



CHAPTER 2

Design Considerations

Revised: October 20, 2010

The Cisco VMDC solution addresses the following design considerations in depth:

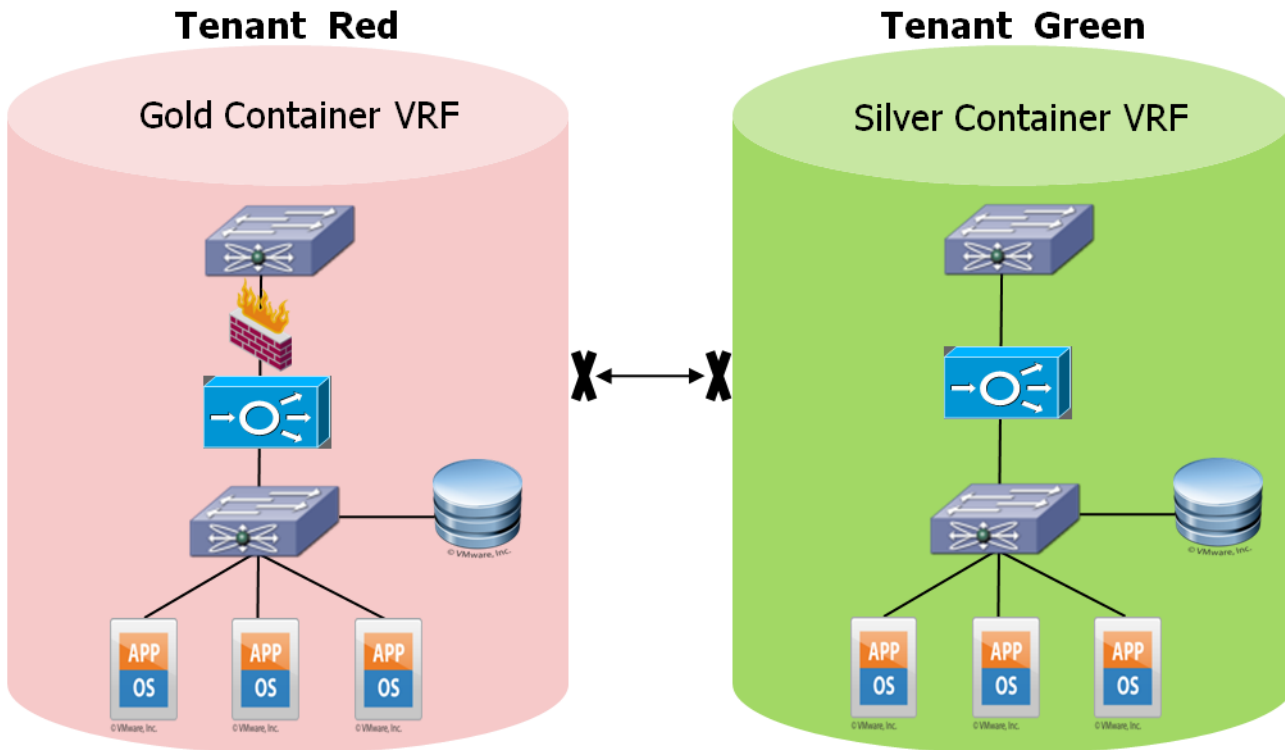
- [Tenant Separation, page 2-1](#)
- [High Availability, page 2-12](#)
- [Performance and Scalability, page 2-24](#)
- [Service Assurance, page 2-41](#)

Tenant Separation

Traditionally, providers deployed a dedicated infrastructure for each tenant that it hosted. This approach, while viable for a multi-tenant deployment model, does not scale well because of cost, complexity to manage, and inefficient use of resources. Deploying multiple tenants in a common infrastructure yields more efficient resource use and lower costs. However, each tenant requires isolation for security and privacy from others sharing the common infrastructure. Therefore, logical separation is a fundamental building block in multi-tenant environments. Virtualization can provide the separation in the network, compute, and storage resources.

[Figure 2-1](#) represents how each tenant can be logically separated within the Cisco VMDC design.

Figure 2-1 Tenant Separation



Network Separation

Each network container requires path isolation and/or logical resource separation at each of the network layers in the architecture.

- **Path Isolation**—The virtualization of the interconnection between devices. This interconnection can be a single or multi-hop. For example, an Ethernet link between two switches provides a single-hop interconnection that can be virtualized using 802.1q VLAN tags.
- **Device virtualization**—The virtualization of the network device, which includes all processes, databases, tables, and interfaces within the device. For example the ACE or FWSM can be virtualized using contexts.

Path Isolation

Path isolation defines independent, logical traffic paths over a shared physical network infrastructure. To define these paths, create VPNs using VRFs and map among various VPN technologies, Layer 2 segments, and transport circuits to provide end-to-end, isolated connectivity between various groups of users. A hierarchical IP network combines Layer 3 (routed), Services (firewall and server load balancing) and Layer 2 (switched) domains. Therefore, the three types of domains must be virtualized, and the virtual domains must be mapped to each other to keep traffic segmented. This mapping combines device virtualization with data path virtualization.

- **Aggregation Layer**—Layer 3 separation (VRF)
- **Services Layer**—Layer 2 separation (VLAN) and Virtual Device Contexts

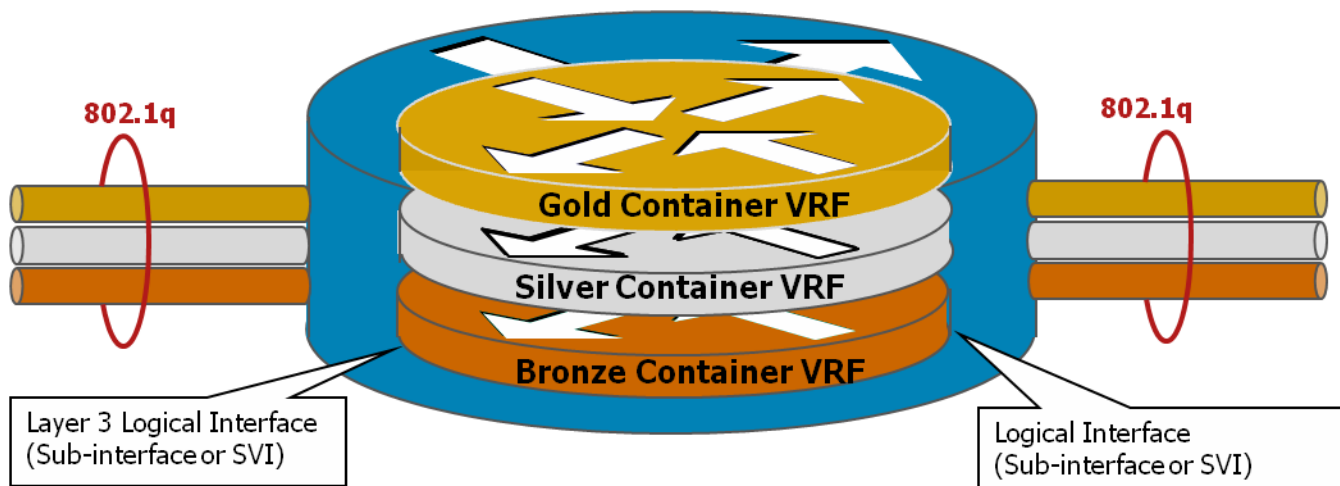
- Sub-Aggregation Layer—Layer 3 separation (VRF) and Layer 2 separation (VLAN)
- Access Layer—Layer 2 separation (VLAN)
- End-to-end virtualization of the network requires separation at each network layer in the architecture:

The virtualized network consists of Layer 2 VLANs and Layer 3 VRFs to provide logical, end-to-end isolation across the network. The number of VRFs matches the number of supported tenants. Each tenant is defined with a unique VRF. VRF information is carried across each hop in a Layer 3 domain, and multiple VLANs in the Layer 2 domain are mapped to the corresponding VRF. Because the Cisco VMDC solution is cloud architecture, this design assumes there is no need to connect the tenant VRFs because each tenant requires isolation and server-to-server communication among tenants is not required.

Layer 3 Separation (Aggregation/Sub-Aggregation)

VRF Lite is a hop-by-hop virtualization technique. Using this technique, each network device and all of its physical interconnections are virtualized. From a data plane perspective, the VLAN tags can provide logical isolation on each point-to-point Ethernet links that connects the virtualized Layer 3 network devices (see [Figure 2-2](#)).

Figure 2-2 Network Device Virtualization with VRF



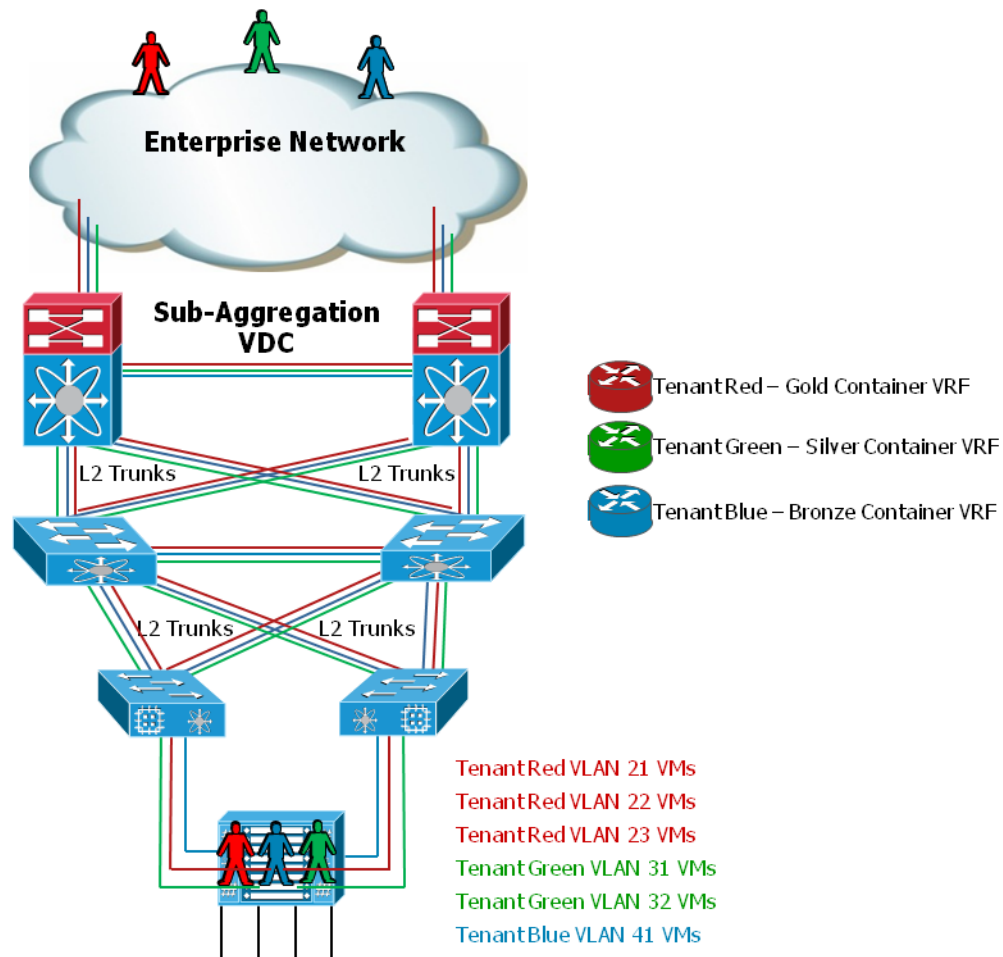
In a multi-tenant environment, Cisco VRF Lite technology offers the following benefits:

- Virtual replication of physical infrastructure—Each virtual network represents an exact replica of the underlying physical infrastructure. This effect results from VRF Lite per hop technique that requires every network device and its interconnections to be virtualized.
- True routing and forwarding separation—Dedicated data and control planes are defined to handle traffic belonging to groups with various requirements or policies. These groups represent an additional level of segregation and security as no communication is allowed among devices belonging to different VRFs unless explicitly configured.

Layer 2 Separation (Access)

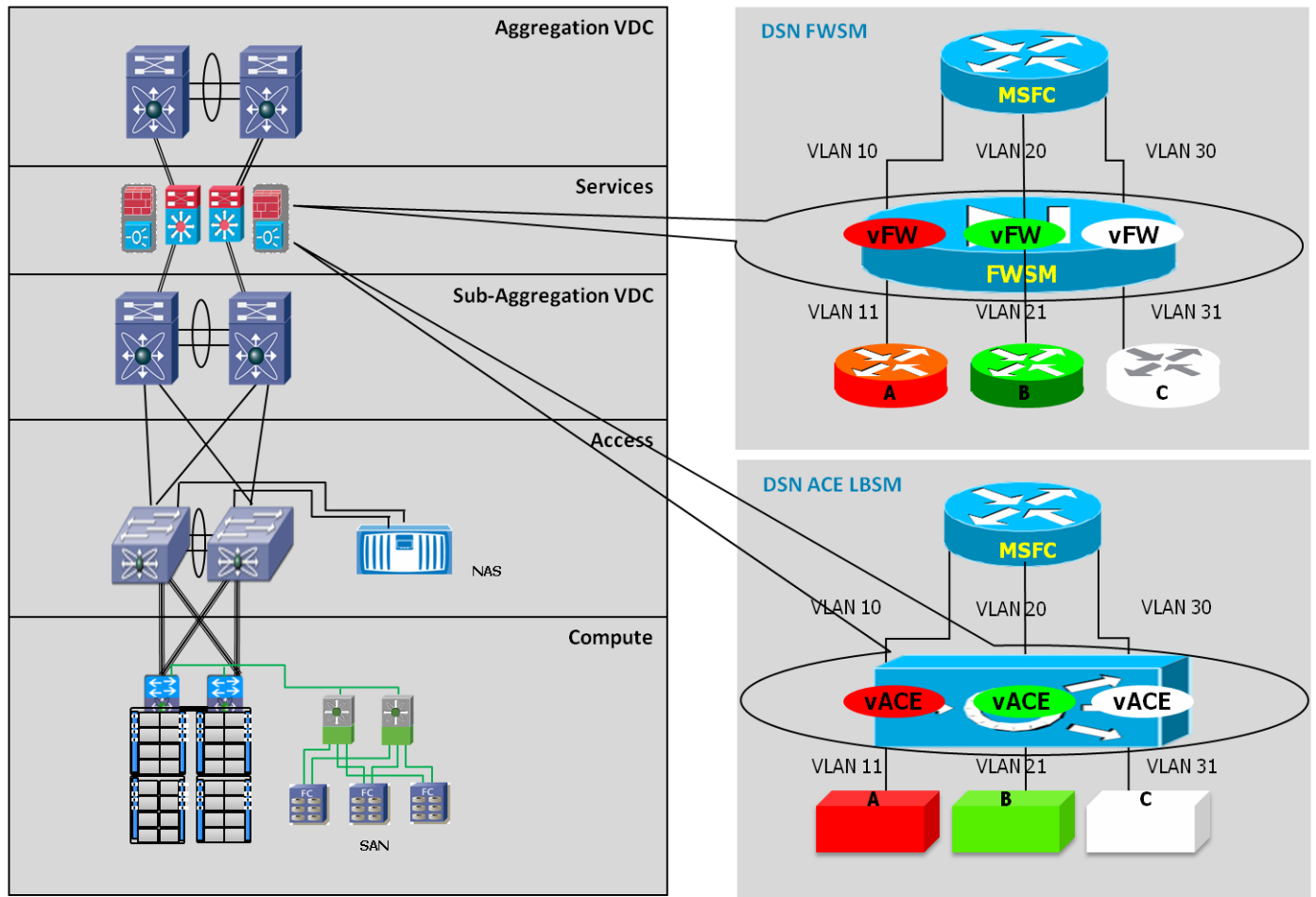
Network separation at Layer 2 is accomplished using VLANs. [Figure 2-3](#) shows how the VLANs defined on each access layer device for Gold network container are mapped to the same Gold VRF at the distribution layer.

Figure 2-3 VLANs to VRF Mapping



Network Services Virtualization

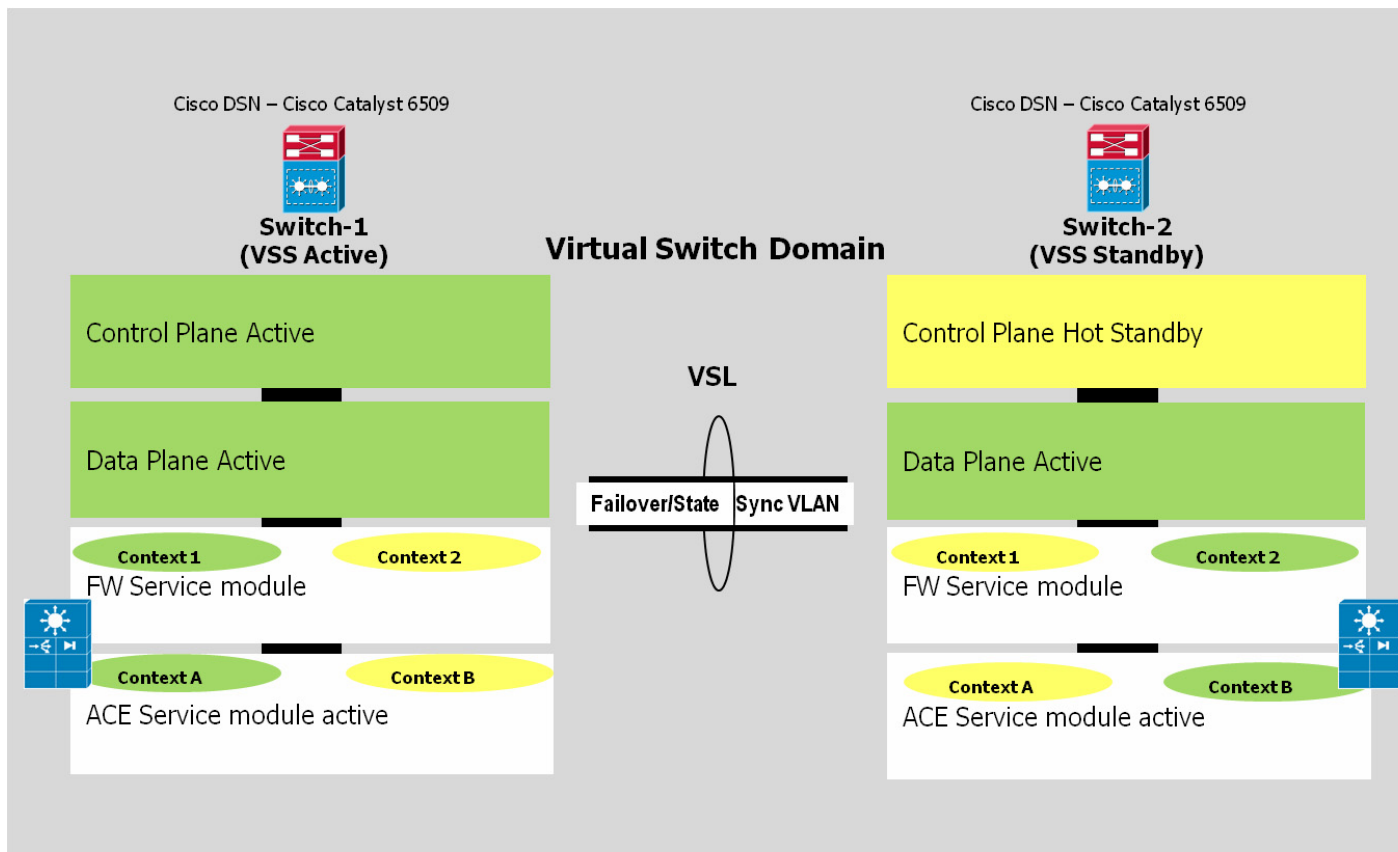
The Cisco Data Center Services Node (DSN) is a Cisco Catalyst 6500 Series Switch with FWSM and ACE service modules dedicated to security and server load balancing functions. To achieve secure separation across the network, the services layer must also be virtualized. [Figure 2-4](#) shows an example of the Cisco DSN directly attached to aggregation layer switches.

Figure 2-4 Virtual Firewall and Load Balancing Services

Using the virtualization features of the Cisco DSN services modules, you can create separate contexts that represent separate virtual devices. The Cisco VMDC solution uses the virtualization features of the Cisco FWSM and Cisco ACE modules to distribute traffic across both Catalyst chassis. As [Figure 2-5](#) depicts, the first Cisco FWSM and Cisco ACE are primary for the first context and standby for the second context. The second Cisco FWSM and Cisco ACE are primary for the second context and standby for the first context. This setup allows modules on both sides of the designs to be primary for part of the traffic, and it allows the network administrator to optimize network resources by distributing the load across the topology.

The Cisco ACE and Cisco FWSM modules balance traffic load per active context. Additional VLANs are carried over the inter-switch link (ISL) to provide fault tolerance and state synchronization. If a Cisco ACE fails, the standby context on its peer module becomes active with little traffic disruption. Active-active design enables traffic load sharing and redundancy.

Figure 2-5 Active-Active Services Chassis with Virtual Contexts



Compute Separation

Virtualization introduces new security challenges and concerns. Traditionally, security policies were applied at the physical server level. However, as physical hosts can now contain multiple logical servers; and therefore, policy must be applied at the VM level. Also, new technologies, such as vMotion, introduced VM mobility within a cluster, where policies follow VMs as they are moved across switch ports and among hosts.

Finally, virtual computing continues to aggregate higher densities of VMs. This high-density model forces us to reconsider firewall scale requirements at the aggregation layer. As a result, high-density compute architectures may require the distribution of security policies to the access layer.

To address some of these new security challenges and concerns, we recommend deploying virtual firewalls at the access layer to create intra-tenant zones. You must also use per-VLAN firewalls at the aggregation layer. Like firewalling at the aggregation layer, layer 2 firewalling can enforce security among the tiers of an application, as described in [Application Tier Separation, page 2-7](#).

VM Security

To provide end-to-end security and traffic isolation for virtual machines, the VMDC solution emphasizes the following techniques:

- **Port profiles.** Port profiles enable VLAN-based separation. Using features found in the Nexus 1000V, you create port profiles and apply them to virtual machine NICs via the VMware vCenter. Each port profile is a policy that can be applied to the VM. The policy settings include VLAN, uplink pinning, security, and policy information.
- **Virtual adapters.** Cisco UCS M81KR Virtual Interface Card (VIC) is a network interface consolidation solution. Traditionally, each VMware ESX server has multiple LAN and SAN interfaces to separate vMotion, service console, NFS, backup, and VM data. In this model, the server requires between four and six network adapters. Using the Cisco VIC, you can create distinct virtual adapters for each traffic flow using a single, two-port adapter.
- **VLAN separation.** Using the Cisco VIC features, you can create virtual adapters and map them to unique virtual machines and VMkernel interfaces through the hypervisor. In a multi-tenant scenario where distinct tenants reside on the same physical server and transmit their data over a shared physical interface, the infrastructure cannot isolate the tenant production data. However, Cisco VIC combined with VN-Link technology can isolate this data via VLAN-based separation. VLAN separation is accomplished when virtual adapters (up to 128) are mapped to specific virtual machines and VMkernel interfaces.

VM Datastore Separation

VMware uses a cluster file system called virtual machine file system (VMFS). An ESX host associates a VMFS volume, which is made up of a larger logical unit (LUN). Each virtual machine directory is stored in the Virtual Machine Disk (VMDK) sub-directory in the VMFS volume. While a VM is operating, the VMFS volume locks those files to prevent other ESX servers from updating them. A VMDK directory is associated with a single VM; multiple VMs cannot access the same VMDK directory.

To maintain tighter control and isolation, architects can map storage LUNs per VM using the raw disk map (RDM) filer system. Each RDM volume maps to a single VM. However, only 255 LUNs can be defined per host; since all resources are in a shared pool, this LUN limitation transfers to the server cluster. In a virtualized environment, this restriction is too limiting. Although a 1:1 mapping of LUNs to tenant VMs is technically possible, it is not recommended because it does not scale and is an inefficient and expensive use of storage resources. In fact, as described in the preceding paragraph, the cluster file system management provided by the hypervisor isolates one tenant's VMDK from another. This coupled with zoning mechanisms and LUN masking isolates tenant datastores within the SAN and at the file system level, serving to limit the effect of VM-based exploits or inadvertent disk corruption.

Application Tier Separation

If a three-tiered application architecture is needed, the tiers can be logically separated on different VLANs. For such requirements, this design proposes using vApp firewalls. This design was validated using VMware vShield for this purpose.

This document addresses the design aspects of the vApp firewalls but does not detail the vShield implementation. For a detailed vShield implementation, refer to VMware documents, such as the *vShield Zones Administration Guide* (www.vmware.com/pdf/vsz_10_admin.pdf).

The Cisco VMDC architecture proposes VLAN separation as the first security perimeter in application tier separation. It proposes that each application reside in separate VLANs within the VRF of tenant. If communication must occur between tiers of an application, the traffic should be routed via the default gateway where security access lists can enforce traffic inspection and access control.

At the access layer of the network, the vShield virtual appliance monitors and restricts inter-VM traffic within and among ESX hosts. Security zones may be created based on VMware Infrastructure (VI) containers, such as clusters, VLANs, or at the VMware Datacenter level. Layer 2, 3, 4, and 7 filters are supported. Security policies can be assured throughout a VM lifecycle, including vMotion events. The vShield Manager organizes virtual machines, networks, and security policies and allows security posture audits in the virtual environment. Monitoring (VM Flow) is performed at the datacenter, cluster, portgroup, VLAN, and virtual machine levels.

A logical construct on the Nexus 1000V, called a virtual service domain (VSD), can classify and separate traffic for vApp-based network services, such as firewalls. Currently, up to 8 VSDs can be configured per host. Up to 512 VSDs can be configured per VSM. A VSD resides on a Service Virtual Machine (SVM), which functions like a “bump in the wire,” serving to segment network traffic. The SVM has three virtual interfaces:

- Management—interface that manages the SVM
- Incoming—guards traffic going into the VSD
- Outgoing—guards traffic exiting the VSD



Note vMotion is not supported for the SVM and must be disabled.

During the vShield agent installation process, the vShield agent vNICs is correlated to the requisite VSD port profiles using the Network Mapping dialog. To bring up the vShield agent, configure the vShield hostname, IP Address and Subnet mask for the vShield VM, and IP Address for the vShield VM's default gateway. Then, the vShield VM is manually added to the vShield Manager inventory.

You can use vCenter to move selected VMs to the member port profile for the VSD. These VMs are protected by the vShield rulesets. vShield allows you to apply two categories of rulesets: L4 (Layer 4) rules and L2/L3 (Layer 2/Layer 3) rules. Layer 4 rules govern TCP and UDP transport of Layer 7 (application-specific) traffic. Layer 2/Layer 3 rules monitor traffic from ICMP, ARP, and other Layer 2 and Layer 3 protocols. These rules are configured at the Data Center level. By default, all Layer 4 and Layer 2/Layer 3 traffic is permitted. These rules are configured on the VM Wall tab. To simplify initial configuration, all vShield firewalls perform stateful inspection and all traffic is permitted by default.

Each vShield agent enforces VM Wall rules in top-to-bottom ordering. A vShield checks each traffic session against the top rule in the VM Wall table before moving down the subsequent rules in the table. This is essentially a first-match algorithm, however, an additional qualification of rulesets exists that uses a hierarchy of precedence levels. This enhancement provides flexibility in terms of applying rulesets at varying VI container level granularity.

In the VM Wall table, the rules are enforced in the following hierarchy:

1. Data Center High Precedence Rules
2. Cluster Level Rules
3. Data Center Low Precedence Rules (in other words, “Rules below this level have lower precedence than cluster level rules” when a data center resource is selected)
4. Default Rules

VM Wall offers container-level and custom priority precedence configurations:

- *Container-level precedence* recognizes the datacenter level as a higher priority than the cluster level. When a rule is configured at the data center level, all clusters and vShield agents within the clusters inherit that rule. A cluster-level rule applies only to the vShield agents in the cluster. These rules must not conflict with higher precedence rules, such as Data Center High Precedence rules.

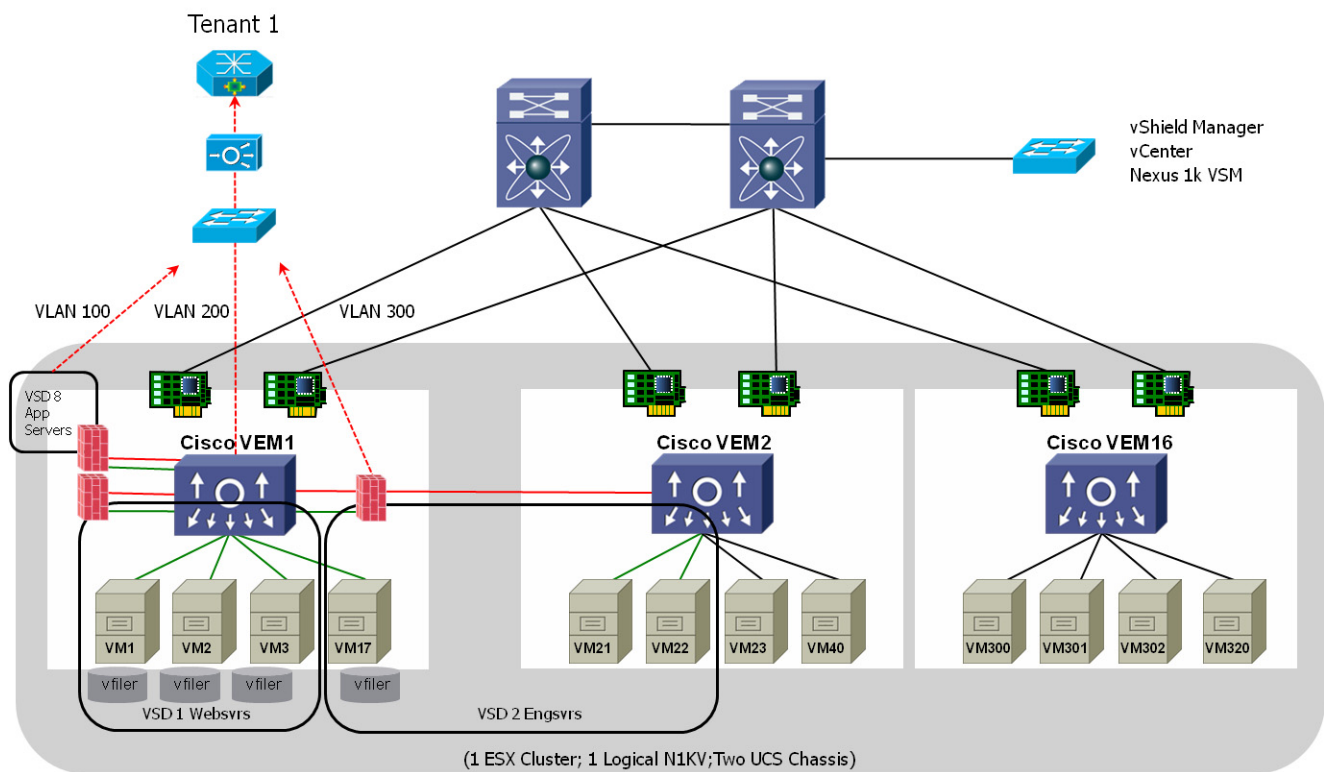
- *Custom priority precedence* allows you to assign high or low precedence to rules at the datacenter level. High precedence rules work like those in container-level precedence. Low precedence rules include the default rules and the configuration of Data Center Low Precedence rules. This option allows you to recognize multiple layers of applied precedence.

**Note**

A key difference exists between the default security stance of vShield firewalls and Cisco firewall: to ease of operation and initial installation, vShield uses an implicit “accept all” packets while Cisco firewall solutions use an implicit “deny all” packets to facilitate highly secure configuration and operation. Therefore, a best-practice security recommendation is to first define “deny all packets” rule for vShield, and then explicitly define rules to allow desired traffic through the firewall.

Figure 2-6 shows how a set of three VSDs/vShields is applied to segment server traffic for a specific tenant. In this example, the ESX cluster extends across two chassis of B200-M1 blade servers installed with M71KR-E mezzanine cards, and the VMs are in a single Nexus 1000V virtual switch. Green lines are protected virtual ports, whereas red lines are unprotected.

Figure 2-6 vFWs in the VMDC Architecture



In VMDC, Nexus 1000V Service Virtual Machines (SVMs) and VSDs with vShield virtual firewalls were implemented to:

1. Define multiple groups of VSD policy zones and apply them to groups of servers for a specific tenant.
2. Create a “Data Center” level rule that denies all traffic. Create a higher precedence rule set that allows specific traffic into the VSD zones and between server VMs/client VMs across the zones. Certain protocols and applications use dynamically allocated port ranges (FTP, MS-RPC, and so

forth). vShield tracks end-point mapper requests and learns the dynamic ports that the VMS are listening on to punch holes in this ephemeral port range only for trusted endpoints. As ephemeral port ranges above 1,024 are often used by botnets and rogue services, the VMDC design advocates using this feature to lock down these ports and to define allow rules only for specific ports for trusted endpoints.

3. Use application-port pair mapping to create application aware rulesets.
4. Validate movement of a vShield firewall policy to another vShield, following the movement of a VM due to a vMotion event, also confirming that the VSDs for the affected vShield continued to operate as expected.

Private VLANs (PVLANS) can complement vFW functionality, effectively creating sub-zones to restrict traffic between VMs within the same VLAN.

This type of distribution of firewall services and policy to the access layer increases scale in hyper-dense compute environments and leverages VMware cluster HA technology to enhance firewall service availability. However, it also presents challenges: the need to scale policy management for larger numbers of enforcement points, and the fact that vApp-based firewalls are relatively new in terms of understanding and managing firewall performance.

Storage Separation

To extend secure separation to the storage layer, we considered the isolation mechanisms available in a SAN environment. Separation occurs both at the switches and the storage arrays connected to the physical fabric.

Storage Area Network (SAN)

MDS Separation

Cisco MDS SAN networks offer many features that make it ideal for the Cisco VMDC solution. These features include true segmentation mechanisms that typically adhere to Fibre Channel protocol guidelines. They offer MDS-specific features that work with zoning to provide separation of services and multiple fabrics on the same physical topology. In addition, the administrative user interfaces reduce troubleshooting costs and increase productivity.

The following features are leveraged for the Cisco VMDC solution:

- **Virtual SANs (VSANs).** By incorporating VSANs in a physical topology, you can include high availability on a logical level by separating large groups of fabrics, such as departmental and homogenous-OS fabrics, without the cost of extra hardware. The VSAN profile is set up so that fabric wide Fibre Channel services, such as name server and zone server, are fully replicated for each new VSAN. This replication ensures that failures and interruptions from fabric changes affect only the VSAN in which they occur.
- **Zoning.** Zoning is another storage security and separation feature available within an MDS VSAN. A zone isolates on per-pWWN basis. A host contains HBAs that act as initiators and are mapped to the port world wide name (pWWN) of a target storage array fabric adapter (FA) port. A host can only communicate with targets in the same zone as the initiating HBA/pWWN. However, a host can associate each HBA/pWWN with a unique zone.

SAN zoning can restrict visibility and connectivity between devices connected to a common Fibre Channel SAN. It is a built-in security mechanism available in a FC switch that prevents traffic leaking between zones. FC zones segment and separate tenants at the physical host level in the SAN network. By default, the MDS does hard zoning.

Hard zoning is enforced by the switch at either ingress or egress. Frames directed to devices outside of the originator's zone are dropped by the switching fabric. In hard zones, the switch does not pass frames from one zone to another. Hard zoning has its limitations. It is designed only to prevent devices from communicating with other unauthorized devices. It is a distributed service common throughout the fabric. Therefore, any configuration changes to a zone disrupt the entire connected fabric. When a zoneset resets, a slight disruption can occur as fabric state change notifications are sent to all switches in the fabric. Rarely, end device connectivity can drop without restoring. However, the disruption caused by configuration changes occurs on a VSAN level for Cisco MDS switches running VSANs. It only affects the VSAN on which the zoneset resides.

Storage Array Separation

On the EMC Symmetrix V-Max storage array, key software features provide for the secure separation of SAN data. Thin pools, device mapping, and LUN masking work together to take extending separation down to the physical disks.

- **Thins Pools.** Thins pools are logical groups of like data devices. Pools take on the same device configuration as the data devices contained within them. For example, RAID5 (7+1) data devices convert a thin pool to a RAID5 (7+1) pool. Therefore, different pools can have different RAID configurations that provide separation at the disk RAID level.
- **Device Mapping.** Device Mapping provides the next layer of separation on a Symmetrix array. Using this feature takes devices and creates a mapping of the device to the front-end ports on the array that connect into the SAN fabric. This mapping creates the equivalent of an access control list within the storage array where the devices can only access ports they are mapped to.
- **LUN Masking.** LUN Masking works with device mapping to further separate the storage array. Three groups are configured; initiators (hosts), storage (devices), and ports (array front end ports) that contain at least 1 member per group. Masking ties all these groups together and the host is then able to view the storage LUN. During this process, the user can define a device specific LUN number that is presented to the host.

NAS

NFS Access Security

NFS servers often need to authenticate that the source of an NFS request is an authorized client system. The NAS can be configured to use any one of several techniques to provide security for end host connectivity. Since an authorized system is designated by hostname or IP address, the NFS server must perform the following steps to verify that an incoming request is from an authorized client:

1. The requesting IP address is obtained from the IP headers in the incoming network packets.
2. The list of authorized IP addresses is obtained from the current server configuration.
3. The requesting IP address is compared against the list of authorized IP addresses to determine if there is a match.

Source IP address is the only method that was validated in VMDC 2.0 since everything is performed using IP addresses (or network numbers with subnet masks); however, some additional techniques may be performed if the configuration on the server is specified using hostnames.

NetApp MultiStore

The NetApp MultiStore software allows the creation of separate and private logical partitions on a single storage system. Each virtual storage partition maintains separation from every other storage partition, so multiple tenants can share the same storage resource without compromise to privacy and security.

High Availability

High Availability is key for building a virtualized cloud environment. Eliminating planned downtime and preventing unplanned downtime are key aspects in the design of the multi-tenant shared services infrastructure. This section covers availability design considerations and best practices related to compute, network, and storage. See [Table 2-1](#) for various methods of availability.

Table 2-1 *Methods of Availability*

Network	Compute	Storage
- EtherChannel - vPC - Device/Link Redundancy - MAC Learning - Active/Passive VSM	- UCS Dual Fabric Redundancy - vCenter Heartbeat - VMware HA - vMotion - Storage vMotion	RAID-DP Virtual Interface (VIF) NetApp HA Snapshot SnapMirror and SnapVault

Service Availability

Service availability is calculated using the following formula:

$$\text{Availability} = \frac{(T_{\text{period}} - T_{\text{without service}} \times 100)}{(T_{\text{period}})}$$

This formula provides a measurement of the percentage of time, for the period T (such as a month, a quarter, a year), in which the service was available to one's tenants. It is common for IaaS public cloud providers to offer an SLA target on average of 99.9% or 3 nines availability. This level equates to a downtime of no more than 8.76 hours per year.

[Table 2-2](#) lists applied availability components common in the IaaS SLA context. A few components, such as managed security services, may be more applicable to the public cloud services context.

Table 2-2 *Applied SLA Availability Components*

Availability Component	Performance Indicators
Portal Availability	Portal service availability; information accuracy, successfully processed service requests
Virtual Machine Availability	Percentage of service availability (% Availability)

Table 2-2 *Applied SLA Availability Components (continued)*

Availability Component	Performance Indicators
Virtual Machine RTO	Recovery Time Objective for restore of a virtual machine in the event of a server crash.
Storage Availability	% Availability
Network Availability	% Availability
Firewall Availability	% Availability (of a vApp vFW or virtual context in the FWSM)
Load Balancer Availability	% Availability (of a virtual context in the ACE)
Backup Data Reliability	% (Scheduled) Successful data backup attempts: this can refer to actual datastore backups or successful clone or mirror attempts.
Managed Security Service Availability	A managed security service is a general term for a number of possible services: these include VPNs (SSL, IPSec, MPLS), IPS, deep packet inspection, DDoS mitigation and compliance (file access auditing and data or datastore encryption) services. Performance indicators will vary depending on how these services are deployed and abstracted to upper layer service level management software.

In addition to the availability components in [Table 2-2](#), service performance components can include incident response time and incident resolution objectives. The latter varies based on the type of service component (VM, network, storage, firewall, and so forth).

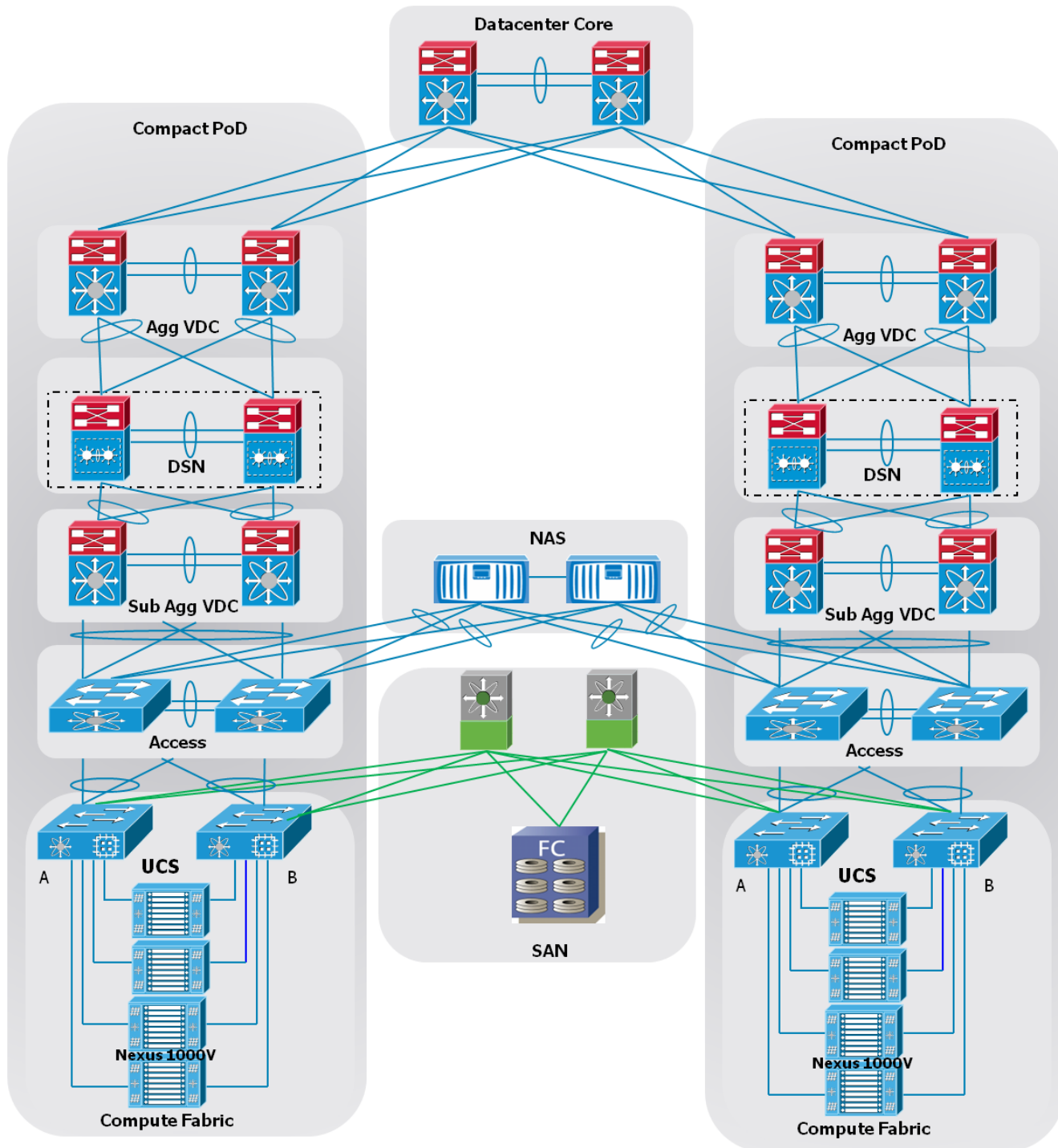
The VMDC architecture addresses the availability requirements of IaaS SLAs for all of the criteria listed in [Table 2-2](#) through 1:1, 1:N or N:N VM, network and storage redundancy, and data security and data protection mechanisms.

Network Availability

Network availability is paramount to any organization running a virtualized data center service. It is strategic for disaster planning, as well as everyday operations, and ensures that tenants can reliably access application servers.

Highly Available Physical Topology

Figure 2-7 Physical Topology



Core Availability

The data center core is meant to be a high-speed Layer 3 transport for inter- and intra-datacenter traffic. High availability at the core is an absolute requirement in any datacenter design. Using the technologies available in the Nexus 7000, it can be achieved in the following ways:

- **Device redundancy**—The core is typically composed of two devices, each with a connection to outside of the data center and a connection back to the aggregation layer of the data center.
- **Supervisor redundancy**—To account for hardware failures within the Nexus 7000, redundant supervisors can be installed.
- **Path redundancy**—With the core comprised of Layer 3 links, this is done primarily using redundant routed paths. The core should have redundant paths to the campus and WAN as well as the aggregation layer VDC of any Compact Pods deployed in the datacenter.
- **Fast Convergence**—Bidirectional Forwarding Detection (BFD) provides fast peer failure detection times across all media types, encapsulations, and topologies, and for multiple routing protocols, including BGP, EIGRP, IS-IS, and OSPF. It sends rapid failure detection notices to the configured routing protocols on the local router to initiate the routing table recalculation process. Once BFD is enabled on the interfaces and for the appropriate routing protocols, a BFD session is created, BFD timers are negotiated, and the BFD peers exchange BFD control packets at the negotiated interval.

For example, if BFD with OSPF protocol is enabled between the pair of aggregation switches and a BFD neighbor session with its OSPF neighbor router goes down, BFD notifies the OSPF process that the BFD neighbor is no longer reachable. To reduce failure recovery times, OSPF removes the neighbor relationship to that router and looks for an alternative path without waiting for the hold timer to expire.

This VMDC 2.0 Compact Pod does not specify a core design as there are several ways it could be deployed. There may be a separate core layer or the core could be collapsed onto the aggregation layer in a single Compact Pod datacenter design.

Aggregation Layer VDC Availability

To achieve high availability in the aggregation layer, many of the features used for availability in the core layer are utilized in addition to some key features available with the Nexus 7000:

- **Multi-Chassis Ether Channels (MEC)**—Multi-Chassis Ether Channels were used to connect the aggregation layer to the services layer. MEC allows for redundant paths between entities while simultaneously removing the sub-optimal blocking architecture associated with traditional spanning tree designs.
- **Virtual Route and Forwarding (VRF)**—Redundant pairs of VRF instances provide Layer 3 services for their associated tenant VLAN segments.
- **First hop redundancy**—HSRP was used to provide gateway redundancy for the services (ACE) devices in the data center. Each switch can become an HSRP peer in however many groups are required for the design. Each of the network containers with ACE services would have an HSRP group.

Services Availability

High availability for the services layer can be achieved whether using appliances or the service chassis design. Appliances can be directly attached to the aggregation switches or to a dedicated services switch, usually a Cisco 6500 series switch. The service chassis design involves using modules designed for the 6500 series chassis.

If using appliances, high availability can be achieved by logically pairing two physical service devices together. The pairs can be used in an active/standby model or an active/active model for load balancing, depending on the capabilities of each appliance pair. Certain service appliances, such as the ASA and ACE, can load balance by dividing their load among virtual contexts that allow the appliance to act as multiple appliances. This can also be achieved with their service module counterparts. This can be particularly valuable in a tenant environment where it is desirable to present each tenant with their own appliance to ensure separation.

The same can be achieved using the service chassis design, but HA would also need to be implemented at the service chassis level. An ideal way to implement this would be to use the Virtual Switching System (VSS). With VSS, redundant modules can be paired across different chassis, but management can be simplified by presenting the two chassis as one to the administrator.

The VMDC 2.0 design uses a services chassis design with VSS implemented. Both service modules, the Firewall Services module and the Application Control Engine module, support virtual contexts and are implemented in an active/active failover pair. EtherChannel load balancing is configured on the 6500 switch to insure link resiliency.

- **Multi-Chassis EtherChannels (MEC)**—Multi-Chassis EtherChannels were used to connect the aggregation layer to the services layer. MEC allows for redundant paths between entities while simultaneously removing the sub-optimal blocking architecture associated with traditional spanning tree designs.

Sub-Aggregation Layer VDC Availability

To achieve high availability in the aggregation layer, many of the features used for availability in the core layer are utilized in addition to some key features available with the Nexus 7000:

- **Multi-Chassis Ether Channels (MEC)**—Multi-Chassis Ether Channels were used to connect the sub-aggregation layer to the services layer. MEC allows for redundant paths between entities while simultaneously removing the sub-optimal blocking architecture associated with traditional spanning tree designs.
- **Virtual Port Channels (VPC)**—Virtual port channels were used to connect the aggregation layer to the access layer. VPC allows for redundant paths between entities while simultaneously removing the sub-optimal blocking architecture associated with traditional spanning tree designs.
- **Virtual Route and Forwarding (VRF)**—Redundant pairs of VRF instances provide Layer 3 services for their associated tenant VLAN segments.
- **First hop redundancy**—HSRP can be used to provide gateway redundancy for the edge devices in the data center. Each switch can become an HSRP peer in however many groups are required for the design. Each network container VLAN would have an HSRP group.

Access Layer Availability

Access layer is designed with the following key design attributes in Nexus 5000:

- Enables loop-less topology via Virtual Port-Channel (vPC) technology. The two-tier vPC design is enabled such that all paths from end-to-end are available for forwarding.
- Nexus 7000 to Nexus 5000 is connected via a single vPC between redundant devices and links. In this design four 10Gbps links are used, however for scalability one can add up to eight vPC members in the current Nexus software release.
- The design recommendation is that any edge layer devices should be connected to Nexus 5000 with port-channel configuration.

- RPVST+ is used as spanning tree protocol in VMDC 2.0 Compact pod. MST may be used if the VLAN and host scale requirements are much larger. For example, MST was used in the VMDC 2.0 Large pod architecture. The Sub-aggregation VDCs are the primary and secondary root for all VLANs. The HSRP priority is matched to the root to ensure optimal traffic flows.

Compute Availability

To provide high availability at the compute layer, the Cisco VMDC solution relies on the following features:

- UCS End-host mode
- Cisco Nexus 1000V and Mac-pinning
- Redundant VSMs in active-standby mode
- High availability within the cluster
- Automated disaster recovery plans

UCS End-Host Mode

Unified Computing System (UCS) fabric interconnect running in end host mode do not function like regular LAN switches.

When UCS Fabric Interconnects operate in End-Host Mode (as opposed to Switch Mode), the virtual machine NICs (VMNICs) are pinned to UCS fabric uplinks dynamically or statically. (VMNICs are logical names for the physical NICs in the server.) These uplinks connect to the access layer switch to provide redundancy toward the network. The fabric interconnect uplinks appear as server ports to the rest of the fabric. When End-Host Mode is enabled, STP is disabled and switching between uplinks is not permitted. End-Host Mode is the default and recommended when the upstream device is a Layer 2 switch. Key benefits of End-Host Mode include the following:

- Reduced STP and Layer 2 forwarding and improved control plane scale-required to learn MAC addresses local to the fabric.
- Active-active uplinks-with STP enabled to block the spanning tree loop, one of the links will be in the in STP alt/block state.

Cisco Nexus 1000V and Mac-Pinning

The Cisco UCS system always load balances traffic for a given host interface on one of the two available fabrics. If a fabric fails, traffic fails over to the available fabric. Cisco UCS only supports port ID- and source MAC address-based load balancing mechanisms. However, Nexus 1000V uses the mac-pinning feature to provide more granular load-balancing methods and redundancy.

VMNICs can be pinned to an uplink path using port profiles definitions. Using port profiles, the administrator can define the preferred uplink path to use. If these uplinks fail, another uplink is dynamically chosen.

If an active physical link goes down, the Cisco Nexus 1000V Series Switch sends notification packets upstream of a surviving link to inform upstream switches of the new path required to reach these virtual machines. These notifications are sent to the Cisco UCS 6100 Series Fabric Interconnect, which updates its MAC address tables and sends gratuitous ARP messages on the uplink ports so the data center access layer network can learn the new path.

Deploy Redundant VSMs in Active-Standby Mode

Always deploy the Cisco Nexus 1000V Series VSM (virtual supervisor module) in pairs, where one VSM is defined as the primary module and the other as the secondary. The two VSMs run as an active-standby pair, similar to supervisors in a physical chassis, and provide high availability switch management. The Cisco Nexus 1000V Series VSM is not in the data path so even if both VSMs are powered down, the Virtual Ethernet Module (VEM) is not affected and continues to forward traffic.

Each VSM in an active-standby pair is required to run on a separate VMware ESX host. This requirement helps ensure high availability even if one VMware ESX server fails. You should also use the anti-affinity feature of VMware ESX to help keep the VSMs on different servers.

VMware HA for Intra-Cluster Resiliency

The VMDC architecture uses VMware HA for intra-cluster resiliency. VMware HA provides 1:N failover for VMs in a cluster, which is better than the 1:1 failover between a primary and secondary VM in a cluster provided by VMware Fault Tolerance. To indicate health, VMware HA agent on each server maintains a heartbeat exchange with designated primary servers in the cluster. These primary servers maintain state and initiate failovers. Upon server failure, the heartbeat is lost and all the VMs for that server restart on other available servers in the cluster's pool. A prerequisite for VMware HA is that servers in the HA pool must share storage; virtual files must be available to all servers in the pool. Also, in the case of FC SANs, adapters in the pool must be in the same zone.

VMware HA

For VMware HA, consider the following:

- The first five ESX hosts added to the VMware HA cluster are primary nodes; subsequent hosts added are secondary nodes. Primary nodes are responsible for performing failover of virtual machines in the event of host failure. For HA cluster configurations spanning multiple blade chassis (that is, there are more than eight nodes in the cluster) or multiple data centers in a campus environment, ensure the first five nodes are added in a staggered fashion (one node per blade chassis or data center).
- With ESX 4.0 Update 1, the maximum number of virtual machines for an eight-node VMware HA cluster is 160 per host, allowing for a maximum of 1280 virtual machines per cluster. If the cluster consists of more than eight nodes, the maximum number of virtual machines supported for failover is 40 per host.
- Host Monitoring can be disabled during network maintenance to prevent against “false positive” virtual machine failover.
- Use the “Percentage of cluster resources reserved as failover spare capacity” admission control policy as tenant virtual machines may have vastly different levels of resource reservations set. Initially, a Cloud administrator can set the failover capacity of 25%. As the environment reaches steady state, the percentage of resource reservation can be modified to a value that is greater than or equal to the average resource reservation size or amount per ESX host.
- A virtual machine's restart priority in the event of ESX Server host failure can be set based on individual tenant SLAs.
- Virtual machine monitoring sensitivity can also be set based on individual tenant SLAs.

VMware vShield

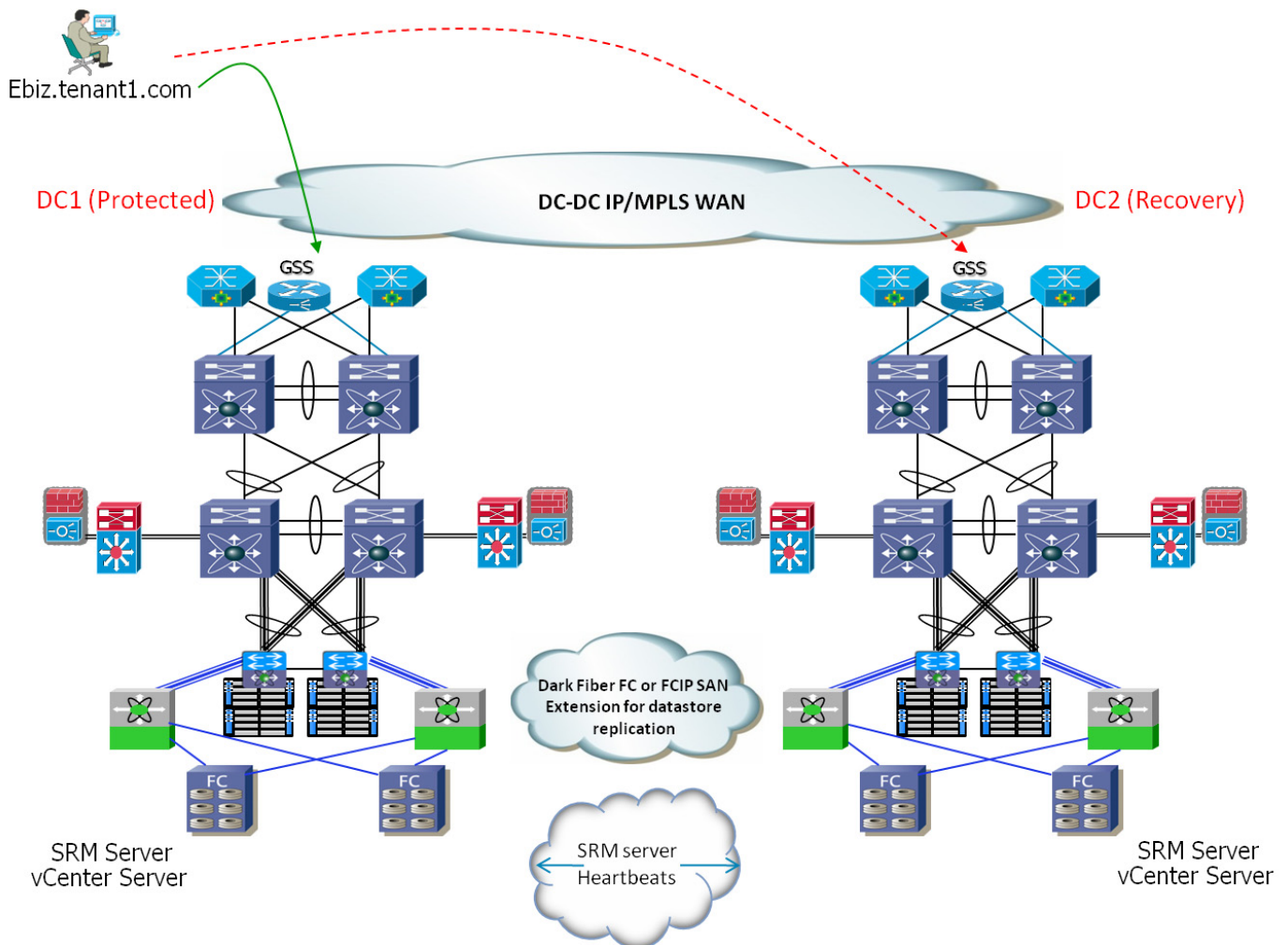
For VMware vShield:

- The vShield virtual machine on each ESX host should have the “virtual machine restart priority” setting of “disabled” as an instance of vShield running on another ESX host will take over the policy enforcement for the virtual machines after HA failover automatically.

Create Automated Disaster Recovery Plans

Tools such as VMware's Site Recovery Manager, coupled with Cisco's Global Site Selection for DNS redirection and synchronous or asynchronous datastore replication solutions such as EMC's SRDF may be used to create automated recovery plans for critical groups of VMs. SRM allows for the specification of source/target resource pairing and thus in contrast to vMotion does not rely on layer two LAN extension as a prerequisite for VM and datastore replication. The VMDC model examines the use of SRM in this manner to provide DR for a selected subset of VMs from the Gold and Silver service tiers, assuming an active/standby relationship between the primary and secondary Data Centers (see [Figure 2-8](#)). Implementation details are described in another module of this solution document set, specific to the topic of DR.

Figure 2-8 Virtual Machine DR Between Active/Standby Data Centers



Storage Availability

In the storage layer, the design is consistent with the high availability model implemented at other layers in the infrastructure, which include physical and path redundancy. [Table 2-3](#) lists the redundancy methods for storage.

Table 2-3 Storage Layer Redundancy Methods

Redundancy Method	Details
Link redundancy	- Redundant links distributed across Line cards using Port-Channels - Multipathing
Hardware redundancy	- Redundant adapter ports (such as CNAs and HBAs) per server - Dual supervisors (MDS) - Dual storage controllers and redundant directors (NetApp NAS and EMC SAN) - RAID 1 and RAID 5 redundant arrays
Node redundancy	Redundant storage devices (MDS switches and SAN fabrics)

Storage Area Network (SAN) Availability

High availability within a FC fabric is easily attainable via the configuration of redundant paths and switches. A given host is deployed with a primary and redundant initiator port which is connected to the corresponding fabric. With a UCS deployment, a dual port mezzanine card is installed in each blade server and a matching vHBA and boot policy are setup providing primary and redundant access to the target device. These ports access the fabric interconnect as N-ports which are passed along to a northbound FC switch. Zoning within the redundant FC switches is done such that if one link fails then the other handles data access. Multipathing software is installed dependent on the operating system which ensures LUN consistency and integrity.

When designing SAN booted architectures, considerations are made regarding the overall size and number of hops that an initiator would take before it is able to access its provisioned storage. The fewer hops and fewer devices that are connected across a given interswitch link, the greater the performance of a given fabric. A common target ratio of hosts across a given switch link would be between 7:1 or 10:1, while an acceptable ratio may be as high as 25:1. This ratio can vary greatly depending on the size of the architecture and the performance required.

SAN Connectivity should involve or include:

- The use of redundant VSANs and associated zones
- The use of redundant interswitch links ISLs where appropriate
- The use of redundant target ports
- The use of redundant fabrics with failover capability for fiber channel SAN booted infrastructure

Fault Isolation

Consolidating multiple areas of storage into a single physical fabric both increases storage utilization and reduces the administrative overhead associated with centralized storage management. The major drawback is that faults are no longer isolated within individual storage areas. Many organizations would like to consolidate their storage infrastructure into a single physical fabric, but both technical and business challenges make this difficult.

Technology such as virtual SANs (VSANs) enables this consolidation while increasing the security and stability of Fibre Channel fabrics by logically isolating devices that are physically connected to the same set of switches. Faults within one fabric are contained within a single fabric (VSAN) and are not propagated to other fabrics.

Dual Fabric Design - Link and Node Redundancy

The VMDC solution leverages best practices for SAN high availability that prescribe full hardware redundancy at each device in the I/O path from host to SAN. Hardware redundancy begins at the server with dual port adapters per host. Redundant paths from the hosts feed into dual, redundant MDS SAN switches (with dual supervisors) and then into redundant SAN arrays with tiered, RAID protection. RAID 1 and 5 are deployed as two most commonly used levels; however the selection of a RAID protection level depends on the balance between cost and the criticality of the stored data.

PortChannel

A PortChannel can be configured without restrictions to logically bundle physical links from any port on any Cisco MDS 9000 Family Fibre Channel Switching Modules. This feature allows you to deploy highly available solutions with great flexibility. In case of a port, ASIC, or module failure, the stability of the network is not affected because the logical PortChannel remains active even though the overall bandwidth is reduced. The MDS PortChannel solution scales to support up to 16 ISLs per PortChannel and aggregates 1-, 2-, 4-, 8-, or 10-Gbps Fibre Channel links. This feature aggregates up to 20,400 MB of application data throughput per PortChannel for exceptional scalability. The MDS PortChannel solution neither degrades performance over long distances nor requires specific cabling. MDS PortChannel uses flow-based load balancing to deliver predictable and robust performance independent of the distance covered.

UCS Link Redundancy

Pending the upcoming availability of FC port channels on UCS FC ports and FC Port Trunking, multiple individual FC links from the UCS 6120s are connected to each SAN fabric, and VSAN membership of each link is explicitly configured in the UCS. In the event of an FC (NP) port link failure, affected hosts login again using available ports in a round-robin fashion. FC port channel support, when available will mean that redundant links in the port channel will provide active/active failover support in the event of a link failure. Multipathing software from VMware or the SAN storage vendor (such as EMC Powerpath software) further enhances HA, optimizing the available link bandwidth use and load balancing across multiple active host adapter ports and links for minimal disruption in service.

PowerPath

PowerPath on the host side allows efficient load balancing across all available paths to the storage array, maximizing the efficiency for I/O throughput, as well as increasing convergence time when a failure occurs.

Storage Backup and Restoration

EMC Symmetrix arrays offer local and remote data backup and protection. The TimeFinder feature allows for local backups in the form of snaps, for storage with limited data resources that require a less costly backup solution, or clones, for a full blown local copy of critical data. To accompany the local protection provided by TimeFinder, EMC's Symmetrix Remote Data Facility (SRDF) feature enables site to site protection. SRDF works synchronously or asynchronously depending on the distance between the sites.

Network Attached Storage (NAS) Availability

A cluster of two independent NAS appliances are deployed and linked together for failover. NAS clusters include built-in, high-availability functionality that permits the cluster to detect and respond to failures of network interfaces, storage interfaces, servers and operating systems within the cluster. The NAS cluster builds on this infrastructure to provide transparent failover of NFS or CIFS sessions.

NetApp Clustering

NetApp HA pairs provide continuous data availability for multi-tenant solutions. The deployment of an HA pair of NetApp controllers ensures the availability of the environment both in the event of failure and in the event of upgrades.

Storage controllers in an HA pair have the capability to seamlessly take over its partner's roles in the event of a system failure. These include controller personalities, IP addresses, SAN information, and access to the data being served. This is accomplished using cluster interconnections, simple administrative setup, and redundant paths to storage. In the event of an unplanned outage, a node assumes the identity of its partner with no re-configuration required by any associated hosts. HA pairs also allow for non-disruptive upgrades for software installation and hardware upgrades. A simple command is issued to takeover and giveback identity.

The following considerations should be made when deploying an HA pair:

- Best practices should be deployed to ensure any one node can handle the total system workload.
- Storage controllers communicate heartbeat information using a cluster interconnect cable.
- Takeover process takes milli-seconds.
- TCP sessions to client hosts are re-established following a timeout period.
- Some parameters must be configured identically on partner nodes.

Storage Network Connectivity (VIFs) using LACP

NetApp provides three types of Virtual Interfaces (VIFs) for network port aggregation and redundancy:

- SingleMode
- Static MultiMode
- Dynamic MultiMode

The Secure Cloud environment uses Dynamic MultiMode VIFs due to the increased reliability and error reporting, as well as compatibility with Cisco Virtual Port Channels. A Dynamic MultiMode VIF uses Link Aggregation Control Protocol (LACP) to group multiple interfaces together to act as a single logical link. This provides intelligent communication between the storage controller and the Cisco Nexus allowing for load balancing across physical interfaces as well as failover capabilities.

Data Availability with RAID Groups and Aggregates

RAID groups are the fundamental building block when constructing resilient storage arrays containing any type of application data set or virtual machine deployment. There exists a variety of levels of protection and costs associated with different RAID groups. A storage controller that offers superior protection is an important consideration to make when designing a multi-tenant environment as hypervisor boot, guest VMs, and application data sets are all deployed on a shared storage infrastructure. Furthermore, the impact of multiple drive failures is magnified as disk size increases. Deploying a NetApp storage system with RAID DP offers superior protection coupled with an optimal price point.

RAID-DP is a standard Data ONTAP feature that safeguards data from double disk failure by means of using two parity disks. With traditional single-parity arrays, adequate protection is provided against a single failure event such as a disk failure or error bit error during a read. In either case, data is recreated using parity and data remaining on unaffected disks. With a read error, the correction happens almost instantaneously and often the data remains online. With a drive failure, the data on the corresponding disk has to be recreated, which leaves the array in a vulnerable state until all data has been reconstructed onto a spare disk. With a NetApp array deploying RAID-DP, a single event or second event failure is survived with little performance impact as there exists a second parity drive. NetApp controllers offer superior availability with less hardware to be allocated.

Aggregates are concatenations of one or more RAID groups that are then partitioned into one or more flexible volumes. Volumes are shared out as file level (NFS or CIFS) mount points or are further allocated as LUNs for block level (iSCSI or FCP) access. With NetApp's inherent storage virtualization, all data sets or virtual machines housed within a shared storage infrastructure take advantage of RAID-DP from a performance and protection standpoint. For example, with a maximum UCS deployment there could exist 640 local disks (two per blade) configured in 320 independent RAID-1 arrays all housing the separate hypervisor OS. Conversely, using a NetApp array deploying RAID-DP, these OSES could be within one large aggregate to take advantage of pooled resources from a performance and availability perspective.

Much as an inferior RAID configuration is detrimental to data availability, the overall failure of the storage controller serving data can be catastrophic.

Storage Backup and Restoration

NetApp storage controllers support various mechanisms for backup and restoration of data, which is of particular importance in a multi-tenant architecture consisting of shared infrastructure. This section discusses the concepts supported by Data ONTAP with respect to data retention and recovery. It should be noted that existing backup solutions are often in place and the NetApp software suite offers seamless integration for many of these applications. In light of this, the following section illustrates the options and flexibility available in backing up and restoring files, volumes, and aggregates.

The primary methods available from NetApp to backup, replicate, and restore data in the Secure Cloud are as follows:

- Snapshots (Aggregate and Volume level) and SnapRestores of the primary file system
- SnapMirror and SnapVault

Snapshots

Aggregate snapshots provide a point-in-time view of all data within an entire aggregate including all contained flexible volumes. A restoration of an aggregate snapshot restores all data in all flexible volumes contained within that aggregate to the same point-in-time, overwriting the existing data.

Volume-Based Snapshots are taken at the volume level, as the associated applications are contained within a volume. Here are some considerations to be made for Volume Snapshots:

- There can only be 255 active snapshots in a volume.
- The snapshot is read-only. Snapshots are scheduled on the primary copy of the data.
- All efforts should be made to ensure data is in a consistent state before creating a snapshot.
- Snapshot Autodelete can be configured to remove older Snapshots to save space.
- Application owners can view their own read-only Snapshots.
- Snapshots can easily be backed up to tape or virtual tape.

Snapshots can be triggered by a number of means; the primary methods are:

- Scheduled snapshots (asynchronous), setup by the storage administrator.
- Remote authenticated Snapshots using ZAPI (an XML protocol over HTTPS).
- Isolated Snapshots by Proxy Host on a per-application basis.

SnapMirror and SnapVault

SnapMirror is replication software intended for disaster recovery solutions or for the replication of volumes to additional controllers or vFiler units. The mirror is an exact replica of data on the primary storage, including all the local Snapshot copies, and can be mounted read-write to recover from failure. If a Snapshot backup is deleted on the source, it goes away on the mirror at the next replication. Here are some considerations to be made:

- A SnapMirror can easily be backed up to tape/virtual tape.
- A SnapMirror provides a means to perform a remote enterprise-wide online backup.
- SnapMirrors can be mounted read-write for failover or maintenance of the primary system.

SnapVault, in contrast, is intended for disk-to-disk backup. A separate Snapshot retention policy is specified for the target environment, allowing long-term archiving of Snapshot backups on secondary storage. Secondary copies managed only by SnapVault cannot be mounted read-write. Backups must be recovered from secondary storage to the original or to an alternative primary storage system to restart.

Like SnapMirror, SnapVault can easily be backed up to tape or virtual tape. Here are some considerations to be made in regards to SnapVault:

- SnapVault can be used in conjunction with SnapMirror for a multi-tiered archive workflow.
- SnapVault can not be mounted read-write as it only stores block-level changes of Snapshots.

Performance and Scalability

Performance is a measure of the speed at which a computer system works. Scalability is the ability to grow in size or complexity without showing negative effects. Problems in either area may expose the enterprise to operating inefficiencies and potential failures of critical business components. Testing, monitoring, and tuning the environment ensures optimal performance and user satisfaction.

Network Performance and Scalability

A challenge of the VMDC architecture is the ability to function well as tenants needs change in size or volume. The following section highlights some of the key scalability variables for each layer of the network.

Layer 3 Scalability

The following features enable Layer 3 scalability in the Compact Pod design for Cisco VMDC:

- [Virtual Routing and Forwarding \(VRF\) Instances, page 2-25](#)
- [Hot-Standby Router Protocol \(HSRP\), page 2-25](#)
- [OSPF Network Scalability, page 2-26](#)
- [Bidirectional Forwarding Detection \(BFD\) for OSPF, page 2-27](#)
- [IP Route Summarization, page 2-28](#)

Virtual Routing and Forwarding (VRF) Instances

In VMDC 2.0, each network container (Gold, Silver, Bronze) uses a unique VRF instance. A tenant may be allocated more than one container depending on their requirements. Compact Pod

A VRF instance consists of:

- an IP routing table
- a derived forwarding table
- a set of interfaces that use that forwarding table
- a routing protocol that determines what reachability goes into the forwarding table

The Cisco VMDC solution Compact Pod was tested and validated using 32 VRFs in the Aggregation Virtual Device Context (VDC) and 32 VRFs in the Sub-Aggregation VDC. [Table 2-4](#) lists the Cisco verified limits for Nexus switches running Cisco NX-OS Release 5.x.

Table 2-4 Cisco NX-OS Release 5.x VRF Configuration Limits

Feature	Verified Limit	VMDC 2.0 Compact Pod Scale
VRFs	1000 per system	64 per Nexus 7010
	250 maximum on each VDC (with 4 VDCs)	32 in Aggregation VDC
		32 in Sub Aggregation VDC

Hot-Standby Router Protocol (HSRP)

Common guidance for optimization of HSRP for fast failover is to reduce the hello and hold timers from their defaults of 3 and 10 seconds, respectively. NX-OS does support HSRP, version 2 with millisecond timers; however, a hello timer of 1 second and hold timer of 3 seconds provides fast failover without creating a high control plane load in networks with a large number of VLAN interfaces. Also, when using hello and hold timers that match those of the routing protocol, the default gateway services failover with similar timing to the IGP neighbor relationships. HSRP hello and hold timers of 1 and 3 seconds are recommended for fast failover, and they were validated in the VMDC Compact Pod architecture.

[Table 2-5](#) lists the Cisco verified limits and maximum limits for switches running Cisco NX-OS Release 5.x.

Table 2-5 Cisco NX-OS Release 5.x HSRP Configuration Limits

Feature	Verified Limit	VMDC 2.0 Compact Pod Scale
HSRP	2000 IPv4 groups per system, with 3s/10s timers.	180 HSRP groups in Sub-Aggregation VDC
	500 HSRP groups per physical interface or VLAN interface.	32 groups per port channel in Aggregation VDC
	100 HSRP groups per port-channel interface.	

OSPF Network Scalability

A routing protocol must be configured for each network containers (VRF) to exchange reachability information between the aggregation and sub-aggregation layers. The Cisco VMDC solution Compact Pod was tested and validated using Open Shortest Path First (OSPF) in each of the 32 network containers.

The ability to scale an OSPF internetwork depends on the network structure and address scheme. Network scalability is determined by the utilization of three resources: memory, CPU, and bandwidth.

- **Memory**—An OSPF router stores the link states for all of the areas that it is in. In addition, it can store summaries and externals. Careful use of summarization and stub areas can reduce memory use substantially.
- **CPU**—An OSPF router uses CPU cycles when a link-state change occurs. Keeping areas small and using summarization dramatically reduces CPU use and creates a more stable environment for OSPF.
- **Bandwidth**—OSPF sends partial updates when a link-state change occurs. The updates are flooded to all routers in the area. In a quiet network, OSPF is a quiet protocol. In a network with substantial topology changes, OSPF minimizes the amount of bandwidth used.

Table 2-6 lists the Cisco maximum OSPF limits for switches running Cisco NX-OS Release 5.x.

Table 2-6 Cisco NX-OS Release 5.x OSPF Configuration Limits

Feature	Maximum Limit	Cisco VMDC Compact Pod Scale
OSPF	200 interfaces	64 interfaces in Aggregation VDC 32 interfaces in Sub-Aggregation VDC
	1000 routers	5 routers per VRF in Aggregation VDC 3 routers per VRF in Sub Aggregation VDC
	300 adjacencies	5 adjacencies per VRF in Aggregation VDC (total 160 per Nexus 7010) 3 adjacencies per VRF in Sub Aggregation VDC (total 96 per Nexus 7010)
	200,000 LSAs	23 LSAs per Gold VRF 23 LSAs per Silver VRF 23 LSAs per Bronze VRF
	4 instances per VDC	1 per VDC
	Up to the system maximum VRFs in an OSPF instance.	32 VRFs per OSPF instance

Consider the following guidance when deploying OSPF:

- **OSPF Neighbor Adjacencies**—OSPF floods all link-state changes to all routers in an area. Routers with many neighbors do the most work when link-state changes occur.
- **Number of Areas**—A router must run the link-state algorithm for each link-state change that occurs for every area in which the router resides. Every area border router is in at least two areas (the backbone and one area). To maximize stability, do not place a router in more than three areas.
- **Designated Router Selection**—In general, the designated router and backup designated router on a local-area network (LAN) have the most OSPF work to do. It is recommended to select routers that are not loaded with CPU-intensive activities to be the designated router and backup designated router.
- **OSPF Timers**—OSPF timers can be selectively configured to provide convergence as a differentiated service per tenant. This configuration is not recommended in a high scale environment because of the increase of control plane load on the CPU due to the high number of adjacencies inherent in the VMDC Multi-VRF topology. To minimize control plane load and provide faster convergence, deploy the default OSPF Hello and Hold timers and configure Bidirectional Forwarding Detection (BFD) to detect link failures.
- **OSPF Throttle Timers**—In Cisco NX-OS, the default SPF timers have been significantly reduced. Common deployments of NX-OS platforms are in a high-speed data center requiring fast convergence, as opposed to a wide area network (WAN) deployment with lower speed links where slower settings might still be more appropriate. To further optimize OSPF for fast convergence in the data center, manually tune the default throttle timers in NX-OS.

Bidirectional Forwarding Detection (BFD) for OSPF

BFD is a detection protocol that provides fast forwarding-path failure detection times for media types, encapsulations, topologies, and routing protocols. You can use BFD to detect forwarding path failures at a uniform rate, rather than the variable rates for different protocol hello mechanisms. BFD makes network profiling and planning easier, and it makes re-convergence time consistent and predictable.

BFD can be less CPU intensive than protocol hello messages because some of the BFD load can be distributed onto the data plane on supported modules. For example, if BFD with OSPF protocol is enabled between the pair of aggregation switches and a BFD neighbor session with its OSPF neighbor router goes down, BFD notifies the OSPF process that the BFD neighbor is no longer reachable. To reduce failure recovery times, OSPF removes the neighbor relationship to that router and looks for an alternative path without waiting for the hold timer to expire.

IP Route Summarization

Route summarization keeps routing tables small for faster convergence and better stability. In the data center hierarchical network, summarization can be performed at the data center core or the aggregation layer. Summarization is recommended at the data center core layer if it is dedicated and separate from the enterprise core. The objective is to keep the enterprise core routing table as concise and stable as possible to prevent route changes occurring elsewhere in the network from impacting the data center, and vice versa. Summarization is recommended at the data center aggregation layer, which is the OSPF area border router (ABR) of the pod.

To summarize the routes into and out of an OSPF area, Cisco VMDC uses NSSAs and summary ranges.

OSPF Not-So-Stubby Area

Not-so-stubby areas (NSSAs) are an extension of OSPF stub areas. Stub areas prevent the flooding of external link-state advertisements (LSAs) into NSSAs, relying instead on default routes to external destinations. NSSAs are more flexible than stub areas in that a NSSA can import external routes into the OSPF routing domain. The Compact Pod design uses NSSAs to limit the number of routes advertised from the aggregation layer to the sub-aggregation layer.

OSPF Summary Range

The OSPF area range command is used only on an ABR. In the Compact Pod design, it is used to consolidate or summarize routes advertised from the local OSPF area (network container) to the rest of the tenant core, campus, and WAN network. The result is that external to the local area a single summary route is advertised to other areas by the ABR.

Layer 2 Scalability

The maximum ICS Layer 2 scale depends on the choice of node at the aggregation layer. The base Compact ICS can support up to 64 servers; however, this number can scale to 256 servers for the Compact Pod without losing required functionality. The following recommendations should be considered at the aggregation layer:

- [Virtual PortChannel, page 2-28](#)
- [VLAN Scale, page 2-30](#)
- [MAC Address Scale, page 2-30](#)
- [vNICs per Distributed Virtual Switch, page 2-30](#)

Virtual PortChannel

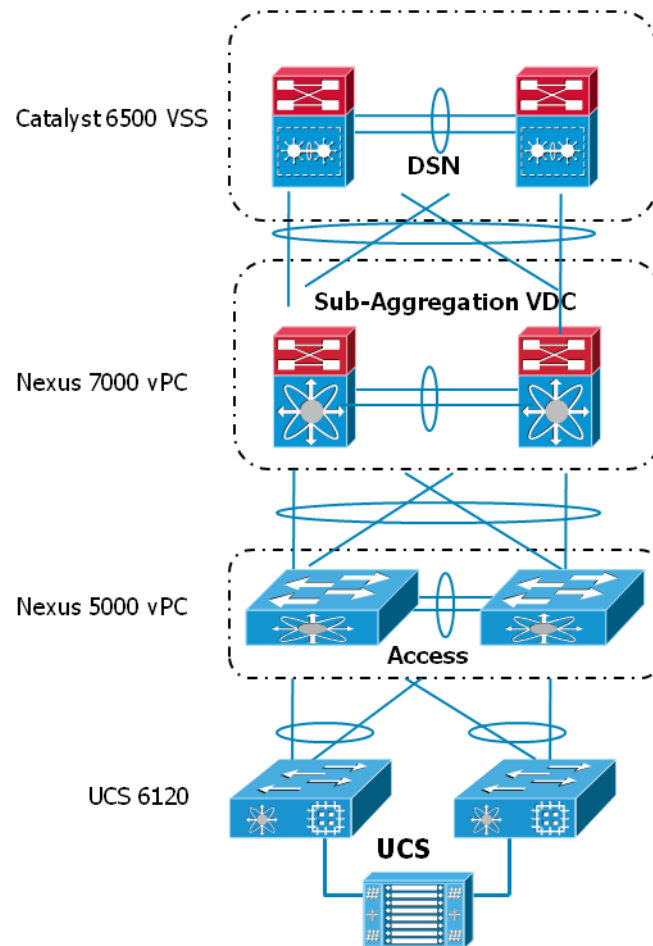
A virtual PortChannel (vPC) allows links that are physically connected to two Cisco Nexus devices to appear as a single PortChannel to any other device, including a switch or server. This feature is transparent to neighboring devices. A vPC can provide Layer 2 multipathing, which creates redundancy via increased bandwidth, to enable multiple active parallel paths between nodes and to load balance traffic where alternative paths exist.

When deployed either between access and aggregation layers or between the Cisco UCS and access layer devices in the Cisco VMDc design, a vPC provides the following benefits:

- Allows a single device to use a PortChannel across two upstream devices
- Eliminates Spanning Tree Protocol blocked ports
- Provides a loop-free topology
- Uses all available uplink bandwidth
- Provides fast convergence if either the link or a device fails
- Provides link-level resiliency
- Helps ensure high availability

Figure 2-9 presents a vPC deployment scenario. In this scenario, the Cisco UCS 6120s connect to the Cisco Nexus 5000 access layer switches, which connect to Cisco Nexus 7000 aggregation layer switches using a vPC link. This configuration makes all links active, and it achieves resiliency and high throughput without relying on STP to provide Layer 2 redundancy.

Figure 2-9 Compact Pod vPC Connectivity



For details on the vPC link concepts and use, refer to the following:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html#w p1053500

VLAN Scale

The VMDC Compact Pod architecture terminates server VLANs at the sub aggregation layer. The total number of VLANs in the VMDC Compact Pod sub aggregation/access layer was 180, defined as 6 Gold x 2 VLANs (12), 10 Silver x 4 VLANs (40), and 16 Bronze x 8 VLANs (128).

The Compact Pod architecture enables vMotion and Layer 2 clustering between ICSs because their VLANs extended to the Sub-aggregation Layer.

Cisco recommends using a distinctive range of VLANs for each ICS in the data center so they are uniquely identified and pruned from the trunks connected to non-local pods. In addition, only allow VLANs that require Layer 2 communication between ICSs and manually prune all other VLANs.

MAC Address Scale

When deploying a virtualized environment, Layer 2 network scale is determined by the number of VMs per server blade, which translates to the number of MAC addresses at each access and sub-aggregation switch on the network. Different vNICs with unique MAC addresses are required for each VM data and management networks, as well as NICs on the ESX host itself. Therefore, the VMDC Compact Pod solution assumes four MAC addresses per VM which translates to roughly 5,760 MAC addresses per Compact ICS.

The Nexus 7000 platform is validated for the role of aggregation/core device with a theoretical maximum of 128,000 MAC addresses.

The Nexus 5000 platform is validated for the role of access device with a theoretical maximum of 16,000 MAC addresses.

vNICs per Distributed Virtual Switch

Within VMware both the Nexus 1000v and UCSM are Distributed Virtual Switches (DVS). VMware supports up to 4,096 DVports per Distributed Virtual Switch regardless DVS type. In each Compact ICS, a total of 1,440 VMs were enabled with four virtual machine NICs (vNICs) each (5,760 MAC addresses). The number of vNICs differs based on the virtual switch used.

Nexus 1000V DVS

Each Nexus 1000v supports 2 Virtual Supervisor Modules and 64 Virtual Ethernet Modules. Each VEM can support up to 216 vEthernet ports (or vNICs) and the Nexus 1000v switch has a DVport limitation of 2,048. Each VMKernel and vNIC interface consumes a single DVport instance on the DVS.

UCSM DVS

When deploying the UCSM DVS with the M81KR VIC in Pass Through Switching (PTS) mode the limitation of vNICs needs to be monitored on a per adapter basis. Each M81KR VIC adapter has 128 virtual interfaces, or VIFs, available. When configuring for use with the UCSM the amount of usable VIFs is dropped to 64 since each VIF has the ability to failover to the redundant fabric interconnect. Each VMkernel and virtual machine NIC interface consumes one of these VIFs.

Compute Performance and Scalability

Compute performance is established by the underlying CPU and Memory speed and their architecture. Each UCS server's CPU performance is based on the type of installed Intel Xeon 5500, Xeon 5600, Xeon 6500 or Xeon 7500 Series processors. The different models of server also allow for different memory configurations, and each models scalability is dependant on the number of slots available. These

numbers range from 12 to 32 slots depending on model of UCS server. The memory available also ranges in speed and size with DIMMS operating at 800Mhz, 1066 MHz and 1333Mhz in density of 4 GB, 8 GB, and 16 GB.

UCS

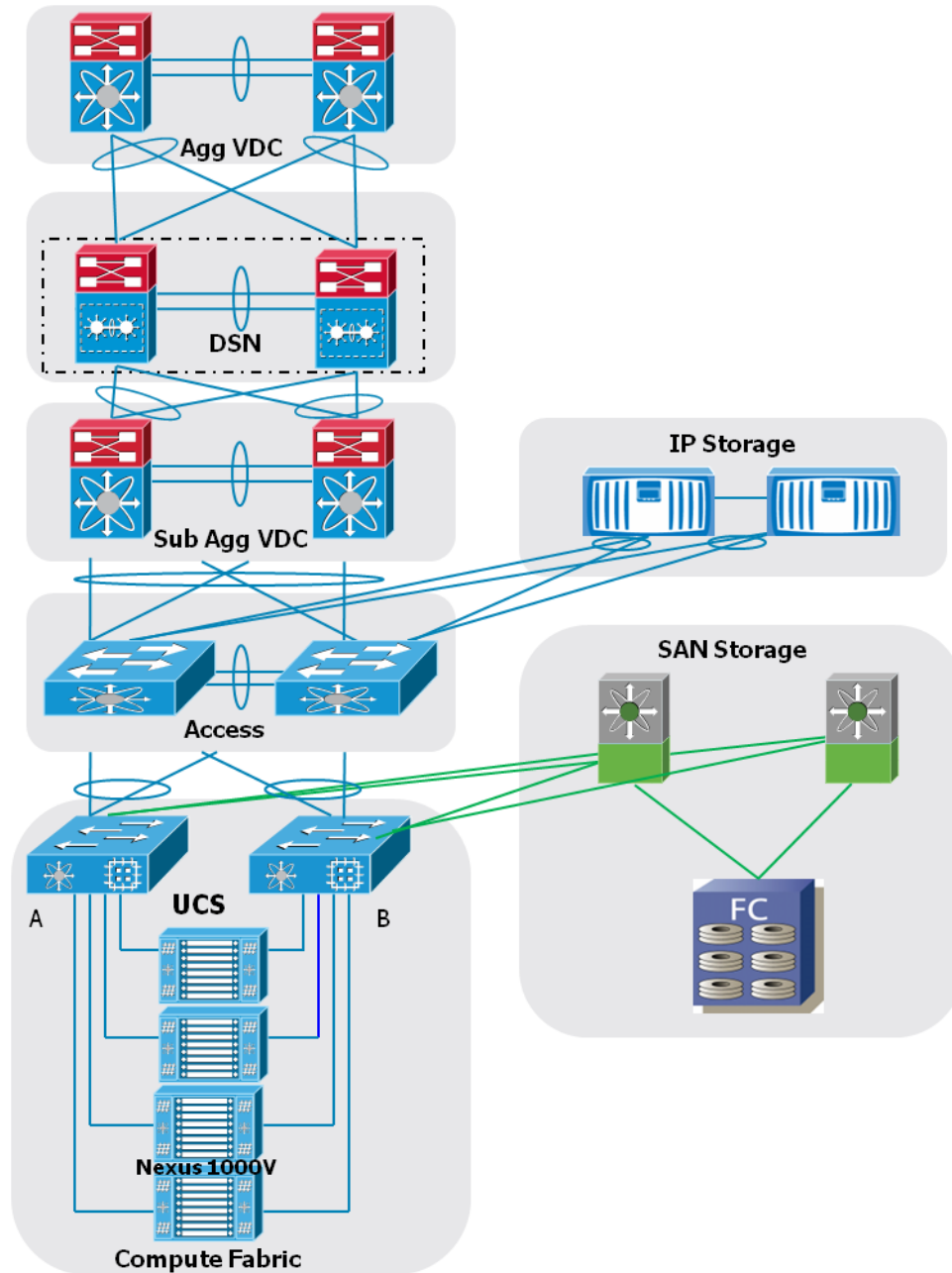
Each UCS 6120 has 20 ports and each UCS 6140 has 40 ports available to connect to servers/chassis if 4 of the onboard ports are used as uplinks. Each chassis connects to the UCS 6120 via 4 links for maximum bandwidth availability thus allowing for a maximum of 5 and 10 chassis to connect to the UCS 6120 and 6140, respectively. It should be noted that only a single link from each UCS chassis is required allowing the total numbers of chassis supported by each fabric interconnect to increase to a maximum of 20 and 40.

Each chassis can hold up to 8 half height server blades or 4 full height server blades thus allowing the environment to scale out based on the number of chassis connected to the fabric interconnects of the pod.

VMs per CPU Core

Server virtualization provides the ability to run multiple virtual servers in a single physical blade. Cisco UCS B Series blade servers has two CPU sockets with 4 cores per CPU for a total of 8 cores (8 vCPUs) per blade. The number of VMs allocated to each blade server depends on the application CPU and memory requirements. If we consider only a low-end server deployment where four VMs are allocated per core, we have 32 VMs per blade server, which totals 2,048 VMs across all 64 blade servers in a Compact Pod.

In the VMDC solution, a single Compact Pod ICS can scale to a maximum of 64 Cisco Unified Computing Services (UCS) servers. The Cisco VMDC design was validated at scale within the pod and from the ICS to the sub-aggregation layer.

Figure 2-10 Compact Pod

The VMDC solution addresses the general compute cloud data center deployments. [Table 2-7](#) identifies how many VMs were enabled during Compact Pod validation. As described in [Tiered Service Models, page 1-19](#), different workload requirements occur in a typical cloud model. In the VMDC architecture, we refer to the Small, Medium, and Large workload sizes. [Table 2-7](#) identifies how many workloads can be implemented on a Compact ICS with a workload mix of 50% Small, 30% Medium, and 20% Large.

Table 2-7 Cisco UCS Configuration by Tier

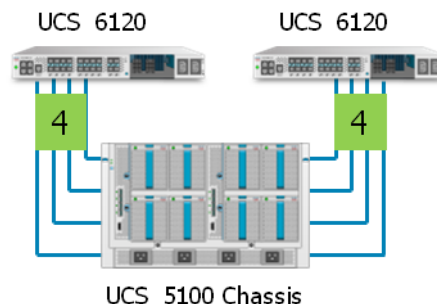
VM Size	Mix Ratio	Blades/Chassis	Cores/Blade	VMs /Core	Total VMs
Small Workload	50%	32	8	4	1,024
Medium Workload	30%	20	8	2	320
Large Workload	20%	12	8	1	96
Total	100%	64	—	—	1,440

In [Table 2-7](#), 20% of the available 64 Cisco UCS blade servers are reserved for Large workloads, which is 12 blade servers. A UCS blade server has two CPU sockets that accept a quad-core CPU; therefore, a total of 96 vCPUs are available to the Large workloads. Allocating one vCPU per large VM yields a total of 96 large workloads. If we calculate the same for each VM size, we find 1,440 general compute VMs available in a Compact Pod.

UCS 6120 Network Oversubscription

Network architects want non-blocking, line-rate bandwidth for each blade server. However to reduce costs, we often build in some level of network oversubscription. Network oversubscription is the level of bandwidth consolidation where the ingress bandwidth exceeds the egress bandwidth. We assume that all servers connected to the access layer switches will not transmit traffic simultaneously at line rate directed toward the access layer uplinks. Therefore, we can safely build in some level of oversubscription without creating a communication bottleneck among end nodes. Obviously, the applications deployed help determine what level of network oversubscription is acceptable. The VMDC solution primarily looks at the general compute deployment. A 4:1 oversubscription ratio is considered for server-to-server communication and client-to-server communication flows.

[Figure 2-11](#) represents a UCS chassis with four uplinks between each fabric extender and the fabric interconnect. It depicts 8 10-Gb uplinks available from each UCS chassis into the UCS fabric. Each UCS chassis contains up to 8 blades, which means each blade has 10-Gb bandwidth available for upstream traffic forwarding. Server virtualization enables multiple logical server instances within a single blade, which could increase the potential bandwidth on the network interface card of the blade. Each UCS B200 blade has 10-Gb bandwidth available; however, that is shared among the virtual servers enabled on the blade.

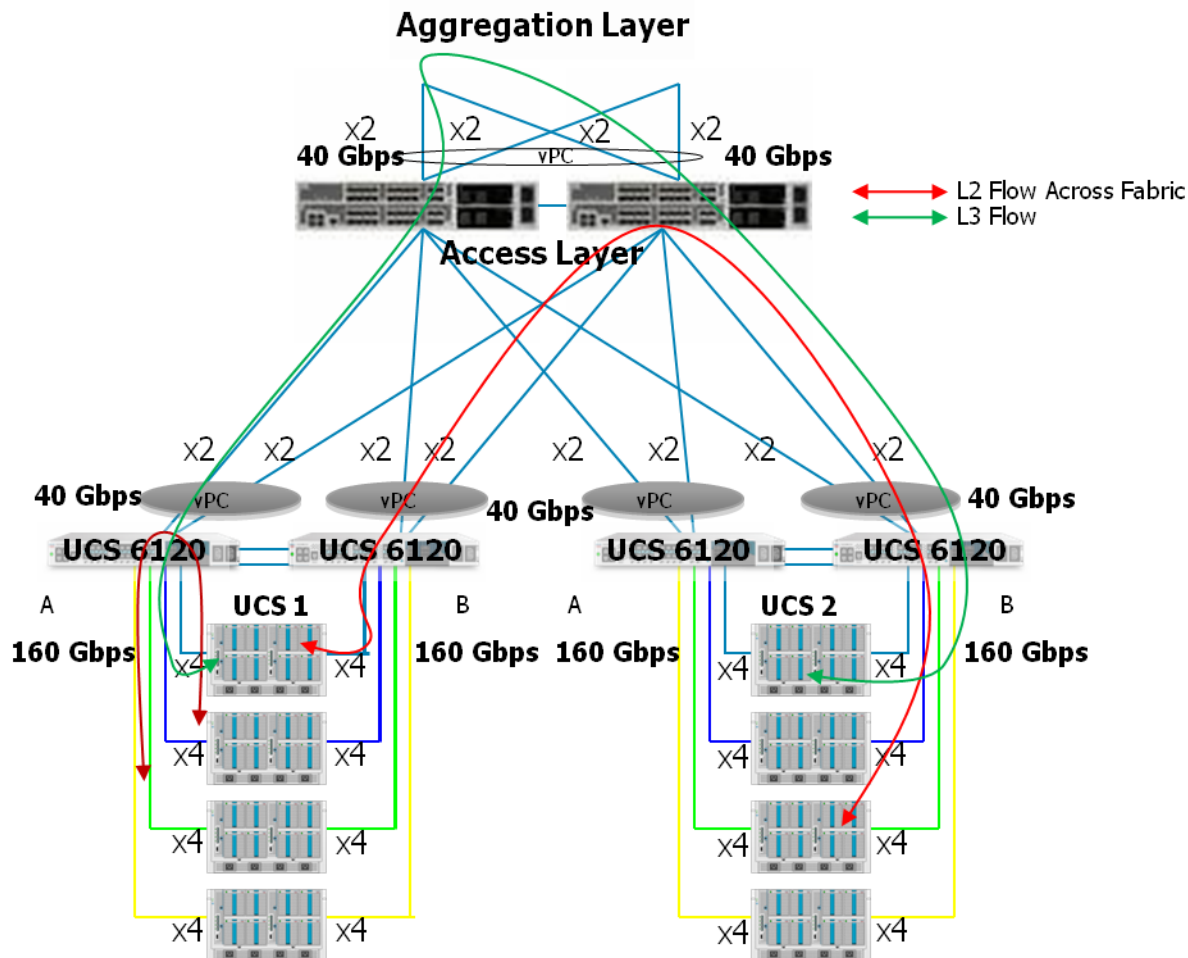
Figure 2-11 UCS Connectivity

Network architects must consider likely traffic flows within the logical topology that have been created on top of the physical topology. Multi-tier application flows create a portion of traffic that does not pass from the server farm to the aggregation layer. Instead, it passes directly between servers.

Application-specific considerations can affect the utilization of uplinks between switching layers. For example, if servers that belong to multiple tiers of an application are located on the same VLAN in the same UCS fabric, their traffic flows are local to the pair of UCS 6120s and do not consume uplink bandwidth to the aggregation layer. Some traffic flow types and considerations are as follows:

- **Server-to-server Layer 2 communications in the same UCS fabric.** Because the source and destinations reside within the UCS 6120 pair belonging to the same UCS fabric, traffic remains within the fabric. For such flows, 10 Gb of bandwidth is provisioned.
- **Server-to-server Layer 2 communications between different UCS fabrics.** As depicted in [Figure 2-12](#), the End-Host Ethernet mode should be used between the UCS 6120s (fabric interconnects) and aggregation layer switches. This configuration ensures that the existence of the multiple servers is transparent to the aggregation layer. When the UCS 6120s are configured in End-host mode, they maintain the forwarding information for all the virtual servers belonging to their fabric and perform local switching for flows occurring within their fabric. However, if the flows are destined to another pair of UCS 6120s, traffic is sent to the access layer switches and eventually forwarded to the servers by the correct UCS 6120.
- **Server-to-server Layer 3 communications.** If practical, you should keep multiple tiers of an application in the same UCS fabric to provide predictable traffic patterns. However, if the two tiers are on the same UCS fabric but on different VLANs, routing is required between the application tiers. This routing results in traffic flows to and from the aggregation layer to move between subnets.

Figure 2-12 Server-to-Server Traffic Flow Types



When deployed in a data center, the majority of traffic flows in a multi-tier application are inter-server. These traffic flows do not pass from the server farm toward the core. Instead, they occur server-to-server over a common fabric. For security purposes, multiple tiers of an application often belong to different VLANs. As such, network architects must consider the characteristics of the application and server architecture being deployed to determine a reasonable oversubscription rate in the network. In this VMDC design, a 4:1 network oversubscription for inter-server traffic is considered for general compute deployment (see [VMs per CPU Core](#), page 2-31).

This concept is illustrated in [Figure 2-12](#), where each UCS chassis 8 eight blades connected to the UCS 6120s using 40 Gb of bandwidth. When all 4 chassis are connected, 160 Gb of bandwidth is aggregated at each UCS 6120. Also, each UCS 6120 is configured in vPC host mode, and its 4 10-Gb uplinks form a port-channel where both links are forwarding to the access layer over 40 Gb of bandwidth. This configuration defines a ratio of 160 Gb /40 Gb, an oversubscription ratio of 4:1 at the access layer when all links are active. Similarly, the oversubscription ratio of 4:1 is provisioned at the aggregation layer when the all links are active.

There will be flows where external clients access the servers. This traffic must traverse the access layer switch to reach the UCS 6120. The amount of traffic that passes between the client and server is limited by the WAN link bandwidth availability. In metro environments, enterprises may provision between 10

and 20 Gb for WAN connectivity bandwidth; however, the longer the distance, the higher the cost of high bandwidth connectivity. Therefore, WAN link bandwidth is the limiting factor for end-to-end throughput.

Alignment of VM Partitions and VMFS to Storage Arrays

Virtual machines store their data on virtual disks. As with physical disks, these virtual disks contain storage partitions and file systems, which are created by the VM's guest operating system. In order to make sure of optimal disk I/O within the VM, you must align the partitions of the virtual disks to the block boundaries of VMFS and the block boundaries of the storage array. Failure to align all three of these items results in a dramatic increase of I/O load on a storage array and negatively affects the performance of all virtual machines being served on the array.

NetApp, VMware, other storage vendors, and VMware partners recommend that the partitions of VMs and the partitions of VMFS datastores are to be aligned to the blocks of the underlying storage array.

Storage Area Network Performance and Scalability

The capacity planning of the I/O subsystem is one of the most important planning steps—the I/O subsystem response time, throughput, and IOPS (I/Os per second) are critical to the overall work done by the application. Typically, the I/O subsystem is the slowest component of the computing environment. It needs to address multiple I/O profiles relating to numbers of I/O operations, I/O sizes, I/O latency, and total I/O throughput. All of these I/O characteristics are closely intertwined.

Port Density and Topology Requirements

The single most important factor in determining the most suitable SAN design is determining the number of end ports—both now and over the anticipated lifespan of the design. As an example, the design for a SAN that will handle a network with 100 end ports will be very different from the design for a SAN that has to handle a network with 1500 end ports.

From a design standpoint, it is typically better to overestimate the port count requirements than to underestimate them. Designing for a 1500-port SAN does not necessarily imply that 1500 ports need to be purchased initially, or even ever at all. It is about helping ensure that a design remains functional if that number of ports is attained, rather than later finding the design is unworkable. As a minimum, the lifespan for any design should encompass the depreciation schedule for equipment, typically three years or more. Preferably, a design should last longer than this, because redesigning and reengineering a network topology become both more time-consuming and more difficult as the number of devices on a SAN expands.

Where existing SAN infrastructure is present, determining the approximate port count requirements is not difficult. You can use the current number of end-port devices and the increase in number of devices during the previous 6, 12, and, 18 months as rough guidelines for the projected growth in number of end-port devices in the future.

For new environments, it is more difficult to determine future port-count growth requirements, but once again, it is not difficult to plan based on an estimate of the immediate server connectivity requirements, coupled with an estimated growth rate of 30% per year.

Traditionally as SANs grow, the switches required increases to accommodate the port count needed. This is particularly true in legacy bladecenter environments as each fibre channel I/O module would constitute another switch to be managed with its own security implications. Additionally, from a performance perspective, this is a concern as each switch or VSAN within an environment has its own domain ID, adding another layer of translation. N-port ID Virtualization or NPIV is a capability of the fibre channel

protocol that allows multiple N-ports to share a single physical port. NPIV is particularly powerful in large SAN environments as hosts that log into an NPIV-enabled device would actually be presented directly to the north-bound fabric switch. This improves performance and ease of management. NPIV is a component of the Fabric Interconnect within a UCS deployment and a requirement of any northbound FC switch.

The fan-in characteristics of a fabric is defined as the ratio of host ports that connect to a single target port while fan-out is the ratio of target ports or LUNs that are mapped to a given host. Both are performance indicators, with the former relating to host traffic load per storage port and the latter relating storage load per host port. The optimum ratios for fan-in and fan-out are dependent on the switch, storage array, HBA vendor, and the performance characteristics of I/O workload.

Device Performance and Oversubscription Ratios

Oversubscription is a necessity of any networked infrastructure and directly relates to the major benefit of a network-to share common resources among numerous clients. The higher the rate of oversubscription, the lower the cost of the underlying network infrastructure and shared resources. Because storage subsystem I/O resources are not commonly consumed at 100 percent all the time by a single client, a fan-out ratio of storage subsystem ports can be achieved based on the I/O demands of various applications and server platforms. Most major disk subsystem vendors provide guidelines as to the recommended fan-out ratio of subsystem client-side ports to server connections. These recommendations are often in the range of 7:1 to 15:1.

When considering all the performance characteristics of the SAN infrastructure and the servers and storage devices, two oversubscription metrics must be managed: IOPS and network bandwidth capacity of the SAN. The two metrics are closely related, although they pertain to different elements of the SAN. IOPS performance relates only to the servers and storage devices and their ability to handle high numbers of I/O operations, whereas bandwidth capacity relates to all devices in the SAN, including the SAN infrastructure itself. On the server side, the required bandwidth is strictly derived from the I/O load, which is derived from factors including I/O size, percentage of reads versus writes, CPU capacity, application I/O requests, and I/O service time from the target device. On the storage side, the supported bandwidth is again strictly derived from the IOPS capacity of the disk subsystem itself, including the system architecture, cache, disk controllers, and actual disks.

In most cases, neither application server host bus adapters (HBAs) nor disk subsystem client-side controllers are able to handle full wire-rate sustained bandwidth. Although ideal scenario tests can be contrived using larger I/Os, large CPUs, and sequential I/O operations to show wire-rate performance, this is far from a practical real-world implementation. In more common scenarios, I/O composition, server-side resources, and application I/O patterns do not result in sustained full-bandwidth utilization. Because of this fact, oversubscription can be safely factored into SAN design. However, you must account for burst I/O traffic, which might temporarily require high-rate I/O service. The general principle in optimizing design oversubscription is to group applications or servers that burst high I/O rates at different time slots within the daily production cycle. This grouping can examine either complementary application I/O profiles or careful scheduling of I/O-intensive activities such as backups and batch jobs. In this case, peak time I/O traffic contention is minimized, and the SAN design oversubscription has little effect on I/O contention.

Best-practice would be to build a SAN design using a topology that derives a relatively conservative oversubscription ratio (for example, 8:1) coupled with monitoring of the traffic on the switch ports connected to storage arrays and Inter-Switch Links (ISLs) to see if bandwidth is a limiting factor. If bandwidth is not the limited factor, application server performance is acceptable, and application performance can be monitored closely, the oversubscription ratio can be increased gradually to a level that is both maximizing performance while minimizing cost.

Control Plane Scalability

A SAN switch can be logically divided into two parts: a data plane, which handles the forwarding of data frames within the SAN; and a control plane, which handles switch management functions, routing protocols, Fibre Channel frames destined for the switch itself such as Fabric Shortest Path First (FSPF) routing updates and keepalives, name server and domain-controller queries, and other Fibre Channel fabric services.

Control plane scalability is the primary reason storage vendors set limits on the number of switches and devices they have certified and qualified for operation in a single fabric. Because the control plane is critical to network operations, any service disruption to the control plane can result in business-impacting network outages. Control plane service disruptions (perpetrated either inadvertently or maliciously) are possible, typically through a high rate of traffic destined to the switch itself. These result in excessive CPU utilization and/or deprive the switch of CPU resources for normal processing. Control plane CPU deprivation can also occur when there is insufficient control plane CPU relative to the size of the network topology and a network-wide event (for example, loss of a major switch or significant change in topology) occurs.

FSPF is the standard routing protocol used in Fibre Channel fabrics. FSPF automatically calculates the best path between any two devices in a fabric through dynamically computing routes, establishing the shortest and quickest path between any two devices. It also selects an alternative path in the event of failure of the primary path. Although FSPF itself provides for optimal routing between nodes, the Dijkstra algorithm on which it is commonly based has a worst-case running time that is the square of the number of nodes in the fabric. That is, doubling the number of devices in a SAN can result in a quadrupling of the CPU processing required to maintain that routing.

A goal of SAN design should be to try to minimize the processing required with a given SAN topology. Attention should be paid to the CPU and memory resources available for control plane functionality and to port aggregation features such as Cisco PortChannels, which provide all the benefits of multiple parallel ISLs between switches (higher throughput and resiliency) but only appear in the topology as a single logical link rather than multiple parallel links.

Ratio of IOPS to Block Size

For small block I/O applications, the critical factor is I/Os per second (IOPS), not bandwidth. Depending on the application block size, the same number of IOPS may have significantly different bandwidth requirements. For example, an application requiring 5000 IOPS with a 4KB block size results in a bandwidth requirement of about 20MB/sec. (5000 IOPS x 4KB blocks). In contrast, an application that uses 16KB blocks with the same number of IOPS needs significantly higher bandwidth: 5000 IOPS x 16KB blocks = 80MB/sec.

Consider that as frame size increases, the number of IOPS decrease and MB/s increases. Therefore you should see the best IOPS performance with small frame sizes and the best bandwidth (MB/s) performance with large frame sizes.

Thin Pool Write Rebalancing

Thin pool write rebalancing normalizes the used capacity levels of data devices within a virtual data pool after new data drives are added or existing data drives are drained. It is a background optimization task that scans the used capacity levels of the data devices within a virtual pool and perform movements of multiple track groups from the most utilized pool data devices to the least used pool data devices. The process can be scheduled to run only when changes to the virtual pool composition make it necessary and user controls exist to specify what utilization delta will trigger track group movement.

Zero Space Reclamation

Zero reclaiming allows data blocks that only contain zeros to become available in the thin pool of available data space to ensure the maximum amount of space is used before the need for adding physical disks. Zero space reclamation frees, or de-allocates, storage extents found to contain all zeros. You can reclaim both allocated/unwritten extents, as well as extents filled with host-written zeros within a thin pool. It is a non-disruptive process that can be executed with the targeted thing device ready and red/write to operating systems and applications.

N-Port ID Virtualization (NPIV)

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 are 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 be managed independently from the perspectives of zoning, aliasing, and security. In a Cisco MDS 9000 family environment, each host connection can log in as a single virtual SAN (VSAN).

N-Port Virtualizer (NPV)

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

Network Attached Storage Performance and Scalability

Traditional NAS systems have a capacity limit to the amount of file system space they can address. These systems also have a “head” or controller. These controllers traditionally have a limit to the performance they can achieve, dictated by the type and number of processors and cache used in each system. As unstructured data has grown, several approaches to overcoming these limitations of traditional NAS have evolved. One approach is to add acceleration hardware in front of NAS systems in the form of additional cache or to use NAND capacity as cache.

NetApp Flexcache Software

FlexCache software creates a caching layer in your storage infrastructure that automatically adapts to changing usage patterns, eliminating performance bottlenecks. The benefits are:

- Eliminates storage bottlenecks automatically, without tedious administration
- Improves read performance in your distributed application and testing environments
- Lets you simplify tiered storage

NetApp Flash Cache (PAM II) Modules

The NetApp Flash Cache (PAM II) modules improve performance for workloads that are random read intensive without adding more high-performance disk drives. Moreover, PAMII uses three caching modes and can be sized up to 512GB, so the application for PAMII is quite broad. The benefits are:

- Optimize performance at a lower cost
- Automatically tiers active data to higher performance storage
- Get the IO throughput, without impacting data center square footage with extra drives
- Grow capabilities, without impacting cooling output and power consumption
- Good for engineering applications, file services, databases, and virtual infrastructures
- Solid state Flash Cache modules use no additional rack space and consume 95% less power than a shelf of 15k RPM disk drives.

NetApp Deduplication

NetApp deduplication can be used broadly across many applications, including primary data, backup data, and archival data. NetApp deduplication combines the benefits of granularity, performance, and resiliency to provide you with a significant advantage in the race to provide for ever-increasing storage capacity demands.

Data deduplication helps control data proliferation. The average UNIX(r) or Windows(r) disk volume contains thousands or even millions of duplicate data objects. As data is created, distributed, backed up, and archived, duplicate data objects are stored unabated across all storage tiers. The end result is inefficient utilization of data storage resources.

By eliminating redundant data objects and referencing just the original object, an immediate benefit is obtained through storage space efficiencies. The following benefits result:

- **Cost Benefit**—Reduced initial storage acquisition cost, or longer intervals between storage capacity upgrades. primary data, backup data, and archival data can all be deduplicated with nominal impact on data center operations.
- **Management Benefit**—The ability to store “more” data per storage unit, or retain online data for longer periods of time. Select which datasets to deduplicate to evaluate those datasets and identify the areas that will provide the greatest return. Perform a full byte-for-byte validation before removing any duplicate data for worry-free deduplication.

NetApp Rapid Clone Utility

In Cisco VMDC, NetApp Rapid Clone Utility (RCU) 3.0 was used to clone and deploy the VMs. RCU 3.0 is a VMware vSphere plug-in that allows you to quickly and efficiently create, deploy and manage the lifecycle of virtual machines (VMs) from an easy-to-use interface integrated into VMware vCenter 4.0 and later.



Note

RCU 3.0 is a new release that is supported only with VMware vCenter Server 4.0 and later. VMware vCenter Server 2.5 installations should use Rapid Cloning Utility 2.1.

RUC can be used to:

- Create clones of templates, virtual machines, or virtual machine snapshots, and deploy them into new or existing NetApp NFS and VMFS (FCP/iSCSI) datastores
- Apply guest customization specifications to the resulting virtual machines
- Provision, resize, deduplicate, and deallocate datastores
- Deploy virtual machines for both server and desktop use
- Re-deploy virtual machines from a baseline image

- Monitor storage savings
- Import virtual machines into virtual desktop infrastructure connection brokers and management tools

Thin Provisioning

Thin provisioning the LUNs at the storage level enables efficient use of available space on the storage array and hot expansion of the storage array by simply adding data devices to the thin pool. Normally, when a 50 GB VMDK is created, it immediately eats up 50 GB of disk space on the Virtual Machine File System (VMFS) volume. Since application owners often demand more space than they truly need, there is a lot of expensive storage area network (SAN) disk capacity dedicated to these applications that will never be used. When you thin-provision a VMDK, storage is not allocated to the VMDK unless it is actually used. As long as only 10 GB of the allocated 50 GB disk is used, only 10 GB is claimed.

Thin provisioning, in a shared storage environment, is an optimized use of available storage. It relies on on-demand allocation of blocks of data versus the traditional method of allocating all the blocks up front. This methodology eliminates almost all whitespace which helps avoid the poor utilization rates, often as low as 10%, that occur in the traditional storage allocation method where large pools of storage capacity are allocated to individual servers but remain unused (not written to). This traditional model is often called “fat” or “thick” provisioning.

With thin provisioning, storage capacity utilization efficiency can be automatically driven up towards 100% with very little administrative overhead. Organizations can purchase less storage capacity up front, defer storage capacity upgrades in line with actual business usage, and save the operating costs (electricity and floorspace) associated with keeping unused disk capacity spinning. Previous systems generally required large amounts of storage to be physically preallocated because of the complexity and impact of growing volume (LUN) space. Thin provisioning enables over-allocation or over-subscription.

Service Assurance

Service assurance is generally defined as a set of procedures that optimize performance and provide management guidance in communications networks, media services, and end-user applications. Service assurance involves quality assurance, quality control, and service level management processes. Quality assurance and control processes ensure that a product or service meet specified requirements, adhering to a defined set of criteria that fulfill customer or client requirements. Service level management involves the monitoring and management of key performance indicators of a product or service. The fundamental driver behind service assurance is to maximize customer satisfaction.

Quality of Service

Today, SLAs often emphasize service availability. Differentiated service levels requirements exist because specific applications or traffic may require preferential treatment within the cloud. Some applications are mission critical, some are interactive, while others are bulk or utilized simply for dev-test purposes. In cloud context, service levels could be end to end, from cloud resources (hosts, datastores) to the end user. These service levels are embodied in the tenant subscription type (Gold, Silver, and Bronze) described in [Tiered Service Models](#), page 1-19.

Quality of Service functions are key to network availability service assurance because they enable differential treatment of specific traffic flows. This differentiated treatment ensures that in the event of congestion or failure conditions, critical traffic is provided sufficient amount bandwidth to meet throughput requirements. Traditionally, an SLA framework includes consideration of bandwidth, delay, jitter, and loss per service class.

The QoS features leveraged in this design are as follows:

- QoS classification and marking
- Traffic flow matching
- Bandwidth guarantees
- Rate limits

QoS Classification and Marking

The process of classification is one of inspecting different fields in the Ethernet Layer 2 header, along with fields in the IP header (Layer 3) and the TCP/UDP header (Layer 4), to determine the level of service that should be applied to the frame as it transits the network devices. The process of marking rewrites the COS in the Ethernet header or the Type of Service bits in the IPv4 header.

As per established best practices, classification and marking are applied at the network edge, close to the traffic source. In this design, the edge is represented by the Nexus 1000V virtual access switch for traffic originating from hosts and VMs and at the C6500 WAN edge for traffic entering the DC infrastructure from the public IP-NGN, Internet, or private WAN backbone. Additionally, the Nexus 5000 provides another edge for traffic originating from the attached NetApp NAS. The following code is an example of marking incoming control traffic for a CoS value of 6 at the Nexus 1000V:

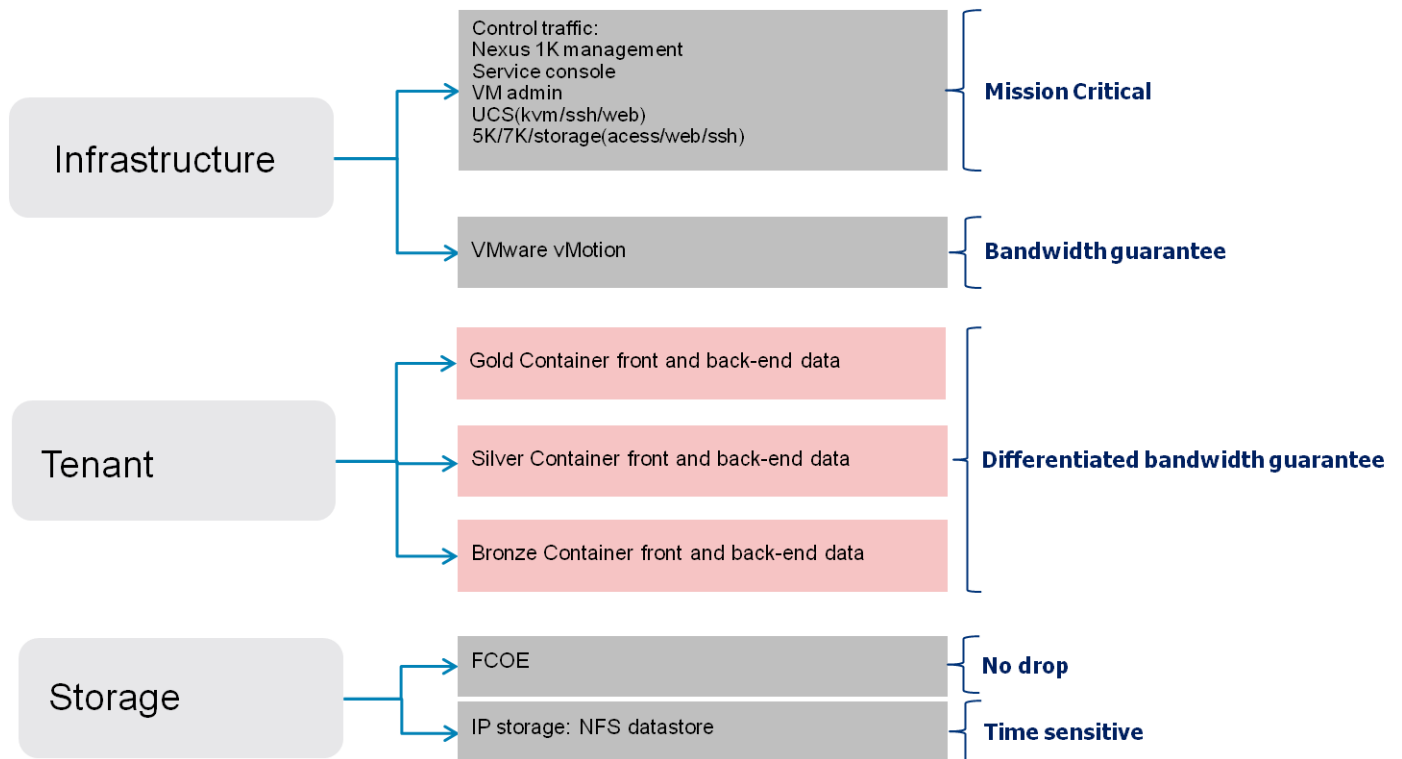
```
policy-map type qos mark-control-packet-vlans
  class n1k-control-vlan
    set cos 6
```

Figure 2-13 illustrates the traffic flow types defined in the VMDC architecture. These types are organized in infrastructure, tenant, and storage traffic categories.

- Infrastructure traffic comprises management and control traffic, including VMware service console and vMotion communication. This is typically set to the highest priority in order to maintain administrative communications during periods of instability or high CPU utilization.
- Tenant traffic is differentiated into Gold, Silver and Bronze service levels and may include VM to VM or VM to storage (back-end) traffic as well as VM to tenant (front-end) traffic. Gold tenant traffic is highest priority, requiring low latency and high bandwidth guarantees; Silver traffic requires medium latency and bandwidth guarantees; and Bronze traffic is delay-tolerant, requiring low bandwidth guarantees.
- The VMDC design incorporates both FC and IP-attached storage. As indicated below, storage requires two sub-categories, since these traffic types are treated differently through the network. FC traffic by definition requires a “no drop” policy, while NFS datastore traffic is sensitive to delay and loss.

When discussing QoS implementations in a cloud data center, you must consider an end-to-end perspective. Figure 2-13 presents the features required to implement QoS correctly.

Figure 2-13 Traffic Flow Types



Matching of Trusted Traffic Flows

Classification and marking of traffic flows creates a trust boundary within the network edges. Within the trust boundaries, received CoS or DSCP values are accepted and matched rather than remarked. For example, for a trusted control traffic flow of CoS 6, the classification process is a simple match of the received value:

```
class-map type qos control-sensitive-qos-class
  match cos 6-7
```

QoS Bandwidth Guarantees End to End

When a packet is ready to be switched to its next hop destination, the switch places the Ethernet frame into an appropriate outbound (egress) queue for switching. The switch performs buffer (congestion) management on this queue by monitoring the utilization. To provide differentiated treatment per defined traffic class in the event of buffer congestion, use the **service-policy** command to specify a minimum bandwidth guarantee and apply it to an interface, subinterface, or virtual circuit.

Rate Limiting

A bandwidth guarantee does not provide bandwidth reservation. If a particular traffic class is not using its configured bandwidth, any unused bandwidth is shared among the other classes. Specific traffic classes can be rate limited (policed) to ensure they do not starve the other classes of bandwidth. When

application throughput requirements are well understood, rate limiting can be used as a safety precaution, protecting the other classes from unexpected failures that lead to adversely high load for a particular class.

Table 2-8 presents the VMDC SLA framework for QoS. Implementation specifics vary due to differences in connectivity, interface types, and QoS scheduling and queueing capabilities across specific platforms in the infrastructure.

Table 2-8 VMDC Service Level Agreement Framework

Traffic Type	UCS Classes	Network Classes	CoS Marking (N1k/M81KR)	BW% UCS	BW% N5k Downlink	BW% N5k Uplink	BW% N7K	DCSP (N7k/C6k)	BW% C6k
Control - i.e, Nexus 1000V control/management, NFS Data store system control,	Platinum (UCS only)	Control	6	10%	10%	10%	10%	—	10%
vMotion	Silver	vMotion	4	—	20%	—	—	—	—
Gold Tenant (front/backend)/ IP Storage	Gold	Gold	5	40%	40%	40%	40%	40	40%
Silver Tenant (front/backend)	Bronze	Silver	2	15	15%	30%	30%	16	30%
Bronze Tenant (front/backend)	Best Effort	Bronze	0/1	0%	15%	20%	20%	0	20%
FCOE		FCOE	3	15%	—	—	—	—	—