CHAPTER 2

Cisco VMDC Design and Deployment

This chapter discusses the design and identifies specific implementation concerns of the VMDC solution. It is divided into the following broad categories:

Compute Virtualization

In this solution, compute virtualization is driven by integration between vSphere's Virtual Center and the Nexus 1000V distributed virtual switch. This section describes the role of these components in the solution and describes how

vSphere 4.0

VMware vSphere 4.0 builds on the Virtual Center Server software available with ESX 3.5 and ESX 4.0 to support cloud computing across the Internet and to secure the VM access layer. Figure 2-1 identifies key vSphere 4 features and their relationship to vCenter.

Figure 2-1 vSphere Within the Cloud

©VMware Inc

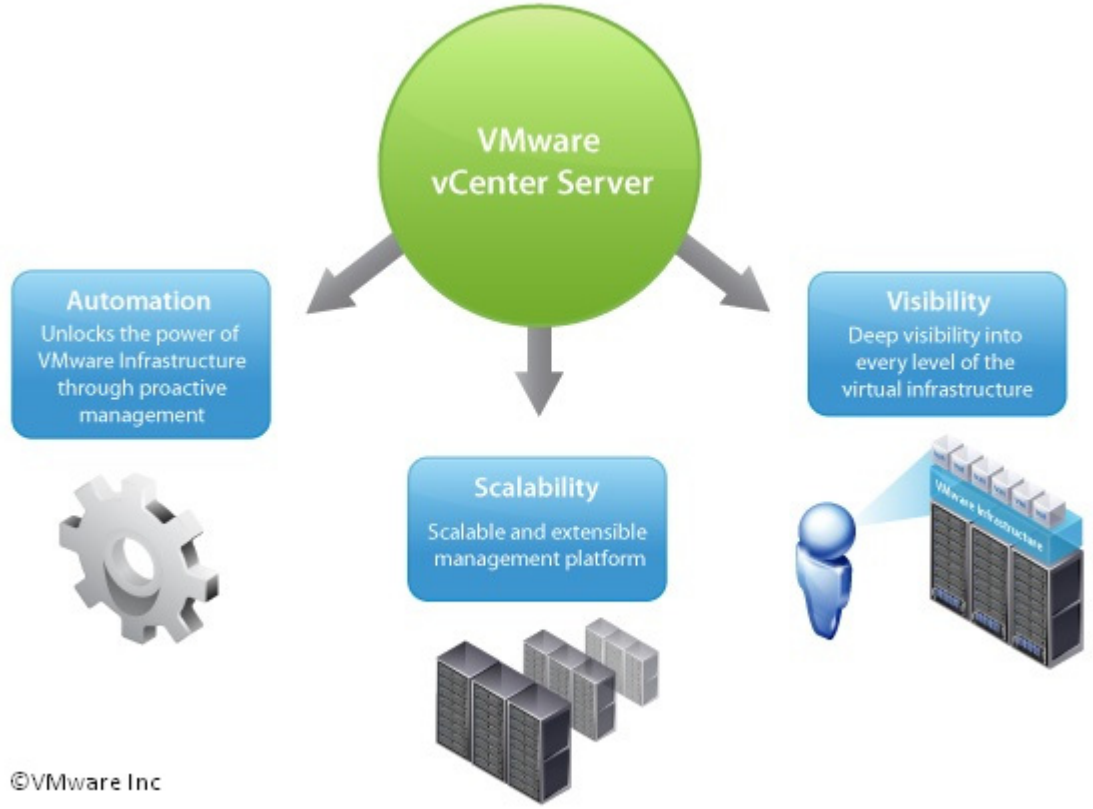
VMware vSphere provides features to support availability, security, scalability, and integration with the Cisco Nexus 1000V distributed virtual switch (DVS). In previous releases of ESX and vCenter, network configuration was performed with the vCenter Virtual Infrastructure interface, which supported the basic access layer features of the VMware vSwitch. Server administrators controlled both the configuration of servers in the virtualized environment and the access layer configuration. With the VMware vSwitch at the access layer, deployments could not guarantee security and their monitoring features remained in the traditional data center where each physical server was physically connected to a Cisco switch. The vSphere 4.0 software bridges the gap between server and network administrators by incorporating the Nexus 1000V software switch, allowing network administrators to control port access, security, and monitoring features via the Cisco NX-OS CLI at the VM access layer.

Virtual Center Server

VMware vCenter Server manages VMware vSphere and all virtual machines and resources hosted in the data center. vSphere combines management and performance data monitoring for all ESX hosts and VMs in a single console where you can access status information for clusters, hosts, VMs, storage, and the guest operating system.

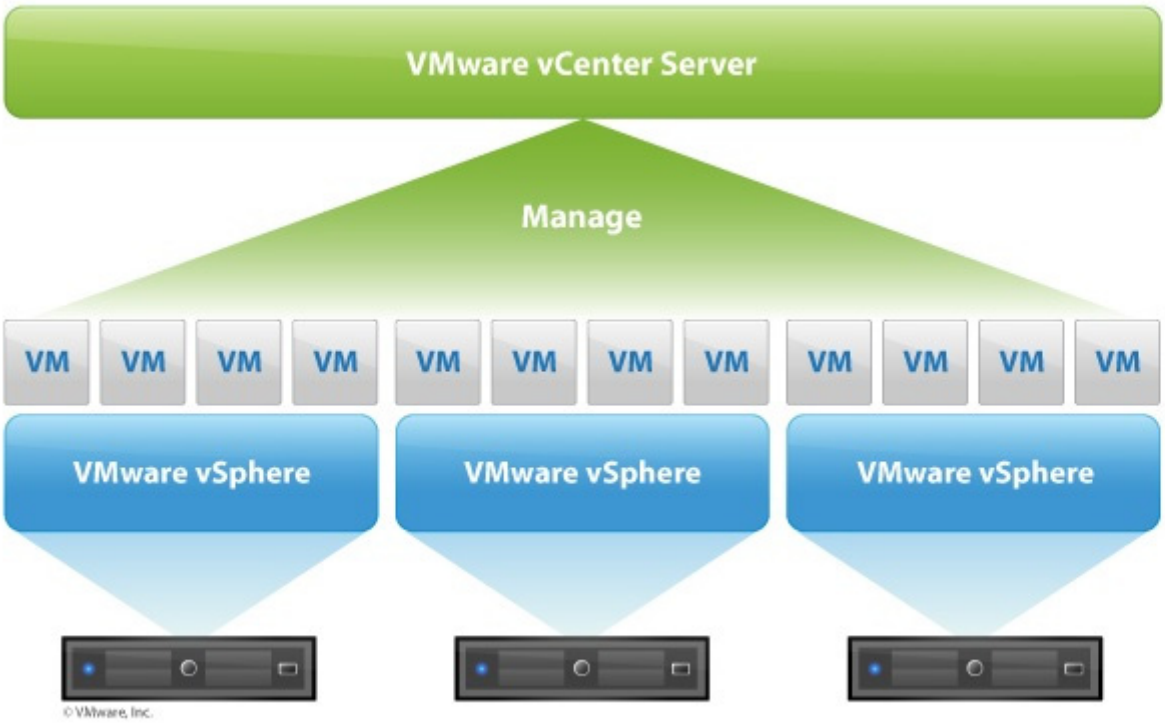
VMware vCenter servers enable fast provisioning of VMs and hosts via templates and cloning. Its snapshot feature allow you to create backups of virtual machines and restore them if a guest operating system fails. VMware vCenter Server also provides control over key capabilities, such as VMware VMotion, distributed resource scheduler, high availability, and fault tolerance.

Figure 2-2 WMware vCenter Eases Deployments



As you extend virtualization throughout the data center, VMware vCenter Server supports the management of many physical hosts and thousands of virtual machines from a single console.

Figure 2-3 Virtual Center Eases Management



You can install Virtual Center Server on several Microsoft Windows platforms, including Windows Server 2003 R2, Microsoft Windows XP Professional, and Microsoft Windows Server 2008.

The vCenter Server software runs as a process on a virtual machine or a physical server. For this design, we recommend using a virtual machine hosted on a server outside of the cluster hosting tenant virtual machines (for example, you could host it on a virtual machine in an infrastructure cluster).

For deployment guidance, refer to the *ESX and vCenter Server Installation Guide* at http://www.vmware.com/pdf/vsphere4/r40/vsp_40_esx_vc_installation_guide.pdf.

Note When installing ESX, you must provide a SQL database instance. While you can select a SQL Express target, it only supports up to five ESX hosts. Deployments of greater than five ESX hosts require Microsoft Enterprise SQL 2005 or Oracle 10g (or later).

Nexus 1000V Switch

Cisco Nexus 1000V Series switches are virtual machine access switches for implementation in VMware vSphere environments running the Cisco NX-OS software operating system. The switch is certified compatible with VMware vSphere, vCenter, ESX, and ESXi, and many VMware vSphere features. It operates inside the VMware ESX hypervisor to enable Cisco VN-Link server virtualization and provide virtual machines with mobile network and security policies to restrict connectivity.

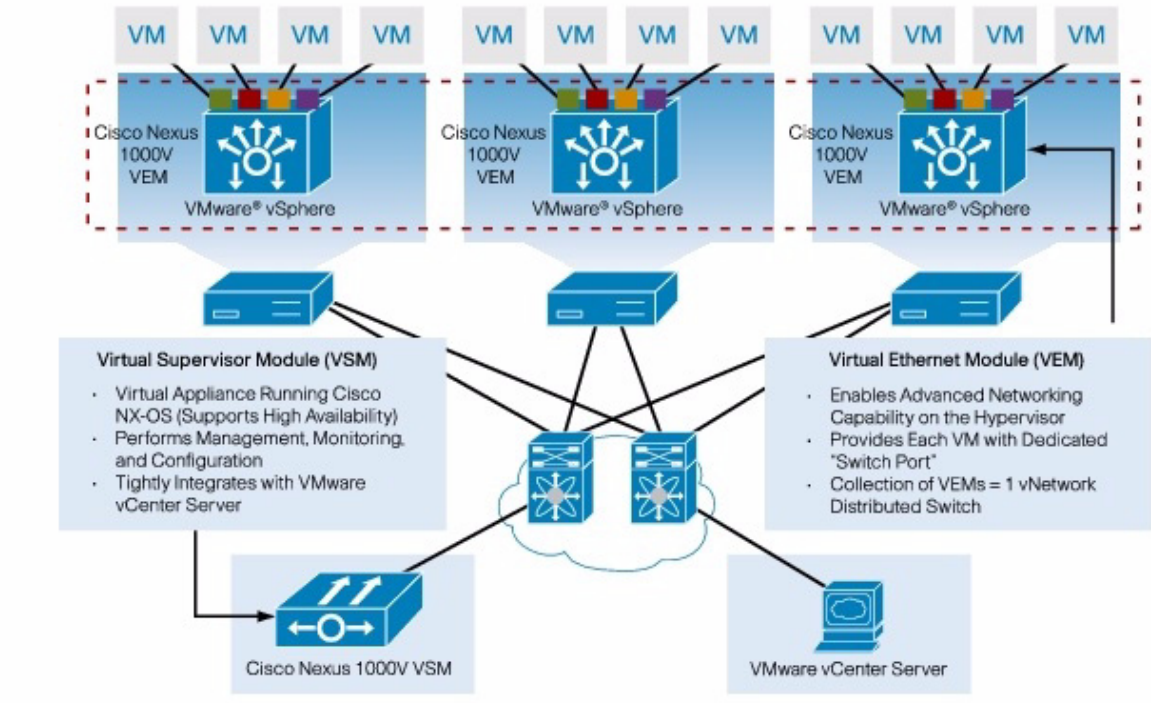
Typically during server virtualization in the data center, the virtual servers are not managed in the same way as physical servers. Server virtualization is a special, longer deployment that requires coordination among server, network, storage, and security administrators. The Cisco Nexus 1000V Series switches enable a consistent networking feature set and provisioning process from the virtual machine access layer to the core network infrastructure. Virtual servers can use the same network configuration, security policy, diagnostic tools, and operational models as their physical server counterparts attached to dedicated physical network ports. Virtualization administrators can access predefined network policy that follows mobile virtual machines to ensure proper connectivity and leave more time for virtual machine administration. These capabilities help you deploy server virtualization faster to gain its benefits sooner.

The Cisco Nexus 1000V switch is a virtual access software switch for vNetwork Distributed Switches that work with VMware vSphere 4.0. The Cisco Nexus 1000V switch has the following components:

- **Virtual Supervisor Module (VSM).** The control plane of the switch and a virtual machine that runs the Cisco NX-OS operating system.
- **Virtual Ethernet Module (VEM).** A virtual line card embedded in each VMware vSphere ESX host that is a part of the distributed switch. The VEM resides partly inside the kernel of the hypervisor and partly in a user mode process, called the VEM Agent.

Figure 2-4 illustrates the basic relationship between the Nexus 1000V components.

Figure 2-4 Nexus 1000V Component Relationship



The VEM executes inside the Cisco Nexus 1000V Series switch hypervisor. The external VSM manages the Cisco Nexus 1000V Series switch VEMs.

Virtual Ethernet Module

The Cisco Nexus 1000V VEM executes as part of the VMware ESX or ESXi kernel and replaces VMware Virtual Switch functionality. The VEM leverages the VMware vNetwork Distributed Switch (vDS) API, developed jointly by Cisco and VMware, to provide advanced networking capability to virtual machines. The Cisco Nexus 1000V Series switch is fully aware of server virtualization events, such as VMware VMotion and Distributed Resource Scheduler (DRS). The VEM pulls configuration details from the VSM to perform the following switching and advanced networking functions:

- Port Channels
- Quality of service (QoS)
- Security: private VLAN, access control lists, port security
- Monitoring: NetFlow, Switch Port Analyzer (SPAN), Encapsulated Remote SPAN (ERSPAN)

In the event communication with VSM fails, the VEM uses Nonstop Forwarding to switch traffic based on the last known configuration. It provides advanced switching with data center reliability in the server virtualization environment.

Virtual Supervisor Module

The Cisco Nexus 1000V Series VSM controls multiple VEMs as one logical, modular switch. Instead of physical line card modules, the VSM supports multiple VEMs running in software on the physical servers. Configuration is performed through the VSM and is automatically propagated to the VEMs. Instead of configuring soft switches inside the hypervisor on a host-by-host basis, administrators define configurations for all VEMs managed by the VSM from a single interface. Cisco NX-OS provides the Cisco Nexus 1000V Series switch with the following benefits:

- **Flexibility and Scalability.** Port profiles provide configuration of ports by category to enable the solution to scale to a large number of ports. Common software can run all areas of the data center network, including LAN and storage area networks.
- **High availability.** Synchronized, redundant VSMs enable rapid, stateful failover and ensure a high availability virtual machine network.
- **Manageability.** The Cisco Nexus 1000V Series switch managed using the command-line interface (CLI), Simple Network Management Protocol (SNMP), or CiscoWorks LAN Management Solution (LMS). The VSM also integrates into VMware vCenter Server to enable network configuration of the Cisco Nexus 1000V switch from a common interface.

Note Active and Standby VSM modules can run on the same physical ESX server and provide redundancy. However, supervisors on two separate ESX servers improves HA of server failure.

Control and Packet VLANs

The control VLAN and the packet VLAN support communications between the VSM and the VEMs in a switch domain. The packet VLAN is used by protocols such as CDP, LACP, and IGMP. The control VLAN is used for:

- VSM-to-VEM configuration commands and responses
- VEM-to-VSM notifications, such as the attachment or detachment of ports to the DVS

Note The control VLAN and packet VLAN should be separate VLANs, and they should be separate from those VLANs that carry data.

Port Profiles

A port profile is a set of interface configuration commands that can be dynamically applied to either physical (uplink) or virtual interfaces. A port profile can define attributes for the following functions:

- VLAN
- Port channels
- Private VLAN (PVLAN)
- ACL
- Port security
- NetFlow
- Rate limiting
- QoS marking

You define port profiles in the VSM. When the VSM connects to vCenter Server, it creates a DVS, and each port profile is published as a port group on the DVS. Then, you can apply those port groups to specific uplinks, VM vNICs, or management ports, such as virtual switch interfaces or VM kernel NICs.

A change to a VSM port profile propagates to all ports associated with that port profile. You can use the Cisco NX-OS CLI to alter a specific interface configuration relative to its port profile configuration. For example, you can shut down a specific uplink or apply ERSPAN to a specific virtual port without affecting other interfaces using the same port profile.

System VLANs

When a server administrator adds a host to the DVS, the VEM in that host must be configurable by the VSM. Since the ports and VLANs for this communication are not in place, system port profiles and system VLANs are configured to meet this need. VSM sends minimal configuration to the vCenter Server, which propagates that configuration to the VEM when the host is added to the DVS.

A system port profile establishes and protects vCenter Server connectivity. This port profile should be configured with system VLANs designed to facilitate transport for the following scenarios:

- System VLANs (control and packet) or vNICs used to bring up ports before communication is established between the VSM and VEM.
- Management uplinks used for VMware vCenter Server connectivity or SSH or Telnet connections. More than one management port or VLAN can be defined. For example, one dedicated for vCenter Server connectivity, one for SSH, one for SNMP, a switch interface, and so forth.
- VMware kernel NIC used to access VMFS storage over iSCSI or NFS.

Note No system VLAN was configured for VMFS over iSCSI or NFS as those protocols were not used for storage in this design.

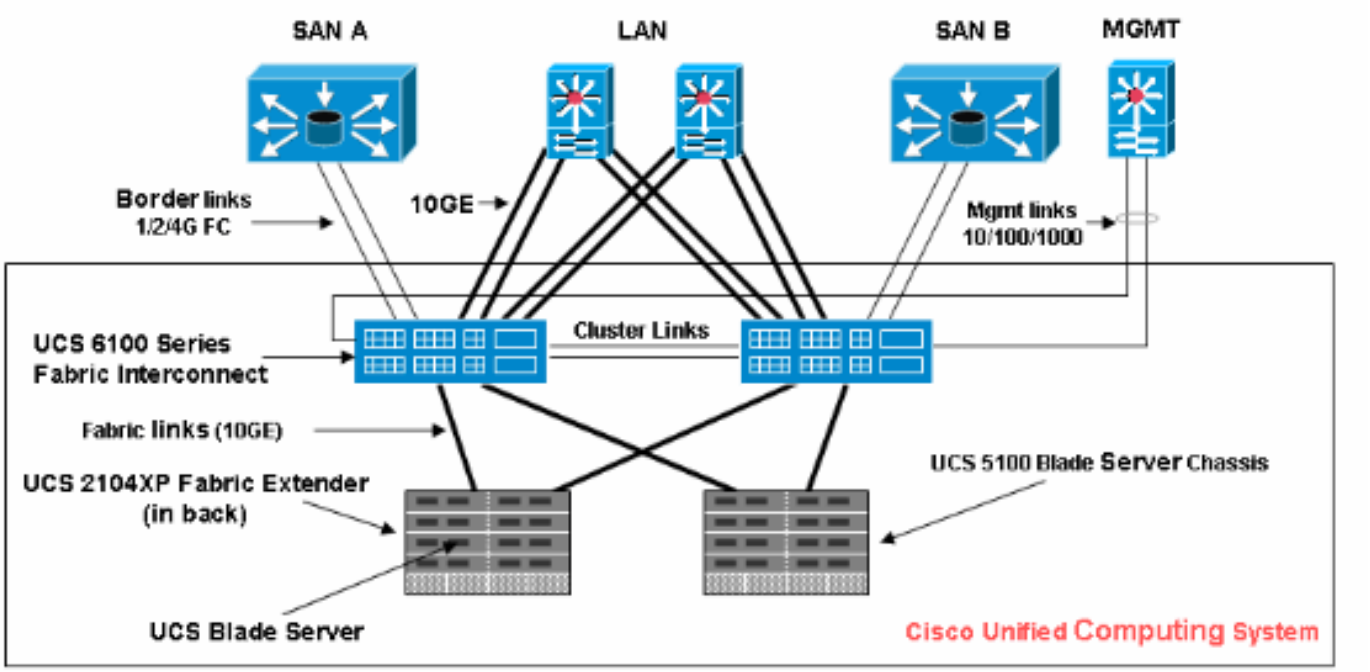
Unified Computing System Design

The Cisco Unified Computing System (UCS) combines compute, network, storage access, and virtualization in a an integrated, multi-chassis platform where all resources are managed in unified domain. The Cisco UCS Manager (UCSM) enables storage, network, and server administrators to collaboratively define service profiles for applications.

Service profiles are logical representations of desired physical configurations and infrastructure policies. They help automate provisioning and increase business agility, allowing data center managers to provision resources in minutes instead of days. With service profiles, server attributes are no longer tied to physical hardware, which guarantees seamless server mobility.

Figure 2-5 shows a typical UCS deployment topology. The components of the UCS and how they are interconnected are explained in the following sections.

Figure 2-5 Cisco UCS Typical Topology



The six-port 8G Fibre Channel (FC) expansion module for UCS 6100 series is orderable configured (N10-E0060) or as spare (N10-E0060=). The ports provide 1/2/4G FC with SFP transceivers or 2/4/8G FC with SFP+ transceivers.

This section presents the following topics:

Physical Hardware Layout

This section presents the following topics:

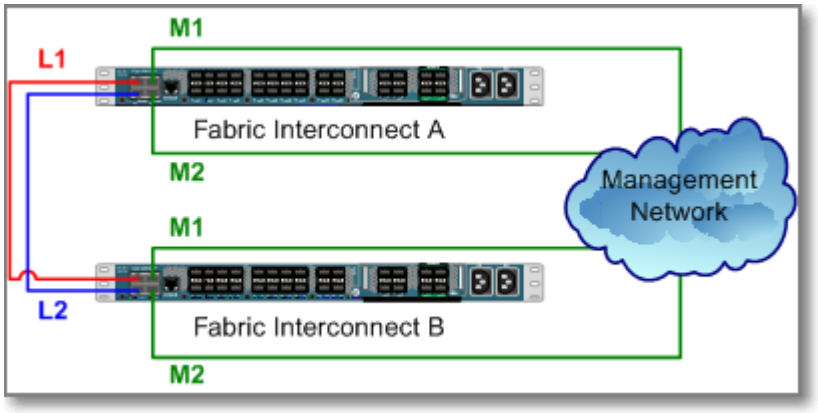
UCS Cluster

Typically, Cisco UCS is deployed in as a high availability clustered for management plane redundancy and increased data plane bandwidth.

Note The cluster configuration provides redundancy for the management plane only. Data plane redundancy depends on the user configuration and may require a third-party tool for support.

The cluster configuration requires two Cisco UCS 6100 Series Fabric Interconnects directly connected with Ethernet cables between the L1 (L1-to-L1) and L2 (L2-to-L2) ports. This connection allows each Fabric Interconnect to continuously monitor the other's status for immediate alert of failure. If one Fabric Interconnect becomes unavailable, the other takes over automatically.

Figure 2-6 UCS Cluster



UCS Chassis

The Cisco UCS 5108 Blade Server Chassis incorporates unified fabric and fabric-extender technology, resulting in fewer physical components and improved energy efficiency compared to traditional blade server chassis. It eliminates the need for dedicated chassis management and blade switches, and it reduces cabling.

The Cisco Unified Computing System scales up to 40 chassis without adding complexity. Each chassis is six rack units (6RU) high, can mount in an industry-standard 19-inch rack, and uses standard front-to-back cooling. The front of each chassis accommodates up to eight UCS B200 M1 blade servers for a maximum of 320 per system (8 per chassis at 40 chassis) or four UCS B-250 M1 blade servers for a maximum of 160. The front also accommodates four power supplies. On the back, the chassis has slots for eight fans, two fabric extenders, and the power entry module.

Server Blades

The Cisco UCS B-200 M1 Blade Servers adapt processor performance to application demands and intelligently scale energy use based on utilization. Each server uses network adapters for consolidated access to the unified fabric. This access reduces the number of adapters, cables, and access-layer switches required for LAN and SAN connectivity, which in turn reduces capital expenses and administrative, power, and cooling costs.

Features of the Cisco UCS B-200 M1 include the following:

- Up to two Intel® Xeon® 5500 Series processors, which automatically and intelligently adjust server performance according to application needs, increasing performance when needed and achieving substantial energy savings when not.
- Up to 96 GB of DDR3 memory in a half-width form factor for mainstream workloads to balance memory capacity and overall density.
- Two optional Small Form Factor (SFF) Serial Attached SCSI (SAS) hard drives available in 73GB 15K RPM and 146GB 10K RPM versions with an LSI Logic 1064e controller and integrated RAID.
- One dual-port mezzanine card for up to 20 Gbps of I/O per blade. Mezzanine card options include a Cisco UCS VIC M81KR Virtual Interface Card, a converged network adapter (Emulex or QLogic compatible), or a single 10GB Ethernet Adapter.

The Cisco UCS B250M1 Extended Memory blade server maximizes performance and capacity for demanding virtualization and large data set workloads. It supports up to 384GB of memory to provide a more cost-effective memory footprint for less-demanding workloads. The server is a full-width, two-socket blade server with increased throughput and more than double the memory compared to traditional two-socket x86 servers.

Features of the Cisco UCS B-250 M1 include the following:

- Up to two Intel® Xeon® 5500 Series processors, which automatically and intelligently adjust server performance according to application needs, increasing performance when needed and achieving substantial energy savings when not.
- Up to 384 GB of DDR3 memory for demanding virtualization and large data-set applications. Alternatively, this technology offers a more cost-effective memory footprint for less-demanding workloads.
- Two optional Small Form Factor (SFF) Serial Attached SCSI (SAS) hard drives available in 73GB 15K RPM and 146GB 10K RPM versions with an LSI Logic 1064e controller and integrated RAID.
- Two dual-port mezzanine cards for up to 40 Gbps of I/O per blade. Mezzanine card options include a Cisco UCS VIC M81KR Virtual Interface Card, a converged network adapter (Emulex or QLogic compatible), or a single 10GB Ethernet Adapter.

Blade Server Memory Configuration

When populating DIMM memory on UCS B200-M1 blade server, populate the DIMMs in balanced configuration:

- Populate DIMMs in multiples of six balanced configuration based on three memory channels per direct-linked CPU socket memory architecture
- Use identical DIMM types throughout the system; same size, speed, and number of ranks
- Populate same DIMMs for each channel, and each CPU socket
- Populate DIMMs to maximize number of channels to achieve highest degree of memory interleaving for highest memory bandwidth

Note When a UCS B200-M1 blade server has just one CPU populated in CPU0 socket, only DIMMs in slots A1, A2, B1, B2, C1, and C2 associated with CPU0 are visible to the operating system. There is no need to populated DIMMs in slot D1, D2, E1, E2, F1, and F2.

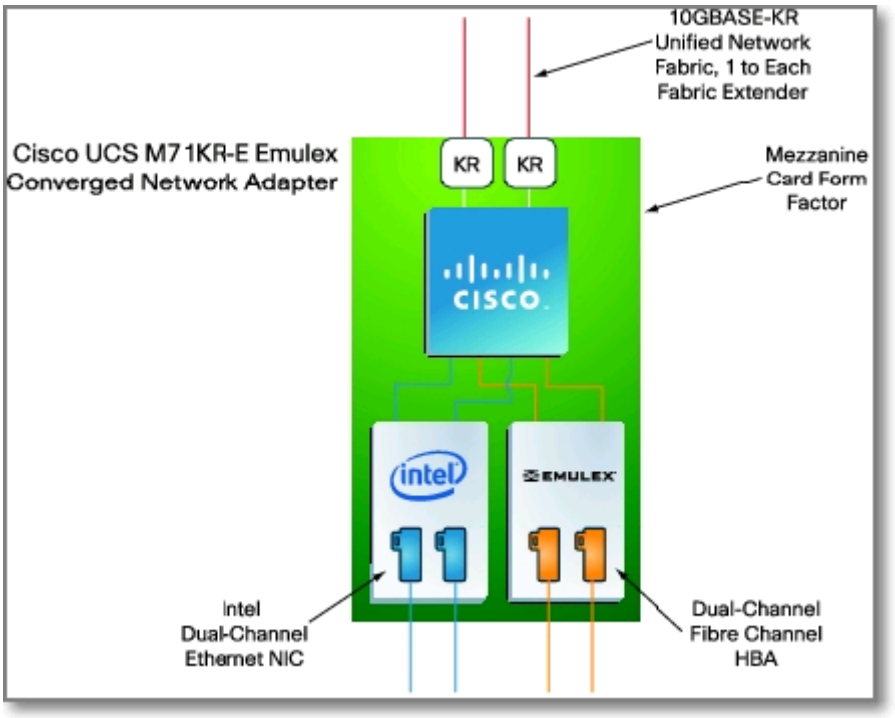
For details and install instruction on memory for the UCS B200-M1 blade server, refer to http://www.cisco.com/en/US/docs/unified_computing/ucs/hw/chassis/install/blade.html

For details and install instruction on memory for the UCS B250-M1 blade server, refer to http://www.cisco.com/en/US/docs/unified_computing/ucs/hw/chassis/install/fullblade.html

Mezzanine Card Options

The Cisco UCS M71KR-E Emulex converged network adapter is supported on the UCS blade server to provide a dual-port connection to the mid-plane of the blade server chassis. The Cisco UCS M71KR-E uses an Intel 82598 10 Gigabit Ethernet controller for network traffic and an Emulex 4-Gbps Fibre Channel controller for Fibre Channel traffic on the same mezzanine card. Cisco UCS M71KR-E adapters present two discrete Fibre Channel host bus adapter (HBA) ports and two Ethernet network ports to the operating system.

Figure 2-7 UCS M71KR-E Adapter



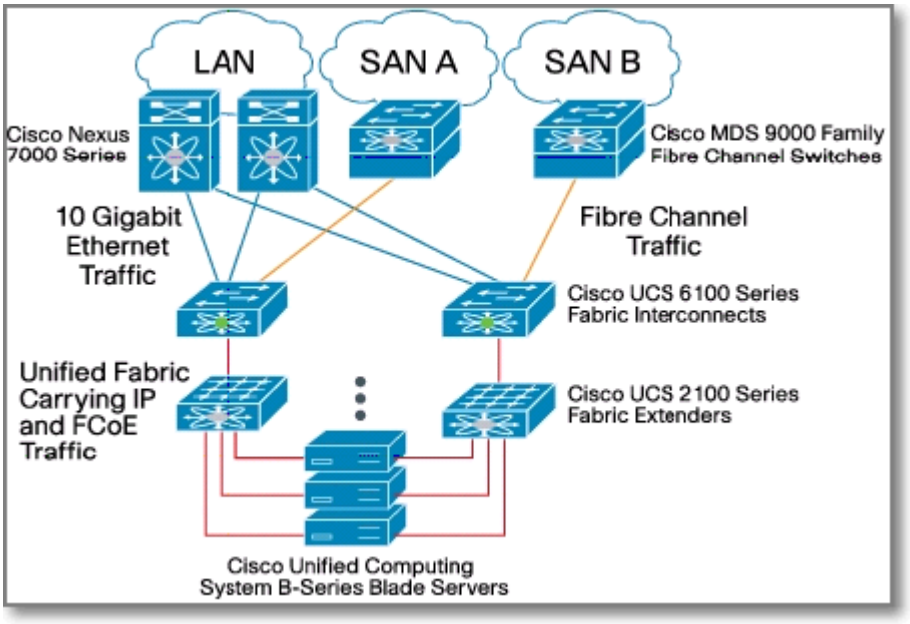
Cisco UCS 6100 Series Fabric Interconnect

The Cisco UCS 6100 Series fabric interconnect is a core part of the Cisco Unified Computing System that provides network connectivity and management capabilities. It provides a unified fabric that consolidates I/O, supporting Ethernet/IP and Fibre Channel traffic in the system though wire-once 10-Gigabit Ethernet and FCoE downlinks and flexible 10-Gigabit Ethernet and 1/2/4-Gbps Fibre

Channel uplinks. Out-of-band management, including switch redundancy, is supported through dedicated management and clustering ports. The interconnects feature front-to-back cooling, redundant front-plug fans and power supplies, and rear cabling that facilitates efficient cooling and serviceability. Typically deployed in active-active redundant pairs, the fabric interconnects provide uniform access to networks and storage, to support a fully virtualized environment with a flexible, programmable pool of resources.

Figure 2-8 shows a unified fabric that carries multiple traffic streams to Cisco UCS 6100 Series fabric interconnects, where Ethernet and Fibre Channel traffic splits into separate networks.

Figure 2-8 Unified Fabric to UCS 6100 Fabric Interconnect



In this solution, the switch provides the following features:

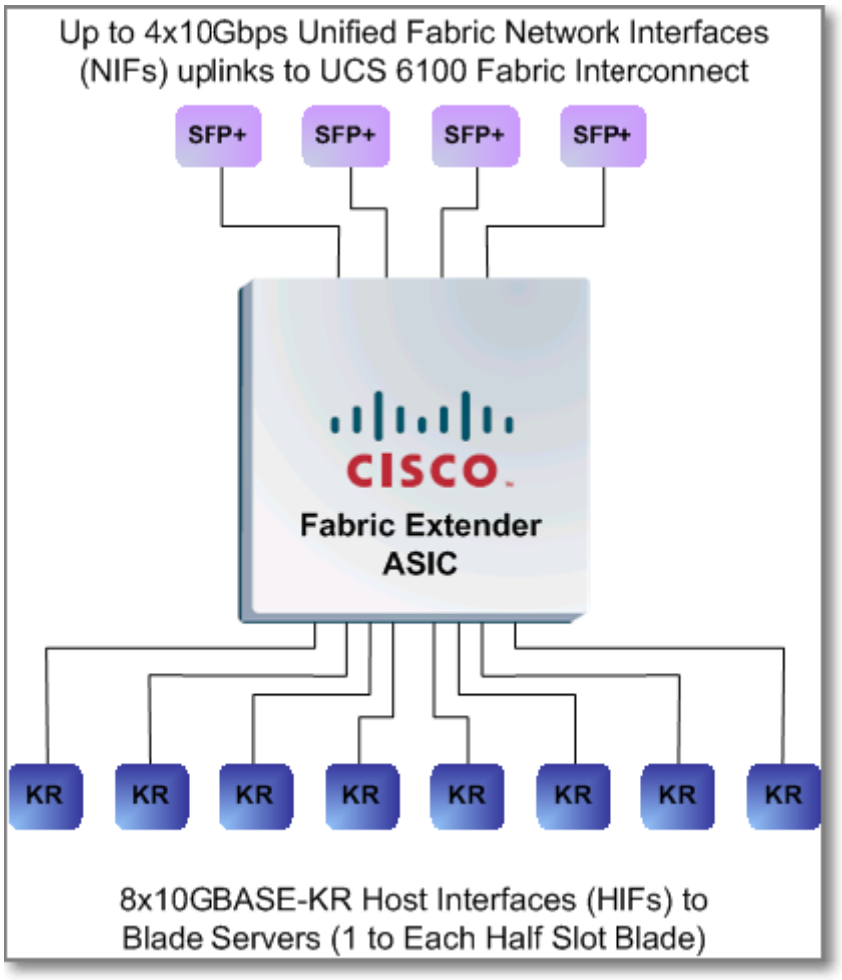
- 10 Gigabit Ethernet, FCoE capable, SFP+ ports
- 20 or 40 fixed port versions with expansion slots for additional Fiber Channel and 10-Gigabit Ethernet connectivity
- Up to 1.04 Tb/s of throughput
- Hardware based support for Cisco VN-Link technology
- Can be configured in a cluster for redundancy and failover capabilities
- Hot pluggable fan and power supplies, with front to back cooling system

Cisco UCS 2104 Fabric Extender

The Cisco UCS 2104 fabric extender brings the I/O fabric to the blade server chassis and supports up to four 10-Gbps connections between blade servers and the parent fabric interconnect, simplifying diagnostics, cabling, and management. The fabric extender multiplexes and forwards traffic using a cut-through architecture over one to four 10-Gbps unified fabric connections. Traffic is passed to the parent fabric interconnect, where network profiles are managed by the fabric interconnects. Each of up to two fabric extenders per blade server chassis has eight 10GBASE-KR connections to the blade chassis

mid-plane, with one connection to each fabric extender from each of the chassis' eight half slots (see Figure 2-9). This configuration gives each half-width blade server access to each of two 10-Gbps unified fabric connections for high throughput and redundancy.

Figure 2-9 Fabric Extender ASIC



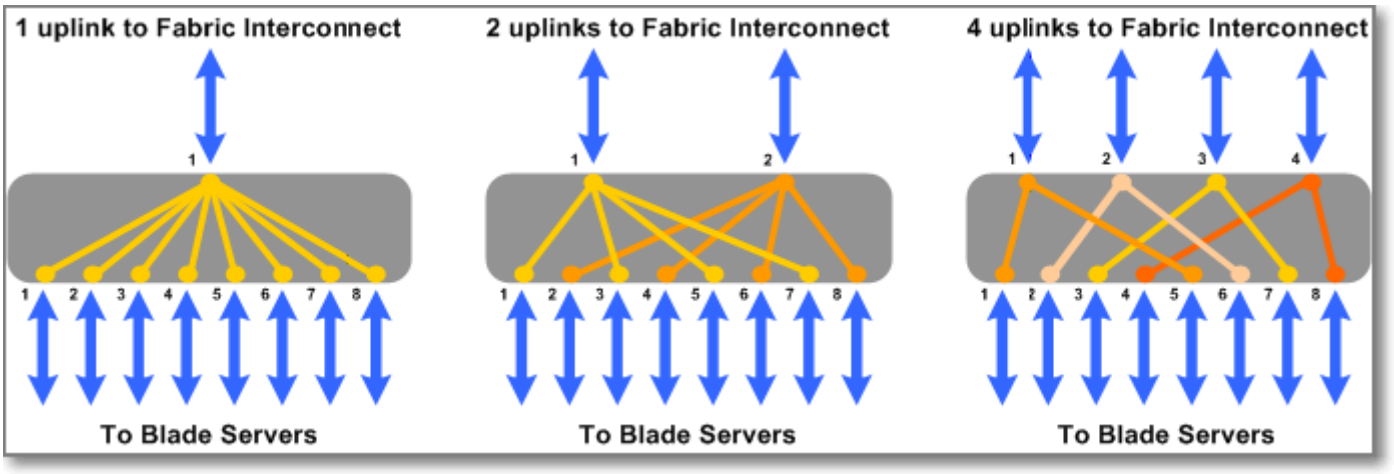
On the UCS 2104 fabric extender, the blade server facing host interfaces (HIFs) are statically pinned to the network interfaces (NIFs) that uplink to the UCS 6100 Series fabric interconnects. One, two, or four NIFs from the UCS 2104 fabric extender to the UCS 6100 Series fabric interconnect are supported; depending the number of NIFs, the HIFs are statically pinned according to Table 2-1:

Table 2-1 HIF Static Pin

NIF-1	NIF-2	NIF-3	NIF-4	Blade 1	Blade 2	Blade 3	Blade 4	Blade 5	Blade 6	Blade 7	Blade 8
On	Off	Off	Off	NIF-1	NIF-1	NIF-1	NIF-1	NIF-1	NIF-1	NIF-1	NIF-1
On	On	Off	Off	NIF-1	NIF-2	NIF-1	NIF-2	NIF-1	NIF-2	NIF-1	NIF-2
On	On	On	On	NIF-1	NIF-2	NIF-3	NIF-4	NIF-1	NIF-2	NIF-3	NIF-4

Figure 2-10 shows HIFs to NIFs static pins of Table 2-1 in graphical format:

Figure 2-10 HIFs to NIFs Static Pins



Note

- A 3-NIFs setup is not supported. The system uses two NIFs if you try to configure a 3-NIFs setup.
- The HIFs to NIFs pinning is fixed and not user configurable. As such, when the chassis is not fully populated with blade servers, the intended use and placement of a blade server in a slot need to be considered to make the best of use of the available bandwidth. In other words, some servers may share the bandwidth as depicted in Figure 2-10.
- If the number of links between the fabric interconnect and the fabric extender is changed, the chassis must be re-acknowledged to have the traffic rerouted or repinned with the new links setup.
- There is no dynamic repinning of HIFs to NIFs; when one NIF fails, all HIFs that are pinned to that NIF are brought down as well.

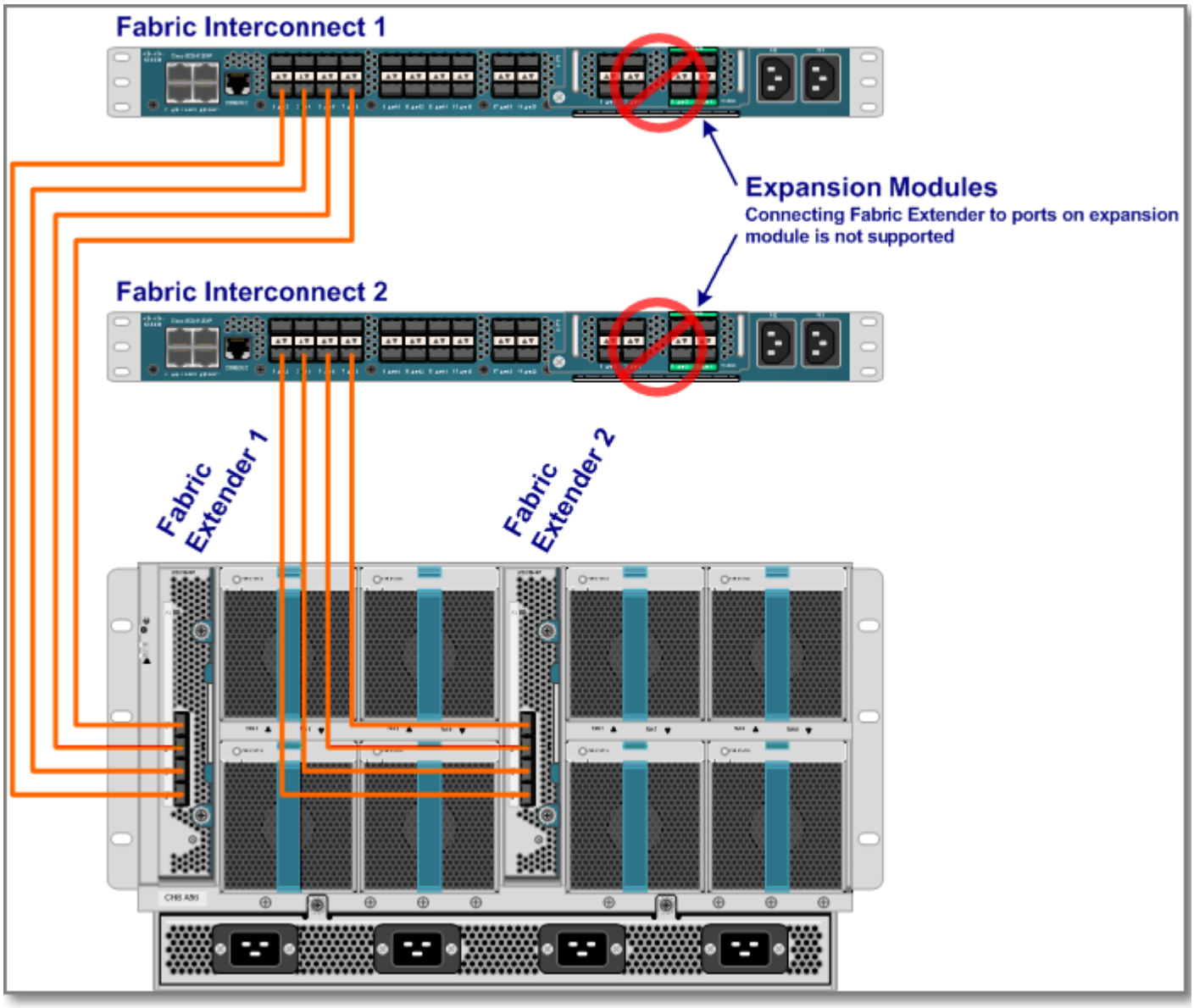
The number of links between the UCS 2104 fabric extender and UCS 6100 Series fabric interconnect depends on bandwidth requirements of applications running on blade servers. Table 2-2 summarizes the available bandwidth for a blade server chassis and for each individual blade server if they are evenly loaded on all server links. Figure 2-10 assumes that the blade server chassis is fully populated with eight half-slot blade servers, each with one network adapter; and each blade server is configured to use both network interfaces for forwarding data traffic.

Table 2-2 UCS LAN

Number of Uplinks from each Fabric Extender to Each Fabric Interconnect	1-link	2-link	4-link
Available bandwidth per blade server chassis with:			
• One Fabric Extender	10 Gbps	20 Gbps	40 Gbps
• Two Fabric Extenders (in active-active mode)	20 Gbps	40 Gbps	80 Gbps
Available bandwidth per blade server with:			
• One Fabric Extender	1.25 Gbps	2.5 Gbps	5 Gbps
• Two Fabric Extenders (in active-active mode)	2.5 Gbps	5 Gbps	10 Gbps

Each UCS 2100 fabric extender in a UCS 5100 blade server chassis is connected to a 6100 Series fabric interconnect for redundancy and bandwidth aggregation. Each fabric extender provides four 10 GE ports to the UCS 6100 Series fabric interconnect. Figure 2-11 shows the physical location of the UCS 2104 Fabric Extender within the back of UCS 5100 blade server chassis. Figure 2-11 also shows the proper way to connect the UCS 2104 fabric extenders to two separate UCS 6100 Series fabric interconnects; each UCS 6100 Series fabric interconnects must be connected to the chassis through its own fabric extender.

Figure 2-11 Proper Connection Between the Server Chassis and two UCS 6100 Series Fabric Interconnects



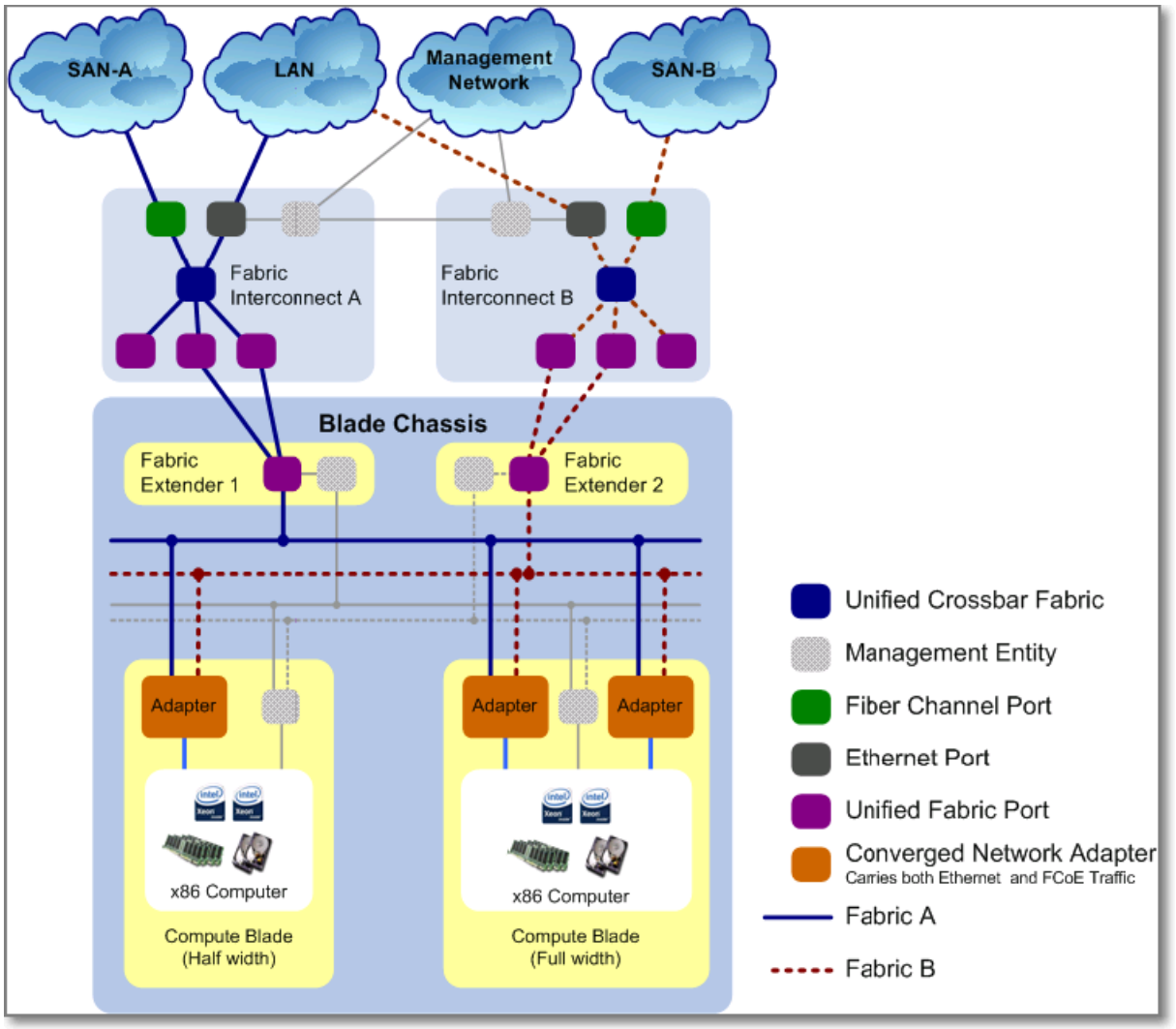
Note

- All ports of a UCS 2104 fabric extender can be connected to one UCS 6100 Series fabric interconnect only. If a connection to a second UCS 6100 Series fabric interconnect is required, a second UCS 6100 Series fabric extender must be added to the server chassis. Furthermore, all ports of the second UCS 2104 fabric extender can be connected to the second UCS 6100 Series fabric interconnect only.
- Connecting the UCS 2104 fabric extender to ports on the expansion modules of UCS 6100 Series fabric interconnect is not supported. While similar in appearance to the other fixed ports on the UCS 6100 Series fabric interconnect, the expansion modules are never used for direct server chassis connections.

Overall Network Architecture

The UCS components integrate into a fully redundant system. A fully redundant UCS is composed of two distinct, independent unified fabric planes ('Fabric A' and 'Fabric B'), with each fabric plane comprised of fabric interconnect connected to a fabric extender in each blade server chassis. These two fabric planes in UCS are completely independent of each other with respect to data forwarding, failure domain, and management perspective. The two unified fabric planes share no operational states. All network end-points such as host adapters and management entities are dual attached to both fabric planes, working in active-active manner, making full use of both fabric planes. A UCS does not need both fabric planes to be operational. It can function with just one fabric plane in case the other fabric plane is either not provisioned or is down for some reason. Figure 2-12 shows a logical view of the two unified fabric planes within a UCS.

Figure 2-12 Components of a Fully Redundant Unified Computing System



Ethernet Switching Mode

The Ethernet switching mode determines how the fabric interconnect behaves as a switching device between the servers and the network. The Fabric Interconnect operates in either of the following Ethernet switching modes:

- Switch Mode, page 2-17
- End-Host Mode, page 2-18

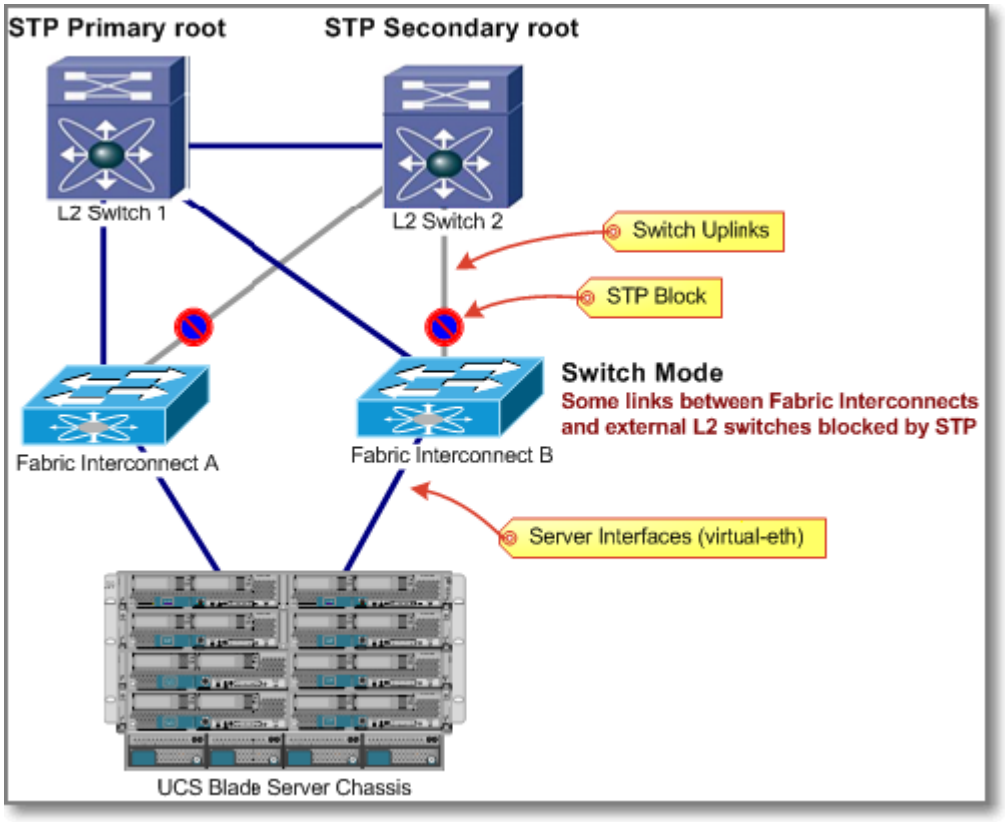
Switch Mode

Switch mode is the traditional Ethernet switching mode. Use switch mode when the Fabric Interconnect is directly connected to a router or when Layer 3 aggregation or VLAN in a box is used upstream. Switch mode has the following attributes:

- The Fabric Interconnects run Spanning Tree Protocol with connected external switches to avoid forwarding loops; broadcast and multicast packets are handled in the traditional way.

- The Fabric Interconnects run Per VLAN Spanning Tree Plus (PVST+), which cannot be changed to other STP such as MST.
- You cannot configure STP parameters, such as bridge priority and hello timers.
- MAC address learning and aging are enabled on both the servers links and the uplinks, operating like a typical layer2 switch.
- Some uplinks connected to external switches may be blocked by STP rules.

Figure 2-13 Fabric Interconnect Operating in Switch Mode



End-Host Mode

End-Host mode allows the Fabric Interconnect to act as an end-host with multiple network adapters to an external Ethernet network. End-host mode is the default Ethernet switching mode and should be used when following are upstream:

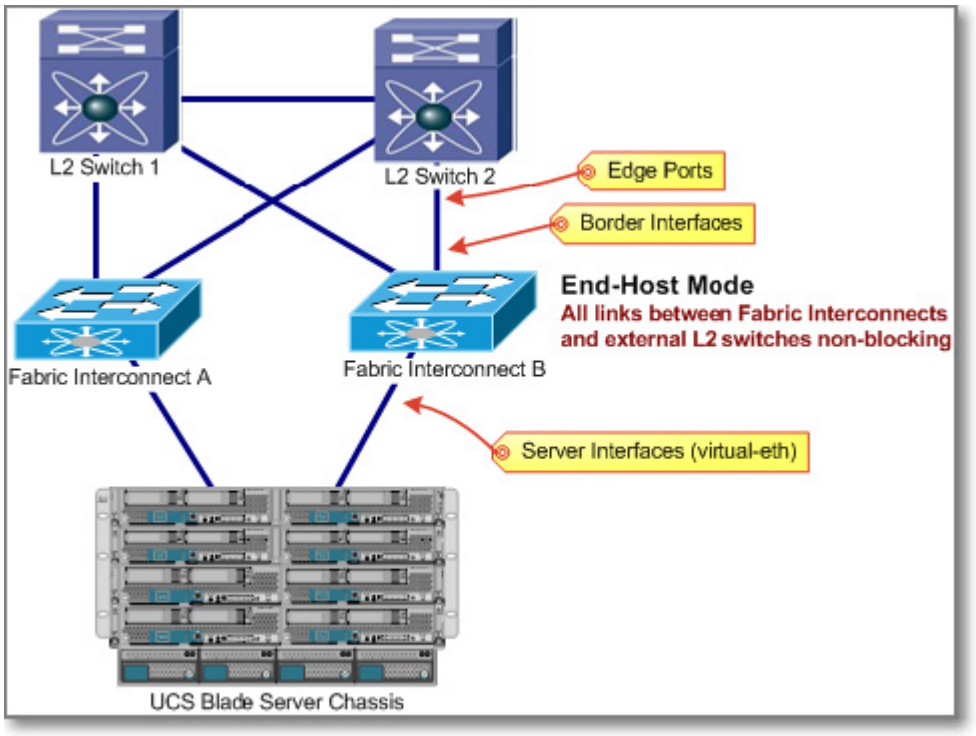
- Layer 2 switching for Layer 2 Aggregation
- Virtual Switching System (VSS) aggregation layer
- Virtual Port-channel

Some of the features of End-Host mode are:

- Spanning Tree Protocol is not use on Ethernet ports on the Fabric Interconnect for loop prevention (STP is still running on the Fabric Interconnect, but the system neither sends BPDU and processes received BPDU).
- The Ethernet ports on the Fabric Interconnect are not configured by default, network administrator must explicitly configure the Ethernet ports to be one of the two types:

  - Uplink Port/Link (Border Interface) - connect to upstream Layer 2 network.
  - Server Port/Link (Server Interface) - connect to blade servers.
- MAC address learning on the uplink ports is disabled; MAC address learning is enabled only on the server ports; each MAC address learned over server interface is pinned to a border interface, which provides redundancy toward the network and makes the uplink port appear as end host to the rest of the network.
- Learned MAC addresses never age out until the server port goes down or is deleted.
- MAC address move is fully supported within the same Fabric Interconnect and across different Fabric Interconnects.
- Fabric Interconnect operating in End-Host mode still switches locally connected servers.
- The Fabric Interconnect listens for broadcasts and multicasts only on a single uplink port per Fabric Interconnect
- Traffic forwarding among uplink ports is prevented.
- All uplink ports should be connected to the same Layer 2 network.
- All uplink ports are used for traffic forwarding; active/active use of uplink ports irrespective of the number of uplink Layer 2 switches connected.
- More scalable than switch mode since the control plane is not stressed as an Layer 2 switch.

Figure 2-14 Fabric Interconnect Operating in End-Host Mode

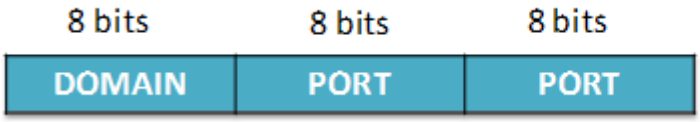


N-Port Virtualization Mode

Server virtualization uses virtual machine technology to prevent proliferation of physical servers in the data center. To to be managed as a unique entity on the storage area network, each virtual server requires a separate address on the fabric. The N-Port virtualization feature supports independent management and increased scale of virtual machines in the data center.

With the increased use of blade center deployments and top-of-rack aggregation devices in customer storage area network (SAN) environments, the deployment and use of aggregation switches is becoming more widespread. Because of the nature of Fibre Channel technology, several concerns must be addressed when deploying large numbers of edge switches. One major concern when designing and building Fibre Channel-based SANs is the total number of switches or domains that can exist in a physical fabric. As the edge switch population grows, the number of domain IDs becomes a concern. The domain is the address of a physical switch or logical virtual fabric; the domain ID is the most significant byte in an endpoint Fibre Channel ID.

Figure 2-15 Fibre Channel ID



The switch uses this Fibre Channel ID to route frames from a given source (initiator) to any destination (target) in a SAN fabric. This 1 byte allows up to 256 possible addresses. Some domain addresses are used for well-known addresses, and others are reserved for future expansion. The Fibre Channel standard allows for a total of 239 port addresses; however, a fabric of that size has not been qualified.

Another design concern is interoperability with third-party switches. In the past, different SAN fabric vendors interpreted the Fibre Channel addressing standard differently. In addition, some vendor-specific attributes used for switch-to-switch connectivity (or expansion port [E-Port] connectivity) made connection of switches from different vendors challenging, leading customers to implement edge switch technology that matched the core director type in the fabric. Management of this complex system becomes a concern as smaller form-factor devices are used to aggregate multiple host connections. Typically, this complexity leads to shared responsibility between platform engineering or server operations and fabric operations. The delineation of management responsibilities is blurred.

N-Port ID Virtualization (NPIV) and N-Port Virtualizer address these concerns.

N-Port ID Virtualization

NPIV allows a Fibre Channel host connection, or N-Port, to be assigned multiple N-Port IDs or Fibre Channel IDs (FCIDs) over a single link. All FCIDs assigned can now be managed on a Fibre Channel fabric as unique entities on the same physical host. Different applications can be used in conjunction with NPIV. In a virtual machine environment where many host operating systems or applications are running on a physical host, each virtual machine can now be managed independently from zoning, aliasing, and security perspectives.

Figure 2-16 NPIV-Aware Server Host Connection

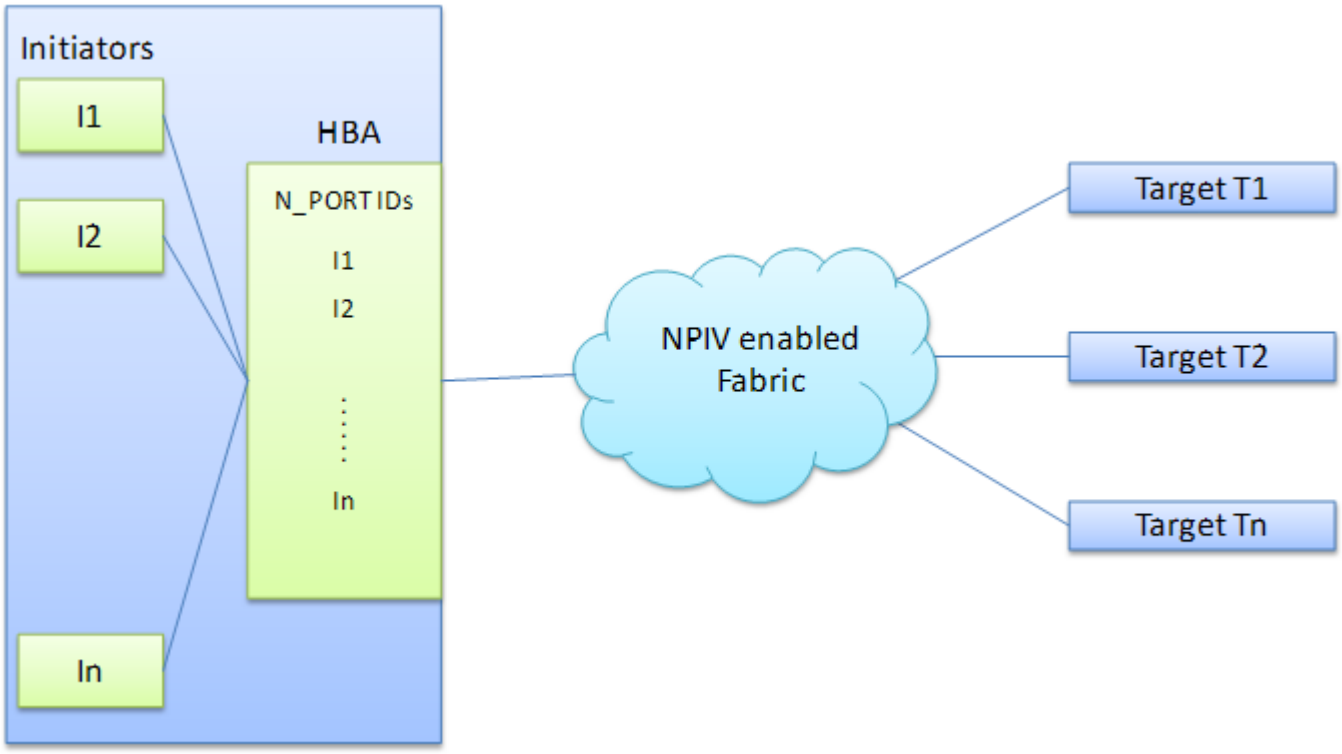


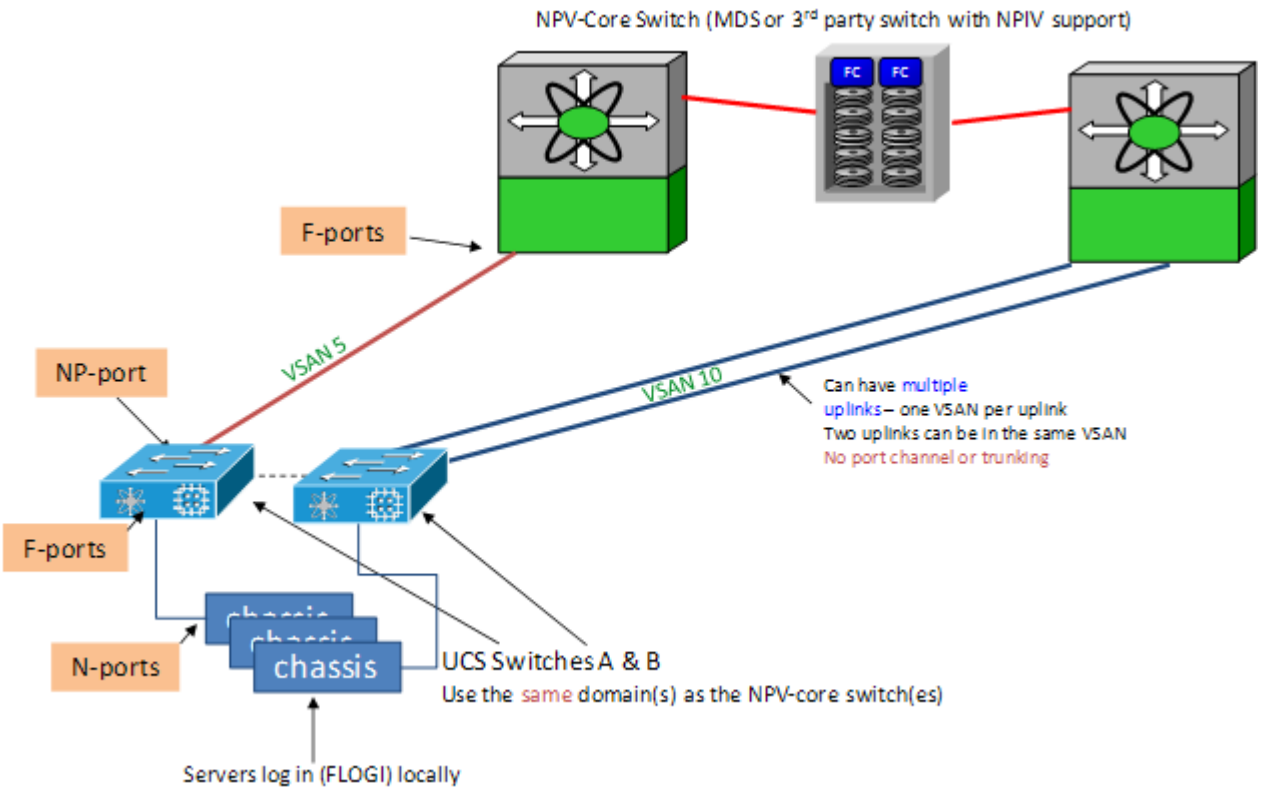
Figure 2-16 depicts intiators (virtual machines in an ESX environemnt) being mapped to a single link to communicate through the fabric to the target storage arrays.

In a Cisco MDS 9000 family environment, each host connection can log in as a single virtual SAN (VSAN). A VSAN is a logical partition of a larger group of physical ports that share a set of fabric services from a management domain. The VSAN architecture is analogous to VLAN deployments in Ethernet networks. When using NPIV in this environment, each subsequent FDISC login is a part of the same VSAN as the original fabric login.

N-Port Virtualizer

An extension to NPIV is the N-Port Virtualizer feature. The N-Port Virtualizer (NPV) feature allows the blade switch or top-of-rack fabric device to behave as an NPIV-based HBA to the core Fibre Channel director. The device aggregates locally connected host ports or N-Ports into one or more uplinks (pseudo-interswitch links) to the core switches.

Figure 2-17 NPV Overview



When you enable NPV on the edge switch, it is turned into a Fibre Channel passthrough to the NPIV core switch, which handles the FibreChannel Protocol (FCP). In this mode, the edge switch only informs the fabric of state changes.

Several NPV uplinks can be connected to either a single or multiple core directors. To date, the Cisco implementation allows connectivity from each edge switch to multiple core directors. The core NPIV directors must be part of the same fabric.

Cisco implementation recommends dynamic and VSAN-based load balancing.

In dynamic load balancing (all uplinks in the same VSAN), as hosts log in to the NPV device, the FDISC logins are sent down the links in a round-robin fashion. The I/O from the end device always follows the NPV path down to the same core switch to which its FDISC login was sent. In a stable fabric, each link should be utilized equally. For example, if there are eight hosts and two uplinks to core switches, each link has four initiators using each link. Should an uplink fail, hosts that were logged in and using this uplink as an uplink to the core switch are logged out of the fabric and go through the FDISC login procedure down the remaining link. In the initial release of the N-Port Virtualizer feature, there is no preferred uplink configuration that dynamically sends the host back down its original link. The host must be reinitialized for the host to use the recovered path. This manual assignment of hosts to preferred uplinks will be implemented in a later release of firmware.

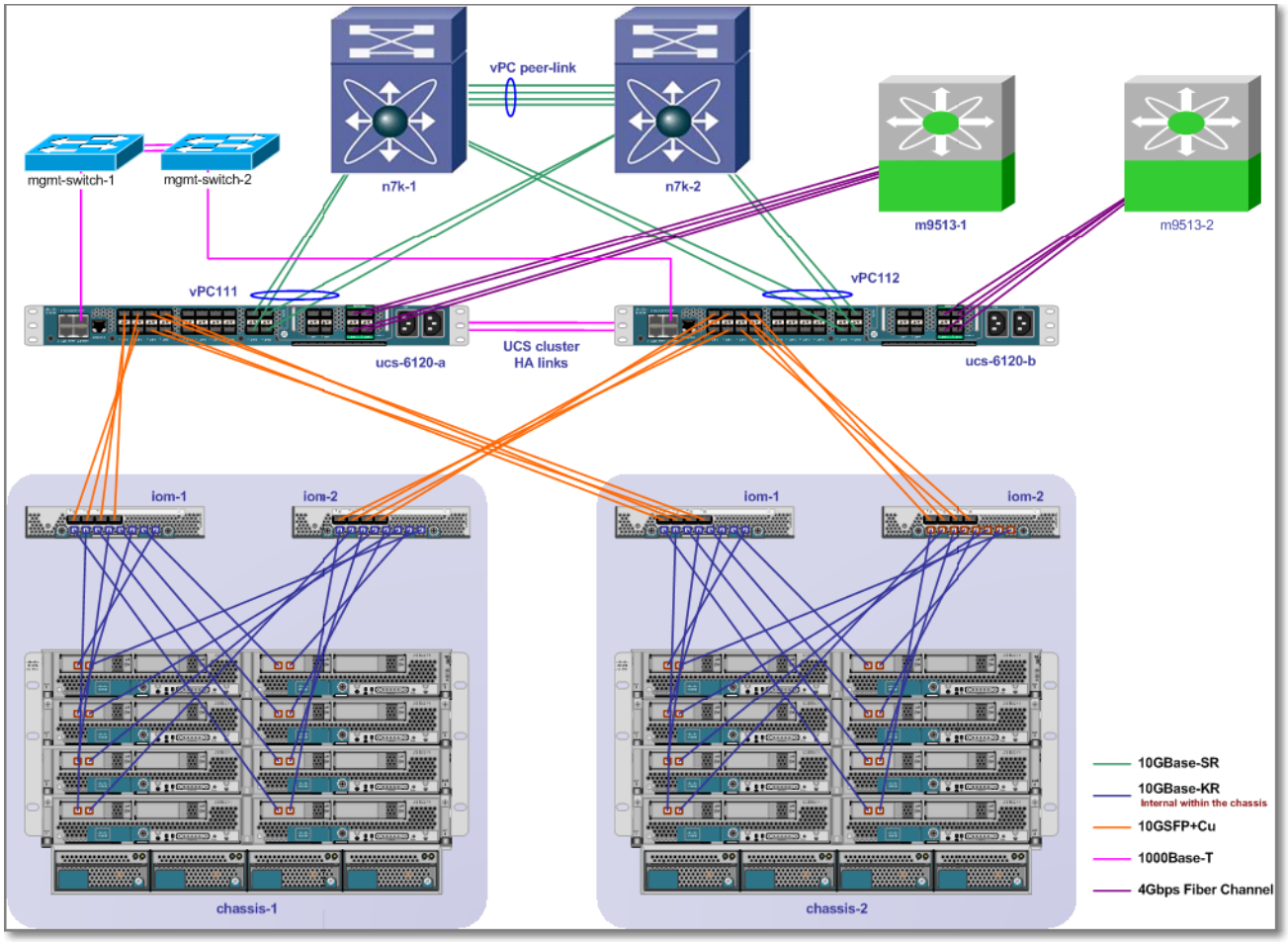
NPV edge devices can be connected to a core Cisco MDS 9000 family switch in multiple VSANs. Each N-Port follows the login procedure for that specific VSAN's Fibre Channel name server. The host ports on the switch are configured to match the VSAN that contains the uplink. All hosts FDISC logins are be sent down the matching VSAN tag of the uplink port. If there are multiple uplinks in the same VSAN, the FDISCs are round-robin load balanced down those links that match the VSAN ID. The N-Port

Virtualizer feature allows transparent connectivity to any core switch that supports the NPIV feature. If the core switch follows the standard NPIV implementation, the interoperability of different switch vendors is no longer a concern.

LAN Design

Figure 2-18 displays the physical network layout for the UCS cluster; for clarity, the fabric extenders (also known as I/O Modules or IOM) within the UCS chassis are displayed outside the chassis to show the internal 10Base-KR links between the Fabric Extenders and the blade servers.

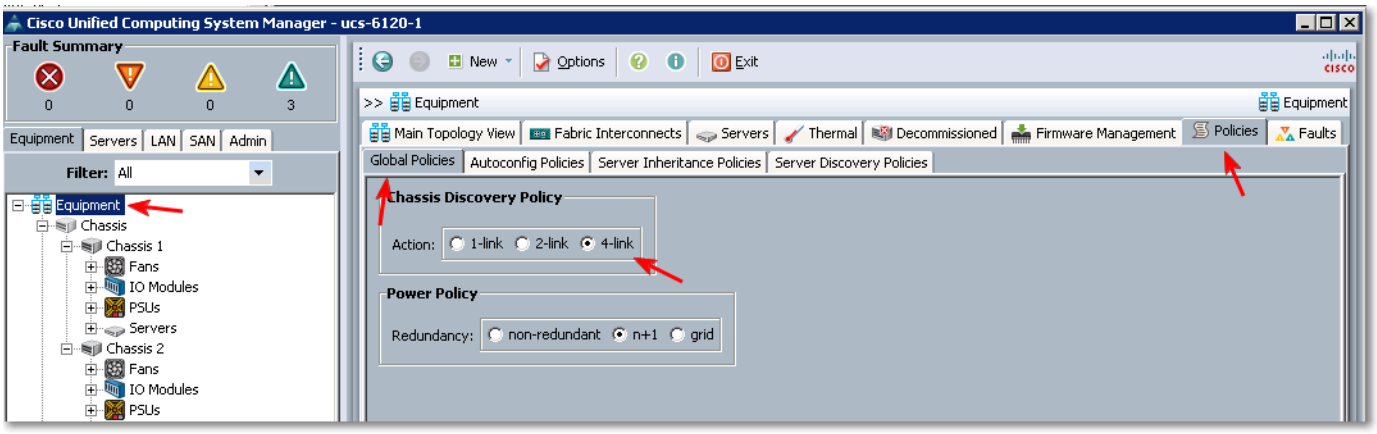
Figure 2-18 Physical Network Layout



4-link Fabric Extender Setup

The setup has two chassis. Each chassis is populated with eight Cisco-M200-M1 half-slot blade servers and two Fabric Extenders for and bandwidth redundancy. Each Fabric Extender has four uplinks to Fabric Interconnect. By default, the UCSM is configured with 1-link between Fabric Interconnect and Fabric Extender for the purpose of chassis discovery. To change the Chassis Discovery Policy to 4-link, navigate to **Equipment > Policies Tab > Global Policies** sub-tab, and select **4-link**.

Figure 2-19 Configuring Chassis Discovery Policy



Note The Chassis Discovery Policy determines how the system responds when a new chassis is added to the cluster. The Chassis Discovery Policy automatically configures the chassis for the number of links between the chassis and the Fabric Interconnect defined in the policy.

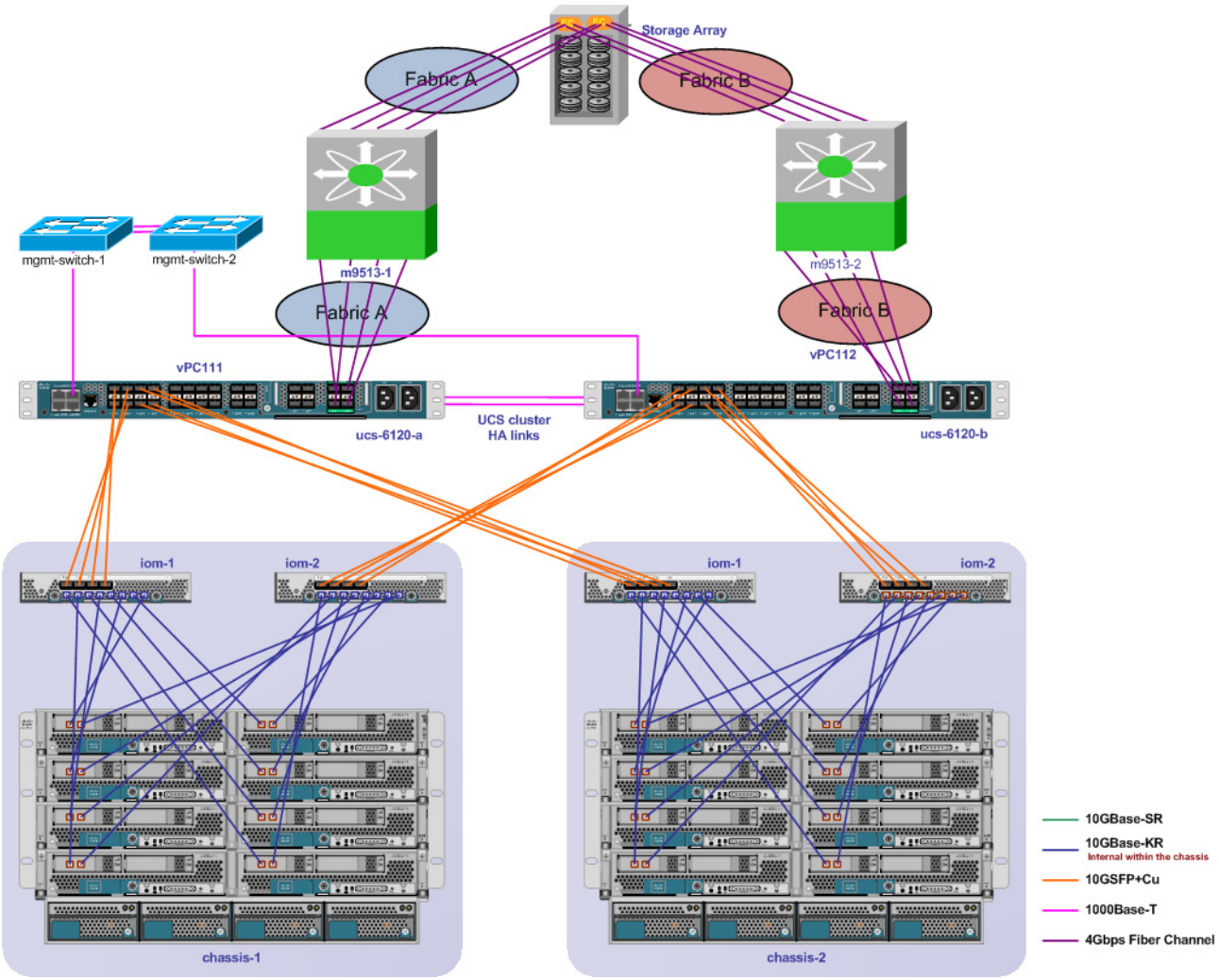
SAN Design

UCS blade servers connect to the SAN fabric via the Cisco UCS 6120XP Fabric Interconnect,which uses an 8-por,t 4-Gb Fibre Channel Expansion Module to access the SAN. These ports require SPF+ adapters. The ports can operate at 1/2/4 Gbps. Figure 2-20 shows the physical connectivity between the UCS 6120XP and the Cisco MDS 9513 Multilayer Director storage switch. Each UCS 6120XP is connected to one MDS 9513 to form their own fabric. Four 4-Gbps FC links connect the UCS 6120 to MDS 9513. The MDS 9513 are then connected to the storage controllers. In this topology, the storage array has two controllers. Each MDS 9513 has two connections to each FC storage controller. These dual connection provide redundancy if a FC controller fails and the MDS 9513 is not isolated.

Note Four additional expansion modules are available for the 6120 XP,

- Ethernet module that provides 6 ports of 10 Gigabit Ethernet using the SFP+ interface
- Fibre Channel plus Ethernet module that provides 4 ports of 10 Gigabit Ethernet using the SFP+ interface; and 4 ports of 1/2/4-Gbps native Fibre Channel connectivity using the SFP interface
- Fibre Channel module that provides 8 ports of 1/2/4-Gbps native Fibre Channel using the SFP interface for transparent connectivity with existing Fibre Channel networks
- Fibre Channel module that provides 6 ports of 1/2/4/8-Gbps native Fibre Channel using the SFP or SFP+ interface for transparent connectivity with existing Fibre Channel networks

Figure 2-20 SAN Layout



The FC portion of the expansion modules runs a snapshot of the Cisco NX-OS. The FC ports run in NPV and NPIV modes only. It is always attached to an NPIV enabled northbound fabric switch. Directly attached storage (DAS) is not supported. By default, all FC ports are enabled when the UCS 6120XP is powered on. There is no inherent HA designed into the UCS for FC. In this solution, the SAN design must include a dual, redundant fabric design and multipathing drivers (VMware) to enable HA.

VSAN setup

The basic steps to configuring the VSAN on the UCS are as follows:

1. Create the VSAN on UCS and the MDS
2. Associate the VSAN to the UCS/MDS physical interface
3. Assign the vHBA service profiles to the VSANs

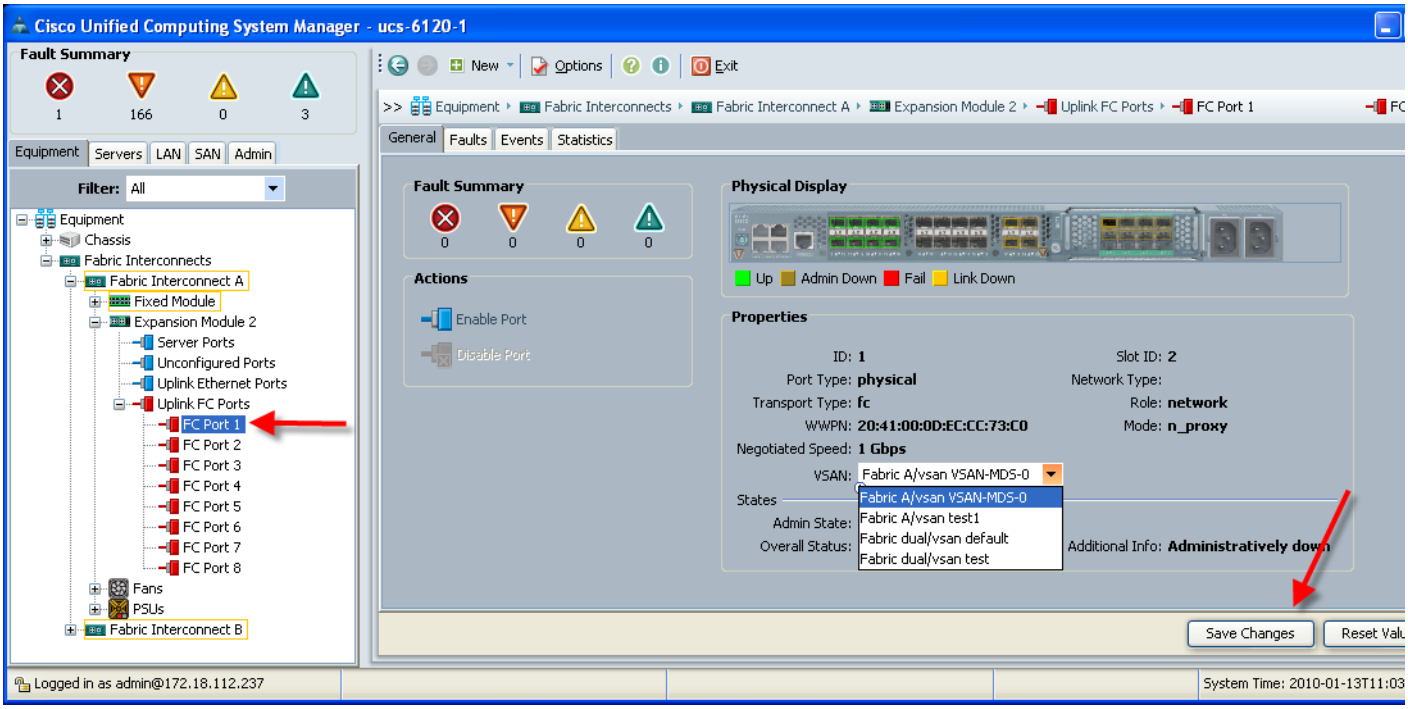
Create the VSAN definition for each fabric. For each VSAN created, a separated VLAN is also required. Fibre Channel is carried over IP (FCoE), which requires a unique VLA.

From the UCS Manager GUI, the VSANs are created on the SAN tab. In the SAN tab, select **FC Uplinks > Fabric A > VSANs**. Right-click the **VSANs** and select **Create VSAN**. On the screen that appears, select where the VSAN will reside. To build dual fabrics for redundancy, select **Fabric A** for this VSAN.

Associating the VSANs to Physical FC ports

To assign a VSAN to the FC port from the UCS Manager, select **Equipment > Fabric Interconnects > Fabric Interconnect A > Expansion Module 2 > Uplink FC ports**. Select the FC port that needs VSAN assignment. Thee VSANdrop down shows common VSANs and those on this fabric. Select the VSAN and click **Save changes**. Repeat for other FC ports in Fabric Interconnect B. UCS 6120XP does not support F-port trunking or port-channels. Assign VSANs on a per port basis and use two FC ports in a VSAN for redundancy.

Figure 2-21 Assigning VSANs to FC ports



vHBA Template

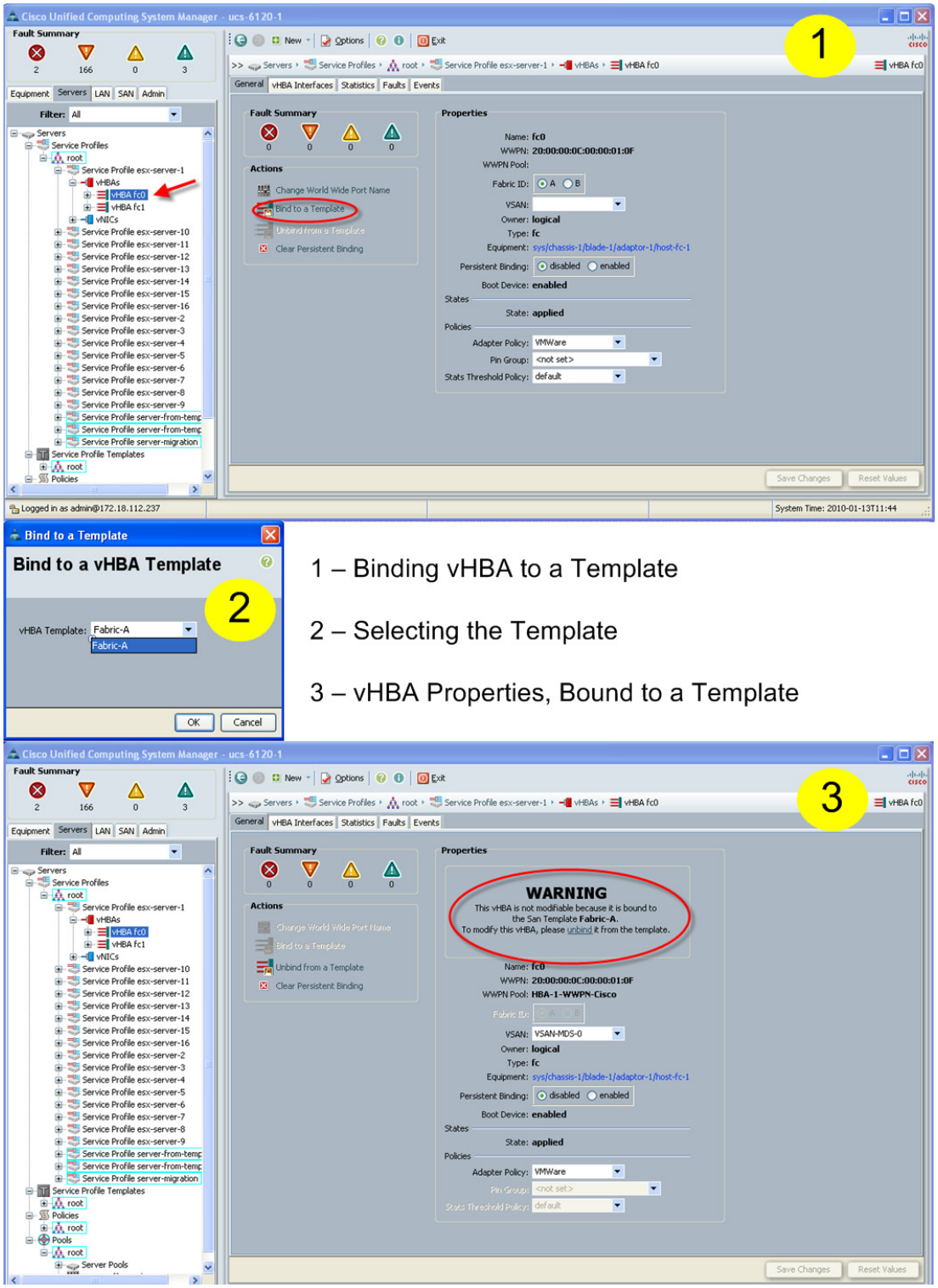
vHBA templates are useful if the solution has a large number of vHBAs that have similar configuration, such as VSANs and WWWN pool. To create a vHBA template, navigate to SAN tab and click **Policies > vHBA templates**. Right-click and select **Create vHBA template**. Enter the details such as Name, description, VSAN, and WWWN pool. Save the changes. Repeat to create a template for Fabric B.

Assigning a vHBA template to the vHBA

To assign the vHBAs to the templates, select **Servers > Service Profiles > root > 'an existing service profile' > vHBAs > vHBA interface** from Service tab. In the right pane, click **Bind to a Template** and then select the vHBA template and click **OK**. Once the vHBA template is assigned to a vBHA interface,

a warning message states that the vHBA properties cannot be changed in there. To change the properties, edit the vHBA template or unbind the vHBA from the template to make changes. Changes to the vHBA template affects all vHBA interfaces bound to that template.

Figure 2-22 Assigning a vHBA template to vHBA



Note If the vHBA is already created, make the vHBA template an Updating Template. Once the vHBA template is assigned to a vHBA and the template is an Updating Template, changes made to the template are passed to the attached vHBA immediately.

Licensing

Every UCS 6120XP comes with FC and FCoE license enabled at no additional cost. However, the fixed ports are licensed. By default, only the first 8 ports are licensed to connect to the UCS chassis. Usings these ports, we can connect two UCS chassis to the 6120XP using a 4-link configuration. If additional UCS chassis need to be connected to the 6120XP, additional port licenses must be purchased. In the case of UCS 6140XP, the first 16 ports are licensed. Expansion modules include all licensing (ports and features).

UCS Manager

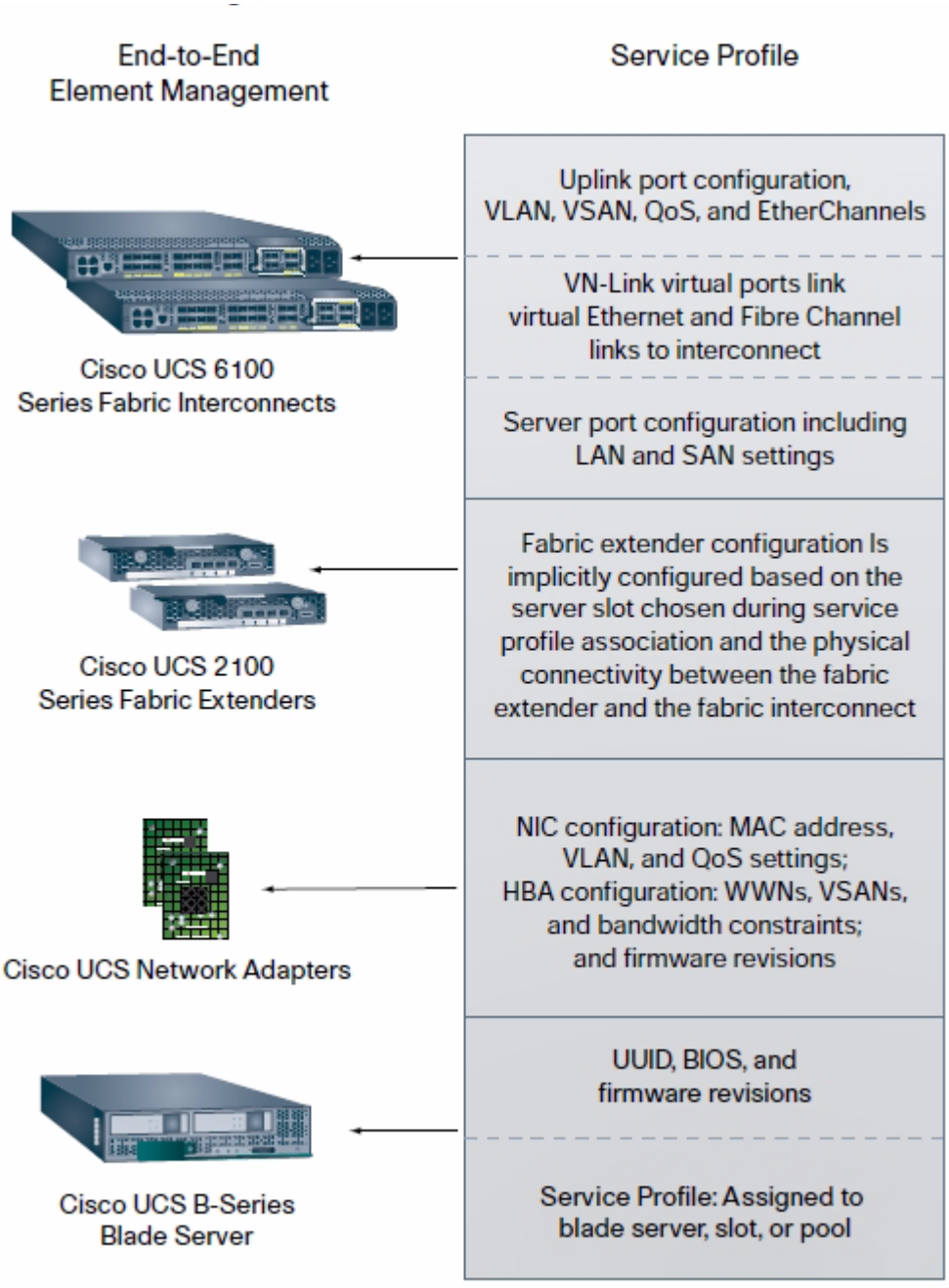
Cisco UCS Manager centralizes management, creates a unified management domain, and serves as the central administrative interface for the Cisco Unified Computing System. Cisco UCS Manager is embedded device-management software that manages the system from end to end as a single logical entity through a GUI, the command-line interface (CLI), or an XML API.

In Cisco UCS, you can use multi-tenancy to divide up a large physical infrastructure into logical entities called organizations. You can achieve a logical isolation between organizations without providing a dedicated physical infrastructure for each organization.

In a multi-tenant environment, you can assign unique resources to each tenant through the related organization. These resources can include different policies, pools, and quality of service definitions. You can also assign locales to Cisco UCS user privileges and roles by organization to restrict access to specific organizations. Cisco UCS Manager implements role and policy-based management using service profiles and templates.

If you set up a multi-tenant environment, all organizations are hierarchical. The top-level organization is always root. The policies and pools you create in root are system-wide and are available to all organizations in the system. However, policies and pools created in other organizations are only available to those organizations below it in the same hierarchy.

Figure 2-23 Cisco UCS Service Profile Scope



Cisco UCS Manager resides on a pair of Cisco UCS 6100 Series fabric interconnects using a clustered, active-standby configuration for high availability. The Manager participates in server provisioning, device discovery, inventory, configuration, diagnostics, monitoring, fault detection, auditing, and statistics collection. It can export the system's configuration information to configuration management databases (CMDBs), facilitating processes based on Information Technology Infrastructure Library (ITIL) concepts. Cisco UCS Manager's XML API also facilitates coordination with third-party provisioning tools that can deploy virtual machines and install operating systems and application software on servers configured by Cisco UCS Manager.

Administration

The UCS Manager offers role-based management that helps organizations make more efficient use of administrator resources. Server, network, and storage administrators maintain responsibility and accountability for domain policies within a single integrated management environment eliminating the need for manual coordination among multiple disciplines. Administrators define the policies to provision compute infrastructure and network connectivity to automate basic server configuration. Roles and privileges in the system can be easily modified and new roles quickly created.

Role Based Access Control

With the Role-Based Access Control (RBAC) function of the Cisco Unified Computing System, you can control user access to actions and resources in the UCS. The UCS RBAC allows access to be controlled based on assigned user roles.

For more information on configuring RBAC and the UCS RBAC model, refer to:

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080ae0fd7.shtml

Server IP KVM Availability

The KVM console is a video over IP representation of the video output on the blade. With it, Cisco UCS offers access to blades similar to other KVM consoles in the blade industry. To use the KVM console to access the blade server, you must assign a pool of IP addresses as a management interface into the server blades. These IP addresse must be externally routable for remote access to servers via the KVM console.

Note The IP addresses are used in reverse order, i.e. if your IP address range is from 10.0.0.1 to 10.0.0.10, the first IP KVM address is 10.0.0.10. Once the block is used up, a new block must be defined and IP addresses are once again pulled from the end of the range.

It is not necessary to keep track of the IP addresses given to each server since the UCSM automatically uses the IP address it was given when a connection attempt is made.

IP KVM IP addresses must be on the same subnet as the management port of the UCSM because there is no tagging or L3 for connections from that port.

For more information on configuring IP KVM, refer to:

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080aefd13.shtml

Syslog Setup

Cisco Unified Computing System provides several diagnostic tools to aid in troubleshooting and monitoring the environment. Syslog is the mechanism for processes and scripts to write log entries. Callers can fully specify characteristics of log entries. A syslog daemon in the system captures logs and saves them in a rotating buffer. These logs can be viewed internally or exported to syslog collectors. Multiple syslog collectors can be specified through configuration.

For more information on configuring syslog with Cisco UCS Manager, refer to:

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080ae0f24.shtml

Note Although three syslog servers are configurable, the highest logging level set is applied to the host and sent out to each syslog server that has been configured.

Servers

Cisco UCS Manager uses service profiles to provision servers and their I/O properties. Service profiles are created by server, network, and storage administrators and are stored in the Cisco UCS 6100 Series fabric interconnects. They maintain configuration information about the server hardware, interfaces, fabric connectivity, and server and network identity.Service profiles are centrally managed and stored in a database on the fabric interconnect.

Note Every server must be associated with a service profile.

Service profile templates simplify the creation of service profiles, helping ensure consistent policies within the system for a given service or application. This automation reduces the number of manual steps that need to be taken, reducing human error, improving consistency, and reducing server and network deployment times.

Sservice profiles also dissociate hardware specific attributes from the design. If a specific server in the deployment is replaced, the service profile associated with the old server is applied to the newly installed server allowing for near seamless replacement of hardware if needed.

Service Profiles

Service profiles are central to blade management in the Cisco Unified Computing System (UCS). A service profile represents a logical view of a single blade server without the need to know exactly which blade you are talking about. Cisco UCS supports two types of service profiles:

- Service Profiles that Inherit Server Identity
- Service Profiles that Override Server Identity

The engineer deploying servers must decide which type of service profiles to use.

Timesaver For large deployments, service profiles should be created from templates for quick and consistent provisioning of many servers.

For details on the difference between types and how to configure service profiles, refer to :

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080af7515.shtml

For details on configuring service profiles from templates, refer to:

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080ae0642.shtml

Service Profile Policies

With UCS Manager, you can configure multiple polices that are specific to the service profile. Each policy applied is specific to the service profile it is configured for, although similar or identical policies can be applied to each service profile within the system. These policies include the following:

- **Server pool**—A server pool contains a set of servers that share the same characteristics. Those characteristics can be their location in the chassis or an attribute, such as server type, amount of memory, local storage, type of CPU, or local drive configuration. You can manually assign a server to a server pool or use server pool policies and server pool policy qualifications to automate the assignment.
- **vNIC connection**—A vNIC is configured through the LAN tab within the UCS Manager. In the design, a vNIC template was used to apply properties to the NIC hardware within each server. More information on how this was applied to the system can be found in the UCS Manager LAN section of this document.
- **Local disk configuration**—Local disk policies are configured and applied to define the type of local storage is to be applied to the server associated with the service profile. The options available are as follows:
  - No Local Storage
  - No RAID
  - RAID Mirrored
  - RAID Striped

  In this design, no local disks were used, and the No Local Storage policy was applied to each server because all servers used boot from SAN storage. If local disks are present for use, other policies may be preferred.
- **Serial over LAN (SoL)**—This policy sets the configuration for the serial over LAN connection for servers associated with service profiles that use the policy. By default, the serial over LAN connection is disabled. In this design, Serial Over LAN was not used.
- **Firmware**—Firmware policies are configurable and are used to set the overall server firmware policy for BIOS, NIC, HBA, and RAID controller. These policies are provided as a bundle and used when updating firmware settings for hardware installed and used in the server.
- **IPMI profile**—You can use this policy to determine whether IPMI commands can be sent directly to the server using the IP address. For example, you can send commands to retrieve sensor data from the BMC. This policy defines the IPMI access, including a username and password that can be authenticated locally on the server, and whether access is read-only or read-write.
- **Statistics**—Statistic threshold policies are configurable through the LAN tab within the UCS Manager. These policies can be configured to report error once a certain statistic has exceeded a certain set threshold value.
- **WWNN**—The WWNN policy allows configuration of the server WWNN to come from a pool or be derived from the hardware defaults.
- **Scrub policy**—This policy determines what happens to local data on a server during the discovery process and when the server is disassociated from a service profile. This policy can ensure that the data on local drives is erased at those times. In this design, the scrub policy was disabled, which optimized the time required to dissociate service profiles.

Warning **In a production environment, scrub policies should be used with caution.**

vNIC Configuration

You can view the vNIC configuration to be applied to each server in the server tabs service profile scope of UCS Manager. In this view, you can add or deletel vNICs.

vHBA Configuration

You can view the vHBA configuration applied to each server in the server tabs service profile scope if UCS Manager. In this view, you can add or delete vHBAs.

UCS Best Practice Resources

Best Practices in Deploying Cisco Nexus 1000V Series Switches on Cisco UCS Systems

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/white_paper_c11-558242.html

Unified Computing System Firmware Management Best Practices

http://www.cisco.com/en/US/products/ps10281/products_configuration_example09186a0080aee43e.shtml

Data Center Layer 2 Design

In this solution, the Layer 2 design includes the following aspects:

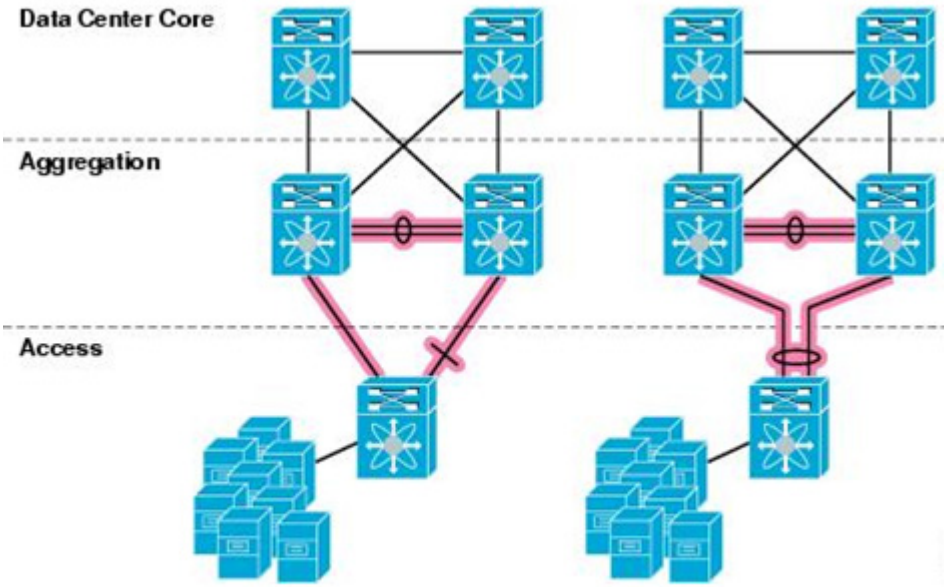
Access, Distribution, Core, and Services Layer Design

A data center composes building block replicated to achieve the desired scale. This replication provides redundancy through duplication of devices in each position. The traditional network design uses a pair of aggregation switches connected to access switches in a redundant way, is a typical example of redundancy. The two main drawbacks of this traditional solution are as follows:

- There is no Layer-2 multi-pathing for a particular VLAN, and the per-VLAN load balancing that enables the use of both uplinks of an access switch needs user configuration. There is no way to escape this constraint, as it dictated by the way bridging requires a spanning tree in the data plane.
- Using STP in the data plane introduces a convergence delay and potential risks.

Port channel technology solves many issues with the interconnection of two switches. Link aggregation alone cannot provide a fully redundant data center, as it does not protect against the failure of a single switch. Two technologies address this limitation. VSS (on the Catalyst 6500 Series switch) and vPC (on the Nexus 7000 Series switch) enable the creation of a Layer 2 port channel interface distributed across two switches. This limited step-up in channeling provides the building block required to build a data center with no dependency on the spanning tree model. Figure 2-24 shows this solution at a high level.

Figure 2-24 Data Center Designs



The left side illustrates the traditional design, where STP handles redundancy. One uplink on the access switch is in STP blocking state. The right side represents the VMDC solution that distributes the end of a channel across the two aggregation switches. The logical view shows that redundancy is hidden from STP and the STP blocked port is eliminated.

However, we recommend keeping STP on as a backup mechanism. Even if the redundancy is hidden to STP, it remains at a lower layer. It is just handled by a different mechanism. STP protects against a configuration error that breaks a channel into individual links.

In the VMDC design, Layer 2 domain spreads across access, collapsed distribution/core, and service layers. All links that connect these layers are port-channels and trunk ports carrying customer VLANs and control/management VLANs.

A key difference between this design and the traditional data center design is the use of virtual port-channels (vPCs) and multi-chassis EtherChannel (MCEC). vPC provides a loop-free topology and is configured on Nexus 7000 series switches used in the collapsed distribution/core layer. MCEC is configured on the VSS system.

The following sections provide an overview of the VLAN and spanning-tree (STP) designs in VMDC.

VLAN Design

Platforms used in this design support normal and extended VLANs in the range of 1-4094. VLANs 3968-4047, 4094 are reserved VLANs allocated for internal use. You cannot create, delete, or modify VLANs in this reserved VLAN range. VLANs excluding reserved VLANs are available for use as data, control, or management VLAN.

Table 2-3 lists VLANs used in the VMDC solution and their purpose:

Table 2-3 VLAN Designin VMDC

vlan	Description
10	Routed Vlan between VSS and 7600
20	Routed Vlan between VSS & 7600
100	fcoe vlan
191	VM PXE Boot
192	Device Management
193	ESX management
194	VM Management
195	vmotion
196	Nexus 1000V control
197	Nexus 1000V packet
198	VMknic Erspan
888	Management Servers
1100-1180	Bronze customer vlans
1301-1308	Bronze customer vlans
1401-1405	Silver customer1-5 vlans
1611-1613	Gold customer1 vlans
1621-1623	Gold customer2 vlans
1631-1633	Gold customer3 vlans
1641-1643	Gold customer4 vlans
1651-1653	Gold customer5 vlans
3001	VM Mgmt isolated

VLAN allowed lists on the trunks vary from device to device in the data center.

STP Design

In VMDC, Layer 2 switches support rapid PVST+ and multiple spanning-tree (MST). With rapid PVST+, a single spanning-tree topology is created and maintained for each VLAN in the Layer 2 domain. With a large number of VLANs in the Layer 2 domain, Rapid PVST+ is not the right choice for the following reasons:

- Rapid PVST+ has additional overhead involved in generating and processing BPDUs. The number of BPDUs transmitted by MST does not depend on the number of VLANs, as in Rapid-PVST.
- Rapid PVST+ has longer convergence during topology changes.

Therefore, the VMDC design implements multiple spanning trees (MST) as the spanning tree protocol. Layer 2 switches in the data center are configured in the same MST region with one MST instance. VLANs, except the reserved internal VLANs, are mapped to the same MST instance.

Tip When you change VLAN to MST instance mapping, the system restarts MST. The recommendation is to map VLANs (1-3967,4048-4093) to the MST instance at the time of initial configuration.

For more information on MST, refer to:

- Configuring Multiple Spanning Tree on Nexus 7000
- Configuring Multiple Spanning Tree on VSS

Many STP extensions are used in this solution. For more information on bridge assurance and other STP extensions, such as BPDU guard, BPDU filter, and root guard, refer to:

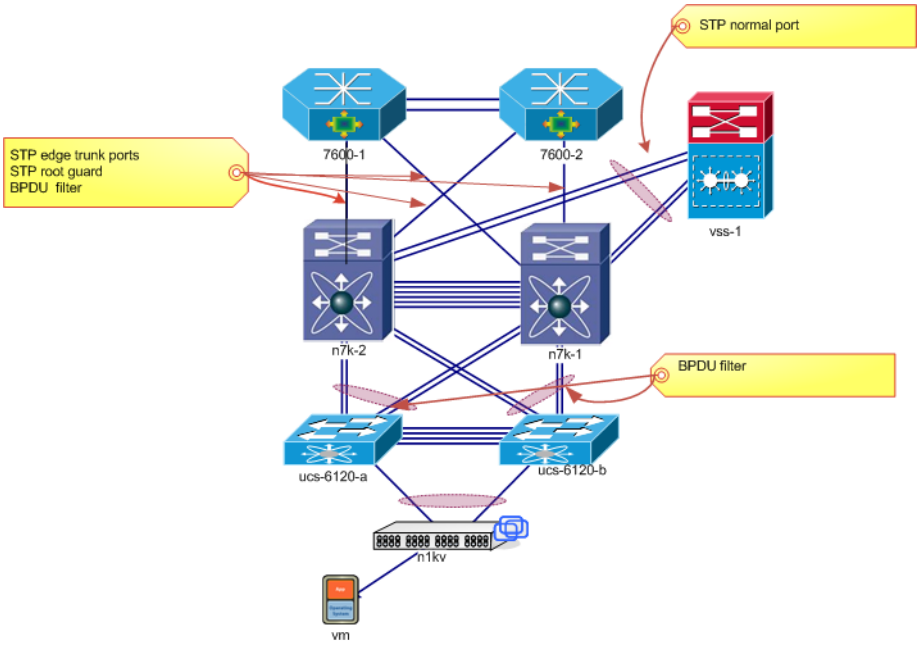
- Configuring STP Extensions on Nexus 7000

STP Logical Topology

The logical STP topology include the Nexus 7000 pair and the VSS chassis, which consists of a pair of Catalyst 6500 operating in virtual switch mode. The pair of UCS 6120 Fabric Interconnects are configured to operate in End host switching mode. In this mode, the fabric interconnects do not speak STP. To achieve a quick Layer 2 convergence during failure, the STP root is situated at the collapsed distribution and core layer, which in this design is the Nexus 7000 pairs. One of the Nexus 7000 devices acts as the primary root of the STP topology and the other acts as the secondary root of the topology.

Figure 2-25 gives an overview of the logical Layer 2 STP topology in this solution.

Figure 2-25 STP Logical Topology



Port-channel ports to the VSS are configured as normal STP ports. Port-channels to the fabric interconnects do not participate in STP. BPDUs should not be received on the port-channels from the Nexus 7000 to the fabric interconnects. The trunk uplinks on Nexus 7000 that connect to Cisco 7600 have STP edge port configuration with root guard and a BPDU filter.

Nexus 1000V Layer 2 Design

The Cisco Nexus 1000V is a software switch that uses the VMware vSphere framework to improve integration between server and network environments. The software switch was introduced with the release of vSphere 4.0 and brought many Cisco specific features not available in the traditional VMware vSwitch down to the software access layer. With policy-based connectivity, network security features, and network monitoring features in the software switch, you can deploy secure, tightly monitored, virtualized environments.

Typically, the access layer is direct link from physical server to physical switch. As the paradigm shifts from physical server to virtualized server, this connection must shift, which brings the access layer into software. The following differences between the Cisco Nexus 1000V and a physical switch exist:

- Joint management by network and server administrators
- **External fabric.** Supervisors and line cards in a physical switch have a shared internal fabric over which they communicate. The Cisco Nexus 1000V switch uses external fabric.
- **No switch backplane.** Line cards in a physical switch can forward traffic to each other on the switch's backplane. Since the Nexus 1000V switch lacks such a backplane, a VEM cannot directly forward packets to another VEM. Instead, it must forward the packet via some uplink to the external fabric, which then switches it to the destination.
- **No Spanning Tree Protocol.** The Nexus 1000V does not run STP because it deactivates all but one uplink to an upstream switch, preventing full utilization of uplink bandwidth. Instead, each VEM is designed to prevent loops in the network.
- **Port channels only for uplinks.** Uplinks in a host can be bundled in a port channel for load balancing and high availability. The virtual ports are not bundled into a port channel.

When implementing Cisco Nexus 1000V switch, consider the following:

- VMotion of the VSM VM is not supported. In particular, do not enable DRS for the VSM VM. VMotion and DRS are supported for other VMs connected to the Cisco Nexus 1000V switch.
- VMware fault tolerance is not supported for the VSM VM. It is unnecessary for the VSM as redundancy is present. It is supported for other VMs connected to Cisco Nexus 1000V switch.
- The snapshot of the VSM VM does not contain the configuration changes made since the snapshot was taken. Therefore, restoring the VSM VM from a snapshot may require care.
- Do not assign more than one uplink on the same VLAN without port channels. It is not supported to assign more than one uplink on the same host to a profile without port channels or port profiles that share one or more VLANs.

Nexus 7000 Layer 2 Design

A pair of Nexus 7000 switches are used as collapsed distribution/core devices aggregating Layer 2 traffic from server access PODs. Both Nexus 7000 switches have dual supervisor engines and three fabric modules for high availability and redundancy. Nexus 7000 switches are connected to other layers via 10Gig trunk uplinks or vPC port-channels.

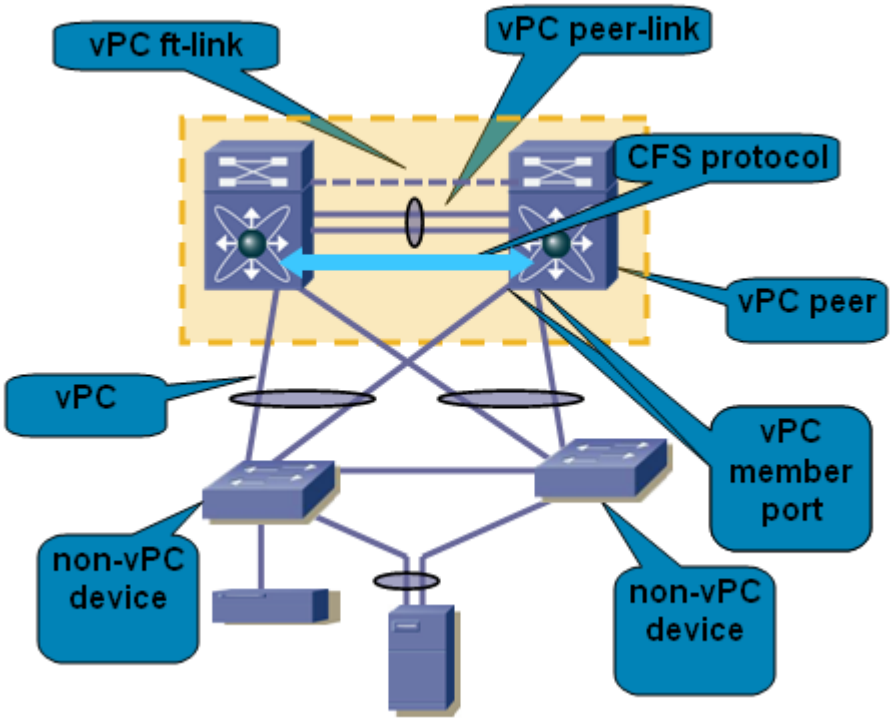
The Nexus 7000 device, n7k-1, is configured as the MST primary root bridge, and n7k-2 is the secondary root bridge. The uplinks to Cisco 7600 routers and the vPC links between 6120 and Nexus 7000 devices are configured as STP edge trunk ports. vPC link to the VSS switch is configured as STP normal port.

Virtual Port Channel Design

In virtual port channel (vPC) design, the Nexus 7000 series switch is used as a Layer 2 only device. A pair of Nexus 7000 series switches act as core devices to inter-connect the Layer 2 aggregation layer, service layer, and backbone routers. All switchports on the Nexus 7000 series switch are trunk ports carrying multiple VLANs. All switchports are also in the default virtual device context (VDC). Switch ports connected to the access layer and the service layer are part of vPC. Ports connected to backbone routers are single Ethernet trunk ports. Another vPC can carry management VLANs over a seperate switch to provide back-door access to the VM.

For advanced redundancy and high availability features, redundant supervisors are used in Nexus 7000 systems. Management ports on all supervisors are in a separate management vPC and are connected to a management switch for redundancy.

Figure 2-26 vPC Design



Nexus 7000 switches used in this design require a minimum of two 10-Gigabit modules (N7K-M132XP-12) to support a pair of Cisco UCS 6120 Fabric Interconnects, which in turn are connected to a customer POD with UCS hosting ESX servers. Ports are configured in dedicated mode to make use of the full 10G bandwidth. There are eight dedicated mode ports on a N7K-M132XP-12 module.

Although vPC provides loop free topology, STP is configured to prevent potential spanning tree loops during mis-configuration or device malfunction. Using MST minimizes spanning-tree related complexities and overhead with a large number of VLANs.

All port channels are created using link aggregation control protocol (LACP). LACP is a conditional service in Cisco NX-OS that must be enabled before commands can be issued. LACP supports up to 16 links in a port channel with eight links active at a time.

All vPC links connecting to the aggregation layer and service layer have two links from each Nexus 7000 switch. The vPC peerlink that connects between the vPC peers has four links. All port channels are distributed across 10G modules to provide link and module redundancy.

A vPC allows links that connected to two Cisco Nexus 7000 to appear as a single port channel to a third device. The third device can be a switch, server, or any other networking device that supports port channels.

The benefits of using a vPC are as follows:

- Allows a single device to use a port channel across two upstream devices
- Eliminates STP blocked ports
- Provides a loop-free topology
- Uses all available uplink bandwidth
- Provides fast convergence during link failure and device failure
- Assures high availability

Note Cisco NX-OS, version 4.1.5 or later is required for vPC support.

The vPC feature is built into the Cisco NX-OS image but not enabled by default. You must enable the vPC feature before you can configure or run the vPC functionality. To enable the vPC feature, issue the **feature vpc** command in global configuration mode.

Virtual Switching System Layer 2 Design

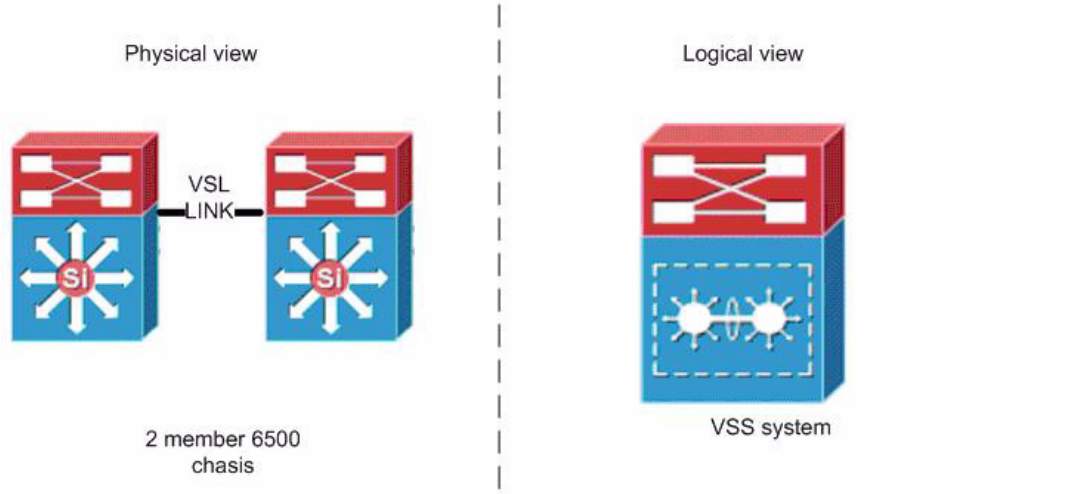
The Virtual Switching System (VSS) merges two Cisco Catalyst 6500 chassis (that meet software and hardware requirements) into a single logical chassis with the combined capability of both. The merge occurs on the control and data planes. As a result, the VSS provides redundancy and high availability, and its data forwarding capacity increases to a maximum of 1.44 Tbps.

To meet requirements of current data center deployments, where you position the VSS in a data center environment is critical. In a typical data center deployment and based on requirements, the Catalyst 6500 switch can be positioned at any of the following positions:

- Access layer so the VSS is closer to the compute environment
- Aggregation layer to provide uplinks to access switches
- Service layers where the VSS running appropriate service modules can provide Layer 4 - 7 services
- Core layer.

For details on positioning the VSS at these various layers, refer to VSS by Architecture Layer, page 2-41.

Figure 2-27 Virtual Switching System



Advantages of using the VSS include the following:

- Effective virtualization since all devices upstream and downstream to the VSS detect the VSS as a single logical chassis.
- Increased data forwarding capacity to the theoretical throughput of 1.44 Tbps.
- Effective redundancy and high availability. Because the VSS system requires only one supervisor to be active and handling control Layer 2 and Layer 3 operations, the other operates as a standby in case of failure or disaster recovery.

VSS Operation

The operation of the VSS requires that one switch acts as the active and the other as the standby. The active to carries Layer 2 and Layer 3 control operation, such as spanning tree operation and routing protocol operation. The active supervisor provides lookup for data forwarding to the line card Distributed Forwarding Card (DFC) and the Policy Feature Card (PFC) of the standby supervisor. This lookup information allows line cards in the active and standby chassis to forward traffic as needed.

Devices upstream and downstream to the VSS detect a single logical device. Thus at Layer 2, the VSS system has a single MAC address. This address is derived from the active chassis and populated across all line cards in both chassis. The two Catalyst 6500 chassis are merged via a virtual switch link (VSL).

For details on a typical VSS configuration, refer to:

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

During initialization of the VSS, active and standby roles are delegated based on individual switch configuration. These switch configurations are exchanged between switches through the VSL using the virtual switch link protocol (VSLP). The active switch can be chosen based on the switch id or the priority configured on the switch, with the switch having the highest priority chosen as the active.

If no issues arise during the setup of the VSS, the resulting redundancy mode is stateful switchover (SSO). Otherwise, the mode is route processor redundancy (RPR) mode. For full operational benefits of the VSS, the redundancy mode must be SSO.

Two Catalyst 6500 switches can successfully form a VSS if they meet hardware and software requirements

VSS Hardware Requirements

The supervisor engine that supports the VSS feature is the Supervisor 720-10G VSS (VS-S720-10G-3C/XL). This supervisor has the PFC3C/XL and the VSL-capable interfaces integrated on the module.

Table 2-4 Hardware Requirements

Hardware	Count	Requirements
Chassis	2	The VSS is available on chassis that support VS-S720-10G supervisor engines and WS-X6708-10GE switching modules. **Note** The two chassis need not be identical.
Supervisor Engines	2	The VSS requires Supervisor Engine 720 with 10-Gigabit Ethernet ports. You must use either two VS-S720-10G-3C or two VS-S720-10G-3CXL supervisor engine modules. The two supervisor engines must match exactly.
Switching Modules	2+	The VSS requires 67xx series switching modules. The VSS does not support classic, CEF256, or dCEF256 switching modules. In virtual switch mode, unsupported switching modules remain powered off.

The VSL EtherChannel supports 10-Gigabit Ethernet ports only. The 10-Gigabit Ethernet port can be located on the supervisor engine module or on one of the following switching modules:

- WS-X6708-10G-3C/XL
- WS-X6716-10G-3C/XL

VSS Software Requirement

To bring up the VSS, the minimum supported software is 122-33SXH. With the exception of the network appliance module (NAM), most services modules can only be recognized if 122-33SXI or later is running on both switches that form the VSS.

VSS by Architecture Layer

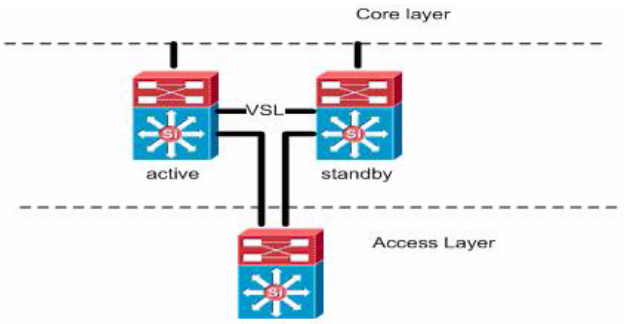
This section discusses the requirements and benefits of positioning the VSS at the access layer, the aggregation layer, the services layer, and the core layer of the data center network.

VSS at the Access Layer

The data center deployment requirements determine where the VSS should be positioned. Port density is the primary benefit of positioning the VSS in the access layer. A recent requirement for any access switch in a data center should be FCoE support, which is supported by the UCS 6100 fabric interconnect switches. If a high port density switch is positioned at the access layer, cabling requirements may be high for an end-of-rack (EoR) approach for the VSS. Also, server deployments prefer the top-of-rack (ToR) approach, where the Nexus 5000 and Nexus 2000 switches are the best candidates for compute environment switches and in the case of a UCS deployment, the 6100 fabric interconnects are made use of.

Despite disadvantages, the VSS system can be positioned at the access layer. More benefits can be obtained if the servers connected to the VSS are capable of NIC teaming. With two 10G ports connected from the server to the VSS, 20G of capable throughput can be achieved. Figure 2-28 shows a typical setup of the VSS connected to a server that is capable of NIC teaming.

Figure 2-28 VSS at the Access Layer

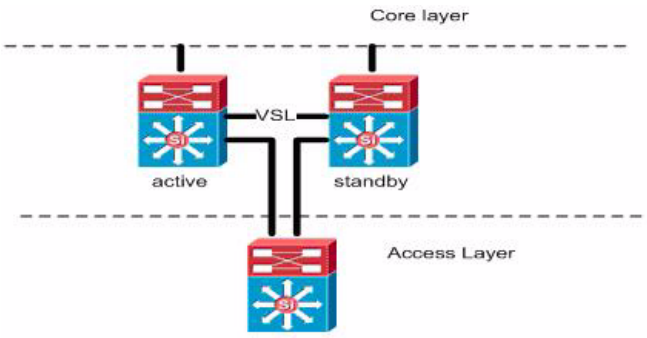


In a setup where a server capable of NIC teaming is connected to a two 10G ports on a VSS, then 20G of capable throughput can be obtained. If any of the VSS members goes offline, that fault is localized to the failed member with no interruption to the Layer 2 network / spanning tree. If any of the uplinks to the aggregation layer fails, the VSL is used if there is need for inter-chassis communication to get to the aggregation layer. Interruption to the Layer 2 network with this setup only occurs if the VSL link fails. Re-convergence time observed should be dependent on the time taken to effectively isolate the previous active chassis from the Layer 2 data path.

VSS at the Aggregation Layer

At the aggregation layer, the MEC can be used to provide uplinks to the access switches. The MEC can comprise of up to eight 10 gigabit (10G) ports connected so that each access switch is dual-homed to the VSS members. Most deployments prefer to have the spanning tree root at the aggregation layer.

Figure 2-29 VSS at the aggregation Layer



Looking at Figure 2-29, if positioned at the aggregation layer, the VSS can be made to be the root of the spanning tree. Spanning tree protocol in play should be dependent on Layer 2 requirement of the data center. The protocol used in this solution test setup was MST due to fast convergence time possible with its use. If the active member of the VSS goes offline, the convergence of the Layer 2 network should be dependent on the time taken for the standby to take control. In most cases, this time is less than the time taken to determine a change in the spanning tree network. With that being the case, no impact should be

seen in the Layer 2 network. If the standby member goes offline, no impact should be seen in the network. Any uplink or downlink failure to the VSS should not impact the operations of the Layer 2 network, since in most cases it is expected that MEC links are used.

VSS at the Services Layer

The function of the services layer is to provide network services to downstream entities, which in this case are the servers. Services provided include Layer 3 - 4 packet inspection (provided by the Firewall Service Module FWSM) and Layers 4 - 7 loads balancing provided by the Application Content Engine (ACE). Table 2-5 shows the minimum hardware and software requirements needed to have the services layer operational on the VSS.

Using the indicated minimum software releases, the following service modules are supported in a VSS:

Table 2-5 Supported Service Modules

Service Module	Minimum Cisco IOS Release	Minimum Module Release
Network Analysis Module (NAM-1 and NAM-2) (WS-SVC-NAM-1 and WS-SVC-NAM-2)	12.2(33)SXH1	3.6(1a)
Application Control Engine (ACE10 and ACE20) (ACE10-6500-K9 and ACE20-MOD-K9)	12.2(33)SXI	A2(1.3)
Intrusion Detection System Services Module (IDSM-2) (WS-SVC-IDSM2-K9)	12.2(33)SXI	6.0(2)E1
Wireless Services Module (WiSM) (WS-SVC-WISM-1-K9	12.2(33)SXI	3.2.171.6
Firewall Services Module (FWSM) (WS-SVC-FWM-1-K9)	12.2(33)SXI	4.0.4

Note Before deploying a service module in VSS mode, upgrade the module to the minimum supported release in standalone mode.

For an understanding of service modules that work on the VSS and the minimum hardware and software requirements, see the following Web site: http://www.cisco.com/en/US/products/ps9336/products_tech_note09186a0080a7c72b.shtml.

At the service layer, all VLANs used must be trunked correctly from the access layer through the aggregation layer up to the core layer. To ease configuration, the arrangement of modules in a VSS member should be a replica of the other VSS member.

The following code is a sample configuration of a services modules on a VSS configured at the service layer:

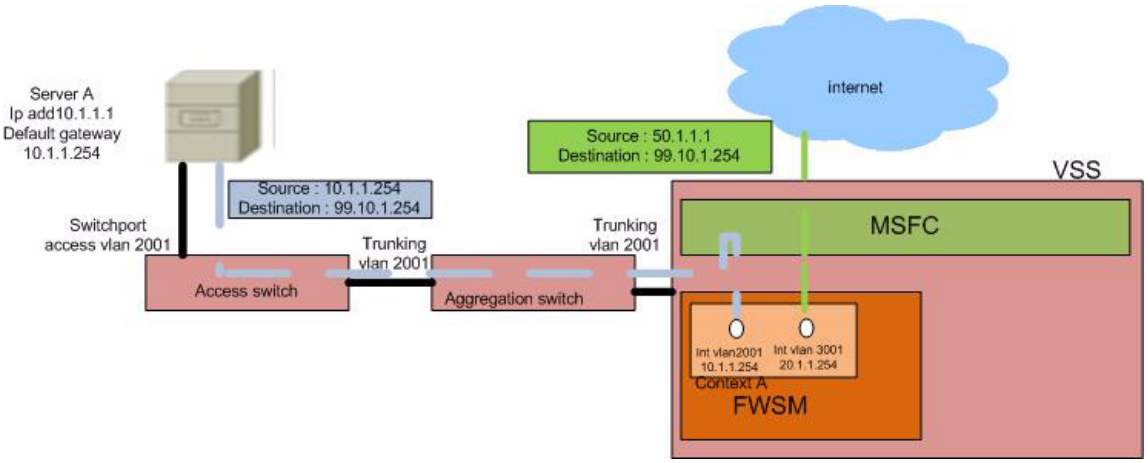
```
vss-1#sh module switch all
 Switch Number:     1   Role:   Virtual Switch Active
---------------------  -----------------------------
Mod Ports Card Type                              Model              Serial No.
--- ----- -------------------------------------- ------------------ -----------
  1    1  Application Control Engine Module      ACE20-MOD-K9       SAD130801KH
  2    8  CEF720 8 port 10GE with DFC            WS-X6708-10GE      SAD113705DL
  3    8  Network Analysis Module                WS-SVC-NAM-2       SAD1049001J
  6    5  Supervisor Engine 720 10GE (Active)    VS-S720-10G        SAL1204E26Z
  7    6  Firewall Module                        WS-SVC-FWM-1       SAD071202DX
```

```
  8    48  CEF720 48 port 10/100/1000mb Ethernet  WS-X6748-GE-TX      SAL1051BJT2


<snip>

Switch Number:     2   Role:  Virtual Switch Standby
---------------------  -----------------------------
Mod Ports Card Type                                  Model              Serial No.
--- ----- -------------------------------------- ------------------ -----------
  1     1  Application Control Engine Module      ACE20-MOD-K9       SAD130902AD
  2     8  CEF720 8 port 10GE with DFC            WS-X6708-10GE      SAL11370AY7
  6     5  Supervisor Engine 720 10GE (Hot)       VS-S720-10G        SAL124052GU
  7     6  Firewall Module                        WS-SVC-FWM-1       SAD083603UD
  8    48  CEF720 48 port 10/100/1000mb Ethernet  WS-X6748-GE-TX     SAL11413C9Z

<snip>
```

How the services layers can be used depends on the requirements in place. The SVI assigned to the services module can act as default gateways for servers that make up the compute environment. Depending on the services to be provided, packets outbound from the compute environment might pass through either of the service module or a combination of the modules. In some instances, there might be no need to provide any service to the packets outbound from the server environment, where in such situations, such packets are routed out without passing through any service module.

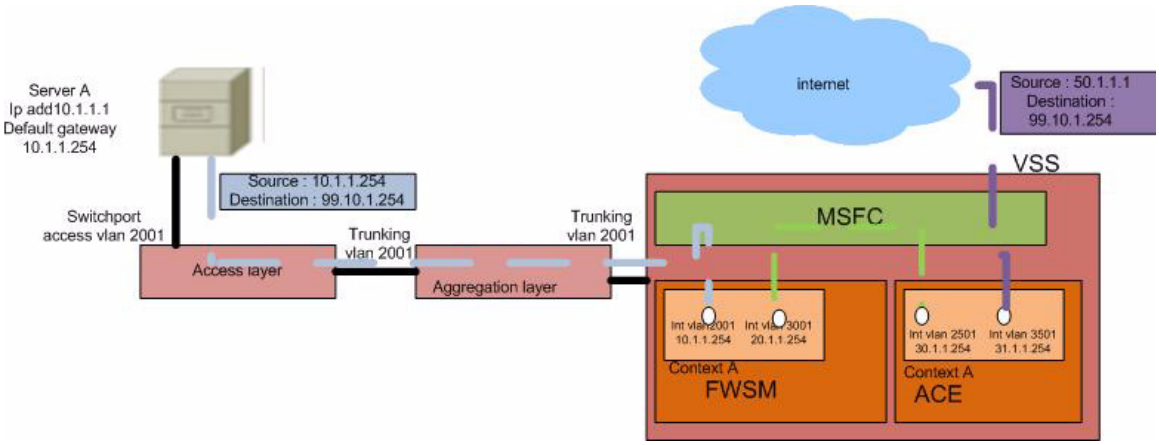
In Figure 2-30, a single service module (FWSM) is needed to provide services to servers in the data center. In this design, firewall services are provided to server A, whose default gateway is an SVI on the FWSM. With correct configuration, the packets outbound from the server should move through the right VLAN (vlan 2001) at the access and aggregation layer to the services and core layers. On reaching the VSS supervisor, packets are intelligently routed to the right service module (FWSM). Firewall services provided, all depend on the requirement that need to be meet. In this design, the FWSM provides packet inspection and translation.

Figure 2-30 Data Path with FWSM on a VSS at the Services Layer



In Figure 2-31, a combination of service modules are used to provide services to servers. This design is an extension of the previous design, with the packets routed back to the second module (ACE) before eventually leaving the VSS. As mentioned previously, services that are provided by each module depend on requirements that need to be meet.

Figure 2-31 Data Path with ACE and FWSM on a VSS at the Services Layer



VSS at the Core Layer

At the core layer, the VSS system can be configured to have routing neighbor adjacencies with domains under or not under the same administration of the operator. With SSO and NSF enabled by default, the VSS system can recover quickly from negative events. By default, NSF is active if OSPF or EIGRP is configured. For BGP, NSF must be manually configured in order for it to be active.

VSS Best Practices and Caveats

Best Practices

The following best practices are recommended for use with a VSS design in a data center environment:

- At least a 10G port of the VS-S720-10G C/CXL VSS supervisor should be member of the Virtual Switch Link (VSL). This ensures that the control channel comes up during either a VSS boot process or a switch member boot process.
- At least 20Gbps should be available for use as a VSL during the operations of the VSS. The higher the data capacity of the VSL, the lower the probability of congestion times during a data center operations peak period. With a higher data capacity of the VSL, benefits associated with adding more service modules to each VSS member can be maximized.
- Full operations of the VSS can be achieved by avoiding configurations that would result in sync failures during bootup of any switch member of the VSS. If a sync failure occurs, that switch undergoing a boot process is isolated. This results in 50% of the operational performance of the VSS.
- To reduce outage times during switchovers or existence of negative events, an effective port hashing algorithm on the MEC should be employed. To achieve lower outage times, the adaptive port hashing algorithm should be used, which reduces the impact addition and deletion of port members to a port-channel has on outage time. Choice of port hashing algorithm in use should be dependent on an option that can closely distribute traffic across all members of a port channel interface.

Caveat

- QoS configurations are not allowed on VSL ports and finally on 1G ports on the supervisor.

EtherChannel Load Balancing

In this solution, virtual port channels (vPCs) are used in the data center for link redundancy and traffic load sharing. However, depending on the nature of the traffic, individual links within a port channel could become polarized, which causes oversubscription and packet loss issues. To minimize the risk of link oversubscription and load balance traffic across port channel links, suitable port channel load balancing technique should be configured on all switches in the data center. There is no guarantee that the default load balancing configuration would provide the best result, hence there may be a need to monitor the link utilization and fine tune the configuration prior to adding new links in the port channel.

An EtherChannel bundles individual Ethernet links of the same type into a single logical link and provides the aggregate bandwidth of up to eight physical links. Depending on the channel protocol and type of hardware used, a system can have up to 16 member links in the port channel although only eight links can be active at any given time.

EtherChannel computes a numerical value that selects a link in the channel to load balance traffic across member links. The hashing algorithm used to compute this value varies depending on the platform used. However, from a user perspective, the main difference is in the default method used and the options available to use as input to the hash. The algorithm can use one or more fields within the packet to derive the hash. Most commonly used fields are IP address, MAC address, and L4 port, but the Nexus 7000 uses VLAN, too.

EtherChannel load-balancing is deterministic since it is based on a hash algorithm. Assuming the same parameters are given as input to the hash, the same EtherChannel member port is selected. A show command is also available to find out the link selected for a given set of input values.

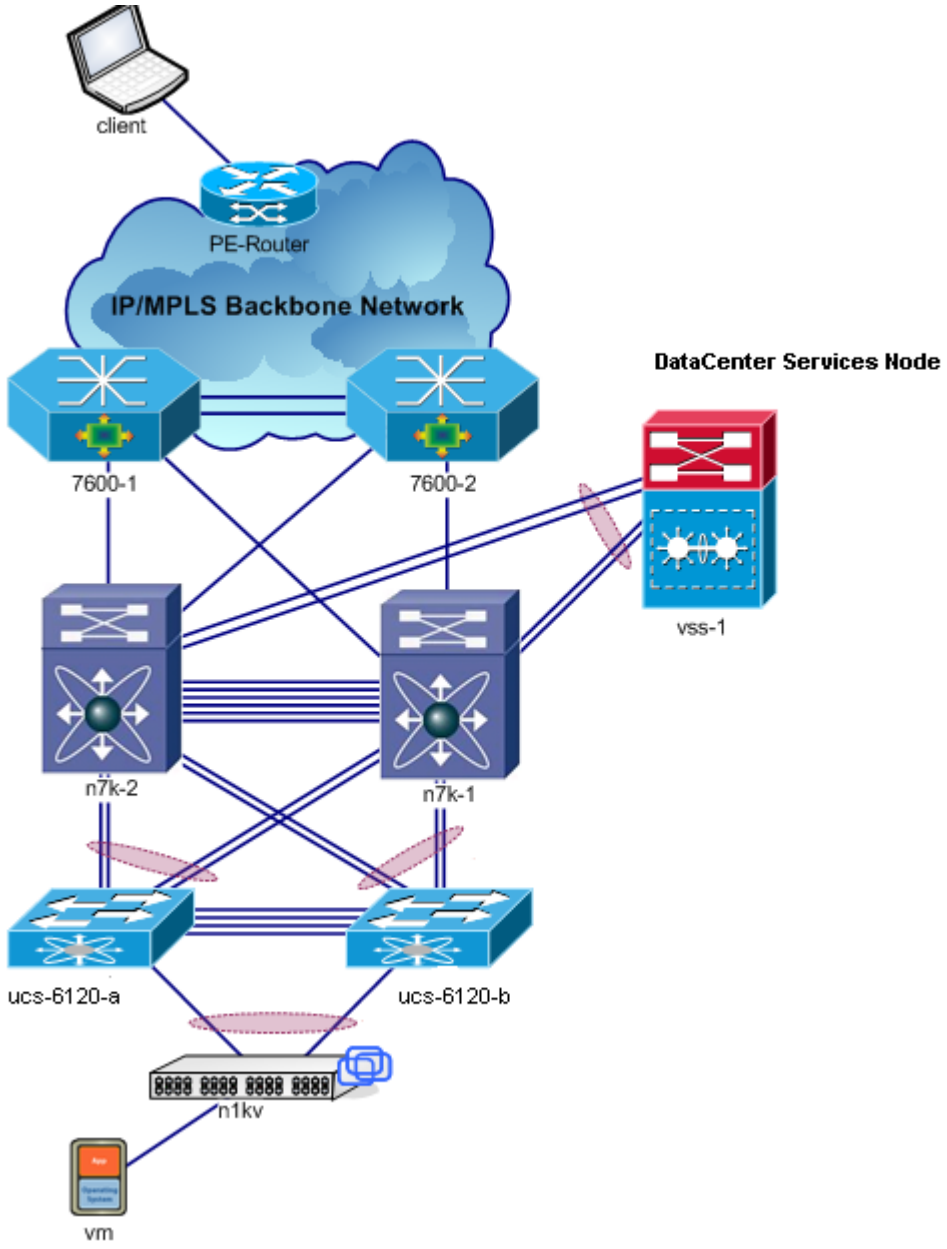
Data Center Layer 3 Design

This section contains s the following topics:

ACE and FWSM

In this topology, the services core is a Data Center Services Node (6500 chassis pair running in VSS mode). Configuring the DSN/VSS at the services layer enables the insertion of service modules that provide Layer 4 - 7 services. Traffic can be configured to bypass these service modules as needed.

Figure 2-32 Topology



Service Class

Three service classes are defined for this solution:

- **Gold**: Includes server load balancer, firewall, multiple virtual machines, best compute and network resource allocations.
- **Silver**: Includes server load balancer, multiple virtual machines, moderate compute and network resource allocations.
- **Bronze**: Includes sets of low compute virtual machine resources, best effort compute and network resource allocations.

Gold - The Gold class is provisioned with server load-balancing (ACE) and the firewall services module (FWSM). Virtual machine traffic is segregated into three VLANs, one each for WEB, APP, and DB services. Each class of service uses the FWSM as its gateway to route between VLANs or the Internet. The FWSM uses a default route that points to the inside VLAN on the ACE. From the ACE point of view, the next hop for Internet-bound traffic is a switched virtual interface (SVI) on the same VLAN as the outside interface on each ACE context.

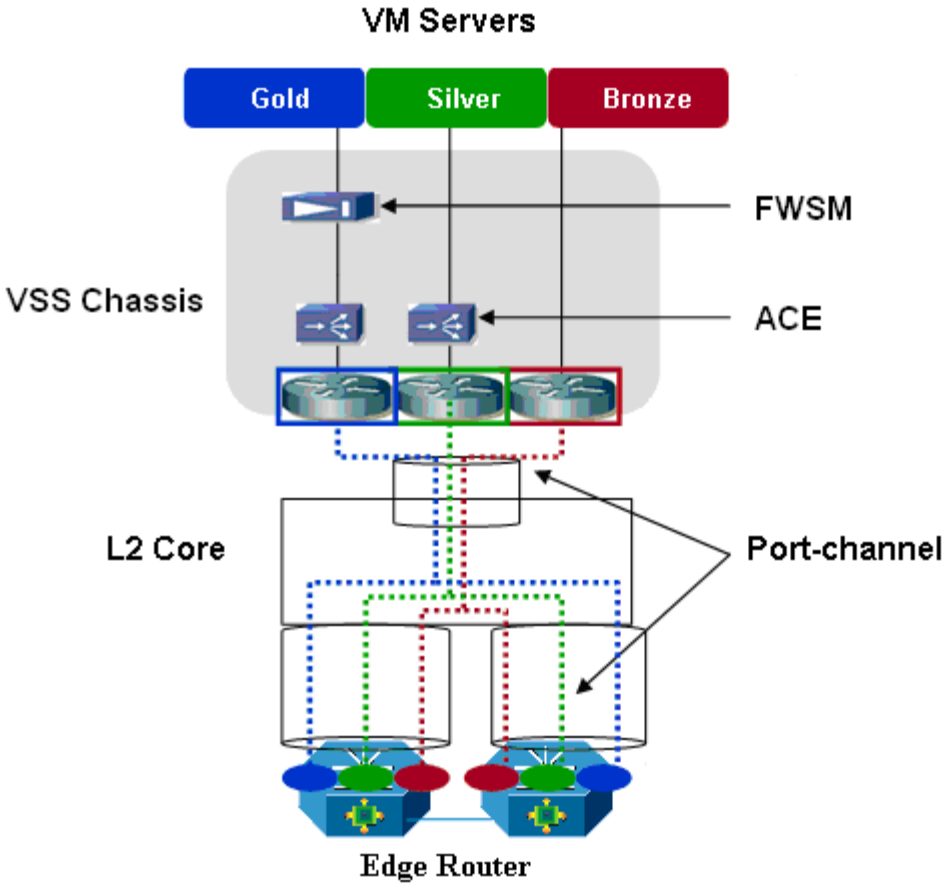
Route health injection (RHI) injects static routes (VIP addresses) from the ACE to the supervisor. The VIPs are static routes injected into the local routing table. These are aggregated to a /8 and injected into iBGP to be used by the host router. These VIPs are used for NAT and SLB traffic, though RHI is not supported for NAT VIP, so static entries must be manually configured on the services core. For L2VPN customers, ACE outer VLANs are Layer 2 carried by the VSS to the HR by way of the Nexus 7000 core. The L2VPN clients use the ACE alias (FT) IP address as a gateway to VIPs that serve for NAT and SLB.

Silver - The Silver class is provisioned with server load-balancing via ACE. Virtual machine traffic for each customer is on a shared VLAN for WEB, APP, and DB traffic. RHI is employed to advertise the VIP address to the supervisor. Using BGP, reachability information on how to get to these VIPs are advertised by the VSS to the hosted routers (HR). Similar to the Gold class, traffic inbound to the ACE is routed to a SVI on the same VLAN as the outside VLAN on each context on the ACE.

Bronze - The Bronze class is not provisioned for L4 - 7 service provisioning, so it bypasses the service modules and routes either to a L3 SVI on the VSS or directly to the Edge routers.

Note The service modules use static routes for routing.

Figure 2-33 Service Class Logical Topology



For more detailed information regarding data center network routing employed in this guide, see Cisco 7600 and Data Center Services Node, page 2-50.

DSN with ACE and FWSM Integration

When the ACE and FWSM service module are installed in a VSS-enabled chassis, you must manually assign VLANs to the modules. The ACE uses the **svclc** command, and the FWSM uses the **firewall** command. In this topology, some VLANs are common (shared) between the ACE and FWSM. To configure common VLANs, designate a particular VLAN-group to carry them and share that group (for example, *vlan-group 1*) across the two modules.

Route Health Injection

Route Health Injection (RHI) is a feature that allows the IP address of the virtual server on the ACE to be injected as a static host route into the routing table. This function is commonly used with routing protocol to enable the MSFC to advertise the availability of a VIP address throughout the network. RHI is used primarily in environments where the VIP is in a subnet that is not directly attached to the MSFC.

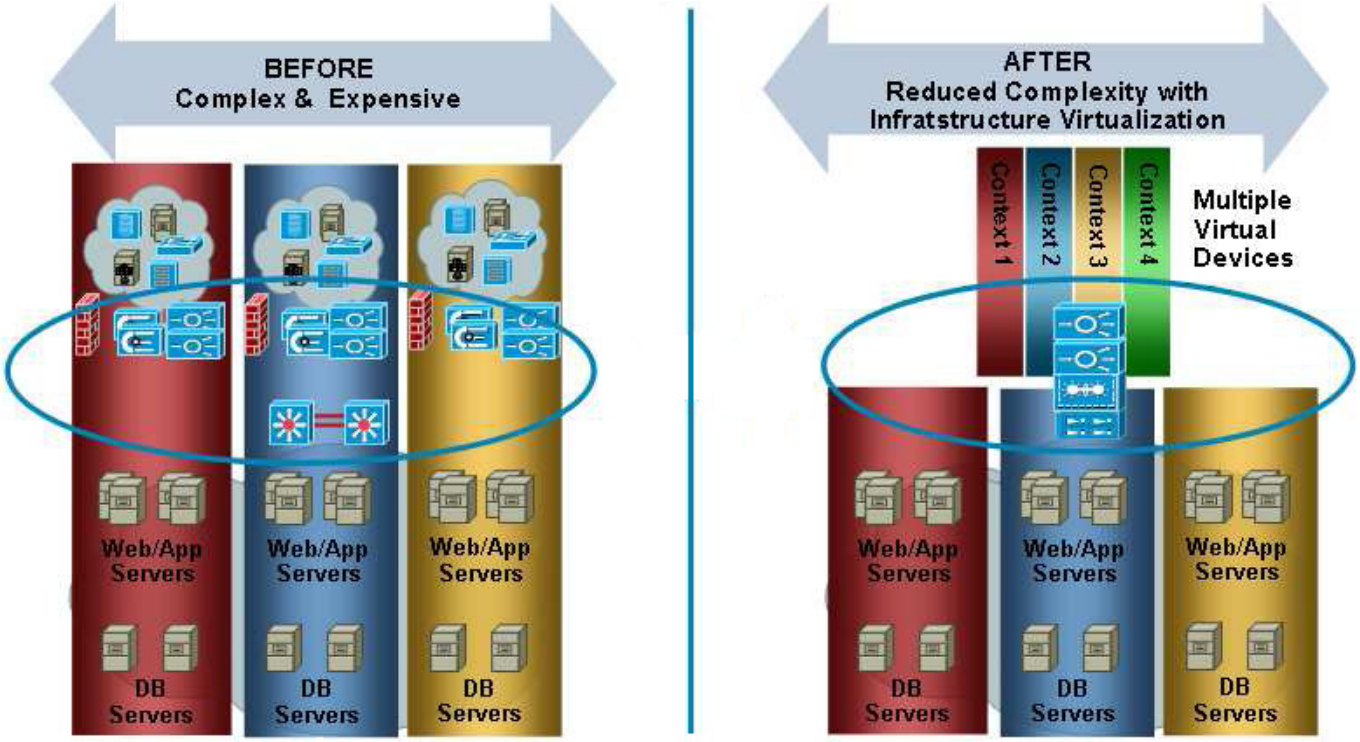
ACE RHI functionality is fully integrated with VSS. RHI notification from the ACE module is always sent to the active virtual switch irrespective of the chassis in which it is present. If there is a stateful switchover (SSO) event, the new active virtual switch sends a control message to the ACE modules to resend the RHI notification. The routes can be viewed through the routing table or `svclc` command.

The DSN advertises customer routes (VLAN interfaces on the VSS, ACE VIPs, and firewall networks) via BGP to the Cisco 7600 routers. The VIPs originate as static host routes injected via RHI. The ACE does not support RHI for NAT VIPs, so they require a manually entered static route on the VSS. The FWSM is not hosting virtual IP addresses. RHI is not currently supported on the FWSM installed in a VSS chassis.

ACE and FWSM Virtualization

In this solution, the ACE and FWSM are configured in multi-context routed mode. By using multiple contexts, virtualization is leveraged, permitting an ACE or FWSM module to behave like an independent device for each customer. Through virtualization, resources can be partitioned giving each context its own ACLs, policies, interfaces, routing, and so on, allowing for customization and isolation for each customer. This flexibility enables you to achieve the best use of resources available in the services switch.

Figure 2-34 Virtualization



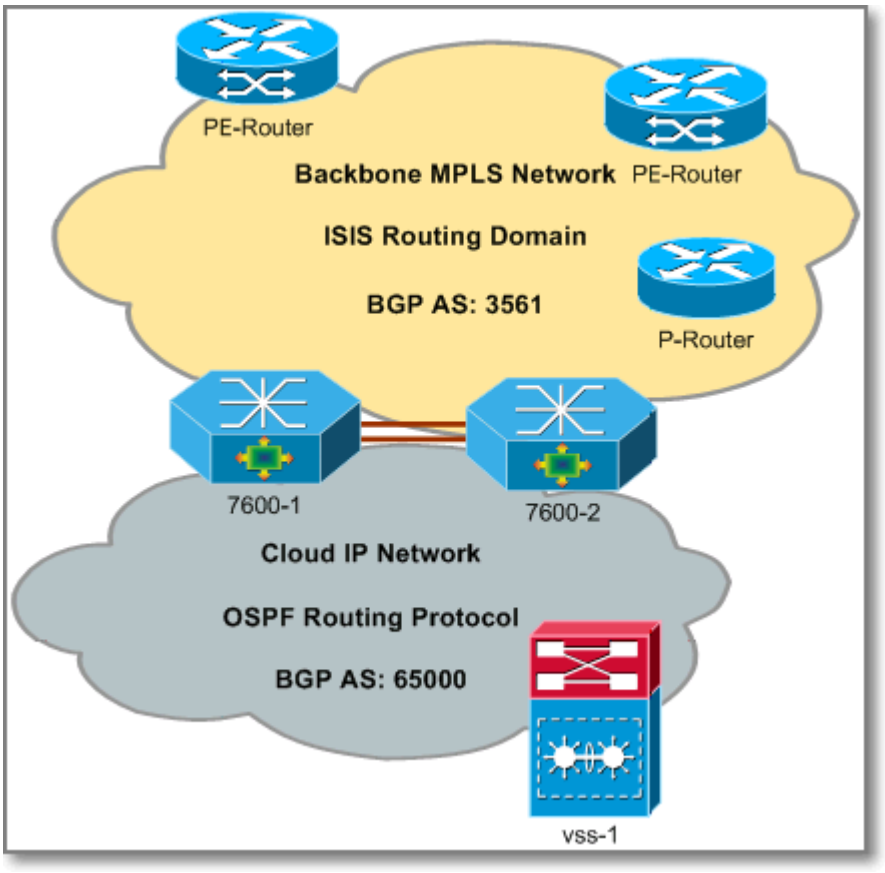
Cisco 7600 and Data Center Services Node

Figure 2-35 shows the logical overview of the network setup. The setup consists of a backbone network domain and a cloud network domain. Cisco 7600 routers serve as the routing gateway between the cloud network and the MPLS-enabled IP backbone network. The following protocols are used in the setup:

- ISIS
- OSPF
- BGP

- MPLS

Figure 2-35 Logical Network Domains



OSPF is configured on network devices within the cloud network to exchange reachability information for the loopback interface IP addresses. For this phase of the solution, a VSS router acting as a DataCenter Service Node (DSN) is deployed within the cloud network. The number of DSN deployments should be dependent on data center requirements.

Within the backbone network, devices use the ISIS routing protocol to carry reachability information for loopback interface IP addresses. MPLS is configured to support L3VPN and L2VPN services.

Cisco 7600 routers serve as the demarcation between the cloud and backbone networks. The cloud network facing links of the Cisco 7600 routers participate in OSPF routing protocol, while the backbone network facing links participate in ISIS routing protocol. The links between the Cisco 7600 routers run OSPF and ISIS routing protocols. The OSPF domain and ISIS domain exist independently with no redistribution between domains.

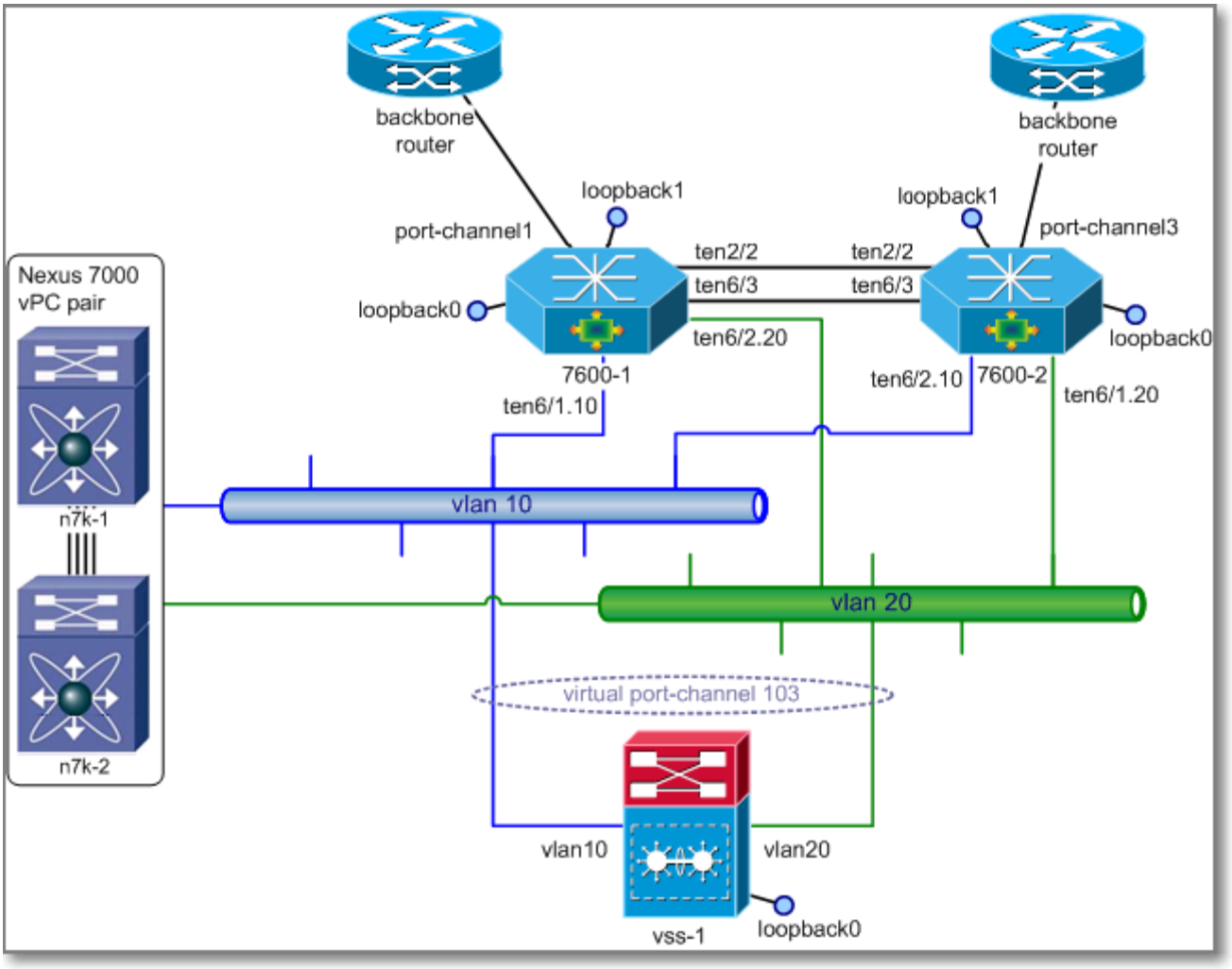
Within the cloud network, BGP carries customer routes (routes to the virtual machines, as well as routes injected into the VSS routing table by the ACE, which advertise reachability to the virtual IP address of the virtual machines) from the Catalyst DSN to Cisco 7600 routers.

On the Cisco 7600 routers, MP-BGP advertises and learns customer routes to and from BGP peers on the backbone network.

Basic Routing Protocols Configuration

Figure 2-36 shows the logical connections of the routers. Each Cisco 7600 router has one physical 10GE connection to each Nexus 7000 switch. The 10GE interface on the Cisco 7600 router is configured as 802.1Q trunk carrying multiple VLANs. Sub-interface VLAN 10 and sub-interface VLAN 20 on each on the physical interface respectively carry IP routed traffic between the cloud network and Cisco 7600 routers; other VLAN sub-interfaces carry L2VPN and L3VPN traffic.

Figure 2-36 Logical IP Network Setup



The DSN has four 10GE connections to the Nexus 7000 vPC pair; the four 10GE interfaces are configured into a port-channel (vPC on the Nexus 7000 pair); the port-channel interface is configured as a switchport carrying multiple VLANs. Interface VLAN 10 and interface VLAN 20 carry the IP routed traffic between the Cloud network and the Cisco 7600 routers.

Configuring ISIS

ISIS routing protocol is configured on the Cisco 7600 pair for exchanging IP reachability information with routers on the backbone MPLS network.

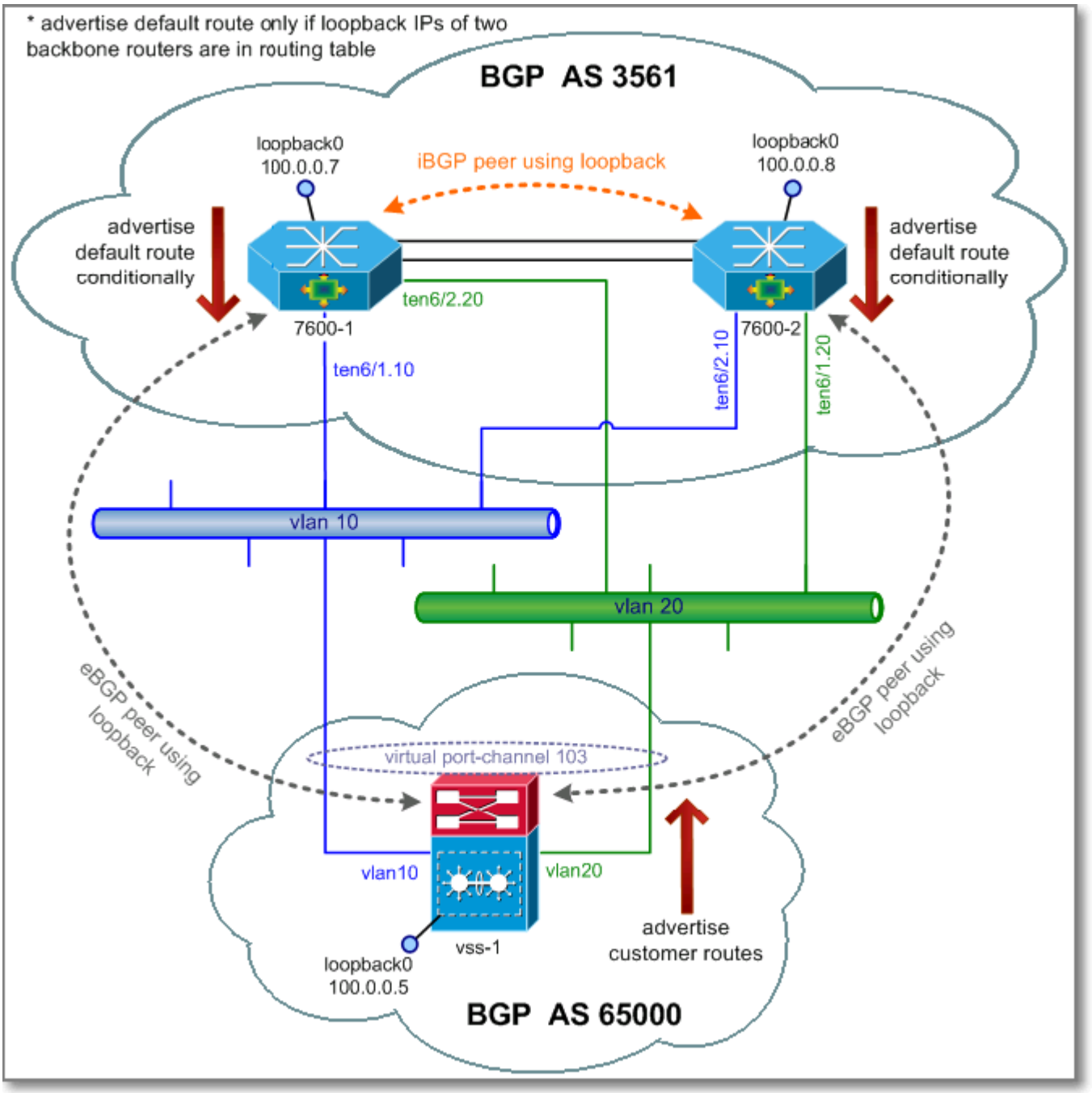
Configuring OSPF

OSPF routing protocol is configured on the Cisco 7600 pair and DSN for exchanging IP reachability information with routers on the cloud network.

Configuring BGP

BGP routing protocol is configured on the Cisco 7600 pair and DSN. The Cisco 7600 routers are in one BGP autonomous system. The DSN is in its own private BGP autonomous system. The BGP peers with its neighbors using loopback interfaces for iBGP and eBGP, for eBGP peering, ebgp-multihop configuration is required.

Figure 2-37 Basic BGP Setup



Each Cisco 7600 router conditionally advertises default route to the DSN as long as it has connectivity to the backbone network. The Cisco 7600 router checks for routes from the backbone network in its routing table (the routes 99.99.99.1/32 and 99.99.99.2/32 are the loopback IP addresses of routers on the backbone network not connected directly to the Cisco 7600 routers; the Cisco 7600 routers learned those routes from ISIS routing protocol) for connectivity to the backbone network.

BGP on the DSN is configured with a private autonomous system number; it is assumed that upstream BGP routers in the backbone network that peer with other autonomous systems will strip the private autonomous system.

The DSN receives BGP default route from each Cisco 7600 router; by default, only the default route advertised by one of the Cisco 7600 routers would be best and installed into the routing table, which effectively made the Cisco 7600 routers into active/standby setup for traffic from the cloud network to the backbone network; the BGP maximum-path configuration allows multiple parallel routes to be considered best and installed into the routing table, change the Cisco 7600 router into active/active setup.

The DSN advertised customer routes (VLAN interfaces on the DSN connected to virtual machines, ACEs, and firewalls networks) via BGP to the Cisco 7600 routers; customer routes are nailed as aggregate static routes to Null0 and a route-map is used the select the routes to be advertised to BGP peers.

MPLS Configuration

MPLS is configured on Cisco 7600 routers on interfaces facing the backbone network. No MPLS is configured on the DSN.

L3VPN Configuration

For this phase of the solution, a simple L3VPN setup is implemented. Cisco 7600 routers serve as L3VPN PE routers. The L3VPN configuration differs depending on what service class the L3VPN termination belongs to. The cloud network provides three service classes:

- **Gold**: The Gold service class includes server load balancer, firewall, multiple virtual machines, best compute, and network resources allocations within the Cloud network.
- **Silver**: The Silver service class includes server load balancer, multiple virtual machines, moderate compute and network resources allocations within the Cloud network.
- **Bronze**: The Bronze service class includes one or two virtual machines and best effort compute and network resources allocation within the Cloud network.

Figure 2-38 show the logical network setup for Gold and Silver service classes. The difference between the Gold service class and Silver service class is the addition of firewall service module below the ACE server load balancer, enabling better protection and security to the virtual machines.

Figure 2-38 Gold Service Class L3VPN Setup

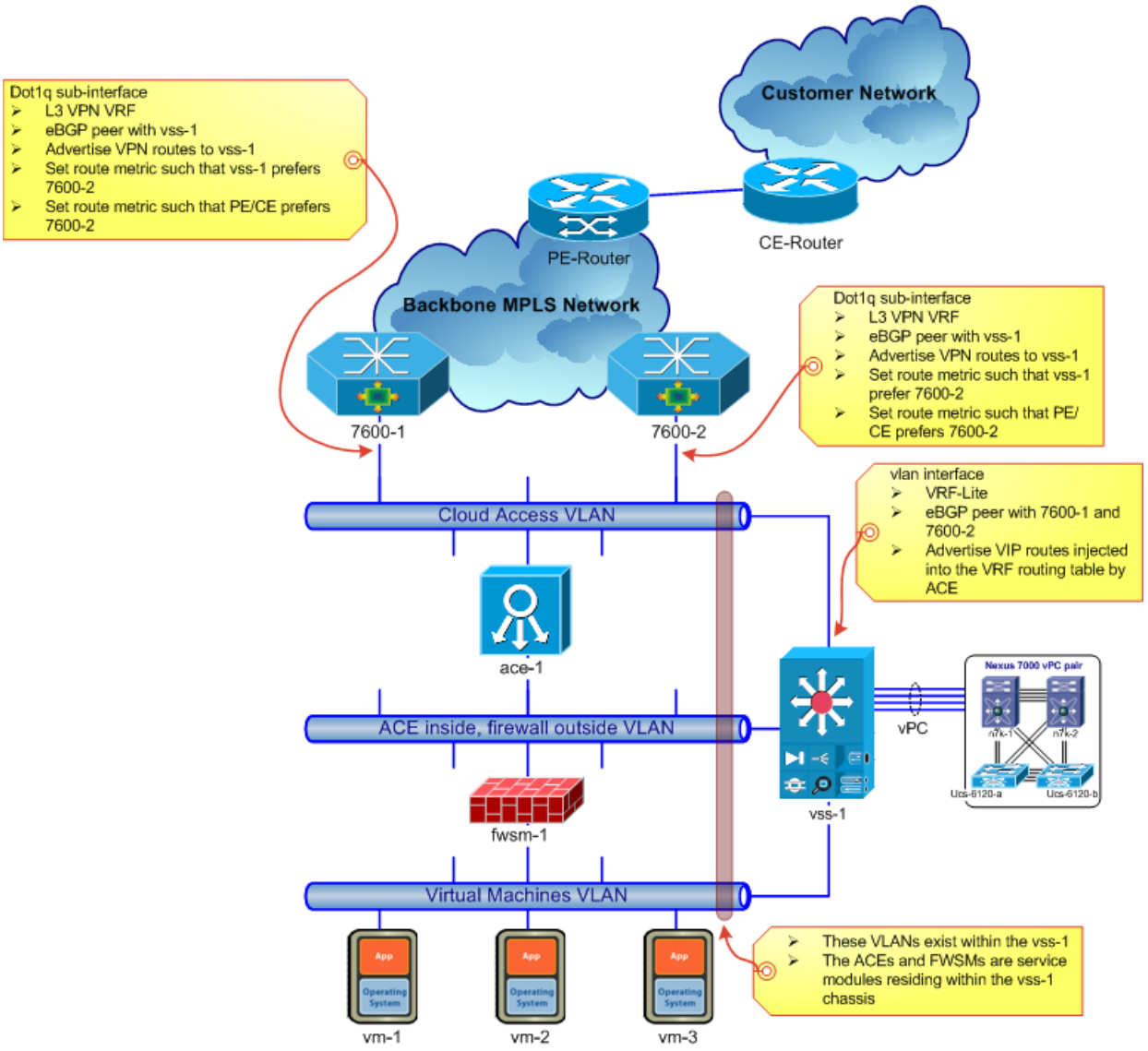
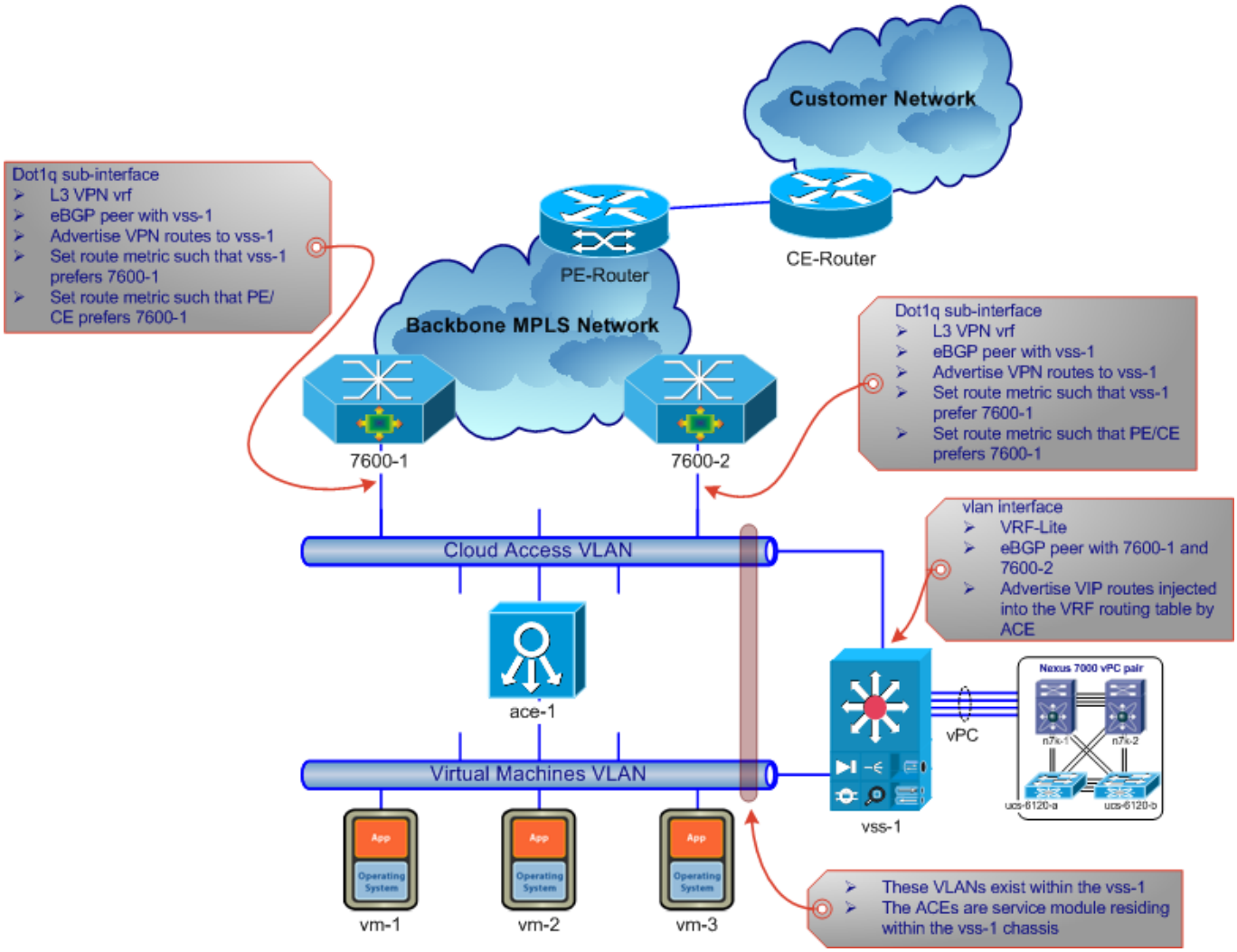


Figure 2-39 Silver Service Class L3VPN Setup



Gold and Silver Service Classes

The L3VPN configuration on Catalyst 7600 routers and the DSN are the same for Gold and Silver service classes. The L3VPN setup resides on the cloud access VLAN (outside interface of ACE). The internal network setup below the ACE is irrelevant to the L3VPN configuration. Each customer has its own cloud access VLAN and VRF, providing total separating between customer networks, communication between virtual machines from different customers is not possible within the cloud and backbone networks.

L3VPN VRF is configured on sub-interface of both Catalyst 7600 routers to guard against router or interface failure. On the DSN, L3VPN VRF-Lite is configured on VLAN interface. The VLAN interface is bound to port-channel switchport, with member links for the port-channel strand across both chassis on the VSS, providing link and device redundancy.

The Catalyst 7600 routers use MP-BGP to advertise routes learned from DSN to the PE/CE router. Each Catalyst 7600 router manipulates the metric of the routes advertised to influence the particular Catalyst 7600 router the PE router would select to send customer data traffic to the cloud network.

BGP within VRF context advertises customer routes learned from PE/CE router to the DSN. Each Catalyst 7600 router manipulates the metric of the routes advertised to influence the Catalyst 7600 router that the DSN would select to send virtual machines data traffic to the customer network.

On Catalyst 7600 routers and the DSN, the interface IP address, rather than the loopback interface, is used for BGP peering, eliminating the need to create a specific loopback interface for each L3VPN customer.

Note On Catalyst 7600 routers, the BGP route metric for each VRF is set for incoming and outgoing directions to ensure symmetric routing.

Note To achieve better redundancy, the main interface for L3VPN sub-interface on each Catalyst 7600 router must be selected so the failure of any downstream Nexus 7000 switch does not bring L3VPN traffic for any customer to a halt.

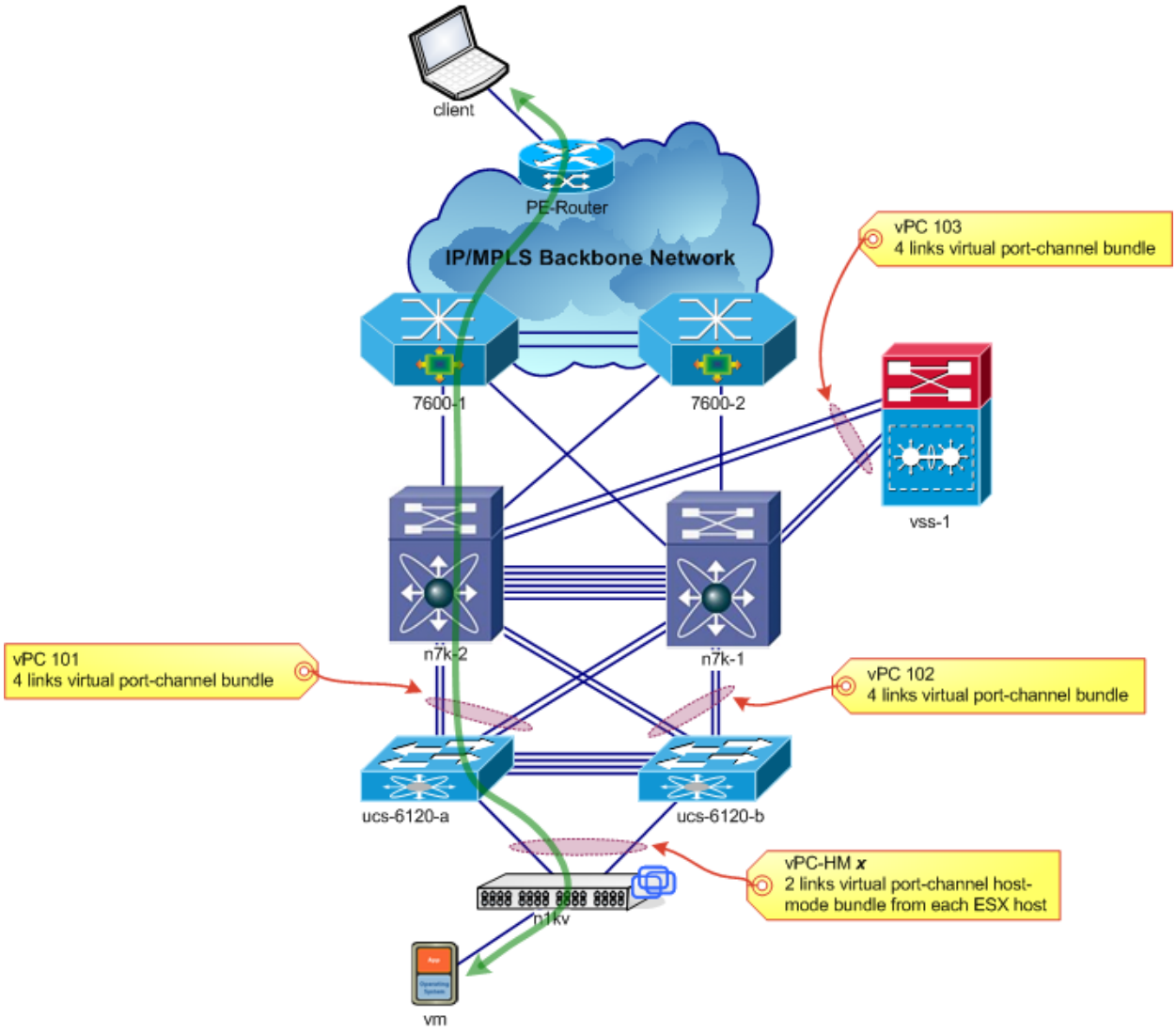
VRF-Lite is configured on the DSN. BGP within VRF context advertises routes injected by ACE into the routing table to Catalyst 7600 routers. The DSN learns customer routes from Catalyst 7600 routers and uses the route metric to select the best route to use to send data traffic.

The PE router implements VRF static routes with the CE router. Static routes from the customer VRF are injected into MP-BGP to be advertised to Catalyst 7600 routers.

Bronze Service Class

The L3VPN setup for Bronze service class is a bit different than those of the Gold and Silver service classes. The Bronze service class does not include the ACE server load balancer and firewall in the service offering; data traffic for Bronze service class does not need to traverse the DSN. Catalyst 7600 routers VLAN sub-interfaces extend directly to the virtual machines network/VLAN via the Nexus 7000 and UCS 6120 fabric interconnect. The Bronze service class virtual machines use Catalyst 7600 routers as the default gateway to reach the respective L3VPN customer network. HSRP is configured on Catalyst 7600 routers VLAN sub-interfaces to provide default gateway redundancy to the virtual machines. Figure 2-40 shows the logical network setup of for Bronze service class.

Figure 2-40 Bronze Service Class L3VPN Setup



The data traffic for each Bronze service class L3VPN customer is manually load-balanced:

- For data traffic from virtual machines to the customer network, HSRP priority is manipulated to prefer one particular Catalyst 7600 router as the outgoing gateway. HSRP pre-emption is configured to ensure that the preferred active HSRP router is always active after recovery from failure.
- For data traffic from the customer network to virtual machines, the BGP metric of the route virtual connected machines network/VLAN is manipulated to prefer one particular Catalyst 7600 router as the incoming gateway.

A simple HRSP tracking mechanism is also implemented for the HSRP group to track the uplink interface on the Catalyst 7600 router; the HSRP active router relinquishes its active role if it loses its uplink connection to the backbone network. HSRP also supports enhanced object tracking, in which the tracking mechanism is separated from HSRP. To learn more about HSRP object tracking, see:

http://www.cisco.com/en/US/docs/ios/ipapp/configuration/guide/ipapp_hsrp_ps6922_TSD_Products_Configuration_Guide_Chapter.html#wp1055048

The cloud network is expected to have multiple Bronze service class customers, each with a VLAN sub-interface on each Catalyst 7600 router. Configuring HSRP group for each sub-interface can have a detrimental impact on network traffic and CPU and memory utilization when the number of sub-interfaces is large.

In a network with two HSRP routers with two physical interfaces each, there are four possible HRSP topologies. Instead of configuring many HSRP groups on the sub-interfaces, only one HSRP group is required on a physical interface for the purposes of electing active and standby routers. This group is known as the master group. Other HSRP groups may be created on each sub-interface and linked to the master group via the group name. These linked HSRP groups are known as client or slave groups.

The HSRP group state of client groups follows that of the master group with a slight random delay so that all client groups do not change at the same time. Client groups do not participate in any sort of router election mechanism. Client groups send periodic messages to refresh their virtual MAC addresses in switches and learning bridges. The refresh message may be sent at a much lower frequency compared with the protocol election messages sent by the master group. For more information, see:

http://www.cisco.com/en/US/docs/ios/ipapp/configuration/guide/ipapp_hsrp_ps6922_TSD_Products_Configuration_Guide_Chapter.html#wp1055821

Note The default mac-refresh timer of the client (slave) HSRP groups is 10 seconds. During a failover, this timer relates to the frequency of HSRP packets used to update the forwarding tables of connected switches. In this configuration, the timer was reduced to 2 seconds to speed up network convergence. In larger HSRP configurations, greater values must be configured.

L2VPN Configuration

Pseudowire Overview

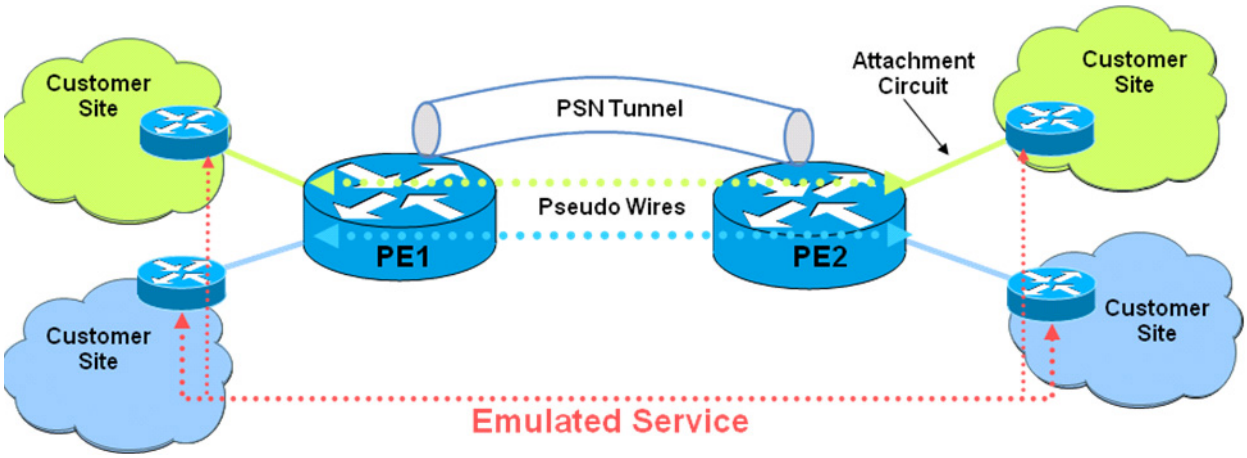
A pseudowire is a connection between two provider edge devices that connects two attachment circuits over a packet-switching network (PSN).

The pseudowire emulates the essential attributes of a service such as ATM, Frame Relay, Ethernet, TDM circuit, and so on, presenting itself as a “transparent wire” to these services.

The packet-switching network that pseudowire operates over may be multi-protocol label switching (MPLS), Internet Protocol (IPv4 or IPv6), or Layer 2 Tunneling Protocol Version 3 (L2TPv3).

Figure 2-41 shows the reference model for pseudowire.

Figure 2-41 Pseudowire Reference Model



A L2VPN pseudowire consists of two unidirectional Label Switched Paths (LSPs). Each is represented by a pseudowire label (also variously known as pseudowire ID, VC ID, and VC label). The pseudowire label is part of the label stack encoding that encapsulates Layer 2 packets going over L2VPN pseudowire.

The following steps explain the procedures to establish a L2VPN pseudowire:

1. A pseudowire is provisioned with an attachment circuit on PE1.
2. PE1 initiates a targeted LDP session to PE2 if none already exists. Both PEs receive LDP keep-alive messages from each other and complete the session establishment. They are ready to exchange pseudowire label bindings.
3. When the attachment circuit state on PE1 transitions to up, PE1 allocates a local pseudowire label corresponding to the pseudowire ID that is provisioned for the pseudowire.
4. PE1 encodes the local pseudowire label into the Label TLV and the pseudowire ID into the FEC TLV. Then it sends this label binding to PE2 in a Label Mapping message.
5. PE1 receives a Label Mapping message from PE2 and decodes the pseudowire label and pseudowire ID from the Label TLV (Type Length Value) and FEC TLV (Forwarding Equivalent Class TLV).
6. PE2 performs Steps 1 through 5 independently.
7. After PE1 and PE2 exchange the pseudowire labels and validate interface parameters for a particular pseudowire ID, the pseudowire with that pseudowire ID is considered established.

If one attachment circuit on one PE router goes down, a Label Withdraw message is sent to the peering PE router to withdraw the pseudowire label that it previously advertised, bring the pseudowire down.

Refer to http://tools.ietf.org/html/rfc3916 for more details.

L2VPN Configuration

For this phase of the solution, a simple L2VPN setup is implemented. The Cisco 7600 routers serve as L2VPN edge routers. The cloud network provides three service classes:

- **Gold**: The Gold service class includes server load balancer, firewall, multiple virtual machines, best compute and network resources allocations within the cloud network.
- **Silver**: The Silver service class includes server load balancer, multiple virtual machines, moderate compute, and network resources allocations within the cloud network.

- **Bronze**: The Bronze service class includes just one or two virtual machines, and best effort compute and network resources allocation within the cloud network.

Figure 2-42 show the logical network setup of for Gold and Silver service classes:

Figure 2-42 Gold Service Class L2VPN Setup

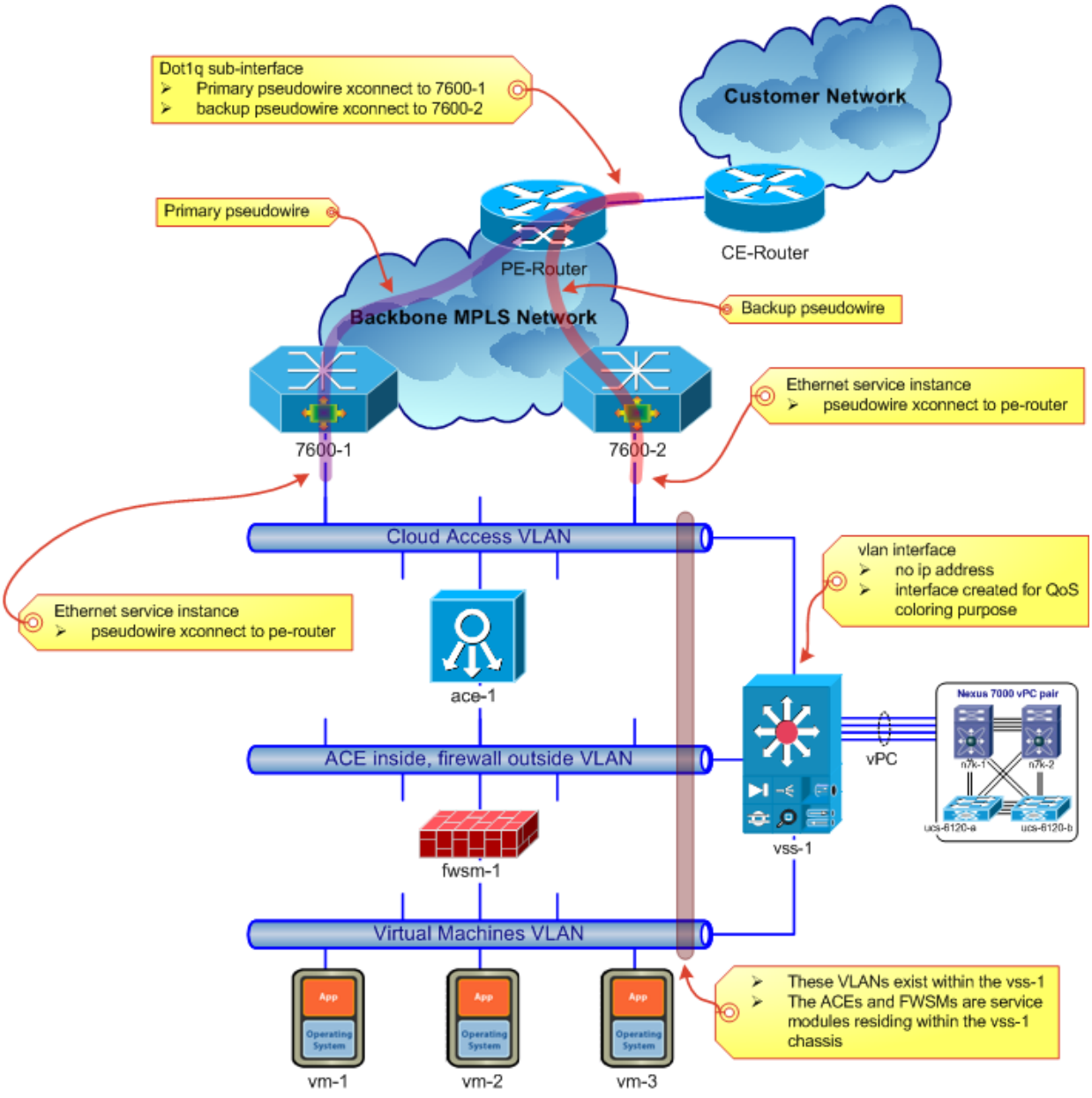


Figure 2-43 Silver Service Class L2VPN Setup

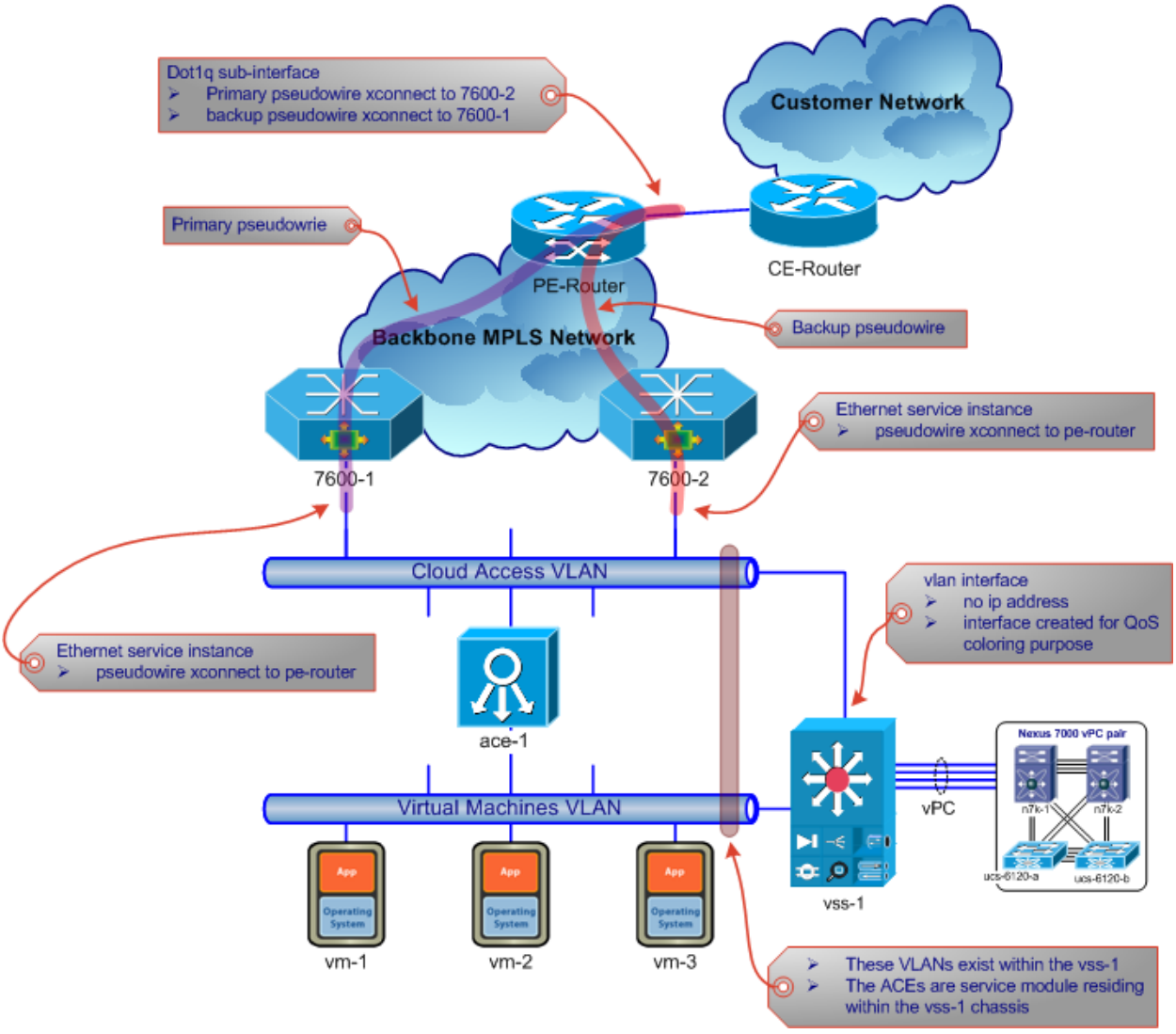
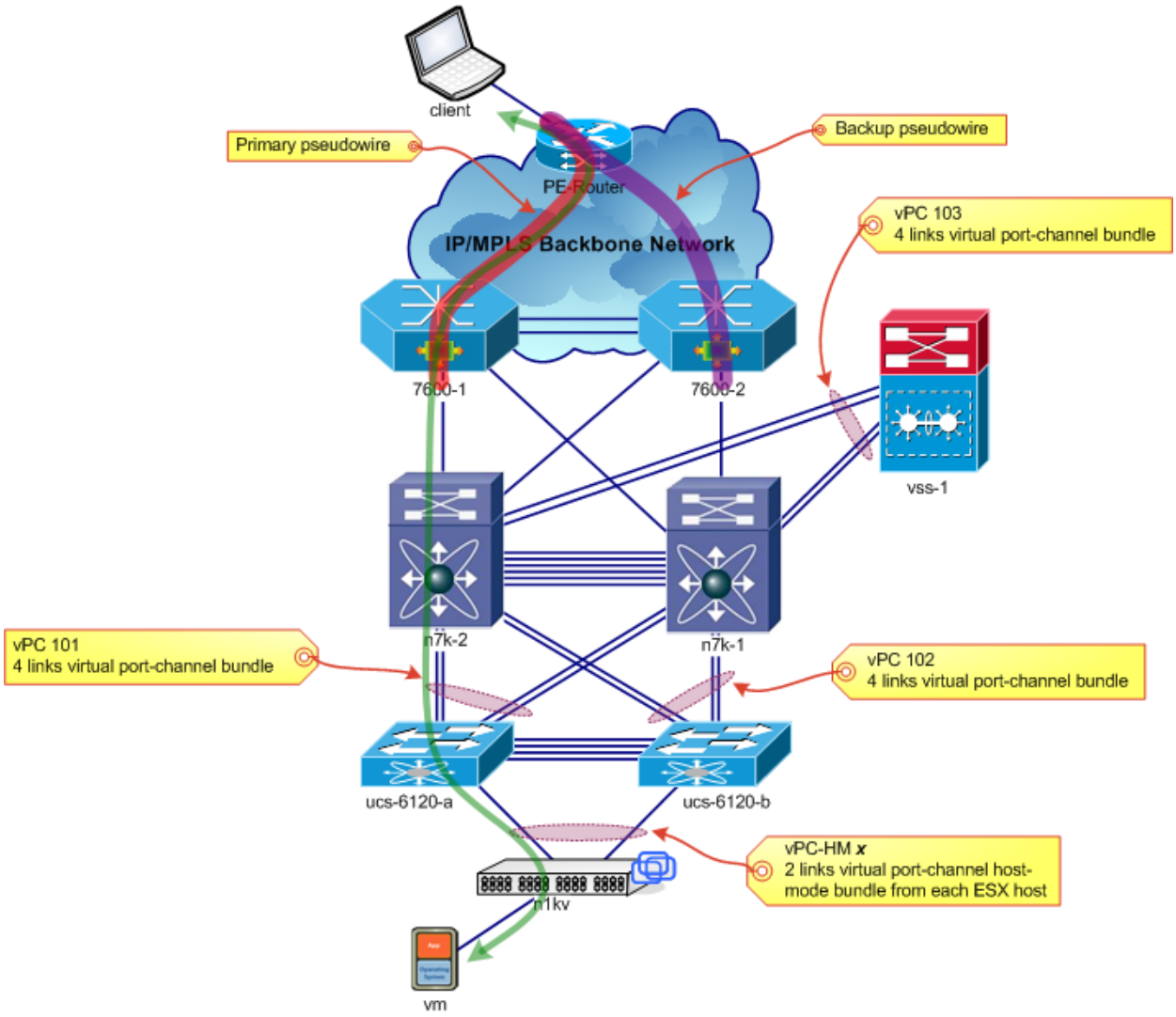


Figure 2-44 Bronze Service Class L2VPN Setup



The configuration for L2VPN is done only on Cisco 7600 routers. The configuration on DSN is limited to creating the VLANs for access to the ACE server load balancers and firewall service modules that reside within the DSN chassis.

Note VLAN interfaces are configured on the DSN for QoS coloring/marking purposes, as the ACE server load balancers and firewall service module do not perform packet coloring/marking.

L2VPN configuration for Gold, Silver, and Bronze service classes on Cisco 7600 routers is the same. The differences between the service classes are in the QoS treatment and inclusion of other auxiliary services, such as server load balancing and firewalling.

Note Each Ethernet service instance can change the existing VLAN tag to be a new VLAN tag by adding, removing, or translating one or two VLAN tags. Flexible VLAN tag rewrite includes three main operations:

- pop (remove an existing tag)
- push (add a new tag)
- translate (change one or two tags to another one or two tags) - this can be seen as a combination of pop and push operations

Theoretically, any existing combination of one or two VLAN tags can be changed to any new combination of one or two VLAN tags by just using a simple line of configuration. Practically, there are some limitations:

- Always use the `symmetric` keyword, although the CLI might not give this impression. Generally rewrite configurations should always be symmetric. For each rewrite on the ingress direction, there should be a reverse rewrite on the egress direction for the same service instance configuration. So, if the outer VLAN tag is popped on ingress direction, the original outer VLAN tag must be pushed back on the egress direction for that same service instance. The system does all this automatically when the `symmetric` keyword is used.
- Due to the mandatory symmetry, some operations can only be applied to a unique tag matching service instance (so they are not supported for VLAN range configurations) or cannot be applied at all.
- One Ethernet service instance can have none or at most one VLAN tag rewrite configuration. If there is no VLAN tag rewrite configuration, existing VLAN tags are kept unchanged. It cannot have more than one VLAN tag rewrite configuration for a particular Ethernet service instance.

On the PE router, L2VPN pseudowire redundancy is implemented to setup backup pseudowire to provide some level of resiliency for network failure. The PE router maintains two pseudowires in active/standby mode to the Cisco 7600 routers. The PE router is not part of this phase of testing. The configuration of the PE router is included here for completeness.

Note PE/CE redundancy is not considered for this phase of the solution.

For more information on L2VPN pseudowire redundancy, see:

http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/fspseudo.html

Data Traffic Routing

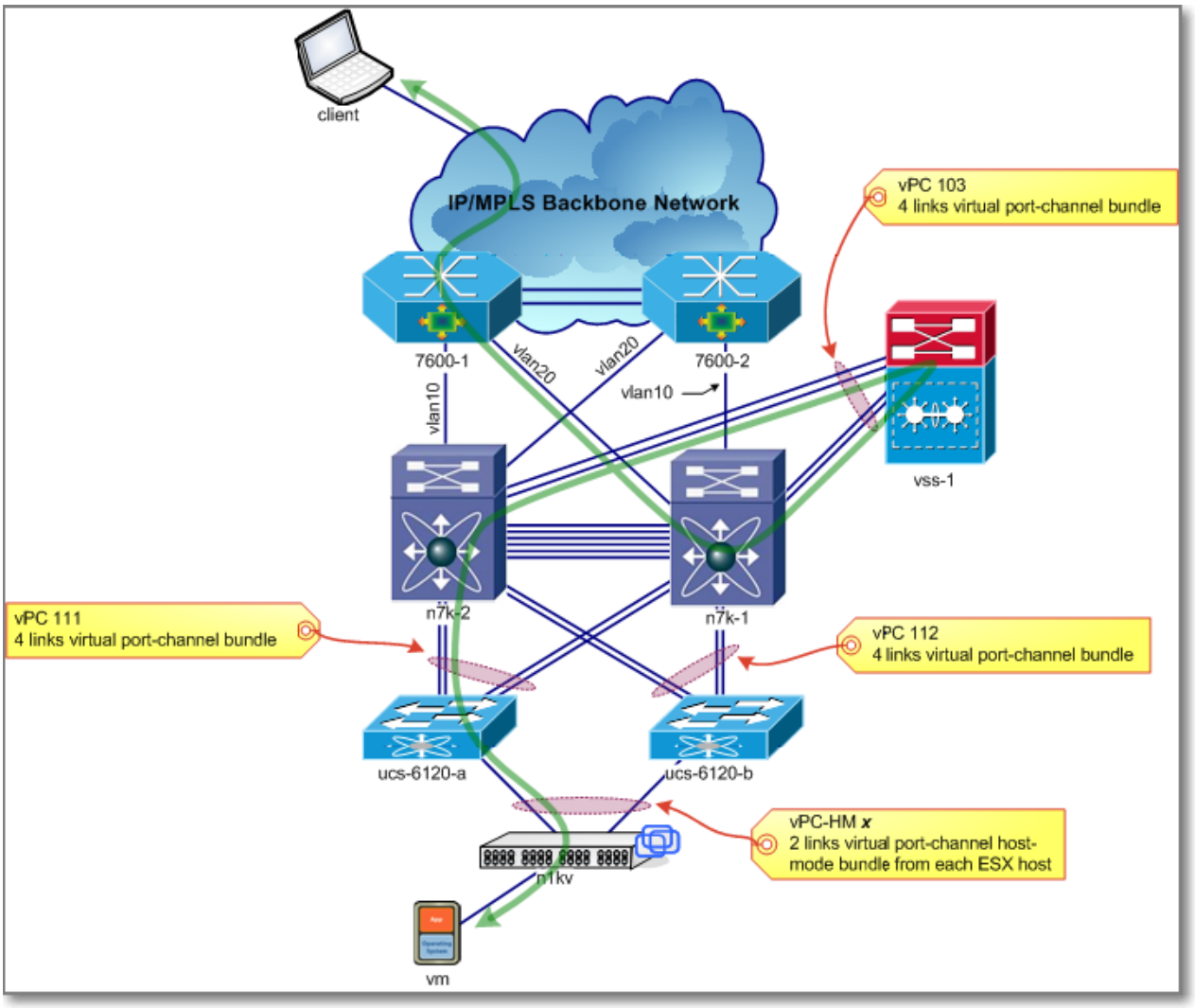
This section discusses the design of IP, L3VPN, and L2VPN data traffic routing in the network cloud.

In the cloud network, the Gold service class includes server load balancer, firewall, multiple virtual machines, best compute, and network resources allocations within the cloud network; the Silver service class includes server load balancer, multiple virtual machines, moderate compute, and network resource allocations within the cloud network; the Bronze service class includes one or two virtual machines and best effort compute and network resources allocation within the cloud network.

IP Routed Data Traffic

For all three service classes, the IP data traffic takes the same basic path as shown in Figure 2-45.

Figure 2-45 IP Routed Data Traffic



Note

- Data traffic for Gold and Silver service classes passes through firewall and server load balancer, but since both the firewall and server load balancer are service modules within the vss-1 chassis, IP data traffic path is not different from those of Bronze service class.
- All Cisco 7600 router links connected to Nexus 7000 switches are configured as non-switchport 802.1Q trunk with routed sub-interfaces.
- Figure 2-45 data traffic takes a symmetric path in both directions. The actual path could be different for each direction. The actual path/link use by data traffic depends on the IP routing table of the device handling the data traffic and the port-channel hashing algorithm configured on the sending device.

- **Virtual Machine to Client Data Traffic**—The Bronze service class virtual machines use the IP address of the VLAN interfaces on the vss-1 router as the default gateway. Gold service class virtual machines use the IP address of the firewall service module inside interfaces as the default gateway.

Silver service class virtual machines use the alias IP address the ACE service module inside interfaces as the default gateway. In all three service classes, outgoing traffic from the virtual machines is routed or forwarded to the vss-1.

The vss-1 router learns one BGP default route from 7600-1 and 7600-2 router respectively. The BGP process on the 7600-1 and 7600-2 peers with the BGP process on the vss-1 using loopback interface and sets the next-hop for the default route they advertise to vss-1 to their respective loopback IP addresses.

The BGP next-hops of the two default routes are loopback interfaces of 7600-1 and 7600-2. The vss-1 router uses recursive lookup to resolve the next-hops. The vss-1 learns routes to the loopback interface of 7600-1 and 7600-2 respectively through OSPF on VLAN 10 interface and VLAN 20 interface.

VLAN 10 and VLAN 20 interfaces are tied to vPC 103 switchport. Outgoing data traffic from vss-1 has four paths to the backbone network, all via vPC 103, to each of the physical link on the 7600-1 and 7600-2 respectively.

- **Client to Virtual Machine Data Traffic**—Each Gold, Silver, or Bronze service class customer has one or more subnets in which the virtual machines reside. The vss-1 router aggregates reachability to these customer-VM subnets and advertises to the Cisco 7600 routers via BGP. The vss-1 router advertises these customer-VM BGP routes using its loopback address as the next-hop.

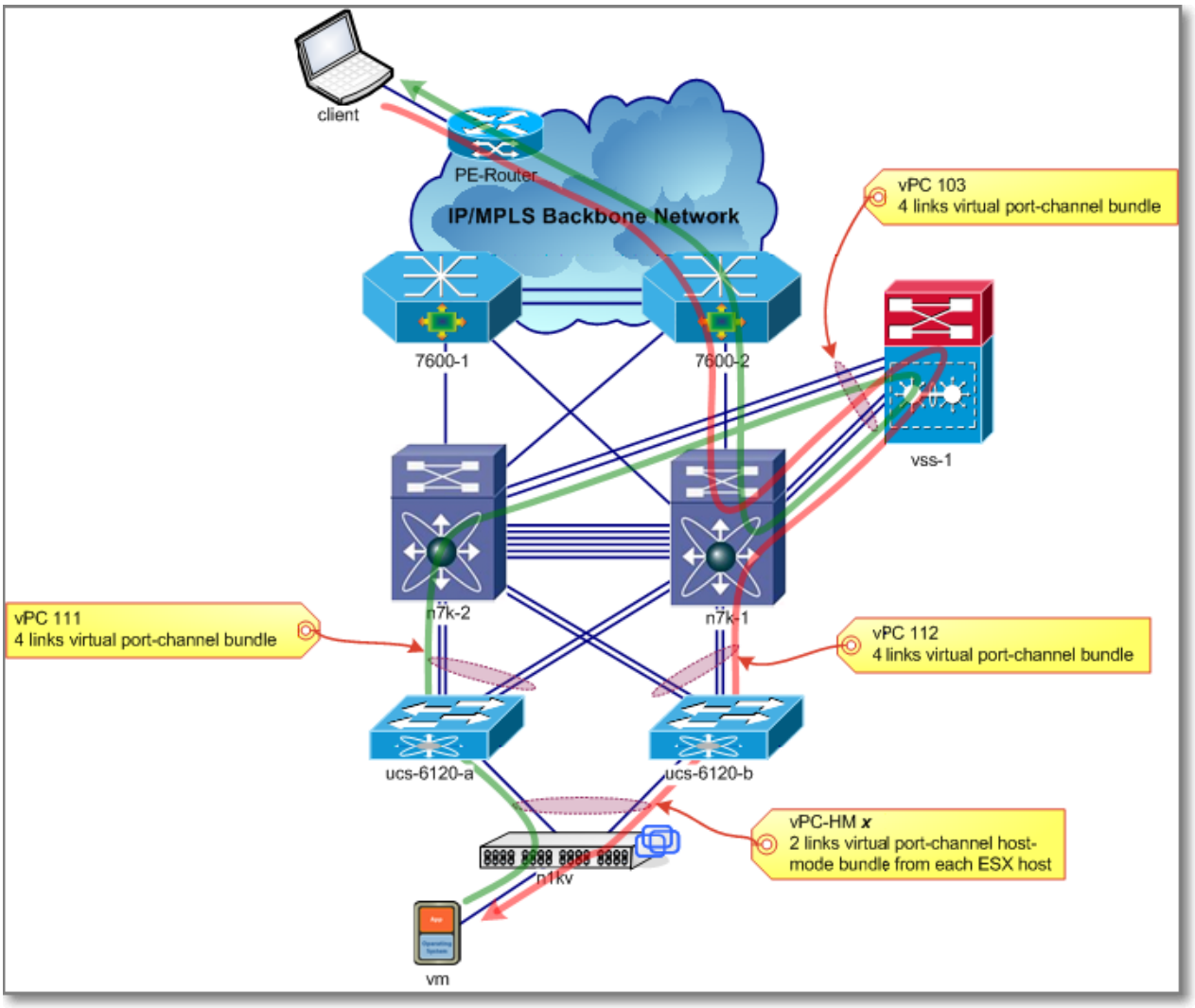
Note Customer-VM subnet (or network/VLAN/route) is the virtual machines (or ACE VIPs) subnet (or network/VLAN) of a particular customer in the cloud network. The notation is used as differentiation to customer network (or route). *Customer network* is the network outside the cloud network where the clients of the virtual machines reside.

The next-hop of the customer-VM BGP routes is not a directly connected interface. 7600-1 and 7600-2 use recursive lookup to resolve the next-hop and rely on OSPF to learn the route to the next-hop on VLAN 10 and VLAN 20 interfaces.

L3VPN Data Traffic

Data traffic for Gold and Silver service classes takes the same basic path as shown in Figure 2-46.

Figure 2-46 L3VPN Gold and Silver Service Classes Data Traffic



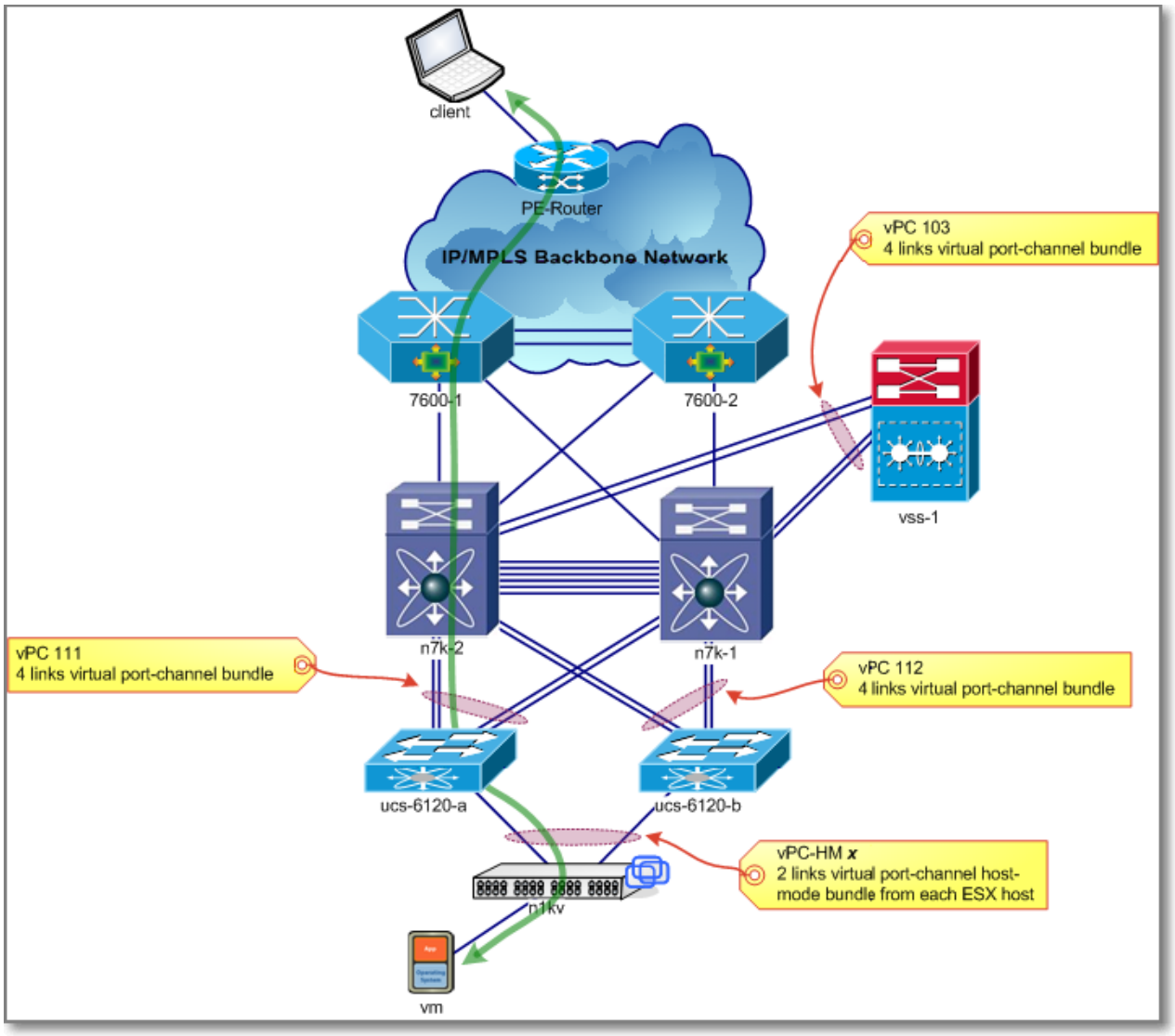
Note

- Data traffic for Gold and Silver service classes pass through firewall and server load balancer, but since both the firewall and server load balancer are service modules within the same vss-1 chassis, IP routing data traffic path is the same for both classes.
- All Cisco 7600 router links connected to Nexus 7000 switches are configured as non-switchport 802.1Q trunk with routed sub-interfaces.
- Unlike the IP routed data traffic, the routed part of the L3VPN data traffic is symmetric, i.e. traffic for the same customer always exits and enters the cloud network using the same link on the same Cisco 7600 router. Figure 2-46 shows the L3VPN data traffic entering and exiting the cloud network via the same link on the same Cisco 7600 router (this is true even when both Cisco 7600 routers are online). This behavior is due to the manual L3VPN routing configuration.

- When L3VPN data traffic traverses the virtual port-channel bundles (vPC 101, vPC 102, and vPC 103), data traffic to and from the virtual machine might use different member links within the port-channel bundle, as shown in Figure 2-46. The actual path/link used by data traffic depends on the port-channel hashing algorithm configured on the sending device.

Bronze service class data traffic takes a different path from that of Gold and Silver classes. Bronze service class data traffic does not pass through the vss-1 router, as shown in Figure 2-47. Bronze service class data traffic is not routed within the cloud network. It is switched by the Nexus 7000, Nexus 5000, and Nexus 1000V switches to the HRSP sub-interface of the Cisco 7600 router.

Figure 2-47 L3VPN Bronze Service Class Data Traffic



Note Figure 2-47 shows data traffic taking a symmetric path for both directions. The actual path could be different for each direction. The actual path/link use by data traffic depends on the port channel hashing algorithm configured on the sending device.

- **Virtual Machine to Client Data Traffic**
  - *Gold and Silver Service Classes*. Gold service class virtual machines use the IP address of the firewall service module inside interfaces as the default gateway. The Silver service class virtual machines use the alias IP address the ACE service module inside interfaces as the default gateway. For both service classes, the outgoing traffic from the virtual machines is routed to the vss-1 router.

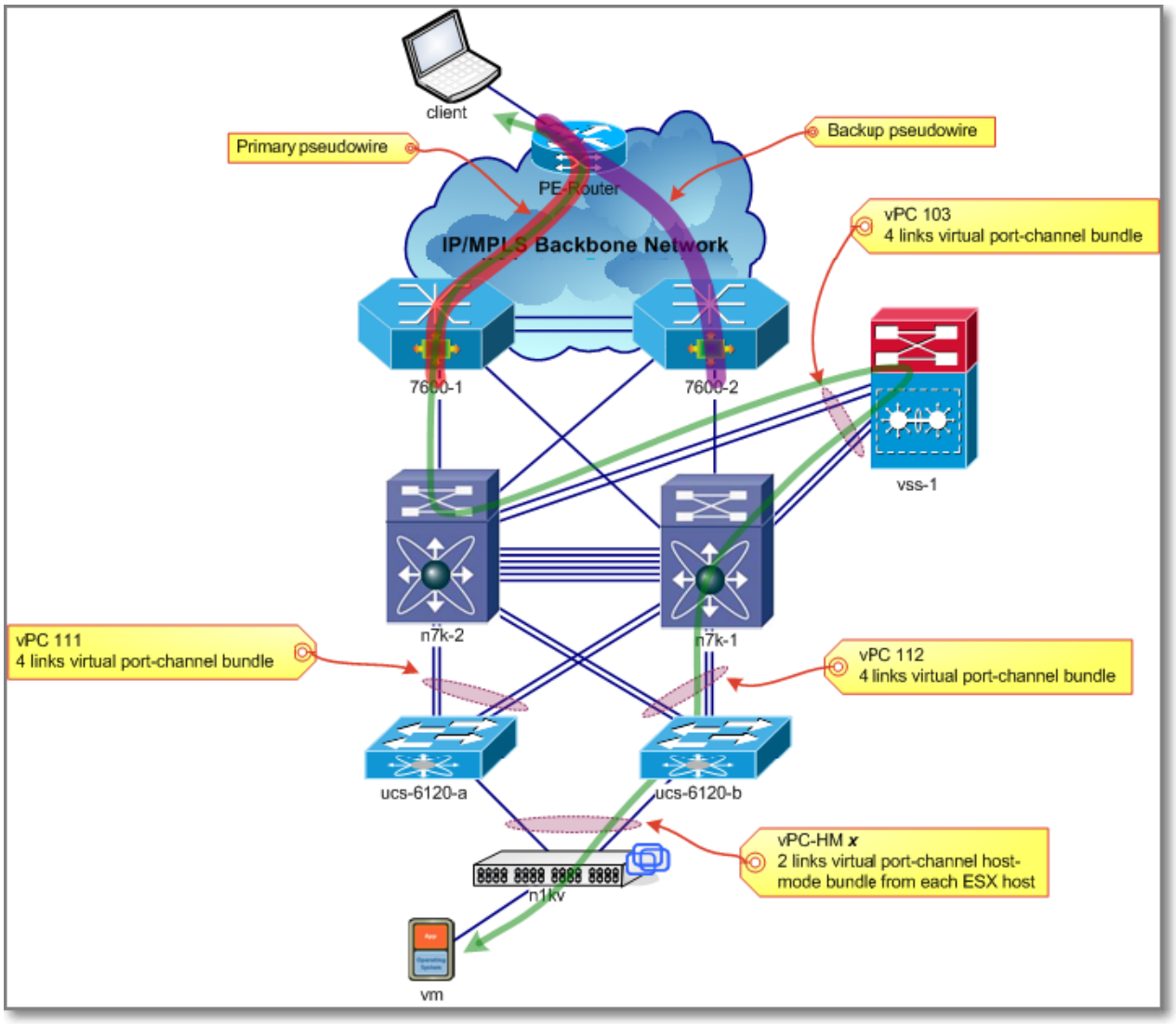
    On the vss-1 router, each customer has a VLAN interface where the ACE outside interface for the customer resides that is placed into a VRF context. This VRF VLAN interface learns L3VPN customer routes 7600-1 and 7600-2 L3VPN VRF advertises over BGP (for L3VPN, BGP peers using the VRF interface). 7600-1 and 7600-2 set different route metrics for the BGP routes they advertise to the vss-1 such that the vss-1 will prefer one router over the other in its route selection. For each L3VPN customer route, the vss-1 router learns two BGP routes, of which only one is preferred because of lower BGP metric.
  - *Bronze Service Class*. Bronze service class virtual machines use Cisco 7600 routers as default gateway. For redundancy, HSRP is configured. The vss-1 router is not involved.

    For each L3VPN customer, one sub-interface is configured on 7600-1 and 7600-2. Each sub-interface is placed into a VRF. Cisco 7600 routers are expected to handle many customers, each with their own sub-interface. Customer sub-interfaces are manually distributed evenly across both Cisco 7600 routers, and the HRSP role for each customer sub-interface manually configured to ensure that no one router (and/or physical interface) is heavily loaded as HSRP active router.
- **Client to Virtual Machine Data Traffic**—Each Gold, Silver, or Bronze service class customer has one or multiple subnets in which virtual machines reside. For Gold and Silver class customers, vss-1 router advertises reachability to virtual IP addresses on ACE server load balancer (client does not hit the real IP address of the virtual machine directly, but rather hits the ACE's VIP) to the Cisco 7600 routers over BGP.
  - *Gold and Silver Service Classes*. For Gold and Silver server customers, each Cisco 7600 router changes the metric of the L3VPN customer-VM BGP routes it receives from the vss-1 router, so that only one Cisco 7600 router is used as the entry point for the customer L3VPN data traffic from client to the virtual machine. One Cisco 7600 router advertises the best customer-VM BGP routes to the PE router via MP-BGP with the changed route metric.
  - *Bronzed Service Class*. For Bronze service class customers, the customer-VM subnets are directly connected to Cisco 7600 routers, and connected customer-VM subnets are redistributed into MP-BGP. Each Cisco 7600 router sets the BGP metric of the customer-VM routes, so that only one Cisco 7600 router is used as the entry point for the customer L3VPN data traffic from client to the virtual machine.

L2VPN Data Traffic

The Gold and Silver service classes, data traffic takes the same basic path as shown in Figure 2-48:

Figure 2-48 L2VPN Gold and Silver Service Classes Data Traffic

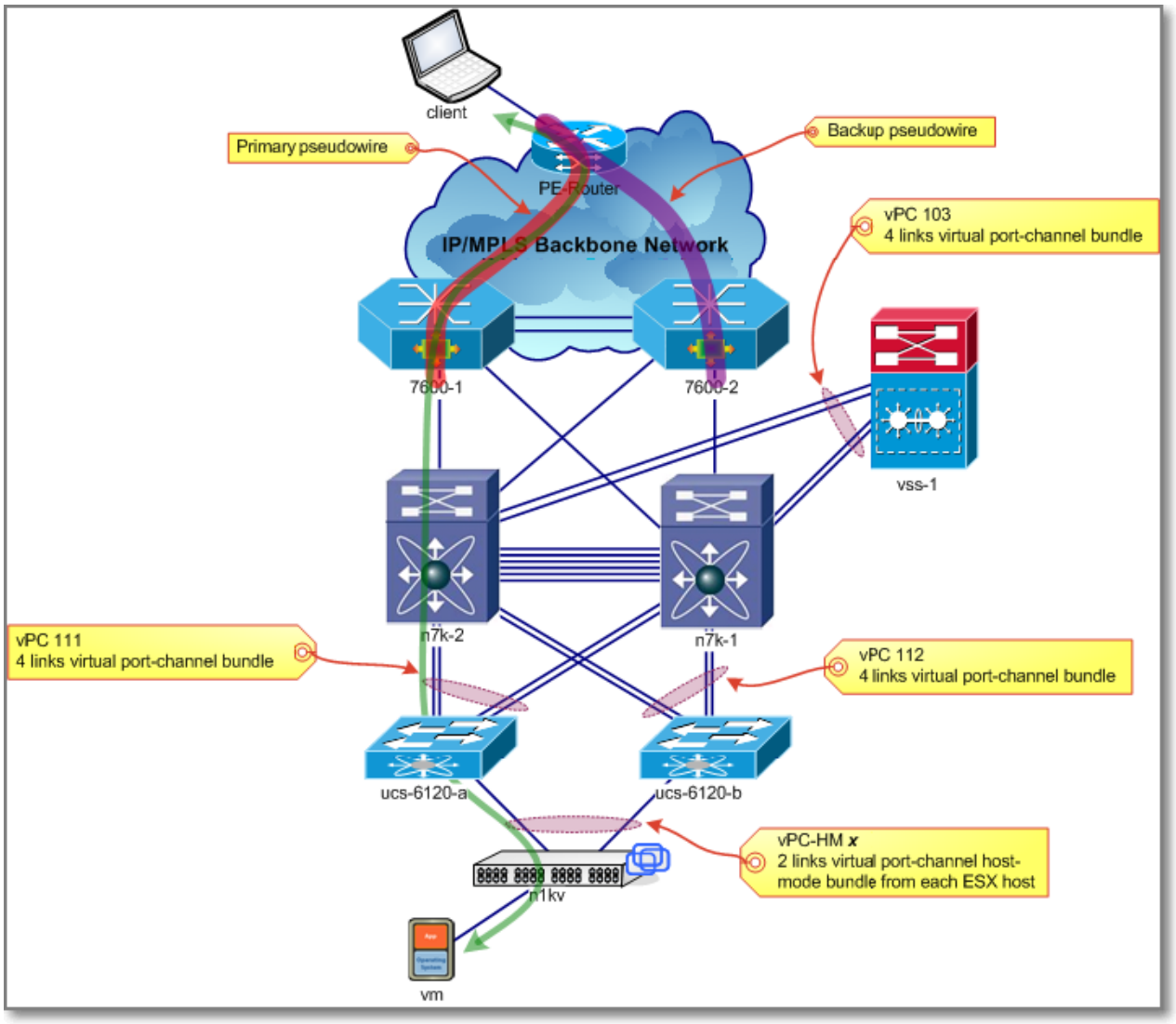


Note

- Data traffic for Gold and Silver service classes passes through firewall and/or server load balancer, but since both the firewall and server load balancer are service modules within the same vss-1 chassis, from routing/forwarding perspective, the data traffic path is the same for both classes.
- All Cisco 7600 routers links connected to the Nexus 7000 switches are configured as non-switchport 802.1Q trunk, with routed sub-interfaces.
- When the L2VPN data traffic traverse the (virtual) port-channel bundles (vPC 101, vPC 102, and vPC 103), the data traffic to and from the virtual machine might use different member link within the port-channel bundle. The actual path/link use by data traffic depends on the port-channel hashing algorithm configured on the sending device.

Bronze service class data traffic takes a different path from that of Gold and Silver classes. Bronze service class data traffic does not pass through the vss-1 router, as shown in Figure 2-49.

Figure 2-49 L2VPN Bronze Service Class Data Traffic



Note Figure 2-49 shows data traffic taking a symmetric path for both directions. The actual path could be different for each direction. The actual path/link use by data traffic depends on the port-channel hashing algorithm configured on the sending device.

Gold service class virtual machines use the IP address of the firewall service module inside interfaces as the default gateway. Silver service class virtual machines use the alias IP address the ACE service module inside interfaces as the default gateway; for both service classes, outgoing traffic from virtual machines is routed/forwarded to the vss-1 router.

On the vss-1 router, the ACE outside VLAN is Layer 2 adjacent to the Ethernet service instance on each Cisco 7600 router, no IP routing between the vss-1 and Cisco 7600 routers, just Layer 2 forwarding.

Each Cisco 7600 router has one pseudowire connection to the same PE router. The PE router maintains the pseudowire in active/backup setting; the backup pseudowire connection does not forward traffic until the active pseudowire fails. Data traffic forwarding paths on Cisco 7600 routers is simplified because of the active/backup pseudowire setup, i.e., only one Cisco 7600 router forwards data traffic for a particular customer at any one time.

Services Layer Design

This section contains the following design topics:

- The services layer is a DataCenter Services Node (6500 chassis pair running in Virtual Switching System mode) with ACE and FWSM installed as service modules. These service modules provide virtualization, health checks, sticky, SSL off-load, inspection advanced services.
- The FWSM and the ACE module are installed in each chassis to support stateful failover in case of single module failure. Up to four FWSM modules and four ACE modules can be installed per chassis. Active-Active redundancy is configured to distribute active contexts and load across fault-tolerant ACE and FWSM service modules.

Three service classes are provisioned in this system. Gold and Silver service classes make use of the services module for advanced services, while the Bronze tier bypasses these modules. In addition, to the levels of advanced services configured, these classes are provisioned with different levels of QoS and virtual machine (VM) resources allowing a structured and flexible solution that provides different levels of customers services. For more information about QoS or VM implementation, see Quality of Service Design and Compute Virtualization.

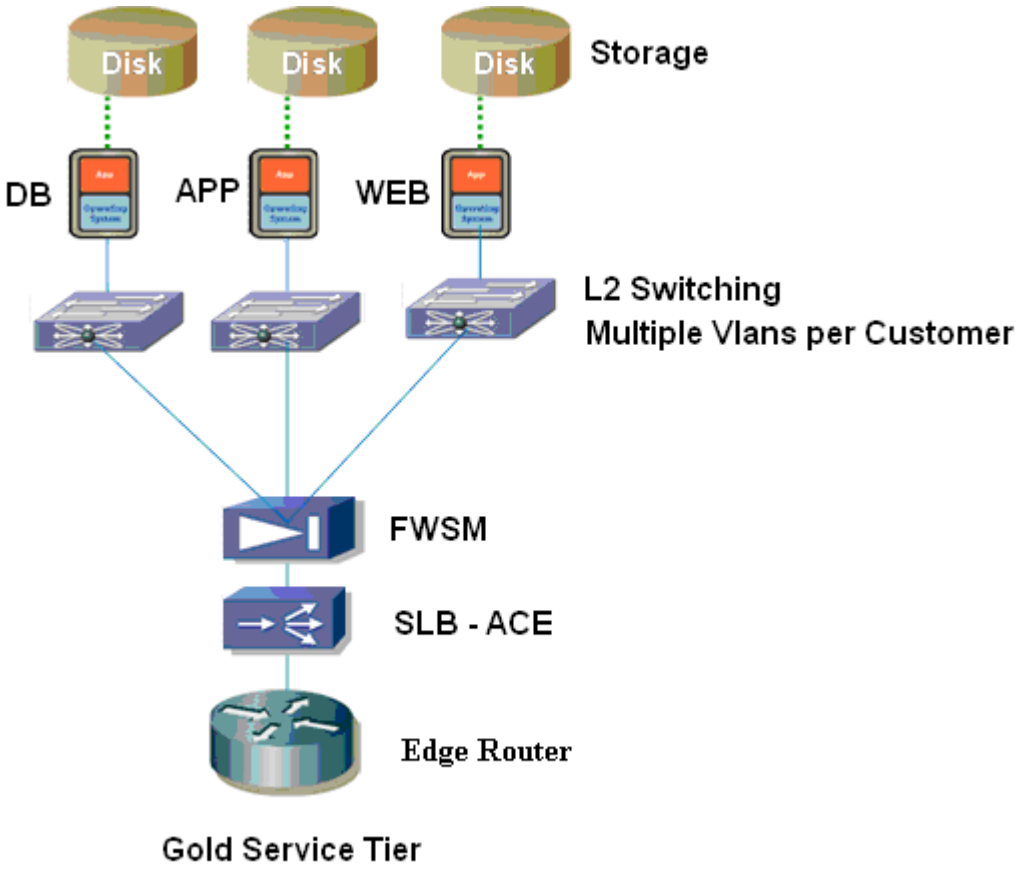
Gold

Gold service class includes server load balancer, firewall, multiple virtual machines, best compute, and network resources allocations within the network. These contexts are configured to use server load-balancing (ACE) and the firewall service module (FWSM).

Virtual machine traffic is segregated into three VLANs, one each for WEB, APP, and DB services. Each class of service uses the FWSM as its gateway to route between VLANs or the Internet. The FWSM uses a default route to inside VLAN on the ACE. The ACE routes to an L3 SVI on the services switch and then to the hosting routers.

The virtual IP (VIP) address used for load-balancing traffic is advertised to the service layer switch using route health injection (RHI). The VIPs injected use separate IP space from that of the outside ACE VLAN SVI. Non-VPN and L3VPN customers use RHI to inject VIP host routes from each customer's outer VLAN subnet on the ACE. The VIPs are static routes injected into the local routing table. These are aggregated to a /8 and injected into iBGP to be used by the hosting router. These VIPs are used for NAT and SLB traffic. RHI is not currently supported for an ACE NAT VIP, so static routes are manually configured on the services switch. For L2VPN customers, the ACE outer VLANs are Layer 2 carried by the VSS to the edge router by way of the Nexus 7000 core. L2VPN clients use the ACE alias (FT) IP address as a gateway to VIPs that serve for NAT and SLB.

Figure 2-50 Gold Service Class



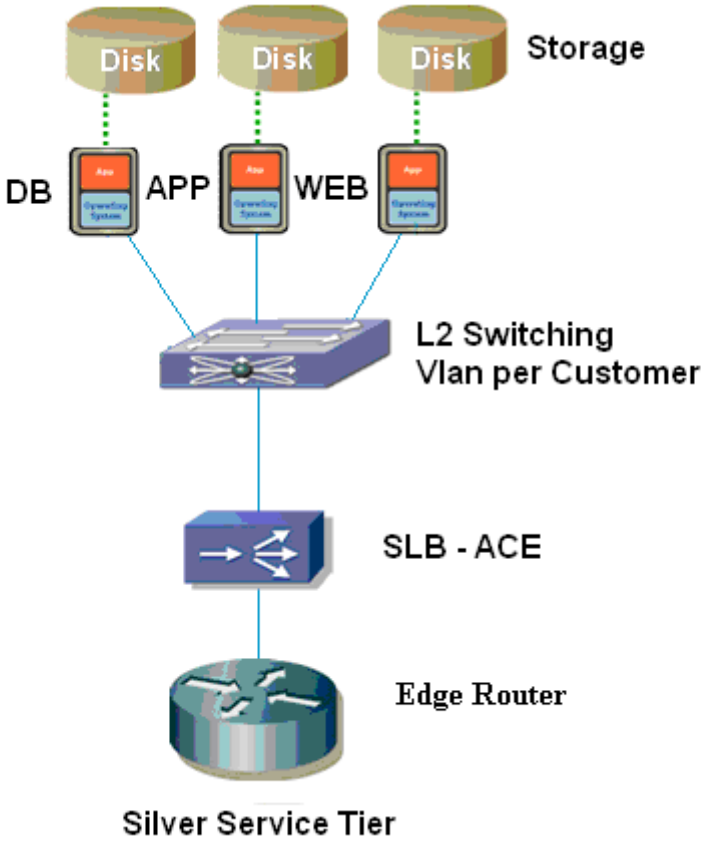
Silver

The Silver service class includes server load balancer, multiple virtual machines, moderate compute, and network resources allocations within the network. These contexts are configured for server load-balancing and do not use the FWSM.

Virtual machine traffic for each customer is on a shared VLAN for WEB, APP, and DB traffic. Traffic is routed to the ACE, L3 SVI on the services switch, and then to hosting routers.

The virtual IP address used for load-balancing traffic is advertised to the services switch using route health injection (RHI). The virtual IP addresses injected use separate IP space from that of the outside ACE VLAN SVI. RHI is used for non-VPN and L3VPN customers to inject virtual IP address host routes from each customer's outer VLAN subnet on the ACE. The VIPs are static routes injected into the local routing table. These are aggregated to a /8 and injected into iBGP to be used by the edge router. These virtual IP addresses are used for NAT and SLB traffic. RHI is not currently supported for an ACE NAT VIP, so static routes are manually configured on the services switch. For L2VPN customers, the ACE outer VLANs are Layer 2 carried by the VSS to the edge router by way of the Nexus 7000 core. The L2VPN clients use the ACE alias (FT) IP address as a gateway to virtual IP addresses that serve as NAT and SLB.

Figure 2-51 Silver Service Class

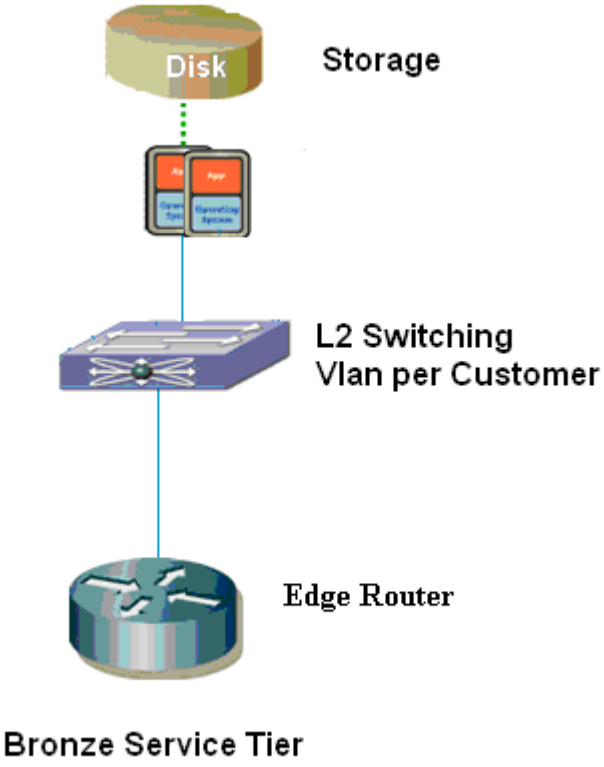


Bronze

The Bronze service class includes one or two virtual machines and best effort compute and network resources allocation. The Bronze class is not provisioned for advanced services, so it bypasses the service modules and routes to a L3 SVI on the VSS or directly to the edge / hosting router.

Virtual machine traffic for customer L3 SVI VLANs is routed first to the VSS and then to hosting routers. L2VPN bound destinations are dedicated per-customer VLANs, extended at L2, up to the edge router, bypassing the VSS. The L3 gateway for Layer 2 extended VPN customers is at the edge router.

Figure 2-52 Bronze Service Class



Data Center SAN Design

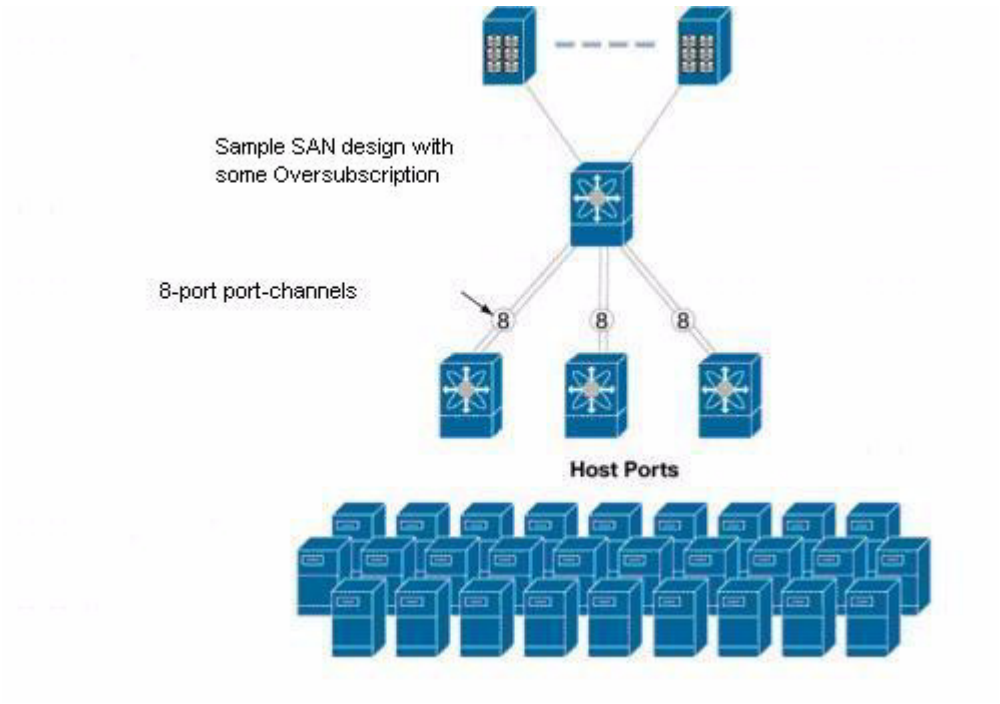
To understand SAN, it is best to break it up into the components that comprise the whole picture. First, the host component can vary from real servers to virtual machines, with operating systems as varied as UNIX/LINUX based operating systems, Microsoft Windows, and others. These *initiators* attach to the SAN by means of a host bus adapter (HBA) or converged network adapter (CNA), typically Emulex or QLogic brand, which takes up a PCI slot on the physical host. Another component of the SAN, storage devices, can be as small as JBODs or tape drives to the larger and more complex block storage arrays. Also known as *targets*, these devices connect to the SAN through means of HBAs/CNAs similar to the hosts. The final part of a SAN are the switches that move the data from one end to the other. On the MDS, features allow data encryption, IP wrapping, data mobility managing, write acceleration, and so on, to occur to optimize traffic speeds or fulfill requirements specific to an individual customer.

It is common practice in SAN environments to build two separate, redundant physical fabrics (Fabric A and Fabric B) in case a single physical fabric fails. When designing for large networks, most environments fall into two types of topologies within a physical fabric:

- Two-tier: Core-edge design
- Three-tier: Edge-core-edge design

Within the two-tier design, servers connect to the edge switches, and storage devices connect to one or more core switches (see Figure 2-53). This allows the core switch to provide storage services to one or more edge switches, thus servicing more servers in the fabric. Interswitch links (ISLs) must be designed so that the overall fabric maintains the fan-out ratio of servers to storage and the overall end-to-end oversubscription ratio.

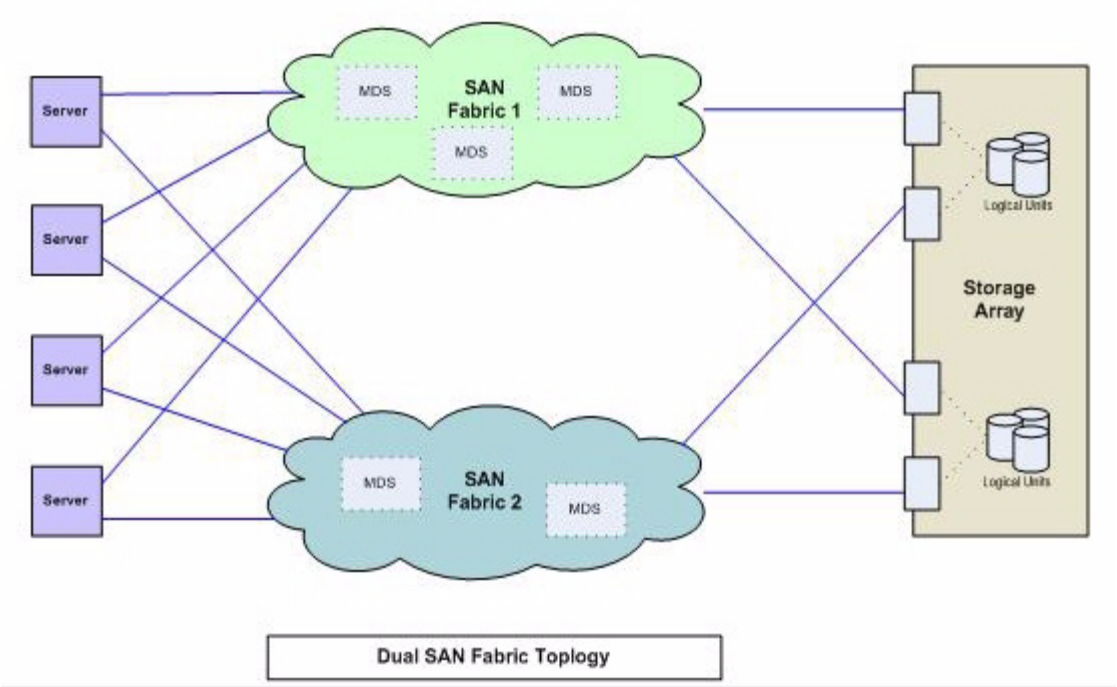
Figure 2-53 SAN Design



Fabric Redundancy

Another area that requires attention in a Fibre Channel SAN is the fabric itself. Each device connected to the same physical infrastructure is in the same Fibre Channel fabric. This opens up the SAN to fabric-level events that could disrupt all devices on the network. Changes such as adding switches or changing zoning configurations could ripple through the entire connected fabric. Therefore, designing with separate connected fabrics helps to isolate the scope of such events. A common practice is to build multiple parallel fabrics and multi-homed hosts and disks into the parallel, physically isolated fabrics. Generally, the primary reason for this isolation is to help ensure that fabric services such as the name service are isolated within each fabric. If a fabric service fails, it does not affect the other parallel fabrics. Therefore, parallel fabrics provide isolated paths from hosts to disks.

Figure 2-54 Dual SAN Setup



VSAN

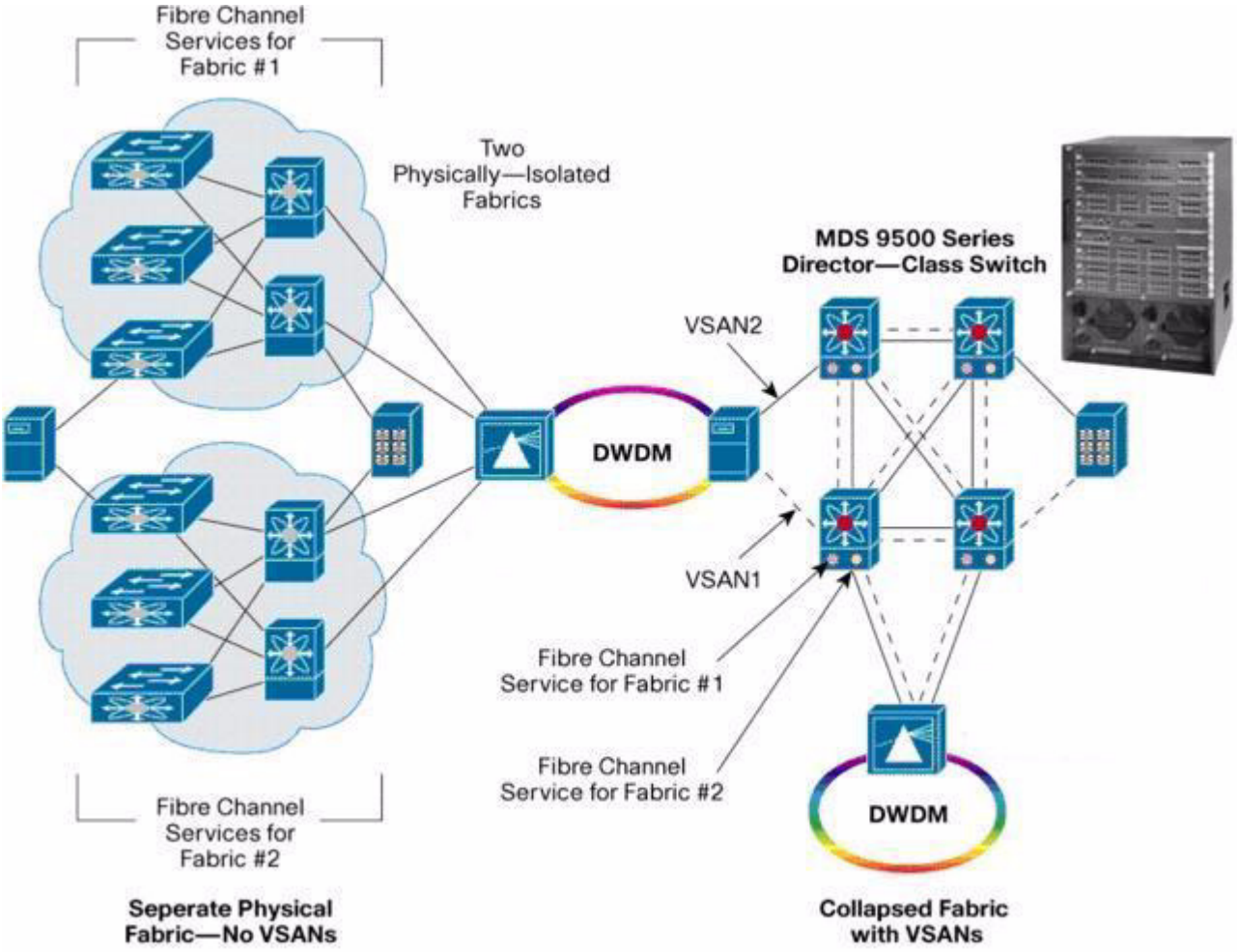
To help achieve the same isolated environments while eliminating the added expense of building physically separate fabrics, Cisco has introduced the VSAN within the Cisco MDS 9000 family. A VSAN provides the ability to create separate virtual fabrics on top of the same physical infrastructure. Each separate virtual fabric is isolated from the others using a hardware-based frame-tagging mechanism on ISLs. An EISL is an enhanced ISL that includes added tagging information for each frame and is supported on links interconnecting any Cisco MDS 9000 family switch product. Membership in a VSAN is based on physical port, and no physical port can belong to more than one VSAN. Therefore, a node connected to a physical port becomes a member of that port's VSAN.

The Cisco MDS 9000 family of products supports 1024 VSANs per physical infrastructure. Each VSAN can be selectively added to or pruned from an EISL to control the VSAN's reach. In addition, special traffic counters are provided to track statistics per VSAN.

Probably the most highly desired characteristic is the high availability profile of VSANs. Not only do VSANs provide strict hardware isolation, but also a full replicated set of Fibre Channel services is created for each new VSAN. Therefore, when a new VSAN is created, a completely separate set of services, including name server, zone server, domain controller, alias server, and login server, is created and enabled across those switches that are configured to carry the new VSAN. This replica of services provides the ability to build the isolated environments needed to address high availability concerns over the same physical infrastructure. For example, an installation of an active zone set within VSAN 1 does not affect the fabric in any way within VSAN 2. VSANs also provide a method to interconnect isolated fabrics in remote data centers over a common long-haul infrastructure. Because the frame tagging is done in hardware and is included in every EISL frame, it can be transported across transports such as dense wavelength-division multiplexing (DWDM) or coarse wavelength-division multiplexing (CWDM). Therefore, traffic from several VSANs can be multiplexed across a single pair of fibers and

transported a greater distance and yet still remain completely isolated. VSANs bring scalability to a new level by using a common redundant physical infrastructure to build flexible isolated fabrics to achieve high-availability goals.

Figure 2-55 Comparison of Physically Separate Fabrics to Collapsed VSAN Fabric



Zoning

In each VSAN, there is only one active zoneset that contains one or more zones. Each zone consists of one or more members to allow for communication between members. Zoning provides access control for devices within a SAN. Cisco NX-OS software supports the following types of zoning:

- N port zoning—Defines zone members based on the end-device (host and storage) port.
  - WWN
  - Fibre Channel identifier (FC-ID)
  - Fx port zoning---Defines zone members based on the switch port
  - WWN
  - WWN plus interface index, or domain ID plus interface index

- Domain ID and port number (for Brocade interoperability)
- iSCSI zoning—Defines zone members based on the host zone
  - iSCSI name
  - IP address
- LUN zoning—When combined with N-port zoning, LUN zoning helps ensure that LUNs are accessible only by specific hosts, providing a single point of control for managing heterogeneous storage-subsystem access.
- Read-only zones—An attribute can be set to restrict I/O operations in any zone type to SCSI read-only commands. This feature is especially useful for sharing volumes across servers for backup, data warehousing, etc.
- Broadcast zones—An attribute can be set for any zone type to restrict broadcast frames to members of the specific zone.

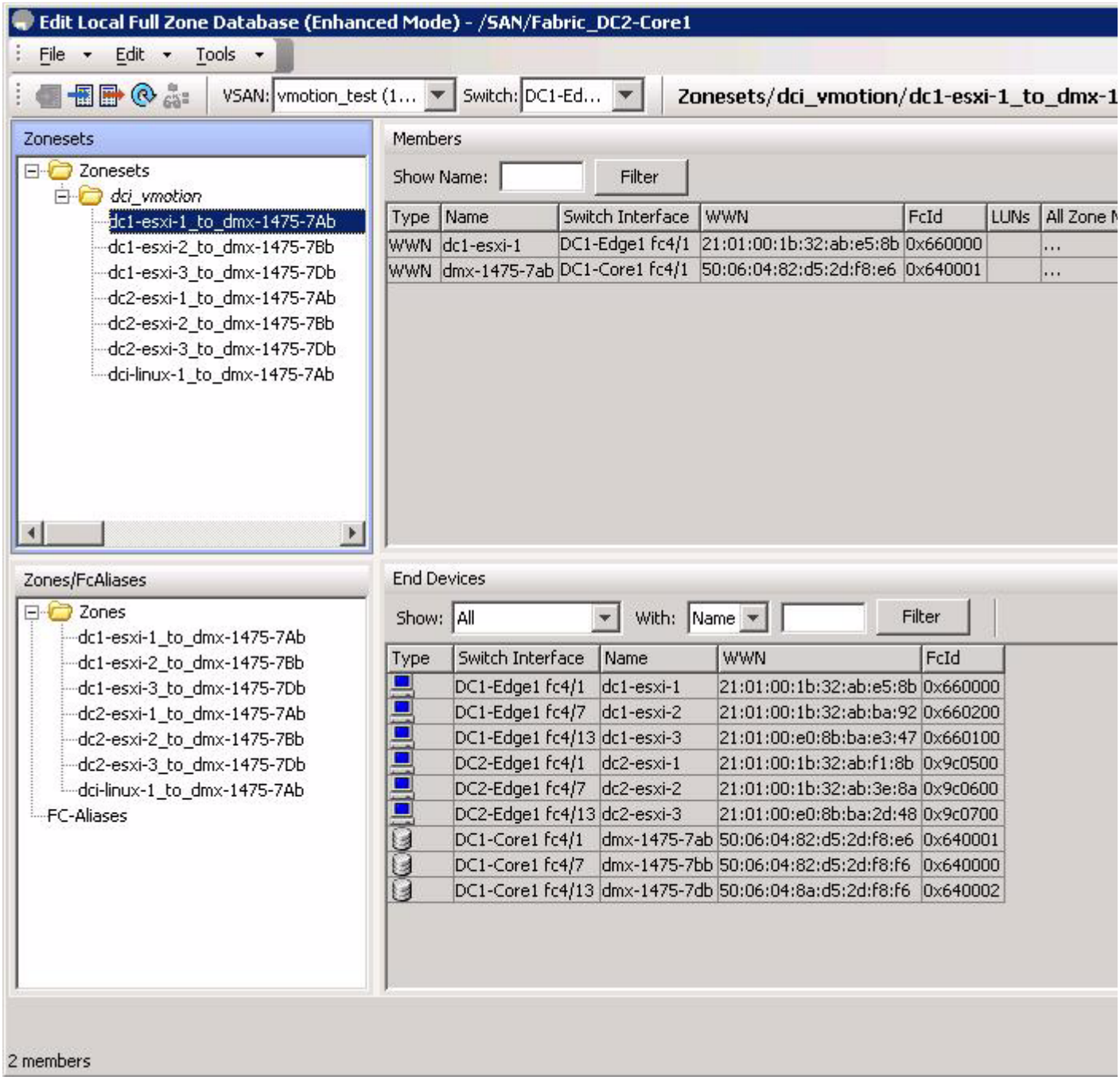
To provide strict network security, zoning is always enforced per frame using access control lists (ACLs) that are applied at the ingress switch. All zoning polices are enforced in hardware, and none of them cause performance degradation. Enhanced zoning session-management capabilities further enhance security by allowing only one user at a time to modify zones.

Depending on the requirements of the environment, choosing the type of zone members is a matter of preference. A recommended best practice is to create a `device-alias` for end devices when managing the network. The device-alias provides an easy-to-read name for a particular end device. For example, a storage port World Wide Name (pWWN) 50:06:04:82:bf:d0:54:52 can be given a device-alias name of Tier1-arrayX-ID542-Port2. In addition, with device-alias, when the actual device moves from one VSAN (VSAN 10) to a new VSAN (VSAN 20) in the same physical fabric, the device-alias name will follow that device. So there is no need to re-enter the device-alias for the new VSAN.

Cisco MDS switches support up to 8000 zones and 20,000 zone members in a physical fabric. There are things to consider for very large environments that may reach this limit.

The following excerpt shows what the user sees in the CLI when viewing the configuration of an active zoneset. The statement in bold is the CLI command to view an active zoneset for VSAN 100. In the output, the user sees the zoneset name for VSAN 100 and zones that are members of the active zoneset. The zones have their own names (given by the user) and are followed by their members. The members that have an asterisk (*) at the beginning of the line show that they logged into the fabric. Next for the zone members are the pwwns that were added and, if configured, the device-alias name given to each pWWN in the brackets.

Figure 2-56 Fabric Manager View of Zoning



The following sample code shows a CLI view of zoning:

```
mds-9222i-1# show zoneset vsan 10
zoneset name infrastructure vsan 10
  zone name spdc-esx01 vsan 10
    pwwn 10:00:00:00:c9:76:f2:1c [spdc-esx01-vhba1]
    pwwn 50:0a:09:81:88:0c:74:aa [Storage Array]
    pwwn 10:00:00:00:c9:76:f2:1d [spdc-esx01-vhba2]

  zone name spdc-esx02 vsan 10
```

```
    pwwn 50:0a:09:81:88:0c:74:aa [Storage Array]
    pwwn 10:00:00:00:c9:76:f2:80 [spdc-esx02-vhab1]
    pwwn 10:00:00:00:c9:76:f2:81 [spdc-esx02-vhba2]

  zone name spdc-esx03 vsan 10
    pwwn 50:0a:09:81:88:0c:74:aa [Storage Array]
    pwwn 10:00:00:00:c9:76:ff:64 [spdc-esx03-vhba1]
    pwwn 10:00:00:00:c9:76:ff:65 [spdc-esx03-vhba2]
```

Quality of Service Design

This section contains the following topics:

Nexus1000v QoS Configuration

The Nexus1000v DVS (distributed virtual switch) supports the following QoS features:

- Traffic classification
- Traffic marking, and
- Traffic policing

By default, the Nexus1000v does not perform any QoS actions on packets passing through the VEMs. QoS actions are only implemented when QoS policy is applied to an interface. Both input and output QoS policy attachments are supported by the Nexus1000v. QoS policy can be attached to either physical or virtual interfaces.

The Nexus1000v supports the use of inheritance via the port-profile construct. Port-profiles are used to configure interfaces. A port-profile can be assigned to multiple interfaces giving them all the same configuration, i.e. the interfaces inherit the configuration of the port-profile. Via inheritance, changes to the port-profile are propagated automatically to the configuration of any interface assigned to it. Inherited port-profiles are automatically configured by the Nexus1000V when the ports are attached on the ESX hosts. This is done by matching up the VMware port-group assigned by the system administrator with the port-profile that created it.

Note In the VMware vCenter Server, a port-profile is represented as a port-group.

Multiple levels of inheritance are supported by port-profile, one port-profile can be configured to inherit the policies from another port-profile. The characteristics of the parent port-profile become the default settings for the child.

Note Not all characteristics of a port-profile can be inherited, refer to Nexus1000v documentation for details.

Port-profiles allow for easier QoS policies administration and maintenance; rather than applying QoS policy to interfaces retroactively after the interfaces are created/attached to ESX host, the used of port-profiles allow QoS policies to be automatically applied to the interfaces at the moment of creation/attachment.

Since the solution only make use of the traffic classification and marking features of the Nexus1000v, only the configuration of those two features would be explored. Please refer to Nexus1000v documentation for more details on traffic policing feature and configuration.

All traffic egressing from the Nexus1000v will be classified and marked according to the class of service accorded to the respective traffic. Since the core network is made up of Layer 2 switches, only the 802.1p CoS bits of the packets/frames would be marked, the IP DSCP/TOS bits would be left in tact. The upstream switches and routers based their QoS actions solely on the CoS bits of the packets/frames. It is assumed that ingress traffic to the Nexus1000v has been properly classified and QoS treatments applied by the upstream switches and routers, no ingress QoS actions are required on the Nexus1000v.

The Nexus1000v needs to classify and mark the following classes of traffic:

- Control traffic
- Management traffic
- User data traffic to/from the virtual machines

The Nexus1000V DVS must classify and mark control traffic, management traffic, and user data traffic to and from the virtual machines.

Table 2-6 Traffic Classification

Traffic class	Sub class	CoS marking
Control	Packet VLAN	CoS=6
	Control VLAN	CoS=6
Management	ESX Host Service Console	CoS=5
	VMotion	CoS=4
User Data	Gold service	CoS=2
	Silver service	CoS=1
	Bronze/Best-effort service	CoS=0

UCS Hardware QoS

The hardware and software components in the UCS support the unified fabric of Cisco, which allows multiple types of data center traffic over a single physical Ethernet network. This Data Center Bridging technology reduces the cabling, management, and cost with the combination of the host bus adapters (HBAs) and network interface cards (NICs) into a single adapter called the converged network adapter (CNA). This adapter can carry LAN and SAN traffic on the same cable.

Cisco UCS uses Fibre Channel over Ethernet (FCoE) protocol to carry Fibre Channel (FC) traffic inside Ethernet frame. Cisco UCS also adheres to multiple 802.1 standards to provide the DCE underlying the FCoE needs to transport those frames effectively. The fabric interconnect separates LAN and SAN traffic from the Ethernet frames and forwards them to the appropriate network ports. This gives the flexibility to deploy this technology without the need to implement the unified fabric solution across the entire data center network.

Cisco UCS blade installed with Cisco UCS CNA M71KR - E Emulex converged network adapter or Cisco UCS CNA M71KR - Q QLogic converged network adapter can handle FC and IP simultaneously. The converged network adapter presents an Ethernet interface and a Fibre Channel interface to the operating system. The operating system is completely unaware of the encapsulation taking place in the Ethernet segment. The operating system needs only the appropriate drivers to recognize the CNA hardware.

At the fabric interconnect, the server-facing Ethernet port receives Ethernet and Fibre Channel traffic. The fabric interconnect, which uses Ethertype to differentiate the frames, separates the two traffic types. Ethernet frames and Fibre Channel frames are switched to their respective uplink interfaces.

The Cisco UCS follows the FCoE protocol as defined by the ANSI T11 Standards Committee. The FC traffic encapsulated inside this Ethernet requires the same lossless network characteristics that are found in a fabric network. Instead of the buffer-to-buffer (B2B) credit system used in native fabric topologies, the FCoE relies on a new set of Ethernet standards that were developed to enhance the Ethernet protocol to ensure lossless transport of the FCoE traffic.

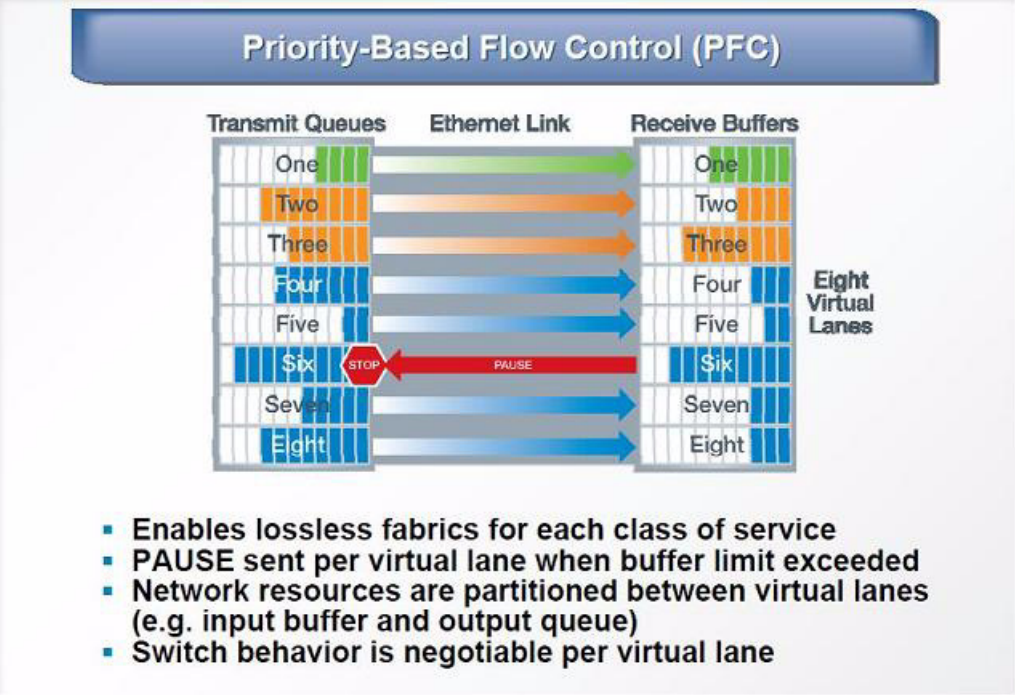
The Ethernet links on the system support the following Ethernet enhancements to ensure lossless transport for the FCoE traffic:

- Priority Flow Control (PFC) IEEE 802.1Qbb is an extension of the PAUSE (802.3x) mechanism. PFC creates eight virtual links in each physical link and allows any of these links to be paused individually without affecting the flow of traffic in the other links.
- Enhanced Transmission Selection (ETS) IEEE 802.1Qaz is a scheduling mechanism in hardware that allows a two-level Deficit Weighted Round Robin (DWRR) with strict priority support. This allows control not only of bandwidth, but also of latency.
- Data Center Bridge eXchange (DCBX) is a discovery and capability exchange protocol to verify that both ends are configured properly to support the DCE traffic. It can provide basic configuration if one of the two sides is not configured properly.
- Congestion notification is optional for DCE (Cisco UCS will support it in the future releases). Congestion Notification IEEE 802.1.Qau is a form of traffic management that shapes traffic as close to the edge as possible to reduce or eliminate congestion. Provides end-to-end congestion management for protocols that do not already have built-in congestion control mechanisms.

Priority Flow Control (PFC)

PFC creates eight separate virtual links on the physical link and allows any of these links to be paused and restarted independently. This approach enables the network to create a no-drop class of service for an individual virtual link that can coexist with other traffic types on the same interface. PFC allows differentiated quality-of-service (QoS) policies for the eight unique virtual links. PFC is crucial for I/O consolidation and also plays a primary role when used with an arbiter for intraswitch fabrics, linking ingress ports to egress port resources.

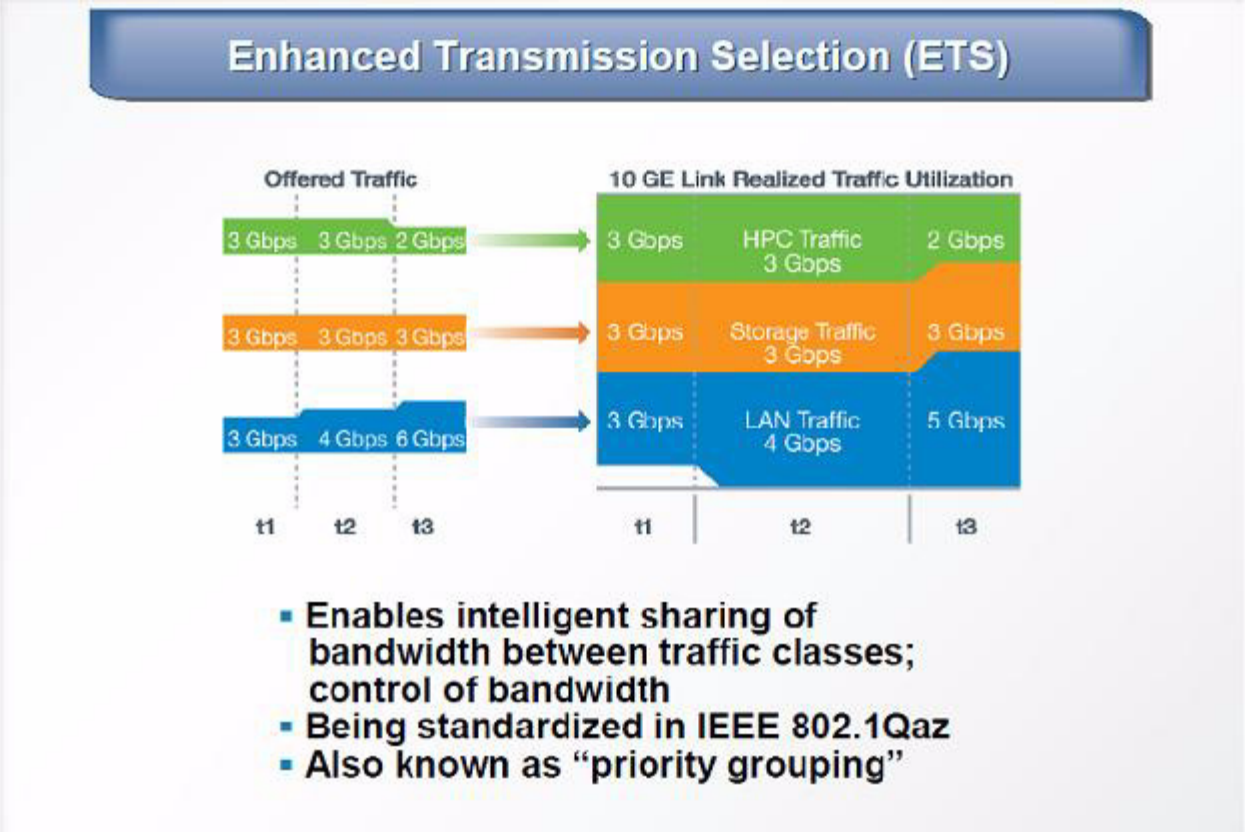
Figure 2-57 Priority Flow Control



Enhanced Transmission Selection (ETS)

ETS provides prioritized processing based on bandwidth allocation, low latency, or best effort, resulting in per-group traffic class allocation. ETS allows differentiation among traffic of the same priority class, thus creating priority groups. Today's IEEE 802.1p implementation specifies a strict scheduling of queues based on priority. With ETS, a flexible, drop-free scheduler for the queues can prioritize traffic according to the IEEE 802.1p traffic classes and the traffic treatment hierarchy designated within each priority group. The capability to apply differentiated treatment to different traffic within the same priority class is enabled by implementing ETS.

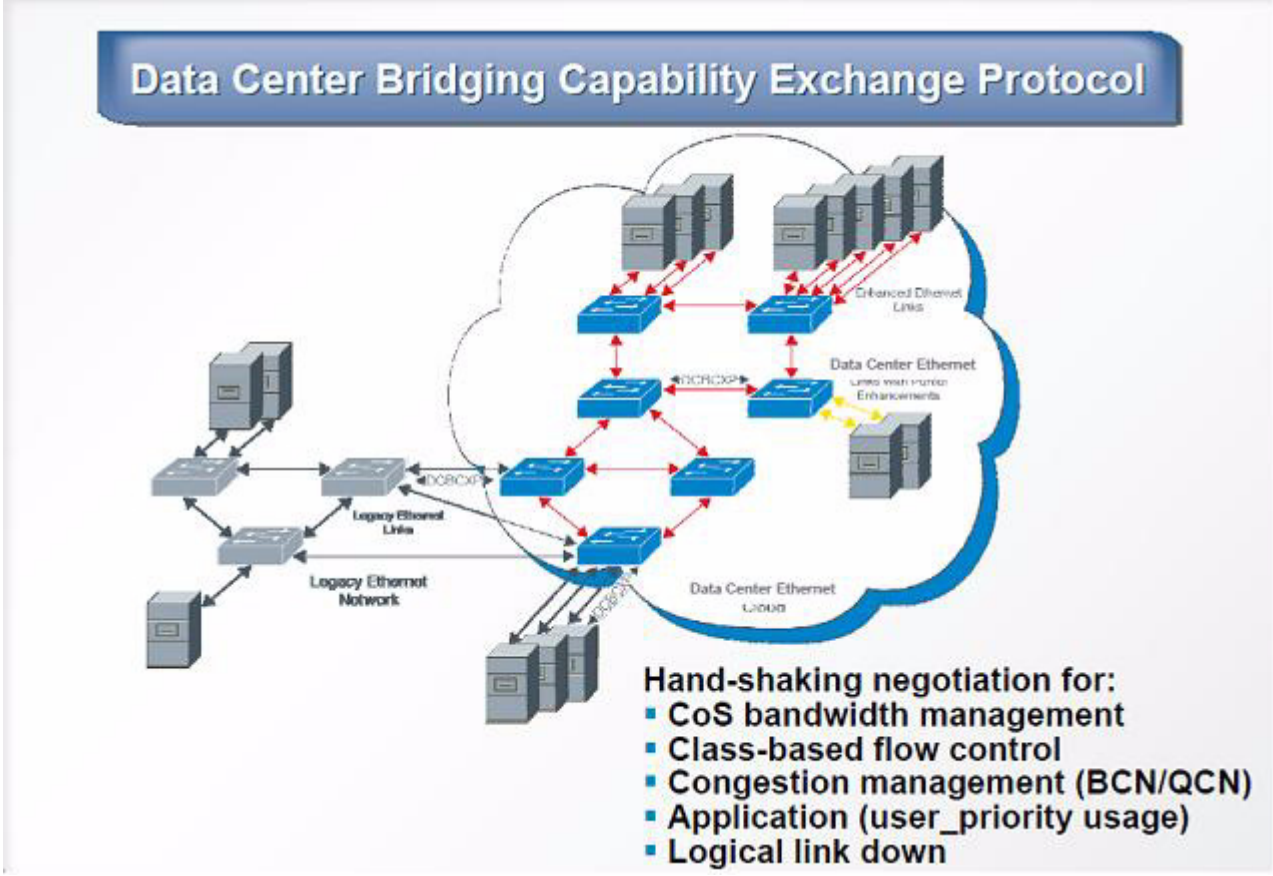
Figure 2-58 Enhanced Transmission Selection



Data Center Bridge eXchange (DCBX)

DCBX is a discovery and configuration protocol that guarantees that both ends of an Ethernet link are configured consistently. DCBX discovers the capabilities of the two peers at each end of a link: it can check for consistency, it can notify the device manager of configuration mismatches, and it can provide basic configuration where one of the two peers is not configured. DCBX can be configured to send conflict alarms to the appropriate management stations.

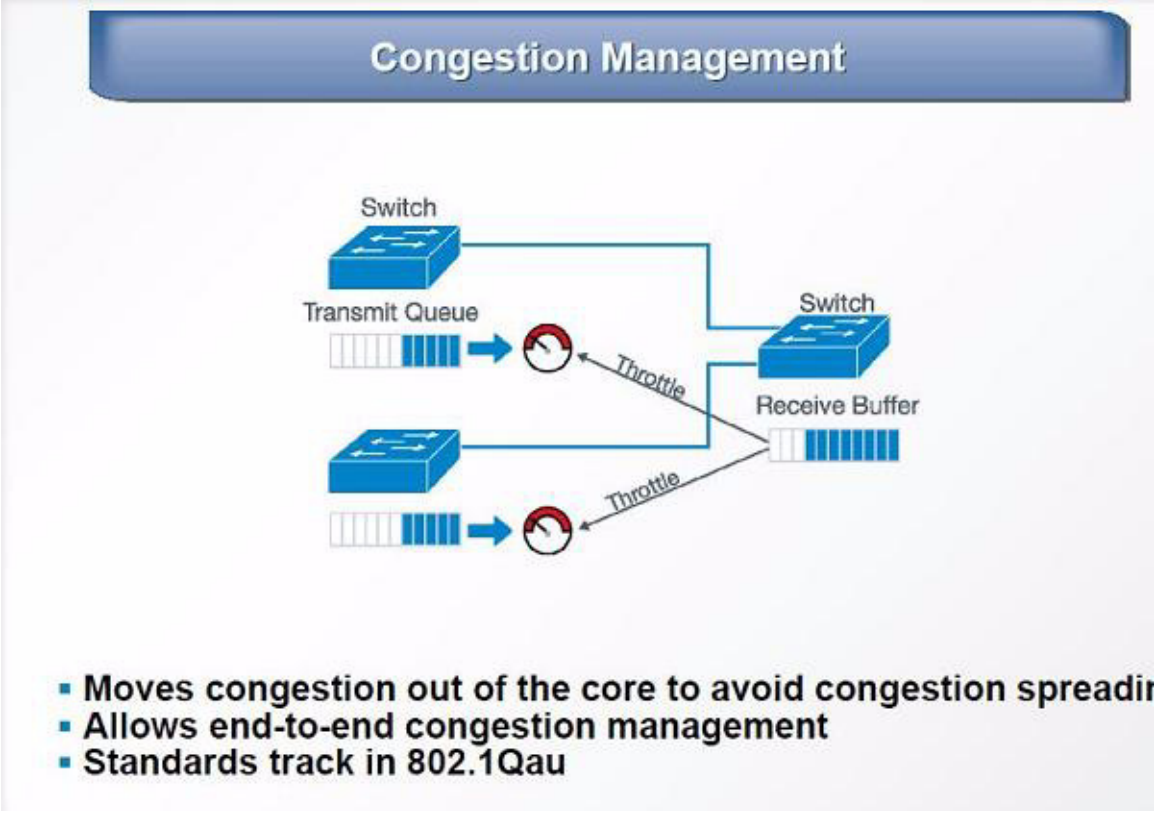
Figure 2-59 Data Center Bridge Exchange



Congestion Notification IEEE 802.1.Qau (Optional)

Congestion Notification is traffic management that pushes congestion to the edge of the network by instructing rate limiters to shape the traffic causing the congestion. Congestion is measured at the congestion point, and if congestion is encountered, rate limiting, or back pressure, is imposed at the reaction point to shape traffic and reduce the effects of the congestion on the rest of the network. In this architecture, an aggregation-level switch can send control frames to two access-level switches asking them to throttle back their traffic. This approach maintains the integrity of the network's core and affects only the parts of the network causing the congestion, close to the source.

Figure 2-60 Congestion Notification IEEE 802.1Qau (Future)

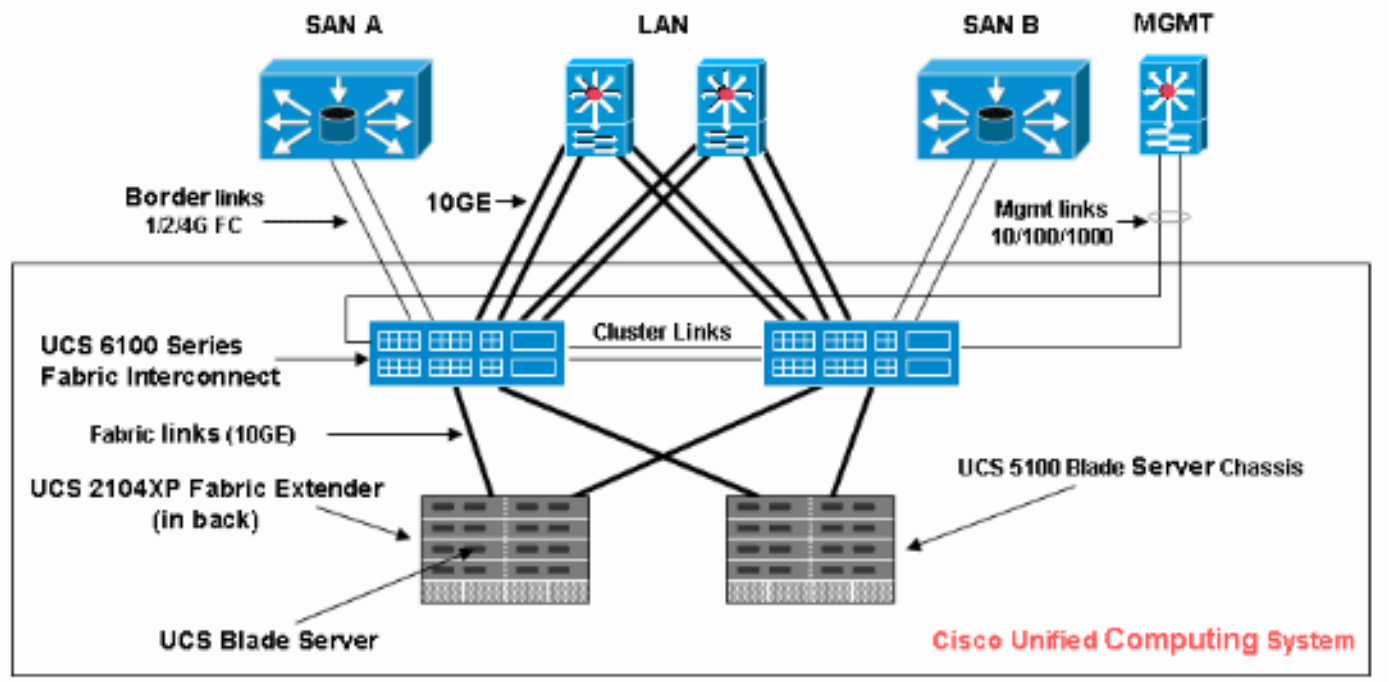


UCS QoS Mapping

Figure 2-61 shows a typical topology used with the Cisco UCS. The 6 port 8G FC expansion module for UCS 6100 series is now orderable either configured (N10-E0060) or as spare (N10-E0060=).

The ports will do 1/2/4G FC with SFP transceivers or 2/4/8G FC with SFP+ transceivers.

Figure 2-61 Cisco UCS Diagram



- For Network to Host direction traffic, Fabric Extender (FEX) use PFC to push the congestion to Fabric Interconnect where there is more buffer.
- For Host to Network direction Fabric Extender use Classic pause (802.3x) or PFC toward host to avoid dropping which is required if there is an oversubscription on uplinks.
- All the drop class will be mapped into one queue, and all the configured no-drop class will be mapped to the other queue.
- Deficit Weighted Round Robin (DWRR) between the 2 queues or one can be configured as strict priority.

UCS Control Traffic Protection

The following topics discuss UCS control traffic protection:

- Fabric Interconnect to FEX Control Traffic Protection, page 2-89
- FEX to Fabric Interconnect Control Traffic Protection, page 2-90
- Control Traffic to and from Adapter, page 2-90

Fabric Interconnect to FEX Control Traffic Protection

In Fabric Interconnect to FEX direction, Fabric Interconnect has a separate queue for data traffic. Control traffic from Switch to FEX share the same queue as data traffic on FEX. This does not cause any issue since

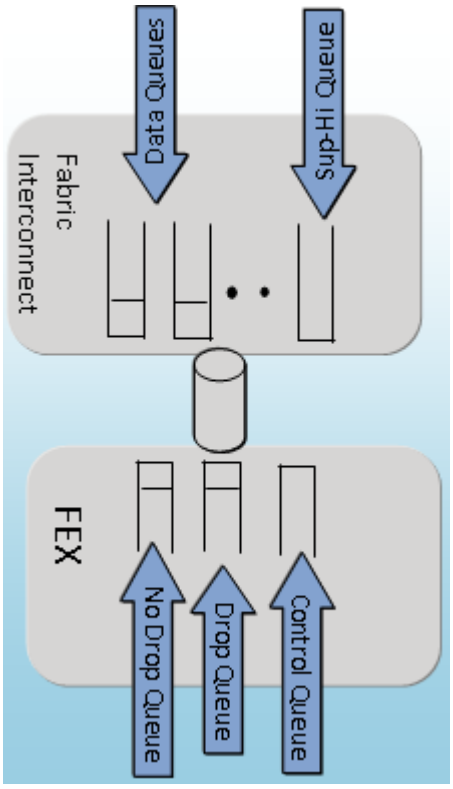
- On Fabric Interconnect, the control traffic is put in Sup-Hi queue and are scheduled as strict priority compared to data traffic.
- The link between Fabric Interconnect and FEX is a no-drop link due to PFC therefore control traffic from Fabric Interconnect to FEX are guaranteed to be delivered.

- As long as Switch to FEX link is not permanently paused, control traffic is guaranteed to get to FEX.
- Control traffic stays in Sup-Hi class on the Fabric Interconnect.

FEX to Fabric Interconnect Control Traffic Protection

In FEX to Fabric Interconnect direction, FEX has separate queue for all control traffic. Traffic from FEX to Fabric Interconnect cannot share the same queue as data traffic since when the congestion occur there is no guarantee that the control traffic will not be dropped. A separate queue is used for control traffic on FEX, this allows control traffic not being affected by data traffic congestion.

Figure 2-62 UCS Traffic Queues



Control Traffic to and from Adapter

Link pause on FEX is Rx-OFF to avoid head of line blocking caused by link PAUSE sent from adapter. Adapter sends and receives control traffic on strict priority queue. Fabric Interconnect and adapter send control traffic with CoS 7. CoS 7 is reserved for control traffic.

QoS Features of the UCS 6120/6140XP Fiber Interconnects

The following QoS features are available:

- Layer 2 IEEE 802.1p (CoS)
- Eight hardware queues per port

- Per-port QoS configuration - QoS policies can assign system classes for individual vNICs or vHBAs.
- CoS trust - When enabled, the associated QoS class is configured on the fabric interconnect and can be assigned to a QoS policy.
- Per-port virtual output queuing - Each port supports up to eight virtual queues, six for system classes and two reserved for control traffic that are not accessible for user configuration.
- CoS-based egress queuing - The CoS values are used to classify a specific type of traffic into different priority queues.
- Egress strict-priority queuing
- Egress port-based scheduling: Weighted Round-Robin (WRR)

Cisco Unified Compute System provides a high-speed, low-latency connectivity and a reliable, robust foundation for unifying LAN and SAN traffic on a single transport. UCS priority flow control (PFC) simplifies management of multiple traffic flows over a single network link and supports different classes of service, enabling lossless and classic Ethernet on the same fabric. Its system wide bandwidth management facilitates consistent and coherent quality-of-service (QoS) management throughout the system. For detailed UCS QoS design in this solution, refer to Unified Compute System QoS.

UCS Software QoS

The following sections describe how software enablement of QoS features can be structured to support service tiers:

- System Classes, page 2-91
- QoS Implementation, page 2-92

System Classes

Cisco UCS uses Data Center Bridging to handle all traffic inside a Cisco UCS system. This industry standard enhancement to Ethernet divides the bandwidth of the Ethernet pipe into eight virtual lanes. Six virtual lanes are user visible as the system classes, and two are reserved for internal traffic. System classes determine how the DCE bandwidth in these virtual lanes is allocated across the entire Cisco UCS system.

Each system class reserves a specific segment of the bandwidth for a specific type of traffic. This provides a level of traffic management, even in an oversubscribed system. For example, you can configure the Fibre Channel Priority system class to determine the percentage of DCE bandwidth allocated to FCoE traffic.

Table 2-7 describes the system classes:

Table 2-7 UCS System Classes

System Class	Description
Platinum Priority Gold Priority Silver Priority Bronze Priority	A configurable set of system classes that you can include in the QoS policy for a service profile. Each system class manages one lane of traffic. All properties of these system classes are available for you to assign custom settings and policies.
Best Effort Priority	A system class that sets the quality of service for the lane reserved for Basic Ethernet traffic. Some properties of this system class are preset and cannot be modified. For example, this class has a drop policy that allows it to drop data packets if required.
Fibre Channel Priority	A system class that sets the quality of service for the lane reserved for Fibre Channel over Ethernet traffic. Some properties of this system class are preset and cannot be modified. For example, this class has a no-drop policy that ensures it never drops data packets.

QoS Implementation

In the current design, all available system classes are used to classify network traffic and allocate corresponding bandwidth. FCoE traffic uses the default value of cos 3 and classified into the Fibre Channel Priority. VMotion traffic is classified into the Gold Priority, and it matches on cos 4. Control traffic (cos 6) get classified into the Platinum Priority. Customer data traffic is classified into Gold class (cos 2) in the Silver Priority, Silver class (cos 1) in the Bronze Priority, and default class (cos 0) in the Best Effort Priority. Six system classes map to different types of traffic with specified bandwidth in the Table 2-8.

Table 2-8 UCS System Classes Traffic Classification and Bandwidth Allocation

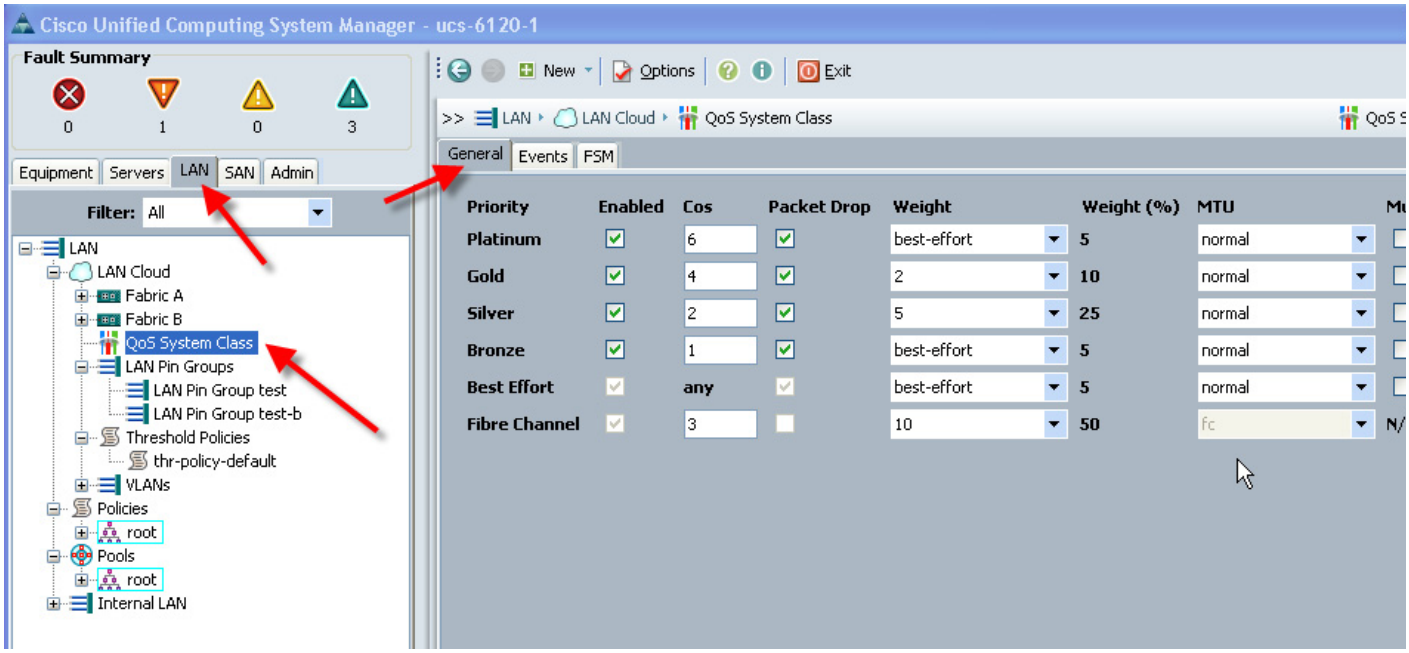
System Classes	Traffic Description	CoS	Assured Bandwidth
Platinum Priority	Control	6	5%
Gold Priority	VMotion	4	10%
Silver Priority	Gold	2	25%
Bronze Priority	Silver	1	5%
Best Effort Priority	Bronze	any	5%
Fibre Channel Priority	FCoE	3	50%

The solution assures 50% bandwidth for FCoE (cos 3) traffic and 10% bandwidth for VMotion (cos 4) traffic, then control traffic (cos 6) is allocated for 5%, Gold (cos 2)traffic is allocated for 25%, Silver (cos 1) traffic is allocated for 5%, and Bronze (cos 0) traffic is allocated for 5% bandwidth. Refer to section [UCS QoS Configuration] for the detailed configurations.

This System Class defines overall bandwidth allocation for each traffic class on the system for server downlink, Ethernet and Fibre Channel uplinks. Since there is no FCoE (Cos3) traffic presented on UCS ethernet uplink, the remaining five classes should be assured up to double the amount of bandwidth on the uplink.

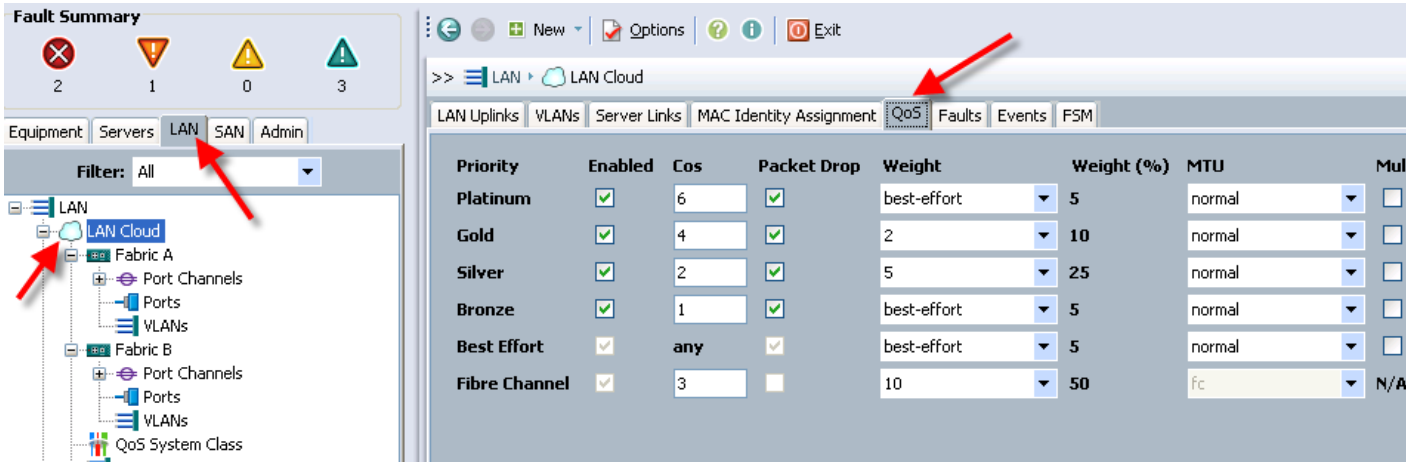
To configure the UCS QoS System Class, navigate to LAN > LAN Cloud > QoS System Class, click the General tab on the right side of the panel, and update the properties for the system class that you want to meet the traffic management needs of the system.

Figure 2-63 UCS QoS System Class Configuration



System Class also can be viewed and configured from LAN > LAN Cloud, click QoS tab on the right side of the panel.

Figure 2-64 UCS QoS System Class Configuration



The QoS System Class configuration allows users to enter the weight (values range from 1 to 10) on a per class basis. The bandwidth percentage is then calculated based on the individual class weight divided by the sum of all of the weights. Currently there is no ability to enter an exact percentage in a field under the weight (%) column.

The solution assigns weight value 10 to Fibre Channel Priority to assure 50% bandwidth for FCoE (cos 3) traffic and weight value 2 (10% bandwidth) for VMotion (cos 4) traffic in the Gold Priority. Control traffic (cos 6) in Platinum Priority has weight value 1 (best-effort) for 5% bandwidth. Gold (cos 2) traffic in Silver Priority has weight value 5 to allocate 25% bandwidth, Silver (cos 1) traffic in Bronze Priority has weight value 1 to allocate 5% bandwidth, and Bronze (cos 0) traffic in Best Effort Priority has weight value 1 for 5% bandwidth.

Nexus 7000 Series Switch QoS

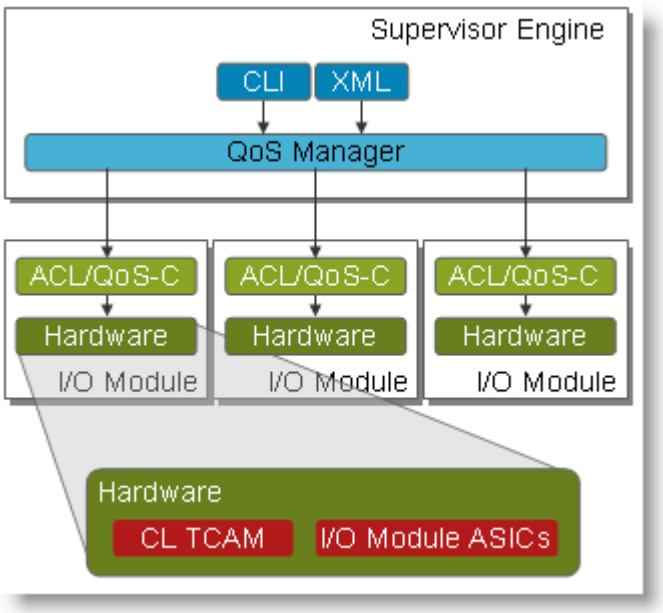
The Nexus 7000 Series switch offers a wide variety of QoS features; however, this design uses a small subset of QoS features.

Nexus 7000 Series switch uses modular QoS CLI (MQC) to apply queuing and traditional QOS policies. The default MQC object type is `qos`. Class maps are used to match traffic. Policy maps define the actions to take on these classes. Service policy ties policy maps to an interface in input and output directions. QoS policies include marking and policing features, while queuing policies include queuing and scheduling features and a limited set of marking features.

For more information about Nexus 7000 QoS features, refer to Nexus 7000 QoS Configuration Guide.

Nexus 7000 Series Switch QoS Detailed Design

Figure 2-65 QoS Architecture Diagram



The QoS manager running on the supervisor engine receives the service policy configuration in the command line or XML interface. QoS manager in turn distributes these policies to ACL/QoS client process running on the line cards. ACL/QoS clients perform ACL merge and program the classification TCAM or queuing ASICs on the line card depending on the type of policy.

Nexus 7000 Series Switch Classification and Marking

Classification is used to partition traffic into classes based on port characteristics such as CoS field or packet header fields such as IP precedence, and DSCP. Classification is done using a match criterion. Packets that fail to match any class map are assigned to a default class called class-default.

Marking is used to identify the traffic type for use in another QoS operation such as policing or queuing.

Note The Nexus 7000 Series switch in this design is a Layer 2 switch with no specific requirement to perform QoS classification at ingress. By default, all ports on the Nexus 7000 Series switch are trusted ports. Hence, CoS values are preserved during transit.

Nexus 7000 Series Switch Queuing, Scheduling, and Congestion Management

The queuing and scheduling process enables you to control the bandwidth allocated to traffic classes, so you achieve the desired trade-off between throughput and latency. Congestion management algorithm provides proactive queue management by dropping traffic based on the cos field. Weighted Random Early Detection (WRED) is used for congestion management at the ingress and egress modules.

Note By default, the Nexus 7000 Series switch enables a system default queuing policy on each port and port channel. When you configure and apply a new queuing policy to an interface, the new policy replaces the default queuing policy.

To change the default queuing behavior, configure the following:

- Configure class maps that define cos-to-queue mapping
- Configure queuing policy maps and define how each class is treated
- Apply queuing service policy to the physical or port-channel interface

Default queue mapping at ingress and egress port is cos-to-queue. It cannot be changed.

Table 2-9 shows the default cos-to-queue mapping on a 32-port 10G module.

Table 2-9 Cos-to-Queue Mapping on a 32-Port 10G Module

Class Map Queue Name	Description	Default CoS Values
8q2t-in-q1	Ingress queue 1 of 8q2t type	5-7
8q2t-in-q2	Ingress queue 2 of 8q2t type	-
8q2t-in-q3	Ingress queue 3 of 8q2t type	-
8q2t-in-q4	Ingress queue 4 of 8q2t type	-
8q2t-in-q5	Ingress queue 5 of 8q2t type	-
8q2t-in-q6	Ingress queue 6 of 8q2t type	-
8q2t-in-q7	Ingress queue 7 of 8q2t type	-
8q2t-in-q-default	Ingress default queue of 8q2t type	0-4
1p7q4t-out-pq1	Egress priority queue of 1p7q4t type	5-7
1p7q4t-out-q2	Egress queue 2 of 1p7q4t type	-
1p7q4t-out-q3	Egress queue 3 of 1p7q4t type	-
1p7q4t-out-q4	Egress queue 4 of 1p7q4t type	-

Table 2-9 Cos-to-Queue Mapping on a 32-Port 10G Module

Class Map Queue Name	Description	Default CoS Values
1p7q4t-out-q5	Egress queue 5 of 1p7q4t type	-
1p7q4t-out-q6	Egress queue 6 of 1p7q4t type	-
1p7q4t-out-q7	Egress queue 7 of 1p7q4t type	-
1p7q4t-out-q-default	Egress default queue of 1p7q4t type	0-4

Ingress queuing policy on a 10G module (N7K-M132XP-12) is applied only at the first-stage buffer (i.e., per-port buffer at 4:1 Mux chip). Second stage buffer behavior is fixed. In the current design, 10G modules are used in dedicated mode; hence no specific queuing policy is used in the ingress direction.

Egress queuing policy on a 10G module (N7K-M132XP-12) is applied at the port ASIC level.

Note When egress queuing policy is applied on a 10G port on N7K-M132XP-12 module, the policy is propagated to the remaining ports in the port group even if only the first port of the port-group is used. Only the first port in the port-group is used in dedicated port mode.

Two types of egress policies are used in the current design. These policies differ by the cos-to-queue mapping and associated bandwidth allocation.

Nexus 7000 Series Switch Downlink Queuing Policy

This policy applies to the peer link and vPC links to the aggregation layer. VMotion traffic marked with cos4 is carried over these links, hence a separate queue is used to map this traffic and provide assured bandwidth of 1Gig. For cos-to-queue mapping and assured bandwidth configuration, refer to Table 2-10. Any cos value that is not explicitly mapped is assigned to the default output queue.

Table 2-10 Cos to Queue Mapping and Assured Bandwidth Configuration

Class Map Queue Name	Traffic description	Assured Bandwidth	CoS Values
1p7q4t-out-q2	Silver	8%	1
1p7q4t-out-q3	Gold	80%	2
1p7q4t-out-q4	Control	1%	6-7
1p7q4t-out-q5	VMotion	10%	4
1p7q4t-out-q-default	Bronze	1%	0

Nexus 7000 Series Switch Uplink Queuing Policy

The Nexus 7000 Series switch uplink queuing policy applies to the port-channel interface that connects to the service layer and trunk ports that connect to the backbone routers. These ports do not carry VMotion (Cos 4) traffic, hence it falls into the default queue with Cos 0, Cos 3 and Cos 5.

Table 2-11 Uplink Queuing Policy

Class Map Queue Name	Traffic description	Assured Bandwidth	CoS Values
1p7q4t-out-q2	Silver	10%	1
1p7q4t-out-q3	Gold	88%	2

Table 2-11 Uplink Queuing Policy

Class Map Queue Name	Traffic description	Assured Bandwidth	CoS Values
1p7q4t-out-q4	Control	1%	6-7
1p7q4t-out-q-default	Bronze	1%	0

VSS QoS

Typically, all packets are treated with a best effort approach where equal treatments are given to packet servicing irrespective of packet priority or importance. With a best effort approach, a packet is serviced based on arrival times. With such an approach, during congestion times, high priority packets are lost, resulting in processes using such packets not working properly. If such processes do not work correctly, network instability can result.

QoS ensures that in times of congestion, packets are serviced based on priority or importance. To prevent starving packets of services times, QoS ensures that a limit is set for servicing high priority packets. With QoS, high priority processes can work in the presence of network congestion for continuous network stability.

On a 6500 Series switch, QoS features are distinguished as follows:

- **Ingress QoS features:** Features that are provided to packets inbound to the 6500 Series switch, including port trust, which ensures that inbound packets maintain QoS labels that they carry; CoS remarking based on default or configured table maps; and CoS-based congestion avoidance based on configured queuing techniques.
- **PFC and DFC QoS features:** Features that the supervisor line card applies to packets during processing. MQC, which involves packet classification based on QoS labels and other Layer 2 and 3 parameters and eventually application of policies, is the most important of these features.
- **Egress QoS features:** CoS remarking based on DSCP marking, and CoS and DSCP based congestion avoidance features that are provided to packets outbound from the Catalyst 6500 Series switch.

For a complete understanding of QoS on the Catalyst 6500 platforms, see http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/qos.html.

On a VSS system, QoS becomes important, since in its absence packets that are necessary for a VSS operation might be lost. The loss of such packets can cause the VSS to introduce instability into the network.

At congestion periods, the VSL becomes important. Typically, the VSL could comprise two to eight 10G ports. If VSLP packets are lost, the VSS enters dual active situation. To ensure all VSS control packets are given priority over other packets, a bit in the VSLP packet is set. During congestion periods, packets with that bit set are given higher priority in the priority queues.

By default, all ports used to build the VSL have the following characteristics:

- Put into trust cos mode
- QoS configuration cannot be added or modified

To optimize the servicing of packets of different priorities during congestion times, queues are used. These optimization techniques are mostly referred to as congestion avoidance techniques. On a Catalyst 6500 Series switch, these techniques can be implemented both in the inbound and outbound direction on port ASICS using built-in queues. The numbers of queues available depend on the line card used. The 10G ports on the VS-720-10G C/CXL, WS-X6708-10G-3C/XL, and WS-X6716-10G-3C/XL are all

capable of supporting up to eight queues in the inbound and outbound direction. Due to the limitation of QoS configuration modification and addition on the VSL ports, the defined cos mapping to queues should be used to determine the QoS labels with which packets should be associated.

To make use of all available queues on the Catalyst 6500 Series switch, `mls qos 10g-only` should be configured. This command requires that all 1G interfaces are in shutdown mode as shown in the following code sample:

```
vss-1(config)#mls qos 10g-only
Error: following ports have to be shut to enable 10g-only mode:
Gi2/6/1 Gi2/6/2 Gi2/6/3 Gi1/6/1 Gi1/6/2 Gi1/6/3

 Command Rejected!
```

With that command, the queuing structure in the inbound direction should be 8q4t for all line cards capable of supporting virtual switch operation with the exception of WS-X6716-10G-3C/XL, which can be 1p7q4t if configured with over-subscription mode.

The following Web site provides a complete understanding of queuing structures capable with Catalyst 6500 Series switch line cards:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/qos.html#wp1665946

Redundancy

Redundancy is central tenant of any Cisco data center solution. In this solution, full redundancy is supported across virtualized servers, compute, storage, and network components. While this document has indicated redundancy design characteristics throughout, this section highlights some of the redundancy characteristics enabled by the solution:

Link Redundancy

The following features provide link redundancy in this solution:

- UCS Link Failures
- FC Single Link Failure—The 6120XP is always in NPV mode. This means the FC uplinks are always NP ports. PortChannel configuration is not available at FCS. When the failed NP_port comes back up, the logins are not re-distributed to avoid disruption. There is an option with the Nexus 5000 to enable re-distribution of logins if a previously failed NP_Port comes back online. This feature is not available on the 6120XP.

- Port-channel Uplink Single Link Failure
- Nexus 7000 Series Switch Link Failures—Nexus 7000 Series switches include virtual port-channel peerlink member failures, vPC peerlink total failures, vPC peer-keepalive management switch failures, and vPC peer link and peer-keepalive link double failures.

Note Peer link complete failure has an adverse effect in an environment mixed with non-vPC and vPC ports in the same VDC. Use two or more 10G modules to form distributed port-channels that make the peer-link.

Virtual Switch Link

The virtual switch link plays an important role in initializing and maintaining the VSS. All control packets needed to build the VSS system are sent through the VSL. The Virtual Switch Link Protocol (VSLP) exchanges control messages across the VSL. The VSLP is comprised of the Link Management Protocol (LMP) and the Role Resolution Protocol (RRP).

Data packets are exchanged across a VSL if the following occurs:

- In a Layer 2 design, ports that are members of a VLAN are in the other peer chassis, and a broadcast/multicast in that VLAN is needed.
- In a Layer 2 design, the destination Layer 2 address associated with a frame is reachable on the peer chassis.
- In a Layer 3 design, the destination Layer 3 address is reachable on the peer chassis.
- In a Layer 3 design, broadcast/multicast packets are received of interest to a Layer 3 address on the peer chassis.
- In a service layer design, inter-module communication is needed. This should occur if modules on both chassis are configured with an active/active design. In such a design, context A, for an example, is active on chassis A and standby on chassis B, with context B being standby on chassis A and active on chassis B.
- If in a services layer design, packets are received on a chassis and the service module of interest is located on a different chassis.

Note The PFC on both supervisors match if the VSS works. Changing the PFC mode from a higher to lower mode with the use of platform hardware `vsl pfc mode pfc3c` is only applicable on the line cards.

The VSL ports could be positioned as follows:

- **All VSL ports on the VS-720-10G-C/CXL:** The VSL comes up faster since the supervisor is the first module to come online during bootup. An assumed advantage might be the potential fault location is limited to one module, resulting in the loss of VSL connection if the module goes offline. The likelihood of that happening is too minimal, and if that occurs it is due to a faulty supervisor. If there is a faulty supervisor, irrespective of the VSL ports, the VSS does not come online.
- **All VSL ports on a line card:** WS-X6708-10G-3C/XL and WS-X6716-10G-3C/XL are 10G line cards that support virtual switch mode. Depending on the location of the line cards, the VSL comes online only after the line cards complete the boot process. With this process, the time taken to bring up the VSL is longer than with all ports on the supervisor.

- **VSL ports on the supervisor and on a line card:** With this approach, some VSL ports can be placed on the supervisor, with the rest placed on the line cards. Fault occurrence is not limited to the line cards. If a line card fails, the VSL ports on the supervisor are active with the VSS operating properly

Cisco 7600 Link Failure

The architecture and design of the data center cloud network is compartmentalized and fully redundant. For more information about L2VPN pseudowire redundancy feature, refer to the following:

http://www.cisco.com/en/US/docs/ios/wan/configuration/guide/wan_l2vpn_pw_red_external_docbase_0900e4b1805e9066_4container_external_docbase_0900e4b1806a218a.html

For all customers (Gold, Silver and Bronze service classes) that has L2VPN attachment circuits on the failed link, bi-directional data traffic is disrupted for less than 1 second, while the failed pseudowire is cleaned up and the backup pseudowire takes over.

Chassis Redundancy

Chassis redundancy aims to address failures of the UCS 6100, the UCS chassis, the Nexus 7000 Series switch, or the Cisco 7600 Series router.

- UCS 6100 Failure—The Cisco Unified Computing System supports the use of two interconnected switches (6120XP) for improved management redundancy and increased switch throughput. Each of the two Cisco UCS Manager instances run as either the active or standby instance.
- UCS Chassis Redundancy—Blade Server Failure; The UCS blade servers run VMware ESX, which are configured into VMware high availability) clusters.
- Power Supply and Fan Failure

Nexus 7000 Series Switch Chassis Failure

The Nexus 7000 Series switch in this design is a Layer 2 only switch connecting Layer 2 aggregation layer, service layer, and backbone routers. vPC connects the Nexus 7000 Series switch with the service layer and Layer 2 aggregation layer. Each Nexus 7000 Series switch has an uplink to the backbone 7600 routers configured with switchport trunk mode. The 7600 series routers are single attached to the Nexus 7000 Series switch due to a QoS limitation on ES+ card when configured as port-channel.

If one of the vPC peers fail in this design, the secondary vPC peer becomes operationally primary. All traffic will failover to the remaining vPC peer device.

Convergence time may vary depending on many factors, including the following:

- Routing protocol timers used.
- Mixed vPC and non-VPC links in the same virtual device context (VDC).
- Time taken for MST to converge during STP topology changes during primary STP root failure.

Cisco 7600 Chassis Failure

The architecture and design of the Data Center Cloud Network is compartmentalized and fully redundant.

Nexus 1000V Series Switch VSM Redundancy

The Cisco NX-OS operating system is designed for high availability at the network, system, and service levels.

The Cisco Nexus 1000V DVS is made up of the following:

- VEMs running within virtualization servers represented as modules within the VSM.
- A remote management component, such as VMware vCenter Server.
- One or two VSMs running within virtual machines (VMs).

Nexus 7000 Series Switch SUP Redundancy

For information on HA features on Nexus7000 and Cisco NX-OS, refer to the following:

http://cco/en/US/docs/switches/datacenter/sw/4_2/nx-os/high_availability/configuration/guide/ha_overview.html

VSS Redundancy

A dual supervisor on each member of the VSS is not supported. However, redundancy of the Layer 3 services is provided via ACE and FWSM redundancy.

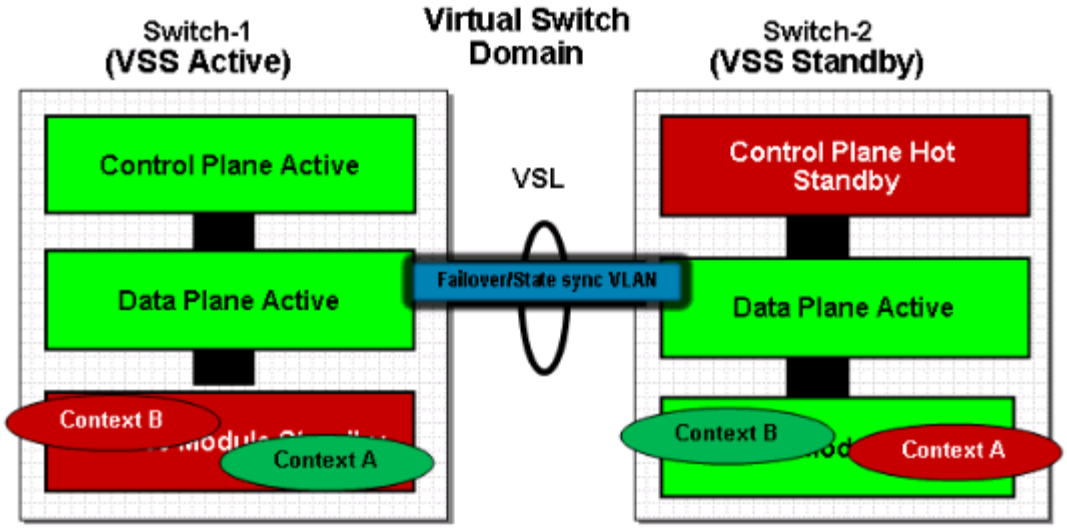
Layer 3 ACE and FWSM Redundancy

In this solution, the ACE and FWSM are configured for active/active redundancy. Using an active/active approach allows for the distribution of active contexts across all service modules on a VSS. This provides for a better distribution of traffic, increased throughput, and effective utilization of module resources. During normal operation, all service modules are used. During a failure, failover of contexts in the failed module is done. Based on the configuration in place, the failover should have little impact to traffic.

Note If stateful traffic is being replicated, it is recommended that 1Gbps of bandwidth be set aside for the ACE and 1-2Gbps for the FWSM.

Figure 2-66 is an example of an active/active context distribution. This is a typical example of even distribution of contexts across all modules on the service module. Figure 2-66 depicts each blade's active context in green. If a module fails, failover of the contexts occurs.

Figure 2-66 Service Module Active/Active Redundancy



VM Infrastructure Redundancy

The following features support redundancy in the VM infrastructure:

- ESX NIC Failure—Upon ESX NIC failure, the second NIC assumes responsibility for all forwarding of packets.
- ESX Failure—Each ESX host supplies its CPU and memory resources for the virtual machines that reside on it. In the event of a physical ESX host failure, these resources are no longer available, and the corresponding virtual machines being supported are powered off. Using VMware high availability, detection of failures occur rapidly, and recovery for downed virtual machines can occur on other hosts residing in the cluster. Core functionality includes host monitoring and virtual machine monitoring to minimize downtime when heartbeats are lost.
- SAN Failure—In a well-designed SAN, you will probably want the host application to access devices over more than one path for better performance and to facilitate recovery of adapter, cable, switch, or GBIC failure.
- vCenter Failure—vCenter clustering is not used in the test design so refer to VMware guidelines.

SAN Redundancy

SAN is created for redundancy with minimal traffic impact. SAN uses MDS 9513 HA capabilities in conjunction with the dual edge-core fabric design.

- SAN FC Link Failures—Multiple links between edge switches and core switches in a SAN allow for failures to happen within a SAN, without forcing a fabric failover. When putting those links in a port-channel together, traffic is load-balanced across each link involved, ensuring efficiency and smoother convergences should link failures occur.

- SAN Fabric Failure—Implementing the dual fabric setup, hosts and storage are connected to primary and secondary fabrics and then zoned in the same fashion on both fabrics. Load balancing is done by the ESX OS sending equal amounts of traffic down both fabrics to maximize bandwidth usage. Traffic is started and verified from the MDS on the primary and secondary fabrics.
- MDS SUP Failure—Using the supervisor high availability, the MDS 9513 can withstand supervisor switchovers and failures without disruption to traffic on either fabric.
- MDS LC Failure—Similar to the supervisor high availability, MDS line cards do not impair traffic flow for ports involved with port groups spread across multiple line cards during a line card failure.
- Storage Controller Failure—Design a HA storage array with fully redundant hardware components and advanced failover features that provide uninterrupted data access.