Cisco Virtualized Multi-Tenant Data Center, Version 2.0
Compact Pod Design Guide

Last Updated: October 20, 2010

Cisco Validated Design

Cisco Virtualized Multi-Tenant Data Center, Version 2.0, Compact Pod Design Guide

CONTENTS

Preface

Revised: April 26, 2011

Overview

The Cisco Virtualized Multi-Tenant Data Center (VMDC) is a reference architecture that brings together core products and technologies from Cisco, NetApp, EMC, VMware, and BMC to deliver a comprehensive cloud solution. The Cisco VMDC 2.0 solution expands on the VMDC 1.1 solution. Key additions in 2.0 are integrated compute stacks as compute and storage building blocks and the validation of two scale points: Compact Pod and Large Pod. The solution also includes two additional implementation modules: a service portal and orchestration component and business continuance (disaster recovery) component. The VMDC 2.0 solution is described in a range of documents:

- Solution and Architecture Overview
- Design and Implementation Guides
  - Compact Pod Design
  - Compact Pod Implementation
  - Large Pod Design
  - Large Pod Implementation
  - Business Continuance Design
  - Service Orchestration Design

This preface contains the following topics:

Purpose of This Document

This document will help you design deployments of a private or public IaaS cloud data center based on the VMDC architecture. This Cisco-driven, end-to-end architecture defines how to create and manage flexible, dynamic pools of virtualized resources that can be shared efficiently and securely among different tenants. An orchestration solution creates a service portal that reduces resource provisioning and improves time-to-market (TTM) for IaaS-based services.

The Compact Pod design focuses on small to medium deployments with up to 32 tenants and up to 4,000 VM instances.

Audience

This document is intended for, but not limited to, network architects, systems engineers, field consultants, advanced services specialists, and customers who want to understand how to deploy a public or private cloud data center infrastructure.

Solution Objectives

The Cisco VMDC 2.0 solution addresses the following problems:

- **Resource utilization**. Traditionally, enterprises design their data centers using dedicated resource silos. These siloed resources include access switches, server racks, and storage pools assigned to specific applications and business units. This siloed approach results in inefficient resource use, where resource pools are customized per application, resulting in few shared resources. This design also cannot harness unused or idle resources, is complex to administer, and is difficult to scale, which results in longer deployment times. For enterprises and public cloud service providers, inefficient resource use increases capital and operational expenses, decreasing revenue margins.
- **Security guarantees**. In a multi-tenant environment, resource access must be controlled among tenants. This control is more challenging when resources are shared. Tenants need to be assured that their data and their applications are secure in highly virtualized systems.
- **Resource provisioning**. Manual provisioning often takes longer than four weeks to provision new resources. In many cases, this lengthy provision time fails to meet business agility and TTM requirements of enterprises and service providers.
- **Complex and expensive administration**. Today, network, server, security, and application administrators must collaborate to bring up new resources for each new or expanding tenant. In highly virtualized systems, collaboration based on manual methods does not scale, resulting in slow responses to business needs. It is complicated and time consuming to streamline tasks, such as manual configuration and resource provisioning. Also, resource churn increases capital and operating expenditures and overhead.

As enterprise IT departments evolve, they want a data center solution that supports rapid provisioning of resources that are efficiently shared and secured. Similarly, service providers want solutions that enable them to reduce TTM for new revenue-generating services and to improve ongoing operational expenses (OpEx). This document introduces an architecture that offers the flexibility to share common infrastructure among tenants while securely separating those tenants and data and enabling per-tenant differentiated services.

Related Documentation

The Cisco VMDC design recommends that general Cisco data center design best practices be followed as the foundation for IaaS deployments. The following Cisco Validated Design (CVD) companion documents provide guidance on such a foundation:

Data Center Design—IP Network Infrastructure

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

Data Center Service Patterns

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

Security and Virtualization in the Data Center

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

Designing Secure Multi-Tenancy into Virtualized Data Centers

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns743/ns1050/landing_dcVDDC.html

Enhanced Secure Multi-Tenancy Design Guide

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/Virtualization/securecldg_V2.html

The following VMDC solution document provide additional details on the solution:

Cisco VMDC 1.1 Design and Deployment Guide

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/vmdcDdg11.pdf

Cisco VMDC Solution Overview

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/solution_overview_c22-602978.html

Cisco VMDC Solution White Paper

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns836/white_paper_c11-604559.html

Vblock Infrastructure Solutions

http://www.acadia.com/solutions/vblock/index.htm

About Cisco Validated Designs

The Cisco Validated Design Program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit www.cisco.com/go/validateddesigns.

ALL DESIGNS, SPECIFICATIONS, STATEMENTS, INFORMATION, AND RECOMMENDATIONS (COLLECTIVELY, "DESIGNS") IN THIS MANUAL ARE PRESENTED "AS IS," WITH ALL FAULTS. CISCO AND ITS SUPPLIERS DISCLAIM ALL WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE. IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THE DESIGNS, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

THE DESIGNS ARE SUBJECT TO CHANGE WITHOUT NOTICE. USERS ARE SOLELY RESPONSIBLE FOR THEIR APPLICATION OF THE DESIGNS. THE DESIGNS DO NOT CONSTITUTE THE TECHNICAL OR OTHER PROFESSIONAL ADVICE OF CISCO, ITS SUPPLIERS OR PARTNERS. USERS SHOULD CONSULT THEIR OWN TECHNICAL ADVISORS BEFORE IMPLEMENTING THE DESIGNS. RESULTS MAY VARY DEPENDING ON FACTORS NOT TESTED BY CISCO.

CHAPTER 1

Architecture Overview

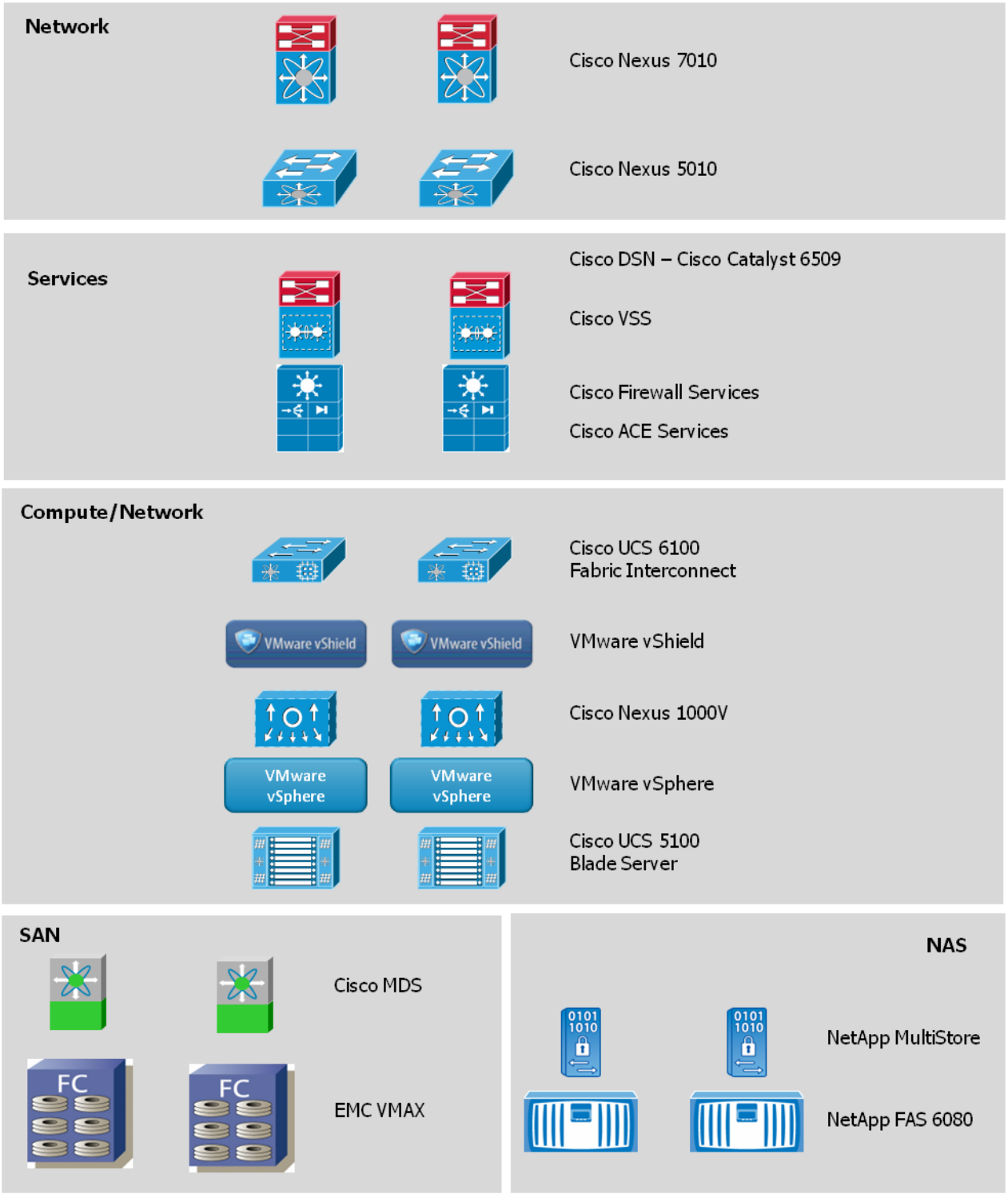
Revised: April 26, 2011

A cloud deployment model differs from traditional deployments in its ability to treat the data center as a common fabric of resources. A portion of these resources can be dynamically allocated and then de-allocated when they are no longer in use. Cisco VMDC leverages basic key building blocks:

- **Shared resource pools**. The resource pools consists of three main components: network, compute, and storage. Each of these components is virtualized so that each cloud tenant appears to have its own set of physical resources.
- **Service orchestration**. Service orchestration uses a set of tools and APIs to automate the provisioning process by using a predefined workflow. Service orchestration is presented as a web portal from which an end user can request specific resources from the cloud.
- **Business continuance**. Business continuity strives to ensure that essential functions can continue during and after a disaster. Business continuance planning seeks to prevent interruption of mission-critical services and to reestablish full functionality as swiftly and smoothly as possible.

Solution Key Components

Figure 1-1 Key Components of the Cisco VMDC 2.0 Solution, Compact Pod Design



Network

The following components were used in the network layer of the VMDC Compact Pod:

Cisco Nexus 7000

The Cisco Nexus 7000 Series is a modular switching system designed to deliver 10 Gigabit Ethernet and unified fabric. Designed for the core and aggregation layers of the data center, it delivers exceptional scalability, continuous operation, and transport flexibility.

It runs the Cisco NX-OS operating system (http://www.cisco.com/en/US/products/ps9372/index.html). For more information, see: http://www.cisco.com/en/US/products/ps9402/index.html.

Cisco Datacenter Services Node (DSN)

The DSN is an orderable option that includes a pair of Catalyst 6509-E chassis with the following modules:

Cisco Catalyst 6500 Virtual Switching System 1440

The Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 merges two physical Cisco Catalyst 6500 Series Switches into a single, logically managed entity. The key enabler of a VSS 1440 is the Virtual Switching Supervisor 720-10G. Once a VSS 1440 is created, it acts as a single virtual Catalyst switch delivering the following benefits:

- **Operational Manageability**. Two Catalyst 6500s share a single point of management, single gateway IP address, and single routing instance eliminating the dependence on First Hop Redundancy Protocols (FHRP) and Spanning Tree Protocols.
- **Availability**. Delivers deterministic, sub-200 millisecond Layer 2 link recovery through inter-chassis stateful failovers and the predictable resilience of EtherChannel.
- **Scalability**. Scales system bandwidth capacity to 1.4 Tbps by activating all available bandwidth across redundant switches.

The VSS platform supports Cisco integrated service modules, such as the Cisco Application Control Engine (ACE) and Firewall Services Module. It supports both 1- and 10-Gbps Ethernet devices allowing for network-based services.

Cisco Firewall Services Module (FWSM)

The Cisco Firewall Services Module (FWSM) is a stateful firewall residing within a Catalyst 6500 switching platform. The FWSM module supports device-level redundancy and scales through multiple virtual security contexts. A virtual security context can be transparent at Layer 2 or addressable at Layer 3. With either deployment model, the security policies associated with each virtual context are consistently applied to protect the related data center networks.

For more information, see:
http://www.cisco.com/en/US/products/hw/modules/ps2706/ps4452/index.html.

Cisco Application Control Engine (ACE)

The Cisco Application Control Engine (ACE) module performs server load balancing, network traffic control, service redundancy, resource management, encryption and security, and application acceleration and optimization. It provides device- and network service-level availability, scalability, and security features to the data center.

- The Cisco ACE offers the following device-level services:
- Physical redundancy with failover capabilities for high availability
- Scalability through virtualization allows ACE resources to be logically partitioned and assigned to meet specific tenant service requirements
- Security via access control lists and role-based access control

Network service levels support the following:

- Application availability through load balancing and health monitoring of the application environments
- Scalability of application load balancing, health monitoring, and session persistence policies as all are locally defined within each ACE virtual partition
- Security services, including ACLs and transport layer encryption (SSL/TLS) between the ACE virtual context, client population, and associated server farm

For more information, see:
http://www.cisco.com/en/US/products/ps5719/Products_Sub_Category_Home.html.

Cisco Nexus 5000

The Cisco Nexus 5000 Series (http://www.cisco.com/en/US/products/ps9670/index.html) switches deliver high performance, standards-based Ethernet and FCoE that enable the consolidation of LAN, SAN, and cluster network environments onto a single Unified Fabric. Each switch contains a single unified crossbar fabric ASIC and multiple unified port controllers to support fixed ports and expansion modules.

The unified port controller provides an interface between the unified crossbar fabric ASIC and the network media adapter and makes forwarding decisions for Ethernet, Fibre Channel, and FCoE frames. The ASIC supports the overall cut-through design of the switch by transmitting packets to the unified crossbar fabric before the entire payload has been received. The unified crossbar fabric ASIC is a single-stage, non-blocking crossbar fabric capable of meshing all ports at wire speed. The unified crossbar fabric implements QoS-aware scheduling for unicast and multicast traffic. Integration of the unified crossbar fabric with the unified port controllers ensures low latency lossless fabric for ingress interfaces requesting access to egress interfaces.

For more information, see: http://www.cisco.com/en/US/products/ps9670/index.html.

Cisco Nexus 1000V

The Nexus 1000V software switch delivers Cisco VN-Link services to virtual machines hosted on the server where the switch resides. Built on the VMware vSphere framework, it integrates server and network environments to ensure consistent, policy-based network capabilities to all servers in the data center. The Nexus 1000V aligns management of the operational environment for virtual machines and physical server connectivity in the data center, which enables a policy to follow a virtual machine during live migration, ensuring persistent network, security, and storage compliance.

For more information, see: http://www.cisco.com/en/US/products/ps9902/index.html.

For more information on Cisco VN-Link technologies see: http://www.cisco.com/en/US/netsol/ns894/index.html.

Compute

The following components were used in the network layer of the VMDC Compact Pod:

- Cisco UCS and UCSM, page 1-5
- VMware vSphere and vCenter Server, page 1-6
- VMware vShield Zones, page 1-6

Cisco UCS and UCSM

The Cisco Unified Computing System (UCS) unites compute, network, storage access, and virtualization into a cohesive system. The system integrates a low-latency, lossless 10-Gigabit Ethernet unified network fabric with x86-architecture servers. All integrated resources participate in a unified management domain, whether it has one or 320 servers with thousands of virtual machines. The Cisco UCS accelerates the delivery of new services via end-to-end provisioning and migration support for both virtualized and non-virtualized systems.

The Cisco Unified Computing System is built from the following components:

- Cisco UCS 6100 Series Fabric Interconnects (http://www.cisco.com/en/US/partner/products/ps10276/index.html) is a family of line-rate, low-latency, lossless, 10-Gbps Ethernet and Fibre Channel over Ethernet interconnect switches.
- Cisco UCS 5100 Series Blade Server Chassis (http://www.cisco.com/en/US/partner/products/ps10279/index.html) supports up to eight blade servers and up to two fabric extenders in a six rack unit (RU) enclosure.
- Cisco UCS 2100 Series Fabric Extenders (http://www.cisco.com/en/US/partner/products/ps10278/index.html) bring unified fabric into the blade-server chassis, providing up to four 10-Gbps connections each between blade servers and the fabric interconnect.
- Cisco UCS B-Series Blade Servers (http://www.cisco.com/en/US/partner/products/ps10280/index.html) adapt to application demands, intelligently scale energy use, and offer best-in-class virtualization.
- Cisco UCS B-Series Network Adapters (http://www.cisco.com/en/US/partner/products/ps10280/index.html) offer a range of options, including adapters optimized for virtualization, compatibility with existing driver stacks, or efficient, high-performance Ethernet.

- Cisco UCS Manager (http://www.cisco.com/en/US/partner/products/ps10281/index.html) provides centralized management capabilities for the Cisco Unified Computing System.

For more information, see: http://www.cisco.com/en/US/partner/netsol/ns944/index.html.

VMware vSphere and vCenter Server

VMware vSphere and vCenter Server provides centralized control and visibility at every level of the virtual infrastructure and provisions service delivery and application service agreements.

VMware vCenter Server provides a scalable and extensible platform that forms the foundation for virtualization management (http://www.vmware.com/solutions/virtualization-management/). VMware vCenter Server, formerly VMware VirtualCenter, centrally manages VMware vSphere (http://www.vmware.com/products/vsphere/) environments.

For more information, see http://www.vmware.com/products/.

VMware vShield Zones

VMware vShield Zones is a centrally managed, stateful, distributed virtual firewall bundled with vSphere 4.x, which takes advantage of ESXi host proximity and virtual network visibility to create security zones. By leveraging various VMware logical containers, it is possible to greatly reduce the number of rules required to secure a multi-tenant environment and therefore reduce the operational burden that accompanies the isolation and segmentation of tenants and applications. This new way of creating security policies closely ties to the VMware virtual machine objects and therefore follows the VMs during vMotion and is completely transparent to IP address changes and network re-numbering. Using vShield Zones within DRS (Distributed Resource Scheduler) clusters ensures secure compute load-balancing operations without performance compromise as the security policy follows the virtual machine.

In addition to being an endpoint and asset aware firewall, the vShield Zones contain microflow-level virtual network reporting that is critical to understanding and monitoring the virtual traffic flows and implement zoning policies based on rich information available to security and network administrators. This flow information is categorized into allowed and blocked sessions and can be sliced and diced by protocol, port and application, and direction and seen at any level of the inventory hierarchy. It can be further used to find rogue services, prohibited virtual machine communication, serve as a regulatory compliance visualization tool, and operationally to troubleshoot access and firewall rule configuration. Flexible user configuration allows role-based duty separation for network, security, and vSphere administrator duties.

The Flow Monitoring feature displays Allowed and Blocked network flows at application protocol granularity. This can be used to audit network traffic and as an operational troubleshooting tool.

For more information, see: http://www.vmware.com/products/vshield-zones/.

Storage

The following components were used in the storage layer of the VMDC Compact Pod:

- Cisco MDS 9513, page 1-7
- Cisco MDS 9134, page 1-7
- Cisco MDS 9148, page 1-7
- Cisco Management Interface, page 1-8

Cisco MDS 9513

The Cisco MDS 9513 Multilayer Director allows you to deploy high-performance SANs using a high-performance, protocol-independent switch fabric. It provides uncompromising high availability, security, scalability, ease of management, and transparent integration of new technologies for extremely flexible data center SAN solutions. The Cisco MDS 9513 is compatible with first-, second-, and third-generation Cisco MDS 9000 Family switching modules.

For more information, see: http://www.cisco.com/en/US/products/hw/ps4159/index.html.

Cisco MDS 9134

The Cisco MDS 9134 Multilayer Fabric Switch is a 1RU chassis that provides line-rate 4-Gbps and 10-Gbps ports. It expands from 24 to 32 ports in 8-port increments and optionally activates 2 10-Gbps ports. The Cisco MDS 9134 offers non-blocking architecture, with all 32 4-Gbps ports and the 2 10-Gbps ports operating at line rate concurrently.

The 10-Gbps ports support a range of optics for connection to the Cisco MDS 9000 family core using 10-Gbps Inter-Switch Link (ISL) connectivity. The Cisco MDS 9134 can also be stacked using copper CX4 X2 transceivers to cost effectively offer up to 64-port densities. The Cisco MDS 9134 supports quick configuration and task wizards that allow it to be deployed quickly and easily in networks of any size. Powered by Cisco MDS 9000 NX-OS/SAN-OS Software, it includes advanced storage networking features and functions and is compatible with Cisco MDS 9500 Series Multilayer Directors and Cisco MDS 9200 Series Multilayer Fabric Switches, providing transparent, end-to-end service delivery in core-edge deployments.

For more information, see: http://www.cisco.com/en/US/products/hw/ps4159/index.html.

Cisco MDS 9148

The Cisco MDS 9148 Multilayer Fabric Switch is a one rack unit (1RU) top-of-rack (ToR) chassis that provides 48 line-rate 8-Gbps ports for storage networking deployments. It can expand from 16 to 48 ports in 8-port increments. The Cisco MDS 9148 delivers a non-blocking architecture, with all 48 1/2/4/8-Gbps ports operating at line-rate concurrently.

The Cisco MDS 9148 supports the Cisco Device Manager Quick Configuration Wizard, which allows it to be deployed quickly and easily in networks of any size. Powered by Cisco MDS 9000 NX-OS Software, it includes advanced storage networking features and functions and is compatible with Cisco MDS 9500 Series Multilayer Directors and Cisco MDS 9200 and other 9100 Series Multilayer Fabric Switches, providing transparent, end-to-end service delivery in core-edge deployments.

For more information, see: http://www.cisco.com/en/US/products/hw/ps4159/index.html.

Cisco Management Interface

The following Cisco management interfaces were used in the storage layer of the VMDC Compact Pod:

- Cisco Device Manager, page 1-8
- Cisco Fabric Manager, page 1-8

Cisco Device Manager

Device Manager is a management solution for Cisco MDS 9000 Family switch chassis. It graphically depicts installed switching modules, the supervisor modules, the status of each port within each module, the power supplies, and the fan assemblies. Device Manager provides two views, Device View and Summary View. Use Summary View to monitor interfaces on the switch. Use Device View to perform the following switch-level configurations:

- Configure zones for multiple VSANs
- Manage ports, port channels, and trunking
- Manage SNMPv3 security access to switches
- Manage CLI security access to the switch
- Manage alarms, events, and notifications
- Save and copy configuration files and software image
- View hardware configuration
- View chassis, module, port status, and statistics

Cisco Fabric Manager

Fabric Manager is a management solution for the MDS family of switches, the Nexus 5000 SAN features, and the UCS Fabric Interconnect with limited support. It provides a robust centralized management station for SAN and unified fabric-enabled devices such as the MDS family of switches and the Nexus 5000. Using Fabric Manager, you can perform the tasks needed during a device's deployment cycle, such as discovery, inventory, configuration, performance monitoring, and troubleshooting.

The tables in the Fabric Manager Information pane correspond to dialog boxes in Device Manager. While Device Manager shows values for a single switch, Fabric Manager shows values for multiple switches. However, for verifying or troubleshooting device-specific configuration, Device Manager provides more detailed information than Fabric Manager.

For more information, see:
http://www.cisco.com/en/US/partner/docs/switches/datacenter/mds9000/sw/5_0/configuration/guides/fund/fm/fmfund_5_0_1.html.

EMC Symmetrix VMAX

EMC Symmetrix VMAX provides high-end storage for the virtual data center. It scales up to 2 petabyte (PB) of usable protected capacity and can be deployed with Flash Drives, Fibre Channel, and Serial Advanced Technology Attachment (SATA) drives, with tiering fully automated with FAST.

For more information, see: http://www.emc.com/products/detail/hardware/symmetrix-vmax.htm.

EMC Powerpath VE

With PowerPath/VE, you can standardize path management across heterogeneous physical and virtual environments. PowerPath/VE enables you to automate optimal server, storage, and path utilization in a dynamic virtual environment. This automation eliminates the need to manually load-balance hundreds or thousands of virtual machines and I/O-intensive applications in hyper-consolidated environments.

For more information, see: http://www.emc.com/products/detail/software/powerpath-ve.htm.

EMC Timefinder

EMC TimeFinder provides local storage replication for increased application availability and faster data recovery.

For more information, see: http://www.emc.com/products/detail/software/timefinder.htm

NetApp FAS6080 Filer

The NetApp FAS6080 provided Enterprise Class Network Attached Storage (NAS) Solution over fully redundant 10 Gigabit and Gigabit Ethernet LANs.

The NetApp FAS6080 Filer system is leveraged for this solution. Through NFS, customers receive an integration of VMware virtualization technologies with WAFL, NetApp's advanced data management and storage virtualization engine. This integration provides transparent access to VM level storage virtualization offerings, such as production-use data deduplication, immediate zero-cost VM and datastore clones, array-based thin provisioning, automated policy-based datastore resizing, and direct access to array-based Snapshot copies.

For more information, see http://www.netapp.com/us/products/storage-systems/fas6000/fas6000.html.

NetApp Snapshot

Snapshot creates point-in-time copies of file systems, which you can use to protect data-from a single file to a complete disaster recovery solution. It supports up to 255 Snapshot copies per volume to create online backups for user-driven recovery.

For more information, see h.ttp://www.netapp.com/us/products/platform-os/snapshot.html.

NetApp Deduplication

NetApp Deduplicaton can be leveraged to oversubscribe of real data storage.

For more information, see http://www.netapp.com/us/products/platform-os/dedupe.html.

NetApp Rapid Clone Utility

NetApp Rapid Clone is a plug-in for VMware vSphere and supports cloning and provisioning of virtual machines.

For more information, see http://blogs.netapp.com/virtualization/2010/02/rcu-30-now-available.html

Service Orchestration

Service orchestration is multi-domain configuration abstraction layer that manages the data center infrastructure. This abstraction layer enables a service catalog/portal-based configuration interface, in which the customer subscribing (application hosting community) to the infrastructure can pick from a limited number of customized service options, and host/place applications as virtual machines. Based upon these picks, configuration actions are executed across multiple domains, and to the device(s) within these domains, that together make up the service as represented within the customer facing portal.

Orchestration (integration across the domain tools) is fundamental as there is no single tool, within the Data Center that can configure the bundled services presented within the service catalog end-to-end. Orchestration coordinates the configuration requirements on top of the domain tools, and insures that all of the services defined within the service catalog/portal, are appropriately sequenced and correctly executed within each specific domain. Moreover, orchestration aggregates all of the individual service components within the service catalog, as a total services pool, and determines if there are sufficient resources across all of the components, to provide the service. The tools required for service orchestration include the following:

- Portal and service catalog (IT service management)
- Configuration management database
- Orchestration (runbook automation)
- Virtualized server provisioning and resource management
- Network provisioning and resource management
- Storage provisioning and resource management

In the the service orchestration layer of VMDC Compact Pod, the following components were used:

- BMC Atrium Orchestrator, page 1-10
- BMC Remedy AR System, page 1-10
- BMC BladeLogic Network Automation, page 1-11
- BMC BladeLogic Server Automation Suite, page 1-11

BMC Atrium Orchestrator

BMC Atrium Orchestrator automates manual tasks. Workflows based on ITIL standards can be built and adapted to match your processes, with components selected from a library of operator actions and workflow templates.

For more information, see:
http://www.bmc.com/products/product-listing/90902406-157022-1134.html.

BMC Remedy AR System

BMC Remedy AR System enables you to automate a broad range of business solutions, from service desk call tracking to inventory management to integrated systems management without learning a programming language or complex development tools. It also acts as a single point of integration, including support for popular API types (such as Java and C), Web Services, ODBC, and utilities such as the BMC Atrium Integration Engine.

For more information see: http://www.bmc.com/products/product-listing/22735072-106757-2391.html.

BMC BladeLogic Network Automation

Using BMC BladeLogic Network Automation, you can implement policy-based automation for managing networks, combining configuration management with compliance assurance. Supported by a robust security model, this network automation solution enables organizations to dramatically reduce operational costs, improve operational quality, and achieve operational compliance. BMC BladeLogic Network Automation automates common tasks of device management, including the following:

- Quick, non-disruptive configuration changes
- Proactive assessment of changes and enforcement of configuration standards
- Rapid deployment of devices from predefined templates
- Simplify provisioning of Service Profiles with Network Containers
- Document planned, unplanned, and unauthorized network changes
- On-demand compliance and Key Performance Indicator reporting

For more information, see:
http://www.bmc.com/products/product-listing/BMC-BladeLogic-Network-Automation.html.

BMC BladeLogic Server Automation Suite

BMC BladeLogic Server Automation Suite enables customers to manage server and application lifecycle events-including Discovery, Inventory, Provisioning, Configuration, Change Control, and Continual Compliance. The BMC solution addresses three functional areas:

- **Configuration**. Configuration management tasks often make up the bulk of the activities performed in a data center-patching, configuring, updating, and reporting on servers, across multiple platforms. The BMC solution, by shielding users from underlying complexity, enables consistency in change and configuration management activities. At the same time, subject to security constraints, it exposes sufficient detail about servers under management to ensure effective and accurate administrative activities.
- **Compliance**. Most IT organizations are required to maintain their server configurations in compliance with some sort of policy-whether regulatory (such as SOX, PCI, or HIPAA), security, or operational. BMC BladeLogic Server Automation achieves and maintains compliance by defining and applying configuration policies. Then, it provides detailed reports on how well servers comply with these policies. If a server or application configuration deviates from policy, the remediation instructions are generated and packaged, and can be either automatically or manually deployed to the server. All operations performed on servers are constrained by the appropriate set of policies, ensuring that servers stay in compliance throughout configuration changes, software deployments, and patches.
- **Provisioning**. BMC BladeLogic Server Automation automates the OS installation and configuration for both physical and virtual servers-delivering rapid, consistent, and reliable server provisioning processes, and ensuring that all servers are set up in compliance with configuration policies.

For more information, see:
http://www.bmc.com/products/product-listing/BMC-BladeLogic-Server-Automation-Suite.html.

Business Continuance

VMware vCenter Site Recovery Manager (SRM) 4.0 provides business continuity and disaster recovery protection for virtual environments. Protection can extend from individual replicated datastores to an entire virtual site.

In a Site Recovery Manager environment, there are two sites involved-a protected site and a recovery site. Protection groups that contain protected virtual machines are configured on the protected site and can be recovered by executing the recovery plans on the recovery site.

Site Recovery Manager leverages array-based replication between a protected site and a recovery site. The workflow that is built into Site Recovery Manager automatically discovers datastores setup for replication between the protected and recovery sites. Site Recovery Manager provides protection for the operating systems and applications encapsulated by virtual machines running on a VMware ESX host. A Site Recovery Manager server must be installed both at the protected and recovery site. The protected and recovery sites must each be managed by their own vCenter Server.

Furthermore, VMware vCenter Site Recovery Manager 4.0 supports VMware vSphere, shared recovery site, and NFS.

The following components were used to provide business continuance for the VMDC Compact Pod:

- VMware Site Recovery Manager, page 1-12

VMware Site Recovery Manager

VMware vCenter Site Recovery Manager is a business continuity and disaster recovery solution that helps you plan, test, and execute a scheduled migration or emergency failover of vCenter inventory from one site to another. It provides the following features:

Disaster Recovery Management

- Discover and display virtual machines protected by storage replication using integrations certified by storage vendors
- Create and manage recovery plans directly from vCenter Server
- Extend recovery plans with custom scripts
- Monitor availability of the remote site and alert users of possible site failures
- Store, view, and export results of test and failover execution from vCenter Server
- Control access to recovery plans with granular role-based access controls

Non-Disruptive Testing

- Use storage snapshot capabilities to perform recovery tests without losing replicated data
- Connect virtual machines to an existing isolated network for testing purposes
- Automate execution of recovery plans
- Customize execution of recovery plans for testing scenarios
- Automate cleanup of testing environments after completing failover tests

Automated Failover

- Initiate recovery plan execution from vCenter Server with a single button
- Automate promotion of replicated datastores for use in recovery scenarios with adapters created by leading storage vendors for their replication platforms
- Execute user-defined scripts and halts during recovery
- Reconfigure virtual machines' IP addresses to match network configuration at failover site
- Manage and monitor execution of recovery plans within vCenter Server

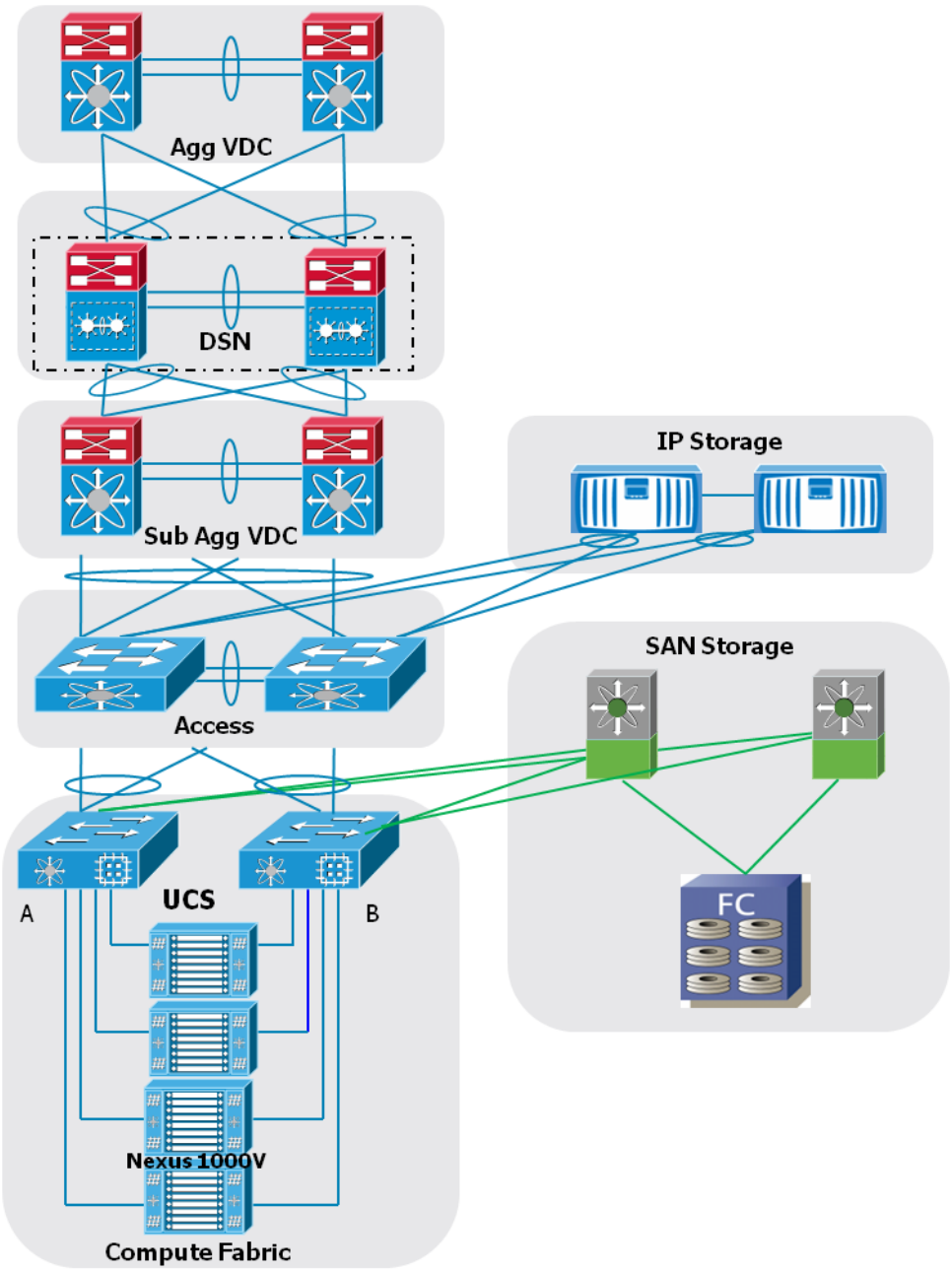
Modular Building Blocks

To scale the data center, Cisco VMDC defines two points of scale: the Point of Delivery and the integrated compute stack. Both repeatable building blocks provide for incremental growth to meet demands. This section defines how these building blocks relate to each other and the data center core and explains how they scale various resources. It contains the following topics:

Pod

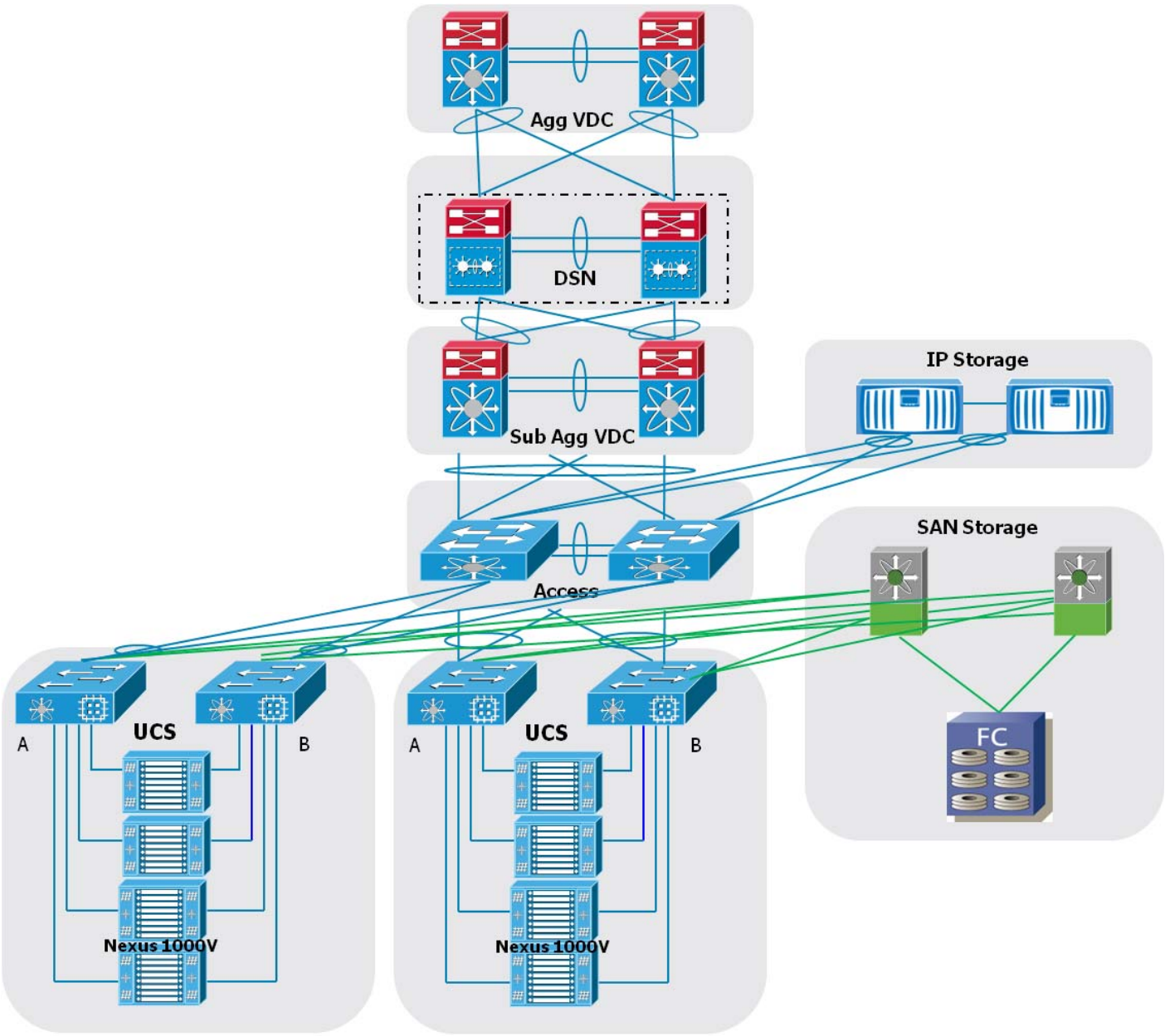
A pod identifies modular unit of data center components. This modular architecture provides a predictable set of resource characteristics (network, compute, and storage resource pools, power, and space consumption) per unit that is added repeatedly as needed. In this discussion, the aggregation layer switch pair, services layer nodes, and one or more integrated compute stacks are contained within a pod (see Figure 1-2).

Figure 1-2 Pod Components



To scale a pod, customers add additional integrated compute stacks (see Figure 1-3). You can continue to scale in this manner until the pod resources are exceeded.

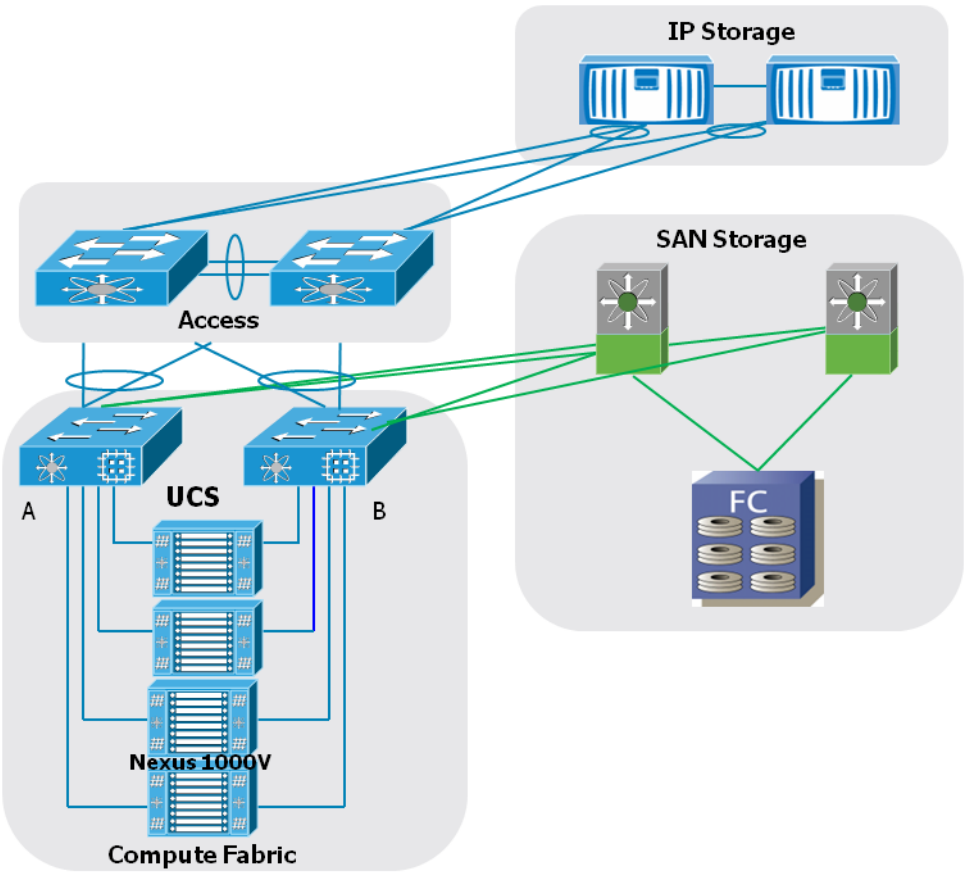
Figure 1-3 Expanding a Pod with Integrated Compute Stacks



Integrated Compute Stack (ICS)

An integrated compute stack can include network, compute, and storage resources in a repeatable unit (see Figure 1-4). In this discussion, the access layer switch pair, storage, and compute resources are contained within an integrated compute stack.

Figure 1-4 Integrated Compute Stack Components



Multi-Tenant Concepts

Multi-tenancy refers to the logical division of a shared pool of network, compute, and storage resources among multiple groups. Cisco VMDC relies on key concepts to deliver a solution that meets the requirements of these groups. This section explains the specific interpretation of multi-tenancy in the VMDC solution with the following topics:

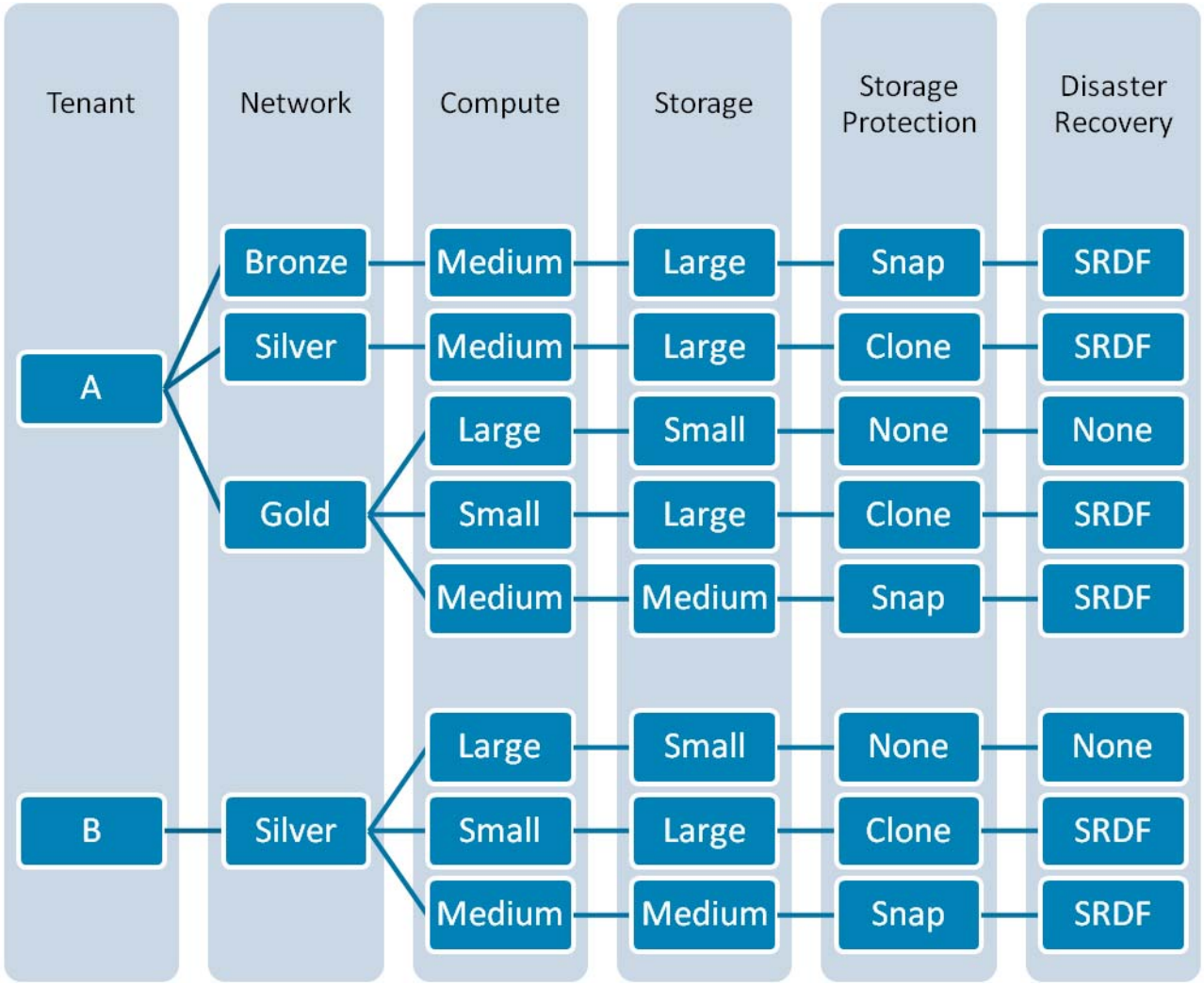
Tenant Defined

In the enterprise private cloud deployment model, the tenant is referenced as a department or business unit, such as engineering or human resources. In the public cloud deployment model, a tenant is an individual consumer, an organization within an enterprise, or an enterprise subscribing to the public cloud services. In either model, each tenant must be securely separated from other tenants because they share the virtualized resource pool.

When a tenant deploys an application or adds a new application, they first select a network container, then a VM size, a storage allocation, a storage protection, and a disaster recovery tier that meets the requirements of the application.

As shown in Figure 1-5, a tenant can select among multiple attributes to define the performance and operation of a virtual server.

Figure 1-5 Tenants and Virtual Servers



Differentiated Services

The cloud is the source of highly scalable, efficient, and elastic services accessed on-demand over the Internet or intranet. In the cloud, compute, storage, and network hardware are abstracted and delivered as a service. End users only consider the functionality and value provided by the service; they do not need to manage the underlying technology. Cloud services are differentiated at three layers in the VMDC solution: network, compute, and storage.

Network Layer

- **Application tiers**. Service tiers can provide differentiated support for application hosting. In some instances, applications may require several application tiers of VMs. For example, a Gold profile could have three application tiers to host web, application, and database (DB) services on different VMs and VLANs. Each tier could provide multiple VMs each for redundancy and provide load balancing. A Silver profile could also have three tiers for web, application, and DB services, but each tier might have multiple VMs on the same VLAN for redundancy and load balancing. A Bronze profile could have three tiers but with the web, application, and DB services residing on the same VM and VLAN.
- **Stateful services**. Customer or employee workloads can also be differentiated by the services applied to each tier. These services can be firewalls, encryption, load balancers, protocol optimization, application firewalls, WAN optimization, advanced routing, redundancy, disaster recovery, and so on. Within a service like firewalls, you can further differentiate among tiers as with inter-VLAN, intra-VLAN, or intra-host inspections. For example, a Gold tier might include firewall inspection, SSL off loading, IPSec encryption, server load balancing, and WAN optimization. A Silver tier might offer only firewall inspection and server load balancing.
- **Quality of Service agreements**. Bandwidth control during periods of network congestion can be key to managing application response time. QoS policies can prioritize bandwidth by service tier. Traffic classification, prioritization, and queuing and scheduling mechanisms can identify and offer minimum bandwidth guarantees to tenant traffic flows during periods of congestion. For example, a Gold service tier might be given the highest priority and a minimum network bandwidth guarantee of 50%. A Bronze service tier might receive best-effort treatment only and no minimum bandwidth guarantees.

Compute Layer

- **Virtual servers**. Typically, cloud providers want to offer multiple service tiers and provide different service level agreements (SLAs). Cloud services can be differentiated into predefined service tiers by varying resource allocation: virtual machine resources. Service profiles can vary based on the size of specific virtual machine (VM) attributes, such as CPU, memory, and storage capacity. Service profiles can also be associated with VMware Distributed Resource Scheduling (DRS) profiles to prioritize specific classes of VMs. For example, a Gold service can consist of VMs with dual core 3-GHz virtual CPU (vCPU), 8 GB of memory, and 500 GB of storage. A Bronze service can consist of VMs with a single core 1.5 GHz vCPU, 2 GB of memory, and 100 GB of storage.

Storage Layer

- **Storage allocation**. Applications require various amounts of disk space to operate. The ability to tune that allocation ensures that applications are not over or under provisioned, which uses resources more intelligently.
- **Storage protection and disaster recovery**. To meet datastore protection, recovery point, or recovery time objectives, service tiers can vary based on provided storage features, such as RAID levels, disk types and speeds, and backup and snapshot capabilities. For example, a Gold service could offer three tiers of RAID-10 storage using 15K rpm Fibre Channel (FC), 10K rpm FC, and SATA drives. While a Bronze service might offer a single RAID-5 storage tier using SATA drives.

The VMDC solution defines options for differentiating levels of resource allocation within the cloud. In this reference architecture, the different levels of services are described as services tiers. Each service tier includes a different set of resources from each of the groups: compute, storage, and network.

Tiered Service Models

The Cisco VMDC architecture allows providers to build service-level agreements (SLAs) that support their tenant or application requirements. The following example is not meant to be a strict definition resource allocation scheme, but to simply demonstrate how differentiated service tiers could be built.

In the Cisco VMDC Compact Pod design we define three service tiers: Gold, Silver, Bronze.

Each service tier is a container that defines different network, compute and storage service levels (see Table 1-1).

Table 1-1 Example Network and Data Differentiations by Service Tier

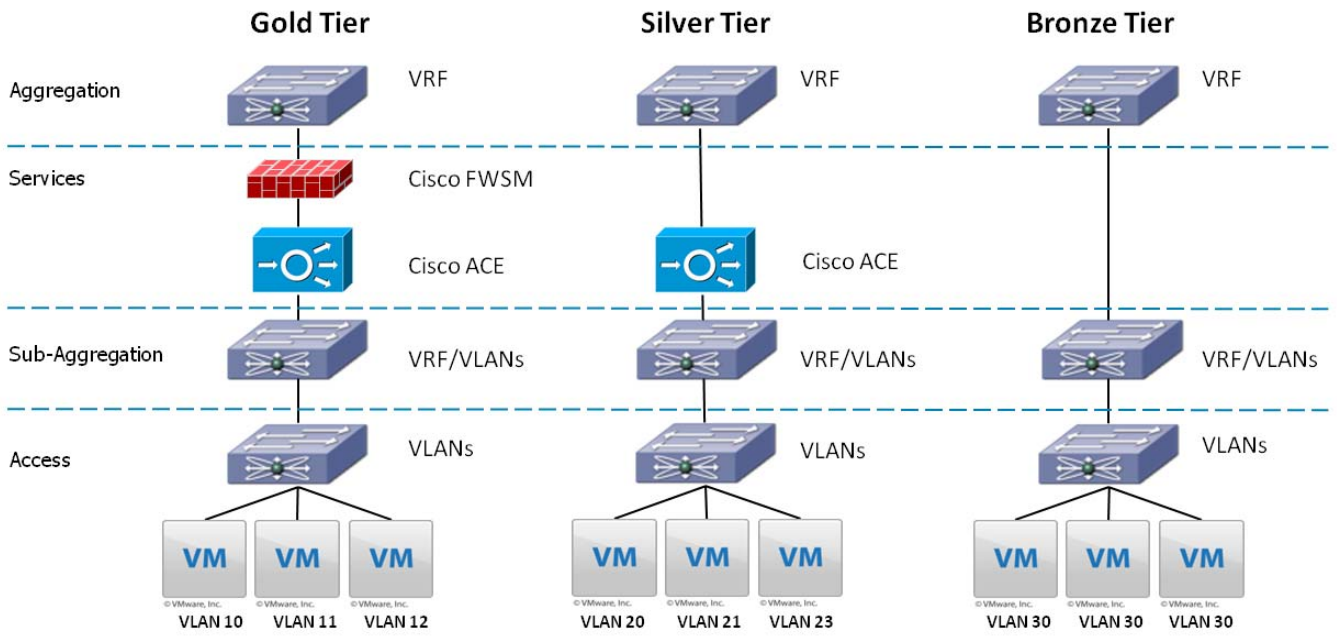
	Gold	Silver	Bronze
Services	Firewall and Load Balancing Services	Load Balancing Services	No additional services
Bandwidth	40%	30%	20%
Segmentation	Single VRF	Single VRF	Single VRF
VLAN	Multiple VLANs per client	Multiple VLANs per client	Single VLAN per client
Data Protection	Clone - Mirror copy (local site)	Snap - Virtual copy (local site)	none
Disaster Recovery	Remote replication (any-point in-time recovery)	Remote replication (With specific RPO/RTO)	none

The following sections identify how the resources differ among the three tiers.

Network Resources

The Cisco VMDC solution leverages Multi-VRF, VLAN, and virtualized services, such as firewall and load balancing contexts, to extend end-to-end network virtualization into the data center. Figure 1-6 depicts the network components assigned to Bronze, Silver, and Gold service tiers in the VMDC solution.

Figure 1-6 Network Resources by Service Tier



Each tenant leverages a number of service tiers to provide a dedicated virtual network (or virtual private data center). Depending upon the tenant application requirements, multiple application tiers can be created within a Gold or Silver container, allowing each separate tier of the application to reside in a separate VLAN within the VRF.

For each Gold and Silver container, a unique VRF and three VLANs are provisioned. The Gold tenant container includes a dedicated virtual firewall and load-balancing instance. The Silver container includes only a virtual load-balancing instance. The Bronze container is assigned a unique VRF and a single VLAN, and no firewall or load-balancing services are provided.

These service tier definitions form a baseline to which additional services may be added for enhanced security, PCI compliance, data store protection, business continuity, or disaster recovery.

Compute Resources

Server virtualization entails running multiple virtual servers on a single physical blade server. The two primary characteristics to consider are vCPU and RAM.

The number of virtual machines (VMs) that can be enabled depends on the workload type being deployed and the CPU and memory capacity of the blade server. Cisco UCS B-series blade servers are two-socket blades based on the Intel Xeon series processor. Each socket has four cores for a total of 8 cores, or 8 vCPUs, per blade.

At the compute layer, service tier differentiation defines three compute workload sizes called Large, Medium, and Small. As Table 1-2 shows, we enabled 32 Small VMs per blade server by allocating 0.25 vCPU for each virtual machine. A Large VM has a dedicated vCPU limiting the total number of Large workloads to 8 per blade server.

Table 1-2 lists the workload options and compute resource sizes.

Table 1-2 Compute Resources by Virtual Server Size

	Virtual Server Options		
	Large	**Medium**	**Small**
vCPUs per VM	1 vCPU	0.5 vCPU	0.25 vCPU
Cores per CPU	4	4	4
VM per CPU	4 VM	16 VMs	32 VMs
VM per vCPU Oversubscription Ratio	1:1 (1)	2:1 (0.5)	4:1 (0.25)
RAM allocated per VM	16 GB dedicated	8 GB dedicated	4 GB from shared pool

Storage Resources

The Cisco VMDC architecture defines three static storage allocation sizes called Large, Medium, and Small. These storage arrays are highly available and reliable.

Table 1-3 lists the storage resources sizes.

Table 1-3 Virtual Server Storage Size Allocation

	Storage Resource Options		
	Large	**Medium**	**Small**
Base storage (GB)	300	150	50
Storage growth increment (GB)	50	50	50

You can further refine the service tiers by differentiating the backup and recovery options. Snap and Clone techniques can create point-in-time consistent copies of tenant volumes. To provide support for disaster recovery, Snap volumes can be replicated to multiple locations. Table 1-4 presents example storage distinctions by service tier.

Table 1-4 Service Tier Distinctions for Storage Backup and Recovery

	Gold	**Silver**	**Bronze**
Backup (retention length options)	1 mo., 6 mo., or 1yr.	1 mo., 6 mo., or 1yr.	1 mo., 6 mo., or 1yr.
Data protection	Clone – Mirror copy (local site) – SNAP copies every 4 hrs.; 36 hr. retention	Snap – Virtual copy (local site) SNAP copies every 8 hrs.; 36 hr. retention	None
Disaster recovery	Remote replication SRDF	Remote replication Symmetrix Remote Data Facility (SRDF)	None

Compact Pod Network Topology

Data center networking technology is currently an area of rapid change. Higher-performance end nodes and the migration to 10-Gigabit Ethernet for edge connectivity are changing design standards, while virtualization capabilities are expanding the tools available to the network architect. When designing the

data center network, the experienced architect relies on a solid hierarchical foundation for high availability and continued scalability. This foundation also provides the flexibility to create different logical topologies utilizing device virtualization, the insertion of service devices, as well as traditional Layer-3 and Layer-2 network configurations. The following section describes the hierarchical network design reference model as applied to meet the requirements and constraints commonly found in today's data centers. As a reference model, this topology is flexible and extensible, and may need to be extended or modified to meet the requirements of a specific enterprise data center network.

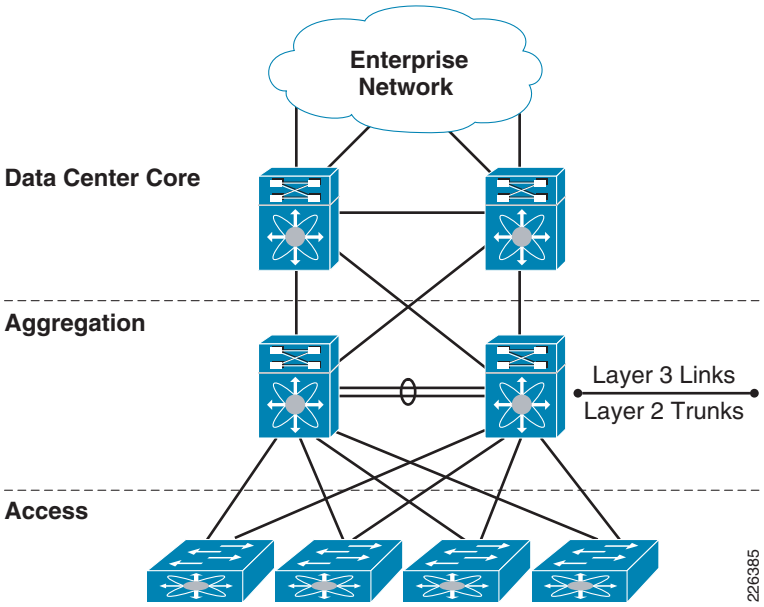
Hierarchical Network Design Reference Model

Hierarchical network design has been commonly used in networking for many years. This model uses redundant switches at each layer of the network topology for device-level failover that creates a highly available transport between end nodes using the network. Data center networks often require additional services beyond basic packet forwarding, such as server load balancing, firewall, or intrusion prevention. These services might be introduced as modules populating a slot of one of the switching nodes in the network or as standalone appliance devices. Each service approach also supports the deployment of redundant hardware to preserve the high availability standards set by the network topology.

A structured data center environment uses a physical layout that correlates tightly to the hierarchy of the network topology. Decisions on cabling types and the placement of patch panels and physical aggregation points must match the interface types and densities of the physical switches being deployed. In a new data center build-out, the two can be designed simultaneously, also taking into consideration the constraints of power and cooling resources. When seeking to avoid significant new investment within an existing data center facility, an architect must consider the pre-existing physical environment of cabling, power, and cooling when selecting switching platforms. Careful planning in conjunction with networking requirements and an eye toward flexibility for the future is critical when designing the physical data center environment. Taking a modular approach to data center design provides flexibility and scalability in both network topology design and utilization of physical resources.

Figure 1-7 illustrates the primary network switching layers of the hierarchical network design reference model for the data center environment. The overall hierarchical model is similar to the reference topology for enterprise campus design, but the term *aggregation layer* replaces the term *distribution layer*. The data center network is less concerned with distributing network access across multiple geographically disparate wiring closets and is focused aggregating server resources and providing an insertion point for shared data center services.

Figure 1-7 Hierarchical Network Design Reference Model



The reference model in Figure 1-7 shows the boundary between Layer-3 routed networking and Layer-2 Ethernet broadcast domains at the aggregation layer. Larger Layer-2 domains increase the physical flexibility of the data center-providing the capability to manually or virtually relocate a server to a different physical rack location with less chance of requiring a change of IP addressing to map to a specific subnet. This physical flexibility comes with a tradeoff. Segregating the network into smaller broadcast domains results in smaller spanning tree domains and failure domains, which improve network stability, reduce convergence times and simplify troubleshooting. When determining how to scale Layer-2 domains, the network architect must consider many factors including the access switching model in the use and nature of the underlying applications being serviced. Cisco has introduced features such as bridge assurance and dispute mechanism into switching products to allow greater scalability of Layer-2 domains with increased stability of the STP.

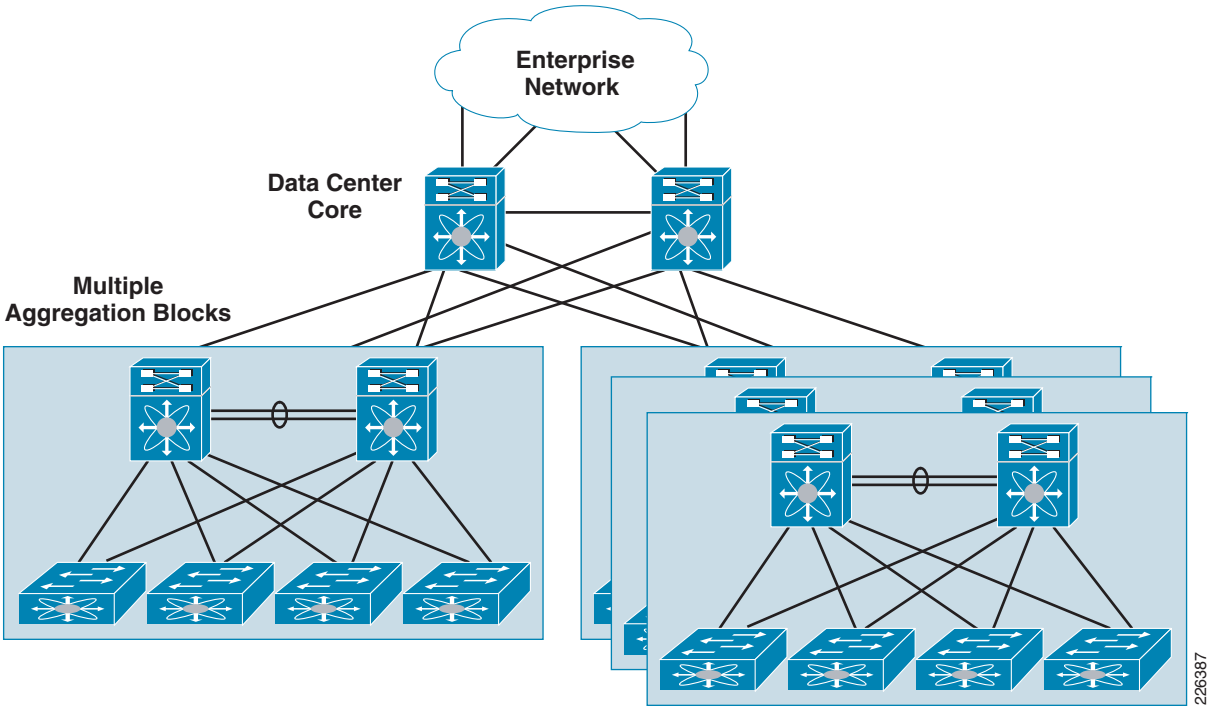
Core Layer

The hierarchical network design model gains much of its stability and high availability characteristics by splitting out switching nodes based on their function, and providing redundant switching units for each functional layer required. The core of a data center network is typically broken out into a pair of high performance, highly available chassis-based switches. In larger or geographically dispersed network environments, the core is sometimes extended to contain additional switches. The recommended approach is to scale the network core continuing to use switches in redundant pairs. The primary function of the data center network core is to provide highly available, high performance Layer-3 switching for IP traffic among the other functional blocks of the network, such as campus, Internet edge and WAN. By configuring all links connecting to the network core as point-to-point Layer-3 connections, rapid convergence around any link failure is provided, and the control plane of the core switches is not exposed to broadcast traffic from end node devices or required to participate in STP for Layer-2 network loop prevention.

In small-to-medium enterprise environments, it is reasonable to connect a single data center aggregation block, or pod, directly to the enterprise switching core for Layer-3 transport to the rest of the enterprise network. Provisioning a separate, dedicated pair of data center core switches provides additional insulation from the rest of the enterprise network for routing stability and also provides a point of scalability for future expansion of the data center topology. As the business requirements expand and

dictate two or more aggregation blocks serving separate pods or zones of the data center, a dedicated data center core network provides for scale expansion without requiring additional Layer-3 interfaces to be available on the enterprise core. An illustration of scaling the data center topology with a dedicated core and multiple aggregation blocks is provided in Figure 1-8.

Figure 1-8 Scaling the Data Center with a Dedicated Core



Cisco's premier switching platform for the data center core is the Nexus 7000 Series switch. The Nexus 7000 Series has been designed from the ground up to support the stringent uptime requirements of the data center. The Nexus 7000 Series switches are optimized for support of high density 10-Gigabit Ethernet, providing scalability in the 18-slot chassis up to 128 wire rate 10-Gigabit Ethernet interfaces when ports are configured in a dedicated mode using the N7K-M132XP-12 I/O Module. The Nexus 7000 Series hardware is coupled with Cisco NX-OS, a modular operating system also designed specifically for the requirements of today's data center networks. NX-OS is built on the industry-proven SAN-OS software-adding virtualization, Layer-2, and Layer-3 features and protocols required in the data center environment. NX-OS includes high availability features, such as granular process modularity, In-Service Software Upgrade (ISSU), and stateful process restart, that are specifically targeted at the service-level requirements of the enterprise or service provider data center.

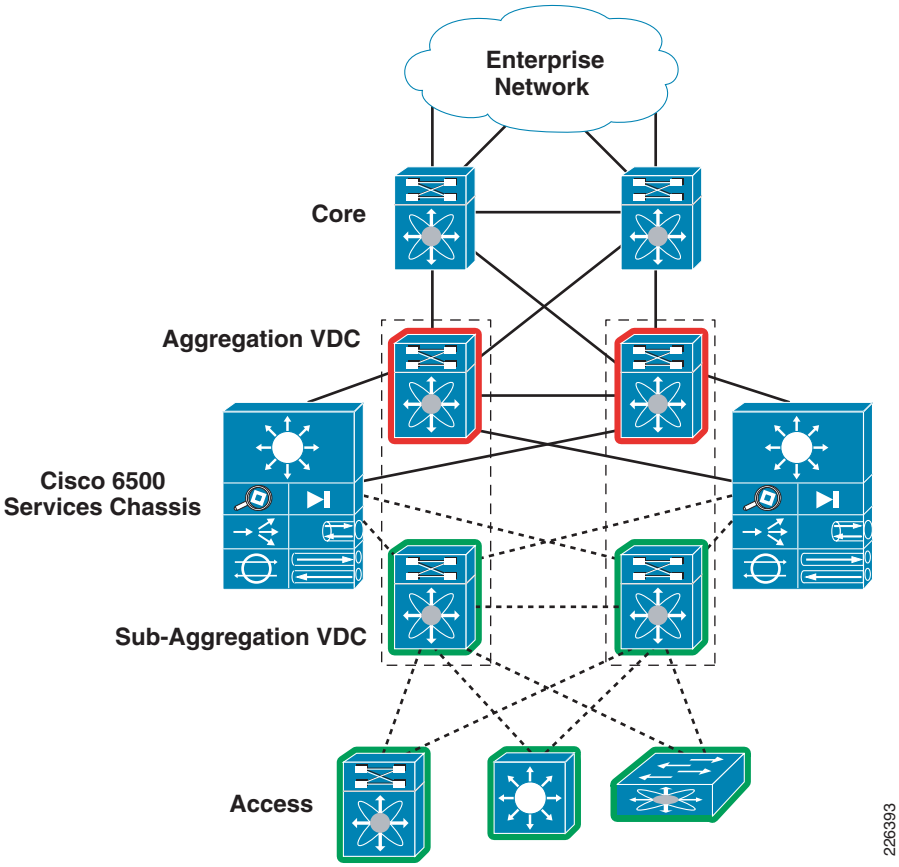
When choosing switching platforms to provision layers of the data center network topology, the network architect must be aware of specific features and interface types required by the network design. The Nexus 7000 Series offers unique virtualization features such as Virtual Device Contexts (VDCs) and Virtual Port Channels (vPCs). The Nexus 7000 Series switches also have excellent high availability features, throughput, and 10-Gigabit Ethernet port densities; however, NX-OS does not support some of the features found in Cisco IOS-based switching platforms. Another Cisco switching platform commonly found in the core of today's data centers is the Cisco Catalyst 6500. The Catalyst 6500 offers software features such as support of Multi Protocol Label Switching (MPLS), VLAN Mapping, and Q-in-Q multiple-level VLAN tagging that may be required in specific designs. The Cisco Catalyst 6500 also offers greater diversity of physical interface types and support for services modules directly installed within the chassis.

Services VDC Sandwich Model

Data center service insertion requirements may include server load-balancing devices, security devices such as firewall and intrusion prevention, and others. Multiple approaches exist for the integration of these services into the data flow. Design decisions include using modules in external Services Chassis, using appliances, and whether to run the service devices in a transparent or routed mode. One very flexible design approach is to use all services in transparent mode, but to insert an additional layer of routing instances between the server farm subnets and the services devices. This approach has been shown in design guidance using VRFs, and the deployment of multiple VRFs also provides the capability to direct traffic independently through multiple virtual contexts on the service devices, leveraging the virtualization of both the routing functions and the services devices in the design.

The VDC capability of the Nexus 7000 Series enables the network architect to leverage another type of virtualization in the design, to improve ease of configuration, supportability, and security. A secondary virtual switching layer called the *sub-aggregation* can be created using VDCs, located between the services devices and the access switches. This topology is referred to as a services VDC sandwich and is the topology used in the VMDC 2.0 Compact Pod architecture. An example of this topology using services modules located in external Catalyst 6500 chassis is shown in Figure 1-9.

Figure 1-9 Services Sandwiched Between VDCs



All the access layer switches shown in Figure 1-9 attach only to the sub-aggregation VDCs. Different classes of servers could also be attached to access-layer switches that connect directly to the main aggregation layer above the Services Chassis, if they either do not require services or are serviced by a different group of services devices. Additional considerations when designing this type of topology include the following:

- Similar designs have been deployed only using a single pair of switches with separate VLANs and VRFs to provide the routing instance below the Services Chassis. The insertion of a separate set of VDCs into the design still represents using a single physical pair of switches to perform these functions but provides better isolation between the routing environments above and below the Services Chassis. This conceptually provides for easier support and configuration, without increasing the impact of a single-switch failure due to the introduction of a second set of VDCs.
- The security model is more robust, since the operating environment of the sub-aggregation VDCs is completely separate from the primary aggregation layer. Instead of being only separate VLANs and VRFs on the same switch, they are separate virtual switches with completely different sets of processes and physical ports.
- Additional interfaces may be required for the VDC sandwich topology as compared with a VRF sandwich topology. The Services Chassis must have separate physical connections into both sets of VDCs as opposed to VLANs sharing the same trunks. Additional interface count must also be provisioned to support the inter-switch link between the two sub-aggregation VDCs.
- This model has been validated by Cisco using Firewall Services Module (FWSM) running in transparent mode and Application Control Engine (ACE) modules running in routed mode, where the two layers of VDCs are direct IP routing peers. Layer 3 control plane load on the VDC below the services may be limited by using static routes pointing to an HSRP address shared between the primary aggregation VDCs to support IP unicast traffic flows. IP multicast traffic is not supported over a combination of static routes and HSRP addresses. If IP multicast is a requirement, then an IGP such as OSPF or EIGRP may be used.
- VDCs provide the distinction between the routing instances of the aggregation and the sub-aggregation layers; however, the use of multiple VRFs in the sub-aggregation layer may be utilized to support additional virtualization capabilities. Distinct VRFs in the sub-aggregation layer may be mapped using VLANs to separate contexts within the virtualized service devices such as the FWSM and ACE, allowing active contexts to be split between both Services Chassis. If services are required between layers of a multi-tier application architecture, placing these tiers in subnets belonging to separate VRFs will allow for powerful, multi-context service insertion between tiers.
- A services VDC sandwich using external Services Chassis provides independent connectivity between the services and both aggregation switches. If the aggregation switch on the left side of the topology fails, the services on the left side have dual connectivity and can maintain a primary role. Service appliances run in transparent mode, such as the Adaptive Security Appliance (ASA) 5580, that only support single connections to carry a given VLAN. In transparent mode, such an appliance will not be dual-homed if attached directly to the aggregation, but it can still be deployed in a highly available manner by using redundant appliances.

Note For more detail on the VDC services sandwich architecture with virtualized services, refer to the *Data Center Service Patterns* document at the following URL: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/dc_serv_pat.html

Aggregation VDC

The primary function of the aggregation VDC is to provide highly available, high performance Layer-3 switching IP traffic from the services layer to the other functional blocks of the network, such as data cetner core (if deployed). In smaller enterprise environments, it is reasonable to collapse the core functionality into the aggregation VDC and provide connections to the campus, Internet edge, and WAN at this layer.

Services Layer - Datacenter Services Node (DSN)

The Cisco VMDC 2.0 DSN design case virtual switching system (VSS) and Cisco FWSM and Cisco ACE virtualization.

Virtual Switching System (VSS)

VSS combines two physical Cisco Catalyst 6500 Series Switches into one virtualized switch. This arrangement enables a unified control plane and also allows both data planes to forward simultaneously. With VSS, multi-chassis EtherChannel (MEC) is introduced, which allows a port channel to be formed across two physical switches. vPC and VSS both provide enhanced system availability through redundant systems, eliminate reliance on Spanning Tree Protocol, achieve faster convergence times, and enable full system availability at all times. For the Cisco DSN use cases, the sub-aggregation layer switches can run in vPC mode and interconnect to the Cisco DSN through MEC, which will be running in the VSS. An additional benefit of integrating VSS with Cisco DSN is that this integration increases the number of supported service modules per chassis from four to eight in a single VSS domain, enabling an active-active highly available service chassis.

Transparent FWSM

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

Routed ACE

The active/standby Services Chassis design leverages the ACE context in a routed server load-balancing mode. The ACE has a single logical interface (VLAN), which exists between itself and the Aggregation VDC. The ACE has a second interface which resides on the Access Layer VLAN where the real servers sit. The ACE routes traffic destined to Enterprise Cloud to an HSRP address on the Aggregation Layer VDC.

The advantage of this deployment model is that the ACE virtual context is exposed to only those flows requiring its services. Non-load-balanced flows traverse in and out of the server farm without being processed by the ACE; while load-balanced flows benefit from dedicated ACE services positioned for optimal performance.

To force traffic back to the ACE context, it is necessary to configure Source NAT on the ACE. Source NAT on the ACE is simple to deploy and enforces symmetric traffic flow by using IP address pools dedicated to the ACE virtual context and basic Layer 2 switching. From a traffic flow perspective, source NAT readily fits into many existing data center designs by introducing the load balancer as a new separate Layer 2 adjacency in the network. To accommodate application-logging requirements, the network administrator may use the ACE's HTTP header manipulation feature by inserting the original source IP address of an HTTP traffic flow into the HTTP header. The disadvantage to this technique is the loss of source IP logging on the server for non-HTTP applications.

Sub-Aggregation VDC

The sub-aggregation layer of the data center provides a consolidation point where access layer switches are connected providing connectivity between servers for multi-tier applications, as well as connectivity across the services, aggregation, and core layers of the data center network to clients residing within the campus, WAN, or Internet. The sub-aggregation layer typically provides the boundary between Layer-3 routed links and Layer-2 Ethernet broadcast domains in the data center. The access switches are connected to the sub-aggregation layer using 802.1Q VLAN trunks to provide the capability of connecting servers belonging to different VLANs and IP subnets to the same physical access switch.

Traditional models of access-layer connectivity include links from each access-layer switch into both switches forming the sub-aggregation-layer redundant switch pair. This approach provides network resiliency in the event of a single link or interface failover or failure of one of the sub-aggregation switches. The inter-switch link between the two sub-aggregation switches is also an 802.1Q trunk that carries all VLANs in use in the server farm. The STP is active independently for each VLAN instance using the Rapid Per VLAN Spanning Tree Plus (RPVST+) model, which blocks redundant ports when they are not needed to avoid network loops. Features such as Virtual Port Channels (vPC) on the Cisco Nexus 7000 Series and Virtual Switching System (VSS) on the Catalyst 6500 series have been introduced to allow both switches in the aggregation pair to act as a single switching unit from a STP and port channel perspective. This approach allows all links between an access switch and the sub-aggregation layer to be active as a single port channel instead of having STP blocking a redundant path.

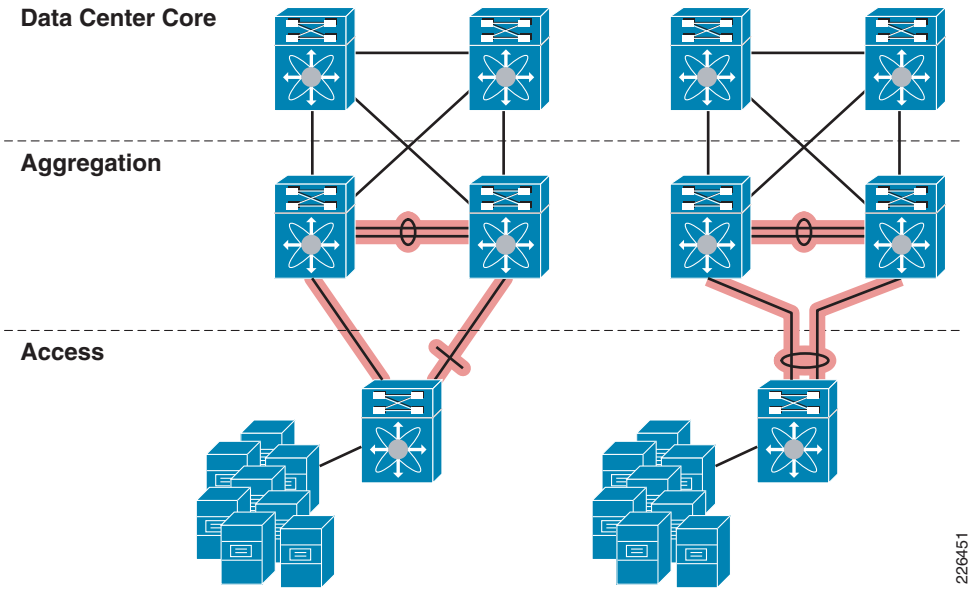
Loop Free Layer 2 Design

A data center is generally made of similar simple building blocks, replicated at will to achieve the desired level of scalability. The solution provides for redundancy, which means that devices in a particular position are at least duplicated. The traditional network design is a typical example of that: a pair of aggregation switches to which as many access switches as necessary are connected in a redundant way. The two main drawbacks of this solution are as follows:

- There is no Layer-2 multipathing for a particular VLAN, and the per-VLAN load balancing that allows using both uplinks of an access switch needs user configuration. There is no way of escaping this constraint as it dictated by the way bridging requires a spanning tree in the data plane.
- The dependency on STP for the creation of a spanning tree topology in the data plane, introducing delay in the convergence and potential risks.

Port channel technology is solving those remaining issues for the very specific case of the interconnection of two switches (see Port Channels/EtherChannels) Alone, link aggregation cannot be used to create a fully redundant data center, as it does not protect against the failure of a single switch. Cisco has recently introduced two technologies that lift this latter limitation. Both VSS (on the Catalyst 6000) and vPC (on the Nexus 7000) allow creating a Layer-2 port channel interface distributed across two different physical switches. This limited step-up in the channeling capability is enough to provide the simple building block required to build an entire data center with no dependency on the spanning tree model. Figure 1-10 shows a high level use of the solution.

Figure 1-10 Loop-Free Network



The left part of Figure 1-10 illustrates the current model, where the redundancy is handled by STP. The right part of Figure 1-10 represents the solution introduced by distributing the end of a channel across the two aggregation switches.

The logical view shows that the redundancy has been hidden from STP. As far as the "Rules for STP Network Stability" are concerned, the right side of Figure 1-10 shows the best solution, where the following are depicted:

- The number of blocked ports has been eliminated
- The freedom of STP has been also entirely removed, as it cannot open a loop even if it wanted to.

However, the recommendation is to keep STP on as a backup mechanism. Even if the redundancy has been hidden to STP, it is still there, at a lower layer. It is just handled by a different mechanism. STP helps protect against a configuration error that breaks a channel into individual links, for example.

Access Layer

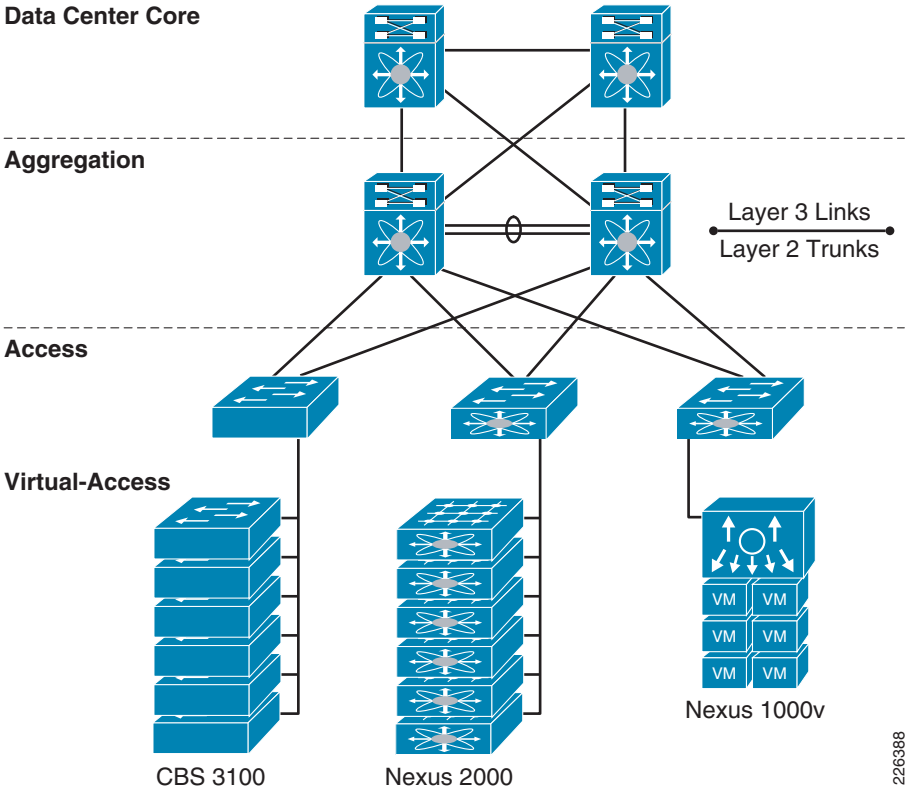
Virtual Access Evolution

The evolution of networking technology in the data center is most evident at the access layer of the network and within the server farm. Several options for building the data center access layer introduce switch virtualization that allows the function of the logical Layer-2 access layer to span multiple physical devices. For example:

- Cisco Nexus 5000 Series switches work in conjunction with the Cisco Nexus 2000 Series Fabric Extenders to act as a single virtual access switch while providing ToR connectivity for servers in multiple racks.
- The software-based switching implementation in the Cisco Nexus 1000V Virtual Distributed Switch also provides virtual access layer switching capabilities designed to operate in server virtualization environments.

Figure 1-11 illustrates these examples of access-layer virtualization in the data center network. The virtual-access sublayer does not represent an additional level of Layer-2 switching; it conceptually exists as virtual I/O modules or line cards extended from a centralized management and control plane. This approach offers many of the benefits of EoR switching, such as reduced aggregation switch port density requirements and fewer points of management, while providing cable-management benefits similar to a ToR model.

Figure 1-11 Data Center Virtual-Access Evolution



The Cisco Nexus 5000 Series switches provide high-density 10-Gigabit Ethernet connectivity and innovative storage integration capabilities for the support of FCoE. With a Layer-2 capable implementation of NX-OS, the Nexus 5000 is optimized for the evolving data center access layer. For customers requiring a density of 1-Gigabit Ethernet server connectivity, the Nexus 2000 Fabric Extenders may be deployed in conjunction with a Nexus 5000 Series switch and treated as a single virtual chassis in the access layer of the data center topology. This approach may be used to provide ToR switching to multiple racks of servers, with all management functions for the Nexus 2000 Fabric Extenders centralized into their associated Nexus 5000 Series switch. The Nexus 5000 Series can also be placed middle-of-row (MoR) to provide 10-Gigabit Ethernet interfaces to nearby servers.

Implementations of hypervisor-based server virtualization systems include software-based logical switching capabilities within the server. The Nexus 1000V virtual distributed switch allows the network architect to provide a consistent networking feature set across both physical servers and virtualized servers. The Nexus 1000V operates as a virtualized chassis switch, with Virtual Ethernet Modules (VEMs) resident on the individual virtualized servers managed by a central Virtual Supervisor Module (VSM) that controls the multiple VEMs as one logical modular switch. The VSM provides a centralized point of configuration and policy management for the entire virtual distributed switch. Both the Cisco Nexus 2000 Fabric Extenders and the Cisco Nexus 1000V represent variations on the evolving capabilities of the data center virtual-access sub-layer.

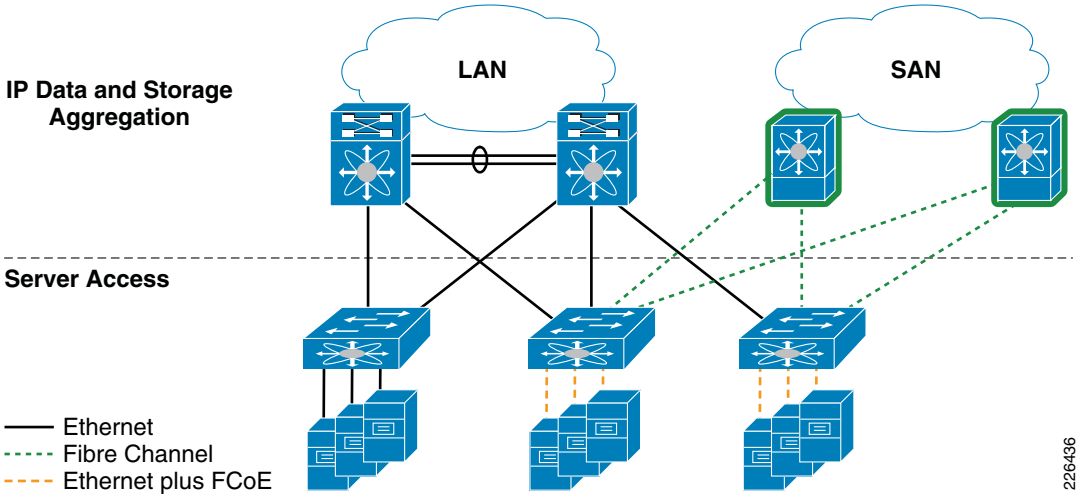
Storage Integration

Another important factor changing the landscape of the data center access layer is the convergence of storage and IP data traffic onto a common physical infrastructure, referred to as a unified fabric. The unified fabric architecture offers cost savings in multiple areas including server adapters, rack space, power, cooling, and cabling. The Cisco Nexus family of switches, particularly the Nexus 5000 Series is spearheading this convergence of storage and data traffic through support of Fibre Channel over Ethernet (FCoE) switching in conjunction with high-density 10-Gigabit Ethernet interfaces. Server nodes may be deployed with converged network adapters (CNAs) supporting both IP data and FCoE storage traffic, allowing the server to use a single set of cabling and a common network interface. The Cisco Nexus 5000 Series also offers native Fibre Channel interfaces to allow these CNA attached servers to communicate with traditional Storage Area Network (SAN) equipment.

At its initial product release, the Cisco Nexus 5000 supports a unified fabric switching approach only at the edge of the data center topology. Over time, the Cisco Nexus family will allow further consolidation of FCoE-based storage traffic into the aggregation layer of the data center. Choosing Cisco Nexus switching platforms for new data center investment today positions the network architect to take advantage of additional I/O consolidation features as they are released across the product family. Figure 1-12 illustrates a topology with CNA-attached servers running both FCoE traffic and IP data traffic over a common interface to a Nexus 5000 switch. The Nexus 5000 splits out the FCoE traffic and provides native Fibre Channel interface connections back into Fibre Channel switches to connect to the shared SAN.

Note In the VMDC design, the UCS 6120 was used to extend the FCoE access layer so that FibreChannel traffic no longer has to pass through the Nexus 5000 to reach the SAN fabric.

Figure 1-12 Access Layer Storage Convergence with Nexus 5000



SAN Design Reference Model

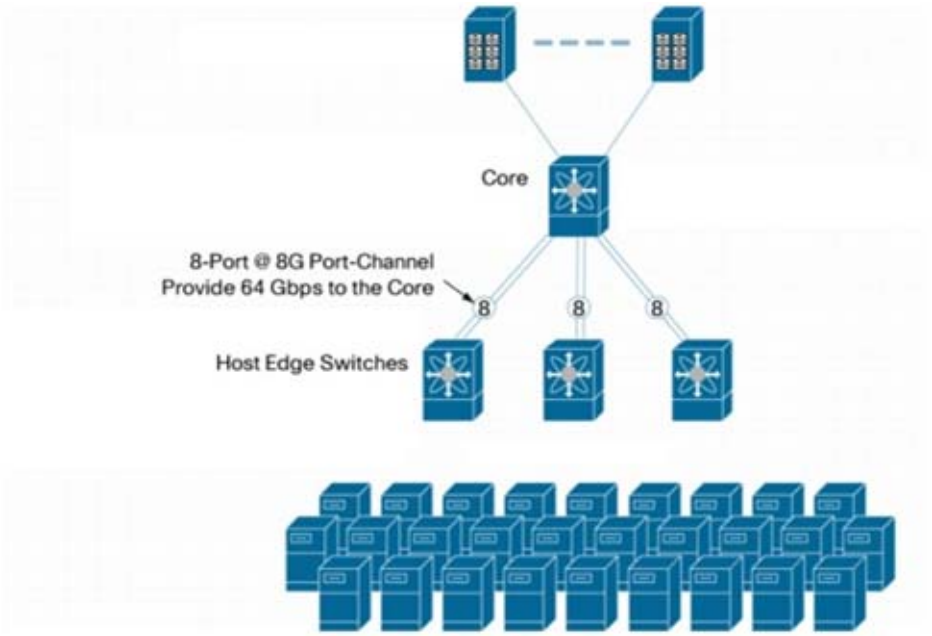
Core/Edge Design Model

It is common practice in SAN environments to build two separate, redundant physical fabrics (Fabric A and Fabric B) in case a single physical fabric fails. Large SAN architectures are classified as one the following topologies in a physical fabric:

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

Within the two-tier design, servers connect to the edge switches, and storage devices connect to one or more core switches (Figure 1-13). This allows the core switch to provide storage services to one or more edge switches, thus servicing more servers in the fabric. The interswitch links (ISLs) will have to be designed so that the overall fabric maintains both the fan-out ratio of servers to storage and the overall end-to-end oversubscription ratio.

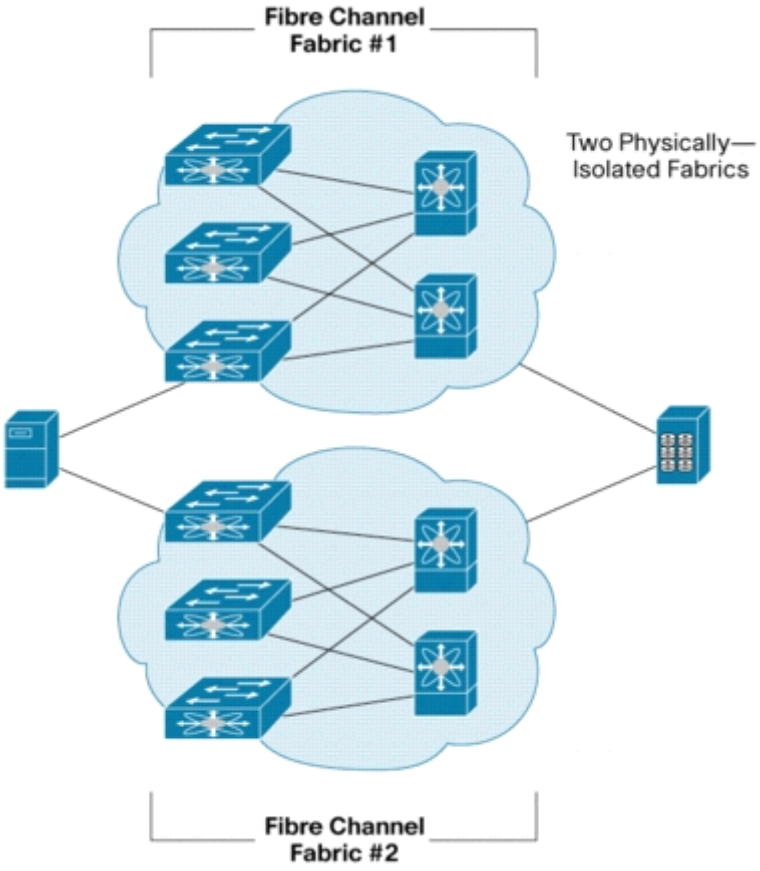
Figure 1-13 Sample Core-Edge Design



Fabric Redundancy

Another area that requires attention in a Fibre Channel SAN is the fabric itself. Each device connected to the same physical infrastructure is in the same Fibre Channel fabric. This opens up the SAN to fabric-level events that could disrupt all devices on the network. Changes such as adding switches or changing zoning configurations could ripple through the entire connected fabric. Therefore, designing with separate connected fabrics helps to isolate the scope of any such events. The Cisco Systems Virtual SAN (VSAN) capability offers a way to replicate this environment, namely, the isolation of events, using the same physical infrastructure. (See Figure 1-14)

Figure 1-14 Designing SANs with Isolated Fabrics



Interswitch links (ISLs)

The connectivity between switches is important as the SAN grows. Relying on a single physical link between switches reduces overall redundancy in the design. Redundant ISLs provide failover capacity if a link fails.

CHAPTER 2

Design Considerations

Revised: April 26, 2011

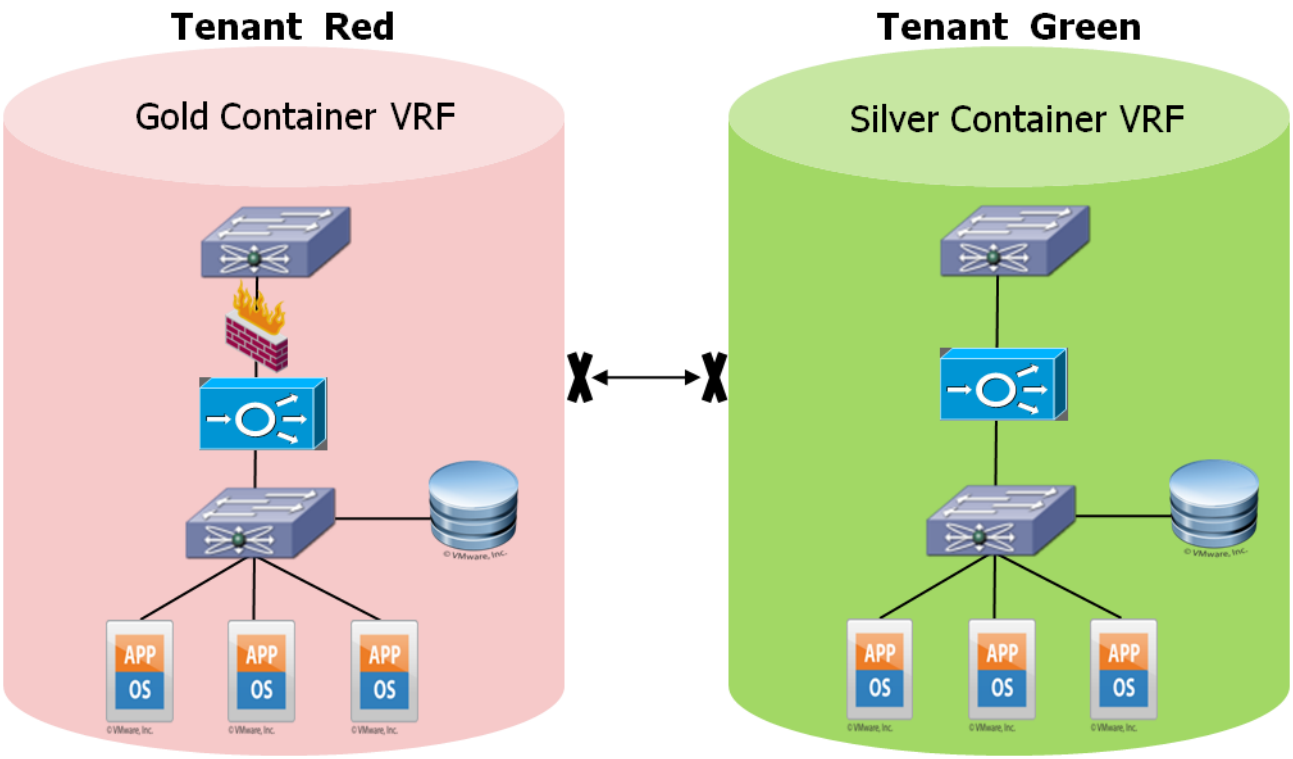
The Cisco VMDC solution addresses the following design considerations in depth:

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

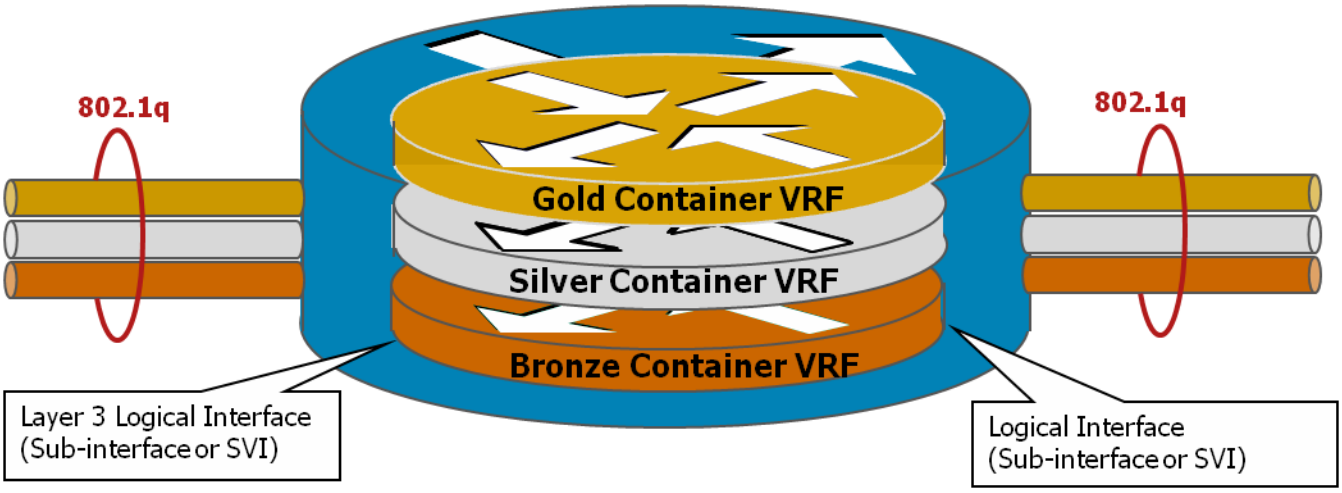
- Sub-Aggregation Layer—Layer 3 separation (VRF) and Layer 2 separation (VLAN)
- Access Layer—Layer 2 separation (VLAN)
- End-to-end virtualization of the network requires separation at each network layer in the architecture:

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



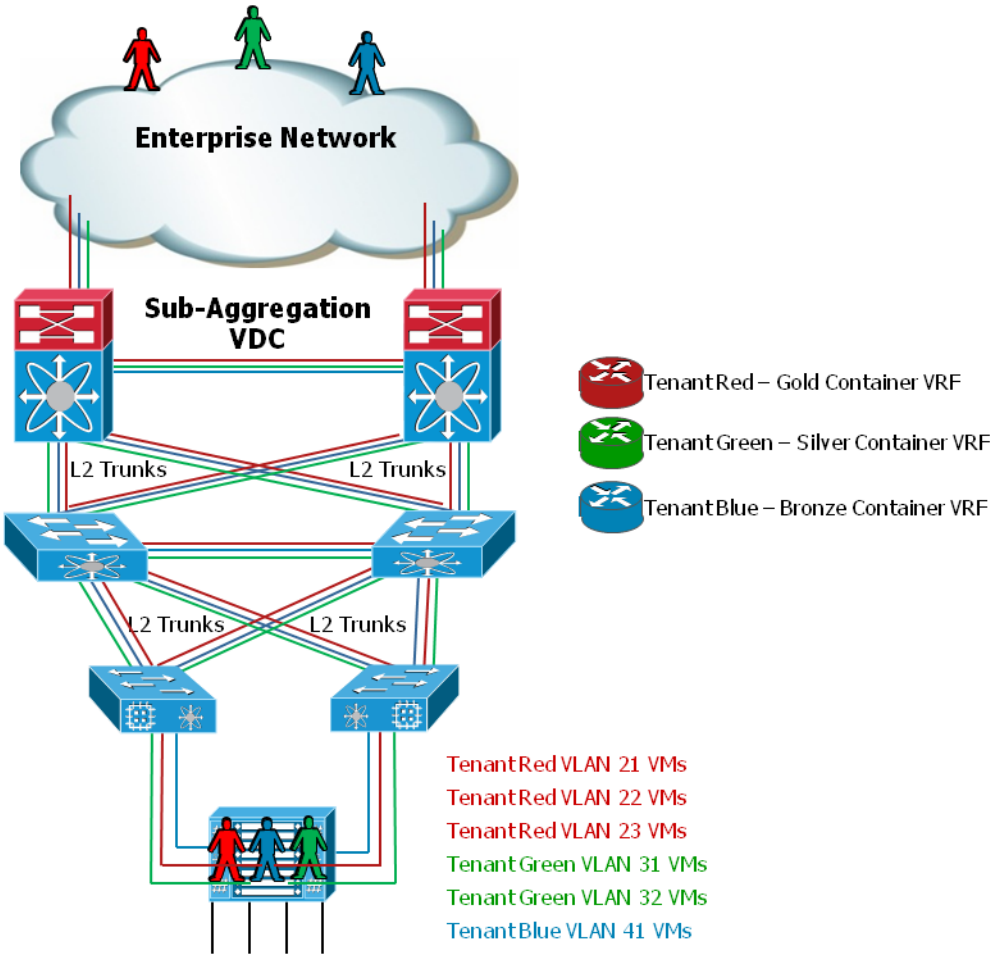
In a multi-tenant environment, Cisco VRF Lite technology offers the following benefits:

- 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.

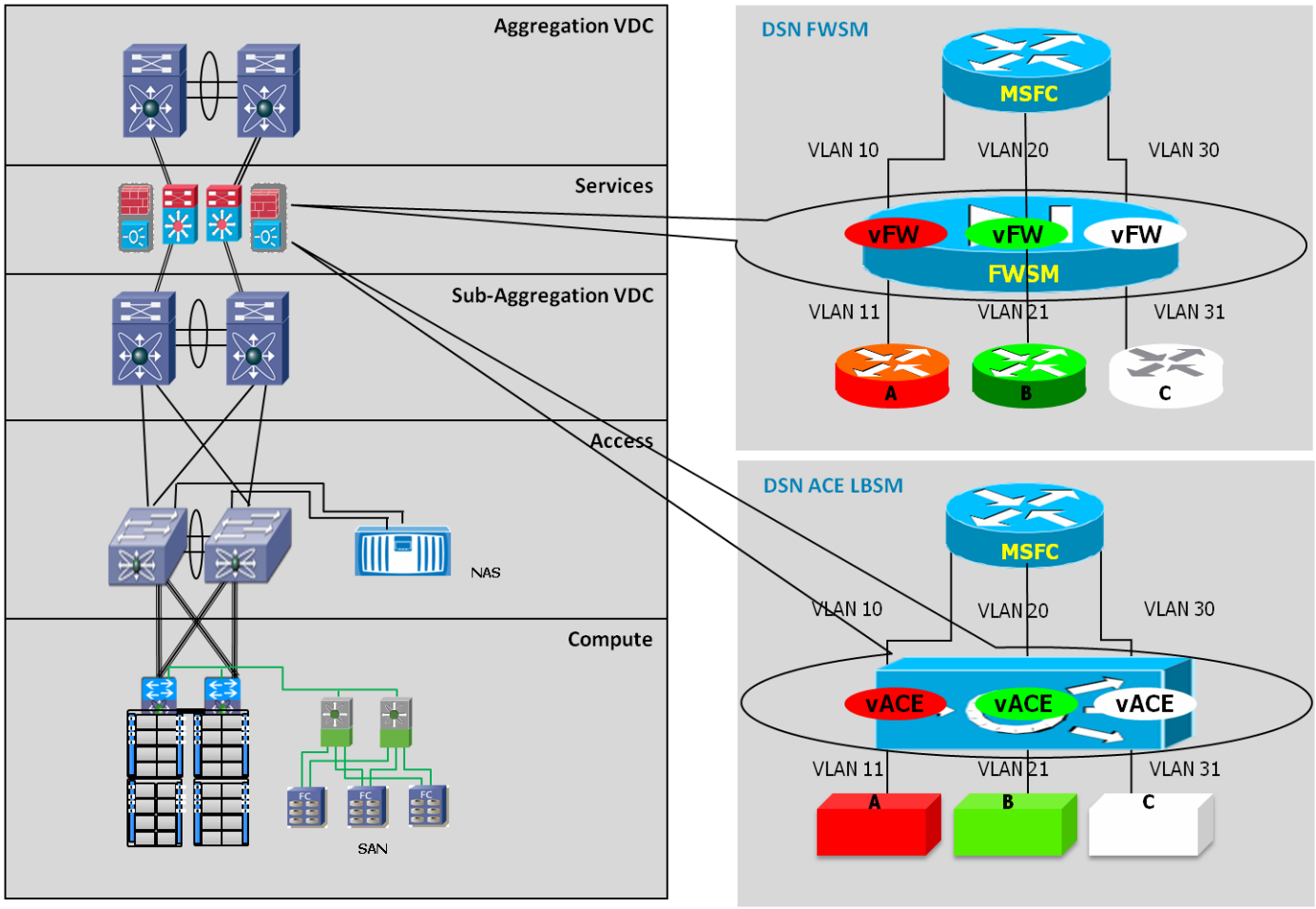
Figure 2-3 VLANs to VRF Mapping



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.

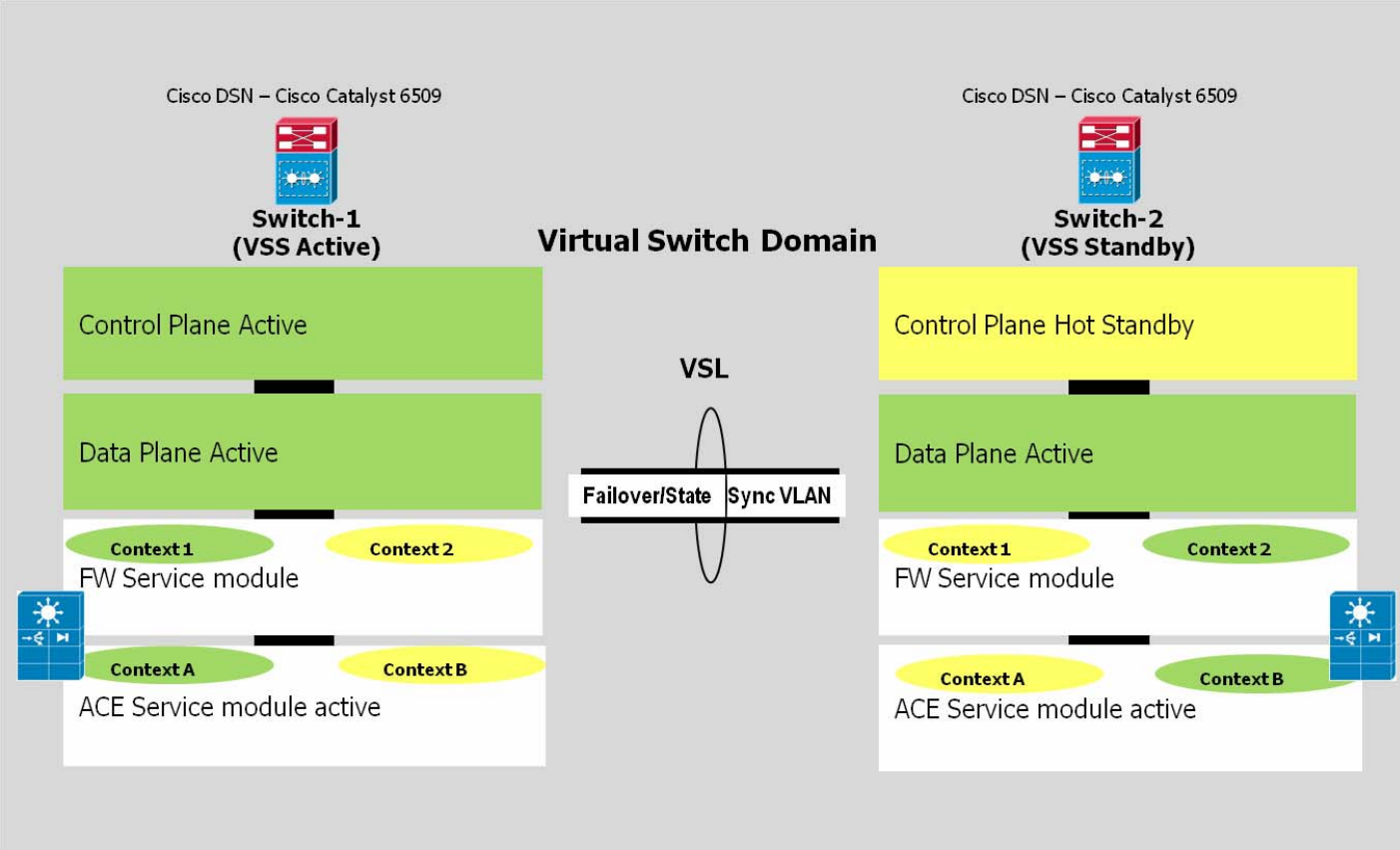
Figure 2-4 Virtual Firewall and Load Balancing Services



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 VMkernal 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 VMkernal 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:

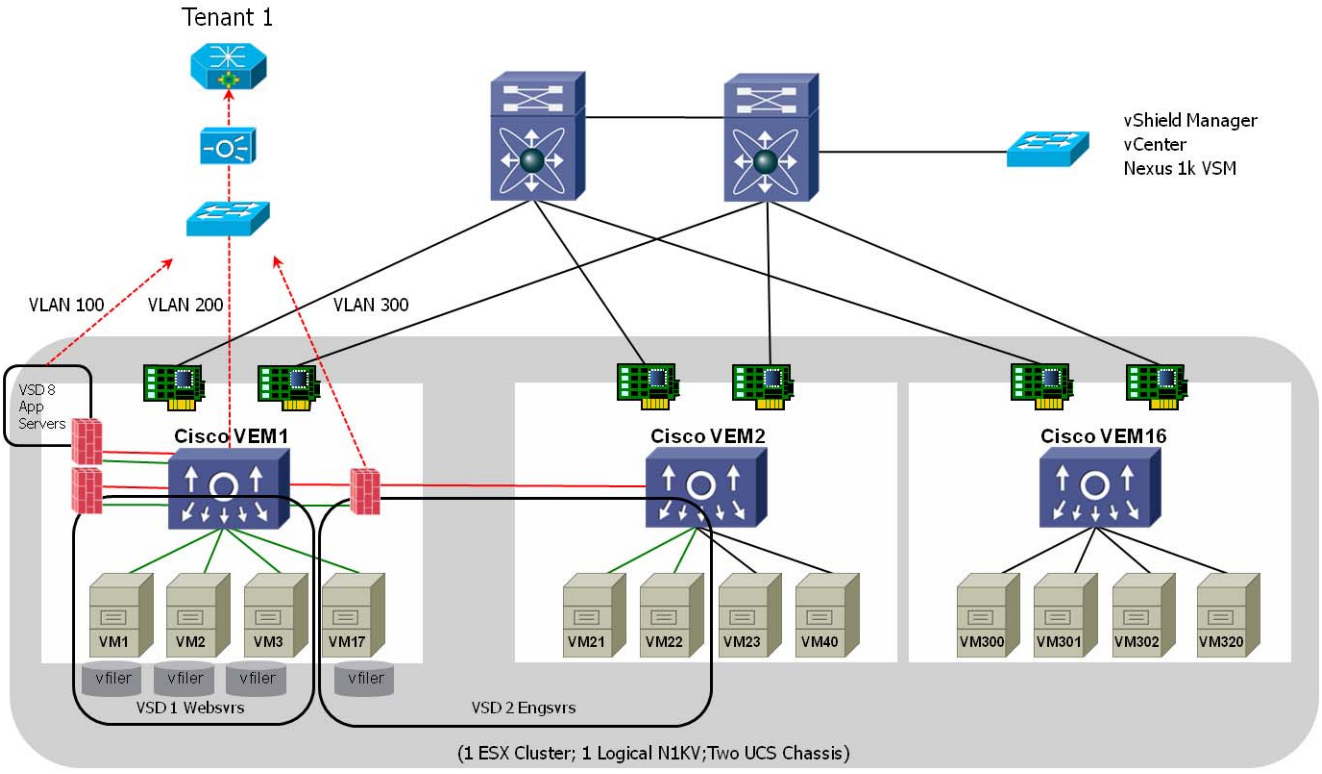
- *Container-level precedence* recognizes the datacenter level as a higher priority than the cluster level. When a rule is configured at the data center level, all clusters and vShield agents within the clusters inherit that rule. A cluster-level rule applies only to the vShield agents in the cluster. 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:

$$\text{Availability} = \frac{\langle \text{Tperiod} - \text{Twithoutservice} \times 100 \rangle}{(\text{Tperiod})}$$

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).

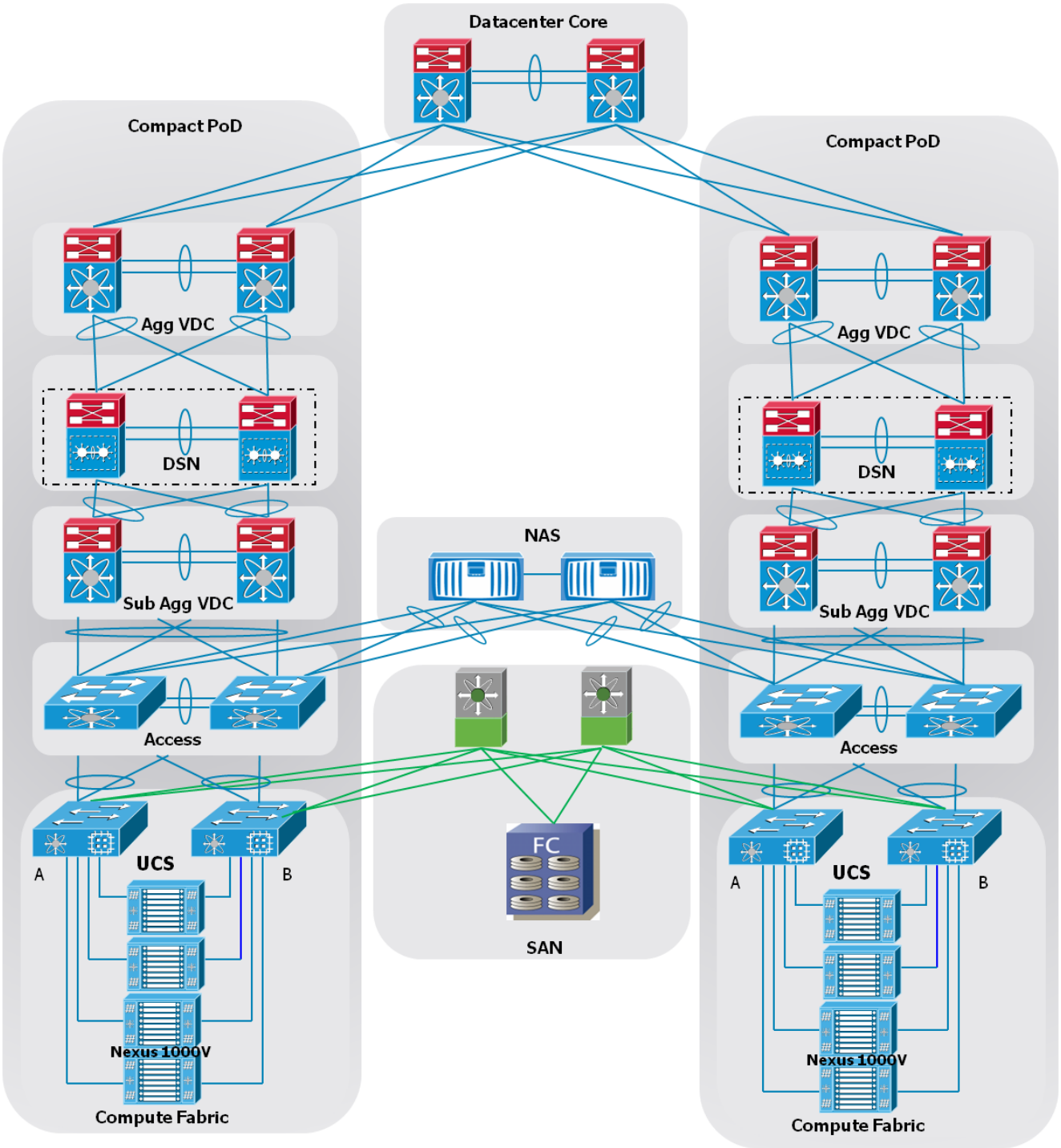
The VMDC architecture addresses the availability requirements of IaaS SLAs for all of the criteria listed in Table 2-2 through 1:1, 1:N or N:N VM, network and storage redundancy, and data security and data protection mechanisms.

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

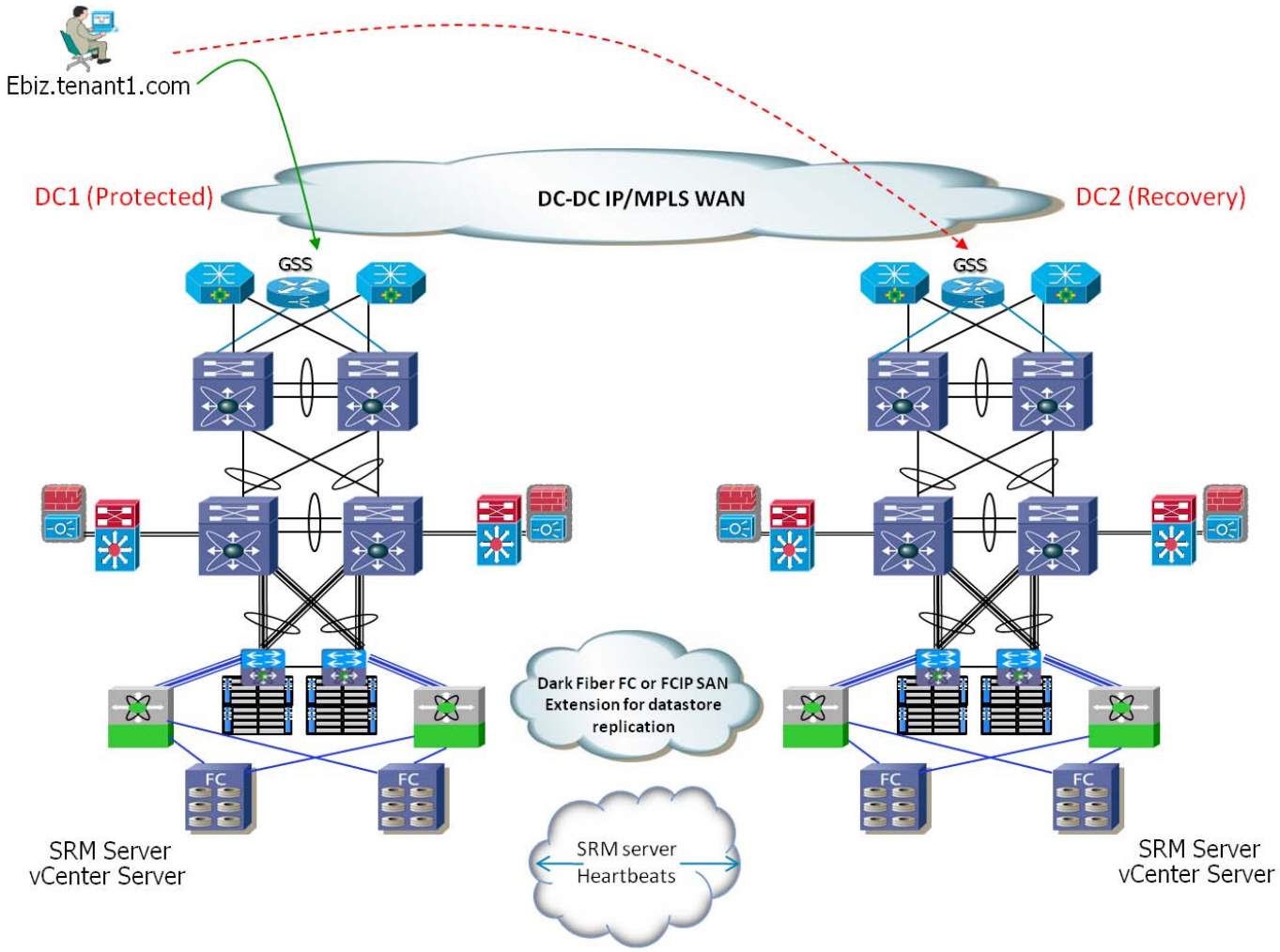
For VMware vShield:

- 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 configure 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 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 architecutre is the ability to function well as tenants needs change in size or volume. The following section highlights some of the key scalablity 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. 100 HSRP groups per port-channel interface.	32 groups per port channel in Aggregation VDC

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:

- **OPSF 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

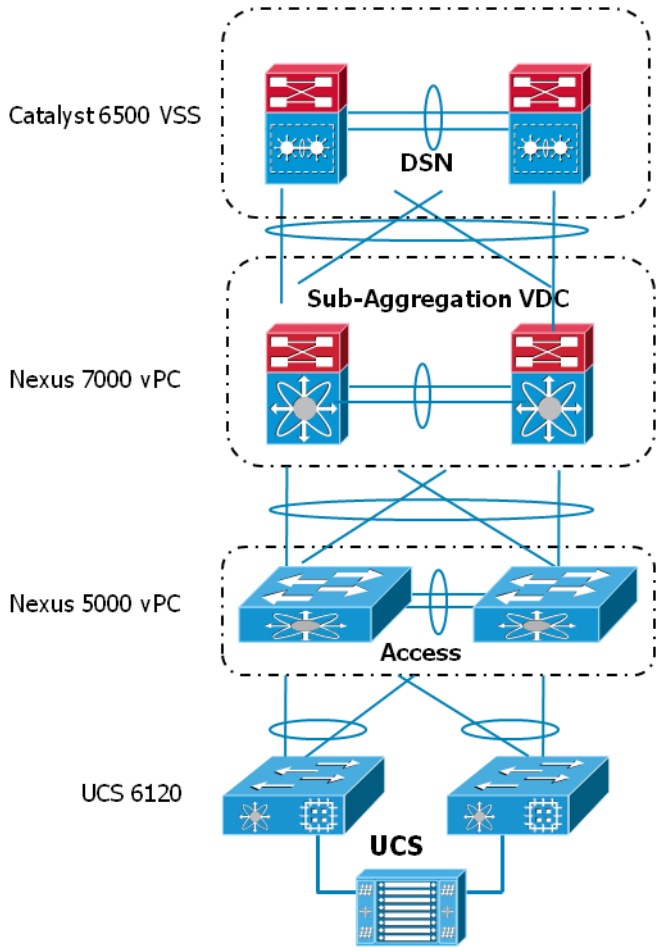
A virtual PortChannel (vPC) allows links that are physically connected to two Cisco Nexus devices to appear as a single PortChannel to any other device, including a switch or server. This feature is transparent to neighboring devices. A vPC can provide Layer 2 multipathing, which creates redundancy via 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 resilience 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#wp1053500

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 increases to a maximum of 20 and 40.

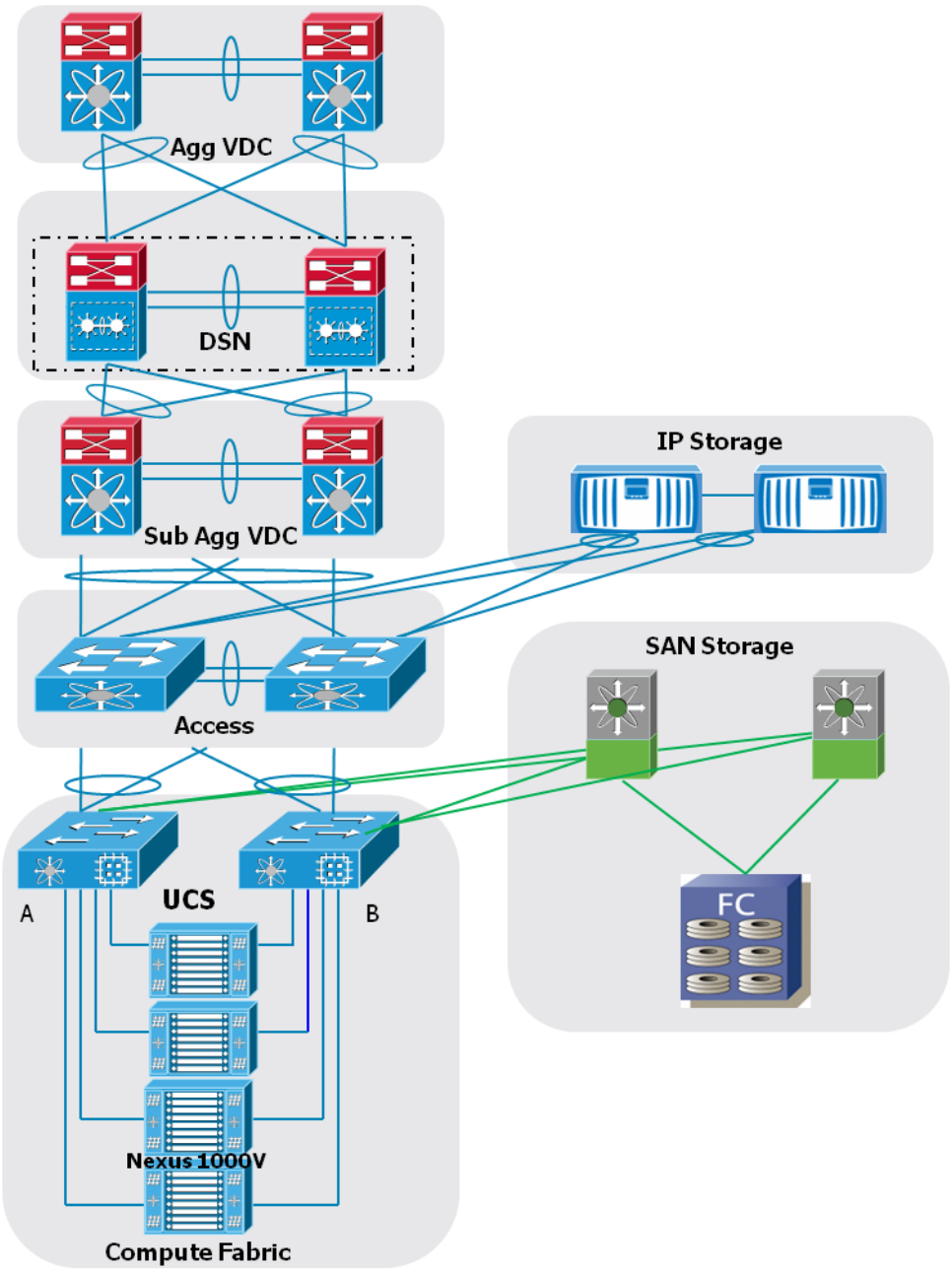
Each chassis can hold up to 8 half height server blades or 4 full height server blades thus allowing the environment to scale out based on the number of chassis connected to the fabric interconnects of the pod.

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

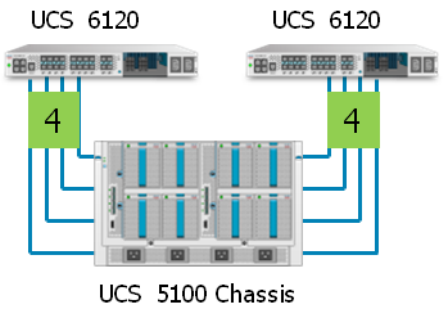
In Table 2-7, 20% of the available 64 Cisco UCS blade servers are reserved for Large workloads, which is 12 blade servers. A UCS blade server has two CPU sockets that accept a quad-core CPU; therefore, a total of 96 vCPUs are available to the Large workloads. Allocating one vCPU per large VM yields a 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

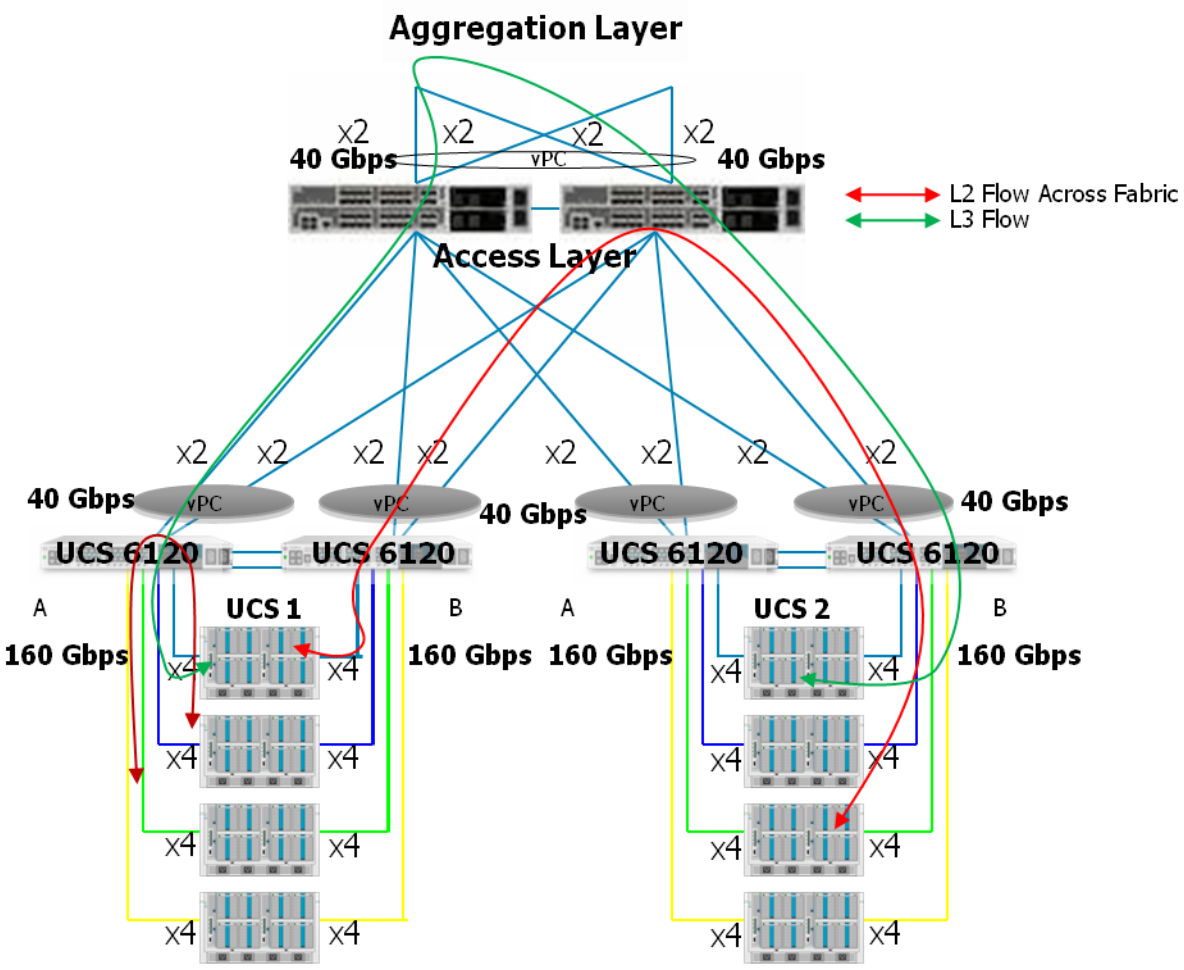


Network architects must consider likely traffic flows within the logical topology that have been created on top of the physical topology. Multi-tier application flows create a portion of traffic that does not pass from the server farm to the aggregation layer. Instead, it passes directly between servers.

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:

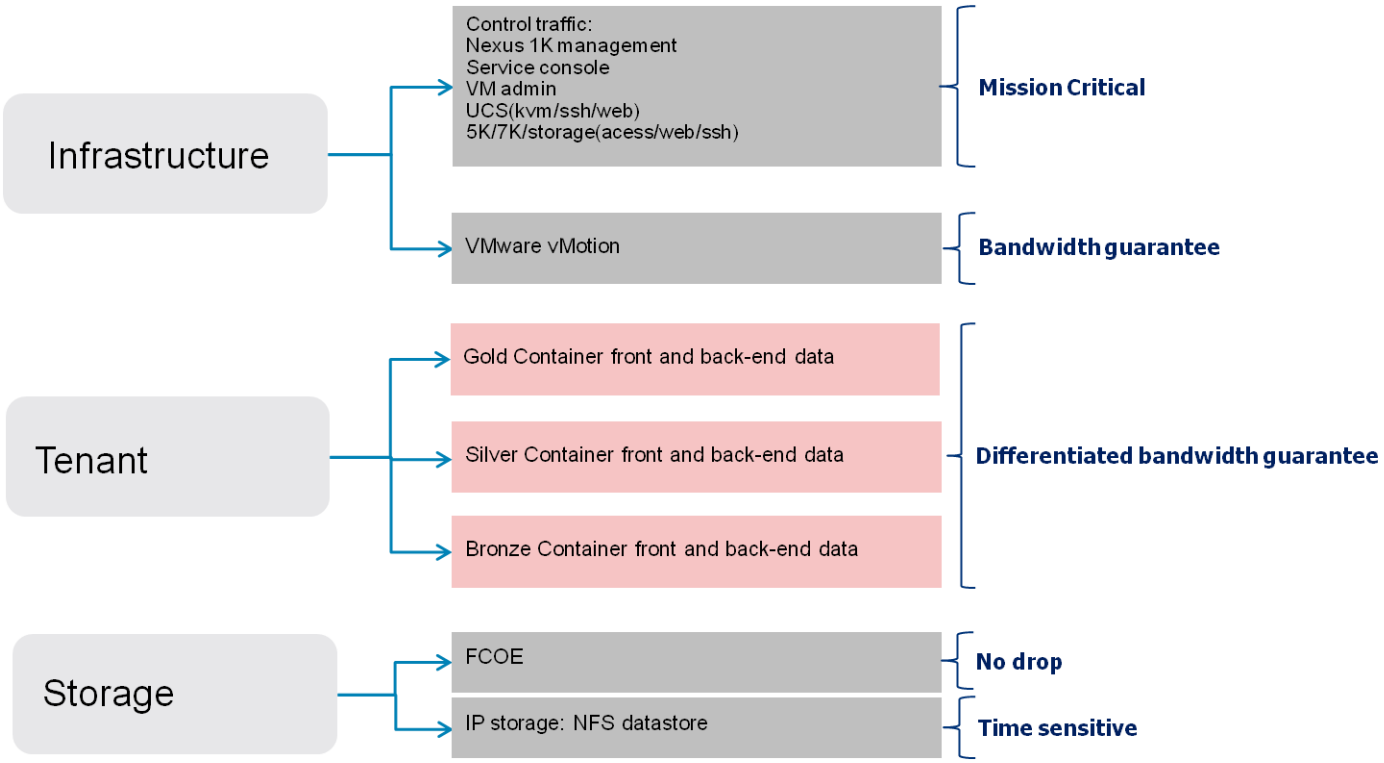
```
policy-map type qos mark-control-packet-vlans
  class n1k-control-vlan
    set cos 6
```

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/C 6k)	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%	—	—	—	—	—

APPENDIX A

Bill of Materials As Validated

Revised: April 26, 2011

Table A-1 presents the product part numbers for the components in the Compact Pod, as well as the network infrastructure required to build out the Cisco VMDC solution, version 2.0.

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration

Part Number	Description	Quantity
Cisco Nexus 7010		
N7K-C7010-BUN	Nexus 7010 Bundle (Chassis,SUP1,(3)FAB1,(2)AC-6KW PSU)	2
N7K-SUP1	Nexus 7000 - Supervisor, Includes External 8 GB Log Flash	2
N7K-M132XP-12	Nexus 7000 - 32 Port 10GbE, 80 G Fabric (req. SFP+)	8
SFP-10G-SR	10GBASE-SR SFP Module	64
CON-OSP-C7010	ONSITE 24X7X4 10 Slot Chassis, No Power Supplies, Fans	2
DC3 COMPACT FLASH	Compact Flash Option	2
N7K-CPF-2GB	Nexus Compact Flash Memory 2 GB (Expansion Flash - Slot 0)	2
N7KS1K9-50	Cisco NX-OS Release 5.0	2
CON-OSP-N7FAB	ONSITE 24X7X4 Nexus 7000 - 10 Slot Chassis	6
CON-OSP-N732XP	ONSITE 24X7X4 Nexus 7000 - 32 Port 10 GbE, 80G Fabric	8
CON-OSP-N7SUP1	ONSITE 24X7X4 Nexus 7000 - Supervisor, Includes Ext	4
Cisco Nexus 5020		
N5k-C5020P-BF	N5000 2RU Chassis no PS 5 Fan Modules 40 ports (req SFP+)	2
N5K-M1008	N5000 1000 Series Module 8xFC 4/2/1 G (req SFP)	2
SFP-10G-SR	10GBASE-SR SFP Module	
N5K PS OPT	Nexus 5 K Power Supply Options	2
N5K-PAC-1200W	Nexus 5020 PSU module, 100-240VAC 1200W	4
N5K CAB OPT	Power Cables	4
CAB-C13-C14-JMPR	Recessed receptacle AC power cord 27	4
N5K EXPAND	N5K Expansion Class	2
N5K-M1-BLNK	N5000 1000 Series Expansion Module Blank	2

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration (continued)

Part Number	Description	Quantity
N5KUK9-413N1.1	Nexus 5000 Base OS Software Rel 4.1(3)N1(1)	2
N5020-ACC-KIT	Generic accessory kit for Nexus 5020	2
SFP-H10GB-CU3M or SFP-H10GB-CU5M	• 10GBASE-CU SFP+ Cable 3 Meter or • 10GBASE-CU SFP+ Cable 5 Meter	64
Cisco Nexus 1000V and VMware (Supporting 1 PoD)		
L-N1K-VLCPU-32	Nexus 1000V eDelivery CPU License Qty 32 (1YR Min Service)	2
VMW-VCS-3A	VMware vCenter Server Standard, 3yr 24x7 support	1
VMW-VS-ENTP-3A	VMware vSphere Enterprise Plus (1 CPU), 3yr 24x7 support	64
Cisco Nexus 2148		
N2K-C2148T-1GE	N2K 1GE FEX, 1PS, 1 Fan Module, 48x1G-BaseT+4x10GE (req SFP+)	2
SFP-10G-SR	10GBASE-SR SFP Module	4
CAB-AC-250V/13A	North America, NEMA L6-20 250V/20A plug-IEC320/C13 receptacle	4
N2K-PAC-200W	N2K-C2148 Series FEX 200W AC Power Supply	2
SFP-H10GB-CU3M or SFP-H10GB-CU5M	• 10GBASE-CU SFP+ Cable 3 Meter or • 10GBASE-CU SFP+ Cable 3 5 Meter	4
Cisco UCS C200 Support 2 Clusters 16 Servers/Total 32 Servers (1-Gbps Compute Layer)		
R200-1120402	UCS C200M1 Rack Srvr w/1PSU,DVD, w/o CPU, mem, HDD, PCIe crd	32
C200 CPU OPT	Processor Options	32
N20-X00006	2.66GHz Xeon X5550 95W CPU/8MB cache/DDR3 1333MHz	64
C200 MEM OPT	Memory Options	32
N01-M304GB1	4GB DDR3-1333MHz RDIMM/PC3-10600/dual rank 1 GB DRAMs	192
C2XX PCI EXP	PCI Express Card Options	32
N2XX-AQPCI03 or N2XX-AEPCI03	• Qlogic QLE2462, 4Gb dual port Fibre Channel Host Bus Adapter or • Emulex LPe 11002, 4Gb Fibre Channel PCIe Dual Channel HBA	32
C2XX RAIL KIT ACC	Rail Kit Accessories	32
R250-SLDRAIL	Rail Kit for the UCS 250 M1 Rack Server	32
R250-CBLARM	Cable Management for UCS 250 M1 Rack Server	32
C200 PWR SUP	Power Supply	32
R2X0-PSU2-650W	650W power supply unit for UCS C200 M1 or C210 M1 Server	32
C2XX PWR CAB	Power Cables	32
CAB-C13-C14-JMPR	Recessed receptacle AC power cord 27	32
C2XX EXPAND OPT	Expansion (Hidden)	32
R200-BBLKD	HDD slot blanking panel for UCS C200 M1 Rack Servers	128

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration (continued)

Part Number	Description	Quantity
R200-BHTS1	CPU heat sink for UCS C200 M1 Rack Server	64
R200-PCIBLKF1	PCIe Full Height blanking panel for UCS 200 M1 Rack Server	32
R200-SATACBL-001	Internal SATA Cable for a base UCS C200 M1 Server	32
Cisco UCS B-Series - Support 1 Full PoD- 4 Clusters - 64 Servers (10-Gbps Compute Layer)		
N10-S6100	UCS 6120XP 20-port Fabric Interconnect/0 PSU/2 fans/no SFP+	4
B SW IMG OPT	Software Image Options	4
N10-MGT001	UCS Manager v1.0.1	4
B SLOT 0 OPT	Slot Options	4
N10-E0080	8-port 4Gb FC/Expansion module/UCS 6100 Series	4
B PWR SUP OPT	Power Supply Options	4
N10-PAC1-550W	550W power supply unit for UCS 6120XP/100-240VAC	8
B PWR CAB OPT	Power Cables	4
CAB-C13-C14-JMPR	Recessed receptacle AC power cord 27	8
B ACC KIT OPT	Accessory Kit Options	4
N10-SACCA	Accessory kit for UCS 6120XP Fabric Interconnect	4
N20-C6508	UCS 5108 Blade Server Chassis/0 PSU/8 fans/0 fabric extender	8
SC IO OPT	I/O Module Addons	8
N20-I6584	UCS 2104XP Fabric Extender/4 external 10Gb ports	16
SC PWR SUP OPT	Power Supply	8
N20-PAC5-2500W	2500W power supply unit for UCS 5108	32
SC PWR CAB OPT	Power Cables	8
CAB-C19-CBN	Cabinet Jumper Power Cord, 250 VAC 16A, C20-C19 Connectors	32
SC EXPANSION OPT	Expansion Options (Hidden)	8
N20-BBLKD	HDD slot blanking panel for UCS B-Series Blade Servers	128
N20-FAN5	Fan module for UCS 5108	64
N01-UAC1	Single phase AC power module for UCS 5108	8
N20-FW001	UCS 5108 Blade Server Chassis FW package/DO NOT PUBLISH	8
BLADE 0	Blade Options 0	64
N20-B6620-1	UCS B200 M1 Blade Server w/o CPU, memory, HDD, mezzanine	64
G PROC OPT	Processor Options	64
N20-X00001	2.93GHz Xeon X5570 95W CPU/8MB cache/DDR3 1333MHz	128
G MEM OPT	Memory Options	64
N01-M304GB1 or N01-M308GB2	• 4GB DDR3-1333MHz RDIMM/PC3-10600/dual rank 1Gb DRAMs or • 8GB DDR3-1333MHz RDIMM/PC3-10600/dual rank 2Gb DRAMs (12 slots per blade)	64 * 12 = 768

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration (continued)

Part Number	Description	Quantity
G MEZZ OPT	Mezzanine Options	64
N20-AC0002 or N20-AE0002	• UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb or • UCS M71KR-E Emulex Converged Network Adapter/PCIe/2port 10Gb	64
G EXPAND OPT	Expansion (Hidden)	64
N20-BHTS1	CPU heat sink for UCS B200 M1 Blade Server	128
Cisco Catalyst 6509-VSS Data Center Services Node		
WS-C6509-E	Catalyst 6500 Enhanced 9-slot chassis, 15RU, no PS, no Fan Tray	2
SV33AEK9-12233SXI	Cisco CAT6000-VSS720 IOS ADVANCED ENTERPRISE SERVICES SSH	2
VS-S720-10G-3CXL	Cat 6500 Supervisor 720 with 2 ports 10GbE MSFC3 PFC3C XL	2
VS-F6K-MSFC3	Catalyst 6500 Multilayer Switch Feature Card (MSFC) III	2
VS-F6K-PFC3CXL	Catalyst 6500 Sup720-10G Policy Feature Card 3CXL	2
VS-S720-10G	Catalyst 6500 Supervisor 720 with 2 10GbE ports	2
MEM-C6K-CPTFL1GB	Catalyst 6500 Compact Flash Memory 1GB	2
BF-S720-64MB-RP	Bootflash for SUP720-64MB-RP	2
CF-ADAPTER-SP	SP adapter for SUP720 and SUP720-10G	2
MEM-C6K-CPTFL1GB	Catalyst 6500 Compact Flash Memory 1GB	2
WS-C6509-E-FAN	Catalyst 6509-E Chassis Fan Tray	2
WS-CAC-6000W	Cat6500 6000W AC Power Supply	4
CAB-AC-2500W-US1	Power Cord, 250Vac 16A, straight blade NEMA 6-20 plug, US	8
ACE20-MOD-K9=	Application Control Engine 20 Hardware	2
ACE20 SW OPT	ACE Module Software Options	2
SC6K-3.0.0A16-ACE	ACE 3.0.0A1(6) Software Release	2
ACE20 PERF LIC OPT	Performance License Options	2
ACE-04G-LIC	Application Control Engine (ACE) 4Gbps License	2
ACE-VIRT-050	Application Control Engine Virtualization 50 Contexts	2
WS-X6708-10G-3CXL	C6K 8 port 10 Gigabit Ethernet module with DFC3CXL (req. X2)	4
WS-F6700-DFC3CXL	Catalyst 6500 Dist Fwd Card- 3CXL, for WS-X67xx	4
WS-X6708-10GE	Cat6500 8 port 10 Gigabit Ethernet module (req. DFC and X2)	4
X2-10GB-SR	10GBASE-SR X2 Module	20
GLC-T	1000BASE-T SFP	4
WS-SVC-NAM-2-250S	Cisco Catalyst 6500 and Cisco 7600 Network Analysis Module	2
SC-SVC-NAM-4.2	Cisco NAM 4.2 for Cat6500/C7600 NAM	2
WS-SVC-FWM-1-K9	Firewall blade for 6500 and 7600, VFW License Separate	6
SC-SVC-FWM-4.0-K9	Firewall Module Software 4.0 for 6500 and 7600, 2 free VFW	6
FR-SVC-FWM-VC-T2	Catalyst 6500 and 7600 virtual FW licensing for 50 VF	6

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration (continued)

Part Number	Description	Quantity
SF-FWM-ASDM-6.1F	Device Manager for FWSM 4.0 for Catalyst 6500 and 7600	2
Catalyst 6509 WAN Edge Layer		
WS-C6509-E	Catalyst 6500 Enhanced 9-slot chassis, 15RU, no PS, no Fan Tray	2
SV33AEK9-12233SXI	Cisco CAT6000-VSS720 IOS ADVANCED ENTERPRISE SERVICES SSH	2
VS-S720-10G-3CXL	Cat 6500 Supervisor 720 with 2 ports 10GbE MSFC3 PFC3C XL	2
VS-F6K-MSFC3	Catalyst 6500 Multilayer Switch Feature Card (MSFC) III	2
VS-F6K-PFC3CXL	Catalyst 6500 Sup720-10G Policy Feature Card 3CXL	2
VS-S720-10G	Catalyst 6500 Supervisor 720 with 2 10GbE ports	2
MEM-C6K-CPTFL1GB	Catalyst 6500 Compact Flash Memory 1GB	2
BF-S720-64MB-RP	Bootflash for SUP720-64MB-RP	2
CF-ADAPTER-SP	SP adapter for SUP720 and SUP720-10G	2
MEM-C6K-CPTFL1GB	Catalyst 6500 Compact Flash Memory 1GB	2
WS-C6509-E-FAN	Catalyst 6509-E Chassis Fan Tray	2
WS-CAC-6000W	Cat6500 6000W AC Power Supply	4
CAB-AC-2500W-US1	Power Cord, 250Vac 16A, straight blade NEMA 6-20 plug, US	8
WS-X6708-10G-3CXL	C6K 8 port 10 Gigabit Ethernet module with DFC3CXL (req. X2)	4
WS-F6700-DFC3CXL	Catalyst 6500 Dist Fwd Card- 3CXL, for WS-X67xx	4
WS-X6708-10GE	Cat6500 8 port 10 Gigabit Ethernet module (req. DFC and X2)	4
X2-10GB-SR	10GBASE-SR X2 Module	20
MDS 9513		
MDS 9513	Cisco MDS 9513 Multilayer Director Switch	2
DS-X9224-96K9	MDS 9000 24-Port 8-Gbps Fibre Channel Switching Module with SFP and SFP+ LC connectors	4
DSX9704		6
DS-X9530-SF2-K9	MDS 9500 Series Supervisor-2 module	4
DS-X9304-18K9	18-port Fibre Channel/4-port Gigabit Ethernet Multiservice (MSM-18/4) module	2
MDS 91XX Fiber Channel Switches		
DS-C9148D-8G32P-K9	MDS 9148 with 32p enabled, 32x8GFC SW optics, 2 PS	2
CAB-9K12A-NA	Power Cord, 125VAC 13A NEMA 5-15 Plug, North America	4
DS-SFP-FC8G-SW	8 Gbps Fibre Channel SW SFP+, LC	64
DS-9148-KIT-CSCO	Accessory Kit for Cisco MDS 9148	2
DS-C9134AP-K9	MDS 9134 with 24 ports enabled with 24 SW SFPs - PL PID	2
M91S2K9-5.0.1A	MDS 9100 Supervisor/Fabric-2, NX-OS Software Release 5.0(1A)	2
M9100FMS1K9	MDS 9100 Fabric Manager Server license for 1 MDS 9100 switch	2
CAB-9K12A-NA	Power Cord, 125VAC 13A NEMA 5-15 Plug, North America	4
DS-9134-KIT-CSCO	Accessory Kit for MDS 9134 Cisco Option	2

Table A-1 Cisco VMDC Solution, Version 2.0. Compact Pod 1 Gbps or 10 Gbps Configuration (continued)

Part Number	Description	Quantity
DS-SFP-FC4G-SW	4 Gbps Fibre Channel-SW SFP, LC	48
EMC V-Max		
SB-DE15-DIR	V-MAX 15SLT DR ENCL	16
SB-DB-SPS	V-MAX SB SPS	4
SB-FE80000	V-MAX 8M FC-NO PREM	4
SB-32-BASE	V-MAX BASE-32 GB	1
SB-ADD32NDE	V-MAX ADD ENGINE-32 GB	1
NF4103001B	V-MAX 4G 10 K300GB DRIVE	88
SB-PCBL3DHR	50A 3PH DELTA HBL-RSTOL	2
SB-ACON3P-50	ADPTR AC 3PH 50A W/3/4IN CONDUIT ADPTR	4
SB-CONFIG05	V-MAX CONFIG 05	1
PP-SE-SYM	PPATH SE SYM	1
ESRS GW 100	SECURE REMOTE SUPPRT GW	1
SYMVP-RN-OPN	SYMM VIRTUAL PROV RUNTIME	1
ENGTY-SB-BAS	V-MAX ENGINUITY BASE LICENSE	1
ENGTY-SB-C02	V-MAX ENGINUITY 1TB (15-25TB)	24
M-PRESW-001	PREMIUM SOFTWARE SUPPORT	1
NF4106001BU	V-MAX 4G 10K600GB DRV UPG	48
SYMVP-RN-OPN	SYMM VIRTUAL PROV RUNTIME	1
ENGTY-SB-UPG	V-MAX ENGINUITY BASE UPGRADE	1
SYM-MIGR-BAS	SYMMETRIX MIGRATION PKG BASE LICENSE	1
ENGTY-SB-C04	V-MAX ENGINUITY 1TB (41-60TB)	25
457-100-183	POWERPATH/VE, STD X86 T2 (8+ CPUs)	128
POWERPATH-VE	POWERPATH FOR VIRTUAL ENVIRONMENT	64
NetApp (Network Attached Storage)		
FAS6080A-IB-BS2-R5	FAS6080A,IB,ACT,ACT,HW/SW, 220V,R5	1
X1107A-R6-C	NIC 2-PORT BARE CAGE SFP+ 10GbE SFP+ PCIe, -c	2
DS4243-0724-12A-R5-C	DSK SHLF, 12x2.0TB,7.2K,SATA, IOM3,-C,45	2