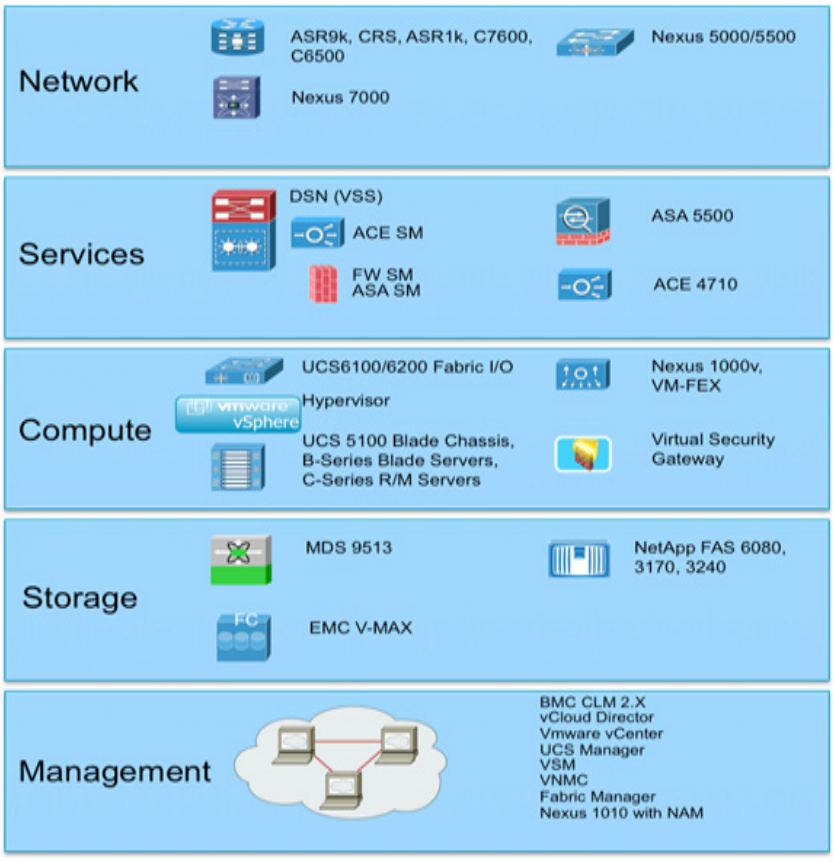
CHAPTER 3

VMDC 3.0.1 Design Details

VMDC Data Center functional layers are shown in Figure 3-1.

Figure 3-1 Functional Layers Within the VMDC Data Center



VMDC Building Blocks

The following functional layers that constitute VMDC component building blocks are introduced:

Network Layer

The Network layer includes the WAN/PE router, which forms the data center perimeter to the enterprise wide area or provider IP/NGN backbone and to the public Internet. These perimeter nodes may be dedicated to Layer 3 routing functions or may be multi-service in nature, providing Layer 2 interconnects between data centers as well as Layer 3 services. WAN/PE routers validated within the VMDC reference system architecture include: the Cisco CRS-1, Cisco ASR 9000, Cisco Catalyst 7600, Catalyst 6500, and ASR 1000.

In this release, the Network layer includes a two-layer Clos spine and leaf arrangement of switching nodes, or the traditional three-layer hierarchical model described in previous releases. In the VMDC 3.0.1 reference architecture, the Network layer comprises of Nexus 7000 systems, serving as the spine and aggregation-edge nodes, and the Nexus 5000 systems as the leaf and access-edge nodes. As shown in Figure 2-3, Figure 2-4, and Figure 2-5, validated VMDC 3.0 topologies feature three variants of the FabricPath models, allowing for fine-tuning of redundancy, port capacity, and bandwidth to the level of service aggregation or access density required by current and anticipated scale requirements.

Services Layer

The Services layer comprises network and security services, such as firewalling, server load balancing, SSL offload, intrusion prevention, network analysis, and gateway functions. A distinct difference arises between the conventional data center services layer and "cloud" data center services layer in that the solution set for the latter must support application of Layer 4 - Layer 7 services at a per-tenant level, through logical abstraction of the physical resources. Centralized services are most useful in applying policies that are broadly applicable across a range of tenants (or workgroups in the private case).

In the VMDC reference architecture, the Data Center Services Node (DSN) provides firewalling and server load balancing services, in a service module form factor (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 DSN fulfills this function, securing remote tenant access to cloud resources.

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 Layer 2 network across multiple physical compute systems. This virtual access switching layer is key as it also logically extends the Layer 2 network to individual virtual machines within physical servers. The feature rich Cisco Nexus 1000V generally fulfills this role within the architecture. Depending on the level of software functionality (such as, QoS or security policy) or scale required, the VM-FEX may serve as a hardware-based alternative to the Nexus 1000V.

A second subsystem is virtual services (vApp-based), which may include security, load balancing, and optimization services. Services implemented at this layer of the infrastructure will complement more centralized service application and uniquely apply to a specific tenant or workgroup and their applications. Specific vApp based services validated within the VMDC architecture as of this writing include the Cisco Virtual Security Gateway (VSG), providing a security policy enforcement point within the tenant virtual data center or Virtual Private Data Center (VPDC). The third subsystem within the Compute layer is the computing resource. This subsystem includes physical servers, hypervisor software providing compute virtualization abilities, and the virtual machines. 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, comprises the compute resources utilized within the VMDC reference architecture.

Storage Layer

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

Management Layer

The Management layer consists of the "back-end" hardware and software resources required to manage the multi-tenant infrastructure. These resources include domain element management systems, as well as higher level service orchestration systems. The domain management systems currently validated within VMDC include Cisco UCS Manager, VMware vCenter, and vCloud Director for compute resource allocation; EMC UIM and Cisco Fabric Manager for storage administration; and Cisco VSM and Virtual Network Management Center (VNMC) for virtual access and virtual services management. Network analysis functionality is provided by Network Analysis Modules (NAMs) residing within Nexus 1010 systems. Automated service provisioning, including cross-resource service orchestration, are provided by BMC Cloud Lifecycle Management (CLM). However, service orchestration functions 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. As a homogenous modular unit of network, compute, and storage resources, the pod concept allows one to address environmental, physical, logical, and application-level requirements in a consistent way. The pod serves as a blueprint for incremental build-out of the Cloud data center in a structured fashion; when resource utilization within a pod reaches a 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.

In practice, the pod concept may serve simply as a framework, with designers defining their own 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 might have an access pod, terminating at the access switching nodes within an infrastructure and one might have a compute pod, addressing only the compute or the compute and storage portions of the infrastructure. Special purpose pods may be defined around application requirements or operational functions. For example, within the VMDC reference architecture, a management pod, referred to as a Virtual Management Infrastructure (VMI) pod is defined to physically and logically separate back-end management resources from production resources.

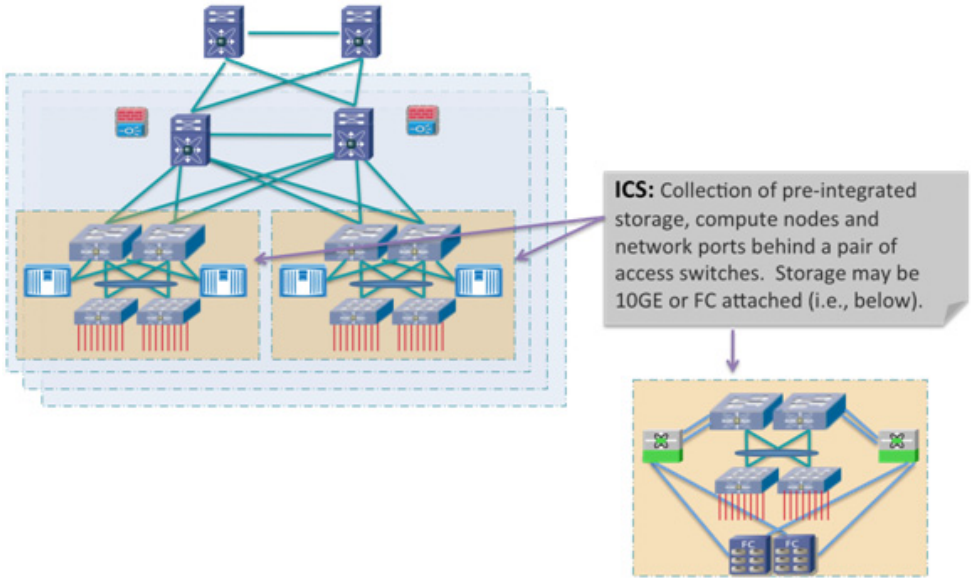
Typically, within the VMDC reference architecture, a general purpose utility compute pod extends from the compute and storage layers to the Layer 2 ports on the aggregation nodes serving as the Layer 2/Layer 3 boundary and up to and including components within the network services layer. Thus in a traditional hierarchical topology model, the port and MAC address capacity of the aggregation nodes are key factors in determining scale, in that they limit the number of pods a single pair of aggregation nodes will support within the Cloud data center. A key benefit of a Clos-type architectural model is that it broadly expands overall port capacity and bandwidth within the layer 2 (pod) domain; however, MAC

address support on the Layer 3 aggregation-edge or access-edge nodes will be a consideration in terms of host scale per pod (where a pod is a single FabricPath domain).

Integrated Compute Stacks

An Integrated Compute Stack (ICS) represents another potential unit of modularity in the VMDC Cloud data center, representing a sub-component within the pod. An ICS is a pre-integrated collection of storage, compute, and network resources, up to and including Layer 2 ports on a pair of access switching nodes. Figure 3-2 shows the location of the ICS within a pod. Multiples of ICSs are deployed like building blocks to fill the capacity of a pod.

Figure 3-2 ICS Concept



Working with eco-system partners, Cisco currently supports two ICS options: a Vblock and a FlexPod.

- A Vblock comprises Cisco UCS and EMC storage systems, offered in several combinations to meet price, performance, and scale requirements.
- Alternatively, a FlexPod comprises UCS compute and NetApp storage resources.

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

Data Center Interconnect

Within the VMDC reference architecture, pods may be interconnected between Data Centers using various data center interconnection methods, such as Overlay Transport Virtualization (OTV), xPLS, or LISP. Though not in scope for this release, these technologies have been tested and the resulting analysis is available as part of the larger body of 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 may be utilized concurrently over the VMDC Nexus-based infrastructure. It should be noted that as FCoE uses FSPF (Fabric Shortest Path First) forwarding, which is not supported over FabricPath today (it uses an IS-IS control plane), FCoE must be transported on separate (classical Ethernet) VLANs. In this VMDC release, 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 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 this as in previous system releases, VMware's vSphere plays a key role, enabling the creation of virtual machines on physical servers by logically abstracting the server environment in terms of CPU, memory, and network touch points into multiple virtual software containers.
2. Unified Computing System (UCS): unifying network, server, and I/O resources into a single, converged system, the Cisco UCS provides a highly resilient, low-latency unified fabric for the integration of lossless 10-Gigabit Ethernet and FCoE functions with x-86 server architectures. The UCS provides a stateless compute environment that abstracts I/O resources and server personality, configuration and connectivity, facilitating dynamic programmability. Hardware state abstraction makes it easier to move applications and operating systems across server hardware.
3. The Cisco Nexus 1000V provides a feature-rich alternative to VMware's Distributed Virtual Switch, incorporating software-based VN-link technology to extend network visibility, QoS, and security policy to the virtual machine level of granularity. This system release uses VMware vSphere 5.0 as the compute virtualization operating system. A complete list of new enhancements available with vSphere 5.0 is available online. Key "baseline" vSphere features leveraged by the system includes ESXi boot from SAN, VMware High Availability (VMware HA), and Distributed Resource Scheduler (DRS). Fundamental to the virtualized compute architecture is the notion of clusters; a cluster consists of two or more hosts with their associated resource pools, virtual machines, and data stores. Working in 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 and/or DRS features are utilized. In general practice, however, the larger the scale of the compute environment and the higher the virtualization (VM, network interface, and port) requirement, the more advisable it is to use smaller cluster sizes to optimize performance and virtual interface port scale. Therefore, in VMDC large pod simulations, cluster sizes are limited to eight servers; in smaller pod simulations, cluster sizes of 16 or 32 are used. As in previous VMDC releases, three compute profiles are created to represent large, medium, and small workload: “Large” has 1 vCPU/core and 16 GB RAM; “Medium” has .5 vCPU/core and 8 GB RAM; and “Small” has .25 vCPU/core and 4 GB of RAM.

The UCS-based compute architecture has the following characteristics:

- It comprises multiple UCS 5100 series chassis (5108s), 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.
- The UCS is a fully redundant system, with two 2200 Series Fabric Extenders per chassis and two 6200 Series Fabric Interconnects per pod.
- Internally, eight uplinks per Fabric Extender feed into dual Fabric Interconnects to pre-stage the system for the maximum bandwidth possible 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 G (16 x 10 G) of uplink bandwidth capacity must be provided per UCS system.
- Four ports from an FC GEM in each 6200 Expansion Slot provide 8 Gig FC to the Cisco MDS 9513 SAN switches (for example, 6200 chassis A, 4 x 8G FC to MDS A and 6200 chassis B, 4 x 8G FC 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

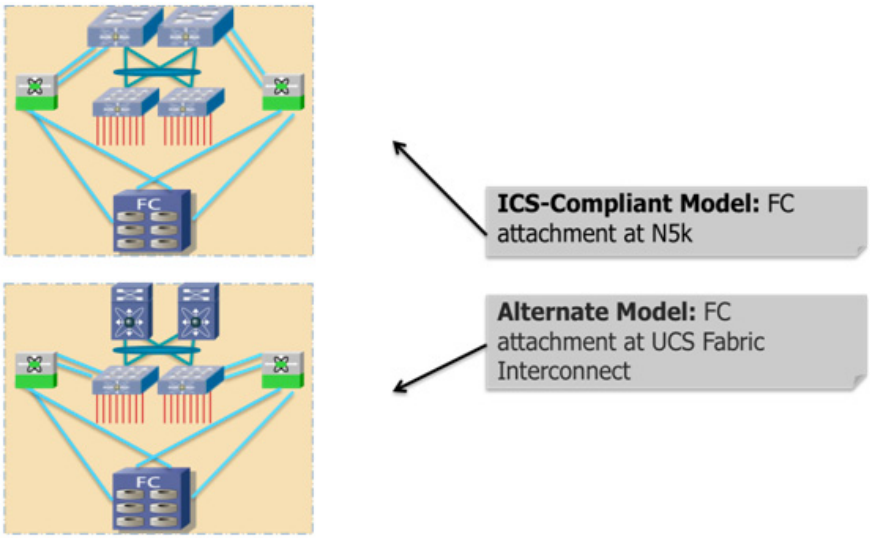
The VMDC SAN architecture remains unchanged from previous (2.0) programs. It follows current best practice guidelines for scalability, high availability, and traffic isolation. Key design aspects of the architecture include:

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

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

1. Model that follows the ICS model of attachment via the Nexus 5000 and/or the Nexus 7000 (depending on ICS type).
2. Model that provides for an attachment at the UCS Fabric Interconnect.

Figure 3-3 SAN FC Attachment



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

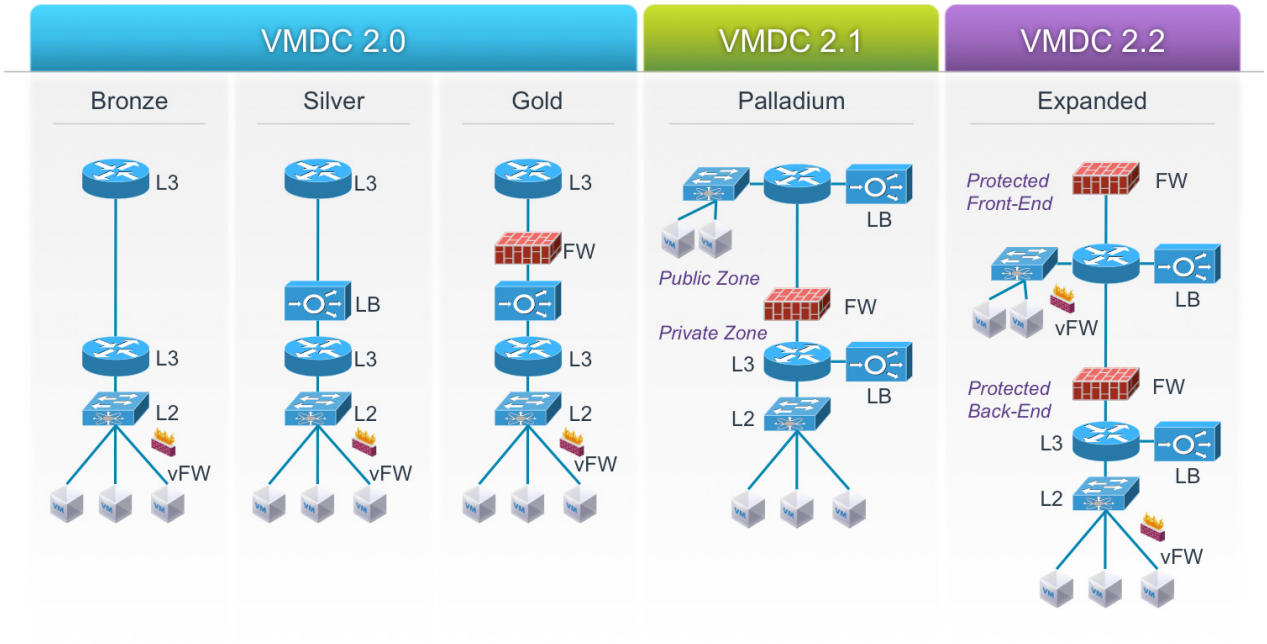
Although Figure 3-3 shows a simplistic SAN switching topology, it is important to note that if greater SAN port switching capacity is required, the architecture supports (and has been validated with) more complex, two-tier core-edge SAN topologies, as documented in the VMDC 2.0 "Compact Pod Implementation 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.

Container Models

Virtualization of compute and storage resources enables sharing across an organizational entity. In contrast, virtualized multi-tenancy, a concept at the heart of the VMDC reference architecture, refers to the logical isolation of shared virtual compute, storage, and network resources. In essence, this is "bounded" or compartmentalized sharing. A tenant is a user community with some level of shared affinity. For example, within an enterprise, a tenant may be a business unit, department, or workgroup. Depending upon business requirements or regulatory policies, a tenant "container" may stretch across physical boundaries, organizational boundaries, and even between corporations.

A tenant container may reside wholly within their private cloud or may extend from the tenant's enterprise to the provider's facilities within a public cloud. The VMDC architecture addresses all of these tenancy use cases through a combination of secured data path isolation and a tiered security model that leverages classical security best practices and updates them for the virtualized multitenant environment. Earlier VMDC releases (2.0 through 2.2) presented six tenancy models. High-level, logical depictions of five of these models are illustrated in Figure 3-4.

Figure 3-4 Validated Cloud Tenancy Models



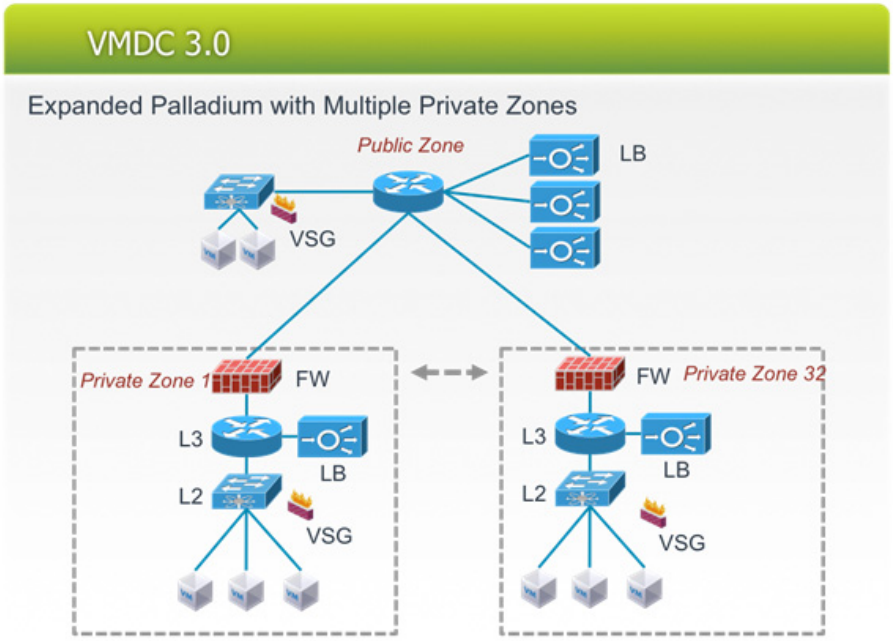
With the notion of a separate front-end and back-end set of zones, each of which may have a different set of network services applied, the Palladium container begins to more closely align with traditional zoning models in use in physical IT deployments, versus earlier VMDC models. Note that a fundamental assumption in this case is that infrastructure security (IPS, etc.) and secured access is implemented north of this tenant container.

As previously mentioned, due to its alignment with private cloud deployment models, this release employs a modified and updated version of the Palladium container. Figure 3-5 shows this modified model, were there is:

- A single, shared (multi-tenant) public zone, with multiple server VLANs and a single ACE context (or multiple contexts) for SLB. This is in the global routing table.
- Multiple, private (unique per-tenant or user group) firewalled zones reachable via the public zone – i.e., the firewall "outside" interface is in the public zone. These private zones include an ACE SLB, and may have 1 to many VLANs.
- A VSG can be applied in a multi-tenant/shared fashion to the public zone.
- A VSG can be applied in dedicated fashion to each of the private zones, providing a second tier of policy enforcement, and back-end (E/W) zoning. Unique VLANs may be used per zone for VLAN-based isolation. However, in validation we assumed the desire to conserve VLANs would drive one to use a single VLAN with multiple security zones applied for policy-based isolation.

An alternative way to view this model is as a single, DC-wide "tenant" with a single front-end zone and multiple back-end zones for (East/West) application-based isolation.

Figure 3-5 VMDC 3.0/3.0.1 Tenant Container



Network

Network considerations are detailed in the following sections:

Layer 3 Design

In VMDC 3.0/3.0.1 a combination of dynamic and static routing is used to communicate reachability information across the layer three portions of the infrastructure. In this design dynamic routing is achieved using OSPF as the IGP. Aggregation-edge nodes functioning as ASBRs use OSPF to advertise learned host routes across the IP core to the WAN Edge/PE routers. To scale IP prefix tables on these aggregation-edge nodes, they are placed in stub areas with the IP core advertising "default route" (Type 7) for reachability. Service appliances (ASA Firewall and ACE) are physically connected directly to the aggregation-edge nodes; reachability to/from these appliances are communicated via static routes. In the case of clustered ASA firewalls (i.e., as in VMDC 3.0.1), for traffic from the ASA(s) to the Nexus 7000 aggregation-edge nodes, a default static route points to the HSRP VIP on the Nexus 7000, while for traffic from the Nexus 7000 aggregation-edge to the ASA, a static route on the Nexus 7000 for server subnets points to the ASA inside IP interface address.

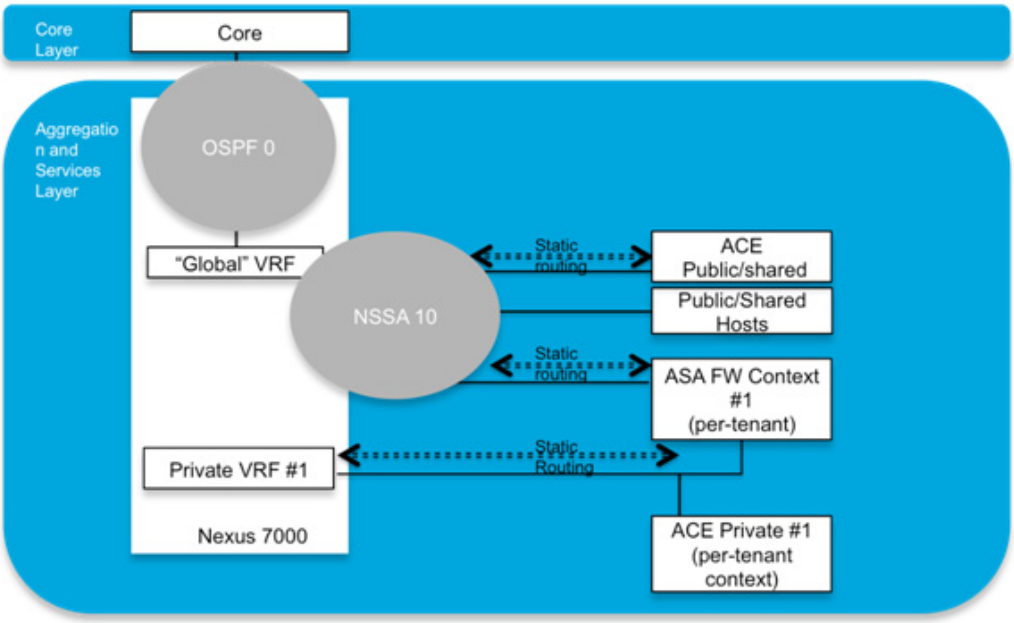
In the "Typical Data Center" design, the ACE appliance is configured in one-arm mode. This has several key benefits: it limits the extension of FabricPath VLANs to the appliances; keeps the VLAN ARP entries off the ACE; and the port-channel method of attachment allows for a separation of failure domains. Source-NAT on the ACE insures symmetric routing. By contrast, in the "Extended Switched

Data Center" the ACE appliance is configured in two-arm mode, providing for more optimal traffic flows (i.e., reducing traffic hairpinning) given that the ACE resides within a C6500 (Data Center Services Node chassis) in this model.

VRF-lite implemented on the aggregation-edge nodes provides a unique per-tenant VRF, serving to further secure and isolate private tenant applications and zones via dedicated routing and forwarding tables. Figure 3-6 shows a high level diagram of the layer three implementation for the Typical DC for a single tenant.

Note Redundant aggregation-edge node and appliances are present, but not shown.

Figure 3-6 Logical Layer 3 Topology (Typical DC)



The Typical Data Center design features a two-node Layer 3 spine (aka aggregation-edge nodes). In this model, active/active gateway routing is enabled through the use of vPC+ on the inter-Spine (FabricPath) peer-link. This creates a single emulated switch from both spine nodes. HSRP thus announces the virtual MAC of the emulated switch ID, enabling dual-active paths from each access-edge switch device, serving to optimize resiliency and throughput, while providing for efficient East/West routing.

In the Extended Data Center Design, GLBP provides four gateways for a 4-wide spine, across which traffic may be load-balanced between pods or between intra-campus buildings. Members of a GLBP group elect one gateway to be the active virtual gateway (AVG) for that group. Other group members provide backup for the AVG in the event that the AVG becomes unavailable. The AVG assigns a virtual MAC address to each member of the GLBP group. Each gateway assumes responsibility for forwarding packets sent to the virtual MAC address assigned to it by the AVG. These gateways are known as active virtual forwarders (AVFs) for their virtual MAC address. GLBP provides load balancing over multiple routers (gateways) using a single virtual IP address and multiple virtual MAC addresses.

In this model, priorities have been set such that aggregation-edge 1 is the AVG for the hosts in pod 1 and the backup for hosts in pod 2 in Building 1; a similar set of priorities is applied to the gateways in Building 2.

This release leveraged GLBP as a "first look" at potential issues regarding use of 4-wide gateways, particularly as two-gateway implementations remain the prevalent deployment model. Looking forward, it should be noted that with NX-OS release 6.2, Anycast FHRP will become the recommended solution for 4 (or more) gateway use cases.

Fabric Path

Cisco FabricPath comprises a new Layer 2 data plane. It does this by encapsulating the 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 to which the frame is heading. For this reason, switch IDs must be unique within the FabricPath domain; these are either automatically assigned (default) or set manually by the administrator (recommended). From there the frame is routed until it reaches the remote switch, where it is de-encapsulated and delivered in its original Ethernet format.

A fundamental aspect of FabricPath is that it uses an IS-IS control plane for establishing Layer 2 adjacencies within the FabricPath core; ECMP is therefore possible and spanning tree is no longer required within this type of layer 2 fabric for loop avoidance. Loop mitigation is addressed via TTL (Time to Live – decremented at each switch hop to prevent looping) and RPF checks for multi-destination traffic. Therefore as previously noted, a common initial use case for FabricPath is as part of a strategy to minimize reliance on Spanning Tree within the Data Center.

Today a FabricPath domain comprises a single logical topology. As part of the establishment of Layer 2 adjacencies across the logical topology, FabricPath nodes create two multidestination trees. These are computed automatically by IS-IS calculations. The highest priority switch is chosen as the root for the first multidestination tree (FTAG1), which is used for broadcasts and flooding and multicast. The second highest priority switch is chosen as the root for the second multidestination tree (FTAG2) which is used for multicast. The designs discussed in this document 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, in which case, two of the dedicated Layer 2 spines serve as roots for the FabricPath domain. Should a root fail, the switch with the next highest priority will take over as root.

If devices that are part of non-FabricPath Layer 2 domains (i.e., 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 are summarized below:

- Two spine nodes, aggregating multiple leaf nodes (i.e., mirroring commonly-deployed hierarchical DC topologies).
- Routing at the Spine (aka Aggregation-edge) nodes. Together with the preceding bullet point, this provides for ease of migration from a traditional hierarchical deployment to a FabricPath deployment. The aggregation nodes now serve not just as the traditional Layer 2/Layer 3 boundary providing routed uplinks for North/South (routed) flows, but also as FabricPath spine nodes.
- FabricPath core ports at the spine (F1s and/or F2/F2Es) provide bridging for East/West intra-VLAN traffic flows.

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

- Leaf nodes (aka access-edge switches) provide pure layer two functions, with FabricPath core ports facing the aggregation layer switches.
- Classical Ethernet edge ports face all hosts.
- Layer 2 resilience design options utilized within this layer of the infrastructure comprise use of vPC+ on the inter-spine peer-link to implement active/active HSRP, ECMP, port-channels between agg-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
- Additionally, in VMDC 3.0.1 VPCs provide resilience from ACE 4710 appliances and clustered ASAs to aggregation-edge nodes.

Key design aspects unique to the FabricPath portion of the Extended Switched Data Center design include:

- Multiple spine nodes, (four Nexus 7000 switches with F1 linecards in the SUT) operating as dedicated Layer 2-only spines. As noted, this provides for dedicated bridging for East/West (inter-pod and/or inter-building) intra-VLAN traffic flows. Since these nodes are used solely for fabric path switching, the FabricPath ports are all "core" ports, meaning that they only send and receive FabricPath-encapsulated traffic, do not run Spanning Tree Protocol (STP), and do not perform MAC learning.
- Two spine (also called aggregation-edge) nodes, aggregating multiple leaf nodes per building (at two buildings). As in the case of the Typical DC design, these four total nodes serve as the Layer 2/Layer 3 boundary for the FabricPath domain. However in this case GLBP rather than HSRP was utilized as the FHRP for gateway redundancy.

Currently, the Nexus 7000 supports three categories of fabric path I/O modules – the N7K-F132XP-15 (NX-OS 5.1), the N7K-F248XP-25 (NX-OS 6.0), and the new N7k-F248XP-25E (NX-OS 6.1). Tcan be used for FabricPath core ports. However, the F1 card only supports Layer 2 forwarding while the F2 and F2E cards support both Layer 2 and Layer 3 forwarding. These considerations are particularly applicable to the spine node configurations in the architectural models discussed in this document.

For example, with respect to the Aggregation-Edge (Layer 3 spine) nodes; in a "Typical DC" design wherein the Nexus 7000 aggregation-edge node is configured with M1 and F1 line cards in a single VDC forming the Layer 2/Layer 3 boundary, F1 cards perform Layer 2 forwarding functions, serving to connect FabricPath "core" VLANs. In this scenario the M1 card provides Layer 2 SVIs or interfaces to the Layer 3 core of the infrastructure via routed ports, also performing proxy routing as required for packets received on F1 interfaces. This mixed-VDC scenario has the benefit of ease of deployment. However, a key consideration as of this writing (i.e., for pre-NX-OS 6.2 code releases) is that such designs will have a maximum MAC address constraint of 16,000. Proxy-Layer 2 learning functionality (targeted at NX-OS release 6.2) will allow the full M1 XL MAC address table size of 128k to be leveraged for M1/F1 mixed scenarios.

F2 or F2E-only scenarios (i.e., 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 again, as of this writing the 16,000 maximum MAC address constraint applies to this model.

In contrast, in the "Extended DC" design the Nexus 7000 aggregation-edge nodes are configured with M1 and F1 (or alternatively, F2 or F2E) line cards, each in their own VDC. This is currently a required deployment model for M1/F2 or M1/F2E scenarios, but also has the advantage of offloading MAC learning from the M1 cards, they will simply learn ARPs on the Layer 3 interfaces. In this model the M1 VDC uses 802.1q with Layer 3 sub-interfaces for the routing function, while the F1/F2 VDC is configured with FabricPath forwarding and segments VLANs on specific SoCs, which are port-channeled to the M1 VDC. The scale is controlled through the number of port-channels created between the two VDCs.

With respect to the Access-Edge (leaf) nodes in the referenced models, Nexus 5548 (or 5596s) with FEX 2200s for port expansion provide TOR access. Alternatively, Nexus 7000s with F1 (or F2) line cards (and 2232 FEX-based port expansion) may perform this function, for EOR access into the fabric.

The Nexus 5500 currently supports up to 24 FEX modules. If using the Nexus 2232PP this would allow. for 768 edge ports per Nexus 5500 edge pair. Thus traffic oversubscription can be impacted greatly with increased FEX usage. Currently 4 FabricPath core facing port-channels with 4 members each are supported on the N5500.

The 6200 series Fabric Interconnects are connected to FabricPath edge nodes via HM-vPC today. FabricPath is on the roadmap but beyond the scope of this release.

As noted, one of the new Layer 2 resilience features introduced with FabricPath is vPC+. This provides a means for devices that do not support FabricPath to be attached redundantly to two separate FabricPath switches without resorting to Spanning Tree Protocol. Like vPC, vPC+ relies on PortChannel technology to provide multipathing and redundancy. Configuring a pair of vPC+ edge nodes creates a single 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.

In both designs, port-channels, rather than single links are used with equal cost multipath (ECMP) for access-edge to aggregation-edge core connections, providing enhanced resilience in the event of a single link member failure. As this is not default behavior after NX-OS 5.2.4, an IS-IS metric must be configured on the port-channel to insure that individual member link failures in port-channels are transparent to the IS-IS protocol.

Services

VMDC 3.0 incorporated physical appliance based services, Data Center Service Node (DSN) service module services, and virtual service form-factor offerings.

VMDC 3.0 considered the implications of appliance-based methods of attachment and the VSS-based model of attachment. VMDC 3.0 validation focused on the first model, whereas VMDC 3.0.1 expanded validation scope to cover the VSS attachment option. Services may be localized (i.e., intra-PoD, as in previous VMDC systems releases), distributed (i.e., across several pods or buildings) or centralized.

The Typical Data Center topology illustrated in Figure 3-8 and the service placement diagram in Figure 3-9 provide examples of localized and centralized service application, while the Extended Data Center topology diagrams illustrate service application which is either centralized (Figure 3-7) or distributed across several pods (Figure 3-8).

Figure 3-7 Extended Switched DC—Centralized Service Attachment

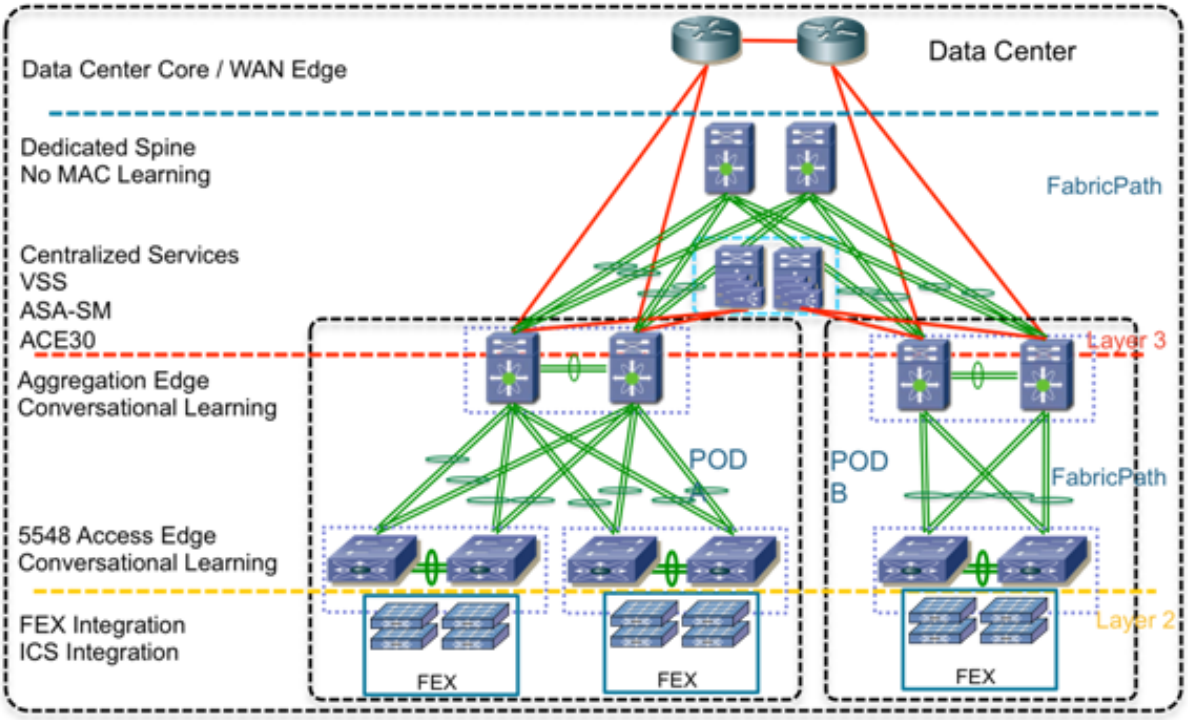
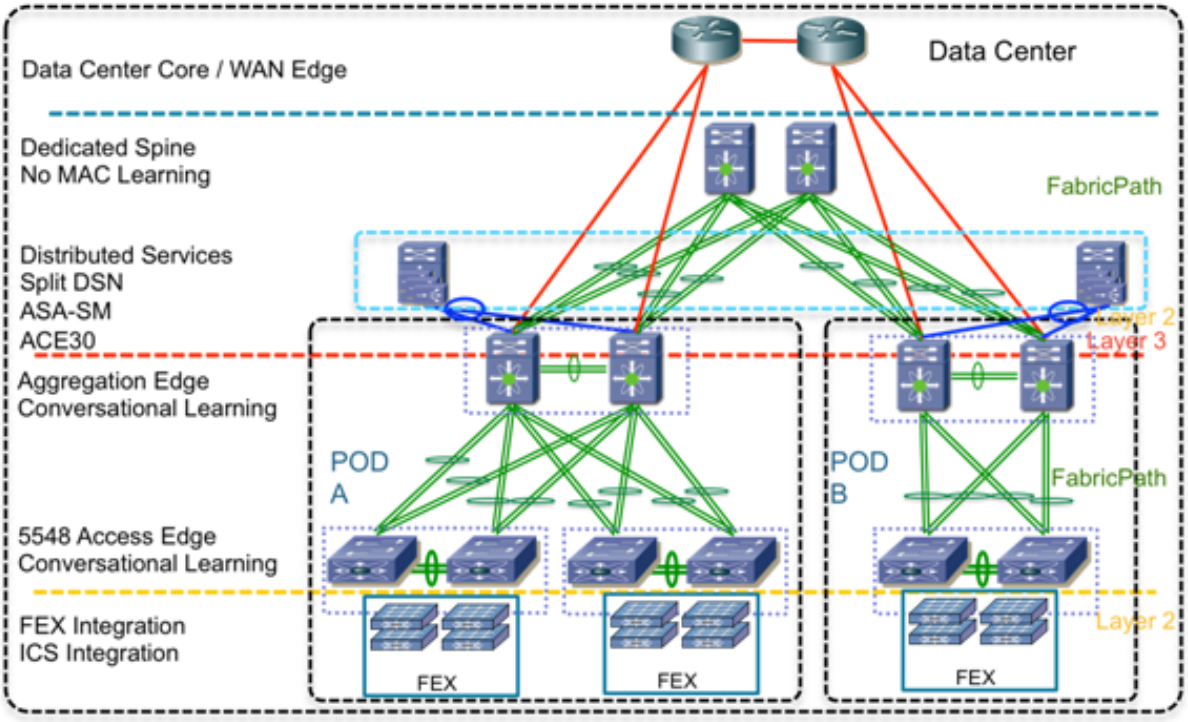


Figure 3-8 Extended Switched DC—Distributed Service Attachment



The ACE 4710 SLB and ASA 5585 firewall appliances were used in the Typical DC (service core) to provide load balancing and front-end/first-tier firewalling. In the context of localized service placement options, a key addition to the architecture in VMDC 3.0.1 is the introduction of clustered ASA Firewalls (Release 9.0+). This feature serves two functions: resilience and capacity and throughput expansion. Up to eight Cisco ASA 5585-X or 5580 Adaptive Security Appliance firewall modules may be joined in a single cluster to deliver up to 128 Gbps of multiprotocol throughput (300 Gbps maximum) and more than 50 million concurrent connections. This is achieved via the Cisco Cluster Link Aggregation Control Protocol (cLACP), which enables multi-system ASA clusters to function and be managed as a single entity. This provides significant benefits in terms of streamlined operation and management, in that firewall policies pushed to the cluster get replicated across all units within the cluster, while the health, performance and capacity statistics of the entire cluster may be managed from a single console.

Clustered ASA appliances can operate in routed, transparent, or mixed-mode. However, all members of the cluster must be in the same mode. As previously noted, in this system release the clustered ASA appliances are deployed and validated in routed mode. However, transparent mode deployment considerations and methodology are discussed in this VMDC white paper: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VMDC/ASA_Cluster/ASA_Cluster.pdf.

Characteristics of localized appliance-based service attachment as implemented in the Typical DC model include:

- Classical Ethernet-based attachment to Nexus 7000 aggregation-edge nodes. For the VMDC 3.0 system release, in both cases, port-channels from each appliance to a single uplink aggregation-edge node were utilized to provide resilient connectivity, though for the ASA vPCs are an alternative and preferable option. In the vPC case, should a single aggregation-edge node fail, the ASA would still be able to pass traffic through the second aggregation-edge node, without requiring an ASA failover.
- In contrast, in VMDC 3.0.1, vPC attachment from clustered ASAs was leveraged to provide enhanced resilience. More specifically, one vPC (across all 4 clustered ASAs) to the N7k aggregation-edge nodes was utilized for data traffic, and multiple port-channels per ASA (to vPCs on the Nexus 7000 aggregation-edge nodes) were used for communication of cluster control link (CCL) traffic. Similarly, with respect to the ACE 4710, vPCs per ACE appliance to both redundant aggregation-edge nodes provided SLB resilience.
- As noted, the ACE is in “one-arm” mode to optimize traffic flows for load balanced and non-load balanced traffic. This limits the extension of FabricPath VLANs to the appliances, and keeps the VLAN ARP entries off the ACE.
- Active/Active Failover between redundant (non-clustered) appliances through configuration of active/standby pairs on alternating (primary/secondary) contexts. In contrast, the clustered resilience functionality available on the ASA is such that every member of the cluster is capable of forwarding every traffic flow and can be active for all flows.
- This design implementation follows current best-practice recommendations, using out-of-band links for FT state communication between redundant appliances. In the context of non-clustered, redundant ASA pairs, interface monitoring is activated to insure proper triggering of failover to a standby interface, only one interface (inside or outside) must be monitored per FT failover group, though monitoring of both is possible. Should it feature higher resilience characteristics, the management path between the redundant ASAs could be a monitoring option. For clustered ASA appliances, the CCL (Cluster Control Link) communicates control plane information between cluster members, including flow re-direction controls. This design follows best practice recommendations for CCL high availability by employing vPCs on the redundant N7k aggregation-edge from port-channels on each ASA in the cluster.

As noted, the appliances are statically routed, and redistributed into NSSA area 10. Key considerations for this service implementation include:

- **Disaster Recovery Implications**—Disaster recovery could be more costly and challenging to implement versus more distributed models, due to the requirement to provide a complete replication of the network services resources and associated subnets at the DR location for failover. Doing this on a pod by pod basis may simply not be economically feasible, unless it is possible to establish and coordinate 1:N failover.
- By definition, this design is optimized for intra-pod communication and traffic flows; inter-pod VLANs with services applied could feature sub-optimal (less efficient) traffic flows.
- On the positive side, resource allocation for compute, storage and network services applied to multi-tiered applications is less complex, in that it is pod-based. This should translate to simpler workflow and resource allocation algorithms for service automation systems.
- Pod-based deployments represent a simple, clear-cut operational domain, for greater ease of troubleshooting and determination of dependencies in root-cause analysis.

In the Extended Switched DC, ACE-30 and ASA service modules were utilized within the DSN to provide load balancing and front-end/first-tier firewalling. To emulate appliance-mode attachment, the redundant DSNs were not configured as a VSS pair, but rather as separate but redundant nodes. Characteristics of the distributed (emulated) appliance attachment as implemented include:

- vPC+ attachment to Nexus 7000 aggregation-edge nodes.
- Active/Active Failover between redundant appliances through configuration of active/standby pairs on alternating (primary/secondary) contexts.
- As previously noted, the ACE is in two-arm mode in this case, to optimize traffic flows for load balanced and non-load balanced traffic flows through the infrastructure nodes.
- Though it is possible to use FabricPath as inter-building transport for FT links as well as data paths, as in the Typical DC case, this design implementation followed the best practice recommendation of out-of-band links for FT state communication between the emulated appliances (in this case, an Layer 2 port-channel was implemented for this purpose between DSN nodes).
- Auto-state configuration on the DSN provides a similar effect to interface monitoring on the ASA, insuring that if there is a DSN failure the ASA state will parallel this.
- Appliances are statically routed.

Key considerations for this approach to service implementation are:

- This is a more complex model versus the localized scenario. With subnets and SVI next-hops distributed across Layer 3 aggregation-edge devices for inter-pod traffic, there is the potential for longer convergence times in failover and recovery, as well as non-optimal service traffic flows, depending on the failure scenario and where server resources are located.
- With respect to the FT path, note that the underlying premise is that a direct (point to point) out-of-band path will be of higher availability than the in-band path. However, depending upon deployment specifics, it may not be feasible to utilize separate fibers (or copper) for this purpose. In this case, preferential QoS treatment should be applied to insure sufficient bandwidth and preferential treatment for FT traffic through the FabricPath nodes in the event of congestion. This will help to minimize latency through the fabric.
- Coupled with inter-building storage replication solutions or distributed storage solutions, such as NetApp MetroCluster or EMC VPLEX, the distributed service implementation approach provides for service failover, in the event of facility outages.
- As stated, another option for service attachment is to centralize service nodes for use by multiple pods in the topology. In this model the Data Center Service nodes (DSNs) are configured as a VSS pair. Thus all services VLANs are contained in the VSS deployment (VSL), conserving the use of FabricPath VLANs (i.e., saving 3 VLANs per tenant). Key characteristics of this deployment model include:

  - The DSN operates in a Layer 3 configuration where all services traffic is routed.
  - Static routes or a dynamic routing protocol (IGP/BGP) may be used to communicate routes to or from the DSN.
  - Contexts can still be configured in active/active operation.

Key considerations for this approach to service implementation are:

- Service VLANs are offloaded from the FabricPath fabric.
- Service control plane traffic is contained within the VSS (VSL).
- Traffic flows from switching systems to service modules are localized. In this case, all traffic to/from services with any number of Pods attached is optimal. This translates to speedy reconvergence and service failover.
- Route Health Injection (RHI) may be leveraged for tuning of data center path selection.

Virtualization Techniques

VMware vSphere 5.0 is utilized as the tenant hypervisor resource in VMDC 3.0 and 3.0.1. Previous program releases leveraged vSphere 4.0 and 4.1. This covers integration 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 scope for this VMDC release, alternate hypervisors may be utilized over the infrastructure provided Cisco UCS is in their prospective Hardware Compatibility List. As of this writing, the Nexus 1000V distributed virtual switch supports only vSphere; however, these alternate hypervisor VMs can connect at the FEX or primary access layer, and participate in appliance based or DSN based services.

System Level Design Considerations

The following system level design considerations are defined:

Scalability

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

- **VLAN Scale**: As of this writing (in NX-OS releases 5.2.5 through 6.1) a maximum of 2000 FabricPath-encapsulated VLANs is supported. This figure will be improved in subsequent releases. However, it is important to note that this by definition is a one-dimensional figure, which does not factor in inter-related (Layer 2 to Layer 3) end-to-end traffic flow considerations such as FHRP constraints per module or per node. In practice, overall system VLAN scaling will be constrained by the effect of ARP learning rates on system convergence and FHRP (HSRP or GLBP) groups per

module or interface, and per node. Regarding the latter, HSRP support per module is currently 500 and 1000 per system, with aggressive timers, or 2000 per system, with default timers; GLBP is 200 per module and 500 per system, with aggressive timers, or 1000 per system, with default timers.

- **Switches per FabricPath Domain**: NX-OS 5.2 supports a maximum of 64 switch ids; NX-OS 6.0 a maximum of 128.
- **Port Density per FabricPath Node**: At 48 ports per module, the F2 line cards provide up to 768 10 or 1 GE ports per switch (N7018), while the F1 cards provide up to 512 10GE ports (N7018). Again, these are uni-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**: All FabricPath VLANs use conversational MAC address learning. Conversational MAC learning consists of a three-way handshake. This means that each interface learns only those MAC addresses for interested hosts, rather than all MAC addresses in the VLAN. This selective learning allows the network to scale beyond the limits of individual switch MAC address tables. Classical Ethernet VLANs use traditional MAC address learning by default, but the CE VLANs can be configured to use conversational MAC learning.

  Despite the advantages of conversational learning for scale within the fabric, MAC address capacity does represent a scale factor on Layer 3 spine (aggregation) or leaf (access) nodes at the edges of the FabricPath domain. This is due to the fact that edge switches maintain both MAC address and Switch ID tables. Ingress switches use the MAC table to determine the destination Switch ID; egress switches may use the MAC table to determine output switchport. This release leveraged two scenarios for implementation of the M1/F1 cards on the Nexus 7000 aggregation-edge switches. In the mixed VDC case (M1/F1 in a single VDC), at present one must consider MAC address scale of 16,000 on each F1 SoC (Switch-on-Chip). There are 16 forwarding engines or SoCs on the F1 card. The MAC scale will be increased in the NX-OS 6.2 release where MAC-Proxy will leverage the XL table sizes on the M1 card and the MAC address capacity will become 128,000.

  The split VDC approach enables higher MAC scaling by separating the functional roles into two separate VDCs (as of this writing the F2 card requires a separate VDC for deployment so this model fits well in F2 deployment cases). The F1/F2 VDC is configured with FabricPath forwarding and segments VLANs on specific SoCs which are port-channeled to the M1 VDC. The scale is controlled with the number of port-channels created between the two VDCs. The M1 VDC uses 802.1q with Layer 3 sub-interfaces for the routing function. The M1 linecards no longer learn MAC addresses, rather ARPs on the Layer 3 interfaces.

  Note The ARP capacity on the M1 card is 1 million. Effectively, in this scenario MAC capacity is limited only by the card distribution and number of ports available for intra-chassis port channels. However, end-to-end MAC capacity must be factored in, and ARP learning rates as well.

  MAC capacity on the Nexus 5500 (layer 2) access-edge nodes is 24,000.
- **ARP Learning Rate**: As noted, ARP learning rates on layer 3 edge nodes affect system convergence for specific failure types. ARP learning rates of 100/second were observed on the Nexus 7000 aggregation-edge nodes during system validation. With tuning, this was improved to 250-300/second.
- **Tenancy**: The tenancy scope for the SUT was 32. However this does not represent the maximum scale of the architecture models. Within the models addressed in this release, several factors will constrain overall tenancy scale. These are - 1) VRFs per system. Currently, up to 1000 VRFs are supported per Nexus 7000 aggregation edge node, but then additional factors include 2) End-to-end VLAN support (i.e., affected by FHRP (HSRP or GLBP) groups per card and per system; and 3) 250 contexts per ASA FW appliance – one may increment this up by adding appliances if needed.

Availability

The following methods are used to achieve High Availability within the VMDC Data Center architecture:

- Routing and Layer 3 redundancy at the core and aggregation/leaf nodes of the infrastructure. This includes path and link redundancy, non-stop forwarding and route optimization.
- In the "Typical Data Center" (2-node spine topology) VPC+ is configured on inter-spine peer-links and utilized in conjunction with HSRP to provide dual-active paths from access edge switches across the fabric.
- Similarly, in the "Extended Switched Fabric Datacenter" topology with 4-wide aggregation-edge nodes, GLBP is utilized to distribute routed traffic over 4 aggregation edge nodes. Looking forward, in NX-OS 6.2, Anycast FHRP will be the preferred option for four or greater redundant gateways.
- Layer 2 redundancy technologies are implemented through the FabricPath domain and access tiers of the infrastructure. This includes ARP synchronization in VPC/VPC+-enabled topologies to minimize flooding of unknown unicast and reconvergence; ECMP; utilization of port-channels between FabricPath edge/leaf and spine nodes to minimize Layer 2 IS-IS adjacency recalculations; and IS-IS SPF tuning, CoPP, GLBP and HSRP timer tuning on aggregation edge nodes, again to minimize system reconvergence.
- Active/Active (active/standby of alternating contexts) on services utilized in the architecture.
- Clustered HA and ECLB (equal cost load balancing) for appliance-based firewall services.
- Hardware and Fabric redundancy throughout.
- (VEM) MCEC uplink redundancy and VSM redundancy within the virtual access tier of the infrastructure.
- Within the compute tier of the infrastructure, port-channeling, NIC teaming and intra-cluster HA through utilization of VMware VMotion.

Security

The security framework from the VMDC 2.1 and 2.2 systems are leveraged for tenancy separation and isolation. Security related considerations include:

- **Aggregation Layer (Layer 3) Separation**—VRF-lite implemented on aggregation-edge nodes at the aggregation layer provides per tenant isolation at Layer 3, with separate dedicated per-tenant routing and forwarding tables on the inside interfaces of firewall contexts. All inter-tenant traffic has to be routed at the outside interfaces on the Firewall that resides in the global VRF. Policies can be applied on the firewall to restrict inter-tenant communication.
- **Access and Virtual Access Layer (Layer 2) Separation**—VLAN IDs and the 802.1q tag provide isolation and identification of tenant traffic across the Layer 2 domain, and more generally, across shared links throughout the infrastructure.
- **Network Services Separation (Services Core, Compute)**—On physical appliance or service module form factors, dedicated contexts or zones provide the means for virtualized security, load balancing, NAT, and SSL offload services, and the application of unique per-tenant policies at the VLAN level of granularity. Similarly, dedicated virtual appliances (i.e., in vApp form) provide for unique per-tenant services within the compute layer of the infrastructure at the virtual machine level of granularity.

- **Storage**—This VMDC design revision uses NetApp for NFS storage, which enables virtualized storage space such that each tenant (application or user) can be separated with use of IPspaces and VLANs mapped to network layer separation. In terms of SANs, this design uses Cisco MDS 9500 and EMC VMAX for Block Storage. This allows for Fiber Channel (FC) access separation at the switch port level (VSAN), Logical path access separation on the path level (WWN/Device Hard Zoning), and at the virtual media level inside the Storage Array (LUN Masking and Mapping).

Manageability

This section addressed service provisioning and orchestration. This architecture leverages BMC Cloud Lifecycle Management for automated Service Orchestration. 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.

Service Assurance and Monitoring

Service assurance is generally defined as the application of policies and processes ensuring that services offered over networks meet a pre-defined service quality level for an optimal subscriber experience. The practice of service assurance enables providers to control traffic flows and identify faults and resolve those issues in a timely manner so as to minimize service downtime. The practice also includes policies and processes to proactively diagnose and resolve service quality degradations or device malfunctions before subscribers are impacted.

In VMDC network service assurance encompasses the following concepts:

Traffic Engineering

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

Port-channels are frequently deployed for redundancy and load sharing capabilities. Since the Cisco Nexus 1000V Series is an end-host switch, the network administrator can use a different approach than can be used on a physical switch, implementing a port-channel mechanism in either of two modes:

- **Standard Port-Channel**—The port-channel is configured on both the Cisco Nexus 1000V Series and the upstream switches.
- **Special Port-Channel**—The port-channel is configured only on the Cisco Nexus 1000V Series, with no need to configure anything upstream. Two options are available, MAC Pinning and vPC Host Mode.

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

For more information on the 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 utilizes 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 may be utilized for a more granular approach to uplink traffic management on the Nexus 1000V. These options are shown in Figure 3-9 and Figure 3-9.

Figure 3-9 Nexus 1000v single Uplink PortChannel Model

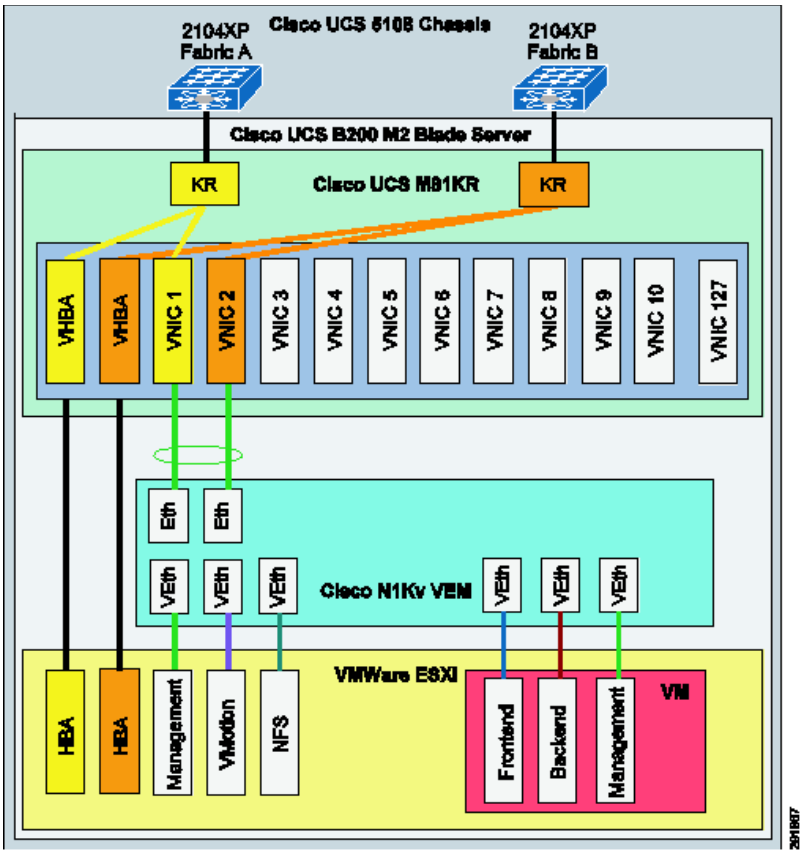
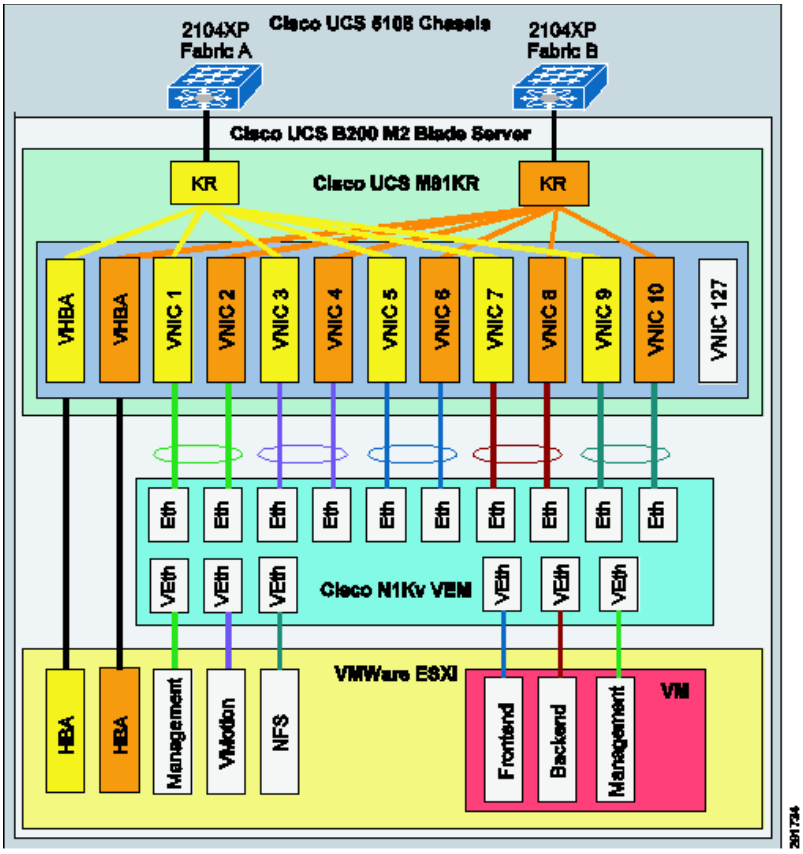


Figure 3-10 Nexus 1000v 5 Uplink PortChannel Model



Traffic engineering can be performed selectively by configuring the Nexus 1000V to select the target uplink with 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, back-end traffic, which has more diversity in terms of bandwidth/response time usage (VM-to-VM - inter-fabric traffic flows, vMotion, backup, and so forth) may benefit by selecting a path such that it allows VM-to-VM traffic to remain within a single fabric where the Fabric Interconnect switches the traffic locally. Table 3-1 lists key architectural features of VMDC 3.0.

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
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

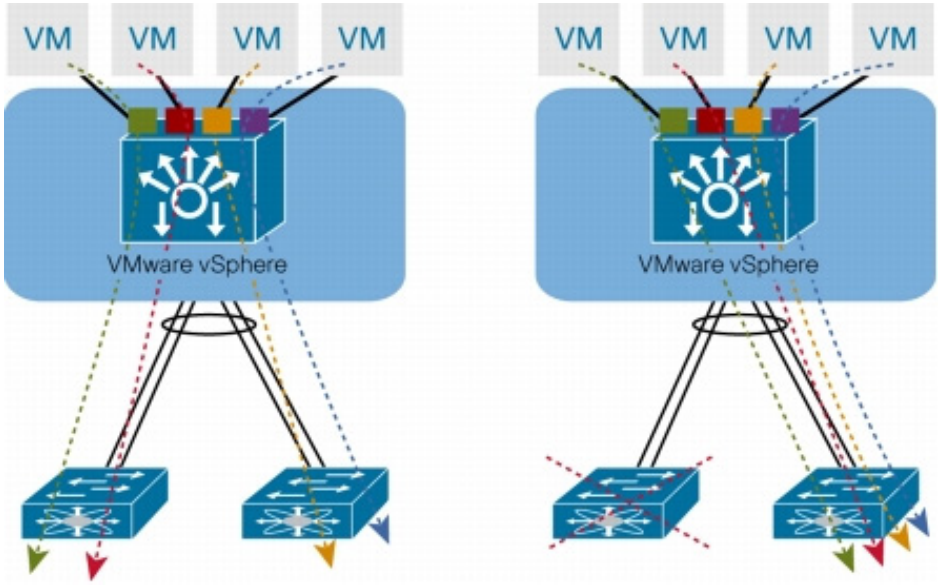
MAC pinning defines all the uplinks coming out of the server as standalone links and pins different MAC addresses to those links in a round-robin fashion. This approach helps ensure that the MAC address of a virtual machine will never be seen on multiple interfaces on the upstream switches. Therefore, no upstream configuration is required to connect the Cisco Nexus 1000V Series VEM to the upstream switches (Figure 3-11).

Furthermore, MAC pinning does not rely on any protocol to distinguish the different upstream switches, making the deployment independent of any hardware or design.

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 Address Resolution Protocol (ARP) packet to alert the upstream switch that the MAC address of the VEM learned on the previous link will now be learned on a different link, enabling failover in less than a second.

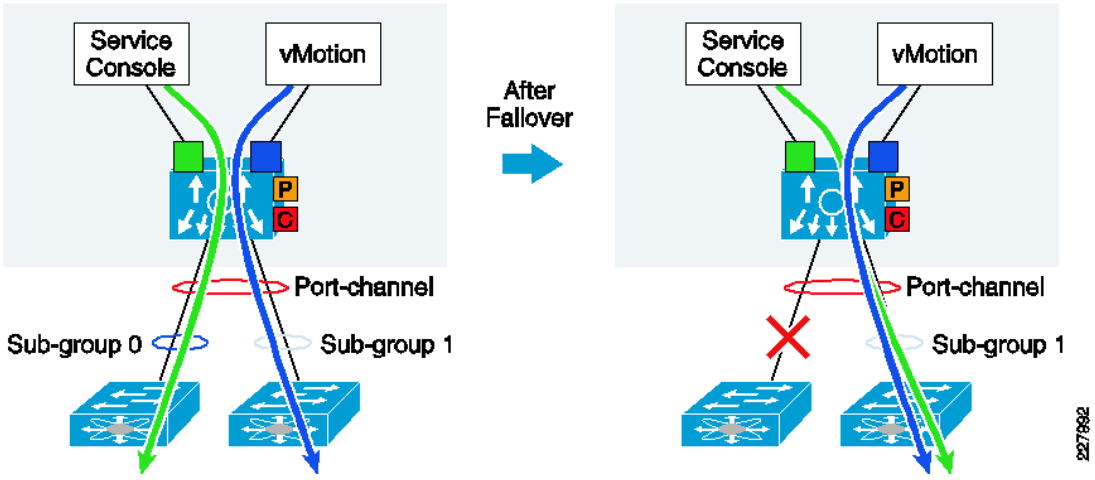
MAC pinning enables consistent and easy deployment of the Cisco Nexus 1000V Series since it does not depend on any physical hardware or any upstream configuration, and it is the preferred method for deploying the Cisco Nexus 1000V Series if the upstream switches cannot be clustered.

Figure 3-11 MAC-Pinning Details



In the case of a fabric failure the Nexus 1000 selects the available remaining fabric to recover the traffic. Figure 3-12 shows the fabric failover with sub-group mac-pining.

Figure 3-12 MAC-Pinning Failover



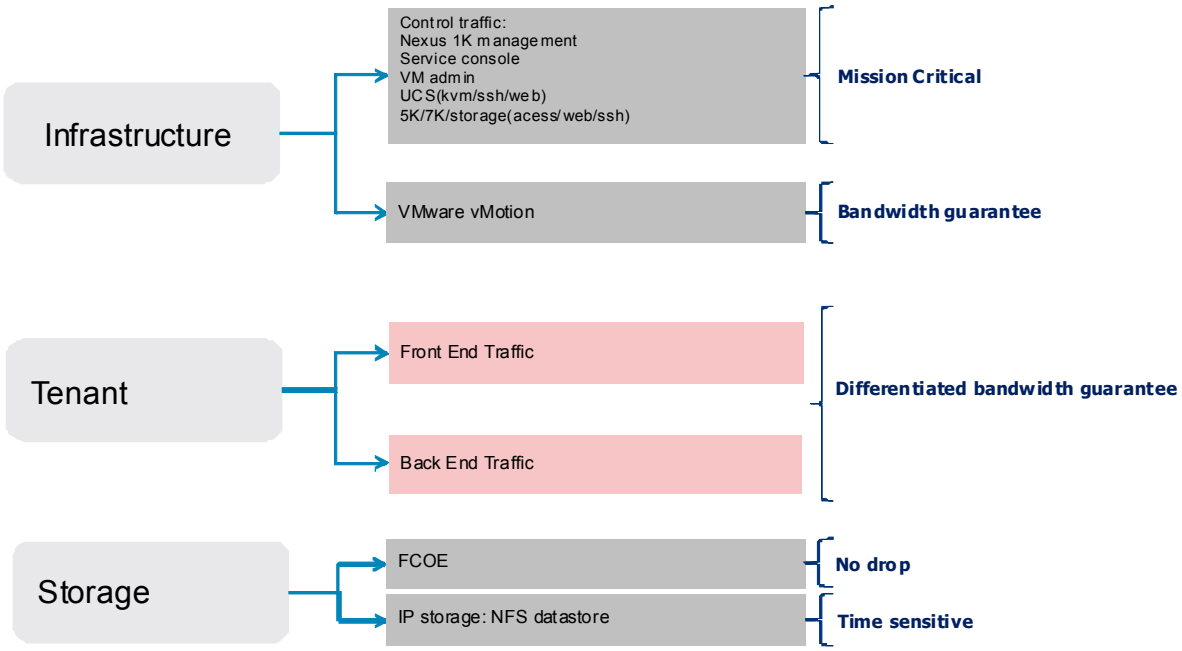
Quality of Service Framework

Quality of QQoS is a key to service assurance because it enables differentiated treatment of specific traffic flows. This differentiated treatment ensures that in the event of congestion or failure conditions, critical traffic is provided sufficient amount bandwidth to meet throughput requirements.

Figure 3-13 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 communications during periods of instability or high CPU utilization.
- Tenant traffic may be differentiated into Front End and Back End Traffic with service levels to accommodate various types of traffic requirements in each category.
- The VMDC design incorporates both FC and IP-attached storage. As indicated in Figure 3-13, storage requires two sub-categories, since these traffic types are treated differently throughout the network. FC traffic by definition requires a “no drop” policy, while NFS datastore traffic is sensitive to delay and loss.

Figure 3-13 Traffic Flow Types



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

- Traffic Classification and Marking, page 3-25
- Congestion Management and Avoidance (Queuing, Scheduling, and Dropping, page 3-26)
- Traffic Conditioning (Shaping and Policing, page 3-28)

Classification and Marking

Classification and marking allow QoS-enabled networks to identify traffic types based on information in source packet headers (i.e., Layer 2 802.1p CoS and DSCP information) and assign specific markings to those traffic types for appropriate treatment as the packets traverse nodes in the network. Marking (coloring) is the process of setting the value of the DSCP, MPLS EXP, or Ethernet Layer 2 CoS fields so that traffic can easily be identified later, i.e. using simple classification techniques. Conditional marking is used to designate in-contract (i.e., "conform") or out-of-contract (i.e., "exceed") traffic.

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

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

Figure 3-14 shows a more granular breakdown of the requisite traffic classes characterized by their DSCP markings and per-hop behavior (PHB) designations. This represents a normalized view across the VMDC and Hosted Collaboration Solution (HCS) validated reference architectures in the context of an eight-class IP/NGN aligned model.

Figure 3-14 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 datacenter QoS models, CoS 3 is reserved for lossless data (FCoE). However, in older WAN/Campus QoS services models, CoS 3 is used for VOIP signaling. The table above assumes that FCOE traffic will be localized to the UCS and Ethernet-attached Storage systems, thus enabling the use of CoS 3 for VoIP signaling traffic within the DC QoS domain. Classification values may need to be tweaked per traffic characteristics: for example CoS value 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 the network design. However, if the end system is not capable of marking or cannot be trusted, one may mark on ingress to the network. In the 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 will mark traffic, and these markings will be trusted at the nodes within the data center infrastructure; in other words, they will use simple classification based on the markings received from the edge devices. Note that where VM-FEX adapters are utilized, marking is implemented on the 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 the DSCP values will not be altered. Intermediate nodes would ideally support QoS transparency, such 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.

Queuing, Scheduling, and Dropping

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

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

- Weighted bandwidth

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

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

One of the challenges in defining an end-to-end QoS architecture is that not all nodes within a QoS domain have consistent implementations. Within the cloud data center QoS domain, we run the gamut from systems that support 16 queues per VEM (i.e., Nexus 1000V) to four internal fabric queues (i.e., Nexus 7000). This means that traffic classes must be merged together on systems that support less than eight queues. Figure 3-15 shows the class to queue mapping that applies to the cloud data center QoS domain in the VMDC 2.2 reference architecture, in the context of alignment with either the HCS reference model or the more standard NGN reference.

Figure 3-15 VMDC Class to Queue Mapping

VMDC 8 class model	COS	VMDC HCS Aligned 8 Class Model	VMDC NGN Aligned 8 Class Model	VMDC (61x0) 6 class model	HCS 6 class model	4 class model N7k fabric
Network Mgmt + Service control	7	Network Mgmt + VM control	Network Mgmt + VM control	Network Mgmt (COS 7) + Service control (COS 7) + Network control (COS 6)	Network Mgmt (COS 7) + Service control (COS 7) + Network control (COS 6)	Queue 4
Network control	6	Network control	Network control			
Priority #1	5	Voice bearer	Res VoIP / Bus Real-time	Priority #1	Voice bearer	Queue 1
Bandwidth #1 (Priority 2)	4	Interactive Video	Video streaming	Bandwidth #1	Interactive Video	
Bandwidth #2	3	Call Control	Video interactive / FCOE	FCOE (Bandwidth #2)	Call Control	Queue 2
Bandwidth #3 "Gold"	2	FCOE	Bus critical in-contract (COS 2) Bus critical out-of-contract (COS 1)*	Bus critical in-contract (COS 2) Bus critical out-of-contract (COS 1)*	FCOE	
Bandwidth #4 "Silver"	1	Webex collaboration data (interactive)	Silver	Silver	Webex collaboration data + Standard data	Queue 3
Standard (Bandwidth #5) "Bronze"	0	Standard data	Standard data	Standard		

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

Note that the Nexus 2000 Fabric Extender provides only **two** user queues for its quality of service (QoS) support, one for all no-drop classes and one for all drop classes. The classes configured on its parent switch are mapped to one of these two 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 two classes. Further, as of this writing (NX-OS 6.1.3), when connected to an upstream Nexus 7000 switch queueing is not supported on Nexus 2000 Host Interface ports: traffic is sent to the default fabric queue on the Nexus 7000, and queuing must be applied on FEX trunk (Network Interface) ports. Upcoming 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.

Additionally, prior to 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 techniques used to enforce a maximum bandwidth rate on a traffic stream; while policing effectively does this by dropping out-of-contract traffic, shaping does this by delaying out-of-contract traffic. VMDC utilizes policing within 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 Layer 2 traffic is shaped at the aggregate (port) level per class, while policing is utilized to enforce per-tenant aggregates.

Sample bandwidth port reservation percentages are shown in Figure 3-16.

Figure 3-16 Sample Bandwidth Port Reservations

Traffic Class	CoS	N7k M1 Egress Queue	N7k F1 Egress Queue	N5k QoS Class Maps	N5k Queuing Class Maps	N5k QoS Groups	N5k % BW
Mgmt	7	1p7q4t-out-pq1	1p3q1t-8e-out-pq1	class-vmdc-priority	vmdc-pq	5	5
Network Control	6	1p7q4t-out-pq1	1p3q1t-8e-out-pq1	class-vmdc-priority			
VoIP	5	1p7q4t-out-pq1	1p3q1t-8e-out-pq1	class-vmdc-priority			
Video, NAS	4	1p7q4t-out-q2	1p3q1t-8e-out-q2	class-vmdc-p2	vmdc-p2	4	10
Call Control	3	1p7q4t-out-q2	1p3q1t-8e-out-q2	class-fcoe	class-fcoe	1	
Premium Data	2	1p7q4t-out-q3	1p3q1t-8e-out-q3	class-vmdc-p3	vmdc-p3	3	60
" "	1	1p7q4t-out-q-default	1p3q1t-8e-out-q-default	class-vmdc-p4	vmdc-p4	2	1
Standard Data	0	1p7q4t-out-q-default	1p3q1t-8e-out-q-default	class-default	class-default	0	24

Network Analysis

The use of network analysis devices is another service readily available in the VMDC design. The Cisco Nexus 1000v NAM VSB is integrated with the Nexus 1010 Virtual Services Appliance to provide network and performance visibility into the Nexus 1000V switching deployment. The NAM VSB uses

embedded instrumentation, such as Netflow and Encapsulated Remote SPAN (ERSPAN) on the Nexus 1000V switch as the data source for traffic analysis, application response time, interface statistics, and reporting.

For more information on the Cisco Prime NAM for Nexus 1010 deployment follow the link below:

http://www.cisco.com/en/US/docs/net_mgmt/network_analysis_module_virtual_blade/4.2/install/guide/nexus/nx42_install.html

NetFlow

NetFlow was developed by Cisco to provide better insight into the IP traffic on the network. NetFlow defines flows as records and exports these 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 a traditional network environment, but the Nexus 1000v provides this capability within the virtual network environment. Nexus 1000v supports NetFlow v9 and by default will use the management 0 interface as an export source.

Caution The use of advanced features such as NetFlow will consume additional resources (i.e., memory and CPU, of your ESX host). It is important to understand these resource dynamics before enabling any advanced features.

Figure 3-17 shows the Cisco NetFlow Collector reporting application statistics on the virtual Ethernet interfaces that reside on the Nexus 1000v. The Nexus 1000v may also monitor flows from the physical interfaces associated with the platform and VMkernel interfaces including VMotion traffic as seen in Figure 3-18.

Figure 3-17 Cisco Netflow Collector Application Statistics Example

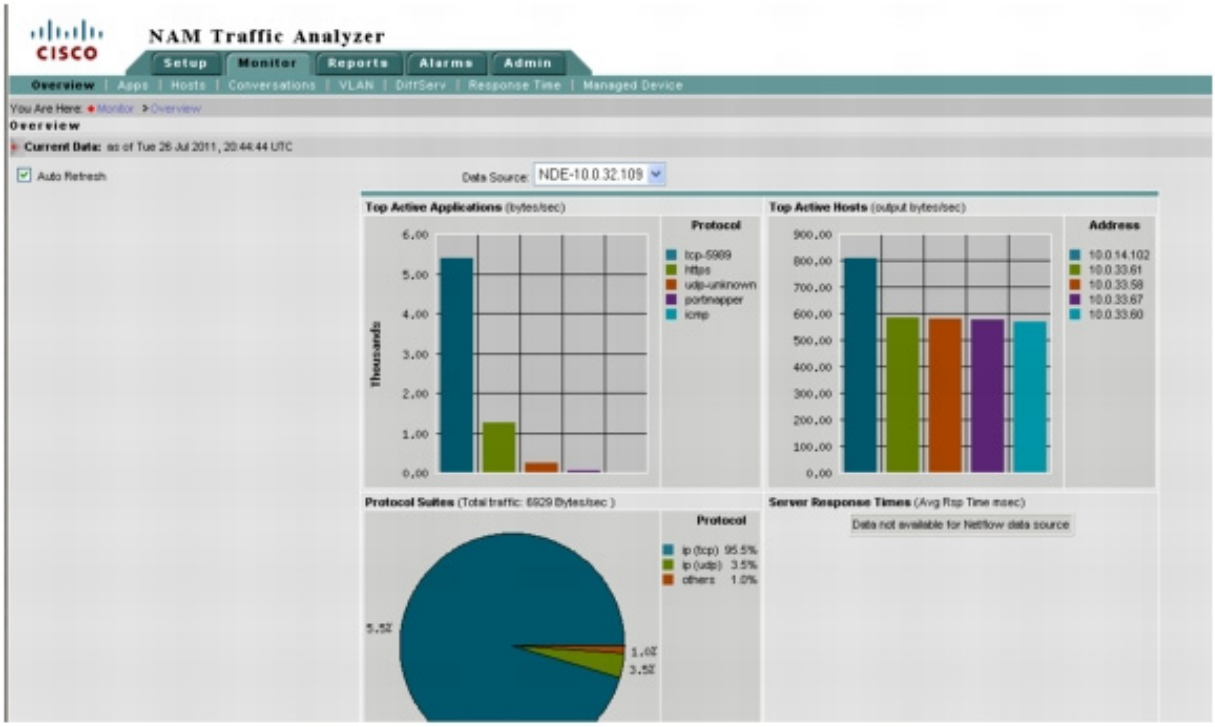
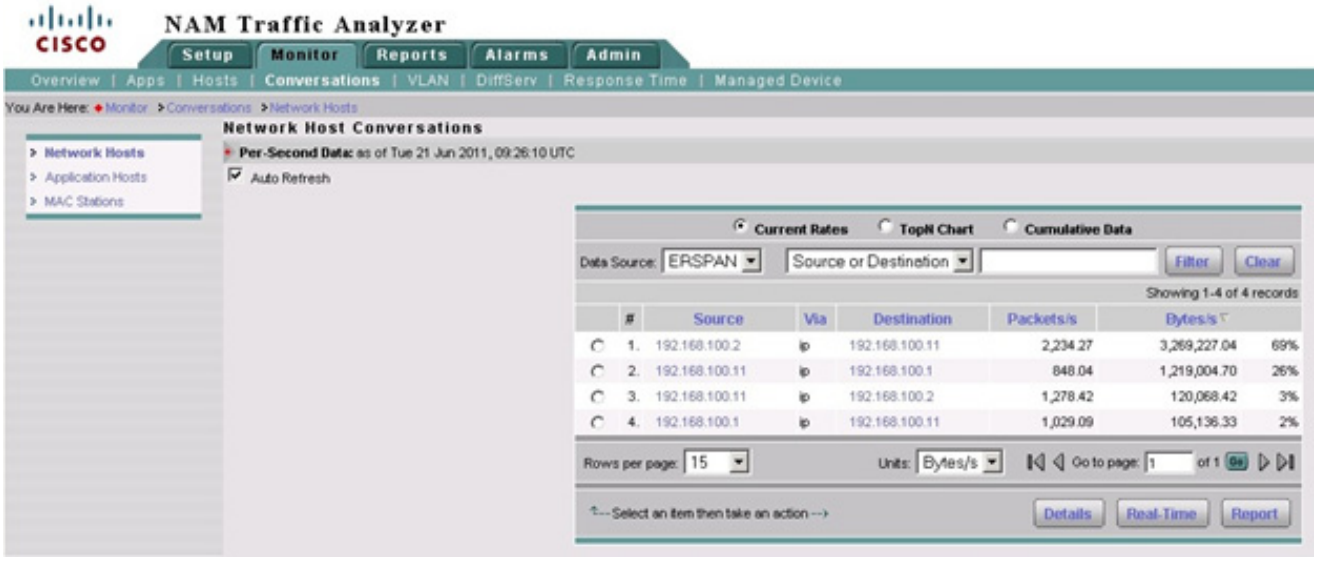


Figure 3-18 Cisco Netflow Collector Nexus 1000v vMotion Results Example



Encapsulated Remote Switched Port Analyzer (ERSPAN)

ERSPAN allows for remote monitoring of network resources. ERSPAN uses GRE tunnels to route traffic to the appropriate destination. The Nexus 1000v supports ERSPAN, allowing network administrators to observe the traffic associated with the following:

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

This flexibility allows the ERSPAN session to not only monitor data associated with virtual machines, but to monitor all traffic associated with the ESX host including VMkernel, VMotion, and service console data. Converging all of these traffic types onto two or a maximum of four CNAs per-ESX host simplifies not only the physical design of the data center but the configuration of the capture points as well.

In the validation of this solution, the final destination for ERSPAN traffic was the Virtual Network Analysis Module (vNAM) resident in Nexus 1010.

For more information on configuring ERSPAN on the Nexus 1000v follow:

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 The use of advanced features such as ERSPAN will consume additional resources (i.e., memory and CPU of the ESX host). It is important to understand these resource dynamics before enabling any advanced features.

Figure 3-19 and Figure 3-20 show examples of a packet decode and application performance metrics available from the ERSPAN data.

Figure 3-19 View of NAM Captured Data from VM NIC

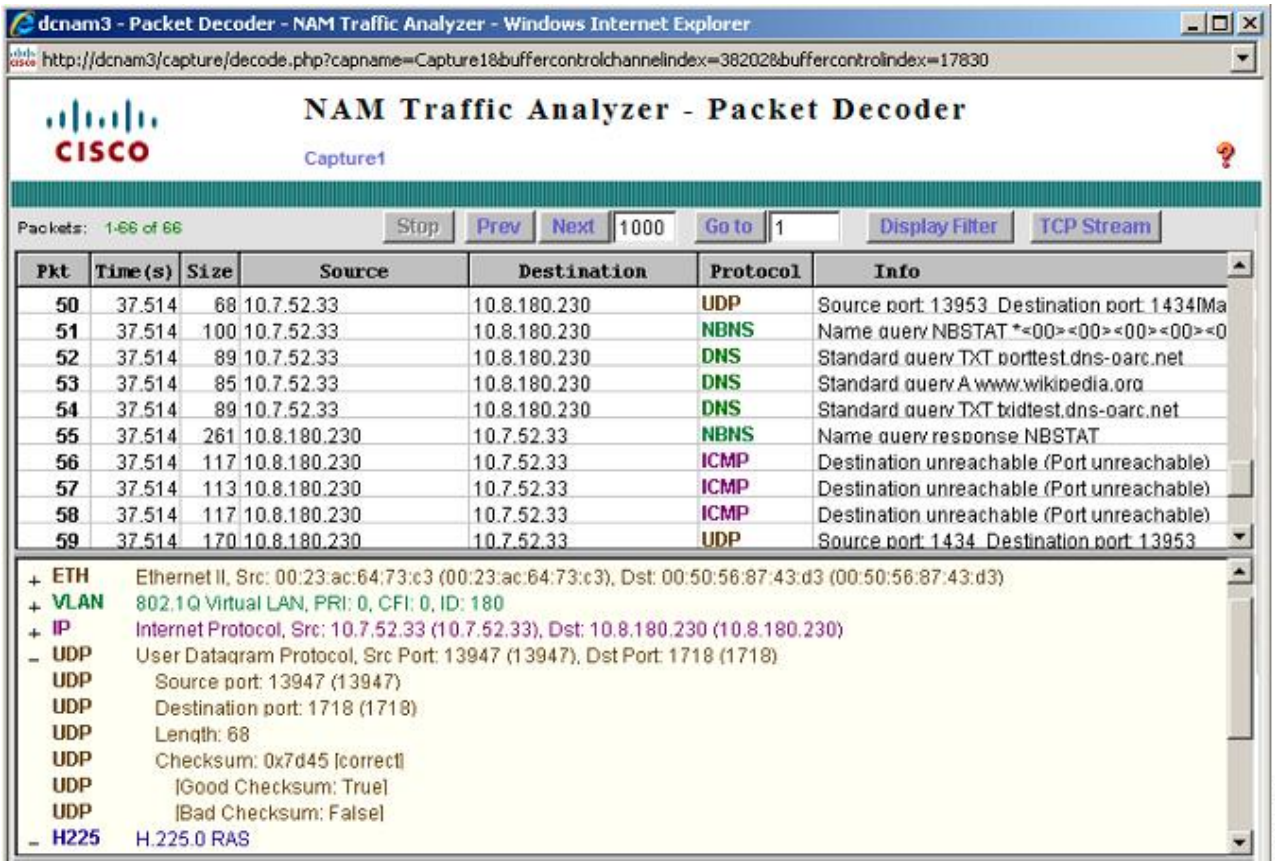
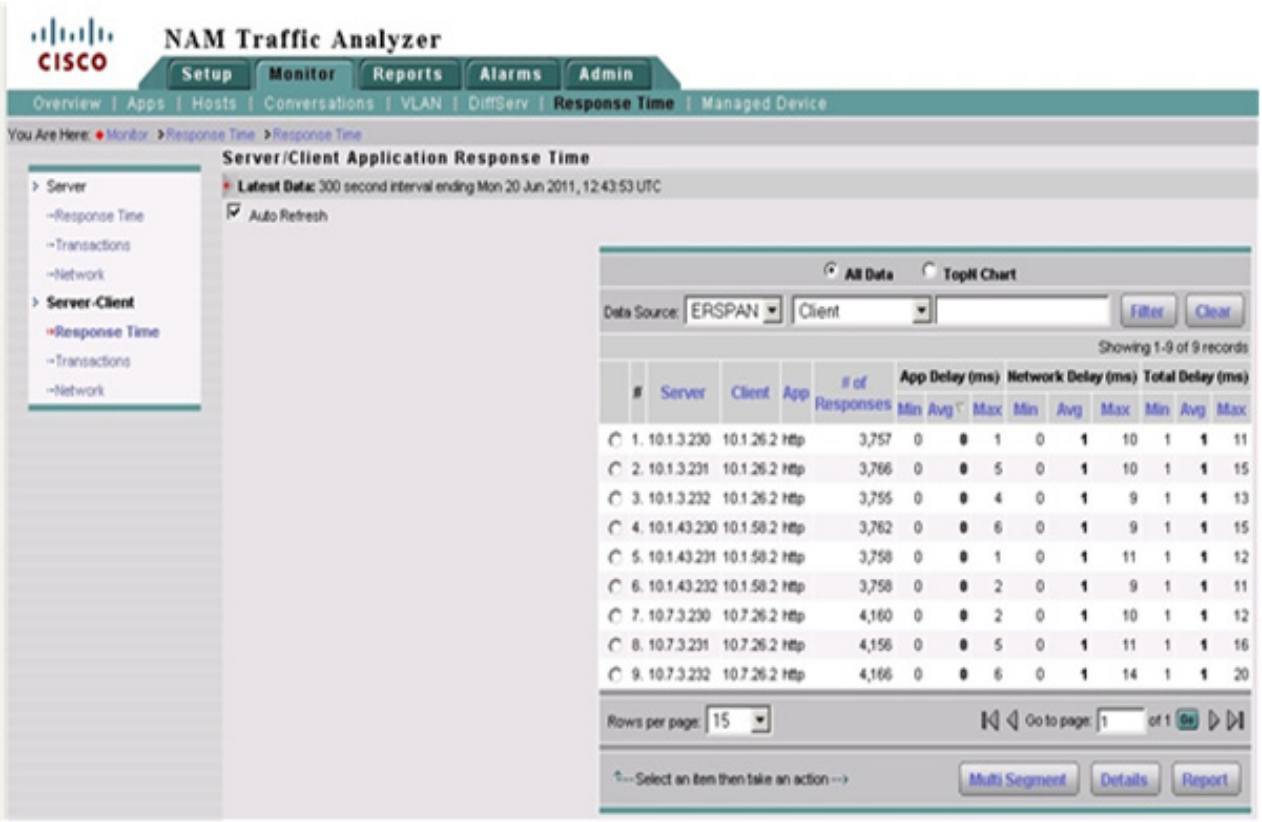


Figure 3-20 Application Response Time Data Collected On N1KV VEM Uplink



CLSA-VMDC (Cloud Service Assurance for VMDC)

Based on the Zenoss Cloud Service Assurance solution, CLSA-VMDC provides a service-impact model-based system providing tenant-based service assurance, including consolidated monitoring of the VMDC infrastructure and simple, easily-deployed plug-ins for customization. The system offers real time aggregated dashboards as well as reporting capabilities. It can be deployed both in centralized and distributed architecture, and allows for incremental deployment growth. While it offers rich functionality for IaaS domains, the solution is lightweight and has open interfaces to allow for simple integration into existing Operations Support System (OSS) and ticketing systems with minimal cost. As such, 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.