Cisco Virtualized Multi-Tenant Data Center, Version 2.2
Design Guide

Last Updated: March 6, 2013

This CVD supersedes the document “Cisco Data Center Infrastructure 2.5 Design Guide” and the document “Data Center Design-IP Network Infrastructure”.

Cisco Validated Design

CCDE, CCENT, CCSI, Cisco Eos, Cisco Explorer, Cisco HealthPresence, Cisco IronPort, the Cisco logo, Cisco Nurse Connect, Cisco Pulse, Cisco SensorBase, Cisco StackPower, Cisco StadiumVision, Cisco TelePresence, Cisco TrustSec, Cisco Unified Computing System, Cisco WebEx, DCE, Flip Channels, Flip for Good, Flip Mino, Flipshare (Design), Flip Ultra, Flip Video, Flip Video (Design), Instant Broadband, and Welcome to the Human Network are trademarks; Changing the Way We Work, Live, Play, and Learn, Cisco Capital, Cisco Capital (Design), Cisco:Financed (Stylized), Cisco Store, Flip Gift Card, and One Million Acts of Green are service marks; and Access Registrar, Aironet, AllTouch, AsyncOS, Bringing the Meeting To You, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, CCVP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Lumin, Cisco Nexus, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Collaboration Without Limitation, Continuum, EtherFast, EtherSwitch, Event Center, Explorer, Follow Me Browsing, GainMaker, iLYNX, IOS, iPhone, IronPort, the IronPort logo, Laser Link, LightStream, Linksys, MeetingPlace, MeetingPlace Chime Sound, MGX, Networkers, Networking Academy, PCNow, PIX, PowerKEY, PowerPanels, PowerTV, PowerTV (Design), PowerVu, Prisma, ProConnect, ROSA, SenderBase, SMARTnet, Spectrum Expert, StackWise, WebEx, and the WebEx logo are registered trademarks of Cisco and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1002R)

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

The Cisco implementation of TCP header compression is an adaptation of a program developed by the University of California, Berkeley (UCB) as part of UCB's public domain version of the UNIX operating system. All rights reserved. Copyright © 1981, Regents of the University of California.

NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED "AS IS" WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Cisco Virtualized Multi-Tenant Data Center, Version 2.2 Design Guide
© 2012 Cisco Systems, Inc. All rights reserved.

CONTENTS

Preface

The Cisco® Virtualized Multi-Tenant Data Center (VMDC) solution provides design and implementation guidance for enterprises deploying private cloud services and service providers building virtual private and public cloud services. The Cisco VMDC solution integrates various Cisco and third-party products that are part of the cloud computing ecosystem.

VMDC 2.2 is an incremental release, leveraging and only slightly modifying the architecture defined in the preceding parent 2.0 release. In this phase of the VMDC solution, we present incremental enhancements to the multi-tenant security models outlined in the previous 2.0 system release, introducing defense-in-depth firewalling utilizing the new Cisco Virtual Security Gateway in combination with the ASA appliance and reworking the end-to-end QoS framework to accommodate multimedia SaaS applications such as the Cisco Collaboration solutions. We also begin to examine issues of hybrid (public/private) interworking from the aspect of VM migration, and look at Service Provider (i.e., intra-organizational) Data Center Interconnect in the context of VPLS transport, focusing on Nexus 7000/ASR 9000 interoperability.

Product screen shots and other similar material in this document are used for illustrative purposes only and are VMAX (EMC Corporation), NetApp FAS3240 (NetApp), vSphere (VMware, Inc.), respectively. All other marks and names mentioned herein may be trademarks of their respective companies. The use of the word "partner" or "partnership" does not imply a legal partnership relationship between Cisco and any other company.

Introduction

Interest in cloud computing over the last several years has been phenomenal. For cloud providers, public or private, it will transform business and operational processes, streamlining customer on-ramping and time to market, facilitating innovation, providing cost efficiencies, and enabling the ability to scale resources on demand.

Cisco's Virtualized Multi-tenant Data Center (VMDC) system defines an end-to-end architecture, which an organization may reference for the migration or build out of virtualized, multi-tenant data centers for new cloud-based service models such as Infrastructure as a Service (IaaS).

The system builds upon these foundational pillars in terms of architectural approach:

- **Secure Multi-tenancy—**Leveraging traditional security best practices in a multi-layered approach to secure the shared physical infrastructure and those logical constructs that contain tenant-specific resources, while applying new technologies to provide security policy and policy mobility to the virtual machine level insures the continued ability to enforce and comply with business and regulatory policies, even in a highly virtualized multi-tenant environment.
- **Modularity—**A pod-based modular design approach mitigates the risks associated with unplanned growth, providing a framework for scalability that is achievable in manageable increments with predictable physical and cost characteristics, and allowing for rapid time-to market through streamlined service instantiation processes.
- **High Availability—**Building for carrier-class availability through platform, network, and hardware and software component level resiliency minimizes the probability and duration of service-affecting incidents, meaning that Private IT and Public Cloud administrators can focus on supporting the bottom line rather than fighting fires.
- **Differentiated Service Support—**Defining logical models around services use cases results in a services-oriented framework for systems definition, insuring that resources can be applied and tuned to meet tenant requirements.
- **Service Orchestration—**Dynamic application and re-use of freed resources is a key aspect of a Cloud-based operations model, thus the ability to properly represent abstractions of the underlying tenant-specific resources and services is a fundamental requirement for automated service orchestration and fulfillment; this is accomplished in the VMDC architecture through continued evolution of network container definitions which can be leveraged by in-house middleware and partner management solutions.

Intended 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.

Related Documents

The following documents are available for reference:

- Cisco Virtualized MultiTenant Data Center Design Guide Release 1.1
- Cisco Virtualized Multi-Tenant Data Center 2.0 Design and Implementation Guide
- Cisco Virtualized Multi-Tenant Data Center, Version 2.1, Implementation Guide
- Design Considerations for Classical Ethernet Integration of the Cisco Nexus 7000 M1 and F1 Modules
- Virtualized Multi-Tenant Data Center New Technologies - VSG, Cisco Nexus 7000 F1 Line Cards, and Appliance-Based Services

- Cisco Virtualized Multi-Tenant Data Center Implementation Guides, Releases 1.0-2.2 (available under NDA) are located at: http://sdu.cisco.com/systems/system.php?sysid=22
- Data Center Interconnect over MPLS, Ethernet or IP Transport documents are located at: http://www.cisco.com/en/US/netsol/ns749/networking_solutions_sub_program_home.html and at: http://www.cisco.com/en/US/netsol/ns975/index.html

About Cisco Validated Designs

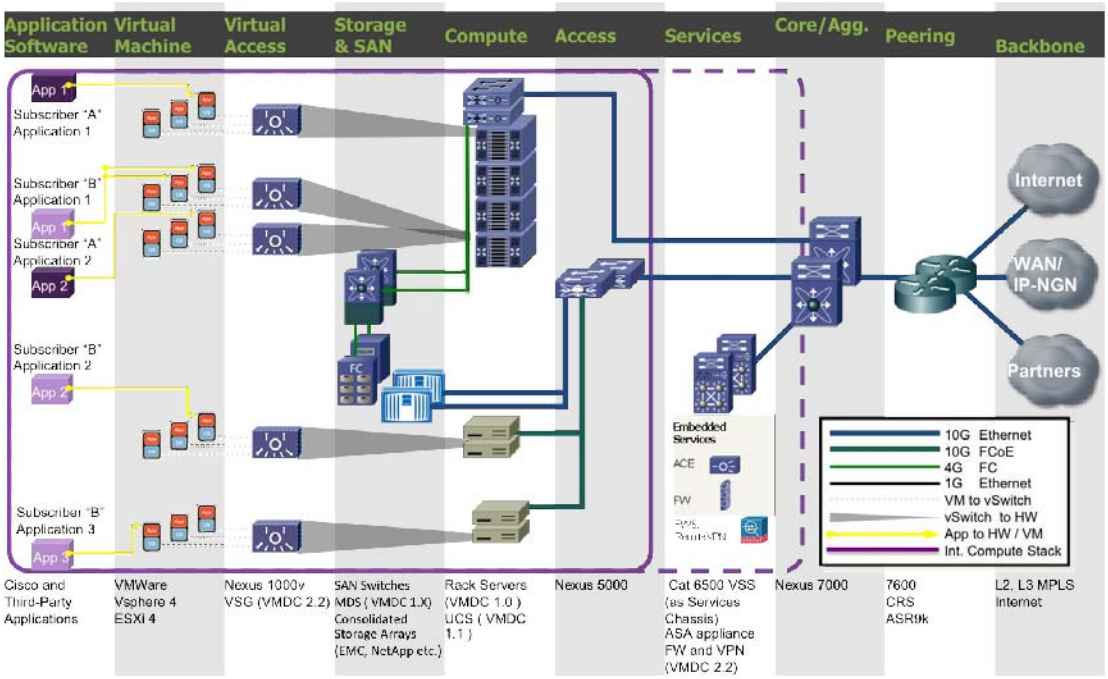
The Cisco Validated Design Program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit http://www.cisco.com/go/validateddesigns.

CHAPTER 1

Architecture Overview

The Virtualized Multi-tenant Data Center (VMDC) reference architecture defines an end-end system suitable for service deployment in a public or private "Cloud" model. Though this document focuses mainly on design considerations specific to aspects of the data center, the "big picture" is that the end-to-end system includes the wide area network (WAN), multiple "provider" data centers, and the tenant's resources on their premise. In the public cloud case, the tenant would typically be located remotely, and have their own data center resources on site in addition to resources within the cloud; in the private case, the tenant could reside locally in another organizational unit logically separated from the IT data center or be located at another facility.

Figure 1-1 System Overview



Cloud Data Center

At a macro level, the Cloud data center consists of network, storage, and compute resources, but it is important to keep in mind that it is part of a larger end-to-end system which includes the following components:

1. **Data Center** (typically interconnected in a system of multiple data centers).
2. **Wide Area Backbone** or IP/Next Generation Network (NGN) (public provider backbone) network.
3. **Public Internet Access**
4. **The Tenant Premise—**In the private case, the "tenant" can be an organization unit with their own compute resources separated logically from the IT data center, or could be accessing their private cloud remotely in a mobile fashion via Secure Sockets Layer (SSL) or IPSec VPN connection, or in a branch or alternative data center/campus In the public case, the "Enterprise-grade" tenant will typically be accessing their resources within the cloud remotely from within their Enterprise environment (i.e., within the Enterprise data center or campus).
5. **Management**—This is a superset of normal data center administration functions over storage, compute, and network resources, including elements which allow for more dynamic resource allocation and automated processes (i.e., an administrative or tenant user portal, service catalog, and workflow automation).

This section discusses the following aspects of the Cloud data center:

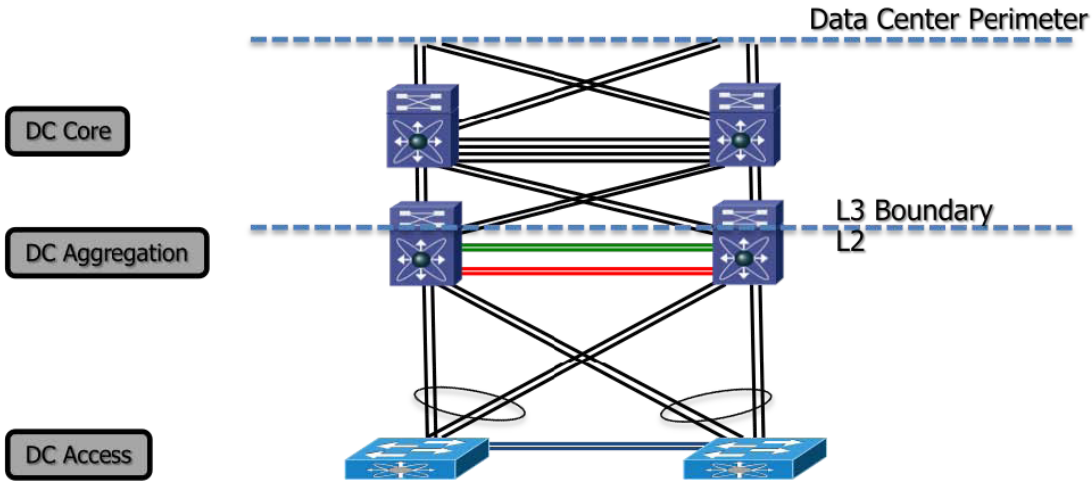
Hierarchical Network Architecture

The data center within the VMDC 2.2 reference architecture is based upon the classic multi-layer hierarchical network model. In general, such a model implements three layers of hierarchy:

- **Core Layer**, characterized by a high degree of redundancy and bandwidth capacity and thus optimized for availability and performance.
- **Aggregation Layer**, characterized by a high degree of high-bandwidth port density capacity and thus optimized for traffic distribution and link fan-out capabilities to access layer switches. Functionally, the nodes in the aggregation layer typically serve as the L2/L3 boundary.
- **Access Layer**, serving to connect hosts to the infrastructure and thus providing network access, typically at Layer 2 (L2) (i.e., LANs or VLANs).

Figure 1-2 shows these three layers of the hierarchical model.

Figure 1-2 Three-Layer Hierarchical Model



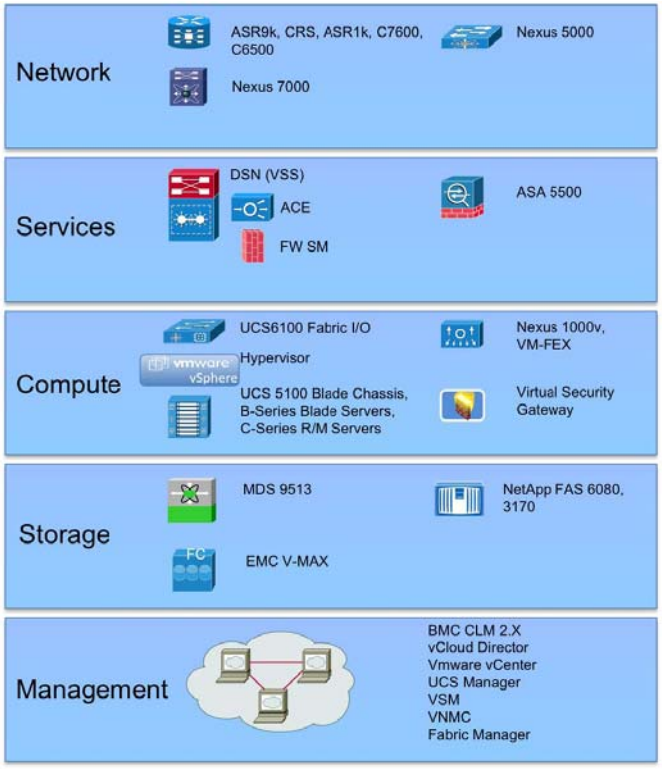
Benefits of such a hierarchical model include scalability, resilience, performance, maintainability, and manageability. The hierarchical design represents a structured approach to building the infrastructure, allowing for relatively easy expansion in modular increments. Redundant nodes and links at each level insure no single point of failure, while link aggregation can be engineered for optimal bandwidth and performance through the aggregation and core layers. Devices within each layer perform the same functions; this consistency simplifies troubleshooting and configuration. The effect is ease of maintenance at lower operational expense.

VMDC Layers

Figure 1-3 shows the following functional layers that comprise the VMDC data center:

- Network
- Services
- Compute
- Storage
- Management

Figure 1-3 Functional Layers Within the VMDC Data Center



The Network layer includes the WAN/PE router, which forms the data center perimeter to the Enterprise wide area or provider IP/NGN backbone and to the public Internet. These perimeter nodes may be dedicated to Layer 3 (L3) routing functions or may be multi-service in nature, providing L2 interconnects between data centers as well as L3 services. WAN/PE routers validated within the VMDC reference system architecture include: the Cisco CRS-1, Cisco ASR 9000, Cisco Catalyst 7600, Catalyst 6500, and ASR 1000. The Network layer also includes the aforementioned, classic three-layer hierarchy of switching nodes. Within the VMDC reference architecture, this portion of the infrastructure is comprised of Nexus 7000 systems, serving as the core (i.e., Nexus 7010) and aggregation (i.e., Nexus 7018) nodes, and the Nexus 5000 system as the access nodes. As shown in Figure 1-10, validated VMDC topologies feature two variants of the three-layer hierarchical model: a collapsed core/aggregation version, and a collapsed aggregation/access version. These allow for fine-tuning of port capacity and bandwidth to the level of aggregation or access density required to accommodate current and anticipated scale requirements.

The Services layer comprises network and security services such as firewalling, server load balancing, SSL offload, intrusion prevention, network analysis, and gateway functions. A distinct difference arises between the conventional data center services layer and "cloud" data center services layer in that the solution set for the latter must support application of L4 - L7 services at a per-tenant level, through logical abstraction of the physical resources. Centralized services are most useful in applying policies that are broadly applicable across a range of tenants (or workgroups in the private case). Within the VMDC reference architecture, the Data Center Services Node (DSN) provides firewalling and server load balancing services, in a service module form factor (i.e., the ACE30 and FWSM or ASA-SM modules); alternatively, these are available in appliance form-factors. This layer also serves as the termination point for remote access IPSec or SSL VPNs; within the VMDC architecture, the Cisco ASA 5580 appliance connected to the DSN fulfills this function, securing remote tenant access to cloud resources.

The Compute layer includes several sub-systems. The first is a virtual access switching layer, which allows for extension of the L2 network across multiple physical compute systems. This virtual access switching layer is of key importance in that it also logically extends the L2 network to individual virtual machines within physical servers. The feature-rich Cisco Nexus 1000V generally fulfills this role within the architecture. Depending on the level of software functionality (i.e., QoS or security policy) or scale required, the VM-FEX may be a hardware-based alternative to the Nexus 1000V. A second sub-system is that of virtual (i.e., vApp-based) services. These may include security, load balancing, and optimization services. Services implemented at this layer of the infrastructure will complement more centralized service application, with unique applicability directly to a specific tenant or workgroup and their applications. Specific vApp based services validated within the VMDC architecture as of this writing include the Cisco Virtual Security Gateway (VSG), providing a security policy enforcement point within the tenant virtual data center or Virtual Private Data Center (VPDC). The third sub-system within the Compute layer is the computing resource. This includes physical servers, hypervisor software providing compute virtualization abilities, and the virtual machines thus enabled. The Cisco Unified Computing System (UCS), featuring redundant 6100 Fabric Interconnects, UCS 5108 Blade Chassis, and B-Series Blade or C-Series RackMount servers, comprise the compute resources utilized within the VMDC reference architecture.

The Storage layer provides storage resources. Data stores will reside in SAN (block-based) or NAS (file-based) storage systems. SAN switching nodes implement an additional level of resiliency, interconnecting multiple SAN storage arrays to the compute resources, via redundant FC (or perhaps Fibre Channel over Ethernet (FCoE)) links.

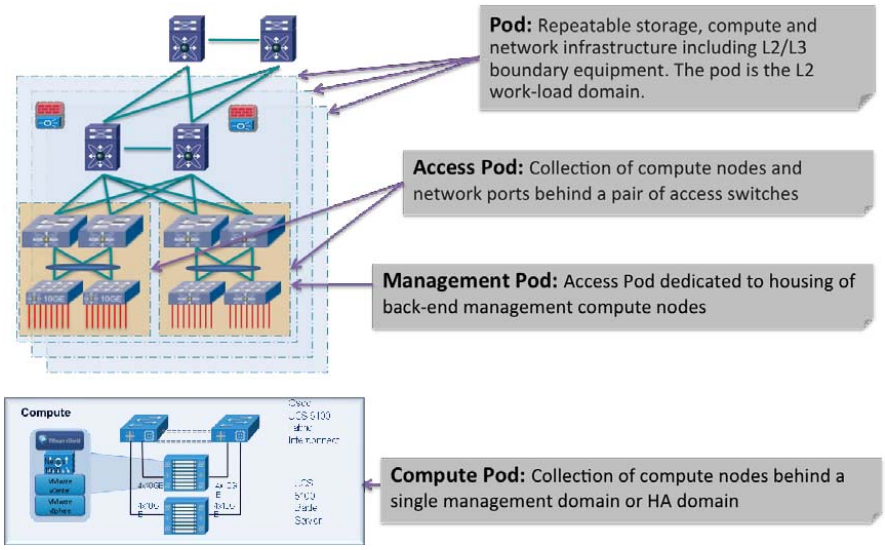
The Management layer consists of the "back-end" hardware and software resources required to manage the multi-tenant infrastructure. These include domain element management systems, as well as higher level service orchestration systems. The domain management systems currently validated within VMDC include Cisco UCS Manager, VMware vCenter, and vCloud Director for compute resource allocation; EMC's UIM and Cisco Fabric Manager for storage administration; and Cisco VSM and Virtual Network Management Center (VNMC) for virtual access and virtual services management. Automated service provisioning, including cross-resource service orchestration functions, are provided by BMC's Cloud Lifecycle Management (CLM) system. Service orchestration functions were not in scope for this VMDC system release.

Building Blocks

The Pod

Previous iterations of the VMDC reference architecture defined resource containers called "pods" that serve as the basis for modularity within the Cloud data center. As a homogenous modular unit of network, compute, and storage resources, the pod concept allows one to address environmental, physical, logical, and application-level requirements in a consistent way. The pod serves as a blueprint for incremental build-out of the Cloud data center in a structured fashion; when resource utilization within a pod reaches a pre-determined threshold (i.e., 70-80%), the idea is that one simply deploys a new pod. From a service fulfillment and orchestration perspective, a pod represents a discrete resource management domain.

Figure 1-4 Pod Concept



In general practice, the pod concept may serve simply as a framework, with designers defining their own variants tuned to specific environmental or performance characteristics. Figure 1-4 shows that a pod can be defined at different levels of modularity, supporting growth in differing increments. However, within the VMDC reference architecture, a general purpose utility compute pod extends from the compute and storage layers to the L2 ports on the aggregation nodes serving as the L2/L3 boundary and up to and including components within the network services layer. The port and MAC address capacity of the aggregation nodes are thus key factors in determining how many pods a single pair of aggregation nodes will support within the Cloud data center.

Special Purpose Pods

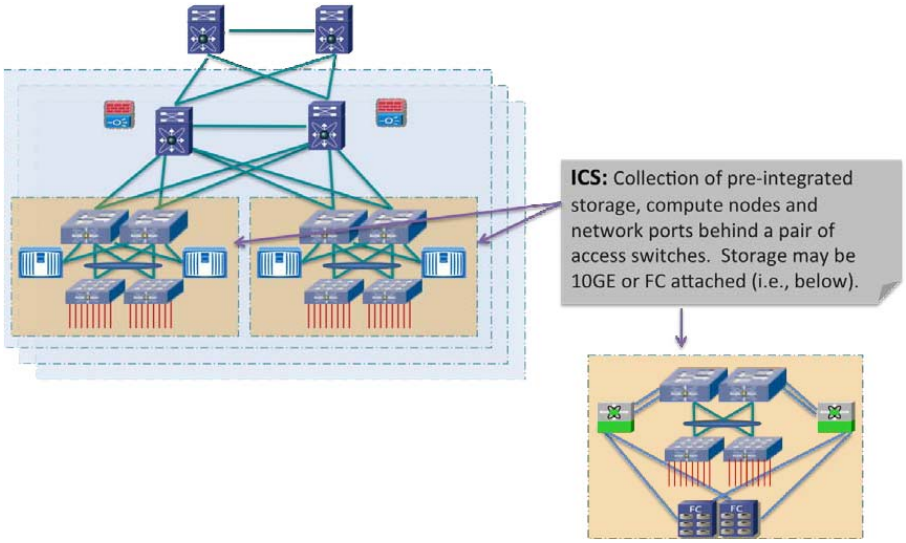
A major premise behind building general purpose homogeneous compute pods and applying logical segmentation overlays to meet business or security policy requirements is that this maximizes utilization of resources. However, in some cases there may be a unique requirement - for ease of operation, special performance tuning, or to meet special security objectives - to physically separate some of the compute nodes out from a general purpose pod and place them in a dedicated, perhaps application-specific pod. The VMDC architecture provides the flexibility to build special purpose pods. Such is the case with the management pod concept.

Back-end management compute nodes may be placed within a general purpose compute pod, and logically isolated and firewalled from production hosts. For smaller, less complex or more streamlined environments, this is an excellent option. However, in larger environments, a separate pod dedicated to back-end management servers (i.e., bare metal and virtualized) is recommended. In the various VMDC 2.X releases, the as-tested systems have in fact included a separate access pod in which servers are dedicated to back-end infrastructure management functions. The benefits of this option include creation of a more discrete troubleshooting domain in the event of instability or failures. The architecture flexibility allows for logical isolation and firewalling or for dedicated firewalls (physical or in vApp form) to be placed on the perimeter of the management container. In practice, role-based access controls (RBAC) tied to directory services would be applied to categorize and limit user access and change control authority as per their functional roles within the organization.

The Integrated Compute Stack

An Integrated Compute Stack (ICS) represents another potential unit of modularity within the VMDC Cloud data center, representing a sub-component within the pod. An ICS is a pre-integrated collection of storage, compute, and network resources, up to and including L2 ports on a pair of access switching nodes. Figure 1-5 shows the location of the ICS within a pod. Multiples of ICSs will be deployed like building blocks to fill the capacity of a pod.

Figure 1-5 ICS Concept



Working with eco-system partners, Cisco currently supports two ICS options:a Vblock and a flexpod. A Vblock comprises Cisco UCS and EMC storage systems, offered in several combinations to meet price, performance, and scale requirements. Similarly, a Flexpod also combines UCS compute and storage resources, however in this case, NetApp storage systems apply. Flexpods are offered in a range of sizes designed to achieve specific workload requirements. The VMDC reference architecture will accommodate more generic units of compute and storage, including storage from other third-party vendors, however the business advantage of an ICS is that pre-integration takes the guesswork out of balancing compute processing power with storage input/output operations per second (IOPS) to meet application performance requirements.

SAN Architecture

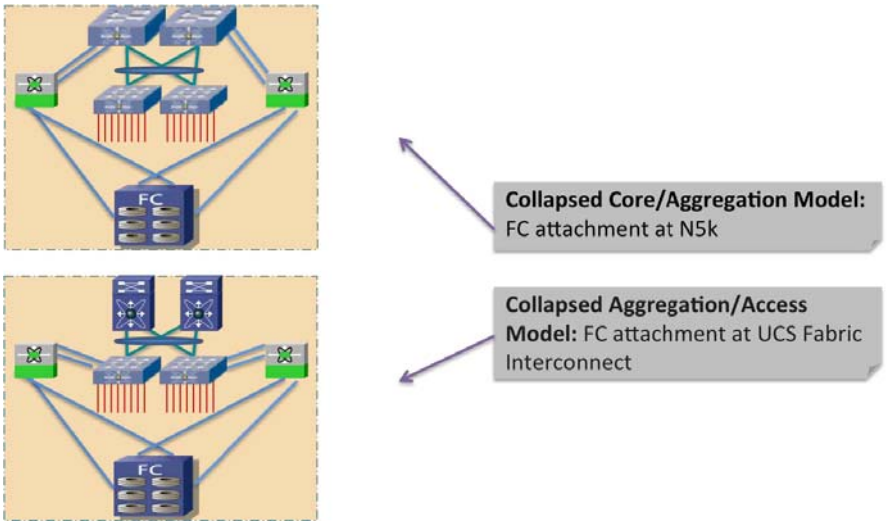
The VMDC SAN architecture remains unchanged from previous (2.0) programs. It follows current best practice guidelines for scalability, high availability, and traffic isolation. Key design aspects of the architecture include:

- Leverage of Cisco Data Center Unified Fabric to optimize and reduce LAN and SAN cabling costs
- High availability through multi-level redundancy (link, port, fabric, Director, RAID)
- Risk mitigation through fabric isolation (multiple fabrics, VSANs)
- Data store isolation through NPV/NPIV virtualization techniques, combined with zoning and LUN masking.

The hierarchical, pod-based infrastructure model described in this document lends itself to two possible attachment points for storage: within the pod and/or at the aggregation nodes - i.e., distributed or centralized. In practice, which option is most suitable for a particular deployment will depend on

application characteristics and anticipated traffic patterns for interactions involving data store access. Companies often employ both options in order to satisfy specific application requirements and usage patterns. In terms of the VMDC validation work, the focus to date has been on consideration of storage as a distributed, pod-based resource. This is based on the premise that in a hierarchical, cloud-type data center model, it is more efficient in terms of performance and traffic flow optimization to locate data store resources as close to the tenant hosts and vApps as possible. In this context, given the two hierarchical topology variants defined (i.e., collapsed core/aggregation and collapsed aggregation/access), we have two methods of attaching Fiber Channel storage components into the infrastructure: the first follows the ICS model of attachment via the Nexus 5000; the second provides for attachment at the UCS Fabric Interconnect, Figure 1-6.

Figure 1-6 FC SAN Attachment Options



In both scenarios, Cisco's unified fabric capabilities are leveraged with converged network adapters (CNAs) providing "SAN-ready" servers, and N-Port Virtualizer on the UCS Fabric Interconnect or Nexus 5000 top-of-rack (ToR) switches enabling each aggregated host to be uniquely identified and managed through the fabric and over uplinks to the SAN systems. Multiple FC links are used from each (redundant) Nexus 5000 or UCS Fabric Interconnect to the MDS SAN switches, in order to match the current maximum processing capability of the SAN system and thus eliminate lack of bandwidth between the SAN components and their point of attachment to the network infrastructure as a potential bottleneck.

Though the diagrams above show simple SAN switching topologies, it is important to note that if greater SAN port switching capacity is required, the architecture supports (is validated with) more complex, two-tier core-edge SAN topologies, documented in the VMDC 2.0 "Compact Pod Implementation Guide," and the Cisco SAN switching best practice guides, available at http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5990/ white_paper_C11-515630.html.

Compute Architecture

The VMDC compute architecture is based upon the premise of a high degree of server virtualization, driven by data center consolidation, the dynamic resource allocation requirements fundamental to a "cloud" model, and the need to maximize operational efficiencies while reducing capital expense (CAPEX). The architecture is thus based upon three key elements:

1. Hypervisor-based virtualization: in this as in previous system releases, VMware's vSphere plays a key role, enabling the creation of virtual machines on physical servers by logically abstracting the server environment in terms of CPU, memory, and network touch points into multiple virtual software containers.
2. Unified Computing System (UCS): unifying network, server and I/O resources into a single, converged system, the Cisco UCS provides a highly resilient, low-latency unified fabric for the integration of lossless 10-Gigabit Ethernet and FCoE functions with x-86 server architectures. The UCS provides a stateless compute environment that abstracts I/O resources and server personality, configuration and connectivity, facilitating dynamic programmability. Hardware state abstraction makes it easier to move applications and operating systems across server hardware.
3. The Cisco Nexus 1000V provides a feature-rich alternative to VMware's Distributed Virtual Switch, incorporating software-based VN-link technology to extend network visibility, QoS, and security policy to the virtual machine level of granularity.

This system release utilizes VMware's vSphere 4.1 as the compute virtualization operating system. A complete list of new enhancements available with vSphere 4.1 is available online. Key "baseline" vSphere functionality leveraged by the system includes ESXi boot from SAN, VMware High Availability (VMware HA), and Distributed Resource Scheduler (DRS).

Fundamental to the virtualized compute architecture is the notion of clusters; a cluster consists of two or more hosts with their associated resource pools, virtual machines, and data stores. Working in conjunction with vCenter as a compute domain manager, vSphere's more advanced functionality, such as HA and DRS, is built around the management of cluster resources. vSphere supports cluster sizes of up to 32 servers when HA and/or DRS features are utilized. In general practice however, the larger the scale of the compute environment and the higher the virtualization (VM, network interface, and port) requirement, the more advisable it is to use smaller cluster sizes in order to optimize performance and virtual interface port scale. Therefore, in VMDC large pod simulations, cluster sizes are limited to eight servers; in smaller pod simulations, cluster sizes of 16 or 32 are utilized. As in the VMDC 2.0 release, three compute profiles (Gold, Silver, and Bronze) are created to represent large, medium, and small workload types. Gold has 1 vCPU/core and 16G RAM; Silver has .5 vCPU/core and 8G RAM, and Bronze has .25 vCPU/core and 4G of RAM.

The UCS-based compute architecture has the following characteristics:

- It comprises multiple UCS 5100 series chassis (5104s), each populated with eight (half-width) server blades.
- Each server has dual 10GigE attachment - i.e., to redundant A and B sides of the internal UCS fabric.
- The UCS is a fully redundant system, with two 2100 Series Fabric Extenders per chassis and two 6100 Series Fabric Interconnects per pod.
- Internally, four uplinks per Fabric Extender feeding into dual Fabric Interconnects pre-stage the system for the maximum bandwidth possible per server; this means that for server to server traffic within the UCS fabric, each server will have 10GigE bandwidth.

- Each UCS 6100 Fabric Interconnect aggregates via redundant 10GigE EtherChannel connections into the access switch (i.e., Nexus 7000). The number of uplinks provisioned will depend upon traffic engineering requirements. For example, in order to provide an eight-chassis system with an 8:1 oversubscription ratio for internal fabric bandwidth to aggregation bandwidth, a total of 80G (8x10G) of uplink bandwidth capacity must be provided per UCS system (Figure 2-20).
- Similarly to the Nexus 5000 FC connectivity in the compact pod design, an eight-port FC GEM in each 6140 Expansion Slot provides 4Gig FC to the Cisco MDS 9513 SAN switches (i.e., 6140 chassis A, 4 x 4G FC to MDS A and 6140 chassis B, 4 x 4G FC to MDS B). In order to maximize IOPS, the aggregate link bandwidth from the UCS to the MDS should match the processing capability of the storage controllers.

The Nexus 1000V functions as the virtual access switching layer, providing per-VM policy and policy mobility.

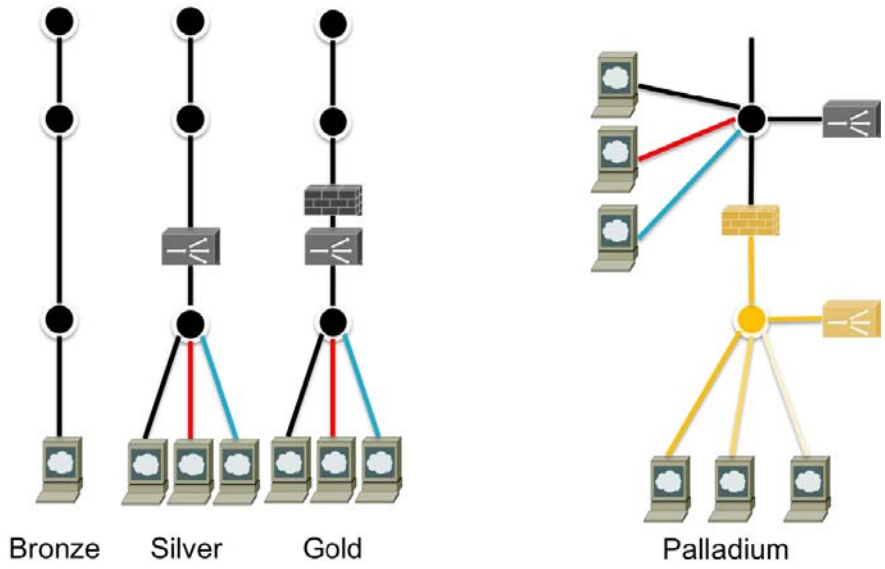
Multi-Tenancy Architecture

Virtualization of compute and storage resources enables sharing across an organizational entity. In contrast, virtualized multi-tenancy, a concept at the heart of the VMDC reference architecture, refers to the logical isolation of shared virtual compute, storage, and network resources. In essence, this is "bounded" or compartmentalized sharing. A tenant is a user community with some level of shared affinity. For example, within an Enterprise, a tenant may be a business unit, department, or workgroup. Depending upon business requirements or regulatory policies, a tenant "compartment" may stretch across physical boundaries, organizational boundaries, and even between corporations. A tenant container may reside wholly within their private cloud or may extend from the tenant's Enterprise to the provider's facilities within a public cloud. The VMDC architecture addresses all of these tenancy use cases through a combination of secured datapath isolation and a tiered security model which leverages classical security best practices and updates them for the virtualized multi-tenant environment.

Tenancy Use Cases

Earlier VMDC releases (2.0 and 2.1) presented four tenancy models. High-level, logical depictions of these models are shown in Figure 1-7. The first three provided a baseline, simple set of tenant containers, which were combined with different levels of network services in a tiered fashion - hence the Bronze, Silver, and Gold nomenclature. The two most interesting containers from this set are Bronze and Gold. Bronze seemingly is the most basic, but simplicity broadens its applicability. One tends to think of these containers as single-tenant in nature, but in practice, a Bronze container may be used to support multiple tenants, with homogenous requirements; i.e., similar workload profiles, QoS, or security policies, or perhaps this is a community of interest using the same application set. A Gold container, with both firewalling and server load balancing applied, assumes a higher degree of security and availability is required as a baseline in order to support the tenant applications. As in the Silver container, multiple VLANs support logical segmentation for N-tiered applications. The idea is that one could combine these tenant containers together in various combinations to support more complex scenarios if desired. The fourth container type demonstrates a further incremental evolution of tenancy models from simple multi-segment containers toward logical approximations of a virtual data center overlay on the physical shared infrastructure. With the notion of a separate front-end and back-end set of zones, each of which may have a different set of network services applied, the Palladium container begins to more closely align with traditional zoning models in use in physical IT deployments.

Figure 1-7 Initial Four VMDC Tenancy Models

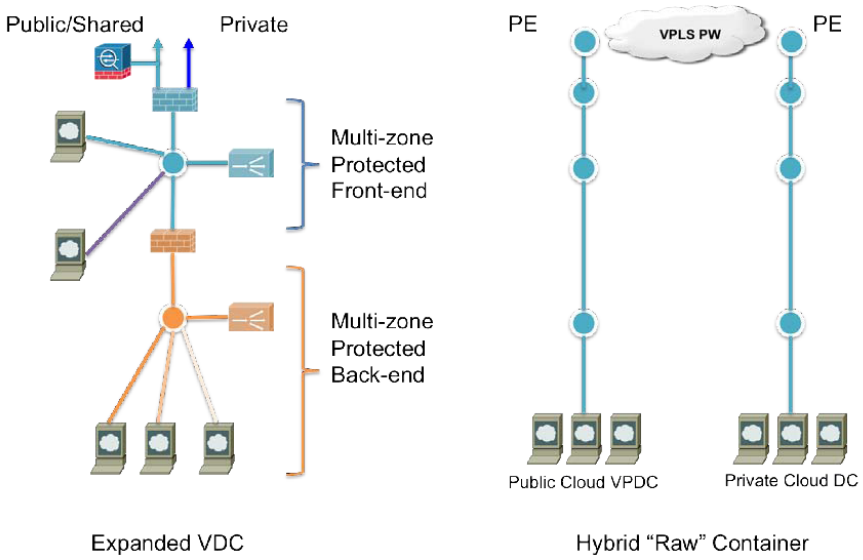


New Tenancy Models Introduced in VMDC 2.2

Two new tenancy models are introduced in VMDC 2.2. The first incrementally evolves the virtual data center concept, providing more expansion of protected front-end and back-end zones while furthering the notion of separate public (i.e., Internet) or shared (i.e., campus/inter-organizational) access from private access. It also includes secured remote IPSec or SSL VPN access. In this case, the term "private" can mean that the virtual data center is routed over the private Enterprise WAN or through the public Cloud provider's IP/NGN via a private MPLS VPN. In the public cloud scenario, this type of virtual data center linked to the tenant Enterprise via an L2 or L3 MPLS VPNs is commonly termed a virtual private data center (VPDC). MPLS VPNs are often used by public Cloud providers as transport for hybrid managed cloud services. As indicated in the left model in Figure 1-8, such services may include IP addressing, security (i.e., firewalling, managed DMZ, zoning, secured remote VPN access), and server resiliency solutions.

In contrast, the second container model represents a "raw" container, so-called because in this case, the tenant provides and manages their own network services and IP addressing within their container, with the public provider offering a seamless extension of the tenant's data center within the public cloud. This is effectively two "Bronze" containers connected via an L2 VPN - i.e., an extended Ethernet. The key benefit of this hybrid tenancy model is that the Enterprise maintains control of their resources within the public cloud; virtual machine migration is controlled by the Enterprise and may be accomplished with minimal modifications to server or virtual machine configurations, as IP readdressing is not required.

Figure 1-8 Two New Tenancy Models



Cloud Services

Another concept at the heart of the VMDC reference architecture is the notion of differentiated service tiering; simply put, tenants may have unique requirements in terms of network throughput, compute processing, storage performance, or data store privacy characteristics, and a successful multi-tenant deployment must be able to address these needs.

Differentiated Services

By definition, in a cloud-based model, compute, storage, and network infrastructure are abstracted and delivered "as a service." To tailor workload characteristics or application performance to specific needs, the cloud administrator has various methods at hand for providing differentiated service tiers and insuring that tenant privacy and service level agreement (SLA) objectives are met:

- **Tiered Workload Definitions—**The secret to building a cloud-ready infrastructure is in categorizing the set of applications that must be supported and distilling these into their basic workload characteristics. Once these are reasonably well-understood, they can in most cases be addressed by a set of standard service profiles. For example, characteristics which apply to the ICS include virtual machine attributes (CPU ratio, memory and associated storage capacity); storage attributes (RAID levels, disk types and speeds, and protection mechanisms); and support for various degrees of application tiering.
- **Availability Mechanisms—**Availability mechanisms may be applied at various layers of the infrastructure to insure that communication requirements are met. For example, within a vSphere cluster, DRS and vMotion or Fault Tolerance may be used to provide optimal resource allocation, even in the event of server failure. Similarly, within the SAN, data protection mechanisms such as snapshots, cloning, and backup archiving help to insure that data store integrity is preserved through various types of failure scenarios. Network services, such as server load balancing, encryption, advanced routing and redundancy, can further help to achieve availability targets. The larger the shared domain (ICS, pod, or entire data center level), the broader the impact of the availability

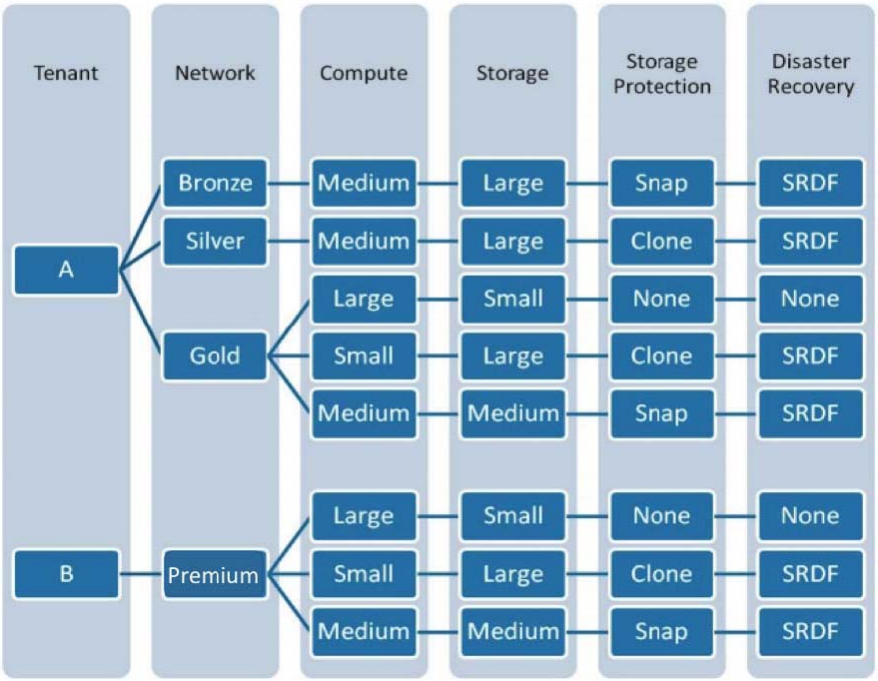
mechanisms utilized at that particular layer of the hierarchy. As these typically do not come without added cost, the goal would be to insure that broadly scoped availability methods meet minimum targeted requirements for the entire tenant community.

- **Secure Isolation—**In a multi-tenant environment, the ability to securely contain and isolate tenant traffic is a fundamental requirement, protecting tenant resources and providing risk mitigation in the event that a specific tenant's privacy is breached. Like availability, isolation mechanisms are applied in a multi-layered fashion in order to implement the requisite infrastructure protection and security zoning policies on a per-tenant basis. In practice, techniques fall into two categories of physical and logical isolation mechanisms. However, VMDC analysis focuses mainly on logical mechanisms. These include various L2 and L3 mechanisms, such as multiple vNICs (i.e., for specific control or data traffic), 802.1q VLANs, MPLS VRFs, VSANs, combined with access control mechanisms (i.e., RBAC and directory services, IPSec or SSL VPNs), and packet filtering and firewall policies.
- **Service Assurance Mechanisms—**Service assurance is a function of availability and QoS policies. The implementation of QoS policies allows for differentiated classification and treatment of traffic flows per tenant per service tier during periods of congestion.
- **Management—**The ability to abstractly represent per-tenant resources and services in the form of a service catalog is a prerequisite for automated service fulfillment and service assurance functions; i.e., the "day 1" and "day 2" management tasks which are so essential to operating under an Infrastructure as a Service (IaaS) model. The service catalog is effectively the highest level of abstraction for the underlying cloud resources. Accurate representations of these resources as policy-based tenancy models to the service catalog rely on interactions directly with domain element managers or middleware management layers via standardized interfaces (i.e., APIs, MIBS, etc.). The more intelligent the middleware layer, the less work has to be done at higher levels in the management framework to understand the tenancy models and commission or decommission resources on a per-tenant basis.

Service Tiering

Previous VMDC releases were modeled based on three baseline categories of tenant network services tiers—Bronze, Silver, and Gold— represented in terms of firewalling, server load balancing, SSL offload, and QoS policy (i.e., hree data classes of service), combined with three workload models, each with specific compute attributes, associated storage characteristics, and business continuance services. Figure 1-9 shows a high-level conceptual illustration of these models, demonstrating a variety of ways in which these resources and services can be applied in combination to meet business or application requirements in a tiered fashion.

Figure 1-9 VMDC Service Tiers

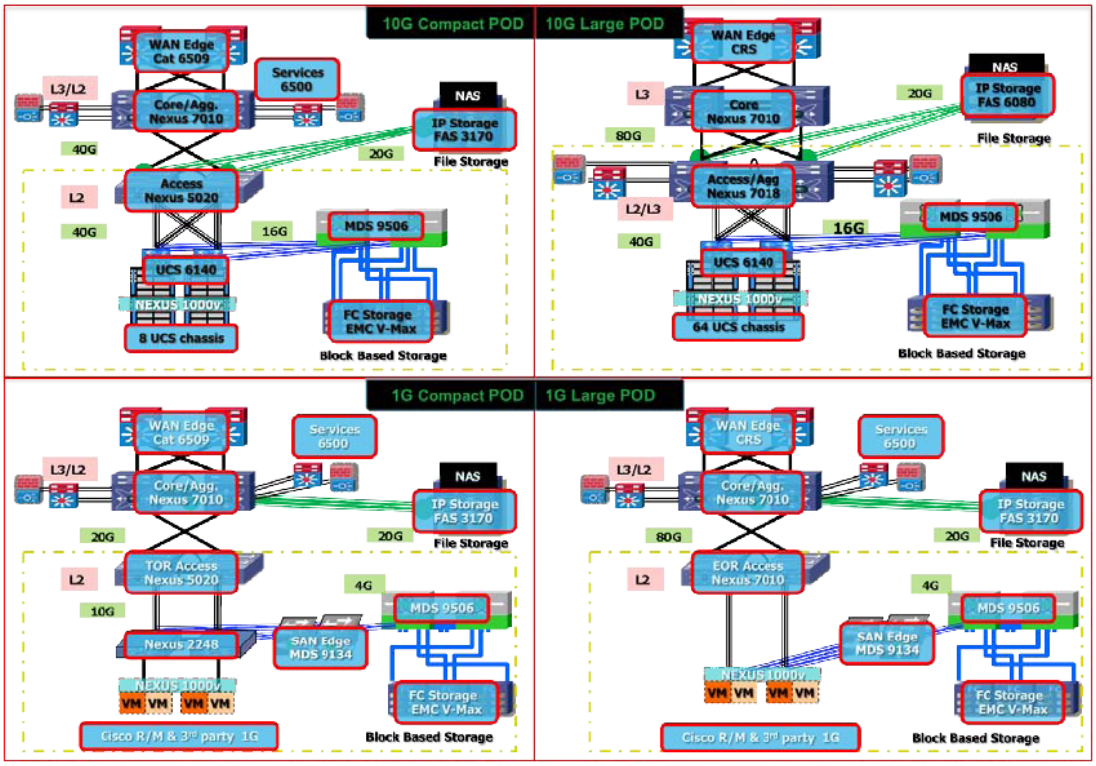


Solution Architecture

This system release leverages the end-to-end architecture defined in VMDC 2.0. This document revisits foundational principals of high availability and modular growth, and describes enhancements to the system in the areas of tenant isolation and security in general, and introduces a new QoS framework for accommodation of multimedia and collaboration applications.

As a brief recap, in VMDC 2.0, we defined two system topologies mainly focused on 10 Gb compute farms, each featuring a variant for 1 Gb compute farms (Figure 1-10). In the first, UCS systems with B-series blade servers provided the compute resource. In the second, C-series rackmount servers were the focus, though it is possible to add other third-party servers. These can be used in combination to meet various storage and server use cases. Note for example, the use of the 2248 Fabric Extender to expand the Nexus 1G port capacity (lower left quadrant Figure 1-10). In this release, we focus on enhancements to the "10G Large Pod" variant displayed in the upper right quadrant in Figure 1-10.

Figure 1-10 VMDC 2.X Topology Options

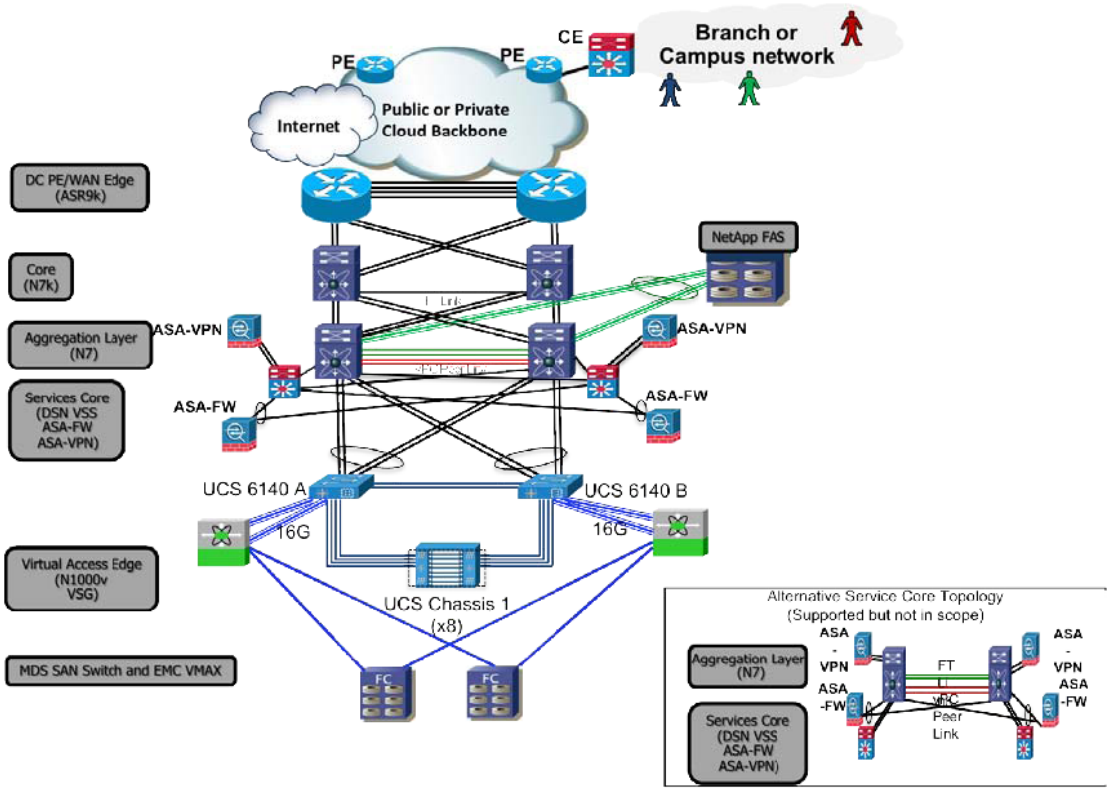


End-End Topologies

Physical Topology

The end-to-end physical topology characterized in this release is depicted in Figure 1-11.

Figure 1-11 Physical Topology View

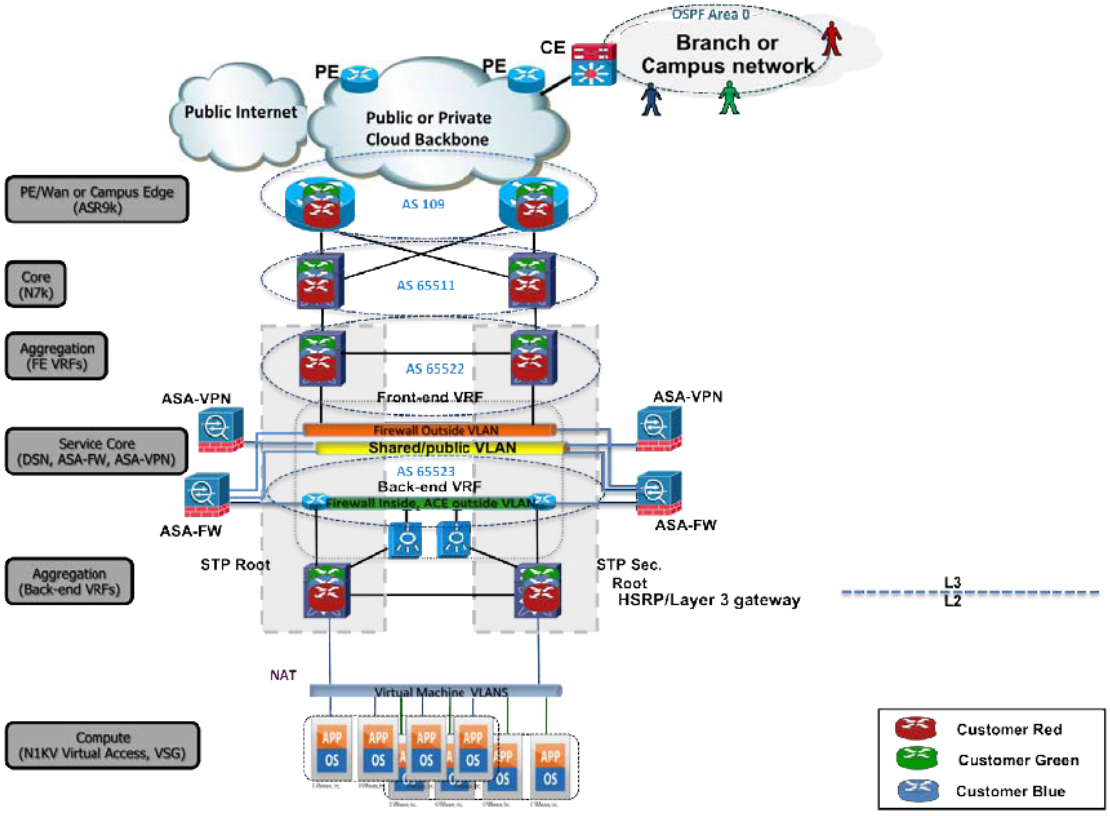


Logical Topology

Figure 1-12 provides a high-level view of the corresponding logical topology.

Note This architecture provides flexibility for mixing and matching services to specific groups of tenants (this diagram does not depict all service flows), and not all tenants will require server load balancing and hybrid firewalling.

Figure 1-12 Logical Topology Overview

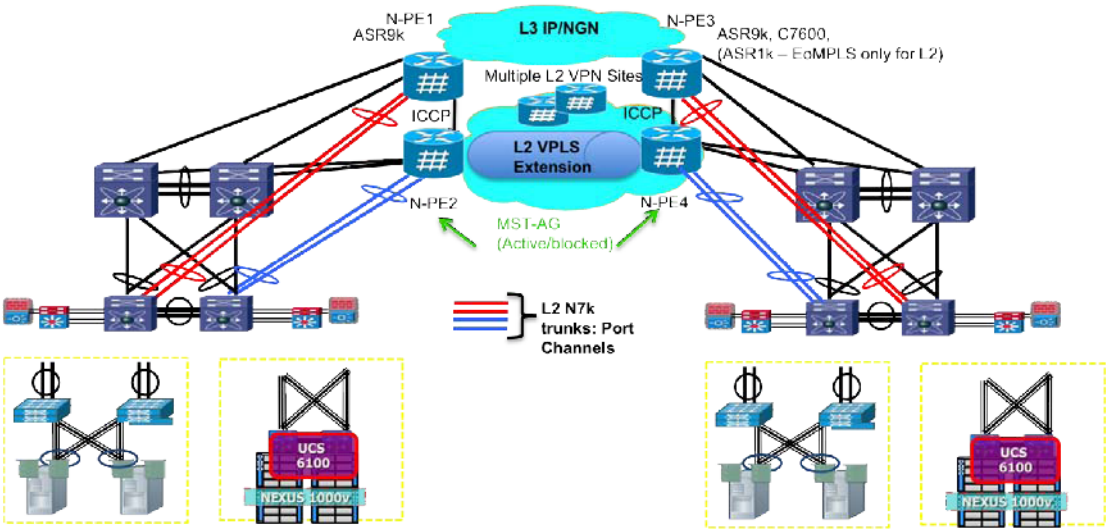


L2 Multi-site VPN Topology

Also included in this release is an analysis of hybrid cloud interconnect use cases and associated L2 transport scenarios. The ASR 9000 DC Edge devices were in this instance serving as multiservice nodes for L2VPN and L3VPN transport between data centers. The diagram is included here for completeness, however, design and implementation guidance are provided in the following two documents:

- VMDC 2.2 VPLS Implementation Guide
- VMDC 2.2 EoMPLS DCI for Hybrid Cloud with vCloud Director Design and Implementation Guide Supplement

Figure 1-13 L2VPN Topology



Solution Components

Figure 1-14 VMDC 2.2 Component Matrix

Function	Components
Network	Cisco ASR9k, C7600, ASR1k
	Cisco Nexus 7010, 7018 -(M1 series 10Gb Ethernet cards)
	Data Center Services Node 6509-E (VSS)
	Application Control Engine (ACE-30) Server Load Balancer
Security Services	Cisco ASA Appliance (5585)
	Cisco Virtual Security Gateway
	Cisco Nexus 1000v
	Cisco MDS soft zoning and VSANs
Compute	Cisco Unified Computing System (UCS) Cisco UCS 6140 Fabric Interconnect UCS5108 Blade Server Chassis UCSB200-M1 Blade Server UCS M71KR-E Converged Network adapter UCS M81KR Virtual Interface card UCS C200 M1 servers (Management Pod)
Virtualization	VMware vSphere VMware ESXi 4.1U1 Hypervisor Cisco Nexus 1000v (virtual access switch)
Storage Fabric	Cisco MDS 9513
Storage Array	EMC 2 Symmetrix V-Max with virtual provisioning NetApp FAS3240 (Management Pod)
Orchestration/ Management*	Domain Management: – UCS Manager – Nexus 1000v Virtual Supervisor Module – Cisco Virtual Network Management Center – VMware vCenter 4.1U1 – VMware VCloud Director – Fabric Manager Orchestration: *BMC CLM 2.1 SP1 not in scope

CHAPTER 2

Design Details

The Virtualized Multi-tenant Data Center (VMDC) 2.2 release continues the design approach outlined in VMDC 2.0, with focus on the following key areas:

Secure Tenant Separation

Traditionally, IT admininistrators deployed dedicated infrastructure for their tenants. Deploying multiple tenants in a shared, common infrastructure optimizes resource utilization at lower cost, but requires designs that address secure tenant separation to insure end-to-end path isolation and meet tenant security requirements. The following design considerations provide secure tenant separation and path isolation:

Network Separation

In order to address the need to support multi-tenancy while providing the same degree of tenant isolation as a dedicated infrastructure, the VMDC reference architecture uses path isolation techniques to logically divide a shared infrastructure into multiple (per-tenant) virtual networks. These rely on both data path and device virtualization, implemented in end-to-end fashion across the multiple hierarchical layers of the infrastructure and include:

- **Network Layer 3 (L3) Separation (core/aggregation layers)**—VRF-lite implemented at core and aggregation layers provides per tenant isolation at L3, with separate dedicated per-tenant routing and forwarding tables insuring that no inter-tenant (server to server) traffic within the data center

will be allowed, unless explicitly configured. A side benefit of separated routing and forwarding instances is the support for overlapping IP addresses; a required feature in the public cloud case or in merger or other situations involving IP addressing transitions in the private Enterprise case.

- **Network Layer 2 (L2) Separation (access, virtual access)**—VLAN IDs and the 802.1q tag provide isolation and identification of tenant traffic across the L2 domain, and more generally, across shared links throughout the infrastructure.
- **Network Services sSparation (services core, compute)**—On physical appliance or service module form factors, dedicated contexts or zones provide the means for virtualized security, load balancing, NAT, and SSL offload services and the application of unique per-tenant policies at the VLAN level of granularity. Similarly, dedicated virtual appliances (i.e., in vApp form) provide for unique per-tenant services within the compute layer of the infrastructure at the virtual machine level of granularity.

Compute Separation

Traditionally, security policies were implemented at the physical server level. However, server virtualization and mobility introduces new security challenges and concerns; in effect, in order to meet these challenges, policy must be implemented at the virtual machine level and be capable of following virtual machines as they move from host to host.

Separation of per-tenant traffic in the compute layer of the infrastructure leverages the following technologies:

- **vNICs—**In the highly virtualized data center, separation of traffic is accomplished via use of multiple vNICs, rather than physical NICs. For example, in VMDC 2.X, multiple vNICs are used to logically separate production (data) traffic from back-end management traffic. This is accomplished with the Cisco UCS Virtual Interface Card (i.e., M81KR VIC in this case), which allows for the creation of virtual adapters and their mapping to unique virtual machines and VMkernal interfaces within the hypervisor.
- **VLANs—**VLANs provide logical isolation across the L2 domain, including the Nexus 1000V virtual access switching domain within the compute tier of the infrastructure.
- **Port profiles—**When combined with Cisco's VN-link technology, port profiles provide a means of applying tenant traffic isolation and security policy at the VLAN and virtual machine (vNIC) level of granularity. Implemented at the virtual access switching domain, these map to vCenter port-groups and thus provide policy mobility through VMotion events.

Storage Separation

In the VMDC reference architecture, separation of virtual machine data stores within the storage domain of the shared infrastructure is accomplished in the following ways:

- **Cluster File System Management—**The vSphere hypervisor's cluster file system management creates a unique Virtual Machine Disk (VMDK) per VM, insuring that multiple VMs cannot access the same VMDK sub-directory within the Virtual Machine File System (VMFS) volume and thus isolating one tenant's VMDK from another.
- **VSANs and FC Zoning—**Segmentation of the shared SAN fabric into smaller logical domains via VSANs and FC zoning provides isolation at the physical host level of granularity.

- **LUN Masking—** Logical Unit Number (LUN) masking creates an authorization process that restricts storage LUN access to specific hosts on the shared SAN. This, combined with VSANs implemented on the Cisco MDS SAN switching systems plus FC zoning, effectively extends tenant data store separation from the SAN switch ports to the physical disks and virtual media within the storage array.
- **vFilers**—Supported on NetApp NAS systems, vFilers provide logical separation of NFS data stores. These may be correlated with IP addresses (IPspaces) and used in combination with 2.1.4 Application Tier Separation 802.1q VLANs and ACL-based security policy enforcement to limit NFS data store access to specific tenants or groups of tenants across the shared infrastructure.

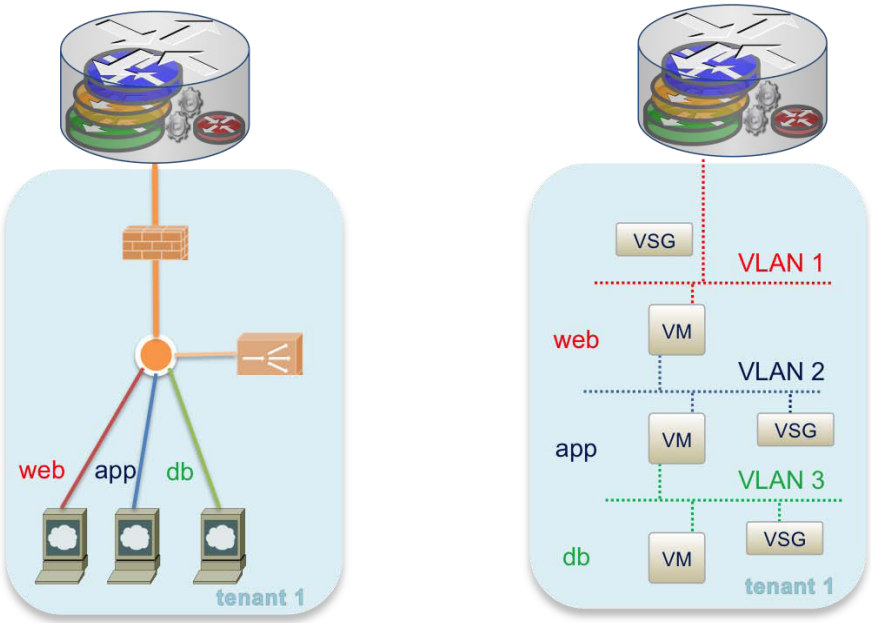
Application Tier Separation

Many applications follow a three-tiered functional model, consisting of web, application, and database tiers. Servers in the web tier provide the public facing, "front-end" presentation services for the application, while servers in the application and database tiers function as the middleware and back-end processing components. Due to this functional split, servers in the web tier are typically considered to be likely targets of malicious attacks, with the level of vulnerability increasing in proportion to the scope of the user community. Applications meant to be accessible over the public Internet rather than simply remain in the Enterprise private cloud or the Enterprise's VPDC in the public cloud would represent the broadest scope and thus a major security concern.

Several methods exist for separation of application tiers:

1. **Network-Centric Method**—This method involves the use of VLANs within the L2 domain to logically separate each tier of servers (left in Figure 2-1).
2. **Server-Centric Method—**This method relies on the use of separate VM vNICs to daisy-chain server tiers together (right in Figure 2-1).

Figure 2-1 VLAN and vNIC Application Tier Separation

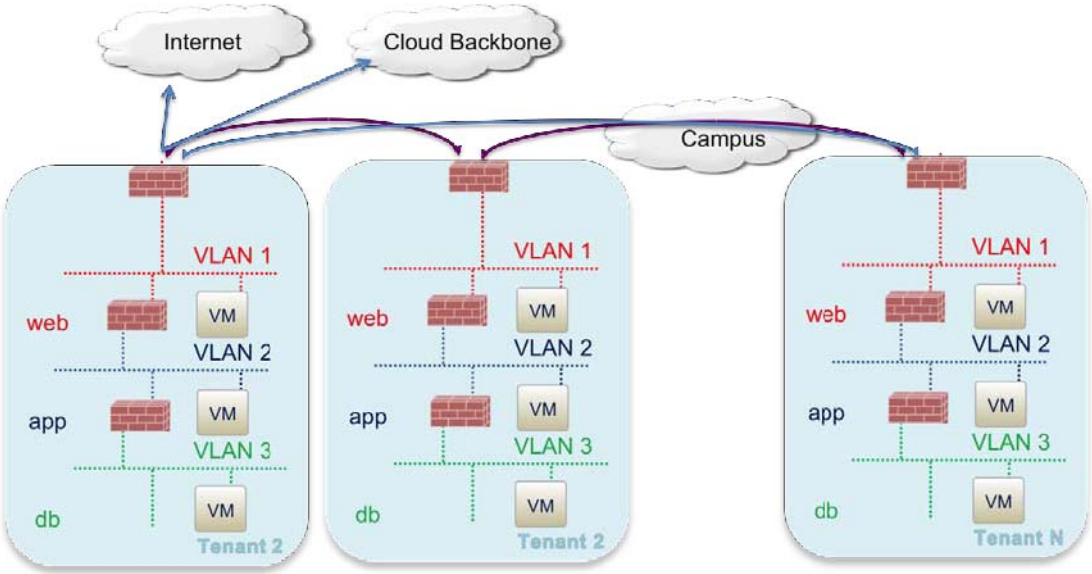


Each method has its pros and cons; which is more desirable will depend on specific deployment characteristics and operational concerns. From an architectural perspective, network service application will be a major factor; the server-centric method naturally lends itself to vApp-based virtualized service insertion, in Cisco's case, leveraging the Nexus 1000V vPath strengths to classify and more optimally redirect traffic flows at the virtual access switching level of the infrastructure. The network-centric method lends itself to designs in which some or all services are applied from Virtual outside the compute tier of the infrastructure, in a services core layer of the hierarchy, with routing of inter-VLAN flows. From an administrative perspective, IT executives must consider expertise across the network and server operations staff together with the available management solutions required to support centralized or highly distributed tenant segmentation or service application models.

The network-centric method is the traditional approach; not all services that one might wish to apply today are available in vApp form, so the current trend is a migration from the network-centric model to hybrid service application scenarios, with some services applied more centrally from the services core and some applied from within the compute layer of the infrastructure. This is particularly true with respect to security services, where from an operational process and business policy enforcement perspective, it may be necessary to hierarchically deploy policy enforcement points, centralizing and more tightly controlling some while distributing others. This trend is the rationale driving consideration of the hybrid approach to security policy enforcement.

In consideration of application separation, it is common for IT administrators to begin by rigorously separating each tier, assuming that minimal communication between servers on each tier is required. This may sometimes translate to a practice of enforcing separation at each tier with firewalls (see Figure 2-2).

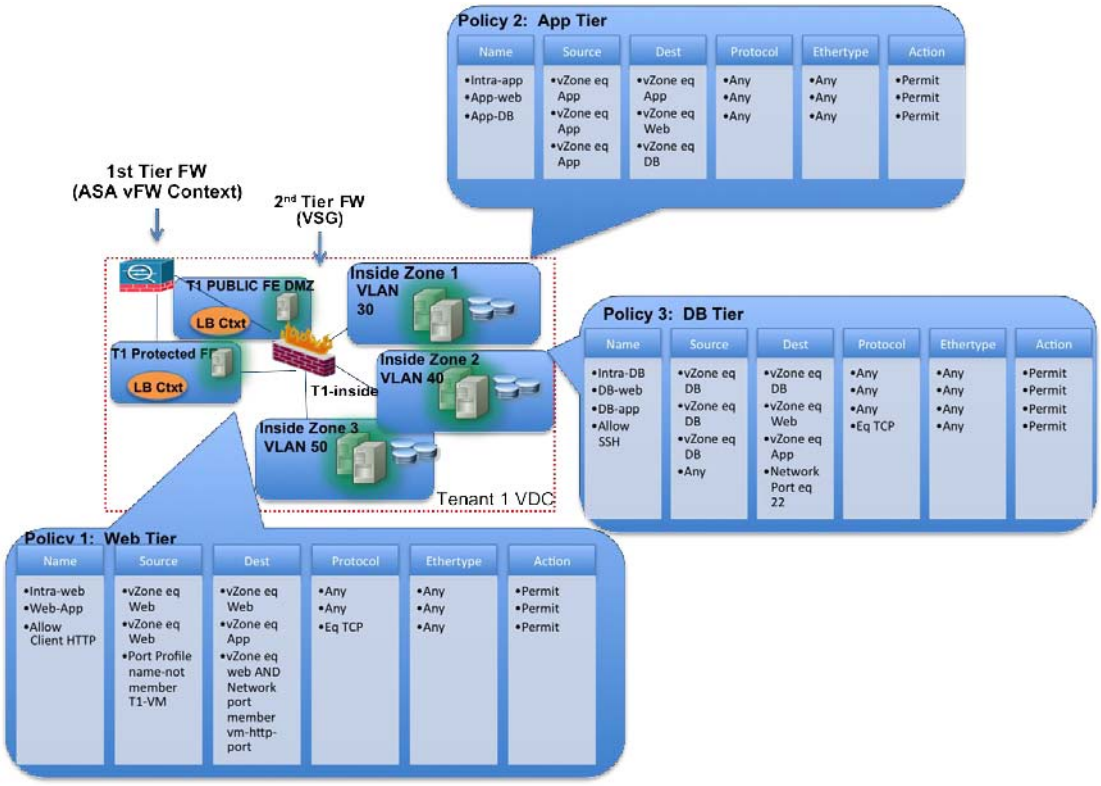
Figure 2-2 Three-Tier Firewall



While this approach seems reasonable in theory, in practice one soon discovers that it is too simplistic. One problem is that applications are complex; applications don't necessarily follow a strict hierarchical traffic flow pattern. Some applications may for example be written to function in a database-centric fashion, with communications flows to the middleware (app) and perhaps presentation (web) tiers from a database core, while others may be written to leverage the middleware layer. Another problem, particularly common for Enterprise scenarios, is that some application flows may need to extend outside of the private cloud tenant or workgroup container, across organizational boundaries and perhaps from

site to site. Finally, application tiers may themselves be distributed, either logically or physically, across the data center or in the private case, across the Enterprise campus. The result is unnecessary and sub-optimal proliferation of policy enforcement points - in which traffic may needlessly be required to traverse multiples of firewalls on the path end-to-end from source to destination. With a hybrid two-tiered firewall model (Figure 2-3), the VMDC architecture seeks to provide a simplified framework that mitigates firewall proliferation over the physical and virtualized infrastructure while allowing for defense-in-depth, as per traditional security best practices. As noted earlier, a benefit of this framework is that it enables hierarchical policy definition, with rigorous, fine-grained enforcement at the outer edges of the tenant container and more permissive, coarse-grained enforcement within the tenant container. This framework also provides a graceful transition from physical to virtual policy enforcement, allowing cloud administrators to leverage existing inventory and expertise.

Figure 2-3 VMDC Two-Tier Hybrid Tenant Firewall Model



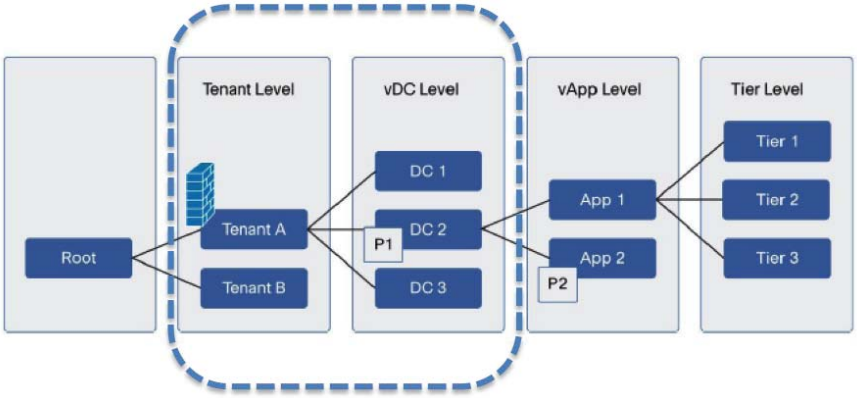
Virtual Security Gateway

The Cisco Virtual Security Gateway (VSG) is a new addition to the VMDC reference architecture. In the VMDC architecture, inter-tenant communication (if allowed) is established through routing at the aggregation layer. However, in Figure 2-3, we see how the VSG virtual security appliance fulfills the functional role of an intra-tenant second tier firewall to filter communication between and within application tiers and from client to server. Tightly integrated with the Cisco Nexus 1000V distributed virtual switch, the Cisco VSG uses the virtual network service path (vPath) technology embedded within the Cisco Nexus 1000V Virtual Ethernet Module (VEM). The vPath capability within the Cisco Nexus 1000V offloads the switching logic directly to the host, providing high performance, seamless interaction with other virtual appliances, and resiliency in case of appliance failure. There is a significant

performance improvement, since most of the packets are offloaded to the hypervisor and processed by the fast path. In addition, the Cisco Nexus 1000V vPath is tenant-aware, which allows for the implementation of security policies within and across multiple tenants.

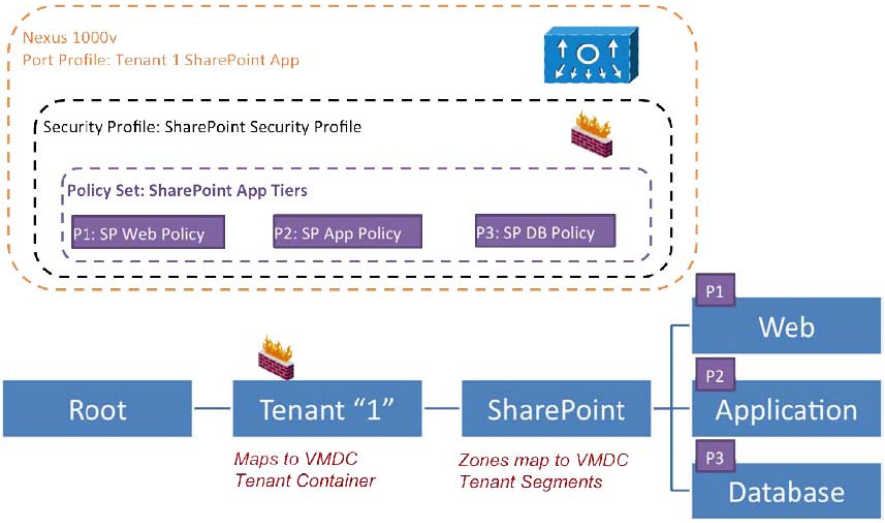
The VSG multi-tenant support relies on a hierarchical policy model (Figure 2-4). This model allows each tenant to be divided into three different sub-levels, which are commonly referred to as vDC, vApp, and tier levels. Security rules and policy definitions can be set at any point in the hierarchy. These rules apply to all VMs that reside at, or below, the enforcement point (tenant level in Figure 2-4). Root-level policies and pools are system-wide and available to all organizations. In a multi-tenant system such as VMDC, to provide proper tenant separation and policy control, a unique instance of VSG must be deployed for each tenant.

Figure 2-4 VSG Hierarchical Policy Model



The VSG hierarchical policy classification is available to be leveraged for more complex policy rulesets, however it is not mandatory to use all policy levels. For example, in the VMDC system reference model, though the VSG policy model allows for sub-tenancy, we commonly envision a tenant container as a single virtual data center with a requirement to support multiple categories of applications, each with multiple application tiers. Figure 2-5 shows this mapping, using the example of a specific application category (i.e., SharePoint). Implementers should follow a practical, "keep it simple" approach that meets their security policy profile requirements without unnecessary complexity.

Figure 2-5 VSG Policy Profile Hierarchy Mapped to VMDC Tenancy



VSG access controls can be applied to network traffic between packet source and destination based on TCP/UDP ports, VM, or even custom attributes, making policy definition much more context-aware than simple legacy stateful packet filtering firewalls. In terms of application separation in the dynamic environment of a cloud-based infrastructure, a key benefit of the VSG is that by moving policy enforcement to the Nexus 1000V DVS, policy zones will automatically follow a VM as it moves from one hypervisor to another within the logical DVS boundary.

As of this writing, Nexus 1000V Release 1.4(a) supports the following policy attributes for source/destination filtering:

- src.net.ip-address
- src.net.port
- dst.net.ip-address
- dst.net.port
- net.ip-address
- net.port net.protocol
- net.ethertype
- src.vm.name
- dst.vm.name
- vm.name
- src.vm.host-name
- dst.vm.host-name
- vm.host-name
- src.vm.os-fullname
- dst.vm.os-fullname
- vm.os-fullname
- src.vm.vapp-name

- dst.vm.vapp-name
- vm.vapp-name
- src.vm.cluster-name
- dst.vm.cluster-name
- vm.cluster.name
- src.vm.inventory-path
- dst.vm.inventory-path
- vm.inventory-path
- src.vm.portprofile-name
- dst.vm.portprofile-name
- vm.portprofile-name
- src.vm.custom.xxx
- dst.vm.custom.xxx
- vm.custom.xxx

Perimeter Security

In traditional security models, it has long been a best practice to apply policy enforcement at defined boundaries between trusted and untrusted user communities or zones. A security zone comprises a logical construct of compute, network, and storage resources which share common policy attributes. One can leverage the common attributes within this construct to create security policies that apply to all the resources within that zone. However, in a highly virtualized system, it may be difficult to determine where these perimeters lie, particularly for the multi-tenant use case. In this system release, we define three perimeters essential for maintaining Enterprise-grade tenant security in a public or private cloud infrastructure:

1. **Front-End Tenant Perimeter**—This is the perimeter between less trusted zones and the interior of the tenant virtual data center within the cloud.
2. **(Intra-VDC) Back-End Tenant Perimeter**—This is the perimeter between a tenant's front-end servers and back-end servers.
3. **Back-End Management Perimeter—**This is the perimeter between the tenant "production" servers and back-end infrastructure management servers.

Between these perimeters, we have the following zones defined:

1. **Public/Shared—**This zone provides a means of entry to the tenant virtual data center from a broader scope of external clients, sourced from either the public Internet, the Enterprise campus, or remote access VPNs (not shown below). This is an untrusted or less trusted zone (versus those within the tenant virtual data center). Note that this zone would also potentially hold a general/shared infrastructure demilitarized zone (DMZ).
2. **Private**—The private zone provides a means of entry to the tenant virtual data center via the cloud backbone; i.e., either the private WAN backbone or the public provider IP/NGN. In the latter case, the expectation is that clients will typically be utilizing a private L2 or L3 MPLS VPN across the public IP/NGN for access.

3. **Tenant DMZ**—This zone provides for a per-tenant DMZ (i.e., versus a more generalized DMZ elsewhere in the Enterprise or public provider infrastructure). It is understood that not all tenant virtual data centers will feature a DMZ zone.
4. **Tenant Front-End (web)**—This provides for a general front-end server zone, suitable for the placement of front-end application presentation servers.
5. **Tenant Back-End**—Minimally, this would include two zones for app and database servers, but could be additional as required to accommodate multiple types of applications and additional application or policy-specific objectives.
6. **Back-End Management—**This zone contains back-end servers used to manage the infrastructure. These could be virtual or bare-metal servers, depending upon the requirements of the management stack solution.

Figure 2-6 and Figure 2-7 shows how this model logically overlays onto the shared virtual and physical infrastructure.

Figure 2-6 Tenant Perimeters and Zones

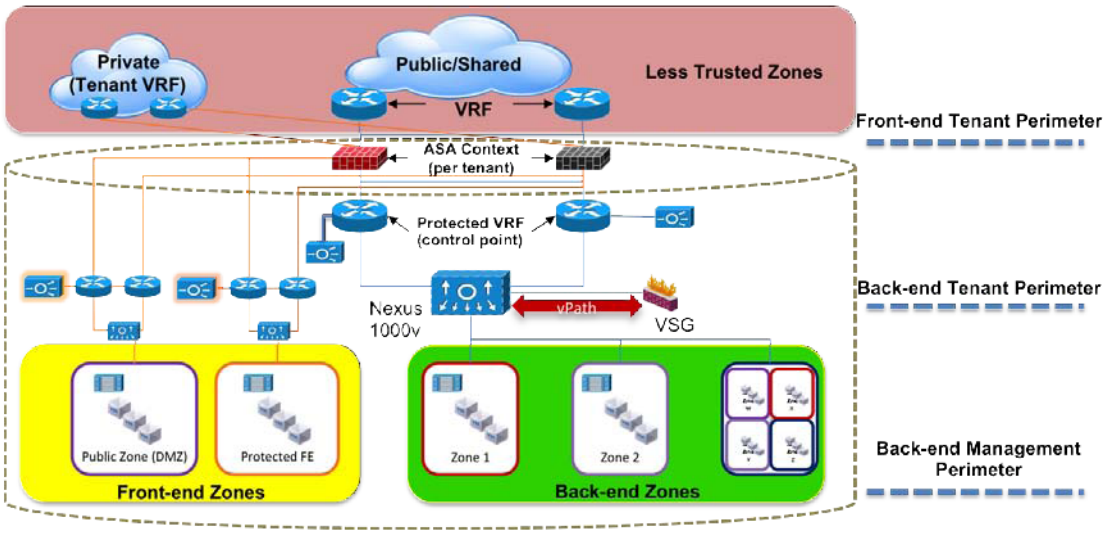
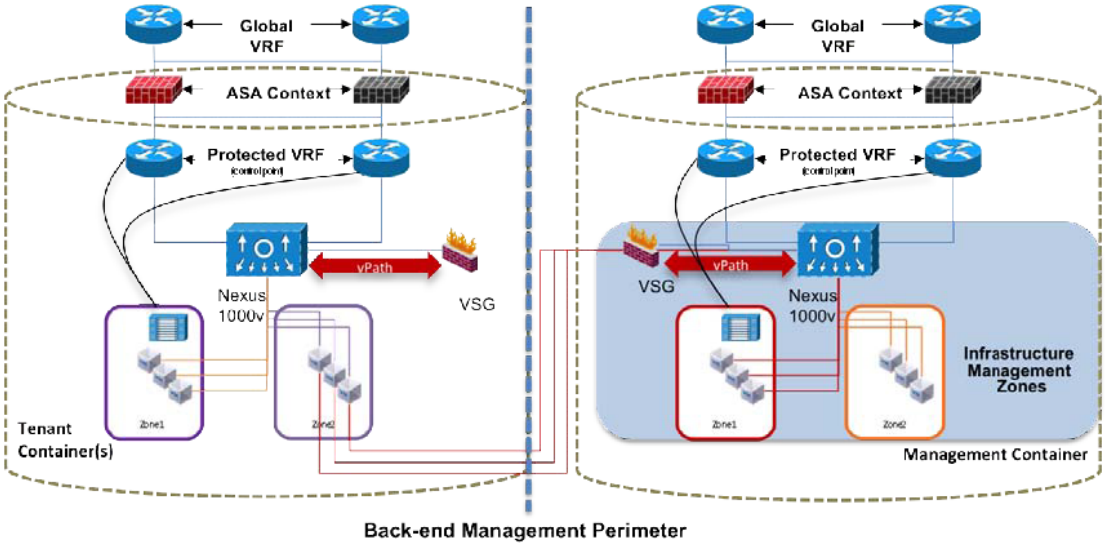


Figure 2-7 Infrastructure Management Zones



In Figure 2-7, a separate set of management vNICs allow tenant VMs to be "dual-homed," with port profiles present on "production" and back-end infrastructure management Nexus 1000V instances. Multiple VSGs may be used in the management container to scale policy enforcement.

This framework provides the flexibility to accommodate a variety of options including:

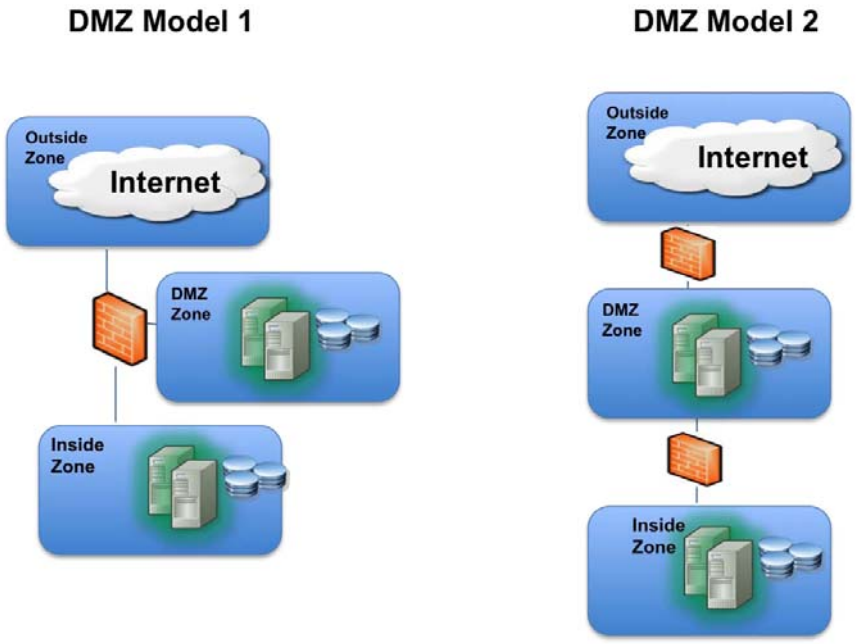
- A provider (infrastructure) DMZ (not shown).
- Additional untrusted zones and nested zones:& rather than a single shared public zone for remote VPN and Internet or campus access, the untrusted zones could be further segmented. Sample use cases applicable to the public provider context would be to provide separate zones for Independent Software Vendor (ISV) access or dedicated per-tenant public access zones.
- Nested front or back-end zones: for example, there could be two nested zones, with different policy rulesets within a single front-end tenant zone, for DMZ servers and more general application presentation servers. Similarly, nested back-end zones could facilitate separation of "production" from "dev-test" back-end servers.
- Accommodation of traditional security best practices: for example, role-based infrastructure or server/VM access control (RBAC), tied to LDAP or radius directories. Though not the focus of this system release, RBAC is a fundamental security requirement. A prerequisite is definition of role categories, to which differing access policies can be applied, i.e., tenant-user, tenant-administrator, administrator-user, etc.

DMZ Zones

A demilitarized zone (DMZ) is a small network inserted as a "neutral zone" between a private "inside" network and the outside public network. The DMZ's role is to prevent outside users from getting direct access to a server that has private data. Often, servers placed within the DMZ enhance perimeter firewall security by proxying requests from users within the private network for access to Web sites or other companies accessible on the public network. The proxy server then initiates sessions for these requests on the public network. However, it is not able to initiate a session back into the private network. It can only forward packets that have already been requested. How would a DMZ be inserted into a tenant virtual data center in the cloud? Two basic models exist for placement of a DMZ. In Model 1 of

Figure 2-8, the DMZ zone is connected to the same network device as the Inside and Outside Zones; in Model 2 of Figure 2-8, the DMZ is in a transit zone between a front-end and back-end firewall. Traditionally, Model 2 is considered to be slightly more secure, the logic being that two firewalls are better than one; this is a defense-in-depth measure, the premise being that if the front-end outside firewall is mis-configured, there is still a measure of security provided by the second firewall. It is this second placement option that the VMDC 2.2 release incorporates into the expanded virtual data center/ VPDC tenancy model.

Figure 2-8 DMZ Placement Options



Note that though this system focuses on the application of a DMZ within the tenant virtual data center, typically there would also be a DMZ on the shared portion of the infrastructure.

High Availability

A highly available infrastructure is the foundation for successful cloud-based services deployment and in particular, for service assurance or SLA guarantees. The VMDC reference architecture is thus modeled for the highest possibility infrastructure availability, to insure no single point of failure. However, resiliency comes at incremental cost and complexity. The ongoing goal of this effort is to model and validate resiliency mechanisms in a multi-dimensional fashion, so that architects and implementers may make informed decisions about which solutions provide the optimal approach for their particular set of business service objectives and technical criteria.

This section presents the following topics:

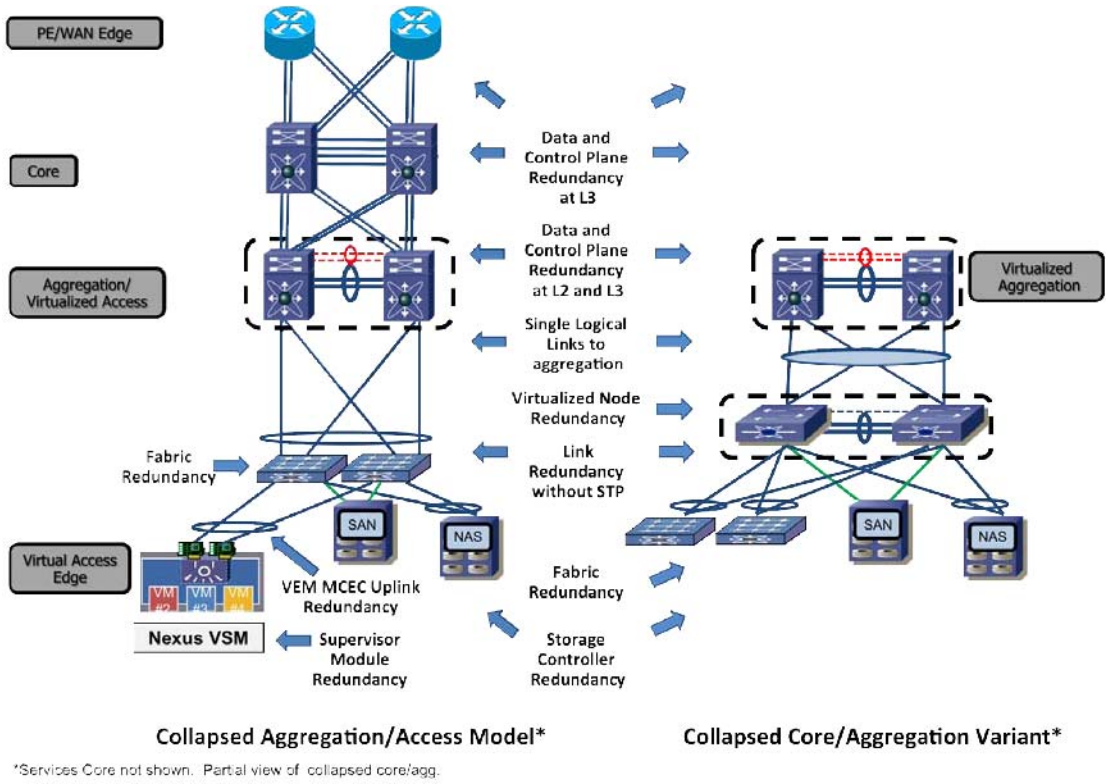
Redundant Network Design

As discussed in depth in VMDC 2.0 and further emphasized in VMDC 2.1, the reference architecture employs a multi-layered approach to infrastructure high availability design.

Figure 2-9 shows how resilience mechanisms are utilized at every level of the infrastructure. These include:

- Redundant links, nodes and paths, end-to-end
- Core layer: redundant L3 paths, links and nodes; redundant supervisors
- Services core (not shown): redundant nodes, redundant data and control plane, redundant supervisors, links and paths
- Aggregation layer: redundant default gateway (Nexus 7000 aggregation nodes); redundant supervisors; redundant links and L3 paths
- Access layer: redundant nodes, supervisors and links
- Compute layer: UCS - redundant fabric and control plane; intra-cluster HA
- Virtual Access: redundant forwarding path (CNA)
- Storage: redundant SAN switching systems (not shown); redundant controllers; RAID
- Management Servers (not shown): intra-cluster HA; clustering or mirroring between management servers; vCenter Server heartbeats; snapshots and cloning

Figure 2-9 Tiered HA Models



L2 Redundancy

The VMDC reference architecture utilizes several key L2 redundancy mechanisms at various points in the infrastructure to provide optimal multipathing.& These are virtual port-channels (vPCs), Multi-Chassis EtherChannel (MEC), and MAC-pinning.

Virtual Port-channels

A Cisco innovation based on port-channel technology (IEEE 802.3ad), virtual Port-Channels (vPCs) allow multiple links to be used between a portchannel-attached device, and a pair of participating switches. The two switches act as vPC peer endpoints, and look like a single logical entity to the device. Traffic is forwarded and load balanced across all the links, but because they are bundled as one logical path, there is no loop created and so there is no requirement for Spanning Tree loop avoidance. With multiple active links comprising the path, vPCs typically provide faster link-failure recovery versus SPT processes, which involve relearning the L2 topology. Combining the benefits of load balancing with hardware node redundancy and port-channel loop management, vPCs offer optimal link bandwidth utilization. For these reasons, vPCs are recommended and leveraged whenever possible within the reference architecture. Specifically, in this release as in previous iterations, vPCs are deployed below the L3/L2 boundary, between the Nexus 7000 aggregation layer, and the Nexus 5000 access nodes or UCS 6100 Fabric I/O modules.

Once again, as in previous releases, we recommend that Spanning Tree Protocol (STP) be enabled over the L2 portion of the infrastructure (i.e., below the aggregation layer) for loop avoidance in the event of mis-configuration.

Multi-Chassis EtherChannel

Another Cisco innovation based on port-channel technology, Multi-Chassis EtherChannel (MEC) is a port-channel that spans the two chassis of a switch; in this case, the Cisco Data Center Service Nodes in the services core of the infrastructure. The portchannel-attached device views the MEC as a standard port channel. Similar to vPCs, MEC allows for optimal link bandwidth utilization across multiple links and redundant hardware nodes. MEC provides resilient routed paths between the Nexus 7000 nodes in the aggregation layer of the infrastructure and the Data Center Service Nodes in the service core layer.

MAC-Pinning

Virtual machine NICs may be pinned statically or dynamically to uplink paths within the UCS. In the reference architecture, MAC-pinning is used in conjunction with the Nexus 1000V to provide more granular load-balancing and redundancy across the system. It does this through the use of notification packets, which in the event of a link failure, inform upstream switches of the new path required to reach destination 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 that the data center access layer network can learn the new path.

L3 Redundancy

HSRP

Hot Standby Router Protocol (HSRP) is a first hop redundancy protocol, enabling the creation of redundant default gateways. HSRP allows two or more routers to act as a single "virtual" router, sharing an IP address and a MAC (L2) address. The members of the virtual router group continually exchange status messages, allowing one router to assume the routing responsibility of another, should it go out of commission for either planned or unplanned reasons. Failover to a standby router in the virtual router group will be transparent to hosts, as they will continue to forward IP packets to the same IP and MAC address. HSRP has been enhanced to gracefully interoperate with vPCs in a quasi "active/ active" state, such that a packet forwarded to the virtual router MAC address is accepted as local by the active and standby HSRP peers, however responses will only be sent from the active HSRP peer. In order to provide default gateway redundancy, HSRP is deployed on the Nexus 7000 nodes within the aggregation layer of the infrastructure - i.e., for all VLANs having their L3 termination on the SVI interfaces of the Nexus 7000 aggregation switches.

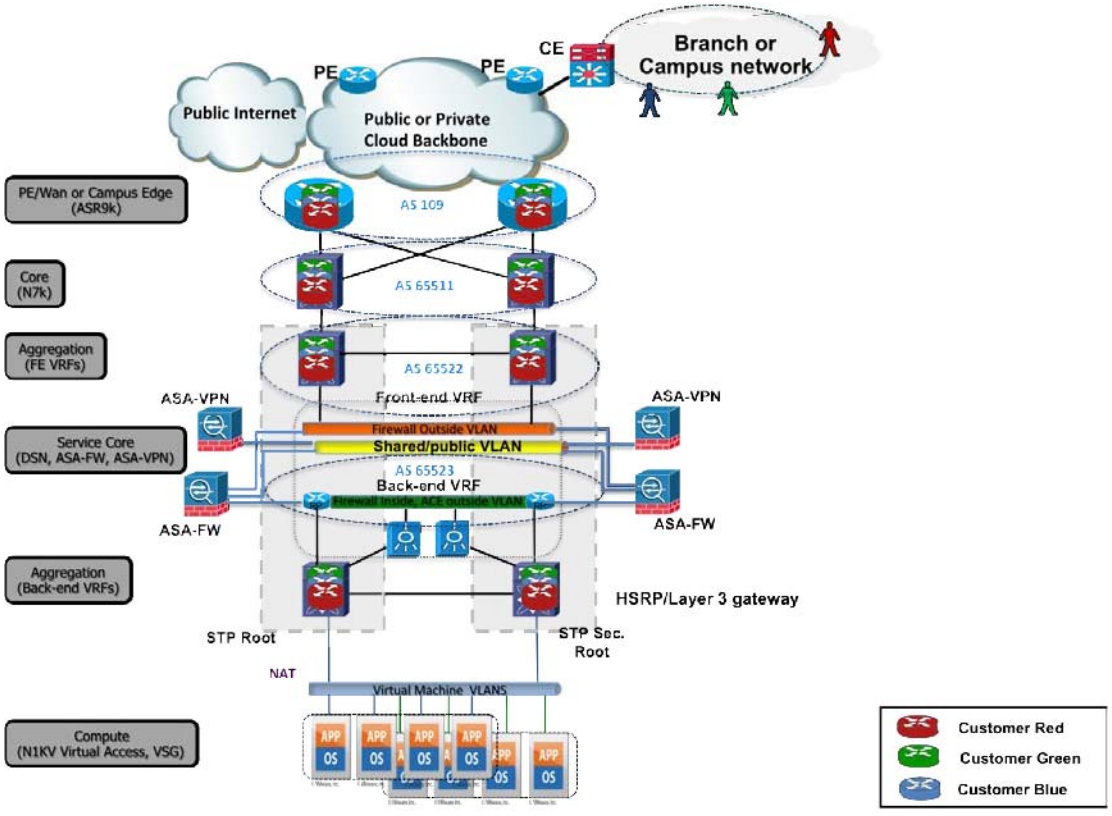
BGP

A L3 IP routing protocol is required in the aggregation, core, and edge layers of the VMDC model. Through various releases, the VMDC solution has been validated with both OSPF and BGP protocols. In this release, OSPF is used as the Interior Gateway Protocol (IGP) between the redundant data center edge routers. As shown in Figure 2-10, BGP is used to establish and maintain IP connectivity within the L3 portions of the infrastructure. In this scenario, eBGP advertises routes between each defined autonomous system, from the services core nodes up to the data center edge nodes, re-routing over redundant L3 paths in the event of a node or link path failure. The use of loopback interface addressing is common in IGPs, including iBGP and for OSPF, insuring that TCP sessions for routed paths are maintained in the event of link failures, while traffic is restored across active links. Loopback interfaces do not apply for eBGP scenarios, where peer interfaces are directly connected, however in the event that peering over interfaces that are not directly connected is required, they can be utilized with additional configuration. More common for this scenario is the use of eBGP multi-hop, which must be used in any case in conjunction with an IGP or static route when the external peering interfaces are not directly connected.

By default, BGP selects one best path if there are several external equal-cost paths available from an AS. In the VMDC 2.2 solution, this would result in utilization of only half of the available infrastructure bandwidth during normal conditions. In order to get the most out of the available bandwidth, traffic is load balanced along the redundant paths. For parallel paths between two eBGP peers, loopback interfaces may be used in conjunction with eBGP multi-hop (and an IGP or static routes to communicate eBGP peer reachability) to load balance traffic. In the case of the VMDC solution, community strings are used to identify and load balance traffic across redundant eBGP paths between the edge and core data center routers.

Additional optimizations for L3 resiliency leveraged in the system include: Cisco Nonstop Forwarding (NSF), Nonstop Routing (NSR), LDP sync, and MPLS graceful restart. More generally, tuning for fast L3 convergence may include the use of BGP graceful restart, BFD, tuning of hello and hold timers, and route summarization.

Figure 2-10 End-to-End Logical Topology



Compute Redundancy

To enable redundancy within the compute layer of the infrastructure, the following features are leveraged and recommended:

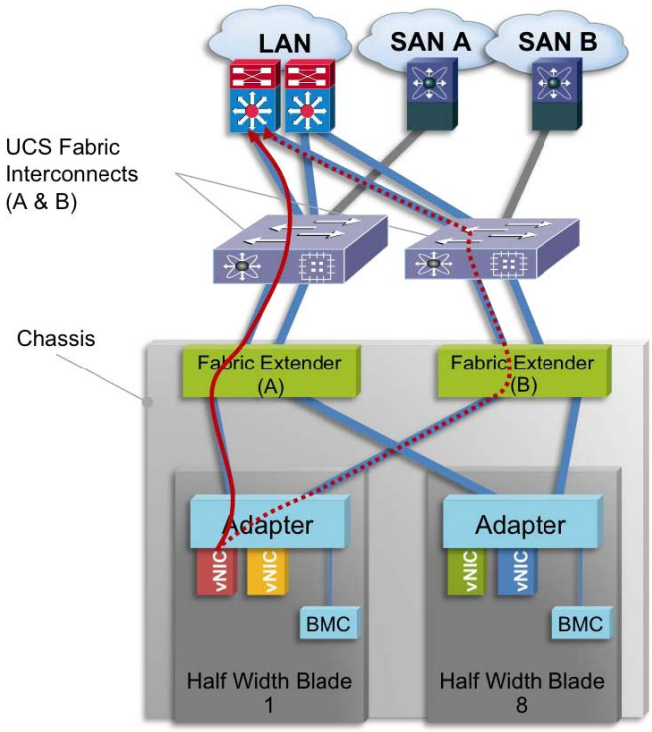
- UCS End-Host (EH) mode
- Nexus 1000V and MAC-pinning (i.e., as previously discussed)
- Redundant VSMs and VSGs in active-standby mode

- VMware High Availability intra-cluster

UCS End-host Mode

The UCS features a highly redundant architecture with redundant power, fabrics (i.e., data plane), control plane and I/O (Figure 2-11).

Figure 2-11 UCS



At this compute layer of the infrastructure, virtual machine NICs are pinned to UCS fabric uplinks dynamically or statically. These uplinks connect to the access layer switching systems, providing redundancy towards the network. In the VMDC solution, UCS Fabric interconnect uplinks operate in EH mode. In this mode, the 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 L2 switching. Key benefits with EH 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

Nexus 1000V and MAC-pinning

The Cisco UCS load balances traffic for a given host interface on one of the two redundant internal fabrics. By default, if a fabric fails, traffic automatically fails over to the available fabric. However, the UCS only supports port-ID and source MAC address-based load balancing mechanisms. As previously discussed, the 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, then another uplink is dynamically chosen.

Active/Standby Redundancy

For high availability, the Cisco Nexus 1000V Series Virtual Supervisor Module (VSM) must be deployed 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 to 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.

VSG redundancy is configured similarly to VSM redundancy; that is, like redundant VSMs, redundant VSGs must be installed on two separate physical hosts. One will be defined as the primary VSG and one as a secondary VSG, operating in active/standby HA mode. As in the VSM case, DRS, VMware HA, and VMware FT should be disabled for the redundant VSG VMs. One may use the anti-affinity feature of VMware ESXi to help keep the VSMs on different servers.

Intra-Cluster High Availability

The VMDC architecture prescribes the use of VMware HA for intra-cluster resiliency. In contrast to VMware FT, 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.

VNMC redundancy is addressed through VMware's HA mechanism, assuming creation of an ESXi cluster in which the redundant VNMC VMs reside.

More generally, this technology is applicable for VMs running back-end management applications.

Additional Considerations

Though not the focus of this release, additional resilience best practices would include the use of application-level clustering, and periodic VM and host backup mechanisms, such as snapshots or cloning and periodic database backups. These are all particularly applicable in terms of insuring HA for back-end management hosts and virtual machines.

To facilitate maintenance operations or business continuance inter-site, the creation of automated disaster recovery plans for groups of virtual machines using scripted tools or utilities such as VMware's Site Recovery Manager may be necessary. This topic is discussed in VMDC 2.0 and Data Center Interconnect systems documentation.

Storage Redundancy

In the storage layer, the high availability design is consistent with the HA model implemented at other layers in the infrastructure, comprising physical redundancy and path redundancy.

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.

Link Redundancy

Multiple individual FC links from the 6140s 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 may optionally be used to optimize use of the available link bandwidth and enhance load balancing across multiple active host adapter ports and links with minimal disruption in service.

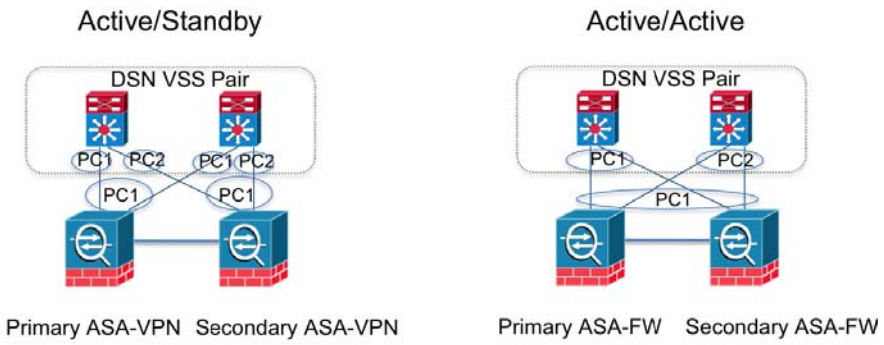
Services Redundancy

As previously noted, in the services layer of the infrastructure, redundancy is employed comprehensively to insure no single point of failure. This includes physical (hardware, links) and logical (i.e., paths, control plane) redundancy.

ASA

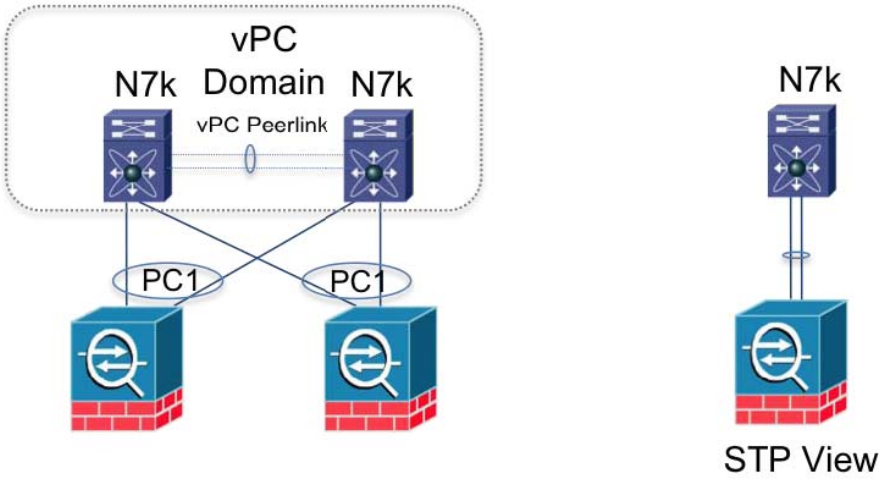
In this system release, two pairs of redundant ASA appliances are utilized for secure VPN remote access and for per-tenant perimeter firewalling. Release 8.4.1 for the ASA introduced support for several key HA features: 802.3ad EtherChannels and stateful failover with dynamic routing protocols, dramatically improving availability for the ASA in vPC or VSS enabled infrastructures. With this release, the ASA systems support configuration of up to 48 EtherChannels; each channel group may consist of up to eight active interfaces. Two failover modes are supported: active/standby and active/ active. If redundant ASAs are configured in active/standby failover mode, two separate EtherChannels must be configured on each upstream switch in the VSS (1 per ASA, as in Figure 2-12). In contrast, in active/active mode, only one EtherChannel is required per switch in the VSS pair. As of this writing, active/active failover is only supported when ASAs are in multi-context mode. Multi-context mode signifies that virtual contexts are configured on the ASA, dividing it into multiple logical firewalls, each supporting different interfaces and policies. Thus in this release, only the ASAs used for firewalling are configured for active/active failover (right in Figure 2-12). In this scenario, best practice recommendations include enabling interface monitoring and low polltime in failover configuration to get better resiliency and faster convergence of traffic traversing port-channels in the event of link failure.

Figure 2-12 ASA Redundancy Modes



Though only validated in this system with MEC on a VSS pair, it is important to note that this scenario will work in a vPC environment as well, for redundant connectivity directly to Nexus 7000 aggregation nodes. In this scenario, vPC allows creating a L2 port-channel between redundant Cisco Nexus 7000 Series devices and each redundant ASA. The concept is slightly different from VSS in that the two Nexus 7000 nodes are still independent switches, with different control and forwarding planes.

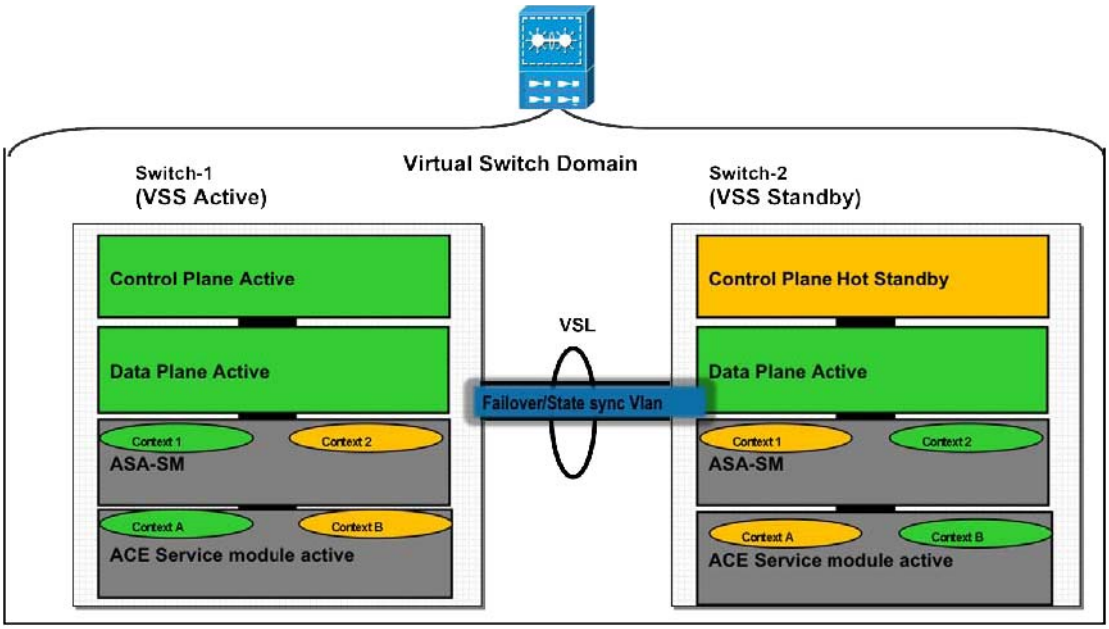
Figure 2-13 ASA Redundancy with Nexus 7000



ACE

Though like the ASA, the ACE Server Load Balancer is available in both service module and appliance form factors, as in previous releases this system was validated only with the service module form factor (i.e., the ACE-30). This conveniently provided an opportunity to contrast HA in the context of appliance-based services (i.e. the ASA case), with HA in service module form factor. In service module form factor, ACE HA is dependent on key functionality provided by the Data Center Services Nodes. When configured as a VSS pair, the nodes form a single virtualized switch domain, with shared redundant control and data planes. Through failover group definitions, redundant service modules placed within each node of the VSS pair are thus able to function in active/active failover mode, per virtual context.

Figure 2-14 VSS and Service Module Redundancy



As described in previous releases, the VSS pair itself relies on MEC and vPC technologies for loop-free redundancy to the aggregation layer.

Service Assurance

Service assurance is generally defined as a set of service level management processes insuring that a product or service meet specified performance objectives tailored to customer or client requirements. These processes involve controlling traffic flows, monitoring and managing key performance indicators to proactively diagnose problems, maintain service quality, and restore service in a timely fashion. The fundamental driver behind service assurance is to maximize customer satisfaction.

Though network service assurance covers a broad spectrum of metrics, including traffic engineering, performance monitoring, and end-to-end system availability, the VMDC 2.2 release focuses specifically on one particular component of service assurance that is key to providing differentiated service level agreements (SLAs): this is Quality of Service (QoS).

In VMDC 2.2, the QoS framework is modified with the following objectives in mind:

- Continued support for Network Control, Network Service, and Network Management traffic classes. Including VMware vMotion, Service Console, and other infrastructure management flows, these are characterized as mission critical categories, essential to maintaining administrative operations during periods of network instability or high CPU utilization.
- Continued support for three data service tiers (i.e., as in all previous VMDC systems releases). In terms of SLAs, these are characterized by two metrics - differentiated bandwidth (i.e., B1, B2 and B3) and availability.
  - In private or public hosted cloud environments, these can be thought of as three utility compute service tiers (i.e., Gold, Silver, and Bronze).

  - In public hybrid inter-cloud environments, these can be part of a more elaborate set of end to-end service tiers, with Gold and Silver classes correlating to business critical (in-contract, out-of-contract) SLAs.

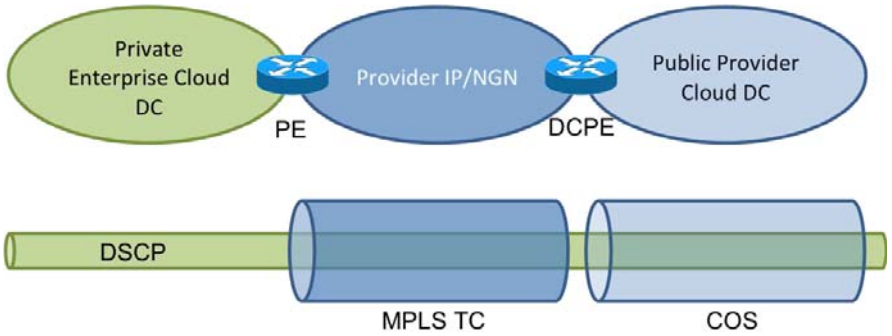
- Support for multimedia, hosted collaboration traffic flows. In terms of SLAs, the low latency traffic classes in this new multimedia service tier (i.e., VoIP bearer and video conference) are characterized by three metrics - bandwidth, delay, and availability. The requisite traffic flows comprise:
  - a new data bandwidth class for Cisco WebEx interactive collaboration
  - VoIP bearer traffic
  - VoIP call control
  - Video conferencing
  - Video streaming (future)
- Support for admission control (future). QoS is a prerequisite for admission control, which may be applicable to future cloud bursting scenarios.
- Support QoS across hybrid public/private domains

In the past, various VMDC system releases have followed either the traditional Cisco Enterprise/ Campus QoS model or the Cisco Service Provider IP/NGN QoS model, depending upon the use case scenarios and targeted audience. These differ slightly in terms of traffic classifications and markings, with the Service Provider model featuring slightly more complexity based on the need to support SLAs end-to-end from public to private QoS domains (Figure 2-15). In consideration of the objectives above, the QoS framework described in this release aligns with the IP/NGN QoS model.

The hybrid prerequisite imposes an additional requirement that has traditionally been unique to the public provider case, but in future as cloud SLAs evolve, may apply to inter-cloud networking scenarios in a private-to-private cloud context. This is the need for QoS transparency. Described in RFC3270, QoS transparency allows a public provider to use their own marking scheme, prioritizing the Enterprise's priority traffic without remarking the DSCP field of the IP packet. With this, the QoS marking delivered to the destination network corresponds to the marking received when the traffic entered the IP/NGN domain.

Any SLAs that are applied would be committed across each domain; thus, public provider end-to-end SLAs would be a concatenation of domain SLAs IP/NGN + public provider data center). Within the public provider data center QoS domain SLAs must be committed from data center edge to edge: at the PE southbound (into the data center), in practice there would thus be an SLA per-tenant per class, aligning with the IP/NGN SLA, and at the Nexus 1000V northbound there would be an SLA per VNIC per VM (or optionally per class per VNIC per VM). As this model requires per tenant configuration at the data center edges only (i.e., PE and Nexus 1000V), ideally there is no per-tenant QOS requirement at the core/agg/access layers of the infrastructure.

Figure 2-15 Hybrid End-to-End QoS Domains



The QoS framework defined in VMDC Release 2.2 follows the "hose" model for point-to-cloud services. This defines a point-to-multipoint (P2MP) resource provisioning model for VPN QoS, and is specified in terms of ingress committed rate and egress committed rate with edge conditioning. In this model, the focus is on the total amount of traffic that a node receives from the network (i.e., tenant aggregate) and the total amount of traffic it injects into the network. In terms of the VMDC architecture, the hose model is directly applicable to the edge QoS implementation at the public provider PE (i.e., the ASR 9000 DC PE in this program phase). Use case scenarios include P2MP VPLS-based transport services (i.e., hybrid DCI use cases), as well as more general VPDC services (i.e., where MPLS L2 or L3 VPNs provide inter-cloud transport).

To provide differentiated services, this release leverages the following QoS functionality:

- Traffic classification and marking
- Congestion management and avoidance (queuing, scheduling, and dropping)
- Traffic conditioning (shaping and policing)

Traffic Classification and Marking

Classification and marking allow QoS-enabled networks to identify traffic types based on information in source packet headers (i.e., L2 802.1p CoS and DSCP information) and assign specific markings to those traffic types for appropriate treatment as the packets traverse nodes in the network. Marking (coloring) is the process of setting the value of the DSCP, MPLS EXP, or Ethernet L2 CoS fields so that traffic can easily be identified later, i.e. using simple classification techniques. Conditional marking is used to designate in-contract (i.e., "conform") or out-of-contract (i.e., "exceed") traffic.

As in previous releases, the traffic service objectives considered in release 2.2 translate to support for three broad categories of traffic:

1. Infrastructure
2. Tenant Service Classes (three data; two multimedia priority)
3. Storage

Figure 2-16 shows a more granular breakdown of the requisite traffic classes characterized by their DSCP markings and per-hop behavior (PHB) designations. This represents a normalized view across the VMDC and hosted collaboration validated reference architectures in the context of an eight-class IP/NGN aligned model.

Figure 2-16 VMDC 2.2 Traffic Classes (Eight-Class Reference)

Traffic Class	EXP/CoS	DSCP	PHB
Utility Compute Data: Bronze-Standard	0	CS0	Default
Utility Compute Data: Silver-Business to Business & Webex Collaboration Data (Interactive)*	1	CS1	AF
Utility Compute Data: Gold – Business Critical	2	CS2	AF
Storage – FCOE & VoIP Call Control	3	CS3	AF42,AF43
Video Streaming (Future)*	4	CS4	AF41
VoIP Bearer & Video Conference	5	CS5	EF
Network Control	6	CS6	AF
Network Mgmt & Service Control	7	CS7	AF

*Webex , Video Streaming and NFS flows not included in 2.2 test scenarios

It is a general best practice to mark traffic at the source-end system or as close to the traffic source as possible in order to simplify the network design. However, if the end system is not capable of marking or cannot be trusted, one may mark on ingress to the network. In the QoS framework defined in this release, the Provider Data Center represents a single QoS domain, with the Nexus 1000V forming the "southern" access edge, and the ASR 9000 forming the "northern" DC PE/WAN edge. These QoS domain edge devices will mark traffic, and these markings will be trusted at the nodes within the data center infrastructure; in other words, they will use simple classification based on the markings received from the edge devices.

Queuing, Scheduling, and Dropping

In a router or switch, the packet scheduler applies policy to decide which packet to dequeue and send next, and when to do it. Schedulers service queues in different orders. The most frequently used are:

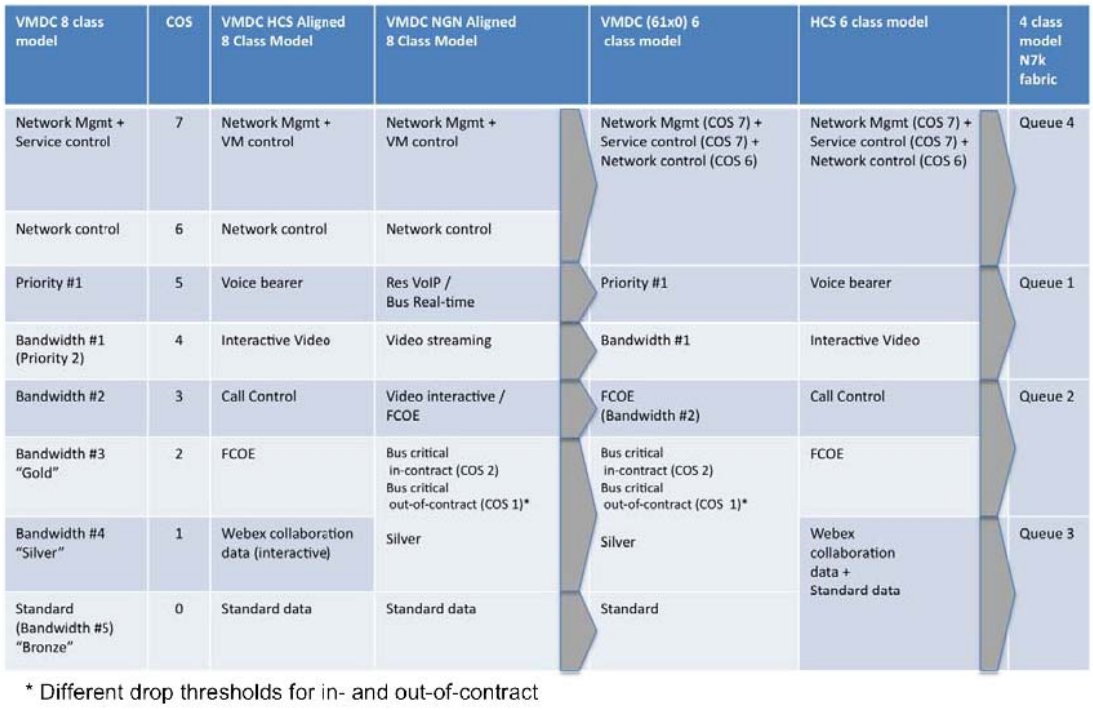
- First in, first out (FIFO)
- Priority scheduling (aka priority queuing)
- Weighted bandwidth

In this release, we use a variant of weighted bandwidth queuing called class-based weighted fair queuing/low latency queuing (CBWFQ/LLQ) on the Nexus 1000V at the southern edge of the DC QoS domain, and at the ASR 9000 northern DC WAN edge, we use priority queuing(PQ)/CBWFQ to bound delay and jitter for priority traffic while allowing for weighted bandwidth allocation to the remaining types of data traffic classes.

Queuing mechanisms manage the front of a queue, while congestion avoidance mechanisms manage the tail end of a queue. Since queue depths are of limited length, dropping algorithms are used to avoid congestion by dropping packets as queue depths build. Two algorithms are commonly used: weighted tail drop (often for VoIP or video traffic) or weighted random early detection (WRED), typically for data traffic classes. In this release, WRED is used to drop out-of-contract data traffic (i.e., CoS value 1) before in-contract data traffic (i.e., Gold, CoS value 2), and for Bronze/Standard traffic (CoS value 0) in the event of congestion.

One of the challenges in defining an end-to-end QoS architecture is that not all nodes within a QoS domain have consistent implementations. Within the cloud data center QoS domain, we run the gamut from systems that support 16 queues per VEM (i.e., Nexus 1000V) to four internal fabric queues (i.e., Nexus 7000). This means that traffic classes must be merged together on systems that support less than eight queues. Figure 2-17 shows the class to queue mapping that applies to the cloud data center QoS domain in the VMDC 2.2 reference architecture, in the context of alignment with either the HCS reference model or the more standard NGN reference.

Figure 2-17 VMDC Class to Queue Mapping



* Different drop thresholds for in- and out-of-contract

Shaping and Policing

Policing and shaping are techniques used to enforce a maximum bandwidth rate on a traffic stream; while policing effectively does this by dropping out-of-contract traffic, shaping does this by delaying out-of-contract traffic.

In this release, policing is utilized within and at the edges of the cloud data center QoS domain to rate limit data and priority traffic classes. At the ASR 9000 data center PE, hierarchical QoS (HQoS) is implemented on egress to the Cloud data center; this uses a combination of shaping and policing in which L2 traffic is shaped at the aggregate (port) level per class, while policing is utilized to enforce per-tenant aggregates.

Sample bandwidth port reservation percentages used in validation to analyze QoS policy effects are shown in Figure 2-18.

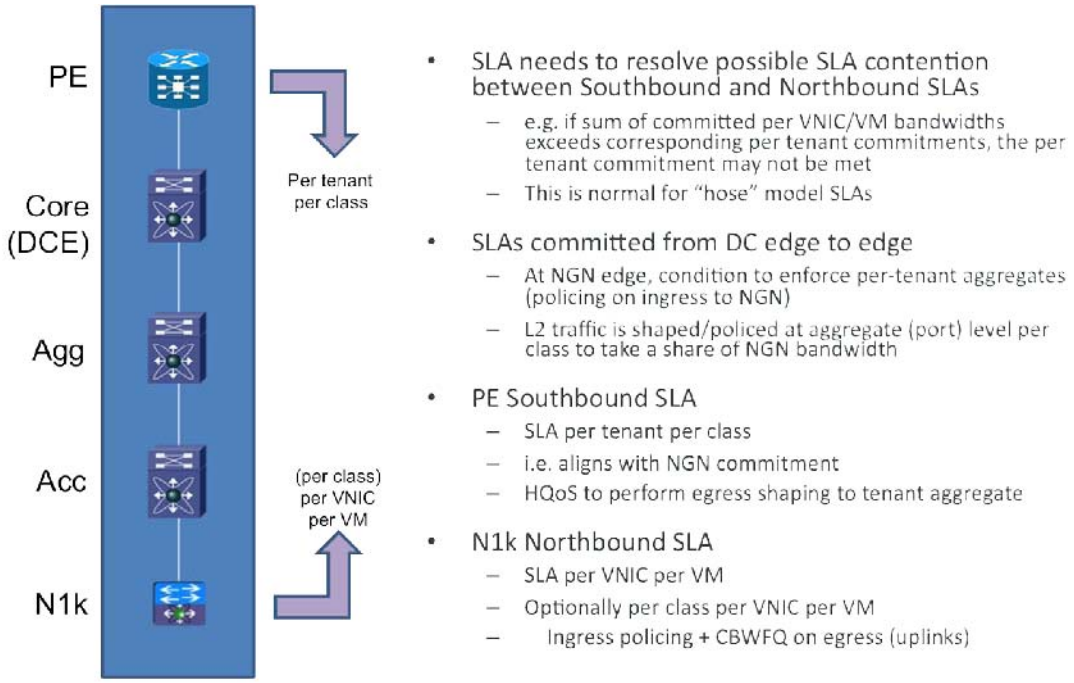
Figure 2-18 Sample Bandwidth Reservations (% of Port)

Traffic Class	EXP/CoS	BW Reserved (Remaining After Priority)	Actions
Utility Compute Data: Bronze-Standard	0	15% (17%)	WRED
Utility Compute Data: Silver-Business & Webex Collaboration Data (Interactive)*	1	60% (70%)	WRED Out of Contract dropped before in contract
Utility Compute Data: Gold – Business Critical	2		WRED
Storage – FCOE & VoIP Call Control	3	3% (4%)	
Video Streaming (Future)*	4	-	
VoIP Bearer & Video Conference	5	15%	Priority, egress policed per tenant
Network Control	6	4% (5%)	
Network Mgmt & Service Control	7	3% (4%)	

*Webex , Video Streaming and NFS flows not included in 2.2 test scenarios

Figure 2-19 provides a high-level synopsis of this end-to-end SLA framework.

Figure 2-19 End-to-End SLA Framework



Scalability

The ability to grow and scale the cloud infrastructure is a function of many factors, ranging from environmental, to physical and logical capacity. Considerations extend beyond the technical scope into the administrative domain.

L2 Scale

Within the L2 domain, several key factors affect scale. These include:

- **Virtual Machine Density**—The number of VMs enabled on each server 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, Communications as a Service (CaaS), which provides raw compute and memory resources on-demand, agnostic to the applications running, is often characterized simply in terms of VMs per CPU core, with packaged bundles of memory options. The number of VMs per CPU core is a significant factor in another way, in that it in turn drives the number of network interfaces (virtual) required to provide access to VMs.
- **VMNics per VM**—Each VM instance requires at minimum two vNICs; in most cases, several are utilized for connections to various types of Ethernet segments, and the ESX host itself will require network interfaces , i.e., for management control interfaces.
- **MAC Address Capacity**—The number of VMs and vNICs per VM will drive MAC table size requirements on switches within the L2 domain. Generally, these tables are implemented in hardware rather than software. So, unless a hardware upgrade is feasible, they will provide an upper bound to the scope of a single L2 domain. In the VMDC system reference architecture, the aggregate number of MAC addresses required within a pod is calculated based on the following formula: (# of server blades per pod) x (# of cores/blade) x (# of VMs/core = 1, 2, 4) x (# of MACs/VM = 4)
- **Cluster Scale**—Cluster sizes are constrained in a number of dimensions, i.e., in terms of number of servers, VMs, and logical storage I/O.
- **ARP Table Size**.
- **VLANS**—VLANs provide logical segmentation within the L2 domain, scaling VM connectivity, providing application tier separation and multi-tenant isolation. Every platform within the L2 and L3 portions of the infrastructure will have VLAN budgets which must be considered when designing tenant containers.
- **Port Capacity**—At the network layer, hardware port density is another physical budgetary constraint. Similarly, this consideration also applies to the compute layer, in terms of logical Ethernet capacity on virtual access edge switches.
- **Logical Failure Domain**—A L2 domain is also a single logical failure domain. From an administrative perspective, operational considerations come into play, in terms of how long it may take to recover from various types of failures if the affected set of resources is quite large.

- **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.

L3 Scale

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**—VRF instances can be used to define a tenant network container. 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.

Resource Oversubscription

Increasing the efficiency of resource utilization is the key driver to oversubscription of hardware resources. This drives CAPEX savings up while still maintaining SLAs.

Network Oversubscription

In considering what network oversubscription ratios will meet their performance requirements, network architects must consider likely traffic flows within the logical and 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

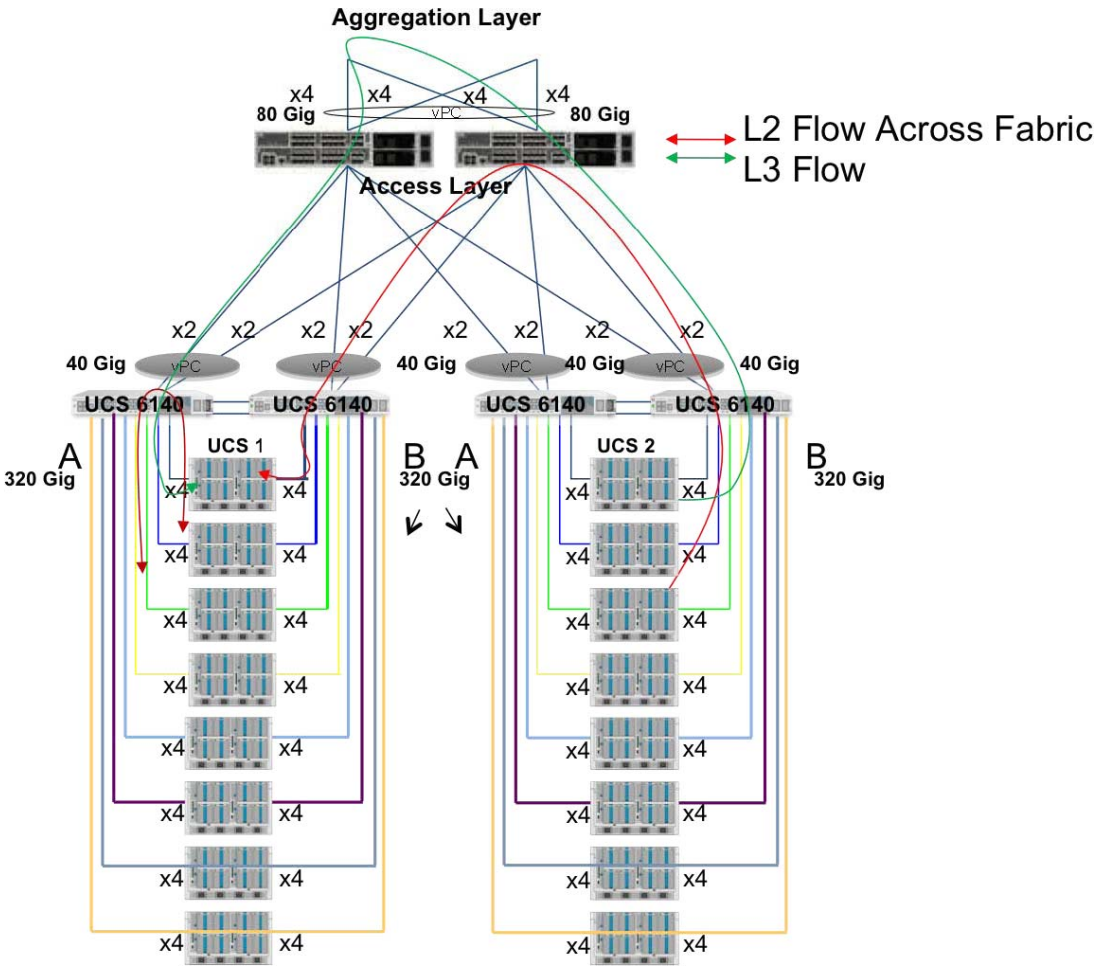
Some traffic flow types and considerations are as follows:

- **Server-to-server L2 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 L2 communications between different UCS fabrics**. As depicted in Figure 2-20, the EH 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 EH 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 L3 communications**. Keeping multiple tiers of an application within the same UCS fabric is recommended if feasible, as it will provide predictable traffic patterns. However, if the two tiers are on the same UCS fabric but on different VLANs, routing is required between the application tiers. This routing results in traffic flows to and from the aggregation layer to move between subnets.

Figure 2-20 Traffic Flows Across the UCS System



In practice, network oversubscription ratios commonly used range from 4:1 to 8:1, depending on use case and level of infrastructure hierarchy. In this VMDC 2.X reference design, an 8:1 network oversubscription for inter-server traffic is considered for general compute deployment. This concept is illustrated in Figure 2-20, where the UCS chassis are connected to each UCS 6140 with 40 Gb (4x10 Gb) of bandwidth. When all eight 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 oversubscription ratio of 8:1 at the access layer when all links are active.

Similarly, the oversubscription ratio of 8:1 is provisioned at the aggregation layer when the all links are active. Oversubscription 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 constrained by WAN link bandwidth. 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.

Compute Oversubscription

Server virtualization 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. In a very simple sense, compute oversubscription may be thought of as the ratio of vCores per VM per server or blade, and in terms of VMs per Gb of memory per blade. Of course, application workloads in real environments have distinct logical footprints of processing, memory, and storage requirements. For this reason, analysis of integrated compute stacks, which includes consideration of IOPS performance is in fact conducted with specific applications generating traffic streams. However, for infrastructure modeling purposes, if IOPS performance is not a test criteria, it is useful to create profiles representing averages of varying workload sizes. In modeling the VMDC infrastructure, three workload profiles are leveraged with the following characteristics:

- Large (20%): 1 vCore/VM (1:1)
- Medium (30%): .5 vCore/VM (2:1)
- Small (50%): .25 vCore/VM (4:1)

Older 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 current generation of B series blade servers supports 12 cores per blade. In an eight-chassis system, this will equate to 64 blades x 12 cores or 768 cores per system. With workload distributions as above, this equates to 2,148 VMs per eight-chassis system, or 17,208 VMs per eight UCS systems of eight chassis each.

Figure 2-21 Sample Workload Profile Distributions

Workload Profile	Distribution	Blades	vCores (8-core)	vCores (12-core)	VMs/ Core	VMs (8-core)	VMs (12-core)
Large	20%	13 (102)	104 (816)	156 (1,224)	1	104 (816)	156 (1,224)
Medium	30%	19 (154)	152 (1,232)	228 (1,848)	2	304 (2,464)	456 (3,696)
Small	50%	32 (256)	256 (2,048)	384 (3,072)	4	1,024 (8,192)	1536 (12,288)
Total 1 UCS/8 chassis (8 UCS/64 chassis)		64 (512)	512 (4,096	768 (6,144)		1,432 (11,472)	2148 (17,208)

Bandwidth per VM

As illustrated in Figure 2-20 and Figure 2-21, a 1:1, 1:2 and 1:4 core:vm ratio for large/medium/small workload types with a 20/30/50 distribution leads to an average of 22 VMs per blade, 1,432 VMs per UCS, and 11,472 maximum per pod. In the case of twelve-core blades, this is 34 VMs per blade, 2,148 VMs per UCS and 17,208 maximum VMs per pod. The network bandwidth per VM can be derived as follows:

The UCS-6140 supports eight uplinks each, so each UCS system can support 80G/1432 = 56M per VM. Oversubscription prunes per VM bandwidth at each layer - aggregation, core, and edge. The core layer provides 1:1 load-balancing (L2 and L3), hence 80G/1432 = 56M per VM within each UCS. Extrapolating to a maximum pod size of 512 servers, this equates to approximately (80G/11472) 7M per VM (eight-core scenario) or (80G/17208) 5M per VM (twelve-core scenario).

Storage Oversubscription

In a shared storage environment, thin provisioning is a method for optimizing utilization of available storage through oversubscription. It relies on on-demand allocation of blocks of data versus the traditional method of allocating all the blocks up front. This methodology eliminates almost all whitespace, which helps avoid poor utilization rates that may occur in the traditional storage allocation method where large pools of storage capacity are allocated to individual servers but remain unused (not written to). In this model, thinly provisioned pools of storage may be allocated to groups of vApps with homogenous workload profiles. Utilization will be monitored and managed on a pool-by-pool basis.

Storage bandwidth calculations for this system can be derived as follows:

There are 4x4G links from each UCS-6140 to MDS (aligning with a VCE Type 2 Vblock). Assuming equal round-robin load-balancing from each ESX blade to each fabric, there is 32G of SAN bandwidth. Inside each UCS system, 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 UCS system, 88,000/1432 equates to 61 IOPS per VM. Extrapolating to a maximum 512 server pod, 88000/11472 provides just under 8 IOPS per VM (eight-core scenario) or approximately 5 IOPS per VM (twelve-core scenario). Of course, one may add more FC and Ethernet ports to increase the per VM Ethernet and FC bandwidth.

DC Scalability

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

- **MAC Address Support on the Aggregation Layer—**The Nexus 7000 platform supports up to 128,000 MAC addresses. For example, considering the modeled distribution mix of small, medium, and large workloads, 11,472 workloads would theoretically be enabled in each large pod, which translates to 11,472 VMs (i.e., on eight-core blades) or 17,208 workloads and VMs on twelve-core B200 series blades. Different vNICs with unique MAC addresses are required for each VM data and management network, as well as NICs on the ESX host itself. The VMDC solution assumes four MAC addresses per VM and this translates to 45,888 (or 68,832) MAC addresses per large pod. In order to optimize intra-pod scale, sharing VLANs between pods is generally 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 pods can be added while still providing network oversubscription ratios that are acceptable for the deployed applications. For example, from a physical port density standpoint (based on the M1 series linecards), the Nexus 7018 could theoretically support up to six large pods, each equating to 512 blades.
- **Control Plane Scalability**—Control plane scalability will vary depending upon the type of encapsulation(s) used to identify tenants, L2 protocols in use (i.e., HSRP, STP), and upon route protocol selection. In the case where VRF-lite is used, 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. This design has been characterized for 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 future, application of LSP and Inter-AS at the core of the infrastructure will serve to further scale this model.