CHAPTER 3

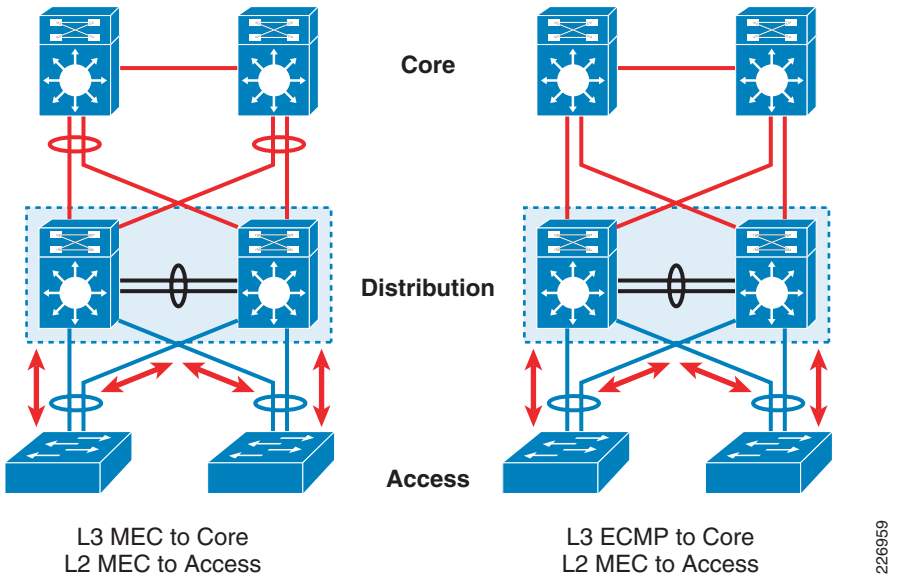
VSS-Enabled Campus Design

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
```

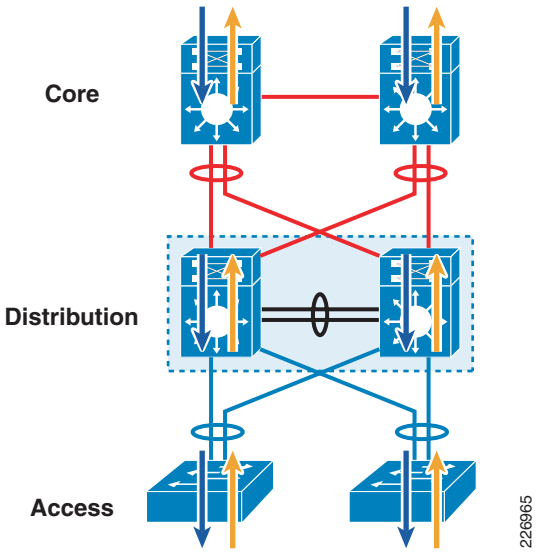
Cisco Catalyst 36xx, Cisco Catalyst 37xx Stack, and Cisco Catalyst 29xx:

```
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.

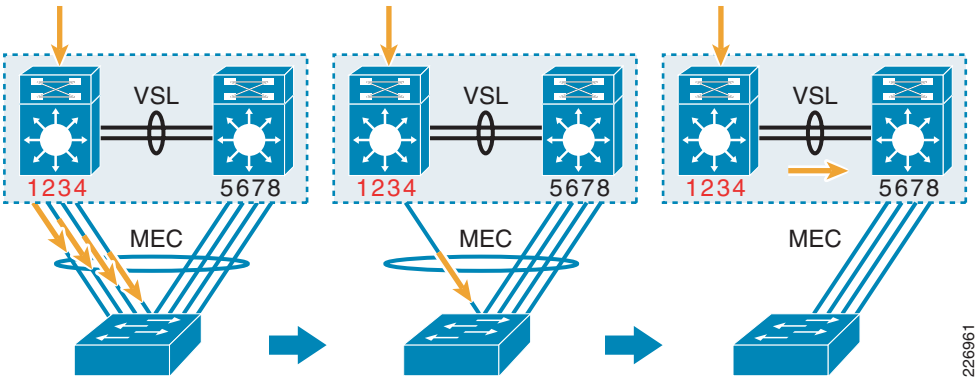
Figure 3-2 VSS Traffic Flow Overview



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

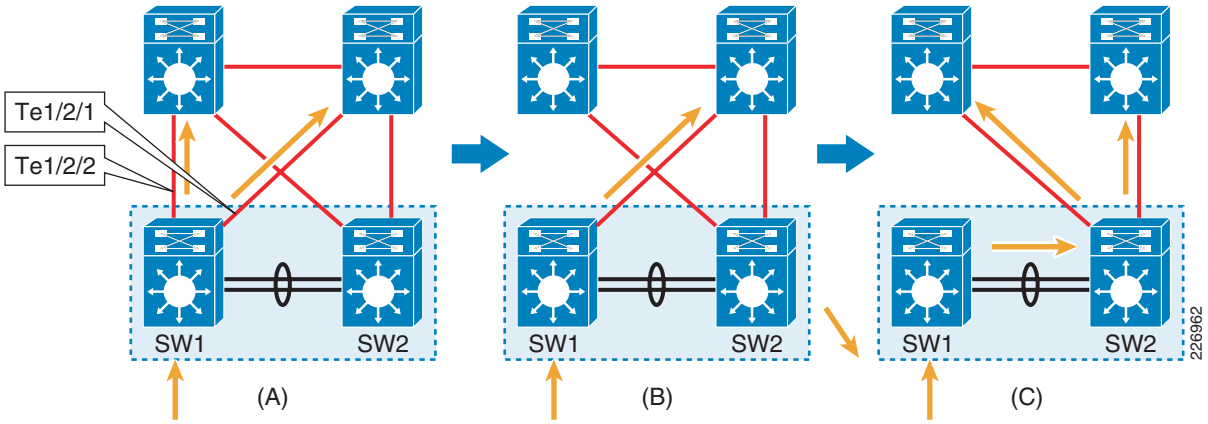
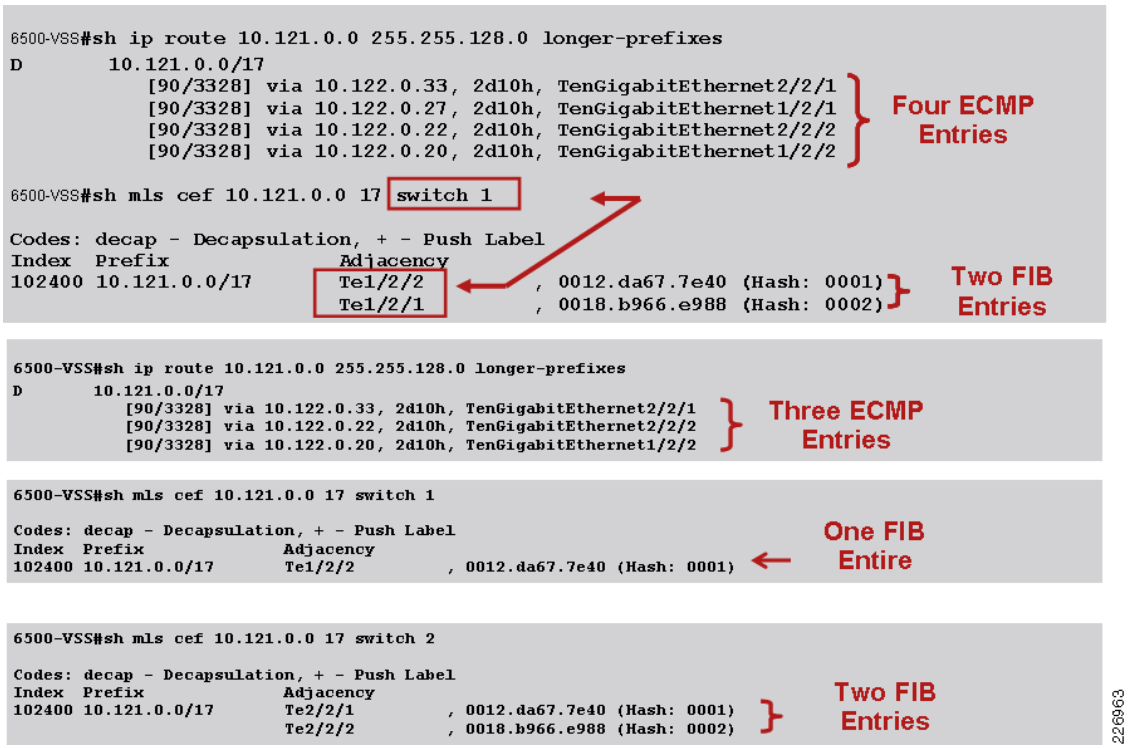


Figure 3-5 ECMP Forwarding Entries—Global and Switch Specifics



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)

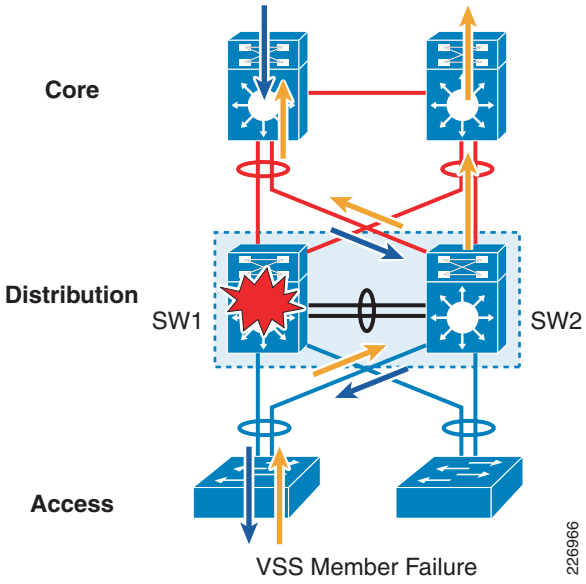
This section covers end-to-end traffic flow behavior with MEC-based topology. The convergence section (multilayer and routing) covers the ECMP-based topology and impact of failures in term packet loss.

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

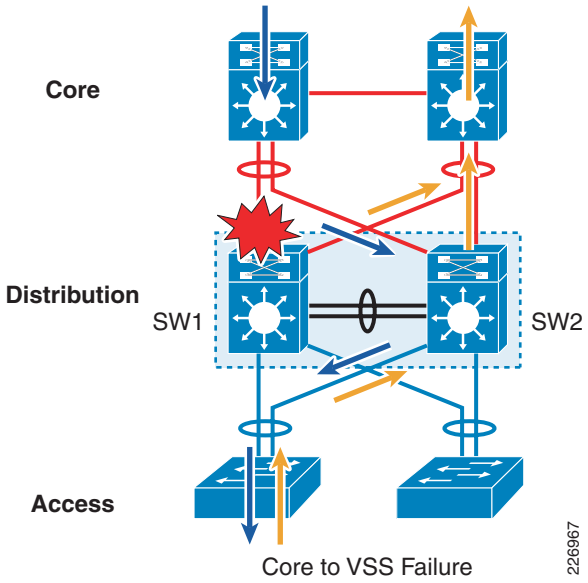


Core to VSS Failure

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

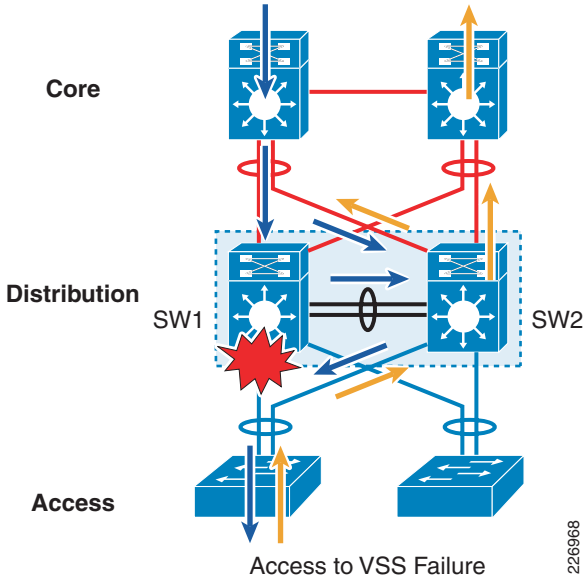


Access Layer-to-VSS Failure

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.

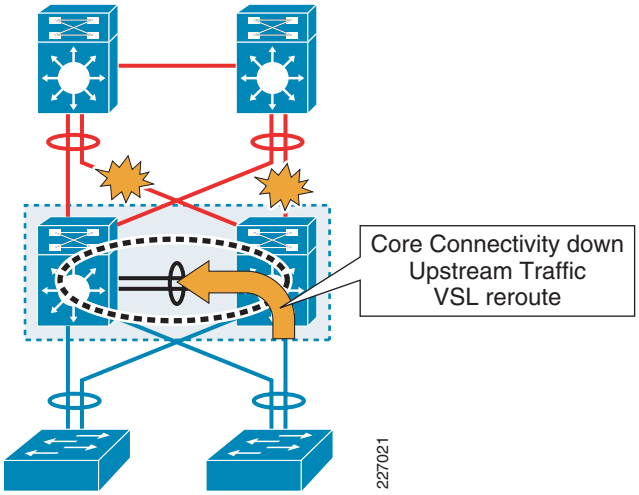
Capacity Planning for the VSL Bundle

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:

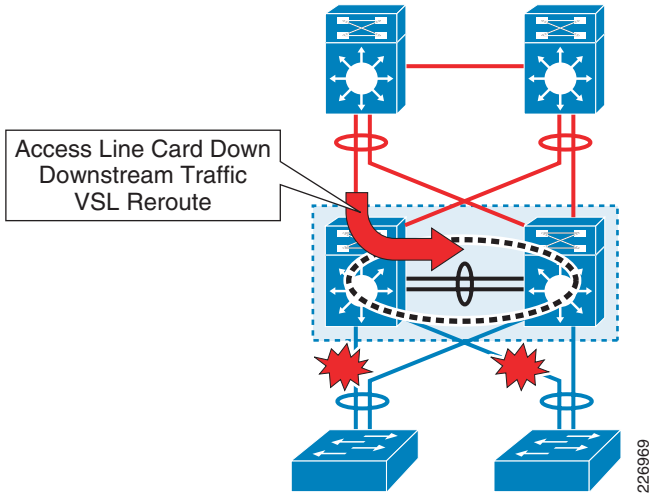
- Failure of all uplinks connected to a member of VSS to the core (Figure 3-9). In this failure, all upstream traffic traverses the VSL bundle. The number and speed of the uplinks limit the maximum traffic that can go over the VSL. Traditionally, in a full-mesh design, each switch member carries 20 Gigabits 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

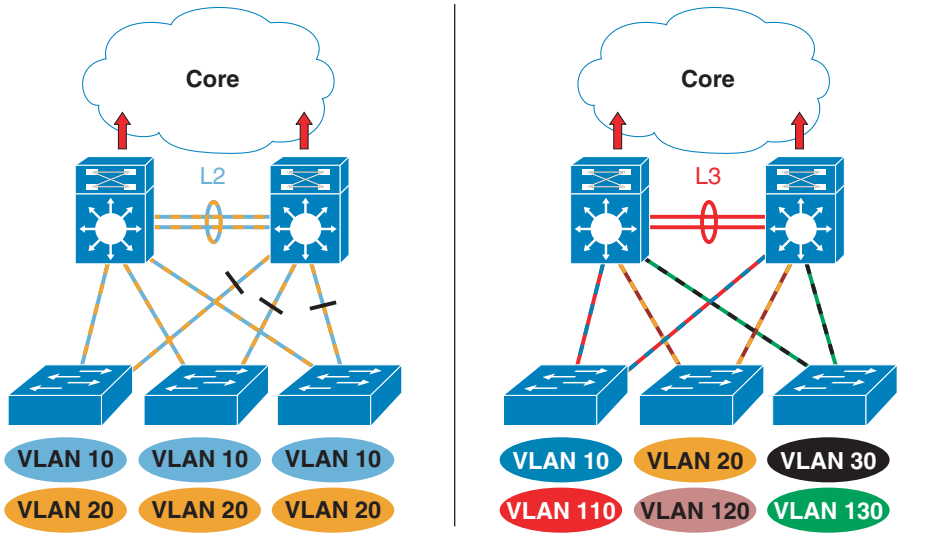
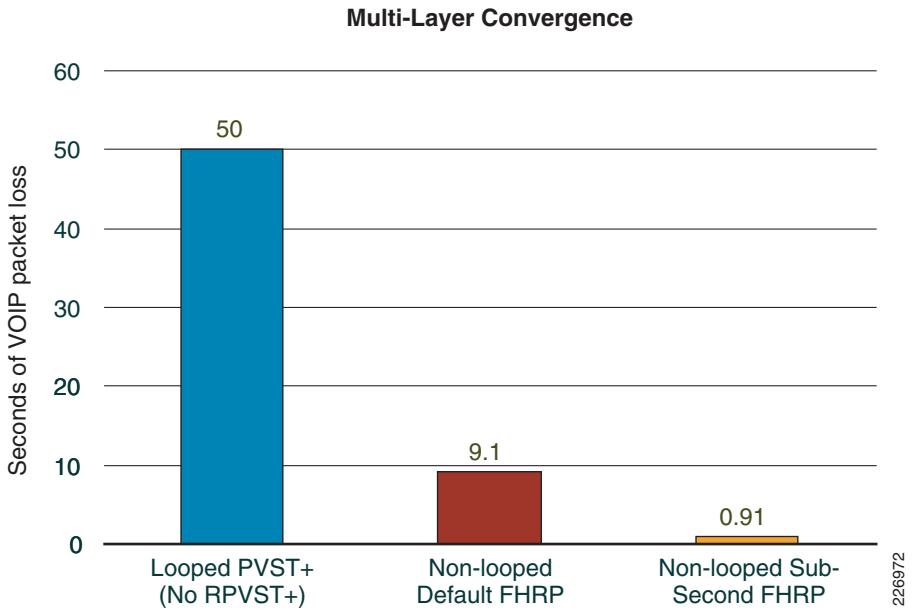


Table 3-1 summarizes the looped and non-looped design environments. Both designs use multiple control protocols, and the consistent application of tuning and configuration options to create a resilient design. In addition, the convergence described in Table 3-1 and illustrated in Figure 3-12 indicate 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

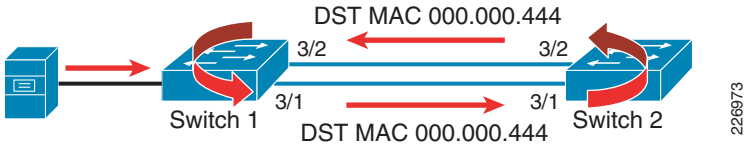
Figure 3-12 Multilayer Convergence Comparison



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

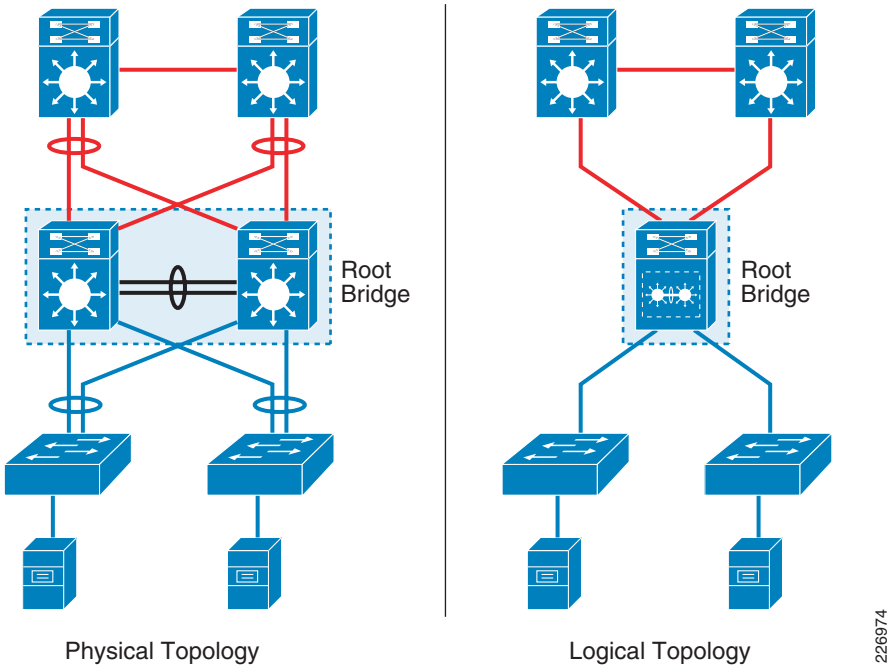
VSS Benefits

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



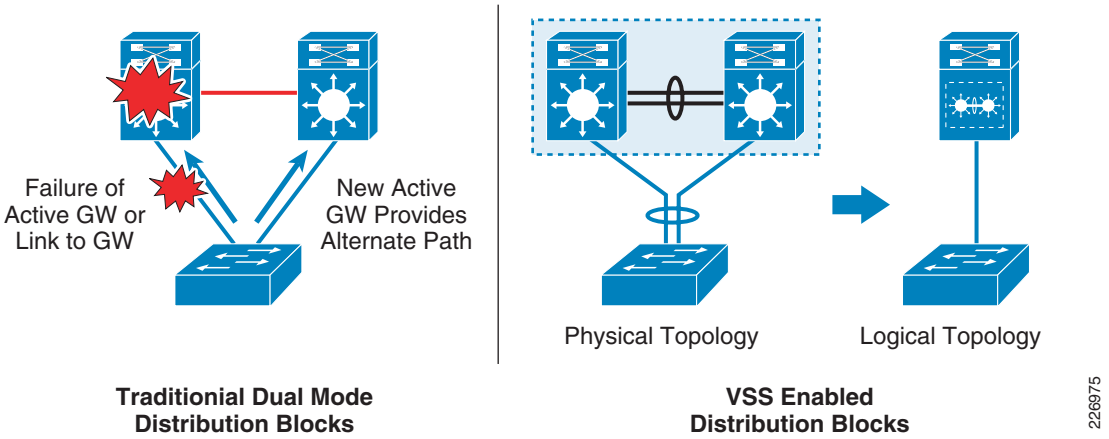
Elimination of FHRP Configuration

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
```

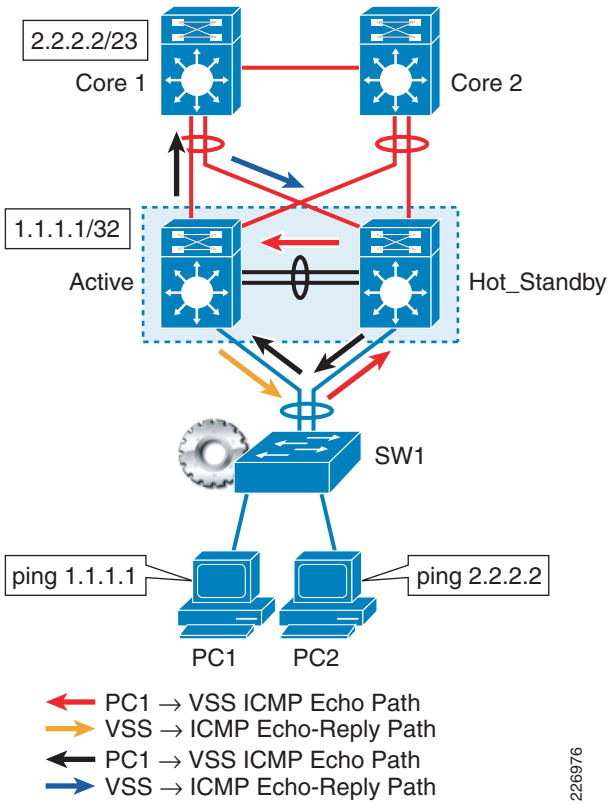
The above configuration would allow the Vlan200 SVI to take ownership of the HSRP group 200 MAC address, while not creating any dependency on HSRP group because it will not link it to the default gateway IP address. After the transition to VSS, hosts will continue sending frames to the HSRP MAC address. 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



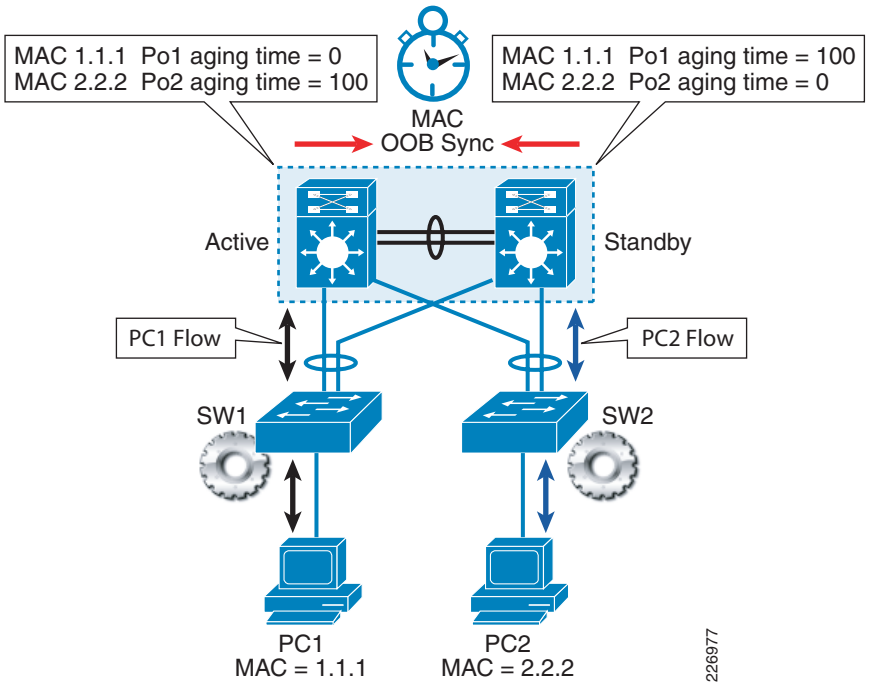
Layer-2 MAC Learning in the VSS with MEC Topology

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*.

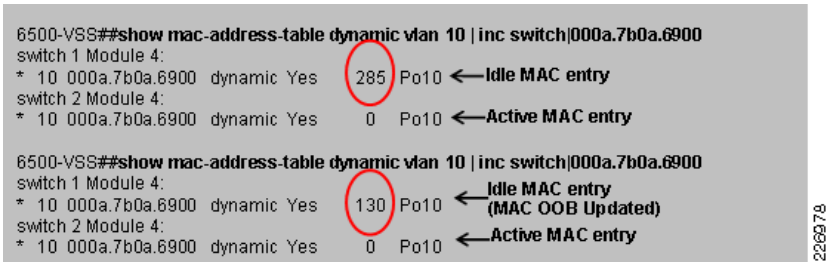
- *MAC Notification (MN)*—This is a hardware-based method for adding or removing a MAC address on a non-primary-entry line card in order to avoid continuous unicast flooding within the system. If the traffic for a MAC destination is presented at the DFC line card level, it will first flood that frame 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.

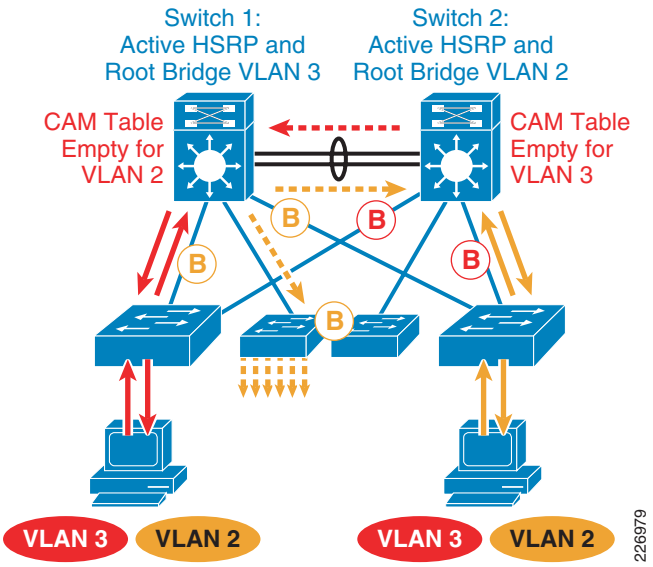
Note By default, the dynamic MAC entry aging-time on the Cisco Catalyst 6500 system with the 6708 module is set to 480 seconds. The MAC aging-timer must be changed from 300 seconds to 480 seconds manually if Cisco Catalyst 6500 with a non-6708 DFC module present in the switch. Starting in IOS Release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

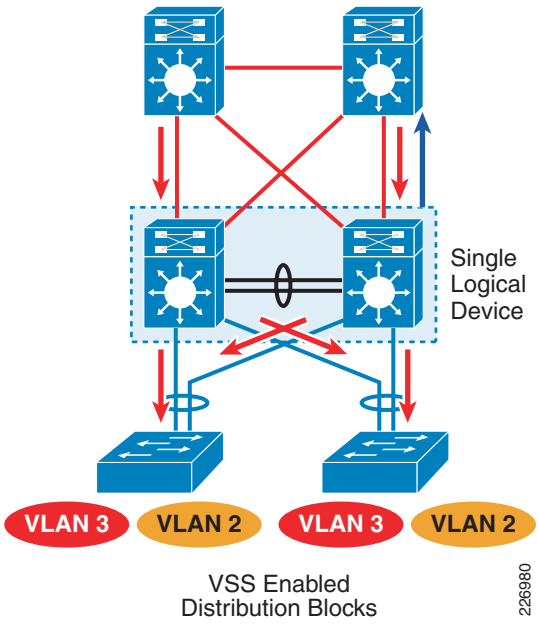


Unicast flooding is more pronounced with VLANs that span multiple access-layer switches. There is no inherent method to stop flooding of the traffic to every port that VLAN traffic exists. In general, there are three methods to reduce the effect of unicast flooding:

- 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

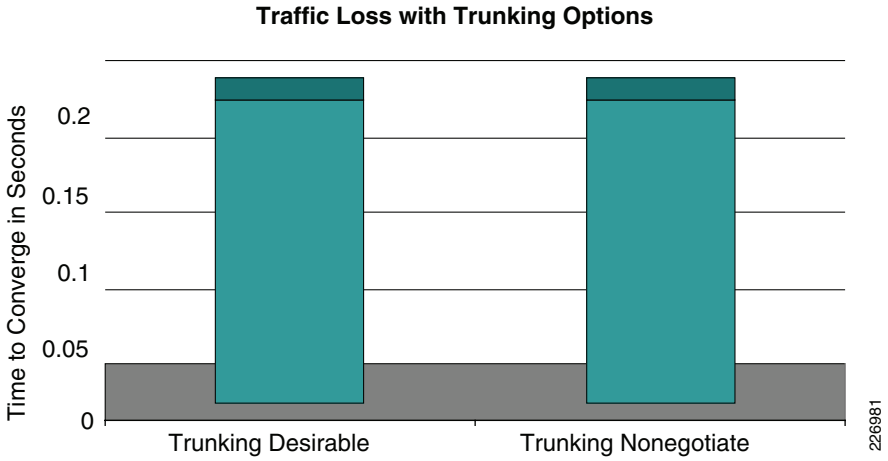


Figure 3-21 compares the convergence losses associated with trunk mode being desirable or nonnegotiable (non-desirable). With either configuration, the losses are less or equal to 200 msec. An additional benefit of running a trunk mode as desirable is that operational efficiency is gained in managing 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.

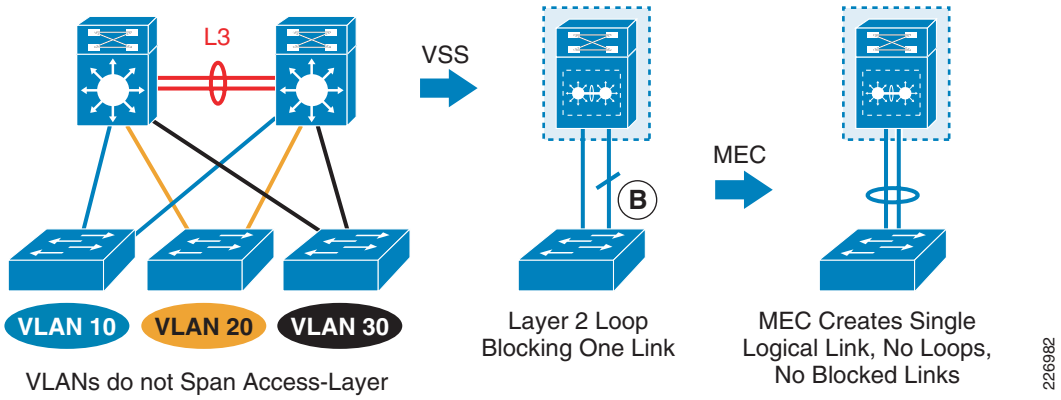
Tip **The aggressive UDLD should *not* be used as link-integrity check, instead use normal mode of UDLD to detect cabling faults and also for the link integrity.**

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:

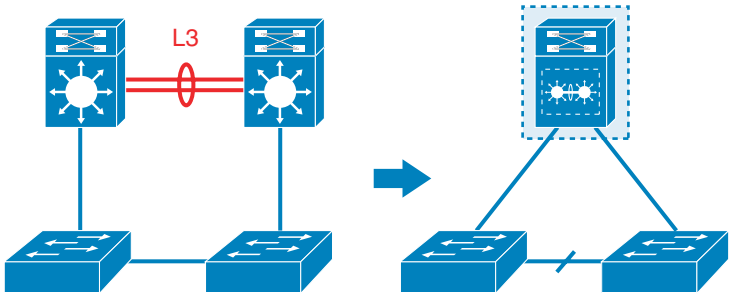
- Unicast flooding
- Looping (blocked link)

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.

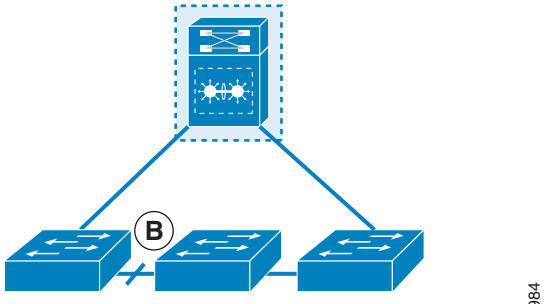
Figure 3-23 U-Shaped VSS Topology Loop Introduction



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.

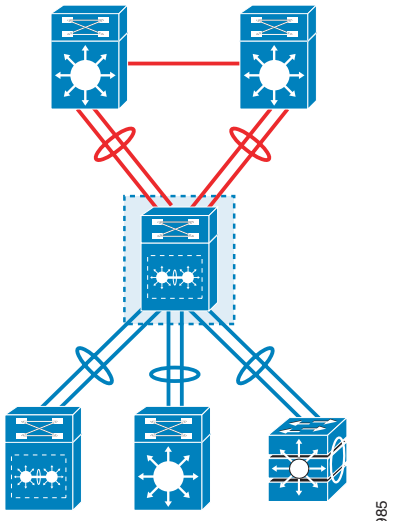
Figure 3-24 STP-blocked Link



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.

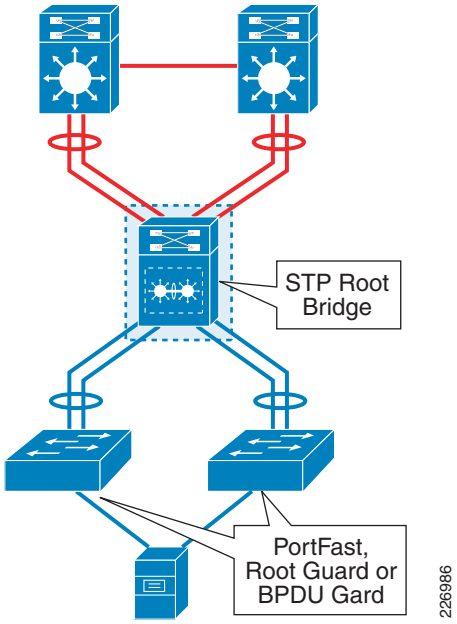
Tip **In the VSS-enabled network, it is critically important to keep the edge port from participating in the STP. Cisco strongly recommends enabling PortFast and BPDU Guard at the edge port.**

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.

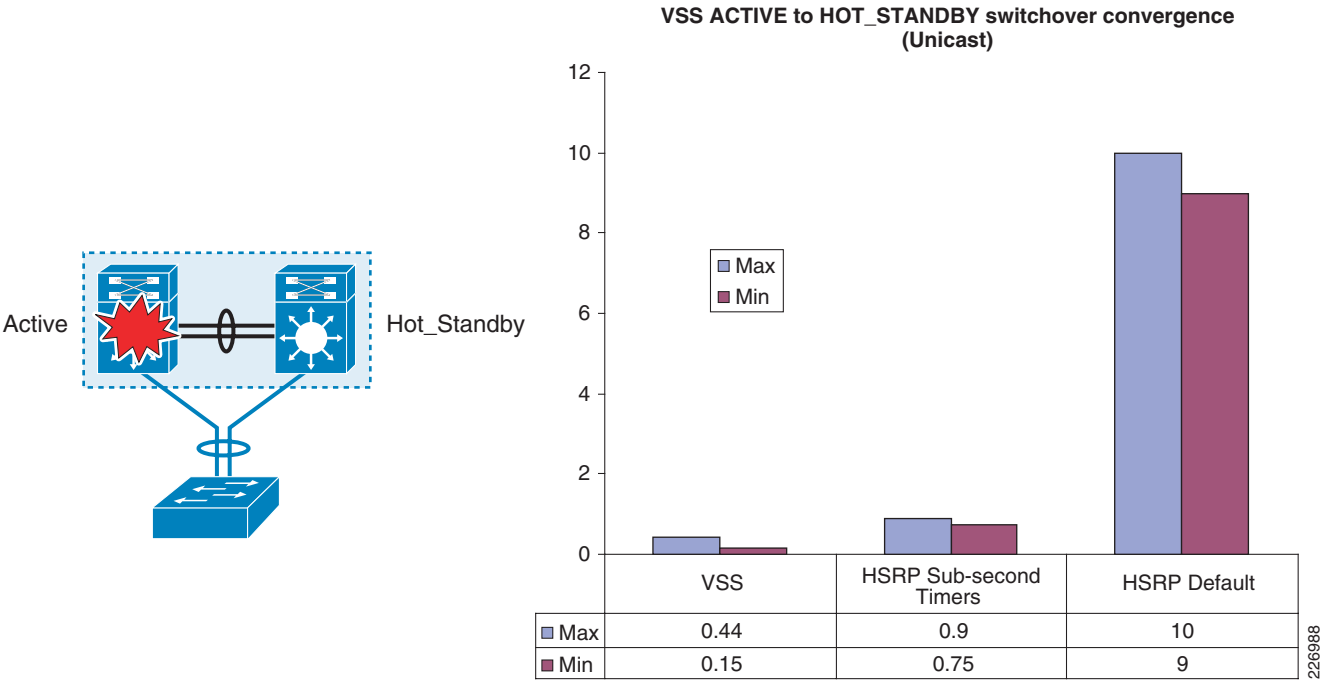
The outlier scope of Layer-2 network span largely remains a design policy exercise and has no single correct size for all enterprise campus networks. The typical recommendation is to use spanning capabilities with the VSS for VLANs assigned for functional use, such as network management, guest access, 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

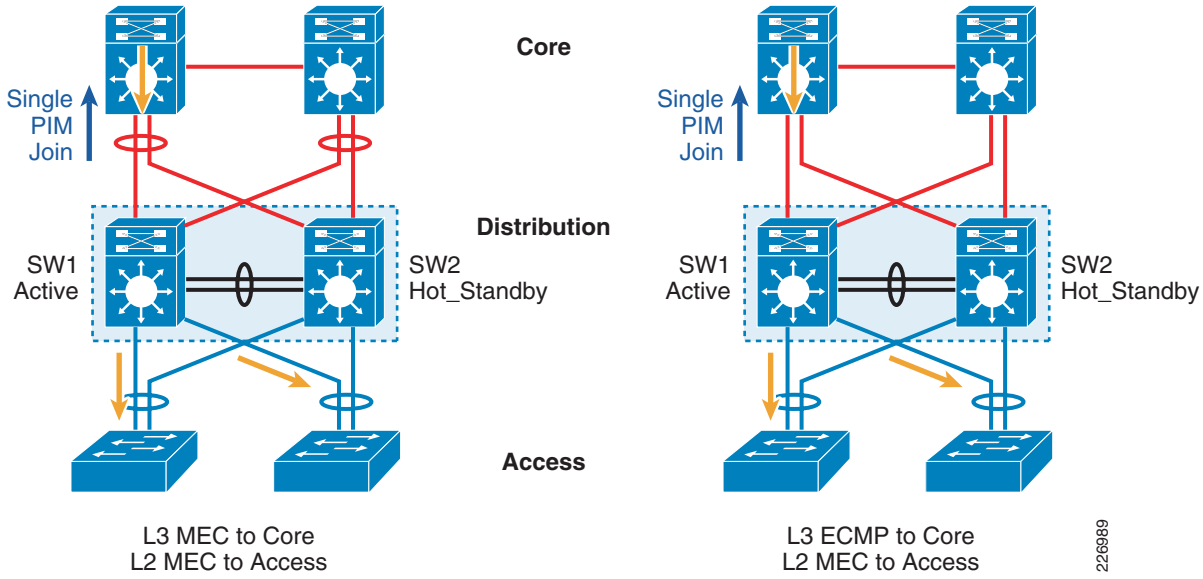
VSS shares all the benefits and restrictions of standalone MMLS technology. Forwarding states are programmed in hardware on both the active and standby supervisor. During a switchover, the hardware continues forwarding multicast data while the control plane recovers and reestablishes PIM neighbor relations. 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

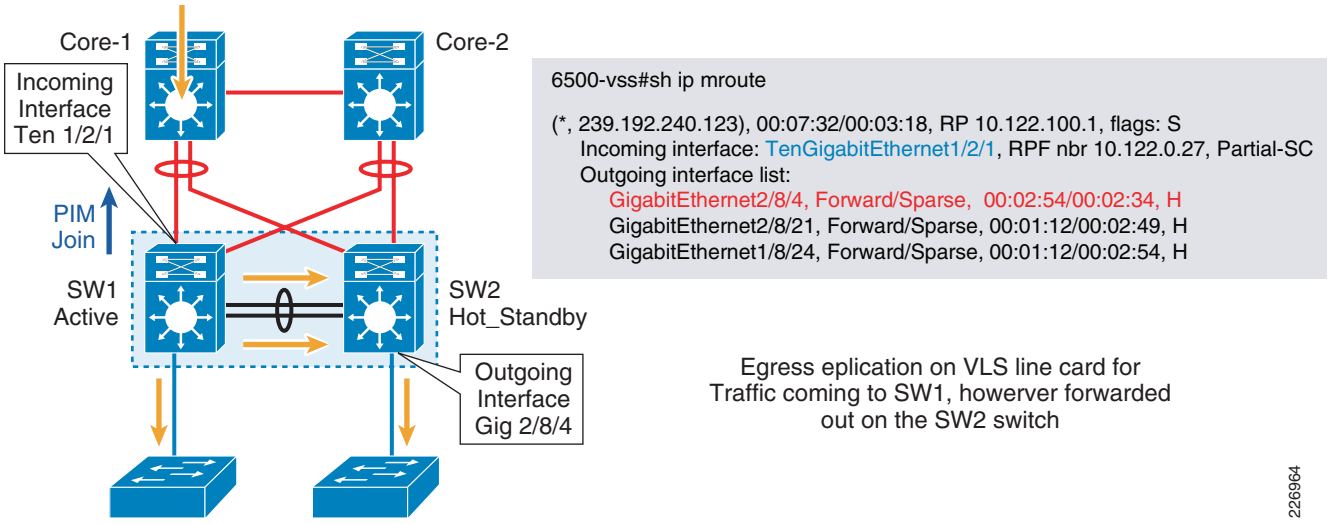


As with unicast data, the control plane for multicast is handled in the active switch. With VSS being a single logical switch, it sends a single PIM join. The PIM join is originated by the active switch, but it can be sent by an interface located on the hot-standby switch (in the case of ECMP). Normally, 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

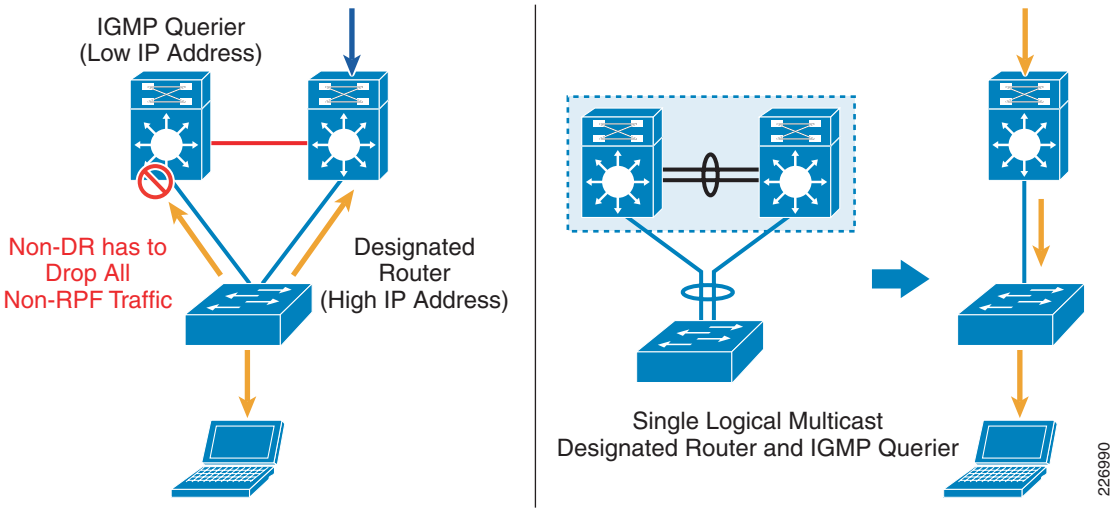
IP multicast uses one router to forward data onto a LAN in redundant topologies. If multiple routers have interfaces into a LAN, only one router will forward the data. There is no load balancing for multicast traffic on LANs. Usually, the designated router (DR)—highest IP address on a VLAN—is chosen 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

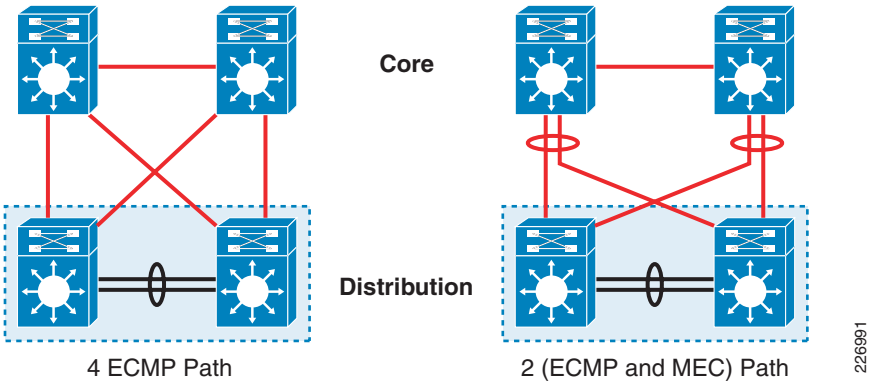
Routing Protocols, Topology, and Interaction

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.”

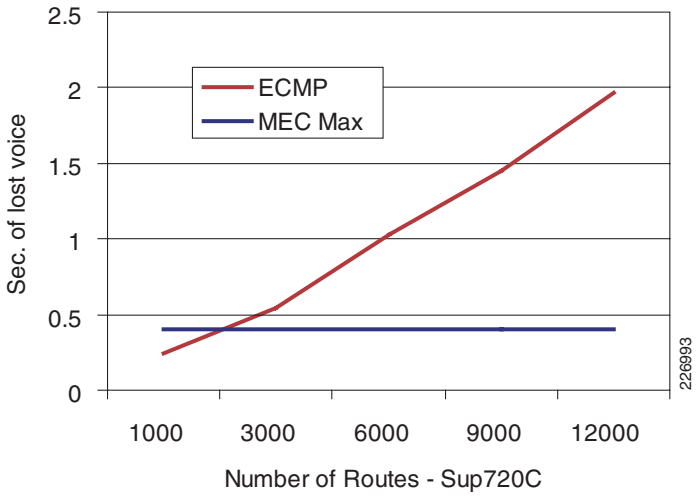
Link Failure Convergence

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

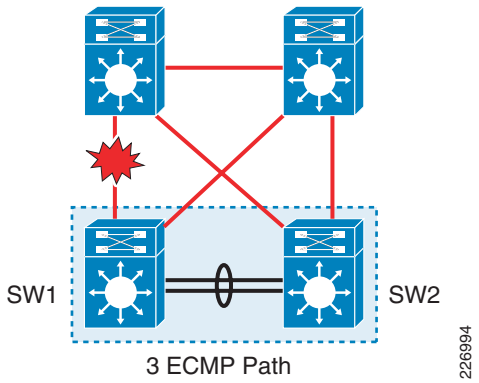
ECMP

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

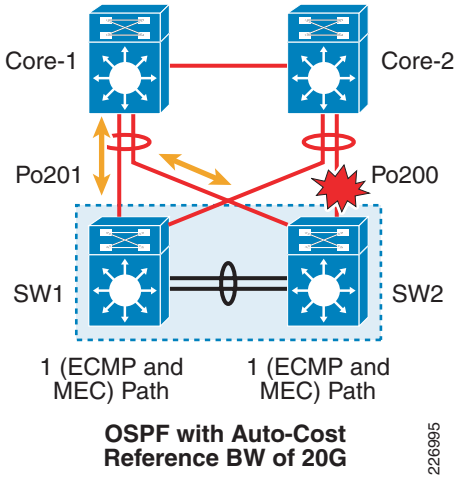


Figure 3-34 illustrates the behavior of metric change associated with OSPF and Layer-3 MEC. The failure of one link from a port-channel (Po200) causes the removal of routes learned from the entire port-channel. As a result, instead of two routes, only one route from one of the core router (core-1) is 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.

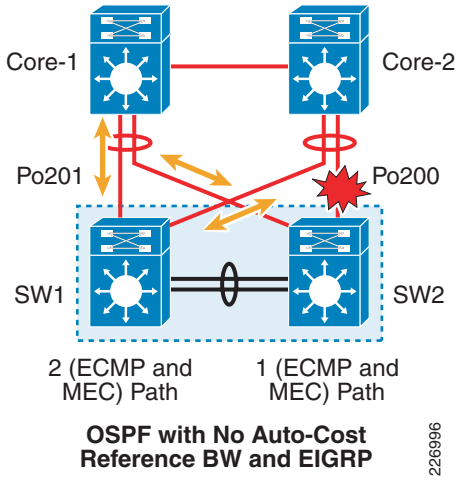
```
6500-VSS# sh ip route 10.121.0.0
Routing entry for 10.121.0.0/16
  Known via "ospf 100", distance 110, metric 13, type inter area
  Last update from 10.122.0.20 on Port-channel201, 00:51:31 ago
  Routing Descriptor Blocks:
  * 10.122.0.20, from 30.30.30.30, 00:51:31 ago, via Port-channel201
    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

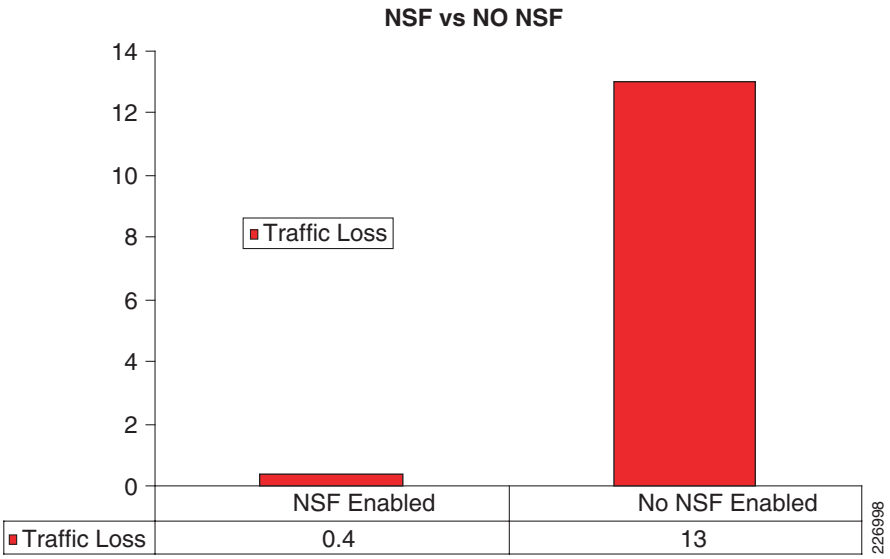
Table 3-4 Comparison of ECMP and Layer-3 MEC Options

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.**

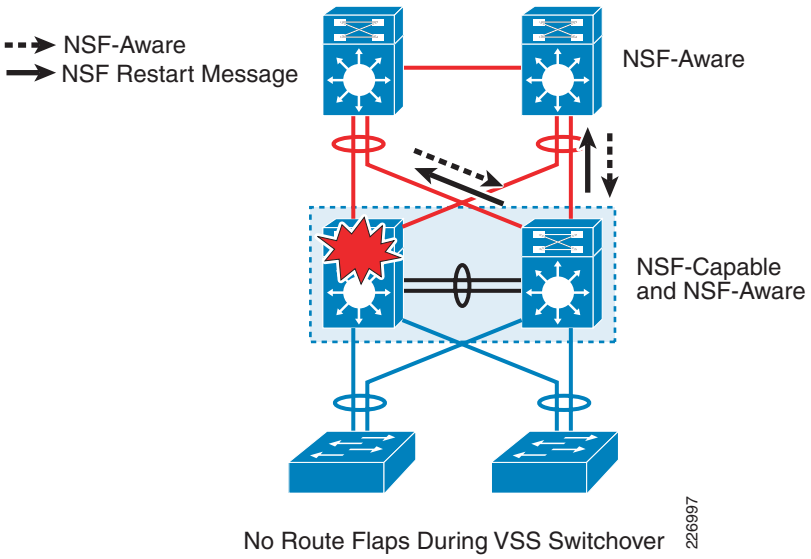
NSF Requirements and Recovery

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

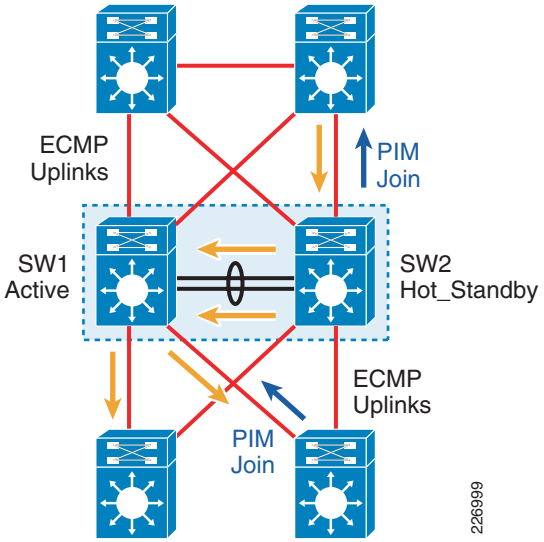
The VSS supports variety of multicast features and related design options, full validation and guidance encompassing all multicast features is beyond the scope of this design guide. This design guide covers the critical design points that can affect multicast traffic within the VSS-enabled campus. There 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

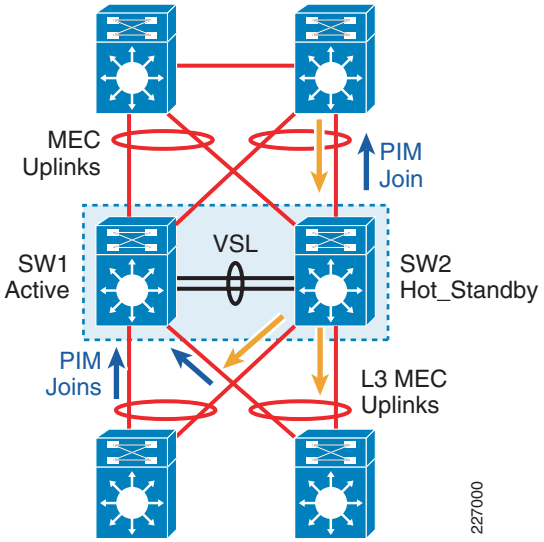
VSS represents a single multicast router. PIM joins are sent based on the highest PIM neighbor IP address (usually a first entry in the routing table for the given source) of the available ECMP paths in the routing table. Because the PIM join are send and received on different switches, the IIL (incoming 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.

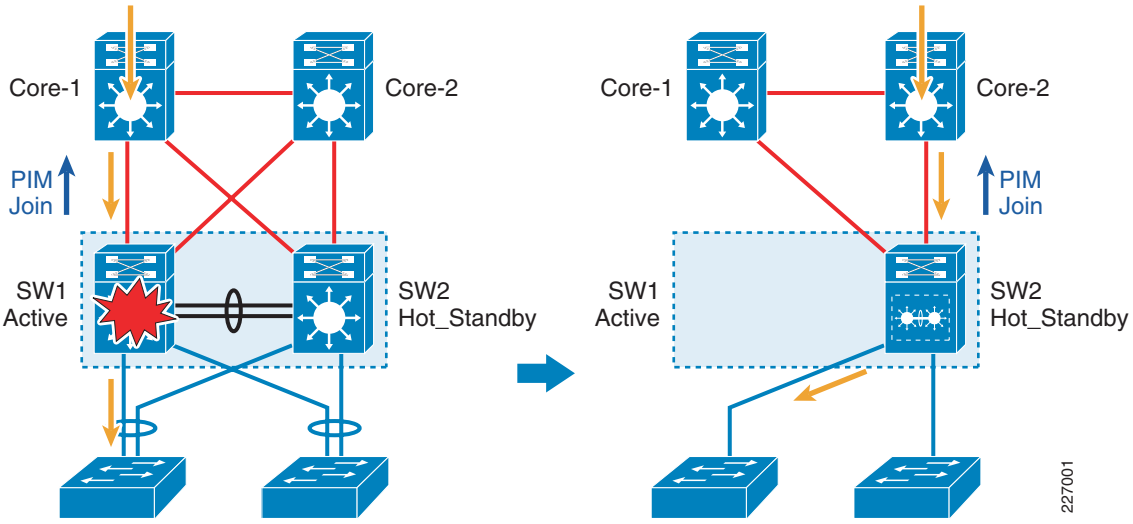
Tip **Cisco recommends using a Layer-3, MEC-based topology to prevent multicast traffic replication over the VSL bundle and avoids delay associated with reroute of traffic over VSL link.**

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

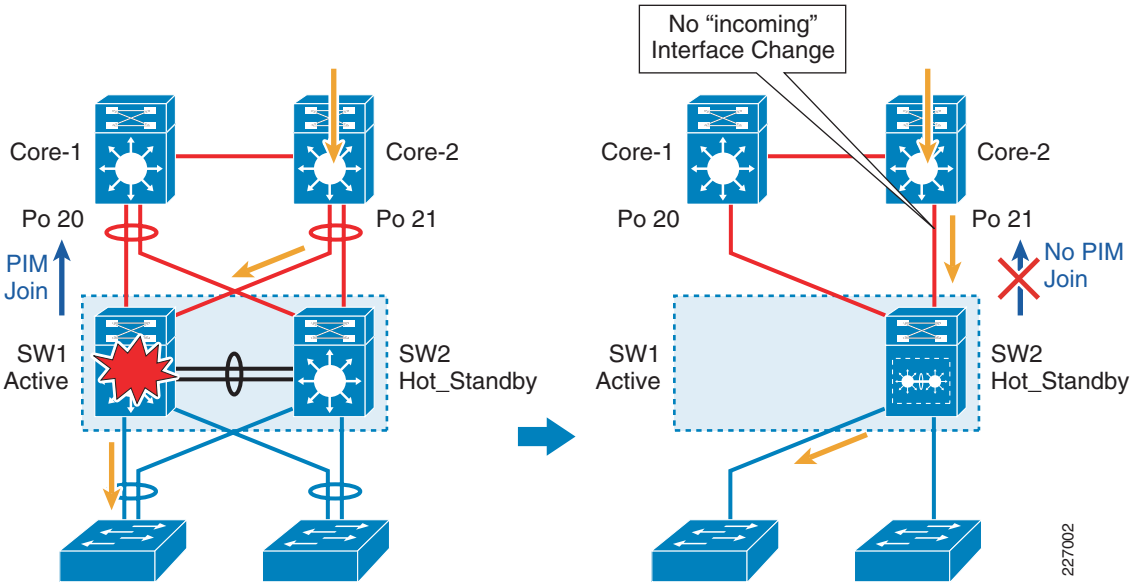
convergence related to this type of failure varies and can range from 2-to-3 minutes or higher depending on many factors such as rendezvous point (RP) reverse path forwarding) (RPF) check, unicast routing protocol convergence, etc.

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

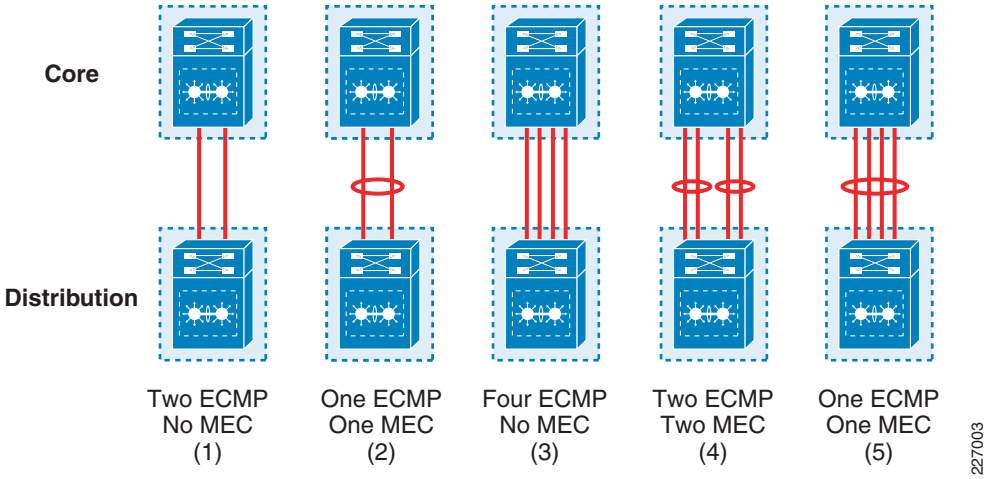
member connected to SW2. The core-2 continues to send multicast flow after the rehash. When SW2 receives the multicast flow, it forwards the traffic in hardware-based MMLS synchronization of multicast entries from old-active.

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.

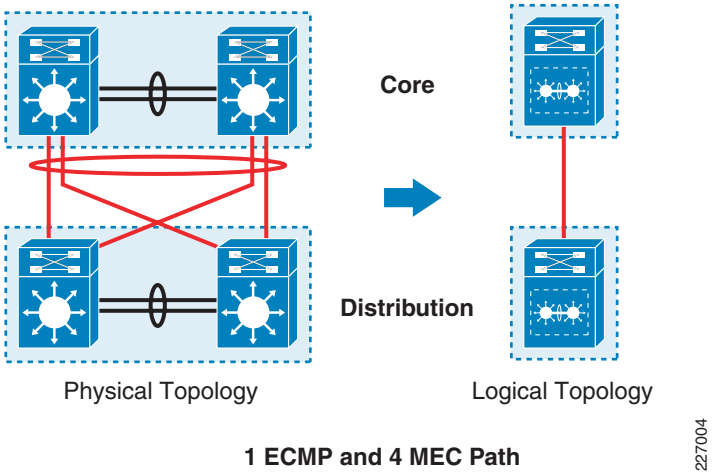
Table 3-6 Design Factors for Topology Options

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:

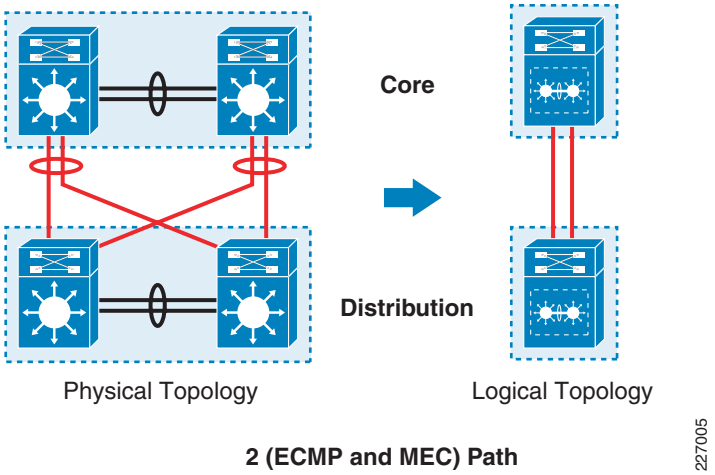
- Design inherently reduces routing control plane overhead in terms of routing topology and neighbor maintenance.
- Link failure or node failure is not dependent on routing table size, so that recovery is consistent.
- Link failure does not force traffic over the VSL bundle, reducing latency and congestion 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

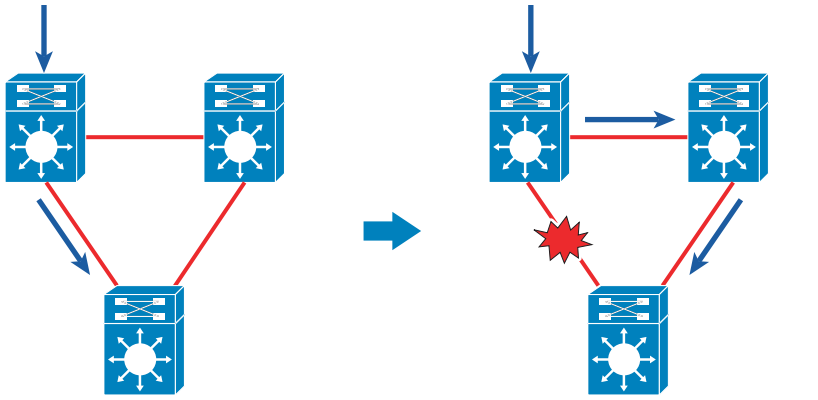
Routed Access

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.

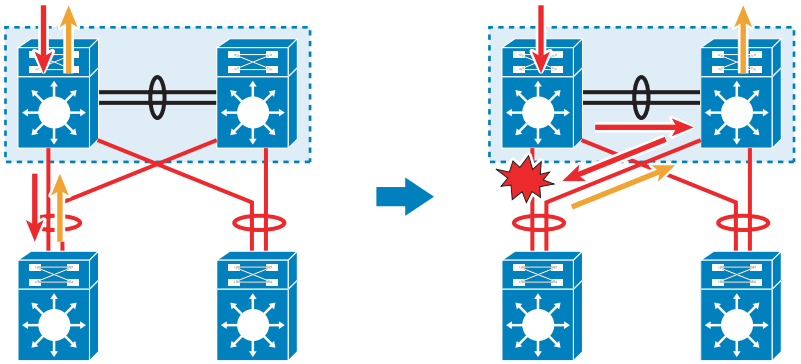
Detail of related failures and configuration guidance are available in the *Routed Access Design Guide* listed in Appendix B, "References."

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

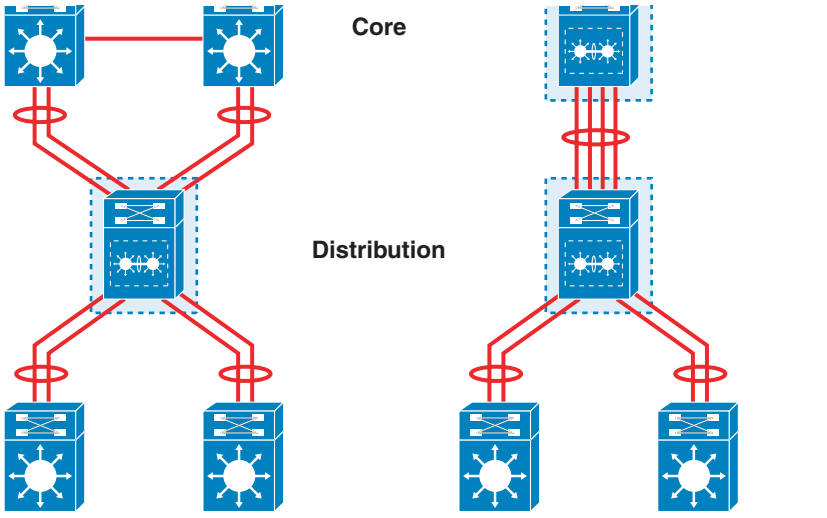


OSPF/EIGRP Downstream Recovery without Route-Change

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



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.