CHAPTER 4

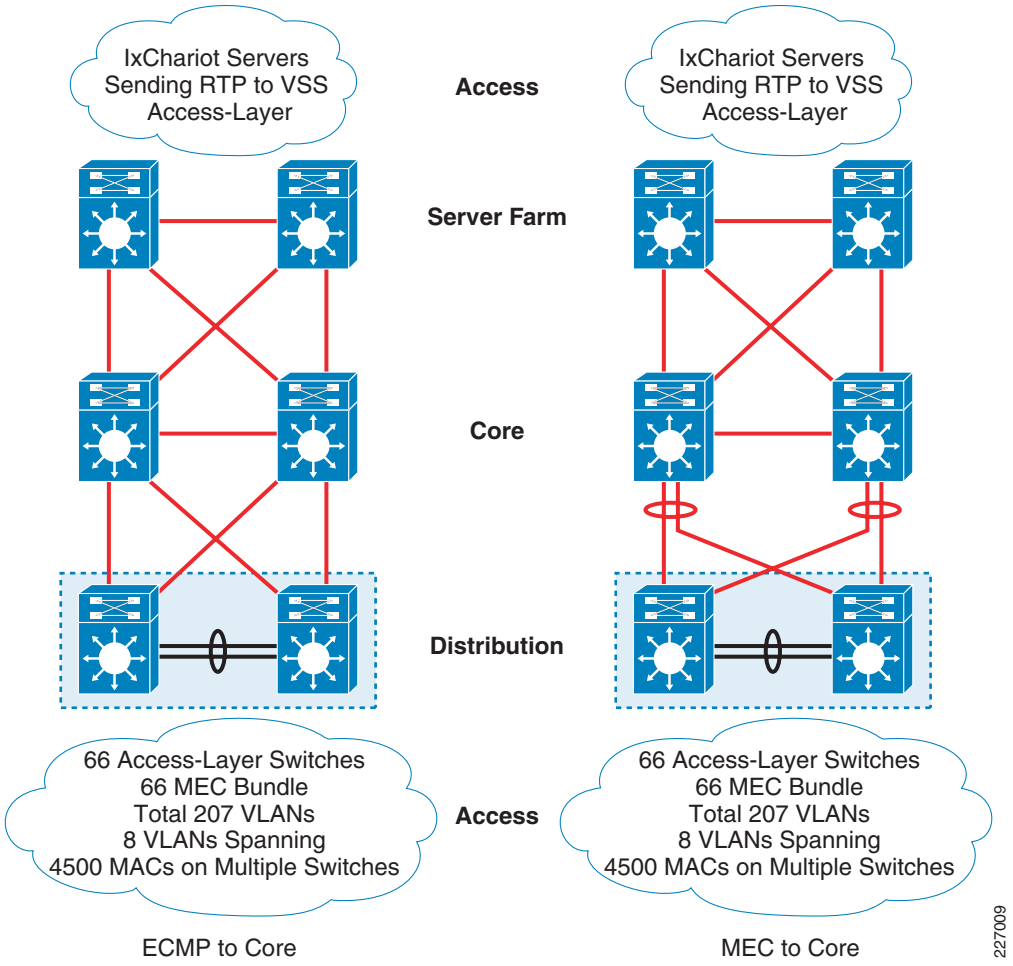
Convergence

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

Solution Topology

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

Figure 4-1 VSS Solution Topology



Software and Hardware Versions

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

Table 4-1 VSS Software and Hardware Summary

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

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

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

VSS-Enabled Campus Best Practices Solution Environment

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

Table 4-2 VSS Environment

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

Table 4-3 Layer-3 Domain

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

Table 4-3 Layer-3 Domain (continued)

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

Table 4-4 Layer-2 Domain

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

Convergence and Traffic Recovery

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

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

VSS Specific Convergence

Active Switch Failover

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

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

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

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

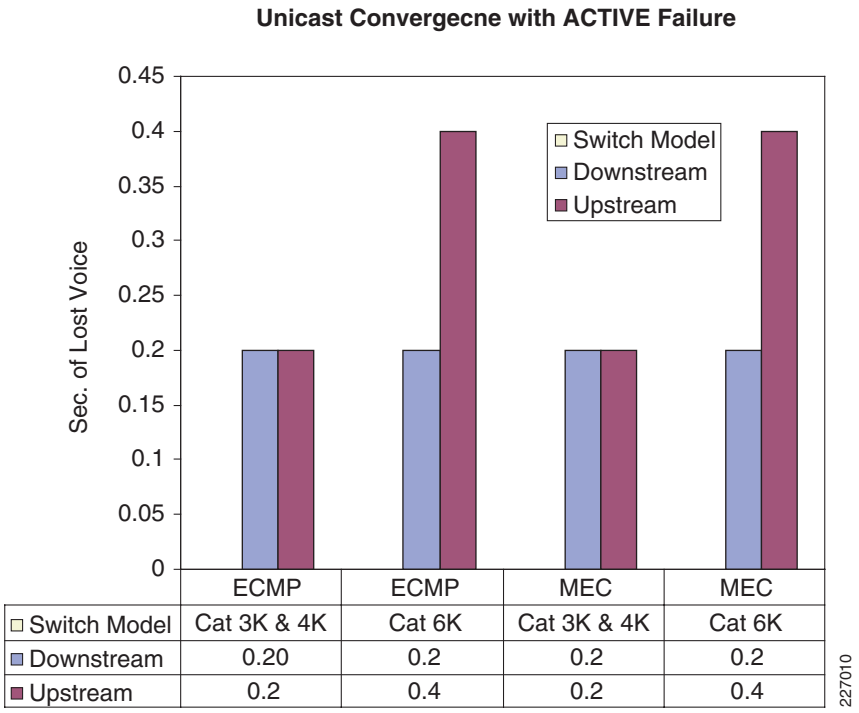
The active failure convergence is validated with Enhanced IGRP and OSPF routing protocol with the topology combination of ECMP or MEC connectivity to the core. The recovery methods for both routing protocol remains the same as summarized in Table 4-5.

Table 4-5 Active Failure Recovery

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

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

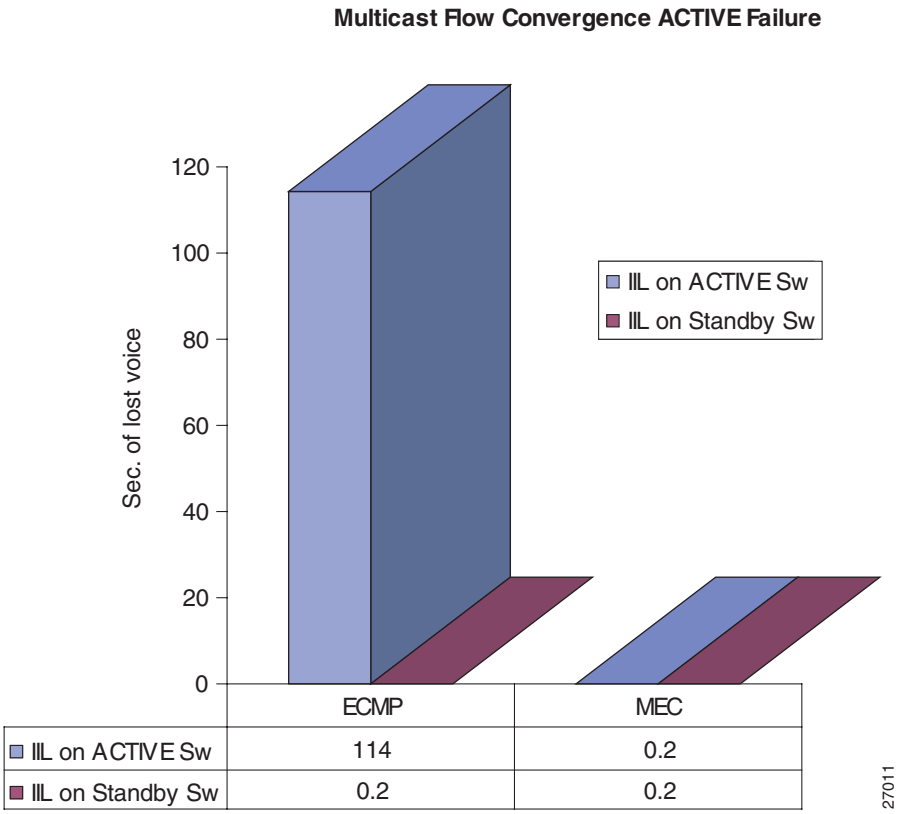
Figure 4-2 Active Failure Convergence



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

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

Figure 4-3 Multicast Convergence



Hot-Standby Failover

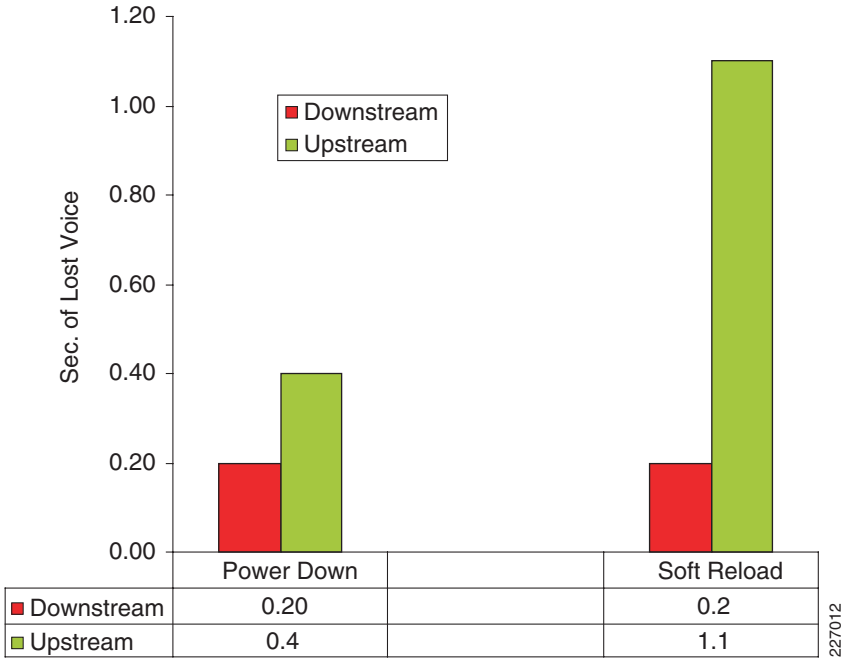
Hot-standby failover does not introduce control plane convergence because it is not actively responsible for managing various protocols and their update. However, in the ECMP topology, the neighbor connected via the hot-standby switch will reset and links connected to hot-standby goes down. The recovery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

Table 4-6 Hot-Standby Failure Recovery

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

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

Figure 4-4 Comparison of Hot-Standby Convergence Characteristics



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

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

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

Hot-Standby Restoration

Traffic recovery depends on two factors:

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

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

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

Table 4-7 Hot-Standby Switch Restoration Recovery

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

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

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

VSL Link Member Failure

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

Line Card Failure in the VSS

A line card failure can occur for the following reasons:

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

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

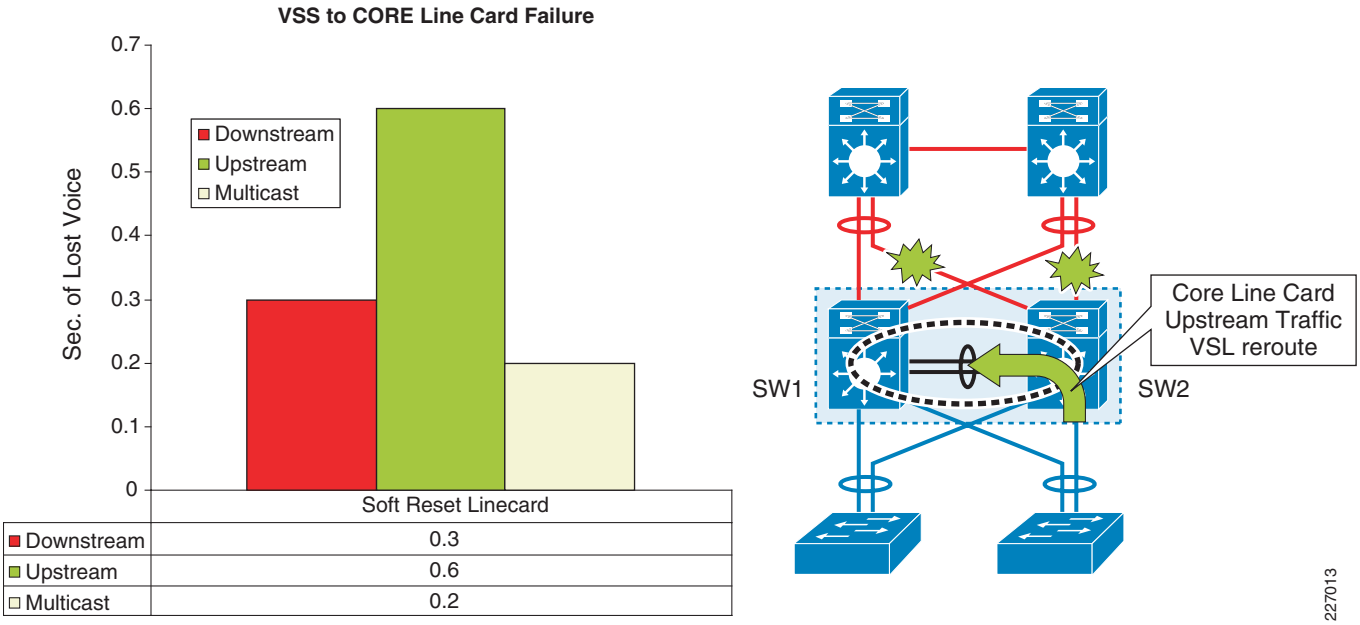
Line Card Connected to the Core-Layer

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

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

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

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



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

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

Line Card Connected to an Access Layer

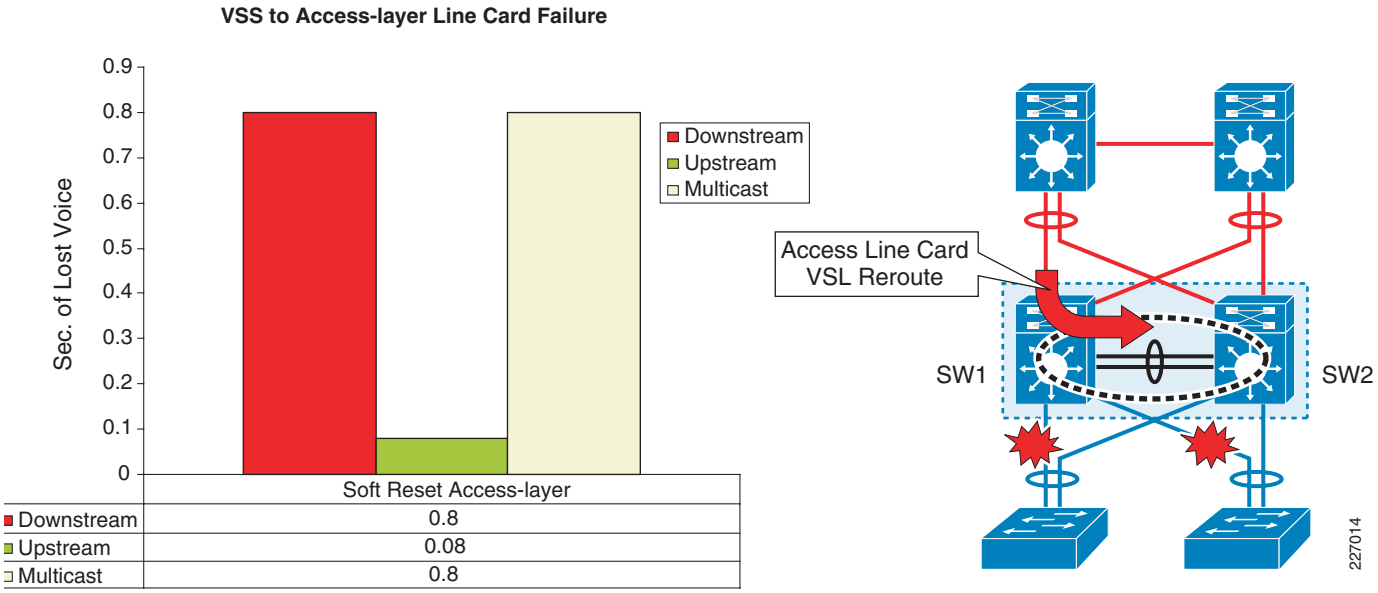
For the access-layer, line-card failures, the traffic flow recovery is the opposite of the core line-card failures. See Table 4-9.

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

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

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

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



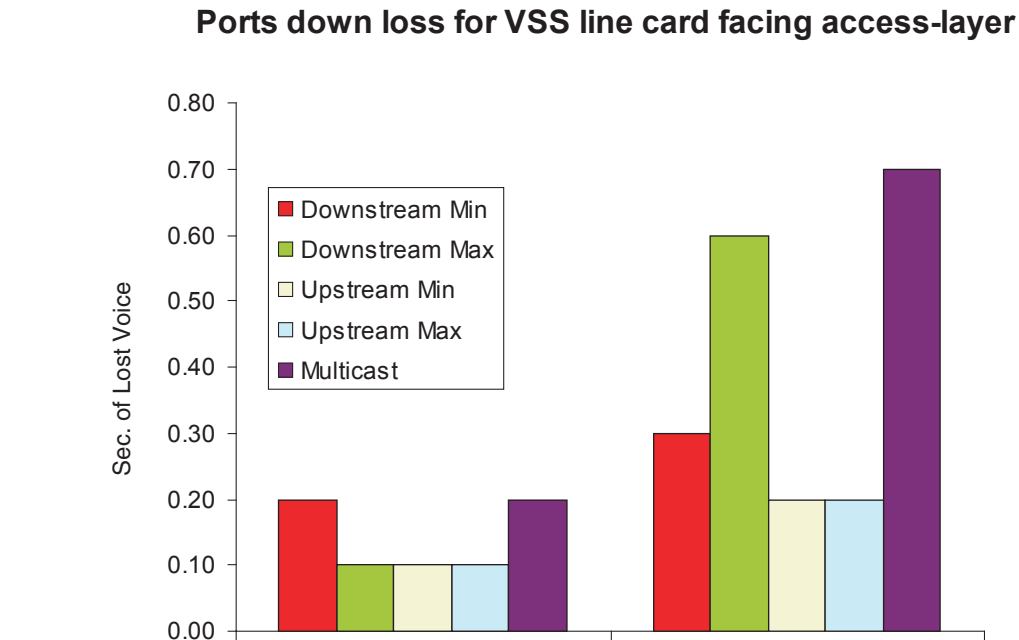
For the multicast traffic flowing downstream, the core device's hashing process results in the selection of the VSS member that will become the forwarder for that flow. For an access-layer line-card failure, the multicast traffic must be rerouted over the VSL link. The peer switch will then forward the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

Port Failures

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

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

Figure 4-7 Recovery Comparison for Port Failures



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

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

Routing (VSS to Core) Convergence

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

Core Router Failure with Enhanced IGRP and OSPF with MEC

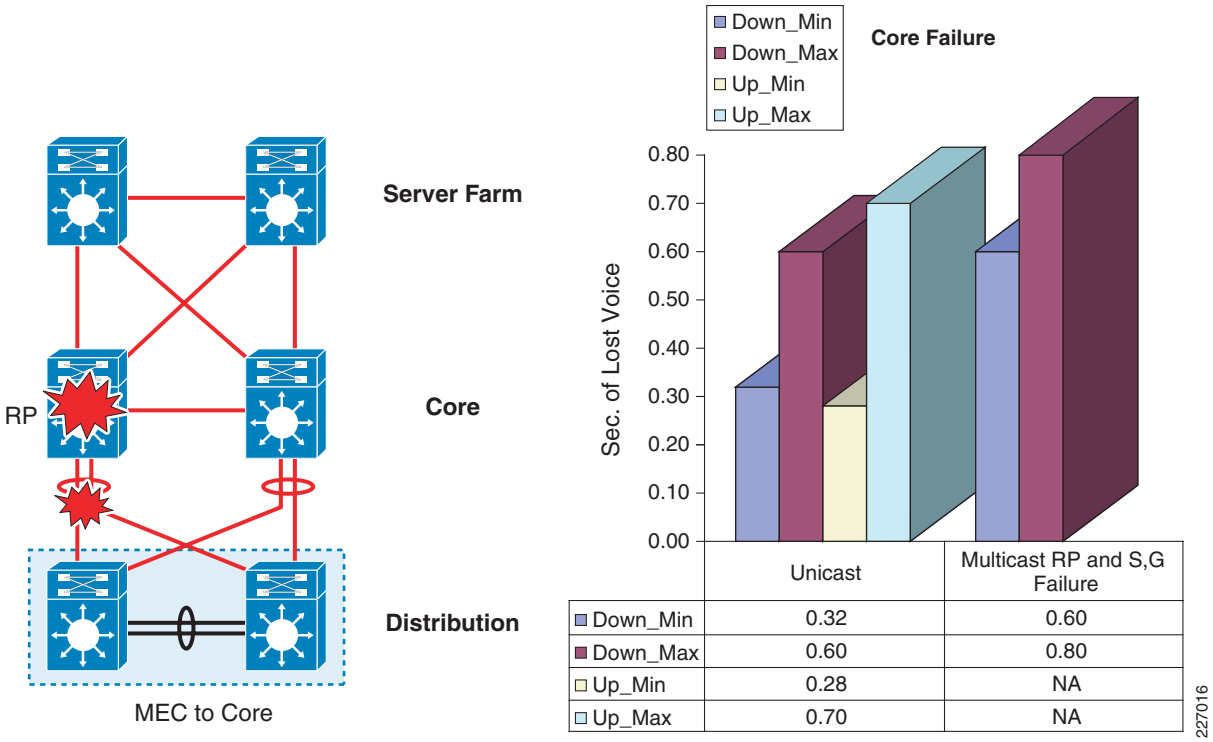
This design guide is validated with standalone core devices. In this design guide, the core is enabled to provide a rendezvous point (RP) for multicast traffic. The design choices surrounding the placement of RP depend on the multicast application and are not covered here. However, the failure of a RP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

Table 4-10 Core Router Failure Recovery Summary

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

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

Figure 4-8 Recovery Comparison for Core Router Failure



Link Failure Convergence

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

MEC Link Member Failure with OSPF

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

Table 4-11 MEC Link Member Failure OSPF Recovery Summary

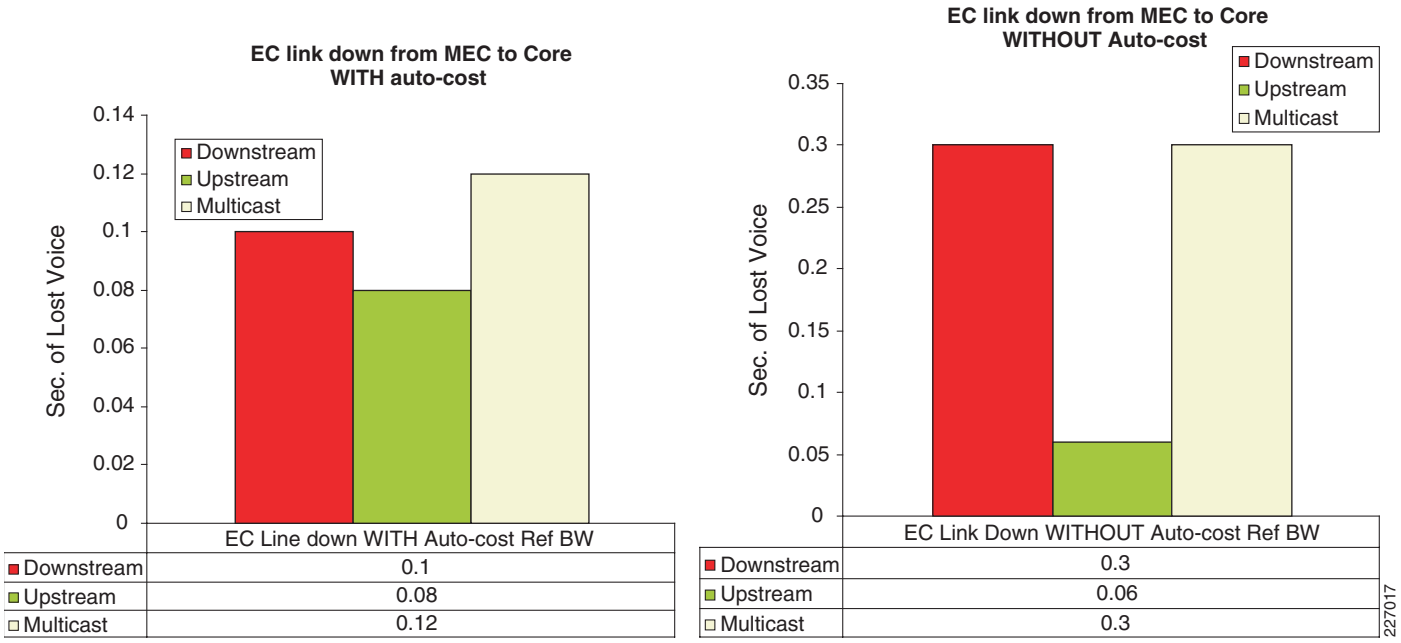
Topology	OSPF with Auto-Cost Ref BW 20G	OSPF without Auto-Cost Reference Bandwidth
Metric change	Yes	No
Resulting bandwidth	Two paths	Three paths
Unicast upstream	Graceful route withdrawal	Local hardware CEF path
Unicast downstream	Graceful route withdrawal	EC recovery at the core
Non-summarized vs.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

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

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

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

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



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

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

MEC Link Member Failure with Enhanced IGRP

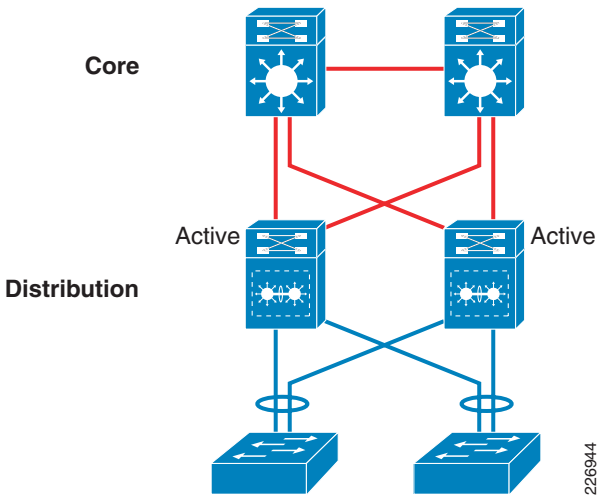
The Enhanced IGRP metric calculation is a composite of the total delay and the minimum bandwidth. When a member link is failed, EIGRP will recognize and use changed bandwidth value but delay will not change. This may or may not influence the composite metric since minimum bandwidth in the path is used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

Campus Recovery with VSS Dual-Active Supervisors

Dual-Active Condition

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

Figure 4-10 Dual Active Topology



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

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

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

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

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

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

Impact of Dual-Active on a Network without Detection Techniques

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

- Control plane disruption
- User data traffic disruption

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

Impact on Layer-2 MEC

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

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

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

Cisco Catalyst 65xx with CATOS:

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

Refer to following URL for details about EtherChannel inconsistencies:

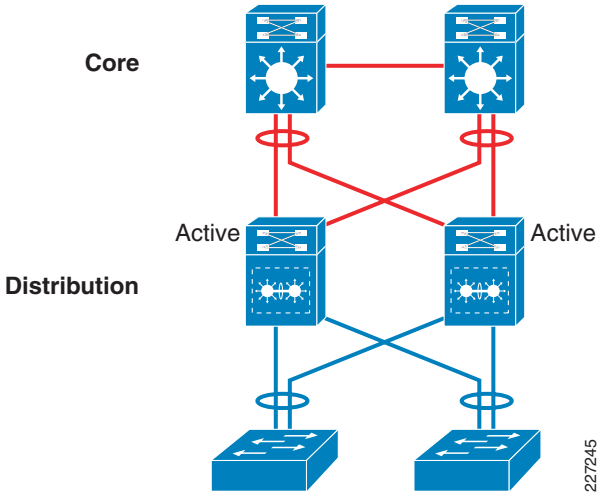
http://www.cisco.com/en/US/tech/tk389/tk213/technologies_tech_note09186a008009448d.shtml

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

Layer-3 MEC with Enhanced IGRP and OSPF

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

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



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

Enhanced IGRP

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

OSPF

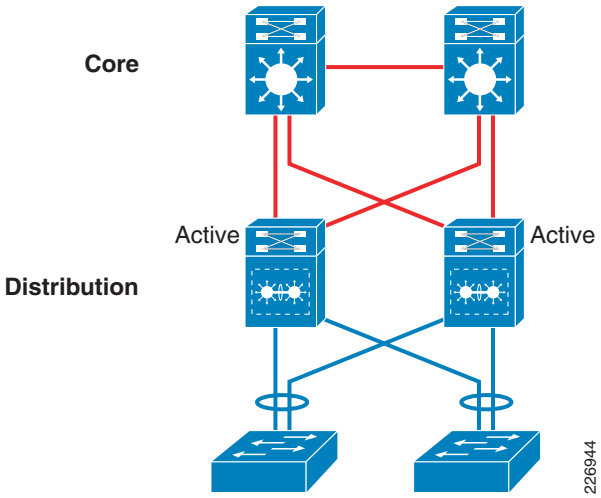
When a dual-active event occurs in an OSPF-based network, the adjacency never stabilizes with either of the active routers. For OSPF to form an adjacency, the protocol requires the bidirectional verification of neighbor availability. With dual-active state, the OSPF neighbor sees multiple messages hashed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

For either routing protocol, adjacency resets or instability leads to the withdrawal of routes and disruption of user traffic. In addition, Layer-2 at the access-layer will be error-disabled as discussed in the "Impact on Layer-2 MEC" section on page 4-19.

Layer-3 ECMP with Enhanced IGRP and OSPF

As illustrated in Figure 4-12, for a normally operational topology, the core router only sees one logical router and two paths for given destination; however, the VSS views four paths for a given destination (two from each core router). With a dual-active condition, the core might see more than one router depending on routing protocol. There is no hashing-related impact on neighbor adjacency and routing update with ECMP (being an independent path) such as the case with the Layer-3 MEC topology.

Figure 4-12 Layer-3 ECMP with Enhanced IGRP and OSPF Topology



Enhanced IGRP

During a dual-active condition, routers do not lose adjacency. Enhanced IGRP does not have any conflicting router IDs (unless Enhanced IGRP is used as redistribution point for some subnets) and each link is routed so that no adjacency change occurs at core routers or from active VSS members. User traffic continues forwarding with virtually no effect on user data traffic. Thus, dual-active condition may not impact Layer-3 connectivity in this topology; however, Layer-2 MEC may get error-disabled causing the disruption of user data traffic.

OSPF

During a dual-active event, two routers with the same IP loopback address announce the duplicate router IDs. Both active routers will announce the same LSA, which results in a LSA-flooding war for the access-layer subnets at the core routers.

Detection Methods

This section goes over various detection techniques and their operation and recovery steps. Following methods are available to detect the dual-active condition:

- Enhanced PAgP
- Fast-Hello—VSLP framework-based hello
- Bidirectional Forwarding Detection (BFD)

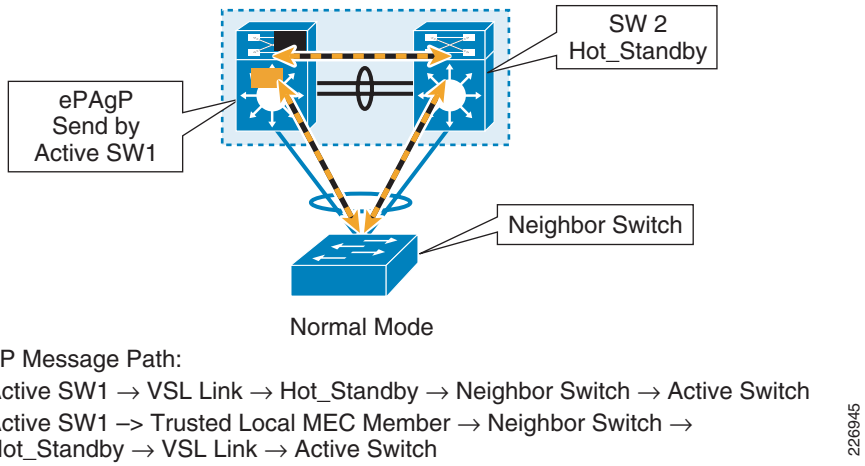
Note The enhanced PAgP and BFD is supported from Cisco IOS Release 12.2(33)SXH, while fast-hello requires Cisco IOS Release 12.2(33)SXI.

Enhanced PAgP

Normal Operation

Enhanced PAgP is an extension of the PAgP protocol. Enhanced PAgP introduces a new Type Length Value (TLV). The TLV of ePAgP message contains the MAC address (derived from the back-plane of an active switch) of an active switch as an ID for dual-active detection. Only the active switch originates enhanced PAgP messages in a normal operational mode. The active switch sends enhanced PAgP messages once every 30 seconds on both MEC link members. The ePAgP detection uses neighbor switches as tertiary connection to detect the dual-active condition. (All detection techniques require a tertiary connection from which the switches can derive the state of a VSL link because neither side can assume that the other is down by simply detecting that the VSL link is non-operational.) The ePAgP messages containing the active switch ID are sent by an active switch on the locally attached MEC-link member as well as over the VSL link. This operation is depicted in Figure 4-13 via black and yellow squares describing ePAgP messages and paths it traverses via neighbor switch. The neighbor switch simply reflects both of these messages via each uplink. This ensures that active the switch independently verifies the bidirectional integrity of the VSL links. For the neighbor switch to assist in this operation, it requires a Cisco IOS software version supporting enhanced PAgP. See Figure 4-13.

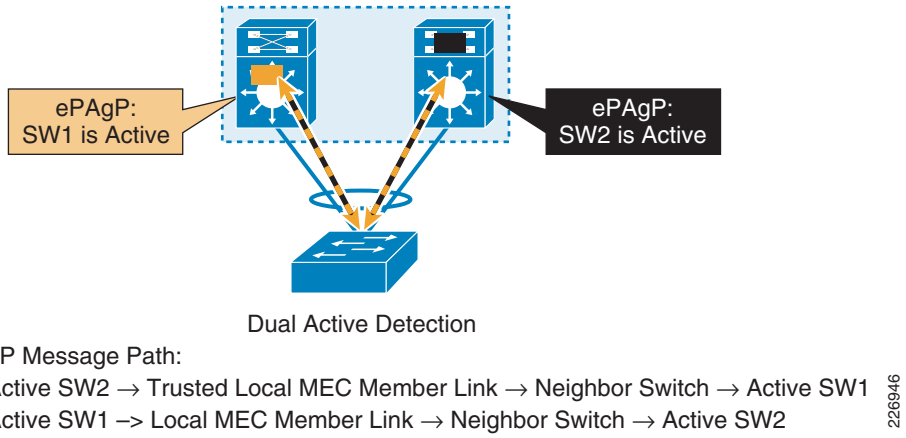
Figure 4-13 Enhanced PAgP Normal Operation



Dual-Active Detection with Enhanced PAgP

With dual-active all the links in a VSL bundle become non-operational, the hot-standby switch (SW2 in Figure 4-14) transitions to active (not knowing the status of remote switch). As SW2 becomes active, it generate its own enhanced PAgP message with its own active switch ID, sending it via the locally-attached MEC link member to SW1 via the neighbor switch. With the VSL link down, the old-active switch (SW1) stops receiving its own enhanced PAgP message and also receives an enhanced PAgP message generated via the remote switch (old hot-standby). These two messages and their paths is shown via yellow and black squares in Figure 4-14. SW1 remembers that it was an active switch and only the previously active SW1 undergoes the detection and recovery from the dual-active condition.

Figure 4-14 Dual-Active Detection with Enhanced PAgP Operation



Once an ePAgP message from SW2 is received by the old-active switch (SW1), SW1 compares its own active switch ID (MAC address derived from local backplane) with new active switch ID. If the received and expected IDs are different, the old-active chassis determines that a dual-active condition is triggered in the VS domain and starts the recovery process. The dual-active condition is displayed by following the CLI that is executed only on the old-active switch because that is where detection is activated.

```
6500-VSS# sh switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: Yes
```

```
No interfaces excluded from shutdown in
 recovery mode

In dual-active recovery mode: Yes
  Triggered by: PAgP detection
  Triggered on interface: Gi2/8/19
  Received id: 0019.a927.3000
  Expected id: 0019.a924.e800
```

Note In Cisco IOS Releases (12.2(33) SXH and 12.2(33) SXI), there is no way to differentiate between old-active versus newly-active switches. Both switches are active and both display the same command prompt. This can pose an operational dilemma when issuing the preceding command. In future releases, the old-active switch prompt may change to something meaningful so that the operator can distinguish between two active switches.

The dual-active condition generates different type of syslogs messages in different switches. The old-active switch (SW1) displays the following syslogs messages:

```
%PAGP_DUAL_ACTIVE-SW2_SP-1-RECOVERY: PAgP running on Gi2/8/19 triggered dual-active
recovery: active id 0019.a927.3000 received, expected 0019.a924.e800
%DUAL_ACTIVE-SW2_SP-1-DETECTION: Dual-active condition detected: all non-VSL and
non-excluded interfaces have been shut down
```

The newly-active switch (SW2) displays the following syslog messages:

```
%VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
%VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/5: Link down
%VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in ACTIVE role
```

The neighbor switch supporting enhanced PAgP protocol also displays the dual-active triggered message:

```
%PAGP_DUAL_ACTIVE-SP-3-RECOVERY_TRIGGER: PAgP running on Gi6/1 informing virtual switches
of dual-active: new active id 0019.a927.3000, old id 0019.a924.e800
```

Caution The neighbor switch also displays this syslog message during normal switchover because it does not know what really happened—it merely detects different switches claiming to be active.

Enhanced PAgP Support

As described in the preceding section, the neighbor switch must understand enhanced PAgP protocol in order to support dual-active detection. That also means enhanced PAgP requires the PAgP protocol to be operational on in the MEC configuration. One cannot disable PAgP and have enhanced PAgP running. Enhanced PAgP can be enabled either on Layer-2 or Layer-3 PAgP MEC members. This means you can run enhanced PAgP between the VSS and the core routers. See Table 4-12.

Table 4-12 Cisco IOS Version Support for Enhanced PAgP

Platform	Software	Comments
Cisco Catalyst 6500	12.2(33)SXH	Sup720 and Sup32
Cisco Catalyst 45xx and Cisco Catalyst 49xx	12.2(44)SG	

Table 4-12 Cisco IOS Version Support for Enhanced PAgP (continued)

Platform	Software	Comments
Cisco Catalyst 29xx, Cisco Catalyst 35xx and Cisco Catalyst 37xx	12.2(46)SE	Cisco Catalyst 37xx stack no support, see the text that follows.
Cisco Catalyst 37xx Stack	Not Supported	Cross-stack EtherChannel only supports LACP

PAgP is supported in all platforms except the Cisco Catalyst 37xx stack configuration in which cross-stack EtherChannel (LACP) is required to have MEC-connectivity with the VSS. Because cross-stack EtherChannel does not support PAgP, it cannot use enhanced PAgP for dual-active detection. A common approach to resolve this gap in support is to use two EtherChannel links from the same stack member. This solution is a non-optimal design choice because it creates a single point-of-failure. If a stack member containing that EtherChannel fails, the whole connectivity from the stack fails. To resolve this single point-of-failure problem, you can put two dual-link EtherChannel group, each on separate stack member connected to VSS; however, it will create a looped topology. The loop-free topology requires a single EtherChannel bundle to be diversified over multiple members which in turn requires LACP.

There are two solutions to the stack-only access-layer requirement:

- Use Fast Hello or BFD as the dual-active detection method (described in the "Fast-Hello (VSLP Framework-Based Detection)" section on page 4-26 or the "Bidirectional Forwarding Detection" section on page 4-30).
- Enhanced PAgP can be enabled either on Layer-2 or Layer-3 PAgP MEC members. This means you can run enhanced PAgP between the VSS and the core routers, although core routers require enhanced PAgP support and implementation of the Layer-3 MEC topology to the VSS.

Enhanced PAgP Configuration and Monitoring

Enhanced PAgP dual-active detection is enabled by default, but specific MEC groups must be specified as trustworthy. The specific CLI identifying MEC group as a trusted member is required under virtual switch configuration. The reason behind not enabling trust on all PAgP neighbors is to avoid unwanted enhanced PAgP members, such as an unprotected switch, unintended vendor connectivity, and so on.

The following conditions are required to enable the enhanced PAgP on EtherChannel:

- MEC must be placed in the administratively disabled state while adding or removing trust; otherwise, an error message will be displayed.
- PAgP protocol must be running on the MEC member. PAgP is recommended to be configured in desirable mode on both sides of the connection.

Fine tuning the PAgP hello-timer from its 30-second default value to one second using the **pagp rate fast** command does not help to improve convergence time for user traffic. This is because dual-active detection does *not* depend on how fast the PAgP packet is sent, but rather on how fast the hot-standby switch is able to generate the enhanced PAgP message with its own active ID to trigger dual-active detection. See Figure 4-15.

Figure 4-15 Enabling Trust on the MEC with PAgP Running

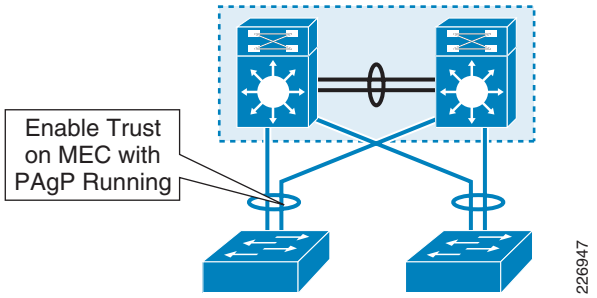


Figure 4-15 shows the trust configuration required for MEC member to be eligible for participating in enhanced PAgP-based dual-active detection. Use the following commands to enable trust on the MEC with PAgP running:

```
6500-VSS(config)# switch virtual domain 10
6500-VSS(config-vs-domain)# dual-active detection pagp trust channel-group 205
```

The enhanced PAgP support and trust configuration can be verified on the VSS switch as well as the enhanced PAgP neighbor the commands shown in the following configuration examples.

VSS switch:

```
6500-VSS# show switch virtual dual-active pagp
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

! << Snip >>

Channel group 205 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
          Dual-Active     Partner             Partner   Partner
Port      Detect Capable  Name                Port      Version
Gi1/8/19  Yes             cr7-6500-3          Gi5/1     1.1
Gi1/9/19  Yes             cr7-6500-3          Gi6/1     1.1
```

Neighbor switch that supports enhanced PAgP:

```
4507-Switch# show pagp dual-active
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

Channel group 4
          Dual-Active     Partner             Partner   Partner
Port      Detect Capable  Name                Port      Version
Te1/1     Yes             cr2-6500-VSS        Te2/2/6   1.1
Te2/1     Yes             cr2-6500-VSS        Te1/2/6   1.1
```

Fast-Hello (VSLP Framework-Based Detection)

Fast-hello is a newest dual-active detection method and is available with Cisco IOS 12.2(33) SXI and newer releases. The primary reasons to deploy fast-hello are as follows:

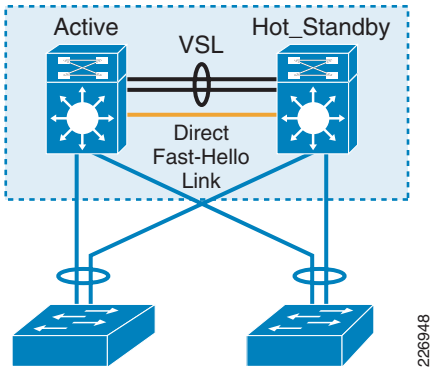
- Whenever enhanced PAgP deployment is not possible, such as in the case of server-access connectivity where servers are connected to the VSS and core connectivity is not Layer-3 MEC-based.
- If the installed-based has Cisco IOS versions that can not support enhanced PAgP.

- The EtherChannel group protocol is LACP.
- Simplicity of configuration is required and as fast-hello is being used as replacement to BFD (see the "Bidirectional Forwarding Detection" section on page 4-30).

Normal Operation

Fast-hello is a direct-connection, dual-active detection mechanism. It requires a dedicated physical port between two virtual-switch nodes in order to establish a session. Fast-hello is a connectionless protocol that does not use any type of handshaking mechanism to form a fast-hello adjacency. An incoming fast-hello message from a peer node with the appropriate TLV information establishes a fast-hello adjacency. See Figure 4-16.

Figure 4-16 Fast-hello Setup



Each dual-active fast-hello message carries the following information in TLVs:

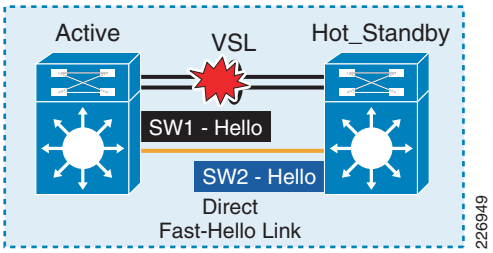
- *VSS Domain ID*—VSS virtual-switch node must carry a common domain ID in each hello message.
- *Switch ID*—Each virtual-switch node advertises the local virtual-switch ID for self-originated hello messages.
- *Switch Priority*—Each virtual-switch node advertises the local virtual-switch priority for self-originated hello messages..

By default, each virtual-switch node transmits fast-hello packets at two-second intervals. Dual-active fast-hello transmit timers are hard-coded and transparent to the end-user in the VSS system. The hard-coded fast-hello timer cannot be configured or tuned, and can only be verified using the **debug** commands. Each virtual-switch node transmits fast-hellos at the default interval to establish a session with its peer node. Any established dual-active fast-hello adjacency will be torn down if either of virtual-switch node fails to receive hellos from its peer node after transmitting five subsequent hello messages. By default, the hold-down timer is hard-coded to 10 seconds. If for any reason the adjacency establishment process fails, either due a problem or misconfiguration, the configured side continues to transmit hello messages at default interval to form a session. The dedicated link configured for fast-hello support is not capable of carrying control-plane and user-data traffic.

Dual-Active Detection with Fast-Hello

Either active or hot-standby switch cannot distinguish between the failures of remote peer or the VSS bundle. The SSO process in a active switch has to react on loss of communication to the hot-standby informing VSS control plane to remove all the interfaces and line cards associated with the hot-standby switch, including the remote port configured as for fast-hello. However, during dual-active, the link that is configured to carry fast-hello is operational (hot-standby is still operational) and exchanges hellos at the regular interval. As a result, the old-active switch notices this conflicting information about the fast-hello link and determines that this is only possible when the remote node is operational; otherwise, the old-active switch would not see the fast hello, implying dual-active has occurred. See Figure 4-17.

Figure 4-17 Dual Active Detection with Fast-hello



The previously active switch initiates the dual-active detection process when all the following conditions are met:

- Entire VSL EtherChannel is non-operational.
- Fast-hello links on each virtual-switch node are still operational.
- The previously active switch has transmitted at least one fast-hello at the regular two-second interval.

Upon losing the VSL EtherChannel, the old-active switch transmits a fast-hello at a faster rate (one per 500 msec) after transmitting at least one fast-hello at the regular interval. This design prevents taking away unnecessary CPU processing power during the active/hot-standby transitional network state. The following **show** command outputs illustrate the dual-active condition and the state of the detection on old-active.

```
6500-VSS# show switch virtual dual fast-hello
Fast-hello dual-active detection enabled: Yes

Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi1/5/1    Link up        Gi2/5/1      Link up

6500-VSS# show switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: Yes
Fast-hello dual-active detection enabled: Yes

No interfaces excluded from shutdown in recovery mode

In dual-active recovery mode: Yes
  Triggered by: Fast-hello detection
  Triggered on interface: Gi1/5/1
```

Syslog messages displayed on the on an old-active switch (SW1) when the dual-active state occurs:

```
Dec 31 22:35:58.492: %EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet1/5/4 left the
port-channel Port-channel1
Dec 31 22:35:58.516: %LINK-SW1_SP-5-CHANGED: Interface TenGigabitEthernet1/5/4, changed
state to down
Dec 31 22:35:58.520: %LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/4, changed state to down
Dec 31 22:35:58.536: %VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
Dec 31 22:35:58.540: %VSLP-SW1_SP-2-VSL_DOWN:   Last VSL interface Te1/5/4 went down
Dec 31 22:35:58.544: %LINEPROTO-SW2_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/1, changed state to down

Dec 31 22:35:58.544: %VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in
ACTIVE role

! << snip >>

Dec 31 22:35:59.652: %DUAL_ACTIVE-SW1_SP-1-DETECTION: Dual-active condition detected: all
non-VSL and non-excluded interfaces have been shut down ! <- Fast-hello triggers recovery
! process and starts recovery process the old active switch.
Dec 31 22:35:59.652: %DUAL_ACTIVE-SW1_SP-1-RECOVERY: Fast-hello running on Gi1/5/1
triggered dual-active recovery
! << snip >>
Dec 31 22:36:09.583: %VSDA-SW1_SP-3-LINK_DOWN: Interface Gi1/5/1 is no longer dual-active
detection capable
```

Syslogs messages on newly-active switch (SW2) when a dual-active state occurs:

```
Dec 31 22:35:58.521: %PFREDUN-SW2_SPSTBY-6-ACTIVE: Initializing as Virtual Switch ACTIVE
processor  Ð Starting NSF Recovery process
! << snip >>

Dec 31 22:36:09.259: %VSDA-SW2_SP-3-LINK_DOWN: Interface Gi2/5/1 is no longer dual-active
detection capable Ð Dual ACTIVE fast-hello link goes down and declares no longer
dual-active detection capable
```

Fast-Hello Configuration and Monitoring

Fast-hello configuration is simple, first enable it globally under virtual switch domain and then define it under the dedicated Ethernet port as follows:

Enable under VSS global configuration mode:

```
6500-VSS(config)# switch virtual domain 1
6500-VSS(config-vs-domain)# dual-active detection fast-hello
```

Enable fast-hello at the interface level:

```
6500-VSS(config)# int gi1/5/1
6500-VSS(config-if)# dual-active fast-hello

WARNING: Interface GigabitEthernet1/5/1 placed in restricted config mode. All extraneous
configs removed!

6500-VSS(config-if)# int gi2/5/1
6500-VSS(config-if)# dual-active fast-hello

WARNING: Interface GigabitEthernet2/5/1 placed in restricted config mode. All extraneous
configs removed!

%VSDA-SW2_SPSTBY-5-LINK_UP: Interface Gi1/5/1 is now dual-active detection capable
%VSDA-SW1_SP-5-LINK_UP: Interface Gi2/5/1 is now dual-active detection capable
```

A link-enabled for fast-hello support carries only dual-active fast-hello messages. All default network protocols, such as STP, CDP, Dynamic Trunking Protocol (DTP), IP, and so on are automatically disabled and are not processed. Only a physical Ethernet port can support fast-hello configuration; any other ports, such as SVI or port-channel, cannot be used as fast-hello links. Multiple fast-hello links can be configured to enable redundancy. The Sup720-10G 1-Gigabit uplink ports can be used if the supervisor is not configured in *10-Gigabit-only* mode. The status of the ports enabled with the fast-hello configuration (active and hot-standby switch) can be known by using the following **show** command output example:

```
6500-VSS# show switch virtual dual-active fast-hello
Fast-hello dual-active detection enabled: Yes
Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi1/5/1    Link up        Gi2/5/1      Link up

6500-VSS# remote command standby-rp show switch virtual dual-active fast-hello
Fast-hello dual-active detection enabled: Yes

Fast-hello dual-active interfaces:
Port       Local State    Peer Port    Remote State
-----------------------------------------------------
Gi2/5/1    Link up        Gi1/5/1      Link up
```

Bidirectional Forwarding Detection

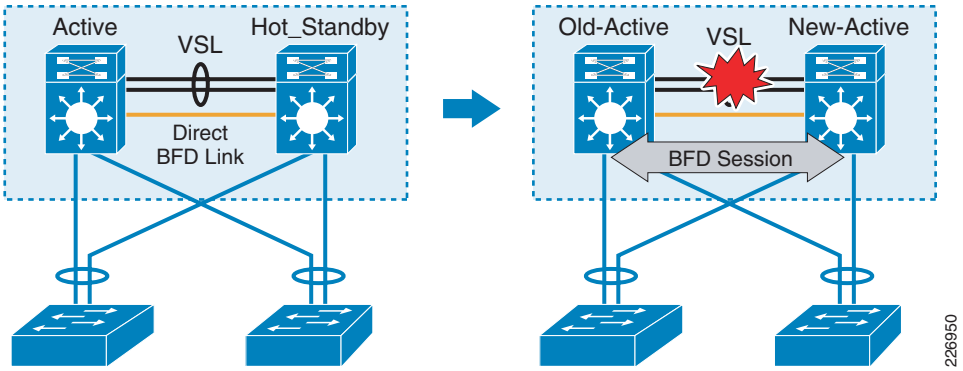
BFD is an alternative method to use when dual-active detection is not possible for the following reasons:

- Enhanced PAgP or fast-hello deployment are not possible due to Cisco IOS version limitations.
- The EtherChannel group protocol is LACP and the Cisco IOS version for fast-hello is not possible.
- Better convergence is required in a specific topology—For example, ECMP-based topologies enabled with OSPF.

Normal Operation

As with fast-hello detection, BFD detection depends on dedicated tertiary connectivity between the VSS member chassis. VSS uses BFD version 1 in echo mode for dual-active detection. Refer to cisco.com for common BFD information. BFD detection is a passive detection technique. When VSS is operating normally, the BFD configured interface remains up/up; however, no BFD session is active on that link. See Figure 4-18.

Figure 4-18 Bidirectional Forwarding Detection (BFD)



Dual-Active Detection with BFD

The BFD session establishment is the indication of dual-active condition. In a normal condition, VSS cannot establish a BFD session to itself as it is a single logical node. When a dual-active event occurs, the two chassis are physically separated, except for the dedicated BFD link that enables the BFD session between the chassis. A preconfigured connected static route establishes the BFD session. (A description of the BFD session-establishment mechanics are described in the "BFD Configuration and Monitoring" section on page 4-32.) BFD sessions last very briefly (less than one second), so you cannot directly monitor the BFD session activity. BFD session establishment or teardown logs cannot be displayed until the debug BFD command is enabled. However, the following syslogs will be displayed on the old active switch.

```
10:28:56.738: %LINK-SW1_SP-3-UPDOWN: Interface Port-channel1, changed state to down
10:28:56.742: %LINEPROTO-SW1_SP-5-UPDOWN: Line protocol on Interface
TenGigabitEthernet1/5/4, changed state to down
10:28:56.742: %VSLP-SW1_SP-3-VSLP_LMP_FAIL_REASON: Te1/5/4: Link down
10:28:56.742: %EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet2/5/4 left the
port-channel Port-channel2
10:28:56.750: %VSLP-SW1_SP-2-VSL_DOWN:   Last VSL interface Te1/5/4 went down
10:28:56.754: %VSLP-SW1_SP-2-VSL_DOWN:   All VSL links went down while switch is in ACTIVE
role
```

The **debug ip routing** command output reveals the removal of routes for the peer switch (hot-standby switch) as the VSL link becomes non-operational and the loss of connectivity for the remote interfaces (from the old-active switch viewpoint) is detected.

```
Jul 31 10:29:21.394: RT: interface GigabitEthernet1/5/1 removed from routing table
Jul 31 10:29:21.394: RT: Pruning routes for GigabitEthernet1/5/1 (1)
```

The following syslogs output shows BFD triggering the recovery process on the old active switch:

```
10:29:21.202: %DUAL_ACTIVE-SW1_SP-1-RECOVERY: BFD running on Gi1/5/1 triggered dual-active
recovery <- 1
10:29:21.230: %DUAL_ACTIVE-SW1_SP-1-DETECTION: Dual-active condition detected: all non-VSL
and non-excluded interfaces have been shut down
```

The following syslog output depicts new active during dual active. Notice the time stamp in bold associated with marker number 2 which is the time compared to the BFD trigger time in the old active switch (see the marker number 1 in the preceding output example).

```
10:28:56.738: %VSLP-SW2_SPSTBY-3-VSLP_LMP_FAIL_REASON: Te2/5/4: Link down
10:28:56.742: %VSLP-SW2_SPSTBY-2-VSL_DOWN:   Last VSL interface Te2/5/4 went down
10:28:56.742: %VSLP-SW2_SPSTBY-2-VSL_DOWN:   All VSL links went down while switch is in
Standby role
```

The following output illustrates the BFD triggering the recovery process on the newly active switch:

```
10:28:56.742: %DUAL_ACTIVE-SW2_SPSTBY-1-VSL_DOWN: VSL is down - switchover, or possible
dual-active situation has occurred <- 2
10:28:56.742: %VSL-SW2_SPSTBY-3-VSL_SCP_FAIL: SCP operation failed
10:28:56.742: %PFREDUN-SW2_SPSTBY-6-ACTIVE: Initializing as Virtual Switch ACTIVE
processor
```

The following output on newly active switch illustrates the installation of the connected route to establish the BFD session with the old active switch:

```
10:28:58.554: RT: interface GigabitEthernet2/5/1 added to routing table
10:29:21.317: RT: interface GigabitEthernet2/5/1 removed from routing table
10:29:21.317: RT: Pruning routes for GigabitEthernet2/5/1 (1)
```

The dual-active detection using BFD takes 22-to-25 seconds (see time stamps in preceding syslogs with markers 1 and 2). BFD takes longer to shutdown the old active switch compared to the fast-hello detection scheme due to the following reasons:

- BFD session establishment is based on IP connectivity. The hot-standby switch requires control plane initialization via SSO before it can start IP connectivity.
- Time required to start IP processes and installing the connected static route;
- Time required for BFD session initialization and session establishment between two chassis.

The impact of longer detection time on user data traffic might not be significant and depends on routing protocol and topology. The BFD-based detection is required in certain topologies for a better convergence. However, the BFD-based detection technique shall be deprecated in future software releases in lieu of improved hello detection of fast-hello. See the "Effects of Dual-Active Condition on Convergence and User Data Traffic" section on page 4-38.

BFD Configuration and Monitoring

BFD configuration requires a dedicated, directly connected physical Ethernet port between the two VSS chassis. BFD pairing cannot be enabled on Layer-3 EtherChannel or on a SVI interface. Sup720-10G one Gigabit uplink ports can be used only if the supervisor is not configured in 10 Gigabit-only mode

BFD configuration for dual-active differs from normal BFD configuration on a standard interface. The BFD session connectivity between switches is needed *only* during dual-active conditions. First, BFD detection must be enabled on global virtual-switch mode. Second, the dedicated BFD interface must have a unique IP subnet on each end of the link. In a normal operational state, the two connected interfaces cannot share the same subnet, yet that sharing is required for BFD peer connectivity during a dual-active event. Once interfaces are paired, the virtual switch self-installs two static routes as a connected route with paired interfaces. It also removes the static route upon un-pairing the interface.

The following commands are required to configure a BFD-based detection scheme.

Enable under VSS global configuration mode.

```
6500-VSS(config)# switch virtual domain 10
6500-VSS(config)# dual-active pair interface gig 1/5/1 interface gig 2/5/1 bfd
```

Enable unique IP subnet and BFD interval on the specific interfaces.

```
6500-VSS# conf t
6500-VSS(config)# interface gigabitethernet 1/5/1
6500-VSS(config)# ip address 192.168.1.1 255.255.255.0
6500-VSS(config)# bfd interval 50 min_rx 50 multiplier 3

6500-VSS(config)# interface gigabitethernet 2/5/1
```

```
6500-VSS(config)# ip address 192.168.2.1 255.255.255.0
6500-VSS(config)# bfd interval 50 min_rx 50 multiplier 3
```

The preceding configuration sequence results in the automatic installation of the required static route. The following messages are displayed on the display console.

```
Console Message:

adding a static route 192.168.1.0 255.255.255.0 Gi2/5/1 for this dual-active pair
adding a static route 192.168.2.0 255.255.255.0 Gi1/5/1 for this dual-active pair
```

Notice that the static route for the subnet configured on switch 1 interface (1/5/1 in above example) is available via interface residing on switch 2 (2/5/1). This configuration is necessary because static routes help establish the BFD session connectivity when chassis are separated during a dual-active event. The BFD protocol itself has no restriction on being on a separate subnet to a establish session.

Note The recommended BFD message interval is between 50 and 100 msec with multiplier value of 3. Increasing the timer values beyond recommended values has shown higher data loss in validating the best practices. Use unique subnet for BFD configuration which does not belong to any part of the IP address range belonging to your organization. Use route-map with redistribute connected (if required for other connectivity) to exclude BFD related connected static route.

BFD detection configuration can be monitored via following CLI commands:

```
6500-VSS# sh switch virtual dual-active bfd
Bfd dual-active detection enabled: Yes

Bfd dual-active interface pairs configured:
  interface-1 Gi1/5/1 interface-2 Gi2/5/1

6500-VSS# sh switch virtual dual active summary
Pagp dual ACTIVE detection enabled: No
Bfd dual ACTIVE detection enabled: Yes
No interfaces excluded from shutdown in recovery mode
In dual ACTIVE recovery mode: No
```

Configuration Caveats

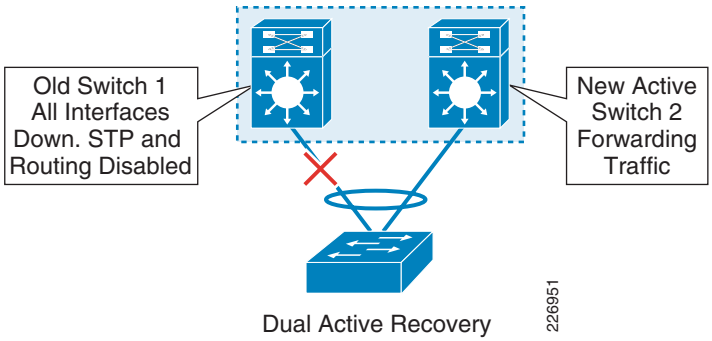
BFD hello timer configuration must be the same on both switches. In addition, any configuration changes related to IP addresses and BFD commands will result into removal of BFD detection configuration under global virtual switch mode and has to be re-added manually. This is designed to avoid inconsistency, because the validity of configuration cannot be verified unless a dual-active event is triggered. The following reminder will appear on the console:

```
6500-VSS(config)# interface gig 1/5/1
6500-VSS(config-if)# ip address 14.14.14.14 255.255.255.0
The IP config on this interface is being used to detect dual-active conditions. Deleting
or changing this config has deleted the bfd dual-active pair: interface1: Gi1/5/1
interface2: Gi2/5/1
deleting the static route 3.3.3.0 255.255.255.0 Gi1/5/1 with this dual-active pair
deleting the static route 1.1.1.0 255.255.255.0 Gi2/5/1 with this dual-active pair
```

Dual-Active Recovery

Once the detection technique identifies the dual-active condition, the recovery phase begins. The recovery process is the same for all three detection techniques. In all cases, the old-active switch triggers the recovery. In the examples presented in this guide, SW1 (original/old-active switch) detects that SW2 has now also become an active switch, which triggers detection of the dual-active condition. SW1 then disables all local interfaces (except loopback) to avoid network instability. SW1 also disables routing and STP instances. The old-active switch is completely removed from the network. See Figure 4-19.

Figure 4-19 Dual Active Recovery.



You can use the exclude interface option to keep a specified port operational during the dual-active recovery process-such as a designated management port. However, the **excluded port** command will not have routed connectivity, because the old-active switch does not have a routing instance. The following is an example of the relevant command:

```
VSS(config-vs-domain)# dual-active exclude interface port_nubmer
```

Note SVI or EtherChannel logical interfaces cannot be excluded during a dual-active event.

It is highly recommended to have console-based access to both the chassis during normal and dual-active conditions.

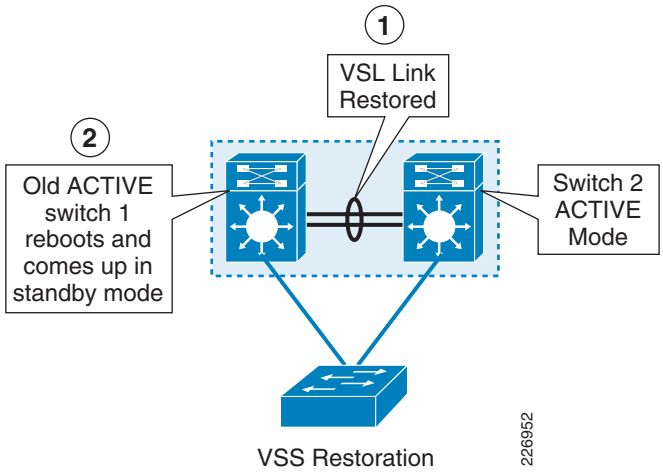
VSS Restoration

The VSS restoration process starts once the VSL connectivity is reestablished. The following events, which are causes of dual-active, restore the VSL connectivity between VSS switch members:

- Restoration of fiber connectivity; this can happen if the network suffered from a physically-severed fiber link.
- Reversal of configuration change, which could have shutdown the VSL bundle.
- Restoration of faulty hardware; this is the least probable event if a resilient design is adopted.

Figure 4-20 provides a high-level summary of the VSS restoration process.

Figure 4-20 VSS Restoration



Once the VSL connectivity is established, role negotiation (via the RRP protocol) determines that the previously active switch (SW1 in Figure 4-20 above) must become the hot-standby switch. There is no reason to change the role of the existing (new) active switch and thus triggering more data loss. This requires SW1 to be rebooted because a switch cannot go directly to the hot-standby state without a software reset. If no configuration mismatch is found, SW1 automatically reboots itself and initializes in the hot-standby mode. All interfaces on SW 1 are brought on line and SW1 starts forwarding packets-restoring the full capacity of the network. For example, the following console messages are displayed during VSL-bundle restoration on SW1 (previously active switch):

```
17:36:33.809: %VSLP-SW1_SP-5-VSL_UP:  Ready for Role Resolution with Switch=2,
MAC=001a.30e1.6800 over Te1/5/5
17:36:36.109: %dual ACTIVE-1-VSL_RECOVERED: VSL has recovered during dual ACTIVE
situation: Reloading switch 1
! << snip >>
17:36:36.145: %VSLP-SW1_SP-5-RRP_MSG: Role change from ACTIVE to HOT_STANDBY and hence
need to reload
Apr  6 17:36:36.145: %VSLP-SW1_SP-5-RRP_MSG: Reloading the system...
17:36:37.981: %SYS-SW1_SP-5-RELOAD: Reload requested Reload Reason: VSLP HA role change
from ACTIVE to HOT_STANDBY.
```

If any configuration changes occur during the dual-active recovery stage, the recovered system requires manual intervention by the use of the **reload** command and manual configuration synchronization. When network outages such as dual-active occur, many network operators may have developed a habit of entering into configuration mode in search of additional command that could be helpful in solving network outage. Even entering and exiting the configuration mode (and making no changes) will mark the configuration as dirty and will force manual intervention. Dual-active condition is also created when accidental software shut down of the VSL port-channel interface. The configuration synchronization process will reflect this change on both chassis. The only way to restore the VSL-connectivity is to enter into configuration mode, which will force the manual recovery of VSS dual-active. Once the VSL bundle is restored, the following syslogs messages will appear only in the old-active switch's console output:

```
11:02:05.814: %DUAL_ACTIVE-1-VSL_RECOVERED: VSL has recovered during dual-active
situation: Reloading switch 1
11:02:05.814: %VS_GENERIC-5-VS_CONFIG_DIRTY: Configuration has changed. Ignored reload
request until configuration is saved
11:02:06.790: %VSLP-SW1_SP-5-RRP_MSG: Role change from Active to Standby and hence need to
reload
11:02:06.790: %VSLP-SW1_SP-5-RRP_UNSAVED_CONFIG: Ignoring system reload since there are
unsaved configurations. Please save the relevant configurations
```

```
11:02:06.790: %VSLP-SW1_SP-5-RRP_MSG: Use 'reload' to bring this switch to its preferred
STANDBY role
```

For the VSS to operate in SSO mode requires that both chassis have exactly identical configurations. The configuration checks to ensure compatibility are made as soon as the VSL link is restored. Any changes (or even simply entering into configuration mode) will mark the flag used for checking the configuration status as dirty—implying possible configuration mismatch. This mismatch might requires one of the following:

- Correcting the mismatched file and reflecting the change so that the configuration on both chassis match
- Saving the configuration to NVRAM to clear the flag

Two types of configuration changes are possible during a dual-active event:

- "Non-VSL Link Configuration Changes" section on page 4-36
- "VSL-Link Related Configuration Changes" section on page 4-36

Both of these configuration change types are discussed in the next sections. Each requires a proper course of action.

Note The behavior of a system in response to configuration changes might depend on the Cisco IOS version implemented. The behavior description that follows applies only to the Cisco IOS Release 12.2(33)SXH.

Non-VSL Link Configuration Changes

For any configuration changes that do not affect the VSL bundle, you must determine to which chassis those changes apply. If changes are on the old-active switch, saving the configuration and manually rebooting the switch will restore the switch in hot-standby mode. If the changes were saved on old-active before the VSL link is being restored, manually rebooting the old-active switch might not be required because saving the configuration clears the dirty status flag. If the changes were made to the active switch, then those changes do not affect dual-active recovery activity. After the recovery (once the VSL link is restored), the new active switch configuration will be used to overwrite the configuration in the peer switch (the old-active switch) when it becomes the hot-standby switch. Changes made to the active switch need not match the old-active switch configuration because the configuration on the old-active switch (now the hot-standby switch) will be overwritten.

VSL-Link Related Configuration Changes

The dual-active condition can be triggered by various events, including the following:

- A user-initiated accidental shutdown of the VSL port-channel
- Changes to the EtherChannel causing all links or the last operational VSL link to be disconnected

When changes to the VSL port-channel are made, the changes are saved in both chassis before the dual-active event is triggered. The only way to restore the VSL-related configuration mismatch is to enter into the configuration mode and match the desired configurations. If the configuration-related to VSL links are not matched and if old-active chassis is rebooted, the chassis will come up in route processor redundancy (RPR) mode. It is only during the old-active switch recovery (in this case manual reboot) that the VSL configurations mismatch syslogs output will be displayed on the new-active switch. The following syslogs output examples illustrate this mismatch output:

```
Aug 28 11:11:06.421: %VS_PARSE-3-CONFIG_MISMATCH: RUNNING-CONFIG
Aug 28 11:11:06.421: %VS_PARSE-3-CONFIG_MISMATCH: Please use 'show switch virtual
redundancy config-mismatch' for details
Aug 28 11:11:06.421: %VS_PARSE-SW2_SP-3-CONFIG_MISMATCH: VS configuration check failed
```

```
Aug 28 11:11:06.429: %PFREDUN-SW2_SP-6-ACTIVE: Standby initializing for RPR mode
Aug 28 11:11:06.977: %PFINIT-SW2_SP-5-CONFIG_SYNC: Sync'ing the startup configuration to
the standby Router.
6500-VSS#show switch virtual redundancy | inc Opera
     Operating Redundancy Mode = RPR
```

VSL-related configuration changes are viewed via the **show switch virtual redundancy config-mismatch** command. The following is an example output:

```
6500-VSS# show switch virtual redundancy config-mismatch

Mismatch Running Config:
Mismatch in config file between local Switch 2 and peer Switch 1:
ACTIVE  : Interface TenGigabitEthernet1/5/4 shutdown
STANDBY : Interface TenGigabitEthernet1/5/4 not shut
In dual-active recovery mode: No
```

In RPR mode, all the line cards are disabled, except where the VSL is enabled. Once the configuration is corrected on the active switch, the **write memory** command will cause the startup configuration to be written to the RPR switch supervisor. The redundancy **reload peer** command will reboot the switch from RPR mode to the hot-standby mode. The following configuration sequence provides an example of this process:

```
6500-VSS# conf t
6500-VSS(config)# int te1/5/4
6500-VSS(config-if)# no shut
6500-VSS(config-if)# end
6500-VSS# wr mem

Aug 28 11:17:30.583: %PFINIT-SW2_SP-5-CONFIG_SYNC: Sync'ing the startup configuration to
the standby Router. [OK]
6500-VSS# redundancy reload peer
Reload peer [confirm] y
Preparing to reload peer
```

If the configuration correction is not synchronized before VSL-link restoration, a VSL-configuration change can cause an extended outage because the only way to determine whether a VSL-configuration mismatch has occurred is after the old-active switch boots up following VSL-link restoration. That means the switch will undergo two reboots. First, to detect the mismatch and then second boot up is required with corrected configuration to assume the role of hot-standby. To avoid multiple reboots, check for a VSL-configuration mismatch *before* the VSL link has been restored. Users are advised to be particularly cautious about modifying or changing VSL configurations.

Tip **The best practice recommendation is to *avoid* entering into configuration mode while the VSS environment is experiencing a dual-active event; however, you cannot avoid configuration changes required for accidental shutdowns of the VSL link or the required configuration changes needed to have a proper VSL restoration.**

Effects of Dual-Active Condition on Convergence and User Data Traffic

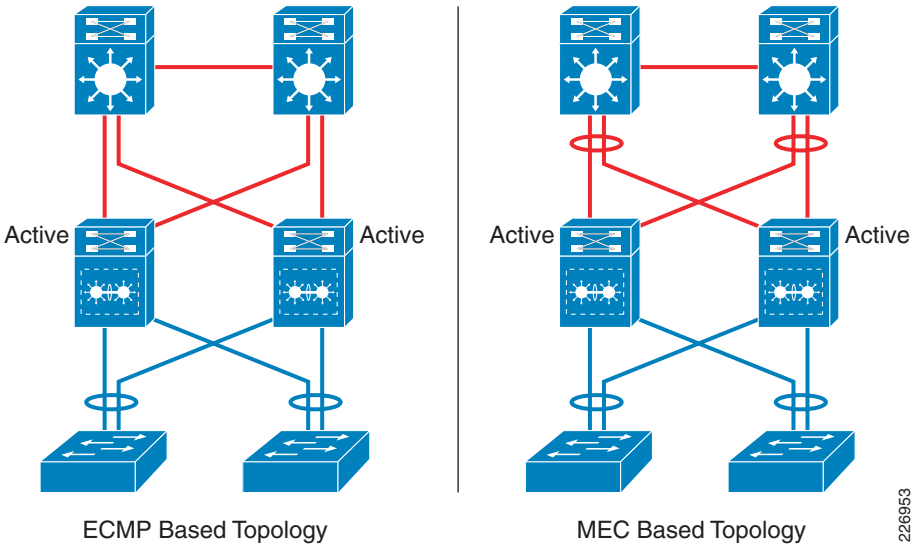
This section covers the effects of the dual-active condition on application and user data traffic. It is important to keep in mind that each detection technique might take a different amount of time to detect the dual-active condition, but this problem's influence upon the speed with which user data traffic is restored depends on many factors. These are summarized in the list of convergence factors that follows. Highly granular details of events (during dual-active) and their interactions with the following convergence factor are beyond the scope of this design guide. Overall, what matters is the selection of a detection technique for a specific environment based on observed convergence data.

User traffic convergence is dependent on the following factors:

- Dual-active detection method—Enhanced PAgP, fast-hello, or BFD
- Routing protocol configured between Layer-3 core to VSS—Enhanced IGRP or OSPF
- Topology used between VSS and Layer-3 core—ECMP or MEC
- SSO recovery
- NSF recovery

Figure 4-21 provides validation topologies with which all the observations with dual-active events and convergence associated with data traffic are measured. It is entirely possible to realize better or worst convergence in a various combination of the topologies such as all Layer-2, Layer-3, or end-to-end VSS. However, general principles affecting convergence remains the same.

Figure 4-21 Comparison of MEC and ECMP Topologies



The ECMP-based topology has four routing protocol adjacencies, where as MEC has two. For the ECMP-based topology, the VSS will send a separate hello on each link, including the hello sent via hot-standby-connected links. This means the that active switch will send the hello over the VSL link to be sent by links connected via hot-standby. The hello sent by core devices will follow the same path as VSS. For MEC-based topology, the hello originated from VSS will always be sent from local links of an active switch. However, the hello originated from core devices may select the link connected to hot-standby based on hashing result. This behavior of link selection from core devices is also repeated for routing update packets. As a result, ECMP and MEC topology exhibits different behavior in the settling the routing protocol adjacency and NSF procedure. In turn, it plays a key role in how fast the convergence of data traffic is possible.

The sequence of events that occur during a dual-active event with a detection technique deployed are generalized below for contextual reference and do not comprise a definitive process definition.

1. Last VSL link is disabled.
2. The currently active switch does not know whether the VSL has become disabled or the remote peer has rebooted. The currently active switch assumes that the peer switch and related interfaces are lost and treat this as an OIR event. As a result, the hot-standby switch interfaces are put into the down/down state, but the remote switch (current hot-standby switch) is still up and running. Meanwhile, local interfaces attached to old-active switch remain operational and continue forwarding control and data traffic. The active switch attached interface may advertise the routing update about remote switch (hot-standby) interfaces status as observed during this event.
3. As a result of all the VSL links being disabled, the hot-standby switch transitions to the active role not knowing whether the remote switch (the old-active switch) has rebooted or is still active. In this case, the situation is treated as a dual-active condition because the old-active switch has not undergone a restart.
4. The new-active switch initializes SSO-enabled control protocols and acquires interfaces associated with local chassis (the line protocol status for each of these interfaces does not become non-operational due to SSO recovery).
5. The newly-active supervisor restarts the routing protocol and undergoes the NSF recovery process, if the adjacent routers are NSF-aware; otherwise, a fresh routing-protocol adjacency restart is initiated. (See the "Routing with VSS" section on page 3-44.)
6. VSS SSO recovery follows the same process of separation of the control and the data plane as with a standalone, dual-supervisor configuration. As a result, the forwarding plane in both the switches remains operational and user traffic is switched in hardware in both switches. This forwarding continues until the control plane recovers. The control plane recovery occurs on both active switches. The old-active switch simply keeps sending routing protocol hello and some update regarding remote switch interfaces Layer-3 status. The newly-active switch restarts the routing protocols and tries to require adjacency with Layer-3 core devices. These parallel control plane activity might lead to adjacency resets, resulting in forwarding path changes (depends on usage of routing protocol and topology) and dual-active detection triggering the shutting down of the old-active switch interfaces.

If implemented, a dual-active detection method determines how fast detection occurs, which in turn triggers the shutting down of the old-active interfaces. Enhanced PAgP and fast-hello take around two-to-three seconds, while BFD takes 22-to-25 seconds. However, detection technique alone does not influence the convergence of user data traffic flow. Even though a faster detection method might be employed, the impact on user data traffic might be greater and vice versa (slower detection method having a better user data convergence).

The convergence validation and traffic-flow characterization during a dual-active event are presented in the following sections—segmented by routing-protocol deployed. Although ECMP- and BFD-based environments are described in general, only the MEC-based topology is used in the detailed descriptions that follow depicting events specific to dual-active conditions for each routing protocol.

Tip **The routing-protocols are recommended to run default hello and hold timers.**

Convergence from Dual-Active Events with Enhanced IGRP

Enhanced PAgP and Fast-Hello Detection

Both ECMP- and MEC-based connectivity to the core delivers user traffic convergence that is below one second.

BFD

The ECMP-based core BFD design provides the same convergence characteristics as enhanced PAgP. The recovery with the MEC-based core is more complex. The MEC core design experiences greater loss because of the destabilization of enhanced IGRP adjacencies leading to the removal routes affecting both upstream and downstream convergence. See Figure 4-22.

Figure 4-22 VSS Dual-Active Convergence with Enhanced IGRP

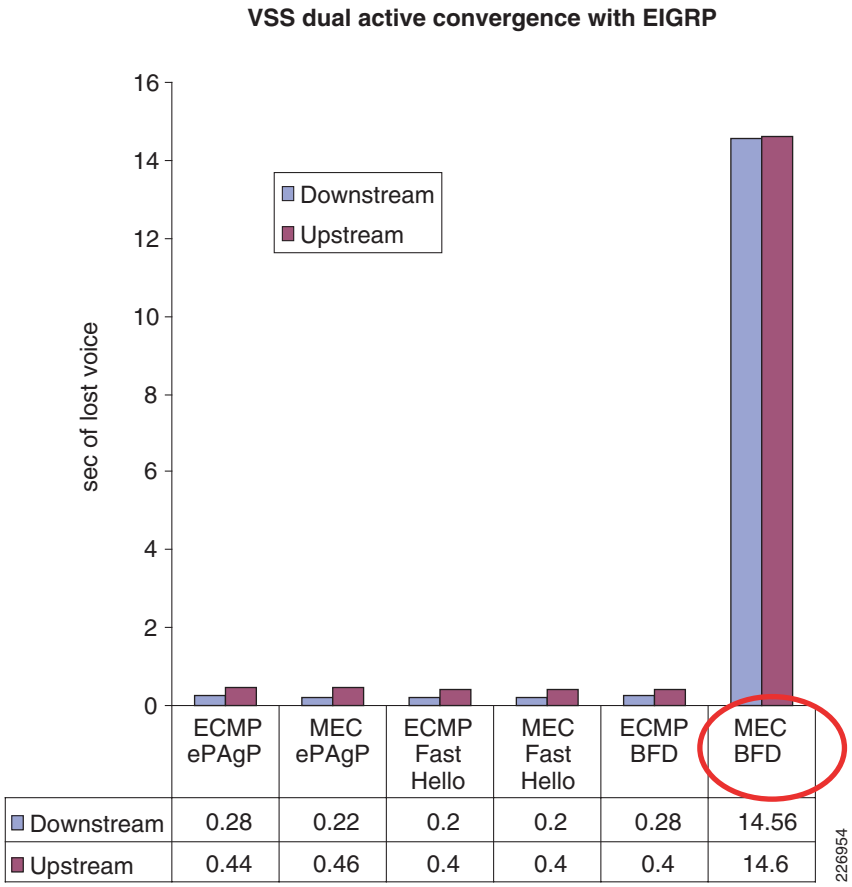


Table 4-13 describes the losses and root causes of the convergence for each combination.

Table 4-13 Convergence Recovery Losses and Causes (Enhanced IGRP Environment)

Dual-Active Detection Protocol	Core-to-VSS Connectivity	End-to-End Convergence	Summary of Recovery and Loss Process During a Dual-Active Event
Enhanced PAgP or Fast-Hello	ECMP	Upstream and downstream: 200-to-400 msec	MEC loses one link from the old active switch viewpoint, but no routes withdrawal notifications are sent or removed during EC link change. Before the new active switch announces the adjacency, the old active switch is shut down (2-to-3 seconds). A clean NSF restart occurs because no routes are withdrawn from the restarting NSF peer while the new active switch announces the NSF adjacency restart (7-to-12 seconds). The only losses related to recovery occur is the bringing down of interface by an old-active switch.
	MEC	Upstream and downstream: 200-to-400 msec	MEC loses one link from old-active viewpoint; however, no routes withdrawal are sent or removed during EC link change. Before the new active announces the adjacency, the old active is shutting down (2-1/2 sec). Clean NSF restart (as no routes being withdrawn from NSF restarting peer while the new active announces the NSF adjacency restart (7-12 sec))
BFD	ECMP	Upstream and downstream: 200-to-400 msec	Same behavior and result as in ECMP with enhanced PAgP (or fast-hello), even though BFD detection takes longer to complete the detection and shut down the old-active switch. Until the shutdown is initiated, the traffic to the old active continues to be forwarded in hardware. The old-active switch removes all the interfaces connected with the peer and sends updates with an infinite metric value to the core. However, the new-active switch interfaces are operational. No routes are withdrawn from the core router because Enhanced IGRP does not perform a local topology calculation until an explicit announcement occurs regarding the routes to be queried via the interface. Meanwhile, the new-active switch undergoes the NSF restart and refreshes adjacency and route updates. No routes or adjacency resets occur during a dual-active event.
	MEC	Upstream and downstream: 4-to-14 seconds	Higher data loss is observed because some Enhanced IGRP neighbor adjacencies might settle on one of the active routers based on the IP-to-Layer-3 MEC member link hashing toward the VSS. The Enhanced IGRP update and hello messages are hashed to different links from the core routers, leading to a loss of adjacency and routes. Either Enhanced IGRP adjacency settles on the old-active or new-active switches. If it settles on the old-active switch, traffic loss is more pronounced because Enhanced IGRP adjacencies must be reestablished with the new-active switch after old-active switch shuts down. As a result, the time required to complete convergence will vary.

Details of BFD Detection with an MEC-Based Topology

As shown in Table 4-13, this combination causes more instability because the detection technique takes longer to complete and adjacency destabilization leads to more traffic disruption. This severity of the destabilization depends on the hash result of source and destination IP addresses and the Enhanced IGRP hello (multicast) and update (unicast) transmissions—which are sent out on a MEC link member that is connected to either the old-active or new-active switch. The Enhanced IGRP packet forwarding path in a normal topology can adopt one of the following combinations in the VSS topology:

- Multicast on the old-active switch and unicast on the new-active switch
- Multicast on the new-active switch and unicast on the old-active switch
- Multicast and unicast on the old-active switch
- Multicast and unicast on the new-active switch

With the preceding combinations in mind, any of the following events can cause an Enhanced IGRP adjacency to reset during dual-active condition:

- Enhanced IGRP adjacency settling on one of the active switches such that the hellos from the core are not received and the hold-timer expires. This occurs because during normal operating conditions, the hash calculation resulted in the sending of hellos to the hot-standby switch that was forwarding packets over the VSL link and now is no longer doing so (VSL link is down). As a result, the old-active switch never sees the hello packet, resulting in adjacency time out. This leads to loss of routes on core routers pertaining to access-layer connected to the VSS and loss of routes on VSS for any upstream connectivity.
- The NSF signal timer expired because remote routers did not respond to the NSF restart hellos from new active switch. This is possible because the remote routers (core) hashing may send hello and NSF hello-ack to old active. Subsequently, the NSF time out will be detected by new active, which will prevent a graceful recovery. As a result, a fresh adjacency restart is initiated.
- The NSF restart process got stuck during the route update process (for example, the update of routes were sent to the old-active switch using unicast hashing) and the new active supervisor declares it adjacency is stuck in the INIT state and forces a fresh restart.

The location of adjacency settlement after the dual-active event determines the variation in convergence:

If the IP addressing is configured such that the adjacency can settle on the new-active switch during dual-active condition where it might not have to undergo a fresh restart of the adjacency, it may lead to a better convergence. However, this will not be consistent during next dual-active condition as the adjacency now settles on old-active, leading to one of trigger conditions described in preceding paragraph.

If the adjacency settles on old-active switch, after 22-to-25 seconds the dual-active event triggers an internal shutdown (different from an administrative shutdown) which then causes the adjacency process to restart with the new active switch that will go through a normal adjacency setup (not an NSF graceful restart). Adjacency resets result in routes from the core and the VSS being withdrawn and subsequent variable packet loss for downstream and upstream paths.

It is possible that on a given network, one may only see partial symptoms. It is difficult to fully and consistently characterize traffic disruption due to convergence because of the many variables involved. The key point is to have a reasonable understanding of the many factors that influence convergence during BFD detection and that these factors can cause convergence following BFD detection to take higher than any other combination.

Convergence from Dual-Active Events with OSPF

OSPF inherently requires unique connectivity for building and maintaining the shortest-path-first (SPF) database. It has two built-in verification check that creates more network visibility under a dual-active condition-router ID and bidirectional verification of neighbor reachability.

Enhanced PAgP and Fast Hello

The ECMP-based core design experiences a higher rate of traffic loss because OSPF removes access-layer routes in the core during a dual-active event. OSPF does this because of the duplicate router IDs seen by the core routers.

The convergence is much better with MEC-based topology. An MEC-based core design does not suffer from route removal in the core because no OSPF route withdrawals are sent (EtherChannel interfaces are still operational). In addition, the detection of dual-active is triggered within 2-to-3 second followed with recovery of control plane on new active switch. During the recovery both switch member interfaces keep forwarding data leading to convergence times that are below one second in duration.

BFD

The ECMP-based core design experiences better convergence compared to the enhanced PAgP design because of the delayed recovery action by BFD that leaves at least one access-layer route operational in the core.

The recovery with a MEC-based core is more complicated. The MEC core design has a higher rate of traffic loss because of adjacency destabilization that leads to the removal of routes. This affects both upstream and downstream convergence. In addition, downstream losses are increased because the core removes routes for the access-layer subnet faster because it detects the adjacency loss. In contrast, the VSS retains the adjacency and the upstream routes. Figure 4-23 compares dual-active convergence given differing configuration options.

Figure 4-23 VSS Dual-Active Convergence with OSPF

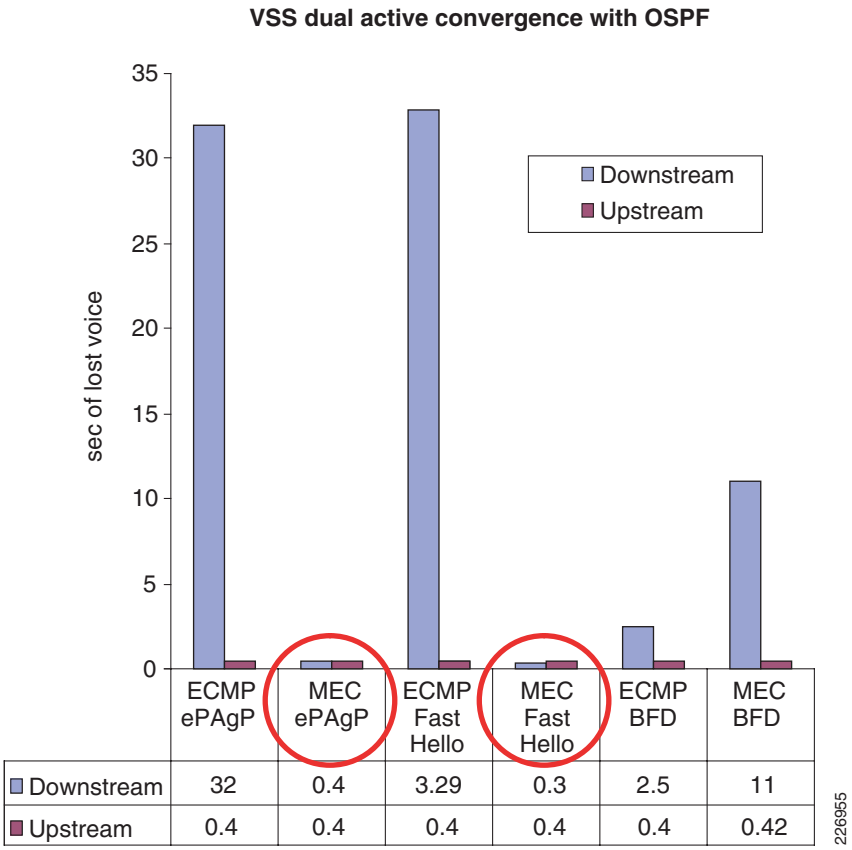


Table 4-14 describes the losses and root causes of the convergence for each combination.

Table 4-14 Convergence Recovery Losses and Causes (OSPF Environment)

Dual-Active Detection Protocol	Core-to-VSS Connectivity	End-to-End Convergence	Summary of Recovery and Loss Process During a Dual-Active Event
Enhanced PAgP	ECMP	Downstream 30-to-32 Second Upstream 200-to-400 msec	Downstream traffic loss is higher because all four routes to the access-layer are withdrawn. The loss of VSL link forces the old active switch to remove all interfaces on the hot-standby switch (the new active switch) as being disabled (although they are actually operational); this route updates is announced to core routers , which causes the withdrawal of two routes learned from the new active switch (even though they are operational). Enhanced PAgP or fast-hello detection shuts down the old active switch interfaces which triggers the withdrawal of a second set of routes from core routing table. Until the new active switch completes its NSF restart and sends its routes to the core routers, downstream traffic will be black holed. The upstream route removal does not happen because no duplicate router ID is seen by the VSS.
	MEC	Downstream and upstream 200-to-400 msec	No routes are removed during EtherChannel link change. Before the new active switch announces the adjacency restart; the old active switch is shut down (2.5 seconds). A clean NSF restart occurs (no routes were withdrawn for the same restarting NSF peer before the new active switch announces the NSF adjacency restart (7-to-12 seconds).
BFD	ECMP	Downstream 2-to-2.5 sec Upstream: 200-to-400 msec	Similar to the preceding enhanced PAgP-ECMP case, but BFD does not disconnect the old active switch for 22-to-25 seconds which keeps at least one route in core for access-layer subnets. Keeping that route helps to prevent traffic being black-holed until the NSF recovery process is completed on new active switch.
	MEC	Downstream: 200 msec to 11 seconds Upstream: 200-to-400 msec	Traffic is affected by several events occurring in parallel. OSPF adjacency might not stabilize until BFD shuts down the old active interfaces (22-to-25 seconds). Depending on which VSS member is active and where the OSPF hello is hashed, the stability of NSF restart might be affected. If the NSF restart occurs cleanly, the convergence can be below one second.

Details of BFD Detection with MEC-Based Topology

As described before, the asymmetrical hashing of hello (multicast) and update (unicast) messages from the core to VSS is possible with MEC in normal operational circumstances, as well as under a dual-active condition. From the VSS to the core, control plane connectivity remains on local interfaces. This combination of behaviors can cause the reset of the adjacency in a dual-active condition. In addition, OSPF adjacency might never stabilize with either of the active VSS routers because of bidirectional neighbor availability verification in the OSPF hello protocol for adjacency formation.

In normal operating condition, the core routes view a single router ID for VSS. During dual-active, core routes will see the same router ID be announced to two active supervisors. The core routers SPF is in state of confusion and will display the duplicate router ID in syslogs if the detail adjacency logging is turned on under OSPF process. For the OSPF adjacency settlement, the core routers will respond to the request coming from either the old or new VSS active switch, not knowing what really happened.

However, the core routers send the multicast hello to either old or new active switch depending on hashing (source IP of the interface address and destination of 224.0.0.5). During a dual-active event, two possibilities arise in which the hello can be sent from the core:

- Link connected to new active switch—While the core is sending the hello to the link connected to new active switch, the old active router is up and continues sending normal OSPF hellos back to core via its local link. At the same time, the new-active router will try to establish adjacency and restart its connection to the core routers by sending special hellos with the RS bit set (NSF restart). This adjacency restart might be continued until the hello from old active without NSF/RS bit set is received via the core router (old active router is up and running because it does not know what happened). This leads to confusion in the core router's NSF aware procedure and that might cause the core router to reset its adjacency. Meanwhile, the old-active router might also time out because it has not received any hellos from the core. Eventually, either of the active VSS switch neighbors will reset the adjacency. Once the adjacency reset is triggered, the core router will retry to establish neighbor adjacency to the new-active router (due to hashing) reaching FULL ADJ, meanwhile the old active router will try to send hello again, this time core routers do not see its own IP address in the received hello as it is an INIT hello from an old active. This will prompt the core to send fast-hello transmissions and new Database Descriptor (DBD) sequence number to the new active router (as it was almost at FULL ADJ with new active router). The new active router complains this with a BAD_SEQUENCE number and resets the ADJ from FULL to EX-START.
- Link connected on old active switch—In this case, hashing turns out to be such that core sends hello messages to the old-active switch and the new active will start first with NSF restart and then INIT hello. The new-active router does not see the response received from the core routers as core routers keep sending response to the old active. As a result, eventually the adjacency restart will be issued by the new active supervisor. This will continue indefinitely if no detection mechanism is employed or (in case of BFD) when the adjacency reset might cause higher packet loss.

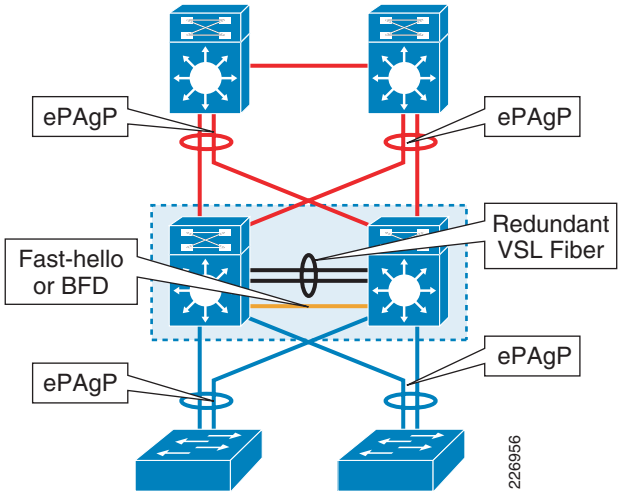
Dual-Active Method Selection

It is obvious that multiple techniques are possible for deployment. Multiple detection techniques deployment is not the replacement for resilient VSS-link configuration. In the presence of multiple detection techniques, enhanced PAgP and fast-hello will be detected first when compared to BFD. In the presence of multiple methods, whoever detects the dual-active first governs the convergence.

The only exception in deploying multiple methods is where OSPF routing enabled with ECMP-based topology. In this topology, the only detection method recommended is BFD, because BFD is the only method that gives the best convergence. If any other method is deployed along with BFD, the BFD will not be the first to detect the dual-active (BFD takes longer time compared to enhanced PAgP or fast-hello).

Enhanced PAgP detection is possible via Layer-2 or Layer-3 MEC. Enhanced PAgP detection might only need to be run on a single neighbor. However, using enhanced PAgP on all interfaces will ensure that, in the worst case, at least one switch is connected to both members of the same VSS pair (assuming that not all cable paths are affected in a failure condition) so that a path will exist for recovery. Figure 4-24 illustrates a high-level view of a topology featuring multiple redundancies to ensure VSL link availability and the placement of detection tools to help reduce traffic disruption under dual-active event conditions.

Figure 4-24 Dual-Active Detection Possibilities



Summary and Recommendations

This section summarize the recommended dual-active detection techniques that one must enable. These recommendations are based on the previous validated convergence behavior possible with the given combination of topology and dual-active detection method. The following recommendations apply:

- More then one detection technique should be deployed at once to avoid failure of a single technique jeopardizing the downtime.
- If ePAgP method is used, enable ePAgP detection on more then one access-layer/core switch.
- ePAgP and Fast Hello detection methods enabled with Layer 3 MEC topology in the core provides sub-second convergence for both OSPF and EIGRP.
- With Pre-12.2(33)SXI3 release, in conjunction with OSPF and ECMP topology where BFD detection should be used, do not deployed any other technique. BFD is being deprecated in future software releases. The BFD detection method should **only** be used for releases prior to 12.2(33)SXI3; see the next recommendation.
- 12.2(33)SXI3 release allows sub-second operation of fast-hello function for rapid detection and recovery (CSCsy30937). This improvement enables sub-second recovery, eliminating topological and routing protocol dependency of dual active detection and recovery.

The new improvements in Fast Hello detection do not alter the convergence results described in this design guide. Refer following design guide for the latest information: http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/Borderless_Campus_Network_1.0/Borderless_Campus_1.0_Design_Guide.html.

Table 4-15 provides the quick summary of these recommendations, where “Good” means recovery below one second and “Ok” means recovery takes more then one second and variable up to 32 seconds.

Table 4-15 Summary of Recovery Comparisons for Convergence Options

	Pre-12.2(33)SXI3			With 12.2(33)SXI3
Dual Active Detection	**ePAGP**	**BFD**	**Fast Hello**	**Sub-second Fast Hello**
EIGRP with ECMP-based topology to the core	Good	Good	Good	Good
EIGRP with L3-MEC-based topology to the core	Good	OK	Good	Good
OSPF with ECMP-based topology to the core	OK	Good	OK	Good
OSPF with L3-MEC-based topology to the core	Good	OK	Good	Good

Note As seen from above and other references discussed in this design guide, MEC-based connectivity to the core enables convergence times below one seconds for both unicast and multicast.