Campus 3.0 Virtual Switching System Design Guide

Cisco Validated Design

January 19, 2011

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Text Part Number: OL-19829-01

Cisco Validated Design

The Cisco Validated Design Program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit www.cisco.com/go/validateddesigns.

ALL DESIGNS, SPECIFICATIONS, STATEMENTS, INFORMATION, AND RECOMMENDATIONS (COLLECTIVELY, "DESIGNS") IN THIS MANUAL ARE PRESENTED "AS IS," WITH ALL FAULTS. CISCO AND ITS SUPPLIERS DISCLAIM ALL WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE. IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THE DESIGNS, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

THE DESIGNS ARE SUBJECT TO CHANGE WITHOUT NOTICE. USERS ARE SOLELY RESPONSIBLE FOR THEIR APPLICATION OF THE DESIGNS. THE DESIGNS DO NOT CONSTITUTE THE TECHNICAL OR OTHER PROFESSIONAL ADVICE OF CISCO, ITS SUPPLIERS OR PARTNERS. USERS SHOULD CONSULT THEIR OWN TECHNICAL ADVISORS BEFORE IMPLEMENTING THE DESIGNS. RESULTS MAY VARY DEPENDING ON FACTORS NOT TESTED BY CISCO.

CCVP, the Cisco logo, and Welcome to the Human Network are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn is a service mark of Cisco Systems, Inc.; and Access Registrar, Aironet, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, Follow Me Browsing, FormShare, GigaDrive, HomeLink, Internet Quotient, IOS, iPhone, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, iQuick Study, LightStream, Linksys, MeetingPlace, MGX, Networkers, Networking Academy, Network Registrar, PIX, ProConnect, ScriptShare, SMARTnet, StackWise, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0711R)

Campus 3.0 Virtual Switching System Design Guide
© <year> Cisco Systems, Inc. All rights reserved.

CONTENTS

Preface

About the Guide

Audience

This design guide is intended for Cisco systems and customer engineers responsible for designing campus networks with Virtual Switching Systems 1440.

Document Objectives

This document provides design guidance for implementing the Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 within the hierarchical campus architecture. Chapter 1, "Virtual Switching Systems Design Introduction" covers traditional design approaches and architectural scope of campus. Chapter 2, "Virtual Switching System 1440 Architecture" introduces the critical components of VSS and provides best-practice design options and recommendation specific in configuring VSS in campus. Chapter 3, "VSS-Enabled Campus Design" discusses the application of VSS in campus and illustrates the traffic flow and campus-specific best practice design recommendation. Chapter 4, "Convergence" illustrates the validated design environment. It includes the convergence characteristics of end-to-end campus enabled with VSS.

Note Throughout the remainder of this guide, the Cisco Catalyst 6500 Series VSS will be referred to as the VSS.

This design guide references and uses accumulated best-practice knowledge documented and available in the documents listed in Appendix B, "References." However, during the development of this design guide, many design choices have been made to replace or update specific options recommended for VSS-specific deployments. This design guide makes explicit references to and/or reaffirms Cisco best practices.

Document Organization

This design guide contains the following chapters and appendices:

Section	Description
This chapter	Provides a brief summary of the content provided in this Campus 3.0 Virtual Switching Systems (VSS) solution publication
Chapter 1, "Virtual Switching Systems Design Introduction."	Provides an overview of VSS design presented in this publications
Chapter 2, "Virtual Switching System 1440 Architecture."	Addresses the architecture and components of Cisco Catalyst 6500 Series VSS 1440.
Chapter 3, "VSS-Enabled Campus Design."	Addresses three major parts of VSS campus design: • EtherChannel optimization, traffic flow and VSL capacity planning • Multilayer design best practices • Routing with VSS
Chapter 4, "Convergence."	Describes the convergence characteristics of end-to-end campus-enabled network with VSS.
Appendix A, "VSS-Enabled Campus Best Practice Configuration Example."	Provides VSS-enabled campus best practice configuration examples
Appendix B, "References."	Provides references and links to related documents

About the Author

Nimish Desai, Technical Lead, CMO Enterprise Systems Engineering (ESE), Cisco Systems.

Nimish currently works as a Technical Leader in the Data Center Application group within ESE. In ESE he was a lead architect on Virtual Switching System Solution development and verification of best practices designs for Cisco Campus networks. Before his work on the ESE Campus solutions team, Nimish worked with Cisco Advanced Services providing design consultation and technical escalation for large Enterprise customers.

Nimish has been working on inter-networking technology for the last 17 years. Before joining Cisco, Nimish developed expertise with large financial institution supporting trading floor, large-scale design of enterprise networks with logistics and insurance companies and product development experience with IBM. Nimish hold MSEE from New Jersey Institute of Technology. Nimish enjoys fishing and outdoor activities including RVing National Parks.

CHAPTER 1

Virtual Switching Systems Design Introduction

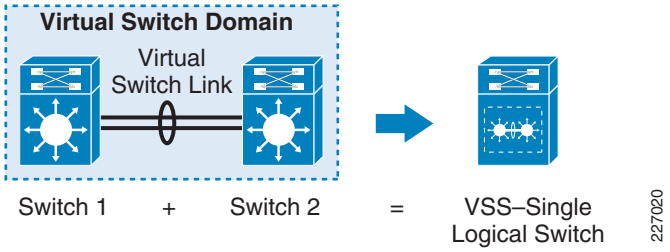
Executive Summary

VSS enables unprecedented functionality and availability of campus network by integrating network and systems redundancy into a single node. The end-to-end campus network enabled with VSS capability allows flexibility and availability described in this design guide.

The single logical node extends the integration of services in a campus network beyond what has been previously possible, without significant compromise. Integration of wireless, Firewall Services Module (FWSM), Intrusion Prevention System (IPS), and other service blades within the VSS allows for the adoption of an array of Service Ready Campus design capabilities. For example, VSS implementation allows for the applications of Internet-edge design (symmetric forwarding), data center interconnection (loop-less disaster recovery), and much more. Though this document only discusses the application of VSS in campus at the distribution layer, it is up to network designer to adapt the principles illustrated in this document to create new applications—and not just limit the use of VSS to the campus environment.

The key underlying capability of VSS is that it allows the clustering of two physical chassis together into a single logical entity. See Figure 1-1.

Figure 1-1 Conceptual Diagram of VSS



This *virtualization* of the two physical chassis into single logical switch fundamentally alters the design of campus topology. One of the most significant changes is that VSS enables the creation of a *loop-free* topology. In addition, VSS also incorporates many other Cisco innovations—such as Stateful Switch Over (SSO) and Multi-chassis EtherChannel (MEC)—that enable non-stop communication with increased bandwidth to substantially enhance application response time. Key business benefits of the VSS include the following:

- Reduced risk associated with a looped topology
- Non-stop business communication through the use of a redundant chassis with SSO-enabled supervisors
- Better return on existing investments via increased bandwidth form access layer

- Reduced operational expenses (OPEX) through increased flexibility in deploying and managing new services with a single logical node, such as network virtualization, Network Admission Control (NAC), firewall, and wireless service in the campus network
- Reduced configuration errors and elimination of First Hop Redundancy Protocols (FHRP), such as Hot Standby Routing Protocol (HSRP), GLBP and VRRP
- Simplified management of a single configuration and fewer operational failure points

In addition, the ability of the VSS to integrate services modules, bring the full realization of the Cisco campus fabric as central to the services-oriented campus architecture.

Virtual Switching System (VSS) Design

To better understand the application of the VSS to the campus network, it is important to adhere to existing Cisco architecture and design alternatives. The following section illustrates the scope and framework of Cisco campus design options and describes how these solve the problems of high availability, scalability, resiliency and flexibly. It also describes the inefficiency inherent in some design models.

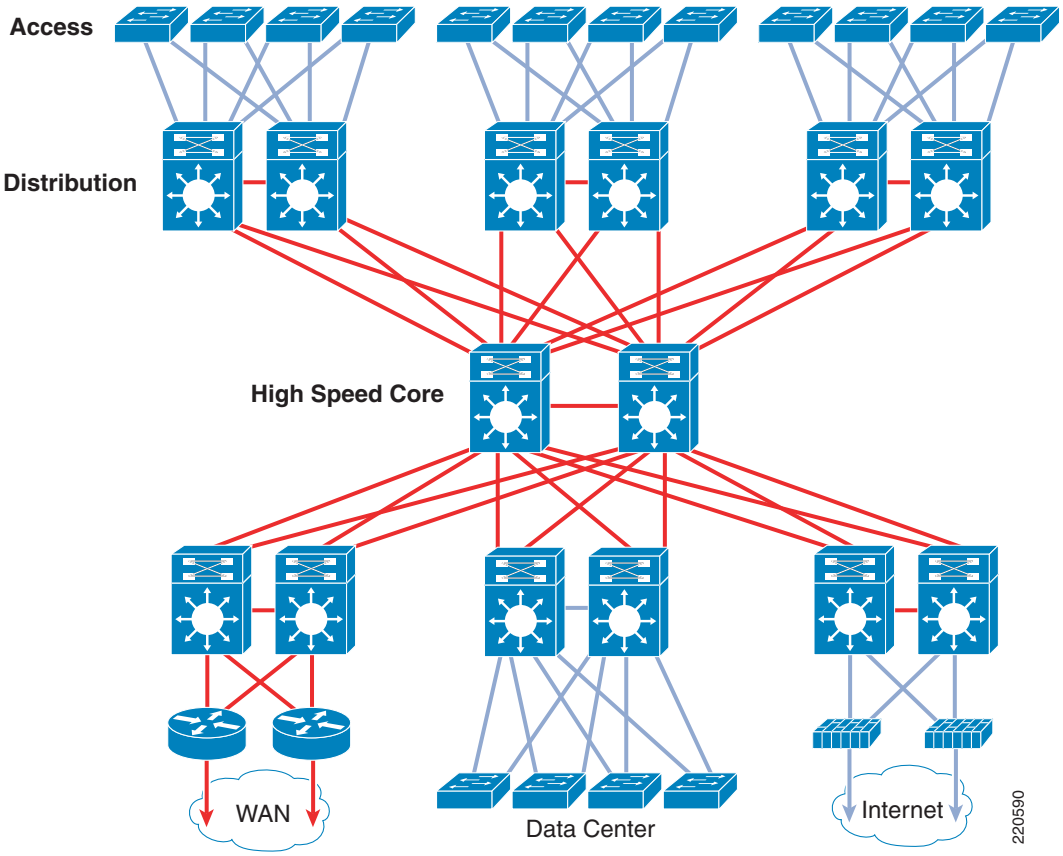
Campus Architecture and Design

The process of designing a campus architecture is challenged by new business requirements. The need for non-stop communication is becoming a basic starting point for most campus networks. The business case and factors influencing modern campus design are discussed in following design framework:

Enterprise Campus 3.0 Architecture: Overview and Framework
http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/campover.html

The use of hierarchical design principles provides the foundation for implementing campus networks that meet these requirements. The hierarchical design uses a building block approach that uses a high-speed routed core network layer to which multiple independent distribution blocks are attached. The distribution blocks comprise two layers of switches: the actual distribution nodes that act as aggregators for building/floors/section and the wiring closet access switches. See Figure 1-2.

Figure 1-2 Hierarchical Campus Building Blocks



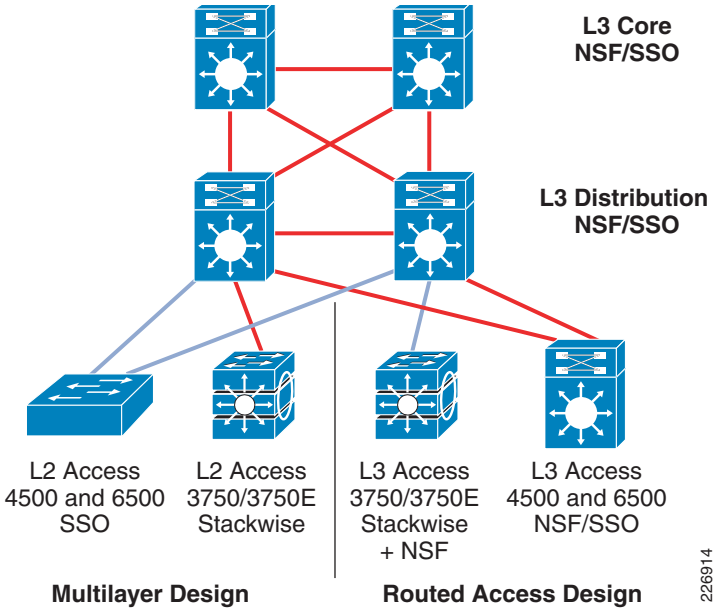
VSS at the Distribution Block

The Campus 3.0-design framework covers the functional use of a hierarchy in the network in which the distribution block architecture (also referred as access-distribution block) governs a significant portion of campus design focus and functionality. The access-distribution block comprises two of the three hierarchical tiers within the multi-tier campus architecture: the access and distribution layers. While each of these two layers has specific services and feature requirements, it is the network topology control plane design choices (the routing and spanning tree protocols) that are central to how the distribution block is glued together and how it fits within the overall architecture. There are two basic design options for how to configure the access-distribution block and the associated control plane:

- Multilayer or multi-tier (Layer 2 in the access block)
- Routed access (Layer 3 in the access block)

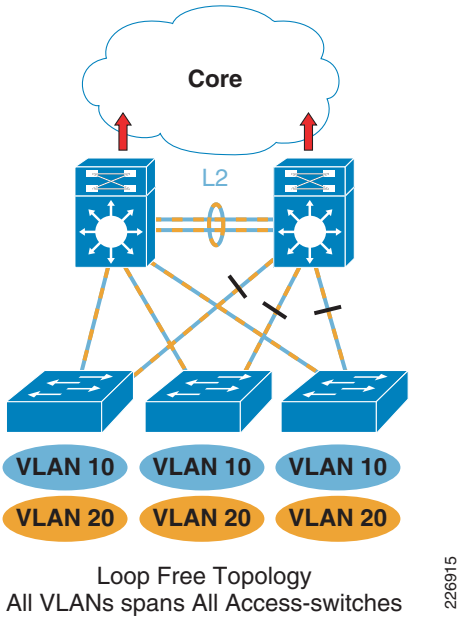
While these designs use the same basic physical topology and cabling plant, there are differences in where the Layer-2 and Layer-3 boundaries exist, how the network topology redundancy is implemented, and how load balancing works—along with a number of other key differences between each of the design options. Figure 1-3 depicts the existing design choices available.

Figure 1-3 Traditional Design Choices



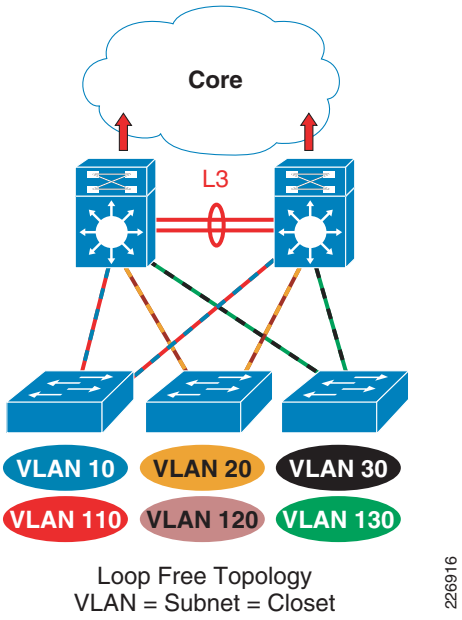
The multilayer design is the oldest and most prevalent design in customer networks while routed access is relatively new. The most common multilayer design consists of VLANs spanning multiple access-layer switches to provide flexibility for applications requiring Layer-2 adjacency (bridging non-routable protocols) and routing of common protocol, such as IPX and IP. This form of design suffers from a variety of problems, such as instability, inefficient resources usage, slow response time, and difficulty in managing end host behavior. See Figure 1-4.

Figure 1-4 Multilayer Design—Looped Topology



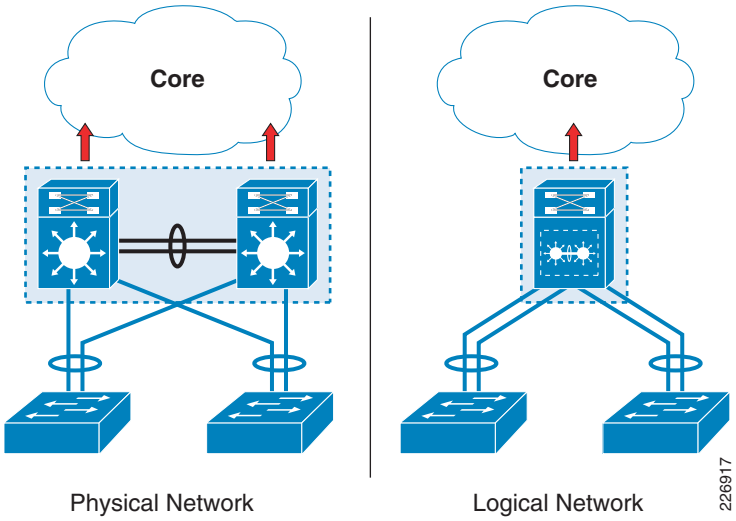
In the second type of multilayer design, VLANs do not span multiple closets. In other words VLAN = Subnet = Closet. This design forms the basis of the best-practice multilayer design in which confining VLANs to the closet eliminate any potential spanning tree loops. See Figure 1-5. However, this design does not allow for the spanning of VLANs. As an indirect consequence, most legacy networks have retained a looped spanning tree protocol (STP)-based topology—unless a network topology adoption was imposed by technology or business events that required more stability, such as implementation of voice over IP (VoIP).

Figure 1-5 Multilayer Design—Loop Free Topology



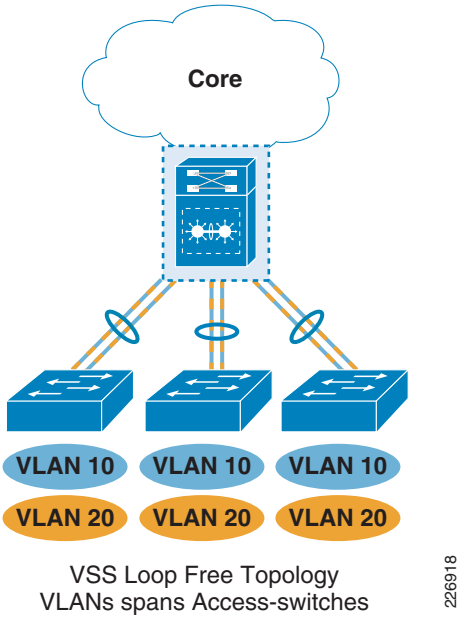
When VSS is used at the distribution block in a multilayer design, it brings the capability of spanning VLANs across multiple closets, but it does so without introducing loops. Figure 1-6 illustrates the physical and logical connectivity to the VSS pair.

Figure 1-6 Virtual Switch at the Distribution layer



With VSS at the distribution block, both multilayer designs transform into one design option as shown in Figure 1-7 where the access layer is connected to single logical box through single logical connection. This topology allows the unprecedented option of allowing VLANs to span multiple closets in loop-free topology.

Figure 1-7 VSS-Enabled Loop-Free Topology



The application of VSS is wide ranging. VSS application is possible in all three tiers of the hierarchical campus—core, distribution, and access—as well as the services block in both multilayer and routed-access designs. However, the scope of this design guide is intended as an application of VSS at the distribution layer in the multilayer design. It also explores the interaction with the core in that capability. Many of the design choices and observations are applicable in using VSS in routed-access design because it is a Layer-3 end-to-end design, but the impact of VSS in multilayer is the most significant because VSS enables a loop-free topology along with the simplification of the control plane and high availability.

Application of VSS

Application of VSS in a multilayer design can be used wherever the need of Layer-2 adjacency is necessary, not just for application but for flexibility and practical use of network resources. Some of the use cases are as follows:

- Application requiring Layer -2 adjacency—Data VLANs spanning multiple access-layer switches
- Simplifying user connectivity by spanning VLANs per building or location
- Network virtualization (guest VLAN supporting transient connectivity, intra-company connectivity, merger of companies, and so on)
- Conference, media room and public access VLANs spanning multiple facilities
- Network Admission Control (NAC) VLAN (quarantine, pasteurization, and patching)
- Outsource group and inter-agency resources requiring spanned VLANs
- Wireless VLANS without centralized controller
- Network management and monitoring (SNMP, SPAN)

Virtual Switching Systems (VSS) Recommended Best Practices—Summary

Throughout this design guide, Cisco recommended best practices have been provided. The key ones are flagged as "Tips" in the sections that discuss the relevant topics. The following table lists all the key Cisco recommended best practices to make it easier for users to see them at-a-glance.

VSS Best Practice Recommendations	Topic
The recommendation is to use a unique domain ID as a best practice, even when you are not connecting multiple VSS domains together.	See "Virtual Domain" for details.
Cisco strongly recommends that you do not modify the default LMP (VSLP) timers.	See "Why Timer Should Not be Modified" for details.
It is recommended to keep the VSL link load-sharing hash method to default (adaptive) as that method is more effective in recovering flows from failed links.	See "Hashing Methods—Fixed versus Adaptive" for details.
Always bundle the numbers of links in the VSL port-channels in the power of 2 (2, 4, and 8) to optimize the traffic flow for load-sharing.	See "Hashing Methods—Fixed versus Adaptive" for details.
Cisco recommends that you do *not* configure switch preemption for the following reasons: • It causes multiple switch resets, leading to reduced forwarding capacity and unplanned network outages. • The VSS is a single logical switch/router. Both switch members are equally capable of assuming the active role because it does not matter which is active—unless required by enterprise policy.	See "Switch Preemption" for details.
The best practice is to keep the PAgP timer settings to default values and to use the normal UDLD to monitor link integrity.	See "Why You Should Keep the PAgP Hello Value Set to Default" for details.
The best practice is to keep the LACP timer settings to the default values and to use the normal UDLD to monitor link integrity.	See "Why You Should Keep the LACP Hello Value Set to Default" for details.
Cisco recommends the configuration of a virtual MAC address for VSS domain using the *switch virtual domain* command.	See "MAC Addresses" for details.
Cisco recommends that you enable and keep the default MAC OOB synchronization activity interval of 160 seconds (lowest configurable value) and idle MAC aging-timer of three times the default MAC OOB synchronization activity interval (480 seconds).	See "Out-Of-Band Synchronization Configuration Recommendation" for detail.
Cisco recommends that trunks at both end of the interfaces be configured using the desirable-desirable or auto-desirable option in a VSS-enabled design.	See "Trunking Configuration Best Practices" for details.
Cisco recommends explicit configuration of required VLANs to be forwarded over the trunk.	See "VLAN Configuration Over the Trunk" for details.
The aggressive UDLD should not be used as link-integrity check, instead use normal mode of UDLD to detect cabling faults and also for the link integrity.	See "Unidirectional Link Detection (UDLD)" for details.

VSS Best Practice Recommendations	Topic
Cisco recommends that you always use a star-shaped topology with MEC (Layer-2 and Layer-3) from each device connected to the VSS to avoid loops and have the best convergence with either link or node failures.	See Topology Considerations with VSS for details.
Cisco recommends that you do *not* enable Loop Guard in a VSS-enabled campus network.	See "Loop Guard" for details.
In the VSS-enabled network, it is critically important to keep the edge port from participating in the STP. Cisco strongly recommends enabling PortFast and BPDU Guard at the edge port.	See "PortFast and BPDU Guard" for details.
Cisco strongly recommends *not* tuning below the values listed in Table 3-5. All other NSF-related route timers should be kept at the default values and should not be changed.	See "NSF Recovery and IGP Interaction" for details.
Cisco recommends using a Layer-3, MEC-based topology to prevent multicast traffic replication over the VSL bundle and avoids delay associated with reroute of traffic over VSL link.	See "Traffic Flow with ECMP versus MEC" for details.
In Layer-2 environment, single logical link between two VSS (option 5) is the *only* topology that is recommended; any other connectivity scenario will create looped topology.	See "VSS in the Core" for details.
Cisco strongly recommends enabling dual-active detection in VSS-enabled environment.	See "Campus Recovery with VSS Dual-Active Supervisors" for details.
The best practice recommendation is to *avoid* entering into configuration mode while the VSS environment is experiencing a dual-active event; however, you cannot avoid configuration changes required for accidental shutdowns of the VSL link or the required configuration changes needed to have a proper VSL restoration.	See "VSL-Link Related Configuration Changes" for details.
The routing-protocols are recommended to run default hello and hold timers.	See "Effects of Dual-Active Condition on Convergence and User Data Traffic" for details.

CHAPTER 2

Virtual Switching System 1440 Architecture

This chapter addresses the architecture and components of Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440. Although this design guide focuses on the deployment specifics of the VSS (and not technology itself), sufficient detail is included for all the necessary components of the VSS that can affect campus design. This includes operational characteristics, design tradeoffs, and best-practice configuration recommendations. For further details about VSS technology, refer to the following document:

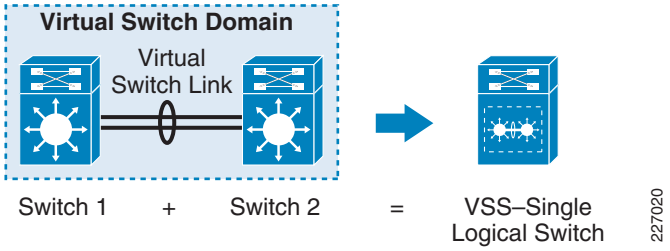
Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 White Paper on VSS technology:

http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/white_paper_c11_429338.pdf

VSS Architecture and Operation

The VSS forms two Cisco Catalyst 6500 switches into a single, logical network entity. Two chassis combine to form a single *virtual switch domain* that interacts with rest of the network as single logical switch and/or router. See Figure 2-1.

Figure 2-1 VSS Conceptual Diagram



The VSS domain consists of two supervisors—one in each member chassis connected via a *Virtual Switch Link* (VSL). A VSL facilitates the communication between two switches. Within the VSS, one chassis supervisor is designated as *active* and the other as *hot-standby.* Both use *Stateful Switch Over (SSO)* technology. The switch containing the active supervisor is called *active switch* and the switch containing hot-standby supervisor is called *hot-standby switch*. VSS operates on a unified control plane with a distributed forwarding architecture in which the active supervisor (or switch) is responsible for actively participating with the rest of the network and for managing and maintaining control plane information. The active switch maintains and updates the hot-standby supervisor with up-to-date information about the states of the system and network protocols via the VSL. If the active supervisor fails, the hot-standby supervisor assumes the active roles in managing the control plane. Both physical chassis do the data forwarding and data forwarding is done in a distributed manner. The devices adjacent

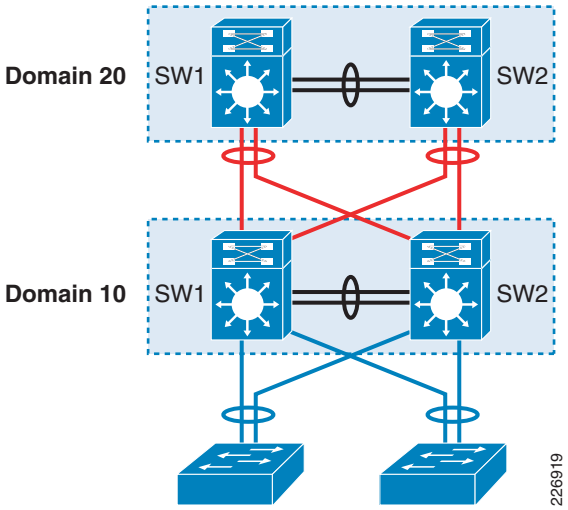
to the VSS are connected via a *Multichassis EtherChannel (MEC)* to form a single logical connection. The single logical switch, in combination with the MEC, forms the foundation of a highly available, loop-free topology. The rest of this section details the operation and best-practice configuration for components of the VSS that influence the deployment of a VSS in the campus distribution block. This design guide does not provide detailed step-by-step instructions about transitioning an existing network to a VSS-based environment, nor does it cover all preparatory work; however, it does cover essential steps that can affect the deployment of a VSS in the campus.

Virtual Switch Domain (VSD) and Switch ID

Virtual Domain

Defining the domain identifier (ID) is the first step in creating a VSS from two physical chassis. A unique domain ID identifies two switches that are intended to be part of the same VSS pair that defines the VSS domain. Assignment of a domain ID allows multiple virtual switch pairs to be connected in a hierarchical manner. Only one VSS pair can participate in a particular domain. The domain ID can have a value ranging from 1 to 255 and must be unique when multiple VSS pairs are connected together. See Figure 2-2.

Figure 2-2 VSS Domain IDs



The domain ID is defined in both physical switches as shown in the following examples.

Standalone Switch 1:

```
VSS-SW1# config t
VSS-SW1(config)# switch virtual domain 10
```

Standalone Switch 2:

```
VSS-SW2# config t
VSS-SW2(config)# switch virtual domain 10
```

The use of a domain ID is important for networks in which VSS is deployed at multiple layers in a manner as shown in Figure 2-2. This unique ID is used in many different protocol and systems configurations—such as virtual Media Access Control (MAC), Port Aggregation Protocol (PAgP), and Link Aggregate Control Protocol (LACP) control packets. If two connected VSS domains contain the same domain ID, the conflict will affect VSS operation.

The following command output example illustrates the domain ID used in the LACP system ID. The last octet—0a (hex)— is derived from the domain ID of 10 (decimal).

```
6500-VSS# sh lacp sys-id
32768,0200.0000.000a <--
```

Tip **The recommendation is to use a unique domain ID as a best practice, even when you are not connecting multiple VSS domains together.**

Switch Identifier

A VSS comprises of pair of physical switches and requires a switch ID to identify each chassis with a unique number. The switch ID can be either 1 or 2 and must be unique on each member chassis. This number is used as part of the interface naming to ensure that the interface name remains the same regardless of the virtual switch role (active or hot-standby switch). Usually the switch ID should be set in configuration mode as shown in the following examples.

Standalone Switch 1:

```
VSS-SW1# config t
VSS-SW1(config-vs-domain)# switch 1
```

Standalone Switch 2:

```
VSS-SW2# config t
VSS-SW2(config-vs-domain)# switch 2
```

However, when a hardware problem on one supervisor prompts the adoption of a new supervisor, you can set the switch ID via the command-line interface (CLI) in enable mode as shown in the following configuration examples.

Standalone Switch 1:

```
6500-VSS# switch set switch_num
```

Standalone Switch 2:

```
6500-VSS# switch set switch_num 2
```

Both methods write the switch identifier in each member chassis ROMMON. The domain ID and switch number via following can be shown via the following CLI example.

```
6500-VSS# show switch virtual
Switch mode                  : Virtual Switch
Virtual switch domain number : 10
Local switch number          : 1
Local switch operational role: Virtual Switch Active
Peer switch number           : 2
Peer switch operational role : Virtual Switch Standby
```

Caution Avoid using the command **write erase** to copy a new startup configuration. This command will erase switch numbers stored in ROMMON and any subsequent reboot will cause both switches to come up in standalone mode. Use the **switch set switch_num** *1/2* command only after both switches are rebooted because the CLI to set the switch number is *not* available in VSS mode.

Virtual Switch Link (VSL)

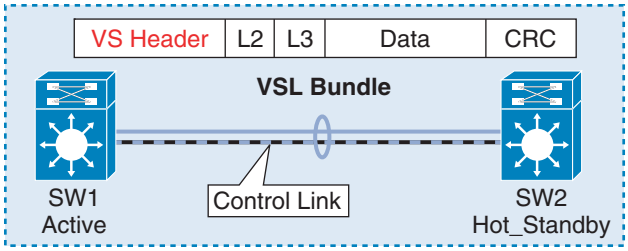
The VSS is made of two physical chassis combined to form a single logical entity. This unification of the control plane is only possible if the system controls, signaling, and backplane exist as a single entity in both chassis. This extension of the system control plane is achieved via a special purpose EtherChannel bundle. The link belonging to EtherChannel is called Virtual Switch Link (VSL). The VSL serves as logical connection that carries critical system control information such as hot-standby supervisor programming, line card status, Distributed Forwarding Card (DFC) card programming, system management, diagnostics, and more. In addition, VSL is also capable of carrying user data traffic when necessary. Thus, the VSL has a dual purpose, supporting system control synchronization and a data link.

The VSL link is treated as a systems control link and encapsulates all traffic into a special system header called the Virtual Switch Header (VSH). This encapsulation is done via dedicated hardware resources and the VSL can only be configured on a 10-Gigabit interface with following hardware ports:

- Sup720-10G, 10-Gbps ports
- WS-X6708
- WS-X6716 (in performance mode only, requires 12.2(33)SXI)

The size of the VSH is the same as that of the internal compact header used by the Cisco Catalyst 6500—it is 32 bytes long. This header is placed after the Ethernet preamble and directly before the Layer-2 header. See Figure 2-3.

Figure 2-3 VSL Header and Control Link



VSL link Initialization and Operational Characteristics

The VSS features a single control plane with a distributed forwarding architecture (see the "Stateful Switch Over Technology" section on page 2-23). Although only one supervisor manages the control plane, both switches participate in learning the control plane information. Any network and system control plane information learned via the hot-standby switch is sent to the active supervisor which in turn updates the hot-standby supervisor. This bidirectional process of learning and updating between switches is carried out over the VSL link.

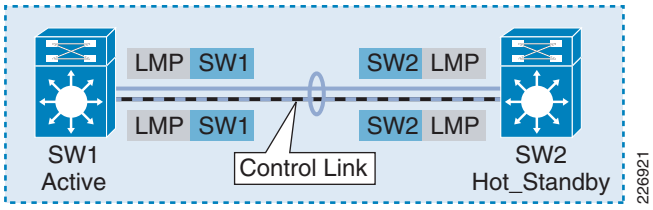
The VSL is an integral component of the VSS because it is the only path through which the two independent systems can become aware of each other during system initialization. This mutual awareness during system initialization is needed to determine the respective role of each physical chassis in becoming either the active or hot-standby virtual switch. For this reason, VSL links are brought up very early in the boot up process—before any other major service or component. To accomplish this task, each switch stores the switch number and other information necessary in ROMMON to activate the VSL link and associated line card booting sequence. During the VSL initialization, the system undergoes a variety of compatibility checks required to form a virtual switch. The VSS requires identical supervisor types and Cisco IOS software versions on both chassis. Please refer to the applicable release note for additional system requirements needed to form a virtual switch. VSL link initialization and maintenance are done through the *VSL Protocol* (VSLP) framework, which consists of two protocols: *Link Management Protocol* (LMP) and *Role Resolution Protocol* (RRP). LMP manages link integrity, while RRP determines the role of each switch member in the virtual switch domain. RRP is covered under "Stateful Switch Over Technology" section on page 2-23.

Link Management Protocol (LMP)

LMP is the first protocol to be initialized, once the VLS line card diagnostics is finished and VSL link comes on line. See Figure 2-4. The LMP is designed to perform following critical functions:

- Establishing and verifying bidirectional communications during startup and normal operation.
- Exchanging switch IDs to detect a duplicate switch ID or member that is connected to another virtual switch. This will be used by the RRP to determine the role of the switch (active or hot-standby).
- Independently transmitting and receiving LMP hello timers to monitor the health of the VSL and peer switch.

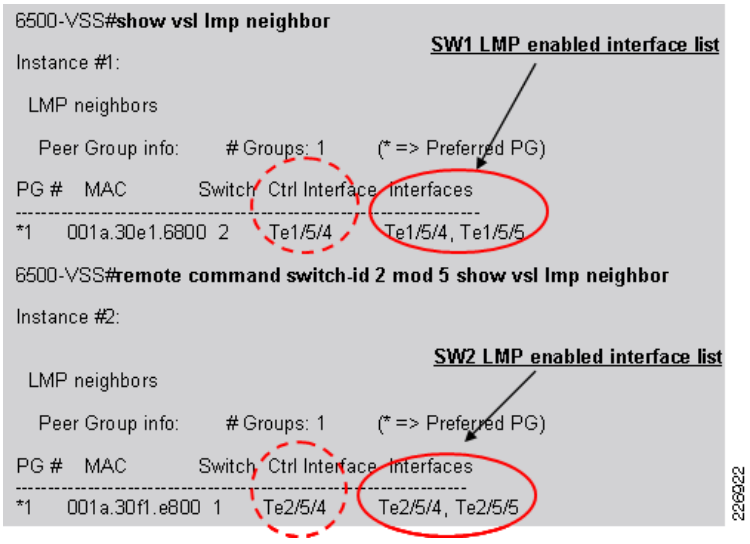
Figure 2-4 Link Management Protocol (LMP)



LMP operates independently on each member switch in the same Virtual Switch Domain (VSD). Unlike other protocols—such as PAgP, LACP and Interior Gateway Protocol (IGP) hello—that use a single control plane over which the active switch originates and terminates the protocol, both switches in the VSD independently originate and terminate LMP control-plane packets on the Switch Processor (SP). See the dotted-red circle highlighting in Figure 2-5. LMP is designed to run on each VSL member link to maintain multiple state machines with the same peer over different ports. In a case in which a unidirectional condition on a port is detected, LMP marks it as *down* and attempts to restart VSLP negotiation—instead of *err-disabling* the port. On each member switch of a VSD, LMP internally forms a single unique peer group (PG) ID with a common set of VSL links. When all VSL interfaces are down, LMP destroys the peer group and notifies RRP to take an appropriate action. The active switch will detach all the interfaces associated with the hot-standby switch. At the same time the hot-standby switch performs switchover, assumes the active role, and detaches all interfaces associated with the previously active switch.

LMP hello packets are Logical Link Control (LLC)/ Sub-network Access Protocol (SNAP)-encapsulated with the destination MAC address matching the Cisco Discovery Protocol (CDP)—01.00.0C.CC.CC.CC. All inter-chassis control plane traffic, including LMP and hello packets, are classified as bridge protocol data unit (BDPU) packets and are automatically placed in transmit priority queue.

Figure 2-5 Output Showing LMP Enabled Interface List



Control Link and Inter-Chassis Control Plane

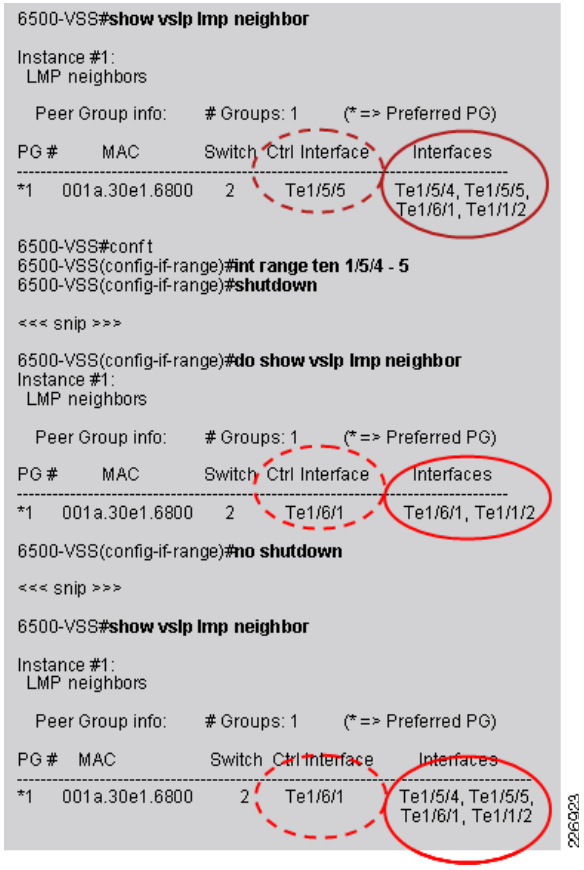
The VSL bundle is special purpose EtherChannel that can have up to eight members. Only one link out of a configured member is selected as the control link and that control link is the only link that can carry the inter-chassis control plane. The control link carries the inter-switch External Out-of-Band Channel (EOBC) control traffic that includes the Switch Control Packet (SCP) for line card communication, Inter-process Communication Packets (IPC), and Inter-Card Communication (ICC) for communicating the protocol database and state—as well as updates to the hot-standby supervisor. In addition, the same link can carry user and other network control traffic—depending how the traffic is hashed (based on source and destination MAC and/or IP addresses). The remaining bundled links carry network control plane and user data traffic, but not the inter-chassis control plane traffic (see the "Traffic Prioritization and Load-sharing with VSL" section on page 2-15). The control link is shown in Figure 2-4

The control-link selection procedure is determined by the VSS system and cannot be managed by the user. During the bootup process, the first VSL link that establishes LMP relationship (state-machine) will be selected as the control link. Based on the Cisco Catalyst 6500 architecture, the supervisor module becomes operational ahead of any other module installed in the switch. If the 10-Gbps port of Sup720-10G module is bundled in the VSL EtherChannel, then it will be selected as control-link interface whenever both switches undergo the boot process.

The **show vslp lmp neighbor** command output illustrated in Figure 2-6 depicts the current control link interface (see the dotted red circle) and list of backup VSL interfaces (which can be used if the current control-link path fails). Backup interfaces are member links of the local virtual switch's VSL EtherChannel. For a highly redundant VSL network design, the VSL EtherChannel must be bundled with multiple, VSL-capable 10-Gbps ports between the switch members. In such a case, the first listed interface under the *Interfaces* column of the **show vslp lmp neighbor** command will immediately become control link interface if the current control link interface fails. When the 10-Gbps port of a

Sup720-10G module is restored and rejoins the VSL EtherChannel, it will be placed as the next available control-link backup path without affecting the current control link interface (see second part of the output illustrated Figure 2-6 after control link is brought back up).

Figure 2-6 Control Link Interfaces Selection



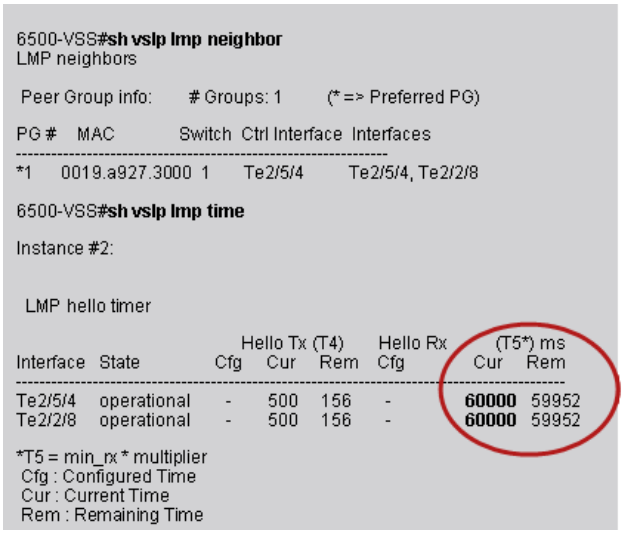
LMP Heart Beat

The LMP heart beat—also referred as the LMP hello timer—plays a key role in maintaining the integrity of VSS by checking peer switch availability and connectivity. Both VSS members execute independent, deterministic SSO switchover actions if they fail to detect the LMP hello message within configured hold-timer settings on the last bundled VSL link. The set of LMP timers are used in combination to determine the interval of the hello transmission applied to maintain the healthy status of the VSL links. The three timers are as follows:

- Hello Transmit Timer (T4)
- Minimum Receive Timer (min_rx)
- T5 Timer (min_rx * multiplier)

Figure 2-7 illustrates an example CLI output depicting the timer values per VSL link member.

Figure 2-7 Timer Values per VLS Link Member



By default, the LMP hello transmit timer (T4) and receive timer (min_rx) are assigned values of 500 msec each. The hold-timer (T5 timer) is derived from a min_rx and default multiplier of 120 (the CLI does not show the default multiplier). By default, a VSL member link time out is detected in 60,000 msec (60 seconds). The expiration of T5 indicates possible instability on the remote peer (active or hot-standby switch). Each switch member will take an independent action if the T5 timer expires. These actions include (but are not limited to) the following:

- If expiration occurs on a VSL port that is the control link, then the switch that detected the problem will force the new control link selection. It is entirely possible that T5 timer would have not yet expired on the remote peer; however, the remote peer switch will respect the request and reprogram internal logic to send control plane traffic to the newly elected VSL port as the control link port.
- If expiration occurs on a non-control link port, the switch that detected the failure selects an available port for user data traffic. Eventually, the remote peer detects a change and removes the link from the bundle.
- If expiration occurs on the last VSL port (a combination control link and user data port) and the timeout is detected on the active switch, then the active switch removes all the peer switch interfaces and announces the change to rest of the network—depending on the configuration on those interface (Layer-2 or Layer-3 protocol). Eventually, the peer switch that is in hot-standby mode will detect the T5 timer expiration and LMP will inform RRP, which forces the hot-standby switch to become the active switch. This triggers a condition known as *dual active* in which both switches declare active roles leading to instability in the network (refer to the "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18 for information about avoiding such conditions).

Why Timer Should Not be Modified

The LMP timer is used primarily for ensuring the integrity of VSS (during high CPU usage or abnormal software behavior). Normal hardware failure detection or switchover (user or system initiated) is invoked via the hardware mechanism known as *Fast Link Notification* (FLN). When an applicable event occurs, FLN informs the firmware of WS-X6708 or Sup720-10G ports to take any necessary action—typically within 50 to 100 msec. FLN is not designed to detect the failure of the remote switch. In most cases, modifying the default LMP timer to a shorter value does not improve convergence. This is because inter-chassis control plane protocol timer (IPC timer) expires before LMP timers and thus supersede the action taken.

Unintended effects can result when VSLP timers are aggressively modified. For example, modifying the VSLP timer to a lower value will add significant instability. An example scenario description follows:

> When configured with a lower VSLP timer value, the VSL will typically fail to establish neighbor relationships between member switches—leading to continuous rebooting (a *crash* in Cisco IOS parlance) of the switch. During the boot up process, the VSS switch must process multiple high-priority activities. When enabled with an aggressive VSLP timer, each member might be unable to maintain the VSLP session due to the lack of resources. When LMP hello times out on either side, LMP removes the VSL member link from VSL EtherChannel. Eventually, all links between the active and hot-standby switches can fail. For the hot-standby switch, this is considered a catastrophic error and the only way to recover is to immediately send a reset signal (crash) and start over. This can continue until at least one LMP session is established and thus can cause significant network instability.

In addition, a VSS member might also fail to send or receive LMP hello message within a configured T5 timer limit due to VSL link congestion or high CPU utilization. In such situations, the VSS system will become unstable, which can lead to a dual-active condition.

Tip **Cisco strongly recommends that you do *not* modify the default LMP (VSLP) timers.**

Role Resolution Protocol (RRP)

RRP is responsible for determining the operational status of each VSS switch member. Based on the configured parameter, a member switch can assume a role of the active, hot standby, or Route Process Redundancy (RPR). The RRP provides checks for Cisco IOS software compatibility. If the software versions are not compatible, RRP forces one of the switches into RPR mode and all line cards are powered off. The "Virtual Switch Role, Priorities and Switch Preemption" section on page 2-27 provides details of the RRP because its application is more relevant to chassis redundancy and SSO operation.

Configuring VSL Bundle

Configuring VSL is the second step in creating a virtual switch (after defining the domain ID). The VSL bundle is a special-purpose port channel. Each standalone switch is required to be configured with unique port-channel interface numbers; before assigning a port-channel number, you must make sure that the port-channel interface number is not used by an existing standalone configuration on either switch. The following configuration steps are required in each switch to convert the standalone switch to a VSS. The detailed conversion process is beyond the scope of this document. For addressing VSS conversion, refer to the *Migrate Standalone Cisco Catalyst 6500 Switch to Cisco Catalyst 6500 Virtual Switching System* document at the following URL:

http://www.cisco.com/en/US/products/ps9336/products_tech_note09186a0080a7c74c.shtml

Standalone Switch 1:

```
VSS-SW1(config)# interface Port-Channel1
VSS-SW1(config-if) #switch virtual link 1

VSS-SW1(config-if)# interface range Ten5/4 - 5
VSS-SW1(config-if)# channel-group 1 mode on
```

Standalone Switch 2:

```
VSS-SW2(config-if)# interface Port-Channel2
VSS-SW2(config-if)# switch virtual link 2

VSS-SW2(config-if)# interface range Ten5/4 - 5
```

```
VSS-SW2(config-if)# channel-group 2 mode on
```

Since VSL EtherChannel uses LMP per member link, the link-aggregation protocols, such as PAgP and LACP, are not required; each member link must be configured in unconditional EtherChannel mode using the **channel-group group-***number* **mode on** command. Once the VSL configuration is completed, using the **switch convert mode virtual** CLI command at the enable prompt will start the conversion process. The conversion process includes changing the interface naming convention from *slot/interface* to *switch_number/slot/interface*, saving the configuration, and rebooting. During switch rebooting, the systems recognize the VSL configuration and proceeds with their respective VSL ports initialization processes. The two switches communicate with each other and determine which will have active and hot-standby roles. This exchange of information is evident through the following console messages:

Standalone Switch 1 console:

```
System detected Virtual Switch
configuration...
Interface TenGigabitEthernet
1/5/4 is member of PortChannel 1
Interface TenGigabitEthernet
1/5/5 is member of PortChannel 1
<snip>
00:00:26: %VSL_BRINGUP-6-MODULE_UP: VSL module in slot 5 switch 1 brought up
Initializing as Virtual Switch active
```

Standalone Switch 2 console:

```
System detected Virtual Switch configuration...
Interface TenGigabitEthernet
2/5/4 is member of PortChannel 2
Interface TenGigabitEthernet
2/5/5 is member of PortChannel 2
<snip>
00:00:26: %VSL_BRINGUP-6-MODULE_UP: VSL module in slot 5 switch 2 brought up
Initializing as Virtual Switch standby
```

A first-time VSS conversion requires that you **must** execute the following command as the final step of accepting the virtual mode operation of the combined switches. If the switch has been converted, or partially converted, you cannot use this command.

```
6500-VSS# switch accept mode virtual
```

The preceding command forces the integration of all VSL-link related configurations from the hot-standby switch and populates the running configuration with those commands. In addition, the startup configurations are updated with the new merged configurations. The following prompt appears:

```
Do you want proceed? [yes/no]: yes
Merging the standby VSL configuration. . .
Building configuration...
[OK]
```

Note Only VSL-related configurations are merged with the conversion step. All other configurations must be managed per your network site implementation requirements. Follow the related Cisco product documentation for further details.

VSL Characteristics

The VSL port channel is treated as an internal systems link. As a result, its configuration, resiliency, mode of operation, quality of service (QoS), and traffic load sharing follow a set of rules that are specific to VSL. This section covers configuration requirements relating to those rules. The logical port channel and its member link both have distinct sets of restrictions.

VSL port-channel logical interface configuration is restricted to VSL-related configuration; all other Cisco IOS features are disabled. The following output illustrates available options:

```
6500-VSS(config)# int po 1
6500-VSS(config-if)# ?
virtual link interface commands (restricted):
  default        Set a command to its defaults
  description    Interface specific description
  exit           Exit from virtual link interface configuration mode
  load-interval  Specify interval for load calculation for an interface
  logging        Configure logging for interface
  mls            mls sub/interface commands
  no             Negate a command or set its defaults
  port-channel   Port Channel interface subcommands
  shutdown       Shutdown the selected interface
  switch         Configure switch link
  vslp           VSLP interface configuration commands
```

VSL member links are in restricted configuration mode once the VSL configuration is applied. All Cisco IOS configuration options are disabled except following:

- EtherChannel
- Netflow configuration
- Default QoS configuration

The following output illustrates available options:

```
6500-VSS(config)# int ten 1/5/4
6500-VSS(config-if)# ?
virtual link interface commands (restricted):
  channel-group   Etherchannel/port bundling configuration
  default         Set a command to its defaults
  description     Interface specific description
  exit            Exit from virtual link interface configuration mode
  load-interval   Specify interval for load calculation for an interface
  logging         Configure logging for interface
  mls             mls sub/interface commands
  no              Negate a command or set its defaults
  priority-queue  Configure priority scheduling
  rcv-queue       Configure receive queue(s)
  shutdown        Shutdown the selected interface
  wrr-queue       Configure weighted round-robin xmt queues
```

When configuring VSL interfaces, only one VSL EtherChannel configuration per virtual-switch is possible. Configuring an additional VSL EtherChannel will result in an error. The following are examples of error messages generated:

```
6500-VSS(config)# interface port-channel 3
6500-VSS(config-if)# switch virtual link 1
% Can not configure this as switch 1 VSL Portchannel since it already had VSL Portchannel
1 configured
6500-VSS(config-if)#
6500-VSS(config-if)# switch virtual link 2
% Can not configure this as switch 2 VSL Portchannel since it already had VSL Portchannel
2 configured
```

In addition, for post VSS conversion, a neighbor switch port cannot be bundled into the local switch's VSL EtherChannel. In the example used here, the Switch 1 VSL EtherChannel cannot bundle a physical port from Switch 2. However, a regular EtherChannel does not have such a restriction. The following example output illustrates the error message generated when attempting to configure this unsupported mode:

```
6500-VSS(config-if)# int te2/1/1
6500-VSS(config-if)# channel-group 1 mode on
VSL bundle across chassis not allowed TenGigabitEthernet2/5/5 is not added to port channel
1
```

The remainder of this section covers design considerations for prioritizing traffic over VSL links, resiliency, and traffic load-sharing options available with VSL.

VSL QoS and Prioritization of Traffic

Today's enterprise application requirements are diverse. Many time-critical services (including voice, video and multicast)—as well as enterprise applications, such as customer relationship management (CRM), SAP, and Oracle—require specific priorities in the campus network. The QoS design guide (available at the URL below) describes an approach for classifying user traffic at the edge and using those priorities intelligently via the Differentiated Services Code Point (DSCP) trust model. This design guide does not cover generic QoS requirements and behavior in VSS; however, this section does address QoS as it applies to a VSL link in terms of default behavior, how the control plane is protected, and implementation options that network designers should consider.

http://www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND_40/QoSCampus_40.html

Hardware Configuration Dependency

Due to the ease of configuration, flexibility, and bandwidth requirement, the distribution layer traditionally leverages line cards for connectivity to the core and access layer, thus the use of supervisor ports is limited. The access layer uses the uplink from either a supervisor or a dedicated port on non-modular switches. The Sup720-10G supervisor offers a new option of using the 10-Gbps port available on the supervisor to provide uplink connectivity to the core. This requires an understanding of the current design choices with the Sup720-10G uplink port with respect to QoS when used as the VSL port and/or uplink to the rest of the network. The VSL link can only be configured on 10-Gbps ports. Choosing the VSL link configuration on either a supervisor port or line card affects the QoS capability on the unused ports of the supervisor. The Sup720-10G supervisor has two 10-Gbps ports and three 1-Gbps ports. The Sup720-10G uplink ports can be configured in one of two modes:

- *Default—Non-10 gigabit-only mode*

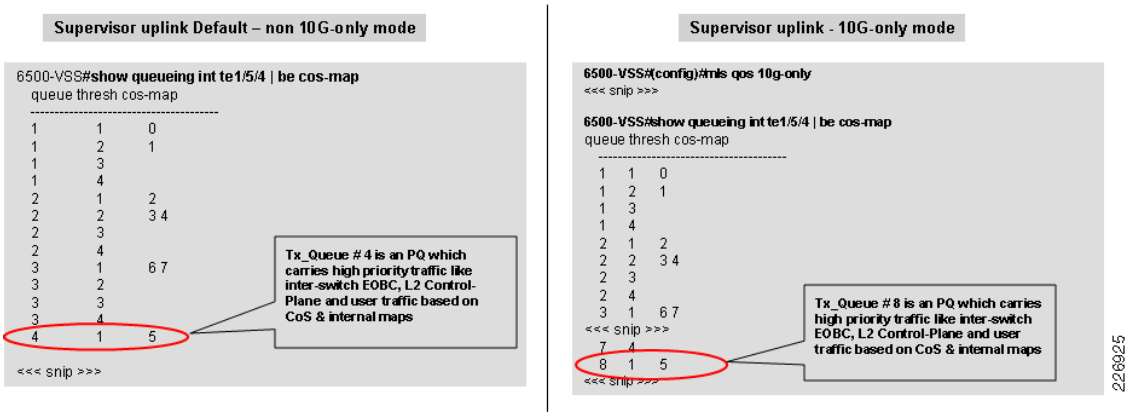
  In this mode, all ports must follow a single queuing mode. If any 10-Gbps port is used for the VSL link, the remaining ports (10 Gbps or 1Gbps) follow the same CoS-mode of queuing for any other non-VSL connectivity because VSL only allows class of service (CoS)-based queuing.

- *Non-blocking—10 gigabit-only mode*

  In this mode all 1-Gbps ports are disabled, as the entire module operates in a non-blocking mode. Even if only one 10G port used as VSL link, still both 10-Gbps port is restricted to CoS-based trust model. 12.2(33)SXI removes this limitation by allowing the unused (non-VSL configured)10-Gbps port be configured based on user preference, including DSCP-based queuing.

Figure 2-8 shows that a 10-Gbps-only mode increases transmit queue from default *1p3q4t* to the *1p7q4t* setting. It also increases the receive queue size from *2p4t* to *8q4t* similar to the WS-X6708 module. However, the default Tx and Rx queue mapping remains CoS-based with or without 10-Gbps-only mode. As a result, there is no real benefit to using an improved queue structure to map each class of traffic in separate queues because the default COS-to-queue mapping cannot be changed.

Figure 2-8 Comparison of Default and Non-Blocking Modes



If one of the WS-X6708 line card port is used as the VSL link, port queuing for that port is limited to being CoS-based; however, the remaining ports can have independent sets of QoS configuration.

The resilient VSL design uses these facts as a design factor. Table 2-1 summarizes the options and restrictions for Sup720-10G uplink ports. The "Resilient VSL Design Consideration" section on page 2-18 describes incorporating these design factors when developing highly resilient and flexible VSL configurations.

Table 2-1 Options and Restrictions for Sup720-10G Uplink Ports

Sup720-10G Uplink Port	10g-only mode	Non 10g-only mode
Queue Structure	Tx– 1p7q4t Rx – 8q4t	Tx –1p3q4t Rx – 2q4t

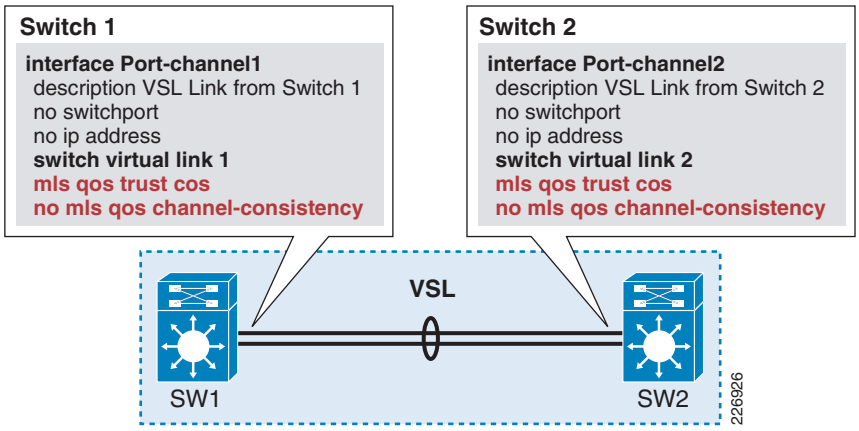
Table 2-1 Options and Restrictions for Sup720-10G Uplink Ports (continued)

Sup720-10G Uplink Port	10g-only mode	Non 10g-only mode
Non-VSS mode (standalone mode) Or VSS mode (no supervisor port used as VSL link)	Only 10-Gbps ports available, all 1-Gbps ports are disabled. Both 10-Gbps ports can have independent sets of QoS configuration including trust model and queuing mode. The DSCP-based queuing allowed.	All ports are available. All uplinks can *only* use a single QoS configuration (maps, thresholds, and queues). Default queuing mode is CoS only. No DSCP based queuing allowed.
10-Gbps ports used for VSL link	If both 10-Gbps ports are used as the VSL link, only CoS-based trust mode supported. Even if only a single 10-Gbps port is used as the VSL link, both 10-Gbps ports are restricted to the CoS-based trust model in Cisco IOS 12.2(33) SXH. Cisco IOS 12.2(33) SXI removes this limitation by allowing the remaining 10-Gbps port to be configured based on user preference, including DSCP-based queuing. The remaining 1-Gbps ports are disabled.	Both 10-Gbps ports are used or a single 10-Gbps port is used as the VSL uplink; you can only define one QoS configuration. Because VSL only allows CoS-based queuing, the remaining non-VSL ports follow the COS-based queuing for any non-VSL connectivity.

Default QOS Setting for VSL

The VSL (by default) uses a CoS-based trust model. This default CoS-trust setting on VSL EtherChannel cannot be modified or removed. Regardless of the QoS configuration in global mode, the VSL EtherChannel is set to CoS-based queuing mode. Figure 2-9 shows the QoS configuration of VSL link on SW1 and SW2. As with a standalone switch, the VSS uses the internal CoS-based queuing with various mapping tables for placing user traffic into the egress queue. Traffic traversing the VSL link uses the same facilities with a CoS-based trust model. The Weighted Round Robin (WRR) queuing scheduler is enabled by default on each VSL member link. The VSL link's QoS configuration does not alter the QoS marking set in the user traffic when that traffic traverses the VSL link.

Figure 2-9 VSL CoS Configuration Example Comparison



The best-practice recommendation for VSL link resiliency is to bundle two 10-Gbps ports from different sources. Doing this might require having one port from the supervisor and other from a Cisco 6708 line card. By default, the Sup720-10G supervisor 10-Gbps port has a non-Distributed Forwarding Card (DFC) 1p3q4t Tx-queue structure in contrast with the WS-X6708 linecard which has a DFC-based 1p7q4t Tx-queue structure. For the conventional configuration of EtherChannel, a bundle between non-congruent queue structures would fail. However, the VSL bundle is by default enabled with the configuration which allows the EtherChannel to be formed. This is enabled via the **no mls qos channel-consistency** command.

The following QoS configuration restrictions apply once the port is bundled in the VSL EtherChannel:

- The CoS-mode setting cannot be changed or removed from VSL port-channel.
- Any QoS configuration, such as DSCP-based trust, receive-queue bandwidth, or threshold limit, cannot be modified on any of the 1-Gbps or 10-Gbps ports of the Sup720-10GE module after bundling the link to the VSL port-channel. Applying such restricted QoS configurations will result in an error.
- User-defined Modular QoS CLI (MQC) service-policies cannot be attached to the VSL EtherChannel.
- Bundling of the 10-Gbps port in the VSL port-channel will fail if unsupported QoS capabilities are preconfigured. All QoS configuration must be removed prior bundling the port into the VSL EtherChannel.

Note The WS-X6708 module supports independent QoS policies on non-VSL configured ports, even if one of its 10-Gbps ports is bundled in the VSL EtherChannel.

Traffic Prioritization and Load-sharing with VSL

This section addresses the following topics:

Prioritizing Specific Traffic Types

The VSL link can potentially carry three types of traffic and uses QoS mapping to distinguish between each. Traffic types carried over a VSL link include the following:

- User data traffic
- Network control plan traffic
- VSS inter-switch communication and Layer-2 per link control traffic

These are described briefly in the sections that follow.

User Data Traffic

In a recommended best-practice configuration implementation, all devices are attached to the VSS via MEC-based connections (see the "MEC Configuration" section on page 2-43). In dual-homed MEC connectivity, pass-through user data traffic does not traverse the VSL link. However, during certain conditions user data must traverse VSL link; applicable conditions include (but are not limited to) the following:

- Failure of uplink from access layer to VSS causing downstream traffic to flow over the VSL link
- Remote Switched Port Analyzer (SPAN) traffic from one VSS switch member to the other
- Service-module traffic flow from FWSM, Wireless Services Module (WiSM), Intrusion Detection System (IDS), and other modules

VSL itself does not alter the QoS marking of user data traffic. It simply categorizes the preceding types of traffic using the CoS-based queuing. As a result, any ingress traffic QoS marking that is not based on CoS must use the internal QoS mapping table to provide the translation to CoS-based queuing. Any end-user application traffic marked with 802.1p CoS 5, DSCP 46, or IPP 5 will be placed in the priority-queue.

Network Control Plane Traffic

The active switch always originates and terminates network control plane traffic from participating adjacent devices. The active switch always uses a locally attached interface (link) to forward the control plane traffic. The network control plane traffic that must traverse the VSL due to either a failure of local links connected to active switch or traffic sourced by a hot-standby switch, including the following:

- *Layer-3 protocol traffic for Layer-3 Equal Cost Multipath (ECMP) links on the hot-standby switch*—Routing protocol control traffic, such as hello, update, database, and so on
- *Traffic intended for the VSS supervisor*—Internet Control Message Protocol (ICMP) responses from other devices, time-to-live (TTL) with value of 1 in increments of hop counts (it must terminate at the active switch), SNMP, Telnet/SSH, and so on.

VSS Inter-Switch Communication and Layer-2 per Link Control Traffic

All communications among VSS member switches are defined as inter-switch traffic. VSS systems automatically classify the following inter-switch control plane protocols as Bridge Protocol Data Unit (BPDU)-type traffic:

- Inter-Switch Communication
  - *Inter-Chassis Ethernet Out Band Channel (EOBC) traffic*— Serial Communication Protocol (SCP), IPC, and ICC
  - *Virtual Switch Link Protocol (VSLP)* —LMP and RRP control-link packets
- *Any Layer-2 per link protocol*—Spanning Tree Protocol (STP) BPDU, Port Aggregation Protocol (PagP)+, LACP, CDP and Unidirectional Link Detection (UDLD), Link Layer Discovery Protocol (LLDP), Root Link Query (RLQ), Ethernet Operations, Administration, and Maintenance (OAM), 802.1x, Dynamic Trunking Protocol (DTP), and so on.

These BPDU packets are automatically placed in transmit priority-queue ahead of any other traffic

Note Network control plane traffic (Layer 2 and Layer 3) is always sent out links connected to the active switch. It will only cross over the VSL either due to a local link being unavailable or the protocol frame needing to be originated from the hot-standby port (e.g. PAgP, LACP or UDLD).

Note Priority-queues are shared by the high-priority user data traffic (marked as expedited forwarding -EF) along with control plane traffic. However, an internal mechanism always ensures that control plane traffic takes precedence over of any other priority-queued traffic. This ensures that user data does not inadvertently affect the operational stability of the entire VSS domain.

Load-Sharing with VSL

As described in the preceding section, the VSL carries multiple types of the traffic over the VSL bundle. From the traffic load-sharing perspective, the VSL bundle is just like any other EtherChannel. It follows the same rules of EtherChannel hashing algorithm available in any given Cisco IOS software. In standalone or virtual-switch mode, a single EtherChannel load-balance hashing is applicable system wide. This means the VSL bundle uses the same configured load-sharing mode that applies to all EtherChannel groups in the VSS. The following output example illustrates load-balancing options:

```
6500-VSS(config)# port-channel load-balance ?
  dst-ipDst IP Addr
  dst-mac                 Dst Mac Addr
  dst-mixed-ip-port       Dst IP Addr and TCP/UDP Port
  dst-port                Dst TCP/UDP Port
  mpls                    Load Balancing for MPLS packets
  src-dst-ip              Src XOR Dst IP Addr
  src-dst-mac             Src XOR Dst Mac Addr
  src-dst-mixed-ip-port   Src XOR Dst IP Addr and TCP/UDP Port
  src-dst-port            Src XOR Dst TCP/UDP Port
  src-ip                  Src IP Addr
  src-mac                 Src Mac Addr
  src-mixed-ip-port       Src IP Addr and TCP/UDP Port
  src-port                Src TCP/UDP Port
```

Note This section only covers the characteristics of EtherChannel related to VSL. For generic EtherChannel design and recommendation refer to the "MEC Configuration" section on page 2-43.

The load-balancing method implemented applies to all network control traffic and user data traffic. The only exceptions to this rule are inter-switch communication traffic (always carried by control link) and LMP hello (sent on every VSL link). The network control plane and user data traffic use source and/or destination MAC addresses and/or the IP address as input to the hash calculation that is used to load-share the traffic. The network control plane traffic that can be load-shared between VSL links includes, but is not limited to, the following (note the difference in QoS categorization of the traffic):

- *Layer-2 protocol traffic*—Broadcast, STP BPDU, PAgP+, LACP, CDP and UDLD , LLDP, RLQ , Ethernet OAM, 802.1x, and DTP
- *Layer-3 protocol traffic for Layer-3 ECMP links on the hot standby switch*—Routing protocol control traffic (such as Hello, Update, and database) and ICMP response
- *Traffic designated to VSS supervisor*—ICMP response from other devices, TTL with 1, and so on

The type of user data traffic crossing VSL links are described in the "Prioritizing Specific Traffic Types" section on page 2-15 section.

Hashing Methods—Fixed versus Adaptive

Traffic across port-channel is distributed based on Result Based Hash (RBH) computation for each port-channel member link. Whenever a port-channel member link is added or removed from a group, RBH must be recomputed for every link in the group. For a short period of time, each flow will be rehashed—causing disruption for the traffic. This hash implementation is called *fixed*.

As of Cisco IOS Release 12.2(33) SXH, Cisco Catalyst 6500 supports the enhanced hash algorithm that pre-computes the hash for each port-channel member link. When the link member fails, dynamic pre-computation of hashing allows new flows to be added to the existing link and also reduces the loss of packets for the flows that were already hashed to the link. This enhanced hash implementation is called *adaptive*. The following example illustrates the hash-distribution options:

```
6500-VSS(config-if)# port-channel port hash-distribution ?
  adaptive  selective distribution of the bndl_hash among port-channel members
  fixed     fixed distribution of the bndl_hash among port-channel members

VSS(config-if)# port-channel port hash-distribution fixed
This command will take effect upon a member link UP/DOWN/ADDITION/DELETION event.
Please do a shut/no shut to take immediate effect
```

By default, the load-sharing hashing method on all non-VSL EtherChannel is fixed. In contrast, the default hash algorithm for the VSL bundle is adaptive because it carries critical inter-switch control plane traffic. The default hashing method can be changed, but the only way to make the hash algorithm effective is to reset the link. Applying this to the VSL will trigger the dual-active condition because both chassis will lose connection with each other when the VSL links bounce (see the "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18).

Tip **It is recommended to keep the VSL link load-sharing hash method to default (adaptive) as that method is more effective in recovering flows from failed links.**

The current EtherChannel hardware can only load-share with three unique binary buckets, thus any combination of links in the EtherChannel bundle that can fill the all the buckets would optimally use all the links in the bundle. This translates into a number of links in the bundle with a formula of the power of 2 for optimal load sharing.

Tip **Always bundle the numbers of links in the VSL port-channels in the power of 2 (2, 4, and 8) to optimize the traffic flow for load-sharing.**

Resilient VSL Design Consideration

The VSL can be configured as single member EtherChannel. Configuration of a resilient VSL link follows the same design principles that apply to deploying a resilient EtherChannel-connected device. Resilient EtherChannel design consists of avoiding any single point-of-failure in terms of line cards or ports. Redundancy of VSL is important so as to avoid the dual-active condition and instability of the VSS. VSL redundancy is useful whether or not the supervisor fails. In the case of an active supervisor failure, the hot-standby switch (supervisor) is ready to take over—VSL resiliency notwithstanding. VSL resiliency *is* important when the supervisor has *not* failed and somehow the systems have lost their VSS links—leading to a dual-active condition.

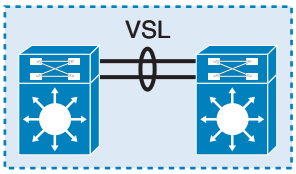
The following key factors should be considered when designing resilient VSL:

- Use port-channels with more than one member in a bundle to reduce the number of potential single points of failure (ports, line cards)

- Use redundant hardware (ports, line cards, and internal resources connecting the port)
- Use diverse fiber paths (separate conduits, fiber terminations, and physical paths) for each VSL links.
- Manage traffic forwarded over the VSL link to avoid single homed devices. This is covered under the "Traffic Flow in the VSS-Enabled Campus" section on page 3-5.
- Since the VSL can only be configured on 10-Gbps port, choices for deploying the VSL bundle are limited to the Sup720-10G, WS-X6708, or WS-X6716 hardware. Besides redundancy, capacity planning also influences number of VSL members per VSL bundle. The capacity planning is explained under the "Capacity Planning for the VSL Bundle" section on page 3-12. There are three design options for avoiding a single point-of-failure:
  - *Use two 10-Gbps ports available with Sup720-10G supervisor*

    Design option 1 (Figure 2-10) is the most common and most intuitive choice. It uses both 10-Gbps ports on the supervisor. However, this option does not provide optimal hardware diversity as both ports are connected to single internal fabric connection. The probability of having both port connections to the fabric backplane having hardware errors is low, but not impossible.

Figure 2-10 VSL over Supervisor Port



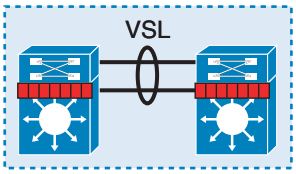
Design 1 – VSL Bundle over Supervisor Port

226927

  - *Use one 10-Gbps port from the Sup720-10G supervisor and another from a VSL capable line card (WS-X6708 or WS-X6716)*

    Design option 2 (Figure 2-11) diversifies the VSL links onto two separate hardware line cards—one port from the Sup720-10G uplink and another from the WS-X6708 line card. This is the best baseline and most practical option for balancing cost and redundancy. This design restricts unused ports on Sup720-10G with CoS-based queuing. Cisco IOS 12.2(33) SXI removes this restriction.

Figure 2-11 Distributed VSL Bundle between Bundle and 67xx Line Card



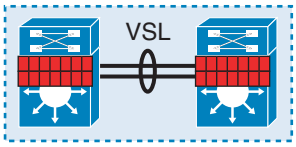
Design 2 – Distributed VSL Bundle between Supervisor and 67xx Line Card

226928

  - Use both 10-Gbps ports from VSL capable line cards (WS-X6708 or WS-X6716)

    Design option 3 (Figure 2-12) uses separate line cards for VSL connectivity. This provides the best flexibility in term of QoS configurations for uplink ports on supervisors, but is not as cost effective as design option 2.

Figure 2-12 Distributed VSL Bundle between Dual 67xx Cards



Design 3 – Distributed VSL Bundle between Dual 67xx Line Cards

226929

Besides avoiding single point-of-failure, the right design selection depends on the availability of the hardware and usage of the uplink ports on the supervisor (see the "VSL QoS and Prioritization of Traffic" section on page 2-12). If one 10-Gbps uplink port is used as the VSL link, then the other 10-Gbps port can only have CoS-based queuing. The Cisco IOS software as of Cisco IOS 12.2(33) SXI removes the CoS-based queuing restriction and allows the other non-VSL 10-Gbps port to be configured for DSCP-based queuing. If any of the Sup720-10G ports are used for connectivity to the core or other network layer in which QoS requirements require flexibility, then design options 2 and 3 are the available choices.

For superior resiliency, having one VSL link on the supervisor and other on a line card is the best option. This design option avoids possible catastrophic disconnection of line cards and allows more ports to be added in the VSL bundle for future growth. Table 2-2 details the design choices for the VSL links. You should be aware that there is no clear choice that emerges from the following options for all implementations; however,Table 2-2 does provide some risk categorization for available choices.

Table 2-2 Design Choices with VSL Links

	Design 1-1 Default - Non 10g-only	Design 1-2 10g-only (mls qos 10g-only)	Design 2-1 Default - Non 10g-only	Design 2-2 10g-only (mls qos 10g-only)	Design – 3
Hardware Configuration	VSL link both on Sup 720-10G uplinks. All uplink ports are available.	VSL link both on Sup 720-10G uplink. Only 10-Gbps ports are available.	One port on Sup 720 and other on 10-Gbps line card. All uplink ports are available.	One port on Sup 720 and other on line card. Only 10-Gbps ports are available.	Both VSL link on different 10-Gbps line cards.
Port Usage and Performance Mode	All uplink ports are available. 10-Gbps and one Gbps are sharing the total 20 Gbps of bandwidth.	Only 10-Gbps ports are available. 10-Gbps ports are non-blocking.	All uplink ports are available. 10-Gbps and one Gbps are sharing the total 20 Gbps of bandwidth.	Only 10-Gbps ports are available. 10-Gbps ports are non-blocking.	All sup uplink ports are available. WS-X6708–2:1 oversubscription.
QoS[1] Configuration Flexibility	Least – all ports can only have CoS based queuing.	CoS-based only since all 10-Gbps ports are used as VSL link. However loses three 1-Gbps ports.	Limited though practical. All uplink ports can only have CoS-based queuing. 10-Gbps line card can have independent QoS.	More Limited then option 2-1. The reminder 10-Gbps port follows the same QoS configuration as VLS port. 12.2(33) SXI removes this restriction. 10-Gbps line card can have independent QoS.	Most Flexible. All sup uplink can be used. All ports on 10 Gbps can have independent QoS in 10g-only mode.

Table 2-2 Design Choices with VSL Links (continued)

VSL Single Point-of-Failure	Possible, due to failure of fabric interconnects. Risk of failure – very low.	Possible, due to failure of fabric interconnects. Risk of failure –very low.	Possibility of losing both ports on distinct hardware is rare.	Possibility of losing both ports on distinct hardware is remote. Risk of failure—very rare.	Possible, due to loss of connectivity to line card in extreme conditions. Risk–Extremely low.
VSS Boot Behavior[2]	Optimal	Optimal	Optimal	Optimal	Delayed
Overall Choice Criteria	Cost effective, efficient, least flexible for QoS configurations on one Gbps uplink. Not recommended due to least diversity of hardware.	Cost effective, performance guarantee. Not recommended due to least diversity of hardware.	Practical. Future enhancement makes it best overall from cost, efficiency and link redundancy. Though reduced QoS flexibility.	Practical. Future enhancement makes it best overall from cost, efficiency and link redundancy.	Most efficient port usage and flexible QoS. Not as cost effective and optimized as design option 1-1 and 1-2.

1. Queue structure and depth is not a factor in VSL since the mapping and ration remains the same. Refer to "VSL QoS and Prioritization of Traffic" section on page 2-12.
2. The VSL link on Sup720-10G ports will come up faster than VSL on line cards since line cards need more time for image download, initialization etc. VSS is optimized for faster booting with VSL link on supervisor port.

VSL Operational Monitoring

This design guide does not cover operational monitoring and troubleshooting of VSS; however, critical information is included to emphasize the need to manage bandwidth utilization—and the health of the VSL port-channel and its member links. Relevant CLI output examples are shown throughout this section.

Troubleshooting of the VSS and VSL port-channel interface might require the port-channel to be spanned to the port to which the network packet decoder is attached. The VSL port-channel can be spanned. However, you can only span a local port-channel to a local destination. Refer to the following CLI command output.

```
6500-VSS# show inteface vsl

VSL Port-channel: Po1
 Port: Te1/5/4
 Port: Te1/5/5
VSL Port-channel: Po2
 Port: Te2/5/4
 Port: Te2/5/5

6500-VSS(config)# monitor session 2 source int po1
6500-VSS(config)# monitor session 2 destination int gi1/4/10

6500-VSS# show monitor session 2
```

```
Session 2
---------
Type                   : Local Session
Source Ports           :
    Both               : Po1
Destination Ports      : Gi1/4/10

Egress SPAN Replication State:
Operational mode       : Centralized
Configured mode        : Centralized (default)
```

As shown in the preceding output, you can monitor the VSL port-channel by spanning it to a switch where that port-channel is local. See the following output example illustrating an attempt to create port monitoring with a destination belonging to the peer (remote) switch. This restriction removes the possibility of looping traffic and avoids over-utilization of VSL links. The port-channel interface numbers are usually created with matching switch number IDs and thus you can easily identify the affinity of a port-channel interface with a switch number.

```
6500-VSS# show interface vsl
VSL Port-channel: Po1
 Port: Te1/5/4
 Port: Te1/5/5

VSL Port-channel: Po2
 Port: Te2/5/4
 Port: Te2/5/5

6500-VSS(config)# monitor sess 1 source int po2
6500-VSS(config)# monitor sess 1 destination int gi1/4/10
% VSL cannot be monitor source with destination on different core
```

Note that the *Giants* counters on the VSL interface (see the output that follows) might lead to an assumption that something is wrong. In fact, this is a normal output. The reason that the interface counters notice the giants is due to fact that VSL inter-switch control frame packets are sent at 1518 bytes + 32 byte of DBUS header between the active and hot-standby switches. Such oversized packets are seen as giants on VSL EtherChannel.

```
6500-VSS# show switch virtual link counters
.
.
! <snip>

Port          Single-Col Multi-Col  Late-Col Excess-Col Carri-Sen     Runts      Giants
Po1                    0         0         0          0         0         0    19788377
Te1/2/8                0         0         0          0         0         0          34
Te1/5/4                0         0         0          0         0         0    19788414
<snip>

Port          Single-Col Multi-Col  Late-Col Excess-Col Carri-Sen     Runts      Giants
Po2                    0         0         0          0         0         0      693910
Te2/2/8                0         0         0          0         0         0          89
Te2/5/4                0         0         0          0         0         0      693821
```

Stateful Switch Over—Unified Control Plane and Distributed Data Forwarding

Stateful Switch Over Technology

The Stateful Switch Over (SSO) technology enables supervisor redundancy in a standalone Cisco Catalyst 6000 Series platform. SSO keeps the necessary control plane and protocol states replicated to the backup supervisor. As a result, if an active supervisor fails, a hot-standby supervisor has enough information about that system and network to continue forwarding packets and to continue in network protocol participation with the rest of the network devices. The dual supervisor-enabled system goes through various states during power-up. During initialization, Cisco IOS determines whether the system has dual supervisors, determines the hardware mode—*simplex* (single supervisor) or *duplex* (dual supervisor), and identifies which supervisor assumes the active or hot-standby role. Cisco IOS software also checks the software version on each supervisor and determines whether the state of the supervisor can enter into SSO or RPR mode. Each supervisor follows the redundancy facility (RF) states described in Table 2-3, depending on the adopted role (active or hot standby). For the supervisor that is elected as *primary,* the supervisor transitions from lowest (disabled) to highest (active) mode after successful SSO startup. The hot-standby supervisor goes through a separate state transition as described in Table 2-3 under the heading *States when becoming Standby-hot.*

Table 2-3 RF States

RF States and Code	RF State Activity
Common States to both supervisor	
RF_UNKNOWN = 0,	Unknown redundancy state; for example, supervisor booting
RF_DISABLED = 1,	Redundancy is disabled; for example, no dual supervisor exists
RF_INITIALIZATION = 2,	First phase of sync between supervisors
RF_NEGOTIATION = 3,	Discovery mode and who becomes active or hot-standby
States when becoming Standby-hot	
RF_STANDBY_COLD = 4,	State on non-active supervisor, peer is active, RPR state
RF_STANDBY_CONFIG = 5,	Sync config from active to hot-standby
RF_STANDBY_FILESYS = 6,	Sync file system from active to hot-standby
RF_STANDBY_BULK = 7,	Protocols (client) state—bulk sync from active to hot-standby
RF_STANDBY_HOT = 8,	Standby ready to be active and getting updates from active
States when becoming ACTIVE	
RF_ACTIVE_FAST = 9,	Immediate notification of hot-standby going active
RF_ACTIVE_DRAIN = 10,	Client clean up—drain queued messages from peer
RF_ACTIVE_PRECONFIG = 11,	Pre-processing configuration, boot environment
RF_ACTIVE_POSTCONFIG = 12	Post-processing the configuration
RF_ACTIVE = 13,	Control and data plane active and participating with network

Among these thirteen states, *13-Active* and *8-Standby-Hot* are critical for determining operational redundancy. These are summarized in the following brief descriptions:

- *State 13-ACTIVE*—In this *active* state, the supervisor is responsible for packet forwarding and managing the control plane. The control plane functions includes handling Layer-3 routing protocol, Layer-2 protocols (STP, BPDU), management—such as Telnet, Simple Network Management Protocol (SNMP), and secure shell (SSH), link and device management (SPAN and CDP), and so on. The active supervisor synchronizes configuration with the secondary supervisor. Finally, the active supervisor synchronizes the state and database of the protocols to the secondary supervisor once the hot-standby supervisor assumes the state of *Standby-HOT* (hot standby).
- *State 8-Standby-Hot*—In this hot-standby state, the supervisor is fully synchronized with the active supervisor and is capable of assuming the active role when needed. This is the final state of the hot-standby supervisor. In this state, each SSO-aware protocol, based on relevant events (such as interface state change, MAC update/change/up/down, and so on), triggers a message from the active supervisor to the hot-standby supervisor. Whenever the primary active supervisor fails for some reason, the protocol state on the hot-standby supervisor goes into the execution (run) state. For example, Cisco Express Forwarding (CEF) is a SSO-aware client. Whenever a change in the CEF's table occurs, the hot-standby supervisor receives an update. This ensures that when the hot-standby unit becomes active, the updated copy of the forwarding information base (FIB) can forward data packet in the hardware, while the control plane undergoes the recovery process. For more information on SSO, refer to following URL:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/nsfsso.html

SSO Operation in VSS

The SSO is the core capability that enables VSS high availability. SSO operational and functional support is similar to standalone node operating with a dual supervisor. The two major differences are as follows:

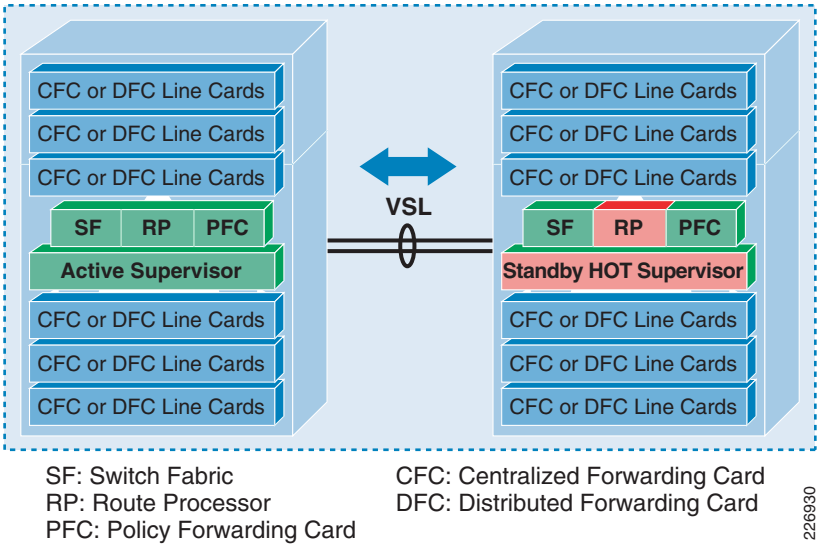
- SSO operation is extended over two chassis, where one supervisor is elected as the active supervisor and the supervisor in the other chassis is designated as the hot standby. This function is defined as *inter-chassis SSO*. See Figure 2-13.
- The packet forwarding occurs on both chassis and supervisors, hence the VSS is a *dual forwarding* solution, although the control plane is managed by only one supervisor.

The SSO operation in the VSS has the same dependency as in the case of a standalone environment. The inter-chassis SSO mode requires identical hardware and Cisco IOS software in both member chassis. Please refer to the URL below for the detailed list of dependencies for the SSO redundancy mode:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/vss.html#wp1059586

Note Cisco IOS 122.(33)SXI removes some of the Cisco IOS software versioning dependencies. Refer to the appropriate release note for more information.

Figure 2-13 Inter-Chassis SSO Operation



Unified Control Plane

As shown in Figure 2-13, one supervisor is actively providing the unified control plane. The chassis carrying the active supervisor is called the *active virtual switch*. In the active switch all three components are active (green)—Switch Fabric (SF), Route Processor (RP), and Policy Forwarding Card (PFC). The active supervisor is also responsible for programming the hardware forwarding information onto all the distributed forwarding cards (DFC) across the entire VSS—as well as programming the policy feature card (PFC) on the hot-standby virtual switch supervisor engine. The unified control plane is responsible for the origination and termination of the traffic types described in the "Network Control Plane Traffic" section on page 2-16—as well as being the sole control-point for maintaining and managing inter-switch communications described in the "VSS Inter-Switch Communication and Layer-2 per Link Control Traffic" section on page 2-16.

Note With the first release of VSS, only single supervisors are supported per physical chassis. Therefore, there is no intra-chassis supervisor engine redundancy. A subsequent software release might offer the ability to add a second supervisor engine into each chassis.

Distributed Data Forwarding

As show in Figure 2-13, both supervisor resources (SF and PFC) are active for user-data forwarding. The Policy Feature Card (PFC) and switching fabric (backplane connectivity for fabric-enabled module) of both supervisors are actively forwarding user data and performing policy functions, such as applying access control lists (ACL) and QoS in hardware. Additionally, all Distributed Forwarding Cards (DFC) can also simultaneously perform packet lookups across the entire VSS. Because the switching fabrics of both switches are also in an active state, the Cisco VSS has the switch fabric capacity of 1440 (720 Mbps x 2) Gbps, or 1.44 Tbps in aggregate.

The active and hot-standby supervisors run in a synchronized mode in which the following system information is synchronized over the VSL link:

- Boot environment
- Synchronization of the running configuration
- Protocol states and the database table—Only protocols capable of supporting SSO redundancy (SSO-aware) are fully capable of supporting SSO-based recovery
- Line card status (interface state table and its capabilities)

During the initialization phase, the hot-standby supervisor undergoes configuration synchronization (RF_STANDBY_CONFIG = 5) with the active supervisor (see Table 2-3). Having an understanding of this configuration synchronization is important when considering switch failures detailed in the "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18.

Both active and hot-standby switches can learn the address simultaneously; however, the active virtual switch manages the network information from adjacent devices (such as MAC, STP, or CEF). Several protocols are SSO-aware such that active switch synchronizes protocol information (database, protocol state) to the hot-standby supervisor. In addition, the active supervisor manages and updates the information about interfaces and line card status on both chassis.

The state of the supervisor's control plane (active, hot-standby, or any other state) can be checked using the following CLI commands. Notice that the fabric state is active in both the chassis, indicative of the dual forwarding state.

```
6500-VSS# show switch virtual
Switch mode : Virtual Switch
Virtual switch domain number : 200
Local switch number : 1
Local switch operational role: Virtual Switch Active
Peer switch number : 2
Peer switch operational role : Virtual Switch Standby

6500-VSS# show switch virtual redundancy
My Switch Id = 1
Peer Switch Id = 2
Last switchover reason = none
Configured Redundancy Mode = sso
Operating Redundancy Mode = sso
Switch 1 Slot 5 Processor Information :
-----------------------------------------------
Current Software state = ACTIVE
Uptime in current state = 3 weeks, 4 days, 9 minutes
Image Version = Cisco IOS Software, s72033_rp
Software (s72033_rp-ADVENTERPRISEK9_WAN_DBG-M), Version
12.2(SIERRA_INTEG_070502) INTERIM SOFTWARE
Synced to V122_32_8_11, 12.2(32.8.11)SR on rainier, Weekly
12.2(32.8.11)SX76
Technical Support: http://www.cisco.com/techsupport
Copyright (c) 1986-2007 by Cisco Systems, Inc.
Compiled Thu 03-May-07 09:46 by kchristi
BOOT = sup-bootdisk:s72033-
adventerprisek9_wan_dbg-mz.SIERRA_INTEG_070502,1;
CONFIG_FILE =
BOOTLDR =
Configuration register = 0x2102
Fabric State = ACTIVE
Control Plane State = ACTIVE
Switch 2 Slot 5 Processor Information :
-----------------------------------------------
Current Software state = STANDBY HOT (switchover target)
Uptime in current state = 3 weeks, 4 days, 8 minutes
```

```
Image Version = Cisco IOS Software, s72033_rp
Software (s72033_rp-ADVENTERPRISEK9_WAN_DBG-M), Version
12.2(SIERRA_INTEG_070502) INTERIM SOFTWARE
Synced to V122_32_8_11, 12.2(32.8.11)SR on rainier, Weekly
12.2(32.8.11)SX76
Technical Support: http://www.cisco.com/techsupport
Copyright (c) 1986-2007 by Cisco Systems, Inc.
Compiled Thu 03-May-07 09:46 by kchristi
BOOT = sup-bootdisk:s72033-
adventerprisek9_wan_dbg-mz.SIERRA_INTEG_070502,1;
CONFIG_FILE =
BOOTLDR =
Configuration register = 0x2102
Fabric State = ACTIVE
Control Plane State = STANDBY
```

Note Cisco IOS software images on both supervisors must match otherwise the standby supervisor will boot in RPR mode and the line card will not be active on that chassis. Refer to following publication for further information on RPR: http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/redund.html

Virtual Switch Role, Priorities and Switch Preemption

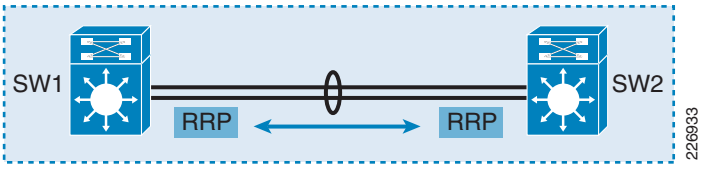
Role Resolution Protocol (RRP)

Having a unified control plane and supporting distributed forwarding with more than one switch requires some form of protocol to determine which switch should (or can) become active, how to change the default setting, and how to deterministically configure which switch member will become active. VSS has a dedicated protocol called Role Resolution Protocol (RRP) that is used to define such behavior. See Figure 2-14.

RRP protocol is used to determine the SSO role (active, hot-standby, or RPR), and to negotiate switch priority and preemption of virtual switch. RRP also checks the software version on each switch which must be the same in order to form a VSS.

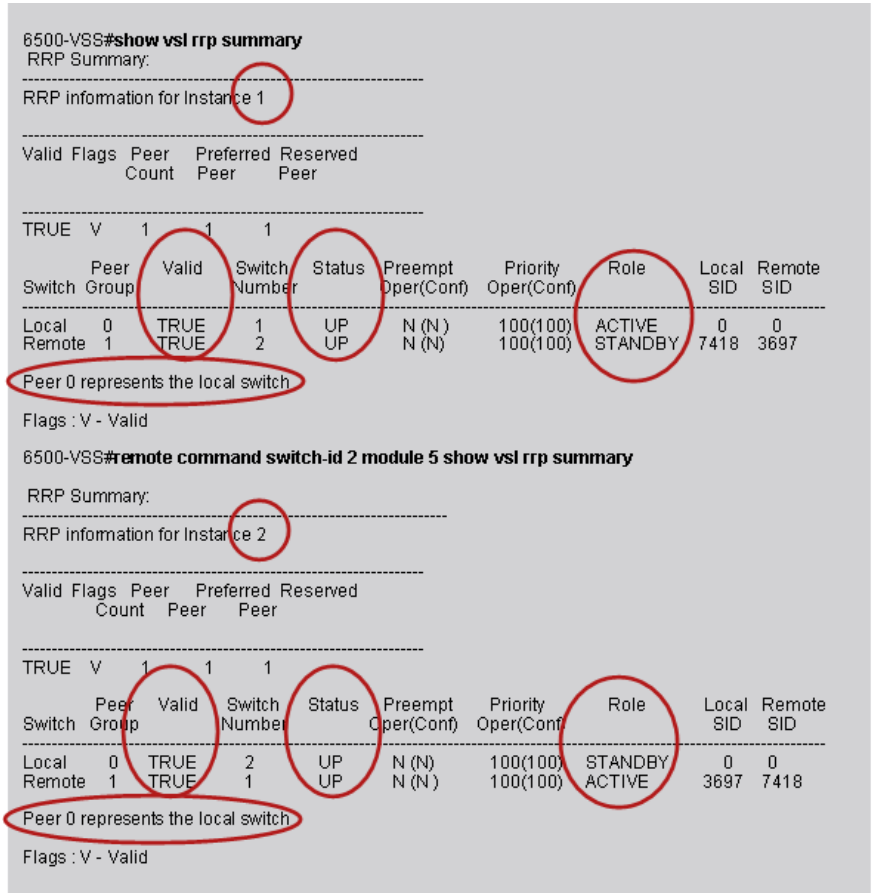
The RRP protocol is initialized once Link Management Protocol (LMP) is fully established on at least one VSL port. The LMP control link is selected by RRP protocol to negotiate the SSO role and switch priority. Each switch member forms local RRP peer group instance and communicate over the control link of the VSL bundle instead running on every VSL link. RRP negotiation packets are encapsulated in the same format as the LMP protocol. Hence, RRP packets are placed into the transmit priority queue in order to prioritize it over data traffic.

Figure 2-14 RRP Negotiation for Role Selection



RRP protocol status can be verified using commands illustrated in Figure 2-15. As shown in the output, the **remote command** *switch-id* command is needed to verify the RRP state of the hot-standby switch. The Peer Group is always 0 on the local switch and 1 for the neighbor switch—regardless of switch ID, priority, preemption, or current SSO role.

Figure 2-15 RRP Protocol Status Information



The RRP session between virtual-switch is negotiated in following conditions:

- When both virtual switches in virtual-switch domain are in bootup mode
- When the hot-standby switch is in the bootup process while the active switch is operational
- When in the dual active recovery phase and the VSL link is restored between virtual switch members

The RRP session is briefly established during the preceding conditions; however, RRP does not maintain a session between two switch members, instead RRP relies on internal notifications from LMP on each switch member. If all VSL member links fail, RRP does not have any control-link path for communication between peers; the LMP protocol on each virtual switch will delete the peer group and notify the RRP process to take SSO-based switchover action. In the absence of VSL link, no RRP action is taken on the active switch if it is still operational; however, the hot-standby switch will transition to the active role because it has no way to determine the remote peer status.

Virtual Switch Priority

The selection of the switch member that is to become the active or hot-standby switch depends on the order and the way in which the switches are initialized. If the switch members boot simultaneously, the switch with the lowest switch ID becomes the active virtual switch. If each switch member boots at different times, then the switch that is initiated first becomes the active virtual switch regardless of switch ID. The VSS can also be configured with switch priority. Usually, the default switch priority (100) need not change; however, you can configure switch priority for two possible requirements:

- You are required to override the default selection of the switch role after initial configuration. This might be due to a need for operational flexibility or requirement for monitoring access. However, the software will not immediately enforce the configured switch priority. Adopting this change requires a reload of both switches in order for the RRP to enable the designated switch to become active. Figure 2-16 illustrates the commands associated with making a change to the switch priority and a syslog message indicating that a reload is required to enforce the new designated priority. If the switch with lower priority comes up first, the software will not force the takeover of the active role when the switch with a higher priority boots up later—unless that switch is configured with preemption.

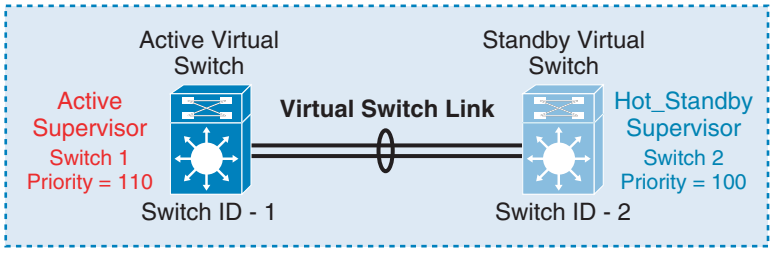
Figure 2-16 Changing Switch Priority

```
6500-VSS#show switch virtual role

Switch  Switch  Status  Preempt      Priority      Role       Session ID
        Number          Oper(Conf)   Oper(Conf)              Local   Remote
-------------------------------------------------------------------------------
LOCAL   1       UP      FALSE(N)     100(100)      ACTIVE     0       0
REMOTE  2       UP      FALSE(N)     100(100)      STANDBY    3071    3108
6500-VSS#conf t
Enter configuration commands, one per line.  End with CNTL/Z.
6500-VSS(config)#switch virtual domain 1
6500-VSS(config-vs-domain)#switch 2 priority 120

Sep 10 11:35:08.945: %VSLP-SW2_SPSTBY-5-RRP_RT_CFG_CHANGE: Configured priority value is different from operational value. Change will take
     effect after config is saved and switch is reloaded.

6500-VSS#show switch virtual role
Switch  Switch  Status  Preempt      Priority      Role       Session ID
        Number          Oper(Conf)   Oper(Conf)              Local   Remote
-------------------------------------------------------------------------------
LOCAL   1       UP      FALSE(N)     100(100)      ACTIVE     0       0
REMOTE  2       UP      FALSE(N)     100(120)      STANDBY    3071    3108
```

226934

- Defining the priority of the switch to match the switch ID will show up in the configuration, which can provide a visual aid in change management. This option is shown in Figure 2-17, where the switch priority is configured to match the switch ID (matching high priority to lowest switch ID).

Figure 2-17 Switch Priority Configured to Match Switch ID

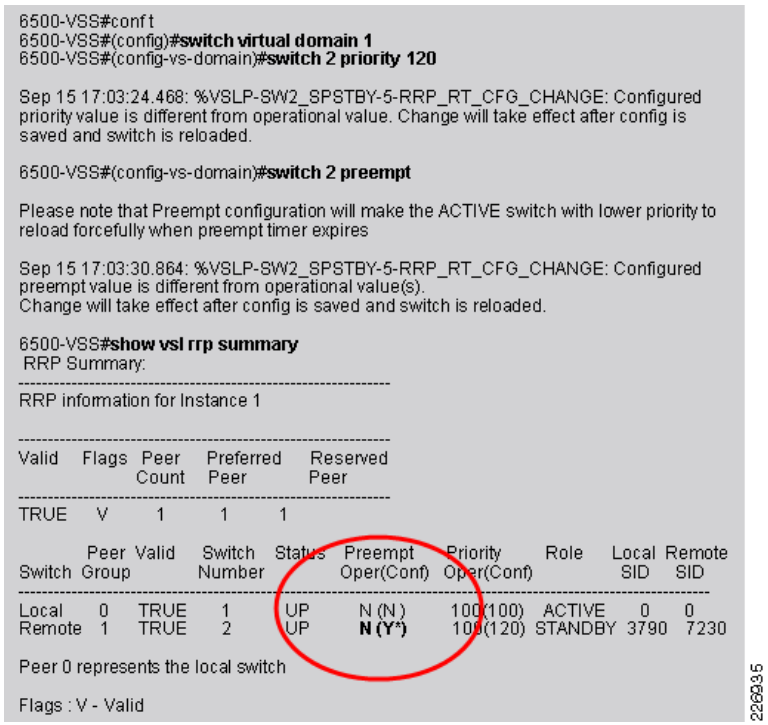


227083

Switch Preemption

If the intention is to select one of the switches to be in the active role in all conditions, then simply increasing switch priority will not achieve this goal. For a deterministic role selection, regardless of the boot order, the desired active switch must be configured with a higher switch ID and switch preemption. See Figure 2-18.The CLI allows the preemption feature to be configured on a low-priority switch. However, the switch preemption setting will not be effective.

Figure 2-18 Configuring Preemption



The implication of switch preemption on the network availability is significant and must be evaluated before deployment. The operational impact of switch preemption should not be compared with widely understood HSRP/GLBP protocol behavior in which preemption only allows relegating the role of being the active HSRP/GLBP forwarder to being in standby mode without much impact to the network (no reload or reset of the chassis).

Switch preemption forces multiple reboots of the VSS member—leading to multiple outages, reducing forwarding capacity while the switches decide which supervisor should assume an active role. For example, if the switch that is configured with preemption fails (or otherwise loses connectivity), then the peer switch will assume the role of being active temporarily. When the preemptive switch boots up and finds that its peer switch is active, the preemptive switch will force the newly active peer switch to give up the role of being active. The only way for the peer switch to give up the active role is by resetting and transitioning to the hot-standby role. Similarly, if the non-preemptive (designated hot-standby) switch somehow comes up first (either due to power failure or delayed user action) and assumes an active role, it will be forced to reset when preemptive switch is brought on line.

Tip **Cisco recommends that you do *not* configure switch preemption for the following reasons:**

- It causes multiple switch resets, leading to reduced forwarding capacity and unplanned network outages.
- The VSS is a single logical switch/router. Both switch members are equally capable of assuming the active role because it does not matter which is active—unless required by enterprise policy.

Virtual Switch Member Boot-Up Behavior

The normal VSS boot-up process consists of diagnostics, VSL link initialization, LMP establishment, and switch role negotiation via RRP. RRP determines the role of each switch leading to the SSO state in which one switch is active and the peer is in the hot-standby state. However, the behavior and the end result are different if there are problems or events leading to VSL interface being inactive/disabled/failed after the RRP negotiation phase in which each switch role was assigned. The VSL link typically becomes disabled for one of two primary reasons:

- Due to the VSL interface and/or line card being non-operational.
- The switch that assumed the role of being active has problems leading to reset of the switch and thus the VSL interface is non-operational.

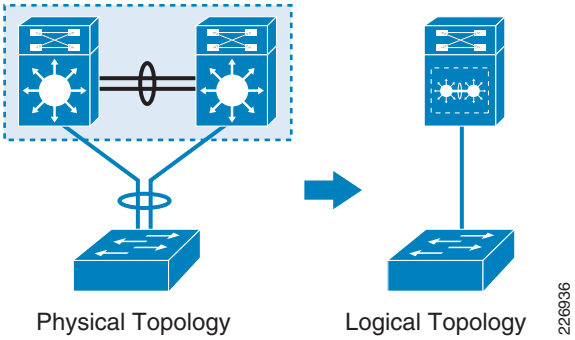
In both cases, the peer switch that has assumed the role of hot standby cannot continue booting, as there is no way to determine (or exchange) the state of the peer switch (due to the VSL being down). The Cisco IOS software will issue a forced reset (described as *crashed* in Cisco IOS parlance) and the peer switch will restart the boot process. If the peer (hot-standby) switch finds that VSL link is still down it will continue booting. In the absence of RRP, it will assume the role of active. This situation can lead to dual-active condition because neither switch has knowledge of the other. For this reason, dual active detection is a mandatory when deploying VSS in campus network. Please refer to the "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18 for more details.

Multi-chassis EtherChannel (MEC)

Traditional EtherChannel aggregates multiple physical links between two switches. MEC is an advanced EtherChannel technology that extends link aggregation to span over two separate switches. VSS allows for distributed forwarding and a unified control plane so that the MEC appears as single port-channel interface existing on both the active and hot-standby switches. Even though the access-layer is connected to a distinct physical chassis via two physical links, from an access-layer switch perspective, this port-channel connection enables a single logical link connected to a single logical switch (referred to as *VSS with MEC*). Figure 2-19 depicts a physical-to-logical transformation in which the logical topology is simplified and (for the spanning tree) VSS with MEC offers no-loops.

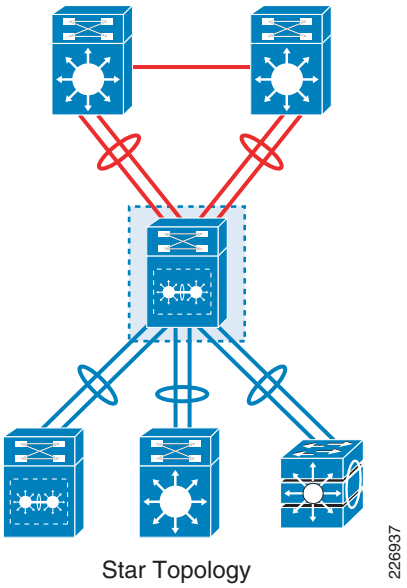
Note MEC configuration is only possible in the VSS; however, access-layer switches requiring connectivity to the VSS are configured with traditional EtherChannel interfaces.

Figure 2-19 MEC—Physical vs. Logical Topology



This capability of spanning EtherChannel over multiple switches as a virtualized, single logical switch creates a topology in which all devices connected via MEC to VSS appear as a star-shaped topology. See Figure 2-20.

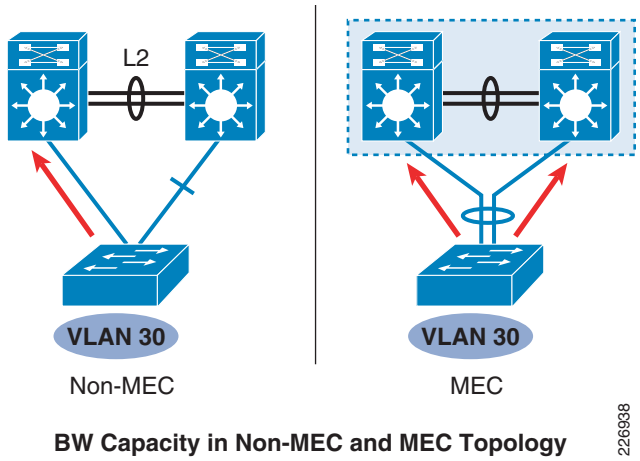
Figure 2-20 Star Topology



The MEC and VSS bring powerful and very effective changes to the campus topology. Two key benefits are as follows:

- *Eliminates loops in multilayer design*—Traditionally, spanning VLANs over multiple closets would create a STP-looped topology because one of the uplinks would be blocked by STP (see Figure 2-21 and Figure 1-4). MEC with VSS together eliminate loops in the campus topology. This is because STP now operates on the EtherChannel logical port and each physical switch appears to be connected via a single logical link to a single logical switch. From the STP viewpoint, this star topology is non-looped. No alternate path exists for STP to be blocked. Figure 2-21 depicts two topologies: The access-layer to VSS topology without MEC in which the uplink is blocked and the access layer to VSS topology with MEC in which both links are forwarding without looping. The advantages to a loop-free network are described in Chapter 3, "VSS-Enabled Campus Design."

Figure 2-21 Bandwidth Capacity in non-MEC and MEC Topologies



BW Capacity in Non-MEC and MEC Topology

- *Doubles the available bandwidth for forwarding*—MEC replaces spanning tree as the means to provide link redundancy. This means that all physical links under the MEC are available for forwarding traffic. The STP can no longer block individual links since its database does not have those links available to calculate a loop-free path. For the network with a looped topology, the total forwarding capacity is half the available bandwidth of physical links. VSS with MEC makes all links available for forwarding and thus doubles the bandwidth available. This offers is an effective change to an existing network in which the lack of a 10-Gbps infrastructure requires choosing a design alternative (EtherChannel on each uplink, routed access, multiple HSRP group, and so on) to efficiently utilize the available links. For any new design, VSS with MEC enables these benefits with simplified topology.

Why MEC is Critical to VSS-Enabled Campus Design

It is important to understand why MEC is critical in a VSS design, before going into detail about MEC functionality, its operation, traffic flow, and implications in campus design. You can have a VSS-enabled network without MEC; however, the resulting topology will consist of either a single point-of-failure (only one link to adjacent network devices) or a looped topology because both links from a given networking device will be connected to a single logical VSS switch in a non-EtherChannel configuration. Either condition reduces the benefits of VSS in any given network. Chapter 3, "VSS-Enabled Campus Design." provides several design proof points to justify the importance of the MEC-enabled design. The following are the major advantages of an MEC-enabled design:

- Enables loop free topology.
- Doubles the available forwarding bandwidth, resulting in reduced application response time, reduced congestion in the network, and reduced operation expenditure.
- Reduces or eliminates control plane activities associated with a single-link failure (either nodal failure or member link failure).
- Allows failure detection in hardware for faster convergence of traffic flows.
- Reduces MAC learning and because the failure of one link does not trigger the Layer-2 control plane convergence.
- Reduces routing protocol announcements—adding efficiency in the Layer-3 network (reduces need for summarization and reduces control plane activity)
- Avoids multicast control plane changes (avoids multicast incoming interface changes), because the failure of one link does not trigger Layer-3 control plane convergence

- Avoids multicast replication over the VSL in an MEC-enabled topology.
- Enables flexibility in deploying dual-active detection techniques (see the "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18).

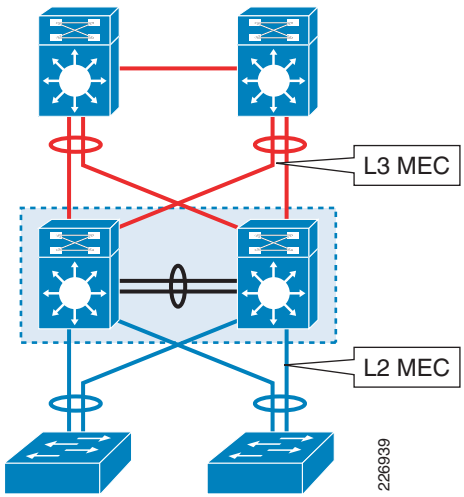
MEC Types and Link Aggregation Protocol

MEC is a distributed EtherChannel environment; however, it inherits all the properties of EtherChannel used in a traditional network. The rest of this section illustrates some, but not all, aspects of EtherChannel technology that apply to MEC. The coverage of EtherChannel here addresses features that might not be obvious or available in other publications—focusing on capabilities that are considered necessary for VSS implementation. Some featured illustrations and protocol configurations are repeated to help readers better understand MEC deployment in the context of a VSS environment.

Types of MEC

Depending upon network connectivity requirements, MEC configuration consists of two modes: Layer 2 and Layer 3. See Figure 2-22.

Figure 2-22 MEC Types



Layer-2 MEC

In a hierarchical, three-layer network, the Layer-2 MEC applies to the connection between the access layer and the distribution layer (where VSS is enabled) as shown in Figure 2-22. In this mode, the Layer-2 MEC is participating in STP, MAC-address learning, and a host of other Layer-2 operations. Layer-2 MEC enabled connectivity can be extended to create large hierarchical Layer-2 topology in which the entire network is loop free. This document does not cover such designs.

Layer-3 MEC

Layer-3 MEC is comprised of a routed port-channel interface. With Layer-3 MEC, the port-channel interface is configured as a routed interface bearing the IP address that participates in Layer-3 functions, such as routing and forwarding using CEF. The natural extension of this type of connectivity is to have multiple, Layer-3 VSS pairs connected together to form the basis for a routed design. The routing application of Layer-3 MEC is discussed in the "Routing with VSS" section on page 3-44.

Link Aggregation Protocol

The EtherChannel is a logical interface. Managing the behavior of a physical member link with (and operational impact on) the rest of the network requires some form of control plane. Two methods available to manage control plane of the underlying link in an EtherChannel group are as follows:

- Port Aggregation Protocol (PAgP)
- Link Aggregation Control Protocol (LACP) or IEEE 802.3ad

Each of these protocols provides the following common functions:

- Ensures link aggregation parameter consistency and compatibility between the VSS and a neighbor switch
- Ensures compliance with aggregation requirements
- Dynamically reacts to runtime changes and failures on local and remote EtherChannel configurations
- Detects and removes unidirectional link connections from the EtherChannel bundle

The EtherChannel is the fundamental building block in a VSS-enabled campus design. The successful deployment of MEC requires operational consistency and interaction with several access-layer switches creating topology that does not introduce unexpected behavior. For these reasons, the MEC interface must be enabled with either PAgP or LACP in order to benefit from functionality described in the preceding discussions. As with traditional EtherChannel, having PAgP or LACP enabled on an MEC provides consistency checks the system configuration and the operational state of the underlying physical links. PAgP and LACP remove a mismatched link from the bundle and thereby provide additional protection against mismatched configurations and operational verification via syslog messages. The following configuration must match in all underlying physical links participating in an MEC bundle.

- Configuration of VLANs on member links
- Trunk type and configuration on member links
- Port status (full- or half-duplex) and QoS support by underlying hardware must be the same in all links

For a complete list of the requirements for forming an EtherChannel, please refer to the individual product release notes and related documentation at www.cisco.com. PAgP and LACP design considerations (as applicable to VSS) that can affect network design are described in "PAgP" and "LACP (IEEE 802.3ad)" sections that follow.

PAgP

PAgP is the mostly widely used link aggregation protocol. PAgP is supported by most Cisco devices and other third-party network interface cards (NIC). In the VSS context, PAgP provides the value-added function of providing assistance in a dual active recovery—in addition to function summarized in the preceding section. PAgP runs on each MEC link member, establishes and manages neighbor adjacency between the VSS and a Layer-2 or Layer-3 partner. PAgP determines multiple attributes on a per member link basis, such as peer-switch state, device name, partner-port, port-aggregation capability, and port-channel bundle group index.

Device ID

The active switch is responsible for origination and termination of PAgP control-plane traffic. PAgP advertises a device-id in order to identify each end of a connection with a unique device status. For VSS, it is a 48-bit number in MAC address format (02.00.00.00.00.*xx*). It consists of a combination of a fixed prefix (02.00.00.00.00) in first five octets and a variable value (*xx*) for last octet. The variable part is the virtual switch domain identifier configured for the system. This PAgP device-id is sent by the two links terminated on two switches that make up a VSS domain. Because the device-id is the same, a remote device assumes that the device-id is coming from the single logical device (the VSS). Even during role switchover in VSS, the device-id remains consistent on each PAgP neighbor to prevent a renegotiation process. This use of the virtual switch domain identifier in PAgP and LACP requires that the identifier to be unique in all VSS domains that are interconnected through the use of MEC. The following command output examples illustrate the device-ID value described in this section, showing the fixed and variable components.

```
6500-VSS# sh run | inc virtual
switch virtual domain 10 ! <-- Device ID
6500-VSS# show pagp neighbor
Flags:S - Device is sending Slow hello.C - Device is in Consistent state.
       A - Device is in Auto mode.        P - Device learns on physical port.

Channel group 10 neighbors
          Partner       Partner          Partner           Partner     Group
Port      Name          Device ID        Port         Age  Flags       Cap.
Gi1/1     6500-VSS      0200.0000.000a   Gi1/4/1      7s   SC          A0001
Gi1/2     6500-VSS      0200.0000.000a   Gi2/4/1      8s   SC          A0001
```

Modes of PAgP Operation

There are many configuration options for PAgP. The best practices for configuring PAgP for EtherChannel are documented in the publication at the following URL:

http://www.cisco.com/en/US/products/hw/switches/ps700/products_white_paper09186a00801b49a4.shtml

This design guide provides selected details from these documents and integrates recommendations that are suited to the VSS-enabled campus environment.

Table 2-4 shows only the best practice-based configuration options. Out of the three choices shown, the recommended mode for PAgP neighbors is *desirable-desirable*. This configuration option enables PAgP on both sides and forces the consistency check mentioned previously in this section. The desirable-desirable option is the best option for ensuring safe and reliable operation of MEC-connected devices. Unlike LACP, PAgP offers strict channel settings and configuration checks prior to bundling the ports. MEC bundle remains in disabled state if PAgP+ detects a configuration mismatch and remains disabled until the error is fixed. This additional measure prevents EtherChannel inconsistency, which would otherwise create operational challenges for managing large-scale EtherChannel deployments.

Table 2-4 Best Practice-based Configuration Options for PAgP

Channel Mode—For both Layer-2 and Layer-3 MEC	VSS	Remote Node	MEC State
	desirable	desirable	operational
	desirable	auto	
	auto	desirable	

The additional configuration options of *silent* or *non-silent* are available based on the network requirements. For most network designs, silent mode is preferred. The silent mode integrates the link into a bundle whether the data is present or not. A non-silent option is used as an indirect measure of link integrity. For the best-practice design, UDLD is the preferred method of link integrity check. As a result, the silent option is more applicable to most network implementations.

Why You Should Keep the PAgP Hello Value Set to Default

By default, PAgP in non-silent mode independently transmits PAgP hello messages at an interval of one per 30 seconds on each member link of an MEC. This is known as *slow-hello* because it takes 105 seconds (3.5 hello intervals) to detect remote partner availability. This timer can be modified so that the PAgP hello is sent every second, which is known as *fast-hello*. Many network designs tend to use this option to accelerate link detection, since UDLD can take longer then fast-hello (3 times 1 second). However, a fast-hello configuration should be avoided in VSS deployment for the following two reasons:

- The VSS control plane might not recover (during the SSO switchover) in 3 seconds, so that the VSS can send a PAgP hello before the remote end declares VSS as being non-responsive. This can lead to false positive
- A fast-hello is sent on a per link basis. For a large-scale deployment, fast-hello transmissions can overrun a switch CPU.

Note Even though one side of the EtherChannel is configured with fast-hello and other side (or device) is configured with slow-hello, operationally they will transmit and receive fast-hellos between devices. This means that even though VSS is configured with slow-hello, improper configuration on remote devices can alter the operational mode of hello messaging.

Tip **The best practice is to keep the PAgP timer settings to default values and to use the normal UDLD to monitor link integrity.**

LACP (IEEE 802.3ad)

LACP is an industry standard port-aggregation protocol that allows for multi-vendor interoperability. LACP is very similar to PAgP in terms of functionality and operation. In VSS, it works for both Layer-2 and Layer-3 MEC interfaces. The following URL provides details of LACP operation and configuration options:

http://www.cisco.com/en/US/products/hw/switches/ps700/products_white_paper09186a00801b49a4.shtml

Device ID

LACP implementation on the VSS uses a pre-computed system-id. The LACP system ID consists of a 48-bit address that is a combination of a fixed prefix in the first five octets (02.00.00.00.00) and variable for last octet (*xx*). The variable part is the virtual switch domain identifier configured for the systems. The following output examples identify the system-id with virtual switch domain number used in last octet of system-id.

```
6500-VSS# sh lacp sys-id
32768,0200.0000.000a ! Device ID info
6500-VSS# sh run | inc virtual
switch virtual domain 10 ! Device ID info
```

LACP Mode of Operation

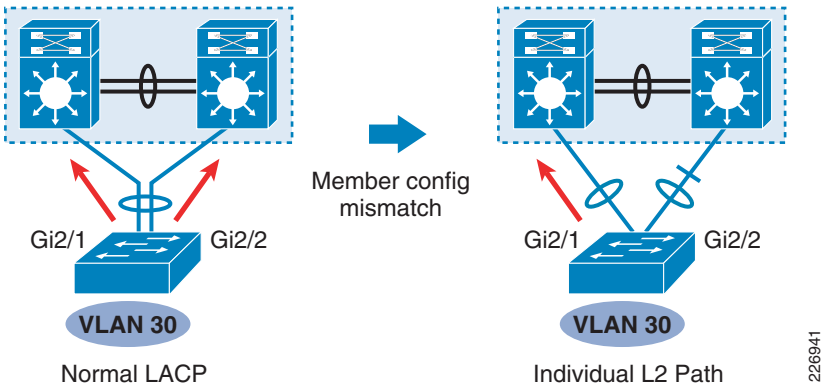
For the same reasons described in preceding PAgP description, Table 2-5 shows only the best practice-based configuration options. Our of the three choices shown, the recommended mode for LACP neighbors is *active-active*

Table 2-5 Best Practice-based Configuration Options for for LACP

Channel Mode—For both Layer-2 and Layer-3 MEC	VSS	Remote Node	MEC State
	active	active	operational
	active	passive	
	passive	passive	

The EtherChannel configured with LACP active-active option allows consistency of configuration on a member link as does PAgP; however, the end result is different. During the EtherChannel bundling process, LACP performs a configuration consistency check on each physical link trying to become a member of the port-channel. If the configuration check fails, a syslog message is generated. In addition, the system generates a special EtherChannel interface and that is assigned with unique alphabetical ID. The system generated LACP MEC will bundle all the physical ports into the MEC that failed the configuration check. See Figure 2-23.

Figure 2-23 LACP Configuration Mismatch Illustration



The following CLI output and configuration process illustrates this behavior. In following output example, the LACP link member is reconfigured with an inconsistent configuration. When a mis-configured interface is attempting to join the port-channel interface, the configuration checks trigger an unique system-generated, port-channel interface.

```
6500-VSS# show etherchannel 20 summary | inc Gi
```

```
Po20(SU)        LACP      Gi2/1(P)       Gi2/2(P)
6500-VSS# show spanning-tree | inc Po20
Po20               Root FWD 3         128.1667 P2p
6500-VSS# config t
6500-VSS(config)# int gi2/2
6500-VSS(config-if)# switchport nonegotiate
6500-VSS(config-if) # shut
6500-VSS(config-if)# no shut
%EC-SPSTBY-5-CANNOT_BUNDLE_LACP: Gi2/2 is not compatible with aggregators in channel 20
and cannot attach to them (trunk mode of Gi2/2 is trunk, Gi2/1 is dynamic)
%EC-SP-5-BUNDLE: Interface Gi2/2 joined port-channel Po20B ! A system generated
port-channel
6500-VSS# show etherchannel 20 summary | inc Gi
Po20(SU)        LACP      Gi2/1(P)
Po20B(SU)       LACP      Gi2/2(P) ! Bundled in separate system-generated port-channel
                                   ! interface

6500-VSS# show spanning-tree | inc Po20
Po20        Root FWD   4         128.1667     P2p
Po20B       Altn BLK   4         128.1668     P2p ! Individual port running STP is blocked
```

This creates two bundles:

- The first bundle is associated with the port that has succeeded in its configuration check.
- The second bundle is system generated and includes the port that did not match configuration.
- As a result, the control plane will be active on both port-channel interfaces (each having one member). The resulting topology consists of two distinct Layer-2 paths created between access-switch and the VSS as shown in Figure 2-23 (this is also true for Layer-3 MEC, but is not shown in the example). The STP topology will consider such network as looped and will block the port with higher STP priority. This is one of the major behavioral considerations for a network topology for LACP compared to PAgP. PAgP offers stricter channel settings and configuration checking prior to bundling the ports. In PAgP, the MEC remains disabled (state is shows as *down*) if PAgP+ detects a configuration mismatch—until the error is fixed.

Why You Should Keep the LACP Hello Value Set to Default

As with PAgP, LACP allows you to configure a hello interval from a default of 30 seconds (slow-hello) to 1 second (fast-hello). Unless the both ends of the LACP neighbor device connection are configured with identical configuration, the LACP hello can be sent at an asymmetric rate from either side. In other words, it is possible that a remote device connected to the VSS can send the LACP with slow-hello and can VSS send the LACP with fast-hello. If the fast-hello method is used for detecting a link failure, the detection time could also be variable based on configuration (30 seconds at one side and three seconds at other end). This is different from PAgP, where the hello transmission rate defaults to fast-hello on both sides (whenever a fast-hello request is made). A fast-hello configuration should be avoided in a VSS deployment for two reasons:

- The VSS control plane might not recover (during the SSO switchover) in 3 seconds (timeout for fast-hello), so that the VSS can send an LACP hello before the remote end declares VSS as being non-responsive. This can lead to false positive
- A fast-hello is sent on a per link basis. For a large-scale deployment, fast-hello transmissions can overrun a switch CPU.
- If the server connected to the VSS is configured based on LACP-MEC with fast-hello, then there is no inherent method to stop the VSS from sending the fast-hello. This can trigger excessive CPU usage when scaled to higher numbers of servers requesting such service.

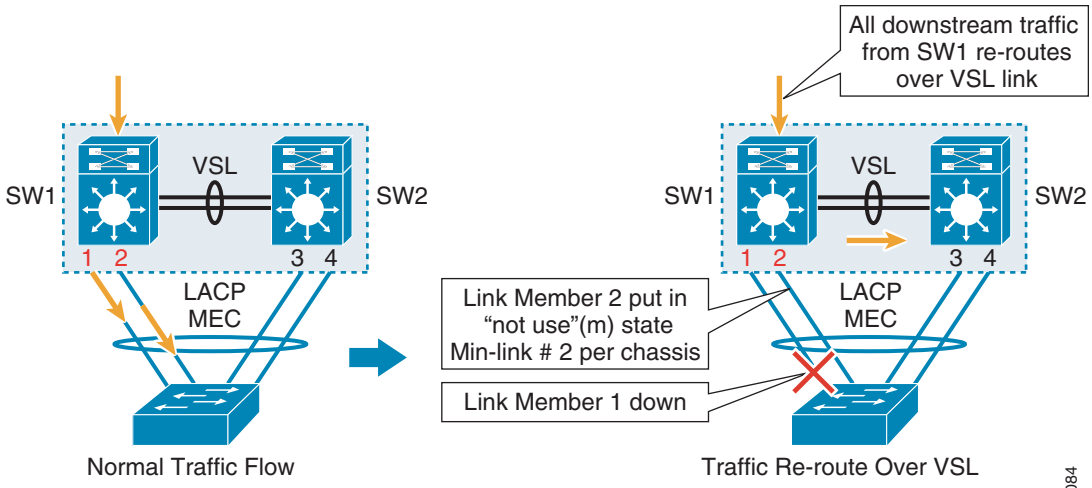
Tip **The best practice is to keep the LACP timer settings to the default values and to use the normal UDLD to monitor link integrity**.

LACP Minimum Link Behavior with VSS

The minimum link feature of LACP is used to maintain a level of bandwidth (in terms of number of interfaces in the *up/up* state) needed for a given application or networking requirement. If the number of links supporting a certain bandwidth falls below a minimum requirement, the bundle is taken out of service. In the standalone design, the minimum link feature is deployed between two separate nodes; the alternate node is available for forwarding traffic as needed. For the VSS, the behavior of the minimum link features is different. In VSS, the LACP EtherChannel minimum link configuration is maintained on a per port-channel basis—although its enforcement is on a per physical chassis basis. The following configuration example and Figure 2-24 illustrate application of this feature.

```
6500-VSS# show etherchannel 10 summary | inc Gi
10     Po10(SU)         LACP      Gi1/4/1(P)     Gi1/4/2(P)     Gi2/4/1(P)     Gi2/4/2(P)

6500-VSS# conf t
6500-VSS(config)# int po10
6500-VSS(config-if)# port-channel min-links 2
6500-VSS(config-if)# int gi1/4/1
6500-VSS(config-if)# shutdown
%LINK-5-CHANGED: Interface GigabitEthernet1/4/1, changed state to administratively down
%LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/4/1, changed state to
down
%LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/4/2, changed state to
down
%EC-SW2_SPSTBY-5-MINLINKS_NOTMET: Port-channel Po10 is down bundled ports (1) doesn't meet
min-links
%EC-SW1_SP-5-MINLINKS_NOTMET: Port-channel Po10 is down bundled ports (1) doesn't meet
min-links
—%LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface GigabitEthernet1/4/2, changed
state to down
%LINK-SW1_SP-5-CHANGED: Interface GigabitEthernet1/4/1, changed state to administratively
down
%LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface GigabitEthernet1/4/1, changed state
to down

6500-VSS# show etherchannel 10 summary
Flags:  D - down        P - bundled in port-channel
        M - not in use, no aggregation due to minimum links not met
        m - not in use, port not aggregated due to minimum links not met

10     Po10(SU)         LACP      Gi1/4/1(D)     Gi1/4/2(m)     Gi2/4/1(P)   Gi2/4/2(P)
```

Figure 2-24 Min-link Configuration Causing Reduced Capacity



Min-link Configuration Causing Reduced Capacity and Undesired Traffic Flow

For a MEC using the LACP control protocol, *minlinks* defines the minimum number of physical links in each chassis for the MEC to be operational. For an example, with the **port-channel min-links 2** configuration on the MEC, each virtual-switch member must match at least two operational local member link ports in order to be associated with the MEC. If one of the member links is down, the other member in the chassis will be put into the *not in use* state. The effect of this enforcement is that one link failure will cause the complete loss of connectivity from one switch member—even though other links are available for forwarding traffic—as shown in the preceding syslog output examples and Figure 2-24. This is will force rerouting of traffic over the VSL link—further congesting the two links that are in the bundle of the other chassis.

For a two-link port-channel configuration (typical access switch-to-VSS network topology), the minimum link feature is not applicable because it looks for the minimum two links per physical chassis and, if configured, will not allow the MEC to be operational. The following output example illustrates the behavior of LACP min-link configuration when a two-member EtherChannel is configured to connect to the VSS with each physical chassis having one member link.

```
6500-VSS# show etherchannel 150 sum
Flags:  D - down        P - bundled in port-channel
Group  Port-channel  Protocol    Ports
------+-------------+-----------+-----------------------------------------------
150    Po150(SU)       LACP      Gi1/7/1(P)     Gi2/7/1(P)

6500-VSS# sh spanning-tree int po 150

Vlan                Role Sts Cost      Prio.Nbr Type
------------------- ---- --- --------- -------- --------------------------------
VLAN0050            Desg FWD 3         128.1667 P2p
VLAN0150            Desg FWD 3         128.1667 P2p

6500-VSS# sh int po 150
Port-channel150 is up, line protocol is up (connected)
  <snip>
  input flow-control is off, output flow-control is off
  Members in this channel: Gi1/7/1 Gi2/7/1
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input never, output never, output hang never
  <<snip>>

6500-VSS# conf t
```

```
Enter configuration commands, one per line.  End with CNTL/Z.
6500-VSS(config)# int po 150
6500-VSS(config-if)# port-channel min-links 2
6500-VSS(config-if)#
%LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/7/1, changed state to
down
%LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet2/7/1, changed state to
down
%LINEPROTO-5-UPDOWN: Line protocol on Interface Port-channel150, changed state to down
%LINK-3-UPDOWN: Interface Port-channel150, changed state to down
%EC-SW1_SP-5-MINLINKS_NOTMET: Port-channel Po150 is down bundled ports (1) doesn't meet
min-links
%SPANTREE-SW1_SP-6-PORT_STATE: Port Po150 instance 50 moving from forwarding to disabled
%SPANTREE-SW1_SP-6-PORT_STATE: Port Po150 instance 150 moving from forwarding to disabled
%EC-SW1_SP-5-MINLINKS_NOTMET: Port-channel Po150 is down bundled ports (1) doesn't meet
min-links
%LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface GigabitEthernet1/7/1, changed state
to down
%LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface GigabitEthernet2/7/1, changed state
to down
%EC-SW2_SPSTBY-5-MINLINKS_NOTMET: Port-channel Po150 is down bundled ports (0) doesn't
meet min-links
%EC-SW2_SPSTBY-5-MINLINKS_NOTMET: Port-channel Po150 is down bundled ports (1) doesn't
meet min-links
```

For an MEC using LACP control protocol, min-links defines the minimum number of physical links in each chassis for a MEC to be operational. With a single member connected to each physical chassis, the configuration violates the minimum link requirements. The following example output illustrates that the port-channel interface is disabled and that the LACP state for each member link is in a wait state. The usage of the min-link feature disables the MEC for the associated connection.

```
6500-VSS# sh int po 150
Port-channel150 is down, line protocol is down (notconnect)
! <<snip>>
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input never, output never, output hang never
  Last clearing of "show interface" counters never
  Input queue: 0/2000/0/0 (size/max/drops/flushes); Total output drops: 0
  Queueing strategy: fifo
  Output queue: 0/2000 (size/max)

6500-VSS# sh etherchannel 150 su
Flags:  D - down        P - bundled in port-channel
        w - waiting to be aggregated

Group  Port-channel  Protocol    Ports
------+-------------+-----------+-----------------------------------------------
150    Po150(SM)       LACP      Gi1/7/1(w)     Gi2/7/1(w)

Last applied Hash Distribution Algorithm: Adaptive

6500-VSS# sh spanning-tree int po 150
no spanning tree info available for Port-channel150
```

Tip **The use of the minimum links feature in the campus environment is not very effective. For all practical deployments of the VSS in the campus (two uplinks from each adjacent network devices), the minimum links feature should not be used.**

MEC Configuration

Configuration requirement for MEC is almost identical with standard EtherChannel interface. This section covers the basic configuration caveats, QoS support, and syslog guidelines for MEC configuration.

The procedure used to configure Layer-2 EtherChannel differs when compared with Layer-3 EtherChannel. Key considerations are as follows:

- Do not create Layer-2 MEC explicitly by defining it via the CLI. Instead, allow the Cisco IOS system to generate the Layer-2 MEC interface by associating to the port-channel group under each member interface.
- Create a Layer-3 MEC interface explicitly via the CLI and associate the port-channel group under each member interface.

Refer to the following URL for more information:

http://cco.cisco.com/en/US/partner/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/channel.html#wp1020478

QoS with MEC

QoS for MEC follows similar procedures as with any standard EtherChannel configuration. For generic QoS support or restrictions related to the VSS, please refer to following white paper QoS chapter:

http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/white_paper_c11_429338.pdf

Monitoring

Because EtherChannel is ubiquitous in a VSS-enabled campus topology, monitoring EtherChannel is more critical in comparison to non-VSS EtherChannel environments. Generic tools available for monitoring a standard EtherChannel interface are fully applicable to an MEC interface. However, this section provides additional details about newer **show** command and specific **logging** commands that should be enabled to enhance operational understanding of EtherChannel connectivity.

Cisco IOS Release 12.2(33)SXH1 for the Cisco Catalyst 6500 platform now supports previously hidden commands for monitoring traffic flow on a particular link member of a MEC or EtherChannel. The following **remote** command will generate output depicting the interface and the port-channel interface being selected for a given source and destination

Hidden commands:

```
Catalyst6500# remote command switch test EtherChannel load-balance interface po 1 ip
1.1.1.1 2.2.2.2
Would select Gi4/1 of Po1
```

Cisco IOS command available for monitoring generic EtherChannel implementations:

```
6500-VSS# show EtherChannel load-balance hash-result interface port-channel 2 205 ip
10.120.7.65 vlan 5 10.121.100.49
Computed RBH: 0x4
Would select Gi1/9/19 of Po205
```

The following syslog configuration (**logging** command) is recommended in VSS with MEC interfaces. These commands can also be applied to the devices connected to the VSS as long as those devices support the respective syslog functionality.

Port-channel interfaces configurations:

```
interface Port-channel20
 logging event link-status
```

```
 logging event spanning-tree status

logging event link-status
%LINK-5-CHANGED: Interface Port-channel220, changed state to administratively down
%LINK-SW1_SP-5-CHANGED: Interface Port-channel220, changed state to administratively down

logging  event spanning-tree status
%SPANTREE-SW1_SP-6-PORT_STATE: Port Po220 instance 999 moving from learning to forwarding
```

Member link configuration:

```
interface GigabitEthernet1/8/1
 description Link member to port-channel
 logging event link-status
 logging event trunk-status
 logging event bundle-status

logging event link-status
Mar 25 11:43:54.574: %LINK-3-UPDOWN: Interface GigabitEthernet1/8/1, changed state to down
Mar 25 11:43:54.990: %LINK-3-UPDOWN: Interface GigabitEthernet2/8/1, changed state to down

logging event trunk-status
%DTP-SW2_SPSTBY-5-NONTRUNKPORTON: Port Gi2/8/1 has become non-trunk
%DTP-SW1_SP-5-NONTRUNKPORTON: Port Gi2/8/1 has become non-trunk

logging event bundle-status
%EC-SW2_SPSTBY-5-BUNDLE: Interface Gi1/8/1 joined port-channel Po220
%EC-SW2_SPSTBY-5-BUNDLE: Interface Gi2/8/1 joined port-channel Po220
```

MEC Load Sharing, Traffic Flow and Failure

Load sharing, traffic flow behavior, and failure conditions are covered in Chapter 3, "VSS-Enabled Campus Design," because the behavior and impact of MEC load sharing is more related to overall design of the campus network.

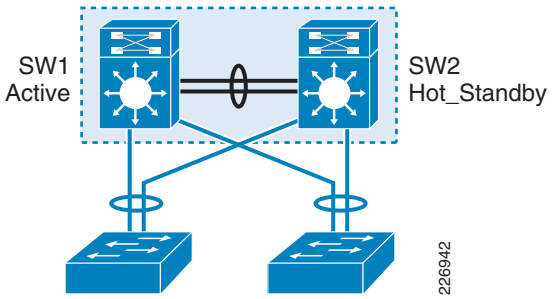
Capacity Planning with MEC

The maximum number of EtherChannels supported by a VSS depends on version of Cisco IOS. A maximum of 128 EtherChannels are supported in Cisco IOS 12.2(33) SXH. This limit is raised to 512 EtherChannels with Cisco IOS 12.2(33) SXI software releases. The scope and scalability as it applies to the VSS in the distribution block is discussed in the "Multilayer Design Best Practices with VSS" section on page 3-14.

MAC Addresses

In a standalone Cisco Catalyst 6500, the MAC addresses used for each interface and control plane is derived from the back plane EEPROM. VSS consists of two chassis (see Figure 2-25). Each physical member of VSS pair consists of pool of MAC addresses stored in backplane EEPROM.

Figure 2-25 VSS Roles



The VSS MAC address pool is determined during the role resolution negotiation. The active chassis pool of MAC addresses is used for the Layer-2 SVI and the Layer-3 routed interface—including the Layer-3 MEC interface. The Layer-2 MEC interface uses one of the link-member MAC addresses. The following CLI output examples reveal the active pool of MAC addresses.

```
6500-VSS# show switch virtual role

Switch  Switch Status  Preempt    Priority  Role
------------------------------------------------------------------
LOCAL    1     UP   FALSE(N )   110(110)  ACTIVE
REMOTE 2     UP   FALSE(N )   100(100)  STANDBY

6500-VSS# show catalyst6000 chassis-mac-addresses
  chassis MAC addresses: 1024 addresses from 0019.a927.3000 to 0019.a927.33ff
```

The MAC address allocation for the interfaces does not change during a switchover event when the hot-standby switch takes over as the active switch. This avoids gratuitous ARP updates (MAC address changed for the same IP address) from devices connected to VSS. However, if both chassis are rebooted together and the order of the active switch changes (the old hot-standby switch comes up first and becomes active), then the entire VSS domain will use that switch's MAC address pool. This means the interface will inherit a new MAC address, which will trigger gratuitous ARP updates to all Layer-2 and Layer-3 interfaces. Any networking device connected one hop away from the VSS (and any networking device that does not support gratuitous ARP), will experience traffic disruption until the MAC address of the default gateway/interface is refreshed or timed out. To avoid such a disruption, Cisco recommends using the configuration option provided with the VSS in which the MAC address for Layer-2 and Layer-3 interfaces is derived from the reserved pool. That takes advantage of the virtual switch domain identifier to form the MAC address. The MAC addresses of the VSS domain remain consistent with the usage of virtual MAC addresses, regardless of the boot order. For the exact formula, see the command and configuration chapter for VSS at www.cisco.com.

Tip **Cisco recommends the configuration of a virtual MAC address for VSS domain using the *switch virtual domain* command.**

```
6500-VSS(config)# switch virtual domain 10
6500-VSS(config-vs-domain)# mac-address use-virtual
```

The individual MAC addresses that reside in each chassis EEPROM are useful for assisting in the dual-active detection process. The following **show** commands illustrate how to find the base address for each chassis.

```
6500-VSS# sh idprom switch 1 backplane detail | inc mac
  mac base = 0019.A927.3000

6500-VSS# sh idprom switch 2 backplane detail | inc mac
```

```
  mac base = 0019.A924.E800
```

The above base addresses located on each chassis are used by the dual-active detection method described in "Campus Recovery with VSS Dual-Active Supervisors" section on page 4-18.

MAC Addresses and MEC

In VSS, the MAC address of the Layer-2 MEC interface is derived from one of the link-member burn-in (bia) interface MAC addresses. In a normal condition, the MAC address of the first interface added to the port-channel interface is chosen for Layer-2 port-channel (MEC) interface. If the interface whose MAC address is used for the port-channel is disabled, the port-channel interface will start using remaining member interface MAC addresses. However, if the interface that has just been disabled is reactivated, the Layer-2 MEC does not reuse the MAC address of that interface.

The process of inheriting the MAC address of the Layer-2 port-channel is shown below. The burn-in address (bia) of the the port-channel is derived from the first interface that was added to the port-channel.

```
6500-VSS#show etherchannel summary  | inc 220
220    Po220(SU)        LACP      Gi1/8/1(P)     Gi2/8/1(P)

6500-VSS#show interface gig 1/8/1 | inc bia
  Hardware is C6k 1000Mb 802.3, address is 0014.a922.598c (bia 0014.a922.598c)

6500-VSS#show interface gig 2/8/1 | inc bia
  Hardware is C6k 1000Mb 802.3, address is 0014.a92f.14d4 (bia 0014.a92f.14d4)
```

Note that the output where port-channel interface MAC address is derived from Gigabit interface 1/8/1.

```
6500-VSS#show interface port-channel 220 | inc bia
  Hardware is EtherChannel, address is 0014.a922.598c (bia 0014.a922.598c)


6500-VSS# conf t
6500-VSS(config)# interface gi1/8/1
6500-VSS(config-if)# shutdown
```

After disabling the link-member Gigabit-interface 1/8/1 whose MAC address was used by Layer-2 MEC, the port-channel starts using the remaining member (Gigabit-interface 2/8/1) burned-in address.

```
6500-VSS#show interface port-channel 220 | inc bia
  Hardware is EtherChannel, address is 0014.a92f.14d4 (bia 0014.a92f.14d4)

6500-VSS#show interface gig 2/8/1 | inc bia
  Hardware is C6k 1000Mb 802.3, address is 0014.a92f.14d4 (bia 0014.a92f.14d4)
```

If the interface is re-added to port-channel bundle, the MAC address of the port-channel does not change. The below CLI output illustrates that behavior.

```
6500-VSS(config)#interface gig 1/8/1
6500-VSS(config-if)#no shutdown


6500-VSS#show interface port-channel 220 | inc bia
  Hardware is EtherChannel, address is 0014.a92f.14d4 (bia 0014.a92f.14d4)

6500-VSS#show interface g 2/8/1 | inc bia
  Hardware is C6k 1000Mb 802.3, address is 0014.a92f.14d4 (bia 0014.a92f.14d4)
```

```
6500-VSS#show interface g 1/8/1 | inc bia
  Hardware is C6k 1000Mb 802.3, address is 0014.a922.598c (bia 0014.a922.598c)
```

In normal condition, the Layer-2 MAC address is used as sources of BPDU frame (addition of link activates the MEC interface and STP operation on that port). If the interface member (whose MAC address is used by Layer-2 MEC) is disabled, the change of source MAC in BPDU frame is detected by the switches connected to the VSS; however, the root MAC of the STP remains the same. This implies that STP topology did not change. For details, refer to "STP Operation with VSS" section on page 3-36.

CHAPTER 3

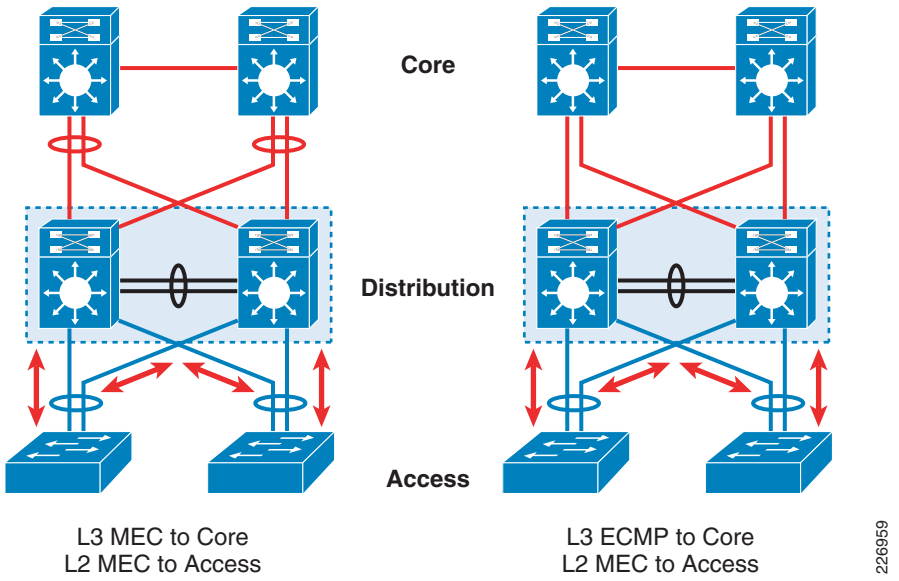
VSS-Enabled Campus Design

VSS-enabled campus design follows the three-tier architectural model and functional design described in Chapter 1, "Virtual Switching Systems Design Introduction," of this design guide. This chapter covers the implementation of VSS in campus design, specifically at the distribution layer addressing all relevant configuration details, traffic flow, failure analysis, and best practice recommendations. The chapter is divided into the following main sections:

EtherChannel Optimization, Traffic Flow, and VSL Capacity Planning with VSS in the Campus

Traditionally, multilayer campus design convergence, traffic-load share and failure characteristics are governed by three key technology factors: STP, FHRP, and topology (looped and non-looped). In VSS-enabled campus, the EtherChannel replaces all three factors and thus is the fundamental building block. The EtherChannel application at Layer-2 and Layer-3 plays a central role in handling user data traffic during stated-state and faulty condition. VSS deployment at the distribution layer does not change any physical topology and connectivity between hierarchical layers—core, distribution, and access. As shown in Figure 3-1, the best practice network retains its redundant systems and links in a fully-meshed topology. For the connectivity between the access layer and VSS, Layer-2 MEC is necessary and integral part of the campus design. The connectivity option from VSS at the distribution (typical Layer-2 and Layer-3 boundary) to Layer-3 domain has two choices: ECMP or Layer-3 MEC. The Layer-3 MEC option is compared to ECMP options in the context of convergence, multicast flows, and dual-active event considerations. For both Layer-2 and Layer-3 options, MEC is ubiquitous in a VSS-based environment so understanding traffic flows and failure behaviors as it relates to MEC in the VSS-enabled design is critically important for both design scenarios. This section also addresses capacity planning associated with a VSL bundle and traffic flow within a VSS campus. The subsequent multilayer and routing sections use this information to develop best-practice recommendations for various failure scenarios.

Figure 3-1 Redundant VSS Environment



Traffic Optimization with EtherChannel and MEC

MEC is the foundation of the VSS-enabled campus. The logical topology created by EtherChannel governs most of the convergence and load sharing of traffic in the VSS environment. The EtherChannel load sharing consists of a highly specific topology, application flow, and user profile. One key concept in traffic optimization in an EtherChannel-base environment is the *hash algorithm*. In general, hash-based mechanisms were devised so that traffic flows would be statistically distributed, based on mathematical functions, among different paths. Consider the following environments and their affects on the effectiveness of a hash-based optimization:

- Core devices carry higher amounts of application flows from various users and application-end points. These flows carry unique source and destination IP addresses and port numbers. These *many-to-many* flows can provide useful input to a hash algorithm and possibly result in better load-sharing with the default setting.
- The access-to-core traffic pattern generally consists of *few-to-few* traffic patterns. This is because the end host communicates to the default gateway. As a result, all traffic flows from hosts on an access-switch have the same destination IP address. This reduces the possible input variation in a hash calculation such that optimal load sharing might not be possible. In addition, the traffic load is asymmetrical (downstream flows traffic is higher then upstream flows).

Due to variations in application deployment and usage patterns, there can be no *one-size-fits-all* solution for the optimization of load sharing via hash tuning. You might need to analyze your network and tune optimization tools based on specific organizational requirements. The following guidelines apply:

- The more values used as an input in the hash calculation, the more likely the outcome of the hash result be fair in link selection.
- Layer-4 hashing tends to be more random than Layer-3 hashing. More input variation, due to diversity in Layer-4 application port numbers, tends to generate better load-sharing possibilities.
- Layer-2 hashing is not efficient when everyone is talking to a single default gateway. Host communicating to default gateway uses the same MEC as destination; as a result only Layer-2-based hash input variation is not optimal.

This design guide does not validate one hash-tuning solution over any other because of the preceding considerations. However, recent advancements associated with EtherChannel traffic optimization are worth understanding and considering while deploying VSS-enabled campus design.

Cisco Catalyst 6500 EtherChannel Options

Cisco offers a variety of EthernetChannel-capable systems. The following options are applicable to both standalone and VSS-enabled Cisco Catalyst 6500s:

Adaptive vs Fixed

As of Cisco IOS Release 12.2(33) SXH, the Cisco Catalyst 6500 supports an enhanced hash algorithm that pre-computes the hash value for each port-channel member link. Adaptive hashing does *not* require each member link to be updated to rehash the flows of the failed link, thus reducing packet loss. The flow will be dynamically rehashed to an available link in the bundle. This enhanced hash implementation is called an *adaptive hash*. The following output example illustrates the options available:

```
6500-VSS(config-if)# port-channel port hash-distribution ?
  adaptive  selective distribution of the bndl_hash among port-channel members
  fixed     fixed distribution of the bndl_hash among port-channel members

6500-VSS(config-if)# port-channel port hash-distribution fixed

This command takes effect when a member link UP/DOWN/ADDITION/DELETION event occurs.
Perform a shutdown and no shutdown command sequences to take immediate effect.
```

By default, the load-sharing hashing method on all non-VSL EtherChannel is *fixed*. The adaptive algorithm is useful for the switches in the access layer for reducing the upstream traffic loss during a link member failure; however, its application or configuration on VSS is only useful if there are more than two links per member chassis. This is because, with two links, the algorithm has a chance to recover flows from the failed links to the remaining locally connected link. With one link in each chassis (a typical configuration), the failed link will force the traffic over the VSL that is not considered to be a member link within the same chassis.

VLAN ID as Hash Variable

For Sup720-3C and Sup720-3CX-enabled switches, Cisco Catalyst switches now support a mixed mode environment that includes VLAN information into the hash. The keyword **enhanced** in the **show EtherChannel load-balance** command output indicates whether the VLAN is included in the hash. Refer to the following output examples:

```
6500-VSS# show platform hardware pfc mode
PFC operating mode : PFC3CXL ! Indicates supervisor capable of VLAN id used as a hash
Configured PFC operating mode : None

6500-VSS# sh EtherChannel load-balance
EtherChannel Load-Balancing Configuration:
        src-dst-ip enhanced ! Indicates VLAN id used as a hash
EtherChannel Load-Balancing Addresses Used Per-Protocol:
Non-IP: Source XOR Destination MAC address
  IPv4: Source XOR Destination IP address and TCP/UDP (layer-4) port number
! << snip >>
```

The VLAN ID can be especially useful in helping to improve traffic optimization in two cases:

- With VSS, it is possible to have more VLANs per closet-switch and thus better sharing traffic with the extra variables in the hash input.
- In situations where traffic might not be fairly hashed due to similarities in flow data; for an example, common multicast traffic will often hash to the same bundle member. The VLAN ID provides an extra differentiator.

However, it is important to understand that VLAN-hashing is only effective if each physical chassis of VSS has more than one link to the access layer. With single link per-chassis to the access layer, there is no load-share form each member switch.

Optional Layer-3 and Layer-4 Operator for Hash Tuning

As of Cisco IOS Release 12.2 (33)SXH, Cisco Catalyst switches support a mixed mode that includes both Layer-3 and Layer-4 information in the hash calculation. The default option is listed below in bold, whereas the preferred option is listed as pointers.

```
VSS(config)# port-channel load-balance ?
  dst-ip                 Dst IP Addr
  dst-mac                Dst Mac Addr
  dst-mixed-ip-port      Dst IP Addr and TCP/UDP Port <-
  dst-port               Dst TCP/UDP Port
  mpls                   Load Balancing for MPLS packets
  src-dst-ip             Src XOR Dst IP Addr
  src-dst-mac            Src XOR Dst Mac Addr
  src-dst-mixed-ip-port  Src XOR Dst IP Addr and TCP/UDP Port <-
  src-dst-port           Src XOR Dst TCP/UDP Port
  src-ip                 Src IP Addr
  src-mac                Src Mac Addr
  src-mixed-ip-port      Src IP Addr and TCP/UDP Port <-
  src-port               Src TCP/UDP Port
```

CLI Showing Flow Hashing over MEC and Standard EtherChannel Interfaces

Refer to "Monitoring" section on page 2-43 for examples that illustrate the use of the switch CLI for monitoring a particular flow. The CLI command is only available for the Cisco Catalyst 6500 platform.

Catalyst 4500 and 3xxx Platform

The EtherChannel hash-tuning options vary with each platform. The Cisco Catalyst 4500 offers similar flexibility in comparison with the Cisco Catalyst 6500, allowing Layer-3 and Layer-4 operators. The Cisco Catalyst 36xx and Cisco Catalyst 29xx Series switches have default hash settings that incorporate the source MAC address that might not be sufficient to enable equal load sharing over the EtherChannel. The configuration examples that follow show default values in bold and preferred options highlighted with an arrow.

Cisco Catalyst 4500:

```
Catalyst4500(config)# port-channel load-balance ?
  dst-ip        Dst IP Addr
  dst-mac       Dst Mac Addr
  dst-port      Dst TCP/UDP Port
  src-dst-ip    Src XOR Dst IP Addr
  src-dst-mac   Src XOR Dst Mac Addr
  src-dst-port  Src XOR Dst TCP/UDP Port <-
  src-ip        Src IP Addr
  src-mac       Src Mac Addr
  src-port      Src TCP/UDP Port
```

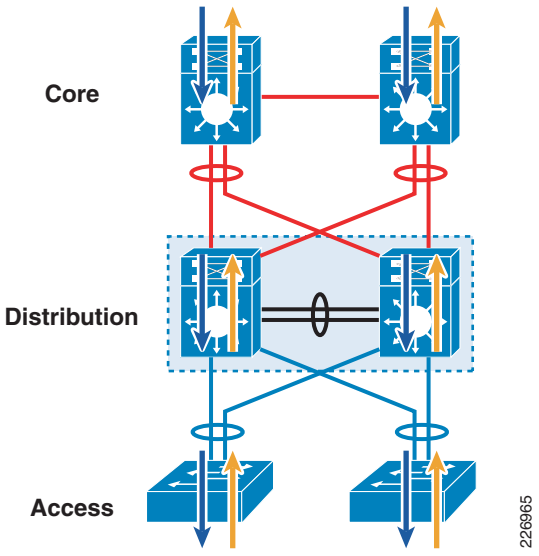
Cisco Catalyst 36xx, Cisco Catalyst 37xx Stack, and Cisco Catalyst 29xx:

```
Catalyst3700(config)# port-channel load-balance ?
  dst-ip       Dst IP Addr
  dst-mac      Dst Mac Addr
  src-dst-ip   Src XOR Dst IP Addr <-
  src-dst-mac  Src XOR Dst Mac Addr
  src-ip       Src IP Addr
  src-mac      Src Mac Addr
```

Traffic Flow in the VSS-Enabled Campus

The VSS environment is designed such that data forwarding always remains within the member chassis. As shown in Figure 3-2, the VSS always tries to forward traffic on the locally available links. This is true for both Layer-2 and Layer-3 links. The primary motivation for local forwarding is to avoid unnecessarily sending of data traffic over the VSL link in order to reduce the latency (extra hop over the VSL) and congestion. Figure 3-2 illustrates the normal traffic flow in a VSS environment where VSS connectivity to the core and the acces layer is enabled via fully-meshed MEC. In this topology, the upstream traffic flow load-share decision is controlled by access layer Layer-2 EtherChannel, and downstream is controlled by the core devices connected via Layer-3 EtherChannel. The bidirectional traffic is load-shared between two VSS member; however, for each VSS member, ingress and egress traffic forwarding is based on locally-attached links that are part of MEC. This local forwarding is a key concept in understanding convergence and fault conditions in a VSS-enabled campus network.

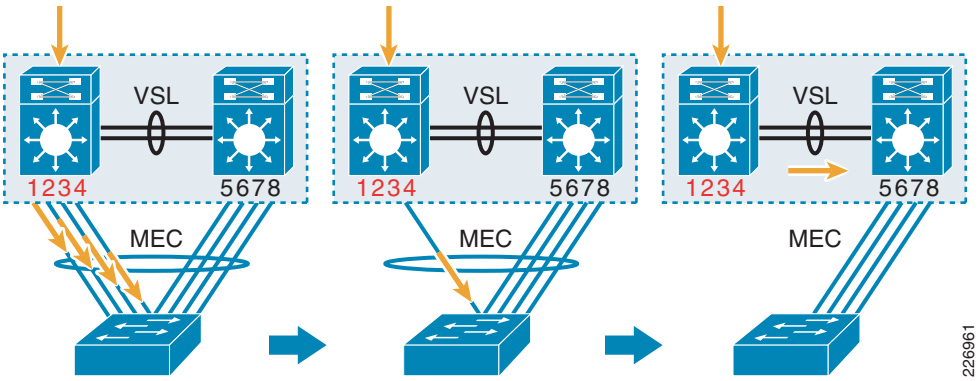
Figure 3-2 VSS Traffic Flow Overview



Layer-2 MEC Traffic Flow

As described above, in a normal mode, the VSS always prefers locally-attached links. This is elaborated for Layer-2 MEC connectivity in Figure 3-3, where the traffic flow behavior is depicted with three different state of the network connectivity. The example describes the fault condition (see center of Figure 3-3) where three out of four links have become non-operational. Since one link is still operational to a VSS member, the downstream traffic still chooses that link, despite the fact that the other VSS member switch has four additional links in the same EtherChannel group reachable via VSL. If all links (1, 2, 3, and 4) fail, the VSS systems detects this condition as an orphaned connectivity, the control plane reprograms all traffic flows over VSL link, and then forwarded via the available MEC links to the access layer.

Figure 3-3 Layer-2 MEC Traffic Flow during Layer-2 MEC Member-link Failure



The case illustrated at the center of Figure 3-3 shows a failure in which a single link is carrying all traffic. In that case, the link can become be oversubscribed. However, this type of connectivity environment is not a common topology. Usually, the access-layer switch is connected via two uplinks to the VSS. In that case, a single link failure forces the traffic to traverse the VSL link. It is important to differentiate the control plane traffic flow from user data traffic flow. The control plane traffic can use either switch to originate traffic. For example, the UDLD, CDP, or any other link-specific protocol that must be originated on a per-link basis will traverse the VSL. However, a *ping* response to a remote device will always choose the local path from the link connected to the peer, because the remote node might have chosen that link from the local hash result—even though the request might have come over the VSL.

Layer-3 MEC Traffic Flow

Layer-3 MEC connectivity form VSS to the core layer consist of a two port-channels. Each port-channel has two links, each on separate physical chassis. When one of the link member of the port-channel fails, the VSS will select another locally available link (which is under distinct port-channel interface) to reroute the traffic. This is similar to ECMP failure, where the path selection occurs based on local system link availability. This type of connectivity also has dependencies on routing protocol configuration and therefore it is described in the "Routing with VSS" section on page 3-44.

Layer-3 ECMP Traffic Flow

Fully-meshed ECMP topology consists of four distinct routing paths (one from each link) for a given destination for the entire VSS. However, each member VSS is programmed with two paths (two links) that translate to two unique Cisco Express Forwarding (CEF) hardware path. For a normal condition, for each member chassis, the traffic from the access layer to the core uses two locally-available links (hardware path). To illustrate the traffic flow behavior, Figure 3-4 is split into three stages. The first stage (see Figure 3-4-(A)), the ingress traffic is load-shared among two equal cost paths. When a single link fails (see Figure 3-4 – (B)), the ingress traffic for SW1 will select remaining link. If all local links fail (see Figure 3-4 – (C)), the FIB is reprogrammed to forward all the flows across the VSL link to another member. The output of the forwarding table is shown in Figure 3-5 and corresponds to the failure status of Figure 3-4.

Figure 3-4 Example Unicast ECMP Traffic Flow

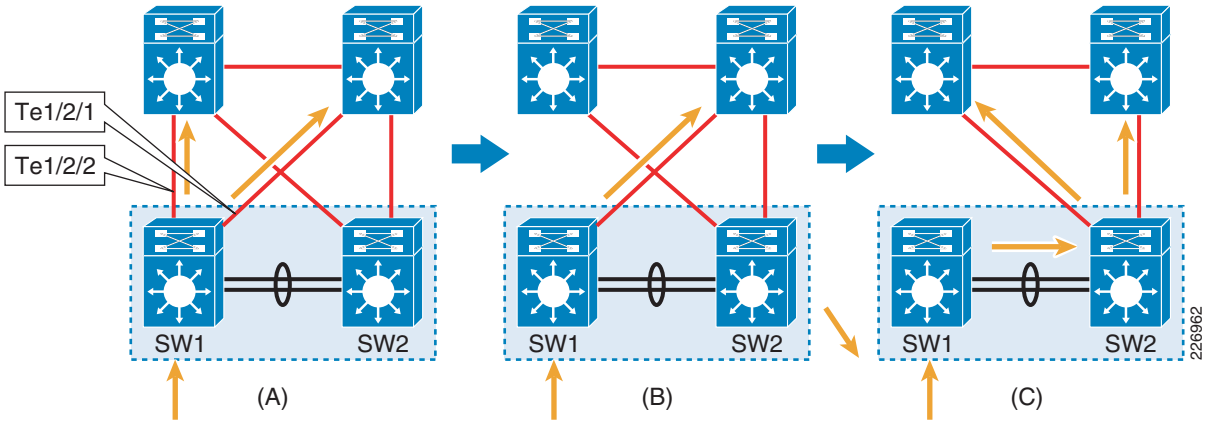
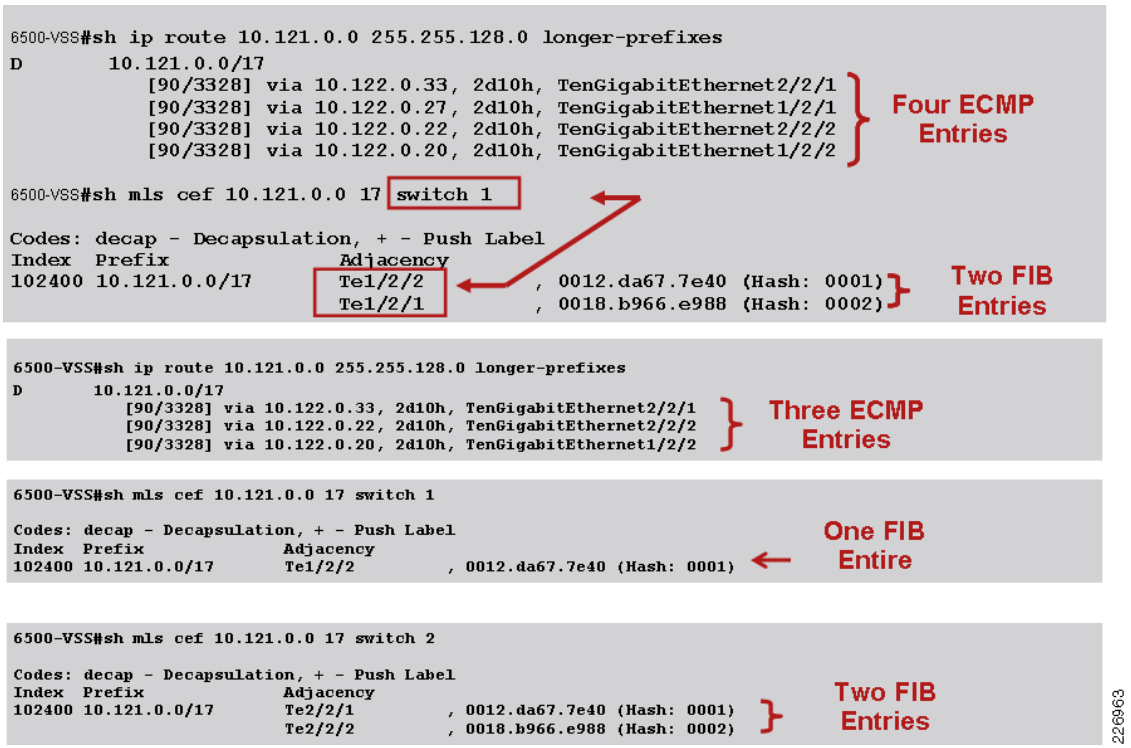


Figure 3-5 ECMP Forwarding Entries—Global and Switch Specifics



Multicast Traffic Flow

VSS shares all the benefits and restrictions of standalone Multicast Multilayer Switching (MMLS) technology. The MMLS enables multicast forwarding redundancy with dual supervisor. Multicast forwarding states include (*,g) and (s,g), which indicate that the incoming and outgoing interface lists for a given multicast flow are programmed in the Multicast Entries Table (MET) on the active supervisor Policy Feature Card (PFC). This table is synchronized in the hot-standby supervisor. During the switchover, the multicast data flows are forwarded in hardware, while the control plane recovers and reestablishes Protocol Independent Multicast (PIM) neighbor relations with its neighbors. The user data traffic flow, which requires replication in the hardware, follows the same rule as unicast as far as VSS forwarding is concerned. VSS always prefers a local link to replicate multicast traffic in a Layer-2 domain. The "Multicast Traffic and Topology Design Considerations" section on page 3-41 covers the Layer-2 related design. The multicast interaction with VSS in a Layer-3 domain includes the behavior of the multicast control plane in building the multicast tree, as well as the forwarding differences with ECMP and MEC-based topology. The "Routing with VSS" section on page 3-44 covers Layer-3 and multicast interaction.

VSS Failure Domain and Traffic Flow

This section uses the behavior of traffic flows described in the preceding section. The traffic flow during a failure in the VSS is dependent on local link availability as well the connectivity options from VSS to the core and access layers. The type of failure could be within the chassis, node, link, or line card. The VSS failures fall broadly into one of three domains:

- VSS member failure
- Core-to-VSS failure (including link, line card, or node)
- Access-layer-to-VSS failure (including link or line card)

This section uses the preferred connectivity method—MEC-based end-to-end—and its failure domains spanning core, distribution (VSS), and access layers. Some failure triggers multiple, yet distinct, recoveries at various layers. For example, a line card failure in the VSS might trigger an EtherChannel recovery at the core or access layer; however, it also triggers the VSL reroute at the distribution layer (VSS). Thus, upstream traffic and downstream traffic recovery could be asymmetric for a given failure. The types of recoveries that can be trigger in parallel are as follows:

- EtherChannel-based recovery
- ECMP or local CEF-based recovery
- Reroute over VSL (A failure that triggers traffic to be rerouted over the VSL bundle)

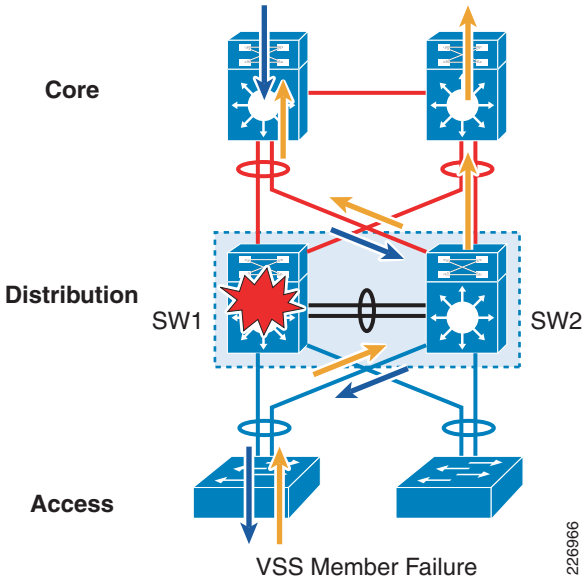
This section covers end-to-end traffic flow behavior with MEC-based topology. The convergence section (multilayer and routing) covers the ECMP-based topology and impact of failures in term packet loss.

VSS Member Failures

An EtherChannel failure can occur either at the nodes that are adjacent to the VSS or at the VSS itself. In both cases, recovery is based on hardware detecting that the link is down and then rehashing the flows from the failed member to a remaining member of the EtherChannel. Depending on the fault, it is possible that you can have only an EtherChannel failure (recovery) at the adjacent node and not at the VSS.

Figure 3-6 depicts the failure of the VSS node. The recovery is based on EtherChannel, as both core and access devices are connected to the VSS via MEC. The traffic in both directions (upstream and downstream) is hashed to the remaining member of EtherChannel at each layer and forwarded to the VSS switch. The VSS switch forwards the traffic in hardware, while the VSS control plane recovers if the failed VSS member was active. The VSS does this with the help of SSO; the switch has a hardware-based CEF that is capable of knowing the next hop and that the switch has a link directly connected to the adjacent node, If the failed member of the VSS is not an active switch, then recovery is simply based on EtherChannel detection at the core and access layers.

Figure 3-6 VSS Member Failure

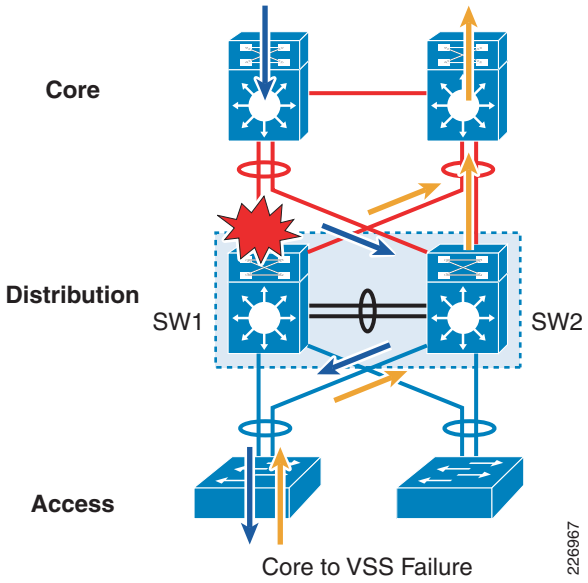


Core to VSS Failure

Figure 3-7 illustrates a failure of one link member of the port-channel between the VSS and the core router. The downstream traffic recovery is based on rehashing of the flow to the remaining member at the core router, which is a EtherChannel-based recovery. The upstream traffic recovery will be based on ECMP with local CEF switching which is triggered at VSS. In the design illustrated in Figure 3-7, there are two port-channel paths (one from each core router) that announce the two routes for the destinations that are used by upstream traffic flows. When a link member of the port-channel interface fails, from a VSS (single logical router) perspective, the available routing path may remain the same depending on routing protocol configuration. That is, from a VSS perspective we might still have two routes for the destinations, each route using each of the port-channels. However, from a member switch perspective local CEF switching is triggered and this means that the loss of a link within the port-channel represents reselection of the alternate path (as each switch in Figure 3-7 has two logical routed port-channel interfaces). This happens since one of the port-channels does not have any local link to that member switch. If the physical switch (SW1) has an alternate locally-attached link, that path will be used for packet forwarding and convergence will be based on local CEF adjacencies update on ECMP path. Otherwise, a member switch will reroute the traffic over a VSL link. As discussed previously, the recovery has a dependency on routing protocol configuration. Those dependencies and design choices associated with the VSS-to-core design are addressed in the "Routing with VSS" section on page 3-44.

The case in which all connectivity from one of the VSS members to the core layer fails (line card failures or both links being disabled will lead to no local path being available from one of the VSS members. This would force the traffic from the core to the VSS to traverse the VSL. Refer to the "Capacity Planning for the VSL Bundle" section on page 3-12.

Figure 3-7 Core to VSS Failure

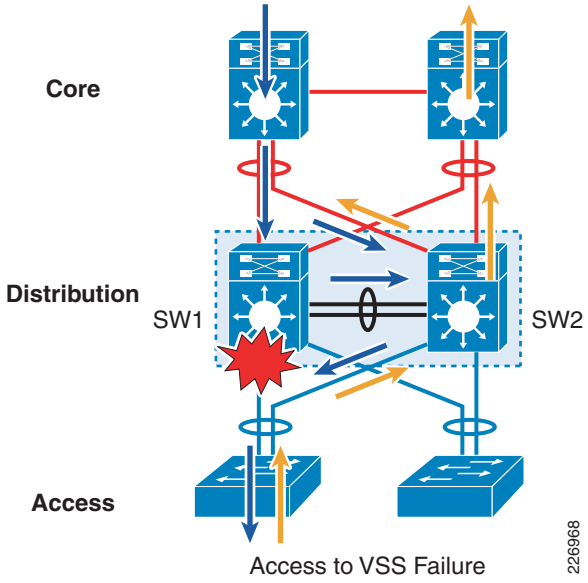


Access Layer-to-VSS Failure

Normally, a MEC-based topology avoids traffic over the VSL bundle. However, several fault scenarios can cause the traffic traverse the VSL bundle as a path of last resort. This is referred as *orphaned devices reroute*.

The entire connectivity to core or access layer failing introduce traffic re-route over VSL. In addition a link failure from an access-layer switch to VSS also introduces traffic reroute over VSL link. This failure is illustrated in Figure 3-8, in which the core routers have no knowledge of the link failure at the access layer. The core routers continue sending downstream traffic to specific the VSS member (SW1). The VSS control plane detects that the local link connected to the access-layer switch has failed; The VSS has knowledge that the Layer-2 MEC connection still has one member connected to SW2. The software at the VSS reprograms those flows such that traffic now goes over the VSL bundle to SW2—finally reaching the access-layer switch. The upstream traffic recovery is based on EtherChannel at the access layer.

Figure 3-8 Access to VSS Failure



In case all connectivity from one VSS member to an access-switch fails, downstream traffic recovery includes the VSL bundle reroute. The case of entire line card failure on VSS is covered in the "Capacity Planning for the VSL Bundle" section on page 3-12.

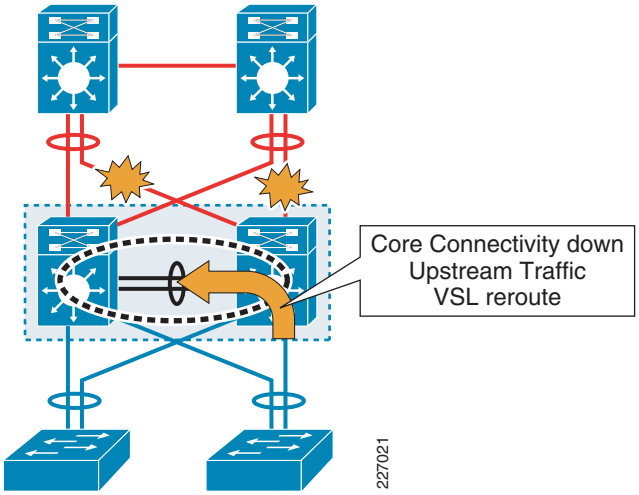
Capacity Planning for the VSL Bundle

In normal condition, the traffic load over the VSL bundle consist of network control-plane and inter-chassis control-plane traffic. In normal condition, both types of the traffic loads are very light and are sent with strict priority. Capacity planning and link sizing for VSS is almost identical to a traditional multilayer design in which the link(s) between two nodes should be able to carry traffic load equivalent of planned capacity during failure conditions.

Two failure points determine the minimum bandwidth requirements for VSL links:

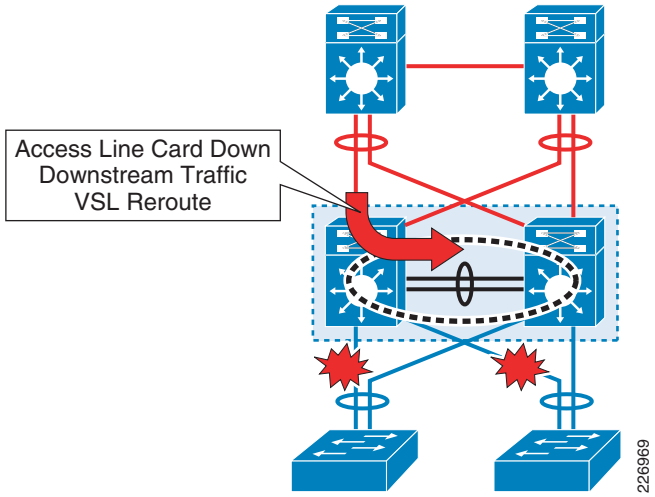
- Failure of all uplinks connected to a member of VSS to the core (Figure 3-9). In this failure, all upstream traffic traverses the VSL bundle. The number and speed of the uplinks limit the maximum traffic that can go over the VSL. Traditionally, in a full-mesh design, each switch member carries 20 Gigabits of bandwidth (two 10-Gigabit links); Thus, the minimum VSL bundle with two links (which is a resilient design) is sufficient.

Figure 3-9 Failure of All Uplinks to the Core



- Failure of all downstream link(s) to access-layer switches from one switch member (Figure 3-10). In this failure all downstream and the inter-access traffic traverses the VSL bundle. Traffic going toward the core is recovered via EtherChannel member at the access layer and need not traverse the VSL because access-layer links connected to a healthy VSS member whose connectivity to the core is intact. The bandwidth and connectivity requirements from the access switch vary by enterprise application need; true traffic capacity during failure is difficult to determine. However, all access-layer switches typically do not send traffic at the line rate at the same time, thus oversubscription for inter-access usually does not exceed the uplink capacity of the single VSS switch. The primary reason is that the traffic flow from the access-switch is typically higher in the direction of the core (WAN, Internet, or data center-oriented) than it is toward the inter-access layer.

Figure 3-10 Failure of a All Downstream Link to the Access-Layer



In both the cases, the normal traffic carrying capacity from each switch is determined by links connected from each switch to the core, because each switch can only forward traffic from locally connected interfaces. Thus, the minimum VSL bundle bandwidth should be at least equal to the uplinks connected to a single physical switch.

Additional capacity planning for VSL links is required due to following considerations:

- Designing the network with single-homed devices connectivity (no MEC) will force at least half of the downstream traffic to flow over the VSL link. This type of connectivity is highly discouraged.
- Remote SPAN from one switch member to other. The SPANed traffic is considered as a single flow, thus the traffic hashes only over a single VSL link that can lead to oversubscription of a particular link. The only way to improve the probability of distribution of traffic is to have an additional VSL link. Adding a link increases the chance of distributing the normal traffic that was hashed on the same link carrying the SPAN traffic, which may then be sent over a different link.
- If the VSS is carrying the services hardware, such as FWSM, WiSM, IDS, and so on, then all traffic that is intended to pass via the services blades may be carried over the VSL. Capacity planning for the services blades is beyond the scope of this design guide and thus not covered.

Multilayer Design Best Practices with VSS

The "VSS at the Distribution Block" section on page 1-3 explains the scope this design guide and summarizes the multilayer design most common in a campus network. The development of Cisco's highly available solution options for campus deployments has resulted in many design and tuning choices (and compromises). Among the key drivers are changing needs of the campus networks requiring support for voice over IP (VoIP) and the many emerging real-time transactional applications. As network designers' assess their own situations and make various deployment choices and compromises, the overall environment selection generally comes down to a choice of one of the two underlying models: looped and loop-free topologies. These models are described in the section that follows and will be used to illustrate the application of VSS at the distribution block in this design publication.

Multilayer Design Optimization and Limitation Overview

Figure 3-11 provides a comparison of a loop-free and a looped multilayer design.

Figure 3-11 Comparison of Looped and Loop-Free Multilayer Designs

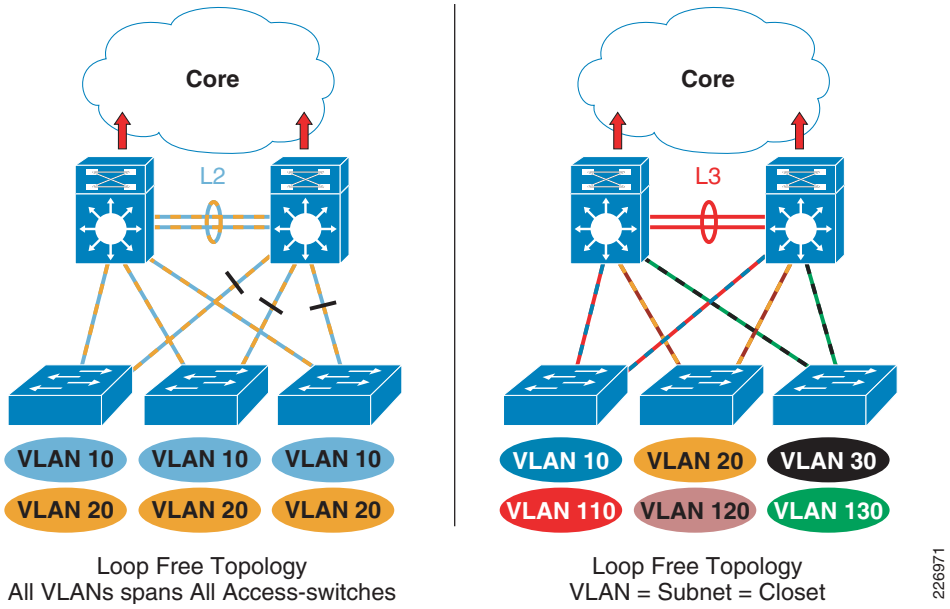
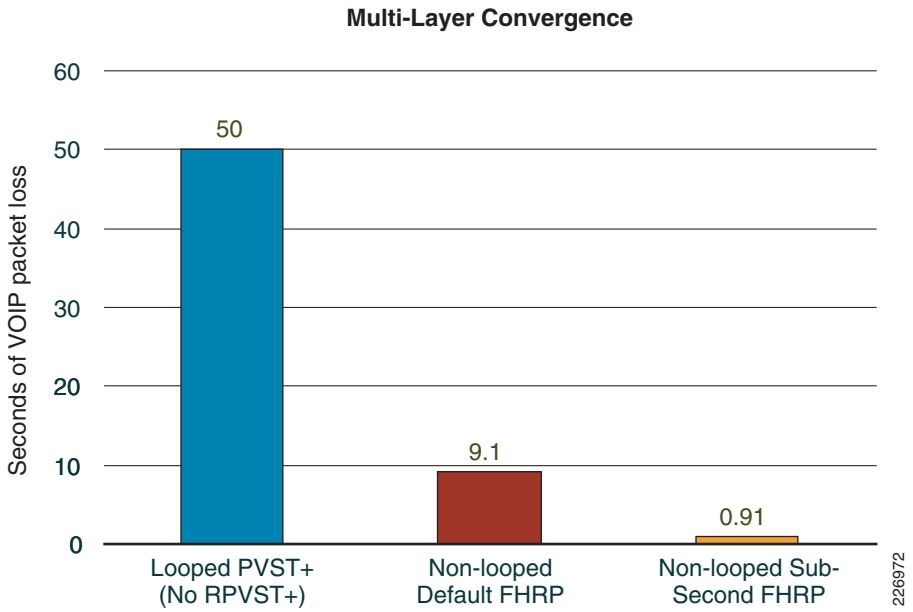


Table 3-1 summarizes the looped and non-looped design environments. Both designs use multiple control protocols, and the consistent application of tuning and configuration options to create a resilient design. In addition, the convergence described in Table 3-1 and illustrated in Figure 3-12 indicate that sub-second convergence requires First Hop Routing Protocol (FHRP), HSRP/GLBP/VRRP timer tuning, and a topology constraint that prevents VLANs to span multiple access switches. In addition, there are additional caveats and protocol behavior that requires further tuning. In either design, the sub-second convergence requires tightly-coupled design where all protocols and layers need to work together.

Table 3-1 Summary of Looped and Non-Looped Multilayer Designs

Looped Topology	Non-Looped Topology
At least some VLANs span multiple access switches	Each access switch has
Layer 2 loops	unique VLANs
Layers 2 and 3 running over link between distribution	No Layer 2 loops
Blocked links	Layer 3 link between distribution
	No blocked links
Application	
User application requires Layer-2 connectivity across access switch	Highly Available Application requirements—VoIP, Trading Floor
Adopting newer technologies solving new business challenges—NAC, Guest Wireless	Eliminate the exposure of loop
Flexibility in move add and change	Controlling convergence via HSRP
Efficient use of subnets	Reduced the side effect
Optimization Requirements	
HSRP and Root Matching	Basic STP Protection—BPDU Guard, Port-security
Load-sharing via manual STP topology maintenance	HSRP Timer Tuning
Unicast Flooding Mitigation—MAC and ARP Timers Tuning	Load-sharing via FHRP groups
Configuration tuning—Trunking, EtherChannel, etc	Trunk configuration tuning
STP—RPVST+ and MST	Layer-3 Summarization configuration
STP Toolkit—Root Guard, Loop Guard, BPDU Guard, Port-security	
Broadcast control	
STP Toolkit—Root Guard, Loop Guard, BPDU Guard, Port-security	
Broadcast control	
Convergence	
PVST – Up to 50 sec	FHRP Default—10 Sec
RPVST + FHRP (default timer)—10-to-11 Sec	FHRP Tuned Timer—900 msec
Other variations apply	Other variations apply

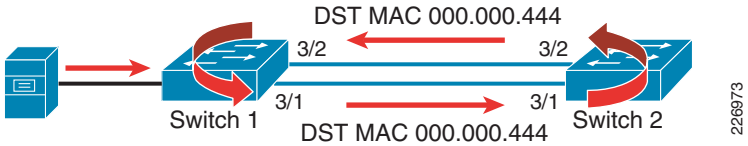
Figure 3-12 Multilayer Convergence Comparison



Loop Storm Condition with Spanning Tree Protocol

The Spanning Tree Protocol (STP) blocks alternate paths with the use of BPDU, thereby enabling a loop-free topology. However, it is possible that STP cannot determine which port to block and, as a result, will be unable to determine a loop-free topology. This problem is usually due to a missed or corrupted BPDU, as a result many devices go active (put the links in forwarding state) to find a loop-free path. If the loss of BPDU event is resolved, then the topology discovery process end; however, if the BPDUs loss continues, then there is no inherent mechanism to stop the condition in which BPDUs continuously circulating where each STP-enabled port tries to fine a loop-free path. See Figure 3-13.

Figure 3-13 General Example of Looping Condition



The only way to stop such a BPDU storm is to shut down network devices one-by-one and then bring the network back up by carefully reintroducing the devices one at a time. Looping can happen in both looped and non-looped topologies because a loop can be introduced by user activity at the access layer. However, the primary reason the looped design is more prone to this problem is that it has more logical paths available for a given VLAN topology.

The following issues can introduce a loop that STP might not be able to block:

- Faulty hardware (GBIC, cabling, CRC, etc) that causes a missed BPDU
- Faulty software that causes high CPU utilization and preventing BPDU processing
- Configuration mistake, for example a BPDU Filer on the forwarding link, causing a BPDU black hole

- Non-standard switch implementation (absorbing, but not sending the BPDU; or dropping the BPDU)
- User creating a topology via laptop networking that causes the BPDU to be missed

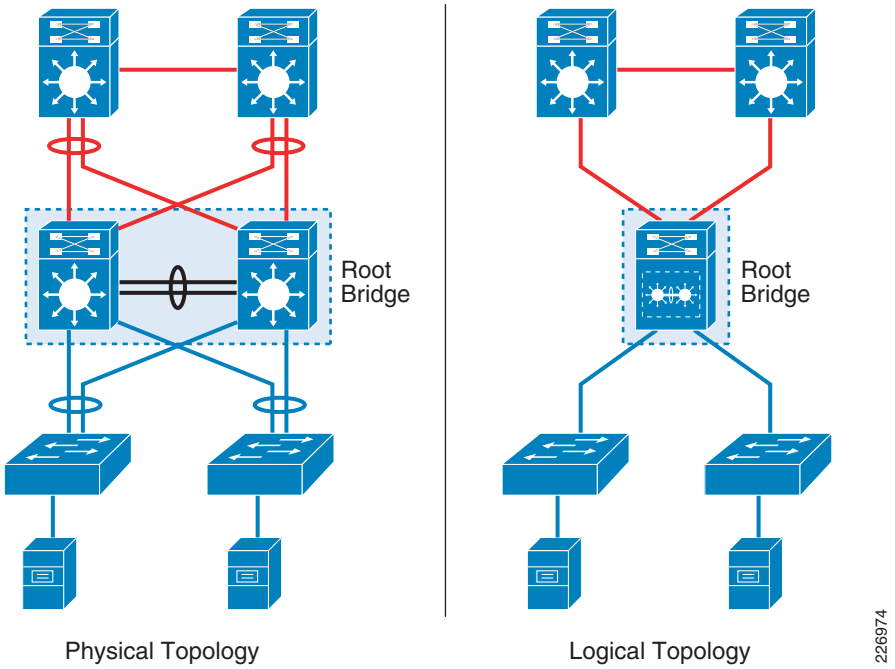
VSS Benefits

The VSS application at the distribution layer in a multilayer campus is designed to create a topology that has the following distinct advantages compared to traditional multi-layer designs:

- Loop-free topology with the use of MEC and unified control plane
- No configuration for default gateway (HSRP, GLBP, or VRRP) and no tuning requirement to achieve sub-second convergence
- Built-in optimization with traffic flow with EtherChannel
- Single-configuration management—consolidation of nodes
- Enables integration of services that requires Layer-2-based connectivity
- Sub-second convergence without the complexity of tuning and configuration

The VSS is applied at the distribution block with physical and logical topology is shown in Figure 3-14. As discussed in Chapter 2, "Virtual Switching System 1440 Architecture," the single logical node and MEC combined offers a star shape topology to STP that has no alternate path, thus a loop-free design is created that does not sacrifice the redundancy of a dual-link, dual-node design.

Figure 3-14 Physical and Logical Topologies



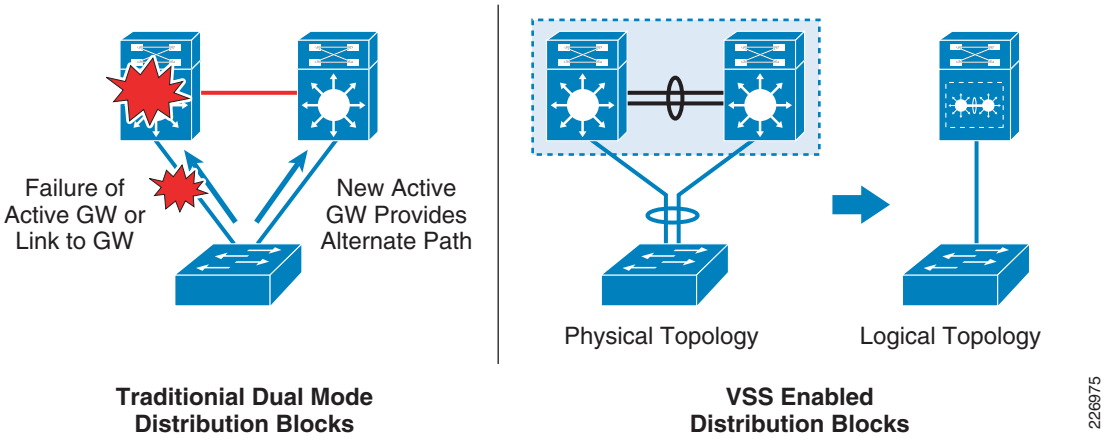
Elimination of FHRP Configuration

As suggested in Figure 3-14, a VSS topology replaces two logical nodes at the distribution layer. This topology eliminates the requirement of default gateway redundancy. This is because the default gateway is now replaced by a single logical node where the interface VLAN IP address is available in both the physical chassis. The convergence behavior of default gateway redundancy is replaced by SSO, as well as EtherChannel. Thus, none of the complexity of FHRP optimization and sub-second tuning is necessary or required.

The VSS appears as single resilient default gateway/first-hop address to end stations. In a non-VSS environment, FHRP protocols would serve as redundancy tools to protect against multiple failures-including distribution-switch or access-layer link failures. In that non-VSS topology, optimization of FHRP would be required to meet sub-second convergence requirements for Cisco Unified Communications. HSRP, GLBP, and VRRP configurations can be quiet complex if they are tuned to meet sub-second convergence as well load-sharing requirements. The optimization required to improve the convergence would include the following:

- Sub-second timer configuration of FHRP Hello
- Preemptive and standby delay configuration
- Dependency on STP convergence in a looped topology
- Platform dependency and CPU capacity of handling sub-second timer for FHRP

Figure 3-15 Elimination of FHRP as a Default Gateway Redundancy



Furthermore, to optimize the load-share of upstream traffic with FHRP would also require the following:

- Multiple HSRP groups defined at each distribution node and the coordination of active and secondary FHRP by even distribution over two routers
- Use of GLBP facilitates automatic uplink load-balancing (less optimal in looped topologies due to alternate MAC address allocation for default gateway)

All of the above required configuration complexities are eliminated by the implementation of the VSS in campus. A practical challenge arises with the elimination of HSRP/GLBP used as a default gateway redundancy. The MAC address of the default gateway IP address is unique and consistent with HSRP/GLBP. In VSS-enabled campus the VLAN interface IP becomes the default gateway. The default gateway IP address remains the same (in order to avoid changes to the end hosts) and is typically carried over to VLAN interface. The VLAN interface MAC address is not the same as HSRP or GLBP MAC address. The VLAN interface MAC is a system generated address. (Refer to "MAC Addresses" section on page 2-44 for more details). Typically, the gratuitous ARP is issued while the IP address in

unchanged, but the MAC address is modified. This change of MAC address can cause disruption of traffic if the end host is not capable or has configuration that prohibits the update of the default gateway ARP entry. End hosts typically cash ARP table entry for the default gateway for four hours.

One possible solution for this problem is to carry HSRP/GLBP configuration to VSS without any neighbor. Keeping the configuration of HSRP/GLBP just to avoid the default gateway MAC address update is not a ideal best practices. This is because the default gateway recovery during the active switch failure is dependent on how fast HSRP/GLBP can be initialized during SSO-based recovery. Thus, one possible alternative is to use default gateway IP address on VLAN interface and temporarily configure a HSRP/GLBP configuration with same group ID as shown below.

```
Interface Vlan200
 ip address 10.200.0.1 255.255.255.0 <-- old HSRP IP
 standby 200 ip 10.200.0.2 <--- old HSRP group id#, but new IP address
```

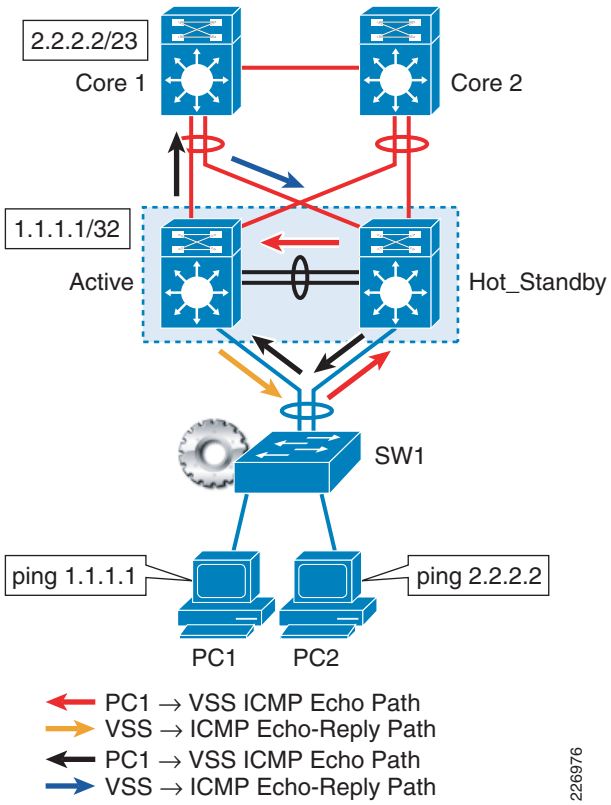
The above configuration would allow the Vlan200 SVI to take ownership of the HSRP group 200 MAC address, while not creating any dependency on HSRP group because it will not link it to the default gateway IP address. After the transition to VSS, hosts will continue sending frames to the HSRP MAC address. As time progresses, packet will enter the VSS destined for these hosts, causing the VSS to send an ARP request for the interesting host. This ARP request will fix the host's ARP entry for the default gateway IP address, causing it to point to the new MAC address. Even for the host for which there is no traffic from VSS so it can trigger the ARP update, it will refresh its ARP table within the next four hours, causing it to then pick up the new IP address of the VSS.

After about four hours have progressed, you can safely remove the HSRP configuration from all SVI's as most likely no hosts are still using the old MAC address. This time can be extended for safety, or the customer can come up with a script that will check the ARP tables of each server before removing HSRP/GLBP configuration.

Traffic Flow to Default Gateway

Figure 3-16 illustrates the flow of a ping from an end host through the default gateway. The upstream path for ICMP traffic is chosen at the access-layer switch based on the hashing decision. If the hash results in the selection of the link connected to the hot-standby switch, the packet traverses the VSL link to reach the active switch for a response. The response from the VSS always takes a local link because the VSS always prefers the local path for user data traffic forwarding. If the ping originates from the VSS, then the VSS first chooses a local path then the responder may chose either link. A ping or data traffic traversing the VSS follows normal forwarding as described in the "Traffic Flow in the VSS-Enabled Campus" section on page 3-5.

Figure 3-16 ICMP Echo-Response Traffic Flow



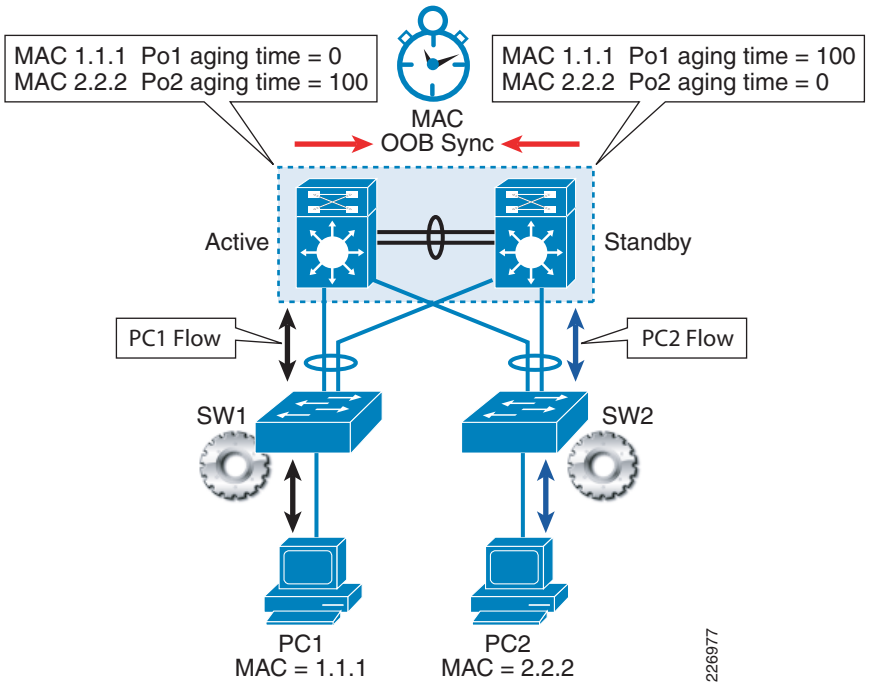
Layer-2 MAC Learning in the VSS with MEC Topology

As in a standalone implementation, a VSS switch member independently employs hardware-based, source MAC-address learning. VSS is also capable of multi-chassis distributed forwarding. In distributed switching, each Distributed Feature Card (DFC) maintains its own Content-Addressable Memory (CAM) table. This means that each DFC learns the MAC addresses and ages them based on the CAM-aging and traffic matching of that particular entry. VSS follows the same timers for maintaining active and aging (idle) timers as with a standalone implementation. Dynamic MAC address entries in the forwarding table have following modes:

- *Active*—A switch considers dynamic MAC entry an *active* entry when a switch is actively switching traffic in the network from the same source MAC address. A switch resets the aging timer to 0 seconds each time it receives traffic from a specific source MAC address.
- *Idle or Aging*—This MAC entry is stored in the forwarding table, but no active flow is present for that MAC. An Idle MAC entry is removed from Layer-2 forwarding table after 300 seconds by default. With distributed switching, it is normal that the supervisor engine does not see any traffic for a particular MAC address for a while, so the entry can expire. There are currently two mechanisms available to keep the CAM tables consistent between the different forwarding engines: DFC, which is present in line modules; and, PFC, which is present in supervisor modules.
- *Flood-to-Frame (FF)*—This is a hardware-based learning method that is triggered every time a new MAC address is presented to the line cards. The MAC address is added to the forwarding table in distributed manner. The line card/port on which the MAC address is first learned is called *primary-entry* or *source-line-card*.

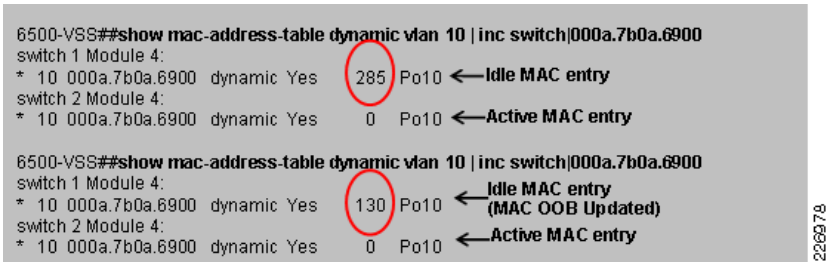
- *MAC Notification (MN)*—This is a hardware-based method for adding or removing a MAC address on a non-primary-entry line card in order to avoid continuous unicast flooding within the system. If the traffic for a MAC destination is presented at the DFC line card level, it will first flood that frame to the entire system because it does not have information about the location of that MAC address in the system. As soon as the primary-entry line card receives the flooded frame, it sends +MN to add a MAC entry on the DFC line card from which this traffic came. A -MN is used to remove an entry from the DFC line card, if it has aged out. See Figure 3-17.

Figure 3-17 MAC Notification



- *MAC Out-of-Band Sync (OOB)*—In a normal condition, the traffic enters and leaves the VSS on a per-chassis basis (as describe in "Traffic Flow in the VSS-Enabled Campus" section on page 3-5). This means that the typical flow will only refresh the MAC entries in a single chassis. Figure 3-17 illustrates that PC1 flow is selecting SW1 based on EtherChannel hashing at the access layer. The PC1 MAC entry will start aging on SW2. Similarly PC2 MAC entry will age out on SW1. Once the idle time is reached, the MAC address is aged out on the non-primary line cards, as well as the peer chassis PFC and its line cards. If the traffic is presented to such a line card, it will have to be flooded to the entire system. In addition, the MEC (being the essential component of VSS) might possibly be operating in distributed EtherChannel mode which would increase the probability of the MAC address being aged out at various line cards. In order to prevent the age out of an entry on a DFC or PFC, the MAC OOB software process mechanism periodically updates the active MAC entry in all line cards and PFC, even if there is no traffic for that MAC address. MAC OOB is designed to prevent an active MAC entry from aging out anywhere in the VSS (as well as standalone system). Only the primary entry module will synchronize the active MAC entries. Idle MAC entries do not get synchronized and are aged out independently. Figure 3-18 show the CLI needed to illustrate the MAC aging and MAC OOB updated entries.. As shown in first CLI output, SW2 module 4 has the active MAC entry as its aging time is zero. Since the flow is hashed to SW2, the same MAC entry on SW1 start aging out as shown in the output in Figure 3-18 where the MAC is aging toward 480 second default timer. The second CLI output is taken after OOB process has synchronized the MAC entry in which the MAC entry timer on SW1 module 4 has reset. Without OOB, the MAC entry on SW1 module 4 would have aged out, potentially causing temporary unicast flooding.

Figure 3-18 MAC OOB Synchronization



Note The MAC synchronization process between virtual-switch nodes is done over the VSL EtherChannel and particularly over the VSL control link.

Out-Of-Band Synchronization Configuration Recommendation

The following CLI is used to enable OOB. The default MAC OOB interval is 160 sec. MAC OOB synchronization is programmed to update active MAC entry's aging-time across all modules at three activity intervals. The idle MAC aging-timer must be set to 480 seconds (MAC OOB interval times three activity intervals).

```
VSS(config)# mac-address-table synchronize activity-time ?
  <0-1275>  Enter time in seconds <160, 320, 640>
% Current activity time is [160] seconds
% Recommended aging time for all vlans is at least three times the activity interval

6500-VSS# show mac-address-table synchronize statistics
MAC Entry Out-of-band Synchronization Feature Statistics:
-----------------------------------------------------------
Switch [1] Module [4]
---------------------
Module Status:
Statistics collected from Switch/Module   : 1/4
Number of L2 asics in this module         : 1

Global Status:
Status of feature enabled on the switch   : on
Default activity time                     : 160
Configured current activity time          : 480
```

The MAC OOB synchronization activity interval settings are applied on a system-wide basis. However, each module independently maintains its own individual aging.

Caution Prior to Cisco IOS Release 12.2(33)SXI, the default idle MAC aging-timer on the RP depicts an incorrect aging of 300 seconds when the MAC OOB synchronization is enabled in Cisco Catalyst 6500 system; however, the SP and DFC modules show the correct value of 480 seconds. Software bug (CSCso59288) resolved this issue in later releases.

Note If WS-6708-10G is present in the VSS system, MAC synchronization is enabled automatically; if not, MAC synchronization must be enabled manually.

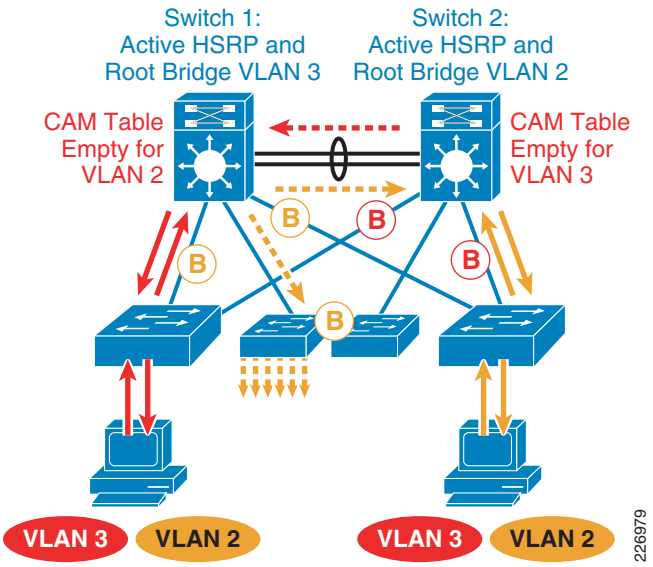
Note By default, the dynamic MAC entry aging-time on the Cisco Catalyst 6500 system with the 6708 module is set to 480 seconds. The MAC aging-timer must be changed from 300 seconds to 480 seconds manually if Cisco Catalyst 6500 with a non-6708 DFC module present in the switch. Starting in IOS Release 12.2(33) SXI, default idle MAC aging-timer is automatically set to 480 seconds.

Tip **Cisco recommends that you enable and keep the default MAC OOB synchronization activity interval of 160 seconds (lowest configurable value) and idle MAC aging-timer of three times the default MAC OOB synchronization activity interval (480 seconds).**

Elimination of Asymmetric Forwarding and Unicast Flooding

Unknown unicast flooding occurs when the upstream and downstream flows has asymmetrical forwarding path. The asymmetric forwarding paths are created in a standalone design, where upstream traffic for a given source MAC always goes to a default gateway; however, the downstream traffic is load-share by core-layer routers reaching both distribution layer gateways. At the start when the source MAC send a first ARP discovery for the default gateway, both the distribution router learns the MAC and ARP-to-MAC mapping is created. Timer for CAM entries expires at five minutes while APR entries at four hours. Since the upstream traffic is directed only at the one of the distribution node, CAM timer for that MAC expires while ARP entries remains at the *standby* distribution router. When a standby router receives the traffic destined for that MAC address, the ARP entry provides the Layer-2 encapsulation, but a corresponding CAM entry does not exist in the CAM table. For any Layer-2 device, this traffic is known as unknown *unicast*, which has to be flooded to all the ports in that VLAN. This problem is illustrated in Figure 3-19 in which two distribution routers have an empty CAM table for the corresponding VLANs. If VLAN 3 and VLAN 2 devices communicate with each other, they do it via the default gateway. For each respective default gateway, this traffic is treated as unknown unicast. In non-looped topology, only one frame is flooded over the link between distribution routers; however, for looped topology, this unknown frame has to be sent to all access-layer switches where interested VLAN exists. If the type of flow is high-volume (FTP, video, etc), this can overwhelm the end devices, leading to extremely poor response time for the critical application. For many networks, this symptom exists but is unknown to the network operation since there are no indication at the network level.

Figure 3-19 Empty CAM Table Example

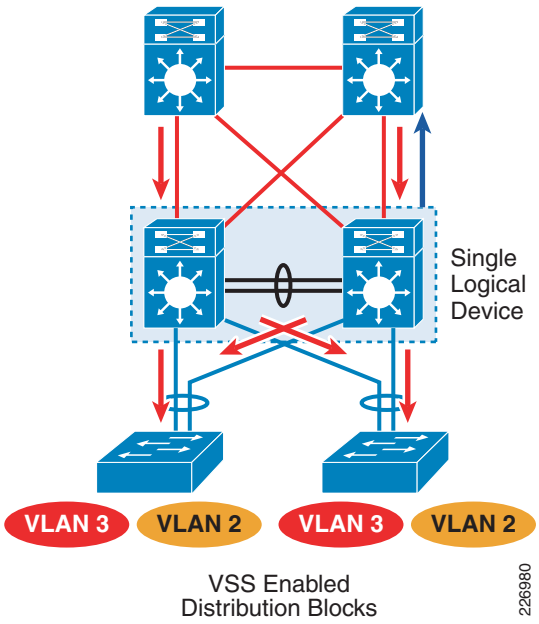


Unicast flooding is more pronounced with VLANs that span multiple access-layer switches. There is no inherent method to stop flooding of the traffic to every port that VLAN traffic exists. In general, there are three methods to reduce the effect of unicast flooding:

- Use a non-looped topology with unique voice and data VLANs per-access switch. Unicast flooding still can occur, but it imposes less user impact because the VLANs are localized.
- Tune the ARP timer to 270 seconds and leave the CAM timer to the default of five minutes. This way, the ARP always times out ahead of the CAM timer and refreshes the CAM entries for traffic of interest. For networks that contain ARP entries in excess of 10,000, you choose to increase the CAM timer to match the default ARP timer to reduce CPU spiking that results from an ARP broadcast.
- Both of the preceding approaches force the selection of a topology or additional configurations. The VSS has built in mechanism to avoid unicast flooding associated with CAM-time outs that does not impose such restrictions. VSS enables a single logical router topology to avoid flooding in a dual-homed topology as shown in Figure 3-20. To avoid unicast flooding, both member switches continuously keep the ARP table synchronized via SSO because ARP is SSO-aware. For the synchronization of MAC addresses on both members of the VSS, the VSS uses three distinct methods:
  - *Flood to frame*—explicit source learning in hardware
  - *MAC notification*— +MN and –MN update of MAC entries to DFC line cards
  - *Out-of-band sync*—globally synchronizes MAC addresses every 160 seconds

These methods are described in the "Layer-2 MAC Learning in the VSS with MEC Topology" section on page 3-20. The unicast flooding is prevented because MAC-address reachability is possible via both the member and the MEC topology enable optimal local-link forwarding. Temporary flooding between switches might occur that will allow relearning of MAC addresses, but this does not affect user application performance.

Figure 3-20 VSS-Enabled Distribution Blocks



Multilayer-Design, Best-Practice Tuning

The deployment of the VSS at the distribution in campus changes some of the traditional configuration best practices. You should be aware that this design guide does not evaluate all the possible configuration options available or prevalent for multilayer campus design. The changes to any best-practice recommendations for a campus multilayer configuration are included in this guide within the context of high-availability network design and VSS deployment.

Trunking Configuration Best Practices

In a traditional multilayer design featuring standalone switches, when Dynamic Trunking Protocol (DTP) and 802.1Q or Inter-Switch Link (ISL) negotiation are enabled, considerable time can be spent negotiating trunk settings when a node or interface is restored. During negotiation, traffic is dropped because the link is operational from a Layer-2 perspective. Up to two seconds can be lost depending on where the trunk interface is being brought up. However, in this configuration, DTP is not actively monitoring the state of the trunk and a misconfigured trunk is not easily identified. There is a balance between fast convergence and your ability to manage your configuration and change control.

In VSS, trunk mode of a port-channel interface being either desirable or undesirable does not exhibit the behavior of standalone node. In VSS, each access-layer is connected via port-channel (MEC), where a link member when brought on line is not a separate negotiation; rather it is an addition to EtherChannel group. The node-related restoration losses are also not an issue when compared to a standalone dual-node design in which each node has a separate control plane that negotiates a separate trunking event. As with VSS, when the node is restored, the link-up event is an additional member link of the MEC and not a trunk interface. See Figure 3-21.

Figure 3-21 Convergence Loss Comparison

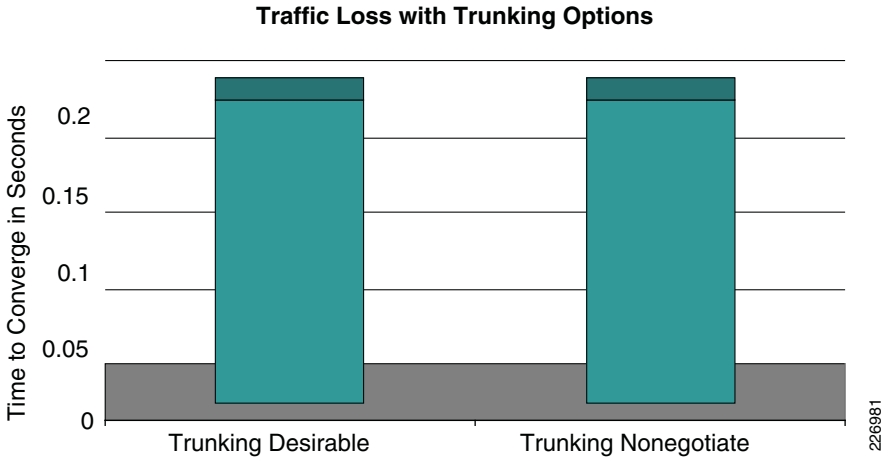


Figure 3-21 compares the convergence losses associated with trunk mode being desirable or nonnegotiable (non-desirable). With either configuration, the losses are less or equal to 200 msec. An additional benefit of running a trunk mode as desirable is that operational efficiency is gained in managing trunked interfaces. VSS enables the ability to span multiple VLANs to multiple access-layer switches. This means that more VLANs are trunked, which increases the possibility of errors occurring during change management that in turn can disrupt the VSS domain. The desirable option reduces the black-holing of traffic because trunking will not be formed if the configuration is mismatched and this option setting causes the generation of syslogs messages in certain mismatch conditions, providing a diagnostic indication of faults to operational staff.

Tip **Cisco recommends that trunks at both end of the interfaces be configured using the desirable-desirable or auto-desirable option in a VSS-enabled design.**

VLAN Configuration Over the Trunk

In a VSS-enabled Layer-2 design in which there are no loops, it is quiet intuitive to allow VLANs proliferation and access to VLAN from any access-layer switches. It is generally accepted best practice to restrict uncontrolled VLAN growth and access policy. The **switchport trunk allowed vlan** command on a trunked port-channel should be used to restrict VLANs to be seen and forwarded to desired switches. This will reduce exposure during moves, adds, and changes and adds clarity when troubleshooting large VLAN domains.

An additional benefit of restricting VLANs on a per-trunk basis is optimizing the use of the STP logical port capability per line card and switch. The STP logical port capacity is determined by how well CPU can handle number of STP BPDU per-VLAN per-physical port be send out during the topology change. The number of logical STP-enabled port capacity is determined by the line card and overall system capability for a given STP domain. These limits are described in the Release Note at the following URL:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/release/notes/ol_14271.html#wp26366

From the VSS's perspective, logical port limits apply per-system, not per-individual chassis because there is only one control plane managing both chassis interfaces. The maximum number of STP logical ports per line cards is dependent on type of line card used. The only types of line cards supported in VSS are the WS-X67xx Series, which means maximum of 1800 logical ports can be configured, but this limit is removed with Cisco IOS Release 122.(33)SXI1. There are two ways to prevent exceeding the STP logical limit per line card:

- Limit the number VLANs allowed per trunk
- Distribute access-layer connectivity over multiple line card on a VSS

The following CLI illustrates how you can determine whether the VSS system is exceeding the STP logical limit per line card:

```
6500-VSS# sh vlan virtual-port switch 1 slot 8
Slot 8 switch : 1
Port        Virtual-ports
--------------------------
Gi1/8/1             8
Gi1/8/2             8
Gi1/8/3             207
Gi1/8/4             207
Gi1/8/5             207
Gi1/8/6             207
Gi1/8/7             207
Gi1/8/8             207
Gi1/8/9             207
Gi1/8/10            207
Gi1/8/11            207
Gi1/8/12            207
Gi1/8/13            207
Gi1/8/14            207
Gi1/8/15            207
Gi1/8/16            7
Gi1/8/17            9
Gi1/8/19            1
Gi1/8/21            9
Gi1/8/23            9
Gi1/8/24            9
Total virtual ports:2751
```

In the preceding output illustration, just over 200 VLANs are allowed on a number of ports which combine to exceed the STP logical port limits. The following **show** command output shows how this number is calculated.

```
6500-VSS# sh int po 221 trunk

Port    Mode         Encapsulation  Status        Native vlan
Po221   desirable    802.1q         trunking      221

Port    Vlans allowed on trunk
Po221   21,121,400,450,500,550,600,900 <-

Port    Vlans allowed and active in management domain
Po221   21,121,400,450,500,550,600,900

Port    Vlans in spanning tree forwarding state and not pruned
Po221   21,121,400,450,500,550,600,900
```

The number of VLAN instances allowed on a trunk equals the logical STP ports consumed. This number is eight for the preceding output (21,121,400,450,500,550,600, and 900 yields eight logical ports).

Now consider an unrestricted port as illustrated in the following output example:

```
6500-VSS# sh int po 222 trunk

Port    Mode         Encapsulation  Status        Native vlan
Po222   desirable    802.1q         trunking      222

Port    Vlans allowed on trunk
Po222   1-4094

Port    Vlans allowed and active in management domain
Po222   1-7,20-79,102-107,120-179,202-207,220-279,400,450,500,550,600,650,900,999 <-

Port    Vlans in spanning tree forwarding state and not pruned
Po222   1-7,20-79,102-107,120-179,202-207,220-279,400,450,500,550,600,650,900,999
```

In the case of the unrestricted port, all VLANs are allowed, which dramatically increased the STP logical port count to 207. (1-7, 20-79,102-107,120-179,202-207,220-279,400,450,500,550,600,650,900, and 999 yields 207 logical ports.)

The total VSS system's STP logical port count can be viewed via the **show vlan virtual-port** command. The system-specific limit calculation is somewhat misrepresented in the VSS CLI output because the command originally was intended for standalone systems. The total logical count for STP ports for a VSS is counted by adding each switch count—even though STP runs on EtherChannel ports—and thus only half of the ports should be counted toward the STP logical port limit. See VSS Release Notes at the following URL:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/release/notes/ol_14271.html#wp26366

Tip **Cisco recommends explicit configuration of required VLANs to be forwarded over the trunk.**

Unidirectional Link Detection (UDLD)

The normal mode UDLD is used for detecting and error-disabling the port to avoid loop broadcast storm triggered by cabling mismatch. The aggressive UDLD is an enhanced form of normal UDLD, traditionally used for detecting a link integrity and faulty hardware. UDLD protocol is used to detect the problem with STP loop, far in advanced for PVST environment where the STP convergence could take up to 50 seconds. The application of an aggressive UDLD as a tool for detecting a looped condition (due to faulty hardware) is limited in VSS-enabled Layer-2 domain, because the VSS is inherently loop-free topology. In addition, typical convergence associated with new STP protocols (RPVST+ and MST) is much faster than aggressive UDLD detection time. The side effect of aggressive UDLD detection is far more impacting than it effectiveness in VSS environment. In VSS (unified control plane), many critical processes requires processing by CPU. Typically, DFC-based line card take longer to initialize. A faulty software may occupy CPU resources such that the aggressive UDLD process does not get a chance to process the hello. These conditions can lead to a false-positive where aggressive UDLD is forced to act in which it will error-disabled connection on the both sides.

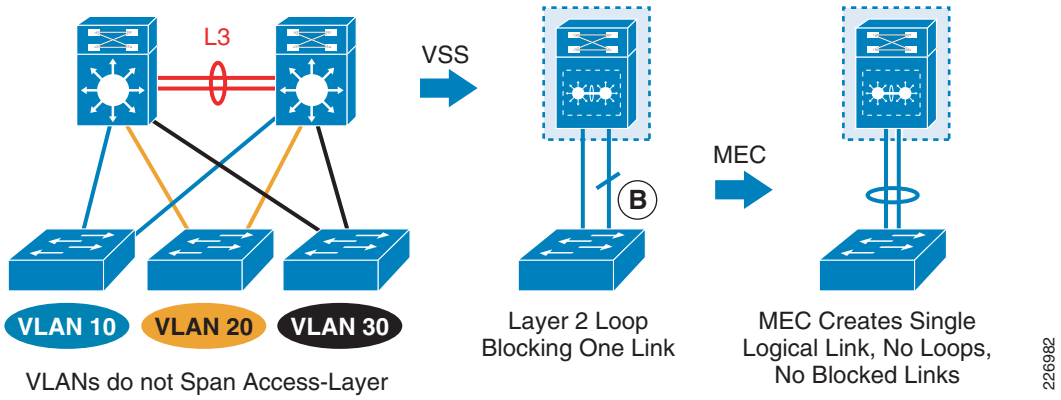
Tip **The aggressive UDLD should *not* be used as link-integrity check, instead use normal mode of UDLD to detect cabling faults and also for the link integrity.**

Topology Considerations with VSS

The impact on the campus topology with the introduction of the VSS (single logical devices) is significant. The devices connected to the VSS (with or without MEC) also play critical roles. Layer-2 and Layer-3 interaction with a given topology determines the topology behavior in a fault condition and thus determines the convergence of user data traffic. This section covers Layer-2 domain, while the "Routing with VSS" section on page 3-44 covers the Layer-3 domain.

Traditionally, many networks have adopted a optimized multilayer topology (V- or U-shaped) with which VLANs do not span closets. Deploying a VSS in such topology without MEC reintroduces STP loops into the networks as shown in Figure 3-22. Use of an MEC is required whenever two Layer-2 links from the same device connect to the VSS. Figure 3-22 illustrates the behavior of VSS-enabled non-looped "V" shape topology with and without MEC.

Figure 3-22 Non-looped Topology Behavior



A daisy-chained access switch topology featuring indirect connectivity presents the following two designs challenges:

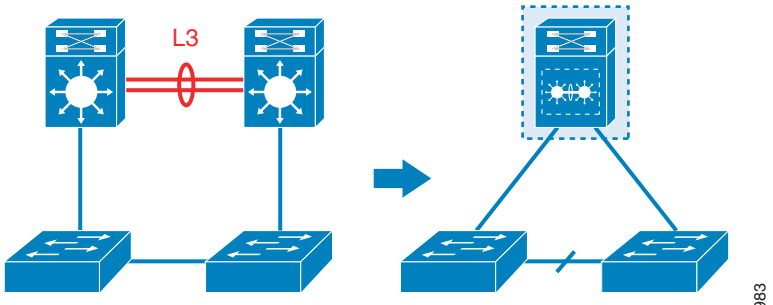
- Unicast flooding
- Looping (blocked link)

The use of a virtual switch in the distribution layer addresses the problem of unicast flooding; however, the network still has a Layer-2 loop in the design with an STP-blocked link. Traffic recovery times are determined by spanning tree recovery in the event of link or node failures.

A U-shaped topology with the VSS is shown in Figure 3-23. It has two undesirable effects:

- It will create a topology with a loop.
- 50 percent of downstream traffic will traverse VSL link if the STP topology is formed such that the uplink connected to the VSS is blocked.

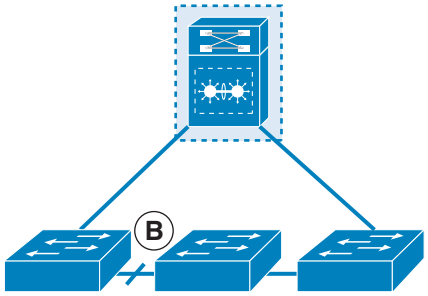
Figure 3-23 U-Shaped VSS Topology Loop Introduction



U Shape Topologies Introduces Loop with VSS

The daisy-chained access topology in which VSS cannot detect indirect connections introduces a Layer-2 loop with an STP blocked link. See Figure 3-24.

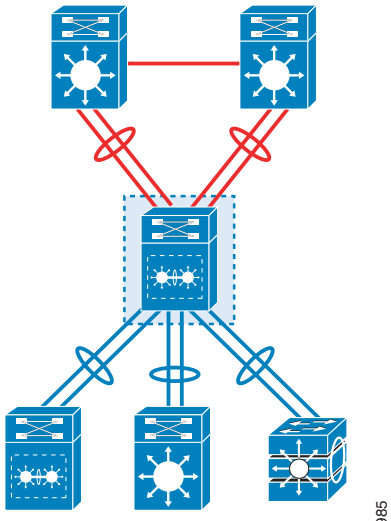
Figure 3-24 STP-blocked Link



Layer 2 Loop Is One Switch Smaller but Still Exists

For either daisy-chained or U-shape topologies, two solutions exist. Either deploy MEC from each switch to avoid making an indirect connection to the access-layer or use a cross-stacked EtherChannel capable switches (Catalyst 37xx stacks) at the access-layer. See Figure 3-25.

Figure 3-25 Star Topology



Star Topology avoids STP Loops

Tip **Cisco recommends that you always use a star-shaped topology with MEC (Layer-2 and Layer-3) from each device connected to the VSS to avoid loops and have the best convergence with either link or node failures.**

Spanning Tree Configuration Best Practices with VSS

Caution One of the benefits of VSS-based design is that it allows the STP be active in the entire Layer-2 domain. The VSS simply offers a loop-free topology to STP. There is no inherent method to offer a topology that is loop-free, unless the topology created by a network designer is star-shaped (MEC-enabled). Spanning tree must be enabled in the VSS-enabled design in order to detect accidental loops created at the access-layer or within the VSS systems (by connecting the same cable back to back to VSS member chassis). In a non-VSS-enabled network, a set of spanning tree tools are available to protect the network from looping storms or reducing the effects of storms so that corrective actions can be taken. For further information about loop storm condition protection in a non-VSS multilayer design, refer to the following URL:
http://www.cisco.com/en/US/products/hw/switches/ps700/products_white_paper09186a00801b49a4.shtml#cg5.

This design guide does not go into the detail about STP tools to reduce the effects of broadcast storms because there are no loops in a VSS-enabled network. However, the following STP-related factors should be considered in the VSS-enabled campus:

- STP Selection, page 3-31
- Root Switch and Root Guard Protection, page 3-32
- Loop Guard, page 3-32
- PortFast on Trunks, page 3-32
- PortFast and BPDU Guard, page 3-35
- BPDU Filter, page 3-36

These are discussed briefly in the following sections.

STP Selection

The selection of a specific STP protocol implementation—Rapid per VLAN Spanning Tree Plus (RPVST+) or Multiple Instance Spanning Tree (MST)—is entirely based on customer-design requirements. For average enterprise campus networks, RPVST+ is the most prevalent and appropriate unless interoperability with non-Cisco switches is required. The additional dependencies and requirements of MST must be evaluated against its advantages of capability to support extremely large number of VLANs and logical ports. However, for majority of campus networks, the 12000-logical port capacity of RPVST+ is sufficient. Refer to the following Release Note URL for VSS:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/release/notes/ol_14271.html#wp26366

Root Switch and Root Guard Protection

The root of the STP should always be the VSS. Use a statically-defined, hard-coded value for the spanning tree root so that no other switches in the network can claim the root for a given spanning tree domain. Use either Root Guard on a link of VSS-facing access-layer switch or enable it at access-layer switch user port (although the later does not prevent someone from replacing access-layer switch with another switch that can take over as root). The root change might not affect forwarding in non-looped designs (root selection matter only when alternate path (loop) is presented to STP); however, the loss of BPDU or inconstancies generated by a non-compliant switch becoming root could lead to instability in the network.

By default, the active switch's base MAC address is used as the root address of the VSS. This root address does on change during SSO switchover so that an access-layer switch does see the root change. For more details, see the "STP Operation with VSS" section on page 3-36.

Loop Guard

In a typical customer network, CPU utilization, faulty hardware, configuration error or cabling problem leads to the absence of BPDUs. This condition causes alternate ports to enter forwarding mode, triggering a looping storm. The BPDU Loop Guard prevents such condition within six seconds (missing three consecutive BPDUs). A Loop Guard normally operates on alternate ports and only works on STP-enabled port. The VSS-enabled with MEC design does not offer a looped topology to STP protocol. As a result, Loop Guard might not be a particularly useful feature in the VSS-enabled network because all ports are forwarding and none are blocking.

If Loop Guard is enabled on both sides of a trunk interface and if the loss of a BPDU occurs, the access-switch's EtherChannel port (where STP is running) state will transition to root-inconsistent. Disabling of the entire EtherChannel is not a desirable outcome to detect a soft error in the design where loop does not exists.

If the Loop Guard is not enabled and the loss of BPDU is observed, the access-layer switch will become the root for VLANs defined in it local database. The user traffic might continue, but after a few seconds either the UDLD or PAgP timer will detect the problem and will error-disable the port.

Use normal mode UDLD or normal hello method of PAgP/LACP to detect the soft errors. UDLD and PAgP/LACP only disable individual link members of the EtherChannel and keeps the access-switch connectivity active. In addition, advances in Cisco IOS Release 12.2(33) SXI incorporate better methods to solve this type of issue. Refer to the following link for more information:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/spantree.html#wp1098785

Tip **Cisco recommends that you do *not* enable Loop Guard in a VSS-enabled campus network.**

PortFast on Trunks

PortFast implementation to trunks immediately places VLANs into the forwarding state without going through listening and learning phases of STP. However, the port-fast state of a trunk becomes a regular STP port as soon as it sees a BPDU from the remote side of a connection, thus its major benefits is to accelerate the forwarding state of STP during initialization.

In traditional multilayer looped networks, the use of the port-fast feature on trunks can lead to temporary loops in highly meshed topologies because it might takes longer to block or discover the alternate paths. Due to this risk, its application has been limited. In the VSS-enabled design, the use of the port-fast capability on trunks is safe because VSS topologies are inherently loop free, thereby eliminating the possibility of temporary loops being created by port-fast feature on a trunk.

With a dual-node design (non-VSS) for an access-layer switch, each interface connects to separate node at the distribution layer. Each failure or initialization of trunk occurs independently and interfaces are not ready to forward traffic (packet losses) either due to the state of STP or trunk negotiations. VSS eliminates this delay because the access-layer is connected with a port-channel where STP operates. Adding another interface effectively adds another EtherChannel member; STP and the trunk state need not be negotiated.

From the following syslogs output examples, you can see that the reduction in initial delay is up to one second when port fast is enabled on a trunk.

Note The impact of this delay on user data traffic is difficult to quantify. Tools cannot accurately time stamp when an interface is initialized and how much data is lost in the interval between restart (**no shutdown** command) and full forwarding. Additionally, a tool must send data before restarting an interface. There is no way to determine the difference between the data sent prior to the interface initialization event. Crude experimentation indicates a connectivity disruption of up to 600 msec without PortFast enabled on the trunk.

PortFast Disabled on the Trunk

The following CLI examples illustrate output showing PortFast as being disables for a trunk.

VSS Syslogs

```
6500-VSS# sh log | inc 106
Oct 22 14:03:31.647: SW2_SP: Created spanning tree: VLAN0106 (5554BEFC)
Oct 22 14:03:31.647: SW2_SP: Setting spanning tree MAC address: VLAN0106 (5554BEFC) to
0008.e3ff.fc28
Oct 22 14:03:31.647: SW2_SP: setting bridge id (which=3) prio 24682 prio cfg 24576 sysid
106 (on) id 606A.0008.e3ff.fc28
Oct 22 14:03:31.647: SW2_SP: STP PVST: Assigned bridge address of 0008.e3ff.fc28 for
VLAN0106 [6A] @ 5554BEFC.
Oct 22 14:03:31.647: SW2_SP: Starting spanning tree: VLAN0106 (5554BEFC)
Oct 22 14:03:31.647: SW2_SP: Created spanning tree port Po206 (464F4174) for tree VLAN0106
(5554BEFC)
Oct 22 14:03:31.647: SW2_SP: RSTP(106): initializing port Po206
Oct 22 14:03:31.647: %SPANTREE-SW2_SP-6-PORT_STATE: Port Po206 instance 106 moving from
disabled to blocking <- 1
Oct 22 14:03:31.647: SW2_SP: RSTP(106): Po206 is now designated
Oct 22 14:03:31.667: SW2_SP: RSTP(106): transmitting a proposal on Po206
Oct 22 14:03:32.647: SW2_SP: RSTP(106): transmitting a proposal on Po206
Oct 22 14:03:32.655: SW2_SP: RSTP(106): received an agreement on Po206
Oct 22 14:03:32.919: %LINK-3-UPDOWN: Interface Vlan106, changed state to up
Oct 22 14:03:32.935: %LINEPROTO-5-UPDOWN: Line protocol on Interface Vlan106, changed
state to up
Oct 22 14:03:32.655: %SPANTREE-SW2_SP-6-PORT_STATE: Port Po206 instance 106 moving from
blocking to forwarding  <- 2
Oct 22 14:03:34.559: %PIM-5-DRCHG: DR change from neighbor 0.0.0.0 to 10.120.106.1 on
interface Vlan106
```

Access-Layer Switch

```
Access-Switch# show logging
```

```
Oct 22 14:03:29.671: %DTP-SP-5-TRUNKPORTON: Port Gi1/1-Gi1/2 has become dot1q trunk
Oct 22 14:03:31.643: %LINK-3-UPDOWN: Interface Port-channel1, changed state to up
Oct 22 14:03:31.647: %LINEPROTO-5-UPDOWN: Line protocol on Interface Port-channel1,
changed state to up
Oct 22 14:03:31.651: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/1,
changed state to up
Oct 22 14:03:31.636: %EC-SP-5-BUNDLE: Interface GigabitEthernet1/1 joined port-channel
Port-channel1
Oct 22 14:03:31.644: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 6 moving from disabled
to blocking
Oct 22 14:03:31.644: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 106 moving from disabled
to blocking <-1
Oct 22 14:03:31.644: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 900 moving from disabled
to blocking
Oct 22 14:03:31.660: %LINK-SP-3-UPDOWN: Interface Port-channel1, changed state to up
Oct 22 14:03:31.660: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface
GigabitEthernet1/1, changed state to up
Oct 22 14:03:31.664: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface Port-channel1,
changed state to up
Oct 22 14:03:31.867: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/2,
changed state to up
Oct 22 14:03:31.748: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 900 moving from blocking
to forwarding
Oct 22 14:03:31.856: %EC-SP-5-BUNDLE: Interface GigabitEthernet1/2 joined port-channel
Port-channel1
Oct 22 14:03:31.868: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface
GigabitEthernet1/2, changed state to up
Oct 22 14:03:32.644: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 6 moving from blocking
to forwarding
Oct 22 14:03:32.644: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 106 moving from blocking
to forwarding <- 2
```

Time to initialize the port-channel interface for a given VLAN is around one second (see markers in the preceding syslog output examples).

PortFast Enabled on a Trunk Port Channel

The following CLI examples illustrate output showing PortFast as being enabled on a trunk port-channel.

VSS Syslogs

```
6500-VSS# sh log | inc 106
Oct 22 14:14:11.397: SW2_SP: Created spanning tree: VLAN0106 (442F4558)
Oct 22 14:14:11.397: SW2_SP: Setting spanning tree MAC address: VLAN0106 (442F4558) to
0008.e3ff.fc28
Oct 22 14:14:11.397: SW2_SP: setting bridge id (which=3) prio 24682 prio cfg 24576 sysid
106 (on) id 606A.0008.e3ff.fc28
Oct 22 14:14:11.397: SW2_SP: STP PVST: Assigned bridge address of 0008.e3ff.fc28 for
VLAN0106 [6A] @ 442F4558.
Oct 22 14:14:11.397: SW2_SP: Starting spanning tree: VLAN0106 (442F4558)
Oct 22 14:14:11.397: SW2_SP: Created spanning tree port Po206 (464F2BCC) for tree VLAN0106
(442F4558)
Oct 22 14:14:11.397: SW2_SP: RSTP(106): initializing port Po206
Oct 22 14:14:11.401: %SPANTREE-SW2_SP-6-PORT_STATE: Port Po206 instance 106 moving from
disabled to blocking   <- 1
Oct 22 14:14:11.401: SW2_SP: RSTP(106): Po206 is now designated
Oct 22 14:14:11.401: %SPANTREE-SW2_SP-6-PORT_STATE: Port Po206 instance 106 moving from
blocking to forwarding  <- 2
Oct 22 14:14:11.769: %LINK-3-UPDOWN: Interface Vlan106, changed state to up
Oct 22 14:14:11.777: %LINEPROTO-5-UPDOWN: Line protocol on Interface Vlan106, changed
state to up
Oct 22 14:14:13.657: %PIM-5-DRCHG: DR change from neighbor 0.0.0.0 to 10.120.106.1 on
interface Vlan106
```

Access-Layer Switch

```
Access-switch# show logging
Oct 22 14:14:04.789: %LINK-SP-3-UPDOWN: Interface Port-channel1, changed state to down
Oct 22 14:14:05.197: %LINK-SP-3-UPDOWN: Interface GigabitEthernet1/1, changed state to
down
Oct 22 14:14:05.605: %LINK-SP-3-UPDOWN: Interface GigabitEthernet1/2, changed state to
down
Oct 22 14:14:05.769: %LINK-SP-3-UPDOWN: Interface GigabitEthernet1/1, changed state to up
Oct 22 14:14:06.237: %LINK-3-UPDOWN: Interface GigabitEthernet1/2, changed state to up
Oct 22 14:14:06.237: %LINK-SP-3-UPDOWN: Interface GigabitEthernet1/2, changed state to up
Oct 22 14:14:09.257: %DTP-SP-5-TRUNKPORTON: Port Gi1/1-Gi1/2 has become dot1q trunk
Oct 22 14:14:11.397: %LINK-3-UPDOWN: Interface Port-channel1, changed state to up
Oct 22 14:14:11.401: %LINEPROTO-5-UPDOWN: Line protocol on Interface Port-channel1,
changed state to up
Oct 22 14:14:11.401: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/1,
changed state to up
Oct 22 14:14:11.385: %EC-SP-5-BUNDLE: Interface GigabitEthernet1/1 joined port-channel
Port-channel1
Oct 22 14:14:11.397: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 6 moving from disabled
to blocking
Oct 22 14:14:11.397: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 6 moving from blocking
to forwarding
Oct 22 14:14:11.397: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 106 moving from disabled
to blocking    <- 1
Oct 22 14:14:11.397: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 106 moving from blocking
to forwarding  <- 2
Oct 22 14:14:11.397: %SPANTREE-SP-6-PORT_STATE: Port Po1 instance 900 moving from blocking
to forwarding
Oct 22 14:14:11.413: %LINK-SP-3-UPDOWN: Interface Port-channel1, changed state to up
Oct 22 14:14:11.913: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/2,
changed state to up
Oct 22 14:14:11.413: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface
GigabitEthernet1/1, changed state to up
Oct 22 14:14:11.413: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface Port-channel1,
changed state to up
Oct 22 14:14:11.901: %EC-SP-5-BUNDLE: Interface GigabitEthernet1/2 joined port-channel
Port-channel1
Oct 22 14:14:11.913: %LINEPROTO-SP-5-UPDOWN: Line protocol on Interface
GigabitEthernet1/2, changed state to up
```

As shown by markers in the preceding syslog output examples, the time between blocking and forwarding is practically zero.

The option to optimize trunk initialization with the Portfast feature should be weighed against the additional configuration requirements. During the initial systems boot up, before a systems can forward the packet, the system requires learning MAC, ARP discovery as well as completing network device discovery via control plane activities (neighbor adjacencies, routing, NSF, and so on). These added steps could nullify the gains of a trunk being able to come online faster; however, it does help when a port is forced to restart or put in-service after the switch/router has become fully operational.

PortFast and BPDU Guard

Protecting and improving the behavior of the edge port in VSS is the same as in any campus design. Configure the edge port with host-port macro to assert the STP forwarding state, reduce Topology Change Notification (TCN) messaging, and eliminate delay caused by other software configuration checks for EtherChannel and Trunk. The VSS is loop-free topology; however, end-user action or access-layer switch miscabling can introduce a loop into the network. The looped network eventually can lead to unpredictable convergence and greatly increase the chance for a loop-based broadcast storm.

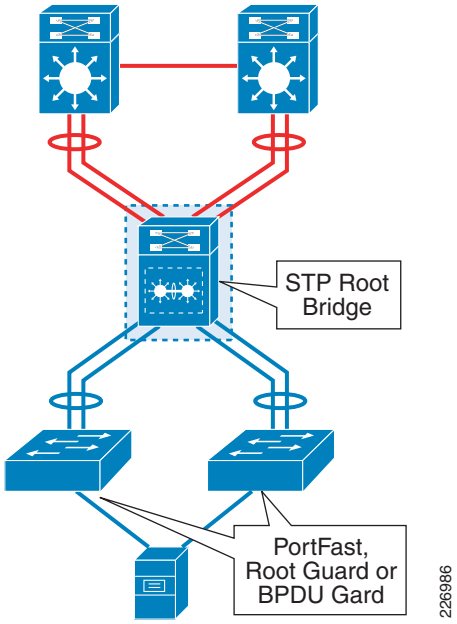
Tip **In the VSS-enabled network, it is critically important to keep the edge port from participating in the STP. Cisco strongly recommends enabling PortFast and BPDU Guard at the edge port.**

When enabled globally, BPDU Guard applies to all interfaces that are in an operational PortFast state. The following configuration example illustrates enabling BPDU Guard:

```
VSS(config-if)# spanning-tree PortFast
VSS(config-if)# spanning-tree bpduguard enable
%SPANTREE-2-BLOCK_BPDUGUARD: Received BPDU on port FastEthernet3/1 with BPDU Guard
enabled. Disabling port.
%PM-4-ERR_DISABLE: bpduguard error detected on Fa3/1, putting Fa3/1 in err-disable state
```

Figure 3-26 illustrates the configuration zone for various STP features.

Figure 3-26 PortFast, BPDU Guard, and Port Security



BPDU Filter

The improper use of the BPDU Filter feature can cause loops in the network. Just as in a traditional multilayer design, avoid using BPDU filtering in VSS-enabled network. Instead, use BPDU Guard.

STP Operation with VSS

VSS is comprised of a single logical switch that advertises single STP bridge-ID and priority, regardless of which virtual-switch member is in the active state. The bridge-ID is derived from the active chassis as shown in the following output:

```
6500-VSS# sh catalyst6000 chassis-mac-addresses
  chassis MAC addresses: 1024 addresses from 0008.e3ff.fc28 to 0008.e400.0027
6500-VSS# sh spanning-tree vla 450
```

```
VLAN0450
  Spanning tree enabled protocol rstp
  Root ID     Priority    25026
              Address     0008.e3ff.fc28
              This bridge is the root
              Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID   Priority    25026  (priority 24576 sys-id-ext 450)
              Address     0008.e3ff.fc28
              Hello Time   2 sec  Max Age 20 sec  Forward Delay 15 sec
              Aging Time 480

Interface           Role Sts Cost      Prio.Nbr Type
------------------- ---- --- --------- -------- --------------------------------
Po202               Desg FWD 3         128.1699 P2p
```

With VSS, spanning tree is SSO-aware. SSO enables STP protocol resiliency during SSO switchover (active failure), the new active switch member maintains the originally-advertised STP bridge priority and identifier for each access-layer switch. This means that STP need not reinitialize and undergo the learning process of the network that speeds the convergence (to sub-second performance).

Similarly, a member-link failure of MEC does not generate the TCN because STP is operating on EtherChannel port. Refer to the following output.

```
6500-VSS# show spanning-tree vl 10 detail | inc Times|Port-channel
  Root port is 1665 (Port-channel10), cost of root path is 3
          from Port-channel10
  Times:  hold 1, topology change 35, notification 2
 Port 1665 (Port-channel10) of VLAN0010 is root forwarding
6500-VSS#show interface port-channel10 | inc Gi
  Members in this channel: Gi1/1 Gi1/2
6500-VSS# conf t
VSS(config)# int gi1/1
VSS(config-if)# shut

6500-VSS# show spanning-tree vlan 10 detail | inc Times|Port-channel
  Root port is 1665 (Port-channel10), cost of root path is 4
          from Port-channel10
  Times:  hold 1, topology change 35, notification 2
 Port 1665 (Port-channel10) of VLAN0010 is root forwarding
6500-VSS#show interface port-channel10 | inc Gi
  Members in this channel: Gi1/2
```

The active switch is responsible for generating the BPDU. The source MAC address of every BPDU frame is derived from a line card upon which the STP port (MEC) is terminated. The MAC address inherited by the MEC port is normally used as a source MAC address for the BPDU frame. This source MAC address can change dynamically due to a node/line or card/port failure. The access switch might see such an event as a new root because the BPDU is sent out with new source MAC. However, this failure does not cause STP topology recomputation in the network because the network is loop-free and the STP bridge-ID/priority remains the same. The **debug** commands below illustrate how you can monitor this behavior on Cisco Catalyst 3560 switches. Note that the source MAC address of the BPDU has changed, but the bridge ID of the root bridge remain the same (VSS is a single logical root).

```
3560-switch# debub spanning-tree switch  rx decode
3560-switch# debug spanning-tree switch  rx process

Apr 21 17:44:05.493: STP SW: PROC RX: 0100.0ccc.cccd<-0016.9db4.3d0e type/len 0032 <-
Source MAC
Apr 21 17:44:05.493:     encap SNAP linktype sstp vlan 164 len 64 on v164 Po1
Apr 21 17:44:05.493:     AA AA 03 00000C 010B SSTP
Apr 21 17:44:05.493:     CFG P:0000 V:02 T:02 F:3C R:60A4 0008.e3ff.fc28 00000000
```

```
Apr 21 17:44:05.493:     B:60A4 0008.e3ff.fc28 86.C6 A:0000 M:1400 H:0200 F:0F00 <- Root
Bridge ID
Apr 21 17:44:05.493:      T:0000 L:0002 D:00A4
Apr 21 17:44:05.544: %DTP-5-NONTRUNKPORTON: Port Gi0/1 has become non-trunk
Apr 21 17:44:06.030: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet0/1,
changed state to down
Apr 21 17:44:06.072: STP SW: PROC RX: 0100.0ccc.cccd<-0016.9db4.d21a type/len 0032 <- New
Source MAC
Apr 21 17:44:06.072:     encap SNAP linktype sstp vlan 20 len 64 on v20 Po1
Apr 21 17:44:06.072:     AA AA 03 00000C 010B SSTP
Apr 21 17:44:06.072:     CFG P:0000 V:02 T:02 F:3C R:6014 0008.e3ff.fc28 00000000
Apr 21 17:44:06.072:     B:6014 0008.e3ff.fc28 86.C6 A:0000 M:1400 H:0200 F:0F00 <- Same
Bridge ID
Apr 21 17:44:06.072:      T:0000 L:0002 D:0014
Apr 21 17:44:06.072: STP SW: PROC RX: 0100.0ccc.cccd<-0016.9db4.d21a type/len 0032
Apr 21 17:44:06.072:     encap SNAP linktype sstp vlan 120 len 64 on v120 Po1
Apr 21 17:42:05.939:      T:0000 L:0002 D:0016
```

The following syslogs appears as symptom of link change, but there is no root-change with link member deactivation.

```
Apr 21 17:39:43.534: %SPANTREE-5-ROOTCHANGE: Root Changed for vlan 1: New Root Port is
Port-channel1. New Root Mac Address is 0008.e3ff.fc28
```

Design Considerations with Large-Scale Layer-2 VSS-Enabled Campus Networks

With the VSS-enabled, loop-free design, network designers are less constrained in designing a network. With a VSS implementation, the network can span VLANs over multiple switches and support multiple VLANs existing on each switch. The primary motivations of such a design are operational flexibility and efficient resource usage (subnets, VLANs, and so on). The obvious questions are as follows:

Q. What is the appropriate STP domain size?

Q. How many VLANs are allowed per-VSS pair?

Q. How many devices can be supported per-VSS pair?

The STP domain sizing and answers to above questions rely on many considerations including non-VSS devices in the STP domain. STP domain consists not only of VSS but also other devices participating in STP topology. However, the key factors affecting spanning convergence must be considered in determining the scope of an STP domain:

- Time-to-converge—Depends on the protocol implemented (802.1d, 802.1s, or 802.1w)
- The number of MAC addresses to be advertised and learned during initialization and failure
- Topology—Number of alternate paths to find a loop-free topology; a deeper topology yields a longer time to find a loop-free path
- MAC address learning—Can be a hardware (faster)- or software-based (slower) function
- MAC address capacity of spanning tree domain—Convergence takes longer with larger numbers of MAC addresses
- Number of VLAN and STP instances governs how many BPDUs must be processed by the CPU. The lower capacity CPU may drop BPDU and thus STP take longer to converge.

- Number of VLANs being trunked across each link – Number of STP logical port on which switch CPU has to send the BDPU
- Number of logical ports in the VLAN on each switch – Overall systems capability

VSS, inherently by design, removes most of the preceding factors that can affect STP convergence by enabling the following:

- Loop-free topology—No topology changes are seen by the spanning tree process, so there is no reaction and no dependency
- No root reelection process upon failure of one to the VSS-member chassis because it is a single logical switch (root)
- VSS supports hardware-based MAC learning
- Key components responsible of STP operation such as STP database, ARP, and port-channel interface state are SSO-aware. Any changes or learning about STP is now synchronized between the active and hot-standby chassis—eliminating the MAC-based topology dependency.

However, the STP domain comprises not just VSS, but also includes access-layer switches and other Layer-2 devices connected to the VSS. Together, these elements put additional constraints on the design that cannot be addressed by the VSS alone and should be considered while designing large Layer-2 networks. The following are examples of related constraints:

- The maximum number of VLANs supported on lower-end switch platforms can be constraining factors. For example, the Cisco 2960 and non-Cisco switches have limited VLAN support.
- MAC-address learning and Ternary Content Addressable Memory (TCAM) capacity for the rest of the switches in the STP domain can be a constraint. Typically, access-switches have limited TCAM capabilities. Exceeding the TCAM capacity can result in MAC addresses becoming available in software, so that the packet destined to a overflow MAC address is switched in software instead of hardware. This can lead to further instability of the switch and thus STP domain.
- The rate of MAC address change offered to the VSS-enabled network might increase the control plane activity. The number of MAC addresses moved, added, or removed per second could be so high as to affect the control plane's capability to synchronize between the active and hot-standby switches.
- The exposure domain for viruses and other infection control policies can be a constraint. With a large-scale Layer-2 network that does not have proper Layer-2 security, a high number of host infections can occur before a mitigation effort is applied.
- The existing subnet structure and VLAN sizing can limit the usage of large Layer-2 VLANs spanning multiple access-layer switches. In a typical campus, an access-layer switch has a subnet scope of 256 hosts (/24 subnet) that naturally map to physical ports available in lower-end switching platforms. It is not always easy to change that structure. Keeping VLANs contained within an access-layer switch provides a better troubleshooting and monitoring boundary.

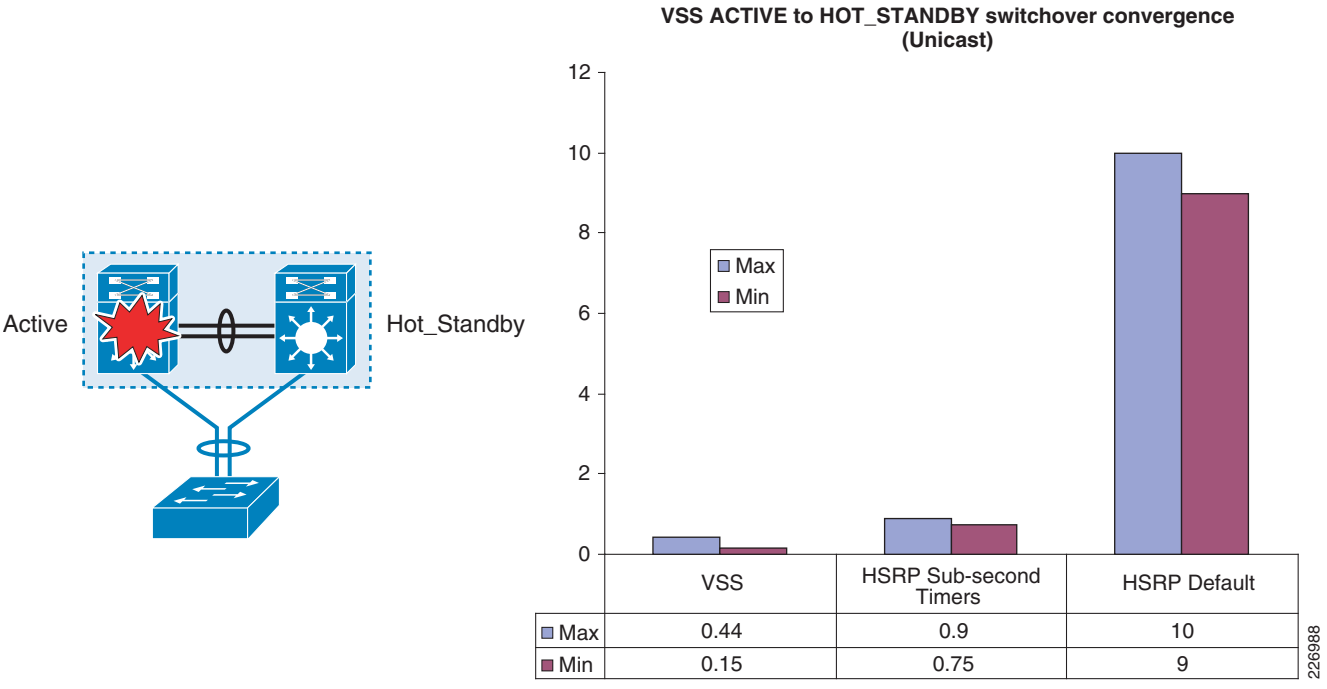
The outlier scope of Layer-2 network span largely remains a design policy exercise and has no single correct size for all enterprise campus networks. The typical recommendation is to use spanning capabilities with the VSS for VLANs assigned for functional use, such as network management, guest access, wireless, quarantine, pasture assessment, and so on. The voice and data traffic can still remains confined to a single access-layer and be expand with careful planning involving the preceding design factors. From the convergence and scalability perspective, the validation performed with this design guide uses a campus networking environment with the attributes summarized in Table 3-2.

Table 3-2 Campus Network Capacity Summary

Campus Environment	Average Capacity and Scope	Validated Campus Environment	Comments
Number of MAC addresses per Distribution Block	4K to 6K	~ 4500	Unique per host to MAC ratio
Average number of access-layer switch per Distribution Block	30 to 50	70	70 MECs per VSS
VLAN Spanned to Multiple Switches	Variable	8 VLANs	Constrained by preceding design factors
MAC Address for Spanned VLANs	Variable	720 MAC/VLANs	
VLAN Confined to Access-layer	Variable	140	Voice and data VLANs restricted per access-layer switch

The convergence associated with an active-to-standby failure with or without spanned VLAN remains same. This design guide validation illustrate the capabilities described in preceding paragraph about VSS eliminating ST- related convergence factors. Figure 3-27 illustrates the convergence when deploying a VSS. The default convergence for standalone enabled with default FHRP is 10 seconds and with tuning the best case convergence is around 900 msec. In contrast, VSS has a 200-msec average convergence time without the requirement of any complex tuning and specialized configuration.

Figure 3-27 Switchover Convergence Comparison



The "Active Switch Failover" section on page 4-5 details traffic flows and events occurring during an active supervisor failure.

Multicast Traffic and Topology Design Considerations

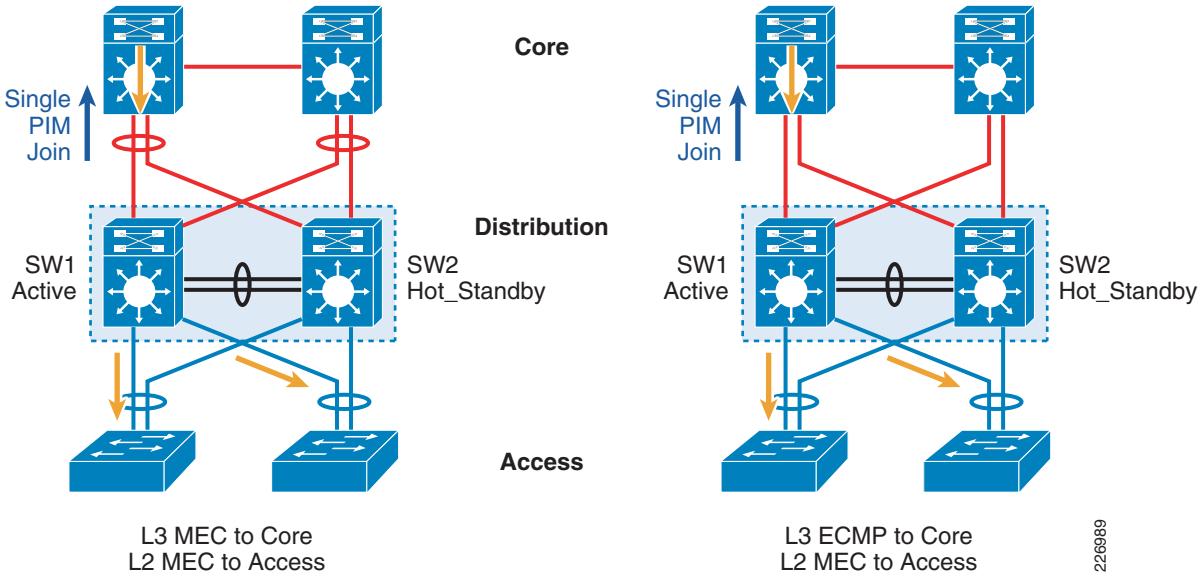
VSS shares all the benefits and restrictions of standalone MMLS technology. Forwarding states are programmed in hardware on both the active and standby supervisor. During a switchover, the hardware continues forwarding multicast data while the control plane recovers and reestablishes PIM neighbor relations. Supervisor Sup720 with all CEF720 fabric line cards running Cisco IOS Release 12.2(18)SXF and later is capable of ingress as well as egress multicast replication. DFC-enabled line cards are recommended for egress replication to avoid multiple PFC lookups by CFC cards. However, for VSS enabled configuration, only multicast egress replication is supported and that is per-physical chassis basis. There is no ingress replication mode for the VSS. The implication of such restriction is that if the multicast flows are required to be replicated over VSL link, every single flow will be replicated for every outgoing interface list that exist over remote peer chassis. The possibility of such flow behavior exists with non-MEC-based design which is illustrated in "Multicast Traffic Flow without Layer-2 MEC" section on page 3-43. The multicast capabilities are illustrated in the following CLI output:

```
6500-VSS# sh platform hardware cap multicast
L3 Multicast Resources
  IPv4 replication mode: egress
  IPv6 replication mode: egress
  Bi-directional PIM Designated Forwarder Table usage: 4 total, 0 (0%) used
  Replication capability: Module                                IPv4        IPv6
                          18                                  egress      egress
                          21                                  egress      egress
                          23                                  egress      egress
                          24                                  egress      egress
                          25                                  egress      egress
                          34                                  egress      egress
                          37                                  egress      egress
                          39                                  egress      egress
                          40                                  egress      egress
                          41                                  egress      egress
  MET table Entries: Module                             Total    Used    %Used
                     18                                 65516       6       1%
                     21                                 65516       6       1%
                     34                                 65516       6       1%
                     37                                 65516       6       1%
Multicast LTL Resources
  Usage:   24512 Total, 13498 Used
```

Multicast Traffic Flow with Layer-2 MEC

Figure 3-28 presents the multicast traffic behavior for a Layer-2 MEC-based network. The Layer-3 connectivity option is included for the reference, but Layer-3 options are addressed in the "Routing with VSS" section on page 3-44.

Figure 3-28 Multicast Traffic Behavior for Layer-2 MEC-based Network

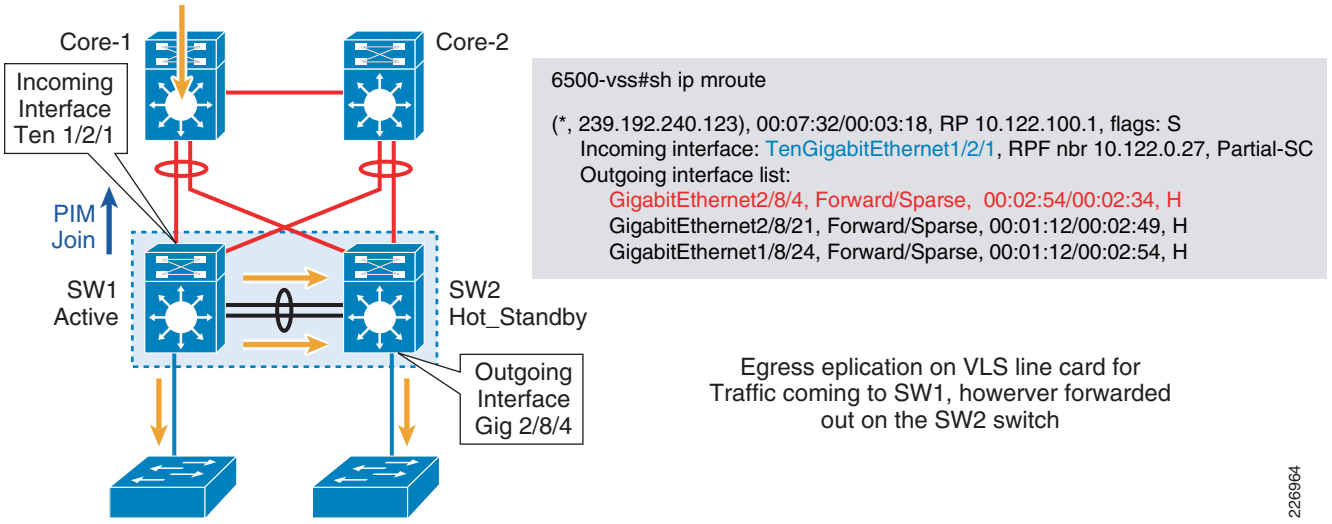


As with unicast data, the control plane for multicast is handled in the active switch. With VSS being a single logical switch, it sends a single PIM join. The PIM join is originated by the active switch, but it can be sent by an interface located on the hot-standby switch (in the case of ECMP). Normally, the highest IP address of the PIM neighbors (which is usually the first entry in a routing table for the destination sourcing the multicast stream) is where the PIM join is sent. Once the incoming interface is selected, the traffic is replicated based on Layer-2 connectivity. For MEC-based connectivity, the switch on which the incoming interface resides will forward the multicast data to a locally connected MEC member and performs egress replication when DFC line cards are available.

Multicast Traffic Flow without Layer-2 MEC

If the Layer-2 connectivity is not Layer-2 MEC-based (has only a single connection to an access-layer such as with single-homed connectivity) or one of the local interfaces is down, then it is possible that incoming and outgoing (replication) interface reside on two different switches. For this type of connectivity or condition, the multicast traffic will be replicated over the VSL link. Egress replication is performed on the SW1 VSL line card for every flow (*,g and s,g) arriving on SW1 for every outgoing interface list (OIL) on the SW2. This can result in high data traffic over VSL. The Layer-2 MEC-based connectivity avoids this non-optimal traffic flow and further supports the necessity of MEC in the VSS-enabled campus. Figure 3-29 illustrates multicast traffic flow without Layer-2 MEC.

Figure 3-29 Multicast Flow Without Layer-2 MEC



VSS—Single Logical Designated Router

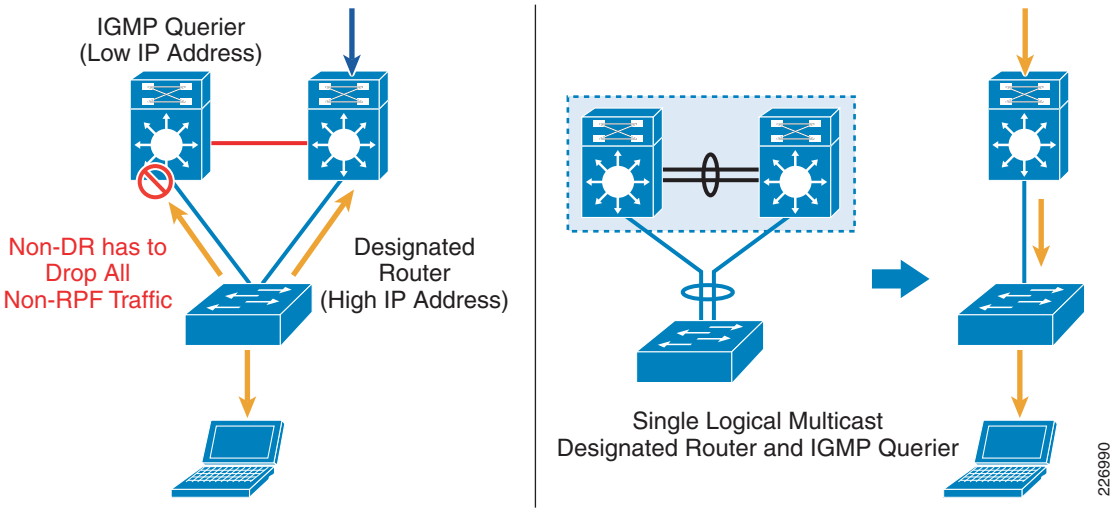
IP multicast uses one router to forward data onto a LAN in redundant topologies. If multiple routers have interfaces into a LAN, only one router will forward the data. There is no load balancing for multicast traffic on LANs. Usually, the designated router (DR)—highest IP address on a VLAN—is chosen to forward the multicast data. If the DR router fails, there is no inherent method for other routers (called backup DRs) to start forwarding the multicast data. The only way to determine whether forwarding has stopped is via continuously receiving multicast data on redundant routers, which means that the access-layer switches forward every single multicast packet uplink. A redundant router sees this data on the outbound interface for the LAN. That redundant router must drop this traffic because it arrived on the wrong interface and therefore will fail the Reverse Path Forwarding (RPF) check. This traffic is called non-RPF traffic because it is being reflected back against the flow from the source. For further information on IP multicast stub connectivity, refer to the following URL:

http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6552/ps6592/whitepaper_c11-474791.html

Non-RPF traffic has two side effects: first, it wastes uplink bandwidth; and second, it causes high CPU usage if proper precautions are not taken based on the hardware deployed.

The topology changes for the multicast in non-VSS versus VSS-enabled campus, as illustrated in Figure 3-30. The VSS is treated as a single multicast router, which simplifies multicast topology as shown in Figure 3-30. Because there is only one node attached to the Layer-2 domain, there is no selection of backup DRs. In VSS for a normal condition, the multicast forwarder is selected based on which, among the VSS switch member links, receives multicast traffic and builds an incoming interfaces list). Because VSS always performs local forwarding, the multicast traffic is forwarded via locally-attached link for that member. If the link connected to Layer-2 domain fails, the multicast data will select the VSL-bundle link in order to forward data to the access-layer. It will not undergo multicast control plane reconvergence. If the incoming interface link fails, then the multicast control plane must reconverge to find an alternate path to the source. This later failure case is addressed under the "Routing with VSS" section on page 3-44.

Figure 3-30 Simplified Multicast Topology



Routing with VSS

This section covers the behavior and interaction of VSS with Layer-3 devices in the context of its overall implementation at the distribution-layer switch. The design guidance and observation are equally applicable VSS implementations in the core and routed-access design. The later part of this section covers the benefits of those VSS designs. Typically, in three-tier architecture, the distribution-layer provides a boundary function between Layer-2 and Layer-3 domain. VSS behavior differs from a standalone node in Layer-3 domain in following ways:

- Routing Protocols, Topology, and Interaction
- Routing Protocol Interaction During Active Failure

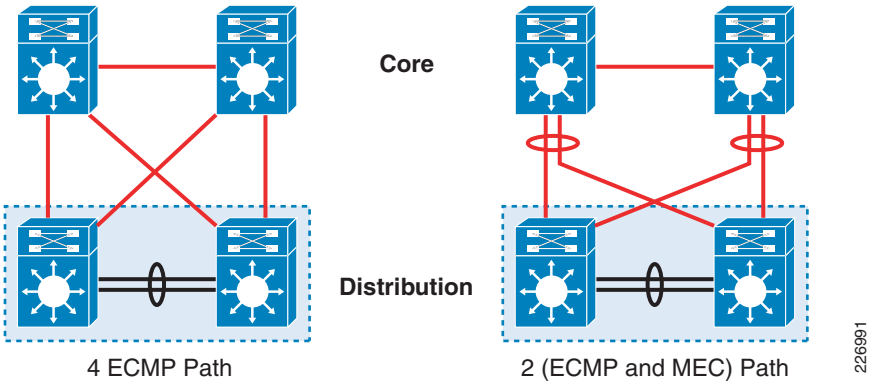
Routing Protocols, Topology, and Interaction

The VSS supports common routing protocols, including Enhanced Internal Gateway Routing Protocol (IGRP), Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), Intermediate System-to-Intermediate System (IS-IS), and Routing Information Protocol (RIP). This design guide only covers Enhanced IGRP and OSPF. The interaction of routing protocols with the VSS core largely depends on the topology deployed. In general, there are two ways to connect VSS to the core devices.

- Equal Cost Multipath (ECMP)
- Layer-3 MEC

Figure 3-31 illustrates the fully-meshed connectivity option with VSS.

Figure 3-31 Fully-meshed Connectivity VSS Option



The traditional best practice campus design recommends that you deploy a full-mesh because the failure of a link or node does not force the traffic reroute to depend of the routing protocol, instead fault recovery depends on CEF path switching in hardware. This recommendation still applies with a VSS deployment, but for a different reason. The failure of link in a non-full mesh (with VSS at the distribution layer and a single physical link connecting each member to their respective core devices) will force the traffic to reroute over the VSL bundle. This is not an optimal design choice because it increases the latency during a failure and possible traffic congestion over VSL links. With a full-mesh design, a VSL reroute is only possible if the link connecting to the core is not diversified over a different line card and the line card fails. It is for this reason that the fully-meshed link should be diversified over the two separate line cards in order to provide fully redundant connectivity.

Unicast traffic takes the optimal path in both ECMP- and MEC-based full-mesh topologies. Each VSS member forwards the traffic to the locally available interfaces and thus no traffic traverses over the VSL bundle under normal operational conditions.

The difference between ECMP- and MEC-based topologies in terms of routing protocol interaction with VSS at the distribution and traditional dual nodes at the core are summarized in Table 3-3.

Table 3-3 Topology Comparison

Topology	ECMP	MEC	Comments
Layer-3 Routed Interfaces	Four point-to-point	Two - Layer-3 Port-channel	
Enhanced IGRP or OSPF Neighbors	Four	Two	

Table 3-3 Topology Comparison (continued)

Topology	ECMP	MEC	Comments
Routing Table entries for a given destination	Four	Two	
VSS originated Hello or routing updates over VSL	Yes, for neighbors on a hot_standby member	No, locally connected interfaces carries hello	Fault condition may change the default behavior
Remote devices originated hello or routing updates over VSL	Yes, for neighbors on a host-standby member	Depends on hashing output, it could traverse over VSL	Fault condition may change default behavior

As shown in Table 3-3, MEC-based connectivity can reduce the neighbor count and reduce routing table entries, thus reducing the CPU load for control plane activity related to a routing protocol. This is especially useful in highly meshed designs with many Layer 3-connected devices, such as in a core or routed-access design.

The topology also determines how neighbor relations are built with other Layer-3 routed devices in the network. For ECMP, it is direct point-to-point connectivity where neighbor hellos are exchanged symmetrically. In ECMP-based topology, the neighbor hello for the links connected to hot-standby switch (from both core devices and VSS) has to traverse VSL link, because the neighbor adjacency has to originate where the point-to-point link is connected.

For MEC-based connections, the neighbor hello connectivity can be asymmetric. In an MEC-based topology, the VSS always prefers to sends hellos and updates (topology or LSAs) from locally-attached link (part of EtherChannel). The hashing always selects locally-available link. The remote devices connected to the VSS via EtherChannel undergo a similar hashing decision process in selecting the link over which to send hellos or routing updates. For each routing protocol, the hello and routing update packet uses different IP address for destination. Based on a type of routing protocol packet, the hash result may select a link that is connected to either active or hot-standby. Therefore, it is entirely possible that hello and update may select a different link to reach the active-switch member of the VSS. This behavior of path selection for neighbor hellos and routing updates plays critical role in determining the stability of neighbors during dual-active failure scenarios.

Design Considerations with ECMP and MEC Topologies

This section covers the effect of the type of topology deployed between the VSS (distribution layer) and the core. Two major design points affect topology selection are as follows:

- Link Failure Convergence
- Forwarding Capacity (Path Availability) During Link Failure

For the traffic flow and convergence behavior, refer to Chapter 4, “Convergence.”

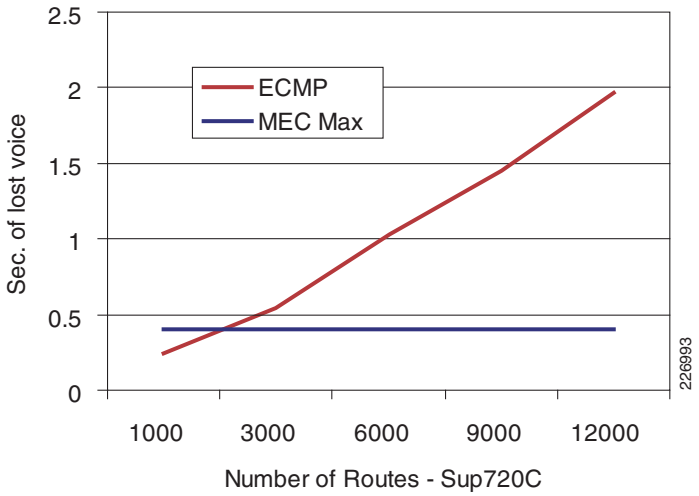
Link Failure Convergence

The link-failure convergence with ECMP and MEC is shown in Figure 3-32. Note that ECMP-based topology convergence depends on routing table size.

ECMP

With the higher numbers of routing entries, higher loss of VoIP data was observed (see Figure 3-32). This is because CEF must readjust the VoIP flows over the failed link. Even though this recovery is not dependent on a routing protocol, the path reprogramming for the destination takes longer, depending on routing table size.

Figure 3-32 Number of Routes vs. Voice Loss



MEC

EtherChannel detection is hardware-based. The failure of the link and adjustment of the VoIP flow over the healthy member link is consistent. The worst-case loss with a MEC-link member failure is about 450 msec. On average, 200-msec recovery can be expected from the Cisco Catalyst 4500 and Cisco Catalyst 3xxx switching platforms.

Forwarding Capacity (Path Availability) During Link Failure

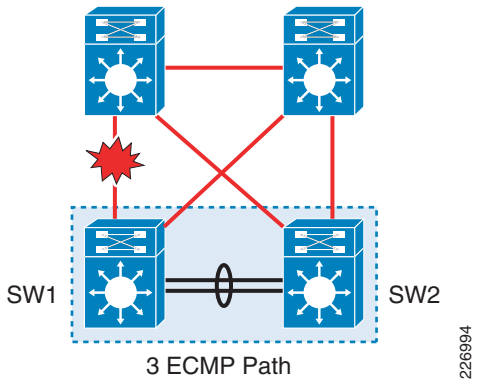
ECMP

For normal operational conditions, the VSS has two paths per-member switch, a total of four forwarding links. The sample routing table entries for a given destination is shown in the following output example:

```
6500-VSS# sh ip route 10.121.0.0 255.255.128.0 longer-prefixes
D       10.121.0.0/17
           [90/3328] via 10.122.0.33, 2d10h, TenGigabitEthernet2/2/1
           [90/3328] via 10.122.0.22, 2d10h, TenGigabitEthernet2/2/2
           [90/3328] via 10.122.0.20, 2d10h, TenGigabitEthernet1/2/2
```

Any single-link failure will result in the ECMP path reprogramming; all three other links remain operational and available for forwarding. See Figure 3-33.

Figure 3-33 Link Failure Effects



As discussed in "Layer-3 ECMP Traffic Flow" section on page 3-7, SW1 will continue forwarding traffic flow to locally-available interfaces. SW2 will have two links and will continue forwarding traffic. This means that, in an ECMP topology, all three paths are available for traffic forwarding, thus logical bandwidth availability remains the same as physical link availability.

MEC

For the MEC-based topology from the VSS to the core, only two logical routed port-channel interfaces are available. This translates to providing only two ECMP paths per-switch member (single logical switch). A link failure's effect is dependent on routing protocol and metric recalculation of the available paths.

OSPF with Auto-Cost Reference Bandwidth

With the integration of high-bandwidth interfaces for campus connectivity, the reference bandwidth for OSPF might require adjustment. If the reference bandwidth is not adjusted, then the OSPF shorted path first (SFP)-based algorithm cannot differentiate the cost of interface bandwidth higher than 100 Mbps. This can result into sub-optimal routing and unexpected congestion. The reference bandwidth is normally adjusted from 100 Mbps default to the highest possible bandwidth available in the network. The reference bandwidth can easily reach 20-Gigabit in the VSS with MEC interfaces bundle comprising of two 10-Gigabit member links. The Cisco IOS allows the metric (OSPF cost) of routed link be reflected by underlying physical link. If the auto-cost reference bandwidth is configured such that the cost of the OSPF-enabled interfaces (routed) changes when the member link of a MEC fails, then forwarding capability can differ from path availability. In this design guide, the VSS is connected via the port-channel interface to the core (two 10-Gigabit links). A MEC member link failure will trigger a cost change to a higher value on one of the port-channel interfaces, resulting in the withdrawal of the route for the given destination. From a physical topology point-of-view, three interfaces are capable of forwarding traffic. However, the effective forwarding capacity is now dependent on the available logical path, which is only one from a single core.

Figure 3-34 OSPF with Auto-Cost Reference Bandwidth

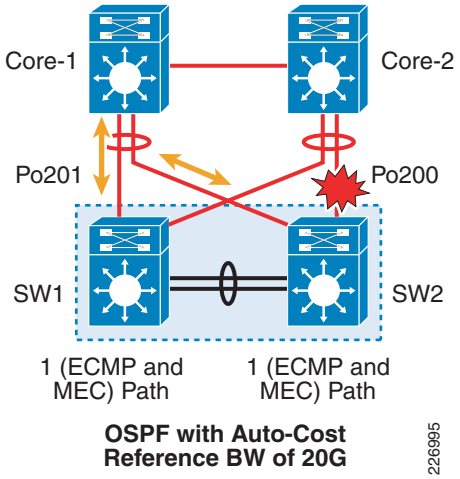


Figure 3-34 illustrates the behavior of metric change associated with OSPF and Layer-3 MEC. The failure of one link from a port-channel (Po200) causes the removal of routes learned from the entire port-channel. As a result, instead of two routes, only one route from one of the core router (core-1) is available. For VSS, this is reflected via one logical path connected to both members. Note that the SW1 has two physical links, but only one link connected to core-1 will be used. The logical path forwarding availability is shown with yellow arrows in the Figure 3-34. The following output from **show** commands depicts the behavior relevant with auto-cost reference bandwidth:

```
6500-VSS# show running-config | begin router ospf
router ospf 100
 router-id 10.122.0.235
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000
 nsf
 area 120 stub no-summary
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 area 120 range 10.125.0.0 255.255.0.0 cost 10
 passive-interface default
 no passive-interface Port-channel200
 no passive-interface Port-channel201
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
 network 10.125.0.0 0.0.3.255 area 120
```

The routing and hardware-CEF path available during normal operational conditions with two port-channel interfaces are presented in the following command output:

```
6500-VSS# sh ip route 10.121.0.0
Routing entry for 10.121.0.0/16
  Known via "ospf 100", distance 110, metric 13, type inter area
  Last update from 10.122.0.20 on Port-channel201, 00:51:31 ago
  Routing Descriptor Blocks:
  * 10.122.0.27, from 30.30.30.30, 00:51:31 ago, via Port-channel200
      Route metric is 13, traffic share count is 1
    10.122.0.20, from 30.30.30.30, 00:51:31 ago, via Port-channel201
      Route metric is 13, traffic share count is 1

6500-VSS#sh mls cef 10.121.0.0 16 sw 1

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108803 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)
```

```
                           Po200           , 0012.da65.5400 (Hash: 7F80)
6500-VSS#sh mls cef 10.121.0.0 16 sw 2

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108802 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)
                           Po200           , 0012.da65.5400 (Hash: 7F80)
```

Hardware-CEF path availability during a member-link failure with two port-channel interfaces is shown in the following output listings. The output shows that only one logical path from each switch is available even though three physical paths are available.

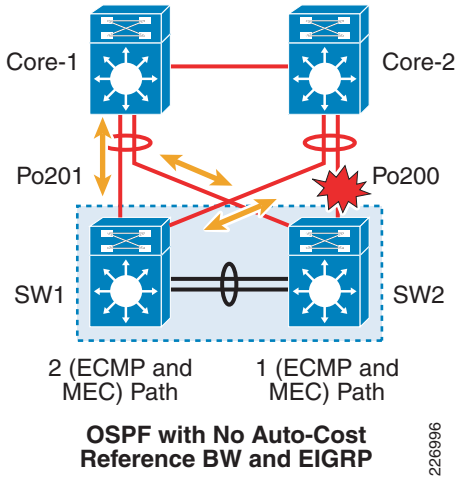
```
6500-VSS# sh ip route 10.121.0.0
Routing entry for 10.121.0.0/16
  Known via "ospf 100", distance 110, metric 13, type inter area
  Last update from 10.122.0.20 on Port-channel201, 00:51:31 ago
  Routing Descriptor Blocks:
  * 10.122.0.20, from 30.30.30.30, 00:51:31 ago, via Port-channel201
      Route metric is 13, traffic share count is 1

6500-VSS# sh mls cef 10.121.0.0 16 sw 1

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108803 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)

6500-VSS# sh mls cef 10.121.0.0 16 sw 2

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108802 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)


6500-VSS# sh ip os ne

Neighbor ID     Pri   State           Dead Time   Address         Interface
10.254.254.8      0   FULL/  -        00:00:36    10.122.0.20     Port-channel201
10.254.254.7      0   FULL/  -        00:00:39    10.122.0.27     Port-channel200

6500-VSS# sh run int po 200
Building configuration...

Current configuration : 378 bytes
!
interface Port-channel200
 description 20 Gig MEC to cr2-6500-1 4/1-4/3
 no switchport
 dampening
 ip address 10.122.0.26 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 logging event spanning-tree status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
end

6500-VSS#sh run int po 201
Building configuration...
```

```
Current configuration : 374 bytes
!
interface Port-channel201
 description 20 Gig to cr2-6500-1 4/1-4/3
 no switchport
 dampening
 ip address 10.122.0.21 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 logging event spanning-tree status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
end
```

OSPF Without Auto-Cost Reference Bandwidth

For the network that has configured OSPF with auto-cost reference bandwidth (100 Mbps), the link member failure does not alter the cost of the routed port-channel interface. This means that route withdrawal does not occur, leaving two routes for the destination in the system. However, this allows you to use all three physical paths (one path in SW2 and two path for SW1). Figure 3-35 illustrates an OSPF-based environment without the auto-cost reference bandwidth.

Figure 3-35 OSPF Without Auto-Cost Reference Bandwidth



The following CLI output illustrates that the state of the routing table and hardware-CEF *before* and *after* a link failure remains the same. The routing and hardware-CEF paths available during normal operational conditions with two port-channel interfaces are illustrated.

```
6500-VSS# sh ip route 10.121.0.0
Routing entry for 10.121.0.0/16
  Known via "ospf 100", distance 110, metric 13, type inter area
  Last update from 10.122.0.20 on Port-channel201, 00:51:31 ago
  Routing Descriptor Blocks:
  * 10.122.0.27, from 30.30.30.30, 00:51:31 ago, via Port-channel200
      Route metric is 13, traffic share count is 1
    10.122.0.20, from 30.30.30.30, 00:51:31 ago, via Port-channel201
```

```
      Route metric is 13, traffic share count is 1
6500-VSS# sh mls cef 10.121.0.0 16 sw 2

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108803 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)
                           Po200           , 0012.da65.5400 (Hash: 7F80)
6500-VSS# sh mls cef 10.121.0.0 16 sw 1

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108802 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)
                           Po200           , 0012.da65.5400 (Hash: 7F80)
```

A link failure keeps the routing path the same and only removes the specific hardware-CEF path for the failed link (SW1), as shown in the following output examples:

```
6500-VSS# sh ip route 10.121.0.0
Routing entry for 10.121.0.0/16
  Known via "ospf 100", distance 110, metric 13, type inter area
  Last update from 10.122.0.20 on Port-channel201, 00:51:31 ago
  Routing Descriptor Blocks:
  * 10.122.0.27, from 30.30.30.30, 00:51:31 ago, via Port-channel200
      Route metric is 13, traffic share count is 1
    10.122.0.20, from 30.30.30.30, 00:51:31 ago, via Port-channel201
      Route metric is 13, traffic share count is 1
6500-VSS# sh mls cef 10.121.0.0 16 sw 2

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108803 10.121.0.0/16       Po201           , 0012.da67.7e40
6500-VSS#

6500-VSS# sh mls cef 10.121.0.0 16 sw 1

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
108802 10.121.0.0/16       Po201           , 0012.da67.7e40 (Hash: 007F)
                           Po200           , 0012.da65.5400 (Hash: 7F80)
```

Enhanced IGRP

The Enhanced IGRP metric calculation is a composite of the total delay and the minimum bandwidth. When a member link fails, EIGRP recognizes and uses the changed bandwidth value but the delay will not change. This may or may not influence the composite metric since minimum bandwidth in the path is used for the metric calculation; therefore, a local bandwidth change will only affect the metric if it is the minimum bandwidth in the path. In a campus network, the bandwidth changed offered between core and VSS is in the order of Gigabits, which typically is not a minimum bandwidth for the most of the routes. Thus, for all practical purposes, Enhanced IGRP is immuned to bandwidth changes and follows the same behavior as OSPF with the default auto-cost reference bandwidth. If there are conditions in which the composite metric is impacted, then EIGRP follows the same behavior as OSPF with auto-cost reference bandwidth set.

Summary

The design choice with the OSPF and Layer-3 MEC topology is that of total bandwidth available during the fault and not the impact on user data convergence since the packet losses are at minimal. For more details, refer to the "Routing (VSS to Core) Convergence" section on page 4-14.

The auto-cost reference bandwidth configuration is required to be the same in the entire network to avoid routing loops. However, you can make an exception of not setting the auto-cost reference bandwidth for the VSS. This is possible because typically the access-layer is configured as totally stubby area in best practiced design. Such a topology does not offer any back-door alternate path from access-layer to the core and vices-versa. The advantage of relaxing the rule of auto-cost is the availability of all paths being used by user data traffic, regardless of the routing protocol and its configuration.

Another insight into selecting this configuration option is the application response time. With the metric change, the forwarding capacity between core and VSS might be reduced. Proper QoS marking should take care of critical traffic such as VOIP and video. The other non-critical traffic may share the bandwidth which may not be sufficient. In the campus network, keeping the default setting of OSPF and EIGRP for the links for Layer-3 MEC-based connectivity is usually a good practice.

Summary of ECMP vs. Layer-3 MEC Options

Overall Layer-3, MEC-based connectivity provides consistent convergence and options of path availability. As a result, implementing Layer-3 MEC to the core is the recommended design. See Table 3-4 for a summary comparison of the ECMP and Layer-3 MEC options.

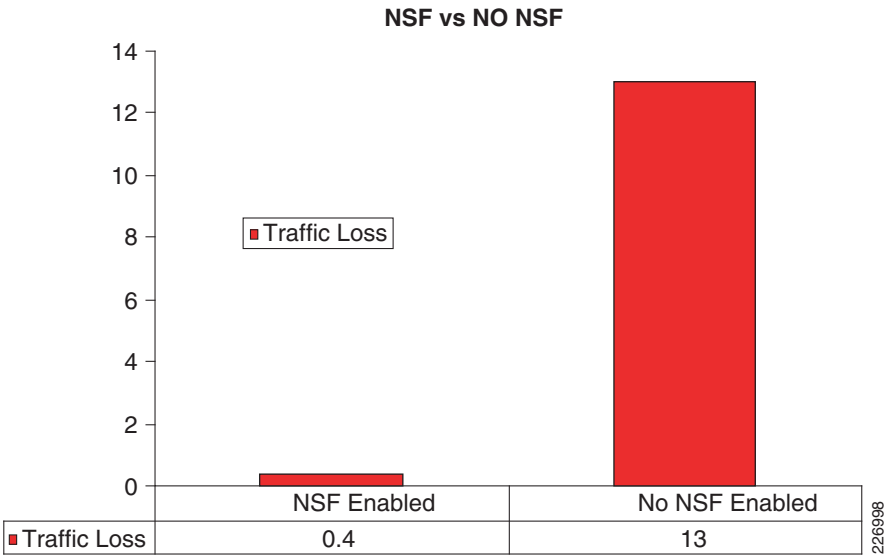
Table 3-4 Comparison of ECMP and Layer-3 MEC Options

Design Factors	ECMP to the Core	Layer-3 MEC to the Core
Recovery method for failure of a link in single VSS member	ECMP path switching to local member	Dependent of routing protocol configurations, route withdrawal-based path selection
Available path during link failure	Three	Dependent on routing protocol configuration—2 or 3

Routing Protocol Interaction During Active Failure

As discussed in the "Stateful Switch Over—Unified Control Plane and Distributed Data Forwarding" section on page 2-23, VSS consists of two supervisors: active and hot-standby. When the active supervisor fails, the SSO-based synchronization helps recover all the protocols that are SSO-aware. Routing protocol resiliency and recovery are not part of SSO. During the switchover, the hot-standby supervisor must reinitialize the routing protocol. As a result, neighboring routers notice the adjacency resets. This has a side-effect of removing routes for the downstream subnets learned from the VSS. In order to avoid such loss, the VSS must be configured with Non-Stop Forwarding (NSF) and the neighboring router must be NSF-aware. The impact of not enabling NSF is illustrated in Figure 3-36, which indicates as much as 13 seconds worth of traffic-loss can occur when NSF functionality is not enabled on the VSS and the adjacent routing devices.

Figure 3-36 NSF versus Non-NSF Voice Loss



Tip **Cisco strongly recommends that you run NSF on the VSS and adjacent routing nodes.**

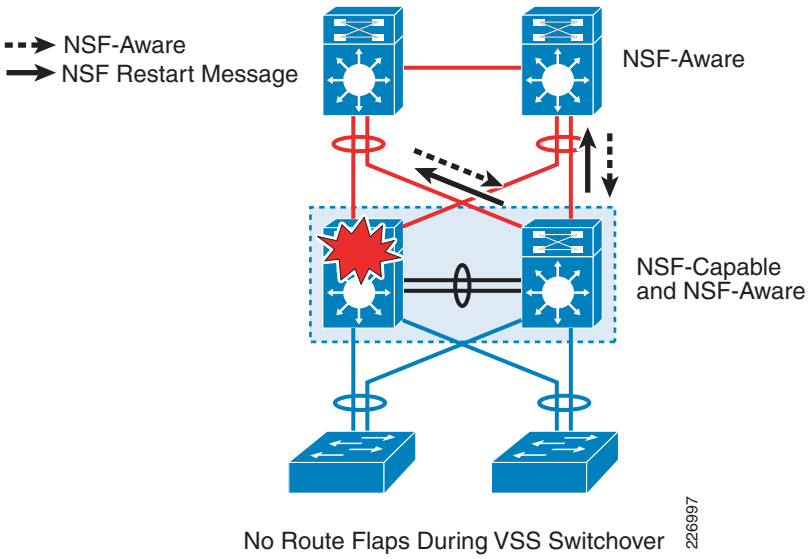
NSF Requirements and Recovery

NSF provides for graceful restart of the routing protocol so that the routing protocol remains aware of the control plane being recovered during the failover and does not react by resetting its adjacencies. NSF allows a router to continue forwarding data along routes that are already known, while the routing protocol information is being restored. SSO provides intelligent protocol recovery when switchover occurs, which is intended for continuous packet forwarding during switchover. However, if the routing protocol reacts to the failed event during the failure, the path through the restarting system is altered and no packet forwarding occurs, reducing the effectiveness of SSO. NSF is specifically designed to reduce the packet loss during switchover by maintaining the routing topology and gracefully updating the hardware forwarding table. Key components for NSF recovery include the following:

- *NSF-capable router*—A router that is capable of continuous forwarding during a switchover is *NSF-capable*. A NSF-capable route is able to rebuild routing information from neighboring NSF-aware or NSF-capable routers.
- *NSF-aware router /NSF helper*—A router running NSF-compatible software that is capable of assisting a neighbor router to perform an NSF restart. Devices that support the routing protocol extensions to the extent that they continue to forward traffic to a restarting (NSF capable) router are *NSF-aware*. A Cisco device that is NSF-capable is also NSF-aware. All Cisco switching platforms supporting routing capability support NSF-awareness.

See Figure 3-37 for a summary of the NSF recovery process.

Figure 3-37 NSF Recovery



NSF recovery depends on NSF-capable and NSF-aware router interaction at the routing protocol level during the supervisor failover. During failover, the routing protocol restarts in the newly active supervisor. Figure 3-37 shows the NSF-capable router undergoing NSF recovery. NSF recovery depends on CEF and routing protocol extensions. During the NSF recovery mode, the NSF-capable router detaches the routing information base (RIB) from CEF. This detachment allows independent control plane recovery while packets continue to be forwarded in the hardware.

As shown in Figure 3-37, the restarting router (hot-standby) notifies its NSF-aware neighbor that they should not reinitialize the neighbor relationship. The router receiving the restart indication puts the neighbor status in hold-down mode. In most cases, the restart indication consists of setting a restart flag in hello packets and sending hello packets at a shorter interval for the duration of the recovery process (this functionality influences the neighbor hello timer value as described in "NSF Recovery and IGP Interaction" section on page 3-56. Non-NSF-aware neighbors ignore the restart indication and disable the adjacency, leading to the removal of routes triggering packets loss. It is strongly recommended that you do not mix non-aware and NSF-aware routers in a network. The process of avoiding adjacency resets helps in route-convergence avoidance because no route recalculation occurs for the network advertised via an NSF-capable router. During NSF recovery, the routing protocol neighbor undergoes a special recovery mode. For more details on the routing protocol neighbor exchange, refer to the following URL: http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper0900aecd801dc5e2.shtml.

The above URL contains a generic NSF/SSO design guideline. The campus-related design choices are detailed in the section that follows.

NSF Recovery and IGP Interaction

The NSF is designed on the premise of convergence avoidance. This fits well with the principle of making the fault domain local and avoids the long route convergence dictated by Interior Gateway Protocol (IGP) timers. IGP neighbor timers are intended to provide alternate-path available via fast detection. For this reason, NSF-enabled environments must determine IGP neighbor dead-timer detection such that failover must avoid adjacency resets. The IGP dead-timer must be greater than the following:

SSO Recovery + Routing Protocol Restart + Time to Send First hello

As soon as the standby supervisor goes active, OSPF sends out fast-hello packets at two-second intervals to expedite the convergence time after a switchover. Enhanced IGRP has an independent mechanism for timer recovery. For more details on this operation, see the following URL:

http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper0900aecd801dc5e2.shtml

The recovery time involved with each event directly controls a lower (minimum) bound on hello timers. SSO recovery involves control plane initialization and executes (run state) protocols with synchronized databases. Routing protocol restart consists of multiple components initialization (start of routing process, rebuild of connected network and interaction with CEF process. Finally, the time it takes for processing and encapsulating the hello packet per-neighbor must be accounted for when sending hello packets with a restart flag. In a VSS, this time ranges between 9 to 13 seconds in a given validated environment.

Recommended timers for OSPF and Enhanced IGRP are shown in Table 3-5. This observation is based on the Cisco Catalyst 6500 Sup720 in the core with 3000 routes in the routing table.

Table 3-5 IGP Timer Requirements for NSF

Routing Protocol	Regular IOS	Modular IOS
Enhanced IGRP hello/hold sec	5/15—Default	5/15 Seconds
OSPF hello/dead sec	10/40—Default	10/60 Seconds

Note that the timer requirement of modular Cisco IOS is higher than that of native Cisco IOS, which may require extending the OSPF dead-timer from the default 40 seconds to a higher value.

Note The design requirement of BGP and IGP interaction is not evaluated with a campus-specific design goal and might require further tuning.

The timers summarized in Table 3-5 represent the minimum requirements for a best practice-based campus network.

Tip **Cisco strongly recommends *not* tuning below the values listed Table 3-5. All other NSF-related route timers should be kept at the default values and should not be changed.**

OSPF

The routing process runs only on the active supervisor. The standby supervisor does not contain OSPF-related routing information or a link-state database (LSDB), and does not maintain a neighbor data structure. When the switchover occurs, the neighbor relationships must be reestablished. The

NSF-capable router must undergo full restart of the neighbor state, but the neighbor router that is NSF-aware undergoes recovery modes with special restart bit signaling from the NSF-capable router in order to avoid the neighbor reset.

The following syslog messages of neighbor adjacency exchange illustrate the NSF restart on the NSF-aware router:

```
%OSPF-5-ADJCHG: Process 100, Nbr 10.120.250.4 on Port-channel6 from FULL to
EXSTART,OOB-Resynchronization
%OSPF-5-ADJCHG: Process 100, Nbr 10.120.250.4 on Port-channel6 from EXSTART to EXCHANGE,
Negotiation Done
%OSPF-5-ADJCHG: Process 100, Nbr 10.120.250.4 on Port-channel6 from EXCHANGE to LOADING,
Exchange Done
%OSPF-5-ADJCHG: Process 100, Nbr 10.120.250.4 on Port-channel6 from LOADING to FULL,
Loading Done
```

The OSPF NSF-aware router does not go from INIT to FULL; instead, it goes from FULL to the EXT-START with OOB-Resynchronization state. However, the NSF-capable peer undergoes full restart-sequencing through all six OSPF neighbor state exchanges.

Enhanced IGRP

Enhanced IGRP has a similar method for informing the neighbor on NSF restart; however, it does not have a transition state like OSPF. The following syslog message shows the NSF restart:

```
%DUAL-5-NBRCHANGE: IP-EIGRP(0) 100: Neighbor 10.120.0.211 (Port-channel2) is up: peer NSF
restarted
```

Configuration and Routing Protocol Support

Cisco NSF is supported on Enhanced IGRP, OSPF, BGP, and IS-IS routing protocols.

Note The design and operational guidance covers only OSPF and EIGRP.

The configuration for NSF-capability in Cisco Catalyst switches is very simple. It is enabled by the **nsf** keyword under each routing protocol instance. NSF-capable switches are automatically NSF-aware. The NSF-aware functionality is built into the routing protocol (if supported) and does not require specific configuration; however, it does require software releases supporting NSF-aware functionality. The following are example commands that illustrate NSF-capability configuration.

Enhanced IGRP:

```
Router(config)# router eigrp 100
Router(config-router)# nsf
```

OSPF:

```
Router(config)# router ospf 100
Router(config-router)# nsf
```

Note The Cisco IOS supports both IETF-based graceful restart extension as well as Cisco's version.

Monitoring NSF

The following **show** command examples illustrate how to observe NSF configuration and states in NSF-capable and NSF-aware routers based on type of routing protocol.

OSPF

The following **show** command output shows that OSPF is NSF-capable. The output statement supports link-local signaling indicates that this router is also NSF-aware. This example illustrates that the NSF is enabled and when the last restart occurred. Note that it indicates how long it took to complete the NSF restart. The NSF restart represents the routing protocol resynchronization—not the NSF/SSO switchover time—and does not represent the data-forwarding convergence time. See the "NSF Recovery and IGP Interaction" section on page 3-56 for the timer guidelines.

```
Router# sh ip ospf
 Routing Process "ospf 100" with ID 10.120.250.4
 Start time: 00:01:37.484, Time elapsed: 3w2d
 Supports Link-local Signaling (LLS)
! <snip>
LSA group pacing timer 240 secs
 Interface flood pacing timer 33 msecs
 Retransmission pacing timer 66 msecs
 Non-Stop Forwarding enabled, last NSF restart 3w2d     ago (took 31 secs)
```

Note that the key output is the Link Local Signaling (LLS) option output. Because OSPF NSF does not maintain OSPF state information on the standby supervisor, the newly active supervisor must synchronize its LSDB with its neighbors. This is done with out-of-band resynchronization (OOB-Resync).

The LR bit shown in the following example output indicates that the neighbor is NSF-aware and capable of supporting NSF-restart on local routers. Note that the first neighbor has undergone NSF recovery with OOB-Resync output. The OOB-Resync message is missing from the second neighbor because it has not gone through NSF recovery.

```
Router# sh ip ospf neighbor detail
 Neighbor 10.122.102.2, interface address 10.120.0.200
    In the area 120 via interface Port-channel6
    Neighbor priority is 0, State is FULL, 7 state changes
    DR is 0.0.0.0 BDR is 0.0.0.0
    Options is 0x50
    LLS Options is 0x1 (LR), last OOB-Resync 3w2d     ago
    Dead timer due in 00:00:07
Neighbor 10.122.102.1, interface address 10.120.0.202
    In the area 120 via interface Port-channel5
    Neighbor priority is 0, State is FULL, 6 state changes
    DR is 0.0.0.0 BDR is 0.0.0.0
    Options is 0x50
    LLS Options is 0x1 (LR)
    Dead timer due in 00:00:05
```

Enhanced IGRP

Enhanced IGRP has a similar recovery method. The supervisor becoming active must initialize the routing process and signals the NSF-aware neighbor with the RS bit in the hello and INIT packet. The Enhanced IGRP NSF capability can be found by using the **show ip protocol** command. The following output indicates that Enhanced IGRP is enabled with the NSF functionality, the default timers, and that it is NSF-aware. See the "NSF Recovery and IGP Interaction" section on page 3-56 for the timer guidelines.

```
Router# sh ip protocol
*** IP Routing is NSF aware ***
Routing Protocol is "eigrp 100 100"
! <snip>
EIGRP NSF-aware route hold timer is 240s
  EIGRP NSF enabled
     NSF signal timer is 20s
     NSF converge timer is 120s
```

Layer-3 Multicast Traffic Design Consideration with VSS

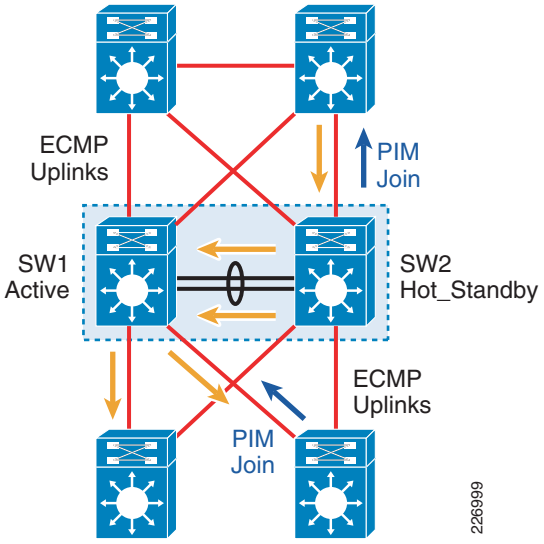
The VSS supports variety of multicast features and related design options, full validation and guidance encompassing all multicast features is beyond the scope of this design guide. This design guide covers the critical design points that can affect multicast traffic within the VSS-enabled campus. There are several other factors that influence multicast traffic behavior that are not addressed in this design, including Rendezvous Point (RP) placement, RP failover, VSS as RP, and so on. Chapter 4, "Convergence" covers important failure scenarios; however, large-scale validation with a multicast topology is beyond of the scope of this design guide. With a VSS-enabled campus, the following important design factors influence multicast traffic interaction at Layer-3:

- Traffic Flow with ECMP versus MEC
- Impact of VSS Member Failure with ECMP and MEC

Traffic Flow with ECMP versus MEC

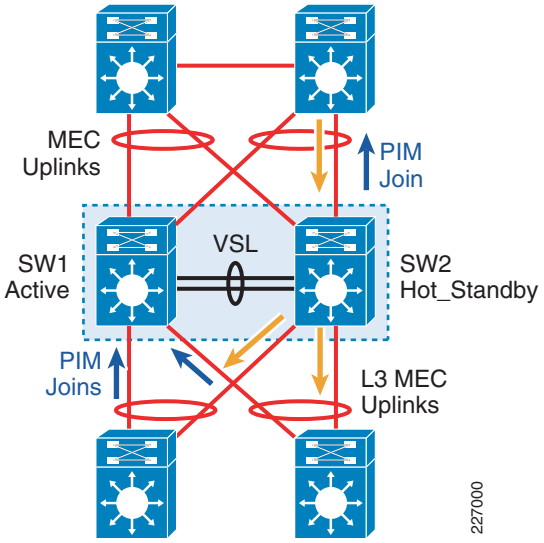
VSS represents a single multicast router. PIM joins are sent based on the highest PIM neighbor IP address (usually a first entry in the routing table for the given source) of the available ECMP paths in the routing table. Because the PIM join are send and received on different switches, the IIL (incoming interface list) and OIL (outgoing interface list) is formed asymmetrically such that the resulting multicast forwarding topology is build in the multicast traffic that can be forwarded over the VSL links. If the PIM joins are not sent to and from the same physical VSS member switch, multicast traffic can be passed across the VSL link as shown in the Figure 3-38.

Figure 3-38 Multicast Traffic Passing Across the VSL Link



In Figure 3-38, the bottom Layer-3 devices send a PIM join based on the highest routing table entry (highest PIM IP address) over the links connected to SW1; however, a PIM join is sent by the VSS on a link connected to SW2. Because only one PIM join is sent via VSS, the incoming interface for the multicast traffic is built on SW2. SW2 does not have an outgoing interface list built locally, SW1 builds the outgoing interface list. Because SW2 is a part of a VSS switch, a unified control plane knows that it has to replicate (egress physical replication) the traffic over the VSL bundle. This replication will be done for every single flow (*,g and s,g) and for every single outgoing interface list entry. This can put an overwhelming bandwidth demand on the VSL links as well as extend delay for multicast traffic. The solution is to use MEC-based connectivity as shown in Figure 3-39.

Figure 3-39 MEC-base Connectivity Option



In the MEC-based topology, it is still possible to have an asymmetrical PIM join process with the incoming interface list (IIL) and outgoing interface list (OIL) on distinct physical switches of the VSS. As shown in Figure 3-39, the IIL is built on SW2 versus OIL is built on SW1. However, both IIL and OIL is built on port-channel interface. The multicast traffic arrives at the SW2; even though PIM join came on SW1 traffic is forwarded by SW2. This is because of the fact that port-channel interface instance exist on both switches and by design. VSS always prefers locally-available interface to forward unicast and multicast traffic. Thus, multicast traffic will be forwarded over the local link instead of being forwarded over the VSL bundle. Due to the MEC configuration, it is possible that multicast traffic can select either of the link members based on the hashing result; however, because the topology implements MEC on either side, the traffic will not go over the VSL bundle. In the event of link failure, multicast traffic will pass across the VSL link and will experience local switch replication. This type of topology is possible in which VSS is deployed as a Layer-3 devices in multiple-tiers (e.g., multi-core or a routed access design). Use MEC uplinks from the access in routed access environments with multicast traffic.

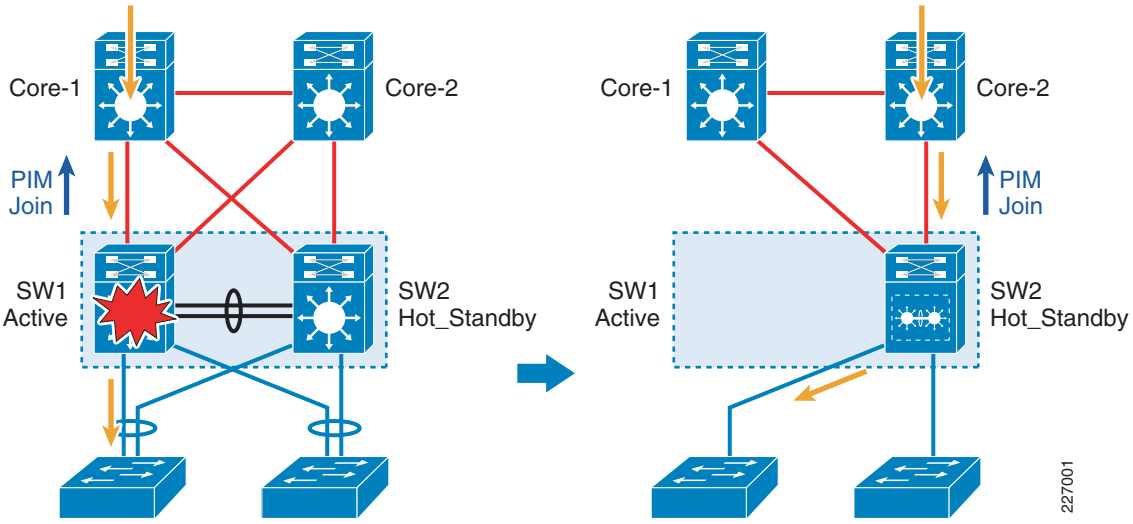
Tip **Cisco recommends using a Layer-3, MEC-based topology to prevent multicast traffic replication over the VSL bundle and avoids delay associated with reroute of traffic over VSL link.**

Impact of VSS Member Failure with ECMP and MEC

ECMP

With an ECMP-based topology, PIM joins are sent based on the highest PIM neighbor IP address (usually a first entry in the routing table for the given source) of the available ECMP paths in the routing table. Any time a PIM neighbor is disconnected (either due to link or node failures), the multicast control plane must rebuild the multicast forwarding tree by issuing a new PIM join on an available interface. For ECMP, multicast convergence depends on location of incoming interfaces. Figure 3-40 illustrates the behavior of multicast traffic flow where incoming interface is built on an active switch.

Figure 3-40 Rebuilding the Multicast Forwarding Tree



If the incoming interface for the multicast stream is built on the SW1 (Figure 3-40) and if SW1 fails before the multicast stream can be forwarded by SW2, SW2 requires the building of a new shortest-path tree (SPT) and the selection of incoming interfaces on newly active switch (SW2). Multicast data delivery will stop until the new path is discovered via multicast control plane convergence. The

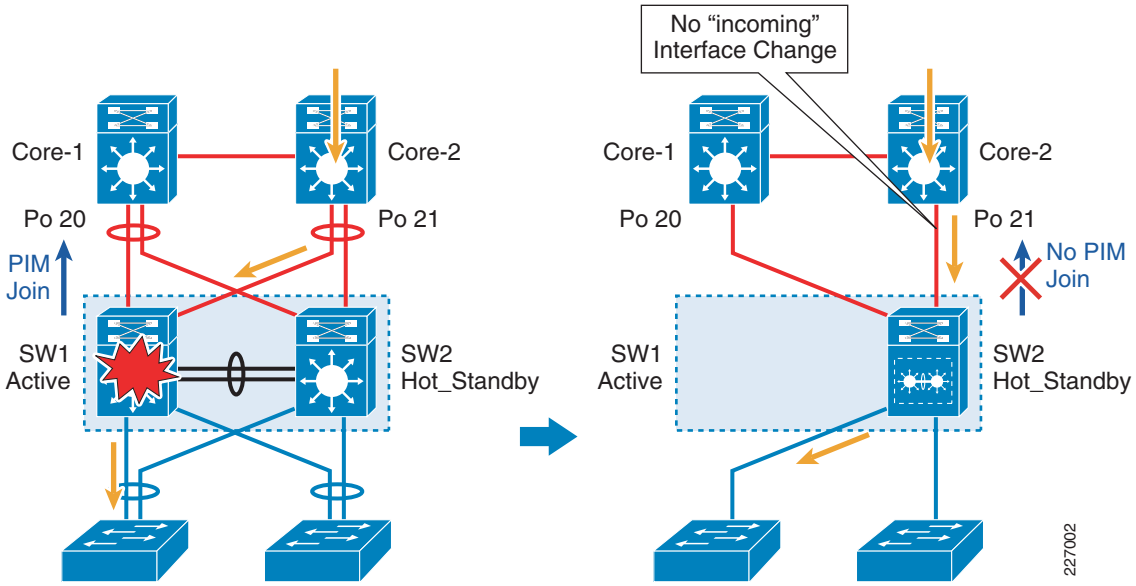
convergence related to this type of failure varies and can range from 2-to-3 minutes or higher depending on many factors such as rendezvous point (RP) reverse path forwarding) (RPF) check, unicast routing protocol convergence, etc.

If the failure is such that switch that is failing is not carrying incoming interfaces for the multicast traffic (not shown here), then the convergence ranges from 200-to-400 msec. In Figure 3-40, the incoming interface for the multicast flow is built on SW1 and if the SW2 fails, the incoming interface does not change; therefore, the multicast control plane does not require sending a new PIM join. Traffic continues forwarding in the hardware.

MEC

For the MEC-based connectivity, the failure of any member switch leaves one EtherChannel member available. The core router forwards the multicast data on the port-channel interface. When the VSS switch member fails, from the core router perspective, the incoming interface for multicast traffic remains unchanged. The core routers only have to rehash the flow to the remaining link member. This implies no state changes are reported to the multicast control plane. MMLS technology keeps the multicast states (*,g and s,g) synchronized, the switch hardware keeps switching multicast traffic. Multicast convergence is consistently in the range of 200-to-400 msec during switchover from active to hot standby. Figure 3-41 illustrates the behavior of multicast traffic flow in a MEC-based connectivity.

Figure 3-41 Failed Switch without Incoming Interfaces for Multicast Traffic



The PIM follows the same rule as ECMP in selecting the routed interface to send a join (highest PIM neighbor address). However, in Layer-3 MEC, the PIM join can select one of the link member based on the hashing value. As a result, the join can reach to either of the core routers. Figure 3-41 illustrate the behavior of multicast traffic flow where PIM join were sent to core-2 and thus incoming interface is built on core-2. The core-2 can choose to forward multicast flow to either SW1 or SW2 based on its hashing of multicast flow source and destination IP address. Figure 3-41 shows that the hashing result selected a link member connected to SW1. When the SW1 fails, both core routers remove one link member from the port-channel. Meanwhile, the SW2 assumes the role of active switch. From core-2’s perspective, the incoming interface for multicast has not changed since port-channel interface is still active with one link

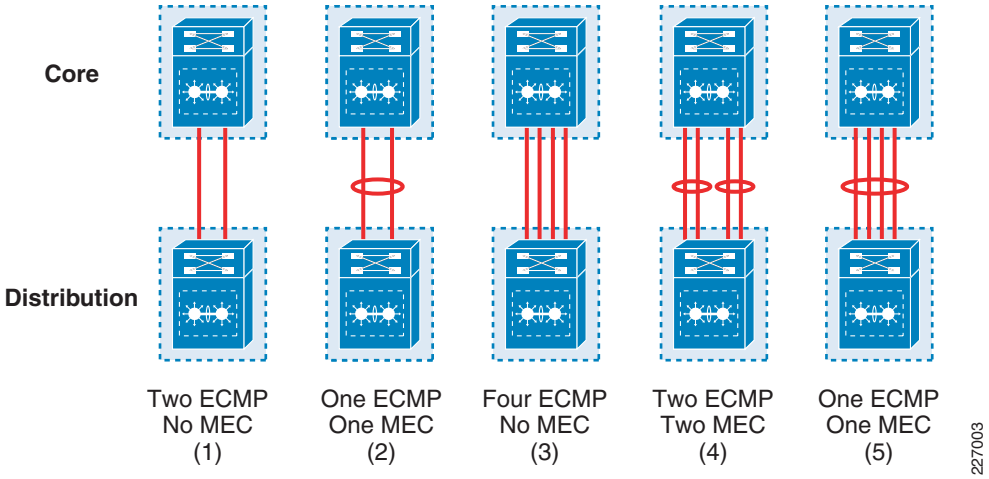
member connected to SW2. The core-2 continues to send multicast flow after the rehash. When SW2 receives the multicast flow, it forwards the traffic in hardware-based MMLS synchronization of multicast entries from old-active.

Multicast convergence is dependent on many factors; however, the key benefit of a VSS with MEC-based topology is that the node failure convergence can be reduced to below one second. In addition, contrast to ECMP-based connectivity, the multicast convergence is not dependent of where the incoming interface is built. In a multicast network with high mroute (*,G and S,G) counts, the recovery of certain failures might not yield convergence below one second.

VSS in the Core

The primary focus of this design guide is the application of VSS at the distribution layer; however, this section briefly covers its application in the core. All the design factors discussed so far also apply to VSS in the core. There are many factors to be considered in designing the core layer. The only design factor considered is the connectivity between core and distribution layer when both layers are using VSS. The connectivity option for VSS in the core and distribution layer consists of five major variations as shown in Figure 3-42. The figure depicts the logical outcome of virtualizing the core, links, and distribution layer.

Figure 3-42 VSS Core and Distribution Connectivity Options



Tip **In Layer-2 environment, single logical link between two VSS (option 5) is the *only* topology that is recommended; any other connectivity scenario will create looped topology.**

Out of the many design factors listed in Table 3-6, the factors highlighted in bold influence the most in deciding the best connectivity options for VSS core and distribution. Options 4 and 5 are discussed in some details in the context of those highlighted in Table 3-6.

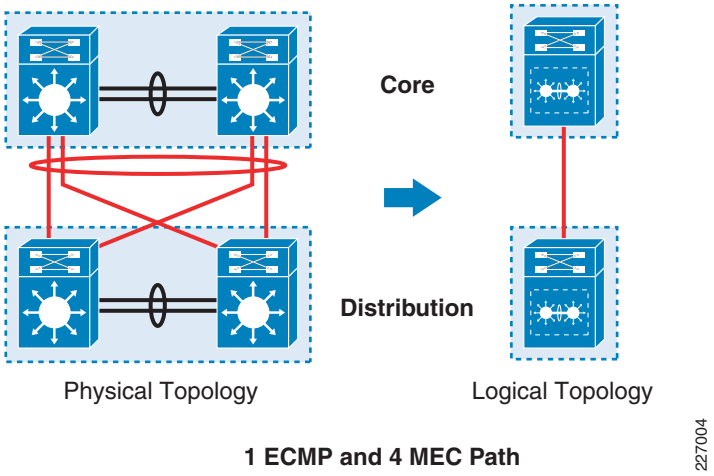
Table 3-6 Design Factors for Topology Options

Design Factors	Topology Options				
	Two ECMP Link, one from each chassis (1)	Two links, one from each chassis, one MEC (2)	Four link, fully meshed, ECMP (3)	Four links, fully meshed, two MEC, two ECMP (4)	Four links, fully meshed, one MEC (5)
Total physical links	2	2	4	4	4
Total logical links	0	1	0	2	1
Total layer 3 links	2	1	4	2	1
ECMP routing path	2	0	4	2	0
Routing overhead	**Double**	**Single**	**Quadrupled**	**Double**	**Single**
Reduction in Neighbor Counts	NO	Yes	NO	Yes	Yes
Single Link Failure Recovery	**Via VSL**	**via VSL**	**ECMP**	**MEC**	**MEC**
Multicast Traffic Recovery	**Variable**	**Consistent**	**Variable**	**Consistent**	**Consistent**
CEF Load-sharing	Yes	No	Yes	Yes	No
MEC Load-sharing benefits	No	Yes	No	Yes	Yes
Mixed Load sharing - CEF and MEC	No	No	No	Yes	No
Dual-Active Trust Support	**No**	**Yes**	**No**	**Yes**	**Yes**
Impact on Metric Change with Link Failure	None	None	None	Yes	None
Configuration and Troubleshooting Complexity	Medium	High	Medium	Medium	Low
Convergence with Single Link Failure	**Variable**	**Variable**	**Variable**	**~ 100 msec**	**~ 100 msec**
Recommended Best Practice Core routing Design	No	No	No	OK	Best

Single Layer-3 MEC—For Fully-Meshed Port-Channel Interface Links—Option 5

Figure 3-43 illustrates a single Layer-3 MEC intended for fully-meshed environments.

Figure 3-43 Single Layer-3 MEC for Fully-Meshed Environments



This design has the following advantages:

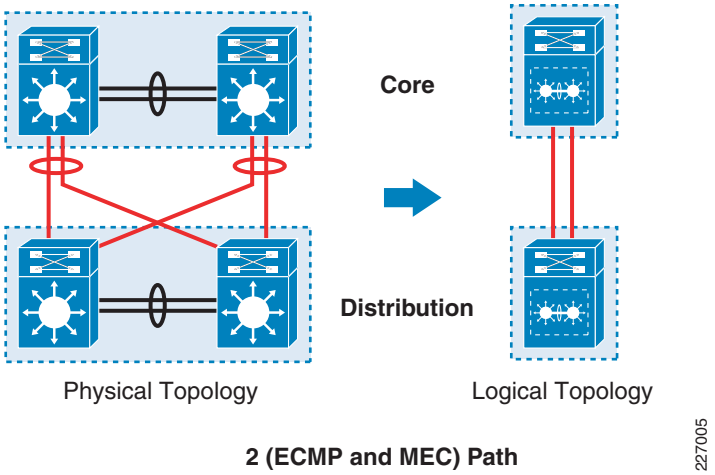
- Design inherently reduces routing control plane overhead in terms of routing topology and neighbor maintenance.
- Link failure or node failure is not dependent on routing table size, so that recovery is consistent.
- Link failure does not force traffic over the VSL bundle, reducing latency and congestion of VSL links.
- Routing updates or metric changes have little effect on path availability regardless of protocol-whether OSPF (auto-cost default or not) or Enhanced IGRP, because the one path through is via the single logical device.
- The convergence with link failure averages around 100 msec, due to the Fast Link Notification (FLN) feature available in WS-6708 hardware.
- Reduces configuration and troubleshooting overhead, due to single logical interface.

Multi-stage load-sharing method is possible if distinct Layer 3 (for CEF load-sharing) and Layer 2 (EtherChannel hash load-sharing) are used. This design option has only one logical Layer-3 path and CEF load sharing cannot be used. If the traffic pattern in the core is such that a finer level of traffic control is required through the use of Layer-3 and Layer-2 load-sharing method, then this design option might not be ideal. In a typical network, the need for such a fine level of granularity is not required, thus a single Layer-3 MEC solutions is usually the preferred choice due to simplicity of configuration, low control-plane overhead and a consistently low convergence time.

Two Layer-3 MEC—Two Layer-3 (ECMP) Port-Channel Interfaces (Each with Two Members)—Option 4

Figure 3-44 illustrates two Layer-3 MECs in an environment featuring two Layer-3 (ECMP) port-channel interfaces.

Figure 3-44 Two Layer-3 (ECMP) Port-Channel Interfaces with Two Members Each



This option has almost the same advantages as the single Layer-3 MEC connectivity, with the following additional considerations:

- Higher routing control-plane overhead
- Higher configuration and troubleshooting overhead
- Performance is dependent on metric changes and routing protocol configuration

The key advantage of this design is that it allows multistage (Layer-3-CEF and Layer-2 EtherChannel load sharing) to provide finer traffic control if required.

ECMP Full Mesh—Option 3

This option is discussed in the first part of this section. It is not generally a preferred option among the available choices.

Square and Non-Full Mesh Topologies—Single Link from Each Chassis—Option 1 and 2

This is a least preferred topology because it has many disadvantages compared with the preceding design options. The dependency of traffic going over the VSL link during link failure is the biggest reason option 1, and 2 are the least preferred options.

Routed Access Design Benefits with VSS

The routed access design is an alternative to multilayer design (see Chapter 1, "Virtual Switching Systems Design Introduction"). Routed access simply extends the Layer-3 boundary to the access-layer. The routed access design in many ways is similar to VSS-enabled campus design. Both models solves the same problems by simplifying topologies and reducing the topology changes introduced with link and nodal failures. The following are some of the common benefits:

- Ease-of-implementation, reducing configuration
- No dependency on FHRP
- No matching of STP/HSRP/GLBP priority
- No Layer-2/Layer-3 multicast topology inconsistencies via single designated router
- Single control plane and ease-of-managing devices and the fault domain
- Consistent convergence time and convergence times not being dependent on GLBP/HSRP tuning

Common benefits of both design approaches leads to an obvious question of application of VSS in routed access design. VSS in Layer-2 domains make a significant contribution in terms of improving multilayer implementations by allowing VLANs to be spanned while removing associated risks. Using the VSS in a routed access design might also be beneficial. The following sections first illustrate the critical components failure recovery in routed access and VSS. Finally, the section summarizes the benefits of deploying VSS in routed access design.

Distribution Layer Recovery

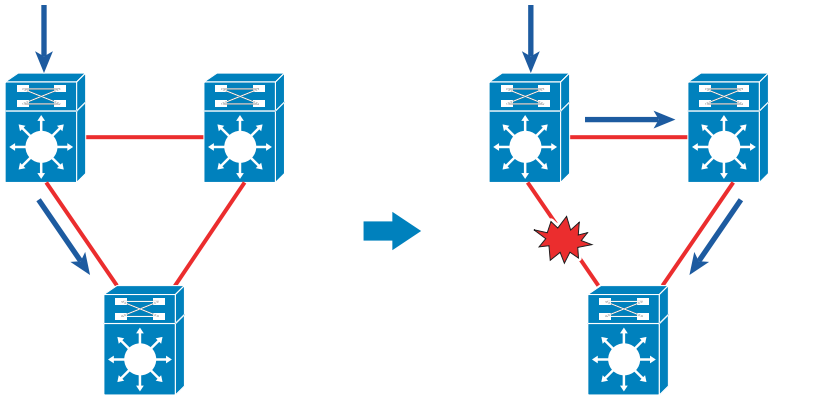
Routed Access

The major benefit of a routed access design is extremely fast convergence (200 msec). In the case of link failure, recovery depends on rerouting over the Layer-3 link between distribution nodes. The recovery is specifically dependent on the following factors:

- Time to detect interface down
- Time for routing protocol to converge in finding alternate route

The detection time depends on type of interfaces (fiber or copper) and physical aspects of the interface configuration. The only component that can be controlled is the behavior of routing protocol. The faster the routing protocol can detect, announce, and then react to the failure event determines the speed of the convergence. Figure 3-45 illustrates the general recovery process.

Figure 3-45 Generalized Routing Protocol Recovery Process



For OSPF, convergence depends on the tuning of the SPF and LSA timers to a level that is below one second, summarization of access-layer subnets, and the type of area defined within access layer. With all three critical components configured properly the, convergence as low as 200 msec can be achieved.

Enhanced IGRP does not have any timer dependency; however, Enhanced IGRP-stub and summarization are critical for convergence.

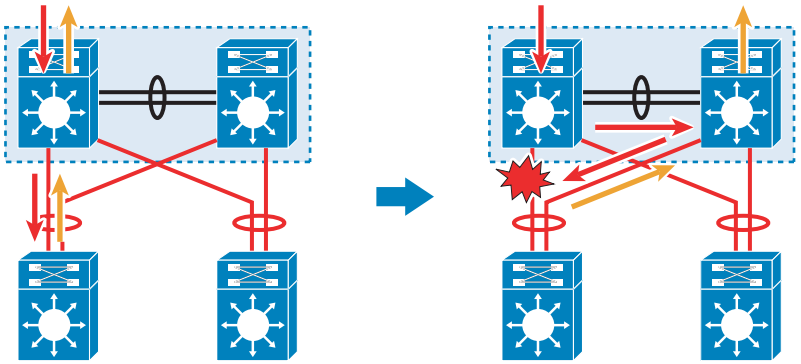
Detail of related failures and configuration guidance are available in the *Routed Access Design Guide* listed in Appendix B, "References."

VSS Recovery

As with routed-access design in which the access-layer link recovery requires rerouting over the Layer-3 link, the VSS has a similar behavior for downstream traffic flow recovery (see "Traffic Flow in the VSS-Enabled Campus" section on page 3-5). In the VSS scenario, the link failure in the access layer will force the downstream traffic to be rerouted over the VSL. The upstream traffic flow will be recovered via EtherChannel (rehashing the flows on the remaining link at the access-layer).

The VSS deployed at the distribution eliminates the dependencies on routing protocol convergence. The benefit of MEC-based topology is that member link failure does not bring down the port-channel interface. When an access-layer link fails, no routing update is required for downstream traffic because the port-channel interface has not been disabled (see Figure 3-46). As a result, the core routers routing topology does not go through convergence.

Figure 3-46 Downstream Recovery without Routing Changes

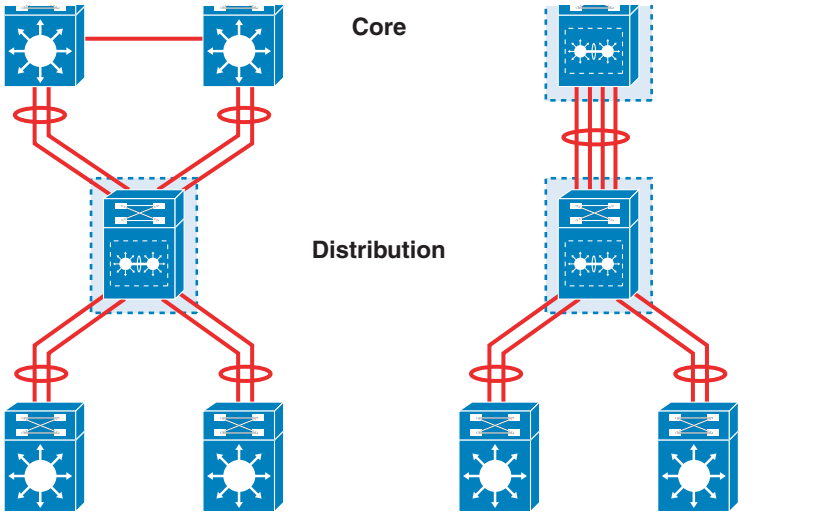


OSPF/EIGRP Downstream Recovery without Route-Change

227007

Use of VSS at the distribution layer enhances the routed-access design goal of creating a simpler, faster-converging environment to a more advanced level. It reduces the complexity of configuration and of fault domain. This design guide has not covered all aspects of the VSS application in routed access campus design; however, validation of critical failure recovery described in preceding section confirms the assertion of its benefits in such design. Figure 3-47 depicts two models of routed access design with VSS. The first one depicts the use of VSS at the distribution layer, the second part depict the use of VSS at distribution as well as in the core, further simplifying topology with Layer-3 MEC.

Figure 3-47 VSS-Enabled Routed Access



VSS at Distribution Layer VSS at Distribution Layer and Core Layer

227008

VSS Enabled Routed Access

Advantages of VSS-Enabled Routed Access Campus Design

The advantages associated with VSS-enabled routed-access campus design include the following:

- Simplified and consolidated configuration reduces operational complexity leaving less to get wrong.
- Routed access has simplified topology and recovery. VSS further simplifies your design because it offers a single logical devices at each layer end-to-end. Single logical routers eliminate most of the dual node inefficiencies with routing at the control plane along with many side benefits, such as a reduced control plane load associated with topology database, peering, and neighbor relationships.
- Allows for greater flexibility in user-adopted design. VSS with routed access gives flexibility in configuring OSPF and Enhanced IGRP tuning requirements. Sub-second timer configuration is not as much of a requirement in the core and distribution layers because access-layer link failures do not require route recalculation. This simplifies configurations and removes the topological dependency of extending timers to non-campus devices
- Link-member failure does not bring down the routed interface. This eliminates the need to advertise routing changes to the core and beyond. In a traditional design, route summarization reduced the unnecessary route churn observed with such events. VSS reduces the need for route summarization In best practice-based design, summarization is still recommended to reduce the control plane instability and to address failures that are might not be solved by VSS. Often in a enterprise network it is difficult to summarize the IP subnet due to installed base and inheritance of legacy planning, VSS with routed access offer a flexibility to existing network as the criticality of summarization is reduced.
- Redundant supervisors provide resiliency via SSO-enabled protocols resulting in consistent recovery during the failover of nodes at the distribution layer. For example, the implementation of OSPF or Enhanced IGRP NSF/SSO eliminates the dependency of convergence on a routing table size (see Figure 3-32).
- A single logical multicast router in the core and distribution layers simplifies the multicast topology resulting in convergence below one second in the core and distribution layers for a nodal failure.

Hybrid Design

The VSS capability of extending Layer-2 domains and providing enhanced functionality in the Layer-3 domain allows you to create a *hybrid* design approach in which multilayer and routed-access designs can be merged into a fully integrated design. The benefits of each design can be exploited to serve specific technology or business requirement. In hybrid design, VLANs requiring spanning multiple closet can be defined at VSS and VLANs that do not required spanning VLANs are routed and thus it is defined at the access-layer. This can be achieved through allowing trunked configurations between the VSS and access-layer where a spanned VLAN are trunked and a non-spanned VLANs are routed. Some of the functional VLANs that require spanning multiple access-layer switches are as follows:

- Network Virtualization (guest VLAN supporting transient connectivity, intra-company connectivity, merger of companies and so on)
- Conference, media room and public access VLANs
- Network Admission Control (NAC) VLAN (quarantine, pasteurization, and patching)
- Outsource group and inter-agency resources requiring spanned VLANs
- Wireless VLANS without centralized controller
- Network management and monitoring (SNMP, SPAN)

Some of the VLANs that can be routed are data and voice VLANs or any other connectivity that is confined to an access-layer switch.

Note The hybrid design approach has not been validated in this release of the design guide.

CHAPTER 4

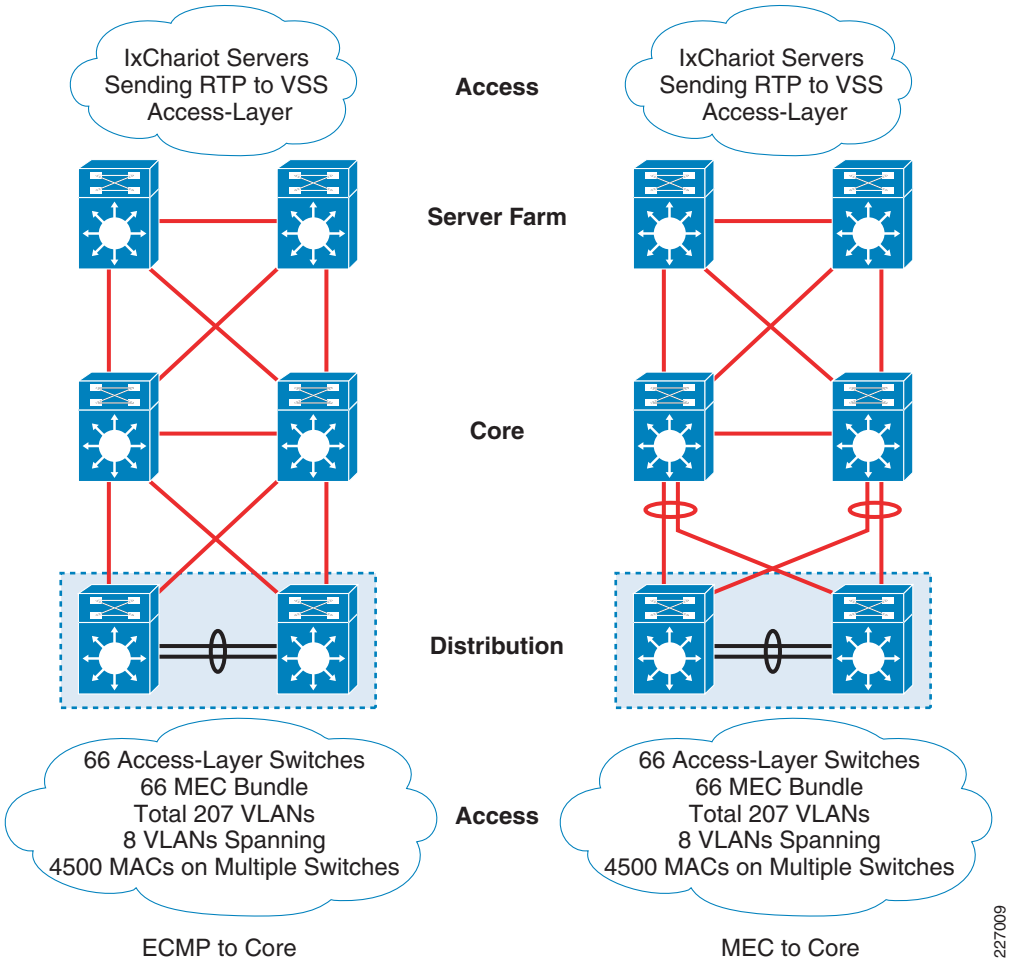
Convergence

This chapter covers convergence results and traffic flow during VSS component failures. It uses and enables all the validated best practices detailed in the previous chapters. The convergence section heavily uses the "VSS Failure Domain and Traffic Flow" section on page 3-9 in identifying the type of components and technology involved affecting convergence. Validation was completed for most failure type combinations and protocol/topology iterations. This chapter does not characterize every failure and every convergence scenario; rather, it addresses critical and common failures in campus networks.

Solution Topology

The VSS solution-validation environment covers ECMP and MEC topologies and related configuration best practices. Figure 4-1 provides a summary of the general VSS environment. This three-tier hierarchy for campus connectivity is derived from best practices established in previous design guides. In terms of connectivity beyond the core, the reference topology consist of serverfarm switches connected via ECMP. The configuration used for convergence results is based on the validated best practices developed during the creation of this design guide. All protocol configurations implement the default timer settings and behavior, except when specifically noted.

Figure 4-1 VSS Solution Topology



Software and Hardware Versions

Table 4-1 summarizes the applicable software and hardware versions associated with the VSS environment that is addressed in this document.

Table 4-1 VSS Software and Hardware Summary

Platform	Software Release	Hardware Configuration	Device Role
Catalyst 6500-E	12.2(33)SXH2(a)	Sup720-10GE	VSS DUT Distribution Layer
		6708-10GE	
		6724-100/1000	
Catalyst 6500-E	12.2(33)SXH1	Sup720 6708-10GE	Core Layer
Access Layer			
Catalyst 6500	Native 12.2(33)SXH	Sup32-8GE	DUT
		6148-GE-TX	

Table 4-1 VSS Software and Hardware Summary (continued)

Platform	Software Release	Hardware Configuration	Device Role
Catalyst 6500	CatOS 8.6	Sup32-8GE 6148-GE-TX	DUT
Catalyst 6500	Modular 12.2(33)SXH	Sup32-8GE 6148-GE-TX	DUT
Catalyst 4500	12.2(40)SG	SupV 10GE	DUT
Catalyst 3750	12.2(40)SE	5 member stack	DUT
Catalyst 3560	12.2(40)SE	Standalone	DUT
Catalyst 3550/3560	12.2(37)SE	60 switches	Control plane load

VSS-Enabled Campus Best Practices Solution Environment

Table 4-2 through Table 4-4 provide a summary of the campus-related VSS implementation best practices that are described in this document.

Table 4-2 VSS Environment

Campus Environment	Validated Campus Environment	Comments
VSL links	Diversified on supervisor port and WS-X6708	
NSF capability configured	Yes	
Topology	ECMP & MEC	
Number of routes	3000	
CEF load-sharing	Yes	
Default VSLP timers	Yes	
Use virtual MAC	Yes	
Port-channel load-share	src-dst-ip enhanced	

Table 4-3 Layer-3 Domain

Campus Environment	Validated Campus Environment	Comments
Routing protocol	Enhanced IGRP and OSPF	
NSF awareness in the core	YES	
Enhanced IGRP hello and hold timers	Default	5/15
OSPF hello and hold timers	Default	10/40
Multicast routing protocol	PIM-SPARSE	
Rendezvous point	ANYCAST IP in CORE	
Number of multicast groups	80	

Table 4-3 Layer-3 Domain (continued)

Campus Environment	Validated Campus Environment	Comments
Multicast SPT threshold	Default	
Topology	ECMP and MEC	
Number of routes	3000	
Route summarization	Yes	
CEF load-sharing	Yes	
Core connectivity	WS-X6708 10G	
Core devices	Standalone 6500	

Table 4-4 Layer-2 Domain

Campus Environment	Validated Campus Environment	Comments
STP	RPVST+	
Number of access-layer switch per distribution Block	66	66 MEC per-VSS
Total VLANs	207	
VLAN spanning	8 VLANs	Multiple switches
Number of network devices per distribution block	~ 4500	Unique per-host to MAC ratio
MAC address for Spanned VLANs	720 MAC/VLANs	
VLAN confined to each access-layer Switch	140	Voice and data per access-layer switch
Unique IP application plows	8000 to 11000	
EtherChannel Technology	PAgP and LACP	
EtherChannel mode—PAgP	Desirable-Desirable	
EtherChannel mode—LACP	Active-Active	
PAgP and LACP timers	Defaults	
Trunking mode	Desirable-Desirable	
Trunking type	802.1Q	
VLAN restricted per-trunk	Yes	
UDLD mode	Normal	
Access-switch connectivity	Supervisor uplink port or Gigabit uplink	

Convergence and Traffic Recovery

In this section, the first part illustrates the failures associated with VSS, the later part includes the failure associated with the routing and core component in VSS-enabled campus. Each failure type includes a table depicting the failure recovery method for both unicast and multicast traffic. The following brief descriptions summarize the traffic pattern and recovery methods associated with VSS:

- *Unicast Upstream Traffic*—Refers to traffic originated at the access-layer and destined for the severfarm switches.
- *Unicast Downstream Traffic*—Refers to traffic originated at the serverfarm switches and destined for the access-layer switch.
- *Multicast Traffic*—Refers to sources connected to serverfarm switches and receiver joins originated at the access-layer. This usually follows the unicast downstream convergence.
- *EC Recovery or Failover*—Refers to EtherChannel link failure and the rehashing of traffic to the remaining member link.
- *ECMP*—Equal Cost Multi-Path (ECMP) refers to fully meshed, routed-interface topology providing a load-sharing CEF path in hardware.
- *Local CEF*—VSS-specific CEF switching behavior with which the local CEF path is preferred over peer switch path.
- *Multicast Control Plane*—Refers to convergence related to multicast components, including (but not limited to) PIM recovery, repopulation of mroute (**,g* and *s,g*) and Reverse Path Forwarding (RPF), building a shortest path tree, and so on.
- *IIL and OIL on Active or Hot-Standby*—Refers to location of incoming multicast traffic (IIL) determined via the RPF check and of the outgoing interface list (OIL) used to switch the multicast traffic on a given VSS member switch. In MEC-based topologies, for a given multicast group, the IIL and OIL is always on same member of a VSS pair. The change in routed interface status usually triggers multicast control plane changes.
- *Stateful Switch Over (SSO)*—SSO refers to a method of recovering from active to hot-standby.
- *Multicast Multilayer Switching (MMLS)*—MMLS refers to the unique method of replicating (**,g* and s,g) entries into the hot-standby supervisor. It allows the multicast traffic to be forwarded in hardware during a supervisor failure or traffic redirection to be triggered during link failure.

VSS Specific Convergence

Active Switch Failover

An active failover is initiated by one of the following actions:

- Application of the redundancy force-failover command
- Physically removing an active supervisor from service
- Powering down an active supervisor

The convergence remains the same for any of the above methods of active failover. The process of switching over from one VSS switch member to the other (active to hot-standby) is influenced by many concepts and design considerations discussed in the preceding sections. The following sequence of events provide a summary of the failover convergence process:

1. Switchover is invoked via software CLI, removing supervisor, powering down active switch, or system initiation.
2. The active switch relinquishes the unified control plane; the hot-standby initializes the SSO control plane.
3. All the line cards associated with active switch are deactivated as the active chassis reboots.
4. Meanwhile, the new active switch (previous hot-standby switch) restarts the routing protocol and starts the NSF recovery process.
5. In parallel, the core and access-layer rehash the traffic flow depending on the topology. The unicast traffic is directed toward the new active switch, which uses a hardware CEF table to forward traffic. Multicast traffic follows the topology design and either rebuilds the multicast control plane or uses the MMLS hardware table to forward traffic.
6. The NSF and SSO functions become fully initialized, start learning routing information from the neighboring devices, and update the forwarding and control plane protocols tables as needed.

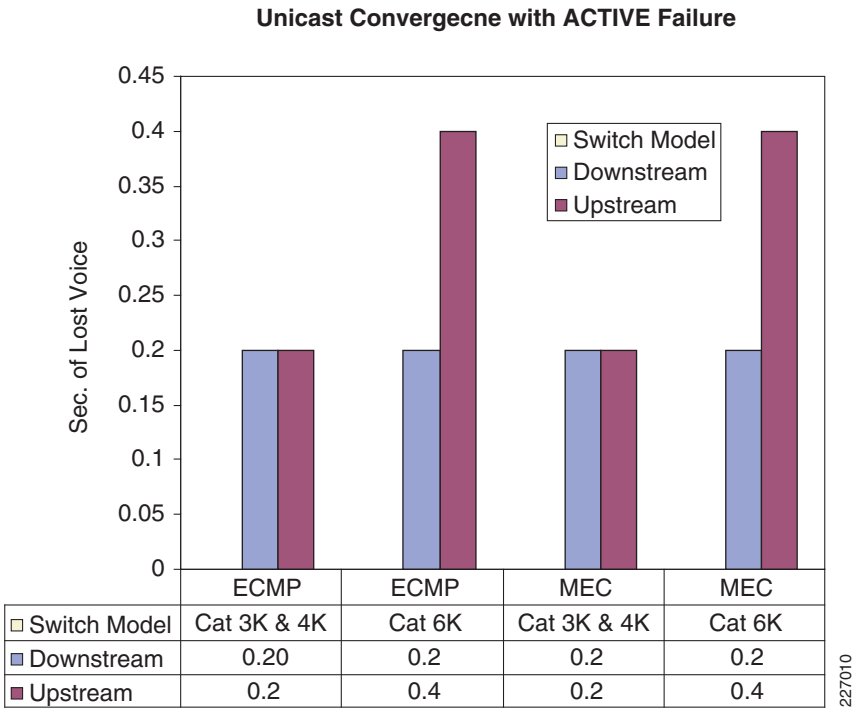
The active failure convergence is validated with Enhanced IGRP and OSPF routing protocol with the topology combination of ECMP or MEC connectivity to the core. The recovery methods for both routing protocol remains the same as summarized in Table 4-5.

Table 4-5 Active Failure Recovery

Topology	ECMP	MEC	Common Recovery
Unicast Recovery Method			
Unicast Upstream	EC failover at access	EC failover at access	SSO
Unicast Downstream	CEF	EC failover at core	SSO
Multicast Recovery Method			
IIL on active switch	Multicast control plane	EC failover at core	MMLS
IIL on hot-standby switch	MMLS	EC failover at core	MMLS

The convergence losses are similar for both Enhanced IGRP and OSPF. Figure 4-2 shows that the average convergence is at or below 200 msec for either the Cisco Catalyst 3xxx or Cisco Catalyst 45xx switching platforms, and around 400 msec for Catalyst 65xx switching platform. One reason that the Cisco Catalyst 6500 has a little higher level of loss is that the distributed fabric-based architecture must consider dependencies before the flows can be rerouted to the available member link.

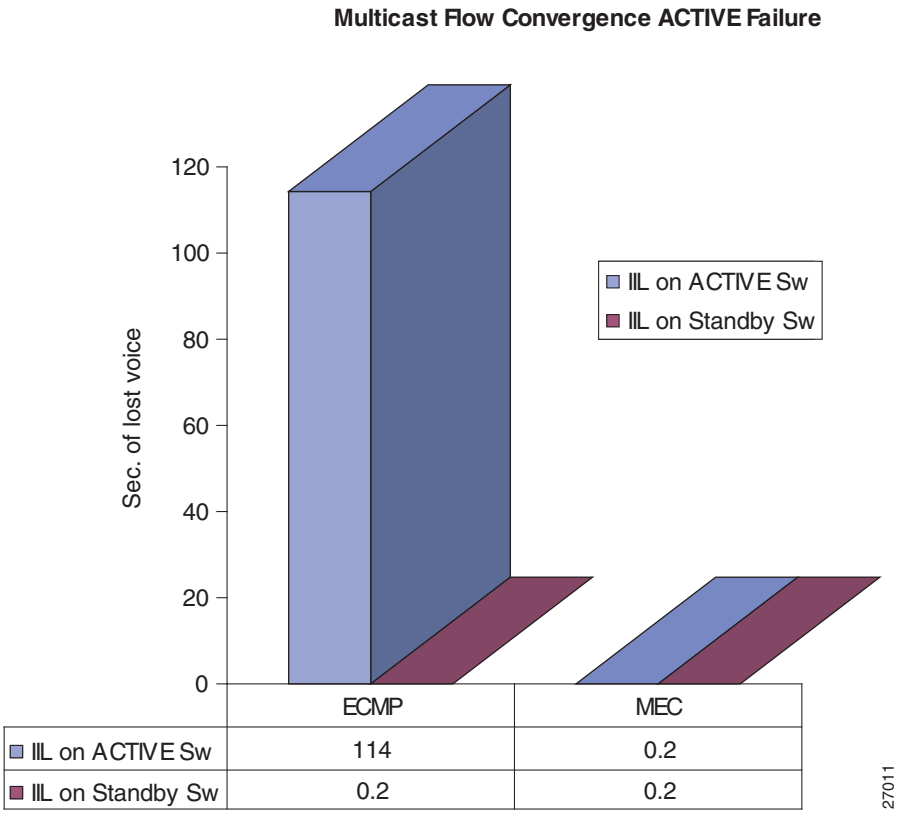
Figure 4-2 Active Failure Convergence



	ECMP	ECMP	MEC	MEC
Switch Model	Cat 3K & 4K	Cat 6K	Cat 3K & 4K	Cat 6K
Downstream	0.20	0.2	0.2	0.2
Upstream	0.2	0.4	0.2	0.4

The multicast convergence shown in Figure 4-3 depends on the topology and where the incoming interface list (IIL) is built. This design choice is discussed in the "Layer-3 Multicast Traffic Design Consideration with VSS" section on page 3-59. Note that the multicast convergence is higher with ECMP, depending on the combination of the IIL list location and switch failure.

Figure 4-3 Multicast Convergence



Hot-Standby Failover

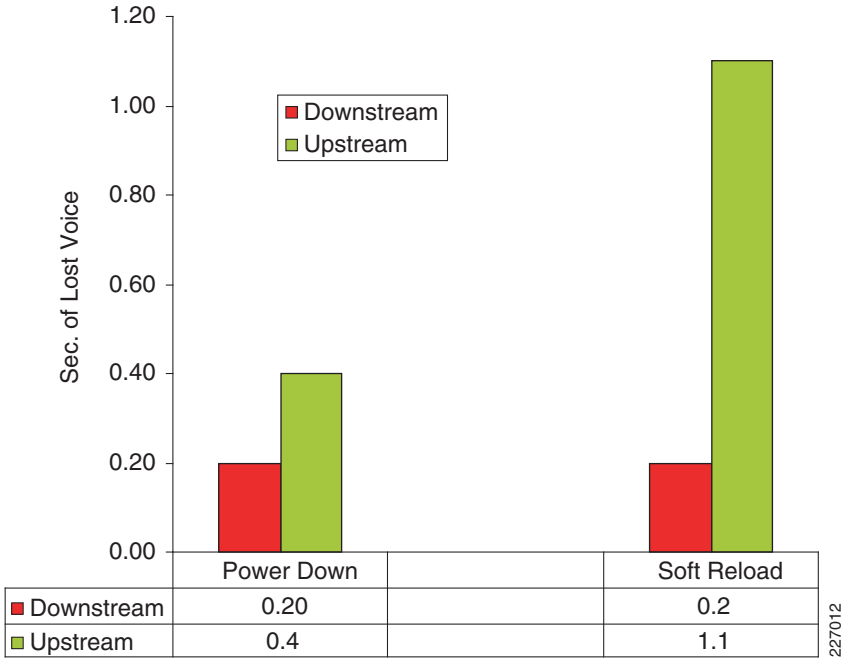
Hot-standby failover does not introduce control plane convergence because it is not actively responsible for managing various protocols and their update. However, in the ECMP topology, the neighbor connected via the hot-standby switch will reset and links connected to hot-standby goes down. The recovery of traffic is the same as an active failure except that the SSO initialization delay is not present. See Table 4-6.

Table 4-6 Hot-Standby Failure Recovery

Topology	ECMP	MEC	Common Recovery
Unicast Recovery Method			
Unicast Upstream	EC failover at access	EC failover at access	
Unicast Downstream	ECMP failover (CEF) at core	EC failover at core	
Multicast Recovery Method			
IIL on active	No impact	No Impact	MMLS
IIL on hot-standby	Multicast control plane	EC	MMLS

As shown in Figure 4-4, the convergence is under one second upon a loss of power, whereas a software failure causes slightly higher packet loss.

Figure 4-4 Comparison of Hot-Standby Convergence Characteristics



The upstream convergence shown in Figure 4-4 is specifics to the way the network connectivity is configured. In a validated design, the uplink connecting the core resides on DFC WS-X6708 line card. The upstream convergence is variable and dependent on the position of line card in a chassis on which given connectivity is configured. Table 4-6 on page 4-8 does not cover the intermittent losses or recovery involved with hot-standby software reload. During the software reload of the hot-standby, the Cisco IOS software removes line card sequentially in ascending slot order after the supervisor card is removed (lower numbered slot is removed first). This behavior is illustrated in the syslogs output below. For the given scenario, slot 2, where the 10-Gigabits connectivity to the core resides, is taken offline. Meanwhile, the access-layer connectivity (slots 7,8, and 9) is still up; therefore, the access-layer switches keep sending upstream traffic to the hot-standby. This traffic is rerouted to VSL as no direct upstream connectivity to the core exists. This contributes to the higher losses associated with the upstream traffic. If you move the core-connected line card to the last slot, the losses will be reversed because the access-line card is powered down first. This forces the access-layer switch to reroutes traffic on remaining link on EtherChannel. However, downstream traffic is still being received at the VSS (until the line card is removed) is rerouted over the VSL link. Therefore, in this case, the downstream losses will be higher.

```
Nov 14 08:43:03.519: SW2_SP:  Remote Switch 1 Physical Slot 5 - Module Type LINE_CARD
removed
Nov 14 08:43:03.667: SW2_SP:  Remote Switch 1 Physical Slot 2 - Module Type LINE_CARD
removed
Nov 14 08:43:04.427: SW2_SP:  Remote Switch 1 Physical Slot 7 - Module Type LINE_CARD
removed
Nov 14 08:43:04.946: SW2_SP:  Remote Switch 1 Physical Slot 8 - Module Type LINE_CARD
removed
Nov 14 08:43:05.722: SW2_SP:  Remote Switch 1 Physical Slot 9 - Module Type LINE_CARD
removed
Nov 14 08:47:09.085: SW2_SP:  Remote Switch 1 Physical Slot 5 - Module Type LINE_CARD
inserted
```

```
Nov 14 08:48:05.118: SW2_SP:  Remote Switch 1 Physical Slot 2 - Module Type LINE_CARD
inserted
Nov 14 08:48:05.206: SW2_SP:  Remote Switch 1 Physical Slot 7 - Module Type LINE_CARD
inserted
Nov 14 08:48:05.238: SW2_SP:  Remote Switch 1 Physical Slot 8 - Module Type LINE_CARD
inserted
Nov 14 08:48:05.238: SW2_SP:  Remote Switch 1 Physical Slot 9 - Module Type LINE_CARD
inserted
```

Hot-Standby Restoration

Traffic recovery depends on two factors:

- *Slot order*—Slot order matters because the line cards power up in a sequential order.
- *Card type*—The type of line card also affects the forwarding state. The DFC line card takes longer to boot.

If the core connectivity is restored first, then downstream traffic has multiple recoveries. The first recovery is at the core layer-either CEF (ECMP)- or EtherChannel-based recovery. Second recovery occurs when the traffic reaches the VSS. Once at the VSS, it must reroute over the VSL link because the line card connected to access-layer will have not yet come online. Similarly, the upstream traffic has multiple recoveries if the access-layer line cards come up first.

Multicast recovery for ECMP has no initial impact because the incoming interface (if it is built on active switch) does not change; however, it is still possible to have new PIM join sent out via the newly added routes (as hot-standby ECMP links recovers triggering RPF check) that will induce multicast control-plane convergence. For MEC-based topologies, recovery is based on the EtherChannel hashing result at the core when a hot-standby-connected link is added to the Layer-3 MEC. It is then possible to reroute traffic at the VSL based on the access-layer line card boot status. Refer to Table 4-7.

Table 4-7 Hot-Standby Switch Restoration Recovery

Topology	ECMP	MEC	Common Recovery
Unicast Recovery Method			
Unicast Upstream	Variable	Variable	See above explanation
Unicast Downstream	Variable	Variable	See above explanation
Multicast Recovery Method			
IIL on active	Variable, Multicast control plane	Variable—EC hashing line card boot status	
IIL on hot-standby	N/A	N/A	Standby restoration

The factors described above can cause the variable convergence in a VSS-based environment. In general, restoration losses are in the range of 700 msec to four seconds. These losses are much better than standalone-node restoration because of the ARP throttling behavior described in the document at the following URL:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/HA_recovery_DG/campusRecovery.html

VSL Link Member Failure

The VSL bundle is a port-channel interface. Its failure convergence and recovery characteristics are similar to MEC-link failure characteristics. The VSL-link failures cause a rehashing of traffic (both user data and control link) over the remaining link. The Layer-3 and Layer-2 MEC topology provides symmetrical local forwarding and failure of the link does not affect the user data reroute over the VSL link. However, in a topology where access-devices are connected to only one member or one of the uplink form access-layer switch fails, VSS will reroute half of the downstream traffic to traverse VSL links. The failure of link in single-homed topology will affect the user data convergence. The convergence of data traffic could be below one second to several seconds. A sub-optimal topology created by non-MEC-based design leads to sub-optimal convergence. Implement the dual-homed MEC design for all the devices connected to a VSS.

Line Card Failure in the VSS

A line card failure can occur for the following reasons:

- *Hardware failure*—requires hardware replacement, typically a planned event.
- *Software failure*—resetting a line card might resolve this issue.

The failure of a line card essentially creates a single-homed or orphaned connectivity link to the VSS as described in the "VSS Failure Domain and Traffic Flow" section on page 3-9. This failure will reroute either upstream or downstream traffic depending on whether the linecard is connected to the access or core layer.

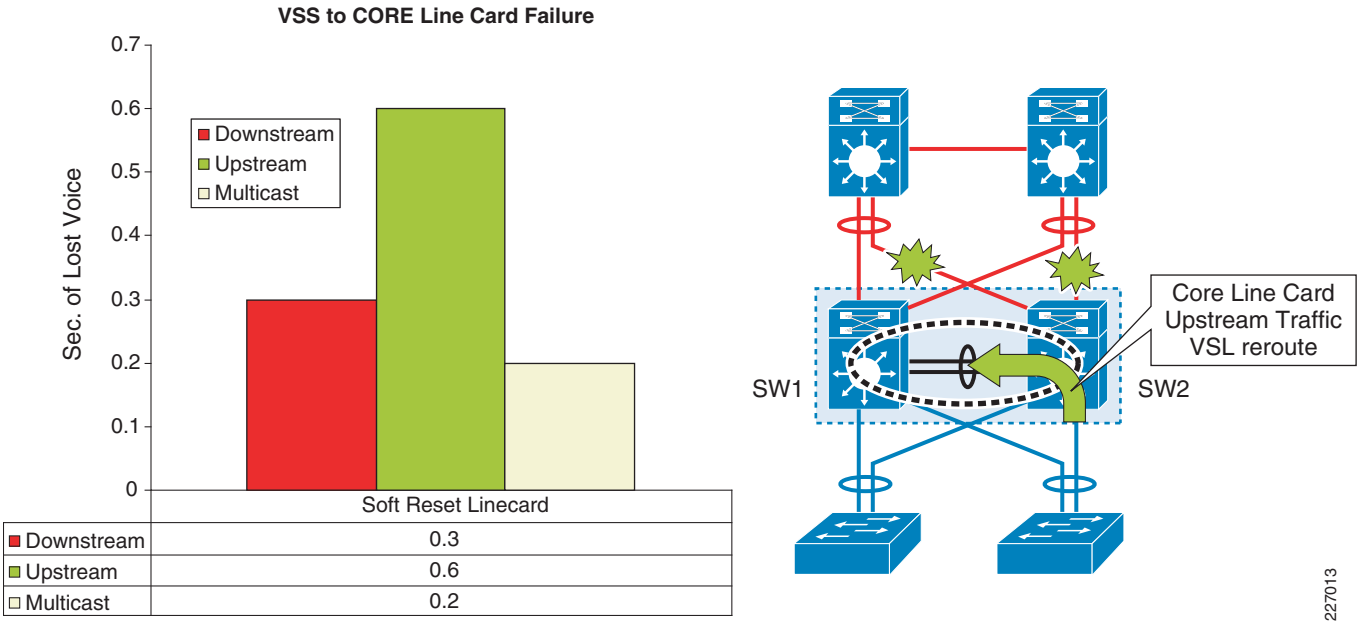
Line Card Connected to the Core-Layer

In order to avoid a single point of failure, the connectivity to the core should employ multiple line cards. The validation applies to a single line card connecting the VSS to the core to illustrate the worst-case loss scenario. See Figure 4-5 for an the convergence and traffic flow when entire connectivity from one of the VSS member switches are down. Table 4-8 lists the core-layer connectivity failure and recovery.

Table 4-8 Core Connectivity (Line Card Failure) Recovery

Topology	MEC	Additional Recovery
Unicast upstream	VSL reroute	
Unicast downstream	EtherChannel failover at the core	
Multicast hashing on failed line card	EtherChannel failover	MMLS

Figure 4-5 VSS-to-Core Single Line-Card Failure and Recovery Convergence



For the multicast traffic flowing downstream, the core device's hashing process result in the selection of the VSS member that will become the forwarder for that flow. In Figure 4-5, it will rehash over the remaining links connected to SW1. The traffic will be forwarded in the hardware by SW1 using MMLS technology that has synchronized the (s,g) pair to the peer switch (SW1 in Figure 4-5).

Caution The multicast data convergence heavily depends on the number of (s,g) pairs (mroute) and several other multicast control-plane functions. For high mroute counts, the convergence might require further validation.

Line Card Connected to an Access Layer

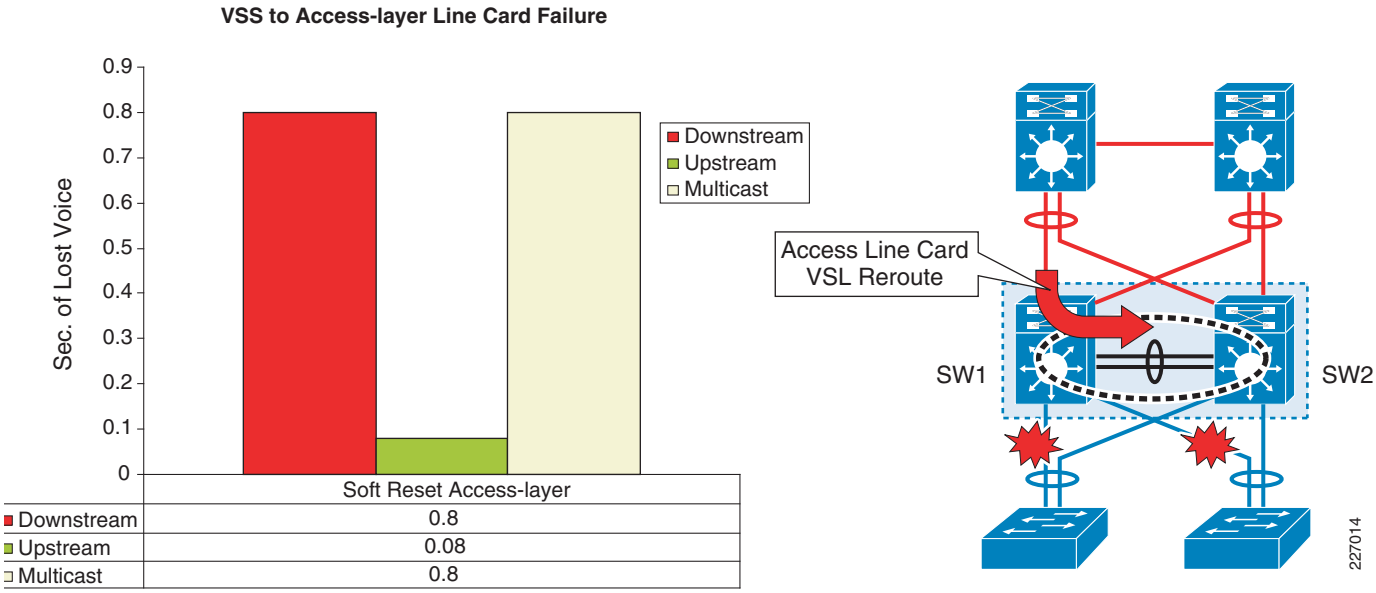
For the access-layer, line-card failures, the traffic flow recovery is the opposite of the core line-card failures. See Table 4-9.

Table 4-9 Failure Recovery Summary for Line Card Connected to the Access Layer

Topology	MEC	Additional Recovery
Unicast upstream	EtherChannel failover at access	
Unicast downstream	VSL reroute	
Multicast hashing on failed line card	VSL reroute	MMLS

Figure 4-6 shows an illustration summarizing failure and recovery for a line-card connected to the access layer.

Figure 4-6 Line Card Connected to Access Layer Failure and Recovery Summary



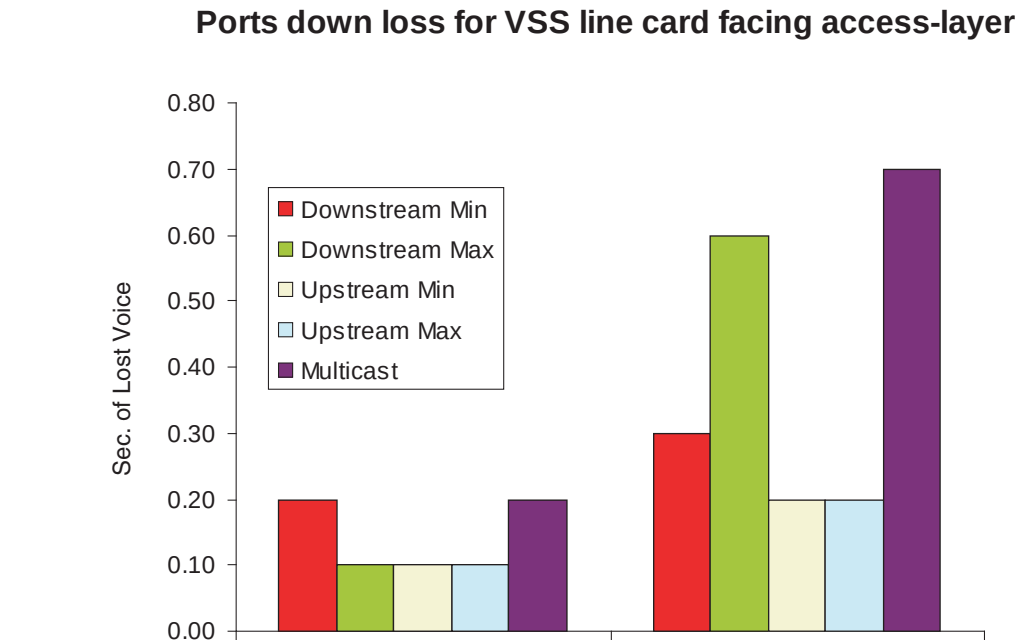
For the multicast traffic flowing downstream, the core device's hashing process results in the selection of the VSS member that will become the forwarder for that flow. For an access-layer line-card failure, the multicast traffic must be rerouted over the VSL link. The peer switch will then forward the traffic via an existing Layer-2 MEC connection to the access-layer switch. For this forwarding, the peer switch uses MMLS that has synchronized the (s,g) pair to the peer switch (SW2 in Figure 4-6).

Port Failures

A failure of a port connected to the access layer is similar to access-layer line card failure in terms of traffic flow recovery. The method and place of introducing a port status change affects convergence.

The convergence illustrated in Figure 4-7 shows that the CLI-induced shutdown of the port has better convergence than physically removing the fiber connection. Carrier delay and software polling of the port to detect the port status add additional convergence delays for physically removing the fiber link.

Figure 4-7 Recovery Comparison for Port Failures



	Soft shut	Pulling fiber
Downstream Min	0.20	0.3
Downstream Max	0.1	0.6
Upstream Min	0.1	0.2
Upstream Max	0.1	0.2
Multicast	0.2	0.7

A port **shutdown/no shutdown** sequence introduced at the access-layer uplink port can cause packet losses in the order of several seconds. The port status detection method in a Cisco Catalyst 6500 system attributes to the root cause of such delay. This behavior is common for a standalone scenario, as well as a VSS-based system. In future Cisco IOS Releases, port status detection optimization might reduce associated packet loss and convergence delays. Operationally, it is better to introduce a change of port status at the VSS and not at the access-layer.

Routing (VSS to Core) Convergence

The design choices with VSS in the Layer-3 domain are described in the "Routing with VSS" section on page 3-44. In that section, a Layer-3 MEC topology is shown to be the most effective way to build a VSS interconnection with routing entities. This section further substantiates this design choice. As a result, ECMP-based convergence is not discussed in this document. In addition, this section details the effects on VSS traffic flow and convergence when core devices fail.

Core Router Failure with Enhanced IGRP and OSPF with MEC

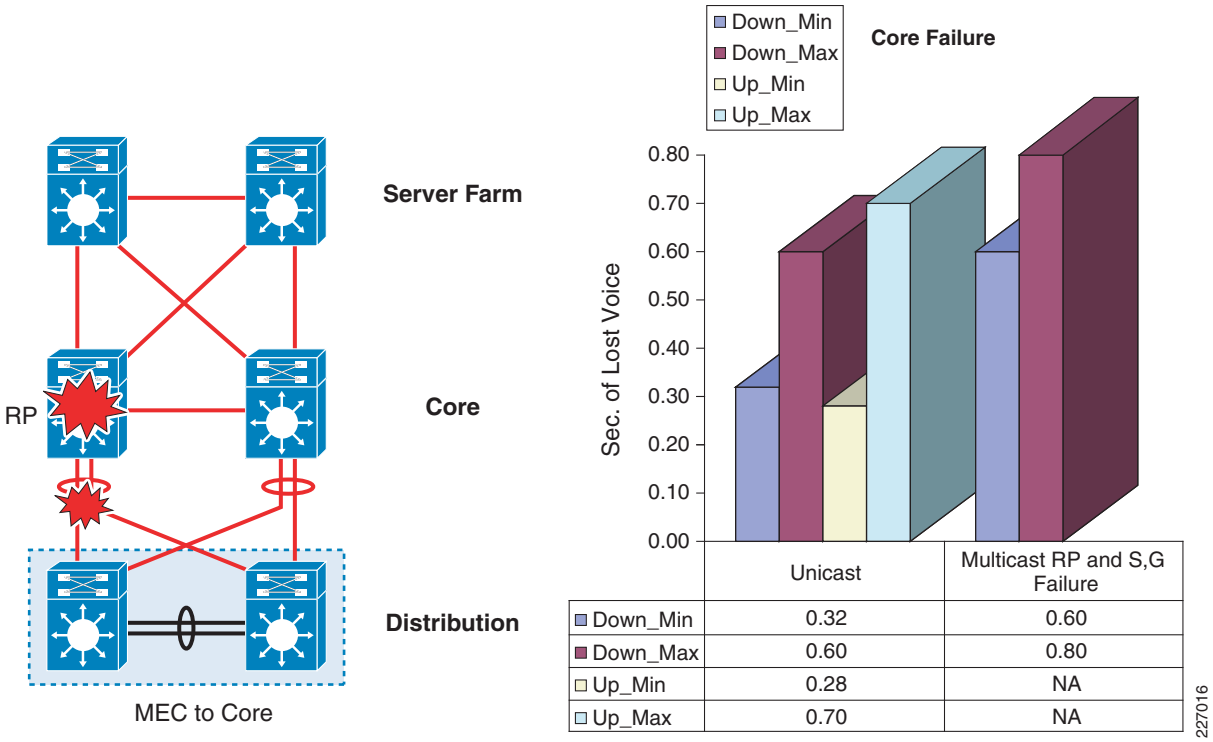
This design guide is validated with standalone core devices. In this design guide, the core is enabled to provide a rendezvous point (RP) for multicast traffic. The design choices surrounding the placement of RP depend on the multicast application and are not covered here. However, the failure of a RP is addressed as part of the description of failed components in core failures. See Table 4-10.

Table 4-10 Core Router Failure Recovery Summary

Topology	MEC	Comments
Unicast upstream	ECMP at VSS	Cisco Catalyst 6500 standalone in the core
Unicast downstream	ECMP at the core server	
IIL on active or OIL on failed core	Multicast control plane	
IIL on hot-standby	N/A	

For both OSPF and Enhanced IGRP, the core failure convergence for unicast traffic remains the same. The core and the VSS are configured with default hello and hold timers. OSPF timers for LSA and SPF are set to the default values. Both upstream and downstream convergence are based on ECMP. Both VSS members have local ECMP paths available for the forwarding and traffic does not traverse the VSL link. For the multicast traffic, a core failure introduces an incoming interface change. The multicast topology must undergo convergence and find a new incoming interface via an alternate port so that the incoming interface for the VSS is also changed. The multicast convergence can vary significantly depending on the topology beyond the core and the location of sources. This convergence is shown in Figure 4-8 for 80 multicast flows. For higher mroutes (higher flows), convergence might vary.

Figure 4-8 Recovery Comparison for Core Router Failure



Link Failure Convergence

The link failure behavior is discussed in the "Design Considerations with ECMP and MEC Topologies" section on page 3-46. The best practice-based configuration is derived from that description. If emphasizes using the MEC topology. This section covers only MEC topology option.

MEC Link Member Failure with OSPF

The dependency of routing protocol and metric change is described in the Forwarding Capacity (Path Availability) During Link Failure, page 3-47. In MEC-based topologies, a link failure might reduce the available forwarding capacity, depending on the routing protocol and associated configuration. The traffic flow recovery and associated attributes are summarized in Table 4-11.

Table 4-11 MEC Link Member Failure OSPF Recovery Summary

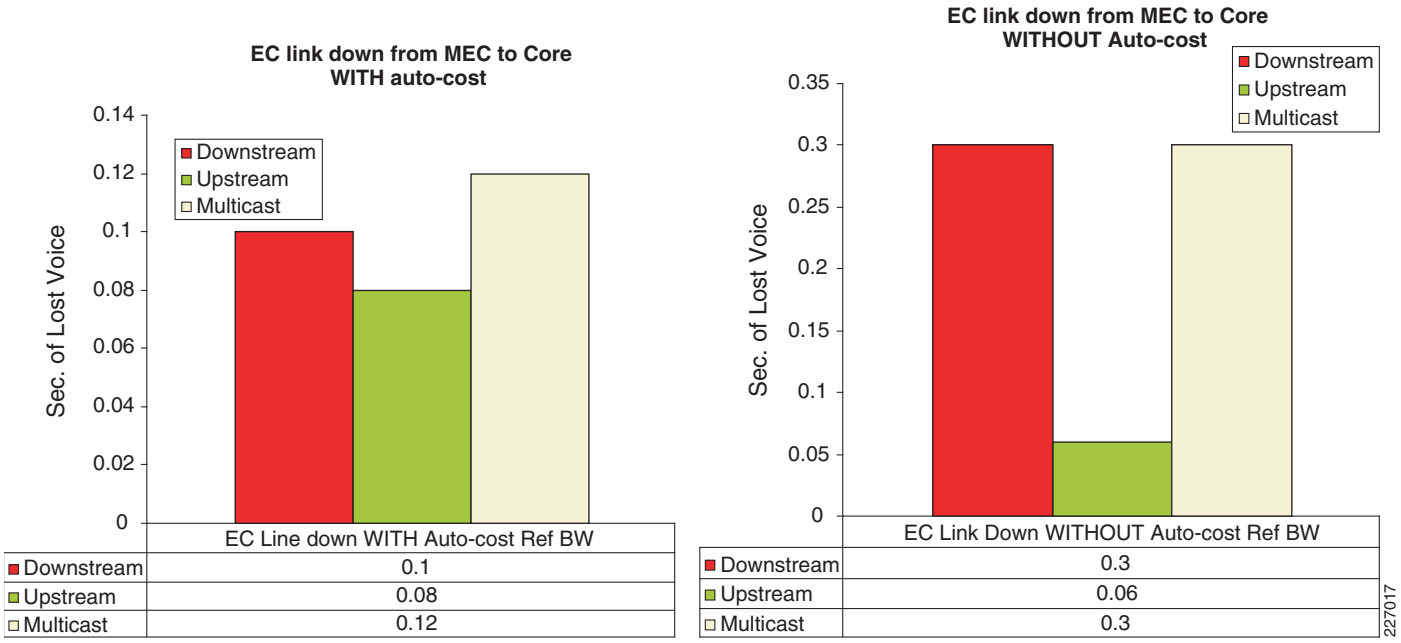
Topology	OSPF with Auto-Cost Ref BW 20G	OSPF without Auto-Cost Reference Bandwidth
Metric change	Yes	No
Resulting bandwidth	Two paths	Three paths
Unicast upstream	Graceful route withdrawal	Local hardware CEF path
Unicast downstream	Graceful route withdrawal	EC recovery at the core
Non-summarized vs. summarized nets	None	None
Multicast	Multicast control plane—Route withdrawal changes outgoing interfaces list at the core	Rehashing of multicast flow if any or no impact

The validation topology includes two Layer-3 MECs at VSS. Each core router is configured with single port-channel with member link connecting to two members of the VSS. The resulting routing topology consists of two ECMP paths (one from each core routers). For OSPF with the auto-cost reference set, a link failure triggers the metric changes on one of the routed port-channel interface. The impact of metric change is the withdrawal of routes learned from the one of the two equal cost paths. This leads to only one routed link from each VSS member being available in the routing table. The recovery for both upstream and downstream depends on graceful route withdrawal. The impact to the user data traffic is extremely low, because the traffic keeps forwarding to the link that is still available on the VSS until the route is withdrawn and because the WS-X6708 line card supports FLN (notification of link status change is hardware-based). See relevant CEF output in Forwarding Capacity (Path Availability) During Link Failure, page 3-47.

For OSPF without the auto-cost reference bandwidth, a link failure does not change the routing information because link failure does not introduce metrics change for the 20 Gigabits aggregate EtherChannel bandwidth. When the port-channel bandwidth changes to 10-Gigabit, the cost remains one because the default auto-cost is 100 MB. The recovery for upstream traffic is based the simple adjacency update in CEF at VSS and not based on ECMP recovery that is triggered when the entire routed interface is disabled (CEF next-hop update). The downstream effect in this topology will depend on the EtherChannel recovery at the core. See the relevant CEF output in Forwarding Capacity (Path Availability) During Link Failure, page 3-47.

See the comparison of recovery performance with and without auto-cost provided in Figure 4-9.

Figure 4-9 Comparison of Auto-Cost and Non-Auto-Cost Recovery



The route withdrawal does not trigger a topology change even though the metric has changed since the the route is learned from the same single logical node.

The only design choice with the OSPF and Layer-3 MEC topology is that of total bandwidth availability during the fault, and not the impact on user data convergence since packet loss is at minimal.

MEC Link Member Failure with Enhanced IGRP

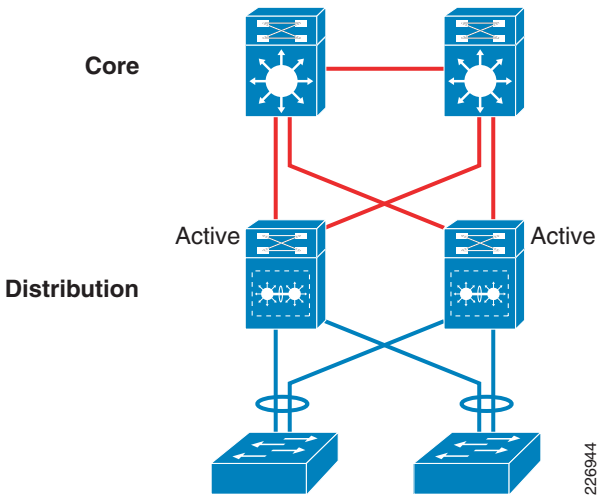
The Enhanced IGRP metric calculation is a composite of the total delay and the minimum bandwidth. When a member link is failed, EIGRP will recognize and use changed bandwidth value but delay will not change. This may or may not influence the composite metric since minimum bandwidth in the path is used for the metric calculation, so a local bandwidth change will only affect the metric if it is the minimum bandwidth in the path (the total delay has not changed). In a campus network, the bandwidth changed offered between the core and VSS is in the order of Gigabits, which typically is not a minimum bandwidth for the most of the routes. Thus, for all practical purposes, Enhanced IGRP is immuned to bandwidth changes and follows the same behavior as OSPF with the default auto-cost reference bandwidth. If there are conditions in which the composite metric is impacted, then EIGRP will follow the same behavior as OSPF with auto-cost reference bandwidth set.

Campus Recovery with VSS Dual-Active Supervisors

Dual-Active Condition

The preceding section described how VSL bundle functions as a system link capable of carrying control plane and user data traffic. The control plane traffic maintains the state machine synchronization between the two VSS chassis. Any disruption or failure to communicate on the VSL link leads to a catastrophic instability in VSS. As described in the "SSO Operation in VSS" section on page 2-24, the switch member that assumes the role of hot-standby keeps the constant communication with the active switch. The role of the hot-standby switch is to assume the active role as soon as it detects a loss of communication with its peer via the VSL link. This transition of roles is normal when either triggered via switchover (user initiated) or the active switch has some trouble. However, during a fault condition, there is no way to differentiate that either remote switch has rebooted or whether the links between the active and hot-standby switches have become inoperative. In both cases, the hot-standby switch immediately assumes the role of an active switch. This can lead to what is known as the *dual-active* condition in which both switch supervisors assume that they are the in-charge of control plane and start interacting with network as active supervisors. Figure 4-10 depicts the state of campus topology in dual-active state.

Figure 4-10 Dual Active Topology



The best way to avoid exposing your network to a dual-active condition is to apply the following best practices:

- Diversify VSL connectivity with redundant ports, line cards, and internal system resources. The recommended configuration options are illustrated under the "Resilient VSL Design Consideration" section on page 2-18.
- Use diverse fiber-optic paths for each VSL link. In the case of a single conduit failure, a dual-active condition would not be triggered.
- Manage traffic forwarded over the VSL link using capacity planning for normal and abnormal conditions. The design guidance for managing traffic over VSL is discussed in Chapter 3, "VSS-Enabled Campus Design."

The best practice-based design implementation significantly reduces the exposure to the dual-active condition, but cannot eliminate the problem. Some of the common causes that can trigger a dual-active problem are as follows:

- Configuration of short LMP timers and improper VSL port-channel configuration.
- User invoked accidental shutdown of VSL port-channel.
- High CPU utilization can trigger a VSLP hello hold-timer timeout, which results in the removal of all the VSL links from the VSL EtherChannel.
- The effects of system watchdog timer failures are similar to high CPU utilization. These might render the VSL EtherChannel non-operational.
- Software anomalies causing the port-channel interface to become disabled.
- Having the same switch ID accidentally configured on both switches during initial setup process or during a change.

Not only is it critical to avoid the dual-active condition, it is also important to detect such a condition and take steps to recover quickly. The rest of the section covers the following:

- Effects of a dual-active condition on the network in the absence of any detection techniques
- Detection options available, their behavior, and recovery
- Convergence expected for specific designs and applicable best practiced options

Impact of Dual-Active on a Network without Detection Techniques

Dual-active condition causes each member chassis to assume the active role, which means that each member acts as a standalone device claiming the same IP and MAC addresses. Network control plane duplication also results, affecting the operation of router IDs, STP root bridge, routing protocol neighbor adjacency, and so on. The impact of a dual-active condition in a production network is two-fold:

- Control plane disruption
- User data traffic disruption

The exact behavior observed for a given network depends on the topology (MEC or ECMP), routing protocol deployed (OSPF or Enhanced IGRP), and type of interconnection (Layer-2 or Layer-3) used. This section addresses the importance of deploying detection techniques. Only critical components and topology considerations are covered.

Impact on Layer-2 MEC

Dual-active triggers two active roots for the same STP domain. Both active chassis generate separate STP BPDUs with different source MAC addresses from each respective line card. The access-layer switch configured with Layer-2 MEC detects multiple MAC addresses claiming to be the source of STP tree. This is detected by PAgP as an EtherChannel inconsistency, which eventually causes the access-layer port-channel interface to enter into an error-disable state. This triggers the generation of syslogs messages; messages seen at the access-layer switches are dependent on the following software versions:

Cisco Catalyst 65xx, Cisco Catalyst 45xx, and Cisco Catalyst 35xx with Cisco IOS:

```
%PM-SPSTBY-4-ERR_DISABLE: channel-misconfig error detected on Gi5/1, putting Gi5/1 in
err-disable state
%PM-SPSTBY-4-ERR_DISABLE: channel-misconfig error detected on Gi5/2, putting Gi5/2 in
err-disable stat
```

Cisco Catalyst 65xx with CATOS:

```
%SPANTREE-2-CHNMISCFG: STP loop - channel 5/1-2 is disabled in vlan/instance 7
%SPANTREE-2-CHNMISCFG2: BPDU source mac addresses: 00-14-a9-22-59-9c, 00-14-a9-2f-14-e4
ETHC-5-
PORTFROMSTP: Port 5/1 left bridge port 5/1-2
```

Refer to following URL for details about EtherChannel inconsistencies:

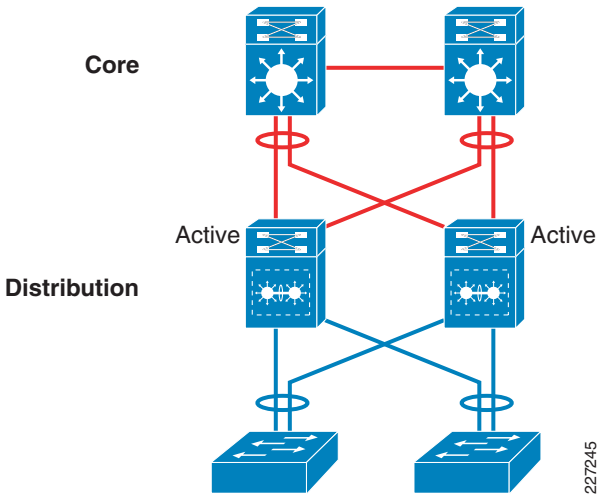
http://www.cisco.com/en/US/tech/tk389/tk213/technologies_tech_note09186a008009448d.shtml

Note The PAgP or LACP protocol itself does not trigger EtherChannel inconsistency in the core or to the access layer. This is because both active routers announce the common PAgP/LACP control plane device-ID information. In dual-active condition the PAgP detects the BPDU sourced with different MAC address, leading to error-disabling of the port-channel.

Layer-3 MEC with Enhanced IGRP and OSPF

During the dual-active condition both active VSS routers keep their respective Layer-3 MEC interfaces in the operational state. However, each active router removes the link member associated with opposite chassis; this is to reflect a condition where each router believes that remote peer has gone down and thus has to remove all the interfaces associated with that peer. The removal of interfaces may trigger a topology update to the core. However, each chassis still physically has all the previous interfaces. Each active router will continue to send neighbor and routing protocol update using those interfaces. See Figure 4-11.

Figure 4-11 Dual-Active State for Layer-3 MEC with Enhanced IGRP and OSPF Topology



For Layer-3 MEC-based topologies, there are only two neighbors in the topology shown in Figure 4-11. For a topology that is operating normally, the core sees one logical router and one path for a given destination; however, the VSS sees two paths for a given destination (one from each core router). With a dual-active condition, the core router might see more than one router, depending on the routing protocol. EtherChannel hashing enables asymmetrical selection of a link for transmitting hello (multicast) and update (unicast) messages. From the core to the VSS flow, the hashing calculation could result in those messages types being transmitted on different EtherChannel link members, while VSS to core connectivity for the control plane remains on local interfaces. During a dual-active event, this could result in adjacency resets or destabilization.

Enhanced IGRP

Depending on how hello and update messages get hashed on one of the member links from the core to one of the VSS active chassis, the adjacency might remain intact in some routers, while others might experience adjacency instability. If instability occurs, the adjacency might reset due to the expiration of either the neighbor hold-timer or NSF signal timer, as well as stuck-in-INIT error.

OSPF

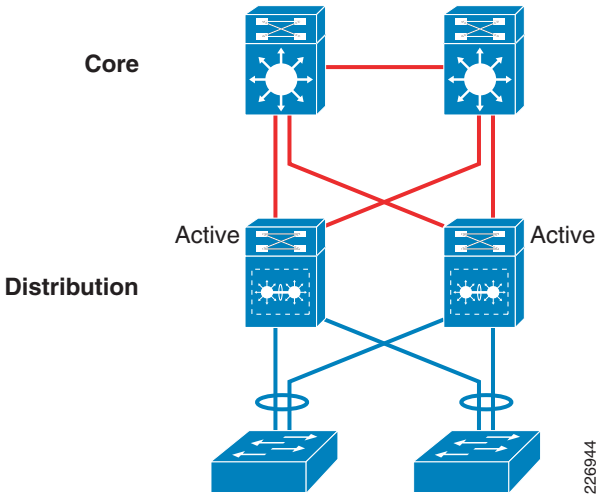
When a dual-active event occurs in an OSPF-based network, the adjacency never stabilizes with either of the active routers. For OSPF to form an adjacency, the protocol requires the bidirectional verification of neighbor availability. With dual-active state, the OSPF neighbor sees multiple messages hashed by the two active routers. In addition, the core routers see two routers advertising with the same router IDs. The combination of duplicate router ID and adjacency resets remove the subnets for the access-layer from the core router's OSPF database.

For either routing protocol, adjacency resets or instability leads to the withdrawal of routes and disruption of user traffic. In addition, Layer-2 at the access-layer will be error-disabled as discussed in the "Impact on Layer-2 MEC" section on page 4-19.

Layer-3 ECMP with Enhanced IGRP and OSPF

As illustrated in Figure 4-12, for a normally operational topology, the core router only sees one logical router and two paths for given destination; however, the VSS views four paths for a given destination (two from each core router). With a dual-active condition, the core might see more than one router depending on routing protocol. There is no hashing-related impact on neighbor adjacency and routing update with ECMP (being an independent path) such as the case with the Layer-3 MEC topology.

Figure 4-12 Layer-3 ECMP with Enhanced IGRP and OSPF Topology



Enhanced IGRP

During a dual-active condition, routers do not lose adjacency. Enhanced IGRP does not have any conflicting router IDs (unless Enhanced IGRP is used as redistribution point for some subnets) and each link is routed so that no adjacency change occurs at core routers or from active VSS members. User traffic continues forwarding with virtually no effect on user data traffic. Thus, dual-active condition may not impact Layer-3 connectivity in this topology; however, Layer-2 MEC may get error-disabled causing the disruption of user data traffic.

OSPF

During a dual-active event, two routers with the same IP loopback address announce the duplicate router IDs. Both active routers will announce the same LSA, which results in a LSA-flooding war for the access-layer subnets at the core routers.

Detection Methods

This section goes over various detection techniques and their operation and recovery steps. Following methods are available to detect the dual-active condition:

- Enhanced PAgP
- Fast-Hello—VSLP framework-based hello
- Bidirectional Forwarding Detection (BFD)

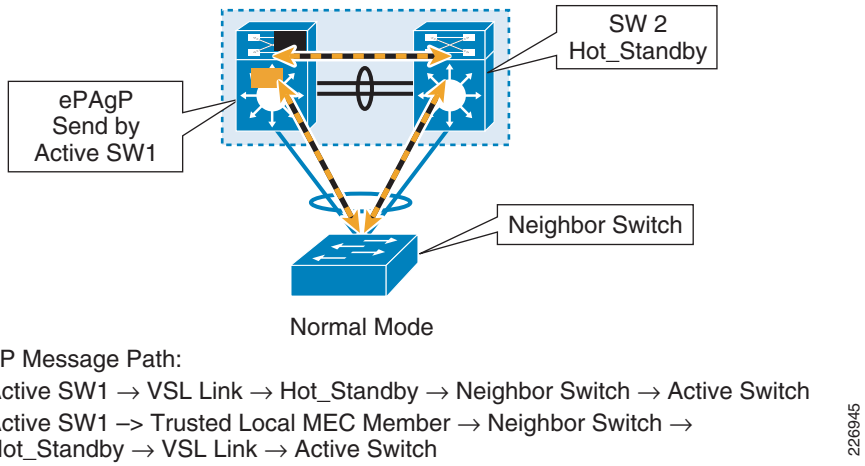
Note The enhanced PAgP and BFD is supported from Cisco IOS Release 12.2(33)SXH, while fast-hello requires Cisco IOS Release 12.2(33)SXI.

Enhanced PAgP

Normal Operation

Enhanced PAgP is an extension of the PAgP protocol. Enhanced PAgP introduces a new Type Length Value (TLV). The TLV of ePAgP message contains the MAC address (derived from the back-plane of an active switch) of an active switch as an ID for dual-active detection. Only the active switch originates enhanced PAgP messages in a normal operational mode. The active switch sends enhanced PAgP messages once every 30 seconds on both MEC link members. The ePAgP detection uses neighbor switches as tertiary connection to detect the dual-active condition. (All detection techniques require a tertiary connection from which the switches can derive the state of a VSL link because neither side can assume that the other is down by simply detecting that the VSL link is non-operational.) The ePAgP messages containing the active switch ID are sent by an active switch on the locally attached MEC-link member as well as over the VSL link. This operation is depicted in Figure 4-13 via black and yellow squares describing ePAgP messages and paths it traverses via neighbor switch. The neighbor switch simply reflects both of these messages via each uplink. This ensures that active the switch independently verifies the bidirectional integrity of the VSL links. For the neighbor switch to assist in this operation, it requires a Cisco IOS software version supporting enhanced PAgP. See Figure 4-13.

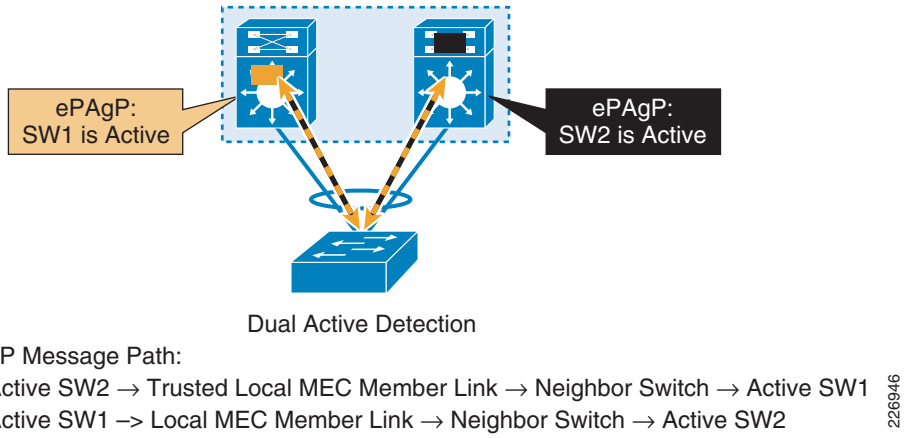
Figure 4-13 Enhanced PAgP Normal Operation



Dual-Active Detection with Enhanced PAgP

With dual-active all the links in a VSL bundle become non-operational, the hot-standby switch (SW2 in Figure 4-14) transitions to active (not knowing the status of remote switch). As SW2 becomes active, it generate its own enhanced PAgP message with its own active switch ID, sending it via the locally-attached MEC link member to SW1 via the neighbor switch. With the VSL link down, the old-active switch (SW1) stops receiving its own enhanced PAgP message and also receives an enhanced PAgP message generated via the remote switch (old hot-standby). These two messages and their paths is shown via yellow and black squares in Figure 4-14. SW1 remembers that it was an active switch and only the previously active SW1 undergoes the detection and recovery from the dual-active condition.

Figure 4-14 Dual-Active Detection with Enhanced PAgP Operation



Once an ePAgP message from SW2 is received by the old-active switch (SW1), SW1 compares its own active switch ID (MAC address derived from local backplane) with new active switch ID. If the received and expected IDs are different, the old-active chassis determines that a dual-active condition is triggered in the VS domain and starts the recovery process. The dual-active condition is displayed by following the CLI that is executed only on the old-active switch because that is where detection is activated.

```
6500-VSS# sh switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: Yes
```

```
No interfaces excluded from shutdown in
 recovery mode

In dual-active recovery mode: Yes
  Triggered by: PAgP detection
  Triggered on interface: Gi2/8/19
  Received id: 0019.a927.3000
  Expected id: 0019.a924.e800
```

Note In Cisco IOS Releases (12.2(33) SXH and 12.2(33) SXI), there is no way to differentiate between old-active versus newly-active switches. Both switches are active and both display the same command prompt. This can pose an operational dilemma when issuing the preceding command. In future releases, the old-active switch prompt may change to something meaningful so that the operator can distinguish between two active switches.

The dual-active condition generates different type of syslogs messages in different switches. The old-active switch (SW1) displays the following syslogs messages:

```
%PAGP_DUAL_ACTIVE-SW2_SP-1-RECOVERY: PAgP running on Gi2/8/19 triggered dual-active
recovery: active id 0019.a927.3000 received, expected 0019.a924.e800
%DUAL_ACTIVE-SW2_SP-1-DETECTION: Dual-active condition detected: all non-VSL and
non-excluded interfaces have been shut down
```

The newly-active switch (SW2) displays the following syslog messages:

```
%VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
%VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/5: Link down
%VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in ACTIVE role
```

The neighbor switch supporting enhanced PAgP protocol also displays the dual-active triggered message:

```
%PAGP_DUAL_ACTIVE-SP-3-RECOVERY_TRIGGER: PAgP running on Gi6/1 informing virtual switches
of dual-active: new active id 0019.a927.3000, old id 0019.a924.e800
```

Caution The neighbor switch also displays this syslog message during normal switchover because it does not know what really happened—it merely detects different switches claiming to be active.

Enhanced PAgP Support

As described in the preceding section, the neighbor switch must understand enhanced PAgP protocol in order to support dual-active detection. That also means enhanced PAgP requires the PAgP protocol to be operational on in the MEC configuration. One cannot disable PAgP and have enhanced PAgP running. Enhanced PAgP can be enabled either on Layer-2 or Layer-3 PAgP MEC members. This means you can run enhanced PAgP between the VSS and the core routers. See Table 4-12.

Table 4-12 Cisco IOS Version Support for Enhanced PAgP

Platform	Software	Comments
Cisco Catalyst 6500	12.2(33)SXH	Sup720 and Sup32
Cisco Catalyst 45xx and Cisco Catalyst 49xx	12.2(44)SG	

Table 4-12 Cisco IOS Version Support for Enhanced PAgP (continued)

Platform	Software	Comments
Cisco Catalyst 29xx, Cisco Catalyst 35xx and Cisco Catalyst 37xx	12.2(46)SE	Cisco Catalyst 37xx stack no support, see the text that follows.
Cisco Catalyst 37xx Stack	Not Supported	Cross-stack EtherChannel only supports LACP

PAgP is supported in all platforms except the Cisco Catalyst 37xx stack configuration in which cross-stack EtherChannel (LACP) is required to have MEC-connectivity with the VSS. Because cross-stack EtherChannel does not support PAgP, it cannot use enhanced PAgP for dual-active detection. A common approach to resolve this gap in support is to use two EtherChannel links from the same stack member. This solution is a non-optimal design choice because it creates a single point-of-failure. If a stack member containing that EtherChannel fails, the whole connectivity from the stack fails. To resolve this single point-of-failure problem, you can put two dual-link EtherChannel group, each on separate stack member connected to VSS; however, it will create a looped topology. The loop-free topology requires a single EtherChannel bundle to be diversified over multiple members which in turn requires LACP.

There are two solutions to the stack-only access-layer requirement:

- Use Fast Hello or BFD as the dual-active detection method (described in the "Fast-Hello (VSLP Framework-Based Detection)" section on page 4-26 or the "Bidirectional Forwarding Detection" section on page 4-30).
- Enhanced PAgP can be enabled either on Layer-2 or Layer-3 PAgP MEC members. This means you can run enhanced PAgP between the VSS and the core routers, although core routers require enhanced PAgP support and implementation of the Layer-3 MEC topology to the VSS.

Enhanced PAgP Configuration and Monitoring

Enhanced PAgP dual-active detection is enabled by default, but specific MEC groups must be specified as trustworthy. The specific CLI identifying MEC group as a trusted member is required under virtual switch configuration. The reason behind not enabling trust on all PAgP neighbors is to avoid unwanted enhanced PAgP members, such as an unprotected switch, unintended vendor connectivity, and so on.

The following conditions are required to enable the enhanced PAgP on EtherChannel:

- MEC must be placed in the administratively disabled state while adding or removing trust; otherwise, an error message will be displayed.
- PAgP protocol must be running on the MEC member. PAgP is recommended to be configured in desirable mode on both sides of the connection.

Fine tuning the PAgP hello-timer from its 30-second default value to one second using the **pagp rate fast** command does not help to improve convergence time for user traffic. This is because dual-active detection does *not* depend on how fast the PAgP packet is sent, but rather on how fast the hot-standby switch is able to generate the enhanced PAgP message with its own active ID to trigger dual-active detection. See Figure 4-15.

Figure 4-15 Enabling Trust on the MEC with PAgP Running

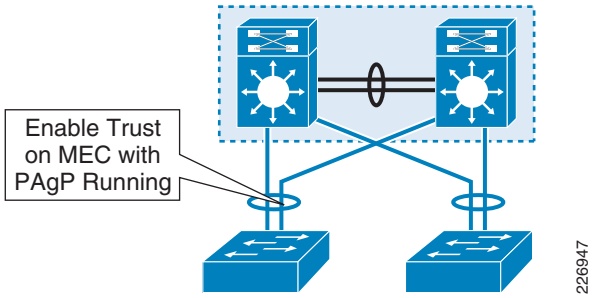


Figure 4-15 shows the trust configuration required for MEC member to be eligible for participating in enhanced PAgP-based dual-active detection. Use the following commands to enable trust on the MEC with PAgP running:

```
6500-VSS(config)# switch virtual domain 10
6500-VSS(config-vs-domain)# dual-active detection pagp trust channel-group 205
```

The enhanced PAgP support and trust configuration can be verified on the VSS switch as well as the enhanced PAgP neighbor the commands shown in the following configuration examples.

VSS switch:

```
6500-VSS# show switch virtual dual-active pagp
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

! << Snip >>

Channel group 205 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
          Dual-Active     Partner             Partner   Partner
Port      Detect Capable  Name                Port      Version
Gi1/8/19  Yes             cr7-6500-3          Gi5/1     1.1
Gi1/9/19  Yes             cr7-6500-3          Gi6/1     1.1
```

Neighbor switch that supports enhanced PAgP:

```
4507-Switch# show pagp dual-active
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

Channel group 4
          Dual-Active     Partner             Partner   Partner
Port      Detect Capable  Name                Port      Version
Te1/1     Yes             cr2-6500-VSS        Te2/2/6   1.1
Te2/1     Yes             cr2-6500-VSS        Te1/2/6   1.1
```

Fast-Hello (VSLP Framework-Based Detection)

Fast-hello is a newest dual-active detection method and is available with Cisco IOS 12.2(33) SXI and newer releases. The primary reasons to deploy fast-hello are as follows:

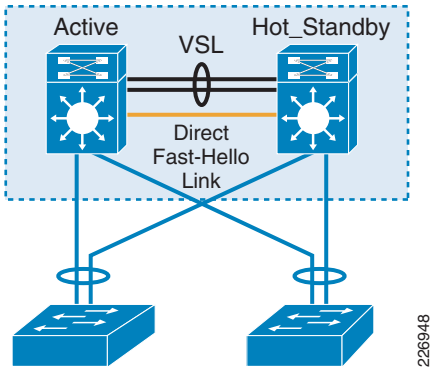
- Whenever enhanced PAgP deployment is not possible, such as in the case of server-access connectivity where servers are connected to the VSS and core connectivity is not Layer-3 MEC-based.
- If the installed-based has Cisco IOS versions that can not support enhanced PAgP.

- The EtherChannel group protocol is LACP.
- Simplicity of configuration is required and as fast-hello is being used as replacement to BFD (see the "Bidirectional Forwarding Detection" section on page 4-30).

Normal Operation

Fast-hello is a direct-connection, dual-active detection mechanism. It requires a dedicated physical port between two virtual-switch nodes in order to establish a session. Fast-hello is a connectionless protocol that does not use any type of handshaking mechanism to form a fast-hello adjacency. An incoming fast-hello message from a peer node with the appropriate TLV information establishes a fast-hello adjacency. See Figure 4-16.

Figure 4-16 Fast-hello Setup



Each dual-active fast-hello message carries the following information in TLVs:

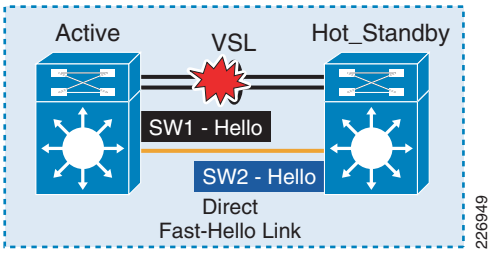
- *VSS Domain ID*—VSS virtual-switch node must carry a common domain ID in each hello message.
- *Switch ID*—Each virtual-switch node advertises the local virtual-switch ID for self-originated hello messages.
- *Switch Priority*—Each virtual-switch node advertises the local virtual-switch priority for self-originated hello messages..

By default, each virtual-switch node transmits fast-hello packets at two-second intervals. Dual-active fast-hello transmit timers are hard-coded and transparent to the end-user in the VSS system. The hard-coded fast-hello timer cannot be configured or tuned, and can only be verified using the **debug** commands. Each virtual-switch node transmits fast-hellos at the default interval to establish a session with its peer node. Any established dual-active fast-hello adjacency will be torn down if either of virtual-switch node fails to receive hellos from its peer node after transmitting five subsequent hello messages. By default, the hold-down timer is hard-coded to 10 seconds. If for any reason the adjacency establishment process fails, either due a problem or misconfiguration, the configured side continues to transmit hello messages at default interval to form a session. The dedicated link configured for fast-hello support is not capable of carrying control-plane and user-data traffic.

Dual-Active Detection with Fast-Hello

Either active or hot-standby switch cannot distinguish between the failures of remote peer or the VSS bundle. The SSO process in a active switch has to react on loss of communication to the hot-standby informing VSS control plane to remove all the interfaces and line cards associated with the hot-standby switch, including the remote port configured as for fast-hello. However, during dual-active, the link that is configured to carry fast-hello is operational (hot-standby is still operational) and exchanges hellos at the regular interval. As a result, the old-active switch notices this conflicting information about the fast-hello link and determines that this is only possible when the remote node is operational; otherwise, the old-active switch would not see the fast hello, implying dual-active has occurred. See Figure 4-17.

Figure 4-17 Dual Active Detection with Fast-hello



The previously active switch initiates the dual-active detection process when all the following conditions are met:

- Entire VSL EtherChannel is non-operational.
- Fast-hello links on each virtual-switch node are still operational.
- The previously active switch has transmitted at least one fast-hello at the regular two-second interval.

Upon losing the VSL EtherChannel, the old-active switch transmits a fast-hello at a faster rate (one per 500 msec) after transmitting at least one fast-hello at the regular interval. This design prevents taking away unnecessary CPU processing power during the active/hot-standby transitional network state. The following **show** command outputs illustrate the dual-active condition and the state of the detection on old-active.

```
6500-VSS# show switch virtual dual fast-hello
Fast-hello dual-active detection enabled: Yes

Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi1/5/1    Link up        Gi2/5/1      Link up

6500-VSS# show switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: Yes
Fast-hello dual-active detection enabled: Yes

No interfaces excluded from shutdown in recovery mode

In dual-active recovery mode: Yes
  Triggered by: Fast-hello detection
  Triggered on interface: Gi1/5/1
```

Syslog messages displayed on the on an old-active switch (SW1) when the dual-active state occurs:

```
Dec 31 22:35:58.492: %EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet1/5/4 left the
port-channel Port-channel1
Dec 31 22:35:58.516: %LINK-SW1_SP-5-CHANGED: Interface TenGigabitEthernet1/5/4, changed
state to down
Dec 31 22:35:58.520: %LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/4, changed state to down
Dec 31 22:35:58.536: %VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
Dec 31 22:35:58.540: %VSLP-SW1_SP-2-VSL_DOWN:   Last VSL interface Te1/5/4 went down
Dec 31 22:35:58.544: %LINEPROTO-SW2_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/1, changed state to down

Dec 31 22:35:58.544: %VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in
ACTIVE role

! << snip >>

Dec 31 22:35:59.652: %DUAL_ACTIVE-SW1_SP-1-DETECTION: Dual-active condition detected: all
non-VSL and non-excluded interfaces have been shut down ! <- Fast-hello triggers recovery
! process and starts recovery process the old active switch.
Dec 31 22:35:59.652: %DUAL_ACTIVE-SW1_SP-1-RECOVERY: Fast-hello running on Gi1/5/1
triggered dual-active recovery
! << snip >>
Dec 31 22:36:09.583: %VSDA-SW1_SP-3-LINK_DOWN: Interface Gi1/5/1 is no longer dual-active
detection capable
```

Syslogs messages on newly-active switch (SW2) when a dual-active state occurs:

```
Dec 31 22:35:58.521: %PFREDUN-SW2_SPSTBY-6-ACTIVE: Initializing as Virtual Switch ACTIVE
processor  Ð Starting NSF Recovery process
! << snip >>

Dec 31 22:36:09.259: %VSDA-SW2_SP-3-LINK_DOWN: Interface Gi2/5/1 is no longer dual-active
detection capable Ð Dual ACTIVE fast-hello link goes down and declares no longer
dual-active detection capable
```

Fast-Hello Configuration and Monitoring

Fast-hello configuration is simple, first enable it globally under virtual switch domain and then define it under the dedicated Ethernet port as follows:

Enable under VSS global configuration mode:

```
6500-VSS(config)# switch virtual domain 1
6500-VSS(config-vs-domain)# dual-active detection fast-hello
```

Enable fast-hello at the interface level:

```
6500-VSS(config)# int gi1/5/1
6500-VSS(config-if)# dual-active fast-hello

WARNING: Interface GigabitEthernet1/5/1 placed in restricted config mode. All extraneous
configs removed!

6500-VSS(config-if)# int gi2/5/1
6500-VSS(config-if)# dual-active fast-hello

WARNING: Interface GigabitEthernet2/5/1 placed in restricted config mode. All extraneous
configs removed!

%VSDA-SW2_SPSTBY-5-LINK_UP: Interface Gi1/5/1 is now dual-active detection capable
%VSDA-SW1_SP-5-LINK_UP: Interface Gi2/5/1 is now dual-active detection capable
```

A link-enabled for fast-hello support carries only dual-active fast-hello messages. All default network protocols, such as STP, CDP, Dynamic Trunking Protocol (DTP), IP, and so on are automatically disabled and are not processed. Only a physical Ethernet port can support fast-hello configuration; any other ports, such as SVI or port-channel, cannot be used as fast-hello links. Multiple fast-hello links can be configured to enable redundancy. The Sup720-10G 1-Gigabit uplink ports can be used if the supervisor is not configured in *10-Gigabit-only* mode. The status of the ports enabled with the fast-hello configuration (active and hot-standby switch) can be known by using the following **show** command output example:

```
6500-VSS# show switch virtual dual-active fast-hello
Fast-hello dual-active detection enabled: Yes
Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi1/5/1    Link up        Gi2/5/1      Link up

6500-VSS# remote command standby-rp show switch virtual dual-active fast-hello
Fast-hello dual-active detection enabled: Yes

Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi2/5/1    Link up        Gi1/5/1      Link up
```

Bidirectional Forwarding Detection

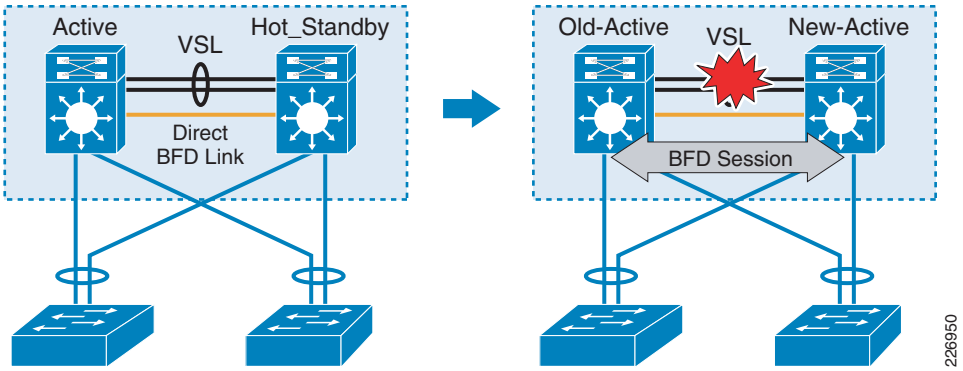
BFD is an alternative method to use when dual-active detection is not possible for the following reasons:

- Enhanced PAgP or fast-hello deployment are not possible due to Cisco IOS version limitations.
- The EtherChannel group protocol is LACP and the Cisco IOS version for fast-hello is not possible.
- Better convergence is required in a specific topology—For example, ECMP-based topologies enabled with OSPF.

Normal Operation

As with fast-hello detection, BFD detection depends on dedicated tertiary connectivity between the VSS member chassis. VSS uses BFD version 1 in echo mode for dual-active detection. Refer to cisco.com for common BFD information. BFD detection is a passive detection technique. When VSS is operating normally, the BFD configured interface remains up/up; however, no BFD session is active on that link. See Figure 4-18.

Figure 4-18 Bidirectional Forwarding Detection (BFD)



Dual-Active Detection with BFD

The BFD session establishment is the indication of dual-active condition. In a normal condition, VSS cannot establish a BFD session to itself as it is a single logical node. When a dual-active event occurs, the two chassis are physically separated, except for the dedicated BFD link that enables the BFD session between the chassis. A preconfigured connected static route establishes the BFD session. (A description of the BFD session-establishment mechanics are described in the "BFD Configuration and Monitoring" section on page 4-32.) BFD sessions last very briefly (less than one second), so you cannot directly monitor the BFD session activity. BFD session establishment or teardown logs cannot be displayed until the debug BFD command is enabled. However, the following syslogs will be displayed on the old active switch.

```
10:28:56.738: %LINK-SW1_SP-3-UPDOWN: Interface Port-channel1, changed state to down
10:28:56.742: %LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/4, changed state to down
10:28:56.742: %VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
10:28:56.742: %EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet2/5/4 left the
port-channel Port-channel2
10:28:56.750: %VSLP-SW1_SP-2-VSL_DOWN:   Last VSL interface Te1/5/4 went down
10:28:56.754: %VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in ACTIVE
role
```

The **debug ip routing** command output reveals the removal of routes for the peer switch (hot-standby switch) as the VSL link becomes non-operational and the loss of connectivity for the remote interfaces (from the old-active switch viewpoint) is detected.

```
Jul 31 10:29:21.394: RT: interface GigabitEthernet1/5/1 removed from routing table
Jul 31 10:29:21.394: RT: Pruning routes for GigabitEthernet1/5/1 (1)
```

The following syslogs output shows BFD triggering the recovery process on the old active switch:

```
10:29:21.202: %DUAL_ACTIVE-SW1_SP-1-RECOVERY: BFD running on Gi1/5/1 triggered dual-active
recovery <- 1
10:29:21.230: %DUAL_ACTIVE-SW1_SP-1-DETECTION: Dual-active condition detected: all non-VSL
and non-excluded interfaces have been shut down
```

The following syslog output depicts new active during dual active. Notice the time stamp in bold associated with marker number 2 which is the time compared to the BFD trigger time in the old active switch (see the marker number 1 in the preceding output example).

```
10:28:56.738: %VSLP-SW2_SPSTBY-3-VSLP_LMP_FAIL_REASON: Te2/5/4: Link down
10:28:56.742: %VSLP-SW2_SPSTBY-2-VSL_DOWN:   Last VSL interface Te2/5/4 went down
10:28:56.742: %VSLP-SW2_SPSTBY-2-VSL_DOWN:   All VSL links went down while switch is in
Standby role
```

The following output illustrates the BFD triggering the recovery process on the newly active switch:

```
10:28:56.742: %DUAL_ACTIVE-SW2_SPSTBY-1-VSL_DOWN: VSL is down - switchover, or possible
dual-active situation has occurred <- 2
10:28:56.742: %VSL-SW2_SPSTBY-3-VSL_SCP_FAIL: SCP operation failed
10:28:56.742: %PFREDUN-SW2_SPSTBY-6-ACTIVE: Initializing as Virtual Switch ACTIVE
processor
```

The following output on newly active switch illustrates the installation of the connected route to establish the BFD session with the old active switch:

```
10:28:58.554: RT: interface GigabitEthernet2/5/1 added to routing table
10:29:21.317: RT: interface GigabitEthernet2/5/1 removed from routing table
10:29:21.317: RT: Pruning routes for GigabitEthernet2/5/1 (1)
```

The dual-active detection using BFD takes 22-to-25 seconds (see time stamps in preceding syslogs with markers 1 and 2). BFD takes longer to shutdown the old active switch compared to the fast-hello detection scheme due to the following reasons:

- BFD session establishment is based on IP connectivity. The hot-standby switch requires control plane initialization via SSO before it can start IP connectivity.
- Time required to start IP processes and installing the connected static route;
- Time required for BFD session initialization and session establishment between two chassis.

The impact of longer detection time on user data traffic might not be significant and depends on routing protocol and topology. The BFD-based detection is required in certain topologies for a better convergence. However, the BFD-based detection technique shall be deprecated in future software releases in lieu of improved hello detection of fast-hello. See the "Effects of Dual-Active Condition on Convergence and User Data Traffic" section on page 4-38.

BFD Configuration and Monitoring

BFD configuration requires a dedicated, directly connected physical Ethernet port between the two VSS chassis. BFD pairing cannot be enabled on Layer-3 EtherChannel or on a SVI interface. Sup720-10G one Gigabit uplink ports can be used only if the supervisor is not configured in 10 Gigabit-only mode

BFD configuration for dual-active differs from normal BFD configuration on a standard interface. The BFD session connectivity between switches is needed *only* during dual-active conditions. First, BFD detection must be enabled on global virtual-switch mode. Second, the dedicated BFD interface must have a unique IP subnet on each end of the link. In a normal operational state, the two connected interfaces cannot share the same subnet, yet that sharing is required for BFD peer connectivity during a dual-active event. Once interfaces are paired, the virtual switch self-installs two static routes as a connected route with paired interfaces. It also removes the static route upon un-pairing the interface.

The following commands are required to configure a BFD-based detection scheme.

Enable under VSS global configuration mode.

```
6500-VSS(config)# switch virtual domain 10
6500-VSS(config)# dual-active pair interface gig 1/5/1 interface gig 2/5/1 bfd
```

Enable unique IP subnet and BFD interval on the specific interfaces.

```
6500-VSS# conf t
6500-VSS(config)# interface gigabitethernet 1/5/1
6500-VSS(config)# ip address 192.168.1.1 255.255.255.0
6500-VSS(config)# bfd interval 50 min_rx 50 multiplier 3

6500-VSS(config)# interface gigabitethernet 2/5/1
```

```
6500-VSS(config)# ip address 192.168.2.1 255.255.255.0
6500-VSS(config)# bfd interval 50 min_rx 50 multiplier 3
```

The preceding configuration sequence results in the automatic installation of the required static route. The following messages are displayed on the display console.

```
Console Message:

adding a static route 192.168.1.0 255.255.255.0 Gi2/5/1 for this dual-active pair
adding a static route 192.168.2.0 255.255.255.0 Gi1/5/1 for this dual-active pair
```

Notice that the static route for the subnet configured on switch 1 interface (1/5/1 in above example) is available via interface residing on switch 2 (2/5/1). This configuration is necessary because static routes help establish the BFD session connectivity when chassis are separated during a dual-active event. The BFD protocol itself has no restriction on being on a separate subnet to a establish session.

Note The recommended BFD message interval is between 50 and 100 msec with multiplier value of 3. Increasing the timer values beyond recommended values has shown higher data loss in validating the best practices. Use unique subnet for BFD configuration which does not belong to any part of the IP address range belonging to your organization. Use route-map with redistribute connected (if required for other connectivity) to exclude BFD related connected static route.

BFD detection configuration can be monitored via following CLI commands:

```
6500-VSS# sh switch virtual dual-active bfd
Bfd dual-active detection enabled: Yes

Bfd dual-active interface pairs configured:
  interface-1 Gi1/5/1 interface-2 Gi2/5/1

6500-VSS# sh switch virtual dual active summary
Pagp dual ACTIVE detection enabled: No
Bfd dual ACTIVE detection enabled: Yes
No interfaces excluded from shutdown in recovery mode
In dual ACTIVE recovery mode: No
```

Configuration Caveats

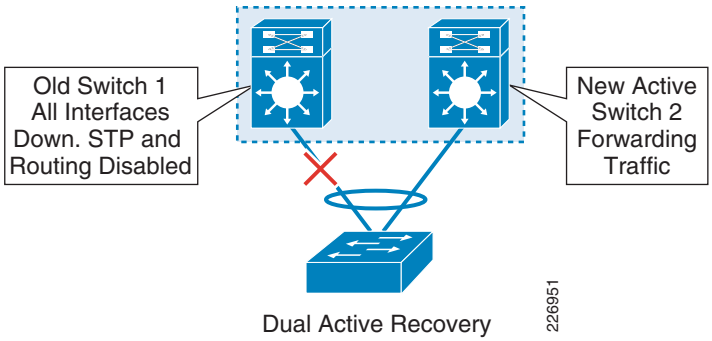
BFD hello timer configuration must be the same on both switches. In addition, any configuration changes related to IP addresses and BFD commands will result into removal of BFD detection configuration under global virtual switch mode and has to be re-added manually. This is designed to avoid inconsistency, because the validity of configuration cannot be verified unless a dual-active event is triggered. The following reminder will appear on the console:

```
6500-VSS(config)# interface gig 1/5/1
6500-VSS(config-if)# ip address 14.14.14.14 255.255.255.0
The IP config on this interface is being used to detect dual-active conditions. Deleting
or changing this config has deleted the bfd dual-active pair: interface1: Gi1/5/1
interface2: Gi2/5/1
deleting the static route 3.3.3.0 255.255.255.0 Gi1/5/1 with this dual-active pair
deleting the static route 1.1.1.0 255.255.255.0 Gi2/5/1 with this dual-active pair
```

Dual-Active Recovery

Once the detection technique identifies the dual-active condition, the recovery phase begins. The recovery process is the same for all three detection techniques. In all cases, the old-active switch triggers the recovery. In the examples presented in this guide, SW1 (original/old-active switch) detects that SW2 has now also become an active switch, which triggers detection of the dual-active condition. SW1 then disables all local interfaces (except loopback) to avoid network instability. SW1 also disables routing and STP instances. The old-active switch is completely removed from the network. See Figure 4-19.

Figure 4-19 Dual Active Recovery.



You can use the exclude interface option to keep a specified port operational during the dual-active recovery process-such as a designated management port. However, the **excluded port** command will not have routed connectivity, because the old-active switch does not have a routing instance. The following is an example of the relevant command:

```
VSS(config-vs-domain)# dual-active exclude interface port_nubmer
```

Note SVI or EtherChannel logical interfaces cannot be excluded during a dual-active event.

It is highly recommended to have console-based access to both the chassis during normal and dual-active conditions.

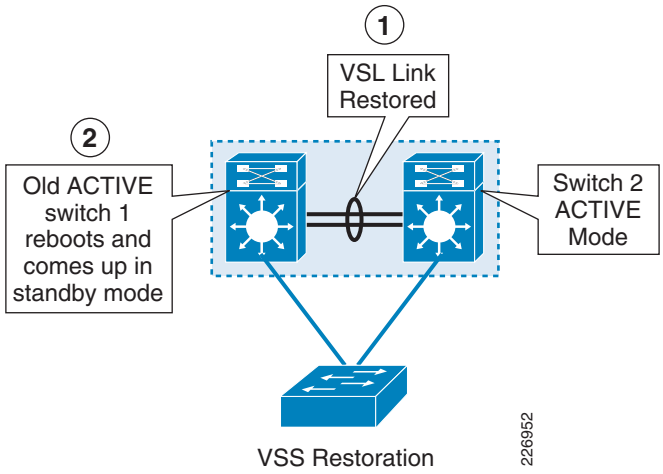
VSS Restoration

The VSS restoration process starts once the VSL connectivity is reestablished. The following events, which are causes of dual-active, restore the VSL connectivity between VSS switch members:

- Restoration of fiber connectivity; this can happen if the network suffered from a physically-severed fiber link.
- Reversal of configuration change, which could have shutdown the VSL bundle.
- Restoration of faulty hardware; this is the least probable event if a resilient design is adopted.

Figure 4-20 provides a high-level summary of the VSS restoration process.

Figure 4-20 VSS Restoration



Once the VSL connectivity is established, role negotiation (via the RRP protocol) determines that the previously active switch (SW1 in Figure 4-20 above) must become the hot-standby switch. There is no reason to change the role of the existing (new) active switch and thus triggering more data loss. This requires SW1 to be rebooted because a switch cannot go directly to the hot-standby state without a software reset. If no configuration mismatch is found, SW1 automatically reboots itself and initializes in the hot-standby mode. All interfaces on SW 1 are brought on line and SW1 starts forwarding packets-restoring the full capacity of the network. For example, the following console messages are displayed during VSL-bundle restoration on SW1 (previously active switch):

```
17:36:33.809: %VSLP-SW1_SP-5-VSL_UP:  Ready for Role Resolution with Switch=2,
MAC=001a.30e1.6800 over Te1/5/5
17:36:36.109: %dual ACTIVE-1-VSL_RECOVERED: VSL has recovered during dual ACTIVE
situation: Reloading switch 1
! << snip >>
17:36:36.145: %VSLP-SW1_SP-5-RRP_MSG: Role change from ACTIVE to HOT_STANDBY and hence
need to reload
Apr  6 17:36:36.145: %VSLP-SW1_SP-5-RRP_MSG: Reloading the system...
17:36:37.981: %SYS-SW1_SP-5-RELOAD: Reload requested Reload Reason: VSLP HA role change
from ACTIVE to HOT_STANDBY.
```

If any configuration changes occur during the dual-active recovery stage, the recovered system requires manual intervention by the use of the **reload** command and manual configuration synchronization. When network outages such as dual-active occur, many network operators may have developed a habit of entering into configuration mode in search of additional command that could be helpful in solving network outage. Even entering and exiting the configuration mode (and making no changes) will mark the configuration as dirty and will force manual intervention. Dual-active condition is also created when accidental software shut down of the VSL port-channel interface. The configuration synchronization process will reflect this change on both chassis. The only way to restore the VSL-connectivity is to enter into configuration mode, which will force the manual recovery of VSS dual-active. Once the VSL bundle is restored, the following syslogs messages will appear only in the old-active switch's console output:

```
11:02:05.814: %DUAL_ACTIVE-1-VSL_RECOVERED: VSL has recovered during dual-active
situation: Reloading switch 1
11:02:05.814: %VS_GENERIC-5-VS_CONFIG_DIRTY: Configuration has changed. Ignored reload
request until configuration is saved
11:02:06.790: %VSLP-SW1_SP-5-RRP_MSG: Role change from Active to Standby and hence need to
reload
11:02:06.790: %VSLP-SW1_SP-5-RRP_UNSAVED_CONFIG: Ignoring system reload since there are
unsaved configurations. Please save the relevant configurations
```

```
11:02:06.790: %VSLP-SW1_SP-5-RRP_MSG: Use 'reload' to bring this switch to its preferred
STANDBY role
```

For the VSS to operate in SSO mode requires that both chassis have exactly identical configurations. The configuration checks to ensure compatibility are made as soon as the VSL link is restored. Any changes (or even simply entering into configuration mode) will mark the flag used for checking the configuration status as dirty—implying possible configuration mismatch. This mismatch might requires one of the following:

- Correcting the mismatched file and reflecting the change so that the configuration on both chassis match
- Saving the configuration to NVRAM to clear the flag

Two types of configuration changes are possible during a dual-active event:

- "Non-VSL Link Configuration Changes" section on page 4-36
- "VSL-Link Related Configuration Changes" section on page 4-36

Both of these configuration change types are discussed in the next sections. Each requires a proper course of action.

Note The behavior of a system in response to configuration changes might depend on the Cisco IOS version implemented. The behavior description that follows applies only to the Cisco IOS Release 12.2(33)SXH.

Non-VSL Link Configuration Changes

For any configuration changes that do not affect the VSL bundle, you must determine to which chassis those changes apply. If changes are on the old-active switch, saving the configuration and manually rebooting the switch will restore the switch in hot-standby mode. If the changes were saved on old-active before the VSL link is being restored, manually rebooting the old-active switch might not be required because saving the configuration clears the dirty status flag. If the changes were made to the active switch, then those changes do not affect dual-active recovery activity. After the recovery (once the VSL link is restored), the new active switch configuration will be used to overwrite the configuration in the peer switch (the old-active switch) when it becomes the hot-standby switch. Changes made to the active switch need not match the old-active switch configuration because the configuration on the old-active switch (now the hot-standby switch) will be overwritten.

VSL-Link Related Configuration Changes

The dual-active condition can be triggered by various events, including the following:

- A user-initiated accidental shutdown of the VSL port-channel
- Changes to the EtherChannel causing all links or the last operational VSL link to be disconnected

When changes to the VSL port-channel are made, the changes are saved in both chassis before the dual-active event is triggered. The only way to restore the VSL-related configuration mismatch is to enter into the configuration mode and match the desired configurations. If the configuration-related to VSL links are not matched and if old-active chassis is rebooted, the chassis will come up in route processor redundancy (RPR) mode. It is only during the old-active switch recovery (in this case manual reboot) that the VSL configurations mismatch syslogs output will be displayed on the new-active switch. The following syslogs output examples illustrate this mismatch output:

```
Aug 28 11:11:06.421: %VS_PARSE-3-CONFIG_MISMATCH: RUNNING-CONFIG
Aug 28 11:11:06.421: %VS_PARSE-3-CONFIG_MISMATCH: Please use 'show switch virtual
redundancy config-mismatch' for details
Aug 28 11:11:06.421: %VS_PARSE-SW2_SP-3-CONFIG_MISMATCH: VS configuration check failed
```

```
Aug 28 11:11:06.429: %PFREDUN-SW2_SP-6-ACTIVE: Standby initializing for RPR mode
Aug 28 11:11:06.977: %PFINIT-SW2_SP-5-CONFIG_SYNC: Sync'ing the startup configuration to
the standby Router.
6500-VSS#show switch virtual redundancy | inc Opera
     Operating Redundancy Mode = RPR
```

VSL-related configuration changes are viewed via the **show switch virtual redundancy config-mismatch** command. The following is an example output:

```
6500-VSS# show switch virtual redundancy config-mismatch

Mismatch Running Config:
Mismatch in config file between local Switch 2 and peer Switch 1:
ACTIVE  : Interface TenGigabitEthernet1/5/4 shutdown
STANDBY : Interface TenGigabitEthernet1/5/4 not shut
In dual-active recovery mode: No
```

In RPR mode, all the line cards are disabled, except where the VSL is enabled. Once the configuration is corrected on the active switch, the **write memory** command will cause the startup configuration to be written to the RPR switch supervisor. The redundancy **reload peer** command will reboot the switch from RPR mode to the hot-standby mode. The following configuration sequence provides an example of this process:

```
6500-VSS# conf t
6500-VSS(config)# int te1/5/4
6500-VSS(config-if)# no shut
6500-VSS(config-if)# end
6500-VSS# wr mem

Aug 28 11:17:30.583: %PFINIT-SW2_SP-5-CONFIG_SYNC: Sync'ing the startup configuration to
the standby Router. [OK]
6500-VSS# redundancy reload peer
Reload peer [confirm] y
Preparing to reload peer
```

If the configuration correction is not synchronized before VSL-link restoration, a VSL-configuration change can cause an extended outage because the only way to determine whether a VSL-configuration mismatch has occurred is after the old-active switch boots up following VSL-link restoration. That means the switch will undergo two reboots. First, to detect the mismatch and then second boot up is required with corrected configuration to assume the role of hot-standby. To avoid multiple reboots, check for a VSL-configuration mismatch *before* the VSL link has been restored. Users are advised to be particularly cautious about modifying or changing VSL configurations.

Tip **The best practice recommendation is to *avoid* entering into configuration mode while the VSS environment is experiencing a dual-active event; however, you cannot avoid configuration changes required for accidental shutdowns of the VSL link or the required configuration changes needed to have a proper VSL restoration.**

Effects of Dual-Active Condition on Convergence and User Data Traffic

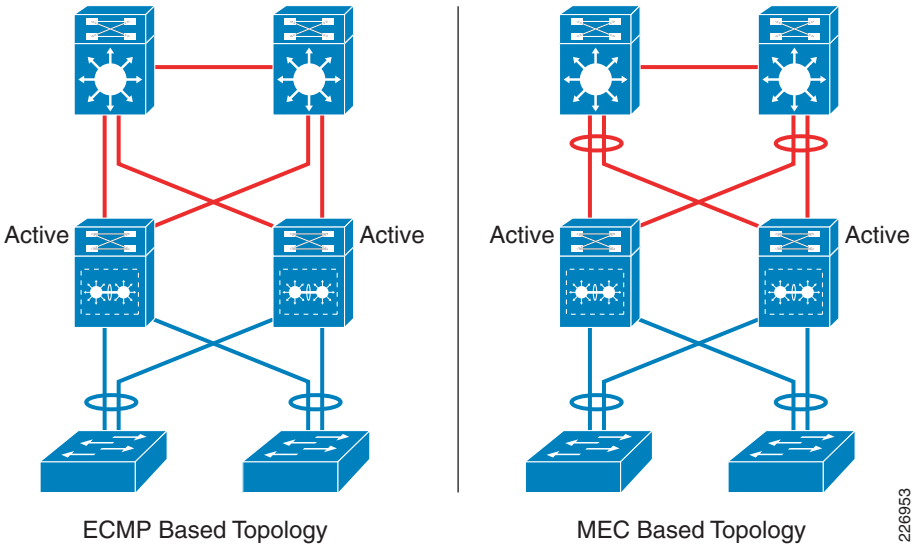
This section covers the effects of the dual-active condition on application and user data traffic. It is important to keep in mind that each detection technique might take a different amount of time to detect the dual-active condition, but this problem's influence upon the speed with which user data traffic is restored depends on many factors. These are summarized in the list of convergence factors that follows. Highly granular details of events (during dual-active) and their interactions with the following convergence factor are beyond the scope of this design guide. Overall, what matters is the selection of a detection technique for a specific environment based on observed convergence data.

User traffic convergence is dependent on the following factors:

- Dual-active detection method—Enhanced PAgP, fast-hello, or BFD
- Routing protocol configured between Layer-3 core to VSS—Enhanced IGRP or OSPF
- Topology used between VSS and Layer-3 core—ECMP or MEC
- SSO recovery
- NSF recovery

Figure 4-21 provides validation topologies with which all the observations with dual-active events and convergence associated with data traffic are measured. It is entirely possible to realize better or worst convergence in a various combination of the topologies such as all Layer-2, Layer-3, or end-to-end VSS. However, general principles affecting convergence remains the same.

Figure 4-21 Comparison of MEC and ECMP Topologies



The ECMP-based topology has four routing protocol adjacencies, where as MEC has two. For the ECMP-based topology, the VSS will send a separate hello on each link, including the hello sent via hot-standby-connected links. This means the that active switch will send the hello over the VSL link to be sent by links connected via hot-standby. The hello sent by core devices will follow the same path as VSS. For MEC-based topology, the hello originated from VSS will always be sent from local links of an active switch. However, the hello originated from core devices may select the link connected to hot-standby based on hashing result. This behavior of link selection from core devices is also repeated for routing update packets. As a result, ECMP and MEC topology exhibits different behavior in the settling the routing protocol adjacency and NSF procedure. In turn, it plays a key role in how fast the convergence of data traffic is possible.

The sequence of events that occur during a dual-active event with a detection technique deployed are generalized below for contextual reference and do not comprise a definitive process definition.

1. Last VSL link is disabled.
2. The currently active switch does not know whether the VSL has become disabled or the remote peer has rebooted. The currently active switch assumes that the peer switch and related interfaces are lost and treat this as an OIR event. As a result, the hot-standby switch interfaces are put into the down/down state, but the remote switch (current hot-standby switch) is still up and running. Meanwhile, local interfaces attached to old-active switch remain operational and continue forwarding control and data traffic. The active switch attached interface may advertise the routing update about remote switch (hot-standby) interfaces status as observed during this event.
3. As a result of all the VSL links being disabled, the hot-standby switch transitions to the active role not knowing whether the remote switch (the old-active switch) has rebooted or is still active. In this case, the situation is treated as a dual-active condition because the old-active switch has not undergone a restart.
4. The new-active switch initializes SSO-enabled control protocols and acquires interfaces associated with local chassis (the line protocol status for each of these interfaces does not become non-operational due to SSO recovery).
5. The newly-active supervisor restarts the routing protocol and undergoes the NSF recovery process, if the adjacent routers are NSF-aware; otherwise, a fresh routing-protocol adjacency restart is initiated. (See the "Routing with VSS" section on page 3-44.)
6. VSS SSO recovery follows the same process of separation of the control and the data plane as with a standalone, dual-supervisor configuration. As a result, the forwarding plane in both the switches remains operational and user traffic is switched in hardware in both switches. This forwarding continues until the control plane recovers. The control plane recovery occurs on both active switches. The old-active switch simply keeps sending routing protocol hello and some update regarding remote switch interfaces Layer-3 status. The newly-active switch restarts the routing protocols and tries to require adjacency with Layer-3 core devices. These parallel control plane activity might lead to adjacency resets, resulting in forwarding path changes (depends on usage of routing protocol and topology) and dual-active detection triggering the shutting down of the old-active switch interfaces.

If implemented, a dual-active detection method determines how fast detection occurs, which in turn triggers the shutting down of the old-active interfaces. Enhanced PAgP and fast-hello take around two-to-three seconds, while BFD takes 22-to-25 seconds. However, detection technique alone does not influence the convergence of user data traffic flow. Even though a faster detection method might be employed, the impact on user data traffic might be greater and vice versa (slower detection method having a better user data convergence).

The convergence validation and traffic-flow characterization during a dual-active event are presented in the following sections—segmented by routing-protocol deployed. Although ECMP- and BFD-based environments are described in general, only the MEC-based topology is used in the detailed descriptions that follow depicting events specific to dual-active conditions for each routing protocol.

Tip **The routing-protocols are recommended to run default hello and hold timers.**

Convergence from Dual-Active Events with Enhanced IGRP

Enhanced PAgP and Fast-Hello Detection

Both ECMP- and MEC-based connectivity to the core delivers user traffic convergence that is below one second.

BFD

The ECMP-based core BFD design provides the same convergence characteristics as enhanced PAgP. The recovery with the MEC-based core is more complex. The MEC core design experiences greater loss because of the destabilization of enhanced IGRP adjacencies leading to the removal routes affecting both upstream and downstream convergence. See Figure 4-22.

Figure 4-22 VSS Dual-Active Convergence with Enhanced IGRP

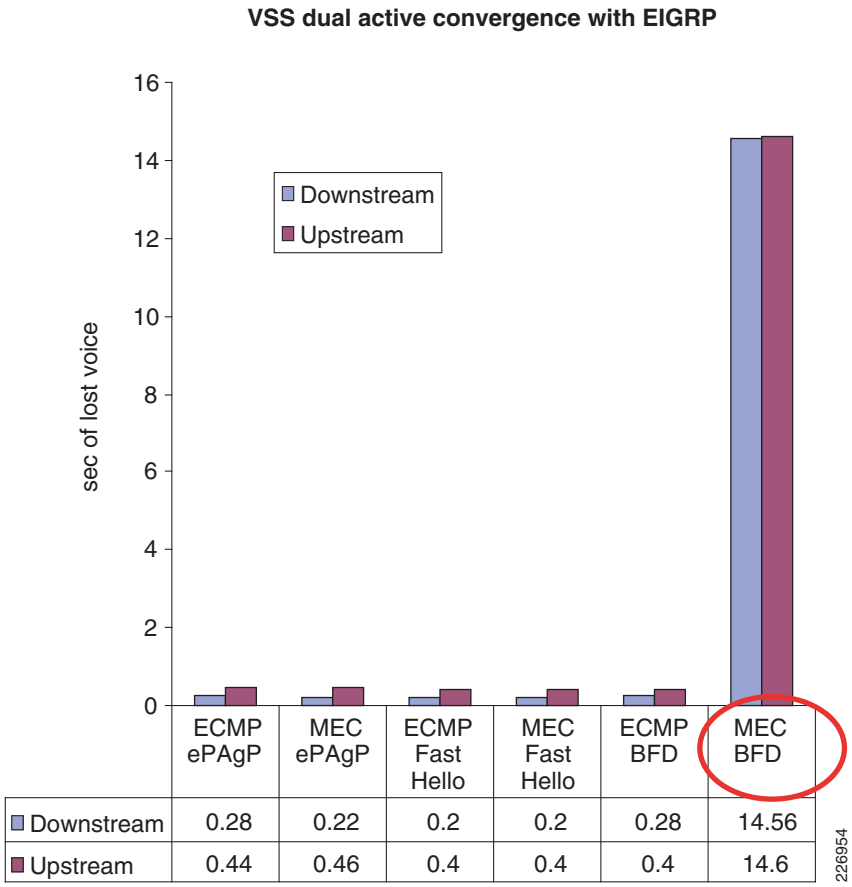


Table 4-13 describes the losses and root causes of the convergence for each combination.

Table 4-13 Convergence Recovery Losses and Causes (Enhanced IGRP Environment)

Dual-Active Detection Protocol	Core-to-VSS Connectivity	End-to-End Convergence	Summary of Recovery and Loss Process During a Dual-Active Event
Enhanced PAgP or Fast-Hello	ECMP	Upstream and downstream: 200-to-400 msec	MEC loses one link from the old active switch viewpoint, but no routes withdrawal notifications are sent or removed during EC link change. Before the new active switch announces the adjacency, the old active switch is shut down (2-to-3 seconds). A clean NSF restart occurs because no routes are withdrawn from the restarting NSF peer while the new active switch announces the NSF adjacency restart (7-to-12 seconds). The only losses related to recovery occur is the bringing down of interface by an old-active switch.
	MEC	Upstream and downstream: 200-to-400 msec	MEC loses one link from old-active viewpoint; however, no routes withdrawal are sent or removed during EC link change. Before the new active announces the adjacency, the old active is shutting down (2-1/2 sec). Clean NSF restart (as no routes being withdrawn from NSF restarting peer while the new active announces the NSF adjacency restart (7-12 sec))
BFD	ECMP	Upstream and downstream: 200-to-400 msec	Same behavior and result as in ECMP with enhanced PAgP (or fast-hello), even though BFD detection takes longer to complete the detection and shut down the old-active switch. Until the shutdown is initiated, the traffic to the old active continues to be forwarded in hardware. The old-active switch removes all the interfaces connected with the peer and sends updates with an infinite metric value to the core. However, the new-active switch interfaces are operational. No routes are withdrawn from the core router because Enhanced IGRP does not perform a local topology calculation until an explicit announcement occurs regarding the routes to be queried via the interface. Meanwhile, the new-active switch undergoes the NSF restart and refreshes adjacency and route updates. No routes or adjacency resets occur during a dual-active event.
	MEC	Upstream and downstream: 4-to-14 seconds	Higher data loss is observed because some Enhanced IGRP neighbor adjacencies might settle on one of the active routers based on the IP-to-Layer-3 MEC member link hashing toward the VSS. The Enhanced IGRP update and hello messages are hashed to different links from the core routers, leading to a loss of adjacency and routes. Either Enhanced IGRP adjacency settles on the old-active or new-active switches. If it settles on the old-active switch, traffic loss is more pronounced because Enhanced IGRP adjacencies must be reestablished with the new-active switch after old-active switch shuts down. As a result, the time required to complete convergence will vary.

Details of BFD Detection with an MEC-Based Topology

As shown in Table 4-13, this combination causes more instability because the detection technique takes longer to complete and adjacency destabilization leads to more traffic disruption. This severity of the destabilization depends on the hash result of source and destination IP addresses and the Enhanced IGRP hello (multicast) and update (unicast) transmissions—which are sent out on a MEC link member that is connected to either the old-active or new-active switch. The Enhanced IGRP packet forwarding path in a normal topology can adopt one of the following combinations in the VSS topology:

- Multicast on the old-active switch and unicast on the new-active switch
- Multicast on the new-active switch and unicast on the old-active switch
- Multicast and unicast on the old-active switch
- Multicast and unicast on the new-active switch

With the preceding combinations in mind, any of the following events can cause an Enhanced IGRP adjacency to reset during dual-active condition:

- Enhanced IGRP adjacency settling on one of the active switches such that the hellos from the core are not received and the hold-timer expires. This occurs because during normal operating conditions, the hash calculation resulted in the sending of hellos to the hot-standby switch that was forwarding packets over the VSL link and now is no longer doing so (VSL link is down). As a result, the old-active switch never sees the hello packet, resulting in adjacency time out. This leads to loss of routes on core routers pertaining to access-layer connected to the VSS and loss of routes on VSS for any upstream connectivity.
- The NSF signal timer expired because remote routers did not respond to the NSF restart hellos from new active switch. This is possible because the remote routers (core) hashing may send hello and NSF hello-ack to old active. Subsequently, the NSF time out will be detected by new active, which will prevent a graceful recovery. As a result, a fresh adjacency restart is initiated.
- The NSF restart process got stuck during the route update process (for example, the update of routes were sent to the old-active switch using unicast hashing) and the new active supervisor declares it adjacency is stuck in the INIT state and forces a fresh restart.

The location of adjacency settlement after the dual-active event determines the variation in convergence:

If the IP addressing is configured such that the adjacency can settle on the new-active switch during dual-active condition where it might not have to undergo a fresh restart of the adjacency, it may lead to a better convergence. However, this will not be consistent during next dual-active condition as the adjacency now settles on old-active, leading to one of trigger conditions described in preceding paragraph.

If the adjacency settles on old-active switch, after 22-to-25 seconds the dual-active event triggers an internal shutdown (different from an administrative shutdown) which then causes the adjacency process to restart with the new active switch that will go through a normal adjacency setup (not an NSF graceful restart). Adjacency resets result in routes from the core and the VSS being withdrawn and subsequent variable packet loss for downstream and upstream paths.

It is possible that on a given network, one may only see partial symptoms. It is difficult to fully and consistently characterize traffic disruption due to convergence because of the many variables involved. The key point is to have a reasonable understanding of the many factors that influence convergence during BFD detection and that these factors can cause convergence following BFD detection to take higher than any other combination.

Convergence from Dual-Active Events with OSPF

OSPF inherently requires unique connectivity for building and maintaining the shortest-path-first (SPF) database. It has two built-in verification check that creates more network visibility under a dual-active condition-router ID and bidirectional verification of neighbor reachability.

Enhanced PAgP and Fast Hello

The ECMP-based core design experiences a higher rate of traffic loss because OSPF removes access-layer routes in the core during a dual-active event. OSPF does this because of the duplicate router IDs seen by the core routers.

The convergence is much better with MEC-based topology. An MEC-based core design does not suffer from route removal in the core because no OSPF route withdrawals are sent (EtherChannel interfaces are still operational). In addition, the detection of dual-active is triggered within 2-to-3 second followed with recovery of control plane on new active switch. During the recovery both switch member interfaces keep forwarding data leading to convergence times that are below one second in duration.

BFD

The ECMP-based core design experiences better convergence compared to the enhanced PAgP design because of the delayed recovery action by BFD that leaves at least one access-layer route operational in the core.

The recovery with a MEC-based core is more complicated. The MEC core design has a higher rate of traffic loss because of adjacency destabilization that leads to the removal of routes. This affects both upstream and downstream convergence. In addition, downstream losses are increased because the core removes routes for the access-layer subnet faster because it detects the adjacency loss. In contrast, the VSS retains the adjacency and the upstream routes. Figure 4-23 compares dual-active convergence given differing configuration options.

Figure 4-23 VSS Dual-Active Convergence with OSPF

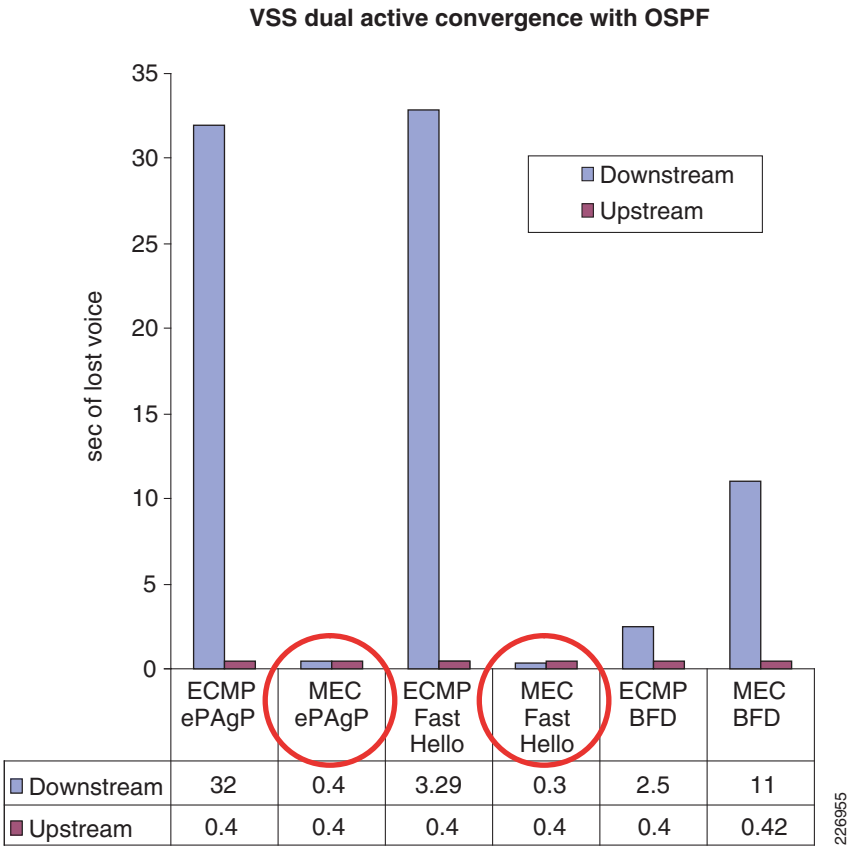


Table 4-14 describes the losses and root causes of the convergence for each combination.

Table 4-14 Convergence Recovery Losses and Causes (OSPF Environment)

Dual-Active Detection Protocol	Core-to-VSS Connectivity	End-to-End Convergence	Summary of Recovery and Loss Process During a Dual-Active Event
Enhanced PAgP	ECMP	Downstream 30-to-32 Second Upstream 200-to-400 msec	Downstream traffic loss is higher because all four routes to the access-layer are withdrawn. The loss of VSL link forces the old active switch to remove all interfaces on the hot-standby switch (the new active switch) as being disabled (although they are actually operational); this route updates is announced to core routers , which causes the withdrawal of two routes learned from the new active switch (even though they are operational). Enhanced PAgP or fast-hello detection shuts down the old active switch interfaces which triggers the withdrawal of a second set of routes from core routing table. Until the new active switch completes its NSF restart and sends its routes to the core routers, downstream traffic will be black holed. The upstream route removal does not happen because no duplicate router ID is seen by the VSS.
	MEC	Downstream and upstream 200-to-400 msec	No routes are removed during EtherChannel link change. Before the new active switch announces the adjacency restart; the old active switch is shut down (2.5 seconds). A clean NSF restart occurs (no routes were withdrawn for the same restarting NSF peer before the new active switch announces the NSF adjacency restart (7-to-12 seconds).
BFD	ECMP	Downstream 2-to-2.5 sec Upstream: 200-to-400 msec	Similar to the preceding enhanced PAgP-ECMP case, but BFD does not disconnect the old active switch for 22-to-25 seconds which keeps at least one route in core for access-layer subnets. Keeping that route helps to prevent traffic being black-holed until the NSF recovery process is completed on new active switch.
	MEC	Downstream: 200 msec to 11 seconds Upstream: 200-to-400 msec	Traffic is affected by several events occurring in parallel. OSPF adjacency might not stabilize until BFD shuts down the old active interfaces (22-to-25 seconds). Depending on which VSS member is active and where the OSPF hello is hashed, the stability of NSF restart might be affected. If the NSF restart occurs cleanly, the convergence can be below one second.

Details of BFD Detection with MEC-Based Topology

As described before, the asymmetrical hashing of hello (multicast) and update (unicast) messages from the core to VSS is possible with MEC in normal operational circumstances, as well as under a dual-active condition. From the VSS to the core, control plane connectivity remains on local interfaces. This combination of behaviors can cause the reset of the adjacency in a dual-active condition. In addition, OSPF adjacency might never stabilize with either of the active VSS routers because of bidirectional neighbor availability verification in the OSPF hello protocol for adjacency formation.

In normal operating condition, the core routes view a single router ID for VSS. During dual-active, core routes will see the same router ID be announced to two active supervisors. The core routers SPF is in state of confusion and will display the duplicate router ID in syslogs if the detail adjacency logging is turned on under OSPF process. For the OSPF adjacency settlement, the core routers will respond to the request coming from either the old or new VSS active switch, not knowing what really happened.

However, the core routers send the multicast hello to either old or new active switch depending on hashing (source IP of the interface address and destination of 224.0.0.5). During a dual-active event, two possibilities arise in which the hello can be sent from the core:

- Link connected to new active switch—While the core is sending the hello to the link connected to new active switch, the old active router is up and continues sending normal OSPF hellos back to core via its local link. At the same time, the new-active router will try to establish adjacency and restart its connection to the core routers by sending special hellos with the RS bit set (NSF restart). This adjacency restart might be continued until the hello from old active without NSF/RS bit set is received via the core router (old active router is up and running because it does not know what happened). This leads to confusion in the core router's NSF aware procedure and that might cause the core router to reset its adjacency. Meanwhile, the old-active router might also time out because it has not received any hellos from the core. Eventually, either of the active VSS switch neighbors will reset the adjacency. Once the adjacency reset is triggered, the core router will retry to establish neighbor adjacency to the new-active router (due to hashing) reaching FULL ADJ, meanwhile the old active router will try to send hello again, this time core routers do not see its own IP address in the received hello as it is an INIT hello from an old active. This will prompt the core to send fast-hello transmissions and new Database Descriptor (DBD) sequence number to the new active router (as it was almost at FULL ADJ with new active router). The new active router complains this with a BAD_SEQUENCE number and resets the ADJ from FULL to EX-START.
- Link connected on old active switch—In this case, hashing turns out to be such that core sends hello messages to the old-active switch and the new active will start first with NSF restart and then INIT hello. The new-active router does not see the response received from the core routers as core routers keep sending response to the old active. As a result, eventually the adjacency restart will be issued by the new active supervisor. This will continue indefinitely if no detection mechanism is employed or (in case of BFD) when the adjacency reset might cause higher packet loss.

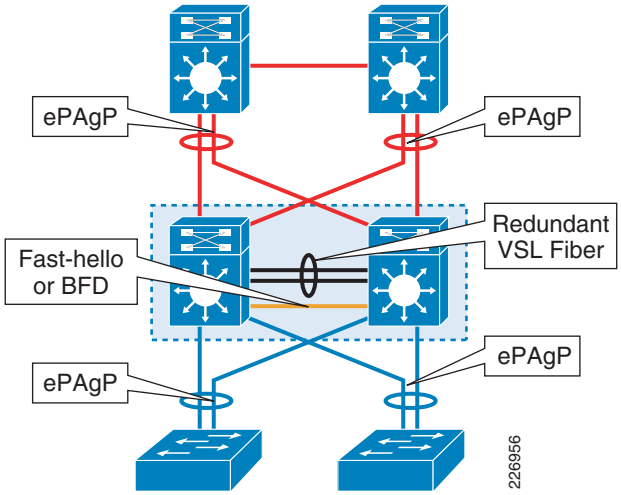
Dual-Active Method Selection

It is obvious that multiple techniques are possible for deployment. Multiple detection techniques deployment is not the replacement for resilient VSS-link configuration. In the presence of multiple detection techniques, enhanced PAgP and fast-hello will be detected first when compared to BFD. In the presence of multiple methods, whoever detects the dual-active first governs the convergence.

The only exception in deploying multiple methods is where OSPF routing enabled with ECMP-based topology. In this topology, the only detection method recommended is BFD, because BFD is the only method that gives the best convergence. If any other method is deployed along with BFD, the BFD will not be the first to detect the dual-active (BFD takes longer time compared to enhanced PAgP or fast-hello).

Enhanced PAgP detection is possible via Layer-2 or Layer-3 MEC. Enhanced PAgP detection might only need to be run on a single neighbor. However, using enhanced PAgP on all interfaces will ensure that, in the worst case, at least one switch is connected to both members of the same VSS pair (assuming that not all cable paths are affected in a failure condition) so that a path will exist for recovery. Figure 4-24 illustrates a high-level view of a topology featuring multiple redundancies to ensure VSL link availability and the placement of detection tools to help reduce traffic disruption under dual-active event conditions.

Figure 4-24 Dual-Active Detection Possibilities



Summary and Recommendations

This section summarize the recommended dual-active detection techniques that one must enable. These recommendations are based on the previous validated convergence behavior possible with the given combination of topology and dual-active detection method. The following recommendations apply:

- More then one detection technique should be deployed at once to avoid failure of a single technique jeopardizing the downtime.
- If ePAgP method is used, enable ePAgP detection on more then one access-layer/core switch.
- ePAgP and Fast Hello detection methods enabled with Layer 3 MEC topology in the core provides sub-second convergence for both OSPF and EIGRP.
- With Pre-12.2(33)SXI3 release, in conjunction with OSPF and ECMP topology where BFD detection should be used, do not deployed any other technique. BFD is being deprecated in future software releases. The BFD detection method should **only** be used for releases prior to 12.2(33)SXI3; see the next recommendation.
- 12.2(33)SXI3 release allows sub-second operation of fast-hello function for rapid detection and recovery (CSCsy30937). This improvement enables sub-second recovery, eliminating topological and routing protocol dependency of dual active detection and recovery.

The new improvements in Fast Hello detection do not alter the convergence results described in this design guide. Refer following design guide for the latest information: http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/Borderless_Campus_Network_1.0/Borderless_Campus_1.0_Design_Guide.html.

Table 4-15 provides the quick summary of these recommendations, where “Good” means recovery below one second and “Ok” means recovery takes more then one second and variable up to 32 seconds.

Table 4-15 Summary of Recovery Comparisons for Convergence Options

	Pre-12.2(33)SXI3			With 12.2(33)SXI3
Dual Active Detection	**ePAGP**	**BFD**	**Fast Hello**	**Sub-second Fast Hello**
EIGRP with ECMP-based topology to the core	Good	Good	Good	Good
EIGRP with L3-MEC-based topology to the core	Good	OK	Good	Good
OSPF with ECMP-based topology to the core	OK	Good	OK	Good
OSPF with L3-MEC-based topology to the core	Good	OK	Good	Good

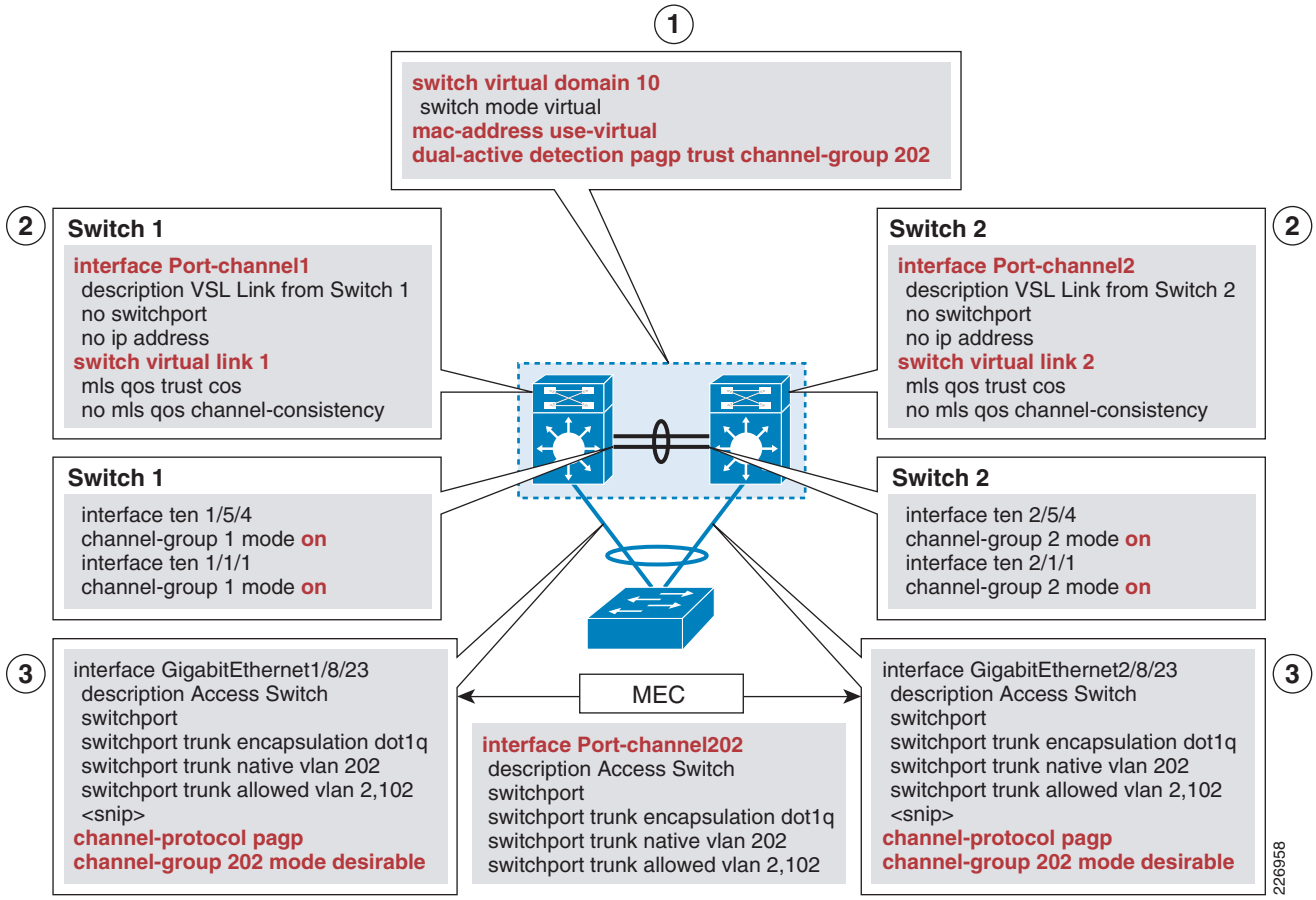
Note As seen from above and other references discussed in this design guide, MEC-based connectivity to the core enables convergence times below one seconds for both unicast and multicast.

APPENDIX A

VSS-Enabled Campus Best Practice Configuration Example

Figure A-1 illustrates the baseline best practice configuration required to set up basic VSS enabled network. The circle indicates the essential steps required to create the VSS systems from standalone. The `red` text highlights the important CLI information with VSS configuration. Comments are provided in *`blue italic`* font.

Figure A-1 Overall VSS-Enabled Campus Best Practice Configuration Summary



End-to-End Device Configurations

The end-to-end devices configuration is categorized into three major sections. Each section configuration contains specific CLI which is a required as part of best practice configuration and corresponding explanation.

- VSS and L2 Domain- Includes above base configuration as well as L2 domain configuration
- Access-layer- Sample L2 domain configuration
- L3 Domain - Includes global L3 configuration for VSS and core routers. Then separate section for specifics topologies (ECMP and MEC) for EIGRP and OSPF. In addition, the core devices configuration shown below are standalone router/devices.

VSS Specific

VSS Global Configuration

```
switch virtual domain 10 ! Must configure unique domain ID
 switch mode virtual
 switch 1 priority 110 ! Not needed, helps in operational mgmt
 switch 2 priority 100 ! Not needed, helps in operational mgmt
 dual-active exclude interface GigabitEthernet1/5/3 ! Connectivity to VSS during dual
active
mac-address use-virtual ! Required for consistent MAC address
dual-active detection pagp trust channel-group 202 ! Enhanced PAgP based dual active
detection

redundancy ! Default SSO Enabled
 main-cpu
  auto-sync running-config
 mode sso
```

Switch 1

```
interface Port-channel1 ! Unique port-channel number for SW 1
 description VSL Link from Switch 1
 no switchport
 no ip address
switch virtual link 1 ! Defines switch ID for SW 1
 mls qos trust cos
 no mls qos channel-consistency

interface ten 1/5/4
channel-group 1 mode on ! EC mode is ON – EtherChannel Managemeent Protocol off
interface ten 1/1/1
channel-group 1 mode on
```

Switch 2

```
interface Port-channel2 ! Unique port-channel number for SW 1
 description VSL Link from Switch 2
 no switchport
 no ip address
switch virtual link 2 ! Defines switch ID for SW 2
 mls qos trust cos
 no mls qos channel-consistency

interface ten 2/5/4
```

```
channel-group 2 mode on ! EC mode is ON - EtherChannel Managemeent Protocoloff
interface ten 2/1/1
channel-group 2 mode on
```

Layer-2 Domain

VSS

```
udld enable
vtp domain campus-test
vtp mode transparent

spanning-tree mode rapid-pvst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree vlan 2-999 priority 24576 !  STP Root

port-channel load-balance src-dst-mixed-ip-port !  Enhanced hash algorithem


vlan 400 !  VLANs spanning multiple access-layer SWs
 name L2_ Spaned_VLAN _400

vlan 450
 name L2_ Spaned_VLAN _450

vlan 500
 name L2_ Spaned_VLAN _500

vlan 550
 name L2_ Spaned_VLAN _550

vlan 600
 name L2_ Spaned_VLAN _600

vlan 650
 name L2_Spaned_VLAN_650

vlan 900
 name NetMgmt_VLAN_900

vlan 999
 name Unused_Port_VLAN_999

vlan 2
 name cr7-3750-Stack-Data-VLAN
!
vlan 102
 name cr7-3750-Stack-Voice-VLAN

interface Vlan2 ! Sample VLAN interface configuration
 ip address 10.120.2.1 255.255.255.0
 no ip redirects
 no ip unreachables
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 hold-queue 150 in
 hold-queue 150 out
!
```

VSS—Multi-Chassis EtherChannel

PAgP

```
interface GigabitEthernet1/8/23  !  Interface on SW 1
 description Access Switch Facing Interface
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
switchport mode dynamic desirable ! Trunk mod dynamic and desirable
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900 !  Only allow need VLANs
for a given trunk
logging event link-status !  Logging for link status
logging event trunk-status !  Logging for trunk status
logging event bundle-status !  Logging for port-channel status
 load-interval 30
 mls qos trust dscp
channel-protocol pagp
channel-group 202 mode desirable !  Define Port-channel, PAgP mode desirable


interface GigabitEthernet2/8/23  !  Interface on SW 2
 description Access Switch Facing Interface
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
switchport mode dynamic desirable
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900
 logging event link-status
 logging event trunk-status
 logging event bundle-status
 load-interval 30
 mls qos trust dscp
 load-interval 30
channel-protocol pagp
channel-group 202 mode desirable

interface Port-channel202  !  Automatically created by defining at interfaces
 description Access Switch MEC
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900
logging event link-status
logging event spanning-tree status ! STP logging enabled on port-channel
 load-interval 30
 mls qos trust dscp
 spanning-tree portfast ! Optional - helps during initialization
 hold-queue 2000 out
```

LACP

LACP Sample Configuration

```
interface GigabitEthernet1/8/23
 description Access Switch Facing Interface
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
switchport mode dynamic desirable
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900
logging event link-status
```

```
logging event trunk-status
logging event bundle-status
 load-interval 30
 mls qos trust dscp
 channel-protocol lacp
 channel-group 202 mode active
 hold-queue 2000 out

interface GigabitEthernet2/8/23
 description Access Switch Facing Interface
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
switchport mode dynamic desirable
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900
logging event link-status
logging event trunk-status
logging event bundle-status
 load-interval 30
 mls qos trust dscp
 channel-protocol lacp
 channel-group 202 mode active
 hold-queue 2000 out

interface Port-channel202 !  Automatically created by defining at interfaces
 description Access Switch MEC
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 202
 switchport trunk allowed vlan 2,102,400,450,500,550,600,650,900
logging event link-status
logging event spanning-tree status
 load-interval 30
 mls qos trust dscp
 spanning-tree portfast ! Optional - helps during initialization
 hold-queue 2000 out
```

Access-Layer Switch

Sample Configuration (Platform Specific Configuration Varies)

```
interface GigabitEthernet0/27
 description Uplink to VSS Switch Gig 1/8/24
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 203
switchport mode dynamic desirable
 switchport trunk allowed vlan 3,103,400,450,500,550,600,650,900
 logging event link-status
 logging event trunk-status
 logging event bundle-status
 carrier-delay msec 0
 srr-queue bandwidth share 1 70 25 5
 srr-queue bandwidth shape  3  0  0  0
 priority-queue out
 mls qos trust dscp
channel-protocol pagp
channel-group 1 mode desirable

interface GigabitEthernet0/28
 description Uplink to VSS Switch Gig 2/8/24
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 203
```

```
 switchport trunk allowed vlan 3,103,400,450,500,550,600,650,900
 switchport mode dynamic desirable
 logging event link-status
 logging event trunk-status
 logging event bundle-status
 carrier-delay msec 0
 srr-queue bandwidth share 1 70 25 5
 srr-queue bandwidth shape  3  0  0  0
 priority-queue out
 mls qos trust dscp
channel-protocol pagp
channel-group 1 mode desirable


interface Port-channel1 !  Automatically created by defining at interfaces
 description EC Uplink to VSS
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 203
 switchport trunk allowed vlan 3,103,400,450,500,550,600,650,900
 switchport mode dynamic desirable
logging event link-status
logging event spanning-tree status
 carrier-delay msec 0
 spanning-tree portfast ! Optional - helps during initialization
```

Layer-3 Domain

The Layer 3 domain represents VSS interconnection to the core-layer. The core-layer devices configuration shown below are standalone router/switch.

Global Configuration

```
mls ip cef load-sharing <option> !  Apply Campus Best Practices
```

Multicast

VSS

```
ip multicast-routing
ip pim rp-address 10.122.100.1 GOOD-IPMC override !  RP mapping with filter

ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```

Core 1 Standalone Router (No VSS)

Core RP ANYCAST - Primary

```
ip multicast-routing

interface Loopback0
 description MSDP PEER INT !  MSDP Loopback
 ip address 10.122.10.1 255.255.255.255
```

```
interface Loopback1
 description ANYCAST RP ADDRESS (PRIMARY) !  Anycast RP Primary
 ip address 10.122.100.1 255.255.255.255

interface Loopback2
 description Garbage-CAN RP
 ip address 2.2.2.2 255.255.255.255

interface Port-channel1 !  Core 1- Core2 L3 for MSDP
 description Channel to Peer Core Node
 dampening
 ip address 10.122.0.18 255.255.255.254
 ip pim sparse-mode
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp

ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255

ip msdp peer 10.122.10.2 connect -source Loopback0 ! MSDP Configuration
ip msdp description 10.122.10.2 ANYCAST-PEER-6k-core-2
ip msdp cache -sa-state
ip msdp originator-id Loopback0
```

Core 2 Standalone Router (No VSS)

```
ip multicast-routing

interface Loopback0
 description MSDP PEER INT
 ip address 10.122.10.2 255.255.255.255

interface Loopback1
 description ANYCAST RP ADDRESS
 ip address 10.122.100.1 255.255.255.255 ! Secondary ANYCAST RP
 delay 600

interface Loopback2
 description Garbage-CAN RP
 ip address 2.2.2.2 255.255.255.255

interface Port-channel1
 description Channel to Peer Core node
 dampening
 ip address 10.122.0.19 255.255.255.254
 ip pim sparse-mode
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp

ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```

```
ip msdp peer 10.122.10.1 connect-source Loopback0
ip msdp description 10.122.10.1 ANYCAST-PEER-6k-core-1
ip msdp cache-sa-state
ip msdp originator-id Loopback0
```

EIGRP MEC

VSS

```
router eigrp 100
 passive-interface default
 no passive-interface Port-channel200
 no passive-interface Port-channel201
 network 10.0.0.0
eigrp log-neighbor-warnings
eigrp log-neighbor-changes
 no auto-summary
 eigrp router-id 10.122.102.1
 eigrp event-log-size 3000
 nsf !  Enable NSF Capability

interface Port-channel200 !  Create L3 MEC Interface first
 description 20 Gig MEC to CORE-1 (cr2-6500-1 4/1-4/3)
 no switchport
 dampening
 ip address 10.122.0.26 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.125.0.0 255.255.0.0 5 !  Summarization for Access-subnets
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface Port-channel201
 description 20 Gig to CORE-2 (cr2-6500-1 4/1-4/3)
 no switchport
 dampening
 ip address 10.122.0.21 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.125.0.0 255.255.0.0 5
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet1/2/1
 description 10 GigE to Core 1
 no switchport
 no ip address
 logging event link-status
 logging event bundle-status
 load-interval 30
 carrier-delay msec 0
```

```
 mls qos trust dscp
 channel-protocol pagp
 channel-group 200 mode desirable
 hold-queue 2000 in
 hold-queue 2000 out
!
interface TenGigabitEthernet1/2/2
 description 10 GigE to Core 2
 no switchport
 no ip address
 logging event link-status
 logging event bundle-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 channel-protocol pagp
 channel-group 201 mode desirable
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet2/2/1
 description to core 1
 no switchport
 no ip address
 logging event link-status
 logging event bundle-status
 logging event spanning-tree status
 load-interval 30
 mls qos trust dscp
 channel-protocol pagp
 channel-group 200 mode desirable
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet2/2/2
 description 10 GigE to Core 2
 no switchport
 no ip address
 logging event link-status
 logging event bundle-status
 load-interval 30
 mls qos trust dscp
 channel-protocol pagp
 channel-group 201 mode desirable
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 1 Standalone Router (No VSS)

```
router eigrp 100
 passive-interface default
 no passive-interface Port-channel1
 no passive-interface Port-channel20
 no passive-interface Port-channel221
 network 10.0.0.0
 no auto-summary
eigrp log-neighbor-warnings
eigrp log-neighbor-changes
eigrp event-log-size 3000

interface Port-channel20
 description 20 Gig MEC to VSS 1/2/1 2/2/1
```

```
 dampening
 ip address 10.122.0.27 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 2 Standalone Router (No VSS)

```
router eigrp 100
 passive-interface default
 no passive-interface Port-channel1
 no passive-interface Port-channel20
 no passive-interface Port-channel221
 network 10.0.0.0
 no auto-summary
eigrp log-neighbor-warnings
eigrp log-neighbor-changes
eigrp event-log-size 3000

interface Port-channel21
 description 20 Gig to VSS 1/2/2-2/2/2
 dampening
 ip address 10.122.0.20 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

EIGRP ECMP

VSS

```
router eigrp 100
 passive-interface default
 no passive-interface TenGigabitEthernet1/2/1
 no passive-interface TenGigabitEthernet1/2/2
 no passive-interface TenGigabitEthernet2/2/1
 no passive-interface TenGigabitEthernet2/2/2
 network 10.0.0.0
 no auto-summary
 eigrp router-id 10.122.102.1
 eigrp log-neighbor-warnings
 eigrp log-neighbor-changes
 eigrp event-log-size 3000
 nsf !  Enable NSF Capability

interface TenGigabitEthernet1/2/1
 description 10 GigE to Core 1
 no switchport
 dampening
```

```
 ip address 10.122.0.26 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface TenGigabitEthernet1/2/2
 description 10 GigE to Core 2
 no switchport
 dampening
 ip address 10.122.0.23 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet2/2/1
 description to Core 1
 no switchport
 dampening
 ip address 10.122.0.32 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface TenGigabitEthernet2/2/2
 description 10 GigE to Core 2
 no switchport
 dampening
 ip address 10.122.0.20 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 1 Standalone Router (No VSS)

```
router eigrp 100
 passive-interface default
 no passive-interface TenGigabitEthernet4/1
 no passive-interface TenGigabitEthernet4/3
 network 10.0.0.0
 no auto-summary
```

```
eigrp log-neighbor-warnings
eigrp log-neighbor-changes
eigrp event-log-size 3000

interface TenGigabitEthernet4/1
 description To VSS Ten1/2/1
 dampening
 ip address 10.122.0.27 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet4/3
 description To VSS Ten2/2/1
 dampening
 ip address 10.122.0.33 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 logging event bundle-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 2 Standalone Router (No VSS)

```
router eigrp 100
 passive-interface default
 no passive-interface TenGigabitEthernet4/1
 no passive-interface TenGigabitEthernet4/3
 network 10.0.0.0
 no auto-summary
eigrp log-neighbor-warnings
eigrp log-neighbor-changes
eigrp event-log-size 3000

interface TenGigabitEthernet4/1
 description To VSS Ten 1/2/2
 dampening
 ip address 10.122.0.22 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out

interface TenGigabitEthernet4/3
 description To VSS Ten 2/2/2
 dampening
 ip address 10.122.0.21 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
```

```
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

OSPF MEC

VSS

```
router ospf 100
 router-id 10.122.0.235
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000 !  Optional
 nsf !  Enable NSF Capability
 area 120 stub no-summary
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 area 120 range 10.125.0.0 255.255.0.0 cost 10
 passive-interface default
 no passive-interface Port-channel200
 no passive-interface Port-channel201
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
 network 10.125.0.0 0.0.255.255 area 120


interface Port-channel200
 description 20 Gig MEC to VSS (cr2-6500-1 4/1-4/3)
 no switchport
 dampening
 ip address 10.122.0.26 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface Port-channel201
 description 20 Gig to VSS (cr2-6500-1 4/1-4/3)
 no switchport
 dampening
 ip address 10.122.0.21 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 1 Standalone Router No VSS)

```
router ospf 100
 router-id 10.254.254.7
 log-adjacency-changes detail ! Helps in NSF Restart Activities
 auto-cost reference-bandwidth 20000 !  Optional
 passive-interface default
 no passive-interface Port-channel1
 no passive-interface Port-channel20
 network 10.122.0.0 0.0.255.255 area 0

interface Port-channel20
 description 20 Gig MEC to VSS 1/2/1 2/2/1
 dampening
 ip address 10.122.0.27 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 2 Standalone Router (No VSS)

```
router ospf 100
 router-id 10.254.254.7
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000 !  Optional
 passive-interface default
 no passive-interface Port-channel1
 no passive-interface Port-channel20
 network 10.122.0.0 0.0.255.255 area 0

interface Port-channel21
 description 20 Gig to VSS 1/2/2-2/2/2
 dampening
 ip address 10.122.0.20 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

OSPF ECMP

VSS

```
router ospf 100
 router-id 10.122.0.235
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000 !  Optional
 nsf !  Enable NSF Capability
 area 120 stub no-summary
```

```
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 area 120 range 10.125.0.0 255.255.0.0 cost 10
 passive-interface default
 no passive-interface TenGigabitEthernet1/2/1
 no passive-interface TenGigabitEthernet1/2/2
 no passive-interface TenGigabitEthernet2/2/1
 no passive-interface TenGigabitEthernet2/2/2
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
 network 10.125.0.0 0.0.255.255 area 120


interface TenGigabitEthernet1/2/1
 description 10 GigE to Core 1
 no switchport
 dampening
 ip address 10.122.0.26 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface TenGigabitEthernet1/2/2
 description 10 GigE to Core 2
 no switchport
 dampening
 ip address 10.122.0.23 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!


interface TenGigabitEthernet2/2/1
 description to Core 1
 no switchport
 dampening
 ip address 10.122.0.32 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
interface TenGigabitEthernet2/2/2
 description 10 GigE to Core 2
 no switchport
 dampening
 ip address 10.122.0.20 255.255.255.254
 ip flow ingress
```

```
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 1 Standalone Router (No VSS)

```
router ospf 100
 router-id 10.254.254.7
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000 !  Optional
 passive-interface default
 no passive-interface GigabitEthernet2/5
 no passive-interface TenGigabitEthernet4/1
 no passive-interface TenGigabitEthernet4/3
 no passive-interface Port-channel1
 network 10.122.0.0 0.0.255.255 area 0


interface TenGigabitEthernet4/1
 description To VSS Ten1/2/1
 dampening
 ip address 10.122.0.27 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
!
interface TenGigabitEthernet4/3
 description To VSS Ten2/2/1
 dampening
 ip address 10.122.0.33 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

Core 2 Standalone Router (No VSS)

```
router ospf 100
 router-id 10.254.254.7
 log-adjacency-changes detail
 auto-cost reference-bandwidth 20000 !  Optional
 passive-interface default
 no passive-interface GigabitEthernet2/5
 no passive-interface TenGigabitEthernet4/1
 no passive-interface TenGigabitEthernet4/3
 no passive-interface Port-channel1
 network 10.122.0.0 0.0.255.255 area 0
```

```
interface TenGigabitEthernet4/1
 description To VSS Ten 1/2/2
 dampening
 ip address 10.122.0.22 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
!
!
interface TenGigabitEthernet4/3
 description To VSS Ten 2/2/2
 dampening
 ip address 10.122.0.21 255.255.255.254
 ip flow ingress
 ip pim sparse-mode
 ip ospf network point-to-point
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
 hold-queue 2000 in
 hold-queue 2000 out
```

APPENDIX B

References

Revised: Month Day, Year, OL-19829-01

The following documents and reference links provide supplemental content supporting the Campus 3.0 VSS design presented in this publication:

- *Enterprise Campus 3.0 Architecture: Overview and Framework*
  http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/campover.html
- *Campus Network for High Availability Design Guide*
  http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/HA_campus_DG/hacampusdg.html
- *High Availability Campus Recovery Analysis*
  http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/HA_recovery_DG/campusRecovery.html
- *High Availability Campus Network Design: Routed Access Layer using EIGRP or OSPF*
  http://www.cisco.com/en/US/docs/nsite/campus/ha_campus_routed_access_cvd_ag.pdf
- *Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 - CCO White Paper on VSS technology*
  http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/white_paper_c11_429338.pdf
- *Best Practices for Catalyst 6500/6000 Series and Catalyst 4500/4000 Series Switches Running Cisco IOS Software*
  http://www.cisco.com/en/US/products/hw/switches/ps700/products_white_paper09186a00801b49a4.shtml
- *Migrate Standalone Cisco Catalyst 6500 Switch to Cisco Catalyst 6500 Virtual Switching System*
  http://www.cisco.com/en/US/products/ps9336/products_tech_note09186a0080a7c74c.shtml