cause any traffic outage, since represents only an optimization of the traffic flows. However, as shown below, a small outage (sub-second) is experienced with ASR1000 deployed as DCI device.
- DC2 to DC1 Layer 2 flows: same as above.

Table 2-6 captures the results achieved with this specific test case.

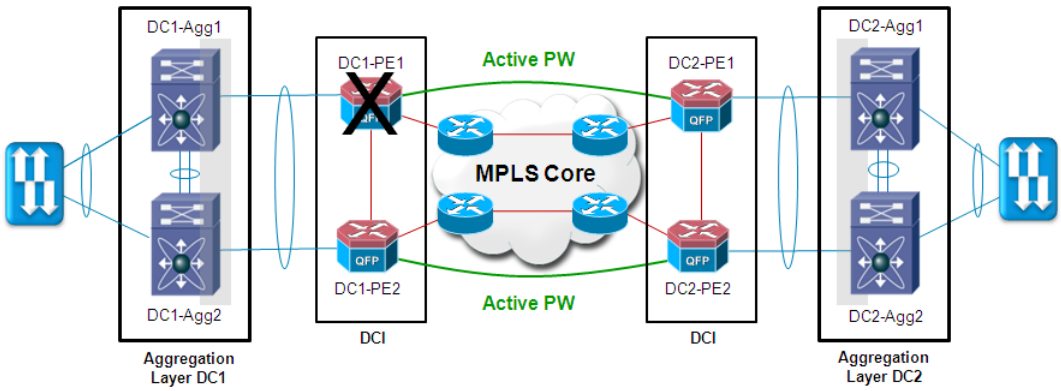
Table 2-6 MPLS Core Brown Out Failure/Recovery Results (Catalyst 6500 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (6500 as PE)	0.17 sec	0.17 sec	0 sec	0 sec
Layer 2 Multicast (6500 as PE)	0.12 sec	0.12 sec	0 sec	0 sec

Test 4a: PE Node Failure/Recovery (ASR1000 Topology)

This failure/recovery scenario is shown in Figure 2-30.

Figure 2-30 PE Node Failure/Recovery (ASR1000 Topology)



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the failure of DC1-PE1 causes the local link connecting to the aggregation layer devices to go down. Therefore, all traffic that DC1-Agg1 receives from the access layer devices (via the vPC) needs to be sent across the vPC peer link (connecting to DC1-Agg2). This is a recovery similar to the local link failure scenario discussed in Test 1a.
- **DC2 to DC1 Layer 2 Flows**: when DC1-PE1 fails, the PW established with DC2-PE1 goes down. This causes DC2-PE1 to bring down (leveraging Remote Ethernet Port Shutdown) the local link connecting to DC2-Agg1. Traffic received by DC2-Agg1 from the access layer switches in DC2 is then sent via the vPC peer link connecting to DC2-Agg2. Once again, this is similar to Test 1a.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: once the PE device is back online and regains connectivity with the MPLS core and the aggregation layer switches, back-to-back connectivity between aggregation devices in the remote sites is reestablished. LACP frames are exchanged again, allowing for the bundling to the vPC of the physical links connecting the aggregation layer to DC1-PE1 (or DC2-PE1).
- **DC2 to DC1 Layer 2 Flows**: the reestablishment of the logical PWs would inform DC2-PE1 to reactivate the links toward the aggregation layer. This is key to be able to re-bundle these to the end-to-end vPC (as discussed above). Therefore, the convergence impact is expected to be mostly the same in both directions.

Table 2-7 captures the results achieved with this specific test case.

Table 2-7 PE Node Failure/Recovery Results (ASR1000 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	1.1 sec	1.1 sec	4.1 sec[1]	4.1 sec[2]
Layer 2 Multicast (ASR1000 as PE)	1 sec	1.1 sec	4 sec[3]	4 sec[4]

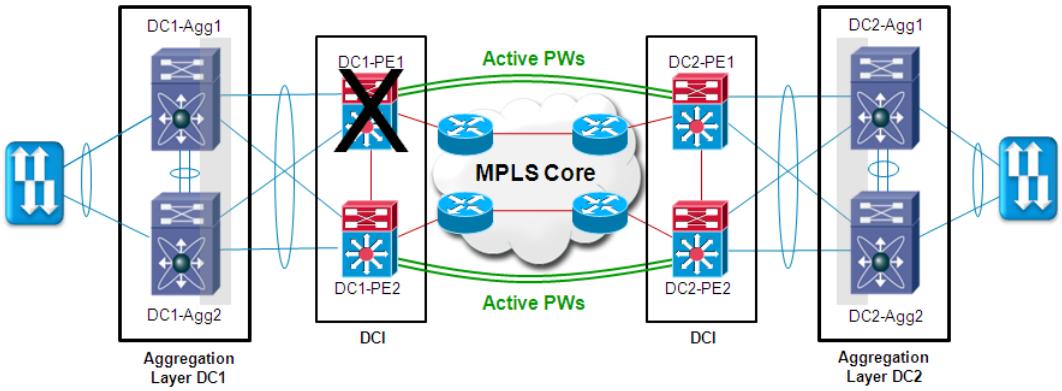
1. As discussed for the recovery scenario in test 1, the highest value obtained with ASR1000 is attributable to the specific not fully mesh connectivity established between the aggregation and the DCI layer. Using full mesh connections reduces the recovery value to the one shown for Catalyst 6500 scenarios and thus it is a recommended best practice.
2. Same as 1
3. Same as 1

4. Same as 1

Test 4b: PE Node Failure/Recovery (Catalyst 6500 Topology)

This failure/recovery scenario is shown in Figure 2-31.

Figure 2-31 PE Node Failure/Recovery (Catalyst 6500)Topology)



Convergence After Failure

- DC1 to DC2 Layer 2 flows: the failure of DC1-PE1 causes the local links connecting to the aggregation layer devices to go down. Therefore, all the traffic that DC1-Agg1 and DC1-Agg2 receives from the access layer devices (via the vPC) will need to be sent across the remaining local vPC link (connecting to DC1-PE2). This is an EtherChannel recovery similar to the local link failure scenario discussed in Test 1 (only now it applies to both aggregation switches).
- DC2 to DC1 Layer 2 flows: when DC1-PE1 fails, both the PWs established with DC2-PE1 go down. This causes DC2-PE1 to bring down (leveraging the EEM script) the local links connecting to both aggregation boxes in DC2. Traffic received by the aggregation devices from the access layer switches in DC2 is then send via the remaining vPC links connecting to DC2-PE2. Once again, this is similar to test 1, only applied to both aggregation layer switches.

Convergence After Recovery

- DC1 to DC2 Layer 2 flows: once the PE device is back online and regains connectivity with the MPLS core and the aggregation layer switches, back to back connectivity between aggregation devices in the remote sites is reestablished. LACP frames are exchanged again, allowing for the bundling to the vPC of the physical links connecting the aggregation layer to DC1-PE1 (or DC2-PE1).
- DC2 to DC1 Layer 2 flows: the reestablishment of the logical PWs would inform DC2-PE1 to reactivate the links toward the aggregation layer. This is key to be able to re-bundle these to the end-to-end vPC (as discussed above). Therefore, it is expected that the convergence impact is mostly the same in both directions.

Table 2-8 captures the results achieved with this specific test case.

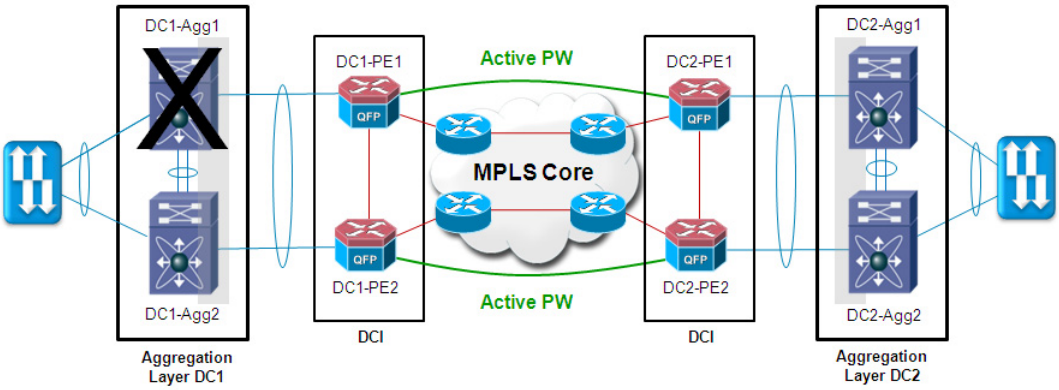
Table 2-8 PE Node Failure/Recovery Results (Catalyst 6500 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (6500 as PE)	0.6 sec	1.16 sec	0.1 sec	0.5 sec
Layer 2 Multicast (6500 as PE)	1 sec	1.13 sec	0.5 sec	0.5 sec

Test 5a: Aggregation Layer Node Failure/Recovery (ASR1000 Topology)

This failure/recovery scenario is shown in Figure 2-32.

Figure 2-32 Aggregation Layer Node Failure/Recovery (ASR1000 Topology)



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: flows originated from devices connected to DC1 access layer and originally sent toward DC1-Agg1 over the vPC logical connection are rehashed on the remaining link connecting to DC1-Agg2. This is the only convergence event required to recover the flows in this direction and it mostly depend on the capability of the access layer switch to perform the re-hashing and obviously independent from the platform deployed in the DCI layer. Testing has shown variable results from 1 to 2.5 seconds.
- **DC2 to DC1 Layer 2 Flows**: the loss of DC1-Agg1 causes the local links toward the DCI layer to fail as well, and this brings down the PW connecting to the remote site. The DCI devices in DC2, leveraging Remote Ethernet Port Shutdown, successively bring down the link to CS2-Agg1. This forces DC2-Agg1 to redirect the traffic on the vPC peer link connecting to DC2-Agg2.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: once the aggregation layer comes back online, the local link between this device and DC1-PE1 also recovers. This allows for the re-establishment of the logical PWs, which causes DC2-PE1 to fully reactivate its local link toward DC2-Agg1. Traffic starts flowing again across the 2 available paths.
- **DC2 to DC1 Layer 2 Flows**: same as DC1 to DC2 Layer 2 Flows.

Table 2-9 captures the results achieved with this specific test case.

Table 2-9 Aggregation Layer Node Failure/Recovery Results (ASR1000 Topology)

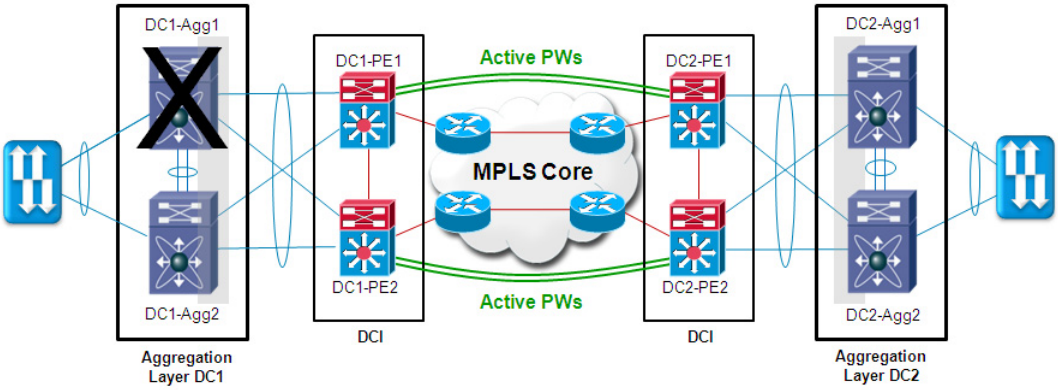
	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	0.9-2.5 sec	1 sec	42 sec[1]	37 sec[2]
Layer 2 Multicast (ASR1000 as PE)	5.5 sec[3]	1 sec	42 sec[4]	37 sec[5]

1. Caused by known N7K recovery issues, which are made worse by not having a full mesh topology between aggregation and DCI layer devices.
2. Same as 1
3. Outage experienced when using Catalyst 6500 in the access layer. This was caused by the time required to reprogram the LTL indexes when sending a high number of Layer 2 multicast flows (1200 in this specific test case).
4. Same as 1
5. Same as 1

Test 5b: Aggregation Layer Node Failure/Recovery (Catalyst 6500 Topology)

This failure/recovery scenario is shown in Figure 2-33.

Figure 2-33 Aggregation Layer Node Failure/Recovery (Catalyst 6500Topology)



Convergence After Failure

- DC1 to DC2 Layer 2 flows: flows originated from devices connected to DC1 access layer and originally sent toward DC1-Agg1 over the vPC logical connection will be rehashed on the remaining link connecting to DC1-Agg2. This is the only convergence event required to recover the flows in this direction and it mostly depend on the capability of the access layer switch to perform the re-hashing.
- DC2 to DC1 Layer 2 flows: the loss of DC1-Agg1 causes the local links toward the DCI layer to fail as well, and this brings down the PWs connecting to the remote site. The DCI devices in DC2, leveraging EEM capabilities, successively bring down the link to their own local aggregation layer devices. This forces a re-hashing of traffic from the aggregation layer to the remaining vPC members connecting to the DCI layer.

Convergence After Recovery

- DC1 to DC2 Layer 2 flows: once the aggregation layer comes back online, the logical PWs eventually are recovered and all the physical links (connecting the DCI and aggregation layers both in DC1 and DC2) are re-bundled together in the end-to-end vPC. Traffic starts flowing again across the 4 available paths.
- DC2 to DC1 Layer 2 flows: same as above.

Table 2-10 captures the results achieved with this specific test case.

Table 2-10 Aggregation Layer Node Failure/Recovery Results (Catalyst 6500 Topology)

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (6500 as PE)	0.9-2.5 sec	2.3 sec	2.4 sec	14 sec[1]
Layer 2 Multicast (6500 as PE)	5.5 sec[2]	2.2 sec	5 sec[3]	14.5 sec[4]

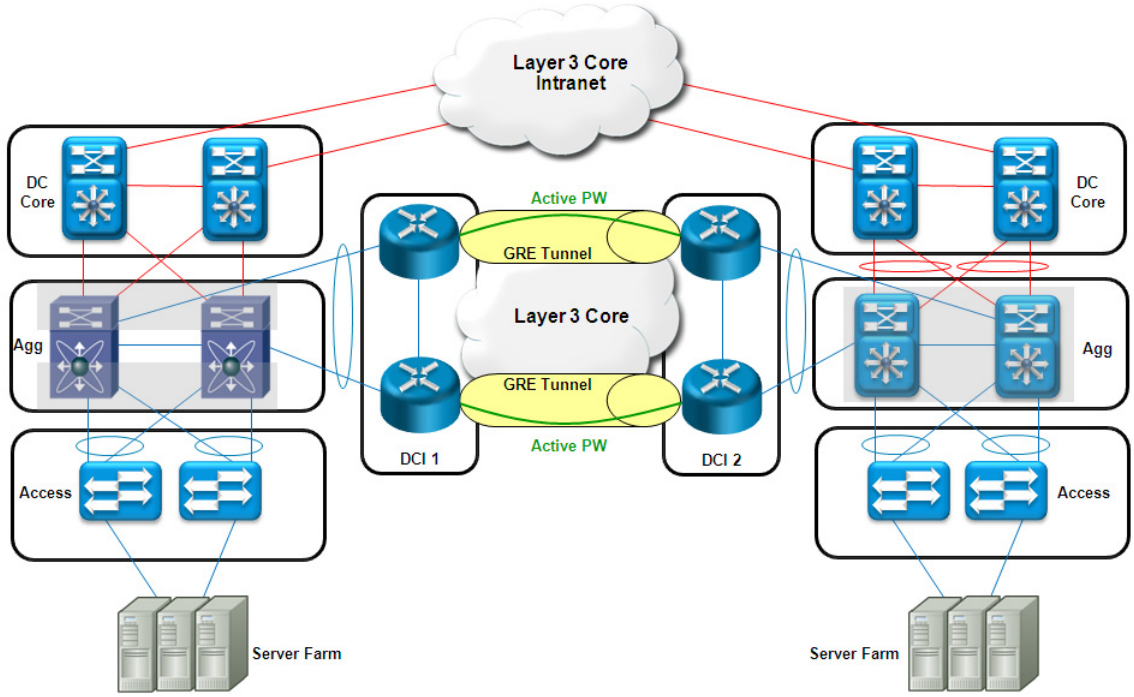
1. Caused by known N7K recovery issues (check with your Cisco representative).
2. Outage experienced when using Catalyst 6500 in the access layer. This was caused by the time required to reprogram the LTL indexes when sending a high number of Layer 2 multicast flows (1200 in this specific test case).
3. Same as 2
4. Same as 2

Deploying Port Based EoMPLS over an IP Core

In some deployment scenarios, the cloud interconnecting the remote data center sites does not allow the creation of native EoMPLS PWs between the enterprise PE devices. This is for example the case when an enterprise does not own the MAN/WAN network infrastructure and acquires network services from a service provider. Even if the SP is actually leveraging an MPLS network to provide connectivity, from an enterprise point of view it offers a Layer 3 service, since typically the enterprise edge devices establish a Layer 3 peering (usually eBGP) with the SP devices.

One common way for the enterprise to work around this problem is to establish a logical overlay between the data center sites and establish the EoMPLS logical PWs across that overlay (instead than across an MPLS enabled cloud). The simplest way of doing so is by leveraging GRE tunnels between the devices in the DCI layer, as highlighted in Figure 2-34.

Figure 2-34 Port Mode EoMPLSoGRE



Note Enabling MPLS connectivity across a GRE tunnel is natively supported on Cisco ASR1000 routers, whereas it requires the use of SIP400 linecard with Catalyst 6500 switches. In the context of this document, only the native solution with ASR1000 devices is discussed.

Comparing Figure 2-34 with Figure 2-3 it is easy to notice how the two designs are essentially identical from the point of view of the LAN extension features:

- Port-based EoMPLS is the technology leveraged to establish the logical PWs
- The connectivity between the DCI layers and the rest of the DC network is the same
- The only difference is represented by that the EoMPLS PW are established by leveraging the logical connectivity provided by the GRE tunnels, instead of by leveraging the underlying physical infrastructure.

The immediate implication is that most of the design considerations discussed for the native MPLS scenario still remain valid here: this applies to end-to-end loop prevention and STP isolation mechanisms, FHRP and inter-DC routing deployments and the use of 802.1AE technology to provide encryption services. Therefore, the focus would be only on the design aspects that uniquely differentiate a GRE-based deployment, starting with the configuration guidelines.

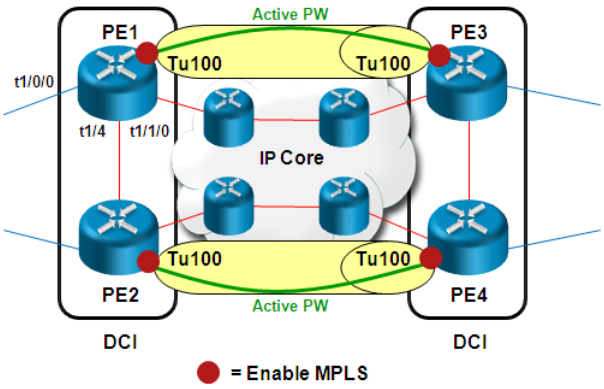
Note Refer to the EoMPLS section of this document for more information about the common design aspects of the solutions.

EoMPLSoGRE Configuration

The MPLS configuration required to enable the EoMPLSoGRE services on the PE devices is fairly straightforward and can be summarized in the following steps:

Step 1 Enable MPLS on the Layer 3 tunnel interfaces established between the PEs in remote DCI layers. This is the main difference when comparing this scenario with the native EoMPLS deployment previously discussed, since the MPLS configuration is now only applied to one logical interface for each PE, as shown in Figure 2-35.

Figure 2-35 Enabling MPLS on GRE Tunnels



This means that the PEs are logically connected back-to-back from an MPLS point of view (there are no devices playing the P role in this scenario).

The required configuration follows:

PE1

```
mpls ldp graceful-restart
mpls ldp session protection
mpls ldp holdtime 15
mpls label protocol ldp
mpls ldp router-id Loopback0 force
!
interface Loopback0
 description LDP connection source
 ip address 11.0.1.31 255.255.255.255
!
interface Loopback100
 description GRE tunnel source
 ip address 12.11.11.11 255.255.255.255
!
interface Tunnel100
 ip address 100.11.11.11 255.255.255.0
 ip mtu 9192
 mpls ip
 tunnel source Loopback100
 tunnel destination 12.11.11.21
```

We recommend leveraging loopback interfaces as source and destination points for establishing the logical GRE connections. The main advantage in doing so is that connectivity across the GRE tunnels can be maintained as long as a physical path connecting the two PE devices is available. In Figure 2-35, should the Layer 3 interface to the IP core fail on the PE1 device, the GRE traffic would be rerouted

across the transit link connecting the two PE routers. As discussed in the "EoMPLSoGRE Failure/Recovery Analysis" section, this would minimize the outage for LAN extension traffic, since the physical link failure would remain transparent to the PW connections.

Note Another scenario where the use of loopback interfaces is useful is when each PE has redundant physical connections toward the IP core.

Step 2 Configure EoMPLS port mode on the PE internal interfaces.

PE1

```
interface TenGigabitEthernet1/0/0
 mtu 9216
 no ip address
 xconnect 11.0.2.31 100 encapsulation mpls
```

The interface configuration is identical to the one already seen for the native MPLS scenario. However, some additional configuration is now required to bind the EoMPLS PW with the GRE tunnels. The simplest way to achieve this is by configuring a static route to specify that the remote loopback IP address should be reached via the logical tunnel interface, as shown in the following configuration:

PE1

```
ip route 11.0.2.31 255.255.255.255 Tunnel100
```

An alternative approach would be to enable a dynamic routing protocol across the GRE tunnels; given that this does not bring any tangible advantage, keep the preceding static configuration.

You can verify how the only LDP neighbors detected in this case is the PE in the remote locations. Notice how the output highlights that the LDP session is established via the GRE tunnel.

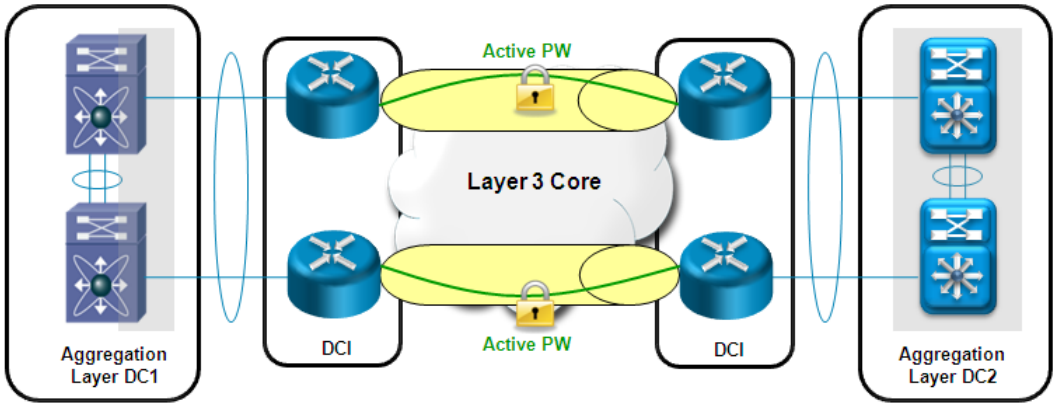
```
PE1#sh mpls ldp neighbor
Peer LDP Ident: 11.0.2.31:0; Local LDP Ident 11.0.1.31:0
    TCP connection: 11.0.2.31.54510 - 11.0.1.31.646
    State: Oper; Msgs sent/rcvd: 18284/18286; Downstream
    Up time: 22:09:37
    LDP discovery sources:
      Tunnel100, Src IP addr: 100.11.11.21
      Targeted Hello 11.0.1.31 -> 11.0.2.31, active, passive
        Addresses bound to peer LDP Ident:
          11.0.2.31       12.11.11.21     22.1.1.1        40.0.20.2
          100.11.11.21
```

IPSec-Based Encryption

When deploying port mode EoMPLS over an IP core, there are essentially two technical alternatives to provide encryption services for LAN extension traffic:

- Using 802.1AE between Nexus 7000 platforms deployed at the aggregation layer of each data center. This is exactly the same deployment already discussed in the "Encryption" section applicable to native MPLS deployments.
- For scenarios where Nexus 7000 is not available on both sites, 802.1AE is not an option and a viable alternative is the use of IPSec, as shown in Figure 2-36.

Figure 2-36 Using IPSec as Encryption Mechanism



The recommended deployment mode for IPSec with ASR1000 platforms is GRE with tunnel protection. The following shows the required configuration:

DC1-PE1

```
crypto isakmp policy 10
 authentication pre-share
crypto isakmp key CISCO address 0.0.0.0 0.0.0.0
!
!
crypto ipsec transform-set MyTransSet esp-3des esp-sha-hmac
crypto ipsec fragmentation after-encryption
!
crypto ipsec profile MyProfile
 set transform-set MyTransSet
!
interface Tunnel100
 ip address 100.11.11.11 255.255.255.0
 tunnel source Loopback100
 tunnel destination 12.11.11.21
 tunnel protection ipsec profile MyProfile
```

Note Another technical alternative would be the deployment of IPSEC Virtual Tunnel Interface (VTI). Given the requirement for a single encrypted tunnel in the context of this specific DCI solution, the scalability characteristics of the VTI approach are not required and the use of tunnel protection is recommended.

MTU Considerations

Every time an IP packet is GRE encapsulated, the overall MTU size of the original packet is obviously increased. This effect is even more marked when deploying IPSec encryption on top of GRE encapsulation. Unfortunately, MPLS MTU configuration and MPLS fragmentation/reassembly are not supported on GRE tunnel interfaces on ASR1000 in 2.5.0 XNE software release, which is the object of this validation.

Support on GRE was added in ASR1000 2.6.0 XNF. Therefore, until a solution is available, the recommendation is to increase the MTU size both on the physical interfaces connecting to the IP core and on the logical tunnel interface. This configuration sample is shown below:

DC1-PE1

```
interface TenGigabitEthernet0/0/0
 mtu 9216
!
interface Tunnel100
 ip mtu 9192
```

EoMPLSoGRE Failure/Recovery Analysis

The testing environment for the EoMPLSoGRE deployment (with or without IPSec) was identical to the one discussed for the native MPLS deployment. This applies to traffic flows established, VLAN and IP routes scalability, and so on. Some of the failure/recovery scenarios are also identical, given that configuration and network topologies from the DCI side toward each DC site is common. This consideration holds true for the following test cases:

- Aggregation to DCI Layer Link Failure/Recovery
- Aggregation Layer Node Failure/Recovery

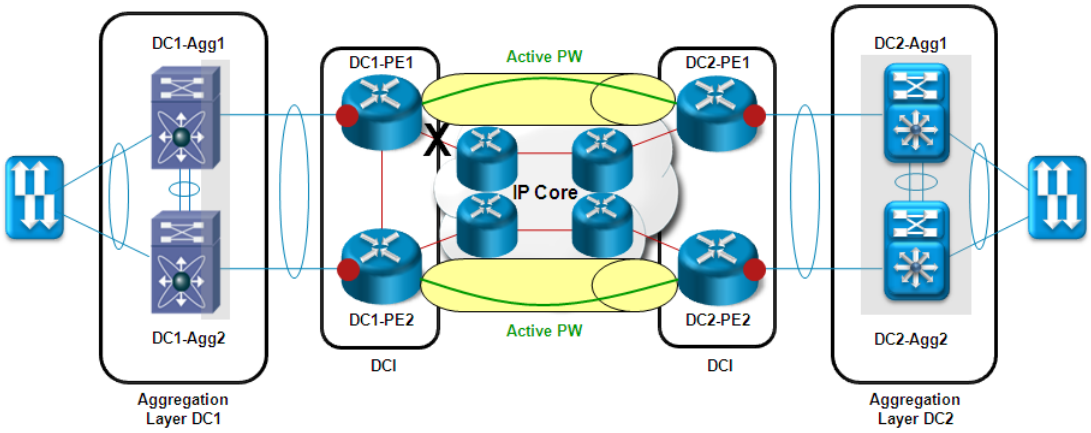
Therefore, only the specific test cases which provide unique design considerations for PW deployments over GRE are discussed below.

Note The results achieved are independent from that IPSec is enabled or not across the GRE logical connections, so the results below apply to both deployment scenarios.

Test 1: DCI Layer to IP Core Failure/Recovery

This failure/recovery scenario is shown in Figure 2-37.

Figure 2-37 DCI Layer to IP Core Failure/Recovery



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the PW established between the top DCI layer devices remain active in this case, since the GRE traffic is rerouted across the transit link interconnecting the two DCI devices in DC1. Therefore, the convergence is exclusively dictated by how fast this traffic rerouting can happen.
- **DC2 to DC1 Layer 2 Flows**: similar considerations are valid in this direction and traffic rerouting is the main responsible for convergence.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: link recovery causes another routing adjustment, so that all GRE traffic is carried across the optimal path (bypassing the transit link between DCI layer devices).
- **DC2 to DC1 Layer 2 Flows**: same as above.

Table 2-11 captures the results achieved with this specific test case.

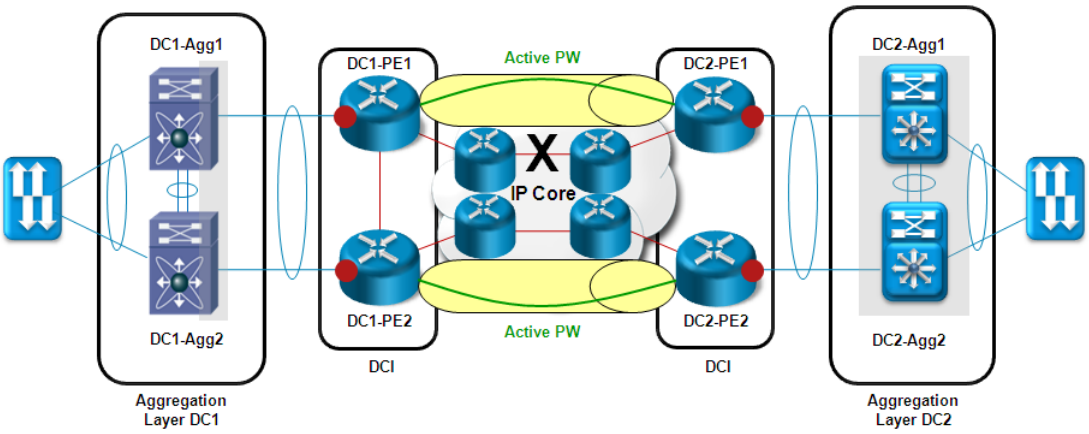
Table 2-11 DCI Layer to IP Core Failure/Recovery Results

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	1.6 sec	0.8 sec	0.8 sec	0.8 sec
Layer 2 Multicast (ASR1000 as PE)	0.8 sec	0.4 sec	0.4 sec	0.4 sec

Test 2: IP Core "Brown Out" Failure/Recovery

This failure/recovery scenario is shown in Figure 2-38.

Figure 2-38 IP Core Brown Out Failure/Recovery



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the failure of the link between core devices would cause a reroute of GRE traffic across the transit link connecting the two PE devices in the DCI layer. This is similar to the test 2 scenario, with the only difference being the DCI layer device triggers an IGP rerouting after receiving an IGP notification from the core router (instead than because of a direct link failure).
- **DC2 to DC1 Layer 2 Flows**: the behavior is identical to what discussed above for the opposite direction.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: link recovery causes in this case another routing adjustment, so that all GRE traffic is carried across the optimal path (bypassing the transit link between DCI layer devices).
- **DC2 to DC1 Layer 2 Flows**: same as DC1 to DC2 Layer 2.

Table 2-12 captures the results achieved with this specific test case.

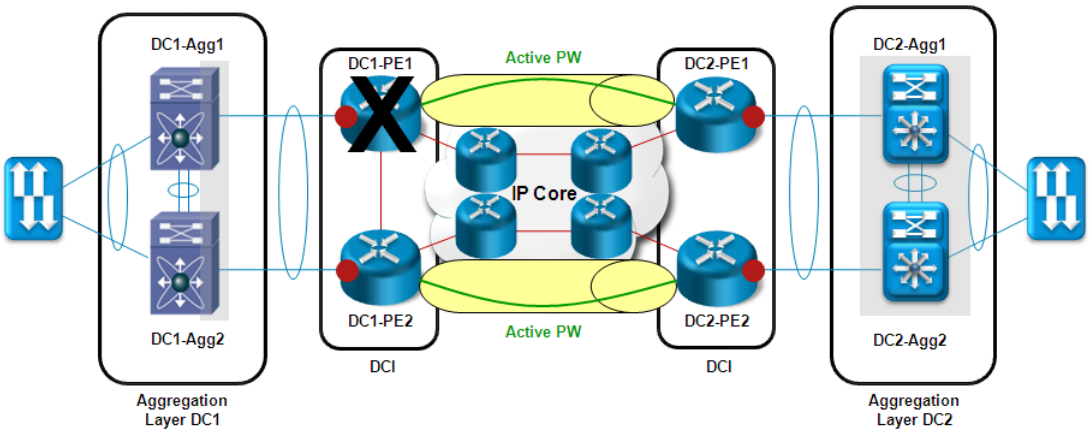
Table 2-12 IP Core Brown Out Failure/Recovery Results

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	1.2 sec	1.1 sec	0.8 sec	0.8 sec
Layer 2 Multicast (ASR1000 as PE)	1.2 sec	1.1 sec	0.4 sec	0.4 sec

Test 3: PE Node Failure/Recovery

This failure/recovery scenario is shown in Figure 2-39.

Figure 2-39 PE Node Failure/Recovery



Convergence After Failure

- **DC1 to DC2 Layer 2 Flows**: the failure of the DC1-PE1 device causes the local link connecting to the aggregation layer devices (DC1-Agg1)to go down. Therefore, all the traffic that DC1-Agg1 still receives from the access layer device (via the vPC) needs to be sent across the vPC peer link (connecting to DC1-Agg2). This is similar to the local link failure scenario discussed in Test 1.
- **DC2 to DC1 Layer 2 Flows**: when DC1-PE1 fails, the PWs established with DC2-PE1 need to go down in order for DC2-PE1 to bring down leveraging Remote Port Shutdown) the local link connecting DC2-Agg1. However, what has been noticed during testing is that the PW remain up until the GRE tunnels stay up. This is independent from having static or dynamic routing configuration across the GRE tunnels. Therefore, it is required to speed up the detection that connectivity across the GRE tunnel is compromised. This can be achieved by tuning the GRE keepalives timers. Unfortunately, the most aggressive setting supported on ASR1000 platforms is 1 second for the keepalives and 3 seconds for the holdtime, as shown in the following configuration:

DC1-PE1

```
interface Tunnel100
 keepalive 1 3
```

The end result is that 3 seconds are added to the total traffic outage.

Convergence After Recovery

- **DC1 to DC2 Layer 2 Flows**: once the PE device is back online and regains connectivity with the IP core and the aggregation layer switch, back-to-back connectivity between aggregation devices in the remote sites is reestablished. LACP frames are exchanged again, allowing for the bundling to the vPC of the physical links connecting the aggregation layer to DC1-PE1 (or DC2-PE1).
- **DC2 to DC1 Layer 2 Flows**: the reestablishment of the logical PW would inform DC2-PE1 to reactivate the links toward the aggregation layer. This is key to be able to re-bundle these to the end-to-end vPC (as discussed above). Therefore, the convergence impact is expected to be mostly the same in both directions.

Table 2-13 captures the results achieved with this specific test case.

Table 2-13 PE Node Failure/Recovery Results

	Failure		Recovery	
Traffic Flows	**DC1 to DC2**	**DC2 to DC1**	**DC1 to DC2**	**DC2 to DC1**
Layer 2 Unicast (ASR1000 as PE)	1.2 sec	5 sec[1]	4.6 sec[2]	4.6 sec[3]
Layer 2 Multicast (ASR1000 as PE)	1.2 sec	5 sec[4]	4.6 sec[5]	4.8 sec[6]

1. As previously mentioned, this was caused by that the EoMPLS PW remained active until the GRE holdtime expired (around 3 seconds).
2. The main reason for these high recovery values was the use of a not fully meshed topology between the aggregation and the DCI layers. As discussed for the native EoMPLS scenario, deploying full mesh connectivity is the recommended design.
3. Same as 2
4. Same as 1
5. Same as 2
6. Same as 2

H-QoS Considerations

Today, more and more enterprise customers connect to their WAN cloud via some type of Ethernet transport, whether it is Ethernet, Fast Ethernet, Gigabit Ethernet or even Ten Gigabit Ethernet. Although the physical interface provides line rate throughput, the reality is that enterprise customers, in many cases, only require (and are willing to pay a service provider for) a fraction of that overall capacity (i.e. a “sub-rate” WAN service).

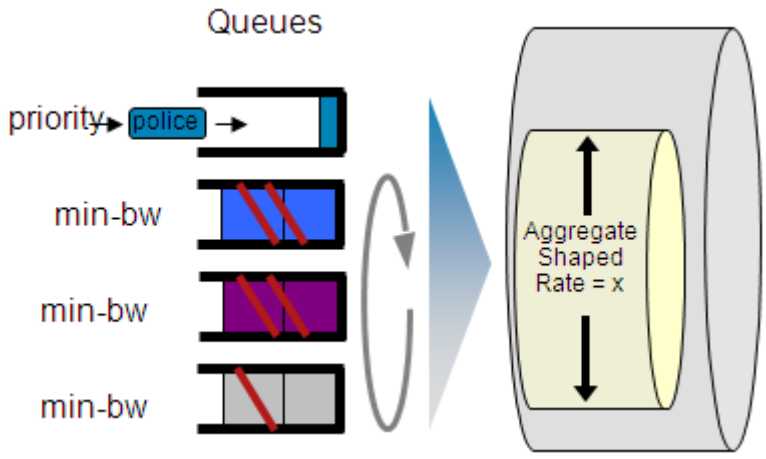
For example, an Enterprise customer requires a 200Mbps WAN access circuit at a given data center site. The WAN circuit may get delivered via a Gigabit Ethernet physical interface connecting to the service provider. The service provider will drop any traffic received from the customer site in excess of the contracted data rate. The main issue is that typically this traffic dropping is done without any distinction for the type of customer traffic involved (i.e. traffic drops will be performed independently of the packet QoS marking done by the customer), unless the customer has specifically contracted a given traffic rate for a given traffic class with the service provider (in which case the service provider may employ class based policer for the customer’s offered traffic).

Enterprise customers need to transport various types of traffic from one data center to another. For example, real time traffic demands low latency and jitter, business critical traffic demands guaranteed delivery, and bulk data transfer traffic demands high throughput and can be bursty. Enterprise customers may want to control prioritization and bandwidth allocation among various types of traffic, control what type of traffic gets delayed or dropped during WAN circuit congestion, and monitor bandwidth usage for this type of traffic.

To address these requirements, Hierarchical Quality of Service (HQoS) is often needed on the DCI WAN edge router to give the Enterprise the traffic control and visibility they desire. As shown in Figure 2-40, a typical HQoS policy has two components:

- A parent shaper leveraged to shape traffic down to the contracted WAN access rate.
- A second QoS child policy used to prioritize traffic and allocate bandwidth to various types of traffic.

Figure 2-40 H-QoS for Sub-rate Access



The QoS capabilities that can be enabled on the DCI WAN edge router depend on the type of encryption technology deployed:

- For 802.1AE deployments, all the traffic sent toward the DCI WAN edge router from the aggregation switches is already encrypted and there is no visibility into the CoS field of the Layer 2 frames or IP precedence/DSCP bits in the IP payload. The consequence is that shaping to contracted sub-rate WAN access speed is still possible in this case, but all the traffic is treated equally from a prioritization perspective.
- For EoMPLS without 802.1AE encryption or EoMPLSoGRE (with or without ipsec protection) deployments, both shaping and prioritization is instead possible and this is the focus of this section.

Deploying H-QoS with ASR1000

The first basic assumption is that traffic is properly marked and classified at the data center access layer. This is usually done setting Layer 2 CoS bits or Layer 3 IP precedence/DSCP bits. Hierarchical QoS capabilities can then be enabled by applying ingress and egress QoS policies.

- An ingress QoS policy can be applied to the interface connecting to the aggregation layer device (in the rest of this section this can also be named the attachment circuit or the xconnect interface). This ingress QoS policy can classify traffic based on Layer 2 CoS or Layer 3 IP precedence/DSCP information. Alternatively, Cisco ASR1000 devices can also reclassify traffic that matches a specifically configured Access Control List (ACL). After successful traffic classification, QoS-group or MPLS EXP bit (or both) can then be set by the ingress QoS policy: this setting can then be used on the egress interface toward the core to classify traffic for the purpose of prioritizing and allocating bandwidth for each traffic type.
- In the egress QoS policy, it is possible to set the MPLS EXP bit for the EoMPLS use case (if this is not done already on ingress) or set the GRE outer header IP precedence/DSCP field in the EoMPLSoGRE (oIPsec) use case in order to preserve QoS marking across the MPLS/IP core. It is

important to highlight that once MPLS encapsulation is added in both EoMPLS and EoMPLSoGRE(oIPSec) scenarios, it is not possible to classify traffic based on the Layer 2 CoS bits or Layer 3 IP precedence/DSCP bits in the original Ethernet frame on the egress interface; this is because only the MPLS header is visible to the egress QoS policy.

During testing with ASR1000, it has been observed that in the EoMPLS port mode deployment, CoS bits from the original Ethernet frames received on the attachment circuit are copied by default to the MPLS EXP bits. In the EoMPLSoGRE port mode use case, the MPLS EXP value is further copied by default to the GRE outer header TOS byte (IP precedence). Given that this default behavior may change over time, it is always recommended to configure an ingress QoS policy to explicitly mark QoS-group or MPLS EXP bits based on the original CoS bits value.

On ASR1000 DCI WAN edge router egress core facing physical interfaces, 2 level Hierarchical QoS policy can be applied to have a parent location shaper policy to shape traffic to contracted sub-rate and a child policy with LLQ for priority traffic and per class bandwidth allocation for non-priority traffic classes.

- For the EoMPLS scenario, classification for the child policy can be based on QoS-group or MPLS EXP bit set in the ingress QoS policy on the Ethernet xconnect interface.
- In the EoMPLSoGRE case, 2 level Hierarchical egress QoS policy can be applied to either the GRE tunnel interface or the core facing physical interface, but not both.

Note Supporting a QoS policy with queuing feature on both GRE tunnel interface and physical interface is planned for ASR1000 future software releases.

If qos-preclassify is enabled on the GRE tunnel interface and an egress QoS policy is attached to the core facing physical interface, classification can be based on qos-group or MPLS EXP bits previously set on the ingress QoS policy (the inner header).

If qos-preclassify is not enabled on the GRE tunnel interface and the egress QoS policy is attached to core facing physical interface, classification can only be based on qos-group set on the ingress QoS policy.

If the QoS policy is attached to the GRE tunnel interface, classification can be based on qos-group or MPLS EXP bits set on the ingress QoS policy.

Setting qos-group on ingress policy and use the qos-group for classification on egress QoS policy is recommended as this works in all the following use cases:

- EoMPLS
- EoMPLSoGRE when QoS is applied to the physical or tunnel interface
- EoMPLSoGRE when pre-classification is enabled on the GRE tunnel interface

Once the traffic made it to the remote data center DCI WAN edge router, the WAN edge router in the remote data center can optionally have an egress QoS policy on the Ethernet xconnect interface (connecting to the aggregation layer device) to prioritize and allocate bandwidth to various types of traffic. You can use the Ethernet frame payload CoS bit or IP packet payload IP precedence/DSCP bits to classify traffic. Alternatively, if desired, remote data center DCI WAN edge router can also remark traffic on coring facing interface based on MPLS EXP bit (in EoMPLS case) or GRE outer header IP precedence or DSCP bit (in EoMPLSoGRE case). However, in most case, this is not really necessary as traffic is going from slower WAN links (e.g. sub-rate GE) to faster LAN links (e.g. 10GE) so congestion is not expected.

IPSec and H-QoS Specific Considerations

When deploying IPSec, it is important to keep in mind that the ASR1000 router supports the crypto Low Latency Queuing (LLQ) functionality, which separates traffic that needs encryption/decryption services into a high priority queue and a low priority queue. Traffic classified into high priority queue in the crypto chip will be given strict priority access for encryption/decryption.

This functionality is important in scenarios where the amount of DCI traffic is likely to oversubscribe the ASR1000 ESP crypto bandwidth. By default, crypto LLQ uses the egress policy from the same physical interface or tunnel interface where encryption is enabled to classify priority traffic. Traffic classified as priority traffic in the egress QoS policy on the physical interface or tunnel interface will be sent to the crypto chip priority queue. All other traffic will be sent to the crypto chip low priority queue.

Attention should be paid when QoS and encryption are not configured on the same interface (physical or logical); this is for example the case when the **tunnel ipsec protection** is configured on the tunnel interface and an egress QoS policy is applied to the physical interface, or when a crypto map is enabled on the physical interface and the egress QoS policy is applied to the tunnel interface. In those scenarios, in order for crypto LLQ to correctly prioritize the traffic, you need to use an ingress QoS policy on the Ethernet xconnect interface and an egress control plane policy for control plan traffic to set the qos-group for important traffic and then use the platform CLI **platform ipsec llq qos-group** <*qos-group*> command to link the qos-group to crypto high priority queue.

Note You can use **multiple platform ipsec llq** commands with different qos-group values to link multiple qos-groups to the crypto high priority queue.

When deploying H-QoS in conjunction with IPSec, it is important to keep in mind the interactions with the IPSec anti-replay functionality that is enabled by default on ASR1000 platforms and it is used to provide anti-replay protection against an attacker duplicating encrypted packets. Basically, the encrypting router assign a unique sequence number to each encrypted packet (in an increasing order). The receiving router keeps track of the highest sequence number (X in this example) attached to frames that it has already decrypted. It also considers a value N, representing the window size (64 packet by default). Any duplicate packet or packet with a sequence number lower than X-N is discarded to protect against replay attacks.

H-QoS can reorder packets after an IPSec sequence number is already assigned to them. This may trigger the anti-replay functionality within IPSec, causing encrypted traffic to be dropped at the receiving IPSec router. As previously mentioned, the anti-replay window size is set to 64-packet by default and this value may not be sufficient when H-QoS is enabled. A couple of workarounds are available for scenarios affected by this specific issue:

- Expand the anti-replay window size on the receiving ASR1000 router by leveraging the global CLI command **crypto ipsec security-association replay window-size** <*window-size*>. This allows the receiving router to keep track of more than 64 packets. ASR1000 supports a max anti-replay window size of 512 packets.
- In some cases, it may be necessary to disable anti-replay completely using the global command **crypto ipsec security-association replay disable** to avoid packet getting dropped due to interactions with H-QoS.

ASR1000 can provide multiple Gbps throughput with both hierarchical QoS and IPSec encryption enabled. ASR1000 supports up to 4000 unique QoS policy and class-maps per system, up to 256 class-maps per QoS policy and up to 128000 hardware queues.

A sample ASR1000 DCI WAN router Hierarchical QoS configuration for both EoMPLS and EoMPLSoGREoIPSec use cases is shown below:

```
class-map match-any COS_REAL_TIME
```

```
  match cos 5
class-map match-any COS_NETWORK_CONTROL
  match cos 7
  match cos 6
class-map match-any COS_BUSINESS_CRITICAL
  match cos 4
class-map match-any COS_BUSINESS
  match cos 3
  match cos 2
  match cos 1
!
class-map match-any PRIORITY_CONTROL TRAFFIC
 match ip precedence 6
class-map match-any QOS_GROUP_6
 match qos-group 6
class-map match-any QOS_GROUP_5
 match qos-group 5
class-map match-any QOS_GROUP_4
 match qos-group 4
class-map match-any QOS_GROUP_3
 match qos-group 3
!
policy-map CoPP
 class PRIORITY_CONTROL TRAFFIC
   set qos-group 6
!
control-plane
 service-policy input CoPP
!
! The following three"platform ipsec llq qos-group" commands are only needed in
!the EoMPLSoGREoIPsec user case
platform ipsec llq qos-group 4
platform ipsec llq qos-group 5
platform ipsec llq qos-group 6
!
policy-map LAN_INGRESS_POLICY
  class COS_NETWORK_CONTROL
    set qos-group 6
    set mpls experimental imposition 6
! You only need to set MPLS EXP bit in the EoMPLS use case
  class COS_REAL_TIME
    set qos-group 5
    set mpls experimental imposition 5
! You only need to set MPLS EXP bit in the EoMPLS use case
  class COS_BUSINESS_CRITICAL
    set qos-group 4
    set mpls experimental imposition 4
! You only need to set MPLS EXP bit in the EoMPLS use case
  class COS_BUSINESS
    set qos-group 3
    set mpls experimental imposition 3
! You only need to set MPLS EXP bit in the EoMPLS use case
  class class-default
    set mpls experimental imposition 0
! You only need to set MPLS EXP bit in the EoMPLS use case
!
policy-map CHILD_POLICY ! Egress Child Policy
  class QOS_GROUP_5
    set ip precedence 5
! You can only set ip precedence in the EoMPLSoGREoIPSEC use case
priority percent 10
! Priority traffic can be policed with an explicit or conditional policer.
! Explicit policer is recommended
  class QOS_GROUP_6
```

```
    set ip precedence 6
! You can only set ip precedence in the EoMPLSoGREoIPSEC use case
bandwidth percent 5
! bandwidth or bandwidth remaining ration statement can also be used. However, you cannot
! mix bandwidth/bandwidth percent statement (for minimum bandwidth gurantee) with
! bandwidth remaining ratio statement(for excessive bandwidth allocation) in the same
! policy, even across different classes.
  class QOS_GROUP_4
    set ip precedence 4
! You can only set ip precedence in the EoMPLSoGREoIPSEC use case
    bandwidth percent 30
  class QOS_GROUP_3
    set ip precedence 3
! You can only set ip precedence in the EoMPLSoGREoIPSEC use case
    bandwidth percent 30
  class class-default
    set ip precedence 0
! You can only set ip precedence in the EoMPLSoGREoIPSEC use case
bandwidth percent 25
!
policy-map WAN_EGRESS_POLICY ! Egress Parent Policy
  class class-default
    shape average 200000000  ! Contracted sub-rate
    service-policy CHILD_POLICY
!
interface GigabitEthernet0/0/0
 description xconnect interface
 service-policy input LAN_INGRESS_POLICY
!
interface GigabitEthernet0/0/1
 description WAN Connection
 service-policy output WAN_EGRESS_POLICY
```

Multipoint Topologies

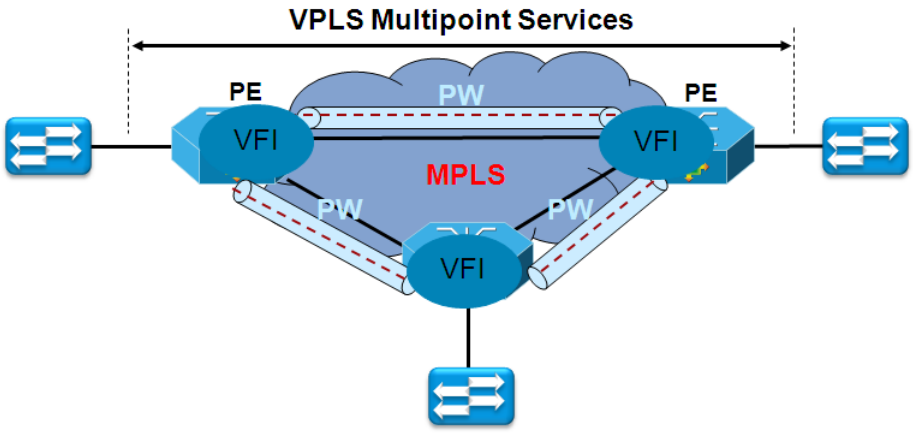
All the solutions discussed until now are usually positioned to provide LAN extension features in point-to-point deployments, where only two remote data center sites are interconnected. Sometimes, these types of services need to be extended to multiple data center locations. Adopting port-based EoMPLS technologies in these cases would essentially mean re-iterating multiple times the deployment previously discussed, with obvious implications on the overall scalability and overall complexity of the solution. In addition, the use of a full mesh of point-to-point EoMPLS PWs would lead to the creation of looped topology, that would require the need to send STP across the logical circuit, not allowing to provide the required STP isolation between data center locations.

The use of Virtual Private LAN Service addresses these issues. VPLS is an architecture that uses MPLS as a transport for multipoint Ethernet LAN services, often referred to as transparent LAN services (TLS), across geographically dispersed locations. To address the issue of multipoint Ethernet capability, the Internet Engineering Task Force (IETF) authored VPLS to describe the concept of linking virtual Ethernet bridges using MPLS Pseudowires (PWs).

At a basic level, VPLS is a group of virtual switch instances (VSI, also called VFI) that are interconnected using Ethernet over MPLS (EoMPLS) circuits in a full-mesh topology to form a single, logical bridge. This functionality provides the multipoint capabilities to this technology. The concept of VSI is similar to a normal bridge device: VPLS forwards Ethernet frames at Layer 2, dynamically learns source MAC address-to-port associations, and forwards frames based on 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 as if a device were attached to an Ethernet switch by linking VSIs using MPLS PWs to form an "emulated" Ethernet switch (Figure 2-41).

Figure 2-41 VPLS Multipoint Services



Because of the flooding behavior, loops in bridged networks are disastrous. There is no counter, such as the Time-To-Live (TTL) field, in an IP header at the frame level to indicate how many times a frame has circulated a network. Frames continue to loop until an entire network is saturated and the bridges can no longer forward packets. To prevent loops in a network, bridges or switches use Spanning Tree Protocol (STP) to block any ports that might cause a loop. In this document, the goal is to limit the use of STP to each local data center site, so we will discuss how to achieve end-to-end loop avoidance by leveraging alternative methods (mostly leveraging the Embedded Event Manager features available on Catalyst 6500 and Nexus 7000 devices).

VPLS consists of three primary components:

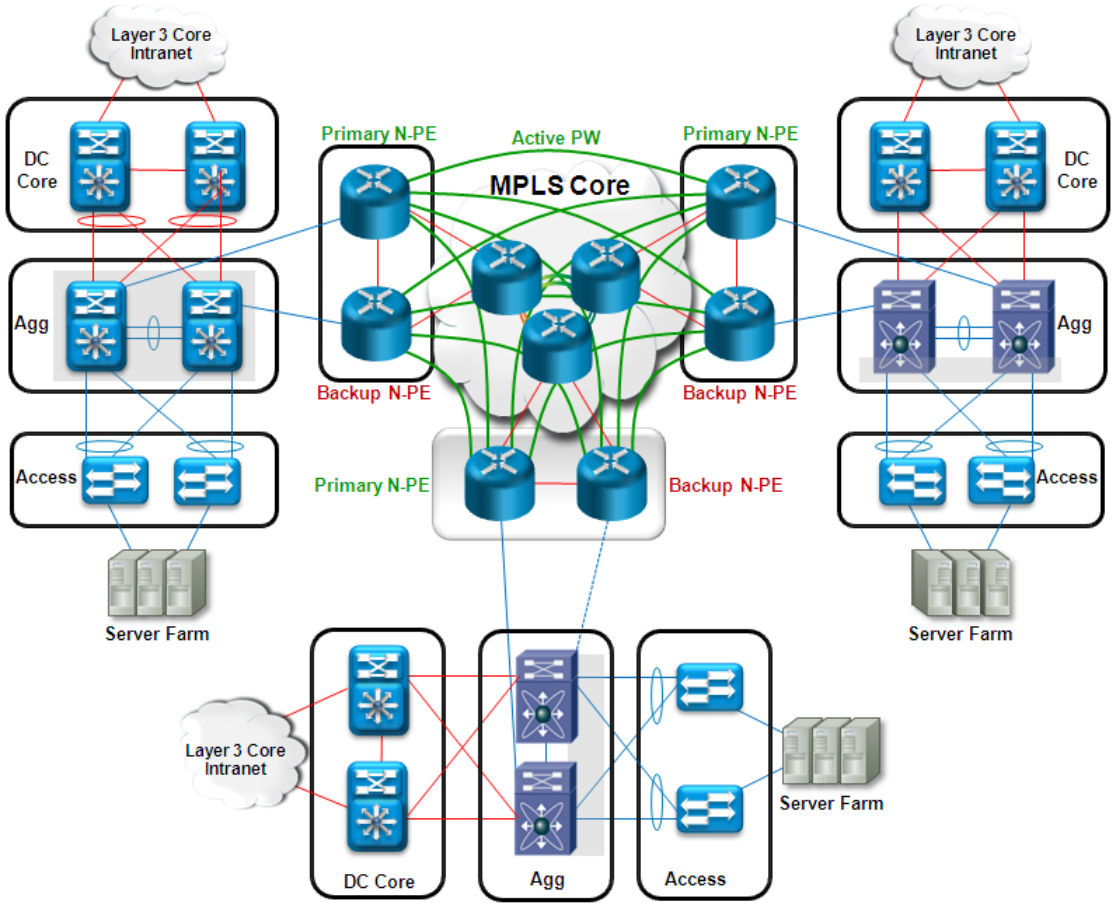
- Attachment circuits: connections between the Network-facing Provider Edges (N-PEs) devices and the customer edge (CE) or aggregation switches. Also, the assumption is that the N-PE are devices under the administrative control of the enterprise that needs the LAN extension features; those devices are not part of the SP network.
- Virtual circuits (VCs or PWs): connections between N-PEs across MPLS network based on draft-martini-l2circuit-trans-mpls-11
- VSI: A virtual Layer 2 bridge instance that connects attachment circuits to VCs.

Note An in depth discussion on VPLS functions is beyond the scope of this document. You can find more information in the Cisco Press book entitled, *Interconnecting Data Centers using VPLS* (ISBN: 978-1-58705-992-6).

Deploying VPLS

Figure 2-42 shows the network topology that was used to validate VPLS as part of DCI phase 2.

Figure 2-42 Validated VPLS Topology



The following are some initial considerations based on the topology in Figure 2-42:

- The PE devices performing the VPLS features are deployed in a dedicated DCI layer. The VLANs that require extension services between the data centers needs to be extended to the DCI layer leveraging Layer 2 trunk connections (the routing point for these VLANs is still positioned at the aggregation layer).
- A full mesh of VPLS pseudowires (PWs) is established between the various PE devices, with the exception of the PW between PEs that belong to the same data center. This additional PW may be required only in scenarios where multiple aggregation layer blocks are deployed inside each data center site; this is not what was validated in the preceding topology so it is beyond the scope of this document.
- All the PWs are active at the same time. Because of this, despite the VPLS default split-horizon behavior, end-to-end loops may be created as discussed in End-to-End Loop Prevention and STP Isolation, page 2-8.
- A Layer 3 link is recommended between the two PEs deployed in the same data center location, to ensure traffic can be rerouted across it in case one PE loses its Layer 3 connection to the core (for additional considerations refer to VPLS Failure/Recovery Scenarios, page 2-80).
- From a platform point of view, the solution validated in the context of this document leveraged Catalyst 6500 switches deployed as PE devices in each data center. VPLS is not supported natively with Sup720 supervisor series. Therefore, SIP modules are leveraged to perform the VPLS

functionality. To achieve 10G connectivity, SIP-600 linecards were leveraged in the topology shown in Figure 2-42 to connect the PE device to the MPLS core and for the transit link between PE devices.

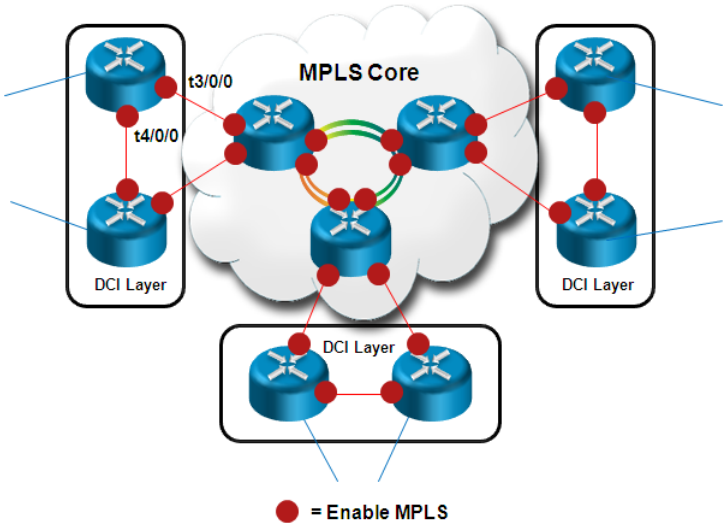
Note VPLS is supported on Catalyst 6500 platforms that also leverage SIP-400 linecards.

VPLS Basic Configuration

The MPLS configuration required to enable the VPLS services on the PE devices is defined in the following steps:

Step 1 Enable MPLS on the Layer 3 links belonging to the PE. This needs to be done both on the physical interface connecting to the MPLS cloud and on the transit link to the peer PE device, as shown in Figure 2-43.

Figure 2-43 Enabling MPLS on Layer 3 Interfaces



The following shows the required CLI commands:

DC1-PE1

```
mpls ldp graceful-restart
mpls ldp session protection
mpls ldp holdtime 60
mpls ldp router-id Loopback0 force
!
interface Loopback0
 description OSPF and LDP Router_ID
 ip address 11.0.1.101 255.255.255.255
!
interface TenGigabitEthernet3/0/0
 description SIP600 Link to the MPLS Core
 mtu 9216
 ip address 99.1.1.1 255.255.255.0
 mpls label protocol ldp
 mpls ip
```

```
!
interface TenGigabitEthernet4/0/0
 description DC1-DCI-6K2 T8/0/0
 mtu 9216
 ip address 99.1.3.1 255.255.255.0
 mpls label protocol ldp
 mpls ip
```

MPLS needs to be enabled only on SIP-600 interfaces (connecting to the MPLS core and to the peer PE device). Turning on MPLS services on the Layer 3 links enables also LDP (Label Distribution Protocol) with the neighbor devices. LDP is required to exchange the MPLS label information required for packets that need to be sent across the MPLS cloud.

Also, since fragmentation within an MPLS network is not allowed, we recommend that you increase the MTU on the MPLS-enabled interfaces. Every IP packet that needs to be sent across an EoMPLS PW requires two MPLS labels, which means that the configured MTU should be at least 8 bytes bigger than the largest IP packet that needs to be sent across the cloud. When possible, we recommend that you configure the maximum MTU size of 9216, as shown in the preceding configuration sample.

Note The MPLS configuration required on the P devices deployed in the MPLS core is similar to the preceding one, since the only basic requirement is to enable MPLS on their Layer 3 interfaces.

When MPLS enabled on the physical Layer 3 interfaces, LDP neighbors are established. You can verify the existence of peer neighbors using the following commands:

```
PE1#sh mpls ldp neighbor
   Peer LDP Ident: 11.0.1.102:0; Local LDP Ident 11.0.1.101:0
    TCP connection: 11.0.1.102.14975 - 11.0.1.101.646
    State: Oper; Msgs sent/rcvd: 252172/252151; Downstream
    Up time: 1w1d
    LDP discovery sources:
      Targeted Hello 11.0.1.101 -> 11.0.1.102, active, passive
      TenGigabitEthernet8/0/0, Src IP addr: 99.1.3.2
        Addresses bound to peer LDP Ident:
          11.0.1.102      11.0.1.11       11.0.1.2        10.0.5.8
          101.15.160.2    99.1.2.1        77.1.2.1        99.1.3.2
   Peer LDP Ident: 99.99.99.1:0; Local LDP Ident 11.0.1.101:0
    TCP connection: 99.99.99.1.51308 - 11.0.1.101.646
    State: Oper; Msgs sent/rcvd: 132123/128125; Downstream
    Up time: 2d04h
    LDP discovery sources:
      TenGigabitEthernet3/0/0, Src IP addr: 99.1.1.2
        Addresses bound to peer LDP Ident:
          99.99.99.1      10.0.5.9        99.0.13.1       99.0.12.1
          99.1.2.2        99.1.1.2
```

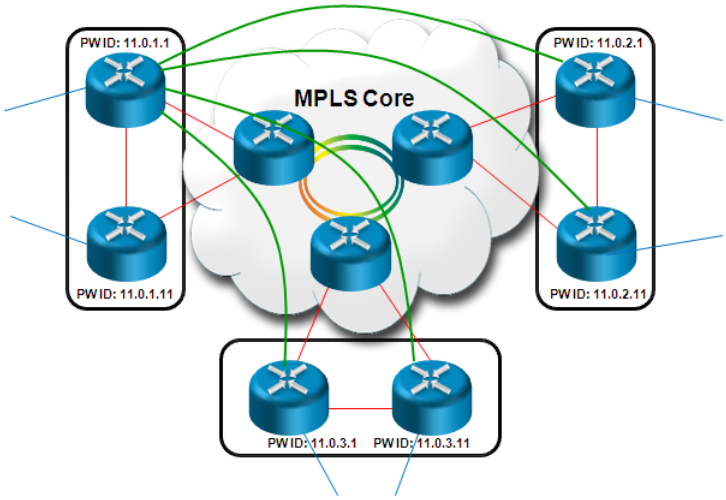
Step 2 To configure VPLS on the Catalyst 6500 platforms, enable the functionality called *xconnect SVI* (interface VLAN). Basically, for each VLAN that needs to be extended, two set of commands are required: the first one allows defining the Virtual Forwarding Instance (VFI) that identifies the group of PWs that form the emulated Ethernet switch interconnecting the remote sites. The seconds set of commands, maps the defined VFI to a specific VLAN.

DC1-PE1

```
xconnect logging pseudowire status
!
interface Loopback100
 description **** PW-loopback ****
 ip address 11.0.1.1 255.255.255.255
!
```

```
l2 vfi vfi_10 manual
 vpn id 10
 neighbor 11.0.3.11 encapsulation mpls
 neighbor 11.0.2.11 encapsulation mpls
 neighbor 11.0.3.1 encapsulation mpls
 neighbor 11.0.2.1 encapsulation mpls
!
interface Vlan10
 mtu 9216
 no ip address
 xconnect vfi vfi_10
```

Figure 2-44 Establishing VPLS PWs between PE Devices



The following considerations are based on the preceding configuration and Figure 2-44:

- Each PE devices defines a specific loopback interface (loopback100) to be used as "PW ID". This is the IP address of the remote PE devices that will be specified during the configuration of the VFI. Because of the use of "LDP session protection," all PEs will establish targeted LDP sessions to the remote loopback IP addresses.
- Each PE establishes 4 PWs with the PE in remote data center sites. A PW between the PEs local to a specific site is not required; as already mentioned, this is because a single aggregation block is deployed inside each data center.
- The VFI and SVI configurations need to be replicated for each VLAN that needs to be extended across the VPLS cloud. This affects the scalability and operational aspects of the solution. Above a certain number of VLANs, we recommend that you migrate to a hierarchical VPLS (H-VPLS) configuration. In our test scenario, we used VPLS configuration to extend up to 300 VLANs between data centers. Discussing H-VPLS alternatives is beyond the scope of this document.
- In contrast to the EoMPLS port mode deployment previously discussed, you must configure the PE internal interfaces connecting to the aggregation layer devices as regular Layer 2 trunk interfaces, carrying all the VLANs that need to be extended via VPLS to the remote locations. From a control plane perspective, all the BPDU frames (STP, CDP, LACP, etc) received from the aggregation layer are handled on the PE itself, so there is no capability of transparently "tunneling" them to the remote locations.

For each VLAN that needs to be extended across the MPLS cloud, a separate VC ID is allocated, as shown in the following output.

```
PE1# show mpls l2transport vc
```

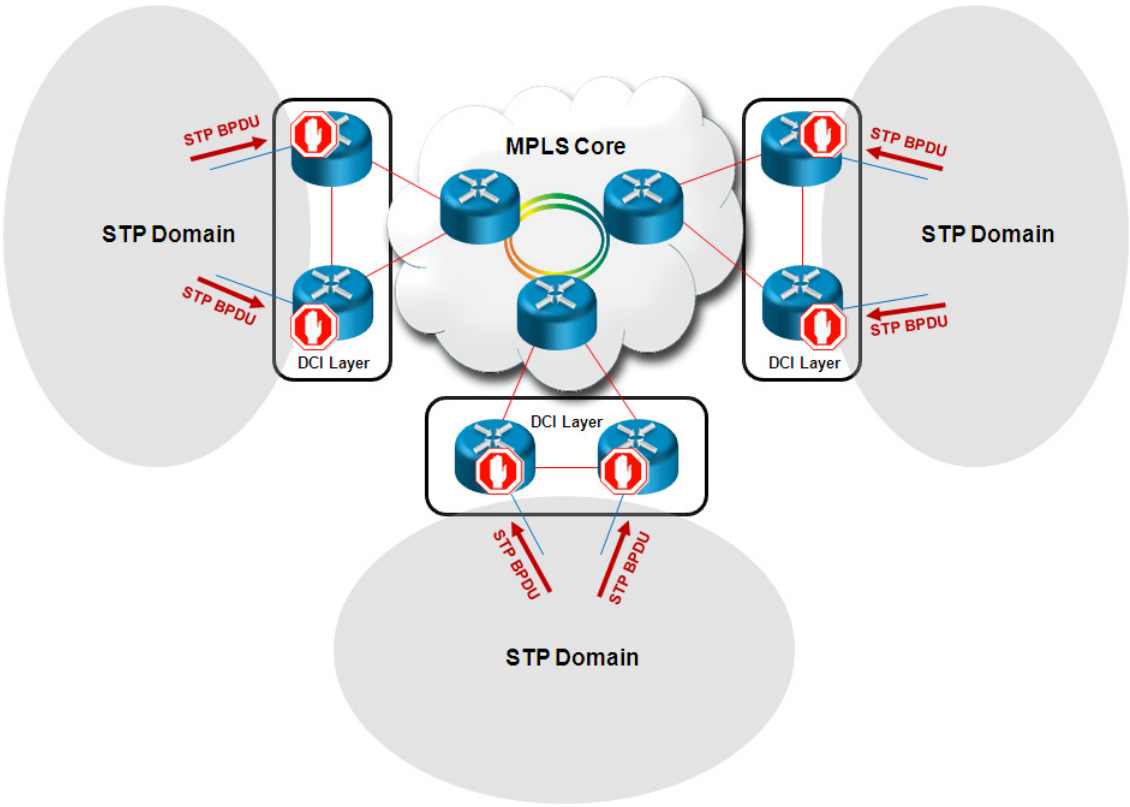
```
Local intf Local circuit Dest address VC ID Status
---------- ------------- ------------ ----- ------
VFI vfi_10 VFI           11.0.2.1     10    UP
VFI vfi_10 VFI           11.0.2.11    10    UP
VFI vfi_10 VFI           11.0.3.1     10    UP
VFI vfi_10 VFI           11.0.3.11    10    UP
VFI vfi_11 VFI           11.0.2.1     11    UP
VFI vfi_11 VFI           11.0.2.11    11    UP
VFI vfi_11 VFI           11.0.3.1     11    UP
VFI vfi_11 VFI           11.0.3.11    11    UP
```

Note The destination address is the same (loopback address of the remote PEs), which means all VCs are carried across the common set of logical PWs.

STP Isolation and End-to-End Loop Prevention

VPLS is a bridging technique that relies on PWs to interconnect point-to-multipoint virtual bridges. Because VPLS is natively built with an internal mechanism known as split horizon, the core network does not require STP to prevent Layer 2 loops. This is the reason why by default STP frames received on an internal interface of a PE are not sent across the VPLS PW, satisfying by definition the STP isolation requirement (Figure 2-45).

Figure 2-45 STP Domain Isolation



Given the behavior exhibited in Figure 2-45, there is no value in sending STP BPDUs between the aggregation and the DCI layer in each local data center site. Therefore, the same BPDU filter configuration already discussed for EoMPLS deployment should be applied on the aggregation layer interfaces facing the DCI layer, as shown in the following configuration:

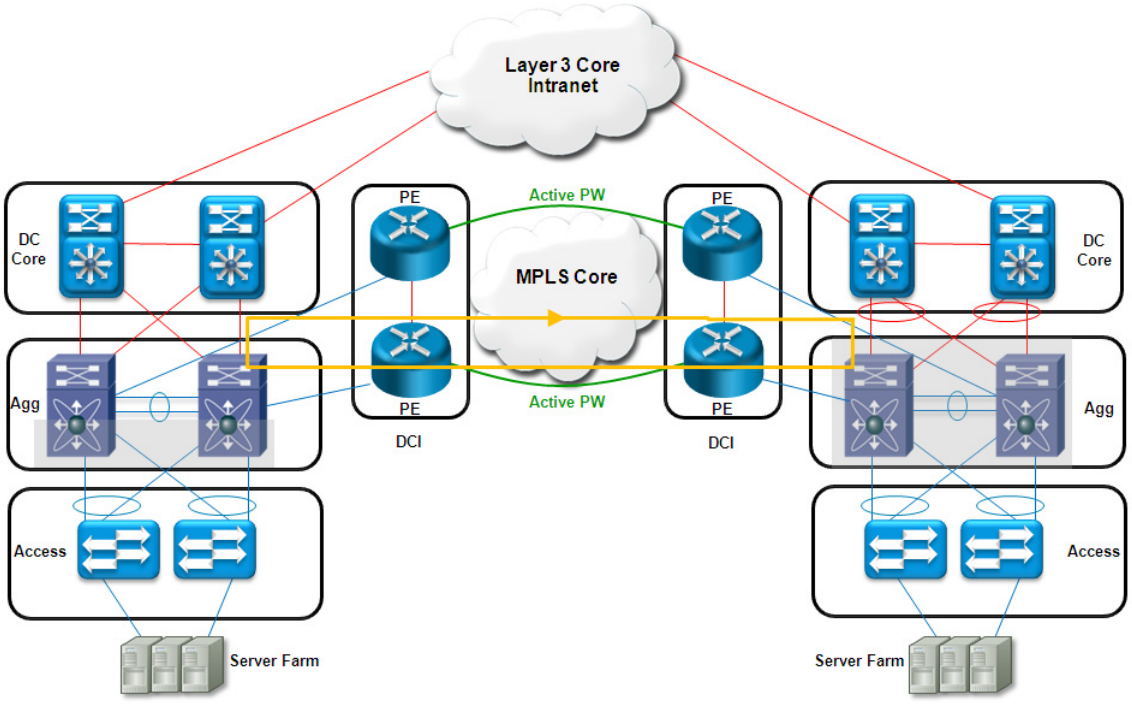
DC1-PE1

```
interface Ethernet1/1
  description L2 Trunk to DCI Layer
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
```

Note Depending on the specific platform deployed at the aggregation layer (Nexus 7000 or Catalyst 6500), the preceding configuration would need to be applied to a physical interface or to a PortChannel, as will be clarified when discussing loop avoidance in Deploying EEM for Loop Avoidance with Nexus 7000 in Aggregation, page 2-60.

You should consider that VPLS split horizon prevents traffic from looping from one PW to another, but this ensures the creation of loop-free topologies only if a single PE is used at each data center site. With multi-homing deployments leveraging redundant PEs in the DCI layer, there is the possibility of creating an end-to-end loop with traffic circling back via the aggregation switches as depicted in Figure 2-46.

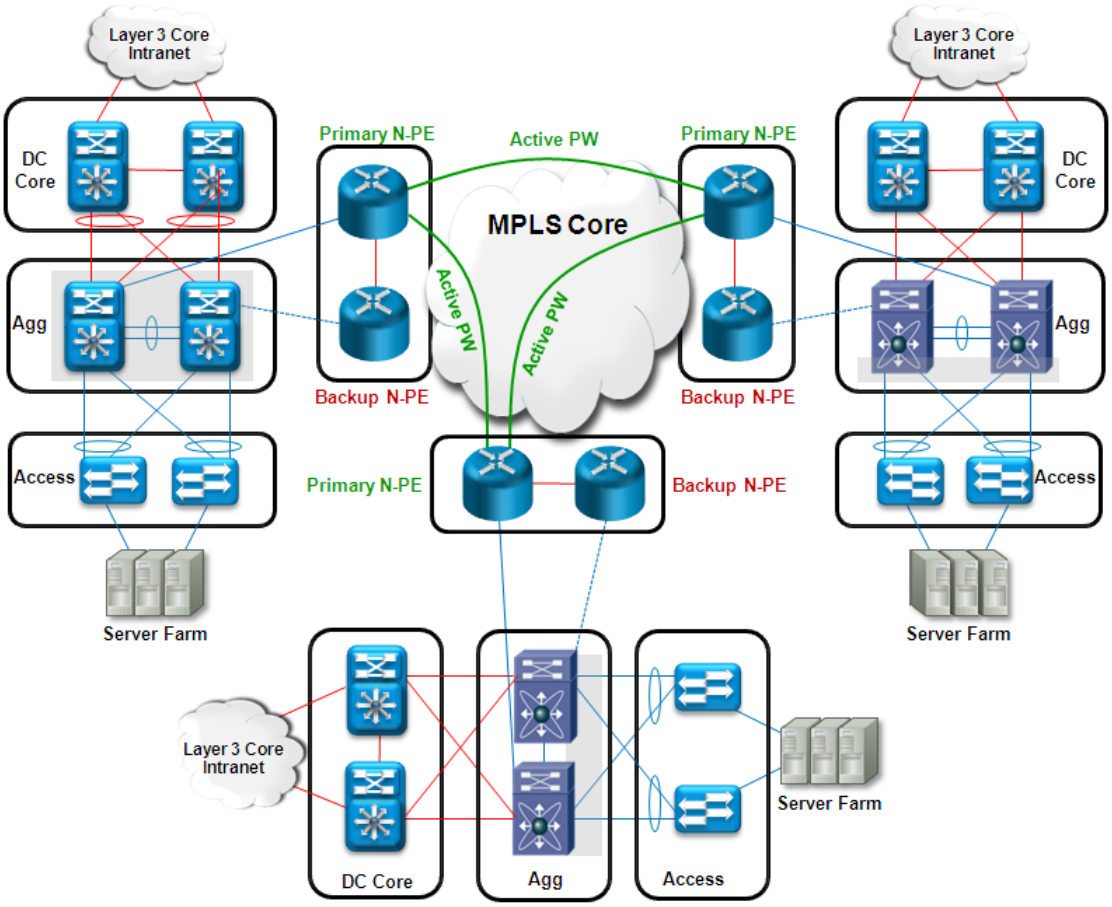
Figure 2-46 Creation of an End-to-End STP Loop



Therefore, an additional mechanism is required to break this end-to-end loop. The approach discussed in this document leverages the Embedded Event Manager (EEM) at the aggregation layer to achieve this purpose.

Figure 2-47 highlights the basic idea behind this approach. In the DCI layer of each data center, one PE actively forwards traffic to/from the VPLS cloud and the idle PE activates only in specific link/node failure scenarios.

Figure 2-47 Creation of Primary and Backup PEs



The following design considerations apply to the scenario depicted in Figure 2-47:

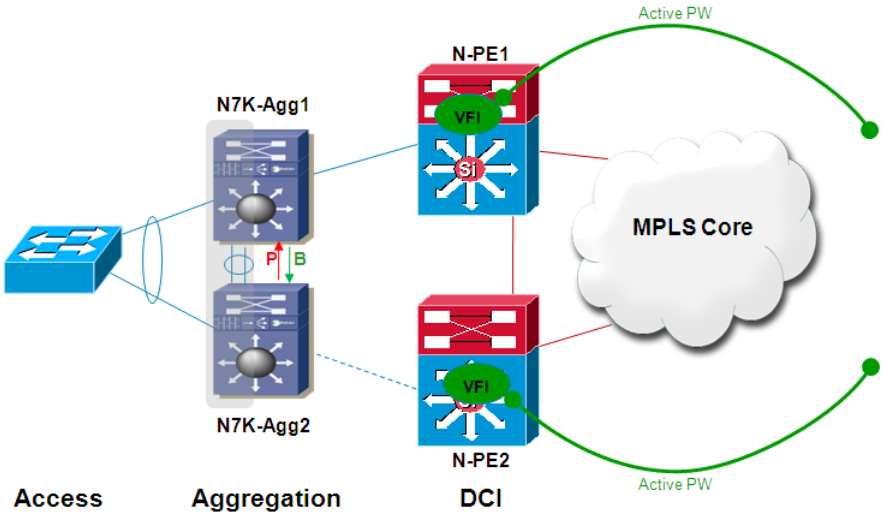
- The Primary/Backup traffic paths can be defined for each VLAN. This provides traffic load-balancing, for example, by dividing in two groups the set of VLANs requiring LAN extension services and by enabling one group on the upper path and the second group on the lower path.
- Despite that a full mesh of VPLS PWs is still established between the PE devices, only the subset connecting the Primary PEs would actually be used to exchange unicast traffic. The other PWs may still be leveraged for delivering unknown unicast, multicast, and broadcast frames, but the EEM configuration would then take care of dropping these frames at the aggregation layer side avoiding the creation of end-to-end loops.
- The active and backup path is not actually dictated by a PE-specific configuration. Instead, it is determined by deploying EEM at the aggregation layer to control which physical links connected to the DCI layer is active (that is, carries the VLANs that are extended to other sites) and which one is idle (that is, does not carry the VLANs until there is a link/node failure that makes the originally active path unable to forward traffic). From a functional perspective, all PEs are actually active and establish PW connections to the remote devices.
- Moving EEM scripts to the aggregation layer provides a loop avoidance solution that is independent of the platform that performs VPLS features in the DCI layer. The validation discussed in this document leveraged Catalyst 6500 switches as PE devices, but the same approach can be followed when deploying other Cisco platforms in that role (ASR9000, CRS-1, and so on).

- The EEM-specific deployment varies with the platform deployed at the aggregation layer: Nexus 7000 and Catalyst 6500 in standalone mode offers independent control plane between peer aggregation switches. Therefore, a semaphore logic is introduced for the two devices to negotiate which should activate the Layer 2 trunk to the DCI layer. In scenarios that leverage Catalyst 6500 in VSS mode, the EEM mode is simpler because the unified control plane between the two VSS member switches does not require the use of the semaphore. More details on these two types of deployments can be found in Deploying EEM for Loop Avoidance with Nexus 7000 in Aggregation, page 2-60, and PE-aggregation links configuration in steady state, page 2-63.

Deploying EEM for Loop Avoidance with Nexus 7000 in Aggregation

When leveraging Nexus 7000 devices in the DC aggregation layer (Figure 2-48), the steady state mode of operation is met when an active Layer 2 trunk connects the primary aggregation node (N7K-Agg1) to a PE device in the DCI layer. Additionally, a secondary Layer 2 trunk is defined between the standby aggregation node (N7K-Agg2) and the DCI layer: none of the VLANs that should be extended across the VPLS PWs are normally carried on this secondary trunk. This prevents an active-active scenario that would create an end-to-end STP loop.

Figure 2-48 Loop Avoidance with Nexus 7000



For the aggregation nodes to agree on their active/backup state, a semaphore logic is introduced: a Primary (P) semaphore is defined on the primary aggregation node (N7K-Agg1). The state of this semaphore indicates whether the Layer 2 trunk that is connected to the primary aggregation node to the DCI layer is active or not. Therefore, the backup aggregation node constantly monitors the state of the P semaphore. Until it is active, the bottom Layer 2 path is kept in idle state. As soon as the secondary node detects that the P semaphore has gone inactive (for whatever reason), it takes over the active role by enabling the VLANs on the idle Layer 2 trunk connection. The secondary node raises at that point a Backup (B) semaphore to indicate that it became the active node.

Therefore, the basic design principle is that the P and B semaphores should never be active at the same time, because that would mean that both aggregation nodes are actively forwarding traffic and this could potentially lead to the creation of an end-to-end STP loop.

Note In reality, for the end-to-end loop to be created, the aggregation layer devices must end up in active-active state in at least two sites at the same time.

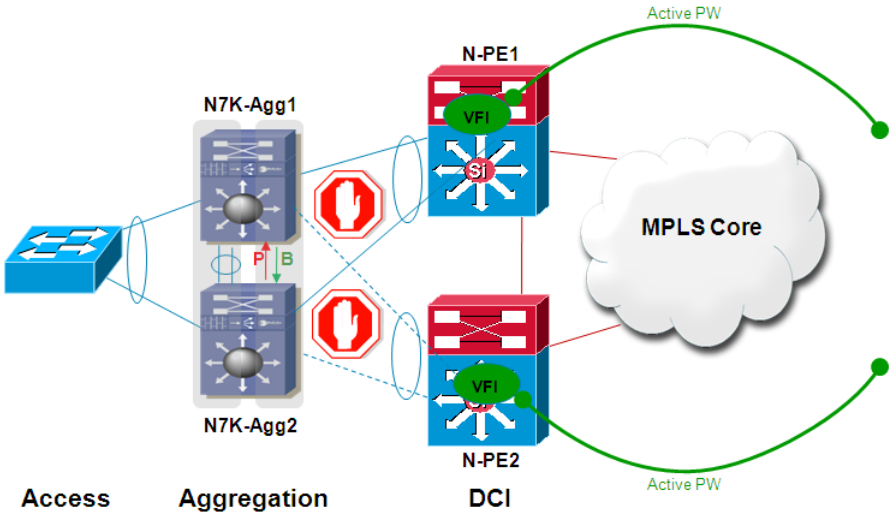
From a deployment perspective, you can detect failure events on the primary aggregation node by tracking a specific route update called *P-semaphore route*. Losing the P-semaphore route would trigger the failover event leading to the activation of the bottom traffic path.

The control plane required to exchange semaphore information between the aggregation nodes is implemented using a dedicated OSPF instance. We recommend that you use an IGP process (or instance) independent from the one that may be used to exchange routing information between the aggregation layer and the DCI layer devices (for more information about routing between sites, see Inter Data Centers Routing Considerations, page 2-24).

Before diving into the details of EEM deployment, you should review some general design considerations that characterize the deployment with Nexus 7000 devices in aggregation.

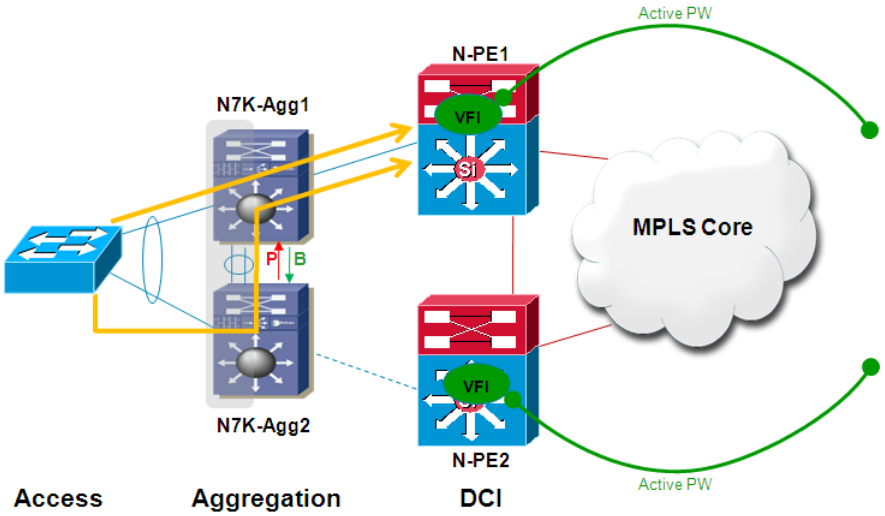
Because of the specific semaphore logic implementation, you cannot use vPC on the aggregation layer devices to front end the DCI layer (Figure 2-49).

Figure 2-49 Lack of Support for vPC between Aggregation and DCI Layer Devices



Since vPC is normally used between the aggregation and the access layer devices, 50% of the outbound traffic flows needs to cross the vPC peer link between aggregation switches (Figure 2-50). This is because at any given time there can only one active Layer 2 path connecting to the DCI layer. Therefore, you must properly dimension the EtherChannel as vPC peer link to handle this additional traffic.

Figure 2-50 Sub-Optimal Traffic Path

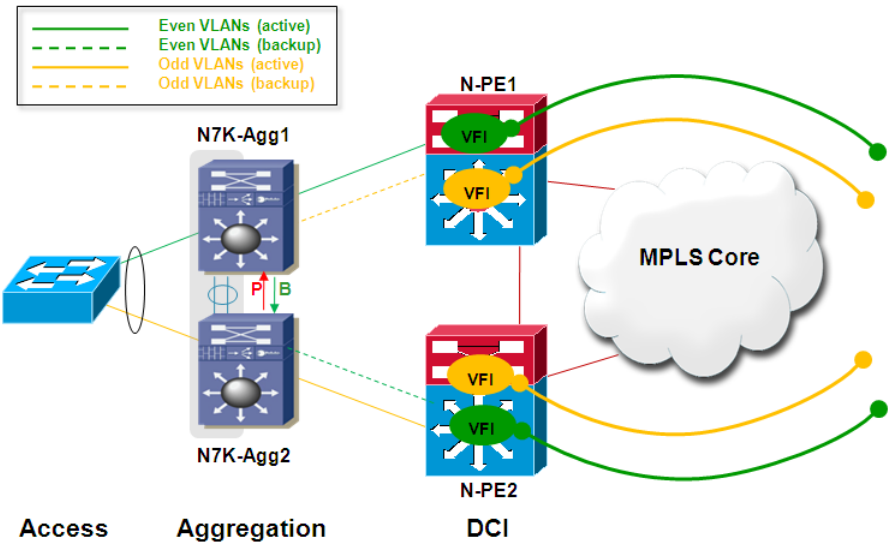


Note The same behavior is not observed for inbound direction, since all the traffic flows received on N7K-Agg1 from the DCI layer is sent toward the access layer by leveraging the vPC link locally available. The peer-link is used for inbound traffic only when this local vPC link toward the access fails.

The behavior described so far can be enabled for each VLAN. If you want to provide load-balancing across different VLANs, we recommend creating two groups (called "Odd" and "Even" for the sake of simplicity) and ensure that "Odd" VLANs uses the upper physical path between aggregation and DCI layer devices, and the "Even" VLANs use the lower path.

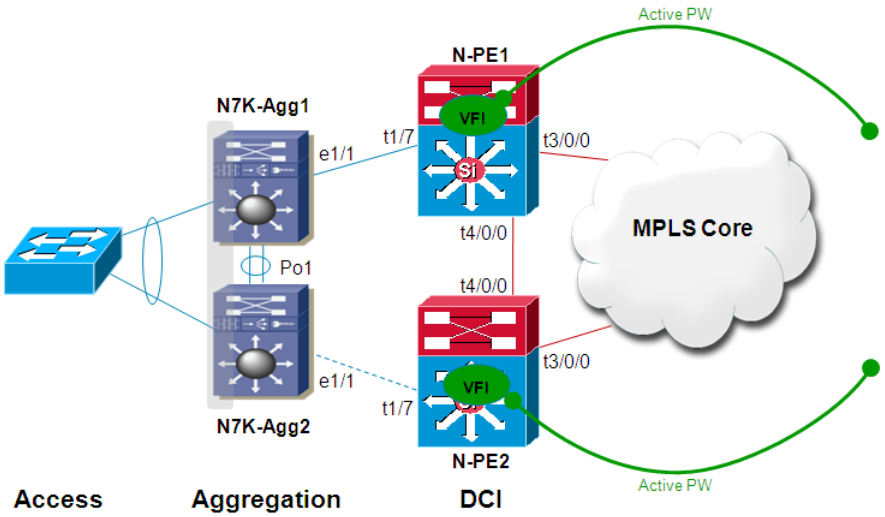
As shown in Figure 2-51, this specific deployment mandates the use of two additional physical links to be enabled between the aggregation and the DCI layers.

Figure 2-51 Load-Balancing across both N-PEs



The deployment of EEM on the Nexus 7000 switches at the aggregation layer ensures the synchronization of state between them. Figure 2-52 defines the topology discussed in PE-aggregation links configuration in steady state, page 2-63, and Semaphore Interface Assignment, page 2-64.

Figure 2-52 Deploying Active and Standby Paths with Nexus 7000



Note The following configurations refer to a scenario where all the VLANs are activated on the upper path. This concept can easily be extended to a scenario where you want load-balancing between upper and lower path.

PE-aggregation links configuration in steady state

The primary path is configured as a regular Layer 2 trunk carrying all the Layer 2 VLANs that are extended to the remote sites. Additionally, you must enable a specific P Semaphore VLAN on the link to activate the P semaphore and indicate that this is the active path. Finally, you must enable a dedicated VLAN on the trunk to establish a Layer 3 IGP peering with the DCI layer device (for more information, see Inter Data Centers Routing Considerations, page 2-24).

N7K-Agg1

```
interface Ethernet1/1
  description Primary Path
  switchport
  switchport mode trunk
  switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN>,<P_SEMAPHORE>
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  mtu 9216
```

The secondary path is also configured as a regular Layer 2 trunk. However, on a steady state, only the optional VLAN for establishing Layer 3 IGP peering is enabled. This is because the B semaphore needs to stay inactive, to signal that this path is not active.

N7K-Agg2

```
interface Ethernet1/1
  description Secondary Path
```

```
  switchport
  switchport mode trunk
  switchport trunk allowed vlan <IGP_VLAN>
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  mtu 9216
```

The Layer 2 trunks connecting the DCI layer devices to the Nexus 7000 in the aggregation layer are configured to carry the required VLANs all the time, independently from which physical path is active. Enabling or disabling a specific path is performed by EEM running at the aggregation layer and no changes need to be applied to the configuration of the PE devices.

N-PE1

```
interface TenGigabitEthernet1/7
 description L2 Trunk to N7K-Agg1
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN><P_SEMAPHORE>
 switchport mode trunk
 mtu 9216
```

N-PE2

```
interface TenGigabitEthernet1/7
 description L2 Trunk to N7K-Agg2
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN><B_SEMAPHORE>
 switchport mode trunk
 mtu 9216
```

Note Adding the P and B semaphore VLANs on the Layer 2 trunks on the PE devices is not strictly required, but it is included in the configuration for completeness.

Semaphore Interface Assignment

As previously mentioned, the use of logic semaphores is required for the two aggregation devices to communicate to each other their state and mode of operation. In the deployment discussed here, each semaphore is represented by a logical VLAN Interface (SVI); the line protocol state (UP or DOWN) of each SVI is used for inter-chassis state communication. To do this, a dedicated IGP instance (in our example an OSPF protocol instance) is used to exchange routing information for the prefixes associated to each defined SVI. The IGP peering is established between the two aggregation devices by leveraging a dedicate VLAN carried across the transit link (vPC peer link), as shown in the following configuration.

N7K-Agg1

```
interface port-channel1
  description vPC Peer Link
  switchport
  switchport mode trunk
  vpc peer-link
  switchport trunk allowed vlan add 4060
  spanning-tree port type network
  mtu 9216
!
interface Vlan4060
  description EEM IGP Peering
  ip address 10.70.60.1/24
  ip router ospf 4060 area 0.0.0.0
```

```
  no shutdown
!
router ospf 4060
  router-id 10.70.60.1
  log-adjacency-changes
  timers throttle spf 100 100 5000
  timers throttle lsa 200 100 5000
```

N7K-Agg2

```
interface port-channel1
  description vPC Peer Link
  switchport
  switchport mode trunk
  vpc peer-link
  switchport trunk allowed vlan add 4060
  spanning-tree port type network
  mtu 9216
!
interface Vlan4060
  description EEM IGP Peering
  ip address 10.70.60.2/24
  ip router ospf 4060 area 0.0.0.0
  no shutdown
!
router ospf 4060
  router-id 10.70.60.2
  log-adjacency-changes
  timers throttle spf 100 100 5000
  timers throttle lsa 200 100 5000
```

Note The vPC peer link is the key connection used to communicate semaphore state information between the two aggregation devices. Thus, we recommend that you increase the reliability of that link by creating a PortChannel that bundles links belonging to different linecard modules. By following this design recommendation, you mitigate the likelihood of vPC peer link failure. However, dual failure scenarios are beyond the scope of this document.

Note We recommend tuning OSPF timers to ensure faster convergence after a failure event. When configuring the **timers throttle spf** and **timers throttle lsa** commands on Nexus 7000 platforms, you must ensure that the first value is larger than 50 msec so as to avoid a specific issue discovered in software release 4.2(3).

Table 2-14 lists all the semaphore SVIs defined on the primary aggregation node.

Table 2-14 Semaphore Definition on Primary Aggregation Node

Interface Name	Description	Purpose	IP Address
SVI 4061	Aggregation Link Down Delay	Used to create a carrier delay to block preemption requests when the aggregation link is flapping.	No

Table 2-14 Semaphore Definition on Primary Aggregation Node (continued)

Interface Name	Description	Purpose	IP Address
SVI 4062	P-semaphore	Signals that the primary aggregation node is operational.	Yes 10.70.62.3/24
SVI 4063	Aggregation Link Up	Requests preemption with delay. It is used to signal to the secondary node that the primary node is ready to preempt and become operational again.	Yes 10.70.63.3/24

The corresponding device configuration follows.

N7K-Agg1

```
interface Vlan4061
  description Aggregation Link Down Delay
  carrier-delay 60
  no shutdown
!
interface Vlan4062
  description P semaphore - L2 trunk is forwarding
  carrier-delay msec 0
  ip address 10.70.62.3/24
  ip ospf passive-interface
  ip router ospf 4060 area 0.0.0.0
  no shutdown
!
interface Vlan4063
  description Aggregation Link Recovery
  carrier-delay 60
  ip address 10.70.63.3/24
  ip ospf passive-interface
  ip router ospf 4060 area 0.0.0.0
  no shutdown
```

Table 2-15 lists the only semaphore required on the secondary aggregation node.

Table 2-15 Semaphore Definition on the Secondary Aggregation Node

Interface Name	Description	Purpose	IP Address
SVI 4064	B-semaphore	Signals that the secondary node is in Active (forwarding) mode.	Yes 10.70.64.3/24

N7K-Agg2

```
interface Vlan4064
  description B semaphore - L2 trunk is forwarding
  carrier-delay msec 0
  ip address 10.70.64.3/24
  ip ospf passive-interface
  ip router ospf 4060 area 0.0.0.0
  no shutdown
```

Note The preceding semaphores are required when all the VLANs are carried on a single physical path. If load-balancing is required between two physical paths, the semaphore definition is duplicated and implemented in a symmetrical fashion.

EEM Related Tracking Objects

As previously discussed, the state of the semaphores is communicated between the aggregation nodes by exchanging routing updates for the IP prefixes associated to each defined SVI. Tracking this information is necessary to trigger EEM scripts to recover connectivity under various link/node failure scenarios. Table 2-16 highlights the tracking objects defined on the primary aggregation node, together with the corresponding EEM script they trigger (for more information on the EEM scripts, see EEM Scripts , page 2-68).

Table 2-16 Tracking Objects on Primary Aggregation Node

Track Object Number	Track Object Description	Condition	Script Triggered
61	Aggregation Link Up Delay State	Track the state of Vlan 4061 (which has configured a 60 seconds carrier delay). This track would come up after the aggregation link recovers, and stays up for 60 seconds (to avoid flapping).	DOWN→UP: L2_PATH_CARRIER_DELAY
64	B-semaphore	Track the B-semaphore prefix on the secondary node. Detecting the B-semaphore going down would trigger the primary node to go back to active mode. Detecting the B semaphore going up would trigger the primary node to go inactive.	UP→DOWN: L2_VLANs_START DOWN→UP: L2_VLANs_STOP

The corresponding configuration follows:

N7K-Agg1

```
track 61 interface Vlan4061 line-protocol
track 64 ip route 10.70.64.3/24 reachability
```

Table 2-17 highlights the tracking objects required on the secondary aggregation node, with the corresponding configuration following.

Table 2-17 Tracking Objects on Backup Aggregation Node

Track Object Number	Track Object Description	Condition	Script Triggered
62	P-semaphore	Track the P-semaphore prefix on the primary node. Detecting the P-semaphore going down would trigger the secondary node to go in active forwarding mode.	UP→DOWN: L2_VLANs_START
63	Primary Node Preemption Request	Track the primary node preemption request to regain the active role. Detecting the semaphore going up would force the secondary node to go inactive (hence to bring down the B-semaphore)	DOWN→UP: L2_VLANs_STOP

N7K-Agg2

```
track 62 ip route 10.70.62.3/24 reachability
track 63 ip route 10.70.63.3/24 reachability
```

EEM Scripts

Changing semaphore states trigger EEM scripts. Table 2-18 lists the different EEM scripts used on the primary aggregation node.

Table 2-18 EEM Scripts on Primary Aggregation Node

Script Name	Description	Details	Triggered By
L2_PATH_CARRIER-DELAY	Aggregation Link Carrier Delay	Triggered when the aggregation link to the PE recovers but only after it is stable for at least 60 seconds (carrier delay of SVI 4061 has expired). The script would enable VLAN 4063 on the aggregation link. Notice that a further carrier-delay is configured for VLAN 4063, which causes an additional 60 seconds delay before SVI 4063 goes up (this is to ensure the network is stable and converged). Line protocol going up on SVI 4063 will send the preemption request message to the secondary node.	Track 61 Up
L2_VLANs_STOP	Backup Node Up	Triggered when the secondary node has gone active and has raised the B-semaphore. It causes the primary node to go inactive.	Track 64 Up

Table 2-18 EEM Scripts on Primary Aggregation Node (continued)

Script Name	Description	Details	Triggered By
L2_VLANs_START	Backup Node Down	Triggered by detecting the B-semaphore going down. This script would make the primary node active.	Track 64 Down
L2_VLANs_BOOT_HOLD	Aggregation Node/Module Reload	Triggered when the linecard used to connect to the N-PE1 device is coming online. The script would keep down the aggregation link for 2 extra minutes before sending the preemption request message to the secondary node (this is done to ensure the network has properly converged and stabilized).	Module Reload

The following are the CLI commands required to implement the preceding EEM logic.

N7K-Agg1

```
event manager applet L2_PATH_CARRIER-DELAY
  event track 61 state up
  action 0.1 cli en
  action 1.0 cli conf t
  action 2.0 cli int eth1/1
  action 2.2 cli sw trunk allow vlan <IGP_VLAN>,4063
  action 4.0 cli int vlan 4061
  action 4.1 cli carrier 0
  action 4.2 cli shut
  action 4.3 cli carrier 60
  action 4.4 cli no shut
!
event manager applet L2_VLANs_STOP
  event track 64 state up
  action 0.1 cli en
  action 1.0 cli conf t
  action 2.0 cli int eth1/1
  action 2.2 cli sw trunk allow vlan <IGP_VLAN>,4061
  action 7.0 cli clear mac add dyn
  action 9.0 syslog msg Primary Node is in Standby
!
event manager applet L2_VLANs_START
  event track 64 state down
  action 0.1 cli en
  action 1.0 cli conf t
  action 2.0 cli int eth1/1
  action 2.1 cli sw trunk allow vlan <L2_VLANs_RANGE>,<IGP_VLAN>,4062
  action 3.0 cli clear mac add dyn
  action 4.0 cli int vlan 4063
  action 4.1 cli carrier 0
  action 4.2 cli shut
  action 4.3 cli carrier 60
  action 4.4 cli no shut
  action 9.1 syslog msg Primary Node is active
!
event manager applet L2_VLANs_BOOT_HOLD
  event module status online module 1
```

```
  action 0.1 cli en
  action 1.0 cli conf t
  action 2.0 cli int e1/1
  action 2.1 cli shut
  action 2.2 cli sw trunk allow vlan <IGP_VLAN>,4061
  action 4.0 syslog msg "Reboot detected, hold L2 trunk down for 2 extra minutes"
  action 5.0 cli sleep 60
  action 6.0 cli sleep 60
  action 7.0 cli int e1/1
  action 7.1 cli no shut
  action 9.0 syslog msg "Start L2 trunk preemption process"
```

Table 2-19 shows the EEM scripts deployed on the secondary aggregation node.

Table 2-19 EEM Scripts on Secondary Aggregation Node

Script Name	Description	Details	Triggered By
L2_VLANs_START	Primary Node Down	This script is triggered when the P-semaphore is detected going down. This script would make the backup node to take the active role. The B-semaphore would be brought up.	Track 62 Down
L2_VLANs_STOP	Primary Node Ready to Preempt	This script is triggered when preemption to the primary node is requested and triggered. This script would take the B-semaphore down, and would switch the backup node to standby mode. This would trigger the primary node to become active.	Track 63 Up

The required EEM script configuration on the secondary node follows:

N7K-Agg2

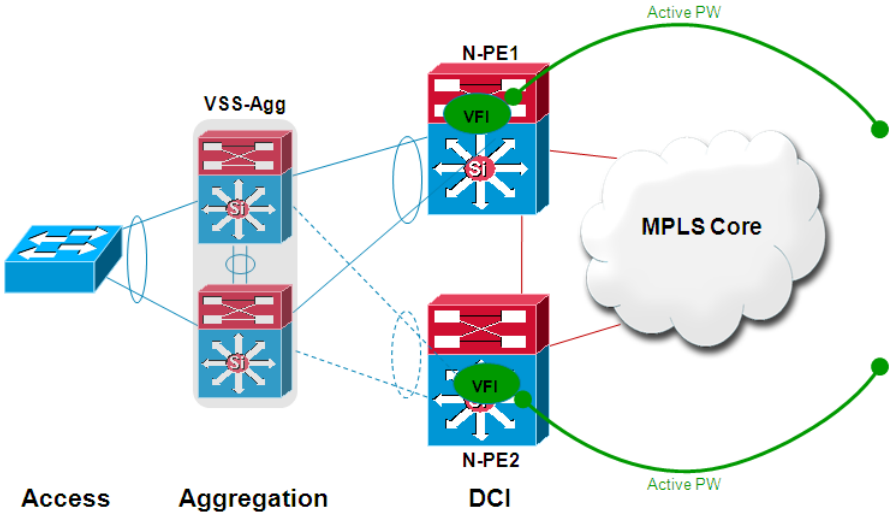
```
event manager applet L2_VLANs_START
  event track 62 state down
  action 0.1 cli en
  action 1.0 cli conf t
  action 4.0 cli int eth1/1
  action 4.1 cli sw trunk allow vlan <L2_VLAN_RANGE>,<IGP_VLAN>,4064
  action 5.0 cli clear mac add dyn
  action 9.0 syslog msg Backup Node is active
!
event manager applet L2_VLANs_STOP
  event track 63 state up
  action 0.0 cli sleep 60
  action 0.1 cli en
  action 1.0 cli conf t
  action 2.0 cli int eth1/1
  action 2.1 cli shut
  action 3.0 cli clear mac add dyn
  action 4.0 cli sw trunk allow vlan <IGP_VLAN>
  action 4.1 cli no shut
  action 9.0 syslog msg Backup Node is standby
```

Note Some minor changes for the `L2_VLANs_START` script on the secondary node are required to improve the recovery of Layer 2 multicast flows. See VPLS Failure/Recovery Scenarios, page 2-80, for more details.

Deploying EEM for Loop Avoidance with Catalyst 6500 (VSS) in Aggregation

The same steady state mode of operation can also be applied by leveraging a pair of Catalyst 6500 switches deployed in VSS mode as aggregation layer devices (Figure 2-53).

Figure 2-53 Loop Avoidance with Catalyst 6500 (VSS)

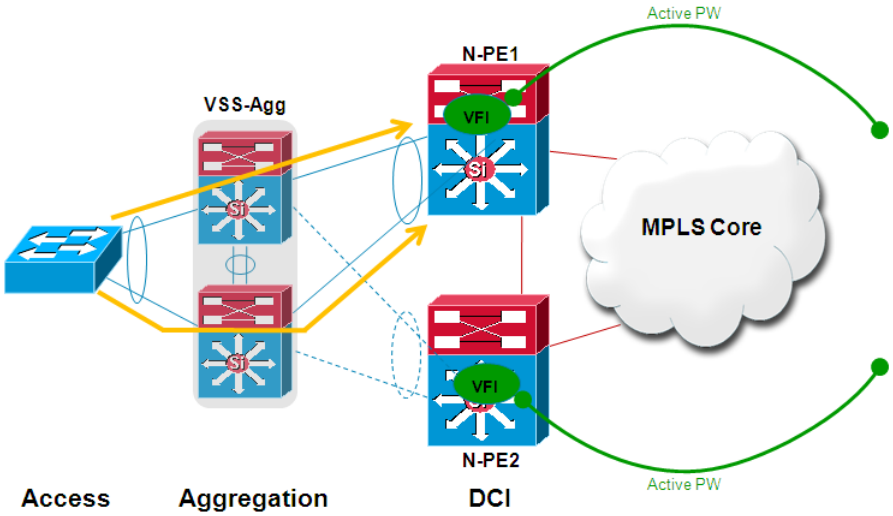


Similar to the Nexus 7000 scenario, the idea is to leverage an active Layer 2 trunk connection between the VSS logical switch and one of the PE devices in the DCI layer (N-PE1). An idle secondary Layer 2 trunk can then be defined toward N-PE2 to be activated under specific failure scenarios.

Use a unified control plane between the two 6500 switches that are members of the same VSS domain provides the following benefits:

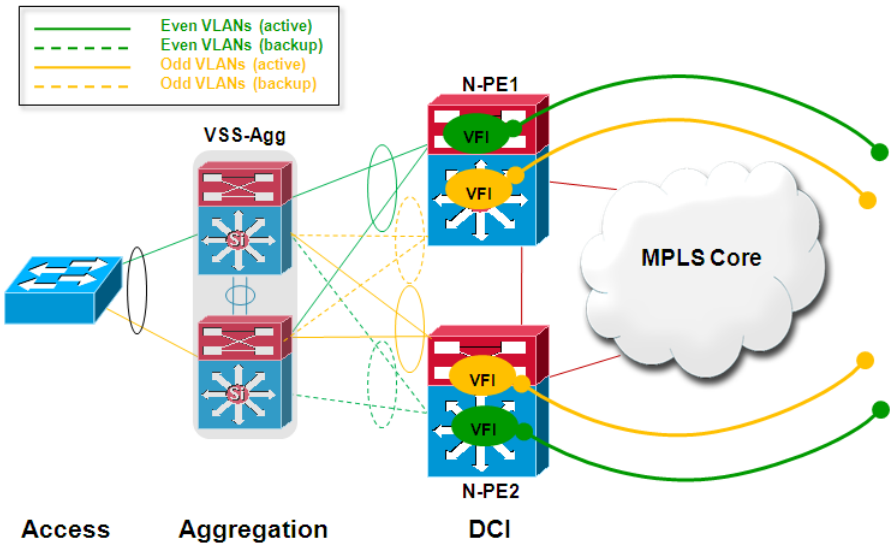
- A Layer 2 trunk EtherChannel can be established between each PE device and the VSS in aggregation. This improves the resiliency of the overall solution. The active Layer 2 path is now represented by the PortChannel interconnecting the VSS with the N-PE1 device, whereas the standby path is a PortChannel toward N-PE2.
- There is no need to introduce a semaphore logic. This simplifies the required EEM configuration and facilitates avoiding active-active scenarios.
- Since Multi Chassis EtherChannels (MCECs) are established toward the access layer and toward the DCI layer, outbound traffic follows an optimal path (that is, no traffic is crossing the VSL link by design), as highlighted in Figure 2-54.

Figure 2-54 Optimal Outbound Traffic Flows



- As previously mentioned for the Nexus 7000 deployment scenario, the behavior described so far can be enabled for each VLAN. If you want to provide load-balancing across different VLANs, you can split two groups of VLANs across two separate physical paths (Figure 2-55) between aggregation and DCI layer devices.

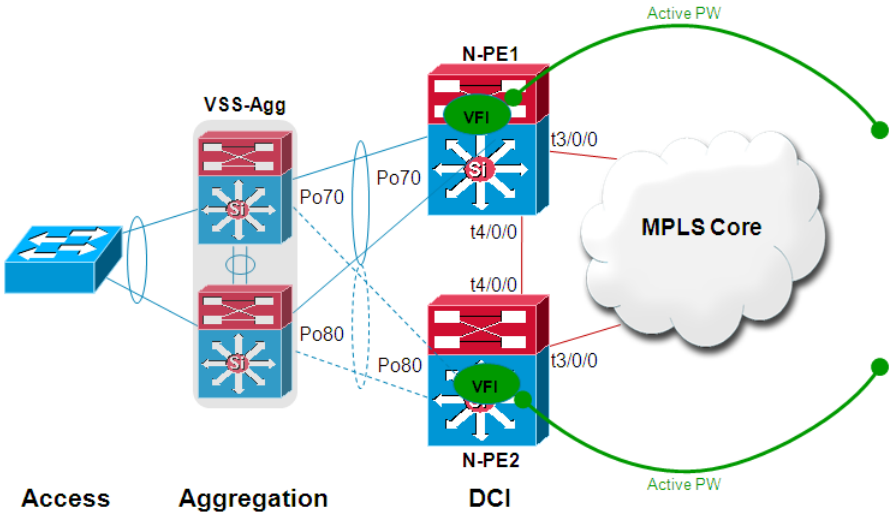
Figure 2-55 VLAN Load Balancing across EtherChannels



As shown in Figure 2-55, this specific deployment mandates the use of eight separate physical interfaces between the aggregation and the DCI layer.

The following sections dive into the details of the deployment of EEM with Catalyst 6500 switches at the aggregation layer. All the considerations below refer to the network diagram in Figure 2-56.

Figure 2-56 Deploying Active and Standby Paths with VSS



Note The following configurations refer to a scenario where all the VLANs are activated on the upper path. This concept can easily be extended to a scenario where load-balancing between upper and lower path is desirable.

PE-aggregation links configuration in steady state

The primary path is configured as an EtherChannel Layer 2 trunk carrying all the Layer 2 VLANs that are extended to the remote sites. In addition to that, a dedicated VLAN may also be enabled on the trunk to establish a Layer 3 IGP peering with the DCI layer device (for more information, see Inter Data Centers Routing Considerations, page 2-24).

The secondary path is an EtherChannel Layer 2 trunk connecting to the second PE device. The only VLAN allowed on that trunk on the steady state is the one used to establish the IGP peering with the DCI layer.

VSS-Agg

```
interface Port-channel70
 description Primary Path
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN>
 switchport mode trunk
 mtu 9216
 spanning-tree portfast edge trunk
 spanning-tree bpdufilter enable
!
description Secondary Path
interface Port-channel80
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <IGP_VLAN>
 switchport mode trunk
 mtu 9216
 spanning-tree portfast edge trunk
 spanning-tree bpdufilter enable
```

The Layer 2 trunks connecting the DCI layer devices to the VSS in the aggregation layer are configured to carry the required VLANs all the time, independently from which physical path is active. Enabling or disabling a specific path is performed by EEM running at the aggregation layer and no changes need to be applied to the configuration of the PE devices.

N-PE1

```
interface Port-channel70
 description L2 Trunk to VSS-Agg
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN>
 switchport mode trunk
 mtu 9216
```

N-PE2

```
interface Port-channel80
 description L2 Trunk to VSS_Agg
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan <L2_VLANs_RANGE>,<IGP_VLAN>
 switchport mode trunk
 mtu 9216
```

EEM Related Tracking Objects

The use of a unified control plane between the two VSS member switches drastically simplifies the EEM configuration and requires one EEM object to be tracked, as highlighted in Table 2-20.

Table 2-20 Tracking Objects on VSS Aggregation

Track Object Number	Track Object Description	Condition	Script Triggered
70	Layer 2 Trunk PortChannel Change of State (Up or Down)	Track the state of the active aggregation link (Layer 2 trunk PortChannel to N-PE1). If the entire logical PortChannel goes down, the connection fails over to the backup PortChannel connected to N-PE2.	UP→DOWN: BACKUP_EC_ACTIVATE DOWN→UP: PRIMARY_EC_ACTIVATE

The corresponding configuration follows:

VSS-Agg

```
track 70 interface Port-channel70 line-protocol
 delay up 180
```

EEM Scripts

The number of EEM scripts required is also reduced in the VSS deployment scenario as shown in Table 2-21.

Table 2-21 EEM Scripts on Primary Aggregation Node

Script Name	Description	Details	Triggered By
BACKUP_EC_ ACTIVATE	Activate the Backup Port-channel to N_PE2	Triggered when the entire logical EtherChannel to N-PE1 has gone down. The script would enable the Layer 2 VLANs to be extended across the backup PortChannel to N-PE2.	Track 70 Down
PRIMARY_EC _ACTIVATE	Activate the Primary Port-channel to N_PE1	Triggered when the primary PortChannel is recovered. The script introduces a 5 minutes delay before re-activating the VLANs on the primary path, to ensure the network has converged and stabilized.	Track 70 Up
BACKUP_EC_ DEACTIVATE	Deactivate the Backup Port-channel to N_PE2	Triggered when the primary PortChannel is recovered. It is run in parallel with the PRIMARY_EC_ACTIVATE script to optimize the convergence achieved in reverting to the primary path.	Track 70 Up

The corresponding configuration follows:

VSS-Agg

```
event manager applet BACKUP_EC_ACTIVATE
 event track 70 state down
 action 0.1 cli command "en"
 action 1.0 cli command "conf t"
 action 2.0 cli command "int po 80"
 action 2.1 cli command "switchport trunk allow vlan <L2_VLANs_RANGE>,<IGP_VLAN>"
 action 3.0 cli command "int po 70"
 action 3.1 cli command "switchport trunk allow vlan <IGP_VLAN>"
 action 9.0 syslog msg "Backup Port-channel activated"
!
event manager applet PRIMARY_EC_ACTIVATE
 event track 70 state up maxrun 1000
 action 0.1 cli command "en"
 action 0.5 cli command "ping 1.1.1.1 time 300 repeat 1"
 action 1.0 cli command "conf t"
 action 3.0 cli command "int po 70"
 action 3.1 cli command "switchport trunk allow vlan <L2_VLANs_RANGE>,<IGP_VLAN>"
 action 5.0 cli command "do clear mac-add dyn"
 action 9.0 syslog msg " Primary Port-channel activated "
!
event manager applet BACKUP_EC_DEACTIVATE
 event track 70 state up maxrun 1000
 action 0.1 cli command "en"
 action 0.5 cli command "ping 1.1.1.1 time 300 repeat 1"
 action 1.0 cli command "conf t"
 action 2.0 cli command "int po 80"
 action 2.1 cli command "shut"
 action 4.1 cli command "switchport trunk allow vlan <IGP_VLAN>"
 action 4.2 cli command "no shut"
 action 9.0 syslog msg " Backup Port-channel deactivated"
```

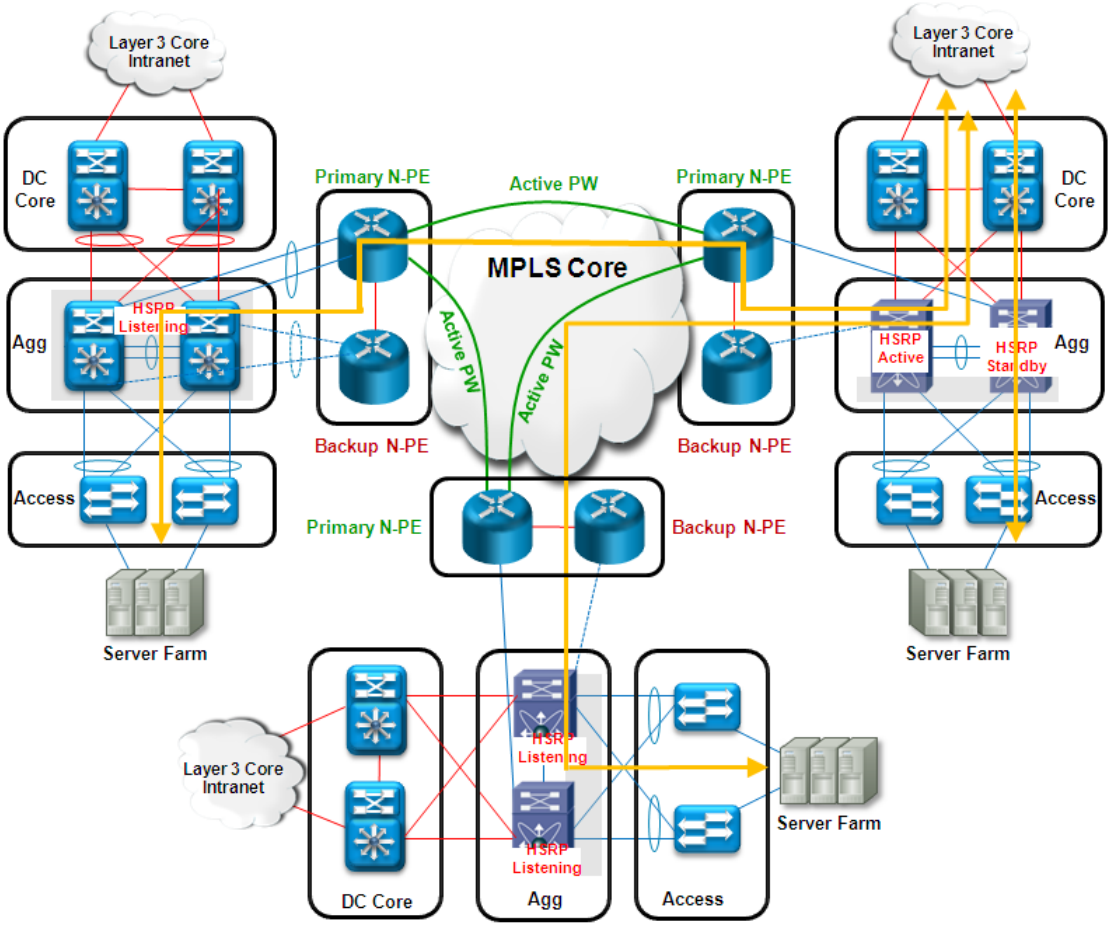
Note Some minor changes for the `BACKUP_EC_ACTIVATE and PRIMARY_EC_ACTIVATE` scripts would be required to improve the convergence of Layer 2 multicast flows. See VPLS Failure/Recovery Scenarios, page 2-80 for more details.

First Hop Redundancy Protocol (FHRP) Deployment

The FHRP deployment that was validated in this phase is similar to the one discussed for point-to-point topologies. The basic idea is to have a single data center as the point of ingress/egress for all the traffic that needs to be routed between the IP subnets extended across the different data center sites.

This means that the aggregation layer devices in the remote sites need to be configured as part of the same HSRP group and they will exchange HSRP messages leveraging the VPLS PW connections. The result is shown in Figure 2-57.

Figure 2-57 HSRP Deployment across DC Sites



In Figure 2-57, all the traffic directed to remote clients that are part of the upper Layer 3 core is taken to the right data center by leveraging the LAN extension services provided by VPLS. In this scenario, routing in and out of the extended IP subnets is performed by the aggregation layer devices in the right

data center. It is important to clarify that this is required only for the IP subnets that are stretched between the data center locations. Subnets that are only defined in a specific location may (and should) still leverage a locally defined default gateway.

Review the configurations and convergence numbers under various failure scenarios discussed in First Hop Redundancy Protocol (FHRP) Deployment, page 2-12 as they apply to this scenario.

Note Improvement to this solution may consist in enabling independent default gateways in each data center for the stretched IP subnets and ensure that also the inbound traffic is directed to the "right" data center (that is, the data center where the destination server actually resides). However, independent default gateways are beyond the scope of this document and will be introduced in future releases of the DCI system testing program.

Inter Data Centers Routing Considerations

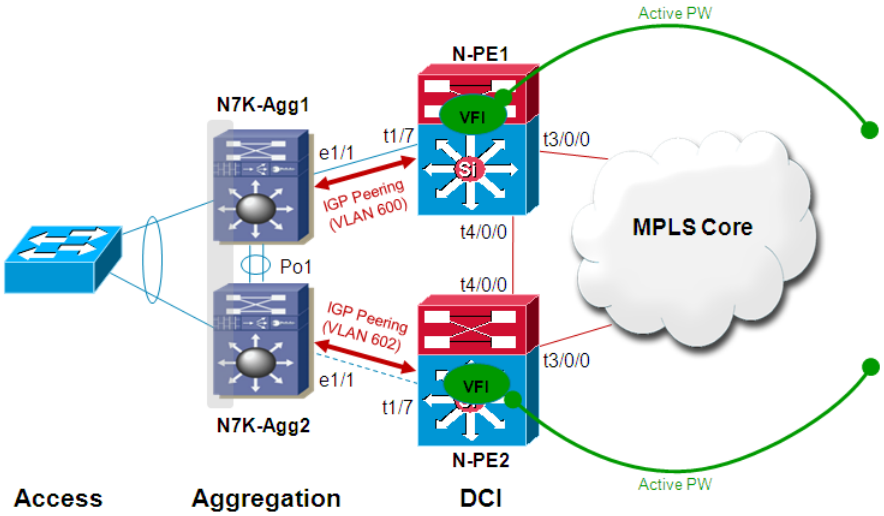
Similarly to what discussed for point-to-point topologies, a couple of different scenarios needs to be considered for Layer 3 communication between remote data centers:

- Routing between remote sites leverages the connectivity across a different Layer 3 cloud that interconnects the remote sites. This is the "Layer 3 Core Intranet" shown in Figure 2-42, where the various DC core devices are connected.
- Routing between data centers happens across the same MPLS cloud used for providing VPLS connectivity. This is usually the case when the bandwidth available via the MPLS cloud is higher than the one through the Layer 3 intranet, or if the MPLS cloud is the only connection available between data center sites.

Focusing on the second scenario, in order for the PE devices to learn the specific DC routes that need to be exchanged across the MPLS core, we recommend enabling dedicated VLANs on the Layer 2 trunks between the aggregation and DCI layers and establish IGP peering between the corresponding SVIs. As noticed in the previous section when discussing the EEM deployment, these VLANs must be enabled on the Layer 2 trunks and are not added or removed by EEM (the scripts only add or removes the Layer 2 VLANs that are extended between remote locations).

Figure 2-58 highlights this idea when using Nexus 7000 in aggregation and it is followed by the corresponding configuration sample. A similar configuration would be required when deploying VSS in aggregation.

Figure 2-58 IGP Peering between Aggregation and DCI Layer Devices



N7K-Agg1

```
interface Ethernet1/1
  description Primary Path
  switchport
  switchport mode trunk
  switchport trunk allowed vlan add 600
!
interface Vlan600
  mtu 9100
  description OSPF peering with N-PE1
  ip address 101.120.0.1/24
  ip router ospf 10 area 0.0.0.0
  no shutdown
!
router ospf 10
  router-id 12.0.2.1
  log-adjacency-changes
  timers throttle spf 100 100 5000
  timers throttle lsa 100 100 5000
N7K-Agg2
interface Ethernet1/1
  description Secondary Path
  switchport
  switchport mode trunk
  switchport trunk allowed vlan add 602
!
interface Vlan602
  mtu 9100
  description OSPF peering with N-PE2
  ip address 101.120.2.1/24
  ip router ospf 10 area 0.0.0.0
  no shutdown
!
router ospf 10
  router-id 12.0.2.1
  log-adjacency-changes
  timers throttle spf 100 100 5000
  timers throttle lsa 100 100 5000
```

N-PE1

```
interface TenGigabitEthernet1/7
 description L2 Trunk to N7K-Agg1
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan add 600
!
interface Vlan600
 description IGP peering with N7K-Agg1
 mtu 9100
 ip address 101.120.0.2 255.255.255.0
!
router ospf 10
 router-id 11.0.2.101
 log-adjacency-changes
 timers throttle spf 50 100 5000
 timers throttle lsa all 100 100 5000
 timers lsa arrival 80
 network 101.0.0.0 0.255.255.255 area 0
```

N-PE2

```
interface TenGigabitEthernet1/7
 description L2 Trunk to N7K-Agg2
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan add 602
!
interface Vlan602
 description IGP peering with N7K-Agg1
 mtu 9100
 ip address 101.120.2.2 255.255.255.0
!
router ospf 10
 router-id 11.0.2.102
 log-adjacency-changes
 timers throttle spf 50 100 5000
 timers throttle lsa all 100 100 5000
 timers lsa arrival 80
 network 101.0.0.0 0.255.255.255 area 0
```

Note We recommend tuning OSPF timers to ensure faster convergence after a failure event. When configuring the **timers throttle spf** and **timers throttle lsa** commands on Nexus 7000 platforms, it is essential to ensure that the first value is larger than 50 msec not to incur in a specific issue discovered in software release 4.2(3).

Few design considerations that apply to this scenario:

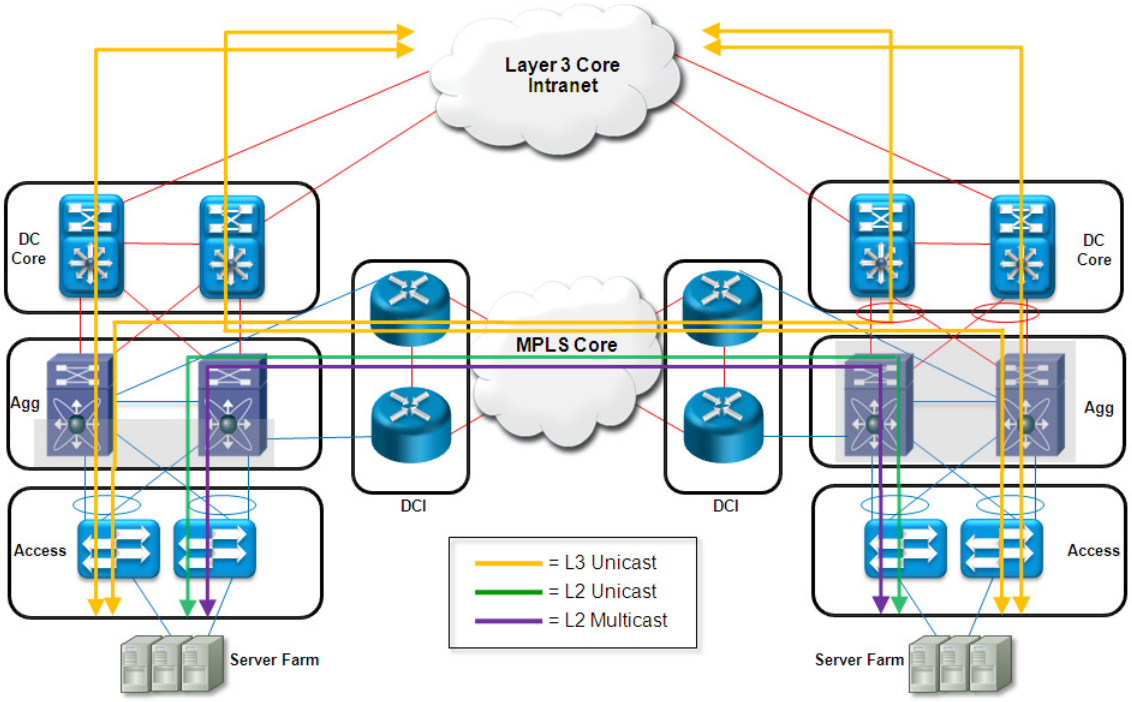
- The IGP used to establish the peering between aggregation and DCI layer devices should be a separate instance from the one previously discussed and used to exchange EEM semaphore information between the aggregation layer switches.
- Depending on the deployment, the same IGP (OSPF 10 in the example) may also be used inside the MPLS core (end-to-end IGP deployment). Alternatively, a separate routing protocol (different IGP instance or BGP) could be used inside the MPLS cloud.
- When possible, summarize the IP prefixes belonging to each specific data center site before injecting them into the MPLS core. This can be achieved in different ways depending on the control plane used in the MPLS core. However, information for doing this is beyond the scope of this document.

VPLS Failure/Recovery Scenarios

Before discussing some of the failure and recovery scenarios that were validated, it is important to clarify the test environment where this validation has been performed.

Figure 2-59 shows the traffic flows that were established across the network under test.

Figure 2-59 Traffic Flows between Data Centers



As shown, a mix of Layer 2 and Layer 3 traffic flows were enabled. The goal was to simulate as much as possible the traffic mix expected in a real deployment. For this reason, frames with different IP size were utilized to build the traffic flows.

From a scalability perspective, the following are the parameters that were validated:

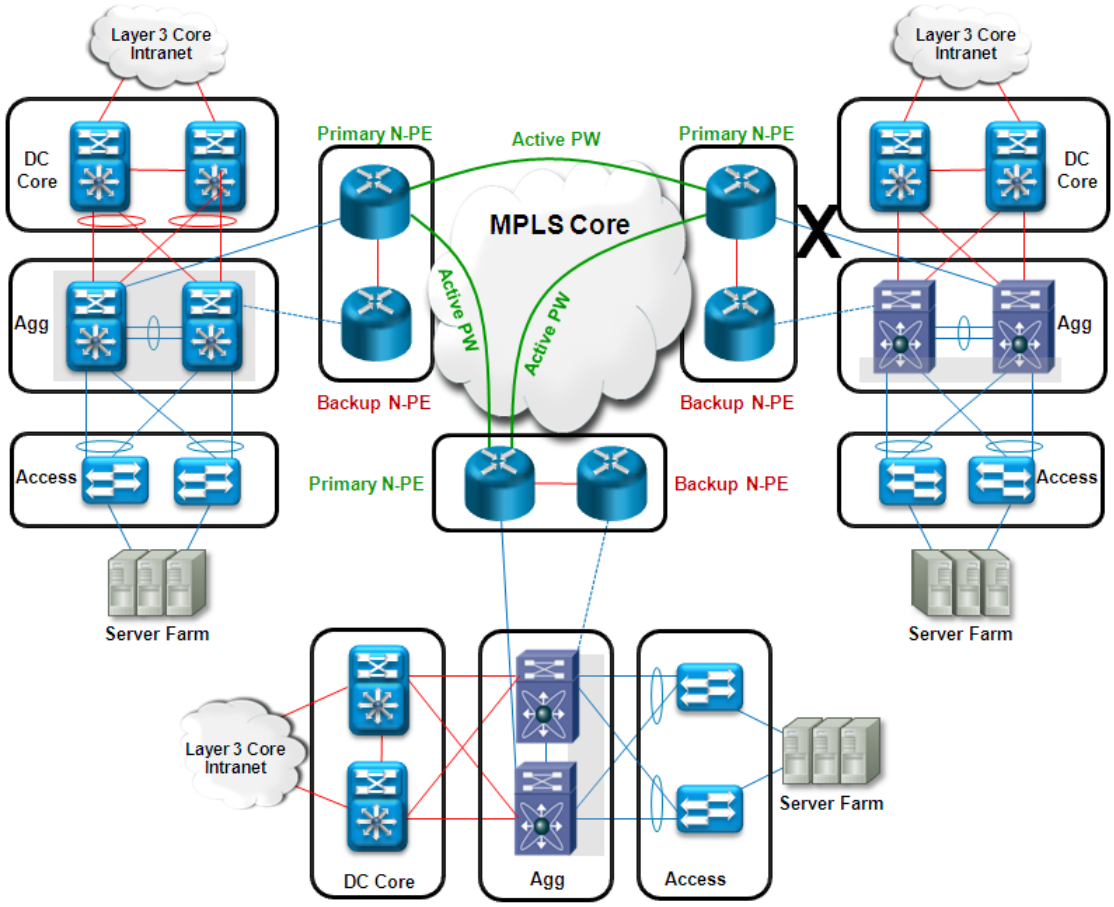
- 300 VLANs extended between data centers
- 3000 routes injected inside each data center
- Around 2.5 Gbps of combined traffic was sent between data centers (in each direction).

Most of the link/box failure scenarios described in the following cause the activation of the EEM scripts described in STP Isolation and End-to-End Loop Prevention, page 2-57. Modifications were made to the original script configuration to improve the behavior of the system under failure scenarios. These changes are described with the relevant test case that required their introduction.

Test 1: Failure/Recovery Active Link between Nexus 7000 and N-PE1

This scenario is highlighted in Figure 2-60.

Figure 2-60 Failure/Recovery Active Link between N7K and N-PE1



Convergence After Failure

- **Outbound Layer 2 Flows**: before the failure, all the Layer 2 traffic flows (unicast and multicast) are carried on that aggregation link connecting the primary aggregation node to N-PE1. The link failure causes the secondary aggregation node to go active by triggering the "L2_VLANs_START" EEM script. As soon as the Layer 2 VLANs are enabled on the secondary trunk, Layer 2 unicast flows are recovered, since the VPLS PWs on N-PE2 are already established to the other two remote data center sites and can be leveraged to deliver traffic (traffic will be flooded on these PWs until N-PE2 re-learns the destination MAC addresses by receiving inbound traffic).

Layer 2 multicast flows differ because IGMP snooping is enabled by default on the Nexus 7000 platforms. With IGMP snooping enabled, Layer 2 multicast traffic is sent out an interface under the following circumstances:

- The switch received an IGMP join from a receiver on that same interface.
- The switch configured that port as an "mrouter" port (port connecting to a multicast enabled router). This can happen dynamically (by receiving a PIM message on that interface) or can be configured statically.

When the secondary path is activated, no IGMP information is available on that link (since the VLANs have just been enabled), To send Layer 2 multicast toward a receiver in a remote data center, you should wait for the next PIM message from a router that would dynamically enable the interface as a mrouter

port. This could potentially lead to up to 60 seconds of worst case multicast outage (PIM messages are sent out by default with that time interval). We recommend that you manually configure mrouter to point to the newly activated trunk. On a Nexus 7000, use the following configurations:

N7K-Agg2

```
vlan 421
  ip igmp snooping mrouter interface Ethernet1/1
```

VLAN 421 in this example has receivers in the remote data center and it is extended across the VPLS cloud. This configuration can be put in place even before the VLAN is enabled on the trunk, so we recommend enabling it from the beginning.

- **Inbound Layer 2 Flows**: recovery of inbound flows is also dependent on the activation of the secondary path via N-PE2. Once again, a full mesh of PWs is established among all PEs. Once the aggregation link fails, the corresponding PWs connecting to the remote sites would go down as well (the aggregation link is the only attachment circuit configured) so the remote PEs will start using the PWs connecting to N-PE2.

Convergence After Recovery

- **Outbound Layer 2 Flows**: once the link recovers, the dynamic preemption process is started on the primary aggregation node (by invoking the "L2_PATH_CARRIER-DELAY" script). Once the carrier delay expires, the secondary aggregation node deactivates the aggregation link (or better, removes the Layer 2 VLANs from it) and this triggers the "L2_VLANs_START" script on the primary node to activate the primary path. Therefore, a similar impact to the traffic flows of what is observed after the failure is expected after recovery and this is the result of having preemption in place. If this is not desired, the preemption mechanism can be removed. In that case the secondary path would remain active even after the primary aggregation link is recovered. For optimizing the convergence of Layer 2 multicast flows, the preceding static mrouter configuration must be applied to the primary aggregation node.

N7K-Agg1

```
vlan 421
  ip igmp snooping mrouter interface Ethernet1/1
```

- **Inbound Layer 2 Flows**: Apply the considerations made for inound flows in Convergence After Recovery above.

Table 2-22 summarizes the convergence results for inbound and outbound traffic flows.

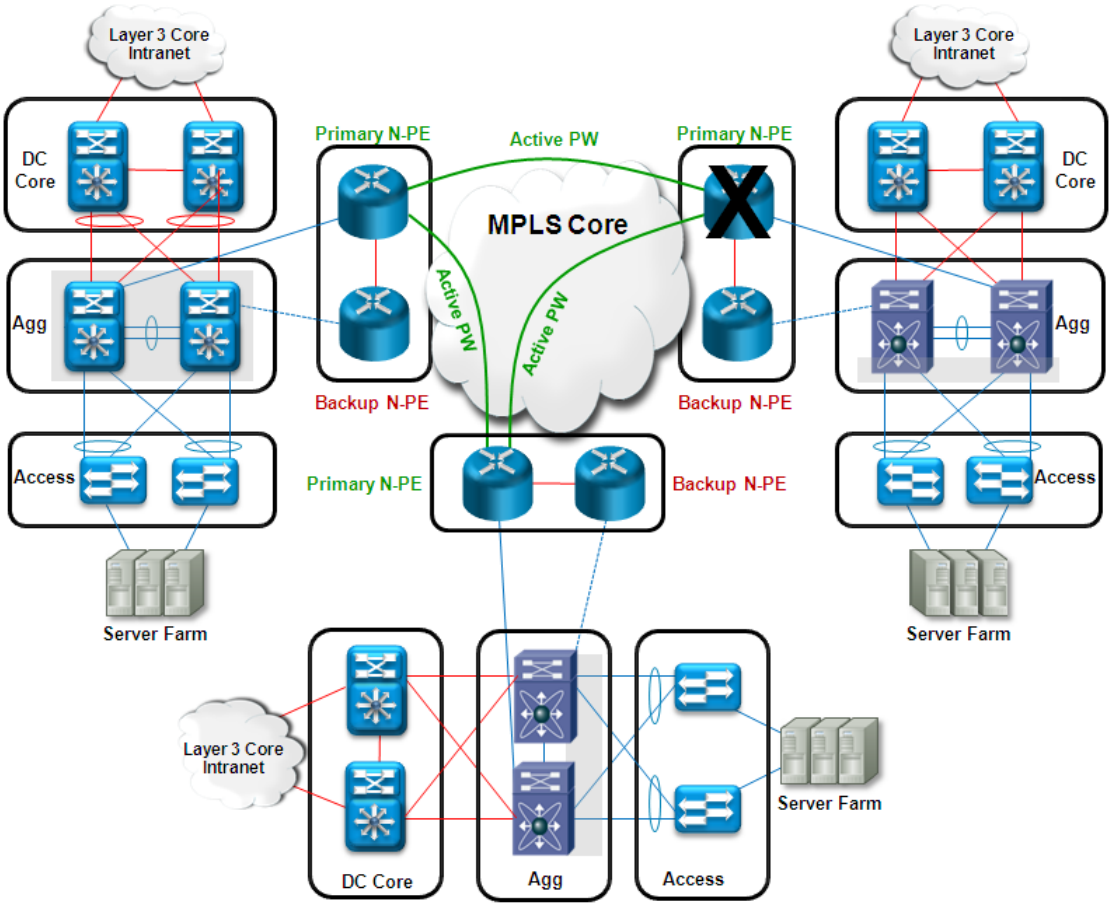
Table 2-22 Aggregation to DCI Layer Link Failure/Recovery Results

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	1.9 sec	1.9 sec	2.2 sec	2.2 sec
Layer 2 Multicast	1 sec	1.9 sec	1.6 sec	2.1 sec

Test 2: Failure/Recovery N-PE Device (using Nexus 7000 in Aggregation)

This scenario is highlighted in Figure 2-61.

Figure 2-61 Failure/Recovery N-PE Device



The same considerations made for the link failure/recovery scenario are valid here. This is because the failure or recovery of the N-PE device essentially is seen as a link failure/recovery from the point of view of the primary aggregation node.

Table 2-23 summarizes the convergence results for inbound and outbound traffic flows.

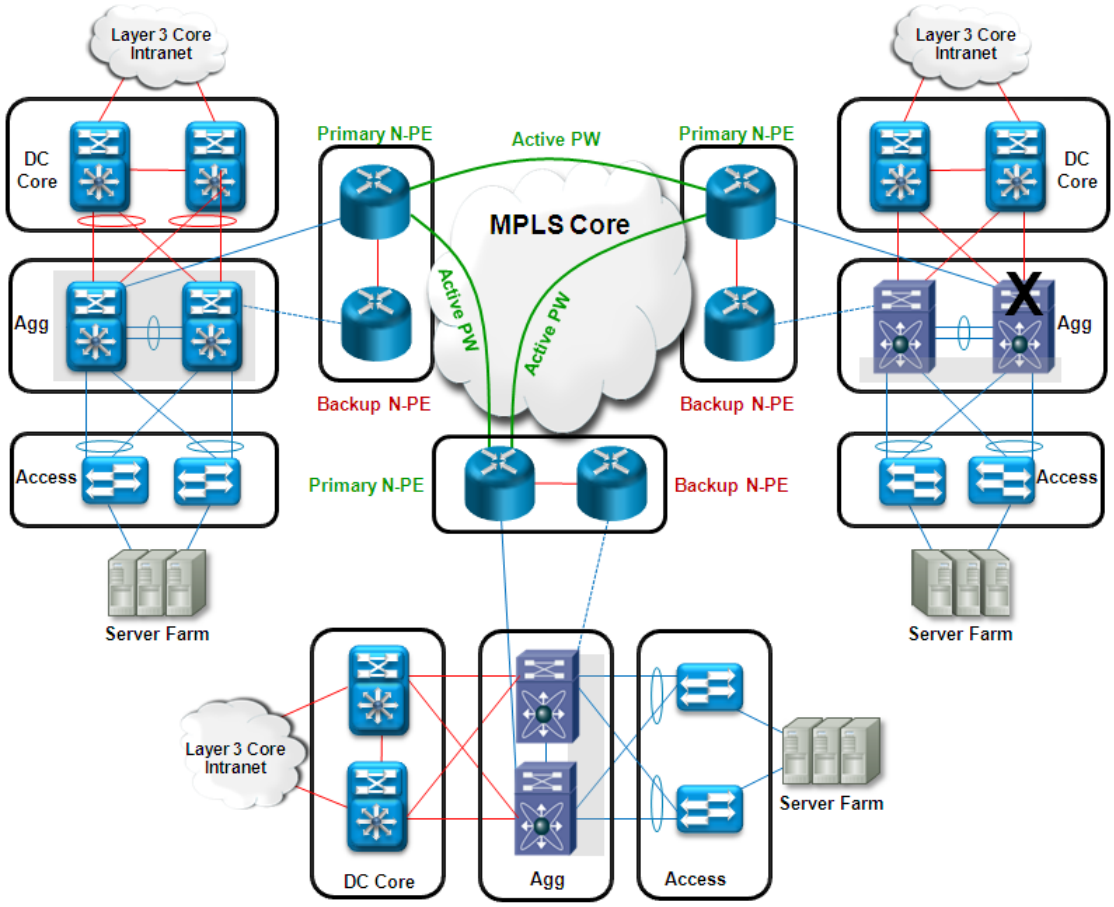
Table 2-23 Failure/Recovery N-PE Device Results

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	2.1 sec	2.1 sec	2.4 sec	2.4 sec
Layer 2 Multicast	1.4 sec	2 sec	1.5 sec	2.3 sec

Test 3: Supervisor Switchover on Nexus 7000 Aggregation Device

This scenario is shown in Figure 2-62.

Figure 2-62 Supervisor Failover on Primary Aggregation Node



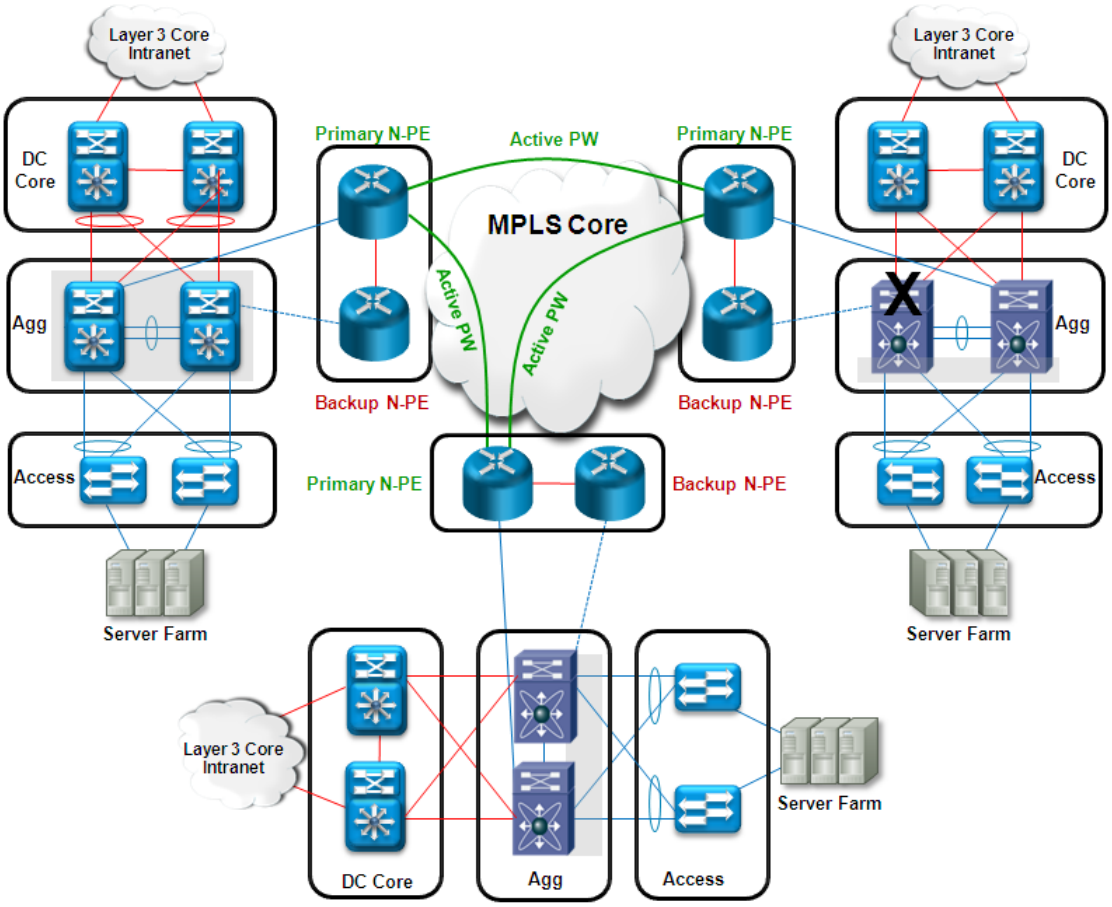
In this case the expectation is no packets should be lost during the supervisor failover. This is because on Nexus 7000 platforms the switching capabilities of the fabric ASIC are independent from the state of the Supervisor. Therefore, in a Supervisor switchover event, the data path between the linecards is not affected at all. The results shown in Table 2-24 confirm this behavior.

Table 2-24 Supervisor Failover on Primary Aggregation Node

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	0 sec	0 sec	0 sec	0 sec
Layer 2 Multicast	0 sec	0 sec	0 sec	0 sec

An interesting consideration needs to be made for the scenario where the supervisor failover is happening on the secondary aggregation node, as shown in Figure 2-63.

Figure 2-63 Supervisor Failover on Secondary Aggregation Node



Despite that there should be no effect on Layer 2 traffic in this case, it was observed during testing that after the supervisor switchover, all routing prefixes (including the P Semaphore) were deleted from the software routing table of the standby supervisor that become active. Since EEM tracks routing information in the software RIB, this event triggered the EEM script to activate the secondary path toward the DCI layer, as a result of the loss of the P Semaphore prefix.

This caused the secondary node to raise the B Semaphore, and therefore, the primary node deactivated the primary path. The end result was a traffic delay (around 2 sec). The same traffic delay was then experienced a few minutes later when the primary node assumed the active role.

The recommended workaround is to tune the "L2_VLANs_START" script on the secondary node. This allows the script to kick in once the P Semaphore prefix disappears from its routing table but only if at the same time the secondary node's routing table contains an IP prefix learned from the N-PE device. In the supervisor failover scenario, both prefixes would disappear from the routing table, and the script would not be triggered (which is the desired behavior).

The following configuration shows the modified EEM script.

N7K-AGG2

```
track 59 ip route 11.0.2.102/32 reachability
track 62 ip route 10.70.62.3/24 reachability
track 100 list boolean and
  object 62 not
  object 59
```

```
!
event manager applet L2_VLANs_START
  event track 100 state up
  action 0.1 cli en
  action 1.0 cli conf t
  action 4.0 cli int eth1/1
  action 4.1 cli sw trunk allow vlan <L2_VLAN_RANGE>,<IGP_VLAN>,4064
  action 5.0 cli clear mac add dyn
  action 9.0 syslog msg Backup Node is active
```

Tracking object 59 represents a specific IP address learned from the PE device (via the IGP peering discussed in Inter Data Centers Routing Considerations, page 2-77). Object 100 triggers the L2_VLANs_START script as described in the recommended workaround.

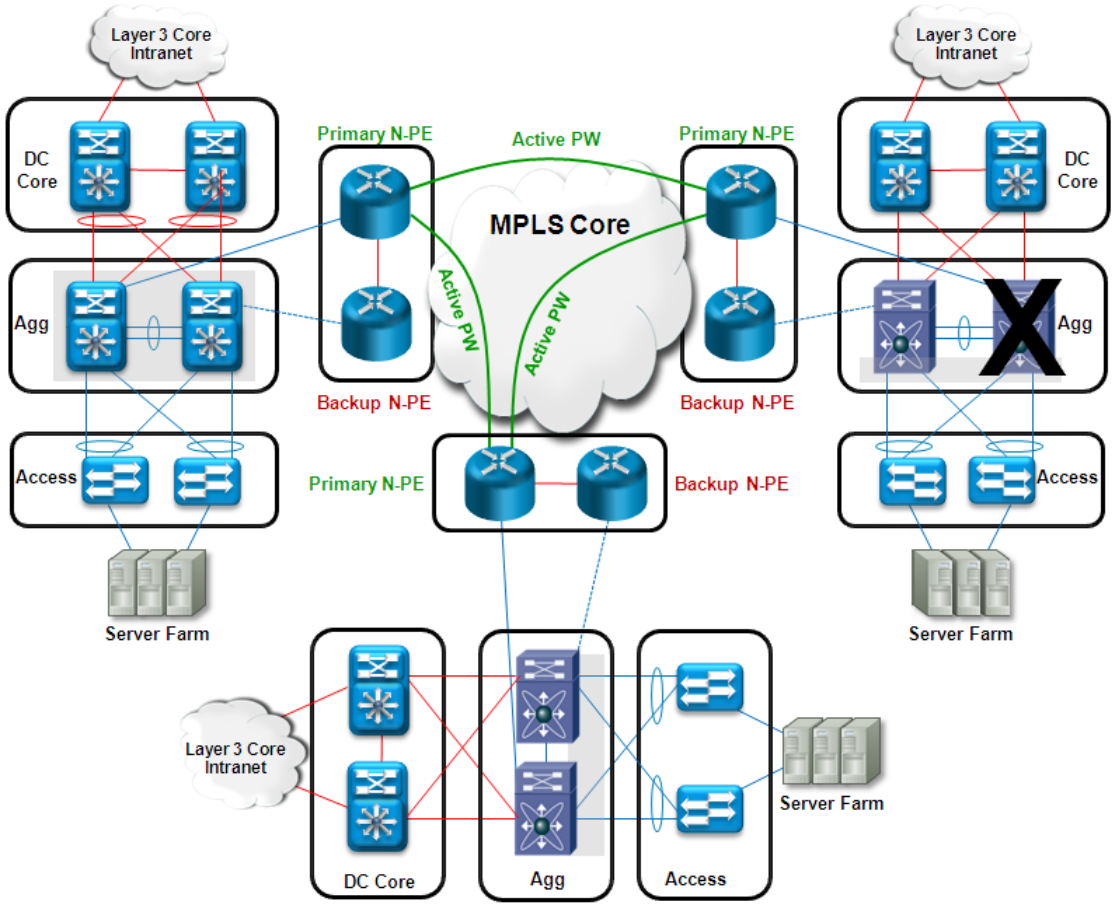
Because of an internal EEM logic, it was observed during testing that the order of the track statements is actually relevant. To achieve the desired goal of not triggering the EEM script after a supervisor switchover event, it is hence recommended to use a lower track ID (track 59) for the IP prefix learned from the DCI device, whereas an higher track number (track 62) should be dedicated to tracking the IP prefix associated to the P semaphore.

Note It is not required to modify the equivalent script on the primary aggregation node, because no semaphore routes are learned at the moment of the supervisor switchover (since the node is already the active one).

Test 4: Failure/Recovery Nexus 7000 in Aggregation

This failure scenario is shown in Figure 2-64.

Figure 2-64 Failure/Recovery Nexus 7000 Aggregation Device



Convergence After Failure

- **Outbound Layer 2 Flows**: when the aggregation switch fails, it triggers the EEM script on the secondary node to activate the secondary path. The only difference is that the notification that the P Semaphore disappeared is now faster, because it is not caused by a routing update but simply by the disappearing of the peer link from which the semaphore prefix was learned.
- **Inbound Layer 2 Flows:** from an inbound traffic perspective, this is the exact same scenario already discussed for the aggregation link failure scenario. Again, it is expected to get a better convergence since the secondary path is activated faster (because of the faster notification that the P Semaphore disappeared).

Convergence After Recovery

- **Outbound Layer 2 Flows**: when the aggregation layer device comes back online, the first EEM script that is activated is "L2_VLANs_BOOT_HOLD" (once the module where the aggregation link belongs comes back online). Once the boot hold timer expires, the primary aggregation node sends the preemption message to the secondary node and regains the active role.
- **Inbound Layer 2 Flows**: the inbound flows are affected when the primary aggregation node regains the active role. The recovery mechanism here is similar to the link recovery previously discussed.

Table 2-25 summarizes the convergence results for inbound and outbound traffic flows.

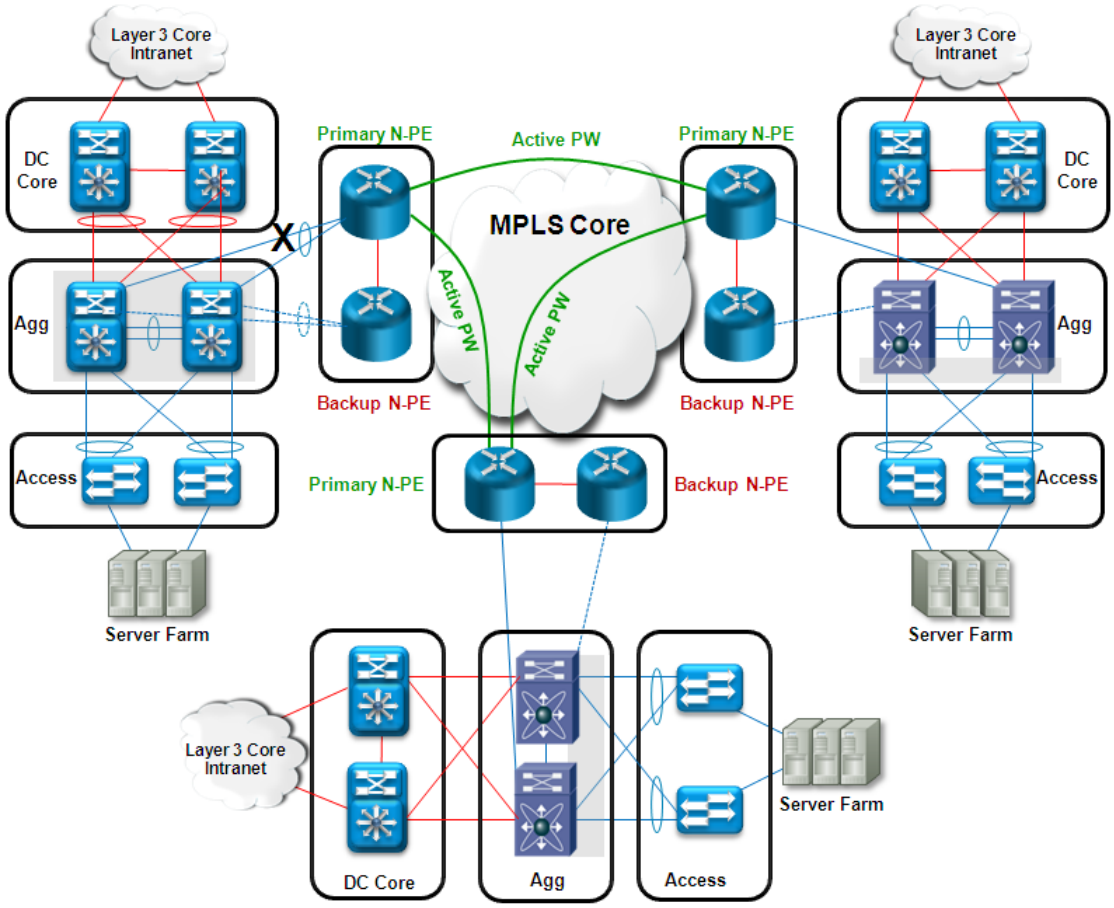
Table 2-25 Failure/Recovery Primary Aggregation Node

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	1.3 sec	1.3 sec	3 sec	2.4 sec
Layer 2 Multicast	3.4 sec	1.3 sec	2.5 sec	2.5 sec

Test 5: Failure/Recovery Port-Channel Link between VSS Aggregation and N-PE Device

This failure scenario is shown in Figure 2-65.

Figure 2-65 Failure/Recovery Port-Channel Link between VSS Aggregation and N-PE Device



The obvious advantage of using a PortChannel as primary traffic path for Layer 2 traffic is that recovery after a physical link failure is driven by rehashing the flows on the only remaining link and does not require EEM to be triggered. This consideration is valid for inbound/outbound traffic and both for link failure and recovery scenarios.

Ensure the reliability of the overall PortChannel by bundling links belonging to separate line-cards. This would ensure the only single point of failure that may cause the entire PortChannel to fail is the failure of the N-PE device (see Test 6: Failure/Recovery N-PE Device (using VSS in Aggregation), page 2-89).

Table 2-26 summarizes the convergence results for inbound and outbound traffic flows.

Table 2-26 Failure/Recovery Port-Channel Link between VSS Aggregation and N-PE Device

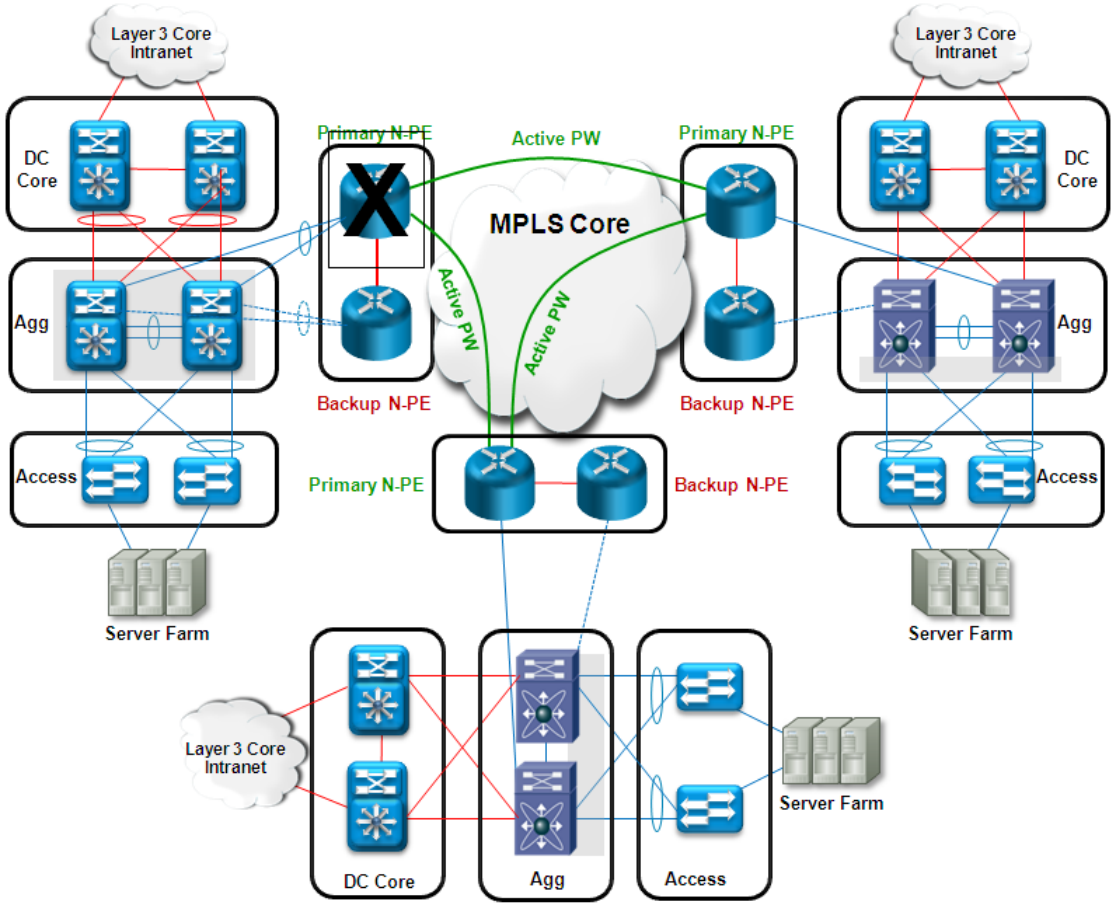
	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	0.5 sec	0.4 sec	0.15 sec	0.23 sec
Layer 2 Multicast	1.8 sec	N/A[1]	N/A[2]	0.11 sec

1. No multicast flows were sent/received on the specific link that failed/recovered.
2. Same as 1

Test 6: Failure/Recovery N-PE Device (using VSS in Aggregation)

This scenario is the same as the one discussed in Test 2: Failure/Recovery N-PE Device (using Nexus 7000 in Aggregation), page 2-82, with the only difference being that a VSS is deployed at the data center aggregation layer (Figure 2-66).

Figure 2-66 Failure/Recovery N-PE Device (VSS in Aggregation)



Convergence After Failure

- **Outbound Layer 2 Flows**: the failure of the N-PE device causes the entire PortChannel (representing the primary Layer 2 traffic path) to fail. This triggers the BACKUP_EC_ACTIVATE script that ensures the activation of the secondary PortChannel that connects the VSS aggregation device to N-PE2 in the DCI layer. The main factor affecting the Layer 2 unicast traffic convergence is how fast the VLANs be enabled on the secondary trunk. During testing, VSS behavior with 12.2(33)SXI3 code was suboptimal, so check with your Cisco representative to determine the best software release to deploy.

  For the recovery of Layer 2 multicast stream, the same considerations identified in Test 1: Failure/Recovery Active Link between Nexus 7000 and N-PE1, page 2-80 apply. Therefore, a static mrouter port configuration is applied, so that as soon as the secondary PortChannel is activated, multicast traffic is sent out to reach receivers located in remote data centers. The main difference is the way this static mrouter port configuration can be applied on Catalyst 6500: instead of being a global command, the command are applied under the SVI where the receivers are placed. The configurations follows:

VSS-Agg

```
interface Vlan412
 ip igmp snooping mrouter interface Port-channel70
```

Also, and most importantly, you cannot specify an interface (like Po70 in the example above) unless the corresponding VLAN is already enabled on it. This implies that in order for the static mrouter port configuration to be applied when the secondary path (Po80) is activated by the EEM script, you must add that specific configuration lines as part of the script. The modified BACKUP_EC_ACTIVATE applet configuration follows.

VSS-Agg

```
event manager applet BACKUP_EC_ACTIVATE
 event track 70 state down
 action 0.1 cli command "en"
 action 1.0 cli command "conf t"
 action 2.0 cli command "int po 80"
 action 2.1 cli command "switchport trunk allow vlan <L2_VLANs_RANGE>,<IGP_VLAN>"
 action 2.2 cli command "interface vlan 412"
 action 2.3 cli command "ip igmp snoop mrouter interface po 80"
 action 3.0 cli command "int po 70"
 action 3.1 cli command "switchport trunk allow vlan <IGP_VLAN>"
 action 4.0 cli command "inter vlan 412"
 action 4.1 cli command "no ip igmp snoop mrouter interface po 70"
 action 9.0 syslog msg "Backup Port-channel activated"
```

In this example the multicast receivers are deployed in VLAN 412.

- **Inbound Layer 2 Flows:** because of the failure of the PE device, the PW established to it from the remote PEs will go down. Therefore, the recovery of inbound flows would also be dictated by the activation of the secondary PortChannel, since the remote PE would initially flood traffic over the PWs connecting to N-PE2.

Convergence After Recovery

- **Outbound Layer 2 Flows**: the recovery of the PE device would cause also the restoration of the primary port-channel. This would trigger the two scripts required to re-activate the primary path (PRIMARY_EC_ACTIVATE) and de-activate the secondary path (BACKUP_EC_DEACTIVATE). Once again, a minor modification to the PRIMARY_EC_ACTIVATE applet is required to apply the proper static mrouter configuration (needed for Layer 2 multicast traffic).

VSS-Agg

```
event manager applet PRIMARY_EC_ACTIVATE
 event track 70 state up maxrun 1000
 action 0.1 cli command "en"
 action 0.5 cli command "ping 1.1.1.1 time 300 repeat 1"
 action 1.0 cli command "conf t"
 action 3.0 cli command "int po 70"
 action 3.1 cli command "switchport trunk allow vlan <L2_VLANs_RANGE>,<IGP_VLAN>"
 action 3.2 cli command "inter vlan 412"
 action 3.3 cli command "ip igmp snoop mrouter interface po 70"
 action 5.0 cli command "do clear mac-add dyn"
 action 6.0 cli command "inter vlan 412"
 action 6.1 cli command "no ip igmp snoop mrouter interface po 80"
 action 9.0 syslog msg " Primary Port-channel activated "
```

- **Inbound Layer 2 Flows**: recovery for inbound flows is mostly dictated by the time required to move the VLANs from the secondary to the primary path.

Table 2-27 summarizes the convergence results for inbound and outbound traffic flows.

Table 2-27 Failure/Recovery N-PE Device (VSS in Aggregation)

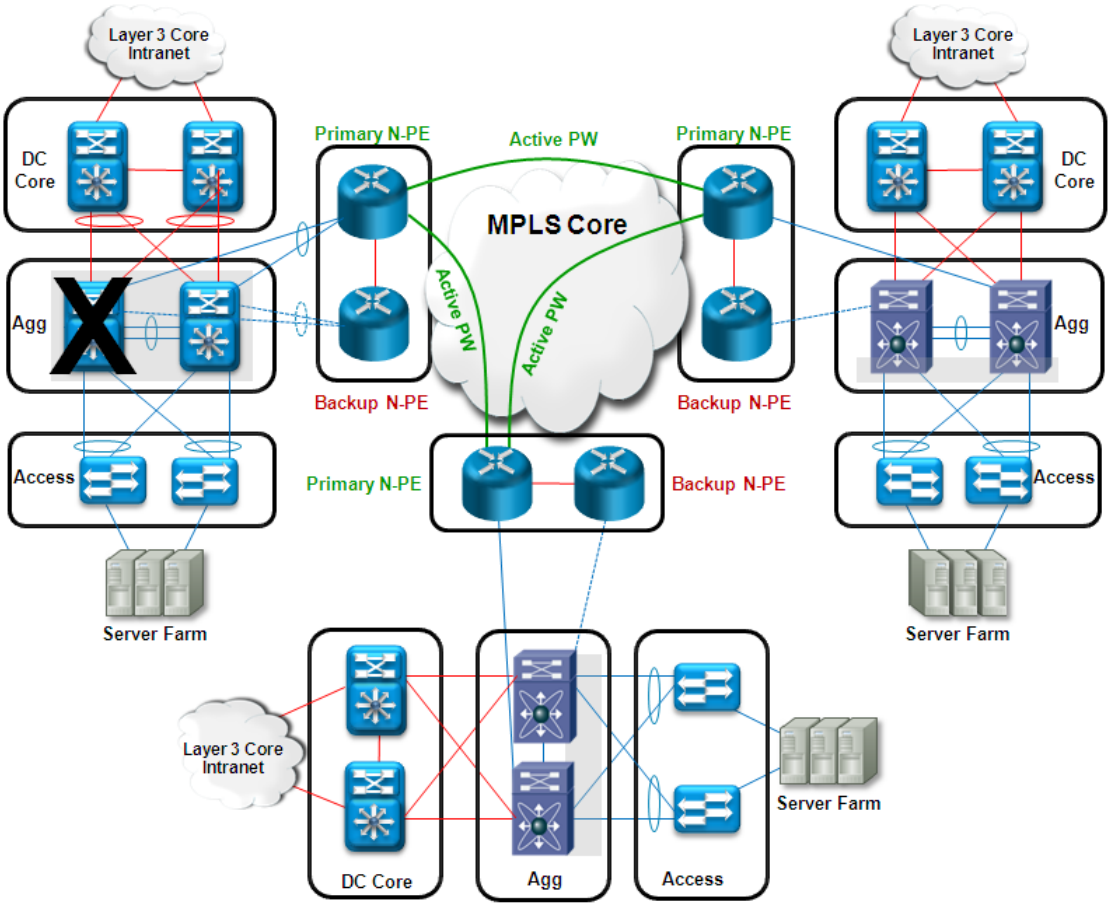
	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	1.9 sec to 5 sec[1]	1.9 sec to 5 sec[1]	1.6 sec to 6.2 sec[1]	2 sec to 7.2 sec[1]
Layer 2 Multicast	1.9 sec[2]	1.5 sec[2]	1.4 sec[2]	1.7 sec[2]

1. A linear recovery was observed across the 300 VLANs extended between sites. The longest outage shown in table above was observed with 300 VLANs carried on the trunk.
2. No linear effect was observed here due to that multicast flows were only activated on a reduced number of VLANs (less than 10).

Test 7: Failure Active VSS Member

Because each VSS member can only support a single Supervisor, both the supervisor failover and the switch reload test cases are shown in Figure 2-67.

Figure 2-67 Failure Active VSS Member



Convergence After Failure

- **Outbound Layer 2 Flows**: the failure of the active VSS member forces the access switches to re-hash the flows to the remaining uplink connecting to the secondary VSS member device. The secondary VSS member would then take care of sending the traffic to the remaining physical link part of the primary PortChannel connecting to the DCI layer (Po70). EEM is not required to take any action in this scenario.
- **Inbound Layer 2 Flows:** recovery for inbound flows depends on the capabilities of the DCI layer device (N-PE1) to re-hash flows on the remaining link of the primary PortChannel that connects to the VSS aggregation device (Po70).

Convergence After Recovery

- **Outbound Layer 2 Flows**: recovery of the VSS member causes the recovery of the uplink between VSS and access layer devices. Traffic is rehashed across both physical members of the PortChannel that connects these devices.
- **Inbound Layer 2 Flows**: recovery of the VSS member causes the recovery of the physical member of the PortChannel that connects the aggregation VSS to the N-PE1 device. Traffic is rehashed across both physical members of the PortChannel that connects these devices.

Table 2-28 summarizes the convergence results for inbound and outbound traffic flows.

Table 2-28 Failure Active VSS Member

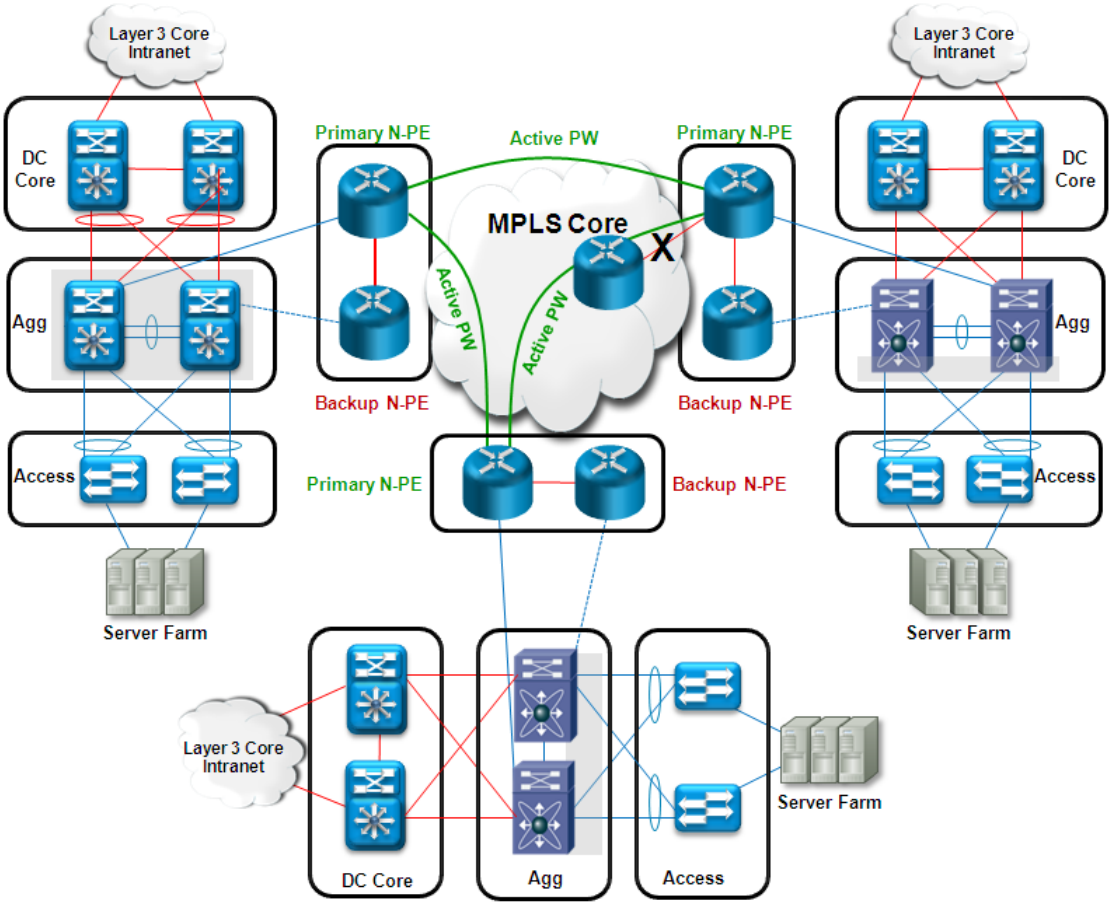
	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	1.2 sec	0.5 sec	0.4 sec	0.7 sec
Layer 2 Multicast	1.2 sec	N/A[1]	0.9 sec	0.9 sec

1. No multicast flows were sent over the link, so the recovery did not cause any outage.

Test 8: Failure/Recovery Core Link N-PE1 (using N7K or VSS in Aggregation)

This scenario is depicted in Figure 2-68. The discussion here applies to both scenarios where a pair of Nexus 7000 (or a VSS switch) are deployed in the aggregation layer.

Figure 2-68 Failure/Recovery Core Link N-PE1



Convergence After Failure

- **Outbound Layer 2 Flows**: the failure of the PE link that connects to the MPLS core does not execute any EEM script. This is because the MPLS traffic can be rerouted across the Layer 3 transit link available between the two PE devices. The overall outage is then mainly dictated by how long

it will take for the Layer 3 rerouting to happen. To minimize the overall traffic outage, it is critical that the logical PWs established with the remote PE devices remain active while the IGP converges and the MPLS traffic is rerouted. To do this, add the following commands on all the PE devices:

```
mpls ldp holdtime 60
ip routing protocol purge interface
```

Without the second command above, the IGP routes originally learned from the MPLS core would be purged from the routing table once the connection to the core fails. The routes would then be added back after the IGP converges, but that would not stop the PWs from going down, causing a longer traffic outage. Applying the command basically allows the IGP protocol to take control, so the routes are not deleted and successively added back but simply modified after IGP convergence is completed.

Note If the Layer 3 link between the two PE was not available, the failure of the connection to the MPLS core would essentially isolate the primary PE. To recover from this failure scenario, enable an EEM script on the PE device to bring down the Layer 2 trunk to the primary aggregation node. This allows the secondary aggregation node to become active and recover traffic through the secondary path. Discussing this scenario is beyond the scope of this document; always connect a Layer 3 transit link between the PE devices.

- **Inbound Layer 2 Flows:** rerouting traffic inside the MPLS core recovers inbound flows. This is because the VPLS traffic directed to N-PE1 loopback address is first delivered from the core to N-PE2 and then reaches N-PE1 via the transit link.

Convergence After Recovery

- **Outbound and Inbound Layer 2 Flows**: once the link to the core is recovered, traffic rerouted to the shortest path in and out of the N-PE1 device.

Table 2-29 summarizes the convergence results for inbound and outbound traffic flows.

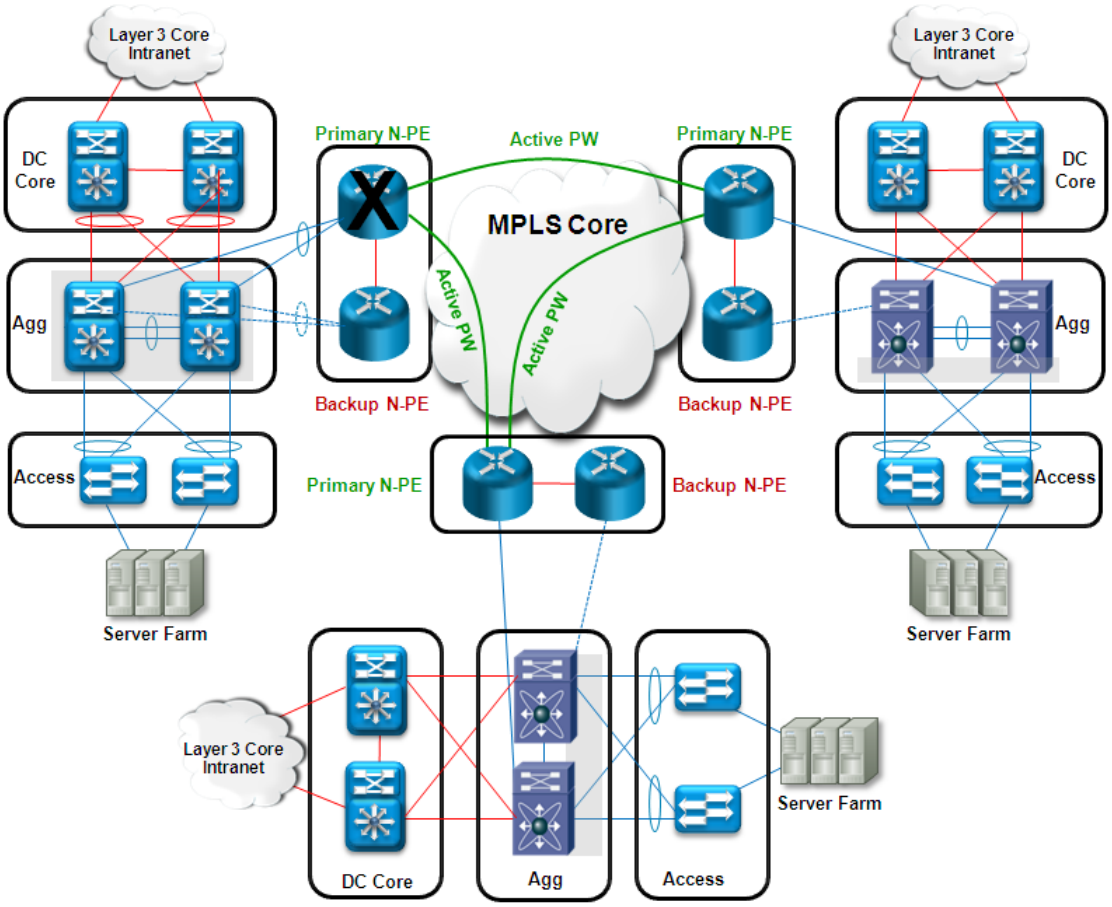
Table 2-29 Aggregation to DCI Layer Link Failure/Recovery Results

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	2.4 sec	2.4 sec	2 sec	2 sec
Layer 2 Multicast	1.5 sec	2.3 sec	1.7 sec	1.9 sec

Test 9: Supervisor Switchover N-PE Device (N7K or VSS in Aggregation)

This final test case is shown in Figure 2-69.

Figure 2-69 Supervisor Switchover N-PE Device



The main issue with this failure scenario is the lack of NSF/SSO support for VPLS on Catalyst 6500 platforms. This means that a supervisor failover would essentially cause all the PW sessions established with the remote PEs to be reset. Traffic black-holed until the PW comes back online, since the primary PortChannel to the VSS in aggregation remains active. Therefore, we recommend not deploying redundant supervisors on the N-PE devices. Instead, rely on physical box redundancy to provide a resilient solution. A supervisor failure in this case would cause the entire PE box to go down, re-creating the failure scenario discussed in Test 2: Failure/Recovery N-PE Device (using Nexus 7000 in Aggregation), page 2-82, and Test 6: Failure/Recovery N-PE Device (using VSS in Aggregation), page 2-89.

Table 2-30 summarizes the convergence results for inbound and outbound traffic flows.

Table 2-30 Supervisor Switchover N-PE Device

	Failure		Recovery	
Traffic Flows	**Outbound**	**Inbound**	**Outbound**	**Inbound**
Layer 2 Unicast	153 sec	153 sec	0 sec	0 sec
Layer 2 Multicast	155 sec	122 sec	0 sec	0 sec

Summary of Design Recommendations

This section summarizes the major design recommendations for Point-to-Point and Multipoint deployment derived from the DCI phase 2.0 validation effort.

- Point-to-Point Deployment Recommendations, page 2-96
- Multipoint Deployment Recommendations, page 2-96

Point-to-Point Deployment Recommendations

We recommend the following when you deploy a point-to-point DCI topology:

- Each aggregation layer device should be connected to both PEs deployed in the DCI layer in a fully meshed fashion.
- Leverage an MPLS enabled Layer 3 link to interconnect the PEs deployed in the same data center location.
- Devices supporting Multi Chassis EtherChannel capabilities (VSS or vPC) should be deployed in aggregation layer. This allows establishing end-to-end PortChannels between the aggregation layer switches deployed in remote data center sites (leveraging EoMPLS PWs as logical extension of the physical links).
- LACP (802.3ad) should be used as control plane replacing STP between data centers (static EtherChannel bundling is not recommended).
- BPDU filtering should be applied on the logical PortChannels established between remote sites to ensure the creation of two completely isolated STP domains.
- EEM should be deployed on Catalyst 6500 PEs to provide the Remote Ethernet Port Shutdown functionality.
- Disable replay-protection on Nexus 7000 802.1AE enabled interface when Catalyst 6500 switches are deployed as PE devices.
- Connect a separate Layer 3 link between each aggregation switch and the PEs in the DCI layer to establish inter-DC Layer 3 connectivity across the MPLS cloud. Xconnect the Layer 3 interface on the PE devices to allow for the establishment of end-to-end IGP adjacencies between the remote aggregation switches.
- Leverage loopback interfaces as source and destination points for establishing the logical GRE connections between remote PE devices.
- Use static route configuration to enforce the establishment of the EoMPLS PW across the GRE tunnel.
- Increase the MTU size both on the physical interfaces connecting to the IP core and on the logical tunnel interface.
- Aggressively tune (1 sec, 3 sec) the GRE keepalive timers.

Multipoint Deployment Recommendations

We recommend the following when you deploy a multipoint DCI topology:

- Establish a full mesh of VPLS pseudowires (PWs) between the various PE devices, with the exception of the PW between PEs belonging to the same data center.

- Leverage an MPLS enabled Layer 3 link to interconnect the PEs deployed in the same data center location.
- Increase the MTU on all the MPLS enabled links (jumbo size when supported).
- Deploy EEM on the aggregation layer devices to protect against the creation of end-to-end STP loops.
- When you want to provide traffic load-balancing across both N-PE devices deployed in the DCI layer, divide in two groups the set of VLANs requiring LAN extension services and enable one group on the upper physical path and the second group on the lower physical path that connects the aggregation layer to the DCI layer.
- Properly increase the available bandwidth on the vPC peer link between the Nexus 7000 in the aggregation layer since it is used by design for 50% of the outbound Layer 2 traffic flows.
- Increase the resiliency of the vPC peer link by bundling together interfaces belonging to different modules.
- Provide a dedicated IGP instance for exchanging semaphore state information between the Nexus aggregation nodes.
- Use a dedicated VLAN on the Layer 2 trunks interconnecting the aggregation layer to the DCI layer devices to create an IGP peering between them.
- Summarize the IP prefixes belonging to each specific data center site before injecting them into the MPLS core.
- Statically configure the active and backup Layer 2 trunks as mrouter ports to optimize Layer 2 multicast traffic convergence.
- Configure the **ip routing protocol purge interface** command on all PE devices.
- Do not use redundant supervisors in the Catalyst 6500 switches deployed in the DCI layer.

Summary

Cisco is the leading global supplier of internetworking solutions for corporate intranets and the global Internet. In light of disaster recovery responsibilities and business continuum, regulatory compliance has emerged as a most challenging issue facing business and enterprise. From Securities and Exchange (SEC) Rule 17a to Sarbanes-Oxley and HIPAA, many legislative requirements now dictate how electronic data is stored, retrieved, and recovered. Organizations failing to meet new mandates face significant penalties and incalculable risk to corporate position and reputation.

Cisco Data Center Interconnect compatible platforms supersede competitor portfolio gaps for DCI profiling by meeting and supporting Layer 2 transport mechanisms, high availability, spanning tree isolation, loop avoidance, multipath load balancing, swift end-to-end convergence and aggregation across WANs, queuing, and interconnect traffic encryption.

The Cisco DCI system release 2.0 is a loop free multi-path solution delivering end-to-end Layer 2 extensions that include a resilient redundant design, VSS and vPC at the aggregation layers, and SIP-400 using MPLS connectivity types EoMPLS, EoMPLS with 802.1AE, and VPLS, and IP connectivity types EoMPLSoGRE and VPLSoGRE.

To this end, the Cisco Data Center Interconnect (DCI) system solution deploys state-of-the-art technology on robust proprietary platforms for strategic customers to extend subnets beyond Layer 3 boundaries of single site data centers, stretching clustered Layer 2 connected node functionality to promote resilient routing flexibility, and virtualization, while offsetting STP loops and broadcast storms, and identifying active IP addresses or subnets.

DCI System Release 2.0 tested two data center (point to point) and multiple data center (multipoint) scenarios. The tests covered the most common to highly unlikely failure scenarios under constant traffic loads. In most scenarios, convergence times were observed as sub-second, highlighting rare failure scenarios that caused a couple seconds of outage. All testing was performed with a mix of EMIX and IMIX traffic and the system was loaded to verify line rate scalability.

This release discovered a wide range of issues during testing. All show stopper and critical issues were resolved and verified within the scope of testing, and unresolved issues are documented as such.