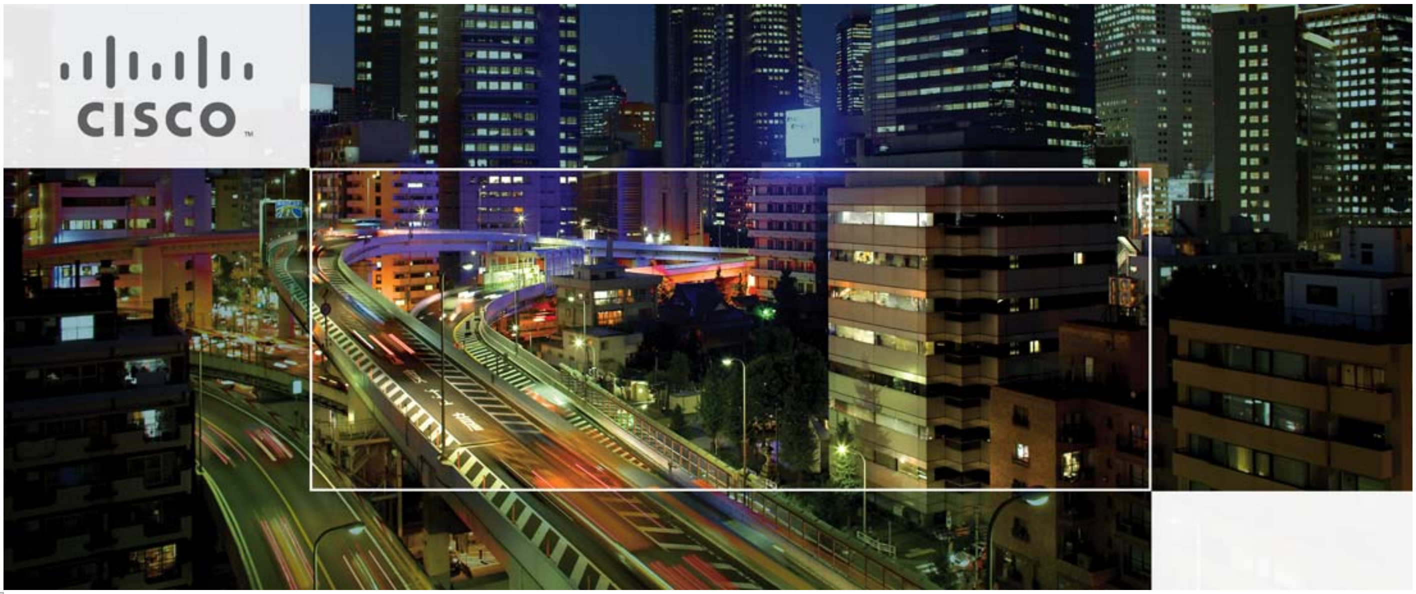


Virtualized Multiservice Data Center (VMDC) Virtual Services Architecture (VSA) 1.0 Design Guide

December 6, 2013

CONTENTS

CHAPTER 1

VMDC VSA 1.0 Introduction

The Cisco Virtualized Multiservice Data Center (VMDC) system provides design and implementation guidance for enterprises deploying private cloud services, and for service providers (SPs) building public and virtual private services. With the goal of providing an end-to-end system architecture, VMDC integrates Cisco and third-party products in the cloud computing ecosystem.

Audience

This guide is intended for, but not limited to, system architects, network design engineers, system engineers, field consultants, advanced services specialists, and customers who want to understand how to deploy a public or private cloud data center infrastructure. This guide assumes that the reader is familiar with the basic concepts of IP protocols, quality of service (QoS), and high availability (HA), and that readers are aware of general system requirements and have a basic understanding of enterprise or SP network and data center architectures.

Overview

VMDC, Cisco's reference architecture for cloud deployment, has been widely adopted by numerous SPs and enterprises worldwide. In this and previous releases, VMDC has provided design guidance for scalable, secure, resilient, public and private cloud infrastructures serving multiple consumers or tenants:

- In the data center portion of the architecture, VMDC 2.X designs were centered on traditional hierarchical infrastructure models incorporating leading Cisco platforms and Layer 2 (L2) resilience technologies such as Virtual Port Channel (vPC), providing network containers or tenancy models of different sizes and service profiles, with necessary network based services and orchestration and automation capabilities to accommodate the various needs of cloud providers and consumers.
- VMDC 3.X systems releases introduced Cisco FabricPath for intra-DC networks, as an optional L2 alternative to a hierarchical vPC-based design. FabricPath removes the complexities of Spanning Tree Protocol (STP) to enable more extensive, flexible, and scalable L2 designs. Customers leveraging VMDC reference architecture models can choose between vPC-based and FabricPath-based designs to meet their particular requirements.

VMDC VSA 1.0 is the first VMDC release dealing specifically with the transition to NFV (Network Function Virtualization) of IaaS network services in the data center. Such services comprise virtual routers, virtual firewalls, load balancers, network analysis and WAN optimization virtual appliances.

In this release, we focus mainly on public provider use cases, building a new logical topology model around the creation of virtual private cloud tenant containers in the shared data center infrastructure. Future releases will incorporate additional cloud consumer models specific to enterprise and private cloud use cases. In particular, future releases will address hybrid consumer models, comprising physical and virtual service appliances, used together as part of a per-consumer or per-tenant service set. These can be implemented on either a 2.X (classical Ethernet) or 3.X (FabricPath) VMDC infrastructure. However, in this release we focus on fundamental implications of an all-virtual approach, and have opted to do so over a simple FabricPath data center topology previously validated in VMDC 3.0.

Problem Statement

The architecture described in this guide addresses the following customer challenges:

1. Tenancy Scale: Previous VMDC systems releases leveraged various abstraction technologies, for example, virtual LANs (VLANs) and virtual routing and forwarding (VRF), for tenant isolation, including separated routing and forwarding. Each abstraction technology impacts logical scale and control plane overhead. In a traditional hierarchical DC network model, the pressure point from a scalability and control plane perspective is at the aggregation layer of the infrastructure, with the number of route peers, VRFs, VLANs, and MAC capacity supported by aggregation nodes presenting key multi-dimensional scalability factors. The virtual services architectural (VSA) model introduced in this release presents an alternative, addressing tenancy scale using a centralized provider edge (PE) and distributed, per-tenant virtual customer edge (vCE) routing model. Tenancy scale is thus increased to the number of eBGP peers (or alternatively, static routes) supported by the PE nodes. As of this writing, this is 5000 per pair of redundant ASR 9000 Series PE routers.
2. Complexity: Current VMDC architecture models feature a relatively high degree of management complexity because service appliances are shared across multiple tenants, and are allocated in logical “slices” (contexts) by automation systems. The VSA model reduces service orchestration complexity, removing cross-tenant dependencies for L4-L7 service allocation. The VSA model represents a simpler logical topology compared to the back-to-back VRF-Lite method employed in VMDC 2.X releases to create rigorous (VRF-based) tenant isolation.
3. Customer Evolution to NFV for IaaS: For years, customers have seen the transition from physical to virtual services as a foundation for an evolution toward “next-gen” data center service-oriented architectures, providing increased flexibility and agility through greater “software definability”.
4. Need for virtual appliance-based multi-tenancy design guidance. VMDC VSA 1.0 is a starting point, representing an opportunity to initially consider one specific deployment model (the vCE model) out of several possible options for an “all-virtual” virtual private cloud instantiation, exploring end-to-end service differentiation, performance and impact on future automation requirements.
5. Need to address logical segmentation constraints of traditional 802.1q VLAN L2 domains through the application of virtual overlays. VMDC VSA 1.0 presents a first look at the use of VXLANs for logical segmentation.

VMDC VSA 1.0 addresses the following use cases:

- Data center and PoD design
- Split N-tiered applications
- Multi-tenancy (including Virtual Extensible LAN (VLXAN)-based logical segmentation)
- Application-centric instrumentation (statistics collection, network analysis, WAN optimization, Performance Agent)

Solution Proposal

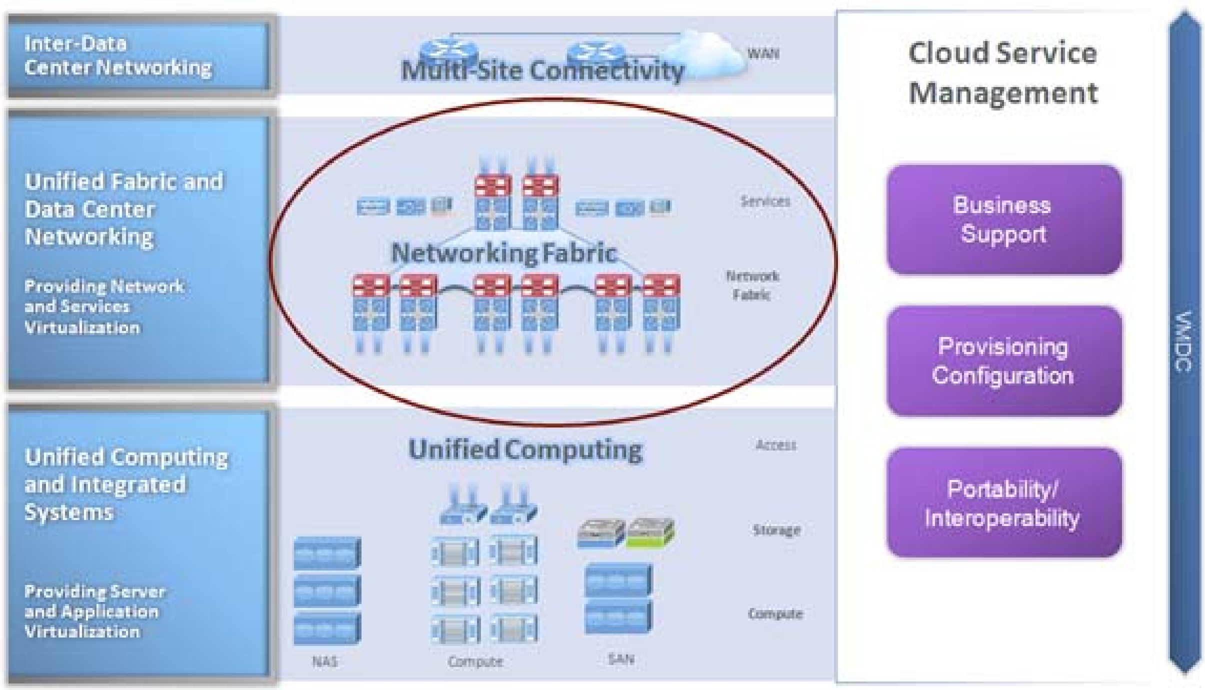
To address the identified requirements, we modified the Unified Computing component of the VMDC architecture, shifting virtualized service functions from the Unified Fabric/Data Center Networking portions of the infrastructure. Figure 1-1 shows a high level view of the overall VMDC system.

In general, the solution comprises three modular layers:

1. Unified Computing and Integrated Systems (UCIS), providing server and application virtualization, typically consisting of FlexPods and Vblocks.
2. Unified Fabric and Data Center Networking (UFDC), providing network and network based services virtualization.
3. Data Center Interconnect (DCI), providing seamless multi-site connectivity.

The solution is complemented by Cloud Service Management components that enable end to end provisioning and orchestration, along with monitoring and assurance.

Figure 1-1 High Level VMDC Solution



VMDC VSA 1.0 shifts network service functions to the UCIS in a way that leverages existing design guidance for the Unified Fabric and DCI layers, along with previous guidance for the UCIS layers in terms of compute and storage for **application** workloads. We maintained the following assumptions:

- Previous design guidance for UCIS (FlexPod, Vblock) components applies. VMDC VSA 1.0 validation was performed on the latest FlexPod release. Applications validated on FlexPod or Vblock continue to function on the overall VMDC architecture.
- Previous design guidance for DCI components applies. VMDC VSA 1.0 focuses mainly on single-site, intra-PoD Virtual Private Cloud (VPC) design alternatives.
- There are no top-level management and orchestration components in VMDC VSA 1.0: orchestration and service assurance are addressed in separate, parallel VMDC programs.

- Cisco (XaaS) applications, such as Unified Communications (UC), Hosted Collaboration Solution (HCS), Media Data Center, Video Surveillance, and TelePresence, use the VMDC architecture as the infrastructure for validation. At this writing, VMDC 2.3 is the latest release used for these validations. No specific Cisco application validations are in the scope of VMDC VSA 1.0. However, given the level of validation already performed, we are confident that these applications will work in VMDC VSA 1.0 infrastructures without major issues.

CHAPTER 2

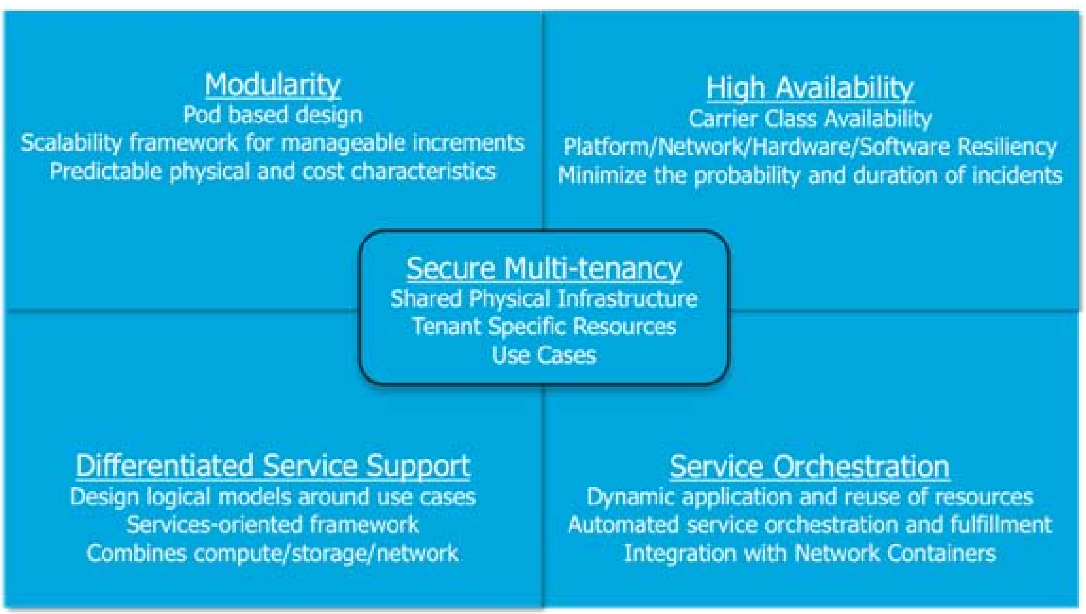
VMDC VSA 1.0 Design Overview

The Virtualized Multiservice Data Center (VMDC) architecture is based on the foundational design principles of modularity, high availability (HA), differentiated service support, secure multi-tenancy, and automated service orchestration (Figure 2-1).

Design Principles

These design principles provide streamlined turn-up of new services, maximized service availability, resource optimization, facilitated business compliance, and support for self-service IT models. These benefits maximize operational efficiency and enable private and public cloud providers to focus on their core business objectives.

Figure 2-1 VMDC Design Principles



Modularity—Unstructured growth is at the root of many operational and CAPEX challenges for data center administrators. Defining standardized physical and logical deployment models is the key to streamlining operational tasks such as moves, adds and changes, and troubleshooting performance issues or service outages. VMDC reference architectures provide blueprints for defining atomic units of growth within the data center, called PoDs.

High Availability—The concept of public and private "Cloud" is based on the premise that the data center infrastructure transitions from a cost center to an agile, dynamic platform for revenue-generating services. In this context, maintaining service availability is critical. VMDC reference architectures are designed for optimal service resilience, with no single point of failure for the shared ("multi-tenant") portions of the infrastructure. As a result, great emphasis is placed upon availability and recovery analysis during VMDC system validation.

Differentiated Service—Generally, bandwidth is plentiful in the data center infrastructure. However, clients may need to remotely access their applications via the Internet or some other type of public or private WAN. Typically, WANs are bandwidth bottlenecks. VMDC provides an end-to-end QoS framework for service tuning based upon application requirements. This release adds consideration of a set of tools for application visiibility, control and optimization, enhancing the ability to provide application-centric differentiated services.

Multi-tenancy—As data centers transition to Cloud models, and from cost centers to profit center, services will naturally broaden in scope, stretching beyond physical boundaries in new ways. Security models must also expand to address vulnerabilities associated with increased virtualization. In VMDC, "multi-tenancy" is implemented using logical containers, also called "Cloud Consumer" that are defined in these new, highly virtualized and shared infrastructures. These containers provide security zoning in accordance with Payment Card Industry (PCI), Federal Information Security Management Act (FISMA), and other business and industry standards and regulations. VMDC is certified for PCI and FISMA compliance.

Service Orchestration—Industry pundits note that the difference between a virtualized data center and a "cloud" data center is the operational model. The benefits of the cloud – agility, flexibility, rapid service deployment, and streamlined operations – are achievable only with advanced automation and service monitoring capabilities. The VMDC reference architectures include service orchestration and monitoring systems in the overall system solution. This includes best-of-breed solutions from Cisco (for example, Cisco Intelligent Automation for Cloud) and partners, such as BMC and Zenoss.

VMDC VSA 1.0 leverages FabricPath as the Unified Data Center fabric. FabricPath combines the stability and scalability of routing in Layer 2 (L2), supporting the creation of simple, scalable, and efficient L2 domains that apply to many network scenarios. Because traffic forwarding leverages the Intermediate System to Intermediate System (IS-IS) protocol, rather than Spanning Tree (STP), the bi-sectional bandwidth of the network is expanded, facilitating data center-wide workload mobility.

For a brief primer on FabricPath technology, refer to:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-687554.pdf
FabricPath benefits include:

Simplified Network, Reducing Operating Expenses

- FabricPath is simple to configure. The only necessary configuration consists of distinguishing core ports, which link the switches, from edge ports, to which end devices are attached. No parameters need to be tuned to achieve operational status, and switch addresses are assigned automatically.
- One control protocol is used for unicast forwarding, multicast forwarding, and VLAN pruning. Networks designed using FabricPath require less combined configuration than equivalent networks based on STP, further reducing the overall management needed for the solution.

- Static network designs make assumptions about traffic patterns and the locations of servers and services. If, as often happens over time, those assumptions become incorrect, complex redesign can be necessary. A fabric switching system based on FabricPath can be easily expanded as needed with additional access nodes in a plug and play manner, with minimal operational impact.
- Switches that do not support FabricPath can still be attached to the FabricPath fabric in a redundant way without resorting to STP.
- FabricPath L2 troubleshooting tools provide parity with those currently available in the IP community for non-fabric path environments. For example, the Ping and Traceroute features now offered at L2 with FabricPath can measure latency and test a particular path's among the multiple equal-cost paths to a destination within the fabric.

Reliability Based on Proven Technology

- Although FabricPath offers a plug-and-play user interface, its control protocol is built on top of the powerful IS-IS routing protocol, an industry standard that provides fast convergence and is proven to scale in the largest service provider (SP) environments.
- Loop prevention and mitigation is available in the data plane, helping ensure safe forwarding unmatched by any transparent bridging technology. FabricPath frames include a time-to-live (TTL) field similar to the one used in IP, and an applied reverse-path forwarding (RPF) check.

Efficiency and High Performance

- With FabricPath, equal-cost multipath (ECMP) protocols used in the data plane can enable the network to find optimal paths among all the available links between any two devices. First-generation hardware supporting FabricPath can perform 16-way ECMP, which, when combined with 16-port 10 gigabits per second (Gbps) port-channels, represents bandwidth of up to 2.56 terabits per second (Tbps) between switches.
- With FabricPath, frames are forwarded along the shortest path to their destination, reducing the latency of the exchanges between end stations compared to a STP based solution.
- FabricPath needs to learn at the edge of the fabric only a subset of the MAC addresses present in the network, enabling massive scalability of the switched domain.

Terminology

FabricPath comprises two types of nodes: spine nodes and leaf nodes. A spine node is one that connects to other switches in the fabric and a leaf node is one that connects to servers. These terms are useful in greenfield scenarios but may be vague for migration situations, where one has built a hierarchical topology and is accustomed to using traditional terminology to describe functional roles.

In this document, we expand our set of terms to correlate fabric path nodes and functional roles to hierarchical network terminology:

- **Aggregation-Edge**—A FabricPath node that sits at the "edge" of the fabric, corresponding to an aggregation node in a hierarchical topology.
- **Access-Edge**—A FabricPath node that sits at the edge of the fabric, corresponding to an access node in a hierarchical topology.

These nodes may perform L2 and/or L3 functions. At times, we also refer to an L3 spine or a L3 edge node to clarify the location of Layer 2/Layer 3 boundaries and distinguish between nodes that are performing Layer 3 functions versus L2-only functions.

FabricPath Topologies

FabricPath can be implemented in a variety of network designs, from full-mesh to ring topologies. In VMDC 3.0.X design and validation, the following DC design options, based on FabricPath, were considered:

- **Typical Data Center Design**—This model represents a starting point for FabricPath migration, where FabricPath is simply replaces older layer 2 resilience and loop avoidance technologies, such as virtual port channel (vPC) and STP. This design assumes that the existing hierarchical topology, featuring pairs of core, aggregation, and access switching nodes, remains in place and that FabricPath provides L2 multipathing.
- **Switched Fabric Data Center Design—**This model represents horizontal infrastructure expansion of the infrastructure to leverage improved resilience and bandwidth, characterized by a Clos architectural model.
- **Extended Switched Fabric Data Center Design—**This model assumes further expansion of the data center infrastructure fabric for inter-PoD or inter-building communication.

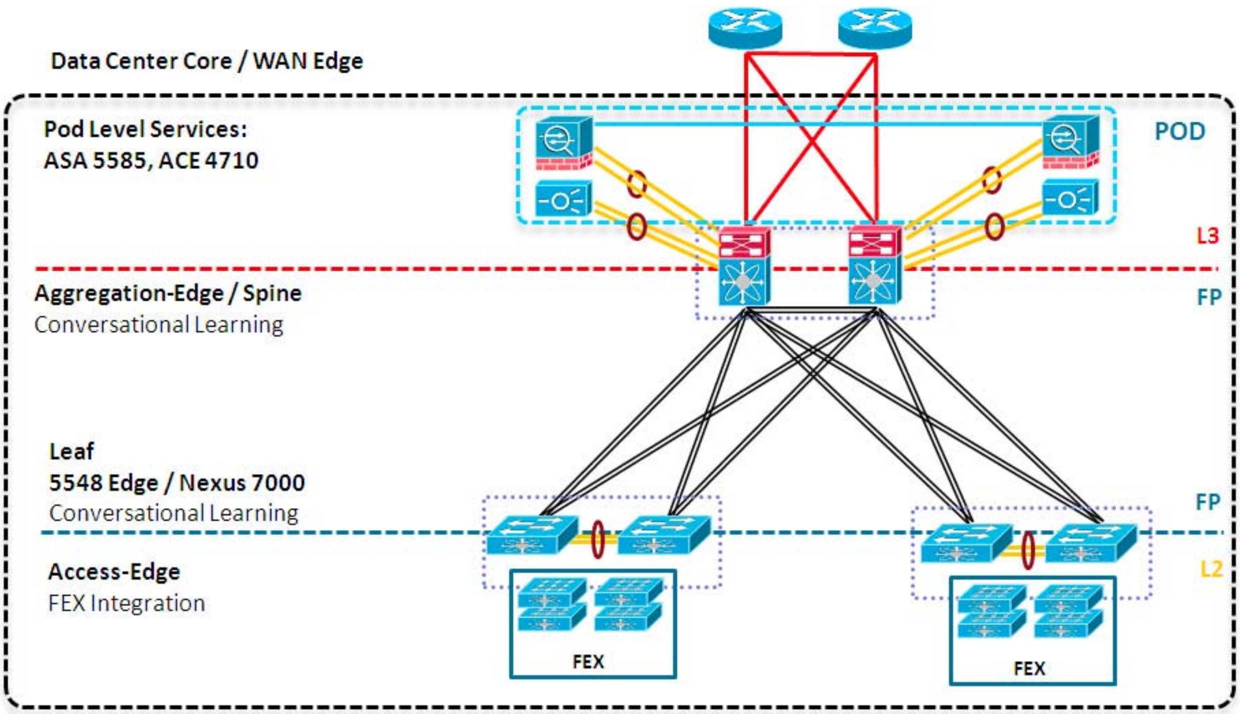
These are discussed in detail in VMDC 3.0 documentation. The Design Guide is publicly available, while the Implementation Guide is available to partners, and to Cisco customers under NDA.

While the logical containers discussed in VMDC VSA 1.0 may be implemented over a traditional classical Ethernet or FabricPath fabric, to constrain scope this release is based solely on the Typical Data Center FabricPath design option previously validated in VMDC 3.0/3.0.1.

FabricPath "Typical Data Center" Model

A Typical Data Center design is a two-tier FabricPath design as shown in Figure 2-2. VMDC architectures are built around modular building blocks called PoDs. Each PoD uses a localized Services attachment model. In a classical Ethernet PoD, vPCs handle L2 switching, providing an active-active environment that does not depend on STP, but converges quickly after failures occur. In contrast, Figure 2-2 shows a VMDC PoD with FabricPath as a vPC replacement.

Figure 2-2 Typical Data Center Design



From a resilience perspective, a vPC-based design is sufficient at this scale, although there are other benefits of using FabricPath, including:

- FabricPath is simple to configure and manage. There is no need to identify a pair of peers or configure port channels. Nevertheless, port channels can still be leveraged in FabricPath topologies if needed.
- FabricPath is flexible. It does not require a particular topology, and functions even if the network is cabled for the classic triangle vPC topology. FabricPath can accommodate any future design.
- FabricPath does not use or extend STP. Even a partial introduction of FabricPath benefits the network because it segments the span of STP.
- FabricPath can be extended easily without degrading operations. Adding a switch or a link in a FabricPath-based fabric does not result in lost frames. Therefore, it is possible to start with a small network and extend it gradually, as needed.
- FabricPath increases the pool of servers that are candidates for VM mobility and thereby enables more efficient server utilization.

Note Certain application environments, especially those that generate high levels of broadcast, may not tolerate extremely large Layer 2 environments.

VMDC 3.0-3.0.1 addressed several methods of resiliently attaching redundant appliance or module-based service nodes in order to optimize service availability and efficient link path utilization, including Ether-channel (for example, as shown previously), vPCs with Multi-Chassis EtherChannel on paired Virtual Switching Systems (VSSs), and vPCs on clustered (Cisco ASA) firewall appliances. However, service node implementation in VMDC VSA 1.0 differs significantly from previous VMDC releases in the following ways:

- **Placement**—In the Compute tier of the infrastructure, instead of the traditional aggregation layer
- **Form-Factor**—vApp, rather than physical
- **Application**—Dedicated per-tenant or organizational entity, rather than shared

These characteristics provide for a "pay as you grow" model with significant CAPEX savings in upfront deployment costs. In terms of redundance implications, dedication of resources to a specific tenant means that strict 1:1 redundancy may no longer be the default mode of operation for these forms of service nodes. Rather, administrators now has greater flexibility to fine-tune redundant services and methods for those tenants or organizations who have mission-critical applications with high availability requirements.

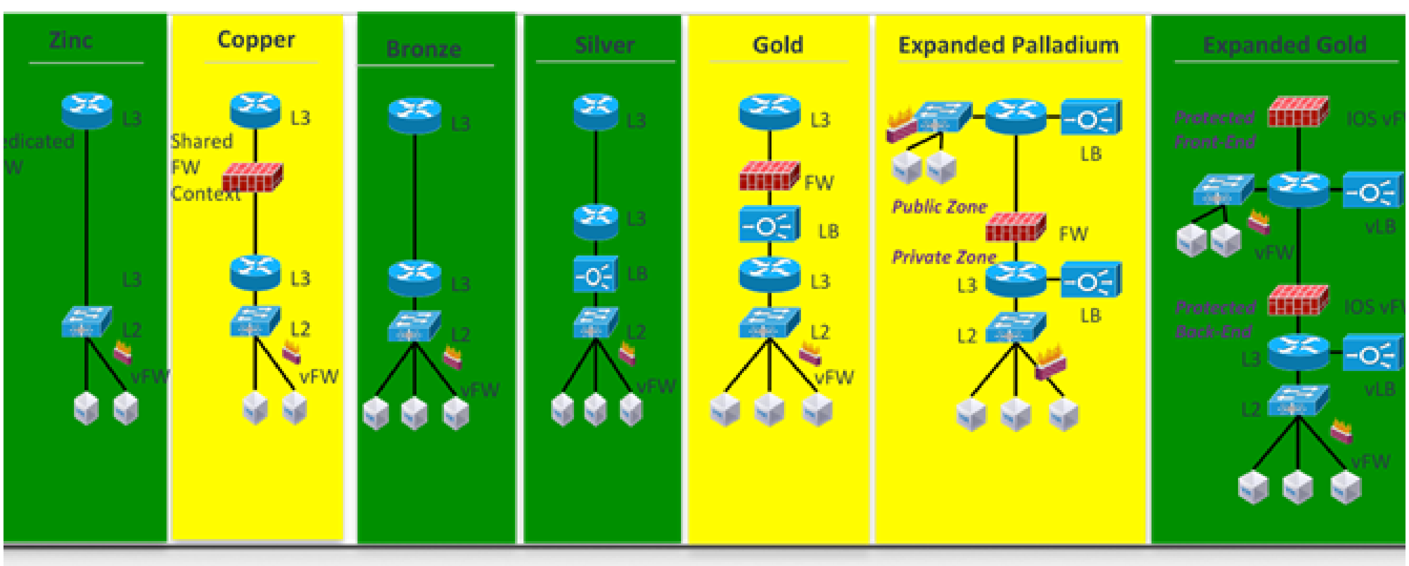
VMDC Virtualized Containers

The VMDC architecture can support multiple virtual containers, referred to as cloud consumer models. These models are described in greater detail later in this document, and in previous release material:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/2.2/collateral/vmdcConsumerModels.pdf

Because this release is based on unique, dedicated per-tenant security, load balancing and optimization services, for validation purposes VMDC VSA 1.0 focuses only on containers that do not feature shared (multi-tenant) security/services zones. High-level representations of these are highlighted in green in Figure 2-3. While the "Gold" container does not feature a shared zone, it is excluded because it is a subset of the "Expanded Gold" container.

Figure 2-3 VMDC Containers



As you move from left to right Figure 2-3, the validated VMDC VSA 1.0 containers, which are based upon real-world, commonly deployed N-tiered application and security models, become increasingly complex, growing from single to multiple security zones and policy enforcement points and from application of single to multiple types of services. VMDC VSA 1.0 features additional dedicated service options, such as network analysis and optimization. Although not shown in Figure 2-3, these are validated as part of the Expanded Gold container.

Solution Components

The following sections describe the network components used in the VMDC VSA 1.0 solutions (summarized in Table 2-1) and provide a snapshot of the intra-DC and overall system end-to-end network topology model validated in VMDC VSA 1.0 (Figure 2-4 and Figure 2-5).

Future VMDC releases will provide the opportunity to consider additional deployment model options featuring hybrid physical and virtual service form-factors.

Table 2-1 VMDC VSA 1.0 Solution Component Matrix

Function	Components
Network	Cisco ASR 9000, ASR 1000, ISRG2 3945, CSR
	Cisco Nexus 7009, 7004 (Nexus 7018 and Nexus 7010 not in SUT but valid architectural option)
	Sup2E, F2E and Sup2, F2 series 1 and 10 Gbps Ethernet cards
	Cisco Nexus 5548
	Cisco Nexus Fabric Extender 2248TPE
Services (vApp Form Factor)	Citrix NetScaler VPX Server Load Balancer
	Cisco vWAAS
	Cisco vNAM
	Cisco Nexus 1100 (Services Chassis)
Security Services (vApp Form Factor)	Cisco IOS XE 3.10 ZBF (for example, on CSR)
	Cisco ASA 1000V
	Virtual Security Gateway
Compute	Cisco Unified Computing System (UCS)
	Cisco UCS 6296UP Fabric Interconnect
	Cisco Fabric Extender 2208XP IO Module
	UCS 5108 Blade Server Chassis
	UCS B200/230/440-M2 and B200-M3 Blade Servers
	C200/240-M2/M3L Servers
	UCS M81KR Virtual Interface card
	UCS P81E Virtual Interface card
	UCS Virtual Interface card 1280, 1240
Virtualization	VMware vSphere
	VMware ESXi 5.1 Hypervisor
	Cisco Nexus 1000V (virtual access switch)
Storage Fabric*	Cisco MDS 9513
	(1/2/4/8 Gbps 24-Port FC Module; 18/4-Port Multiservice Module; Sup-2A; 24-port 8 Gbps FC Module; 18-port 4 Gbps FC Module)
	* *Not Applicable to this release.*

Table 2-1 VMDC VSA 1.0 Solution Component Matrix (continued)

Function	Components
Storage Array	NetApp FAS 6040[1]
Orchestration/ Management*	Domain Management: • UCS Manager • CIMC • Nexus 1000V Virtual Supervisor Module • Cisco Virtual Network Management Center • vWAAS Central Manager (vCM) • VMware vCenter 5.1 • Fabric Manager • Ontap 8.1.2 Service Assurance: * *CLSA VMDC VSA 1.0 not in scope* Orchestration: * *BMC CLM and CIAC not in scope*

1. Refer to NetApp storage array product family information: http://www.netapp.com/us/products/storage-systems/index.aspx

Figure 2-4 VMDC VSA 1.0 Intra-DC Topology

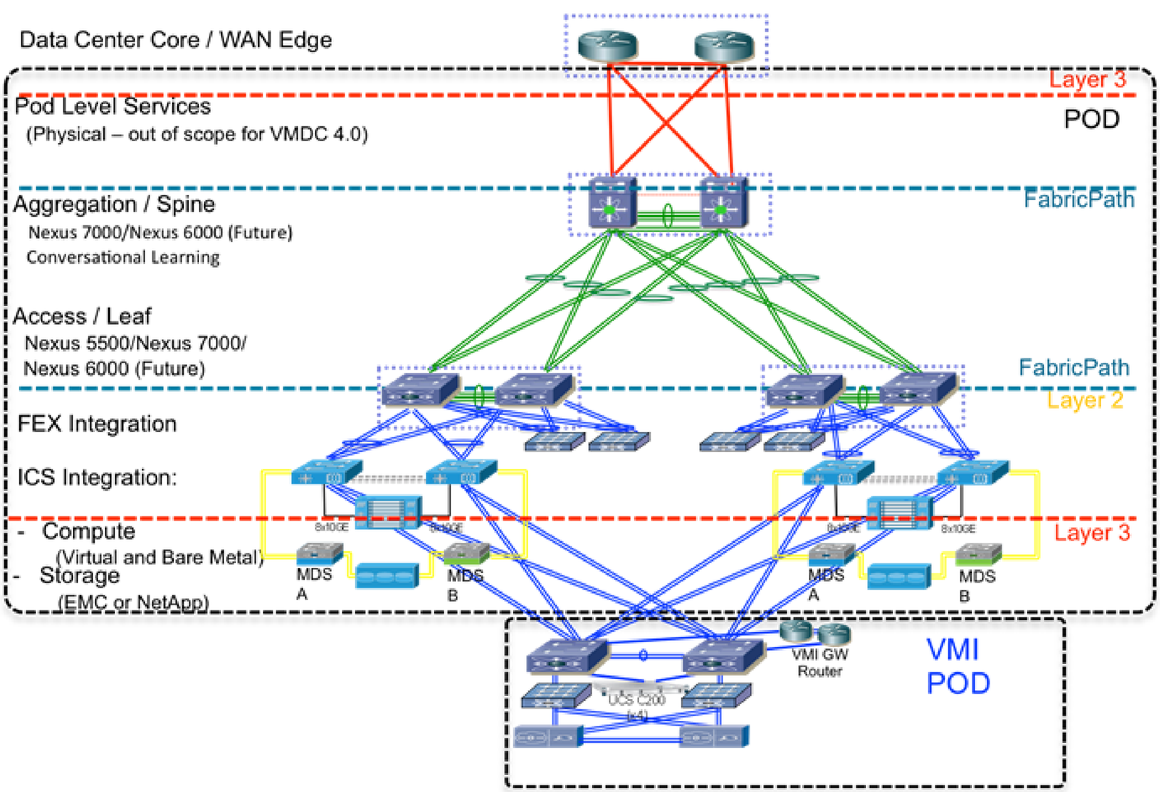
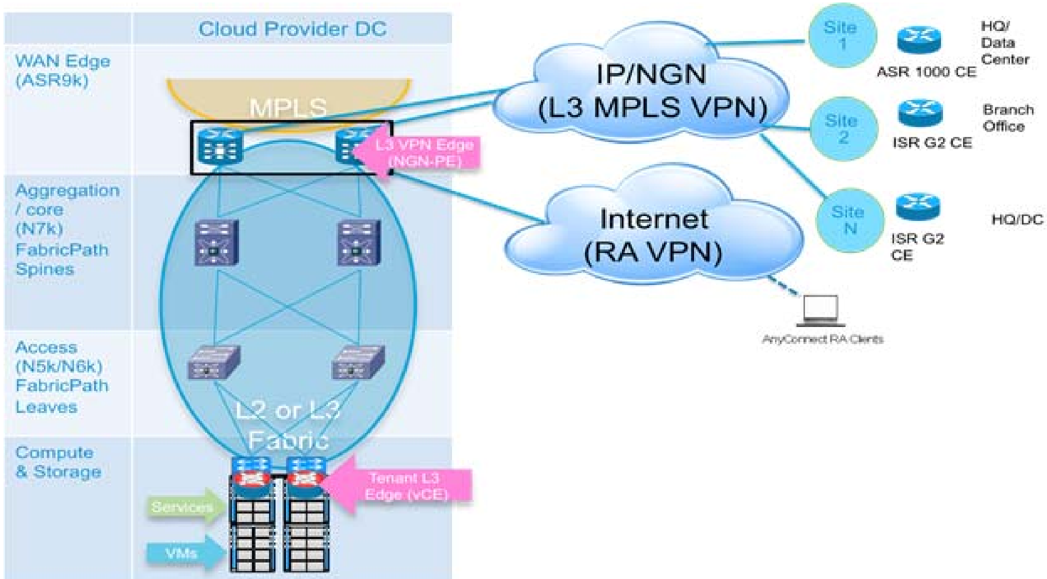
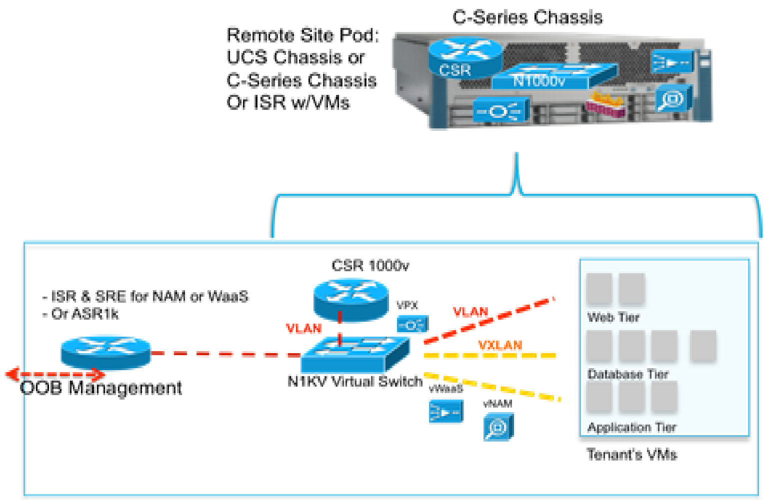


Figure 2-5 ***VMDC VSA 1.0 End-to-End Topology***



Though this document focuses mainly on the public cloud portion of the architecture, it is important to note that the virtual service model may easily be utilized in scaled down form at Enterprise remote sites to provide private cloud services as part of a Public Provider managed service offering. In this case, the remote sites in the preceding diagram in this context would be centrally controlled via out of band management paths. The private clouds can be tailored to fit application and services requirements, ranging in size from a Flexpod or Vblock to a small C-Series chassis "pod-in-a-box" entry point (Figure 2-6).

Figure 2-6 ***Remote Site Private Cloud***



VMDC Change Summary

The following release change summary is provided for clarity.

- **VMDC 1.0, 1.1**—Introduces architecture foundation for deploying virtualized and multi-tenanted data centers for cloud-based services. It supports high availability, elasticity, and resiliency of virtualized compute, network, and storage services.
- **VMDC 2.0**—Expands VMDC 1.1 by adding infrastructure orchestration capability using BMC software's Cloud Lifecycle Management, enhances network segmentation and host security, uses integrated compute stacks (ICS) as building blocks for the PoD, and validates compact and large PoD scale points.
- **VMDC 2.1**—Generalizes and simplifies VMDC 2.0 architecture for a multi-tenant virtualized data center used for private cloud. Improvements include multicast support, simplified network design, jumbo frame support, improved convergence, performance, scalability for private cloud, QoS best practices, and increased design flexibility with multi-tenant design options.
- **VMDC 2.2**—Builds on top of VMDC 2.0 and 2.1 for a common release supporting public, private, and hybrid cloud deployments. Enhancements include “defense in depth” security, multi-media QoS support, and Layer 2 (VPLS) based DCI.
- **VMDC 2.3**—Further expands on topology models in previous 2.X releases, providing a more collapsed architectural model, offering smaller footprint and entry point option. Enhancements include introduction of a new “copper” tenancy container mode.
- **VMDC 3.0/3.0.1**—Introduces FabricPath as an L2 multipathing technology alternative for the intra and inter-pod Data Center Unified Fabric infrastructure, considering the implications of various methods of appliance or service module-based service insertion.

Related Documents

The following documents are available for reference and consideration.

- Cisco Virtualized Multi-tenant Data Center Design and Implementation Guides, Releases 1.0-2.2

  http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns743/ns1050/landing_vmdc.html#~releases
- Design Considerations for Classical Ethernet Integration of the Cisco Nexus 7000 M1 and F1 Modules

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/2.6/vmdcm1f1wp.html
- Virtualized Multi-tenant Data Center New Technologies - VSG, Cisco Nexus 7000 F1 Line Cards, and Appliance-Based Services VPLS and EoMPLS Based DCI Solution with nV Edge and vPC

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/2.6/vmdctechwp.html
- Cisco VMDC 2.2 Design Guide

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/2.2/design_guide/vmdcDesign22.html
- VMDC 3.0.1 Fabric Path-based Design Guide

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/3.0.1/DG/VMDC_3.0.1_DG.html
- Data Center Interconnect over MPLS, Ethernet or IP Transport documents

http://www.cisco.com/en/US/netsol/ns749/networking_solutions_sub_program_home.html

http://www.cisco.com/en/US/netsol/ns975/index.html

- Cloud Service Assurance for VMDC

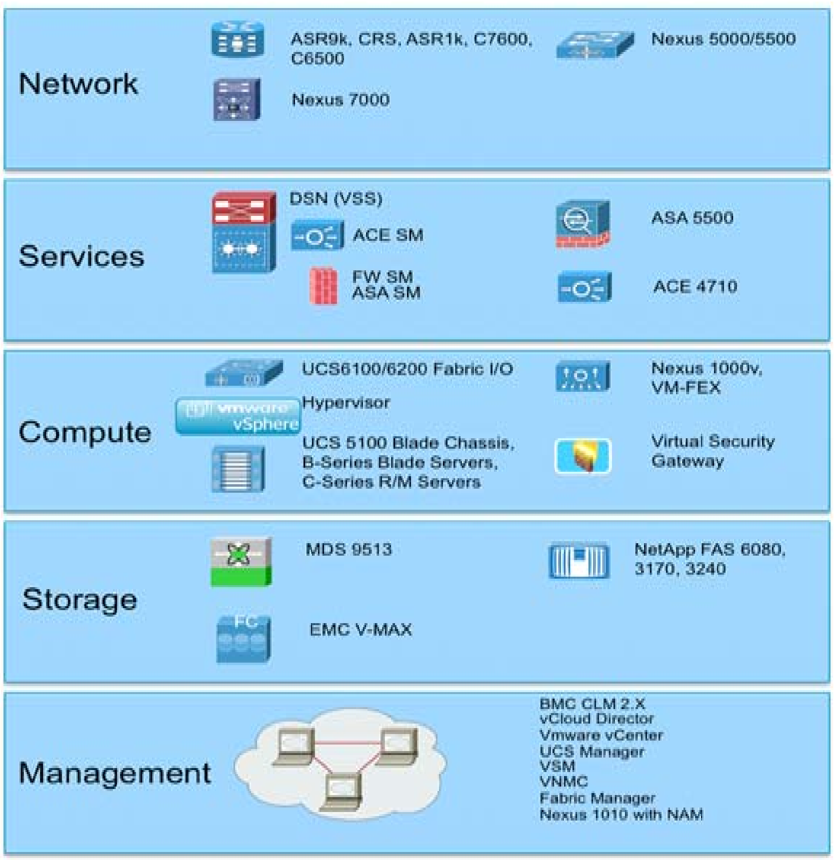
  http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/dz_cloudservice.html

CHAPTER 3

VMDC VSA 1.0 Design Details

Virtual Multiservice Data Center (VMDC) functional layers are shown in Figure 3-1.

Figure 3-1 Functional Layers Within the VMDC Data Center



Note Generally, the Services Functional Layer includes physical firewall and server load balancing (SLB) appliance or service module form factors. However, in VMDC VSA 1.0, this layer is limited to virtual appliance form factors.

VMDC Building Blocks

The following functional layers comprise the VMDC component building blocks:

Network Layer

The Network layer includes the WAN/provider edge (PE) router, which forms the data center perimeter to the enterprise area or service provider (SP) IP/NGN backbone, and to the public Internet. These perimeter nodes can be dedicated to Layer 3 (L3) routing functions, or can be multi-service in nature, providing L2 interconnects between data centers along with L3 services. WAN/PE routers validated in the VMDC reference system architecture include: Cisco CRS-1, Cisco ASR 9000, Cisco Catalyst 7600, Catalyst 6500, Cisco ASR 1000, and Cisco ISRG2.

The Network layer includes either a two-layer Clos spine and leaf arrangement of switching nodes, or the traditional three-layer hierarchical model described in previous (2.X) releases. While the Virtual Services Architecture (VSA) introduced in VMDC VSA 1.0 works with both models, in this release the Network layer comprises Nexus 7000 systems, serving as spine and aggregation-edge nodes, and Nexus 5000 or 7000 systems as leaf and access-edge nodes. As described in VMDC 3.0.1 Design Guide, validated VMDC 3.0 topologies feature several variants, enabling fine tuning of redundancy, port capacity, and bandwidth to the level of service aggregation or access density required by current and anticipated scale requirements.

VMDC VSA 1.0 introduces another network layer functional component, the Cloud Services Router (CSR) which serves as the L3 boundary and logical perimeter for the tenant Virtual Private Cloud container in the multi-tenant/shared cloud data center infrastructure. The CSR is a virtual router, so it resides in the compute tier of the infrastructure. Supporting multiple services, such as IOS zone-based firewalls (ZBFWs), IP security (IPsec) remote access virtual private network (VPN) termination and network address translation (NAT), the CSR provides the flexibility to add additional services without additional CAPEX.

Services Layer

The Services layer comprises network and security services, such as firewalls, SLB, Secure Sockets Layer (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: the solution set for the latter must support L4 - L7 services at a per-tenant level through logical abstraction of 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).

In previous VMDC reference architectures (2.X, 3.0), the Data Center Services Node (DSN) provides firewall and SLB services, in a service module form factor (for example, ACE30 and ASA-SM modules). Alternatively, these services are available in appliance form factors (ACE 4710, ASA 5500). This layer also serves as the termination point for remote access IPsec or SSL VPNs. In the VMDC architecture, the Cisco ASA 5580 appliance connected to the aggregation, aggregation-edge switching nodes or the DSN fulfills this function, securing remote tenant access to cloud resources.

In the all-virtual service scenario of VMDC VSA 1.0, these services and more are embedded in the virtual service subsystem of the Compute layer of the infrastructure.

Compute Layer

The Compute layer includes three subsystems: virtual access, virtual service, and compute. The first subsystem is a virtual access switching layer, which extends the L2 network across multiple physical compute systems. This virtual access switching layer is key because it also logically extends the L2 network to individual virtual machines (VMs) within physical servers. The feature-rich Cisco Nexus

1000V generally fulfills this role within the architecture. Depending upon the level of software functionality (such as quality of service (QoS) or security policy) or scale required, the Cisco VM Fabric Extender (VM-FEX) can serve as a hardware-based alternative to the Nexus 1000V.

A second subsystem is virtual services (vApp-based), which can include security, SLB, network analysis, and optimization services. Services implemented at this layer of the infrastructure complement more centralized service applications, and uniquely apply to a specific tenant or workgroup and their applications. Specific vApp-based services previously validated for the VMDC architecture include Cisco Virtual Security Gateway (VSG), providing a second security policy enforcement point within the tenant virtual data center or Virtual Private Cloud container. Additionally, in this release, IOS-XE ZBF features on the CSR or ASA 1000V provide perimeter firewalling; Citrix NetScaler VPX provides SLB; the CSR or VPX provide NAT; the CSR provides IPsec VPN termination; Virtual Network Analysis Module (vNAM) provides network analysis; and Virtual Wide Area Application Services (vWAAS) provides WAN optimization.

The third subsystem in the Compute layer is the computing resource. This subsystem includes physical servers, hypervisor software providing compute virtualization abilities, and the VMs. The Cisco Unified Computing System (UCS), featuring redundant 6100 or 6200 Fabric Interconnects, UCS 5108 Blade Chassis, and B-Series Blade or C-Series servers, comprise the compute resources in the VMDC reference architecture.

Storage Layer

The Storage layer provides storage resources. Data stores reside in a storage area network (SAN), which is block-based, or in network attached storage (NAS), which is file-based. SAN switching nodes provide an additional level of resiliency, interconnecting multiple SAN storage arrays to the compute resources over redundant FibreChannel or FibreChannel over Ethernet (FCoE) links.

Management Layer

The Management layer comprises the "back-end" hardware and software resources required to manage the multi-tenant infrastructure. These resources include domain element management systems and higher level service orchestration systems. The domain management systems currently validated within VMDC include Cisco UCS Manager, Cisco Integrated Management Controller, VMware vCenter, and vCloud Director for compute resource allocation; EMC UIM and Cisco Fabric Manager for storage administration; vWAAS Central Manager for traffic optimization services management; and Cisco VSM and Virtual Network Management Center (VNMC) for virtual access and virtual services management. Network Analysis Modules (NAMs), residing within Nexus 1010 systems or as vNAMs within the compute layer of the infrastructure, provide network analysis functionality.

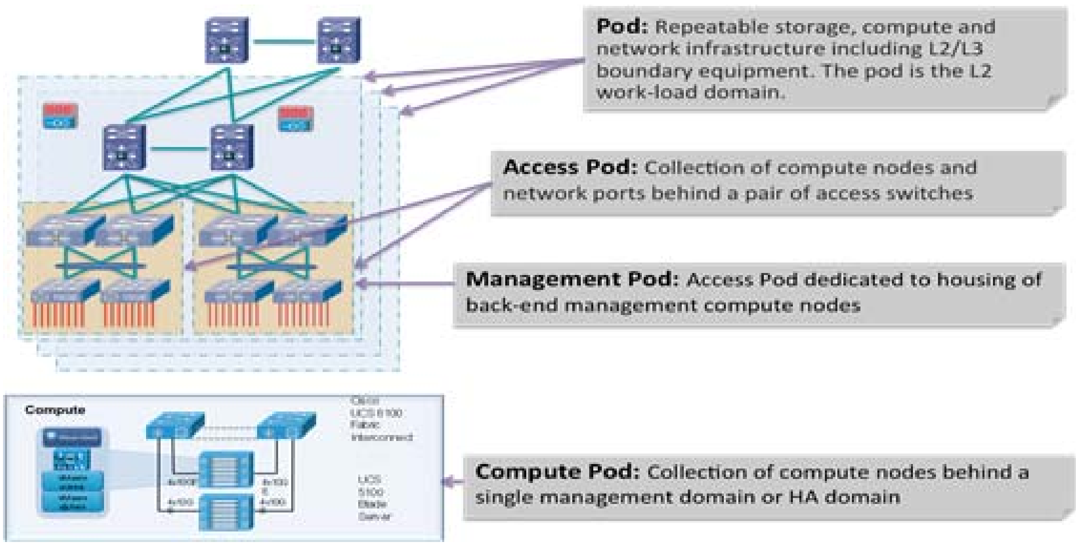
Note Also available and validated as Flexpod domain management components are the NetApp OnCommand Unified Manager and OnCommand System Manager software, NetApp VSC (Virtual Storage Console - a vCenter plug-in that provides end-to-end virtual machine (VM) monitoring, provisioning, B&R and management for VMware vSphere environments running on NetApp storage).

This layer can also include third party NetFlow collectors for aggregating and correlating network statistics. Automated service provisioning, including cross-resource service orchestration, is provided by BMC Cloud Lifecycle Management (CLM) or Cisco Intelligent Automation for Cloud (CIAC). Zenoss Cloud Service Assurance provides “Day 2” service impact visibility and root cause analysis tools. However, service orchestration and assurance solutions were not in scope for this VMDC system release.

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 (Figure 3-2). As a homogenous modular unit of network, compute, and storage resources, the pod concept addresses environmental, physical, logical, and application-level requirements in a consistent way. The pod serves as a blueprint for incremental build-out of cloud data centers in a structured fashion. When resource utilization within a pod reaches a predetermined threshold (for example, 70% to 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 3-2 Pod Concept



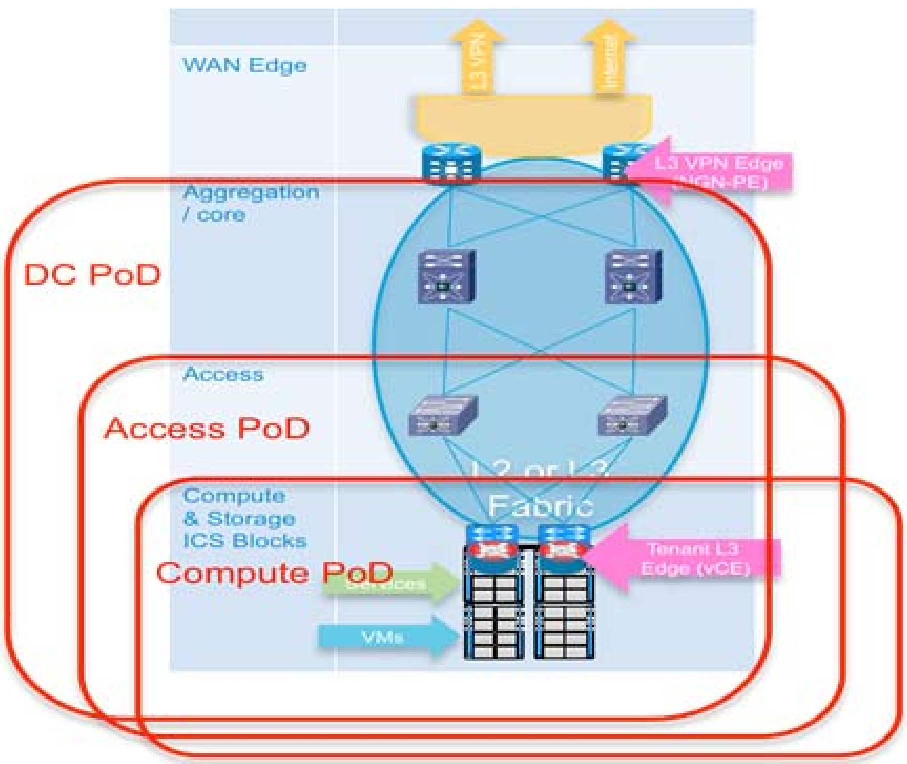
In practice, the pod concept can serve simply as a framework, with designers defining variants tuned to specific environmental or performance characteristics. A pod can be defined at different levels of modularity, supporting growth in differing increments. For example, one could have an access pod, terminating at access switching nodes within an infrastructure; and one could have a compute pod, addressing only the compute or the compute and storage portions of the infrastructure. Special-purpose pods can be defined around application requirements or operational functions. For example, in the VMDC reference architecture, a management pod, called a Virtual Management Infrastructure (VMI) pod, is defined to physically and logically separate back-end management resources from production resources.

Previously in the VMDC reference architecture, a general purpose utility compute pod extended from the compute and storage layers to the L2 ports on aggregation nodes serving as the L2/L3 boundary, up to and including components in the network services layer. In a traditional hierarchical topology model, the port and MAC address capacity of the aggregation nodes were key factors in determining scale. Port and MAC address capacity limited the number of pods that a pair of aggregation nodes could support in a cloud data center.

In contrast, a key benefit of a Clos-type architectural model is that it broadly expands overall port capacity and bandwidth within the L2 (pod) domain. However, where physical service appliances are attached at the FabricPath edge nodes, MAC address support on L3 aggregation-edge or access-edge nodes is a consideration in terms of host scale per pod (e.g., a pod comprising a single FabricPath domain). Furthermore, in the virtual customer edge (vCE) model that is the focus of VMDC VSA 1.0, the logical topology is modified to move the L3 boundaries to the centralized PE/WAN edge router and the per-tenant virtual CE routers in the compute layer. Similarly, service appliances move from the aggregation layer to the compute layer. In this case, one can consider the L3VPN gateway (PE routers) as a pod boundary.

Another option is to define a pod along access switch (leaf node) boundaries. Alternatively, one can define a compute pod, built along UCS system boundaries. In this release, because a tenant footprint is hosted across a number of Compute or Access switching systems, we depict a pod as extending from the compute layer across the entire data center FabricPath domain, up to and including trunks to ports on the PE/WAN edge routers (DC Pod, Figure 3-3).

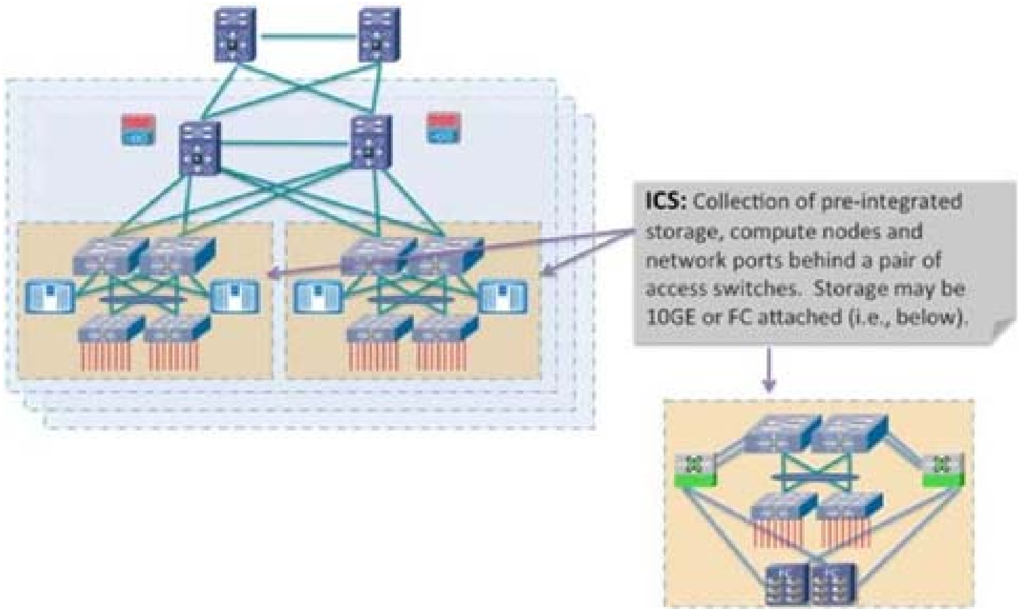
Figure 3-3 DC Pod in VMDC VSA 1.0



Integrated Compute Stacks

An Integrated Compute Stack (ICS) represents another potential unit of modularity in the VMDC cloud data center, representing a subcomponent within the pod. An ICS is an integrated collection of storage, compute, and network resources, up to and including L2 ports on a pair of access switching nodes. Figure 3-4 shows the location of the ICS in a pod. Multiple ICSs are deployed like building blocks to fill the capacity of a pod. This optimizes flexibility and allows Data Center Operations to incur CAPEX costs on a pay as you grow basis.

Figure 3-4 DC Pod in VMDC VSA 1.0



Working with ecosystem partners, Cisco currently supports two ICS options: Vblock and FlexPod.

- A Vblock comprises Cisco UCS and EMC storage systems, offered in several combinations to meet price, performance, and scale requirements.
- A FlexPod comprises UCS compute and NetApp storage resources. FlexPods are offered in a range of sizes designed to achieve specific workload requirements. A FlexPod can be scaled up or scaled out to host the entire workload for a particular pod. Using a FlexPod at the ICS layer provides the flexibility to scale the ICS layer to a Pod. FlexPods are integrated into ICS by attaching at the FabricPath access-edge nodes (for example, Nexus 5500 or Nexus 7000).

More information about Vblocks and FlexPod implementations is available here: http://www.cisco.com/en/US/netsol/ns1137/index.html

The VMDC reference architecture further accommodates generic compute and storage units, including storage from other third-party vendors. However, the business advantage of an ICS is that integration takes the guesswork out of balancing compute processing power with storage input/output operations per second (IOPS) to meet application performance requirements.

Data Center Interconnect

In the VMDC reference architecture, pods can be interconnected between data centers using various data center interconnection methods, such as Overlay Transport Virtualization (OTV), xPLS, or Locator/ID Separation Protocol (LISP). Though not in scope for VMDC VSA 1.0, these technologies have been tested and the resulting analysis is available in VMDC reference documents, Refer to http://www.cisco.com/en/US/netsol/ns975/index.html for details.

Unified Data Center Networking

Past descriptions of a *unified fabric* focused rather narrowly on storage transport technologies, such as FCoE. In a cloud architecture model such as VMDC, the concept of a unified fabric is one of virtualized data center resources (compute, application, storage) connected through a high-bandwidth network that is very scalable, high performing, and enables the convergence of multiple protocols onto a single physical network. In this context, the network is the unified fabric. FCoE, VM-FEX, vPCs and FabricPath are Ethernet technologies that have evolved data center fabric design options. These technologies can be used concurrently over the VMDC Nexus-based infrastructure.

Note FCoE uses FSPF (Fabric Shortest Path First) forwarding, which FabricPath does not yet support (FabricPath uses an IS-IS control plane). FCoE must be transported on separate (classical Ethernet) VLANs. In VMDC VSA 1.0, we assume that FCoE links are leveraged outside of the FabricPath domain—such as within the ICS portions of the FabricPath-based pod—to reduce cabling and adapter expenses and to realize power and space savings.

Compute

The VMDC compute architecture assumes, as a baseline premise, 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). Therefore, the architecture is based upon three key elements:

1. **Hypervisor-based Virtualization**—In VMDC VSA 1.0, as in previous VMDC releases, VMware vSphere plays a key role, logically abstracting the server environment in terms of CPU, memory, and network into multiple virtual software containers to enable VM creation on physical servers. In this release, vSphere VMs provide the foundation for router and service node virtualization.

   Note Separate, interrelated documents address Microsoft Hyper-V and Nexus 1000V integration for application workloads in VMDC FabricPath systems: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/Hyper-V/MS_Hyper_V.html

2. **UCS Network, Server, and I/O Resources in a Converged System**—UCS provides a highly resilient, low-latency unified fabric for integrating lossless 10 Gigabit Ethernet and FCoE functions using x86 server architectures. UCS provides a stateless compute environment that abstracts I/O resources, server personality, configuration, and connectivity to facilitate dynamic programmability. Hardware state abstraction simplifies moving applications and operating systems across server hardware.

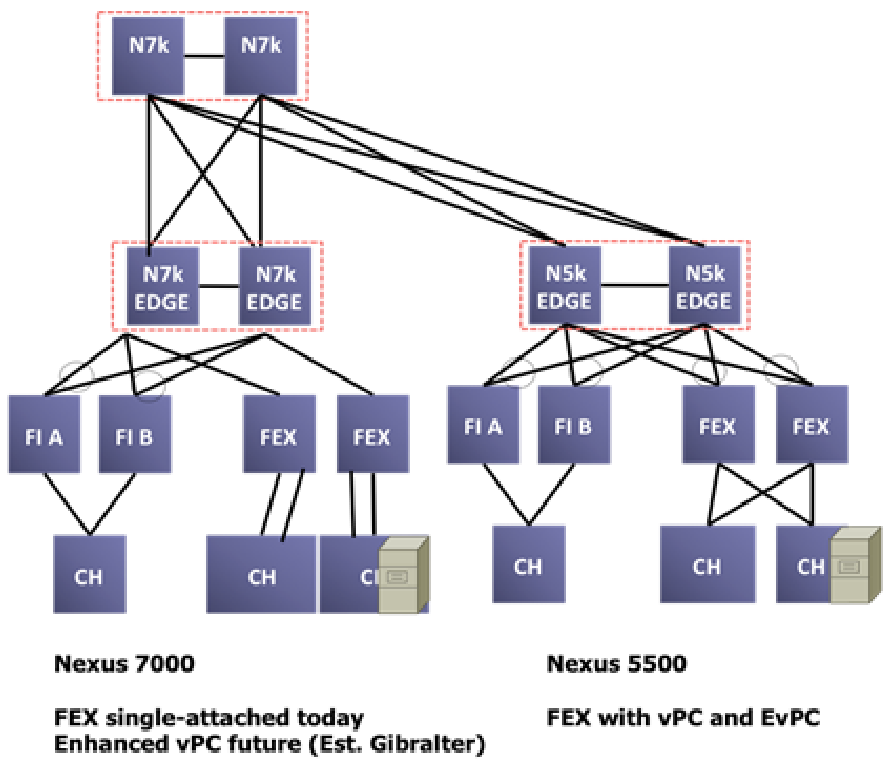
3. **The Nexus 1000V**—This virtual switch, which provides a feature-rich alternative to VMware Distributed Virtual Switch, incorporates software-based VN-link technology to extend network visibility, QoS, and security policy to the VM level. VMDC VSA 1.0 uses VMware vSphere 5.1 as the compute virtualization operating system. A complete list of new vSphere 5.1 enhancements is available online. Key "baseline" vSphere features leveraged by the system include ESXi boot from SAN, VMware High Availability (HA), and Distributed Resource Scheduler (DRS). Basic to the virtualized compute architecture is the notion of clusters; a cluster comprises two or more hosts with their associated resource pools, VMs, and data stores. Working with vCenter as a compute domain manager, vSphere 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 or DRS features are used. In practice, however, the larger the scale of the compute environment and the higher the virtualization (VM, network interface, and port) requirements, the more advisable it is to use smaller cluster sizes to optimize performance and virtual interface port scale and limit the intra-cluster failure domain. Previously in VMDC large pod simulations, cluster sizes were limited to eight servers; in smaller pod simulations, cluster sizes of 16 or 32 were used. For VMDC VSA 1.0, cluster sizes of 16 servers are deployed in the system under test (SUT). As in previous VMDC releases,

three compute profiles are created to represent large, medium, and small application workloads: “Large” has 1 vCPU/core and 16 GB RAM; “Medium” has 0.5 vCPU/core and 8 GB RAM; and “Small” has 0.25 vCPU/core and 4 GB of RAM.

The Nexus 1000V provides additional logical segmentation capabilities using VXLANs. A MAC-in-UDP encapsulation, VXLAN packets feature a 24-bit LAN segment identifier that significantly increases logical scale in the infrastructure. The Nexus 1000V performs VXLAN encapsulation and de-encapsulation, so VXLANs are transparent to infrastructure layers north of this virtual access edge device.

Finally, the compute layer of the infrastructure can include bare metal servers for applications that are unsuitable for virtualization. In VMDC VSA 1.0, bare metal servers are attached via 1 GE interfaces to FEX 2200s attached to Nexus 5500 or Nexus 7000 access-edge (leaf) nodes.

Figure 3-5 Bare Metal Server Placement



The UCS-based compute architecture has the following characteristics:

- It comprises multiple UCS 5100 Series chassis, each populated with eight half-width server blades.
- Each server has dual 10 GigE attachments – in other words, to redundant A and B sides of the internal UCS fabric.
- UCS is a fully redundant system, with two 2200 Series FEX per chassis and two 6200 Series Fabric Interconnects per system.
- Internally, eight uplinks per FEX feed into dual Fabric Interconnects to pre-stage the system for the maximum possible bandwidth per server. This configuration means that each server has 20 GigE bandwidth for server-to-server traffic in the UCS fabric.

- Each UCS 6200 Fabric Interconnect aggregates via redundant 10 GigE EtherChannel connections into the leaf or “access-edge” switch (Nexus 5500). The number of uplinks provisioned will depend upon traffic engineering requirements. For example, to provide an eight-chassis system with an 8:1 oversubscription ratio for internal fabric bandwidth to FabricPath aggregation-edge bandwidth, a total of 160 Gbps (16 x 10 Gbps) of uplink bandwidth capacity must be provided per UCS system.
- Four ports from an FC GEM in each 6200 Expansion Slot provide 8 Gbps Fibre Channel to the Cisco MDS 9513 SAN switches (for example, 6200 chassis A, 4 x 8 Gbps Fibre Channel to MDS A and 6200 chassis B, 4 x 8 Gbps Fibre Channel to MDS B). 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.

Storage

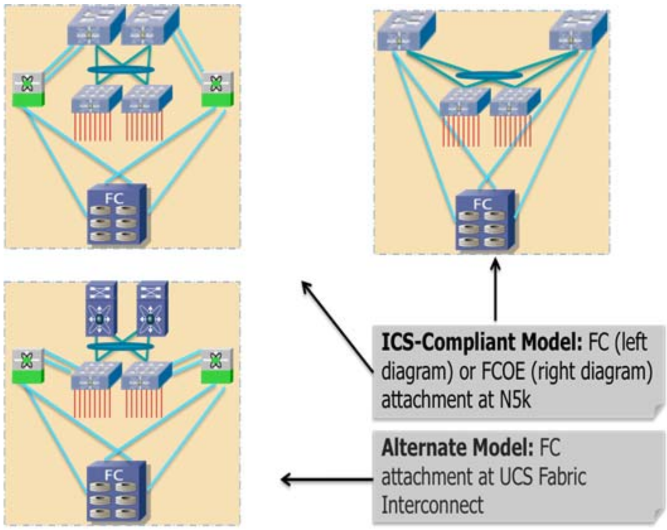
The VMDC SAN architecture remains unchanged from previous (2.0 and 3.0) programs. It follows current best practice guidelines for scalability, high availability, and traffic isolation. Key design aspects of the architecture include:

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

In terms of VMDC validation, the focus to date has been on storage as a distributed, pod-based resource. This is based on the premise that it is more efficient for performance and traffic flow optimization to locate data store resources as close to the tenant hosts and vApps as possible. In this context, we have the following methods of attaching FibreChannel storage components into the infrastructure as shown in Figure 3-6:

1. Models that follow the ICS model of attachment via Nexus 5000 and Nexus 7000, depending upon ICS type.
2. Models that provide for attachment at the UCS Fabric Interconnect.

Figure 3-6 SAN FC Attachment



In these scenarios, Cisco's unified fabric capabilities are leveraged with converged network adapters (CNAs) to provide "SAN-ready" servers, and NPV 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. Multiple FC links are used from each (redundant) Nexus 5000 or UCS Fabric Interconnect to the MDS SAN switches, to match the current maximum processing capability of the SAN and thus eliminate lack of bandwidth as a potential bottleneck between the SAN components and their point of attachment to the network infrastructure.

Similarly, for FCOE, multiple 10 GigE links provide resilience, and performance and cost efficiencies, by consolidating IP data, file and block traffic onto Ethernet. In this case, additional consolidation for smaller infrastructures may be attained by eliminating SAN switching systems, as illustrated.

Although Figure 3-6 shows a simplified SAN switching topology, it is important to note that if greater SAN port switching capacity is required, the architecture supports (and has been validated with) more complex, two-tier core-edge SAN topologies, as documented in the VMDC 2.0 "Compact Pod Implementation Guide," and more generally in Cisco SAN switching best practice guides, available at http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5990/white_paper_C11-515630.html.

NAS Architecture

The VMDC NAS architecture is FlexPod-aligned, following current best practice guidelines for scalability, HA, and traffic isolation. Key design aspects of this portion of the architecture include:

- Infrastructure resiliency through multi-level redundancy of field replaceable unit (FRU) components, multipath HA controller configurations, RAID-DP, and software enhancements that help with failures from a software perspective and a hardware perspective.
- Risk mitigation through fabric isolation and multi-level redundancy of connections (multiple fabrics, vPCs or port-channels, interface groups at the storage layer).
- vPCs address aggregate bandwidth, link, and device resiliency. UCS fabric interconnects and NetApp FAS controllers benefit from the Nexus vPC abstraction, gaining link and device resiliency, and full utilization of a nonblocking Ethernet fabric. From a storage perspective, both standard Link Aggregation Control Protocol (LACP) and the vPC link aggregation technologies play important roles in the FlexPod design.

- Network redundancy in clustered NetApp Data ONTAP is supported by the interconnect and switching fabrics, permitting cluster and data and management network interfaces to fail over to different nodes in the cluster, which extends beyond the HA pair.

For NAS connectivity, the FlexPod architecture leverages the Unified Target Adapter (UTA) and the traditional 10 GigE Ethernet adapter. UTA provides the greatest flexibility when migrating to an end-to-end FCoE design; however, a standard 10 GigE can be used for IP-based storage designs. The vPC links between the Nexus 5548 switches and NetApp storage controller UTAs are converged, supporting both FCoE and traditional Ethernet traffic at 10 Gbps and providing a robust connection between initiator and target. UTAs installed in each NetApp storage controller use FCoE to send and receive Fibre Channel traffic to and from the Nexus switches over 10 GigE. UCS also uses FCoE to send and receive Fibre Channel traffic to and from the various UCS components (for example, UCS B-Series blade servers and UCS C-Series servers). The system provides the option to leverage true end-to-end FCoE, which greatly simplifies network design and reduces application time to market.

Container Models

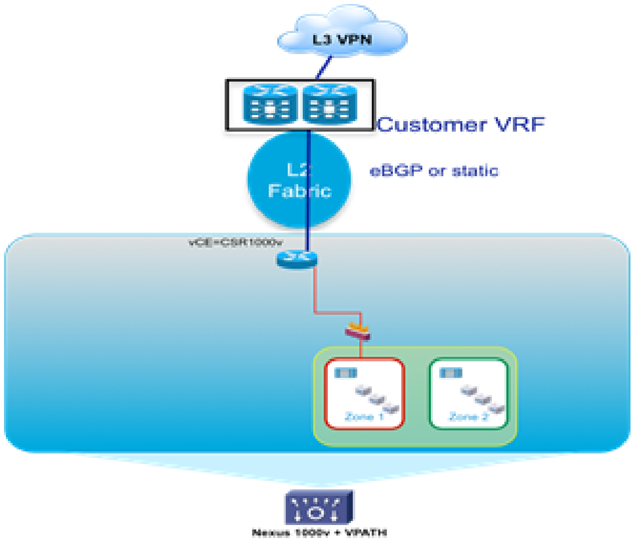
Virtualizing compute and storage resources enables sharing across an organizational entity. In contrast, virtualized multi-tenancy, a concept at the heart of the VMDC reference architecture, refers to the logical isolation of shared virtual compute, storage, and network resources. In essence, this is "bounded" or compartmentalized sharing. A tenant is a user community with some level of common security affinities. For example, in an enterprise, a tenant may be a business unit, department, or workgroup. Depending upon business requirements or regulatory policies, a tenant "container" may stretch across physical boundaries, organizational boundaries, and even between corporations. In large-scale environments, network function virtualization of tenant services provides considerable CAPEX cost savings, enabling a "pay as you grow" infrastructure model.

A tenant container can reside wholly in the private cloud, or can extend from the tenant enterprise to SP facilities in a public cloud. The VMDC architecture addresses these tenancy use cases through a combination of secured data path isolation and a tiered security model that leverages classical security best practices and updates them for the virtualized multitenant environment.

VMDC VSA 1.0 considers the following container models:

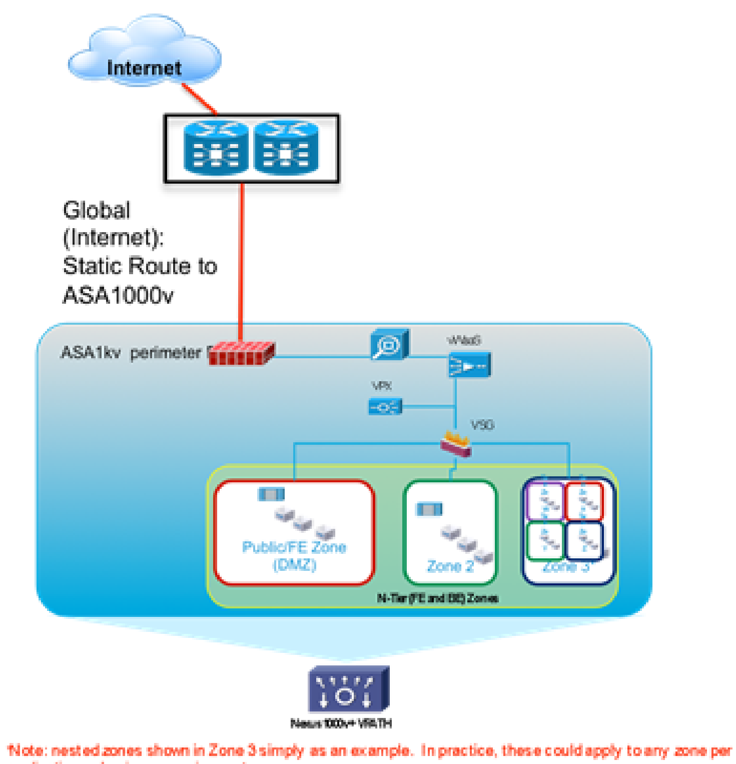
- **Bronze**—The most basic container type, a bronze container features a single logical segment for the attachment of hosts. Optionally, an L2 virtual firewall (for example, Cisco VSG) can be applied to provide security zoning. In VMDC VSA 1.0, CSR provides the L3 boundary, serving as the logical perimeter for this container, and as the default gateway.

Figure 3-7 Bronze Container



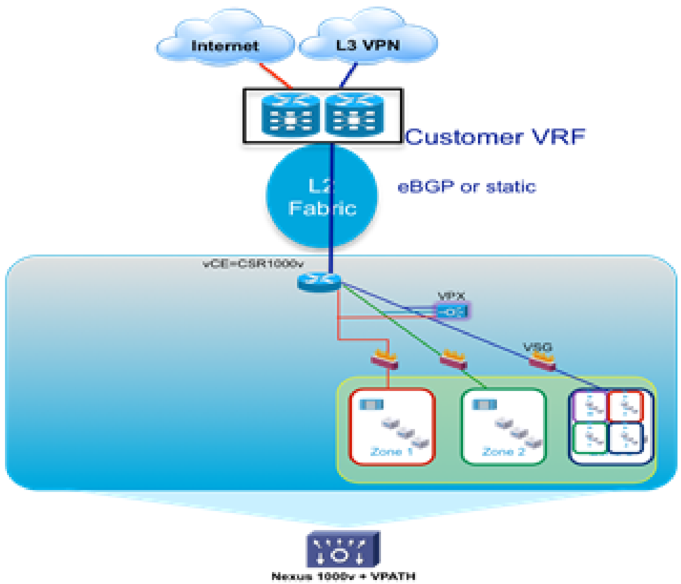
- **Zinc**—A new container in VMDC VSA 1.0, the Zinc container is similar to Bronze in that the zinc container is also a single-segment container. However, the logical perimeter and L3 boundary is the ASA 1000V virtual firewall. With only one "outside" and "inside" interface, a common deployment use case is expected to be protecting servers from client traffic originating from the public Internet. Again, VSG is shown as an optional second L2 policy enforcement point. Additional virtual optimization and network analysis appliances are also options.

Figure 3-8 Zinc Container



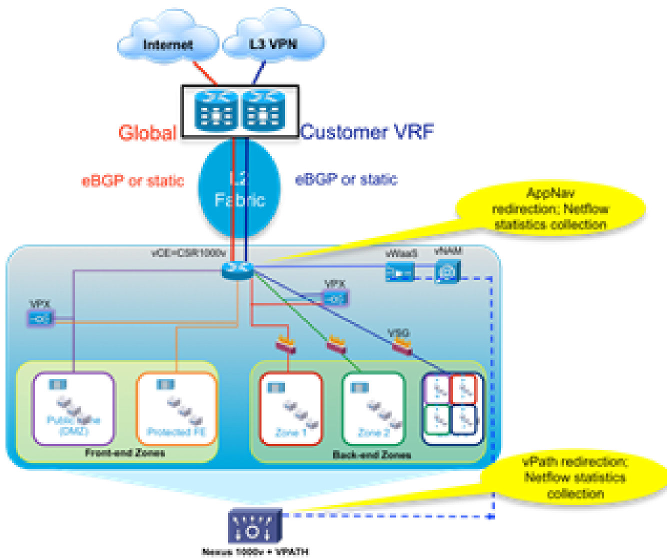
- **Silver**—The silver container expands services, featuring three logical segments and adding SLB. As in any container model, VSG can be added to provide additional zoning. As in the Bronze container, CSR provides the L3 boundary and default gateway.

Figure 3-9 Silver Container



- **Expanded Gold**—This container type is the most complex, providing more expansion of protected front-end and back-end zones while furthering the notion of separating public (Internet or demilitarized zone (DMZ)) or shared (campus/inter-organizational) access from private access. The expanded gold container type can include secured remote IPsec access. Note: the CSR does not support SSL remote access (RA) VPN termination as of this writing. 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 VPN, is commonly termed a virtual private data center (VPDC). Public cloud providers often use MPLS VPNs as transport for hybrid managed cloud services. Such services include IP addressing, security (firewalling, managed DMZ, zoning, and secure remote VPN access), and server resiliency solutions.

Figure 3-10 Expanded Gold Container



It is important to note that because the CSR supports multiple logical interfaces, any virtual containers featuring CSR as the L3 boundary support combined virtual and bare metal hosts, via VLAN stitching, or alternatively, via the VXLAN gateway on the Nexus 1000V.

Network

Network considerations are detailed in the following sections:

- Layer 3 Design, page 3-14
- Fabric Path, page 3-15

Layer 3 Design

In VMDC VSA 1.0, a combination of dynamic and static routing is used to communicate reachability information across the L3 portions of the Data Center infrastructure. In this design, dynamic routing is achieved using External Border Gateway Protocol (eBGP) from dedicated, per-tenant virtual routers (CSRs) functioning as vCE routers to redundant, centralized routers (ASR 9000s or ASR 1000s) functioning as PE routers.

Note: static routes could alternatively be configured for the vCE to PE paths. This may be an acceptable alternative from an operational standpoint if the routes will be configured using automation systems; otherwise manually maintaining static routes could present a challenge in highly scaled environments.

Depending upon the virtual private cloud container model, the CSR has either one (for example, Bronze, Silver) or two (for example, Expanded Gold) northbound interfaces to the PE router: one connects to the tenant private VRF and the second connects to the PE global routing table for routing over the Internet. Because the CSR supports IPsec VPN termination, encrypted IPsec client traffic from the Internet can be routed via the PE router to the CSR, where it is decrypted and routed to destination hosts in the container. For Zinc containers, in which the ASA 1000V is the logical L3 perimeter, static routes communicate reachability from and to the PE routers. In this model, WAN edge/PE routers effectively function as an L3 autonomous system boundary router (ASBR) and MPLS VPN gateway, extending the tenant virtual private cloud container in the public provider Data Center to their IP VPN.

The CSR1000V and ASA1000V are default gateways for all Hosts and routable virtual service appliances within the tenant containers. The ASR 9000 WAN/PE is the gateway to the Internet and private customer networks, for all devices in the data center. For the ASA 1000V in the Zinc container, the ASR9000 is the default gateway to the Internet, via static routing. For the CSR1000V in Silver/Bronze/Gold containers the ASR9000 is the gateway to the customer networks, which the ASR9000 advertises to the CSR1000v via eBGP. The ASR9000 can inject specific prefixes via BGP to the CSR for more granular control of tenant routing. For the CSR1000V in a Gold container with Internet access, the ASR9000 is the Internet gateway, and advertises a default route to the CSR1000V via eBGP on the Internet-facing link. The CSR does not have to learn all Internet routes, but can simply route traffic destined to the Internet toward the default route. Tenant-to-tenant communication may be enabled through leaking of VRF routes at the centralized PE.

Alternative L3 logical models for addressing tenancy scale not addressed in this system release include but are not limited to: 1) implementing MPLS Inter-AS Option B at the aggregation switching nodes, functioning as intra-DC PEs in a traditional hierarchical DC design, and 2) a distributed Virtual PE (vPE) model, described in BGP L3VPN Virtual PE Framework.

It is important to note that the vCE and vPE models are not necessarily mutually exclusive – it is possible that a Provider might run both models concurrently within a Public Data Center, to meet the differing needs of their customers. A practical use case which might lead a Provider to implement a vPE model over a vCE model is one in which the customer or “tenant” requires sub-tenancy – for example, the customer might be an ISV (Independent Software Vendor), and wish to use their slice of the Public Cloud to provide granular, differentiated services to their customers. Other practical deployment considerations include operational consistency and ease of use.

Fabric Path

Cisco FabricPath comprises an L2 data plane alternative to classical Ethernet. FabricPath encapsulates frames entering the fabric with a header that consists of routable source and destination addresses. These addresses are the address of the switch on which the frame was received and the address of the destination switch toward which the frame is heading. For this reason, switch IDs must be unique in the FabricPath domain; the IDs are either automatically assigned (default) or set manually by the administrator (recommended). The frame is routed until it reaches the remote switch, where it is de-encapsulated and delivered in its original Ethernet format.

FabricPath uses an IS-IS control plane to establish L2 adjacencies in the FabricPath core; so equal-cost multipath (ECMP) is supported and Spanning Tree Protocol (STP) is no longer required for loop avoidance in this type of L2 fabric. Loop mitigation is addressed using time to live (TTL), decremented at each switch hop to prevent looping and reverse path forwarding (RPF) checks for multi-destination traffic. As previously noted, a common initial use case for FabricPath is as part of a strategy to minimize reliance on STP in the Data Center.

A FabricPath domain comprises one logical topology. As part of establishing L2 adjacencies across the logical topology, FabricPath nodes create two multi-destination trees. IS-IS calculations compute the trees automatically. The highest priority switch is chosen as the root for the first multi-destination tree (FTAG1), which is used for broadcasts, flooding, and multicast. The second highest priority switch is chosen as the root for the second multi-destination tree (FTAG2), which is used for multicast. The designs described in this guide leverage the current best practice recommendation for root selection, which is to manually define the roots for the FTAG trees. In this case, the logical choice is to set the roots as the spine nodes, as they have the most direct connectivity across the span of leaf nodes. In the

Typical Data Center, there are only two spine nodes, so each serves as a root. In the Extended Switched Data Center, there are multiple spine nodes; two of the dedicated L2 spines serve as roots for the FabricPath domain. Should a root fail, the switch with the next highest priority takes over as root.

If devices that are part of non-FabricPath L2 domains (that is, spanning-tree dependent) are attached to FabricPath edge nodes using classical Ethernet, this design leverages the best practice recommendation to configure edge nodes as spanning tree roots, to avoid inadvertent blocking of redundant paths.

Additional key design aspects of the FabricPath portion of the Typical Data Center design as deployed in this release are summarized below:

- Two spine nodes, aggregating multiple leaf nodes (i.e., mirroring commonly-deployed hierarchical DC topologies).
- Leaf nodes (aka access-edge switches) and spine nodes (aka Aggregation-edge nodes) provide pure layer two functions, providing transit VLANs for vCE to WAN Edge/PE connectivity. This is in contrast to the Typical Data Center model as implemented in VMDC 3.0-3.0.1, where the Spine nodes performed routing functions.
- FabricPath core ports at the spine (F1s and/or F2/F2Es) provide bridging for East/West intra-VLAN traffic flows.
- Classical Ethernet edge ports face all hosts.

> **Note** A FabricPath core port faces the core of the fabric, always forwarding Ethernet frames encapsulated in a FabricPath header.

- L2 resilience design options in this infrastructure layer comprise using ECMP, port-channels between aggregation-edge and access-edge nodes across the FabricPath core; and VPC+ on edge nodes for the following options:
  1. Attaching servers with port-channels.
  2. Attaching other Classic Ethernet Switches in vPC mode.
  3. Attaching FEX in Active/Active mode.

Currently, the Nexus 7000 supports three types of FabricPath I/O modules: N7K-F132XP-15 (NX-OS 5.1); N7K-F248XP-25 (NX-OS 6.0); and the new N7k-F248XP-25E (NX-OS 6.1). These can be used for FabricPath core ports. However, the F1 card supports only L2 forwarding, while the F2 and F2E cards support L2 and L3 forwarding.

F2 or F2E-only scenarios (that is, performing L2 and L3 forwarding, as in VMDC 3.0.1) also provide benefits in terms of ease of deployment, and lower power consumption, but as of this writing, the 16,000 maximum MAC address constraint applies to this model.

With respect to access-edge (leaf) nodes in the referenced models, Nexus 5548 (or Nexus 5596) having FEX 2200s for port expansion provide TOR access. Alternatively, Nexus 7000s having F1 (or F2) line cards (and 2232 FEX-based port expansion) can perform this function, for end-of-row (EOR) fabric access.

The Nexus 5500 supports up to 24 FEX modules. If using the Nexus 2232PP this would support. 768 edge ports per Nexus 5500 edge pair. Traffic oversubscription can be greatly impacted with increased FEX usage. Currently, four FabricPath core facing port-channels having four members each are supported on the Nexus 5500.

Currently, 6200 Series Fabric Interconnects connect to FabricPath edge nodes using vPC host mode (vPC-HM). FabricPath is on the roadmap but beyond the scope of this release.

vPC+ is one of the new L2 resilience features introduced with FabricPath. This enables devices that do not support FabricPath to be attached redundantly to two separate FabricPath switches without resorting to SPT. Like vPC, vPC+ relies on port-channel technology to provide multipathing and redundancy. Configuring a pair of vPC+ edge nodes creates an emulated FabricPath switch ID for the pair. Packets originated by either vPC+ node are sourced with this emulated switch ID. Other FabricPath switches simply see the emulated switch ID as reachable through both switches. Prerequisites include direct connection via peer-link, and peer-keepalive path between the two switches forming the vPC+ pair.

Port-channels, rather than single links, are used with ECMP for access-edge to aggregation-edge core connections, providing enhanced resilience if a link member fails. As this is not default behavior after NX-OS 5.2.4, an IS-IS metric must be configured on the port-channel to ensure that individual member link failures in port-channels are transparent to the IS-IS protocol.

Virtualization Techniques

Previous program releases leveraged VMware vSphere 5.0, 4.0 and 4.1. vSphere 5.1 is the tenant hypervisor resource used in VMDC VSA 1.0. This integrates with Cisco's Nexus 1000V distributed virtual switch, enabling end to end visibility to the hypervisor level for security, prioritization, and virtual services.

Though not in the scope of VMDC VSA 1.0, alternate hypervisors can be used in VMDC reference architectures if UCS is in their prospective Hardware Compatibility List. As of this writing, the Nexus 1000V distributed virtual switch supports only vSphere and Hyper-V. However, alternate hypervisor VMs can connect at the FEX or primary access layer, and participate in appliance-based or Data Center Services Node (DSN) module-based services.

Services

Previous VMDC releases incorporated physical appliance-based and DSN module-based services, and virtual service appliance form factors. From VMDC 2.2 forward, two tiers of security policy enforcement points are featured in the enterprise-grade Expanded Gold container: the first perimeter firewall implemented on a physical form factor, and the second (VSG) implemented as a virtual appliance. The premise was that this hybrid model would best satisfy rigorous security requirements. As is traditional, with the exception of the VMDC 3.0 "Switched Data Center" FabricPath topology model, all physical form factors were attached at the aggregation or aggregation-edge nodes.

VMDC VSA 1.0 departs from tradition in that all IaaS network service functions are virtualized. In this model, services are attached via VLAN stitching at the virtual access edge in the compute layer of the infrastructure. The list of virtual service appliances includes: CSR; Citrix NetScaler VPX for SLB; ASA 1000V; VSG; Virtual Network Analysis Module (vNAM); and the Virtual WAN Acceleration Service Module (vWAAS). Running on general-purpose server hardware, these software-based form factors are ideal for cloud data centers in that they are software-defined and provide flexibility and agility through enhanced programmability.

CSR

Discussed at length in an earlier white paper (*VMDC Virtual Service Architecture with CSR*), the CSR is an x86-based virtual router based on the ASR 1000 product family. Although the ASR 1000 features optimized ASIC-based forwarding, CSR forwarding is software- based. However, the CSR is extremely feature-rich, inheriting much of the ASR 1000 functionality as it leverages IOS-XE (XE3.10 as of this writing). CSR currently offers a maximum forwarding rate of 1 Gbps, and fixed licensing packages presently rate-limit performance to the following throughput options: 250 Mbps, 100 Mbps, 50 Mbps, 25 Mbps, and 10 Mbps. These packages position CSR as the virtual private Cloud perimeter router solution, where routing throughput requirements generally range from 10Mbps to 1Gbps. As of XE3.10, for the 50 Mbps throughput option, resources required to host the CSR are: 1 vCPU, 2.5G RAM, 8G HD. Note: process-intensive functions may require additional DRAM. Check the release notes for more detailed considerations: http://www.cisco.com/en/US/products/ps12559/prod_release_notes_list.html

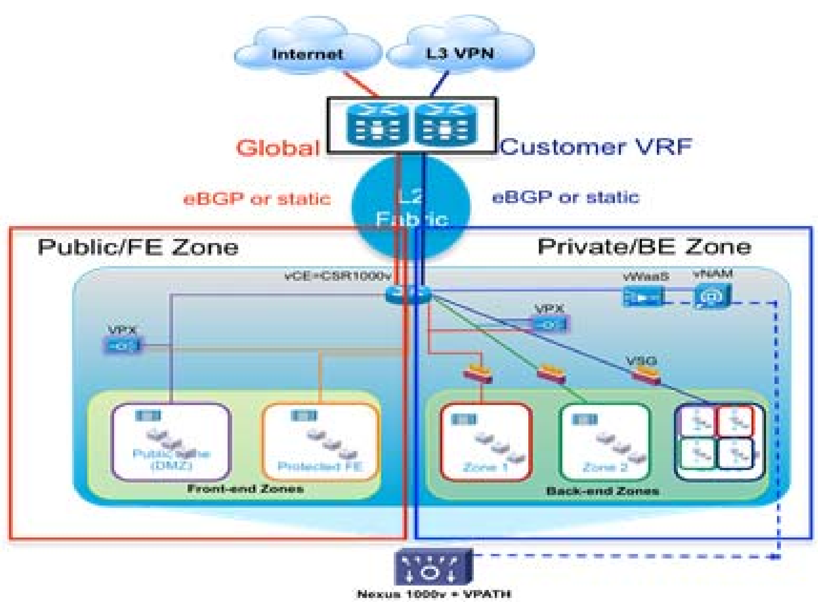
The CSR can provide the following services in the VMDC VSA 1.0:

Virtual Routing—The CSR is implemented in this release as a virtual CE for routing tenant traffic to the Internet or the tenant IP MPLS VPN via the PE L3 gateway. The CSR routes IPv4 and IPv6 packets and is also the L3 (default) gateway for hosts in the logical virtual private cloud container.

It is important to note that the CSR supports full MPLS functionality, although this is not a focus of VMDC VSA 1.0. The CSR provides the flexibility to support alternative logical models for scaling multi-tenancy, such as the virtual PE model.

IOS XE ZBF (Zone Based Firewall)—Collapsing perimeter firewall policy enforcement onto the virtual router appliance provides the opportunity to simplify the virtual private cloud container, reducing CAPEX and OPEX costs. Stateful ZBFs are implemented on logical interfaces (Figure 3-11). By default, only interfaces belonging to the same zone can communicate. Zone pairs must be defined to enable inter-zone communication. ZBFs are supported for IPv4 and IPv6 packets. The vCE model facilitates this use case; fine-grained firewalling is possible because the CSR routes only IP-encapsulated packets. In this example, CSR implements a front-end zone, including all applicable downstream and upstream logical segments, to securely separate public and Internet traffic from devices and logical segments participating in the private, back-end zone.

Figure 3-11 Public and Private IOS Zone-Based Firewall on CSR



IPsec VPN gateway—CSR provides route-based IPsec VPNs, terminating and decrypting IPsec VPN tunnels for secure remote access to resources in the virtual private cloud container via the Internet.

Traffic control and visibility point—CSR provides instrumentation for high-touch application visibility and control with features such as Performance Agent for round-trip response time statistics collection, AppNav traffic redirection (for example, to performance optimization service appliances), Switched Port Analyzer (SPAN), NetFlow, QoS, NAT and Dynamic Host Configuration Protocol (DHCP).

Should redundant CSRs be required, Hot Standby Router Protocol (HSRP) can be used to provide resiliency between CSR pairs. In this case, it is actually the HSRP VIP interface that would be the default gateway for hosts within the container. HSRP route tracking can be defined to insure symmetric traffic flows through each CSR.

Citrix NetScaler VPX SLB

The Citrix NetScaler VPX virtual service appliance performs SLB and SSL offload services in the VMDC VSA 1.0 architecture. As of this writing, the VPX is available in four models, ranging from 200 Mbps to 3 Gbps maximum throughput, suiting a broad range of performance requirements and use cases. This release leverages the 200 Mbps (VPX-200) model. Supported hypervisors as of this writing are: vSphere ESXi, Microsoft Hyper-V, and XenServer. This release is based on the vSphere ESXi hypervisor. The number of logical network interfaces supported by the VPX is determined by hypervisor limits. Currently, for vSphere 5.1 and ESXi hardware version VMX-09, this is a maximum of 10. VPX supports IPv4 and IPv6 packets, and can operate in transparent or routed mode. Required VPX-200 resources are two vCPUs, 2 GB RAM, and 20 GB HD.

Note Available shortly as an alternative SLB is the Cisco Netscaler 1000v. Also available in a variety of license bundles, this version of the Netscaler load balancer differs from the Citrix VPX in the following ways:

- Sold and supported directly by Cisco, providing administrative benefits of a single point of contact.
- Integration into the Nexus 1000v service insertion technology, providing consistent operational experience and flexible service delivery.

More details regarding the Netscaler 1000v are available at the following links:

http://cisco.com/go/ns1000v and

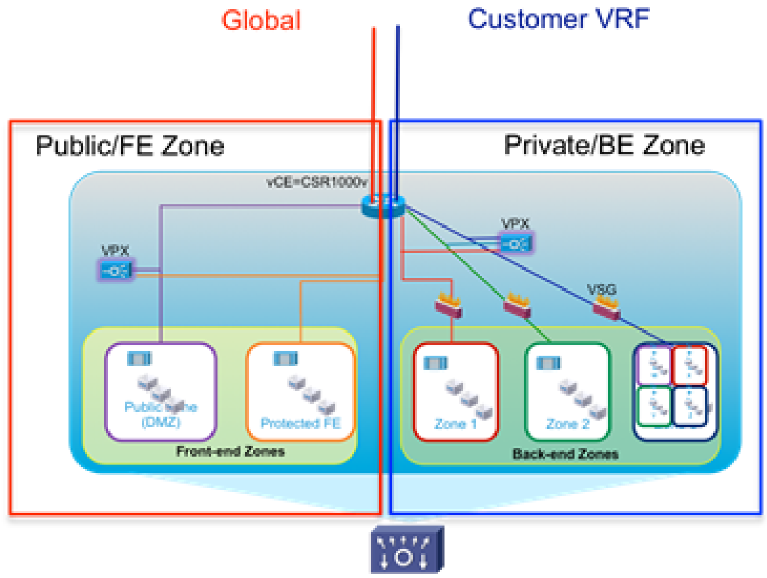
http://www.cisco.com/en/US/products/ps13296/index.html

In this release we focus mainly on the VPX load balancing and resilience capabilities, however the VPX is quite feature-rich, supporting a broad range of use cases and functionality. The VPX may be installed from an OVF and configured via CLI, however further enhancing usability and ease of configuration is the browser-based VPX GUI. More detailed information about the VPX is available online.

Characteristics of the VPX virtual service appliance-based service attachment implemented in this release include:

- Virtual network interface card (vNIC)-based attachment at the DVS or Nexus 1000V virtual access edge switch.
- VPX instance per front-end and back-end zone (zone in Expanded Gold Container).

Figure 3-12 VPX per Zone in Expanded Gold Container



- In Figure 3-12, the CSR creates two firewall zones – a front-end zone, for hosts that can be accessed from clients in the public Internet (in this example, on two subnets), and a back-end zone, for hosts that can be accessed only using the tenant private MPLS VPN. Rather than have a single VPX serving both public and private zones (e.g., with an interface in each set of zones), in this case two VPX instances are used – one in the front-end zone and another in the back-end zone. This reinforces security policy enforcement, insuring that there is no chance of “back-door” access from the public to the private zones.
- In Figure 3-12, each VPX is in multi-subnet “one-arm” mode featuring L2 adjacency to server subnets in order to optimize traffic flows for load balanced and non-load balanced traffic. In contrast, an alternative option would be to use a single vNIC connection to the CSR – another one-arm, multi-subnet implementation, which has the benefit of reducing the number of vNICs required on the VPX. Another important benefit of note is that from an automation perspective, this alternative may be somewhat simpler to orchestrate in terms of adding load balanced subnets, with minimal service impact. However in this case traffic flows and performance would be sub-optimal, as all load-balanced traffic from both hosts and clients would first need to transit through the CSR. Both options will work, however in this release we focused on the illustrated model for end-to-end system validation purposes.
- The CSR is the default gateway, so that all load-balanced traffic is properly routed on to either the Internet or the tenant MPLS VPNs.
- Incoming client traffic accesses the Virtual IP (VIP) address of the VPX. One-arm mode deployments require source-NATing of client requests and server responses to insure symmetry.
- Though not illustrated, a separate subnet having a NetScaler IP (NSIP) address is configured on the VPX to transport back-end management traffic. The NSIP is the IP address utilized for management and general system access to the VPX itself, and for HA communication. As baseline parameters, initial instantiation of a VPX instance simply requires definition of the NSIP, mask and default route.
- Additionally, a Subnet IP Address (SNIP) is defined per load-balanced server subnet, in order to bind these interfaces for server communication.

- For HA scenarios, two redundancy options are available: Active/standby failover between redundant VPX pairs, or clustering. It is important to note that load balancing distribution across multiple VPX appliances is supported only in the clustered case. Given that a virtual appliance is a dedicated rather than shared resource, and that the failure domain is thus minimized, in this release we focused on active/standby failover as the most applicable use case. Setting up HA pairs is fairly simple: one assigns a unique node ID number to the primary and secondary nodes, and points each node to the NSIP (management interface) address of the other node in the pair. In HA mode, heartbeat packets are sent on all active interfaces, eliminating the need for a dedicated peer link between primary and secondary systems. Failover from a primary to a secondary occurs when the dead-interval timer is exceeded, at which time connections are reestablished on the new primary VPX instance. Note: in practice it may also be useful to define a SNIP on the NSIP (management) subnet, in order to allow continued communication with the primary VPX appliance, regardless of whether it is in active or standby state.

Direct Server Return (DSR), also known as “direct routing”, “nPath” or “SwitchBack” is another possible mode of load balancer operation that offers the following benefits versus one-arm mode:

- **Preservation of client source addresses** (e.g., SNAT loses them).
- **Performance**—In many cases, inbound client traffic is typically much smaller than outbound traffic (e.g., 1:8 for Yahoo, per NANOG 2010 reports). In DSR, the load balancer only handles inbound packets, as servers respond directly to clients, bypassing the load balancer. Thus this mode of operation may offer better performance than one-arm mode.

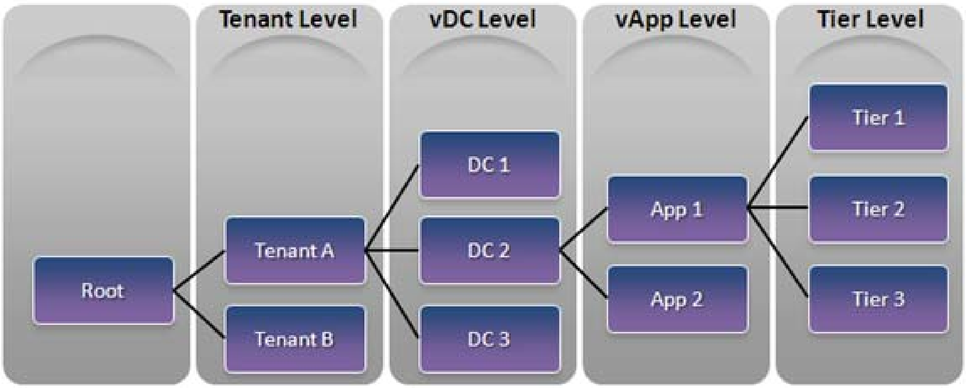
Some limitations of DSR (in layer 2 mode) are that PAT is not possible and servers cannot respond directly to ARP requests for the VIP (e.g., non-ARPing loopback interfaces must be configured on the servers).

ASA 1000V

If only perimeter firewalling is required, without multiple inside and outside interfaces, dynamic routing or other multi-service L3 features, the ASA 1000V provides an alternative to CSR. Like VSG, the ASA 1000V is integrated with the Nexus 1000V DVS, leveraging vPath for service chaining and fast-path traffic offload, and presently supporting up to a maximum 500 Mbps throughput.

Each ASA 1000V instance is installed as a virtual machine with the following characteristics: 1 vCPU at 1 GHz; 1.5 GB vRAM; and 2.5 GB vHD. Four interfaces are provided per virtual appliance: one management, one failover, and two for data (for example, one “inside” protected and one “outside” interface). As with VSG, VNMC provides hierarchical, policy-driven domain management.

Figure 3-13 VNMC Hierarchical Policy



In cases in which multiple tiers of policy enforcement are required, the ASA 1000V may be combined with the VSG, the latter serving to provide additional security zoning within the tenancy container – i.e., below the “inside” interface of the ASA 1000V. This is illustrated in Figure 3-8.

Unlike the physical ASA appliance, the ASA 1000V supports only active/standby failover. To maximize availability, active and standby systems should be placed on separate server blades. Heartbeats are exchanged between the failover pair over a failover link. When a failure is detected, the newly active ASA 1000V accepts all traffic destined for the ASA 1000V. Because of the failover link, the active ASA 1000V already has connection state information for connections that were active before failover. Because only one context exists, preemption is not supported.

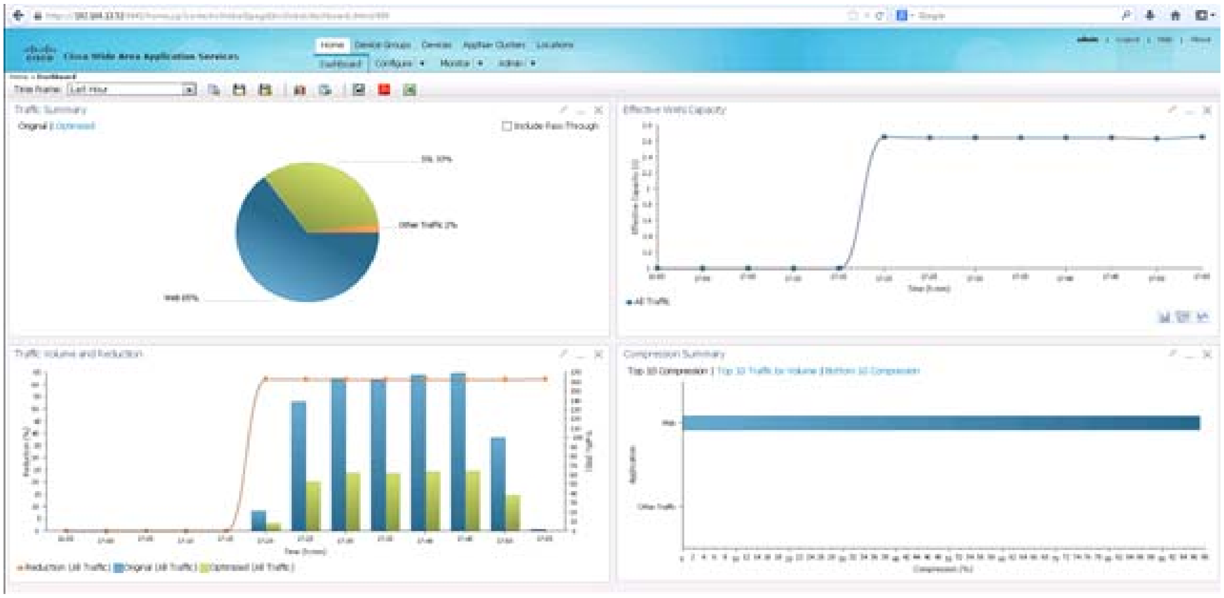
Cisco vWAAS

vWAAS is a key component of Cisco’s AVC (Application, Visibility and Control) product portfolio. Available in numerous form factors sized to fit a wide range of customer deployment requirements, from branch to HQ to data center, vWAAS provides end-to-end application performance optimization to improve response times and minimize the negative impacts of latency and limited bandwidth.

vWAAS 5.2 features enhanced integration with the Nexus 1000V, supporting vPath co-residency with virtual service appliances such as the VSG. In the VMDC VSA 1.0 architecture, release 5.2 version of the vWAAS-750 is used in tenant virtual private clouds with a router-integrated form-factor, the Cisco Services-Ready Engine module, on ISR G2s at remote customer premises. Resource requirements for vWAAS-750 are: 2 vCPU, 4 GB RAM, and 250 GB HD. In the data center Virtual Management Infrastructure (VMI), vWAAS Central Manager (vCM) provides domain management support for 100 to 2000 devices. Out of band (OOB) management over the WAN to the vCM enables management of remote WAAS devices.

In the data center Virtual Management Infrastructure (VMI), vWAAS Central Manager (vCM) provides domain management support for 100 to 2000 devices. Out of band (OOB) management over the WAN to vCM enables management of remote WAAS devices. Figure 3-14 is an example of the application traffic and optimization visibility provided by the system.

Figure 3-14 Sample vWAAS Application Reports



Traffic must be redirected from networking devices to the vWAAS for optimization; there are multiple ways to do so, including policy based routing and Web Cache Coordination Protocol (WCCP). In this system we focused on validation of AppNav redirection from the CSR.

Introduced on CSR in IOS XE 3.8, AppNav provides key benefits over previous technologies such as WCCP, providing a way to scale traffic redirection and improve performance and lower CPU overhead using the following characteristics and techniques:

- Scaling redirection through decoupling flow distribution from flow processing; in AppNav deployments, a flow distribution unit, called the AppNav Controller, and multiple service nodes (1-32) process flows. In VMDC VSA 1.0, CSR functions as an AppNav Controller, and vWAAS functions as a service node. As noted, code prerequisites are IOS-XE 3.8 (CSR) and 5.1 (vWAAS).
- Intelligent redirection of new flows based on the load on each service node.
- Bypass of flows that do not require optimization; service nodes can inform the AppNav Controller to directly pass through non-optimized traffic, minimizing latency and resource utilization.
- Ability to add or remove a service node with minimal impact to traffic.
- For special applications (for example, Messaging API (MAPI)/Exchange and Citrix VDI), ensures that a family of flows is redirected to the same service node for optimal treatment.

vPath redirection from the Nexus 1000v DVS is another option. In this case interception and redirection is based on the Nexus 1000V port-profile configuration. This is a MAC-in-MAC redirection technique, which requires vWAAS to be L2-adjacent to the host toward which traffic is destined (it need not be located on the same ESXi host). In this case, vPath interception is configured, on the port profile of the VM server in both directions, to redirect VM server packets to vWAAS. vPath redirects are transported over the Nexus 1000V service VLAN. vWAAS receives the vPath intercepted packet, performs WAN optimization, and returns the response packet to VEM. vWAAS egress traffic received by VEM is forwarded without additional vPath interception. Management packets are not vPath-encapsulated. The key benefits of vPATH interception are:

- No need to define the direction of interception (in or out)—vPath maintains a flow entry table for each TCP flow that is used to intercept and redirect traffic.

- Automatic bypass of pass-through traffic; vWAAS automatically offload pass-through traffic to vPath.
- Policy-based configuration; policies defined in the Nexus 1000V VSM are propagated to VMware vCenter and applied to the specified VM.
- VM mobility awareness; if a VM is moved, vPath continues to intercept and redirect traffic without requiring network changes.
- Fault-tolerant persistent performance; the vWAAS Data Redundancy Elimination (DRE) cache can deploy in SAN. VMware HA creates a new VM using the same DRE cache storage if vWAAS fails.

vNAM

The vNAM extends the Cisco Prime Network Analysis Module portfolio maximizing deployment flexibility in the virtual/cloud environment. The vNAM combines application awareness, deeper visibility into the network and rich performance analytics to accelerate operational decisions. It can be deployed easily anywhere in the network to improve or assure services levels. For example, vNAM can be deployed in the tenant container to monitor hosted workloads, at remote sites to monitor the end-user experience, or wherever there is a need to eliminate network blind spots. It can be installed on x86 platforms with ESXi and KVM virtualization infrastructures.

The vNAM gathers information from the network in multiple ways:

- Switched Port Analyzer (SPAN), Remote SPAN (RSPAN), encapsulated remote SPAN (ERSPAN) from Cisco switches.
- VLAN access control list (VACL)-based captures, used in conjunction with SPAN when supported by the switching platform.
- Promiscuous mode enabled on VMWare vSwitch for ESXi deployments.
- Cisco WAAS to deliver end-to-end visibility into WAN optimization infrastructure.
- Cisco Performance Agent to extend visibility into remote sites.
- NetFlow (Versions 5 and 9).

Deployed in the tenant network container, Cisco Prime vNAM analyzes the TCP-based interactions for the hosted workload to monitor performance in terms of metrics such as transaction time, server response time, and application delay. Setting performance thresholds helps to proactively detect performance problems, troubleshoot application response time concerns, and minimize the risks of violating service-level objectives. Cisco Prime vNAM also provides insight into network usage by applications, top talkers, and conversations to help optimize use of cloud infrastructure.

Refer to Service Assurance and Monitoring, page 3-27, and NetFlow, page 3-37 for more details.

System Level Design Considerations

The following system level design considerations are defined:

- Scalability, page 3-25
- Availability, page 3-25
- Security, page 3-26
- Manageability, page 3-26
- Service Assurance and Monitoring, page 3-27

Scalability

The following lists the most relevant scale concerns for the models discussed in this system release.

- **BGP Scale**—At this writing the ASR 9000 supports 5,000 BGP peers and functions as the centralized PE router for the virtual CE routers in the pod. For non-redundant CSR scenarios, up to 5000 virtual CE peers are supported per ASR 9000.
- **VLAN Scale**—At this writing (in NX-OS releases 5.2.5 through 6.1), up to 2,000 FabricPath-encapsulated VLANs are supported. This figure will improve in subsequent releases, but this constraint is expected to be eliminated in NX-OS 6.2, which targets 4000 transit VLANs. In future, segmentation scale will increase with the use of alternative encapsulations such as VXLANs.
- **Switches per FabricPath Domain**—NX-OS 5.2 supports up to 64 switch IDs; NX-OS 6.0 up to 128.
- **Port Density per FabricPath Node**—At 48 ports per module, F2 line cards provide up to 768 10 GigE 1 GigE ports per switch (N7018), while F1 cards provide up to 512 10 GigE ports (N7018). These are one-dimensional figures, but serve to give a theoretical maximum in terms of one measure of capacity. Currently, the Nexus 7000 FabricPath limitation is 256 core ports or 256 edge ports.
- **MAC Address (host) Scale—**FabricPath VLANs use conversational MAC address learning comprising a three-way handshake. Each interface learns MAC addresses only for interested hosts, rather than all MAC addresses in the VLAN. This selective learning enables the network to scale beyond the limits of individual switch MAC address tables. Classical Ethernet VLANs use traditional MAC address learning by default, but CE VLANs can be configured to use conversational MAC learning. MAC capacity on Nexus 5500 (L2) access-edge nodes is 24,000.
- **Tenancy—**The tenancy scope for this validation was 2,000. However, this does not represent the maximum scale of the architecture models. For the models we addressed, several factors constrain overall tenancy scale. These are: BGP peers per PE router per DC pod (5,000; end-to-end VLAN support (currently, 2,000 transit VLANs); VLANs per UCS (1,000, although this constraint can be minimized through the use of VXLANs for host connectivity); and Nexus 1000V scale (4,000 ports/128 hosts in release 2.2).

Availability

The following methods are used to achieve HA in the VMDC data center architecture:

- Routing and NV-edge clustered redundancy at the WAN/IP NGN infrastructure edge, including path and link redundancy, non-stop forwarding and route optimization.
- L2 redundancy technologies are implemented through the FabricPath domain and access tiers of the infrastructure. This includes Address Resolution Protocol (ARP) synchronization in vPC/vPC+-enabled topologies to minimize unknown unicast flooding and reconvergence; ECMP; and port-channel utilization between FabricPath edge/leaf and spine nodes to minimize L2 IS-IS adjacency recalculations and system reconvergence.
- Hardware and fabric redundancy throughout.
- VEM: Multi-Chassis EtherChannel (MCEC) uplink redundancy and VSM redundancy in the virtual access tier of the infrastructure.
- In the compute tier of the infrastructure, HSRP (for CSR redundancy), port-channeling, NIC teaming, and intra-cluster HA through the use of VMware vMotion, along with Active/Standby redundant failover for SLB and ASA 1000V virtual appliances.

Security

Security best practices from previous VMDC releases are leveraged for tenancy separation and isolation. The fact that dedicated network service resources are employed simplifies the isolation model so that VRF isolation is not required in the data center.

Security related considerations include:

- **Remote Access**—IPsec and MPLS VPNs provide secure remote access over the Internet or public provider IP/NGN backbone.
- **L3 Separation—**BGP at the WAN edge/PE routing provides per-tenant routing to dedicated per-tenant vCE routers. Policies can be applied on both devices to restrict inter-tenant communication.
- **Access and Virtual Access Layer (L2) Separation—**VXLAN or VLAN IDs and the 802.1q tag provide isolation and identification of tenant traffic across the Layer 2 domain.
- **Network Services Separation (Compute)—**Dedicated per-tenant virtual service appliances or zones provide virtualized security, load balancing, NAT, and SSL offload services, and the application of unique per-tenant policies at VLAN/VXLAN or VM granularity.
- **Storage**—This VMDC design uses NetApp for NFS storage, which enables virtualized storage space so that each tenant (application or user) can be separated using ipspace and VLANs mapped to network layer separation. The vSphere hypervisor's cluster file system management creates a unique Virtual Machine Disk (VMDK) per VM, ensuring 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. In clustered Data ONTAP, a Storage Virtual Machine (SVM) contains data volumes and one or more LIFs (logical interfaces which have IPs) through which it serves data to the clients. An SVM securely isolates the shared virtualized data storage and network, and appears as a single dedicated server to its clients. Each SVM has a separate administrator authentication domain and can be managed independently by a SVM administrator. Secure multi-tenancy is provided by network administration and control that is scoped to a particular SVM. Multiple SVMs can coexist in a single cluster without being bound to any node in a cluster. Additional methods for implementing secure customer separation within a FlexPod unit can be found at: https://tech.netapp.com/internal/03/technet_services_solutions_smt.html. For SANs, this design allows for Fiber Channel (FC) access separation at the switch port level (VSAN), Logical path access separation on the path level (World Wide Name (WWN) or Device Hard Zoning), and at the virtual media level in the storage array (LUN masking and mapping).

Manageability

For service provisioning and orchestration, this architecture leverages Cisco Intelligent Automation for Cloud (CIAC) and BMC Cloud Lifecycle Management (CLM) for automated service orchestration. Information about CIAC can be found here: Intelligent Automation for Cloud. CLM was addressed in previous system releases (VMDC 2.0 and updated in the VMDC 2.2 release). Additional documentation can be found on Design Zone at Cloud Orchestration with BMC CLM.

From a storage/FlexPod automation perspective, OnCommand: Workflow Automation (WFA), NetApp's storage automation product, makes common storage management processes simple and easy. Storage experts can easily define common storage management processes like provisioning, setup, migration, and decommissioning, and make them available for execution by approved users. WFA can leverage the current automation policies to demonstrate the value of a "Storage Service Catalog" and can also integrate with the existing orchestration systems. Refer to https://communities.netapp.com/community/products_and_solutions/storage_management_software/workflow-automation.

Service Assurance and Monitoring

Service assurance is generally defined as the application of policies and processes to ensure that network services meet predefined service quality levels for optimal subscriber experiences. Service assurance enables SPs to control traffic flows, identify faults, and resolve issues in a timely manner to minimize service downtime. Service assurance also includes policies and processes to proactively diagnose and resolve service quality degradations or device malfunctions before subscribers are impacted.

In VMDC VSA 1.0, network service assurance encompasses the following concepts:

- Traffic Engineering, page 3-27
- QoS Framework, page 3-31
- Application Visibility and Control, page 3-37
  - NBAR, page 3-37
  - NetFlow, page 3-37
  - Performance Agent, page 3-41
  - Network Analysis, page 3-41
  - Application Optimization, page 3-41
- Cloud Service Assurance for VMDC, page 3-41

Network service assurance may be used in conjunction with storage service assurance to provide application-centric service differentiation. In this release we leverage the following storage service assurance functions:

- Storage QoS, page 3-42
- Storage Oversubscription, page 3-43
- Storage Service Tiering, page 3-43

Traffic Engineering

Traffic engineering is a method of optimizing network performance by dynamically analyzing, predicting and regulating the behavior of transmitted data.

Port-channels are frequently deployed for redundancy and load sharing. Because the Nexus 1000V is an end-host switch, network administrators can use different approach than those used on physical switches, implementing a port-channel mechanism in one of the following modes:

- **Standard Port-Channel**—The port-channel is configured on the Nexus 1000V and on upstream switches.
- **Special Port-Channel**—The port-channel is configured only on the Nexus 1000V; there is no need to configure anything upstream. Two options are available: MAC pinning and vPC host mode.

Regardless of mode, port-channels are managed using standard port-channel CLI, but each mode behaves differently.

For more information on Nexus 1000V port-channel configurations, follow this link:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/guide_c07-556626.html#wp9000299

The VMDC virtual access layer design uses vPC host mode and then uses MAC pinning to select specific links from the port channel. As discussed in previous system releases, multiple port-channels can be used for a more granular approach for uplink traffic management on the Nexus 1000V. These options are shown in Figure 3-15 and Figure 3-16.

Figure 3-15 Nexus 1000v 5 Uplink Port Channel Model

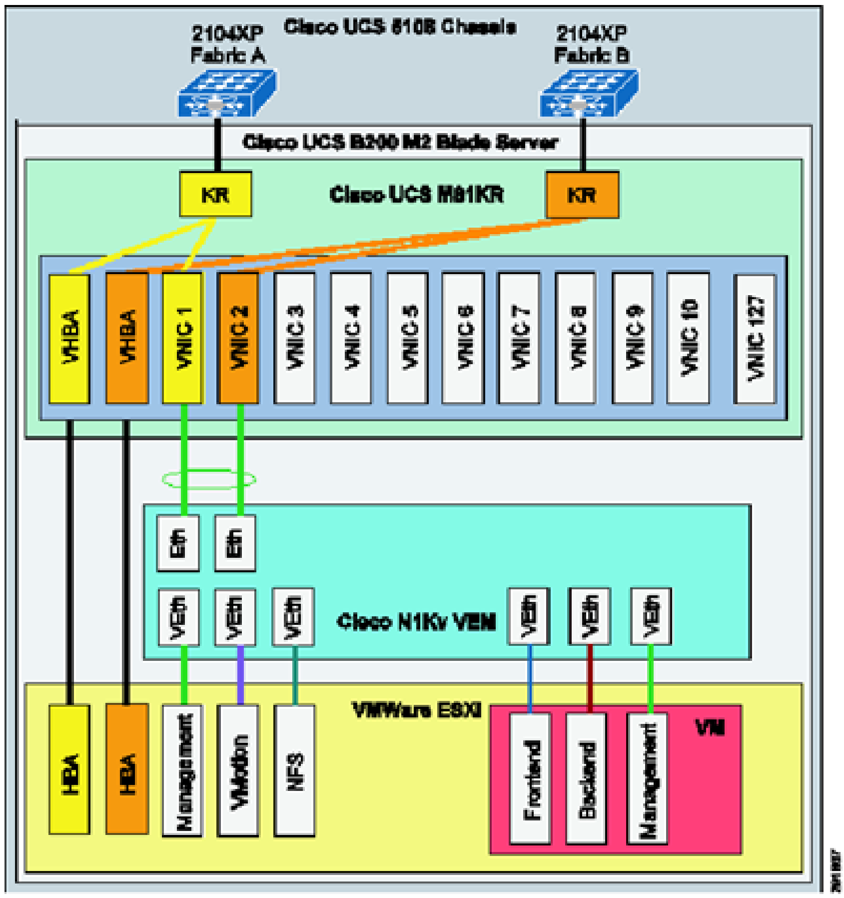
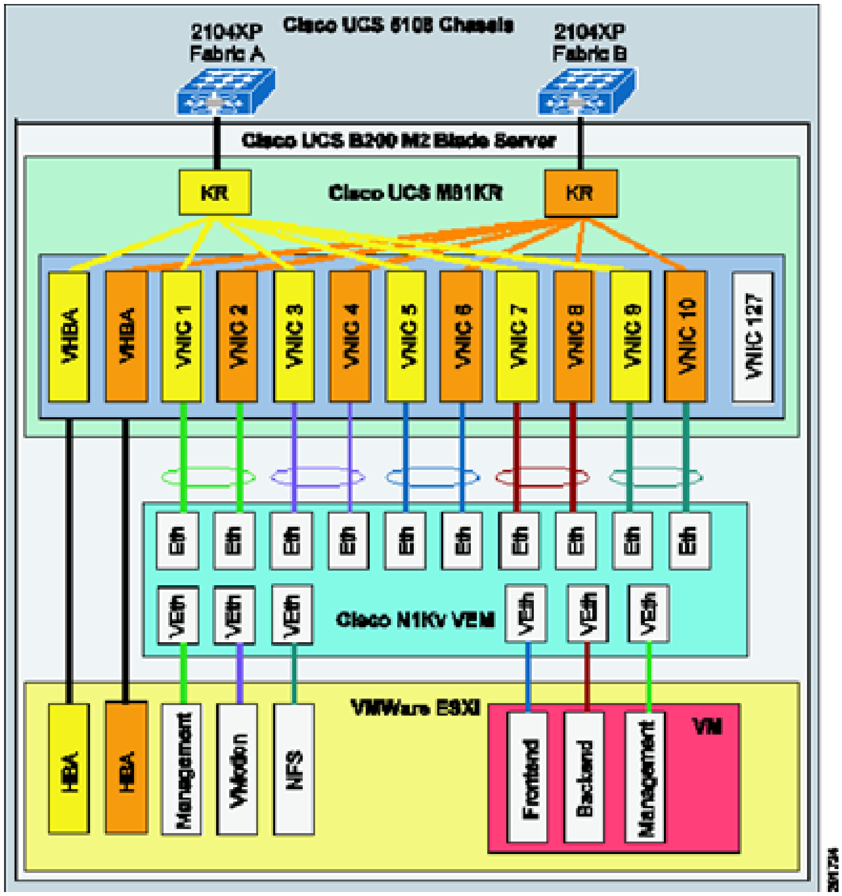


Figure 3-16 Nexus 1000v 5 Uplink PortChannel Model



Traffic engineering can be performed selectively by configuring the Nexus 1000V to select the target uplink using a manual configuration (static pinning) instead of the default. For example, front-end traffic that contains many diversified flows can use both members (fabrics) of the port-channel. On the other hand, backend traffic, which has more diversity in terms of bandwidth/response time (VM-to-VM inter-fabric traffic flows, vMotion, backup, and so on) can benefit by selecting a path that enables VM-to-VM traffic to remain in the same fabric so that Fabric Interconnect switches the traffic locally. Table 3-1 lists key architectural features of VMDC VSA 1.0.

Table 3-1 Traffic Classification Example for MAC Pinning

Traffic Type	Classification	UCS Fabric	Mac-Pining Option	Rational
Front End Traffic	Tenant Data	Fabric A & B	Automatic	Load Share on all available uplinks, most traffic should be exiting the pod through the Aggregation-Edge Nexus 7000

Table 3-1 Traffic Classification Example for MAC Pinning (continued)

Traffic Type	Classification	UCS Fabric	Mac-Pining Option	Rational
Back End Traffic	Tenant Data	Fabric-A	Manual	Keep most back end traffic local switched on one Fabric Interconnect
vMotion	VMkernel/Control	Fabric-B	Manual	Keep vMotion traffic local switched on one Fabric Interconnect

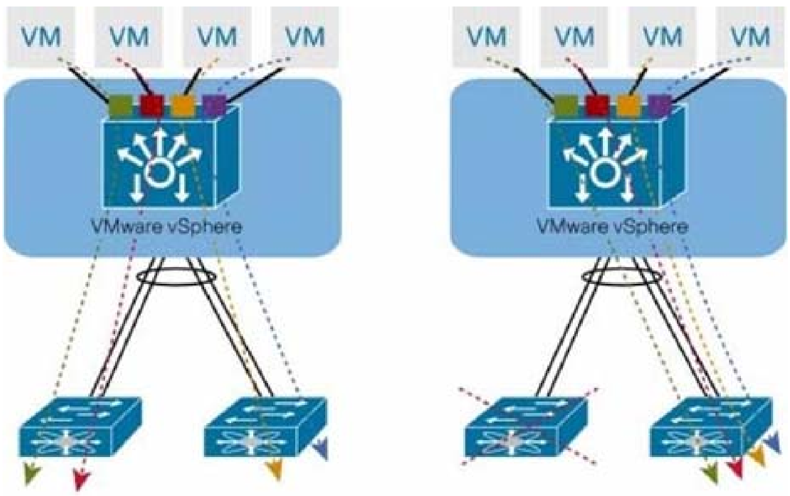
MAC Pinning

MAC pinning defines all uplinks coming out of the server as standalone links and pins different MAC addresses to those links in a round-robin fashion. This approach helps to ensure that the MAC address of a virtual machine is never seen on multiple interfaces on the upstream switches. No upstream configuration is required to connect the Nexus 1000V VEM to upstream switches (Figure 3-17).

MAC pinning does not rely on any protocol to distinguish upstream switches, so the deployment is independent of any hardware or design. MAC pinning enables consistent, easy Nexus 1000V deployment because it does not depend on any physical hardware or any upstream configuration, and it is the preferred method for deploying Nexus 1000V if upstream switches cannot be clustered.

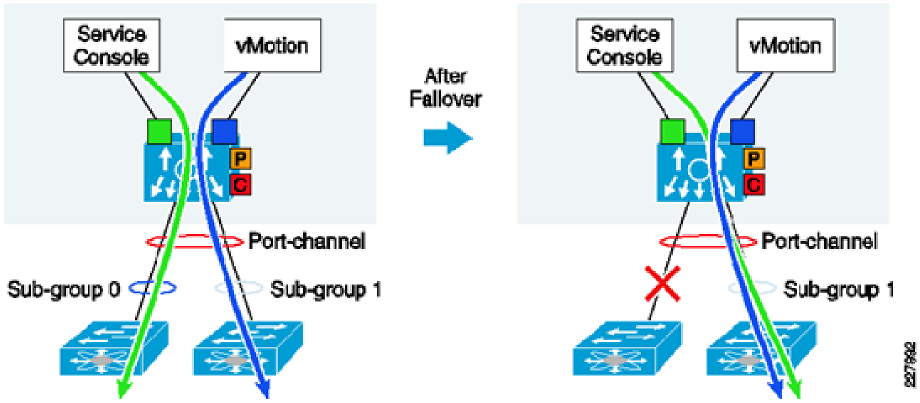
However, this approach does not prevent the Nexus 1000V from constructing a port-channel on its side, providing the required redundancy in the data center in case of a failure. If a failure occurs, the Nexus 1000V sends a gratuitous ARP packet to alert the upstream switch that the MAC address of the VEM learned on the previous link must now be learned on a different link, enabling sub-second failover.

Figure 3-17 MAC-Pinning Details



In the case of a fabric failure, the Nexus 1000V selects the available remaining fabric to recover the traffic. Figure 3-18 shows the fabric failover with subgroup MAC pining.

Figure 3-18 MAC-Pinning Failover



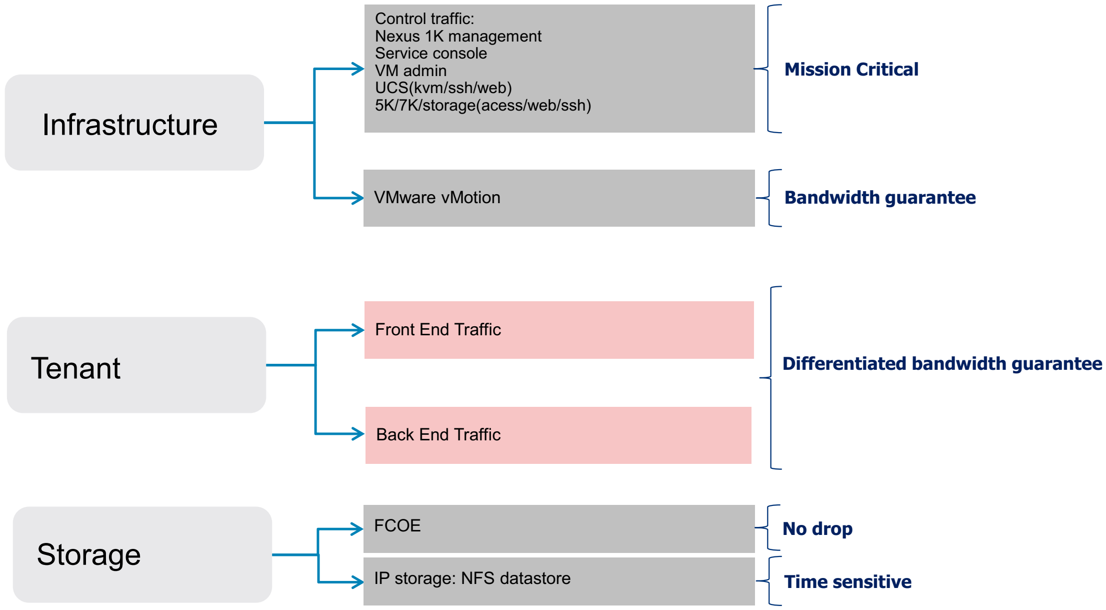
QoS Framework

QoS is a key to service assurance because it enables differentiated treatment of specific traffic flows. Differentiated treatment ensures that critical traffic is provided sufficient bandwidth to meet throughput requirements during congestion or failure conditions.

Figure 3-19 illustrates the different traffic flow types defined in previous VMDC releases. These traffic types are organized in infrastructure, tenant, and storage traffic categories.

- Infrastructure traffic comprises management and control traffic, including VMware service console and vMotion communication. This is typically set to the highest priority to maintain administrative communication during periods of instability or high CPU utilization.
- Tenant traffic can be differentiated into front end and backend traffic, with service levels to accommodate various traffic requirements in each category.
- The VMDC design incorporates Fibre Channel and IP-attached storage. As shown in Figure 3-19, storage requires two subcategories, because these traffic types are treated differently throughout the network. Fibre Channel traffic, by definition, requires a “no drop” policy, while Network File System (NFS) data store traffic is sensitive to delay and loss.

Figure 3-19 Traffic Flow Types



To provide differentiated services, VMDC leverages the following QoS functionality:

- Traffic Classification and Marking, page 3-32
- Congestion Management and Avoidance (Queuing, Scheduling, and Dropping, page 3-33)
- Traffic Conditioning (Shaping and Policing, page 3-35)

Classification and Marking

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

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

1. Infrastructure
2. Tenant service classes (three data; two multimedia priority)
3. Storage

Figure 3-20 provides a more granular description of the requisite traffic classes, characterized by their DSCP markings and per-hop behavior (PHB) designations. This represents a normalized view across validated VMDC and HCS reference architectures in the context of an eight-class IP/NGN aligned model.

Figure 3-20 VMDC Traffic Classes (8-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

Note that in newer data center QoS models, CoS 3 is reserved for lossless data (FCoE). However, in older WAN/campus QoS services models, CoS 3 is used for voice over IP (VoIP) signaling. The table assumes that FCOE traffic is localized to UCS and Ethernet-attached storage systems, enabling the use of CoS 3 for VoIP signaling traffic in the data center QoS domain. Classification values may need to be tweaked per traffic characteristics; for example, CoS 4 could potentially be used for VoIP call control if video streams are not deployed.

It is a general best practice to mark traffic at the source-end system, or as close to the traffic source as possible, to simplify network design. However, if the end system cannot mark or cannot be trusted, marking can be performed on network ingress. In the VMDC QoS framework, the cloud data center represents a single QoS domain, with the Nexus 1000V forming the "southern" access edge, and the ASR 9000 or ASR 1000 forming the "northern" DC PE/WAN edge. These QoS domain edge devices mark traffic, and these markings are trusted at nodes in the data center infrastructure. In other words, they use simple classification based on the markings received from the edge devices. Note that where VM-FEX adapters are used, marking is implemented on UCS Fabric Interconnects; in contrast to the Nexus 1000V implementation, there is no ability to conditionally mark-down CoS in the event of congestion.

In VMDC, the assumption is that DSCP values are not altered. Intermediate nodes would ideally support QoS transparency, so that CoS values would not need to be re-marked. That said, if QoS transparency is not supported on a particular node within the QoS domain, it will be necessary to workaround this gap by re-marking.

In VMDC VSA 1.0, VPX SLB does not support QoS transparency. The insertion of CSR as the tenant virtual private cloud represents a trust boundary. In these cases, it is necessary to classify and remark at the CSR.

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:

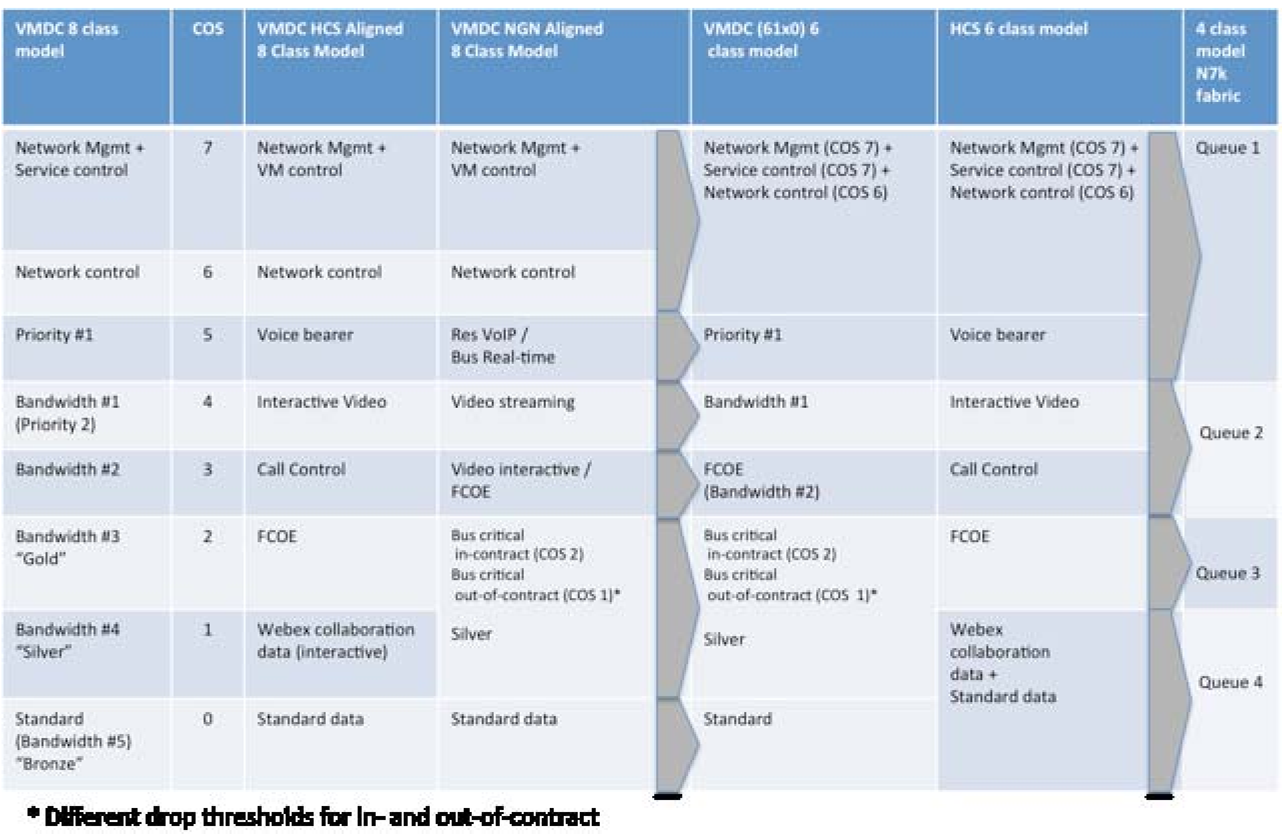
- First in, first out (FIFO)
- Priority scheduling (also called priority queuing)
- Weighted bandwidth

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 data center QoS domain. At the ASR 9000 or ASR 1000 northern data center WAN edge, we use priority queuing (PQ)/CBWFQ to bound delay and jitter for priority traffic while supporting weighted bandwidth allocation for the remaining data traffic classes.

Queuing mechanisms manage the front of a queue, while congestion avoidance mechanisms manage the back of a queue. Because queue depths are limited, dropping algorithms, which drop packets as queue depths build, are used to avoid congestion. Two dropping algorithms are commonly used: weighted tail drop (often for VoIP or video traffic) or weighted random early detection (WRED), typically for data traffic classes. As in previous releases, WRED is used to drop out-of-contract data traffic (CoS 1) before in-contract data traffic (Gold and CoS 2), and for Bronze/Standard traffic (CoS 0) in the event of congestion.

Defining an end-to-end QoS architecture can be challenging because not all nodes in a QoS domain have consistent implementations. In the cloud data center QoS domain, we run the gamut from systems that support 16 queues per VEM (Nexus 1000V) to four internal fabric queues (Nexus 7000). This means that traffic classes must be merged on systems that support less than eight queues. Figure 3-21 shows the class-to-queue mapping that applies to the cloud data center QoS domain in the VMDC reference architecture, in the context of alignment with either the HCS reference model or the more standard NGN reference.

Figure 3-21 VMDC Class-to-Queue Mapping



> **Note** The Nexus 2000 Fabric Extender provides only two user queues for QoS support: one for all no-drop classes and the other for all drop classes. The classes configured on its parent switch are mapped to one of these queues; traffic for no-drop classes is mapped one queue and traffic for all drop classes is mapped to the other. Egress policies are also restricted to these classes. Further, at this writing (NX-OS 6.1.3), queueing is not supported on Nexus 2000 host interface ports when connected to an upstream Nexus 7000 switch. Traffic is sent to the default fabric queue on the Nexus 7000, and queuing must be applied on FEX trunk (network interface) ports. Future NX-OS releases will feature enhanced Nexus 7000 support for FEX QoS, adding network QoS and default queuing policy support on downstream Nexus 2000 host interfaces.

Before NX-OS release 6.1.3, only two ingress queues are supported on the F2/F2E Nexus 7000 line cards. Release 6.1.3 adds support for four ingress queues. These line cards support four egress queues.

Shaping and Policing

Policing and shaping are used to enforce a maximum bandwidth rate (MBR) on a traffic stream; while policing effectively does this by dropping out-of-contract traffic, shaping does this by delaying out-of-contract traffic. VMDC uses policing in and at the edges of the cloud data center QoS domain to rate-limit data and priority traffic classes. At the data center WAN edge/PE, hierarchical QoS (HQoS) may be 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 used to enforce per-tenant aggregates.

Sample bandwidth port reservation percentages are shown in Figure 3-22.

Figure 3-22 Sample Bandwidth Port Reservations

Traffic Class	CoS	N7k F1 Egress Queue	N5k QoS Class Maps	N5k Queuing Class Maps	N5k QoS Groups	N5k % BW	UCS	UCS % BW
Mgmt	7	1p3q1t-8e-out-pq1	class-vmdc-priority	vmdc-pq	5	5	6	
Network Control	6	1p3q1t-8e-out-pq1	class-vmdc-priority					
VoIP	5	1p3q1t-8e-out-pq1	class-vmdc-priority				5	
Video, NAS	4	1p3q1t-8e-out-q2	class-vmdc-p2	vmdc-p2	4	10	4	
Call Control	3	1p3q1t-8e-out-q2	class-fcoe	class-fcoe	1		3	
Premium Data	2	1p3q1t-8e-out-q3	class-vmdc-p3	vmdc-p3	3	60	2	
" "	1	1p3q1t-8e-out-q-default	class-vmdc-p4	vmdc-p4	2	1		
Standard Data	0	1p3q1t-8e-out-q-default	class-default	class-default	0	24	Best Effort	

Figure 3-23 and Figure 3-24 summarize the QoS mechanisms employed intra-DC in this release:

Figure 3-23 Sample Bandwidth Port Reservations

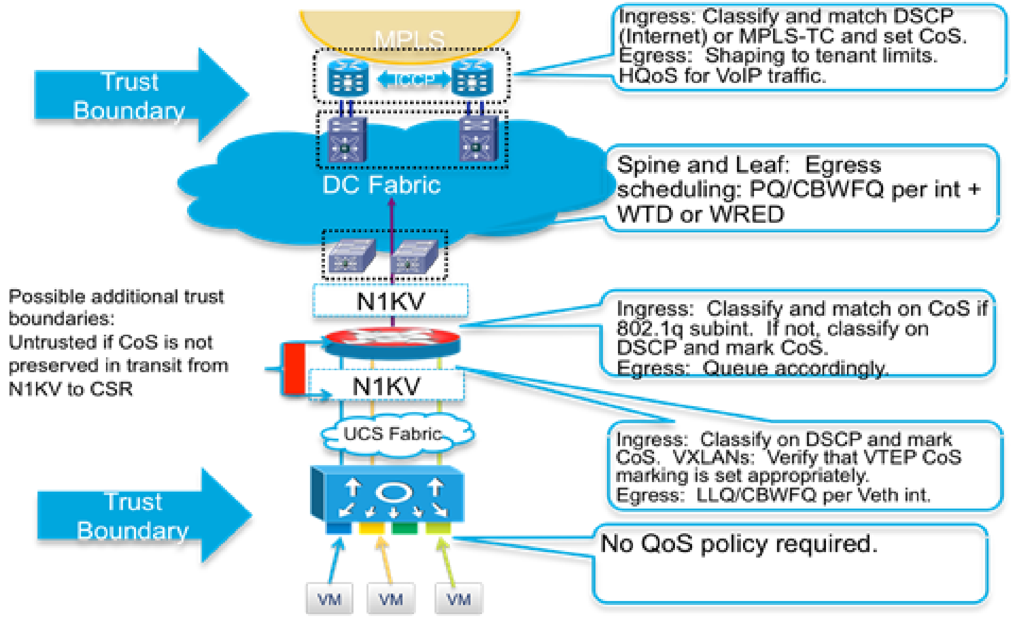
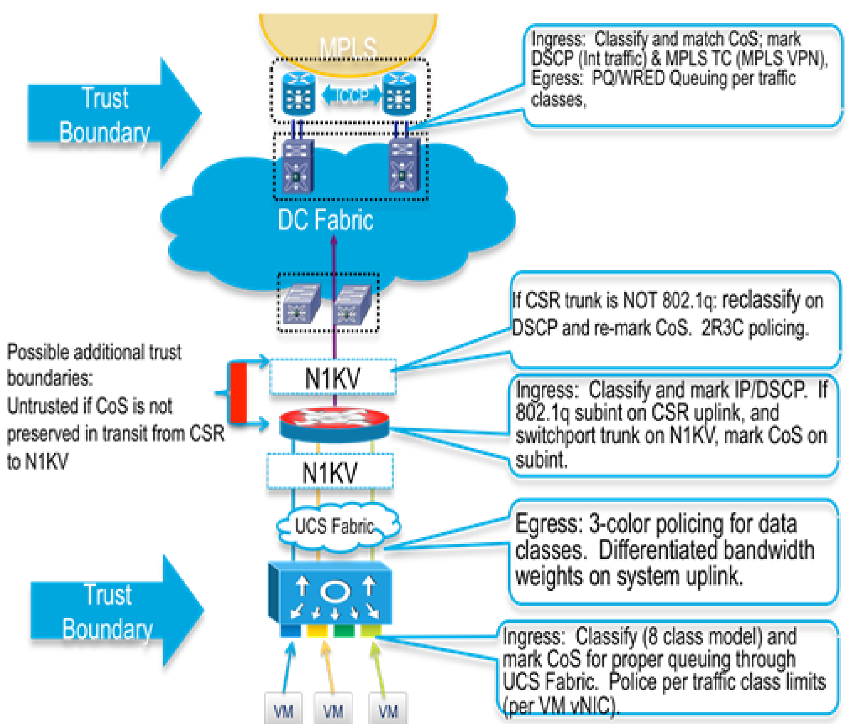


Figure 3-24 Sample Bandwidth Port Reservations



Application Visibility and Control

Cisco's Application Visibility and Control solution is a suite of services for application classification, traffic control and monitoring that support capacity planning and resource management, improve business-critical application performance, and in the public environment, allow for highly tuned service assurance. Technologies such as Quality of Service, previously discussed, Network Based Application Recognition, Netflow, Network Analysis and Application Optimization form the basis of this solution set.

NBAR

Cisco Network Based Application Recognition (NBAR) provides the option of using stateful deep packet inspection for granular, application-level traffic inspection and traffic classification for performance monitoring and tuning or for other use cases, such as application monitoring for security purposes. NBAR currently provides the ability to identify over 1000 application signatures. NBAR is also capable of defining customized application profiles based on ports, URL or even payload values. NBAR is implemented in hardware on physical routers such as the ASR1000 or ISRG2, however on the CSR, NBAR is software-based. This means that caution should be employed in deploying NBAR to understand the impact on forwarding performance and insure the desired packet throughput.

NetFlow

Cisco developed NetFlow to provide better insight into IP traffic. A key component of Cisco's Application, Visibility and Control suite of features and functionality, NetFlow defines flows as records and exports the records to collection devices. NetFlow provides information about the applications in and utilization of the data center network. The NetFlow collector aggregates and assists network administrators and application owners to interpret the performance of the data center environment.

The use of NetFlow is well documented in traditional network environments, but the Nexus 1000V provides this capability in the virtual network environment. Nexus 1000V supports NetFlowv9 and by default uses the management 0 interface as an export source.

Caution Using advanced features such as NetFlow consumes additional ESXi host resources (memory and CPU). It is important to understand these resource dynamics before enabling advanced features.

Figure 3-25 shows the Cisco NAM application traffic statistics on the Nexus 1000V virtual Ethernet interfaces. The Nexus 1000V can also monitor flows from the physical interfaces associated with the platform and VMkernel interfaces including vMotion traffic, as seen in Figure 3-26.

Figure 3-25 Cisco NAM Traffic Summary

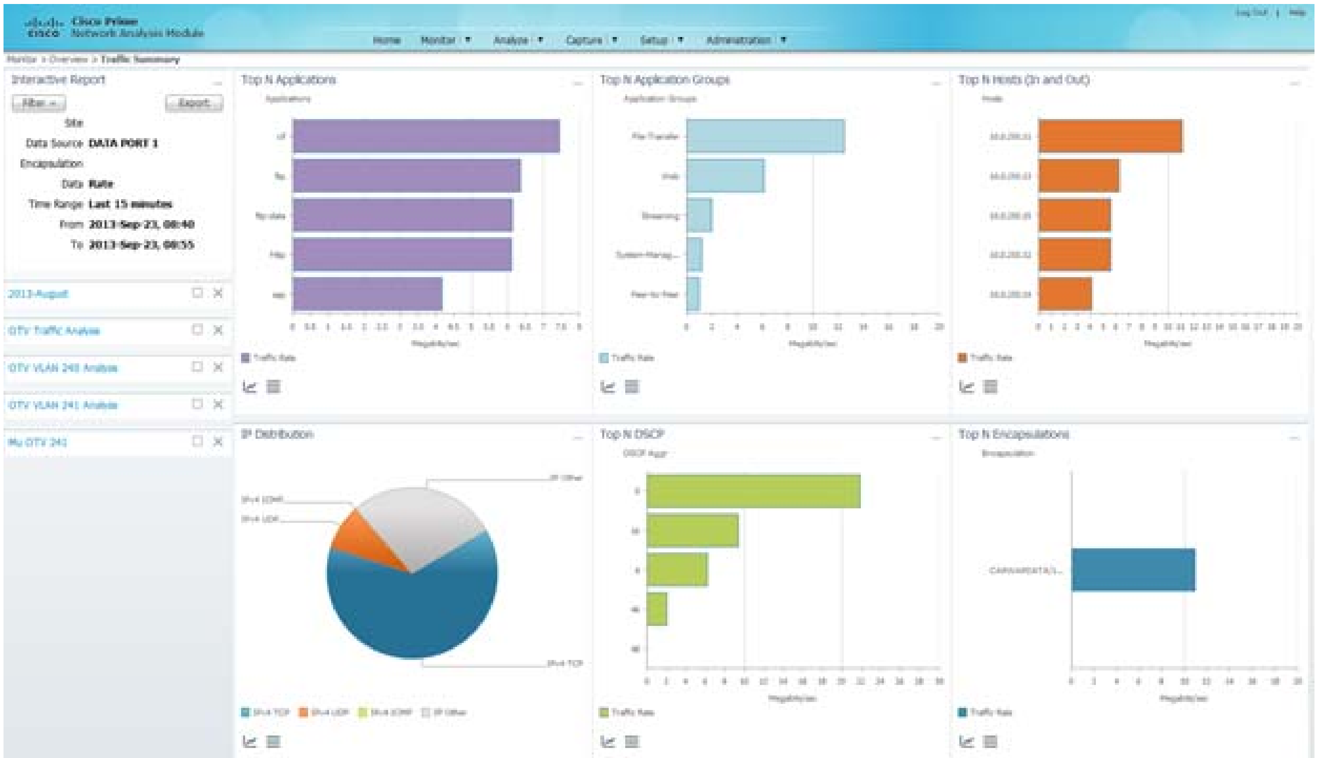


Figure 3-26 Cisco Netflow Collector Nexus 1000v vMotion Results Example

ERSPAN

ERSPAN supports remote monitoring of network resources. ERSPAN uses generic routing encapsulation (GRE) tunnels to route traffic. The Nexus 1000V supports ERSPAN, enabling network administrators to observe traffic associated with:

- The individual vNIC of a VM connected to a VEM
- The physical ports associated with the ESXi host
- Any port channels defined on the VEM

This flexibility enables ERSPAN sessions to monitor data associated with VM's, and to monitor all traffic associated with the ESXi host including VMkernel, vMotion, and service console data. Converging these traffic types onto two or a maximum of four CNAs per-ESXi host simplifies physical data center design and the configuration of capture points.

In the validation of this solution, the final destination for ERSPAN traffic was the Virtual Network Analysis Module (vNAM), resident in the compute layer of the infrastructure in vApp form factor.

For more information about configuring ERSPAN on the Nexus 1000V, follow this link:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus1000/sw/4 0 4 s v 1 2/system_management/configuration/guide/n1000v_system_9span.html

Caution Using advanced features such as ERSPAN consumes additional ESXi host resources (memory and CPU). It is important to understand the resource dynamics before enabling advanced features.

Figure 3-27 and Figure 3-28 show examples of a packet decode and application performance metrics available from the ERSPAN data.

Figure 3-27 View of NAM Captured Data from VM NIC

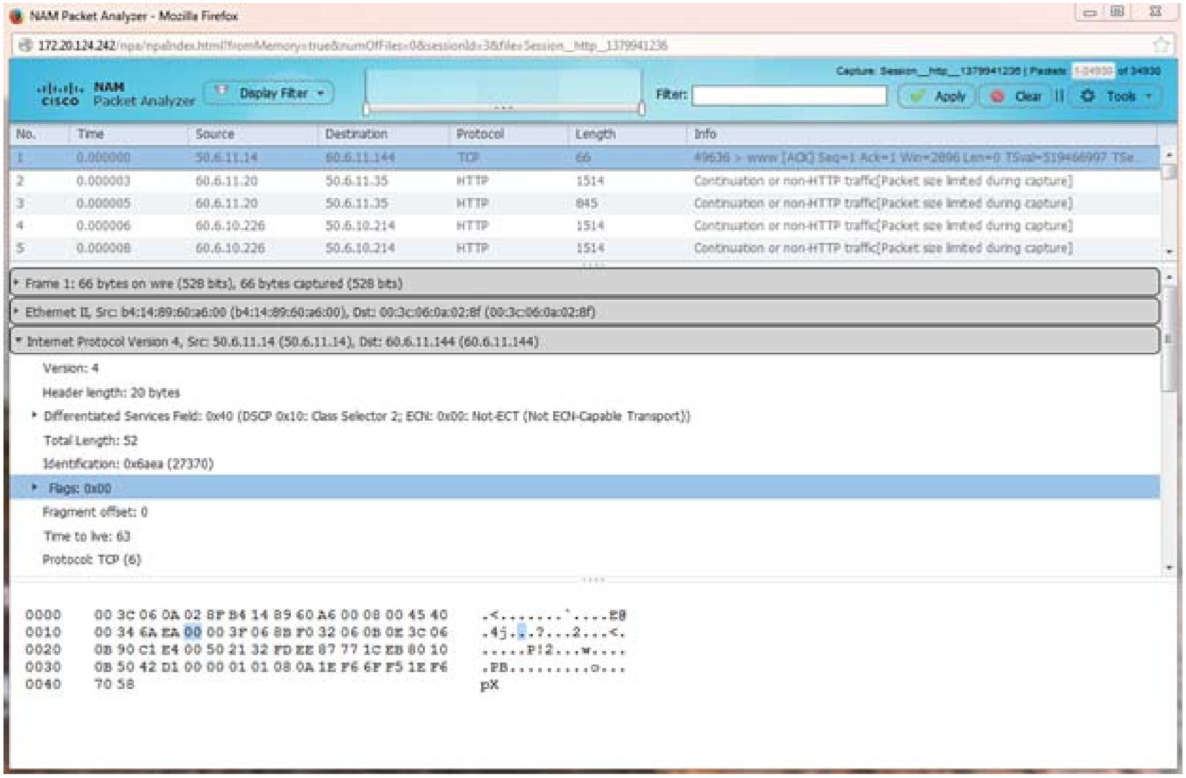
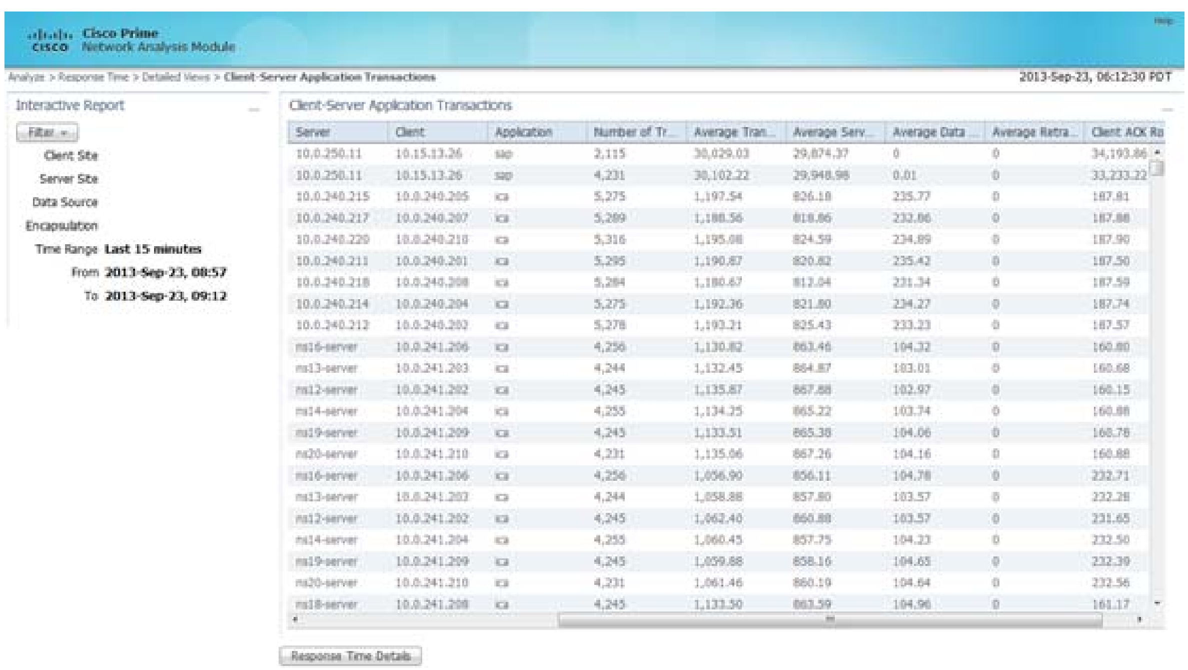


Figure 3-28 Application Response Time Data Collected on Nexus 1000V VEM Uplink



Performance Agent

Cisco Performance Agent is a licensed feature of Cisco IOS Software, providing embedded instrumentation for application analytics. Based upon deep packet inspection capabilities of NBAR (Network Based Application Recognition), Performance Agent offers comprehensive application performance and network usage data to help network administrators accurately assess user experience and optimize the use of network resources. Cisco Performance Agent works with other Cisco IOS Software Application Visibility and Control features, such as Netflow, NBAR and QoS, as well as the Cisco Network Analysis Module (NAM or vNAM) and the WAAS or vWAAS. Once enabled, Performance Agent may be configured to export application performance analytics, traffic statistics, and WAN-optimization statistics, delivering combined metrics gathered from Netflow, NBAR and NAM monitoring if desired. The information may be delivered to third party management tools in a Netflow v9 format or presented to the Cisco Prime NAM or WAAS vCM consoles in GUI reporting fashion. Performance Agent is supported on the ASR1000, ISRG2 and CSR routers utilized in this systems release. More information on Performance Agent is available online.

Network Analysis

The use of network analysis devices is another application visibility service readily available in the VMDC design. The Cisco Nexus 1000V NAM VSB is integrated with the Nexus 1100 Virtual Services Appliance to provide network and performance visibility into the Nexus 1000V switching deployment.

In this release of VMDC we also introduce the vNAM, a new vAPP-based form-factor that leverages the Nexus 1000v for traffic redirection and visibility. Cisco vNAM is introduced with Cisco Prime NAM software release 6.0. This software release introduces several new features enhancing overall application and performance visibility, including enhanced classification capabilities leveraging the Nexus 1000v DVS, and support for VXLAN, LISP and OTV encapsulations. For VMDC VSA 1.0 , which utilizes VXLANs to scale segmentation within the compute layer of the infrastructure, the latter is key to making this type of "overlay" technology feasible from an operational perspective.

The NAM VSB and new vNAM use embedded instrumentation, such as NetFlow and Encapsulated Remote SPAN (ERSPAN) or SPAN on the Nexus 1000V switch as the data source for traffic analysis, application response time, interface statistics, and reporting. Alternatively, in VMDC VSA 1.0, the CSR vCE, as the Virtual Private Cloud boundary point provides another source for SPAN or ERSPAN of aggregated traffic statistics for network analysis.

For more information about NAM for Nexus 1100, follow this link:

http://www.cisco.com/en/US/products/ps10846/index.html

For more information about vNAM, see the following link:

http://www.cisco.com/en/US/products/ps12941/index.html

Application Optimization

Wide Area Application Services or WAAS (WAAS) is another key component of Cisco's Application Visibility and Control suite of products, providing application-centric acceleration services over the wide area. For information on how WAAS was utilized in VMDC VSA 1.0, refer to Cisco vWAAS, page 3-22.

Cloud Service Assurance for VMDC

Cloud Service Assurance for VMDC (CLSA-VMDC), based on the Zenoss Cloud Service Assurance solution, is a service-impact model-based system for tenant-based service assurance. CLSA-VMDC supports consolidated VMDC infrastructure monitoring and simple, easily-deployed plug-ins for customization. CLSA-VMDC offers real-time aggregated dashboards and reporting capabilities. CLSA-VMDC can be deployed in centralized and distributed architectures, and supports incremental deployment growth. While CLSA-VMDC offers rich functionality for IaaS domains, the solution is lightweight and has open interfaces to enable simple integration into existing operations support system (OSS) and ticketing systems with minimal cost. This solution is positioned not as a replacement, but as a complement to existing Manager-of-Manager (MOM) systems (for example, IBM Netcool), ticketing systems (for example, BMC Remedy), and so on. Additional documentation can be found on Design Zone at Data Center and Cloud Service Assurance.

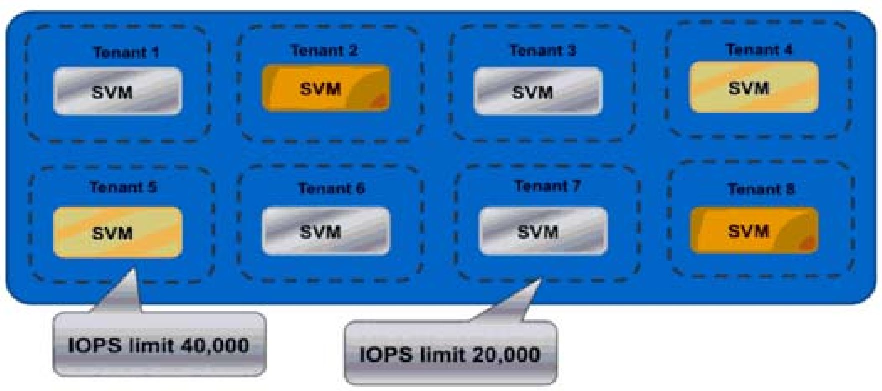
Storage QoS

Tenant workloads in the VMDC datacenter should be prevented from affecting each other. While this is handled by QoS mechanisms at the network layer, those QoS mechanisms do not protect performance at the storage layer. Because all storage I/O traffic is classed the same in the VMDC datacenter, all tenants receive an equal share of the storage performance capacity unless controlled by some other mechanism. Without implementing performance limits on storage I/O, the performance of one tenant workload may suffer due to the overwhelming use of storage I/O by a neighboring tenant workload sharing the same physical hardware.

Additionally early adopter tenants' workloads may experience higher than normal performance early in a PoD lifecycle due to the low number of tenants using the resources. Without setting storage performance caps those tenants may perceive a decrease in storage performance once the tenant capacity for their PoD has been reached. Even though the performance level they are receiving may be within the bounds of the tenant's service level agreement, the tenant may perceive that performance has dropped below the expected level.

NetApp FAS controllers running clustered Data ONTAP prevent workloads from impacting each other through the use of Storage Quality of Service (Figure 3-29).

Figure 3-29 Application of Storage QoS



NetApp Storage QOS allows I/O ceilings do be defined in terms of IOPS. Those performance ceilings can be applied to individual workloads or to groups of workloads. In a multi-tenant environment, a tenant could have IOPS limits set across all workloads hosted within a particular ICS or could have limits set on each deployed workload.

For additional information on NetApp Storage QoS, see http://www.netapp.com/us/media/wp-7191.pdf.

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 white space, 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. 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.

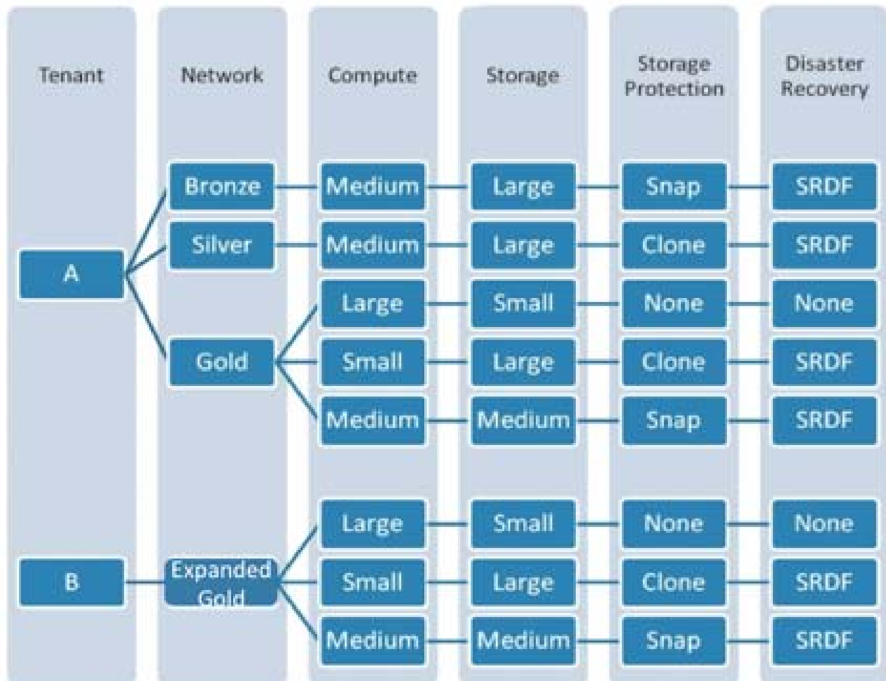
In the Integrated Compute Stack layer of the VMDC architecture, in the context of a FlexPod environment, thin provisioning, data deduplication, and FlexClone thincloning technology are the critical components of the NetApp solution, offering multiple levels of storage efficiency across the virtual desktop OS data, installed applications, and user data. This helps customers save 50% to 90% of the cost associated with shared storage (based on existing customer deployments and NetApp solutions lab validation). Thin provisioning is a method of logically presenting more storage to hosts than is physically available. With thin provisioning, the storage administrator can access a pool of physical disks (known as an aggregate) to create logical volumes for different applications to use, without pre-allocating space to those volumes. The space is allocated only when the host needs it. Thus unused aggregate space is available for the existing thin-provisioned volumes to expand or for use in the creation of new volumes. NetApp deduplication saves space on primary storage by removing redundant copies of blocks in a volume that may be hosting hundreds of virtual desktops. This process is transparent to the application and user and can be enabled and disabled on the fly. Using NetApp deduplication in conjunction with file FlexClone technology can reduce the overall storage footprint of virtual machines.

Storage Service Tiering

Service differentiation in the VMDC reference architecture is a composite of differentiated network services (QoS policy, security, server load balancing, SSL offload, application control/optimization); differentiated compute attributes; and differentiated storage and business continuance characteristics.

Figure 3-30 shows this concept, demonstrating a variety of ways in which these resources and services can be applied in various combinations to meet business or application requirements in a tiered fashion.

Figure 3-30 Tiered Services



With respect to storage resources, the following methods may be utilized to differentiate storage services to meet tenant application requirements and insure service continuity:

- **Storage Tiering**—A mixture of various types of storage drives (SATA, SAS, flash) with differing capacity, performance and reliability characteristics are applied to meet application IOPs requirements.
- **Storage Protection**—Storage protection options in the form of datastore snapshots to provide point in time file copies, or cloning for data volume or dataset replication.
- **Disaster Recovery**—Application of synchronous or asynchronous data replication technologies from source to target storage system for backup and recovery from secondary facilities for mission-critical applications.

In the context of FlexPod and NetApp storage arrays, the following links provide more information on tiering, replication, backup, and DR technologies:

- **Virtual Storage Tiering**: http://www.netapp.com/in/technology/virtual-storage-tier/index.aspx
- **SnapMirror Datasheet**: http://www.netapp.com/in/products/protection-software/snapmirror.aspx
- **SnapMirror Best Practices**: http://www.netapp.com/us/media/tr-4015.pdf
- **SnapVault Datasheet**: http://www.netapp.com/in/products/protection-software/snapvault.aspx
- **SnapVault Best Practices**: http://www.netapp.com/us/media/tr-4183.pdf