Integrating the Virtual Switching System in Cisco Data Center Infrastructure

Contents

Corporate Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Introduction

This document provides reference architectures and configuration guidance using the Cisco Catalyst 6500's Virtual Switching System (VSS) 1440 within an enterprise data center. The classic hierarchical architectures presented in this guide position VSS technology and functionality at the aggregation and access layers of the data center network.

Audience

This document is intended for network engineers and architects who need to understand the design options and configurations necessary for the Cisco Catalyst 6500 virtual switching services in the data center network.

Document Objective

The objective of this document is to provide customers with guidance on how to deploy VSS functionality in a Cisco-based data center. This document is not intended to introduce the user to basic Cisco data center design best practices, but rather to build on these well-documented concepts. The prerequisite Cisco data center design knowledge can be found at the following locations:

- Cisco.com—Data Center:

  http://www.cisco.com/go/dc
- Cisco Validated Design (CVD) Program:

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

For additional information regarding VSS and migration techniques, refer to the following:

- http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/prod_white_paper0900aecd806ee2ed.html

Overview

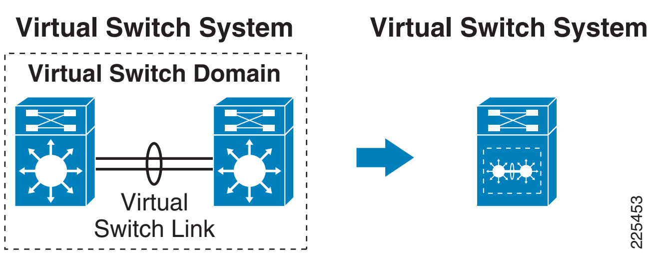
The data center is a critical portion of the enterprise network. The data center network design must address the high availability requirements of any device or link failure. It is also an area in which more intelligence is required from the network in order to perform security and application services. This document describes the introduction of the Cisco Catalyst 6500 VSS in the enterprise data center design. In particular, this publication addresses the implementation of VSS technology at the access and aggregation layers of the data center and its effect on the availability of data center services and applications.

VSS Technology

VSS technology allows for the grouping of two Cisco Catalyst 6500 switches into a single virtual switch. A VSS system provides physical infrastructure redundancy while simultaneously simplifying the logical topology of the data center.

Figure 1 illustrates the concept of VSS. The left side of Figure 1 represents the physical layout of the VSS: two Cisco Catalyst 6500s are physically connected through a virtual switch link (VSL). The two switches are members of a virtual switch domain and, as the right side of the figure shows, this construct forms a single logical switch with a single control plane—a virtual switching system.

Figure 1 Virtual Switch System



Note The VSS is sometimes referred to as the VSS1440, because it provides for 1.4 Tbps of forwarding switching fabric.

The primary benefits of this logical grouping include the following:

- Increased operational efficiency of a simplified network leveraging virtualization
- Increased availability and forwarding performance via Inter-chassis Stateful Switchover (SSO) and Nonstop Forwarding (NSF)
- Increased availability and forwarding performance via Multichassis EtherChannel (MEC)

The enterprise data center can leverage these VSS advantages. The remainder of this document will explore the integration and impact of this technology on the data center and details how these benefits can be achieved.

Note This document focuses on the design aspects of VSS in a data center environment. For more information on VSS technology, refer to the following URLs:
http://www.cisco.com/en/US/products/ps9336/index.html or
http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/prod_white_paper0900aecd806ee2ed.html

Key Concepts

The following section describes the fundamental building blocks of the VSS functionality including:

- Virtual Switch Domain, page 4
- Virtual Switch Link (VSL), page 5
- Multichassis EtherChannel (MEC), page 5

Virtual Switch Domain

A Virtual Switch Domain consists of two Cisco Catalyst 6500s as members that meet the minimum software and hardware requirements to obtain VSS functionality. See the "Required Components" section on page 11 for more details. The virtual switch domain is the boundary of the logical switching system; each domain should be identified by a unique system ID. Currently, the Virtual Switch Domain may consist of only two Cisco Catalyst 6500 platforms. Although the number of member switches is limited to two per domain, the number of domains is not; there are 255 unique domain IDs available.

The VSS employs an active/standby control topology where the active VSS switch performs all control plane functions. The following list highlights some of this control traffic:

- Layer 2—EtherChannel, Port Aggregation Protocol (PAgP), Link Aggregate Control Protocol (LACP), Spanning Tree Protocol (STP)
- Layer 3—Open Shortest Path First (OSPF), Enhanced Interior Gateway Protocol (Enhanced IGRP), Virtual Private Network (VPN) Routing and Forwarding (VRF), and so on
- First-hop Redundancy Protocols—Hot Standby Routing Protocol (HSRP), Virtual Router Redundancy Protocol (VRRP), and so on
- Management Protocols—Simple Network Management Protocol (SNMP), Telnet, Secure Shell (SSH), and so on

The active virtual switch is chosen during the instantiation of the domain using the Role Resolution Protocol (RRP) across the newly active VSL. In addition, the initialization of the domain requires that all hardware and software requirements are met and configurations are synchronized between the virtual switch domain members. These are all functions of the Virtual Switch Link Protocol (VSLP) that runs between the two domain members.

When in a normal operating state, the VSS data plane is active/active and both switches in the domain are forwarding traffic. The inter-chassis Nonstop Forwarding/Stateful Switchover (NSF/SSO) allows the standby switch to forward traffic. Each of the Policy Feature Cards (PFC) of the active and standby supervisors performs forwarding decisions for traffic ingress to the their local switch ports. It should be noted that use of Distributed Forwarding Cards (DFC) features further enhances the forwarding capabilities of the system.

Note VSS domain configuration details can be found in the "Core and Aggregation Layer Features" section on page 14 in the "VSS Design in the Aggregation Layer" section.

Virtual Switch Link (VSL)

As shown in Figure 1, the Virtual Switch Link (VSL) is an inter-switch link (ISL) that forms the backbone of the VSS. The VSL supports control traffic between domain switches allowing the VSS system to form and operate. In addition, normal data traffic may also leverage the VSL connection as a valid forwarding path. The VSL link benefits from the high availability and scalability features of Cisco EtherChannel.

The communication between VSS members across the VSL uses the Virtual Switch Link Protocol (VSLP). The VSLP includes the following protocols:

- Link Management Protocol (LMP)
- Role Resolution Protocol (RRP)

LMP manages the VSL link providing for the exchange of domain member identities and switch parameters necessary to form the VSS. RPR validates the capabilities of the domain members and coordinates the active switch election process. In addition to these functions, VSLP monitors the state of the VSL connection via probe messages.

All traffic traversing the VSL will have a 32-byte VSL header (VSH) inserted between the Ethernet preamble and Layer-2 frame header. Each frame on the VSL is subject to predefined quality of service (QoS) rules that favor VSS control traffic to maintain system stability. The addition of the VSH to the Ethernet frame requires ASIC support and thus the minimum VSS hardware requirements detailed in the sections that follow.

Multichassis EtherChannel (MEC)

EtherChannel allows multiple physical ports in a single switch to be aggregated forming one logical port. This logical port may be defined as a Layer-2 or Layer-3 interface consisting of a maximum of eight ports. The ports comprising the EtherChannel interface are often spread across the line cards of a modular switch to provide a more resilient connection. In general, EtherChannel improves the overall scalability and availability of the network and is a well-documented best-practice within the data center. As Figure 2 illustrates, an EtherChannel may be defined between two switches or a server and switching platform.

Figure 2 Switch and Server EtherChannel Examples

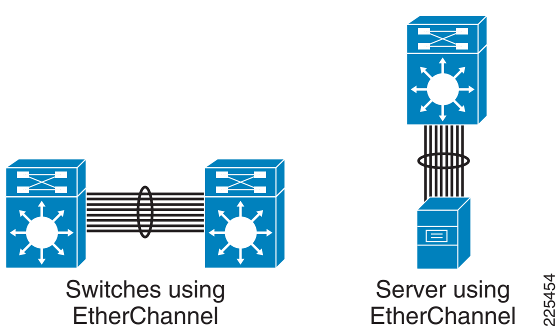


Figure 2 also exemplifies the one-to-one relationship that traditional EtherChannel entails. This is where the VSS system expands the realm of EtherChannel to create a one-to-many relationship using MEC.

VSS allows an EtherChannel to exist across two physical chassis that are logically one. As shown in Figure 3, from a control plane perspective the VSS system is a single switch and thus the traditional switch may leverage EtherChannel. The right side of Figure 3 details that the EtherChannel is supported across two VSS-enabled chassis. This channel is forwarding on all ports; from a logical Layer-2 perspective there is not loop. Convergence is no longer dependent on the implementation of spanning tree, but on the resilience of EtherChannel itself. The forwarding fabric is expanded and the Layer-2 topology simplified.

Figure 3 VSS MEC Logical and Physical Switching Topology

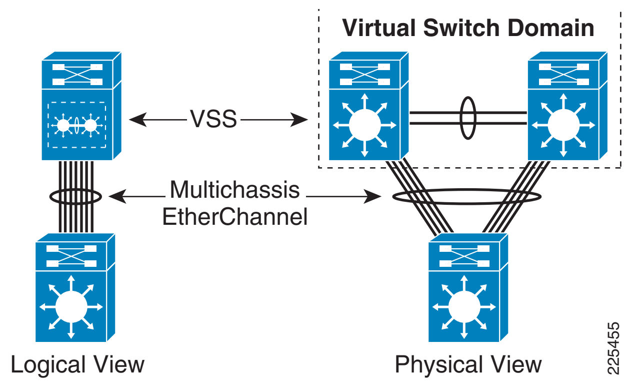
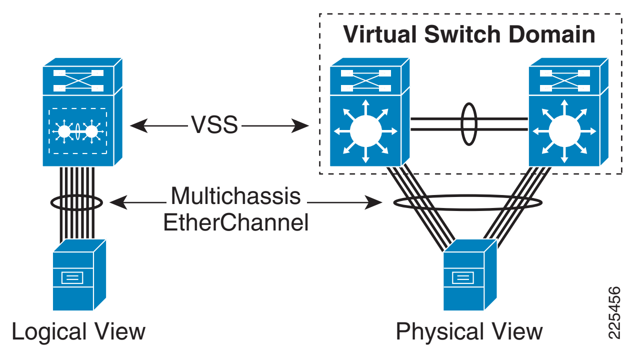


Figure 4 shows that the benefits of MEC are not limited to network switching devices, but extend to the server farm. The left side of Figure 4 shows the logical topology. A single server aggregates ports to the VSS-enabled neighboring switch. This would appear to be a single point-of-failure until one reviews the right side of the slide representing the physical layout of VSS with MEC.

Figure 4 VSS MEC Logical and Physical Server Topology



In both switch and server configurations, a VSS-enabled switching layer enhances the availability and forwarding capabilities of traditional switch and server technologies. This benefit is transparently achieved through the currently available port aggregation techniques of these devices in combination with VSS MEC.

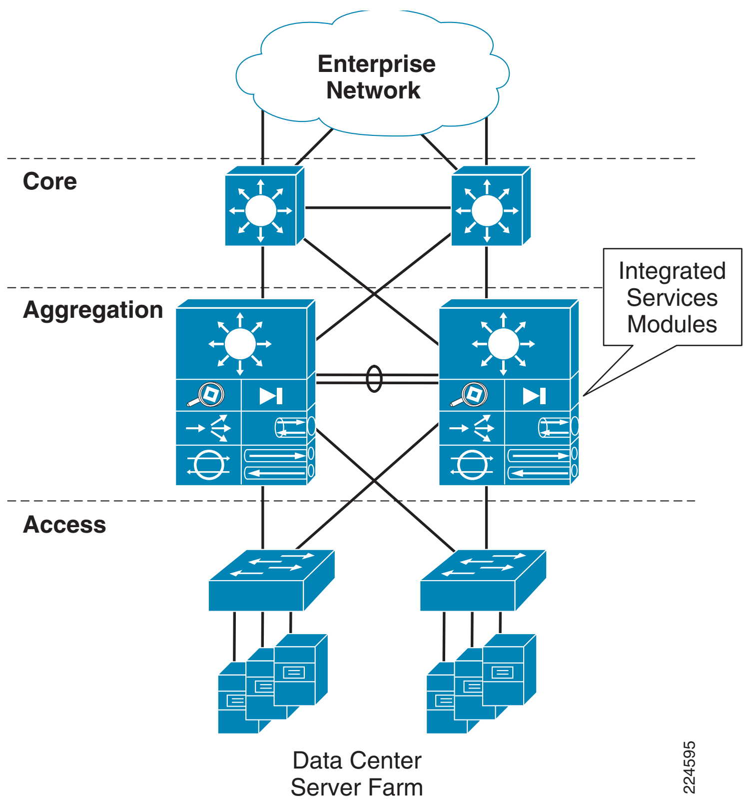
Cisco Data Center Architecture

The following section reviews the current best-practice hierarchical data center designs. These designs form the foundation of the VSS-enabled data center architecture.

Integrated Services Model

The Cisco Catalyst 6500 platform offers the option of integrating service modules directly into card slots within the chassis, thereby conserving valuable rack space, power, and cabling in the data center network. One common design model is to integrate these modules directly into the aggregation-layer switches within the hierarchical network design, as shown in Figure 5. This approach is commonly taken when there are available slots within existing aggregation-layer switches, or chassis slot capacity is planned and allocated to the service modules in the initial design.

Figure 5 Data Center Architecture—Integrated Services Model



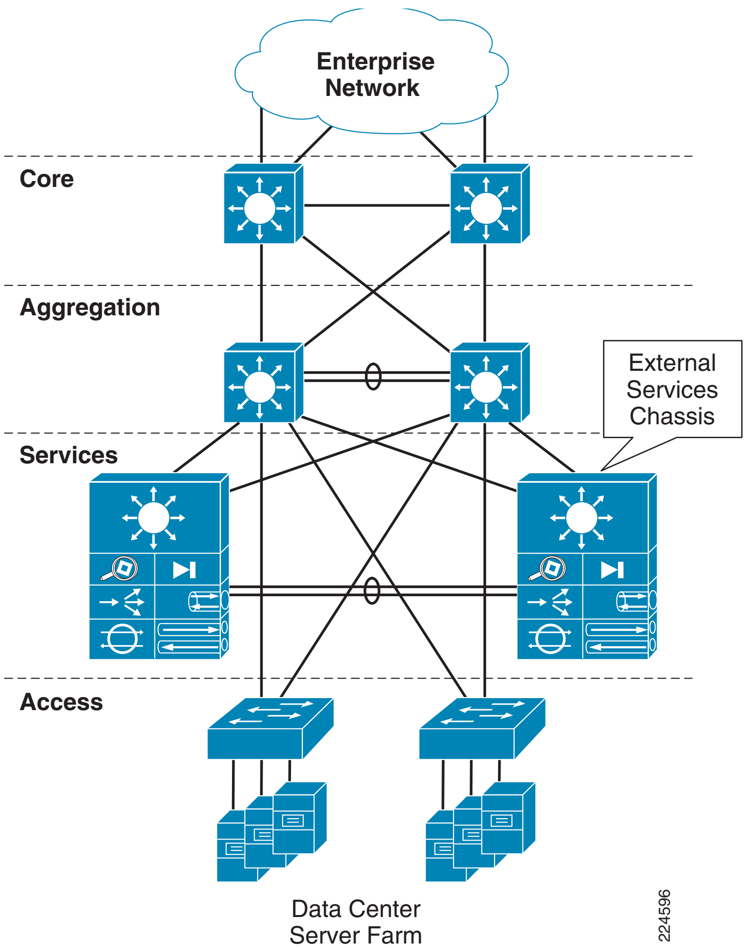
Services Chassis Model

As the data center network grows and needs to scale over time, there can be a requirement to recover the slots that are being consumed by the service modules in order to accommodate greater port density in the aggregation layer. This would allow aggregation of a greater number of access-layer switches without needing to move to a second aggregation block. Other factors might drive the migration away from an integrated services approach, such as the desire to deploy new hardware in the aggregation layer that may not support the Cisco Catalyst 6500 service modules. For example, the Cisco Nexus 7000 Series switches have a different linecard form factor and do not support Cisco Catalyst 6500 service modules.

The initial release of the Cisco Catalyst 6500 VSS 1440 does not support installation of service modules beyond the Network Analysis Module (NAM) in the chassis; this support requires new software that is planned for Cisco IOS Release 12.2(33)SXI.

Since these modules require a Cisco Catalyst 6500 chassis for power and network connectivity, another approach for integrating these devices into the data center network may be considered. One approach is the implementation of an additional pair of Cisco 6500 chassis that are adjacent to the aggregation layer of the data center network. These switches are commonly referred to as *services chassis*. Figure 6 illustrates the physical topology of the services chassis data center design.

Figure 6 Data Center Architecture—External Services Model



Note At the time of writing Cisco IOS 12.2(33)SXI is not available. As a result, all of the testing has been performed leveraging the services chassis model to understand the behavior of VSS in combination with intelligent network services—specifically the Firewall Services Module (FSM) and Application Control Engine (ACE) service module.

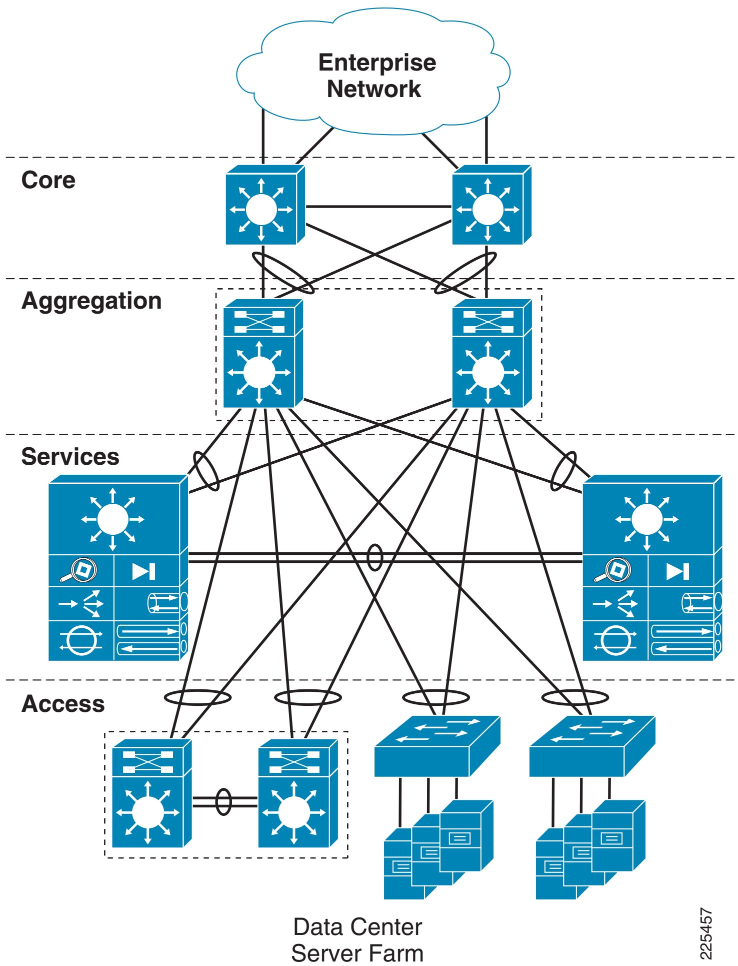
For more information on services chassis design, see the following URL:

- http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/dc_servchas/service-chassis_design.html

Cisco Data Center Architecture Leveraging VSS

This section discusses where VSS might prove beneficial in the enterprise data center. Figure 7 depicts the location of VSS-enabled switches in the tested enterprise data center topology.

Figure 7 Test Bed Topology



Note MEC is leveraged between all entities and the VSS-enabled devices; although it is not a requirement, it is a best practice.

Core Layer

The data center core is traditionally a high-speed, Layer-3 fabric leveraging routing protocols for reliability and rapid convergence. The core employs Equal Cost Multi Path (ECMP) allowing the distribution of load across the data center infrastructure. The data center core at a minimum employs Gigabit Ethernet, but predominately consists of 10 Gigabit Ethernet connections.

When considering the use of VSS at any layer of the data center, you must consider the advantages that VSS provides versus the role you asks it to play. For example, VSS allows you to reduce the complexity of the data center via virtualization. Thus, if you employ VSS at the core, the number of routing instances goes from two distinct devices to one logical entity. This may seem an insignificant reason to migrate to VSS at the core layer, but most benefits of VSS occur at Layer 2.

It is important to remember that current Enhanced IGRP or OSPF deployments in the core of the data center provide a very robust Layer-3 environment that is well-maintained and understood by network administrators. There is little motivation to move away from the traditional core model to VSS for simplicity, manageability, or availability reasons. Therefore, VSS was not enabled at the core layer of the data center for test effort associated with this publication; however, the advantages of VSS at the neighboring aggregation-layer switches is observable at the core with rapid predictable convergence times.

Note There is a growing trend in the data center to support increasingly larger Layer-2 domains for flexible server and application environments. This is a problem VSS can address; however, at this time it is probably best to contain these VLANs within a Layer-3 boundary at the aggregation layer. Extending VLANs through the core is not a best practice at this time and should only be considered an exception to a long existing rule.

Aggregation Layer

The aggregation layer of the data center provides connectivity for the access-layer switches in the server farm and aggregates them into a smaller number of interfaces to be connected into the core layer. In most data center environments, the aggregation layer is the transition point between the purely Layer-3 routed core layer and the Layer-2 switched access layer. 802.1Q trunks extend the server farm VLANs between access and aggregation layers.

As shown in Figure 7, the aggregation layer also provides a common connection point to insert services into the data flows between clients and servers, or between tiers of servers in a multi-tier application. The services chassis are dual-homed into the aggregation layer with 802.1Q trunks similar to the way that access-layer switches are connected.

The aggregation layer is an ideal location to introduce VSS. VSS logically creates a single logical switch from two physical switches. The Services Switches are logically homed to a single aggregation layer switch, but physically dual-homed to the VSS-enabled switches using MEC. Redundant physical paths are maintained while removing the dependence on redundancy protocols such as STP and FHRP. VSS simplifies the Layer-2 and Layer-3 topology of the data center.

Note It is always recommended to enable spanning tree when redundant physical paths exist in the network. It must be understood that, in a VSS environment, the services of spanning tree are not leveraged—but should be enabled.

Figure 7 highlights another best practice when leveraging VSS: Dual-homing of all devices to the VSS, including access, service and core switches. Dual-homing allows the use of MEC throughout the data center, allowing traffic to flow optimally within the data center and to converge at the port level. Orphaned, or single-homed devices, force traffic over the VSL link between VSS switches. This is possible, but proper VSL capacity planning must be taken into account. The VSL link may be comprised of up to eight physical 10 Gigabit Ethernet links. This would appear more than ample, but determining the requirements for the applications residing in each data enter must be taken. Single-homed devices might become "isolated" from the network if the VSS aggregation-layer switch it leverages fails. It is a best practice to physically dual-home all devices to a VSS environment, avoiding the potential of device isolation.

Services Layer

The services layer employs intelligent integrated service modules. At the time of testing, the VSS system officially supported one service module—the Network Analysis Module (NAM). This current product limitation prevents the use of the Application Control Engine (ACE) and Firewall Services Modules (FWSM) from the services-layer test environment. These application and security services are considered critical for today's enterprise data center and therefore preclude the use of VSS at the services layer. The services layer referenced in this document does not employ VSS, but does benefit from the use of VSS at the aggregation layer of the data center.

For more information on services chassis design please go to the following URL:

- http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/dc_servchas/service-chassis_design.html

Access Layer

The access layer of the data center provides entry to the compute power of the data center, namely the server farms. The access layer offers port density, availability and scalability to end nodes. It also presents an opportunity for network administrators to leverage VSS technology—making the most of the highly available and robust forwarding fabric VSS provides. A VSS access layer permits:

- Full forwarding fabric via MEC uplinks to the aggregation layer
- Simplified Layer-2 topology via VSS virtualization
- Full forwarding fabric to servers leveraging MEC enabled edge ports

The remainder of this document focuses on the implementation of VSS in the enterprise data center—specifically the aggregation and access layers. It details the changes required in the neighboring network layer devices to realize the advantages of VSS in the data center.

Required Components

The hardware and software components listed in Table 1 were used in the construction of the validated design models addressed in this publication.

Table 1 Hardware and Software Components

Design Components	Platforms, Line Cards, End Points within Role	Releases
Core Router/Switch	Cisco Catalyst 6500 Series WS-X6724-SFP WS-X6704-10GE VS-S720-10G	12.2(33)SXH2a

Table 1 Hardware and Software Components (continued)

Aggregation Router /Switch	Cisco Catalyst 6500 Series VS-S720-10G WS-X6748-GE-TX WS-X6704-10GE WS-X6708-10GE WS-SVC-NAM-2	12.2(33)SXH2a 3.6(1a)
Services Layer Switch	Cisco Catalyst 6500 Series VS-S720-10G WS-X6704-10GE WS-SVC-NAM-2 WS-SVC-FWM-1 ACE10-6500-K9	12.2(33)SXH2a 3.6(1a) 4.0(1) A2(1.1)
Access Layer Switch	Cisco Catalyst 6500 Series VS-S720-10G WS-X6704-10GE WS-X6748-GE-TX WS-SVC-NAM-2	12.2(33)SXH2a 3.6(1a)
Server Environments	HP ProLiant DL580 G4	Windows 2003 SP2 Red Hat enterprise Linux Server release 5.2 (Tikanga) VMware ESX Server, 3.0.3, 104629

VSS Design in the Aggregation Layer

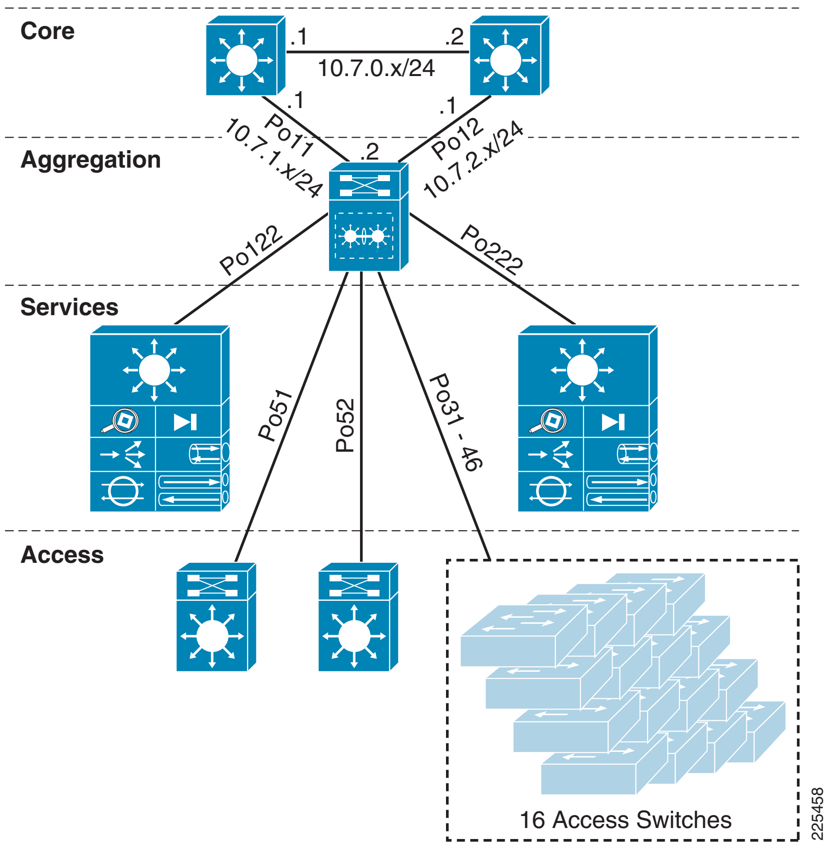
This section addresses VSS design in the aggregation layer by presenting the following sections:

- Infrastructure Description, page 12
- Core and Aggregation Layer Features, page 14

Infrastructure Description

Figure 7 on page 9 illustrates the physical implementation of the tested VSS data center topology. The environment consists of Cisco Catalyst 6500 switching platforms connected via 10 Gigabit Ethernet links. From a physical perspective, this is consistent with the hierarchical data center designs in use today. However, appearances can be deceiving because a VSS-enabled data center introduces a logically streamlined data center environment. Figure 8 is a view of the VSS-enabled data center. As Figure 8 shows, VSS simplifies the aggregation layer by replacing multiple devices with a single VSS identity.

Figure 8 VSS Logical Test Topology



The aggregation layer of the data center is typically defined as the dividing line between Layer-2 and Layer-3 services—being both a routing entity and the primary root for most (if not all) VLANs in the data center. VSS can simplify this deployment.

At the aggregation layer, VSS allows the use of MEC between itself and all neighboring devices. As shown in Figure 8, a Layer-3 MEC, port channel, exists between the core and aggregation layers of the data center, thereby simplifying routes into and out of the data center. Perhaps more significant is the removal of logically redundant paths in the data center. There are no Layer-2 loops with which to contend—mitigating the reliance on spanning tree for path resolution. All links are forwarding, non-blocking from a spanning tree perspective. The advantages of using VSS at the aggregation layer extend to the Layer-2 devices it supports at the access and services layers.

Core and Aggregation Layer Features

Core Layer

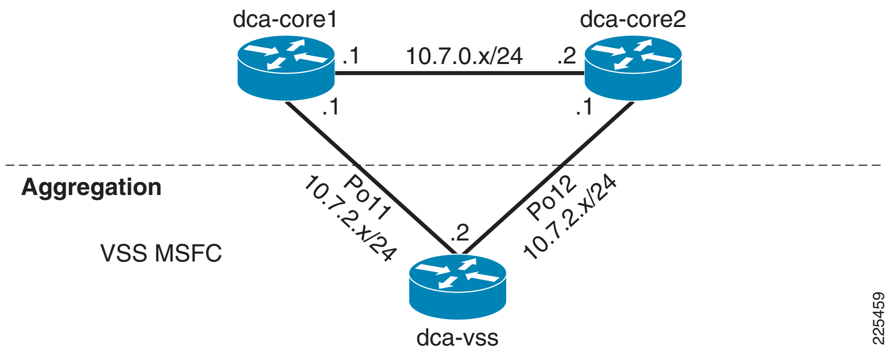
Overview

The core layer consists of two Cisco Catalyst 6500s leveraging Sup720's with 10 Gigabit Ethernet. The core is defined strictly as a Layer-3 domain that is NSF aware. The VSS at the aggregation layer leverages interchassis NSF/SSO, allowing the standby VSS switch to assume control plane responsibilities in the event of the active VSS switch failing. NSF allows for traffic forwarding without requiring the routing topology to converge.

Figure 9 illustrates the Layer-3 configuration between the core routers and the VSS aggregation router; an Interior Gateway Protocol (IGP) mediates the Layer-3 relationship between tiers. Notice that there is a single routing instance defined at the aggregation layer—the VSS-enabled layer. Layer-3 port channels use MEC to connect the core and aggregation routers.

Using MEC and NSF between the routing devices provides for physical path redundancy without requiring a routing topology convergence when the MEC links are failed or recovered. The objective of NSF is to maintain the flow of traffic over a defined, pre-existing routing topology. MEC complements this capability by providing physical path redundancy, obscuring failure events from the routing protocol. The traffic continues to flow in the event of a link or VSS member failure because of the combined benefits of MEC and NSF. The introduction of VSS at the aggregation layer mitigates the impact of routing topology changes at the core by fundamentally eliminating them.

Figure 9 Layer-3 Topology at the Core and Aggregation Layers



As the combined use of NSF and MEC dampens Layer-3 routing changes, network administrators should not implement aggressive IGP timers—namely HELLO and HOLD-TIME timers. The objective of these timers is to detect and act upon state changes of routing neighbors to maintain valid routes. This is the opposite goal of NSF. We currently recommend maintaining default IGP timers in a VSS-enabled environment, allowing NSF/SSO and MEC to provide a stable Layer-3 environment. Testing of various failures between the core and aggregation layers with default IGP timers resulted in near zero failover times.

Note OSPF, Enhanced IGRP, Border Gateway Protocol (BGP) and Intermediate System-to-Intermediate System (IS-IS) are NSF-aware protocols.

Features

The following section describes the configuration of the core switches when neighboring with a VSS-enabled aggregation layer.

Enhanced IGRP Configuration

The use of MEC requires creation of a logical port, or port channel, to represent the aggregated ports on the platform. The port channel may be created leveraging either a statically or dynamically leveraging a link aggregation protocol such as LACP or PAgP.

In the following example based on Figure 9, two 10 Gigabit Ethernet ports are logically combined to form port channel "11" on dca-core1. This port channel uses PAgP to negotiate membership across two VSS member switches. The use of PAgP "desirable" mode means that the interface will be actively seeking membership in a channel.

```
interface TenGigabitEthernet4/1
 description <to VSS Switch 1 >
 no ip address
 channel-protocol pagp
 channel-group 11 mode desirable
!
interface TenGigabitEthernet4/2
 description <to VSS Switch 2>
 no ip address
 channel-protocol pagp
 channel-group 11 mode desirable
!
```

The port channel interface is dynamically created when declared on the first physical interface configured as a member. At this point it is a Layer-2 interface until an IP address has been associated with it. Below port channel "11" is defined as a Layer-3 interface participating in Enhanced IGRP.

Note The use of PAgP is not required, but is desirable if you intend to use Enhanced PAgP to be leveraged as a dual-active detection mechanism for VSS. See the Aggregation Layer, page 16 for more details.

```
interface Port-channel11
 description <<** to VSS **>>
 ip address 10.7.1.1 255.255.255.0
 ip authentication mode eigrp 7 md5
 ip authentication key-chain eigrp 7 eigrp
!
```

The Enhanced IGRP routing instance on the dca-core1 router has the following configuration:

```
router eigrp 7
 network 10.0.0.0
 no auto-summary
 eigrp router-id 1.1.1.1
 nsf
!
```

Note There are no custom timers in the configuration and that the router is NSF aware.

OSPF Configuration

The use of MEC to the core layer allows simplification of the Layer-3 infrastructure. Dynamic PAgP or LACP aggregation allows for individual links in the MEC to fail without requiring a routing topology change; therefore, the timers are set to their defaults to dampen their on the network. Allowing the resiliency of NSF/SSO and MEC to provide a highly available Layer-3 environment.

The following configurations were used in test configurations created for this publication:

```
interface Port-channel11
 description <<** to VSS **>>
 ip address 10.7.1.1 255.255.255.0
 ip pim sparse-mode
 ip ospf authentication message-digest
 ip ospf message-digest-key 1 md5 c1sc0
 ip igmp version 3
!
router ospf 7
 router-id 1.1.1.1
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 0 authentication message-digest
 passive-interface default
 no passive-interface Port-channel11
 network 10.7.0.0 0.0.63.255 area 0
```

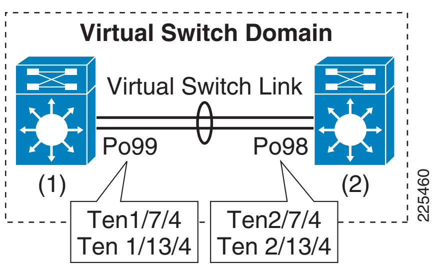
Aggregation Layer

Overview

The aggregation layer is an ideal location to deploy a VSS. As a demarcation point for Layer-2 and Layer-3 services, VSS can readily simplify the logical topology of the data center as shown previously in Figure 8. This section will focus on the implementation of VSS in the data center aggregation layer.

Figure 10 represents the aggregation layer VSS used during testing. The image shows two Cisco Catalyst 6500s with VSS supervisors connected via four 10 Gigabit Ethernet links to form the VSL.

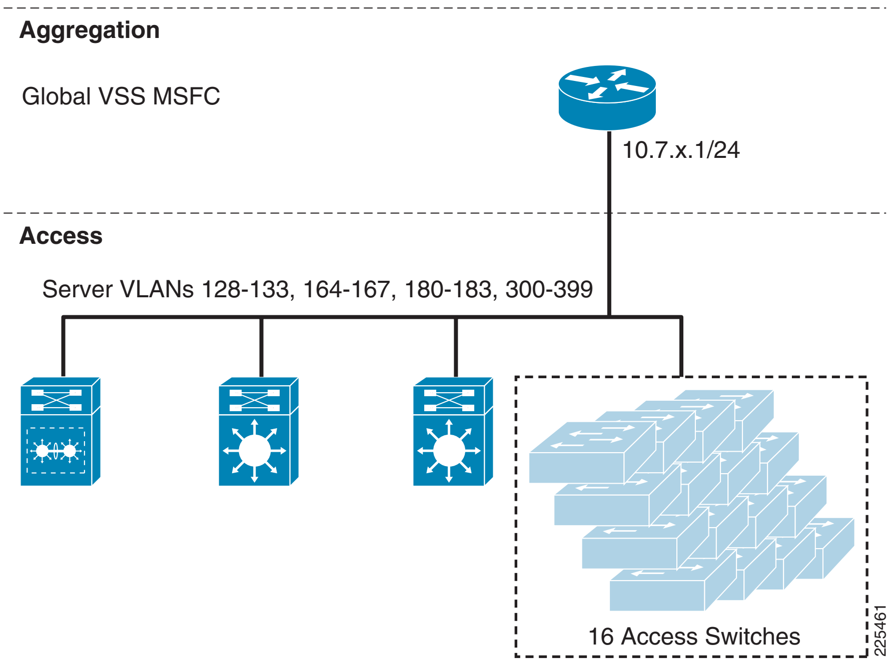
Figure 10 VSS Bidirectional Forwarding Detection (BFD) Example



The VSL is an EtherChannel with interfaces distributed between line cards to provide maximum resilience. Notice that each member of the VSS system uses a unique port channel defined on its local interfaces, the two port channels are referenced as virtual switch link resources to form the VSL. When configuring any EtherChannel it is a best practice to utilize a number of links to the power of two (2, 4, 8) to optimize the channel hashing algorithms. A maximum of eight interfaces may be aggregated to construct the VSL.

Figure 11 is one example of the Layer-2 and Layer-3 topologies tested. In this sample, the VSS routing instances acts as the default gateway to all the server subnets and root for all the associated VLANs. Notice there are no loops introduced using MEC and there is a single routing instance to manage. This model reflects a traditional Layer-2 and Layer-3 data center aggregation and access deployment.

Figure 11 Layer-2 and -3 View of Aggregation Layer



Note The Services Chassis Active/Active, page 30 describes intelligent network services and VRFs within the VSS-enabled aggregation layer.

Features

This section provides guidance on the implementation of VSS in the aggregation layer of the data center. The following features are highlighted:

- Virtualization
- Layer 2
- Layer 3

Virtualization

The first feature of this new data center model that must be addressed is the *virtualization* occurring at the network aggregation layer within the data center.

Note Typically, the introduction of VSS into an existing data center is a migration process. This paper will not document the process of migration. To review VSS migration strategies, refer to the following URL: http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps9336/prod_white_paper0900aecd806ee2ed.html

Configuring Virtual Switching System (VSS)

A VSS environment consists of two member switches. The switches obtain virtual switch IDs during the VSS conversion process. These VSS-enabled switches may then join a larger construct called a *switching domain*. The switching domain forms the boundaries of the virtual switching functionality of VSS. The virtual switch domain configuration used during testing follows.

```
switch virtual domain 100
 switch mode virtual
 switch 1 priority 110
 switch 2 priority 100
 mac-address use-virtual
!
```

Notice that two VSS-enabled switches are referenced by ID and assigned a priority. VSS priority implies that one switch is favored as the “active” VSS member. The default priority value is 100. The higher priority value assumes the active role during role resolution events—such system startup or if preemption is enabled. It is not recommended to use preemption in the data center. Typically, the VSS member switches will leverage identical line card configurations and reside in a dual-homed environment where favoring one system over another is irrelevant.

The **mac-address use-virtual** command permits a virtual media access control (MAC) address to be used by the system. The virtual MAC address command accesses a pool of reserved addresses that incorporate the virtual switch domain. Without this command, the VSS will use the MAC address of the active VSS member at boot time as determined by the active supervisors EEPROM; the standby VSS member will inherit this address upon failure or reboot of the one active switch.

The two members of the VSS domain communicate using the VSL. The VSL is formed using two EtherChannels. In the following example, port channels 98 and 99 are assigned to the VSL via the **switch virtual link** command. The QoS policies are assigned automatically to the VSL link which prioritizes VSS traffic.

```
interface Port-channel98
 description <<**  Etherchannel to Agg2 **>>
 no switchport
 no ip address
switch virtual link 2
 mls qos trust cos
 no mls qos channel-consistency
!
interface Port-channel99
 description <<** Etherchannel to Agg1 **>>
 no switchport
 no ip address
 switch virtual link 1
 mls qos trust cos
 no mls qos channel-consistency
!
```

The physical interfaces leveraged by each port channel reside on one of the two VSS member chassis. The **show etherchannel summary** command confirms that VSS member switch “2” possesses the physical ports comprising port channel 98 and that switch “1” provides the interface resources for port channel 99.

```
dca-vss# show etherchannel summary
…
Group  Port-channel  Protocol    Ports
------+-------------+-----------+-----------------------------------------------
98     Po98(RU)         -        Te2/7/4(P)     Te2/13/4(P)
99     Po99(RU)         -        Te1/7/4(P)     Te1/13/4(P)
```

Note The VSS system interface references include the VSS switch ID as the first identifier for all interfaces. In the preceding example, Te2/7/4 describes the 10 Gigabit Ethernet interface on VSS member switch 2, slot 7, port 4.

The interfaces are configured as statically configured EtherChannels using the default load balancing hash algorithm result of the source and destination IP address. An example configuration follows.

```
interface TenGigabitEthernet1/7/4
 description <<** to DCA-Agg1 **>>
 no switchport
 no ip address
 mls qos trust cos
 channel-group 99 mode on
!
```

The VSL is the medium allowing a single switch supervisor to control two physically independent switching platforms. However, the link is not limited to only VSS system messaging. The VSL may also provide a transport for data traffic in the following circumstances:

- Layer-2 traffic flooded over a VLAN
- Packets requiring software processing by the active supervisor engine where the ingress interface is on the standby chassis
- The packet destination is on the peer chassis, such as the following examples:
  - Traffic within a VLAN in which the known destination interface is on the peer chassis.
  - Traffic replicated for a multicast group when the multicast receivers are on the peer chassis.
  - The known unicast destination MAC address is on the peer chassis.
  - The packet is a MAC notification frame destined for a port on the peer chassis.
  - The downstream local MEC links fail.

In most environments, the use of the VSL link for some data traffic is unavoidable. The network administrator must provide enough capacity to operate under normal and failure conditions. The VSS switch members will always prefer a local forwarding path and it is highly recommended to dual-home all entities to the VSS system via MEC to reduce the presence of data traffic on the VSL. To verify the VSL use the **show switch virtual link** command. In the example the follows, interface Te2/7/4 supports the control traffic between VSS peers.

```
# show switch virtual link
VSL Status : UP
VSL Uptime : 1 day, 44 minutes
VSL SCP Ping : Pass
VSL ICC Ping : Pass
VSL Control Link : Te2/7/4
```

To verify the VSS system configuration, use the **show switch virtual role** command:

```
# show switch virtual role
Switch  Switch Status  Preempt    Priority  Role     Session ID
        Number         Oper(Conf) Oper(Conf)         Local  Remote
------------------------------------------------------------------
LOCAL    2     UP      FALSE(N )   2  (2  )  ACTIVE   0      0
REMOTE   1     UP      FALSE(N )   1  (1  )  STANDBY  2221   6924
```

Notice that the status for the “Local” and “Remote” switch is “UP” and that preempt is not enabled.

VSS High Availability

VSS improves the accessibility of data center resources by being a highly available entity itself. As previously discussed, SSO is a requirement for each peer in the VSS configuration. SSO allows the standby switch to readily assume responsibility for the switching fabric. This feature is enabled via the following configuration statements:

```
redundancy
 mode sso
 auto-sync running-config
!
```

There are three primary failure events to contend with in a VSS deployment:

- Failure of the active VSS switch
- Failure of the standby VSS switch
- Failure of the VSL

The failure of the active VSS peer switch is managed via SSO. The VSL link allows state information to be passed from the active to the hot standby peer supervisor. Upon detection of the active peer failing, the standby VSS assumes the active switching role. There is no spanning tree convergence or routing topology change. See the "Layer 2" section on page 22 and "Layer 3" section on page 25 for details. During testing, the newly active VSS peer assumes control of the aggregation layer switching fabric in less than one second.

The second failure scenario, in which the standby VSS switch fails, is essentially a non-event. There is no convergence as the active VSS peer continues to operate normally. The data center aggregation layer loses capacity, but the forwarding fabric for all dual-homed devices is available.

Note Devices single-homed to the aggregation layer are at risk of isolation; dual-homing to any VSS device is recommended whenever possible.

The final failure scenario involves the loss of the VSL between the VSS peers. As noted earlier, the VSL connection is vital to the operation of the virtual system. Failure of this link without proper detection mechanisms would result in a dual-active scenario in which both VSS member switches assume an active posture creating an unstable Layer-2/-3 environment.

There are two dual-active mitigation techniques:

- Enhanced PAgP (EPAgP)
- Bidirectional Forwarding Detection (BFD)

These are discussed briefly in the sections that follow.

Enhanced PAgP

EPAgP leverages MEC as a medium to detect an active/active VSS environment. EPAgP introduces a new type length value (TLV) to the PAgP messages shared between the switches forming the MEC. The VSS-enabled switches place the ID of the active switch in this field. If the TLV value received differs from that of the active switch, the active switch will enter recovery mode. By default, recovery mode on a VSS switch means all interfaces except for the VSL links will be disabled. EPAgP is the recommended method to address potential dual-active conditions.

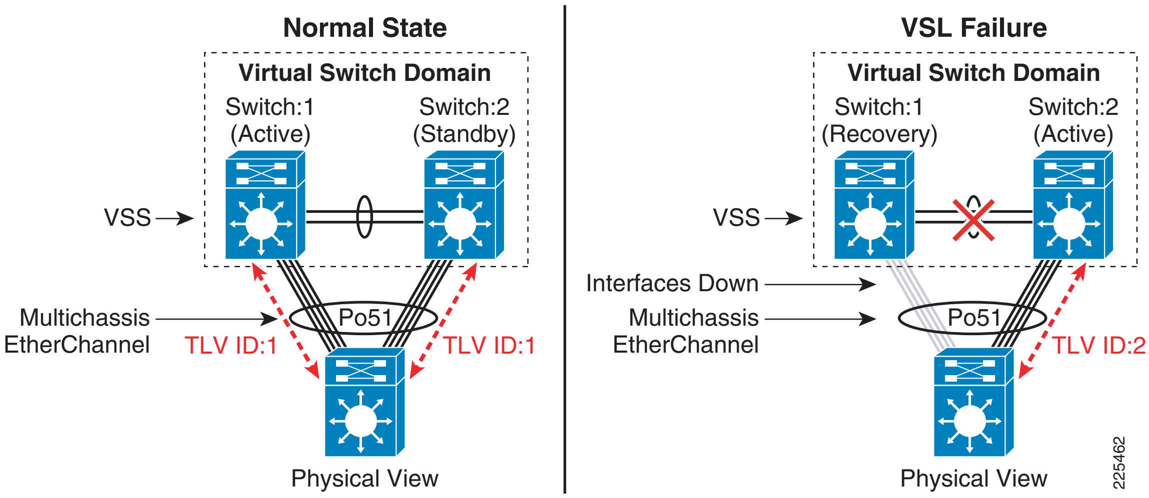
Note EPAgP support is available on the Cisco Catalyst 6500 platforms using IOS image 12.2(33)SXH, it is important to verify EPAgP support on the non-VSS switch being leveraged for dual-active detection.

Figure 12 illustrates an EPAgP implementation where a third Cisco switch supporting EPAgP provides a means for dual-active detection. In this example a MEC, port channel 51, exists between the VSS aggregation switches and an access layer switch. The TLV associated with VSS member switch "1" will be the active ID passed under normal operating conditions. The following sample configuration was used during testing indicating that the trusted MEC is port channel 51.

```
switch virtual domain 100
 dual-active detection pagp trust channel-group 51
```

Upon failure of the VSL, VSS switch 2 assumes the active role and immediately advertises its TLV value as it considers itself the active VSS switch. When the new TLV value reaches the VSS switch 1, switch 1 identifies that a dual-active condition exists and immediately brings down all of its interfaces entering a recovery mode. This aggressive action by the previously active switch 1 provides network stability by simply removing itself from the active/active network topology. Testing indicated convergence times in the 400-500 millisecond ranges.

Figure 12 Dual-Active Mitigation: EPAgP Example



The **show switch virtual dual-active** command provides further insight into the dual-active detection configuration. Note in the example below BFD is not enabled and no interfaces are excluded from the "shutdown" procedure taken by the VSS switch when a dual active condition has been detected. Typically, management interfaces may be excluded from this process to allow access to the VSS platform. The received and expected fields indicate the different TLV values that initiated the dual-active recovery mode.

```
# show switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: No
No interfaces excluded from shutdown in recovery mode
In dual-active recovery mode: Yes
  Triggered by: PAgP detection
  Triggered on interface: Te2/13/1
  Received id: 000d.662e.7d40
  Expected id: 000d.662e.7840
```

Note To manage the VSS system, the test team leveraged console access and MEC to a dedicated out-of-band (OOB) management network that was not excluded from the dual-active recovery mode.

Bidirectional Forwarding Detection

BFD is a protocol written to verify the connectivity between two devices—in this case, two switches in a VSS domain. BFD dual-active detection requires a dedicated interface on each VSS member switch. These interfaces are not active unless the VSL link fails. An example configuration follows.

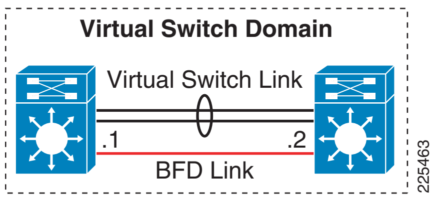
```
interface GigabitEthernet1/5/1
 no switchport
 ip address 10.7.230.1 255.255.255.0
 bfd interval 100 min_rx 100 multiplier 50
!
```

```
interface GigabitEthernet2/5/1
 no switchport
 ip address 10.7.230.2 255.255.255.0
 bfd interval 100 min_rx 100 multiplier 50
!
```

Notice that each interface resides on one of the physical chassis. Figure 13 depicts the physical connectivity of the VSL and BFD links. The VSS domain defines the use of BFD dual-active detection. As the following configuration shows, the Gigabit Ethernet interfaces 1/5/1 and 2/5/1 are trusted by the VSS system for VSL failure detection:

```
switch virtual domain 100
 dual-active pair interface GigabitEthernet1/5/1 interface GigabitEthernet2/5/1 bfd
!
```

Figure 13 VSS BFD Example



BFD is not recommended for dual-active detection unless the EPAgP-based method is unavailable. BFD detection does not occur until the VSL link is declared "down" and the BFD interfaces are enabled. This results in the VSS environment being in an "active/active" state for a longer period of time depending on the BFD timer settings.

Note A new detection mechanism, fast-hellos, are being introduced in Cisco IOS Release 12.2(33)SXI, which offers another choice and improved dual-active detection.

Layer 2

The introduction of VSS at the data center aggregation layer simplifies the Layer-2 network topology. Traditionally, redundant paths or "loops" in the data center have been leveraged to provide a more resilient and available network design. These redundancy benefits come with some risk because network administrators must rely on a spanning tree protocol to logically remove the loops at Layer 2 by blocking. Data center architects often remove loops from their topologies or reduced the size of their Layer-2 domains to remove or mitigate the possibility of forwarding loops in the network. Each of these methods address the problem introduced by loops, but VSS provides another option to improve the availability and scalability of the switching fabric by virtualizing the Layer-2 fabric under the control of one active switching supervisor.

Multichassis EtherChannel

Multichassis EtherChannel Configuration

MEC allows a single EtherChannel to extend beyond a single switch chassis to two VSS switch members. This concept was introduced earlier in the "Multichassis EtherChannel (MEC)" section on page 5. MEC simplifies the traffic patterns in the data center (as Figure 8 illustrates) from a Layer-2 perspective. Despite physically redundant paths, there are no Layer-2 forwarding loops. Cisco MEC is designed to limit VSL link utilization by preferring local MEC interfaces.

Note A maximum of 128 port channels are supported in Cisco IOS 12.2(33)SXH.This number will be increased to 512 for Cisco IOS 12.2(33)SXI.

VSS MEC functionality supports both dynamic and static port aggregation techniques. Dynamic formation of EtherChannels via PAgP or LACP is achieved using a common device identifier between the VSS peers and the remote switch. The remote non-VSS switch is unaware that two distinct switching platforms are truly being leveraged in the EtherChannel.

Cisco devices allocate traffic across members of an EtherChannel bundle using a hash distribution mechanism. Cisco IOS 12.2(33)SXH and later for the Cisco Catalyst 6500 supports an alternative hash-distribution algorithm called the *adaptive algorithm*. Use of the adaptive algorithm eliminates the reset of the port ASIC on each port in the channel when a single physical port is added to or deleted from the channel. The adaptive algorithm was shown to slightly improve network convergence times during single-port EtherChannel failovers during design validation. The adaptive algorithm may be enabled globally or on a per-interface basis. If using a global configuration, ensure that all connected endpoints support the use of the adaptive algorithm.

The following example configuration highlights the use of a Layer-4 hashing algorithm to maximize load distribution across a MEC formed via PAgP and another via LACP. In each case, the port channel is similarly defined supporting the necessary VLANs each access layer switch requires.

```
port-channel hash-distribution adaptive
port-channel load-balance src-dst-mixed-ip-port
interface TenGigabitEthernet1/13/1
 channel-protocol pagp
 channel-group 51 mode desirable
!
interface TenGigabitEthernet1/13/2
 channel-protocol lacp
 channel-group 52 mode active
!
interface Port-channel51
 description <<** To Access Switch 1 **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 128-133,164-167,180-183,300-399
 switchport mode trunk
interface Port-channel52
 description <<** To Access Switch 2 **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 128-133,164-167,180-183,300-399
 switchport mode trunk
```

Note See Figure 8 on page 13 for an illustration.

Spanning Tree Configuration

The use of VSS removes logical loops from the Layer-2 topology of the data center. As shown on the left in Figure 14, the spanning-tree domain becomes a two-switch topology, despite the physical realities that exist on the right hand side of the figure. The use of a MEC port channel creates an abstraction from a spanning-tree viewpoint; a port channel equates to a single spanning-tree interface. This reduces the number of logical interfaces by half in the data center permitting the environment to scale at Layer 2.

Note Spanning tree should not be disabled in a VSS environment. Spanning tree, though not actively impacting the forwarding paths in the data center, should be allowed to contend with inadvertent looping conditions that might occur—usually due to human error. Spanning tree mode RPVST+ is currently recommended for data center environments.

Figure 14 VSS Layer-2 Simplification Example

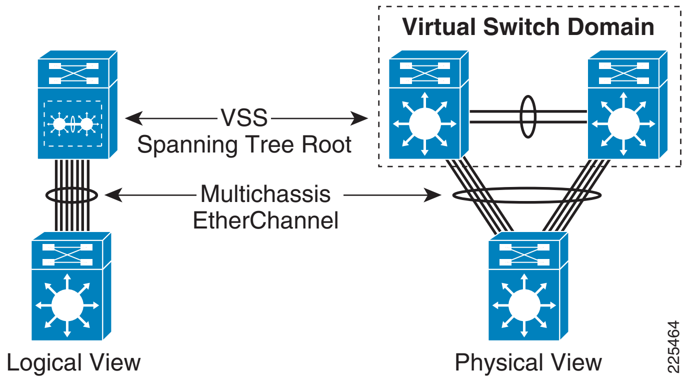


Figure 14 suggests that the VSS-enabled aggregation layer should be the spanning tree root for the VLANs present in the data center. This means that any MEC-attached access layer switches will not process a spanning tree topology change in the event of a single MEC interface failure on its uplink. In fact, the MEC would have to completely fail for the port channel to be declared "down" and spanning tree would then attempt to find an alternate path. The connections are port channels, which abstract the introduction or loss of an interface from the spanning-tree calculations. Since spanning tree is not maintaining the active forwarding topology it is not necessary to implement LoopGuard or RootGuard in a VSS environment. The following spanning tree configurations are recommended on the VSS aggregation layer switches:

```
spanning-tree mode rapid-pvst
spanning-tree extend system-id
spanning-tree pathcost method long
spanning-tree vlan  X-X root primary
```

Note The **root primary configuration** command will set the priority for those VLANs to 24576.

From a spanning-tree perspective, the **show spanning-tree summary totals** command indicates that the 114 VLANs in the test environment access layer are all forwarding. There are no blocking ports to the VSS aggregation layer which allows full utilization of all links between the access and aggregation switches.

```
# show spanning-tree summary totals
Switch is in rapid-pvst mode
Root bridge for: none
EtherChannel misconfig guard is enabled
Extended system ID     is disabled
Portfast Default             is disabled
PortFast BPDU Guard Default  is disabled
Portfast BPDU Filter Default is disabled
Loopguard Default          is disabled
UplinkFast                   is disabled
BackboneFast                 is disabled
Pathcost method used   is long
```

```
Name                  Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
114 vlans                    0         0        0        126        126
```

Layer 3

The aggregation layer is typically the demarcation between Layer-2 and Layer-3 networking services. This same model holds true in a VSS environment except that there is only a single NSF routing instance to manage across the two VSS peers. Combined with the MEC functionality described earlier, the Layer-3 topology of the data center becomes less complex and highly available. This section describes FHRPs, Enhanced IGRP, OSPF, and VRF support.

FHRP

FHRPs such as VRRP and HSRP were designed to allow for a highly available first IP route hop for host systems. FHRPs allow two (or more) distinct routers to share a common IP address providing a redundant Layer-3 default gateway for end nodes. The VSS system creates a single logical router at Layer 3. This VSS routing instance fulfills this first-hop role without the need for a dedicated protocol. The VSS IP route is highly available due to MEC and the resiliency of the VSS system. VSS eliminates the need for FHRP at the aggregation layer of the data center. See Figure 11 on page 17 for a visual reference.

Enhanced IGRP

The VSS system supports Enhanced IGRP. As discussed in the "Core Layer" section on page 9, VSS leverages NSF to mitigate routing topology changes. This means timers and thresholds traditionally implemented to quickly detect route failures should be avoided. As a result of testing, the use of default timers is recommended. Below is a sample configuration of the VSS routing instance at the aggregation layer of the data center. Notice no timers are enabled and only those interfaces with routing neighbors are active.

```
router eigrp 7
 passive-interface default
 no passive-interface Vlan151
 no passive-interface Vlan153
 no passive-interface Vlan161
 no passive-interface Vlan163
 no passive-interface Port-channel11
 no passive-interface Port-channel12
 network 10.0.0.0
 no auto-summary
 !
```

In this example, port channels 11 and 12 are actually MEC-attached to the core-layer routers. Using MEC with NSF allows failures of links in the MEC or one of the VSS member switches without requiring a routing topology change, as NSF will continue to forward packets. Testing revealed improved availability of the routing infrastructure in comparison to the traditional aggregation-to-core layer design (without VSS). Figure 9 on page 14 illustrates the tested Layer-3 topology.

Note The VLAN interfaces in the example form a relationship with a VRF located across integrated network services. More information on this configuration is available in the "Services Chassis Active/Active" section on page 30.

OSPF

The configuration of OSPF on the aggregation layer VSS switch uses NSF and MEC to allow Layer 3 forwarding paths to change without forcing convergence of the routing topology. This implies using default values for the hello and dead timers in a VSS environment. For more information, refer to "Core Layer" section on page 9.

Virtual Routing Forwarding Lite

The VSS system is capable of supporting VRF instances. This capability allows you to design virtual Layer-3 routes within the data center. As noted in the *Data Center Service Integration: Services Chassis Design Guide*, VRF instances may be leveraged to create contained Layer-2 service domains, see the "Services Chassis Active/Active" section on page 30. An example VRF configuration follows.

```
address-family ipv4 vrf servers1
  network 10.0.0.0
  no auto-summary
  autonomous-system 7
  eigrp router-id 5.5.5.2
  nsf
 exit-address-family
interface Vlan163
 mac-address 1234.1234.0163
 ip vrf forwarding servers1
 ip address 10.7.162.7 255.255.255.0
 ip pim sparse-mode
 ip authentication mode eigrp 7 md5
 ip authentication key-chain eigrp 7 eigrp
 ip igmp version 3
end
```

The use of a static MAC address (under the VLAN interface configuration) using the VRF allows traffic to flow smoothly between the transparent service layers. The northern and southern routers in the active/active design required static MAC addresses to provide a Layer-2 path between service and aggregation chassis.

Services Layer

The following section details the implementation of services via a services chassis. The Cisco IOS image used during testing (Cisco IOS 12.2(33)SXH) does not supported integrated services modules besides the NAM. The following services sections leverage the previously documented services chassis designs found at the following URL:

- http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/dc_servchas/service-chassis_design.html

The fundamental difference is the introduction of VSS at the aggregation layer.

Services Chassis Active/Standby

Overview

The active/standby design denotes the use of active and standby services in the data center. Under normal operating conditions, this design results in one of the two services chassis processing the traffic requiring services. This design emphasizes highly available, redundant services over infrastructure utilization.

Physical Topology

Figure 7 on page 9 depicts the physical topology leveraged at the services chassis and aggregation layers. Each of the services chassis are dual-homed to the aggregation layer VSS member switches. In this example, the four connections are all 10 Gigabit Ethernet. The 40 Gbps MEC link between each of the services chassis and aggregation switches supports all data—including fault tolerant and stateful traffic types. This is a significant change from the services chassis design document mentioned earlier that leveraged a dedicated connection between the services chassis for fault tolerant and state traffic.

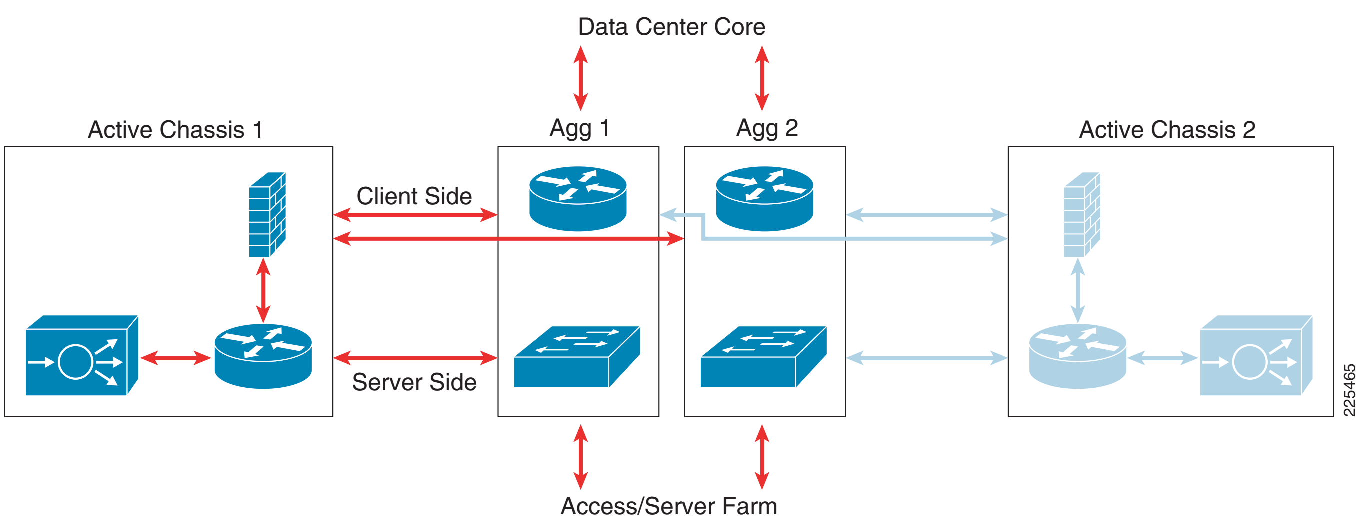
The physical positioning of linecards within the services chassis affects switch recovery. The service modules within the services chassis should be placed in the lower-numbered or upper-level slots of the chassis. The 10 Gigabit Ethernet linecards are positioned in the higher-numbered or lower-level slots in the switch. This arrangement is adopted because these slots are enabled earlier during the services switch initialization allowing the service modules to boot prior to connectivity to the VSS aggregation layer being established. This provides for a smoother recovery of the services switch.

Logical Topology

Traffic Flow

Figure 15 shows the traffic flow in the active/standby services design model. The left side of the figure contains the active services—in this case firewall and load balancing functions. Services on the right side of this model are in standby mode and will only become active if a failure occurs within an equivalent function in the primary path.

Figure 15 Active/Standby Model Traffic Flow



This design model was validated with the following characteristics:

- *Routed FWSM*—A routed service device can be conceptually easier to troubleshoot because there is a one-to-one correlation between VLANs and subnets. In addition, it involves a simplified spanning tree structure since the device is not forwarding Bridge Protocol Data Units (BPDU) between VLANs.
- *One-armed ACE*—The one-armed ACE can be introduced seamlessly into the network and will not be in the path of other traffic that does not need to hit the virtual IP (VIP) addresses. ACE failure or failover only affects traffic that is being load-balanced or that is using other ACE application services—such as secure-socket layer (SSL) acceleration. A traffic-diversion mechanism is required

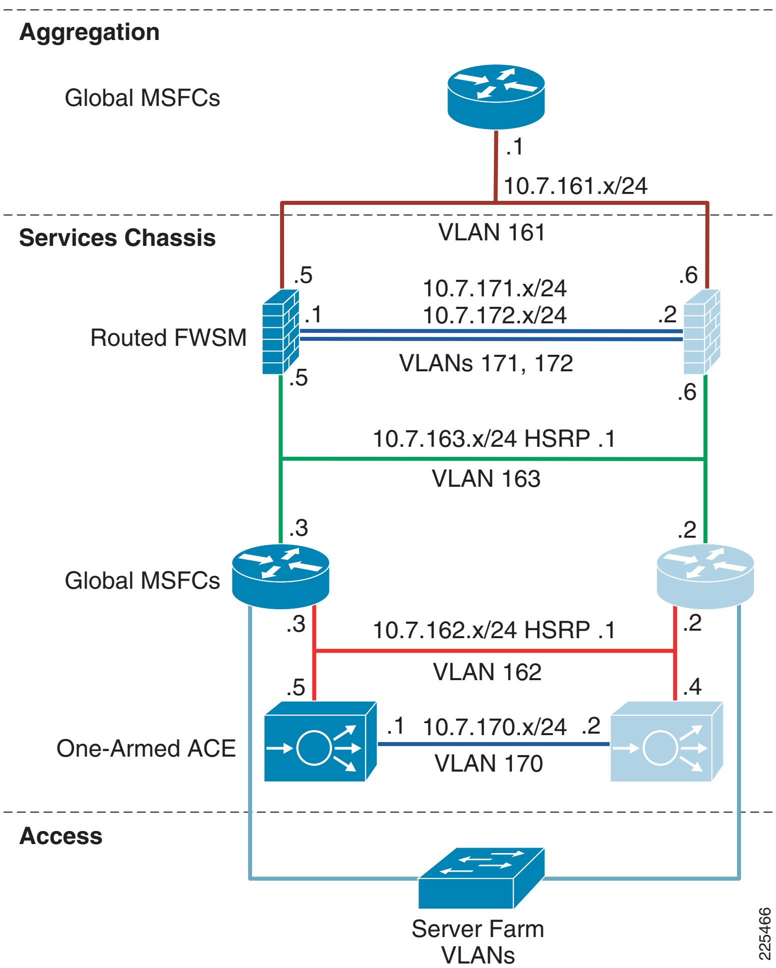
to ensure both sides of an application session pass through the ACE—either Policy-based Routing (PBR) or Source-address Network Address Translation (Source NAT) can be used. Source NAT was chosen for the validation of this design for its ease-of-configuration and support relative to PBR.

- *Services Chassis 6500 Multilayer Switch Feature Card (MSFC) as IP Default Gateway for Server Farm Subnets*—Using the MSFC as default gateway for servers provides for the insertion or removal of services above the MSFC without altering the basic IP configuration of devices in the server farm. It also prevents the need to enable ICMP redirects or have load-balanced traffic traverse the FWSM twice during a session.

Logical Design

The introduction of VSS technology at the aggregation layer simplifies the deployment of a services chassis. As shown in Figure 16, the VSS aggregation layer is the spanning tree root for all VLANs, including the fault tolerant and stateful VLANs leveraged by the services modules. There is no direct connection between each services chassis. The uplink MEC between the services switches and the aggregation layer removes logical loops allowing Layer-2 convergence based on EtherChannel rather than spanning tree.

Figure 16 Active/Standby Logical Design



The services chassis uplink MEC provides a medium for fate sharing as the state of the integrated services modules depends on the state of the channel through autostate. Network administrators should rely on autostate for service availability and consider removing other forms of service tracking such as heartbeats or query interfaces. Autostate messages are triggered when there is a change in Layer-2 forwarding topology. Since the spanning-tree root resides on the VSS, and the connection from the services chassis leverages MEC, autostate messages will only be generated when the MEC completely fails. This means the services chassis is completely isolated and service module failover should occur. Testing showed that the use of MEC and autostate provides faster convergence times for the ACE and FWSM modules within the services chassis versus a traditional deployment.

The following is a brief analysis of the function of each of the VLANs within the logical design (since there are no transparent mode modules in this topology, each VLAN corresponds to a unique IP subnet):

- *VSS Aggregation Global MSFC to routed FWSM*—This is shown as VLAN 161 in Figure 16. This VLAN is extended across the dual-homed physical links (MEC) between the services chassis and aggregation layer. It provides the ingress and egress path for traffic on the client side of the service modules.
- *FWSM Fault Tolerance links*—These are shown as VLAN 171 and 172 in Figure 16 and are extended across the MEC uplinks through the aggregation layer. They carry failover hello packets, state information, and allow the primary and secondary FWSMs to keep their configurations synchronized.
- *Routed FWSM to Services Chassis Global MSFCs*—This is shown as VLAN 163 in Figure 16. This VLAN is extended across the dual-homed physical links between the services chassis and aggregation layer. The services chassis MSFC makes forwarding decisions to direct traffic received on this link directly to the server farm or to the one-armed ACE module if a virtual IP (VIP) address is the destination.
- *Services Chassis Global MSFCs to One-Armed ACE*—This is shown as VLAN 162 in Figure 16. This is both the ingress and egress interface for traffic being serviced by the ACE module. The ACE performs Source NAT, which changes the source address of packets that it is forwarding to the server farm. In this way, the return packets must also pass through the ACE to have their destination addresses translated back to that of the original requesting client node. This VLAN is extended across the dual-homed physical links (MEC) between the services chassis and aggregation layer.
- *ACE Module Fault Tolerance link*—This link is shown as VLAN 170 in Figure 16 and is extended across the aggregation layer via the MEC. This link carries hello traffic and allows configuration synchronization between the two ACE modules.
- *Services Chassis Global MSFCs to Server Farm VLANs*—These VLANs are extended across the dual-homed links to the aggregation layer (MEC), and also extend down into the access layer to support server connectivity. In the reference topology, eight different VLANs carrying different types of serviced traffic—voice, firewalled-only data, and switch-assisted load-balanced (SLB) data—were configured. The actual number and purpose of VLANs deployed will be specific to a customer requirement.

Note Not illustrated in Figure 16 is the possibility of having VLANs that carry non-serviced traffic. For server farm subnets that do not require FWSM or ACE services, a traditional hierarchical design data path may be used with these VLANs terminating on the aggregation layer and their IP default gateway services provided by the VSS aggregation-layer global MSFC.

The following example configuration highlights some of the features leveraged on the services chassis switch. The example implements autostate messaging with the ACE module in slot 5 and the FWSM in slot 4 providing fate sharing with the MEC uplink to the aggregation layer.

```
svclc autostate
svclc multiple-vlan-interfaces
svclc module 5 vlan-group 1,2,152,153,162,163
svclc vlan-group 1  146
svclc vlan-group 2  170
svclc vlan-group 153  153
svclc vlan-group 163  163

firewall autostate
firewall multiple-vlan-interfaces
firewall module 4 vlan-group 1,3,151,152,161,162,
firewall vlan-group 3  171,172
firewall vlan-group 151  151
firewall vlan-group 152  152
firewall vlan-group 161  161
firewall vlan-group 162  162
```

Spanning tree is enabled, but does not impact the forwarding topology as all links to the aggregation layer are forwarding. This is a failsafe mechanism that is also leveraged by autostate.

```
spanning-tree mode rapid-pvst
spanning-tree extend system-id
spanning-tree pathcost method long
```

The previous defined recommendations for the use of the adaptive hash and Layer-4 load balancing across the MEC are enabled.

```
port-channel hash-distribution adaptive
port-channel load-balance src-dst-mixed-ip-port
```

The MEC port channel is defined using PAgP and may also be leveraged for VSL dual-active detection. It should be noted that the services switch MEC may leverage dynamic LACP or static EtherChannel as well.

```
interface Port-channel122
 description <<** to VSS **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 151-153,161-163,170-172
 switchport mode trunk
!
interface TenGigabitEthernet1/4
 description <<** to VSS **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 151-153,161-163,170-172
 switchport mode trunk
 channel-protocol pagp
 channel-group 122 mode desirable
!
```

Note The configuration of the ACE and FWSM service modules were identical to those documented in the *Services Chassis Design Guide* except for the failover tracking mechanisms.

Services Chassis Active/Active

Overview

The active/active services chassis design provides for the utilization of all network and service devices. This design leverages virtualization to optimize the network and its services.

VSS technology virtualizes the aggregation layer simplifying the Layer-2 environment of the data center

ACE and FWSM support multiple virtual contexts. These contexts are deployed in transparent mode, allowing seamless introduction of their services.

The service contexts are not limited to transparent mode. Routed service contexts are supported if this is an application requirement. Virtualization provides service flexibility.

VRF support on the VSS aggregation switches contains the Layer-2 domain creating a transparent services layer in the network when combined with the virtualized network services modules.

Physical Topology

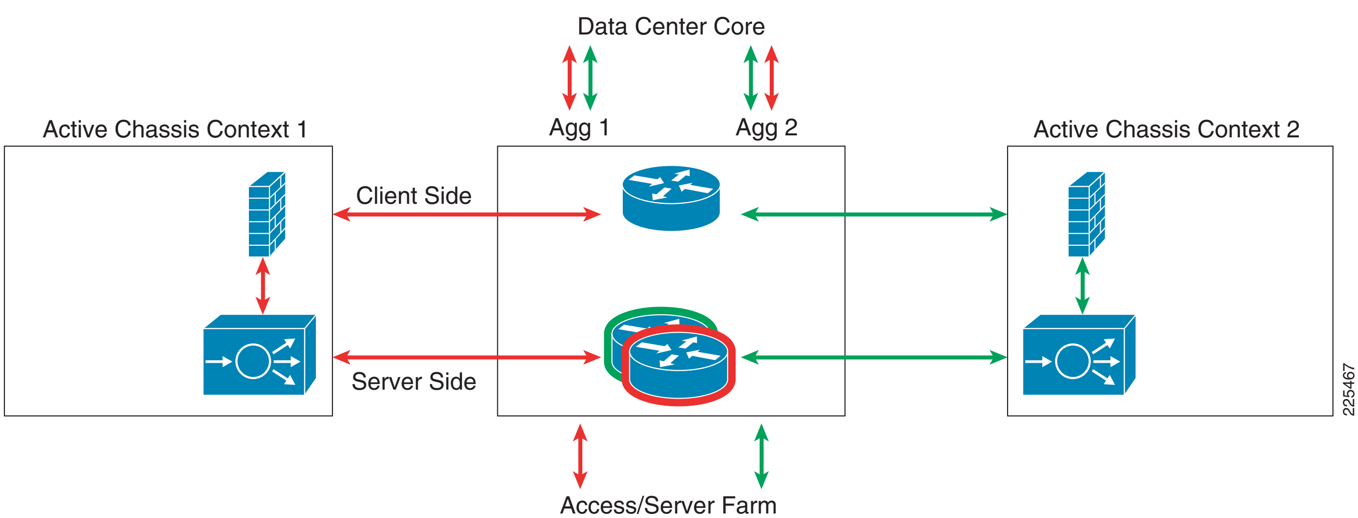
The physical topology of this design is identical to the design described in the "Services Chassis Active/Standby" section on page 26. Please reference that previous section for more details and Figure 7 on page 9.

Logical Topology

Traffic Flow

Figure 17 illustrates the traffic flow of the active/active services chassis design combined with a VSS aggregation layer. In this model, data center ingress and egress traffic flows through the VSS aggregation layer supporting multiple VRFs. Intelligent network services are active on each of the services chassis switches providing load distribution and high availability.

Figure 17 Active/Active Traffic Flow



This design model was validated with the following characteristics:

- *Transparent FWSM*

  A transparent firewall requires less configuration than a routed firewall because there is no routing protocol to configure—or list of static routes to maintain. It requires only a single IP subnet on the bridge-group interface and forwards BPDUs between bridging devices that live on attached segments. In that way, it is truly transparent and not a bridge itself. The VLANs on the different interfaces of the transparent FWSM will carry different VLAN numbers, so a transparent device is often said to be *stitching* or *chaining* VLANs together.

Note The FWSM supports a maximum of eight bridge-group interfaces (BVI) per context.

- *Transparent ACE*

  The transparent ACE implementation works similarly to the FWSM. Multiple VLANs are stitched together to transport one IP subnet and BPDUs are forwarded to allow adjacent switches to perform spanning-tree calculations. Unlike the one-armed ACE approach, a transparent ACE sits inline with traffic and requires no traffic diversion mechanism to ensure that both sides of a protocol exchange pass through the device. The ACE supports a maximum of two Layer-2 interface VLANs per bridge-group and a maximum of two thousand BVIs per system.

- *Dual Active Contexts on the Services Modules*

  With the virtualization capabilities of the Cisco Catalyst 6500 Services Modules, two separate contexts have been created which behave as separate virtual devices. The first FWSM and ACE are primary for the first context and standby for the second context. The second FWSM and ACE are primary for the second context and secondary for the first context. This allows modules in both sides of the design to be primary for a portion of the traffic and allows the network administrator to distribute load across the topology instead of having one set of modules nearly idle in a pure standby role.

Note It is important to note that in an active/active design, network administrators must properly plan for failure events in which one service module supports all of the active contexts. If the total traffic exceeds the capacity of the remaining service module, the potential to lose connections exists.

- *Aggregation Layer VRF Instances as IP Default Gateway for Server Farm Subnets*

  Using VRF instances for the default gateway for servers provides for the insertion or removal of services above the VRF without altering the basic IP configuration of devices in the server farm. It also provides for direct routing between server farm subnets through the aggregation layer without a requirement to drive traffic out to the services chassis for first-hop IP default gateway services. For the active/active design, a separate set of VRF instances was created for each of the two Services Modules contexts in order to keep traffic flows segregated to the proper side of the design.

- *Traffic Flow Between Service Modules and Clients*

  For client/server traffic, ingress and egress traffic on the client side is balanced across the aggregation VSS global MSFC.

- *Traffic Flow Between Service Modules and Server Farm*

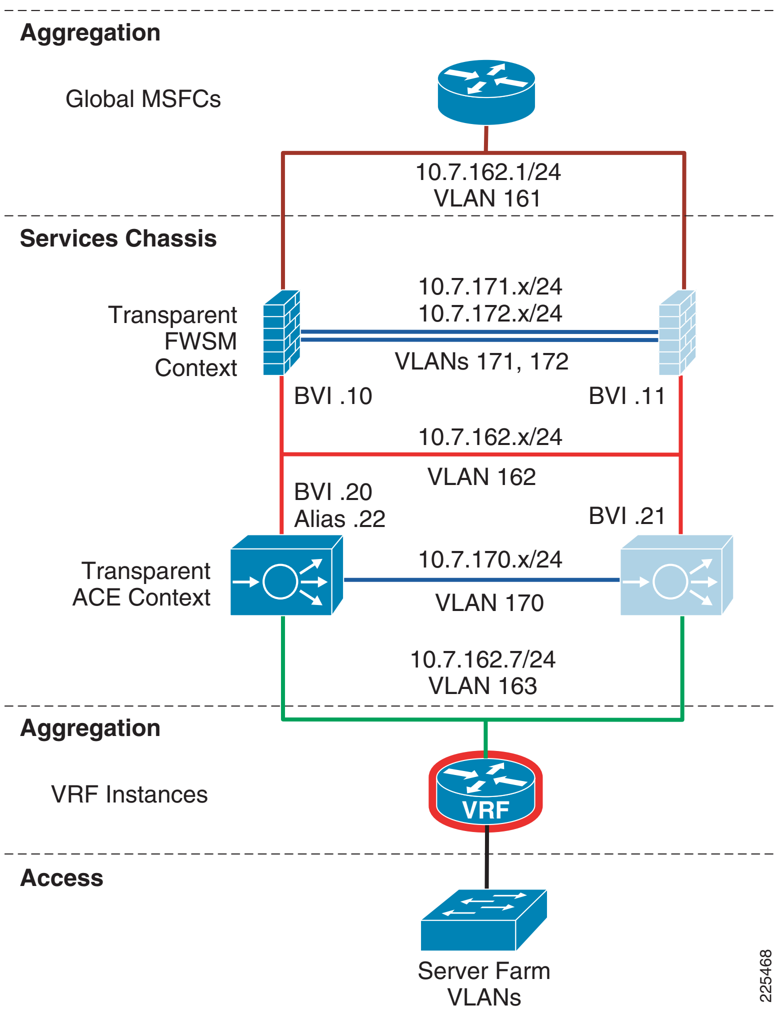
  For client/server traffic, ingress and egress traffic on the server (access layer) side is concentrated to the VSS aggregation-layer switch VRF instance which is configured as the IP default gateway for the server farm subnets.

Logical Design

The VSS aggregation switch simplifies the design at Layer 2. The VSS aggregation-layer switch is the spanning-tree root. The MEC functionality afforded via VSS creates a loopless topology that removes spanning-tree dependencies and provides a full forwarding switching fabric for the services chassis. In addition, the MEC connectivity between the aggregation and services switches provides EtherChannel-based convergence times. With a consistent Layer-2 topology, the integrated service modules in the services chassis may leverage autostate for rapid service convergence.

The VSS aggregation switch provides a simplified Layer-3 forwarding path using its global MSFC and VRFs. Ingress and egress traffic leverage a transparent services layer sandwiched between the VSS Layer-3 instances. Virtualization of Layer-2 and Layer-3 at the aggregation layer optimizes data center services.

Figure 18 Active/Active Services Chassis Logical Model



The following is a brief analysis of the function of each of the VLANs within the logical design. VLANs 161, 162, and 163 represent a single IP subnet because the primary design that was validated for this architecture used transparent mode on both FWSM and ACE contexts.

- *VSS Aggregation Global MSFC to Transparent FWSM*—This is shown as VLAN 161 in Figure 18. This VLAN is extended across the dual-homed physical links between the services chassis and aggregation layer and provides the ingress and egress path for traffic on the client side of the service modules.
- *FWSM Fault Tolerance Links*—These are shown as VLAN 171 and 172 in Figure 18 and are extended across the MEC uplink to the VSS aggregation layer. They carry failover hello packets and state information, and allow the primary and secondary FWSM contexts to keep their configurations synchronized.
- *Transparent FWSM to Transparent ACE Context*—This is shown as VLAN 162 in Figure 18 and is extended across the dual-homed physical links between the services chassis and aggregation layer. The transparent ACE intercepts traffic that is destined for a VIP address and passes other traffic through without altering packets.

- *Transparent ACE Context to Aggregation VRF Instance*—This is shown as VLAN 163 in Figure 18 and is extended across the dual-homed physical links to the aggregation layer. This VLAN carries traffic from the server side of the Services Modules to and from the server farm VLANs by being routed by the aggregation layer VRF instances.
- *ACE Module Fault Tolerance Link*—This link is shown as VLAN 170 in Figure 18 and is extended across the aggregation layer via the MEC uplinks. This link carries hello traffic and allows config synchronization between the two ACE modules.
- *Aggregation Layer VRF Instance to Server Farm VLANS*—These VLANs are south of the aggregation layer VRF in Figure 18. In the active/active services chassis model with VRFs, the server farm VLANs are contained between the aggregation and access layers. They do not need to be extended directly into the services chassis. In the reference topology, eight different VLANs carrying different types of serviced traffic—voice, firewalled-only data, SLB data—were configured. The actual number and purpose of VLANs deployed will be specific to a customer requirement.

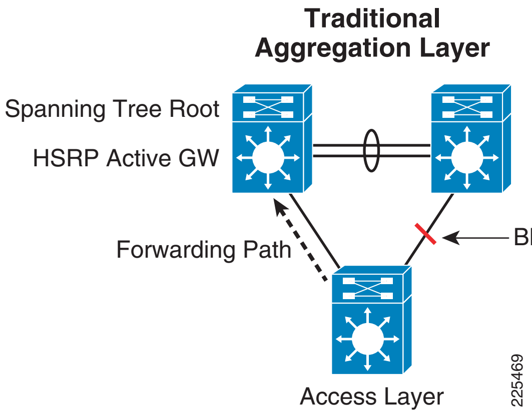
Note Not illustrated in Figure 18 is the possibility of having VLANs that carry non-serviced traffic. For server farm subnets that do not require FWSM or ACE services, a traditional hierarchical design data path may be used with these VLANs terminating on the aggregation layer and their IP default gateway services provided by the VSS aggregation layer global MSFC.

Note The configuration of the services chassis is identical to what is documented in the "Logical Design" section on page 28 (in the "Services Chassis Active/Standby" portion of this publication). The configurations for the ACE and FWSM service modules were identical to those documented in the *Services Chassis Design Guide* except for the failover tracking mechanisms no longer necessary.

Access Layer

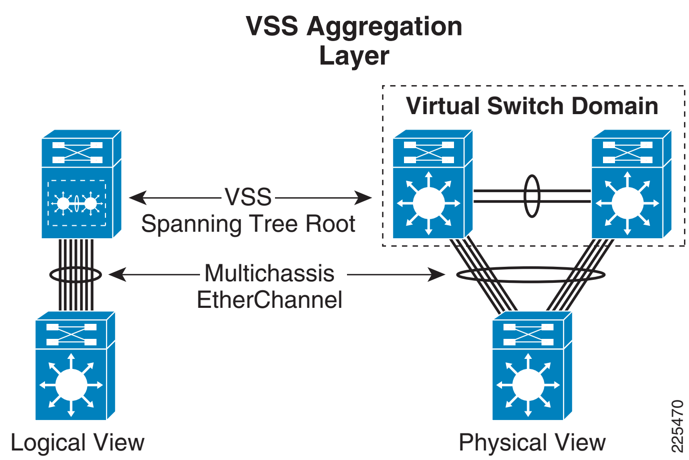
The primary function of the access layer is to provide port density for the server end nodes in the data center. The introduction of VSS at the aggregation layer does not alter this role, but enhances the forwarding capability of the traditional access layer design. Figure 19 depicts the traditional looped "triangle" design where spanning tree accounts for physical loops in the network by blocking one of the uplink paths to the aggregation layer. This model is well documented and highly leveraged in today's data centers; however, half of the uplink capacity at the access layer remains idle. To optimize uplink utilization, network administrators will often split the spanning tree root definitions and the active HSRP gateway between the aggregation switches—which improves performance at the cost of complexity. The traditional access-layer design is dependent on spanning tree to provide a reliable and predictable network topology and rapid convergence in the event of a failure.

Figure 19 Traditional Triangle Access Layer Design



The use of a VSS aggregation layer offers improved uplink utilization with less complexity. As shown in Figure 20, the VSS aggregation-layer design leveraging MEC allows all uplinks to be forwarding. There are no Layer-2 loops because it is logically a direct link between two switches. All of the ports in the MEC port channel are forwarding. Traffic in this design is then load balanced via the EtherChannel hashing algorithm chosen—optimizing uplink utilization. The VSS design removes complexity and relies on EtherChannel rather than spanning tree to provide rapid convergence.

Figure 20 FVSS aggregation layer with traditional Access Layer



The access layer switch used during testing leveraged the following key configurations:

```
port-channel hash-distribution adaptive
port-channel load-balance src-dst-mixed-ip-port
```

The use of the adaptive hash provides a more resilient EtherChannel as described earlier in the document under the "Multichassis EtherChannel Configuration" section on page 22. The use of a Layer 4-based load balancing algorithm is a best-practice recommendation to achieve an even traffic distribution across the available interfaces in the MEC.

```
spanning-tree mode rapid-pvst
spanning-tree pathcost method long
```

Spanning tree is active, but not affecting the forwarding path in the access layer. It is a recommended best practice to always enable spanning tree when redundant physical paths in the data center exist.

```
interface Port-channel51
 description <<**  MEC to VSS aggregation layer Switch **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 128-133,164-167,180-183,300-399
 switchport mode trunk
!
```

The MEC uses the preceding port channel configuration providing a link to the VSS aggregation layer. With the preceding configurations, the forwarding topology from a Layer-2 perspective has been optimized as this spanning tree topology confirms.

```
# show spanning-tree summary totals
Switch is in rapid-pvst mode
…
Name                   BlockingListening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
113 vlans              0          0         0        125        125
```

VSS Design in the Access Layer

Infrastructure Description

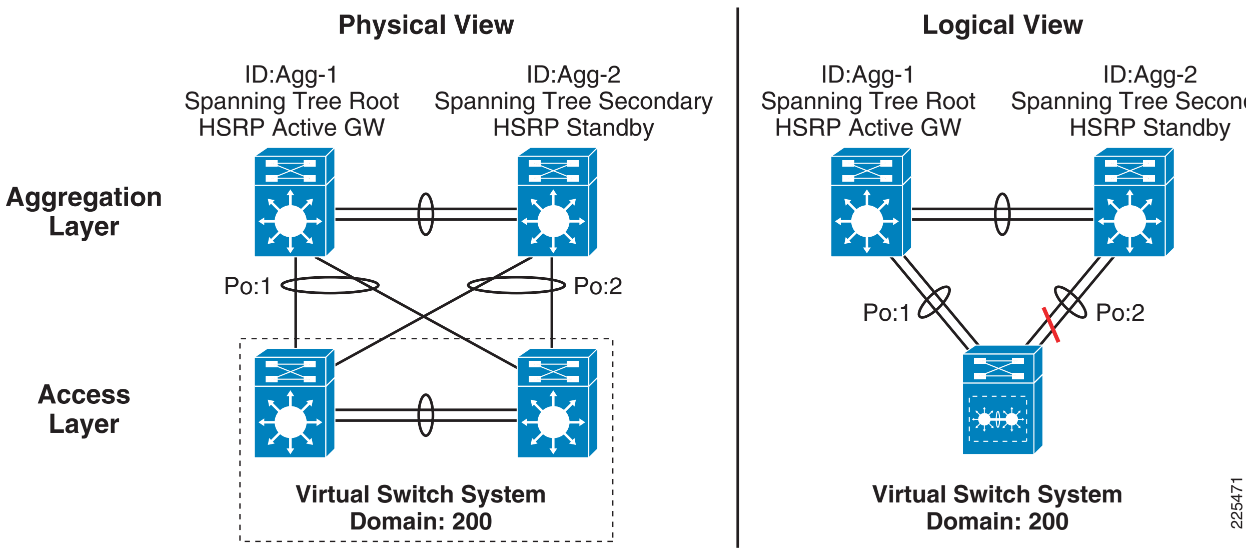
The use of VSS at the access layer allows network and server administrators to benefit from a virtualized network infrastructure. There are two VSS access layers models to consider:

- Traditional Aggregation Layer and VSS Access Layer, page 36
- VSS Aggregation and Access Layers, page 38

Traditional Aggregation Layer and VSS Access Layer

The use of a VSS-enabled access layer in a data center leveraging a traditional aggregation layer mirrors a traditional deployment in many respects. Figure 21 depicts the physical and logical design of a VSS-enabled access layer. The VSS member switches support a VSL link and are dual-homed to the aggregation layer switches. The aggregation switches have an ISL between them and have one switch defined as the Layer-2 and Layer-3 preferred switch, following the current best practices recommendations. In this example, the *Agg-1* and *Agg-2* aggregation switches connect to the VSS access-layer switch using MEC, port channels 1 and 2. As seen on the right, port channel "2" is blocking as spanning tree is required to address the loop in the network between these three logical devices.

Figure 21 Traditional Aggregation Layer with VSS Enabled Access Layer



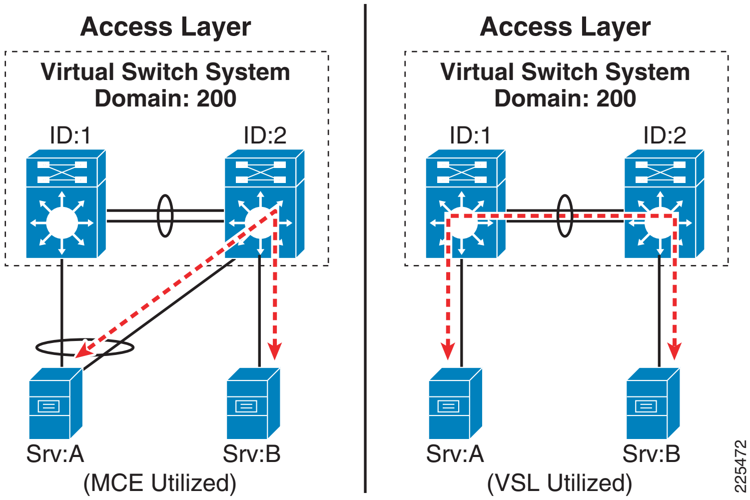
The primary benefits of this design are:

- VSS access switch configuration follows existing best practices for Layer-2 connectivity (spanning tree, edge port configurations)
- Reduced number of access-layer switches to manage without sacrificing number of access ports available
- Reduced spanning-tree domain
- MEC available for dual-homed servers

As the logical view depicted in Figure 21 shows, the VSS switch looks and behaves much like a traditional Layer-2 switch. However, the proper provisioning of the VSL link must always be taken into account. For example, if one of links in port channel 1 fails, the VSL link will be leveraged to forward traffic to the remaining link in the MEC.

As stated earlier, it is best practice to dual-home a VSS environment to other network devices in the design. This same recommendation holds true for endpoints or server connecting to a VSS access-layer switch. Figure 22 illustrates this point as two servers, *A* and *B*, are connected to a VSS access switch. Server *A* is dual-homed using an EtherChannel connection while server *B* is attached to only VSS switch member 2. The VSL link is not utilized with the MEC attached server as the local forwarding path is preferred over the VSL link. The traffic between servers *A* and *B* does not leverage the VSL. The right side of Figure 22 shows that "orphaned" or single-homed ports will utilize the VSL for communication.

Figure 22 VSS Access Layer Server-to-Server Traffic Pattern Example



Note More information on server and VSS connectivity can be found in the "Server Connectivity" section on page 40.

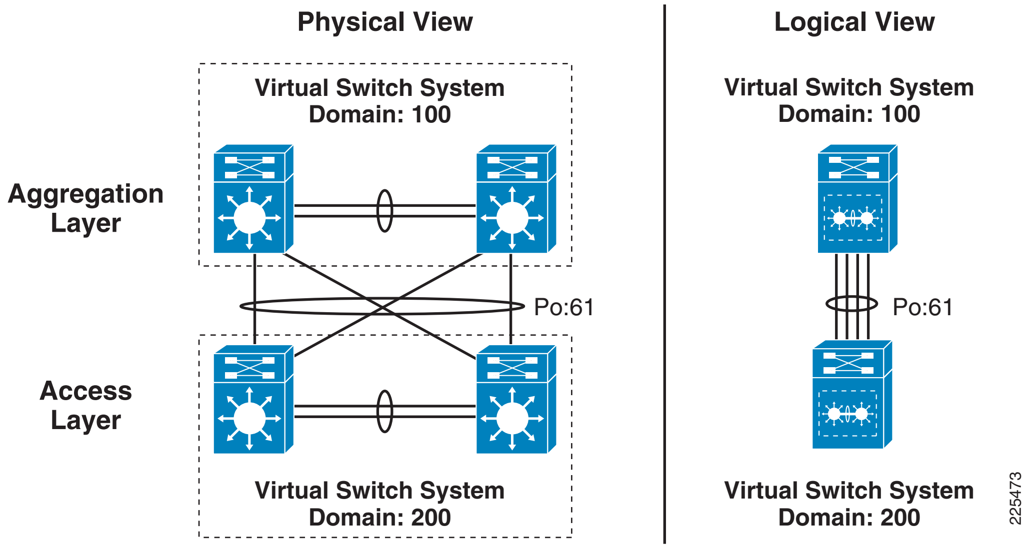
VSS Aggregation and Access Layers

VSS at the aggregation and access layer simplifies the Layer-2 domain while providing a more robust switching fabric. Figure 23 depicts the aggregation-layer and access-layer configuration implemented during testing.

The access switches are dual-homed to the aggregation switches using 10 Gigabit Ethernet links. From a physical perspective, this mirrors the "triangle" topology leveraged in many contemporary data centers—implying that a conversion to VSS might not require re-cabling of an existing data center infrastructure beyond the VSL link functionality.

The right side of Figure 23 shows the logical topology formed when using a VSS-enabled aggregation and access layer. This virtual topology forms a simplistic one–to-one switch link via MEC. There are no loops to contend with; all links will be forwarding and, from a management perspective, there are only two logical devices to configure and monitor. Spanning tree does not create a forwarding topology and convergence is dependent on EtherChannel convergence—not spanning tree.

Figure 23 VSS Access Layer Physical and Logical View



Features

VSS at the access layer allows a network administrator to grow the Layer-2 domain in a very proscribed manner. Through virtualization, the traditional pair of access-layer switches is replaced by a single logical entity. This means the complexity of Layer 2 is reduced while maintaining the available port density and physical redundancy in the network. Figure 23 highlights this advantage because the logical view shows only two logical devices.

Perhaps the greatest advantage is to the servers leveraging the VSS access-layer switch. Given the NIC teaming functionality of current server platforms and the MEC capabilities of VSS, enterprises may deploy a highly available and robust server farm. This portion of the document will address virtualization, Layer 2, and server connectivity in the VSS access layer.

Configuring Virtualization

The use of VSS at the access layer follows the same configuration recommendations as those described for the aggregation layer. The following VSS configuration was used:

```
switch virtual domain 200
 switch mode virtual
 dual-active detection pagp trust channel-group 61
```

Notice that the MEC between the VSS access and aggregation switches is the dual-active detection medium. The **show switch virtual dual-active pagp** command confirms this functionality.

```
# show switch virtual dual-active pagp
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1
Channel group 61 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
          Dual-Active     Partner              Partner   Partner
Port      Detect Capable  Name                 Port      Version
Te1/1/1   Yes             dca-vss              Te2/13/7  1.1
```

```
Te1/1/2   Yes          dca-vss            Te1/13/7  1.1
Te2/1/1   Yes          dca-vss            Te1/13/8  1.1
Te2/1/2   Yes          dca-vss            Te2/13/8  1.1
```

Note For more information on VSS virtualization, see the "Aggregation Layer" section on page 16 section.

Layer 2

The VSS access-layer switch reliably expands the Layer-2 fabric available to end nodes in the server farm by removing logical loops in the network. This reduces ones dependency on spanning tree, but does not mean a network administrator should remove spanning tree completely from the environment. As with the aggregation layer, a VSS switch should have spanning tree enabled as a failsafe mechanism. Remember spanning tree will not be affect the forwarding topology. The following configuration enables RPVST+ at the access-layer VSS switch:

```
spanning-tree mode rapid-pvst
spanning-tree extend system-id
spanning-tree pathcost method long
```

Server Connectivity

This section of the document will focus on the connection of servers to a VSS access-layer switch.

Single-homed Server Configurations

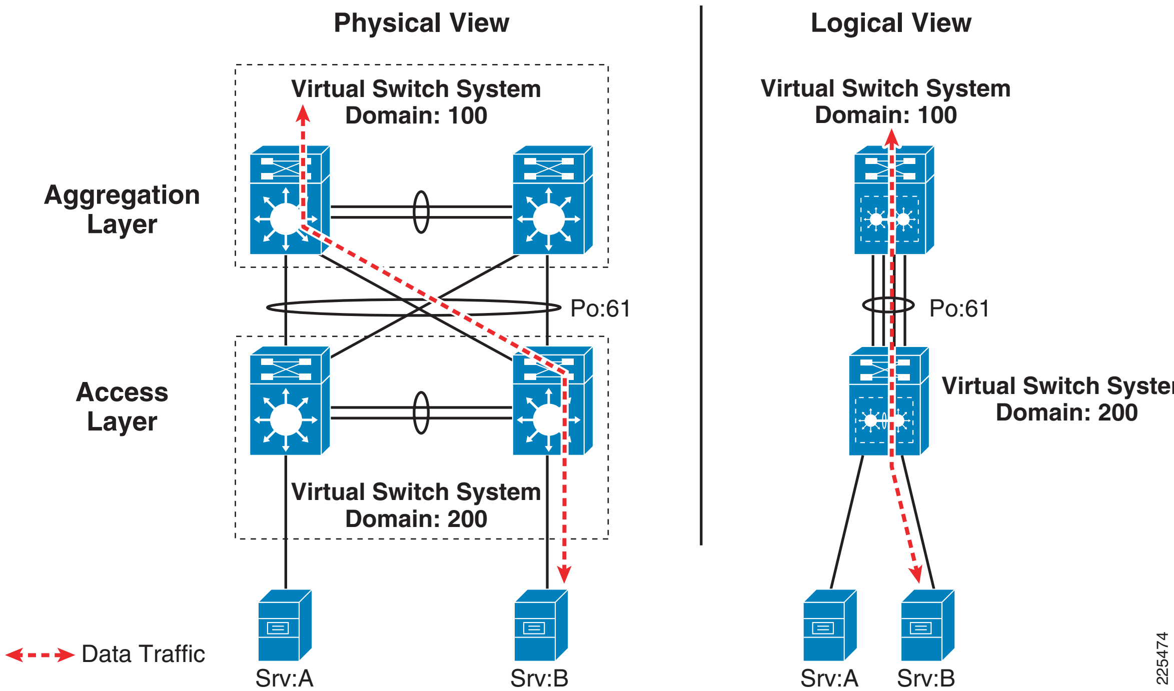
A single server homed to a single switch—VSS-enabled or otherwise—is exposed to a single point of failure. For servers housing less critical applications this is an acceptable and common practice. There are two primary traffic patterns to consider:

- Client-to-server
- Server-to-server

Client-to-Server Traffic

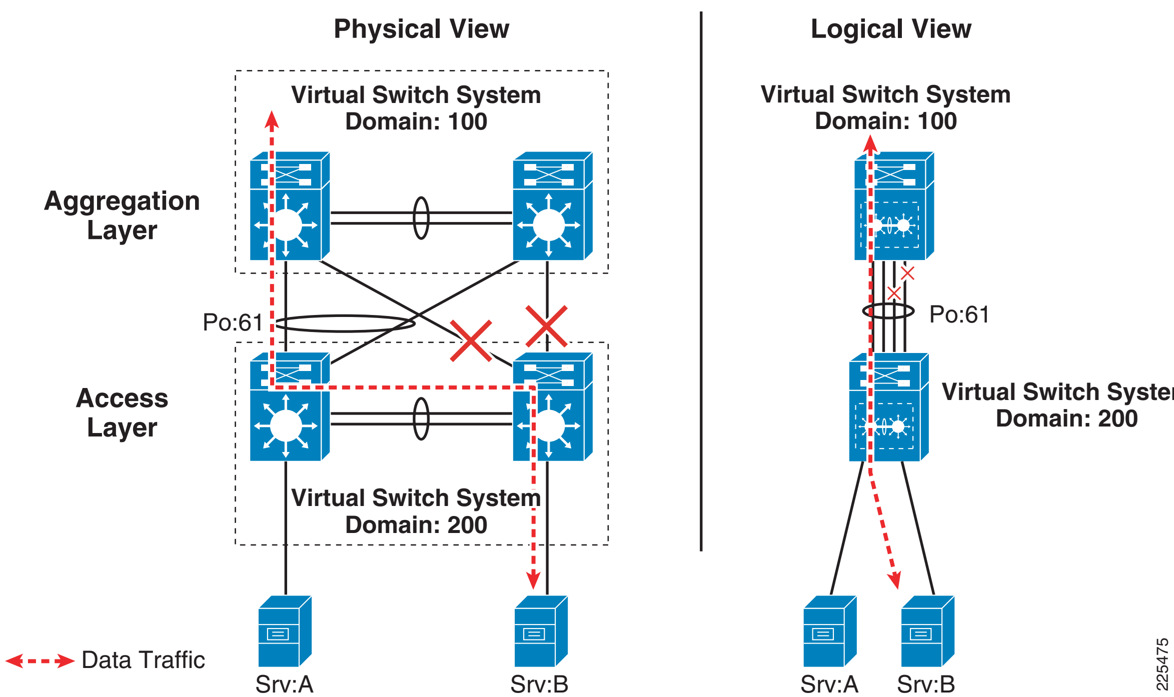
Figure 24 represents the flow of traffic through the VSS aggregation and access layers. The preferred forwarding path in and out of the data center leverages the local MEC interface at each layer.

Figure 24 Client-to-Server Traffic Pattern



Failure of all the local MEC interfaces will force utilization of the VSL at the access layer, but the logical flow remains unchanged as shown in Figure 25.

Figure 25 Client-to-Server Traffic Pattern with MEC Failures

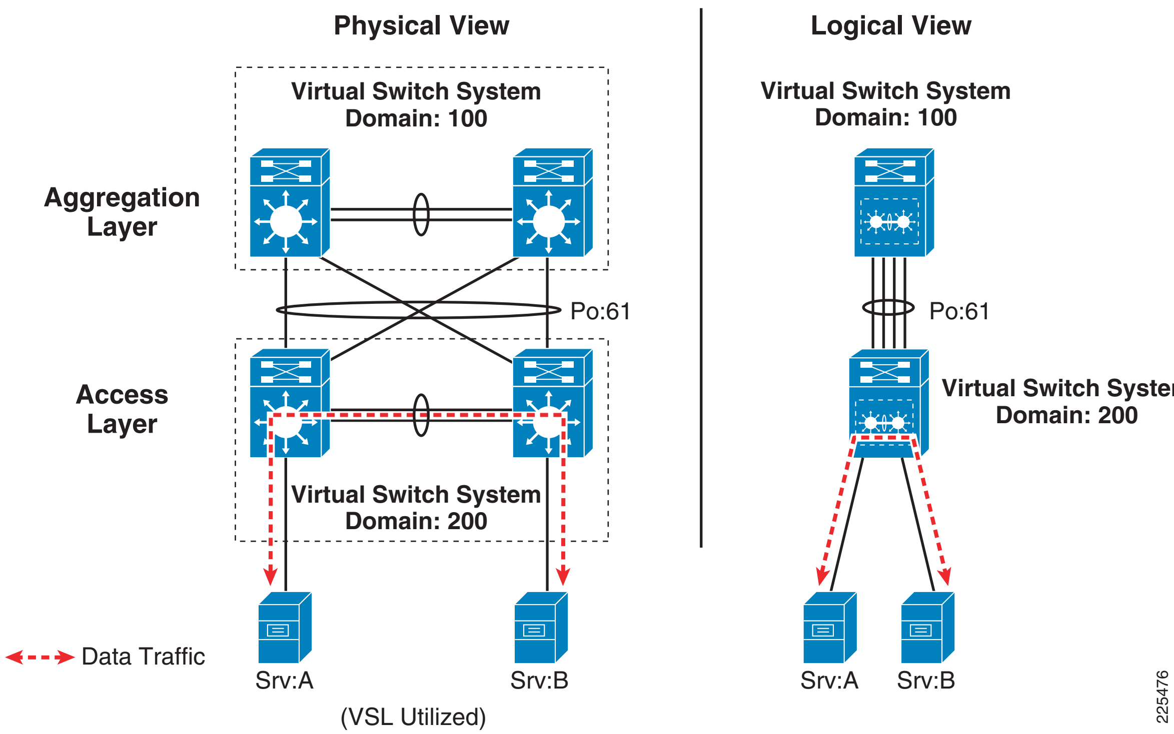


Server-to-Server Traffic

Single homed server-to-server traffic places more pressure on the VSL link under normal conditions. Figure 26 depicts the physical and logical flow of single-homed servers in a VSS access layer. The servers in this scenario are connected to the same VSS switch, but to different members of the domain. This forces traffic across the VSL. Network administrators should consider homing the servers to the same switch to avoid VSL utilization.

The VSS access-layer switch is more than capable of accommodating single-homed servers, but (as stated throughout the document) the VSL link should be provisioned appropriately. The VSL EtherChannel link may be a maximum of eight 10 Gigabit Ethernet connections—providing a highly available transport. However, as with all EtherChannels, the VSL link hashes the traffic based on the EtherChannel load-balancing algorithm favoring local interfaces in the port channel. The default hash is source and destination IP-based, meaning that (if there are many common flows resulting in the same hash result) there could be contention on a single link of the channel. Leverage a Layer 4-based hash to achieve improved load distribution across the system.

Figure 26 Server-to-Server Traffic Pattern



Note The VSL consists of 10 Gigabit Ethernet links; therefore, overrunning a single link might be an unnecessary concern dependent on the traffic load of the data center.

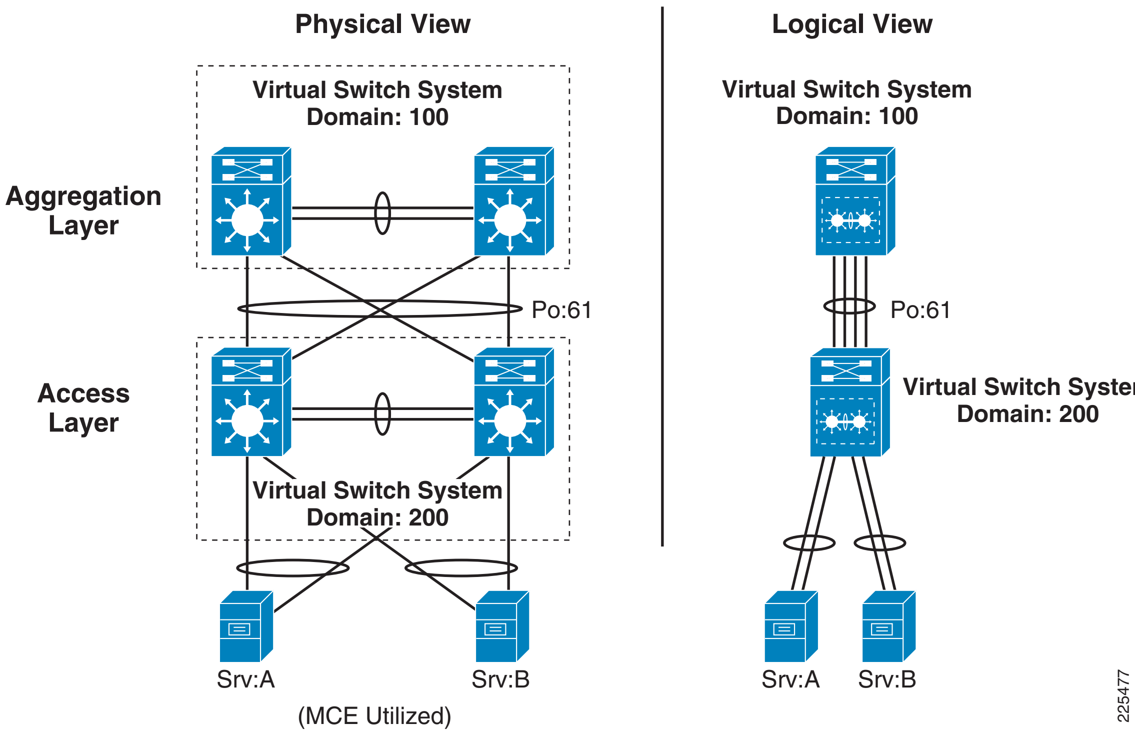
Dual-homed Server Configurations

To achieve the maximum performance capabilities of VSS technology, it is best to leverage MEC throughout the data center—including within the server farm. MEC is compatible with the NIC teaming or bonding technologies available on contemporary server platforms. In fact, NIC teaming and MEC are well suited for one another because NIC teaming implies Layer-2 adjacency and MEC simplifies this requirement via virtualization.

Figure 27 shows the physical and logical configuration of the VSS-enabled access layer. As shown, VSS MEC functionality and server NIC teaming allows for a highly available loop-free design. This section of the document details the configuration of MEC using the following operating systems:

- Windows 2003 Server
- Red Hat Enterprise Linux 5 Update 2
- VMware ESX 3.03

Figure 27 NIC Teaming with MEC



It should be noted that there are many flavors and variations of NIC teaming—including network fault tolerance (NFT), transmit load balancing (TLB), dual channel (flex links), switch assisted, dynamic and static. Taking into consideration the use of a VSS access layer, the testing focused on active/active NIC configurations. As a result, teams only leveraging NFT and TLB active/standby configurations were eliminated from the effort. In addition, the use of dual-channel teams was eliminated because the design stresses the optimization of link utilization—eliminating unused standby links.

Windows Server 2003

The following NIC teaming configurations were tested using a Windows Server 2003 operating system connected to a VSS access layer:

- 802.3ad Dynamic with fault tolerance
- Switch-assisted load balancing (SLB) with fault tolerance

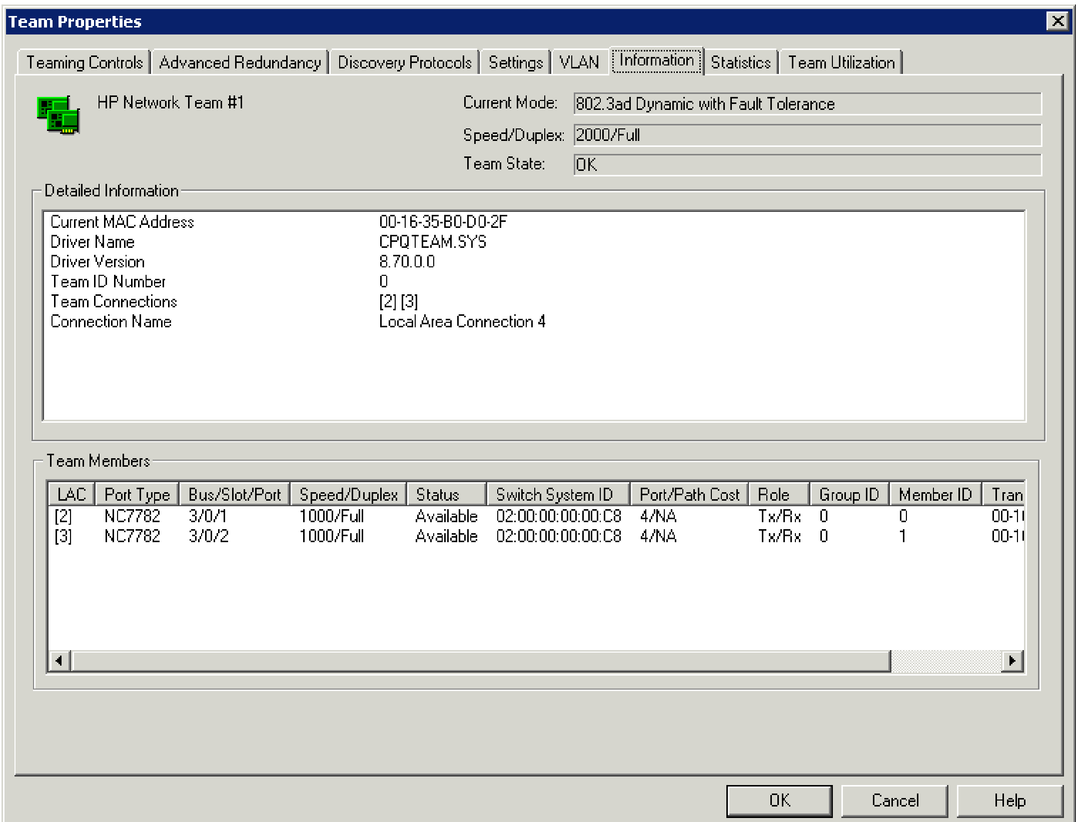
Note The NIC teams were created using a pair of Hewlett Packard NC7782 Gigabit Server Adapters configured using HP Network Configuration Utility (NCU)

802.3ad Dynamic with Fault Tolerance

This type of team leverages LACP to create a EtherChannel where all transmit packets are load balanced across the members of the server team and all received packets are load balanced across the members via the switch. This is an active/active deployment model. There is a single virtual MAC address and IP address shared by the team members. Failure of any links in the team results in the traffic being redistributed across the remaining members of the team.

Figure 28 details the configuration used during testing. Notice that the *Speed/Duplex* setting reflects a 2 Gbps link, indicating successful port aggregation between the server and VSS access-layer switch.

Figure 28 802.3ad Teaming Configuration



225478

The following port channel configuration was used for testing dynamic LACP-based aggregation:

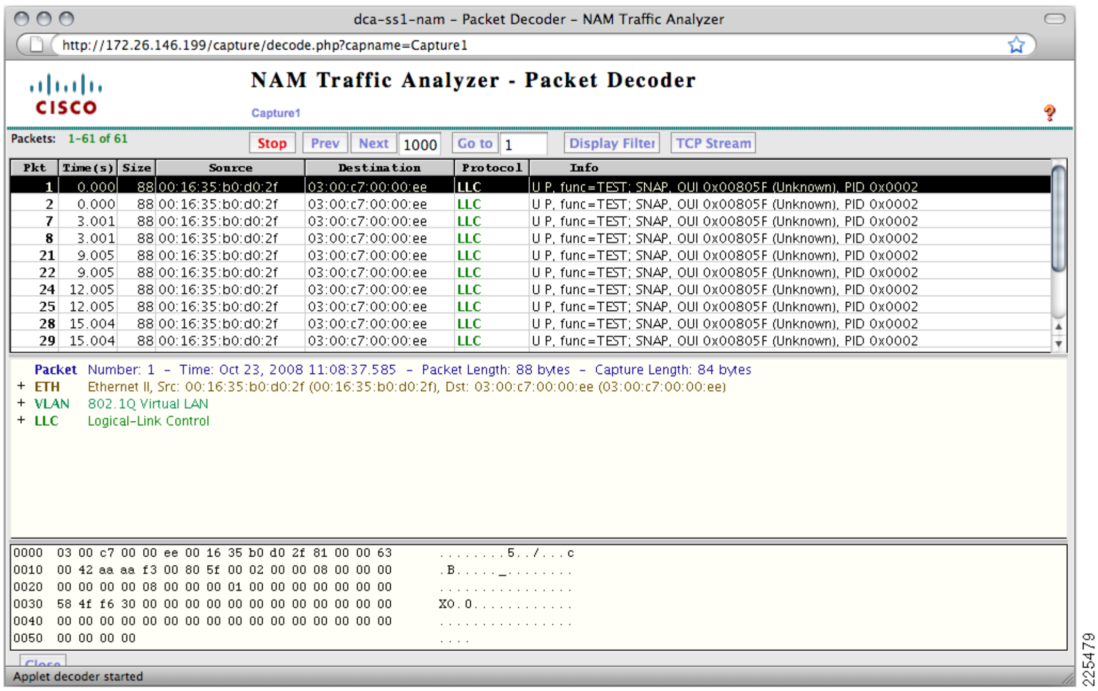
```
interface Port-channel9
 description <<** Windows 2003 HP NIC Team **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 99,133
 switchport mode trunk
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
end
```

The interfaces comprising the port channel leveraged the following configuration to enable LACP:

```
interface GigabitEthernet x/x/x
channel-protocol lacp
channel-group 9 mode active
```

It should be noted that the server NIC teaming "heartbeat" functionality was enabled for transmit and receive path validation. Transmit path validation means the server will attempt to transmit a frame on any link that has been inactive for a period of time defined by the interval setting. The default time is 3 seconds and is configurable in the range of 3-to-60 seconds. This heartbeat is designed to determine physical errors on an interface and will result in the removal of the NIC if the heartbeat cannot be placed on the wire. Figure 29 shows the use of this heartbeat by inactive NICs in the channel. If two heartbeats are missed the interface is declared down with the minimum setting this would result in a maximum of 9 seconds of downtime.

Figure 29 Example Heartbeat Capture



Receive path validation operates inversely; heartbeat frames are sent by the operational NIC team members when one or more of the team members has not received any frames in the interval period. The assumption being that the failed NIC team member will receive a frame and confirm the validity of the forwarding path. The failed NIC will then reset.

Availability and Performance

The 802.3ad NIC team combined with MEC proved to be both highly available and salable. The detection of a NIC team member failing averaged 1.5 seconds while recovering this same interface occurred in less than one second. Providing a highly available solution. The performance gains in terms of server productivity are excellent. Figure 30 illustrates that the amount of traffic supported by the NIC team effectively doubles upon recovery of a failed member NIC.

Figure 30 NIC Team Recovery Throughput Example

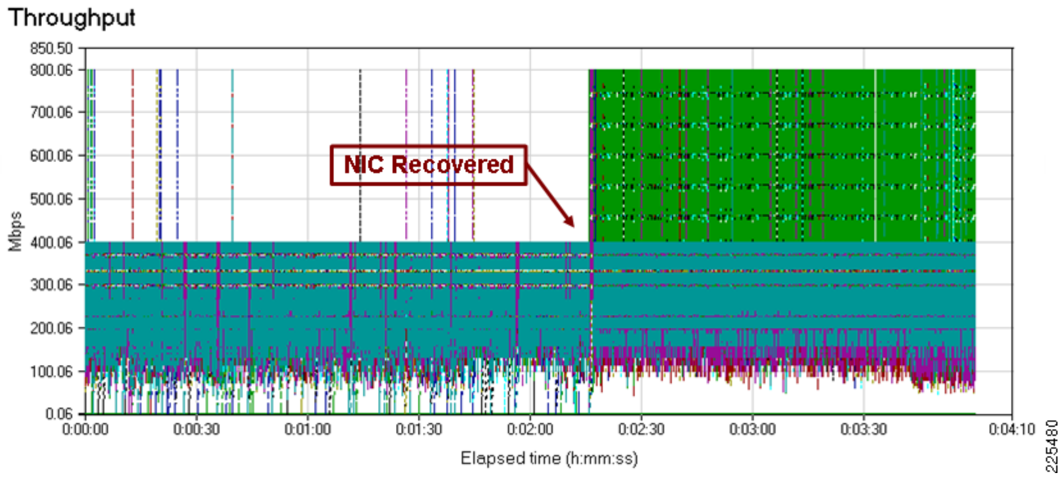


Figure 31 is a visual indication of the team performance. Receive and transmit traffic is distributed across the team members indicating that the dynamic channel between the VSS access layer and server are working. The VSS system is using the following transmit load balancing hashes:

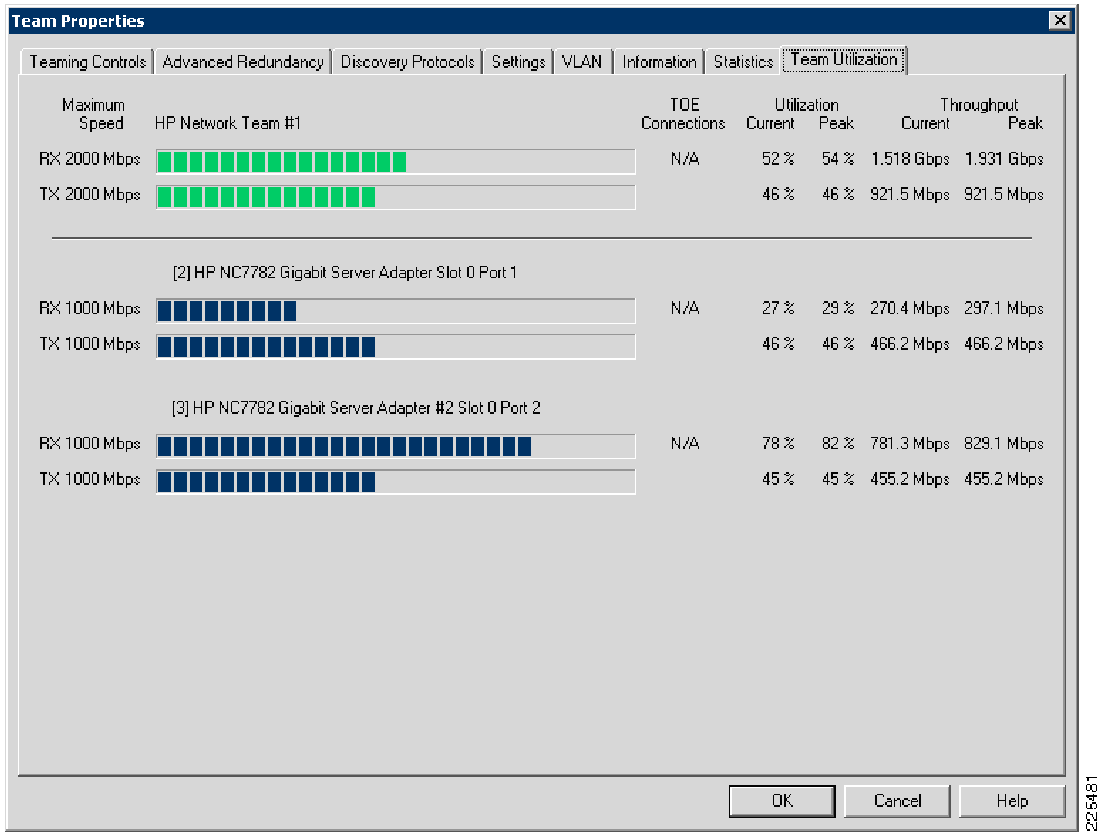
```
# show etherchannel load-balance
EtherChannel Load-Balancing Configuration:
        src-dst-mixed-ip-port enhanced
        mpls label-ip

EtherChannel Load-Balancing Addresses Used Per-Protocol:
Non-IP: Source XOR Destination MAC address
  IPv4: Source XOR Destination IP address and TCP/UDP (layer-4) port number
  IPv6: Source XOR Destination IP address
  MPLS: Label or IP
```

The server is configured for "automatic" load balancing, allowing it to leverage the most granular data points for load distribution:

- Layer-4 port information for TCP connections
- Layer-3 IP source and destination address
- Layer-2 destination MAC address
- Round robin (last resort)

The primary objective is to optimize link utilization without sacrificing the ordering of packets on the wire. Packets received out of order generally degrade the receiving system's performance.

Figure 31 Sample of NIC Teaming Utilization

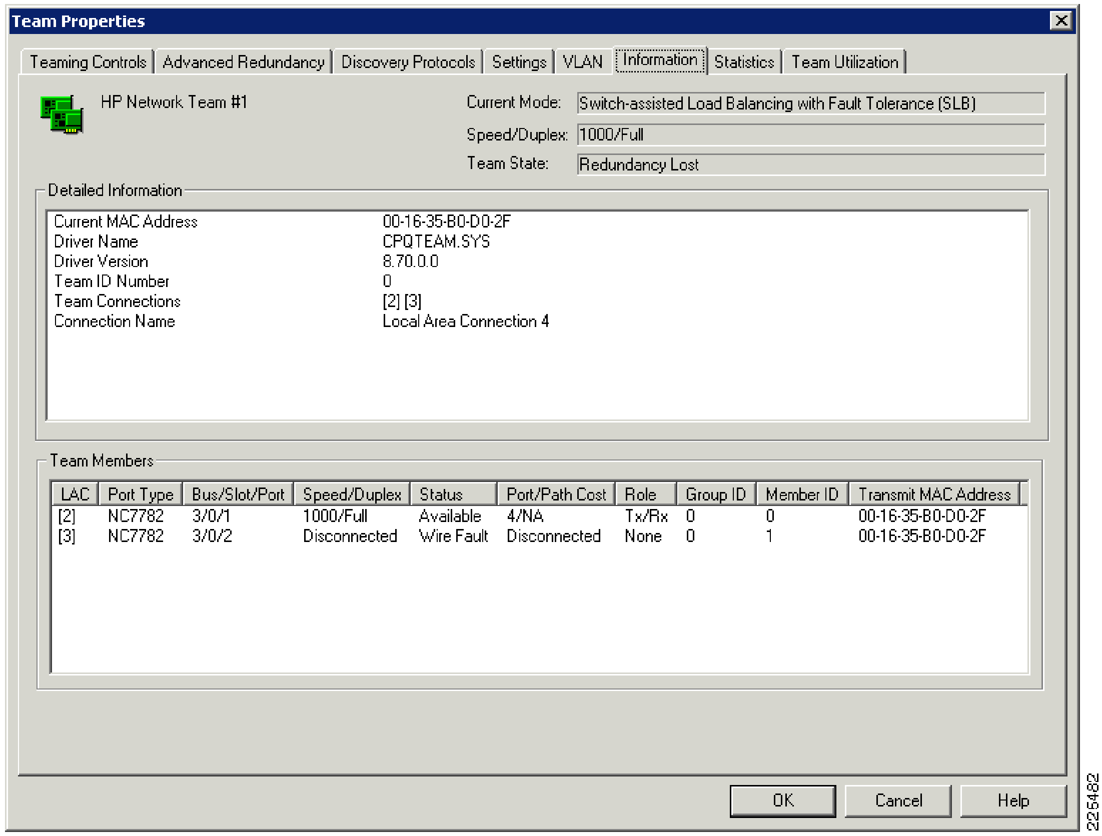


Switch-assisted Load Balancing with Fault Tolerance

The SLB with fault tolerance uses static EtherChannel to aggregate the NICs on the servers and ports on the VSS system. SLB has traditionally been leveraged for redundancy to a single switch; however, with the introduction of VSS at the access-layer, MEC may be leveraged to provide improved availability by introducing switch redundancy via virtualization. The SLB methods discussed in the previous section all apply.

Figure 32 is the teaming configuration used during testing. The performance in terms of throughput of this SLB configuration was equivalent to the 802.3ad configuration previously described; however, there was a difference in terms of availability.

Figure 32 SLB with Fault Tolerance Example



The VSS switch configuration implements a port channel with member interfaces using static EtherChannel as follows:

```
interface GigabitEthernet x/x/x
channel-group 9 mode active
```

Availability and Performance

The NIC team had similar behavior to the dynamic LACP channel in the event of a NIC failure, showing an approximately 1.5 seconds worth of lost connectivity for flows previously supported by the failed NIC. It was during the NIC recovery that the use of static EtherChannel becomes less desirable as the interfaces between the switch and server become active at different times causing approximately 2-to-3 seconds of downtime for those streams assigned to the interface. Specifically, those connections being transmitted by the server are lost as the server immediately leverages the NIC upon detection of physical link while the switch has not yet added the port back into the port channel bundle. This process was observed to take approximately one second on average. For this reason, it is recommended to use 802.3ad dynamic port channels for MEC server connectivity.

Red Hat Enterprise Linux 5 Update 2

NIC teaming in Linux environments, commonly referred to as *bonding*, has seven different modes of operation. These seven modes are as follows:

- Balanced Round Robin (mode=0)
- Active-Backup (mode=1)
- Balance XOR (mode=2)

- Broadcast (mode=3)
- 802.3ad (mode=4)
- Balance TLB (mode=5)
- Balance ALB (mode=6)

802.3ad dynamic bonding is the recommended method to leverage the availability and performance benefits afforded via the VSS implementation of MEC, while maintaining ordered delivery of packets across the network. Issues with the other methods are as follows:

- Balanced Round Robin does not lend itself to the ordered delivery of packets.
- Active-Backup and Balance TLB do not support the full active/active interface design requirements and capabilities of VSS.
- Balance XOR is a consideration for static EtherChannel connectivity, but only if the traffic in the data center is limited to a broadcast domain. Any gateway-destined traffic would be polarized to a single interface in the bond. Server traffic patterns must be well understood using this choice.
- Broadcast mode, while providing redundancy, does not efficiently utilize network or server resources.
- Balanced ALB does provide EtherChannel support and overall link utilization, but it adds a layer of complexity by leveraging the Address Resolution Protocol (ARP) to direct incoming traffic across the slave interfaces in the bond.

For the above reasons, the 802.3ad dynamic bonding mode is recommended for the server farm.

802.3ad (mode=4)

802.3ad is the recommended mode of operation for Linux bonds with a VSS system. This is not to imply that the other modes would not work, but as stated earlier there are many advantages to a dynamic LACP implementation. The following bonding configuration was used during testing:

```
# cat ifcfg-bond1
DEVICE=bond1
BOOTPROTO=none
ONBOOT=yes
BONDING_OPTS="mode=4 miimon=100 xmit_hash_policy=1"
USERCTL=no
```

In this example, the configuration file indicates the name of the bond and the definition of the bond (*bond1*). The bond leverages LACP, mode 4, with interface monitoring via miimon to detect link state. The hash policy is set to "1" for L4 based load balancing.

Using a sub-interface command, network administrators may assign multiple virtual LAN interfaces under the overall bond configuration. In the file output the follows, the VLAN option is enabled and the name of the bond is extended.

```
# cat ifcfg-bond1.129
DEVICE=bond0.129
NETWORK=10.7.129.0
IPADDR=10.7.129.151
NETMASK=255.255.255.0
USERCTL=no
VLAN=yes
```

Finally, assign physical interfaces as slave devices to the master bond configuration. Two interfaces are slave devices of master bond *bond1* in the following configuration:

```
cat ifcfg-eth2
# Intel Corporation 82546GB Gigabit Ethernet
DEVICE=eth2
```

```
BOOTPROTO=none
HWADDR=00:19:BB:E9:6C:92
ONBOOT=yes
TYPE=Ethernet
MASTER=bond1
SLAVE=yes
USERCTL=no
cat ifcfg-eth3
# Intel Corporation 82546GB Gigabit Ethernet
DEVICE=eth3
BOOTPROTO=none
HWADDR=00:19:BB:E9:6C:93
ONBOOT=yes
TYPE=Ethernet
MASTER=bond1
SLAVE=yes
USERCTL=no
```

The **cat /proc/net/bonding/bond1** command will verify the configuration of the bond and the status of its slave interfaces. The following example output indicated the bond mode, hash, and interfaces previously configured. Notice that the *Partner Key* value is equal to that of the associated port channel configuration at the access layer. The port channel configuration is the same as documented in the "Windows Server 2003" section on page 43.

```
/proc/net/bonding/bond1
Ethernet Channel Bonding Driver: v3.2.4

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer3+4 (1)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

802.3ad info
LACP rate: slow
Active Aggregator Info:
        Aggregator ID: 2
        Number of ports: 2
        Actor Key: 17
        Partner Key: 151
        Partner Mac Address: 02:00:00:00:00:c8

Slave Interface: eth2
MII Status: up
Link Failure Count: 2
Permanent HW addr: 00:19:bb:e9:6c:92
Aggregator ID: 2

Slave Interface: eth3
MII Status: up
Link Failure Count: 2
Permanent HW addr: 00:19:bb:e9:6c:93
Aggregator ID: 2
```

The MAC address is shared by the team and learned by the VSS switches in the data center. This MAC address is typically inherited from the first active NIC in the bond but can be manually defined as well.

Note For more information on Linux bonding go to http://lxr.linux.no/linux/Documentation/networking/bonding.txt

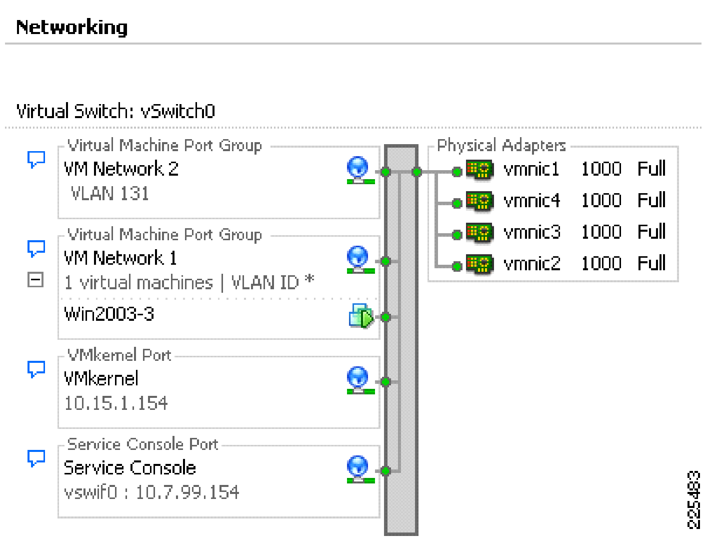
Availability and Performance

The Linux bond showed similar load distribution across the 802.3ad bond member based on the Layer-4 hashing algorithm. Failure of a single bond member interface resulted in an average of 1.5 seconds of traffic loss across flows dedicated to that link of the channel. Recovery was very rapid with the bond member being assimilated into the channel and passing traffic in approximately 10–to-20 milliseconds.

VMWare 3.0.3

The ESX server supports a virtual switch, which is a software construct providing connectivity for virtual machines within the host. Figure 33 is a representation of the ESX virtual switch (*vSwitch0*) leveraged during testing. *vSwitch0* has four Gigabit Ethernet interfaces, known as *vmnics*, assigned as physical resources. The vmnics connect to the VSS access-layer switches via MEC. The virtual switch supports multiple port groups and VLANs, as well as virtual machines. As a result, the four vmnics are configured as trunks connected via a MEC. This configuration allows for active/active interfaces for all VLANs in the channel.

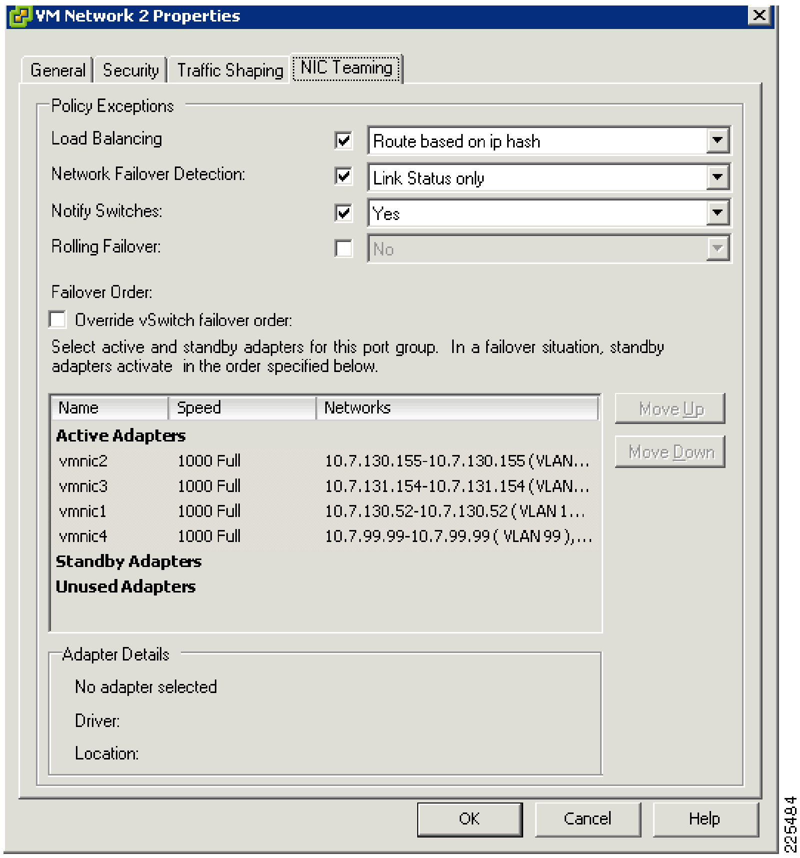
Figure 33 ESX Server Network Configurations



The current best-practice recommendation from Cisco Systems and VMware is to use an IP-hash on the virtual switch for load distribution across the vmnics. This is an active/active design with load balancing based on the hash of the source and destination IP address. ESX server does not currently support dynamic link aggregation; therefore, static EtherChannels are necessary.

Figure 34 shows the configuration under test. Notice there are no failover preferences since all paths in a VSS MEC are equal. Link status determines availability of the vmnic and the interfaces in the port channel.

Figure 34 Virtual Switch NIC Teaming Properties



The following is the port channel configuration on the VSS access layer:

```
interface Port-channel154
 description <<** VMware MEC **>>
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 15,99,130,131
 switchport mode trunk
 switchport nonegotiate
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
 no cdp enable
!
```

Note The interfaces that are members of port channel 154 in this example use static EtherChanneling. This is configured using the **channel-group** <#> command mode.

Note For more details about VMware Infrastructure 3 in a Cisco networking environment, see the following URL: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/vmware/VMware.html

Availability and Performance

The use of the VSS MEC and a Layer-3 hash on the ESX server NIC team showed favorable failover and recovery times. Failing a single vmnic resulted in less than 500 millisecond convergence of flows previously using the failed interface. Recovery of the failed link resulted in convergence times in the 600 millisecond range.

Conclusion

This document discusses using VSS technology in the aggregation and access layers of the data center as a solution to availability, scalability, and infrastructure utilization concerns that exist in many contemporary enterprise data centers. This document detailed the configuration of network and server platforms to provide a robust network environment via VSS.

Additional References

- *Data Center Design—IP Network Infrastructure*

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html
- *Security and Virtualization in the Data Center*

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/dc_sec_design.html
- *Data Center Service Patterns*

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/dc_serv_pat.html