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

February 29, 2012

Cisco Validated Design

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

CONTENTS

Preface

This preface explains the objectives and intended audience of the Virtual Multi-tenant Data Center (VMDC) solution and outlines the organization of the Virtual Multi-tenant Data Center Design Guide.

Purpose

Infrastructure as a Service (IaaS) simplifies application development and implementation by virtualizing underlying hardware resources and operating systems. This allows IaaS users to significantly cut development and deployment times by cloning the environments best suited for an application without having to factor in the underlying hardware environment. Units of this infrastructure, including compute, storage, and networks, collectively form a cloud infrastructure.

This document describes a reference architecture that brings together core products and technologies from Cisco, NetApp, EMC, BMC, and VMware to deliver a comprehensive end-to-end cloud solution. Focused on IaaS cloud deployment, the Cisco VMDC solution provides customers with robust, scalable, and resilient options for cloud data center deployments.

This Cisco driven end-to-end architecture defines how to provision flexible, dynamic pools of virtualized resources that can be shared efficiently and securely among different tenants. Process automation greatly reduces resource provisioning and time to market (TTM) for IaaS-based services. Shared resource pools consist of virtualized Cisco unified compute and virtualized SAN and NAS storage platforms connected using Cisco data center switches and routers.

This solution addresses the following problems:

- **Inefficient Resource Utilization**—Traditionally, Enterprises design their data centers using dedicated resource silos. These silos include access switches, server racks, and storage pools assigned to specific applications and business units. This approach results in inefficient resource use, where resource pools are customized per application, resulting in fewer shared resources. This design cannot harness unused or idle resources, is complex to administer, and is difficult to scale, which results in longer deployment times. For the public cloud Service Provider, inefficient resource utilization translates to higher capital expense and operating expense and decreases revenue margins.

- **Security Guarantees**—In a multi-tenant environment, access to resources must be controlled to ensure isolation and security among users. This becomes more challenging when resources are shared. Tenants need to be assured that in new highly virtualized systems their data and applications are secure.
- **Resource Provisioning and TTM**—Facility consolidation coupled with increased deployment of virtualized servers results in larger, very dense data center systems. Manual provisioning often takes two to four weeks or longer. In many cases, this lengthy duration fails to meet business agility and time to market (TTM) requirements of Enterprises and Service Providers.
- **Complex and Expensive Administration**—Network, server, security, and application administrators must collaborate to bring up new resources for each new or expanding tenant. Collaboration based on manual methods no longer scales in these new highly virtualized systems, resulting in slow responses to business needs due to complex IT operations. It is complicated and time consuming to streamline manual configuration and resource provisioning tasks. It also increases capital and operating expenditures and overhead caused by resource churn.

As Enterprise IT departments evolve, they are looking for a data center solution that is efficiently shared, secured, and rapidly provisioned. Similarly, Service Providers are looking for solutions that enable them to reduce TTM for new revenue-generating services and reduce ongoing operating expense (OpEx). The VMDC infrastructure design provides a model for flexible sharing of common infrastructure, maintaining secure separation of tenant data and enabling per-tenant differentiated services. The VMDC Orchestration Design Section details how to rapidly provision these shared pools.

Audience

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

This design guide assumes that the reader is familiar with the basic concepts of IP protocols, QoS, DiffServ and HA. This guide also assumes that the reader is aware of general system requirements and has knowledge of Enterprise or Service Provider network and Data Center architectures.

Obtaining Documents

Cisco documentation and additional literature are available on Cisco.com. This section explains the product documentation resources that Cisco offers.

Cisco.com

- For the most current Cisco documentation, go to http://www.cisco.com/techsupport
- To access the Cisco web site, go to http://www.cisco.com
- To access international Cisco web sites, go to http://www.cisco.com/public/countries_languages.shtml

Product Documentation DVD

The Product Documentation DVD is a library of technical product documentation on a portable medium. The DVD enables you to access installation, configuration, and command guides for Cisco hardware and software products. With the DVD, you have access to the HTML documentation and some of the PDF files found on the Cisco web site at this URL: http://www.cisco.com/univercd/home/home.htm

The Product Documentation DVD is created and released regularly. DVDs are available singly or by subscription. Registered Cisco.com users can order a Product Documentation DVD (product number DOC-DOCDVD= or DOC-DOCDVD=SUB) from Cisco Marketplace at the Product Documentation Store at this URL: http://www.cisco.com/go/marketplace/docstore

Ordering Documentation

You must be a registered Cisco.com user to access Cisco Marketplace. Registered users may order Cisco documentation at the Product Documentation Store at this URL:
http://www.cisco.com/go/marketplace/docstore

If you do not have a user ID or password, you can register at this URL:
http://tools.cisco.com/RPF/register/register.do

Documentation Feedback

You can provide feedback about Cisco technical documentation on the Cisco Support site area by entering your comments in the feedback form available in every online document.

You can submit e-mail comments about technical documentation to bug-doc@cisco.com.

You can submit comments by using the response card (if present) behind the front cover of your document or by writing to the following address:

Cisco Systems
Attn: Customer Document Ordering
170 West Tasman Drive
San Jose, CA 95134-9883

We appreciate your comments

Cisco Product Security Overview

Cisco provides a free online Security Vulnerability Policy portal at this URL:

http://www.cisco.com/en/US/products/products_security_vulnerability_policy.html

From this site, you will find information about how to do the following:

- Report security vulnerabilities in Cisco products
- Obtain assistance with security incidents that involve Cisco products
- Register to receive security information from Cisco

A current list of security advisories, security notices, and security responses for Cisco products is available at this URL: http://www.cisco.com/go/psirt

To see security advisories, security notices, and security responses as they are updated in real time, you can subscribe to the Product Security Incident Response Team Really Simple Syndication (PSIRT RSS) feed. Information about how to subscribe to the PSIRT RSS feed is found at this URL:

http://www.cisco.com/en/US/products/products_psirt_rss_feed.html

Cisco is committed to delivering secure products. We test our products internally before we release them, and we strive to correct all vulnerabilities quickly. If you think that you have identified a vulnerability in a Cisco product, contact PSIRT:

For **emergencies** only—security-alert@cisco.com

An emergency is either a condition in which a system is under active attack or a condition for which a severe and urgent security vulnerability should be reported. All other conditions are considered nonemergencies.

For **nonemergencies**—psirt@cisco.com

In an emergency, you can also reach PSIRT by telephone:

1 877 228-7302
1 408 525-6532

Tip We encourage you to use Pretty Good Privacy (PGP) or a compatible product (for example, GnuPG) to encrypt any sensitive information that you send to Cisco. PSIRT can work with information that has been encrypted with PGP versions 2.x through 9.x.

Never use a revoked encryption key or an expired encryption key. The correct public key to use in your correspondence with PSIRT is the one linked in the Contact Summary section of the Security Vulnerability Policy page at this URL
http://www.cisco.com/en/US/products/products_security_vulnerability_policy.html

The link on this page has the current PGP key ID in use.

If you do not have or use PGP, contact PSIRT to find other means of encrypting the data before sending any sensitive material.

Product Alerts and Field Notices

Modifications to or updates about Cisco products are announced in Cisco Product Alerts and Cisco Field Notices. You can receive these announcements by using the Product Alert Tool on Cisco.com. This tool enables you to create a profile and choose those products for which you want to receive information.

To access the Product Alert Tool, you must be a registered Cisco.com user. Registered users can access the tool at this URL:

http://tools.cisco.com/Support/PAT/do/ViewMyProfiles.do?local=en

To register as a Cisco.com user, go to this URL:

http://tools.cisco.com/RPF/register/register.do

Obtaining Technical Assistance

Cisco Technical Support provides 24-hour-a-day award-winning technical assistance. The Cisco Support web site on Cisco.com features extensive online support resources. In addition, if you have a valid Cisco service contract, Cisco Technical Assistance Center (TAC) engineers provide telephone support. If you do not have a valid Cisco service contract, contact your reseller.

Cisco Technical Support Website

The Cisco Technical Support Web site (http://www.cisco.com/tac) provides online documents and tools for troubleshooting and resolving technical issues with Cisco products and technologies. The Cisco Technical Support Web site is available 24 hours a day, 365 days a year.

Accessing all the tools on the Cisco Technical Support Web site requires a Cisco.com user ID and password. If you have a valid service contract but do not have a login ID or password, please register at this URL:

http://tools.cisco.com/RPF/register/register.do

To ensure that all cases are reported in a standard format, Cisco has established case priority definitions.

- **Priority 1 (P1)**—Your network is down or there is a critical impact to your business operations. You and Cisco will commit all necessary resources around the clock to resolve the situation.
- **Priority 2 (P2)**—Operation of an existing network is severely degraded, or significant aspects of your business operation are negatively affected by inadequate performance of Cisco products. You and Cisco will commit full-time resources during normal business hours to resolve the situation.
- **Priority 3 (P3)**—Operational performance of your network is impaired, but most business operations remain functional. You and Cisco will commit resources during normal business hours to restore service to satisfactory levels.
- **Priority 4 (P4)**—You require information or assistance with Cisco product capabilities, installation, or configuration. There is little or no effect on your business operations.

Submitting a Service Request

Using the online TAC Service Request Tool is the fastest way to open S3 and S4 service requests. (S3 and S4 service requests are those in which your network is minimally impaired or for which you require product information.) After you describe your situation, the TAC Service Request Tool provides recommended solutions. If your issue is not resolved using the recommended resources, your service request is assigned to a Cisco engineer. The TAC Service Request Tool is located at this URL:

http://www.cisco.com/techsupport/servicerequest

For S1 or S2 service requests, or if you do not have Internet access, contact the Cisco TAC by telephone. (S1 or S2 service requests are those in which your production network is down or severely degraded.) Cisco engineers are assigned immediately to S1 and S2 service requests to help keep your business operations running smoothly.

To open a service request by telephone, use one of the following numbers:

- Asia-Pacific: +61 2 8446 7411
- Australia: 1 800 805 227

- EMEA: +32 2 704 55 55
- USA: 1 800 553 2447

For a complete list of Cisco TAC contacts, go to this URL:

http://www.cisco.com/techsupport/contacts

To ensure that all service requests are reported in a standard format, Cisco has established severity definitions.

- **Severity 1 (S1)**—An existing network is "down" or there is a critical impact to your business operations. You and Cisco will commit all necessary resources around the clock to resolve the situation.
- **Severity 2 (S2)**—Operation of an existing network is severely degraded, or significant aspects of your business operations are negatively affected by inadequate performance of Cisco products. You and Cisco will commit full-time resources during normal business hours to resolve the situation.
- **Severity 3 (S3)**—Operational performance of the network is impaired while most business operations remain functional. You and Cisco will commit resources during normal business hours to restore service to satisfactory levels.
- **Severity 4 (S4)**—You require information or assistance with Cisco product capabilities, installation, or configuration. There is little or no effect on your business operations.

Obtaining Additional Publications and Information

Information about Cisco products, technologies, and network solutions is available from various online and printed sources.

- The Cisco Online Subscription Center is the web site where you can sign up for a variety of Cisco e-mail newsletters and other communications. Create a profile and then select the subscriptions that you would like to receive. To visit the Cisco Online Subscription Center, go to:

  http://www.cisco.com/offer/subscribe

- The Cisco Product Quick Reference Guide is a handy, compact reference tool that includes brief product overviews, key features, sample part numbers, and abbreviated technical specifications for many Cisco products that are sold through channel partners. It is updated twice a year and includes the latest Cisco channel product offerings. To order and find out more about the Cisco Product Quick Reference Guide, go to this URL:

  http://www.cisco.com/go/guide

- Cisco Marketplace provides a variety of Cisco books, reference guides, documentation, and logo merchandise. Visit Cisco Marketplace, the company store at:

  http://www.cisco.com/go/marketplace/

- Cisco Press publishes a wide range of general networking, training, and certification titles. Both new and experienced users will benefit from these publications. For current Cisco Press titles and other information, go to Cisco Press at this URL:

  http://www.ciscopress.com

- Internet Protocol Journal is a quarterly journal published by Cisco for engineering professionals involved in designing, developing, and operating public and private internets and intranets. You can access the Internet Protocol Journal at:

http://www.cisco.com/ipj

- Networking products offered by Cisco, as well as customer support services, can be obtained at:

  http://www.cisco.com/en/US/products/index.html

- Networking Professionals Connection is an interactive web site where networking professionals share questions, suggestions, and information about networking products and technologies with Cisco experts and other networking professionals. Join a discussion at:

  http://www.cisco.com/discuss/networking

- “What's New in Cisco Documentation” is an online publication that provides information about the latest documentation releases for Cisco products. Updated monthly, this online publication is organized by product category to direct you quickly to the documentation for your products. You can view the latest release of “What's New in Cisco Documentation” at:

  http://www.cisco.com/univercd/cc/td/doc/abtunicd/136957.htm

- World-class networking training is available from Cisco. You can view current offerings at:

  http://www.cisco.com/en/US/learning/index.html

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 http://www.cisco.com/go/validateddesigns.

CHAPTER 1

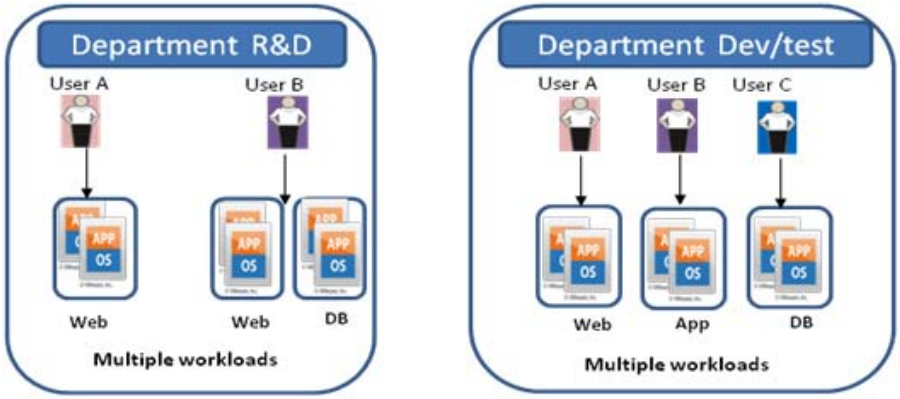
Design Overview

This document presents design and implementation guidance for a private or public IaaS cloud data center. The intended deployment uses a multi-tenant, differentiated service tier model.

Cloud Tenants

A tenant is an entity that subscribes to cloud services. In the Enterprise private cloud deployment model, that entity is a department or sub-organization, such as development, test, research and development, or human resources. Figure 1-1 shows multiple users in the same department belong to the same tenancy. Within the tenancy, multiple workloads can be implemented by different users who belong to the same department.

Figure 1-1 Tenants and Workloads



In the public cloud deployment model, a tenant is an individual consumer, an Enterprise, or a sub-organization within an Enterprise subscribing to the virtual private cloud services hosted by a Service Provider.

Each tenant must be securely separated from other tenants who share the common virtualized resource pool. However, workloads owned by one tenant are visible to others unless firewalls are configured to block communications among different tenant applications.

Differentiated Cloud Services

Cloud providers, whether Service Providers or Enterprises, want an IaaS offering with multiple feature tiers and pricing levels. The cloud is a 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 consider the functionality and value provided by the service only; they do not need to understand or manage the underlying technology.

To tailor workload or application requirements to specific customer needs, the cloud provider can differentiate services with a multi-tiered service infrastructure and quality of service (QoS) settings. Such services can be used and purchased under a variable pricing model. Infrastructure and resource pools can be designed so that end users can add or expand services by requesting additional compute, storage, or network capacity. This elasticity allows the provider to maximize the user experience by offering a custom, private data center in virtual form.

Typically, cloud providers want to offer three, four, or five different service tiers and provide different service level agreements (SLAs). IaaS cloud services can be differentiated into pre-defined service tiers by varying support of the following features:

- **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 Features**—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.
- **Application Tiers**—Service tiers can provide differentiated support for application hosting. In some instances, applications may require several application tiers of VMs. Often, each tier is placed on separate VLANs. For example, a Gold profile could have three application tiers on three separate VLANs to host web, application, and database (DB) services on different VMs. Each tier could provide five 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 two VMs for redundancy and load balancing. In contrast, a Bronze profile could have three tiers but in a less differentiated manner, with the web, application, and DB services residing on the same VLAN or potentially on the same VM.
- **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 a key differentiator. 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.

The VMDC solution defines options for differentiating IT cloud services. In this reference architecture, these cloud services are called service tiers. Typically when we talk about service tiers, we look at the server CPU and storage options. But if a web application is hosted in the cloud model, load balancing and firewall inspection are also required. To achieve secure separation of tenant data, Layer 2 and Layer 3 features, such as virtual routing and forwarding (VRF) and VLANs, must be enabled. With this virtual network separation configured, service tiers contain virtual compute, storage, and network resources.

Deployment Models

The Cisco VMDC solution qualifies a three-tier model of Bronze, Silver, and Gold tiers comprising IaaS services. These tiers define service levels for compute, storage, and network performance (Table 1-1).

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

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

Using this tiered model, you can do the following:

- Offer service tiers with well-defined and distinct SLAs
- Support customer segmentation based on desired service levels and functionality
- Allow for differentiated application support based on service tiers

Cloud Service Tiers

In the Cisco VMDC solution, three service tiers are defined: Bronze, Silver, and Gold. Each service tier is a container that is assigned specific network, compute, and storage resources. In the following sections, we explain how to differentiate service tiers in your cloud and the resources that those tiers can contain.

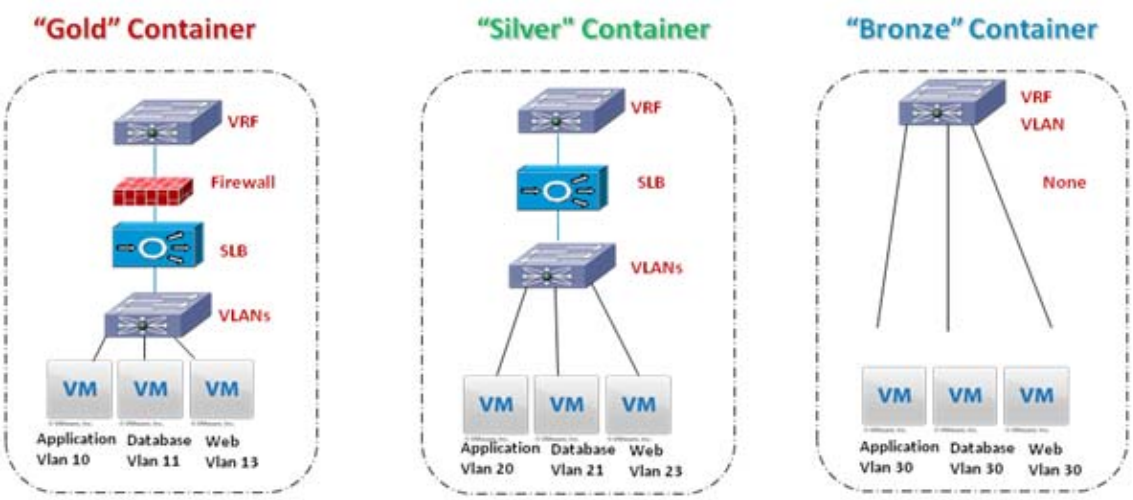
The Cisco VMDC solution allocates network, compute, and storage resources according to the Bronze, Silver, and Gold service tiers. The following sections identify differences in resources among the three tiers:

- Network Resources, page 1-4
- Compute Resources, page 1-4
- Storage Resources, page 1-5

Network Resources

Figure 1-2 depicts the network components assigned to Bronze, Silver, and Gold service tiers in the VMDC solution. The Gold, Silver, and Bronze tiers present predefined baseline network configuration choices through a self-service portal from which a tenant can select the container.

Figure 1-2 Network Resources by Service Tier



These network containers are enabled by end-to-end virtualization of the network infrastructure. The Cisco VMDC solution leverages end-to-end VRF-Lite, VLAN, and virtualized services, such as virtual firewall and load balancing contexts.

Each service tier uses a unique VRF-Lite instance per tenant to provide a dedicated virtual network (or virtual private data center). Depending on the application, multiple application tiers may exist, such as the hosted web, application, and database tiers of an e-commerce application. Each tier of this application resides in a separate VLAN within the VRF-Lite instance. For each Silver and Gold tenant, a unique VRF and three VLANs are provisioned. The Gold tenant is further differentiated by a dedicated virtual firewall and load balancing services, whereas Silver tenants receive only a dedicated virtual load balancing service. Being a best effort service, the Bronze tenant is assigned a unique VRF-Lite instance for each tenant, but all three tiers of an application must share the same 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, datastore protection, business continuity, or disaster recovery.

Compute Resources

In the VMDC 2.0 system, at the compute layer, service tier differentiation was modelled based on three compute workload sizes called Small, Medium, and Large. From an application perspective, key characteristics to consider are vCPU and RAM. Server virtualization runs multiple virtual servers on a single blade server.

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 servers based on the Intel Xeon series processor. Each socket has four cores with a total of eight cores, or 8 vCPUs, per blade. As Table 1-2 shows, 32 Small VMs per physical host were enabled

by allocating 0.25 vCPU for each virtual machine whereas Large has a dedicated vCPU for each VM, limiting the total Large workloads to 8 per blade server. Effectively, this comprises a compute oversubscription factor (OSF) of 4:1, 2:1 and 1:1.

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

Table 1-2 Compute Resources by Size

	Small	Medium	Large
vCPUs per core	0.25 vCPU	0.5 vCPU	1 vCPU
VMs per blade	32 VMs	16 VMs	8 VMs
RAM (GB)	4	8	16

Storage Resources

The Cisco VMDC architecture defines three persistent storage workload sizes called Small, Medium, and Large. Datastore retention and availability is a major concern for customers of cloud-based offerings. Thus, a baseline premise of the system is that a tiered retention, protection, and recovery model will insure that storage availability and reliability may be tailored to meet tenant requirements.

Table 1-3 lists the workload options and storage resources sizes.

Table 1-3 Storage Services by Size

	Small	Medium	Large
Base storage (GB)	50	150	300
Storage growth increment (GB)	50	50	50
Backup (retention length options)	1 mo., 6 mo., or 1yr.	1 mo., 6 mo., or 1yr.	1 mo., 6 mo., or 1yr.
Data protection	None	Snap - Virtual copy (local site) SNAP copies every 8 hrs.; 36 hr. retention	Clone - Mirror copy (local site) - SNAP copies every 4 hrs.; 36 hr. retention
Disaster recovery	None	Remote replication Symmetrix Remote Data Facility (SRDF)	Remote replication SRDF

You can further refine the service tiers by differentiating the backup and recovery options. To ensure data protection and durability, 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

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

Table 1-4 Service Tier Distinctions for Storage (continued)

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

Per Tenant Logical Flow

First, a tenant chooses the network container that provides him a virtual dedicated network within the shared infrastructure. A tenant has three selections to choose from for their network container - the provided network services and level of separation. Cisco VMDC defines the following network container selections: Gold, Silver and Bronze.

Second, a tenant chooses the compute and storage resources to run inside the container. Cisco VMDC defines three workload sizes for compute and storage resources: Small, Medium, and Large. A tenant's choice of a workload size depends heavily on the application being implemented in the container.

Therefore, a network container comprises the services, resources, and service path through the infrastructure. The concept of a network containers is introduced and embodied in the Cisco VMDC solution.

Solution Objectives

The Cisco VMDC solution targets the following objectives:

- Expedite ordering with a single bill of materials (BOM).
- Validate interoperability end to end.
- Support incremental investment via modular build out.
- Provide security at every layer.
- Provide an application aware network.
- Reduce power and space requirements per server.
- Increase performance for network-based backup and data replication.
- Support service tiers to quickly scale up without wasting resources.
- Enable workload portability.
- Support end-to-end service orchestration and automated provisioning.

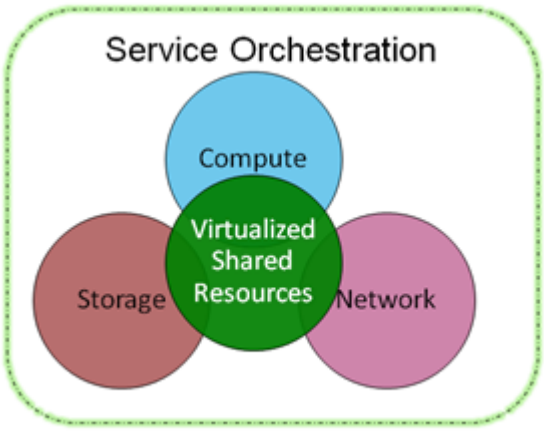
CHAPTER 2

Architecture Overview

A cloud deployment model is distinctive from traditional deployments in its ability to treat the data center as a common fabric of resources available in an on-demand basis. A portion of these pools are dynamically allocated to individual tenants and then deallocated when they are no longer in use. As depicted in Figure 2-1, VMDC defines two key building blocks:

- **Virtualized Shared Resource Pool**—The resource pool consists of network, compute, and storage components. These components are virtualized and used by multiple tenants securely.
- **Service Orchestration**—Service orchestration automates the resource provisioning workflow. It leverages a set of tools and APIs to dynamically provision cloud resources on demand. A tenant initiates the workflow process using a web portal to request specific resources.

Figure 2-1 VMDC Building Blocks



Note This document addresses design aspects of the shared resource pool that a customer must understand before implementing a cloud data center. It does not address service orchestration components and design. A separate module of the VMDC 2.0 document provides service orchestration design and implementation guidance.

When designing an IaaS architecture and the shared resources pools, network architects should consider the following design goals:

- **Secure Separation**—Provides end-to-end tenant path isolation and security. Tenants are isolated from each other via several security techniques at different layers of the network or infrastructure. For example, virtual route forwarding instances (VRFs) are leveraged at the Layer 3 to stop

communication between tenants at Layer 3 domain. Likewise, similar isolation features are leveraged at compute and storage layers to provide complete isolation of tenants in a shared infrastructure.

- **Data Center Scalability**—A pod-based architecture provides network architects the ability to modularize the infrastructure into easily replicable units called pods. Architects can plan for an initial pod, which guarantees a certain scale and performance along with a scalable data center core network. This architecture provides a predictable and homogeneous method for adding self-contained pods as additional resources are needed.
- **High Availability**—Availability ensures that the cloud resources are accessible even during a failure situation. Availability is required to meet the expectations of service-level agreements (SLAs) in a cloud deployment.
- **Service Assurance**—Provides mechanisms to define different service levels and defines how to adhere to them using network QoS techniques during both steady and non-steady states. To differentiate IaaS service tiers, network architects can reserve and guarantee certain network bandwidths based on their subscription rules for the tier. For example, a Gold tenant could be guaranteed with 1 Gbs of bandwidth per VM whereas a Silver tenant only gets 0.5 Gbs per VM.

Table 2-1 presents example storage distinctions by service tier.

Table 2-1 Components of the Large Pod Resource Pool

Features	Components
Network	Cisco CRS Cisco Nexus 7010 Cisco Nexus 7018 Data Center Services Node 6509-E (VSS) Firewall Service Module Application Control Engine Module
Compute	Cisco Unified Computing System (UCS) • UCS5108 Blade Server Chassis • UCSB200-M1 Blade Server • UCS M71KR-E Converged Network adapter • UCS M81KR Virtual Interface card • Cisco UCS 6120, Cisco UCS 6140 fabric interconnect
Virtualization	VMware vSphere VMware ESXi 4.0U1 Hypervison Cisco Nexus 1000V (virtual access switch)
Security	Cisco Firewall Services Module (FWSM), ACE Application Control Engine VMware vShield NetApp vFiler and Virtual Service Domains MDS soft zoning and VSANs Cisco Nexus 1000V
Storage Fabric	Cisco MDS 9513

Table 2-1 Components of the Large Pod Resource Pool (continued)

Features	Components
Storage Array	EMC 2 Symmetrix VMAX with virtual provisioning NetApp FAS3170

Orchestration/Management	Domain Management:
	• UCS Manager
	• Nexus 1000V VSM
	• VMware vCenter
	• Fabric Manager
	BMC Cloud Lifecycle Management (CLM):
	• BMC BladeLogic Server Automation
	• BMC BladeLogic Network Automation
	• BMC Remedy Action Request Suite
	  • Service Request Manager (SRM)
	  • Atrium Core
	  • Atrium Orchestrator
	  • Remedy AR System Server
	  • Cloud Extension Pack
	• BMC Remedy Change Management

Discrepancy analysis	Determines if configuration deltas exist within the Cisco CSGs defined in the server farm Provides a report with found discrepancies; shows any diverging Cisco IOS ® Software CLIs between Cisco CSGs

End-to-End Topologies

Figure 2-2 shows the end-to-end logical topology for Gold, Silver, and Bronze service classes.

Figure 2-2 End-to-End Logical Topology (Gold, Silver, Bronze)

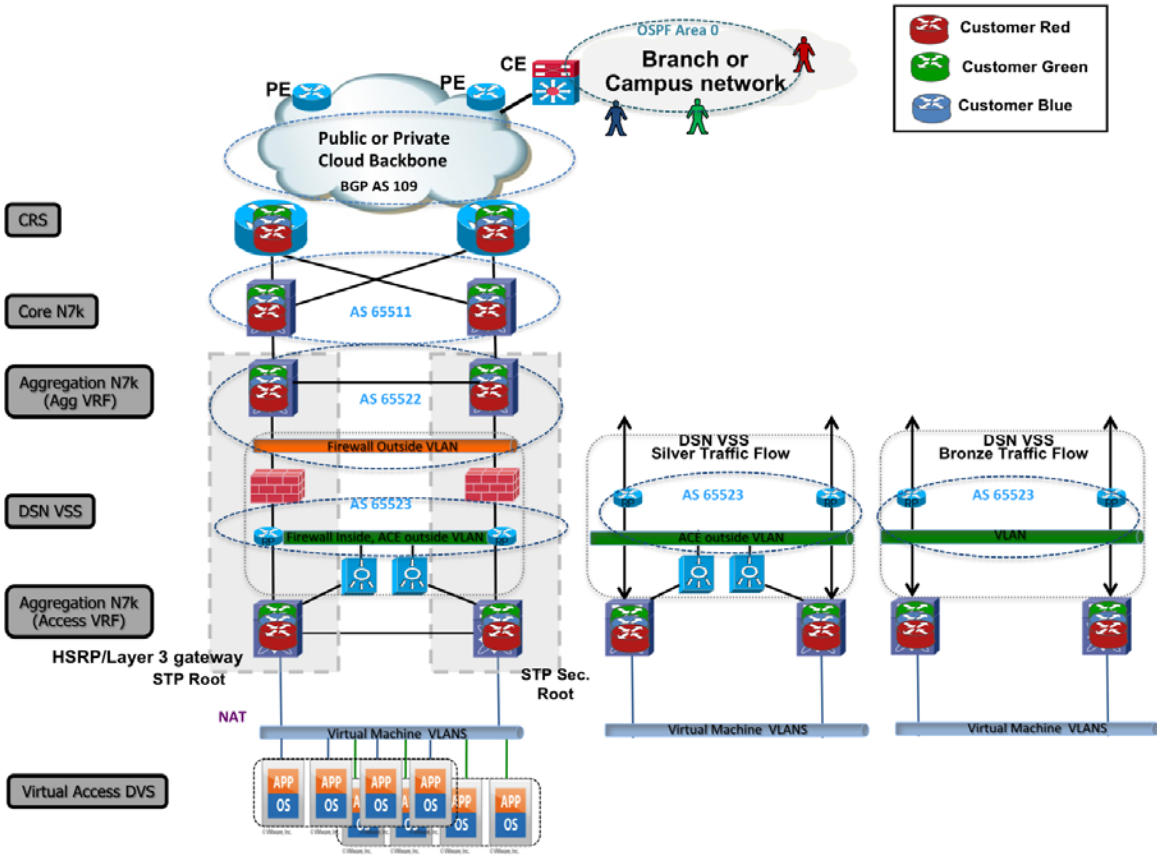
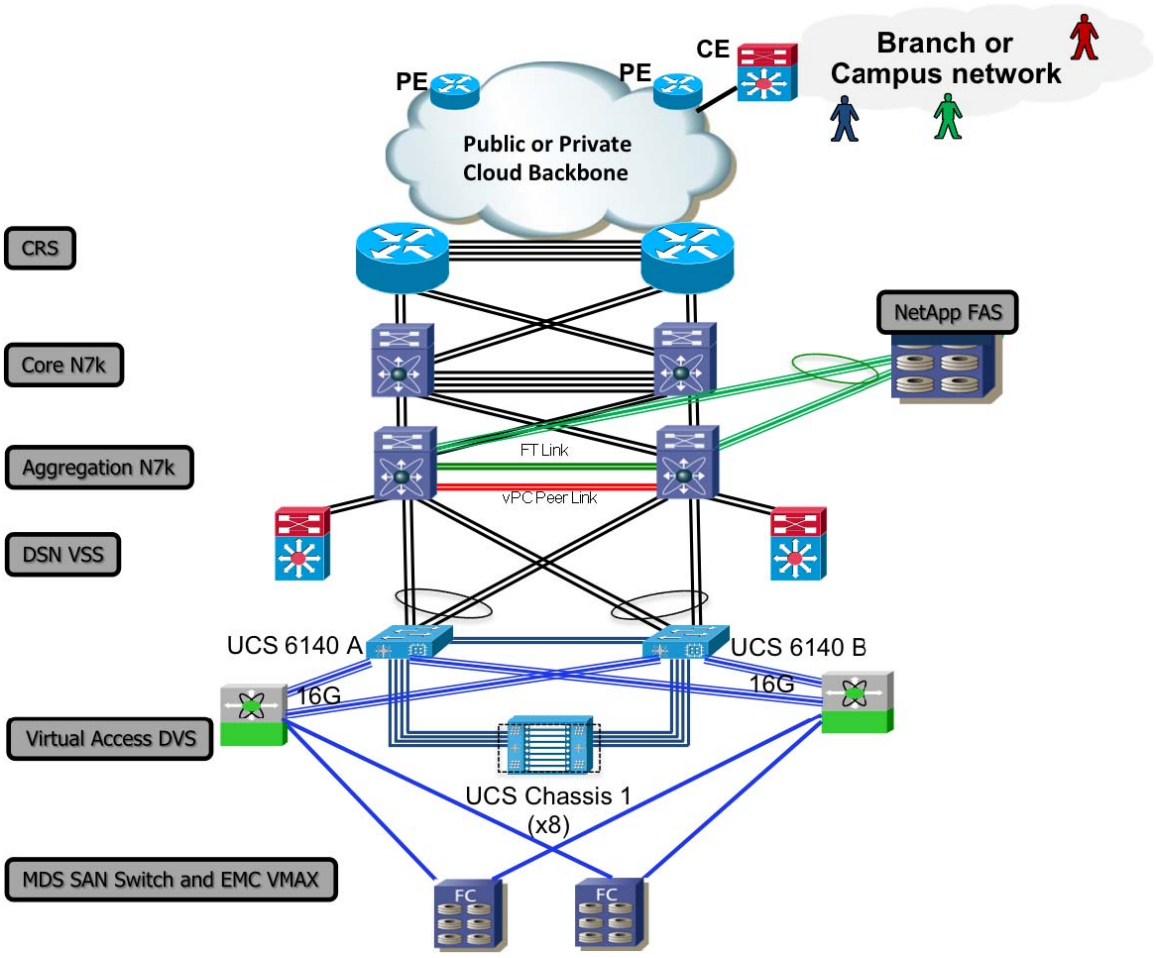


Figure 2-3 the end-to-end physical topology.

Figure 2-3 End-to-End Physical Topology (Gold, Silver, Bronze)



CHAPTER 3

Design Considerations

In the Cisco VMDC solution, service tiers are the basis of the end-to-end design of the cloud data center. When deploying such a data center, the following design requirements must be considered:

Data Center Scalability

Data center scalability is a multi-dimensional attribute that refers to the capability of a platform within the system and the ability to provision, manage, and operate the system. As such, some of the key considerations for scaling the data center include the following:

Multi-Tenancy—The key to scaling data center resources without increasing the capital expense (CAPEX) and operating expense (OPEX) of the data center depends on the ability to virtualize hardware resources and support a multi-tenant environment. Multi-tenancy requires a flexible granular resource container definition that allows for adding services and capacity on-demand.

Elasticity—Elasticity is the ability to scale resources up or down in a short period based on service-level agreements (SLAs). Elasticity enables resources on demand and scale resource utilization as needed. The cloud architecture is based on the resource container, where a dedicated segment of cloud infrastructure provides guaranteed access to compute, storage capacity, and relevant services. Server-based elasticity allows tenants access to the compute and memory pool when required and releases those resources when not in use, providing a system where the tenant pays only for resources utilized.

Redundancy and Fault Recovery—When scaling a data center, redundancy and failure recovery procedures add reliability and business continuity for tenants hosted on the data center. This ability should be introduced in the resource container so that it is a completely self-sufficient unit of the data center itself.

Automation Management and Provisioning—To efficiently scale the data center once the resource container is defined, proper procedures to provision and manage individual elements of the resource container should also be considered. This refers to the management plane provided by the service orchestration piece, which is outside the scope of this document.

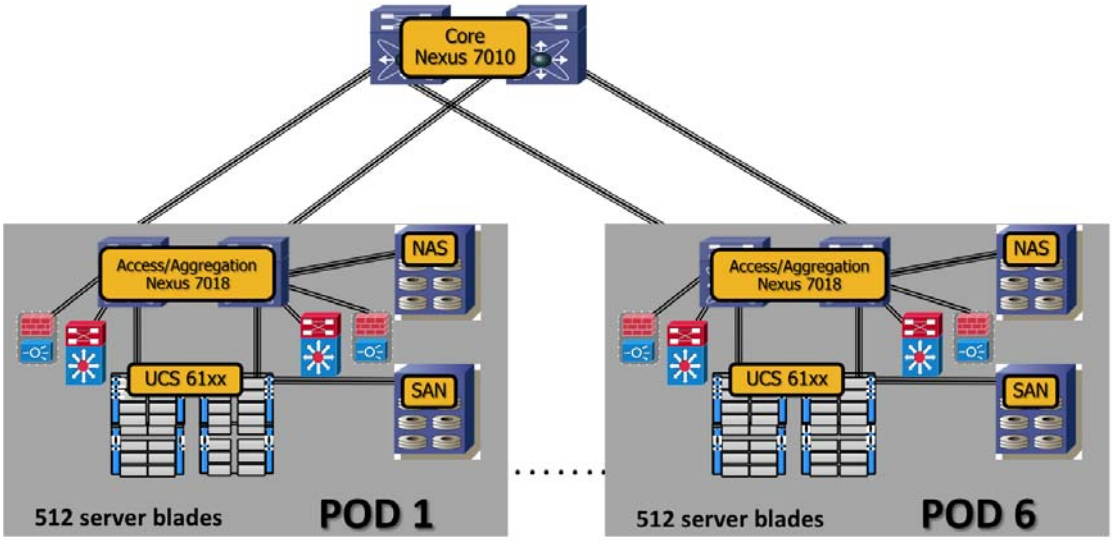
Typically, a resource container allows scaling server instances and storage logical unit number (LUN) sizes. The network fabric connecting the compute and storage pools should be able to accommodate the network bandwidth and port density scale required for demanding growth. Capacity planning is an

important aspect of the cloud data center design. As no customer can build resources at infinite scale, the optimal and cost-effective solution adds resource pools to existing data centers based on projected usage without disrupting the existing network. The resource container concept, called pod, achieves elasticity, simplifies capacity planning, and does not disrupt the existing environment.

Pod

A pod is a discrete, homogeneous, modular unit of data center components. Because they are homogeneous and modular, pods support templates for incremental build-out of the data center that address environmental, physical, logical, and application requirements. This modular architecture provides a predictable set of resource characteristics per unit that is added repeatedly as needed. In some cases, a multi-tier Layer 2 access model (or virtualization features within the access layer) may define the boundaries of a pod differently. Many possible variations exist for using the pod concepts to scale the data center topology. In this discussion, the access-layer switch pair, storage, and compute resources are in the data center pod (Figure 3-1).

Figure 3-1 Pod Expansion Concept

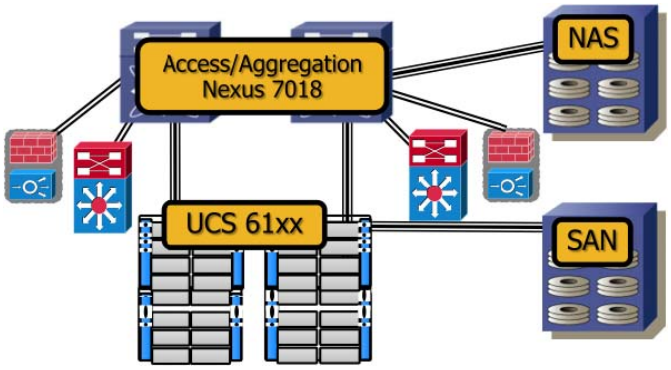


Initially, a customer implements the data center using a base pod and adding more pods to expand the data center. In the data center, pod-based architectures provide predictable resource pools, power, and space consumption. In Figure 3-1, the core layer is common to multiple pods. As additional pods are needed, they are connected to the network via the aggregation layer.

Using the pod-based architecture, you can scale a data center in a predictable manner. Once the architect is familiar with the base pod performance and scalability characteristics, additional pods can be integrated with confidence without having to re-validate the new pod. The Cisco VMDC design takes into consideration network, compute, and storage scale requirements of a base pod and reflects what components must be repeatable to satisfy cloud data center scale requirements. The design was then validated at scale within the pod and from the pod to the core layer.

Figure 3-2 presents the access-layer network, compute, and storage (IP and block) resources in the data center pod. Pod resources must be consistently repeated to achieve predictable performance. We recommend consistent sizing of the network, compute, and storage resources within a pod to achieve predictable performance. When between 80 and 90% of the resources in the pod are consistently used at capacity, add a second matched pod. Do not extend the resources in the second pod relative to the first.

Figure 3-2 Data Center Pod Components



Large Pod Scale

In the VMDC solution, a single large pod can scale to a maximum of 512 Cisco Unified Computing Services (UCS) servers. Network architects may start their cloud with a smaller size pod with less attached risk, and they can also grow when needed by adding additional large pods to the existing core network. The total number of pods supported per pair of infrastructure core nodes is essentially driven by the port capacity and MAC address scale of the nodes at the aggregation layer.

The VMDC solution addresses the general compute cloud data center deployments. Table 3-1 identifies the number of workloads that can be enabled in the large pod. As described in Cloud Service Tiers, page 1-3, different workload requirements occur in a typical cloud model. In the VMDC architecture, we refer to small, medium, and large workload sizes. Table 3-1 identifies the number of workloads that can be implemented using a large pod with a sample workload mix of 50% small, 30% medium, and 20% large.

Table 3-1 lists the workload options and storage resources sizes.

Table 3-1 Cisco UCS Configuration by Tier

Workload Type	Mix Ratio	No. of Blades	No. of Cores	VMs per core	VMs	Storage per VM
Small	50%	256	256*8 = 2,048	4	8,192	50GB
Medium	30%	154	154*8 = 1,232	2	2,464	150GB
Large	20%	102	102*8 = 816	1	816	300GB
TOTAL	100%	512	—	—	11,472	—

Of the available 512 Cisco UCS blades in Table 3-1, 20% are reserved for the large workload, which is 102 blades. A Xeon 5570-based UCS B Series blade has two CPU sockets that accept a quad-core CPU; therefore, assuming a vCPU per core, a total of 816 vCPUs are available to the large workloads. Allocating one vCPU per workload yields a total of 816 large workloads. If we calculate the same for each workload type, we find 11,472 general compute workloads available in a large pod.

Large Pod Scalability Considerations

To ensure scalability of the large pod, Cisco VMDC 2.0 addresses the following design considerations:

- Layer 2 Scale, page 3-4
- Layer 3 Scale, page 3-5
- Access or Aggregation Layer Platform, page 3-6

Layer 2 Scale

Efficiency of resource utilization and multi-tenant solutions are directly dependent on the amount of virtualization implemented in a data center. Scale of the VMs drives the scale requirement of the network components in terms of port densities, and L2 and L3 capacity.

VMs per CPU Core

Server virtualization provides the ability to run multiple server instances in a single physical blade. Essentially, this involves allocating a portion of the processor and memory capacity per VM. Processor capacity is allocated as "Virtual CPUs" (vCPUs) by assigning a portion of the processor frequency. In general parlance a vCPU is often equated to a blade core. Cisco UCS B Series blade servers have two sockets, each supporting four to eight cores. B Series blade servers equipped with the Xeon 5570 processors support four cores per socket or eight total cores. The number of VMs enabled on each blade depends on the workload type and the CPU and memory requirements. Workload types demand different amounts of compute power and memory, e.g. desktop virtualization with applications such as web browser and office suite would require much less compute and memory resources compared to a server running a database instance or VoIP or video service. Similarly, CaaS provides another dimension to the VM scale requirements, since it provides raw compute and memory resources on-demand, agnostic to the applications running.

VMs per CPU core drive the number of network interfaces (virtual) required to provide access to VMs.

Note In this document, we assume a vCPU per core. However, this only partially represents what type of VM is presented to a guest OS. Depending on the hypervisor functionality, it may be possible to specify both vCPUs per VM and cores per socket per VM. As an example, one could allocate eight vCPUs per VM and four "cores" per socket to present a virtual form of a two quad-core processor system to the Guest OS. Similarly, one could allocate two "cores" and eight vCPUs, presenting 4 dual-core processors to the guest OS. This allows the flexibility to increase processor allocation per VM while staying within licensing limits.

VMNics per VM

Each instance of VM uses a virtual NIC, which is an instance of the physical NIC. This virtualization allows the abstraction of the physical interface from the network function (L2 and L3 logic). Cisco UCS M81KR provides 128 virtual NICs on a single mezzanine card. The vNICs can be used as Ethernet or fiber-channel interfaces. VM-link implementation on the vNIC enables packet forwarding for the VMs using the same physical interface in hardware.

MAC Addresses

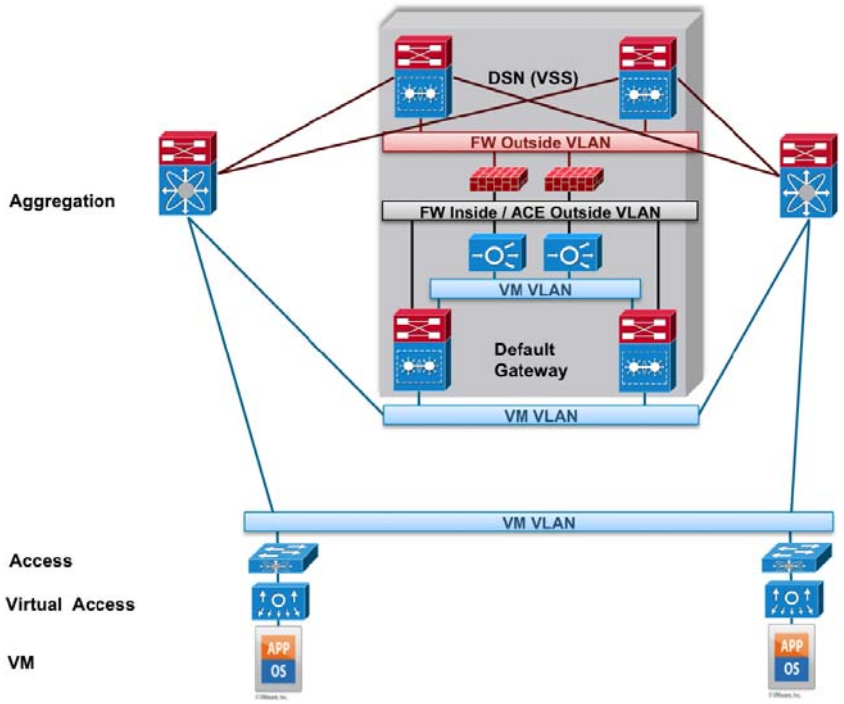
Typically, each VM has multiple vNICs; a minimum of two vNICs is recommended: one for data and one for VM management. Besides these, there are vNICs on the ESX host for VMkernal traffic, such as vMotion, service console, NFS, and DRS/HA. When calculating the scalability needs of the access layer, a minimum of two MAC addresses per VM and 2 per blade is recommended. In modelling of the MAC address requirements for the VMDC architecture a more conservative number of four MAC addresses per host was used. Applied to a large pod, MAC addresses expected at the access layer switch range between 16K-64K, depending on the workload type as per Table 3-1. The calculation to estimate this number is as follows:

(no. of server blades=512) x (no. of cores=8) x (number of VMs/core=1,2,4) x (no. of MACs/VM=4) = total no. of MACs per large pod.

VLANs

VLANs are a L2 option to scale the VM connectivity, providing application tier separation and multi-tenant isolation. They are also utilized to chain services in the path of the packet flow to provide SLA guarantees, as shown in Figure 3-3.

Figure 3-3 VLAN Usage to Provide Service Chaining



L2 Control Plane

When building L2 access/aggregation layers, the L2 control plane also must be designed to address the scale challenge. Placement of the spanning-tree root is key in determining the optimum path to link services as well as providing a redundant path to address network failure conditions. In order to provide uniformity in the network virtualization independent of equipment connected to the L2 network, it is important to support a variety of spanning-tree standards, including 802.1ad, RSTP, MSTP, and PVST.

Layer 3 Scale

The large pod design includes use of VRF-Lite in the core and aggregation layers to enable segmentation of tenants hosted on the common physical infrastructure. VRF-Lite completely isolates L2 and L3 control and forwarding planes of each tenant, allowing the flexibility in defining an optimum network topology for each tenant.

A VRF-Lite instance uses mp-BGP to advertise VPN information between core/edge and aggregation layers. Each VPN contains a set of VLANs and corresponding HSRP interfaces to provide a L3 path. The service segment extending from the service node to the DC-edge also uses an L3 path which is VRF-aware and hence consumes L3 resources.

Scaling the L3 domain depends on the following:

BGP Peering—Peering is implemented between the edge, core and the aggregation layers. The edge layer terminates the IP/MPLS VPNs and the Internet traffic in a VRF and applies SSL/IPSec termination at this layer. The traffic is then fed to the core layer via VRF-Lite. Depending on the number of data centers feeding the edge layer, the BGP peering is accordingly distributed. Similarly, depending on the number of pods feeding a data-center core layer, the scale of BGP peering decreases as we descend the layers.

HRSP Interfaces—Used to virtualize and provide a redundant L3 path between the services, core, edge, and aggregation layers.

VRF Instances—A VRF instance can be used to define a single network container representing a service class as defined in Table 3-1. The scaling of VRF instances depends on the sizing of these network containers.

Routing Tables and Convergence—Though individual tenant routing tables are expected to be small, scale of the VRF (tenants) introduces challenges to the convergence of the routing tables upon failure conditions within the data center.

Services—Services consume IP address pools for NAT and load balancing of the servers. Services use contexts to provide tenant isolation.

Access or Aggregation Layer Platform

Large pod scale is based on extending the L3 to the aggregation layer and using L2 to scale the access layer. Characteristics such as core-to-VM ratio, service class (Gold, Silver, and Bronze) distribution, bandwidth over-subscription ratio, VRF and MAC scale, and aggregation layer port density determine scale models that can be implemented.

Table 3-1 illustrates a simple scale model. Using an eight-chassis UCS cluster with B200 M1 blades and all Bronze VMs with 1:4 core:VM ratio; leads to 32 VMs per blade, 2048 per cluster, and 16k per pod of 512 blades. With a 1:1, 1:2 and 1:4 core:vm ratio for Gold/Silver/Bronze, and 20/30/50 distribution between service classes (Gold, Silver, Bronze); leads to an average of 21 VMs per blade, 1344 VMs per cluster, 10752 per pod.

The network bandwidth per VM can be derived as follows:

- The UCS-6140 supports eight uplinks each, so each UCS cluster can support 80G/1344 = 59M per VM. Over-subscription prunes per VM bandwidth at each layer - aggregation, core, edge. The core layer provides 1:1 load balancing (L2 and L3), hence 80G/1344 = 59M per VM within each cluster. Extrapolating this to a full pod of 512 servers, this equates to (80G/10752) 7.4M per VM.

Storage bandwidth calculations can be derived as follows:

- There are 4x4G links from each UCS-6140 to MDS (aligns to the VCE Vblock). Assuming equal round-robin load balancing from each ESX blade to each fabric, there is 32G of SAN bandwidth. Inside each UCS cluster there is (160G/2) 80G FCoE mapped to 32G on the MDS fabrics. On the VMAX, eight FA ports are used for a total (both fabrics) of 32G bandwidth. EMC's numbers for IOPS are around 11,000 per FA port. Using eight ports, we get a total of 88,000 IOPS. Considering a cluster, 88,000/1344 equates to 65 IOPS per VM. Extrapolating to a full pod, 88000/10752 = 8 IOPS per VM.

Of course, you can add more FC and Ethernet ports to increase this per VM Eth and FC bandwidth.

VRF and MAC Scale

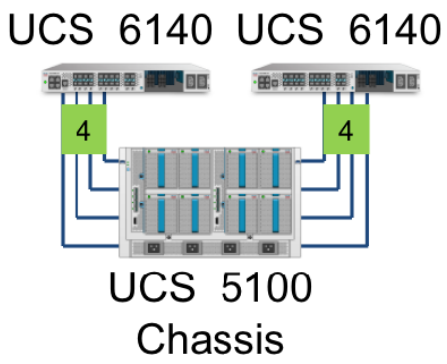
In this design, the collapsed aggregation/core layer must support a large number of MAC addresses, Layer 3 features (such as HSRP, OSPF, and BGP), and 10-Gb ports to aggregate the bandwidth required for client-to-server communications and server-to-server communications between pods. The Nexus 7000 platform is validated for the role of aggregation/core device with a maximum of 128,000 MAC addresses.

Network Over-Subscription

Increasing the efficiency of resource utilization is the key driver to oversubscribe hardware resources. This drives CAPEX savings up while still maintaining SLAs. In order to determine an efficient model of over-subscription, a study of the traffic pattern is essential to determine the amount of over-subscription possible. Typically, an 8:1 over-subscription ratio is considered while performing network capacity planning. The modeling depends on the amount of server-server and server-client traffic generated. Accordingly, bandwidth requirements can be estimated as the business grows.

Figure 3-4 represents a UCS chassis with 4x uplinks between each fabric extender and the fabric interconnect. It depicts 8x10-Gb uplinks available from each UCS chassis into the UCS fabric. Each UCS chassis contains up to eight 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 3-4 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 6140s and do not consume uplink bandwidth to the aggregation layer. Some traffic flow types and considerations are as follows:

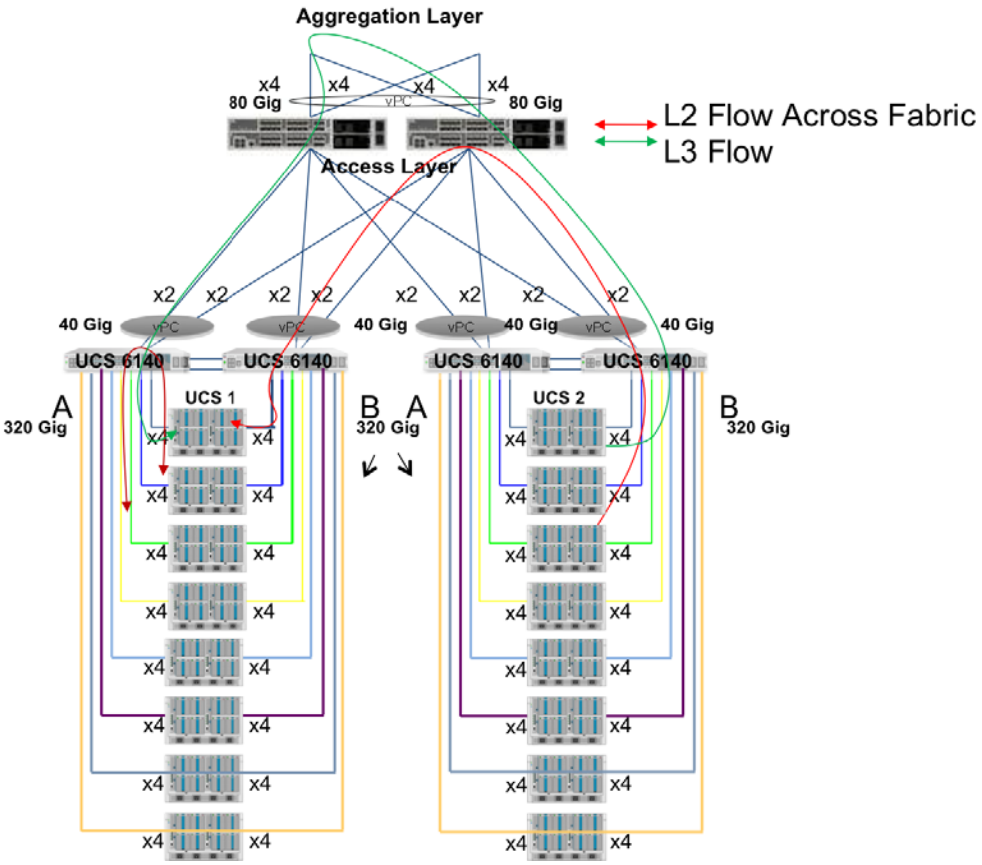
Server-to-Server Layer 2 Communications in the Same UCS Fabric—Because the source and destinations reside within the UCS 6140 pair belonging to the same UCS fabric, traffic remains within the fabric. For such flows, 10 Gb of bandwidth is provisioned.

Server-to-Server Layer 2 Communications Between Different UCS Fabrics—As shown in Figure 3-5, the End-Host Ethernet mode should be used between the UCS 6140s (fabric interconnects) and aggregation layer switches. This configuration ensures that the existence of multiple servers is

transparent to the aggregation layer. When the UCS 6140s 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 6140s, traffic is sent to the access layer switches and eventually forwarded to the servers by the correct UCS 6140.

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 3-5 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 over-subscription rate in the network. In this VMDC design, an 8:1 network over-subscription for inter-server traffic is considered for general compute deployment.

Figure 3-5 shows where the UCS chassis are connected to each UCS 6140 with 40 Gb of bandwidth. When all 8 chassis are connected, 320 Gb of bandwidth is aggregated at each UCS 6140. The four 10-Gb uplinks from each UCS 6140 form a port-channel where both vPC trunks are forwarding to the access layer over 40 Gb of bandwidth. This configuration defines a ratio of 320 Gb /40 Gb, an over-subscription

ratio of 8:1 at the access layer when all links are active. Similarly, the over-subscription ratio of 8:1 is provisioned at the aggregation layer when the all links are active. The over-subscription at the aggregation layer depends on the amount of traffic expected to exit the pod.

There will be flows where external clients access the servers. This traffic must traverse the access layer switch to reach the UCS 6140. 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.

Data Center Scalability with Large Pod

The data center scalability based on the large pod is determined by the following key factors:

- **MAC Addresses Support on the Aggregation Layer**—The Nexus 7000 platform supports up to 128,000 MAC addresses. Also, when the mix of small, medium, and large workloads are used, 11,472 workloads can be enabled in each large pod, which translates to 11,472 VMs. 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 solution assumes four MAC addresses per VM and 45,888 MAC addresses per large pod. Sharing VLANs between pods is discouraged unless it is required for specific purposes, such as application mobility. Filtering VLANs on trunk ports stops MAC address flood.
- **10 Gig Port Densities**—Total number of 10-Gig ports supported by the access/aggregation layer platform dictates how many additional compact PoDs can be added while still providing network over-subscription ratios that are acceptable to the deployed applications. The Nexus 7018 supports up to six large pods, equating to 512 blades.
- **Control Plane Scalability**—Each tenant VRF deployed on the aggregation layer device must maintain a routing adjacency for its neighboring routers. These routing adjacencies must maintain and exchange routing control traffic, such as hello packets and routing updates, which consume CPU cycles. As a result, control plane scalability is a key factor in determining the number of VRFs (or tenants) that can be supported using large pod growth. This design verified a minimum of 150 tenants.

A data center based on a large pod design can provide a minimum of 256 tenants and a range of workloads from 8,192 and up, depending on workload type. It can be expanded further by adding additional large pods to the existing core layer. In the VMDC solution, given the scalability limitation of the platforms leveraged, network architects can safely support six additional large pods beyond the base pod.

High Availability

Availability is defined as the probability that a service or network is operational and functional as needed at any point in time. Cloud data centers offer IaaS to either internal Enterprise customers or external customers of Service Providers. The services are controlled by using SLAs, which can be stricter in Service Provider deployments than in an Enterprise. A highly available data center infrastructure is the foundation of SLA guarantee and successful cloud deployment.

An end-to-end, highly available network infrastructure design provides predictable operational continuity.As organizations must satisfy SLAs made for business application uptime, they cannot afford to lose any connectivity due to an equipment failure. Therefore, the data center design must ensure that

a single failure either in hardware or software in the infrastructure does not affect the cloud subscribers' service. The following sections describe the design considerations that Cisco VMDC leverages to enable highly available and resilient end-to-end infrastructure:

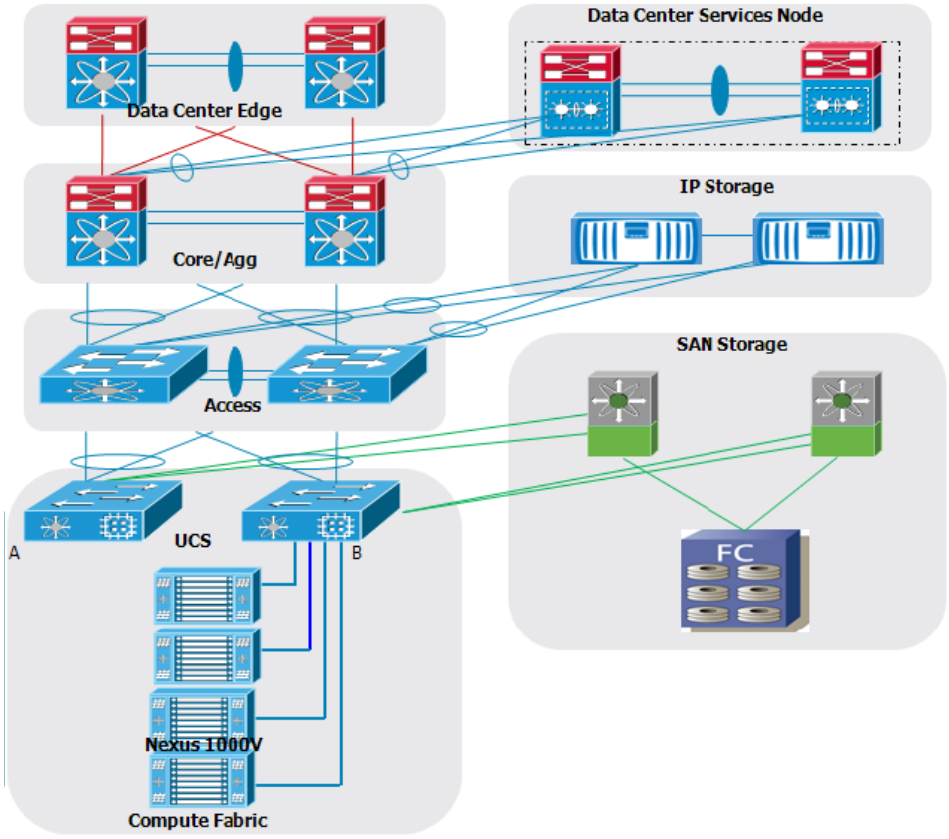
Physical Redundancy Design Consideration

To build an end-to-end resilient design, hardware redundancy is the first layer of protection that provides rapid recovery from failures. Physical redundancy must be enabled at various layers of the infrastructure as described in Table 3-2. For example, using node redundant features, such as VSS, if a device fails due to a supervisor reload, its redundant peer with a synchronized forwarding database forwards traffic without interruption and without any Layer 3/2 protocol convergence. In addition to recommending node-level redundancy, we also recommend implementing intra-node redundancy—redundant supervisors and line cards within each node of the pair (Figure 3-6).

Table 3-2 Physical Redundancy Layers

Physical Redundancy Method	Details
Node redundancy	Redundant pair of devices, for example, VSS and vPC
Hardware redundancy within the node	• Dual-supervisors • Distributed port-channel across line cards • Redundant line cards per VDC
Link redundancy	• Distributed port-channel across Line cards • vPC • VSS MEC

Figure 3-6 End-to-end High Availability



In addition to physical layer redundancy, the following logical redundancy considerations help provide a highly reliable and robust environment.

Virtual Port-channel (vPC)

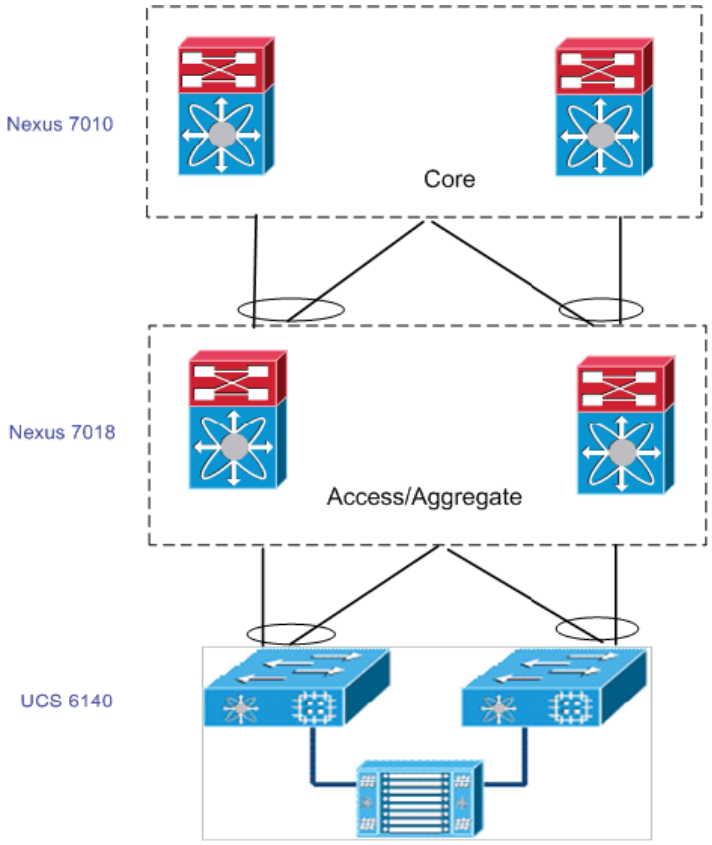
A virtual port-channel (vPC) allows links that are physically connected to two Cisco Nexus devices to appear as a single port-channel to any other device, including a switch or server. This feature is transparent to neighboring devices. A vPC can provide Layer 2 multipathing, which creates redundancy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

A vPC provides the following benefits when deployed either between access and aggregation layers or between the Cisco UCS and access layer devices in the Cisco VMDC design:

- Allows a single device to use a port-channel across two upstream devices
- Eliminates Spanning Tree Protocol (STP) 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 3-7 shows a vPC deployment scenario in which the Cisco UCS 6140s connect to Cisco Nexus 7000 access layer switches that 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 3-7 Large Pod End-to-end 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

Hot Standby Router Protocol

For VLANs having their Layer 3 termination on the SVI interfaces of Data Center Aggregation Switches, Hot Standby router Protocol (HSRP) will be configured as first hop redundancy protocol (FHRP). FHRP provides redundancy for IP clients that are configured with a default gateway. A group of routers function as one “virtual” router that provides redundancy in the event of a network failure. Both Data Center aggregation switches will become HSRP neighbors. A number of enhancements have been made to the vPC solution in order to integrate with the Layer 3 features and specifically HSRP. In HSRP, the improvement has been made to the forwarding engine to allow local Layer 3 forwarding at both the active HSRP peer and at the standby HSRP peer and thus provides an active/active HSRP configuration.

HSRP control protocol, however, still acts like an active/standby pair, such that only the active device responds to ARP requests, but a packet destined to the shared HSRP MAC address is accepted as local on either the active or standby HSRP device. In addition, HSRP control protocol can be configured active/standby on per VLAN basis, for example, HSRP can be active on even VLANs on the aggregation switch 1 and HSRP can be active on odd VLANs on the aggregation switch 2 such that both devices load balance ARP responses and reducing the control plane burden from a single active device having to respond to all ARP requests.

Note Large numbers of VLANs and VRFs are required to support multiple tenants in a multi-tenant architecture. Careful consideration is required when an aggressive implementation of HSRP timers are required because this may affect control plane stability.

Spanning Tree Design Considerations

VMDC large pod architecture uses Multiple Spanning Tree (MST) protocol, which maps multiple VLANs into a spanning tree instance, with each instance having a spanning tree topology independent of other spanning tree instances. This architecture provides multiple forwarding paths for data traffic, enables load balancing, and reduces the number of STP instances required to support a large number of VLANs. MST also improves the fault tolerance of the network because a failure in one instance (forwarding path) does not affect other instances (forwarding paths).

MST provides rapid convergence through explicit handshaking as each MST instance uses the IEEE 802.1w standard, which eliminates the 802.1D forwarding delay and quickly transitions root bridge ports and designated ports to the forwarding state. As Rapid Per-VLAN Spanning Tree (Rapid PVST) is the default spanning tree mode in Cisco NX-OS, the global configuration command in Cisco NX-OS to enable MST is as follows:

```
spanning-tree mode mst
```

In large pod architecture, the following MST features are used in access layer:

Spanning Tree Edge Ports—Spanning-tree edge trunk ports are configured on the Nexus 7000 ports connected to UCS 6140s. Conceptually, edge ports are related to the spanning-tree port-fast feature, which implies that a port directly connects to an end station, cannot create any bridging loop. To provide faster convergence, edge port quickly transitions into the forwarding state, skipping the listening and learning states. Neither edge trunk ports nor port-fast enabled ports generate any topology change notification (TCN) when these ports experience any changes.

Bridge Assurance—Bridge assurance is enabled only on vPC peer link. Bridge Assurance can be used to protect against certain problems that can cause bridging loops in the network. Specifically, you use Bridge Assurance to protect against a unidirectional link failure or other software failure and a device that continues to forward data traffic when it is no longer running the spanning tree algorithm. With Bridge Assurance enabled, BPDUs are sent out on all operational network ports, including alternate and backup ports, for each hello time period. If the port does not receive a BPDU for a specified period, the port moves into the blocking state and is not used in the root port calculation. Once that port receives a BPDU, it resumes the normal spanning tree transitions. This bidirectional hello mechanism helps prevent looping conditions caused by unidirectional links or a malfunctioning switch. Bridge Assurance can be enabled only on spanning tree network ports that are point-to-point links. Finally, both ends of the link must have Bridge Assurance enabled. In Cisco NX-OS, the following global configuration command sets the port type to network:

```
spanning-tree port type network
!
```

Routing Protocols

All Layer 3 devices in a network must learn IP prefixes in order to establish Layer 3 communications. Therefore, a Layer 3 IP routing protocol is required in the aggregation, core, and edge layers of the VMDC model. In the large pod VMDC architecture, BGP, an exterior routing protocol, is used to establish IP connectivity. BGP is an advanced path-vector routing protocol.

Routing Protocol Timers

Though enabling sub-sec protocol timers are supported via Cisco devices, this is not advisable in a VMDC environment where tenants are separated with VRFs and VLANs. Additionally, multi-tenant virtual cloud architecture requires large numbers of VLANs and VRF instances compared to a traditional deployment model. As a result, it is recommended to deploy default BGP Hello and Hold timers to provide faster convergence, yet to minimize control plane load.

NSF for BGP

Cisco Nonstop Forwarding (NSF) with Stateful Switchover (SSO) is a Cisco innovation for routers and switches with dual route processors (RP) or supervisors (SUP). Cisco NSF with SSO allows a router or switch, which has experienced a hardware or software failure of an active RP or SUP, to maintain data link layer connections and continues forwarding packets during the switchover to the standby RP or SUP. This forwarding can continue in spite of the loss of routing protocol adjacencies with other devices. Routing information is recovered dynamically in the background, while packet forwarding proceeds uninterrupted.

Cisco has implemented a new capability, i.e. graceful restart capability in BGP to enable NSF in NX-OS, IOS, and IOS-XR releases. In the large pod VMDC, NSF or BGP graceful restart feature is recommended in core, aggregate, and edge layers.

Bidirectional Forwarding Detection (BFD) for BGP

BFD can detect peer link failure very quickly regardless of media, encapsulations, topologies, and underlying routing protocols such as BGP, EIGRP, IS-IS, and OSPF. Once BFD has been enabled on the interfaces and at the appropriate routing protocols, a BFD session is established, BFD timers are negotiated, and the BFD peers begin to send BFD control packets to each other at the negotiated interval. In the event of a failure, BFD immediately notifies its consumers such as local routing process such as BGP process, to take necessary actions. For example, If BFD is enabled for BGP neighbors between a peer of aggregation switches and BFD detects any failure, BFD will immediately notify local BGP process that BFD neighbor is no longer reachable. As a result, BGP will tear down its neighbor relationship and look for alternative paths immediately without waiting for the hold timer to expire and thus provides faster failure recovery. VMDC large pod architecture will enable BFD in a future release.

Route Summarization

With IP route summarization, route flaps and instabilities are isolated, and convergences are improved. Route summarization also reduces the amount of routing traffic, the size of the routing table, and the required memory and CPU for the routing process. Large pod VMDC architecture uses BGP as routing protocol for multi-tenancy. As each multi-tenant customer maintains it own routing table, which is isolated from one another, the number of routing entries in the routing table are very small. Moreover, a 16-bit network mask is used for each tenant. As a result, route summation does not provide any added benefits and thus route summarization is not used in large pod VMDC architecture.

Aggregation Layer HA Design Considerations

In the large pod VMDC design, the access and aggregation layer functionality are combined into one layer as opposed to a three tier data center architecture which provides segmentation and scale. The choice of the collapsed access/aggregation model versus a 3-tier architecture depends on the port density and capacity of the node at the aggregation layer. The targeted scale for the large pod is 64 servers to start with and scaling this up by a factor of 8 (512 servers) is achievable through the combined core and aggregation layer without losing required functionality. The following recommendations should be considered at the aggregation layer:

- Spanning Tree Protocol (STP) Recommendations, page 3-15
- Routing Protocol Recommendations, page 3-16

Spanning Tree Protocol (STP) Recommendations

Port-channeling techniques such as virtual port-channel (vPC) and Multi-Chassis EtherChannel (MEC) can be used to enable loop-free topology in Layer 2 without using spanning-tree protocol, STP. Typically, STP provides loop free topology by eliminating redundant paths in a Layer 2 network. However, this may result in longer convergence due to the timer-based nature of the protocols even if STP is deployed with advanced features such as Rapid PVST+ and MST. Even though STP is a well-versed protocol that is heavily used in Layer 2 networks, no tools exist to easily troubleshoot a Layer 2 loop or broadcast storm. Further, it does not provide Layer 2 multi-path for a given VLAN. However, with user configuration advanced spanning tree protocols such as Rapid PVST+ or MST supports Layer 2 multi-path. Nevertheless, we recommend enabling STP to prevent any configuration errors and using MEC or vPC port-channeling techniques on the links connect to the Core and Data Center Services layer devices. MEC and vPC form port-channels across multiple aggregation devices by treating a pair of aggregation layer devices as a single device from the control plane perspective. Port channeling technique provides Layer 2 multi-path and a loop free Layer 2 topology and it does not depend on STP.

We recommend the following STP settings at Aggregation/Access layer of VMDC architecture:

- Explicitly specify Root Bridge on the aggregation layer. To provide the optimal topology, configure the root bridge on the default gateway, which is in this design a part of aggregation layer switches. The root bridge is the bridge with the smallest bridge ID in the network. The most significant 16 bits of the bridge ID can be set using the priority command. The last 48 bits of the bridge ID are MAC address from the switch chassis. Deterministically, the root bridge can be specified by configuring the best bridge priority on the aggregation layer switches. This configuration ensures that no downstream switch becomes active, which would result in suboptimal forwarding and STP failures. To configure the root bridge, use the following command:

```
Spanning-tree vlan 1 priority 24576
```

- Configure the vPC peer switch feature. For software releases prior to NX OS 5.0 (2), you must enable STP primary and secondary root on the pair of Nexus 7000 switches configured as aggregation layer switches. Thus, if a primary root fails, the secondary root becomes active though STP convergence. However, for software releases Cisco NX-OS 5.0(2) and later, an enhancement exists to make the pair of Nexus 7000 switches appear as a single logical STP root so if a primary root fails, there is no STP root convergence. This feature improves convergence for primary STP root failures and eliminates the need to define the STP primary and secondary roots. Network administrators must still specify identical primary root priority on both Nexus 7000 switches, but there is no need to enable the secondary root as both vPC switches act a single root device in the vPC peer switch configuration. The following is an example of a vPC peer switch configuration in global configuration mode:

```
vpc domain 100
    peer-switch
```

Cisco NX OS 5.0(2) or later is required to enable the vPC peer switch feature.

Routing Protocol Recommendations

In a VMDC environment, tenants are separated with VRFs and VLANs, and multi-tenant virtual cloud architecture requires large numbers of VLANs and VRF instances compared to a traditional data center deployment model.

Scalability is one of the key aspects of the VMDC solution. With the current NX-OS release i.e. 5.0.3, BGP is the only routing protocol that can provide the desired scalability demanded by Service Providers. As a result, BGP is the recommended routing protocol for large pod VMDC architecture. For more information about BGP recommendation, see the "Routing Protocols" section on page 3-14.

Core Layer High Availability Design Considerations

The core layer of the Large PoD VMDC is purely a Layer 3 environment. As a result, the core layer is free of spanning tree and vPC. However, core layer ensures HA by Layer 3 port-channeling and device level redundancy. BGP is used as a routing protocol in this layer. For more information about BGP, see "Routing Protocols" section on page 3-14.

Compute Layer High Availability Design Considerations

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

UCS End-Host Mode

Fabrics interconnect operating as a host called End-Host, EH mode. Virtual machine NICs are pinned to UCS fabric uplinks dynamically or statically. These uplinks connect to the access layer switch providing redundancy towards the network. The fabric interconnect uplinks appear as server ports to the rest of the fabric. When this feature is enabled, STP is disabled; switching between uplinks is not permitted. This mode is the default and recommended configuration if the upstream device is Layer 2 switching. Key benefits with End-Host mode are as follows:

- All uplinks are used
- Uplinks can be connected to multiple upstream switches
- No Spanning-tree is required
- Higher Scalability because the Control Plane is not occupied

- No MAC learning on the uplinks
- MAC move is fully supported within the same switch and across different switches

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 Pair of VSMs in Active-Standby Mode

Always deploy the Cisco Nexus 1000V Series virtual supervisor module (VSM) 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 ESXi host. This requirement helps ensure high availability even if one VMware ESXi server fails. You should also use the anti-affinity feature of VMware ESXi to help keep the VSMs on different servers.

Fault Tolerance

VMware Fault Tolerance (FT) provides continuous availability by utilizing VMware vLock-step technology, which is a record and playback technique. In a FT deployment, primary and secondary virtual machines (VMs) are engaged in a virtual lockstep process. Virtual lockstep maintains a process that replicates instruction-for-instruction and memory-for-memory states of a primary VM into a secondary VM. In this way, secondary VM maintains identical state as the primary VM and thus provides fault tolerance. In fact, FT provides 1:1 redundancy for VMs.

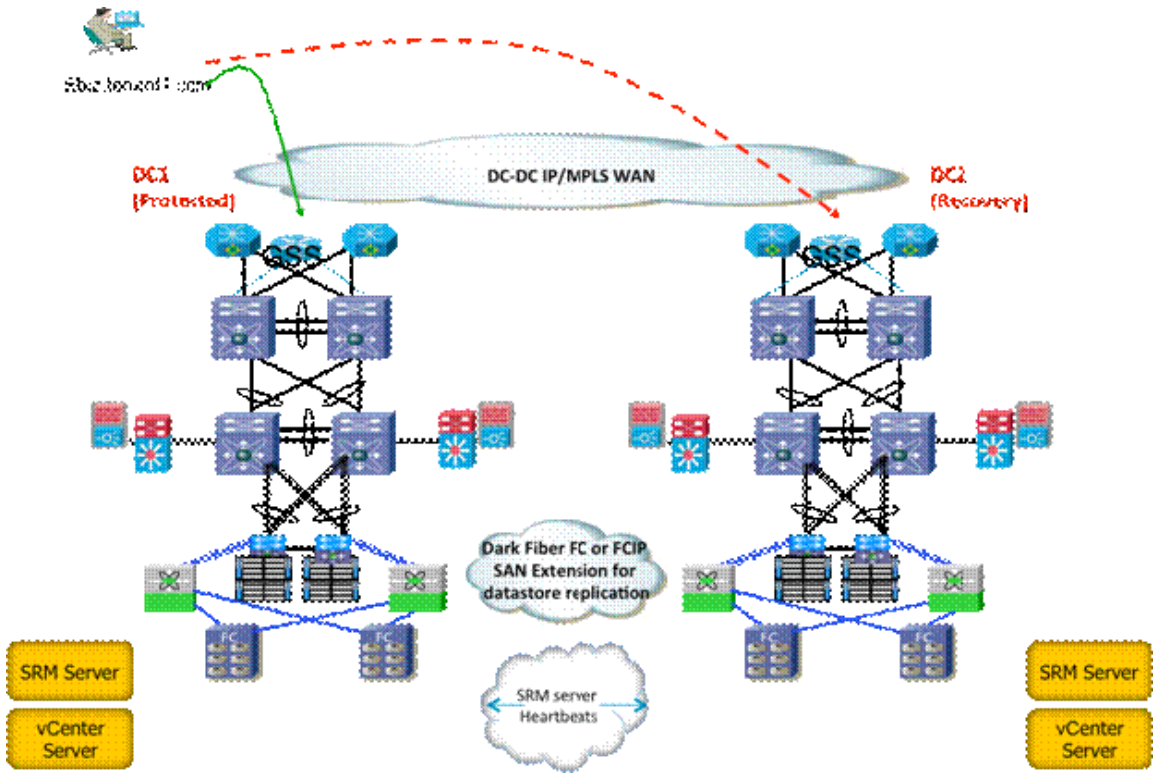
Utilize Cluster High Availability

The VMDC architecture prescribes the use of VMware HA for intra-cluster resiliency. In contrast to VMware Fault Tolerance, which provides a 1:1 failover between a primary and secondary VM within a cluster, VMware HA provides 1:N failover for VMs within a single cluster. In this model, an agent runs on each server and maintains a heartbeat exchange with designated primary servers within the cluster to indicate health. These primary hosts maintain state and initiate failovers. Upon server failure, the heartbeat is lost, and all the VMs for that server are automatically restarted on other available servers in the cluster pool. A prerequisite for VMware HA is that all servers in the HA pool must share storage: virtual files must be available to all hosts in the pool. All adapters in the pool must be in the same zone in the case of FC SANs.

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 data store 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 data store replication. For example, SRM could be used in the VMDC model to optionally provide disaster recovery 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 (Figure 3-8). This use case was tested as a proof-of-concept and details are provided in a separate application note.

Figure 3-8 Virtual Machine Disaster Recovery Between Active/Standby Data Centers



Storage Layer High Availability Design Considerations

In the storage layer, the HA design is consistent with the HA model implemented at other layers in the infrastructure, comprising physical redundancy and path redundancy. Table 3-3 lists the storage layer redundancy methods.

Table 3-3 Storage Layer Redundancy Methods

Redundancy Method	Details
Link redundancy	• Redundant links distributed across Line cards using Port-Channels • Multi-pathing
Hardware redundancy	• Redundant adapter ports (i.e., CNAs, HBAs) per server • Dual supervisors (MDS) • Dual storage controllers (NetApp NAS and EMC SAN) • RAID 1 and RAID 5 redundant arrays
Node redundancy	• Redundant storage devices (MDS switches, SAN fabrics)

Link Redundancy

Pending the upcoming availability of FC port-channels on UCS FC ports and FC port trunking, multiple individual FC links from the 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 will re-login in a round-robin manner using available ports. 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. Multi-pathing software from VMware or the SAN storage vendor (i.e., EMC Powerpath software) further enhances HA, optimizing use of the available link bandwidth and enhancing load balancing across multiple active host adapter ports and links with minimal disruption in service.

Hardware and Node Redundancy

The VMDC architecture leverages best practice methodologies for SAN HA, prescribing full hardware redundancy at each device in the I/O path from host to SAN. In terms of hardware redundancy this begins at the server, with dual port adapters per host. Redundant paths from the hosts feed into dual, redundant MDS SAN switches (i.e., with dual supervisors) and then into redundant SAN arrays with tiered, RAID protection. RAID 1 and 5 were deployed in this particular instance as two more commonly used levels; however the selection of a RAID protection level will depend on a balancing of cost versus the critical nature of the data that is stored.

Service Assurance

Service assurance is generally defined as a set of procedures intended to 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 insure that a product or service meet specified requirements, adhering to a defined set of criteria which 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.

SLA criteria in the IaaS context focuses on service availability. It is important to note that several types of XaaS cloud services have been defined by NIST. In addition to “infrastructure as a service”, these are “platform as a service” and “software as a service.” As one goes up the stack the scope of service level criteria is broader, including more focus on middleware, operating system and application performance. This document focuses on IaaS service level criteria.

- Service Availability, page 3-20
- Quality of Service, page 3-21

Service Availability

Service availability is generally calculated using the following formula:

$$\%\text{Availability} = [\{T_{\text{Period}} - T_{\text{Without service}}\} / T_{\text{Period}}] * 100$$

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

What are the components of service availability in the IaaS context? Table 3-4 provides a synopsis of frequently applied availability SLA constituents. A few of these, such as managed security services for example, may be more applicable to the public cloud services context.

Table 3-4 Frequently Applied Availability SLA Constituents

Availability Component	Performance Indicators
Portal Availability	Portal service availability; information accuracy, successfully processed service requests
Virtual Machine Availability	Percentage of service availability (% Availability)
Virtual Machine RTO	Recovery Time Objective for restoring of a virtual machine in the event of a server crash.
Storage Availability	% Availability
Network Availability	% Availability
Firewall Availability	% Availability (i.e., of a vApp vFW or virtual context in the FWSM)
Load Balancer Availability	% Availability (i.e., of a vApp vFW or virtual context in the FWSM)
Backup Data Reliability	% (Scheduled) Successful data backup attempts: this can refer to actual data store 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 (i.e., file access auditing and data or data store 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 parameters described above, service performance parameters may also include incident response time and incident resolution objectives. The latter may vary depending on the type of service component (VM, network, storage, firewall, etc.).

The VMDC architecture addresses the availability requirements of IaaS SLAs for all of the criteria listed in the preceding table, through 1:1, 1:N or N:N VM, network and storage redundancy, and data security and data protection mechanisms. These have been outlined in the HA, service tier and multi-tenant isolation sections of the document.

Quality of Service

Though in general, IaaS SLAs emphasize service availability, the ability to offer differentiated service levels is effectively a reflection of the fact that specific applications or traffic may require preferential treatment within the cloud. Some applications could be mission critical, some could be interactive, and the rest could be bulk or utilized simply for dev-test purposes. In an IaaS context service levels are end to end, from cloud resources (hosts, data stores) to the end user. These service levels are embodied in the tenant subscription type (i.e., Gold, Silver, and Bronze) described earlier in this document.

Quality of service functions are a key aspect of network availability service assurance, in that they enable differential treatment of specific traffic flows, helping to insure that in the event of congestion or failure conditions, critical traffic is provided with a sufficient amount of the available bandwidth to meet throughput requirements. Traditionally, an SLA framework generally includes consideration of bandwidth, delay, jitter, and packet loss per service class.

The QoS features leveraged in this design are as follows:

- QoS classification and marking
- Traffic flow matching
- Bandwidth reservation

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 is nothing but rewriting the COS in the Ethernet header or the type of service bits in the IPv4 header.

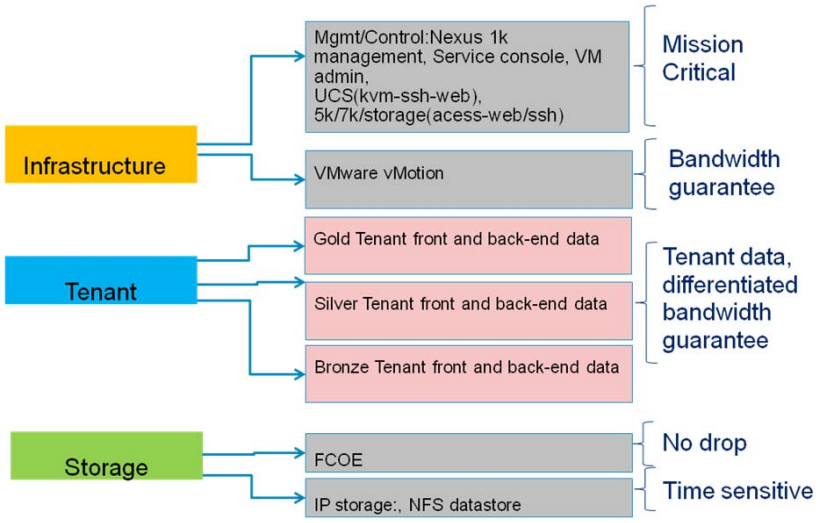
As per established best practices, classification and marking are applied at the network edge, close to the traffic source: in this design, at the Nexus 1000 virtual access switch for traffic originating from hosts and VMs and at the CRS-1 WAN edge for traffic entering the DC infrastructure from the public IP-NGN, Internet, or private WAN backbone. At the Nexus 1000 switch then, marking incoming control traffic for a CoS value of 6 would appear as follows:

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

Figure 3-9 shows the traffic flow types defined in the VMDC architecture. These break down into 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 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.

Figure 3-9 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 simply accepted and matched rather than remarked. Classification and marking are applied at the network edge, close to the traffic source, in this design, at the Nexus 1000V virtual access switch for traffic originating from hosts and VMs and at the CRS-1 WAN edge for traffic entering the DC infrastructure from the public IP-NGN, Internet, or private WAN backbone. The trust boundary in this design is at the Nexus 7000 Access/Aggregation device connecting to the UCS (and Nexus 1000V), and on the Nexus 7000 DC Core router connecting to the CRS-1 WAN edge router. 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 Reservation 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. In order to provide differentiated treatment per defined traffic class in the event of buffer congestion, it is possible to use the service-policy command to specify a minimum bandwidth guarantee and apply it to an interface, subinterface or virtual circuit.

One challenge of using a bandwidth guarantee is that it 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 may need to be rate limited (policed) to ensure that they do not starve the other classes of bandwidth. When application throughput requirements are well understood, rate limiting may be used more as a safety precaution, protecting the other classes in the case of unexpected failures that lead to adversely high load for a particular class.

The VMDC SLA framework for QoS is represented in Table 3-5. Specific implementation details vary due to differences in connectivity, interface types, QoS scheduling, and queuing capabilities across specific platforms in the infrastructure. These are discussed and documented in the Large PoD Implementation Guide.

Table 3-5 Cisco UCS Configuration by Tier

Traffic Type	Traffic Classes	CoS Marking (N1k/ M81KR)	BW% UCS	BW% N7K	BW% C6k
Control - i.e, Nexus 1000V control/management, NFS Data store system control	Control	6	5%	10%	10%
vMotion	Control	6	incl. above	incl. above	N/A
Tenant (front/backend)	Gold	5	20%	20%	20%
IP Storage	NAS	5	incl. above	incl. above	N/A
Tenant (front/backend)	Silver	2	15%	30%	30%
Tenant (front/backend)	Bronze	0	10%	20%	20%
FCOE	FCoE	3	50%	N/A	N/A

Secure Tenant Separation

The cloud deployment model calls for common resource pools to be efficiently shared among multiple tenants. In a private cloud, a tenant is an Enterprise department. In a public cloud, a tenant is an individual customer or Enterprise who subscribes to cloud resources. In both public and private cloud scenarios, multi-tenancy provides better resource utilization but requires designing of secure tenant separation to ensure end-to-end security and path isolation within the shared infrastructure.

Traditionally, customers deployed dedicated infrastructure for each tenant. This approach poses serious limitations on management of costs and complexity, and inefficiency in 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, secure separation is a fundamental requirement in multi-tenant environments. To achieve secure tenant separation and path isolation, we configure end-to-end virtualization of network, compute, and storage resources.

The following design considerations provide secure tenant separation and path isolation:

- Network Separation, page 3-23
- Compute Separation, page 3-32
- Storage Separation, page 3-32
- Application Tier Separation, page 3-34
- VM Security, page 3-38
- Fiber Channel Zones, page 3-38

Network Separation

End-to-end virtualization of the network requires separation at each network layer in the architecture:

- Network Layer 3 separation (core/aggregation)
- Network Layer 2 separation (access)
- Network services separation (firewall and load balancing services)

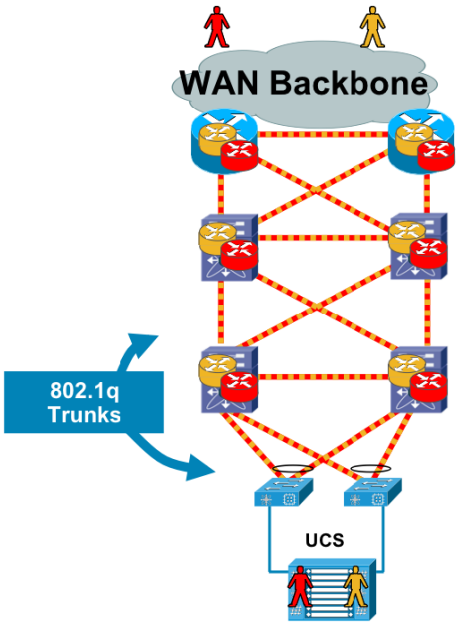
- Tenant path isolation
- Client-server traffic separation

Additionally, virtual private remote access in the form of SSL and IPSec or MPLS VPNs serves to further secure tenant traffic over the shared public or private cloud.

Network Layer 3 Separation (Core/Aggregation)

As shown in Figure 3-10, an end-to-end VRF-Lite deployment can provide Layer 3 separation among tenants.

Figure 3-10 VRF-Lite End-to-end



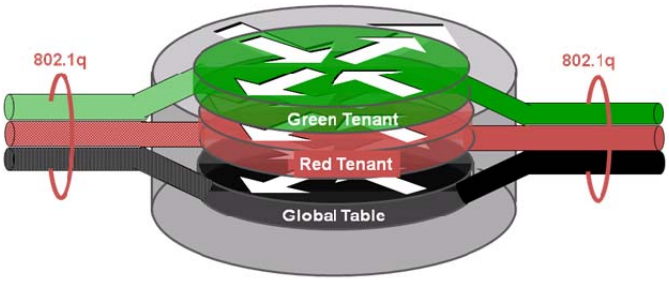
A VRF-Lite instance represents a tenant container. Depending on the service tier, the container could include services such as load balancer, firewall, IPSec/SSL off loading; VLANs representing the application tiers (Web, application, database).

A VRF instance consists of the following:

- An IP routing table
- A derived forwarding table
- A set of interfaces that use the forwarding table
- A set of rules and routing protocols that determine what goes into the forwarding table

VRF-Lite uses a Layer 2 separation method to provide path isolation for each tenant across a shared network link. Typically, a dot1q tag provides the path separation.

Figure 3-11 Network Device Virtualization with VRF



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

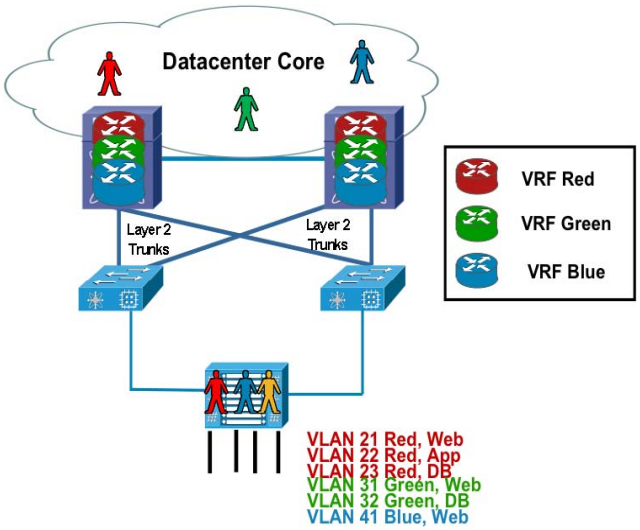
- **True Routing and Forwarding Separation**— Dedicated data and control planes are defined to handle traffic belonging to groups with various requirements or policies. These groups represent an additional level of segregation and security as no communication is allowed among devices belonging to different VRFs unless explicitly configured.
- **Uniform Deployment Across Cisco Nexus and Catalyst Products**—VRF-Lite is natively supported in hardware across specific Cisco Nexus and Catalyst switches to achieve higher throughput and performance.
- **Virtualized Physical Infrastructure**—Each virtual network represents a replica of the underlying physical infrastructure. Depending on the service tier (Gold, Silver, Bronze), the virtual network varies in the number of links included in the VRF context. The end result is a physical infrastructure shared among multiple tenants.

Network Layer 2 Separation (Access)

Network separation at Layer 2 is accomplished using VLANs. While VRFs are used to identify a tenant, VLAN-IDs are used to provide isolation at Layer 2.

In addition to the tenant VLAN, it is required to provide differentiated access to each tenant, to provide isolation between application and service tiers. In large deployments, you may need to enable the extended range of VLANs.

Figure 3-12 VLANs to VRF Mapping



In Figure 3-12, a tenant using vrf Red is assigned 3 VLANs (21,22,23) for its web application as per the Gold service tier while vrf Green is assigned two VLANs as per the Silver service tier, and vrf Blue is assigned 1 VLAN as per the Bronze service tier. This design assumes there is no need to connect the tenant VRFs. However, alternatives exist to enable this capability, as discussed in the Cisco VMDC Implementation Guide, version 2.0.

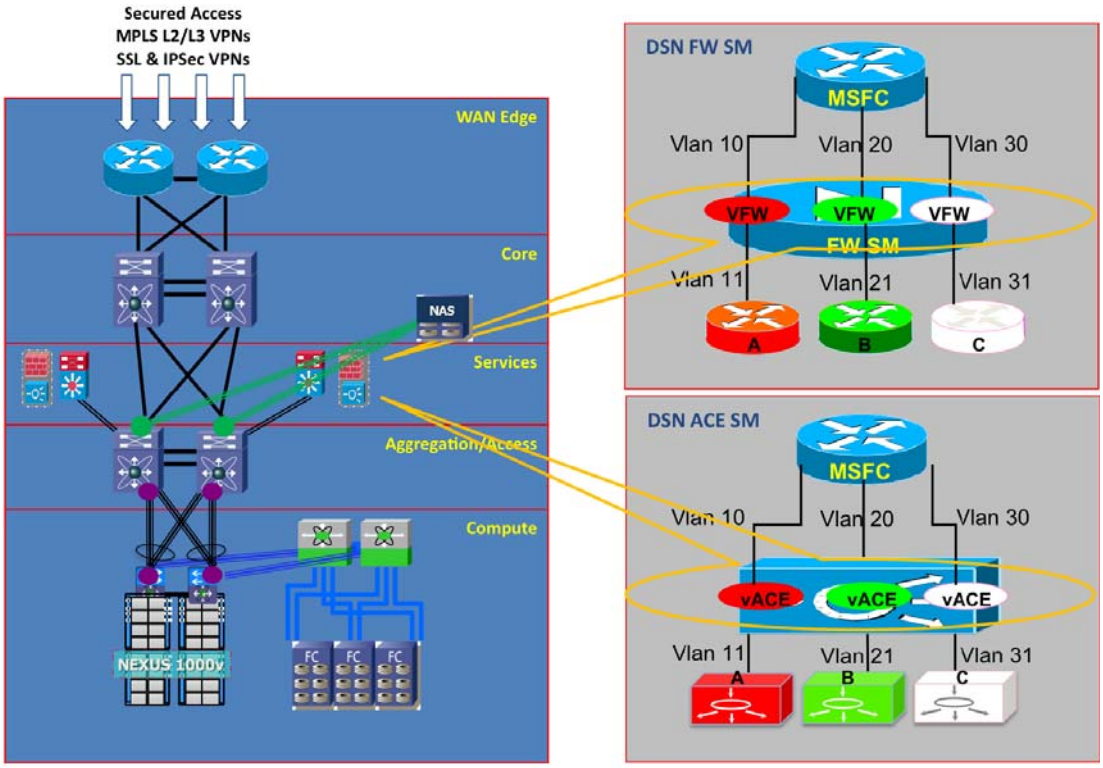
Network Services Virtualization (Firewall and Load Balancing Services)

Data center networks also require intelligent services, such as firewall and load balancing of servers and hosted applications. In this cloud solution, these services are provided with the Gold, Silver, and Bronze network containers. In the Cisco VMDC architecture, firewalls and load balancers are included in the Gold and Silver service tiers. The Cisco Firewall Services Module (FWSM) provides Layer 2 and Layer 3 firewall inspection, protocol inspection, and network address translation (NAT). The Cisco ACE service module provides server load balancing and protocol (IPSec, SSL) off-loading.

To achieve secure separation across the network, the services layer must also be virtualized. Security contexts and/or zones with independent policies are used to virtualize services. The contexts are stitched to a vrf via VLANs to provide services.

We recommend a dedicated services layer to isolate intelligent services into their own layer. The Cisco Catalyst 6500 chassis can integrate service modules in card slots to conserve rack space, power, and cables. The Cisco Data Center Services Node (DCSN) is a Cisco Catalyst 6500 Series Switch with FWSM and ACE service modules dedicated to security and server load balancing functions. Figure 3-13 shows the Cisco DCSN directly attached to aggregation layer switches.

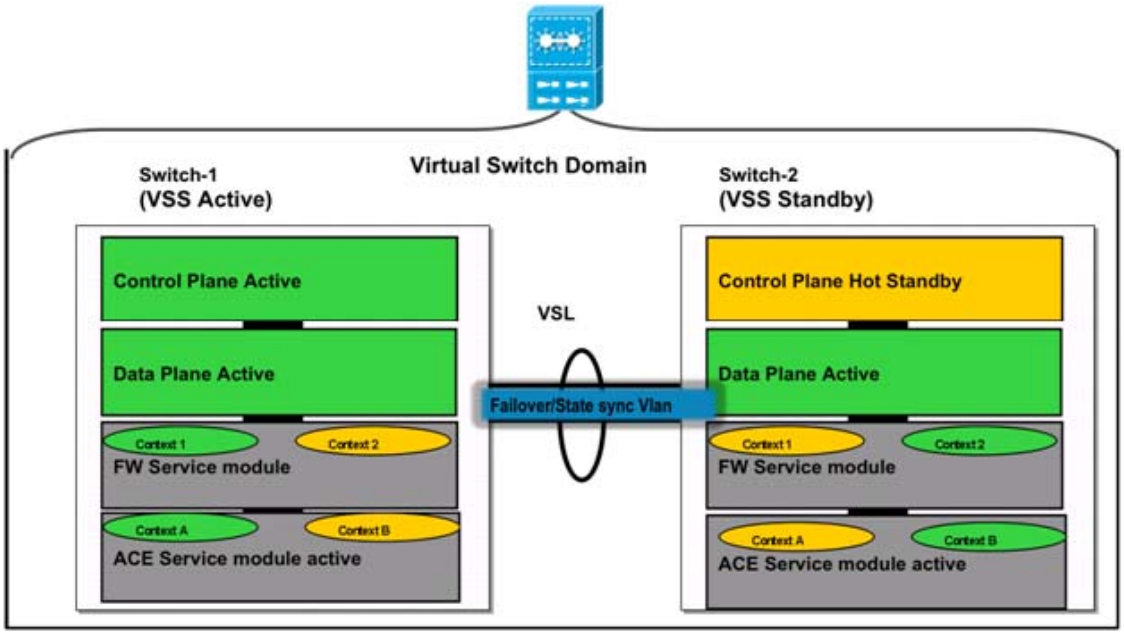
Figure 3-13 Virtual Firewall and Load Balancing Services



Using the virtualization features of the Cisco DCSN 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 3-14 show, 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 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 3-14 Active-Active Services Chassis with Virtual Contexts



Tenant Path Isolation

As the cloud data center requirements shifted from dedicated infrastructure for each tenant to a shared resource pool model, network architects began to consider alternative technologies that supported sharing the infrastructure and provided the same level of tenant isolation as a dedicated infrastructure. The Cisco VMDC design uses path isolation techniques to logically divide a shared infrastructure into multiple virtual networks.

Path isolation defines independent, logical traffic paths over a shared physical network infrastructure. These paths use VRFs, VLANs and contexts/zones as resources to provide tenant isolation.

The goal of segmenting the network is to support a multi-tenancy on the physical infrastructure; however, it must ensure resiliency, scalability, and improved security.

A hierarchical 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 enable end-to-end traffic segmentation. Finally, the server and storage virtualization is integrated with this virtualized path.

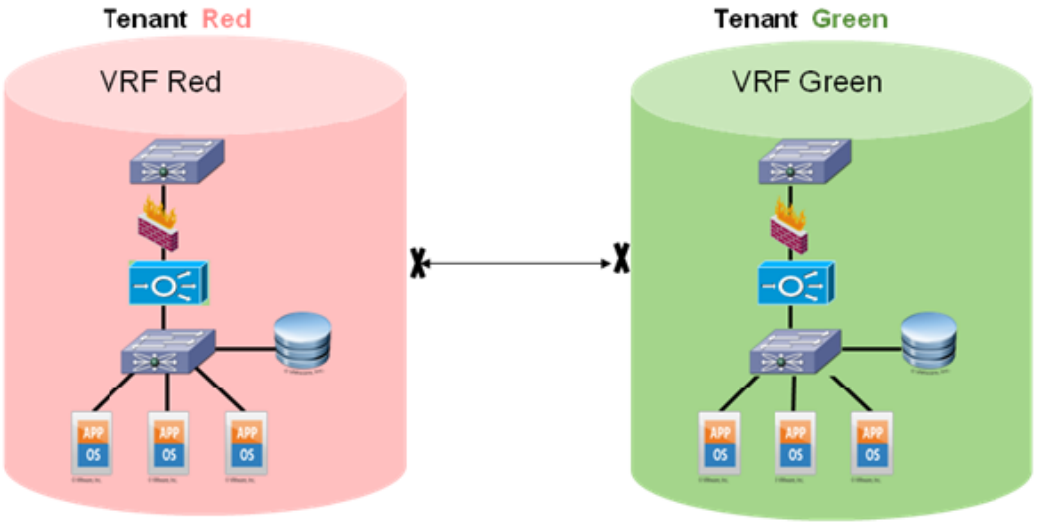
Following are the levels of virtualization required and implemented in this design:

- **Device Virtualization** The virtualization of the network device, which includes all forwarding/routing processes, databases, tables, and interfaces within the device.
- **Data Path Virtualization** 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.

Figure 3-15 shows how each tenant can be logically segmented using end-to-end path isolation. In the Cisco VMDC design, path isolation provides the following:

- Tenant Isolation and boundary
- Per-Tenant Security and Policy Control

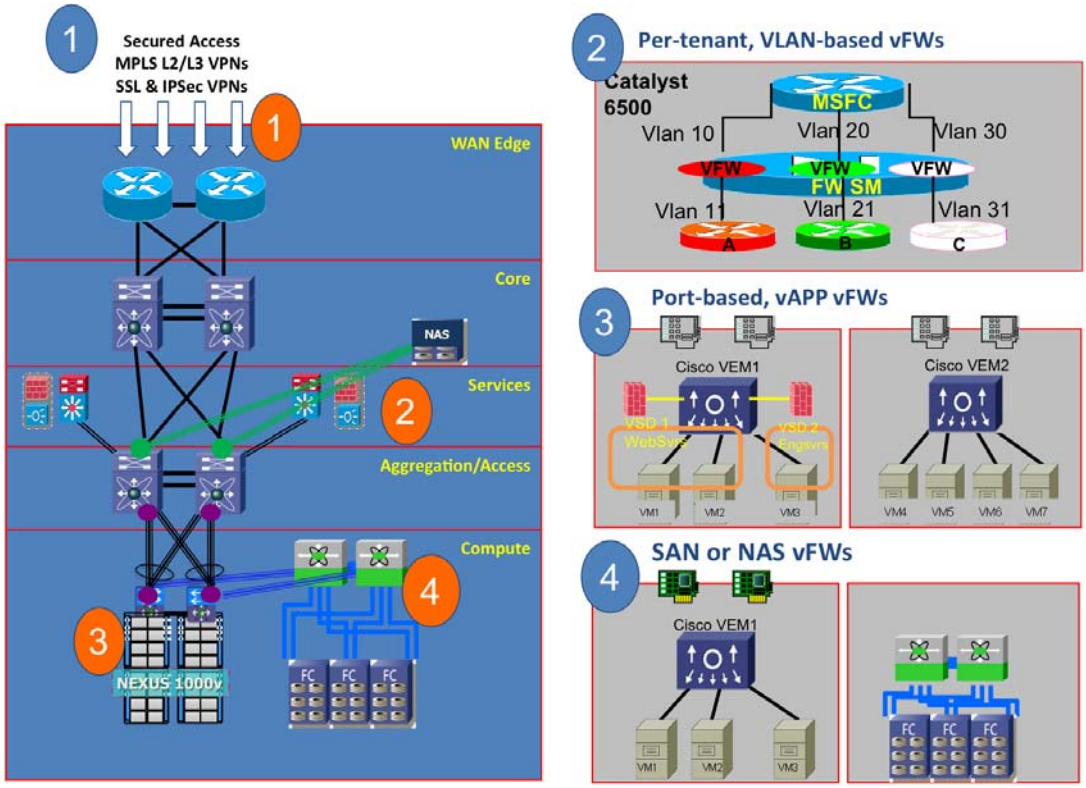
Figure 3-15 Tenant Path Isolation



Client-Server Traffic Separation

The path isolation provided by VRFs helps secure tenants against overlapping network architectures. However, cloud administrators must protect against security exploit attempts by external users who access the cloud resources. Consider the case where a web client accesses the web services inside the cloud. For a given web application, we must secure the server farm and the communication between the web server and back-end servers, such as middleware and database servers.

Figure 3-16 Firewall Inspections of Tenant Traffic



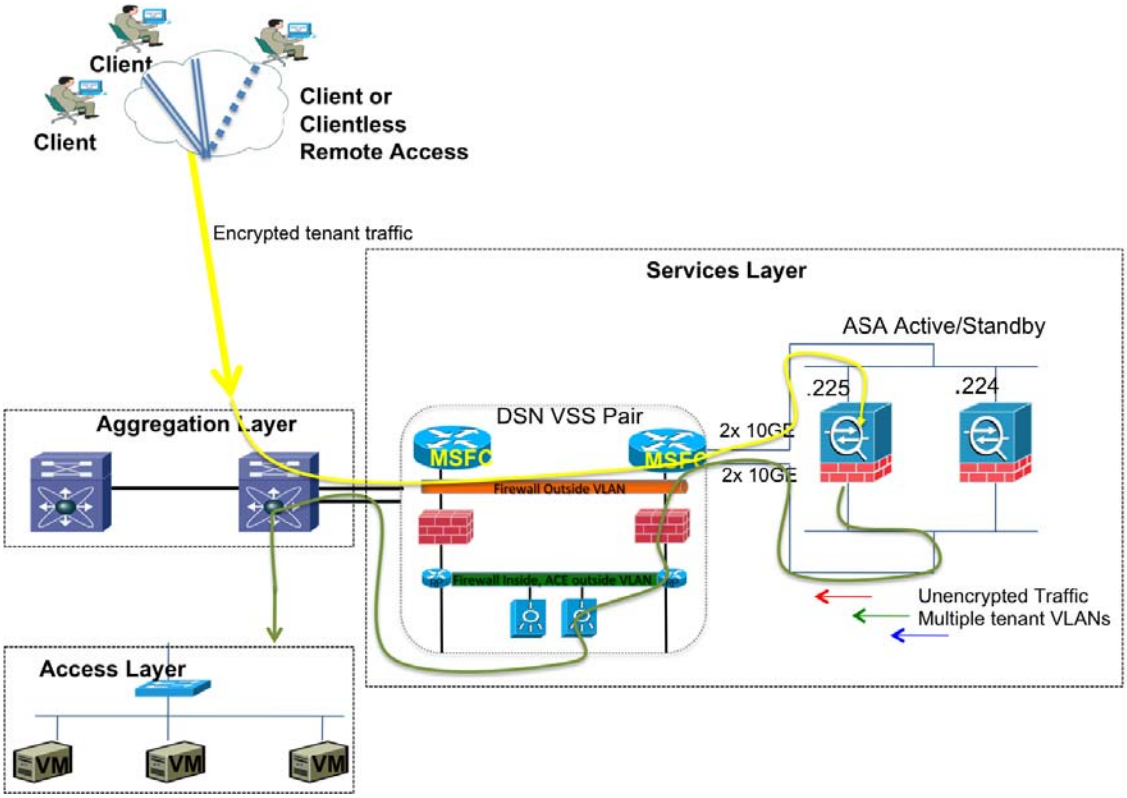
The following technologies are included in the VMDC design to address external users:

- Dedicated virtual firewall context on the firewall module that belongs to a particular tenant is used to provide traffic inspection
- DOS attack prevention
- L4-7 protocol inspection
- ACLs to control traffic permitted into the network

The VMDC firewalling model effectively employs a tiered model (tenant access, services, application and storage). Figure 3-16 shows the mapping of these tiers to secured access mechanisms.

The design incorporates 3 methods of secured remote access: MPLS, IPSec and SSL VPNs. MPLS VPN access was incorporated in phase 1 of the VMDC architecture; design guidance is available in the VMDC Release 1 documents and so will not be a focus of this discussion. SSL and IPSec VPN access are new components of the architecture model in this release, provided as part of the ASA 5500 functionality in the services layer of the network (Figure 3-17).

Figure 3-17 SSL and IPSec VPN Logical Flow



Providing easy-to-manage full-tunnel network access through both SSL VPN and IPSec VPN client technologies, advanced clientless SSL VPN capabilities, and network-aware site-to-site VPN connectivity, the ASA5500 enables Providers and Enterprises to create secure connections across public networks to mobile users, remote sites, contractors, customers and business partners.

The ASA supports all popular authentication mechanisms, including but not limited to Local user database, RADIUS, Windows NT LAN Manager (NTLM), Active Directory Kerberos, Native RSA SecurID, RADIUS with Expiry, one-time password (OTP) via RADIUS (State/Reply message attributes), Lightweight Directory Access Protocol (LDAP) with password expiry capabilities (including pre-expiry warning), digital certificates (including X.509), smartcards, SSO and SPNEGO. ASA supports CRL and OCSP for certification revocation checks. It also supports AAA and Certificate authentication simultaneously. Additionally, the ASA can look at fields in the certificate (including Extended Key Usage Extensions) to make additional policy decisions. Optionally, the ASA can be configured to request certificates only for a limited group of users. Finally, the ASA can also act as a certificate authority.

The ASA is designed to bind granular policies to specific users or groups across multiple identity management systems via Dynamic Access Policies (DAP). DAPs are created by setting a collection of access control attributes associated with a specific user tunnel or session. These attributes address issues of multiple group membership and endpoint security. In the VMDC solution, these user group definitions and associated policies are key, serving to create and support “multi-tenant” remote secured access. Specifically, in the VPN termination case, tenants are differentiated based on group name and domain name, rather than by unique virtual contexts as in the firewall use case.

Figure 3-17 shows the logical connectivity. In this model, tenants are dropped first into a common remote access VLAN, authenticated, and from there traverse unique per-tenant VLANs, with DAP filters and filters on their unique ACE or Firewall contexts insuring that another tenant may not access their

servers or VMs. Typically in the case of SSL VPN access, tenants would utilize a link in their portal, that takes them to the pop up window of the ASA, where they will login and be authenticated by a directory server such as an LDAP or Radius server (not shown). To scale beyond AAA Auth-server group limits, a Radius server may be used in conjunction with an LDAP server. In this case the Radius server will proxy to the LDAP server which then authenticates tenant access. The RADIUS will return option 25, which will carry a unique tenant identifier, for example, "VPDC_ID", and an ASA DAP rule will then force the tenant to their specific home page. With these policy restrictions in place, the tenant would only be able to access their own particular home page.

Similarly, in the case of IPSec access, each client belongs to a group. The remote client or router contacts the IPSec VPN gateway, providing authentication credentials. The central site device checks the provided credentials and pushes a configuration out to the remote device, thus establishing the VPN. QoS, firewall, ACL and NAT policies are established on a per-tunnel basis.

The VMDC 2.0 firewalling model includes per-tenant, VLAN-based virtual firewalls implemented at the services layer of the network. At this layer of the network, the abstracted virtual firewalls (vFWs) use stateful protocol and application inspection (via virtual device contexts running on FWSM or a standalone ASA appliance) to control and restrict traffic flows among groups of tenant VMs. ACLs filter L4-7 protocols, such as TCP, UDP, ICMP, SMTP, FTP, SNMP and HTTP, matching on port ID or source and destination IP addresses. If needed, NAT may be employed to further mask servers and VMs.

Zone-based firewalls can be used at the compute layer of the network, using differentiated "protected versus unprotected" virtual ports in Nexus 1000V Virtual Service Domains in conjunction with a virtual appliance firewall. In a multi-tenant context, use cases could include enabling users to self-administer their own zones, creating front and back-end server groups in order to separate their DMZ VMs from restricted back-end VMs, or simply to further scale policy application. This is described in greater detail in Application Tier Separation, page 3-34. Again, ACLs configured on the Nexus 1Kv may be employed for stateful L4-7 protocol inspection and filtering. PVLANs may further augment firewall capabilities at this tier, providing intra-VLAN, inter-VM traffic control.

Finally, SAN or NAS virtual firewalling mechanisms are employed to restrict and control access to tenant data stores. Described in greater detail in "Storage Separation," these include WWPN or WWNN zoning and LUN masking and the use of separation and abstraction techniques such as NetApps vFilers.

According to the SAFE Network Foundation Protection guidelines, Layer 2 switching security includes the following best practices:

- Restriction of broadcast domains
- STP Security
- DHCP protection
- ARP spoofing protection
- IP spoofing protection
- MAC flooding protection
- Use of VLANs to securely separate specific traffic types (i.e., to separate management and control traffic from data traffic)

The VMDC design builds upon these foundational Layer 2 switching security best practices. While these foundational security features have been part of Cisco's switching platforms for some time and are well understood, several of them are new to the Nexus 1000V in the 4.0(4)SV1 software release utilized in this phase of the VMDC design validation. These are DHCP snooping, dynamic ARP inspection, and IP source guard. These three features are focused on improving DHCP security and mitigating the possibility of DHCP DDoS and ARP "man in the middle" poisoning attacks by providing a means to differentiate between trusted ports (i.e., connected to DHCP servers) and untrusted ports (i.e., connected to clients). The virtual switch then functions as a mini-firewall, inspecting and blocking DHCP server-client messages from untrusted ports, ARP packets with invalid IP to MAC address bindings, and

IP packets with invalid source MAC or source IP addresses. As per established SAFE best practices, the VMDC design recommends these anti-spoofing features be applied uniformly across the switching infrastructure to function effectively.

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, policy must be applied 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 of the network. The result is that high-density compute architectures may require the distribution of security policies to the access tier.

To address some of these new security challenges and concerns, virtual firewalls may be deployed at the access layer of the data center infrastructure to create intra-tenant policy zones. Firewalls at the aggregation layer of the network enable one to define perimeter and per-tenant policies. Like firewalling at the aggregation layer, access layer firewalling can enforce security among the tiers of an application, as described in the next section, Application Tier Separation, page 3-34.

Storage Separation

To extend secure separation to the storage layer, we considered the isolation mechanisms available in a SAN environment. Cisco MDS storage area networks (SANs) offer true segmentation mechanisms, similar to VLANs in Ethernet. These mechanisms are called VSANs and work in conjunction with Fiber Channel (FC) zones. However, VSANs do not tie into the virtual host bus adapter (HBA) of a VM. VSANs and zones associate to a host, not to a VM. All VMs running on a particular host belong to the same VSAN or zone. Since it is not possible to extend SAN isolation to the VM, VSANs or FC Zones are used to isolate hosts from each other in the SAN fabric. To keep management overhead low, we do not recommend deploying a large number of VSANs. Instead, the Cisco VMDC solution leverages Fiber Channel soft zone configuration to isolate the storage layer on a per-host basis, and it combines that method with zoning via pWWN/device alias for administrative flexibility.

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.

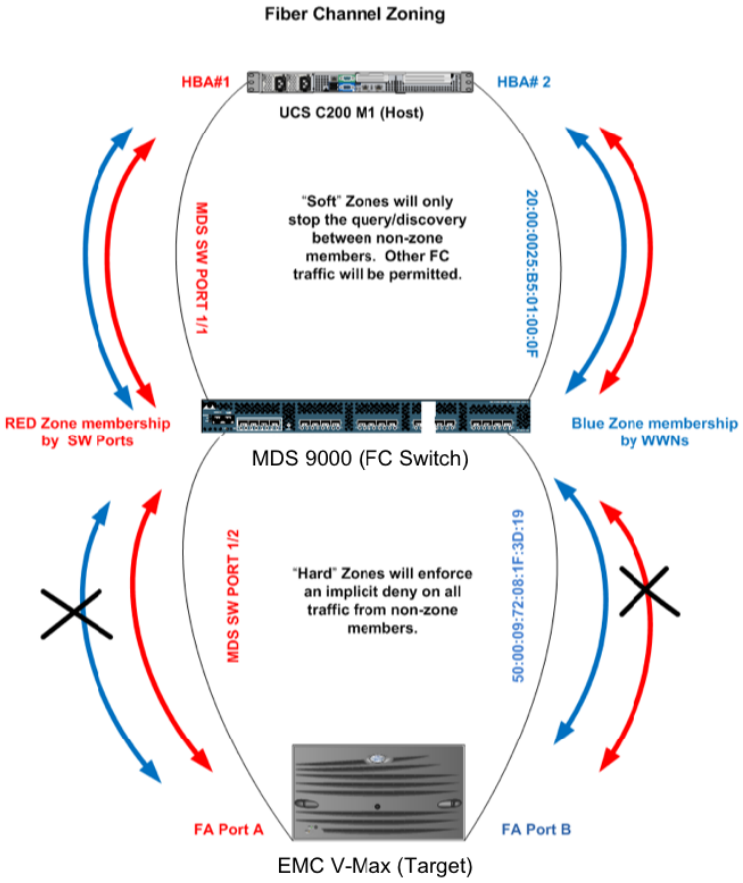
Fiber Channel Zones

SAN zoning can restrict visibility and connectivity between devices connected to a common Fiber 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.

Two methods of zoning exist (Figure 3-18):

- Soft zoning
- Hard zoning

Figure 3-18 Hard and Soft Zoning Example



Soft Zoning

Soft zoning is enforced only at the name server. When a device directs a frame toward a device that is not in the same zone, a soft zone implementation passes the frame. Therefore, soft zoning is less secure than hard zoning as it relies on the device to behave nicely and only send frames to members in the same zone. The benefit of soft zoning is that administrators can shift devices among ports without changing the zoning configuration. However, soft zones can be difficult to manage in large numbers, and it is difficult to track a device's location in the SAN because it can be easily moved between zones.

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, 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.

VM Data Store 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. Each virtual machine directory is stored in the Virtual Machine Disk (VMDK) sub-directory in the VMFS volume. While a VM is in operation, 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.

It is recommended to implement logical unit number (LUN) masking, an authorization process that makes a LUN available only to specific hosts on the EMC SAN as further protection against misbehaving servers corrupting disks belonging to other servers. This complements the use of zoning on the MDS, effectively extending zoning from the front end port on the array to the device that the physical disk resides on.

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 data stores within the SAN and at the file

Application Tier Separation

Many cloud-based applications follow a three-tiered component architecture. The tiers are typically web, application, and database. The web, or customer facing, tier communicates with the application and database tiers to process transactions. The application and database tiers serve as the middleware and backend components for the web tier. To ensure a higher degree of security, Enterprises deploy firewall security in front of the web tier and to inspect intra-application inter-tier communications. For such requirements, this design proposes using vApp firewalls. This design was validated using VMware's vShield for this purpose.

This design guide focuses on the design aspects of vApp firewalls rather than exhaustive details of vShield functionality. Feature implementation details are available in VMware documentation such as the vShield Zones Administration Guide.

The Cisco VMDC architecture proposes VLAN separation as the first degree of security in application tier separation. The design suggests that each application resides in separate VLANs within a tenant VRF. If communication must occur between tiers of an application, the traffic is routed through the default gateway where security access lists can enforce traffic inspection and access control.

At the access layer of the network, the vShield virtual firewall appliance monitors and restricts inter-VM traffic within and between ESX hosts. Security zones may be created based on abstracted VMware Infrastructure (VI) containers, such as clusters and VLANs, or at the VMware “Data Center” 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 provides a view of virtual machines, networks, and security policies and allows security posture audits in the virtual environment. Monitoring (VM Flow) is performed at the data center, cluster, portgroup, VLAN, and virtual machine levels.

A logical construct on the Nexus 1000 called a virtual service domain (VSD) allows one to classify and separate traffic for vApp-based network services such as firewalls. As of this writing, 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, which functions like a “bump in the wire”, serving to segment network traffic. The Service Virtual Machine (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.

To configure a VSD, three port profiles must be created: “Inside” (protected), “Outside” (unprotected) and “Member” (where the individual protected VMs reside). Following is an example configuration for creation of an inside port profile for VSD VSD1-customer1websrvrs (see Figure 3-13):

```
n1000v# config t
n1000v(config)# port-profile  _*customer1websrvrs*_  _*-inside*_
n1000v(config-port-prof)# switchport mode trunk
n1000v(config-port-prof)# switchport trunk allowed vlan  _*200*_
n1000v(config-port-prof)# virtual-service-domain  _*VSD1-customer1websrvrs*_
n1000v(config-port-prof)# no shut
n1000v(config-port-prof)# vmware port-group  _*customer1*_  _*websrvrs*_
_*-inside-protected*_
n1000v(config-port-prof)# service-port inside default-action forward
n1000v(config-port-prof)# state enabled
An outside port profile for the VSD "VSD1-customer1websrvrs " defined above would be
configured as follows:
n1000v(config)# port-profile  _*customer1websrvrs*_  _*-outside*_
n1000v(config-port-prof)# switchport mode trunk
n1000v(config-port-prof)# switchport trunk allowed vlan  _*200*_
n1000v(config-port-prof)# virtual-service-domain  _*VSD1-customer1websrvrs*_
n1000v(config-port-prof)# no shut
n1000v(config-port-prof)# vmware port-group  _*customer1*_  _*websrvrs*_
_*-outside-unprotected*_
n1000v(config-port-prof)# service-port outside default-action forward
n1000v(config-port-prof)# state enabled
```

The corresponding member port profile configuration for the VSD follows:

```
n1000v(config)# port-profile  _*customer1*_  _*websrvrs*_  _*-member*_
n1000v(config-port-prof)# switchport mode access
n1000v(config-port-prof)# switchport access vlan  _*200*_
n1000v(config-port-prof)# virtual-service-domain  _*VSD1-customer1websrvrs*_
n1000v(config-port-prof)# no shut
n1000v(config-port-prof)# vmware port-group  _*customer1websrvrs*_  _*-membervms*_
n1000v(config-port-prof)# state enabled
```

Correlation of the vShield agent vNICs to the requisite VSD port profiles is achieved during the vShield agent installation process, using the “Network Mapping” popup window. That is: source network vsmgmt would be mapped to destination network vsmgmt; source network protected would be mapped to customer1websrvrs-inside, and source network unprotected would be mapped to Customer1websrvrs-outside for the example above.

Additional parameters must be configured to bring up the vShield agent, including the vShield hostname, IP Address and Subnet mask for the vShield VM and IP Address for the vShield VM's default gateway. The vShield VM is then manually added to the vShield Manager inventory.

vCenter may be used to move selected VMs to the member port profile for the VSD. These will be protected by the vShield rulesets. vShield allows for application of two categories of rulesets: these are L4 (Layer 4) 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 may only be configured at the Data Center level. By default, all Layer 4 and Layer 2/Layer 3 traffic is allowed to pass. These are configured via a tab called the “VM Wall” tab. Note also that all vShield firewalls perform stateful inspection by default. All traffic is allowed by default, for ease of initial configuration.

Note that this default behavior is the reverse of default behavior for Cisco firewall solutions, wherein packets are implicitly denied in order to maximize secure configuration and operation. Thus, a best-practice security recommendation would be to first set a deny all packets rule, and then only allow specific traffic through the firewall.

Each vShield agent enforces VM Wall rules in descending order. 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, there is an additional qualification of rulesets using a hierarchy of precedence levels. This provides additional 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:

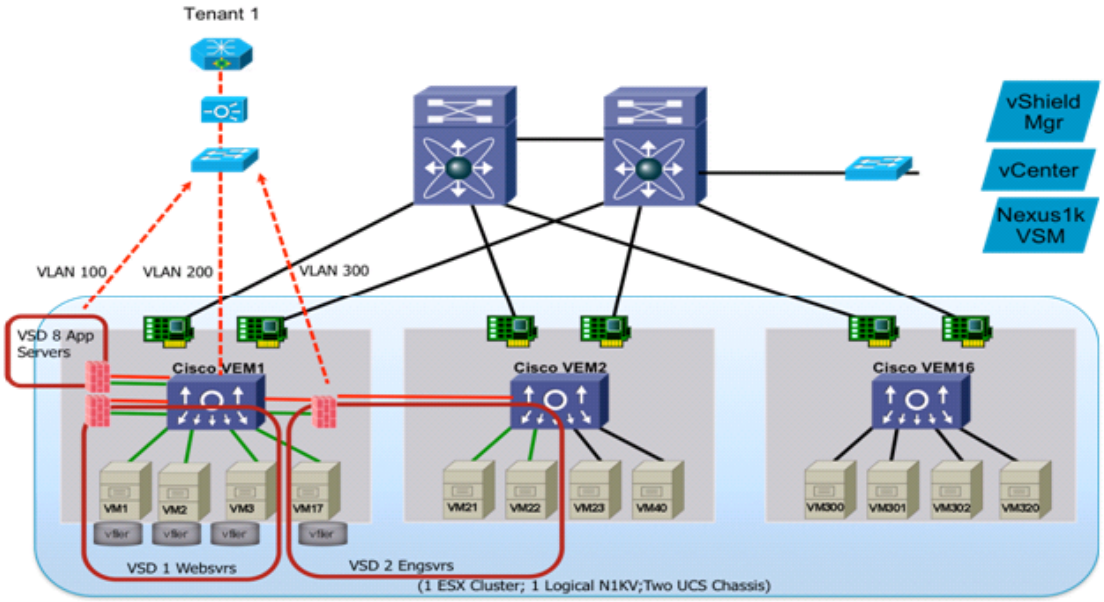
1. Data Center High Precedence Rules
2. Cluster Level Rules
3. Data Center Low Precedence Rules (i.e., “Rules below this level have lower precedence than cluster level rules” when a data center resource is selected)
4. Default Rules

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

- Container-level precedence refers to recognizing the data center level as being higher in priority than the cluster level. When a rule is configured at the data center level, all clusters and vShield agents within the clusters inherit it. A cluster-level rule is only applied to the vShield agents within the cluster. These must not conflict with higher precedence rules (i.e., Data Center High Precedence rules).
- Custom priority precedence refers to the option of assigning high or low precedence to rules at the data center level. High precedence rules work as noted in the container-level precedence description. Low precedence rules include the default rules and the configuration of Data Center Low Precedence rules. This allows one to recognize multiple layers of applied precedence.

“Vlan 200" from the configuration example above is illustrated in the following diagram. This shows application of a set of three VSDs/vShields for segmentation of server traffic for a specific tenant. In this example, the ESX cluster extends across two chassis of Menlo server blades, and the VMs are in a single N1KV virtual switch. Green lines indicated protected versus red, unprotected virtual ports.

Figure 3-19 vFWs in the VMDC Architecture



In the VMDC design, N1KV Service Virtual Machines and VSDs with vShield virtual firewalls were implemented to do the following:

- Define multiple groups of VSD policy zones and apply them to groups of servers for a specific tenant.
- Create a "Data Center" level rule that denies all traffic, then a higher precedence set of rules allowing only 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, etc). vShield tracks end-point mapper requests and also can learn which dynamic ports are listening on VMs and punch holes in this ephemeral port range only for trusted endpoints. Since ephemeral port ranges above 1024 are often used by botnets and rogue services, the VMDC design advocates using this feature to lock down these ports and set up "allow" rules only for specific ports for trusted endpoints.
- Use application-port pair mapping to create application aware rulesets.
- Validate movement of a vShield firewall policy to another vShield, following the movement of a VM due to a vMotion event, confirming that the VSDs for the affected vShield continued to operate as expected.

It is important to note that PVLANs may be utilized to complement vFW functionality, effectively creating sub-zones that may be used to restrict traffic between VMs within the same VLAN.

This type of distribution of firewall services and policy to the access tier of the network has distinct advantages, allowing for increased scale in hyper-dense compute environments and leveraging of VMware cluster HA technology to enhance firewall service availability. However, there are challenges: first and foremost is the need to scale policy management for larger numbers of enforcement points, and secondly, is the fact that vApp-based firewalls are relatively new paradigms, particularly in terms of understanding and managing firewall performance.

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 switch, 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 four to 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 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 data over a shared physical interface, the infrastructure cannot isolate the tenant production data. However, the Cisco VIC combined with VM-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.

Storage Separation—To extend the concept of secure separation all the way to the storage layer, this solution looked at possible isolation mechanisms available in SAN environments. True segmentation mechanisms such as VLANs in the Ethernet world are available in SAN networks. They are called VSANs and FC zones. However, VSANs do not have ties into the virtual HBA of a VM. VSANs associate to a host but not to VM granularity. All VMs in a particular host belong to the same VSAN or zone. In order to minimize management overhead it is not recommended to deploy VSANs in a large number. This solution particularly leveraged Fiber Channel pWWN soft zoning for isolation on a per host basis.

Fiber Channel Zones

SAN zoning provides a means of restricting visibility and connectivity between devices connected to a common Fibre Channel SAN. It is a built in security mechanism available in an FC switch. There is no traffic leaking between zones.

Zone provides effective segmentation and tenant separation in the SAN network at a physical host level. Two methods of zoning are possible:

- Soft zoning
- Hard zoning

Soft Zoning

Soft zoning means that the switch will place WWNs of devices in a zone, and it doesn't matter what port they're connected to. If WWN Q, for example, lives in the same soft zone as WWN Z, they will be able to talk to each other. Likewise, if Z and A are in a separate zone, they cannot see each other. It simply means that the enforcement relies on the WWN of the node in the fabric. The benefit to using soft zones is that you can connect to any port on a switch, and know that you'll have access to the other nodes you're supposed to see. At the same time soft zoning may pose difficulties in terms of managing large number of zones. It may also be challenging to keep track of device location in the SAN because they can be easily moved around based on WWN.

Hard Zoning

Hard zoning is applied to either switch ports or end-station name. By port zoning, you're configuring a particular device to be part of a zone. A zone provides a logical boundary preventing this port from talking to un-authorized ports in different zones. Similarly name zoning restricts access by a device's World Wide Name.

Zoning does however have its limitations. Zoning was designed to do nothing more than prevent devices from communicating with other unauthorized devices. It is a distributed service that is common throughout the fabric. Any installed changes to a zoning configuration are therefore disruptive to the entire connected fabric. A fabric reset will cause everyone to re-login at the same time, and fabric updates get sent to everyone.

Zones provide isolation on per host basis. A host contains HBAs that are the initiators, identified by WWN name. These are associated with a target storage array FA port. The host can only communicate with other initiators and targets within its particular zone.

VMware uses a cluster file system called virtual machine file system or VMFS. The VMs on an ESX host associate with virtual machine disks (VMDKs) in a VMFS volume which is made up of one large logical unit (LUN). Each virtual machine directory is stored in a unique VMDK sub-directory in the VMFS volume. When a VM is operating VMFS has a lock on its files so that other ESX servers cannot update them. When a VMDK sub-directory is associated to a VM it is tied to that VM only and multiple VMs do not have access to the same VMDK directory.

It is recommended to implement LUN masking, an authorization process that makes a Logical Unit Number (LUN) available only to specific hosts on the EMC SAN as further protection against misbehaving servers corrupting disks belonging to other servers. This complements the use of zoning on the MDS.

Tighter control and isolation of tenant data stores can be achieved by mapping storage LUNs per VM using raw disk map filer system (RDM). Each RDM volume is mapped to a single VM. However there is physical upper limit of 255 LUNs per cluster. Thus while technically possible, a 1:1 mapping of LUNs to tenant VMs is not recommended, as this approach 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 data stores within the SAN and at the filesystem level, serving to limit the effect of VM-based exploits or inadvertent disk corruption.

APPENDIX A

Related Documentation

The design recommends that general Cisco data center design best practices are leveraged as the foundation for implementing Infrastructure as a Service deployment. The following CVD companion documents provide information about these best practices:

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#wp1043848

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