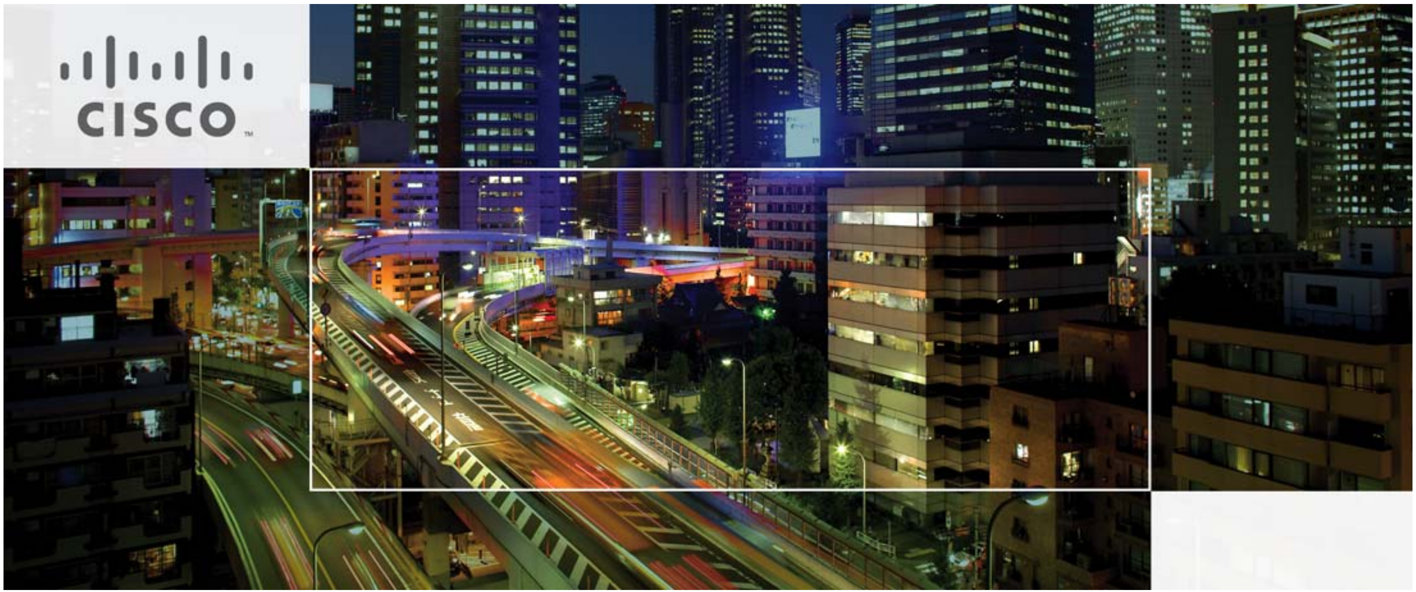


Virtualized Multiservice Data Center (VMDC) 2.3 Implementation Guide

April 6, 2013

Cisco Validated Design

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

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

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

Virtualized Multiservice Data Center (VMDC) 2.3 Implementation Guide

CONTENTS

Preface

This preface explains the objectives and intended audience of the Cisco® Virtualized Multiservice Data Center (VMDC) 2.3 solution and outlines the organization of the VMDC 2.3 Implementation Guide.

Product screen shots and other similar material in this document are used for illustrative purposes only and show trademarks of EMC Corporation (VMAX), NetApp, Inc. (NetApp FAS3240), and VMware, Inc. (vSphere). All other marks and names mentioned herein may be trademarks of their respective companies.

Use of the word "partner" or "partnership" does not imply a legal partnership relationship between Cisco and any other company.

Implementation Guide Purpose

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

This guide describes implementation details for a reference architecture that brings together core products and technologies from Cisco, NetApp, EMC, BMC, and VMware to deliver a comprehensive end-to-end cloud solution. Focused on IaaS cloud deployment, the Cisco VMDC solution provides customers with robust, scalable, and resilient options for cloud Data Center (DC) deployments.

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

This solution addresses the following problems:

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

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

As Enterprise IT departments evolve, they are looking for a DC solution that is efficiently shared, secured, and rapidly provisioned. Similarly, Service Providers are looking for solutions that enable them to reduce TTM for new revenue-generating services and reduce ongoing OPEX. The Cisco VMDC infrastructure design provides a model for flexible sharing of common infrastructure, maintaining secure separation of tenant data and enabling per-tenant differentiated services.

This guide provides the implementation and configuration details, feature overviews, and best practices needed to deploy a VMDC 2.3-based cloud DC.

Audience

This guide is intended for, but not limited to, system architects, network/compute/storage design engineers, systems engineers, field consultants, advanced services specialists, and customers who want to understand how to deploy a public or private cloud DC infrastructure. This guide assumes that the reader is familiar with the basic concepts of IP protocols, Quality of Service (QoS), High Availability (HA), Layer 4 (L4) - Layer 7 (L7) services, DC platforms and technologies, SAN and VMware hypervisor. This guide also assumes that the reader is aware of general system requirements and has knowledge of Enterprise or Service Provider network and DC architectures and platforms and virtualization technologies.

Document Organization

Table 1-1 provides the organization of this document.

Table 1-1 Document Organization

Topic	Description
Chapter 1, “Implementation Overview”	This chapter provides an overview of the VMDC solution.
Chapter 2, “Compute and Storage Implementation”	This chapter discusses compute and storage implementation for this solution.
Chapter 3, “Layer 2 Implementation”	This chapter discusses Layer 2 implementation for this solution.
Chapter 4, “Layer 3 Implementation”	This chapter discusses Layer 3 implementation for this solution.
Chapter 5, “Services Implementation”	This chapter discusses services implementation for this solution.
Chapter 6, “QoS Implementation”	This chapter discusses QoS implementation for this solution.
Chapter 7, “Resiliency and High Availability”	This chapter discusses resiliency and redundancy for this solution.
Chapter A, “Best Practices and Caveats”	This appendix lists the recommended best practices and caveats when deploying this solution.
Chapter B, “Related Documentation”	This appendix lists the Cisco Validated Design (CVD) companion documents for this solution.
Chapter C, “Configuration Templates”	This appendix lists the configurations per-tenant type for this solution.
Chapter D, “Glossary”	This glossary provides a list of acronyms and initialisms used in this document.

About Cisco Validated Designs

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

CHAPTER 1

Implementation Overview

The cloud is the source of highly scalable, efficient, and elastic services accessed on-demand over the Internet or intranet. In the cloud, compute, storage, and network hardware are abstracted and delivered as a service. End users enjoy the functionality and value provided by the service without the need to manage the underlying technology. A cloud deployment model differs from traditional deployments in its ability to treat the Data Center (DC) as a common fabric of resources. A portion of these resources can be dynamically allocated and de-allocated when they are no longer in use.

The Cisco Virtualized Multiservice Data Center (VMDC) solution is the Cisco reference architecture for Infrastructure as a Service (IaaS) cloud deployments. This Cisco IaaS cloud architecture is designed around a set of modular DC components consisting of building blocks of resources called pods. A pod, or Point of Delivery, is comprised of the Cisco Unified Computing System (UCS), SAN and NAS storage arrays, Access (switching) layers, Aggregation (switching and routing) layers, Services layer devices (firewalls, load balancers), and multiple 10 GE fabric using highly scalable Cisco network switches and routers.

The VMDC solution utilizes compute and pod building blocks consisting of shared resource pools of network, compute, and storage components. Each of these components is virtualized and used by multiple tenants securely, so that each cloud tenant appears to have its own set of physical resources. Cloud service orchestration tools automate the resource provisioning workflow within the cloud DC. Orchestration leverages a set of tools and APIs to dynamically provision cloud resources on demand. The VMDC solution is targeted towards Enterprises building private clouds and Service Providers building IaaS public clouds and virtual private clouds. There have been several iterations of the VMDC solution, with each phase encompassing new platforms, versions, and technologies. The most recently released versions of VMDC are the VMDC 2.2 system release, which is based on traditional Layer 2 (L2) fabric architecture with Virtual Port-Channels (vPC), and the VMDC 3.0 system release, which is based on an extended L2 DC fabric utilizing Cisco FabricPath. Both systems utilize an end-to-end VRF-Lite Layer 3 (L3) model within the DC.

For more information about the VMDC solution, refer to the following documentation on Cisco.com Design Zone:

- Cisco VMDC 2.2 Design Guide
- Cisco VMDC 2.2 Implementation Guide
- Cisco VMDC 3.0 Design Guide
- Cisco VMDC 3.0 Implementation Guide
- Cloud Ready Infrastructure Smart Solutions Kits Accelerate Design and Deployment of Unified DC

This document is the implementation guide for the VMDC 2.3 solution. The VMDC 2.3 architecture is based on the prior VMDC 2.2 architecture, with some major changes focusing optimizing the design to achieve higher tenancy scale, while lowering the cost and footprint of the solution. The key changes in the VMDC 2.3 solution are listed below.

- Use of ASA and ACE services appliances to connect directly to the Aggregation layer Nexus 7000, instead of using service modules on the Catalyst 6500 Data Center Services Node (DSN).
- Collapsed Core/Aggregation layer, instead of a separate core layer to interconnect Pods. The Pods will be interconnected through the DC-PE layer in VMDC 2.3.
- Nexus 7004 with F2 module and SUP2 supervisor, as the Aggregation layer (Other Nexus 7000 form-factors and linecards can also be utilized).
- Cisco Aggregation Services Router (ASR) 1006 as Data Center Provider Edge (DC-PE) (other ASR platforms or form-factors can also be utilized).
- Optimized tenancy models (Gold, Silver, Bronze network containers) to consume less resources (VRF, HSRP, VLAN, BGP etc) on the platforms, and thus provide higher tenancy scale.
- Introduction of a Copper tenant container, with shared outside global routing and shared firewall context, but separate inside routing contexts. The Copper container is suitable for Internet-based cloud access for Small/Medium Business (SMB) customers, who typically require a few Virtual Machines (VMs) on a single VLAN.

This document is the implementation guide for the VMDC 2.3 solution. Refer to the Cisco VMDC 2.3 Design Guide for the design considerations that were used for this architecture.

This chapter presents the following topics:

Solution Architecture

The VMDC 2.3 solution utilizes a hierarchical network design for High Availability (HA) and scalability. The hierarchical or layered DC design uses redundant switches at each layer of the network topology for device-level failover that creates a highly available transport between end nodes using the network. DC networks often require additional services beyond basic packet forwarding, such as SLB, firewall, and intrusion prevention. These services might be introduced as modules populating a slot of one of the switching nodes in the network or as standalone appliance devices. Each service approach also supports the deployment of redundant hardware to preserve HA standards set by the network topology. This layered approach is the basic foundation of the VMDC design to provide scalability, performance, flexibility, resiliency, and service assurance. VLANs and Virtual Routing and Forwarding (VRF) instances are used to provide tenant isolation within the DC architecture, and routing protocols within the VRF instances are utilized to interconnect the different networking and service devices.

This multi-layered VMDC DC architecture is comprised of WAN, Core, Aggregation, Services, and Access layers. This architecture allows for DC modules to be added as demand and load increases. It also provides the flexibility to create different logical topologies utilizing device virtualization, the insertion of service devices, and traditional Layer 3 (L3) and L2 network configurations. The previous VMDC 2.0/2.2 architectures included all of the above layers. The VMDC 2.3 architecture has been optimized by utilizing a collapsed Core/Aggregation layer, so the architecture does not include a separate Core layer. Instead, the different pods in the VMDC 2.3 DC are interconnected at the WAN layer.

These layers in the VMDC 2.3 architecture are briefly described below.

- **WAN/Edge—**The WAN or DC Edge layer connects the DC to the WAN. Typically, this provides IP or Multiprotocol Label Switching (MPLS)-based connectivity to the Internet or intranet. The ASR 1006 is used as an MPLS PE router in the VMDC 2.3 design, providing L3VPN connectivity to the provider IP/MPLS network. It also provides aggregation of all the DC pods as they connect directly to the ASR 1006 PE.
- **Aggregation—**The Aggregation layer of the DC provides a consolidation point where Access layer switches provide connectivity between servers for multi-tier applications and across the core of the network to clients residing within the WAN, Internet, or campus. The VMDC 2.3 design utilizes Nexus 7004 switches as the Aggregation layer. This Aggregation layer provides the boundary between L3 routed links and L2 Ethernet broadcast domains in the compute cluster, and is also the connection point for the DC Services layer (firewalls, load balancers, and so forth). A **VRF sandwich** design is used on the Nexus 7004 aggregation device to insert the firewall layer for Gold and Copper tenants. An Outside VRF instance on the aggregation Nexus 7004 interconnects the DC-PE to the Perimeter Firewall layer. An Inside VRF on the aggregation Nexus 7004 is used to connect the Firewall layer to the compute cluster. This inside VRF on the aggregation Nexus 7004 switch is the default gateway for the VMs. For Silver and Bronze tenants, a single VRF design is used, and each tenant has a VRF instance where the default gateway for the VMs is implemented.
- **Access—**The Access layer of the network provides connectivity for server farm end nodes in the DC. The Nexus 5548 is utilized as the Access layer switch in the VMDC 2.3 design. The Nexus 5548 connects to multiple UCS fabrics - UCS 6200 Fabric Interconnects and UCS 5100 Blade Chassis with UCS B-series blade servers). Typically, the Nexus 5500, UCS Fabric-Interconnects, UCS Blade-Chassis along with storage resources are bundled together in Integrated Compute Stacks (ICS) such as the VCE Vblock and Cisco/NetApp FlexPod.
- **Services—**Network and security services such as firewalls, server load balancers, intrusion prevention systems, application-based firewalls, and network analysis modules are typically deployed at the DC Services layer. In the VMDC 2.3 design, these services are implemented by appliances connected directly to the aggregation Nexus 7004 switches. The SLB service is provided by one or more pairs of ACE 4710 appliances. A pair of ASA 5585-X security appliances connected to the Nexus 7004 aggregation switches provides firewall services. A pair of ASA 5555-X security appliances connected to the Nexus 7004 aggregation switches provides secure remote access (IPsec-VPN and SSL-VPN) services, enabling remote clients to securely connect to the cloud resources. In addition, the Cisco Virtual Services Gateway (VSG) working in conjunction with the Cisco Nexus 1000V Distributed Virtual Switch (DVS) provides security services in the Compute layer, thereby providing intra-VLAN and inter-VLAN protection to the VMs.
- **Integrated Compute and Storage—**This is the Compute and Storage block, such as FlexPod or Vblock. This typically consists of racks of compute based on UCS and storage, and have a pair of Nexus 5500 switches aggregating the connections out of the block. The Nexus 5500 Access switch within the ICS provides connectivity both for the LAN (via 10GE Ethernet links) and SAN (via dedicated FC links), and also connects to the storage for the ICS stack.
- **Virtual Access—**Access switch virtualization allows the function of the logical L2 Access layer to span multiple physical devices. The Nexus 1000V DVS running on top of the VMware ESXi hypervisor is used in the VMDC 2.3 design.

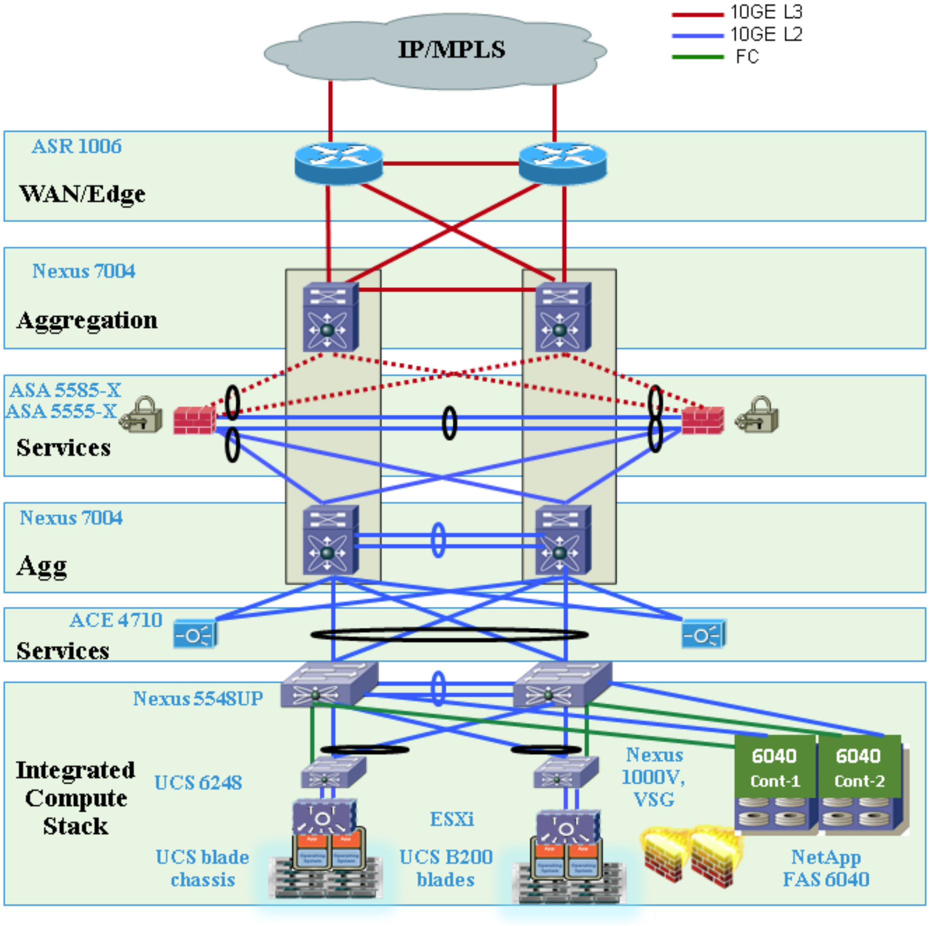
The Compute and Storage layer in the VMDC 2.3 design has been validated with a FlexPod-aligned implementation using the following components:

- **Compute—**Cisco UCS 6248 Fabric Interconnects (FIs) with UCS 5108 blade chassis populated with UCS B200 half-width blades. VMware vSphere 5.0 ESXi is the hypervisor for virtualizing the UCS blade servers.

- **SAN—**Cisco Nexus 5500 switches provide Fibre Channel (FC) connectivity between the UCS compute blades and the NetApp FAS 6040 storage array.

Figure 1-1 provides a logical representation of the VMDC 2.3 architecture.

Figure 1-1 VMDC 2.3 Solution Architecture



Multitenancy and VRF Separation

Multitenancy refers to the virtualization or logical division of a shared pool of network, compute, and storage resources among multiple tenants or groups. This logical separation is used instead of requiring dedicated physical resources for each tenant, thereby reducing cost and increasing utilization of resources. In the Enterprise private cloud deployment model, the tenant is referenced as a department or business unit, such as engineering or human resources. In the public cloud deployment model, a tenant is an individual consumer, an organization within an Enterprise, or an Enterprise subscribing to the public cloud services. In either model, each tenant must be securely separated from other tenants because they share the virtualized resource pool.

In the VMDC 2.3 solution, VLANs and VRF instances are used to provide traffic isolation between tenants. Each tenant has its own VRF and associated set of VLANs and sub-interfaces. VRF instances allow multiple routing configurations in a single L3 switch using separate virtual routing tables. By default, communication between VRF instances is not allowed to protect the privacy of each tenant.

Service appliances like the ASA and ACE are also virtualized into virtual contexts for each tenant to provide traffic isolation. VRF-Lite is used throughout the DC L3 domain to securely isolate the tenants. BGP is used as the routing protocol within the DC, interconnecting the tenant VRF instances on the DC layers/devices. Per-VRF Border Gateway Protocol (BGP) is configured between the WAN/ Edge ASR 1006 router and aggregation Nexus 7004 device. For Gold and Copper tenants, a VRF sandwich design is used, and there are static routes utilized between the firewall context and Nexus 7004 switch for routing between the ASA context to and from the inside as well as outside VRF instances.

Modular Building Blocks

VMDC 2.3 provides a scalable solution that can address the needs of small and large Enterprise and Service Provider cloud data centers. This architecture enables customers to select the design that best suits their immediate needs while providing a solution that can scale to meet future needs without re-tooling or re-designing the DC. This scalability is achieved using a hierarchical design with two different modular building blocks: Pod and Integrated Compute Stack (ICS).

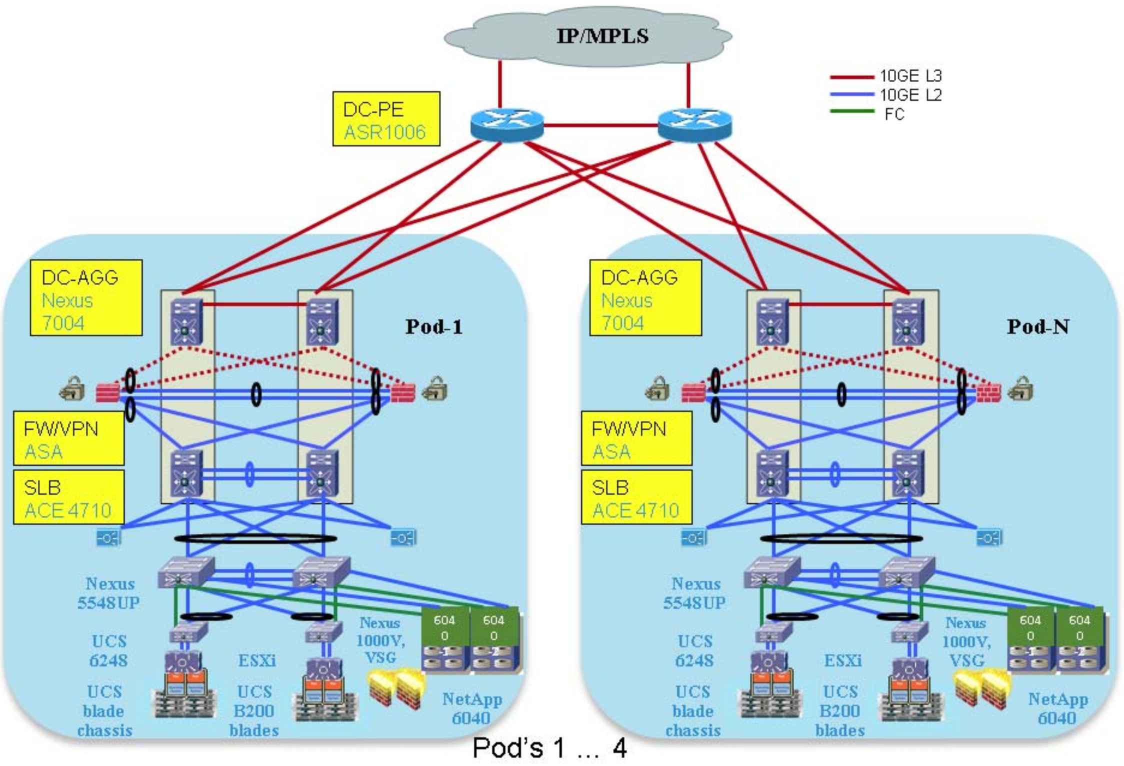
Pod

The modular DC design starts with a basic infrastructure module called a pod, which is a physical, repeatable construct with predictable infrastructure characteristics and deterministic functions. A pod identifies a modular unit of DC components and enables customers to add network, compute, and storage resources incrementally. This modular architecture provides a predictable set of resource characteristics (network, compute, and storage resource pools, power, and space consumption) per unit that is added repeatedly as needed.

In the VMDC 2.3 design, the Aggregation layer switch pair, Services layer nodes, and one or more integrated compute stacks are contained within a pod. The pod connects to the WAN/PE layer devices in the DC. To scale a pod, providers can add additional integrated compute stacks and can continue to scale in this manner until the pod resources are exceeded. To scale the DC, additional pods can be deployed and connected to the WAN layer devices.

Figure 1-2 illustrates how pods can be used to scale compute, network, and storage in predictable increments within the DC.

Figure 1-2 VMDC Pods for Scaling the Data Center



Integrated Compute Stack (ICS)

The second modular building block utilized is a generic ICS based on existing models, such as the VCE Vblock infrastructure packages or the Cisco/NetApp FlexPods. The VMDC 2.3 architecture is not limited to a specific ICS definition, but can be extended to include other compute and storage stacks.

An ICS can include network, compute, and storage resources in a repeatable unit. In this solution, storage and compute resources are contained within an ICS. To scale a pod, customers can add additional integrated compute stacks and can continue to scale in this manner until pod resources are exceeded. The storage and SAN resources within each pod can be interconnected to build a hierarchical storage area network.

Note The VMDC modular architecture has been designed to accommodate different types of iIntegrated compute stacks. Previous versions of VMDC have been validated with VCE Vblocks and Cisco/ NetApp FlexPods, and can support alternative ICS like Hitachi UCP, or other custom built compute stacks. While the VMDC 2.3 system can also accommodate Vblocks, FlexPods or other flavors of ICS, it has been validated with the Cisco/NetApp FlexPod.For more information on FlexPod, please refer to the following link.

Figure 1-3 illustrates how ICS can be used to scale the pod.

Figure 1-3 VMDC ICS for Scaling the Pod

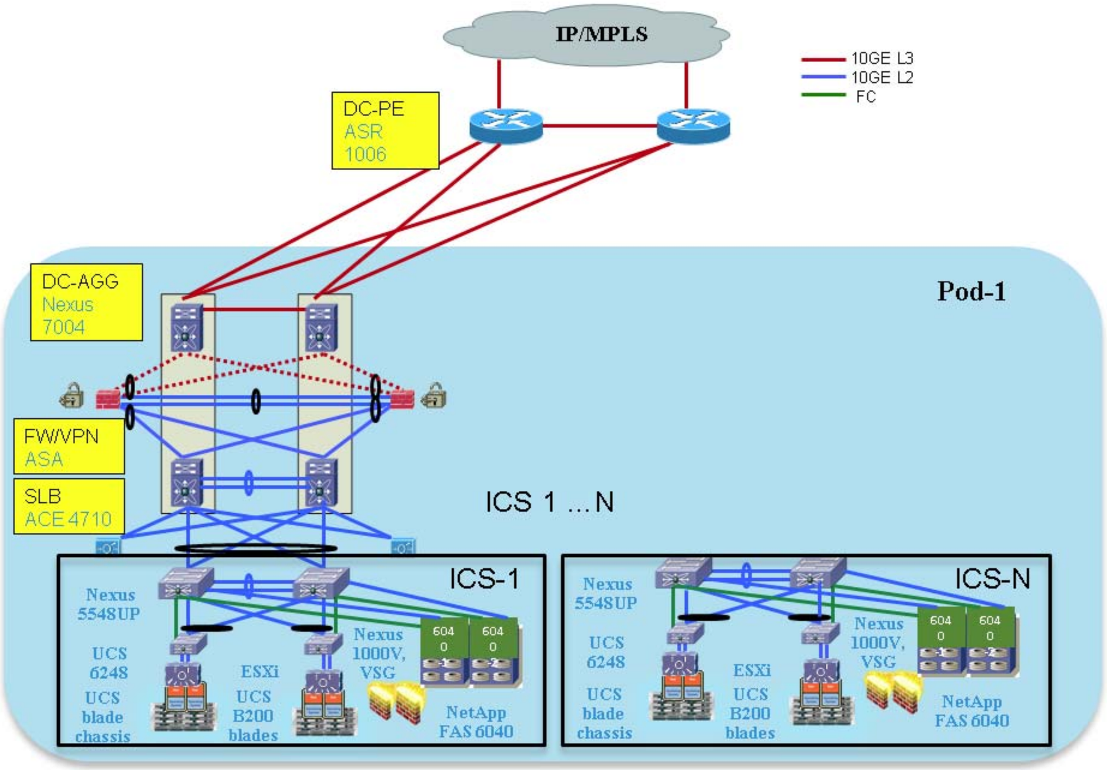


Table 1-1 lists some of the key architectural features of VMDC 2.3.

Table 1-1 VMDC 2.3 Architecture

	VMDC 2.3 Design
DC Layers	• DC WAN Edge (MPLS-PE) • Aggregation • Services • ICS Access • Virtual Access
Tenant Isolation	VRF-Lite, VLANs, virtual contexts
Routing Protocol in DC	BGP
Perimeter Firewall for Gold tenants	• 2 VRF instances per Gold tenant for private access (L3VPN) (outside and inside VRF) on Nexus 7004 Aggregation, separated by the private firewall context • 1 Demilitarized Zone (DMZ) VRF per Gold tenant for public access (via Internet) protected via DMZ firewall context • Interconnect between Private and DMZ firewall context
Shared Firewall for Copper tenants	One common firewall context, protecting all Copper tenants from Internet and connecting to per-tenant inside VRF
No Firewall tenants	Single VRF for Bronze and Silver tenants

Table 1-1 VMDC 2.3 Architecture (continued)

Perimeter Firewall Layer	Routed (L3) ASA 5585-X
SLB	Routed (L3) ACE 4710 in one-arm mode sitting on the server VLANs (one pair per 20 Silver tenants, and one pair per 10 Gold tenants - total of 4 ACE 4710s for 20 Silver + 10 Gold tenants)
Secure Access to Cloud	IPsec-VPN, SSL-VPN on ASA 5555-X
Max VMs per Pod	6000 VMs - based on MAC Scale with F2 modules, which provide high port-density (can support higher number of VMs (MACs) if using M series modules)
Max VMs per DC	24,000 VMs - based on 4 Pods and F2 modules on Nexus 7004 Agg) (can support higher number of VMs if using M series modules on Nexus, or when using more than 4 pods)
ICS size	64 half-width UCS blades, 8 UCS 5108 chassis, 2 UCS 6248 FIs in a FlexPod aligned topology (number of chassis and blades used can vary based on VM, Application and IOPs requirements)
ICS per Pod	3 (can support more based on VM/App sizing and network over-subscription ratios, and using larger Nexus 7000 form factors for higher port density)
Pod Compute Size	192 half-width UCS blades - with 3 ICS, and 8 chassis per ICS, 8 blades per chassis
Pods per DC	4 - with ASR 1006 as PE (can support more Pods if using larger ASR 1000 form-factors for higher 10G port-density)
Tenants per Pod	500 (10 Gold, 10 Silver, 220 Bronze, 250 Copper validated), can vary for different mix
Tenants per DC	2000 - based on a mixed-tenancy scale of 500 tenants (250 Copper, 220 Bronze, 20 Silver, 10 Bronze) per pod (tenancy numbers will depend on the ratio of Gold/Silver/Bronze/Copper tenants, which will also determine number of ASA and ACE appliances and virtual contexts needed).

Service Tiers

Cloud providers, whether Service Providers or Enterprises, want an IaaS offering with multiple feature tiers and pricing levels. To tailor workload or application requirements to specific customer needs, the cloud provider can differentiate services with a multi-tiered service infrastructure and Quality of Service (QoS) settings. The Cisco VMDC architecture allows customers to build service level agreements that support their tenant or application requirements. Such services can be used and purchased under a variable pricing model. Infrastructure and resource pools can be designed so that end users can add or expand services by requesting additional compute, storage, or network capacity. This elasticity allows the provider to maximize the user experience by offering a custom, private DC in virtual form.

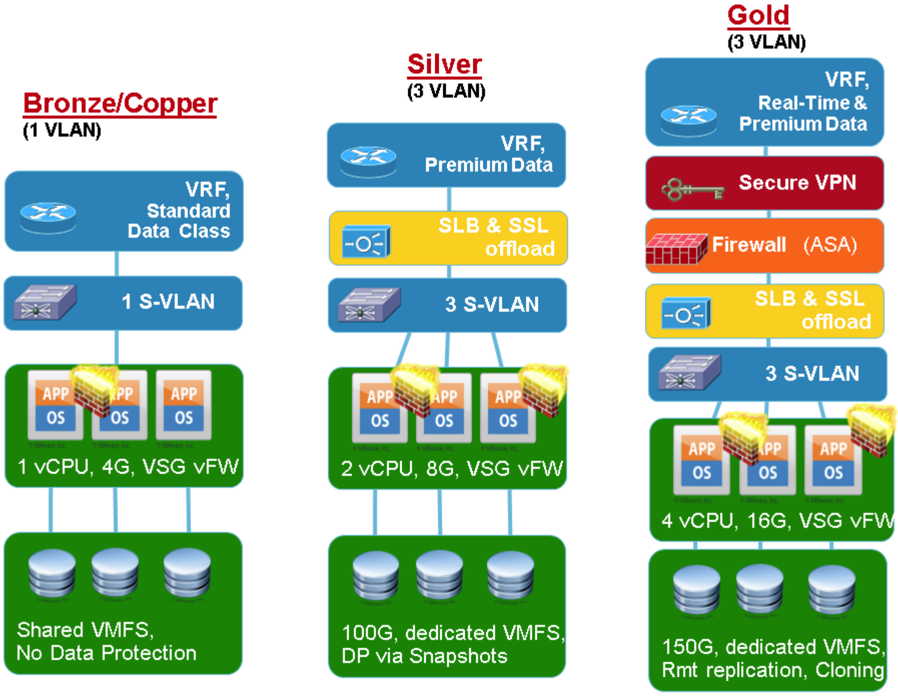
The VMDC 2.3 solution defines a reference multi-tier IaaS service model of Gold, Silver, Bronze, and Copper tiers. These service tiers define resource and service levels for compute, storage, and network performance. This is not meant to be a strict definition of resource allocation, but to demonstrate how differentiated service tiers could be built. These are differentiated based on the following features:

- **Network Resources—**Differentiation based on network resources and features.

  - **Application Tiers—**Service tiers can provide differentiated support for application hosting. In some instances, applications may require several application tiers of VMs (web, application, database). VMDC 2.3 Gold and Silver services are defined with three application tiers on three separate VLANs to host web, application, and database services on different VMs. The Bronze and Copper service is defined with one VLAN only, so if there are multi-tiered applications, they must reside on the same VLAN or potentially on the same VM (Linux, Apache, MySQL, PHP, Perl, and Python (LAMP)/Windows Apache, MySQL, PHP, Perl, and Python (WAMP) stack).
  - **Access Methodsand Security**—All four services, Gold, Silver, Bronze, and Copper, are defined with separate VRFs to provide security and isolation. Gold offers the most flexible access methods - over Internet, L3VPN, and secure VPN access over the Internet. Also, the Gold model has multiple security zones for each tenant. The Silver and Bronze models do not support any perimeter firewalling and only have VRF instances for isolation. The Copper model supports access over Internet, protected by a shared firewall to a private inside VRF per tenant.
  - **Stateful Services—**Tenant workloads can also be differentiated by the services applied to each tier. The Gold service is defined with an ASA 5585-X virtual firewall, ACE 4710 Virtual Server Load Balancer (vSLB), and secure remote access (IPsec-VPN and SSLVPN) on the ASA 5555-X. The Silver tier is defined with an ACE 4710 vSLB. The Bronze tier is defined with no services. The Copper tier has a shared perimeter firewall across all tenants. In addition, for all four service tiers, security can be provided in the Compute layer by utilizing the VSG, in conjunction with the Nexus 1000V DVS.
  - **QoS—**Bandwidth guarantee and traffic treatment can be a key differentiator. QoS policies can provide different traffic classes to different tenant types and prioritize bandwidth by service tier. The Gold tier supports VoIP/real-time traffic, call signalling and data class, while the Silver, Bronze, and Copper tiers have only data class. Additionally, Gold and Silver tenants are given bandwidth guarantee, with Gold getting more bandwidth (2x) than Silver.
- **VM Resources—**Service tiers can vary based on the size of specific VM attributes, such as CPU, memory, and storage capacity. The Gold service tier is defined with VM characteristics of 4 vCPU and 16 GB memory. The Silver tier is defined with VMs of 2 vCPU and 8 GB, while the Bronze and Copper tier VMs have 1 vCPU and 4 GB.
- **Storage Resources—**To meet data store protection, the recovery point, or the recovery time objectives, service tiers can vary based on provided storage features, such as Redundant Array of Independent Disks (RAID) levels, disk types and speeds, and backup and snapshot capabilities. The Gold service is defined with 15k FC disks, the Silver tier on 10k FC disks, and the Bronze tier on Serial AT Attachment (SATA) disks.

Figure 1-4 shows the three service tiers defined and validated in the VMDC 2.3 solution which are similar to the ones offered in VMDC2.2. Additionally, a Copper container is also offered, which has the same parameters as Bronze, but has only Internet-based access and no L3VPN-based access. These are reference service tiers that have been defined as part of the VMDC 2.3 reference architecture. Cloud providers can use this as a basis and define their own custom service tiers, based on their own deployment requirements.

Figure 1-4 VMDC 2.3 Service Tiers



The following tables provide details on the VMDC 2.3 three-tier service model. Table 1-2 depicts the network resource differentiation, Table 1-3 depicts the compute resource differentiation, and Table 1-4 depicts the storage resource differentiation.

Table 1-2 VMDC 2.3 Service Tiers—Network Resources

	Gold Tier	Silver Tier	Bronze Tier	Copper Tier
Tenant Isolation	VLANs and VRFs	VLANs and VRFs	VLANs and VRFs	VLANS and VRFs
Tenant Zones	2 PVT and DMZ	1 PVT	1 PVT	1 PVT
VRFs	3 (PVT-outside, PVT-inside and DMZ)	1 (PVT)	1 (PVT)	1 (PVT)
Server VLANs	4 (3 in PVT and 1 in DMZ)	3	1	1
Services	Virtual Firewall (vFW) with ASA-5585-X60, vSLB with ACE 4710, Secure Remote Access (IPsec & SSL VPN) with ASA 5555-X, Virtual Security with Nexus 1000V VSG	vSLB with ACE 4710, Virtual Security with Nexus 1000V VSG	Virtual Security with Nexus 1000V VSG	Shared Perimeter Firewall, Virtual Security with Nexus 1000V VSG

Table 1-2 VMDC 2.3 Service Tiers—Network Resources (continued)

QoS Traffic Classes	Real-time traffic (for ex., VoIP) + Premium Data traffic (bandwidth guaranteed)	Premium Data traffic (bandwidth guaranteed)	Standard traffic (available bandwidth)	Standard traffic (available bandwidth)
Sample QoS SLA CIR / PIR per tenant *	500 Mbps / 3 Gbps	250 Mbps / 2 Gbps	100 Mbps / 1 Gbps(down) 100 Mbps rate-limit up	100 Mbps bandwidth reserved for downstream for all Copper together available bandwidth (up)

Note This is a sample SLA that was validated in the testbed, based on the number of VMs, tenants, and link density.

Table 1-3 VMDC 2.3 Service Tiers—Compute Resources

	Gold Tier	Silver Tier	Bronze & Copper Tier's
VM per CPU core	1	2	4
(VM:Core oversubscription)	(1:1)	(2:1)	(4:1)
Number of VMs per	8	16	32
UCS B200-M1 blade			
(8 cores)			
RAM per VM	16 GB	8 GB	4 GB
vNIC per VM	2	2	2
	(one for data, one for management)	(one for data, one for management)	(one for data, one for management)

Table 1-4 VMDC 2.3 Service Tiers—Storage Resources

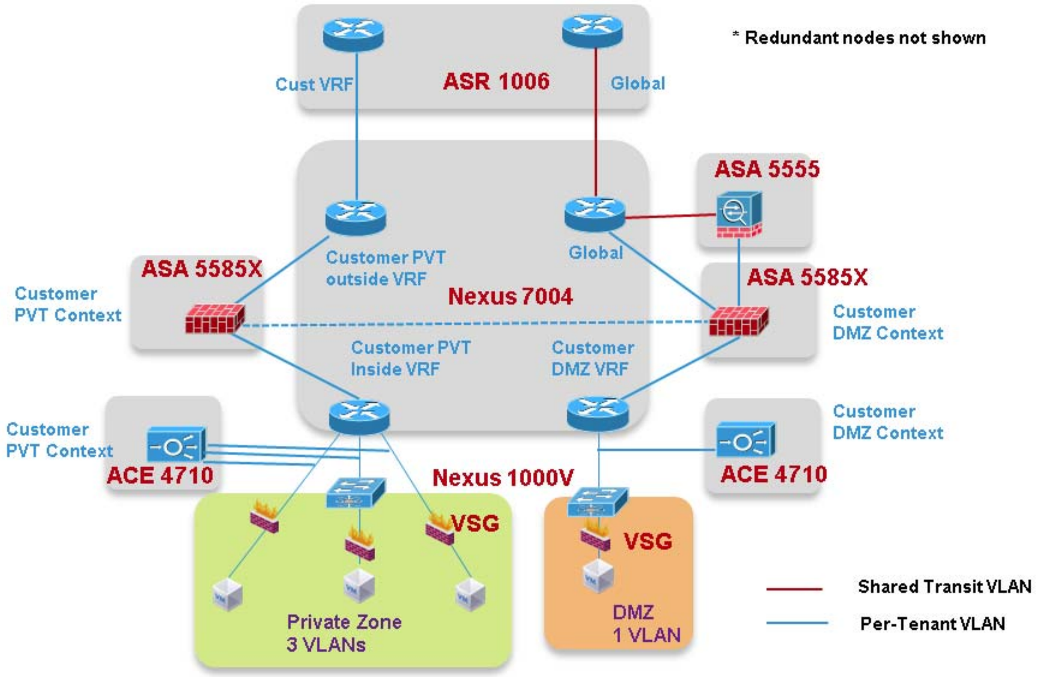
	Gold Tier	Silver Tier	Bronze/Copper Tier's
Disk size per VM	300 GB	150 GB	50 GB
Disk size increment	50 GB	50 GB	50 GB
Disk Type	15k FC drives	10k FC drives	7200 RPM SATA drives
Data Protection	Clone - mirror copy (local site)	Snap - virtual copy (local site)	None
Disaster Recovery	Remote replication	Remote replication	None

Note Table 1-4 is shown as a sample representation of storage differentiation.

VMDC 2.3 Gold Service

Figure 1-5 shows a logical representation of a VMDC 2.3 Gold service tier network container.

Figure 1-5 VMDC 2.3 Gold Service Tier Logical Topology—PVT VRF (Zone)



The network container is a logical (virtual) segment of the shared (common) physical network resources (end-to-end through the DC) that represents the DC network domain carrying tenant traffic. The physical infrastructure is common to all tenants, but each network device (routers, switches, firewalls, and so forth) is virtualized such that each tenant's virtual network container is overlaid on the common physical network.

The Gold tenant gets two network (and compute/storage) zones to place workloads into. Each zone has its own set of VLANs, VRF instances, and firewall/load balancer contexts. Figure 1-5 shows a logical representation of a two-zone VMDC 2.3 Gold network container.

This Gold service tier provides the highest level of sophistication by including secure remote access, firewall, and load balancing to the service. The vFW (on the ASA 5585-X60) provides perimeter security services, protecting tenant VMs. The vSLB (ACE 4710 appliance) provides load balancing across VMs in each tier of the tenant. The ASA 5555-X provides virtualized secure remote access (IPsec-VPN and SSL-VPN) to tenant VMs from the Internet. The ACE and ASA service module/ appliance are utilized in routed (L3) virtual mode in the VMDC 2.3 design. The Gold service tier also includes the Nexus 1000V VSG for providing virtual security services to the VMs. The Gold service provides higher QoS SLA and three traffic classes - real-time (VoIP), call signaling, and premium data.

The two zones can be used to host different types of applications, to be accessed through different network paths. The two zones are discussed in detail below.

- **PVT Zone.** The PVT, or Private, Zone and its VMs can be used for cloud services to be accessed through the customer MPLS-VPN network. The customer sites connect to the provider MPLSCore and the customer has their own MPLS-VPN (Cust-VRF). The VMDC DC ASR 1000 PE connects to the customer sites through the MPLS-VPN (Cust-VRF in Figure 1-5). This CustVRF is extended through the VMDC network to the Nexus 7004 Aggregation switch. On the Agg/Access Nexus 7004, the Cust-VRF-outside connects to the ASA Cust-vFW, and then is connected back into a Cust-PVT-inside VRF on the Nexus 7004 Agg/Access device (VRF sandwich to insert service nodes), and then to the Compute layer on the UCS. For the VMDC 2.3 Gold tenant, the PVT zone is defined with three server VLANs. In addition, each tenant is assigned a separate Nexus 1000V VSG instance. The tenant is defined as an ORG in the VSG (VNMC), with the three VLANs placed into separate VSG sub-zones. The VSG is used to provide security policies to monitor and protect traffic between the VLANs (sub-zones).
- **DMZ Zone.** The VMDC 2.3 Gold container supports a DMZ Zone for tenants to place VMs into a DMZ area, for isolating and securing the DMZ workloads from the PVT workloads, and also to enable users on the Internet to access the DMZ-based cloud services. The ASR 1000 PE WAN router is also connected to the Internet, and a shared (common) VRF (usually global routing table) exists for all Gold tenants to connect to (either encrypted or unencrypted). Encrypted (SSL or IPsec Remote Access VPN) traffic is sent to an ASA 5555-X, and based on the VPN policy, is mapped to a particular tenant and the corresponding tenant VPN VLAN. The tenant VPN VLAN then connects to the tenant DMZ-vFW (different vFW context on the ASA 5585-X than the tenant PVT-vFW), then to the tenant DMZ-VRF (different VRF on the Nexus 7004 Agg/ Access than the tenant PVT-VRF), and then to the Compute layer for the DMZ Zone. Similarly, unencrypted traffic from the Internet, based on the destination VM/VIP address, is sent to the tenant DMZ-vFW, then to the DMZ-vSLB, DMZ-VRF, and the DMZ Compute Zone. The DMZ Zone can be used to host applications like proxy servers, Internet-facing web servers, email servers, etc. The DMZ Zone consists of one server VLAN in this implementation.

In VMDC 2.3, a Gold tenant can choose to have only the PVT Zone, only the DMZ Zone, or both the PVT and DMZ Zones. If the tenant has both PVT and DMZ Zones, then the Gold tenant will consume three VRF instances (Cust-PVT-outside, Cust-PVT-inside, and Cust-DMZ), two VFW instances, two vSLB instances, two VSGs, and four server VLANs. To facilitate traffic flows between the DMZ and PVT Zones (for example, proxy or web servers in the DMZ Zone, application and database servers in the PVT Zone), the DMZ-vFW and PVT-vFW are interconnected, with the appropriate security policies.

Load-balanced traffic for all tiers of Gold tenants is implemented using the ACE 4710, which has one interface in each of the tiers.

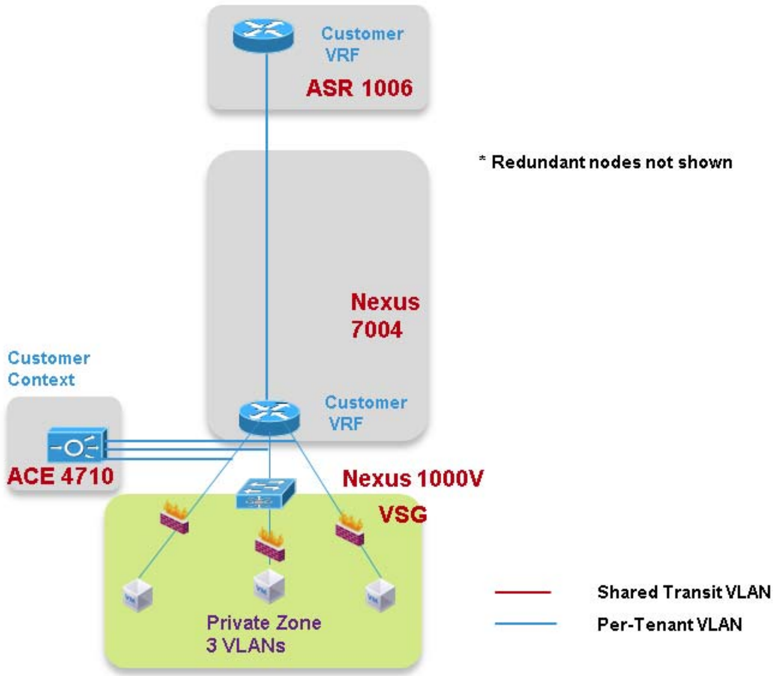
The following cloud traffic services flows can be enabled in the VMDC 2.3 two-zone Gold service tier:

- MPLS-VPN to PVT Zone
- Unsecured (clear) Internet to DMZ Zone
- Secure (Remote Access SSL/IPsec VPN) Internet to DMZ Zone
- DMZ to PVT Zone
- MPLS-VPN to DMZ Zone
- PVT to Internet Zone is via an HTTP proxy hosted in the DMZ Zone

VMDC 2.3 Silver Service

Figure 1-6 shows a representation of a VMDC 2.3 Silver network container.

Figure 1-6 VMDC 2.3 Silver Service Tier Logical Topology

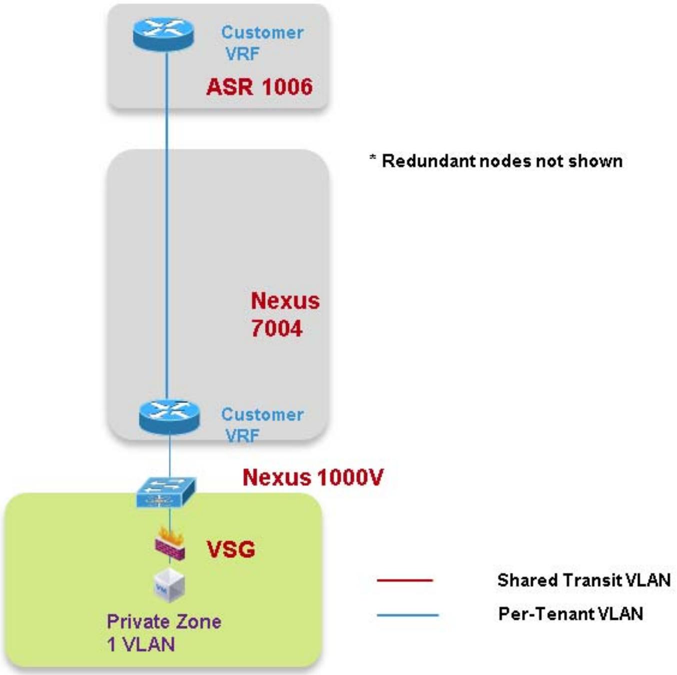


The Silver service tier includes one VRF instance per Silver tenant and three server VLANs (three-tiered applications) for each tenant. The Silver service includes a load-balancing service for more sophistication over the Bronze tier. The vLB (ACE 4710 appliance) provides load balancing across VMs in each tier of the tenant. The ACE service load balancer is utilized in one arm, routed (L3), virtual mode in the VMDC 2.3 design, and one context is used per Silver tenant. The context has links on each of the server VLANs and works in one-arm mode. The Silver service tier also includes the Nexus 1000V VSG to provide virtual security services to the VMs. The Silver service provides medium QoS SLA and one traffic class, premium data.

VMDC 2.3 Bronze Service

Figure 1-7 shows a representation of the VMDC 2.3 Bronze network container.

Figure 1-7 VMDC 2.3 Bronze Service Tier Logical Topology

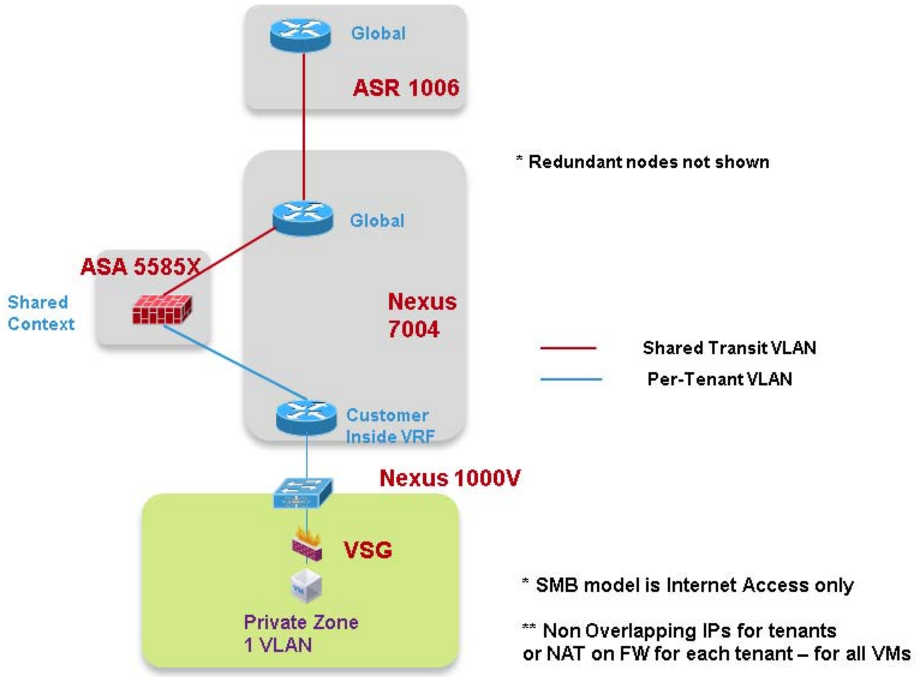


The Bronze service tier includes one VRF instance and one server VLAN for each tenant. The Bronze service is the least sophisticated tier and does not include any perimeter security services. The Bronze service tier does include the Nexus 1000V VSG for providing virtual security services to the VMs. The Bronze service provides lower QoS SLA and one traffic class, standard data.

VMDC 2.3 Copper/SMB Service

Figure 1-8 shows a representation of the VMDC 2.3 Copper network container.

Figure 1-8 VMDC 2.3 Copper Service Tier Logical Topology


Global
ASR 1006
* Redundant nodes not shown
Global
ASA 5585X
Shared Context
Nexus 7004
Shared Transit VLAN
Per-Tenant VLAN
Customer Inside VRF
Nexus 1000V
VSG
Private Zone
1 VLAN
* SMB model is Internet Access only
** Non Overlapping IPs for tenants
or NAT on FW for each tenant – for all VMs

The Copper service tier, also referred to as the Small/Medium Business (SMB) tier, includes one VRF instance and one server VLAN for each tenant. The Copper service is the least sophisticated tier and shares a single firewall context with all other Copper tenants for access from the outside interface connecting to the Internet. The inside interface for each tenant is connected to a VRF per tenant, and server VLANs are connected to these VRF instances. This allows multiple VLANs for tenants if required. The Copper service tier does include the Nexus 1000V VSG to provide virtual security services to the VMs. The Copper service provides the lowest QoS SLA and one traffic class, standard data.

Solution Components

The VMDC 2.3 solution comprises a collection of Cisco and third-party hardware, software and management components. Table 1-5 highlights the components validated as part of the VMDC 2.3 solution.

Table 1-5 VMDC 2.3 Solution—Cisco and Third-Party Components

Product	Description	Hardware	Software
Cisco ASR 1000	WAN (MPLS-PE) Router	ASR 1006, RP-2/ ESP-40/SIP-40/ SPA-10GE-V2	IOS XE 3.7.1S (15.2. (4)S1-based)
Cisco ACE 4700	ACE (Server Load Balancer)	ACE 4710-MOD-K9	A 5(2.1)
ASA 5555-X	IPsec & SSL VPN remote access	ASA 5555-X	9.0.1
ASA 5585-X	Firewall Services Appliance	ASA 5585-X with SSP-60	9.0.1
Cisco Nexus 5500	Integrated Access Switch	Nexus 5548UP	NX-OS 5.2(1)N1(2)

Table 1-5 VMDC 2.3 Solution—Cisco and Third-Party Components (continued)

Cisco UCS	Unified Compute System	UCS 5108 blade chassis, UCS 6248 Fabric Interconnect B200-M2 and M3 server blades, Cisco VIC 1240, Cisco VIC 1280, Cisco UCS M81KR Adapters	2.0(4b)
Cisco Nexus 1010	Virtual Service Appliance		4.2(1)SP1(5.1)
Cisco Nexus 1000V	Distributed Virtual Switch		4.2(1)SV2(1.1)
Cisco VSG	Nexus 1000V Virtual Security Gateway		4.2(1)VSG1(4.1)
Cisco VNMC	Virtual Network Management Center		2.0(3f)
NetApp FAS	SAN Storage Array (for Management Pod)	FAS 6040, FAS 3240	ONTAP 8.1.1
VMware vSphere ESXi	Hypervisor		5.0.0 Build 804277
VMware vSphere vCenter	Virtualization Manager		5.0.0 Build 821926
VMware vSphere Auto Deploy			5.0.0.3392

Table 1-4 lists the component versions validated as part of the VMDC 2.3 solution, which is a reference architecture consisting of a set of hardware and software components that have been validated together at a point of time. It is possible that by the time customers are ready to deploy a VMDC 2.3-based cloud DC, some of these hardware and software versions could be end-ofsale or end-of-life, or there could be newer versions of these available and recommended by the product teams. In such situations, the newer or recommended releases or platforms should be used for deployments.

Note 1. The VMDC 2.3 solution was validated with the ASA 55555-X for IPsec and SSL VPN remote access. For higher performance and throughput, you can also use the ASA 5585-X with SSP-60.

2. The NetApp FAS6040 is used as the SAN storage array in the VMDC 2.3 compute pod to host production (data) VMs. The NetApp FAS3240 is used in the VMDC 2.3 management pod to host management VMs (VMware Virtual Center, Nexus 1000V VSM, VNMC, test tools, BMC CLM orchestration applications, and other management applications).

CHAPTER 2

Compute and Storage Implementation

The Virtualized Multiservice Data Center (VMDC) 2.3 solution uses modular blocks for Compute and Storage, generically referred to as Integrated Compute and Storage (ICS) stacks. A number of these stacks can be attached to a pod, providing compute and storage scale. The limiting factor in terms of the number of Virtual Machines (VMs) supported in an ICS, pod, and Data Center (DC) is usually multi-dimensional, and in this design, the key parameter for the per-pod limit is the number of MAC addresses on the Nexus 7000 aggregation. With the low cost design based on F2 modules on the Nexus 7004, this is limited to 5000 to 6000 VMs assuming dual-Network Interface Card (NIC) VMs. The other limitation is the number of ports on the Nexus 7004 connecting to each ICS stack, and affecting the bandwidth for north-south as well as east-west routed traffic. In this design, three ICS stacks can be connected to the Nexus 7004, with 4x 10G links per aggregation switch, for a total of 80G to the ICS switch layer. Refer to the Cisco VMDC 2.3 Design Guide for more discussion of the scaling factors.

For validation purposes, a smaller footprint ICS was built as listed in Table 2-1. The ICS design is FlexPod-like and uses the NetApp 6040 as the storage for both SAN and NAS. The details of the test build out are covered in the subsections below.

Table 2-1 ICS Stack

Tenant Type	Number of Tenants	Number of VLANs per Tenant	Number of VMs per VLAN [1]	Total VMs
Gold	10	Private (PVT) Zone - 3 Demilitarized Zone (DMZ) - 1	3 per VLAN (12 VMs)	120
Silver	20	3	3 per VLAN (9 VMs)	180
Bronze	10	1	6	60
Copper	10	1	6	60
Total	50	-	-	420

[1] In addition to these VMs, test tool and traffic generator VMs were also configured.

The following sections show the considerations and configurations while implementing compute and storage for VMDC 2.3, including the virtual Switching layer based on the Nexus 1000V:

- Cisco Unified Computing System Implementation, page 2-2
- Storage Implementation Overview, page 2-8
- Hypervisor vSphere ESXi Implementation, page 2-26
- Nexus 1000V Series Switches, page 2-31
- 2.5 Compute and Storage Best Practices and Caveats, page 2-38

Cisco Unified Computing System Implementation

This section presents the following topics:

- UCS Configuration, page 2-2
- UCS Uplinks Configuration, page 2-4

UCS Configuration

Figure 2-1 shows an overview of the Cisco Unified Computing System (UCS) setup.

Figure 2-1 UCS Physical Layout

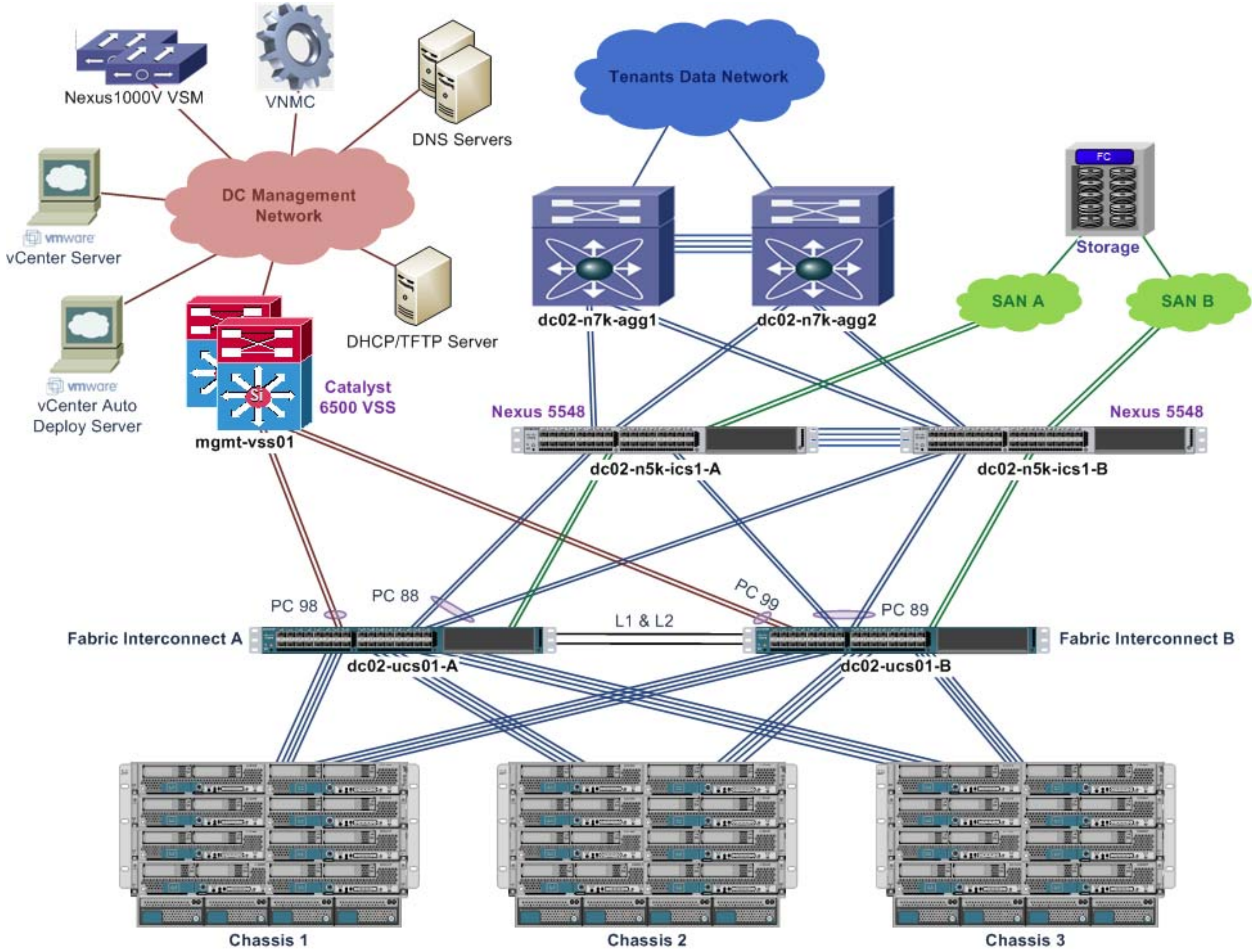


Table 2-2 details the UCS hardware used for the compute infrastructure.

Table 2-2 UCS Hardware

Component	Product Name	Quantity
Fabric Interconnect (FI)	Cisco UCS 6248UP	2
Chassis	Cisco UCS 5108	3
I/O Module	Cisco UCS 2208XP	6
Blade Server	Cisco UCS B200 M3 (2 x 8 cores CPU, 96GB Memory)	8
Blade Server	Cisco UCS B200 M2 (2 x 6 cores CPU, 96GB Memory)	16
Adapter	Cisco UCS VIC 1280	8
Adapter	Cisco UCS VIC 1240	8
Adapter	Cisco UCS M81KR	8

1. For the test implementation, only a small scale, three-chassis compute infrastructure is set up. More chassis and blade servers can be added to support larger scale deployment.
2. The test implementation includes both UCS B200-M2 and B200-M3 blades servers, as both types are supported. The UCS B200-M3 blade server has more CPU resources.
3. The test implementation includes three types of Cisco virtualized adapters. All three types of virtualized adapters used the same driver on vSphere ESXi and are supported.

The following list highlights the Cisco Unified Computing System Manager (UCSM) configuration:

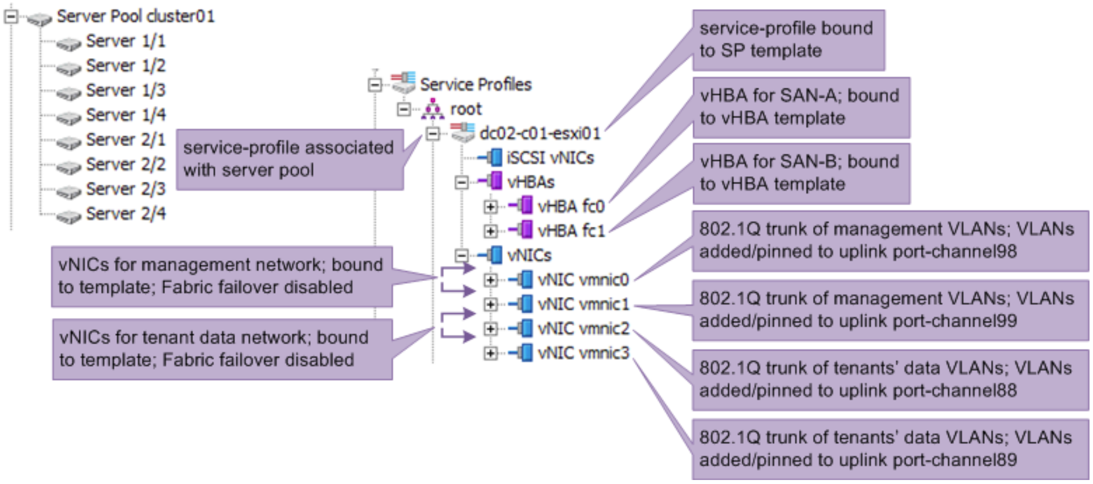
- The UCS FIs are configured into a cluster to provide active/standby management plane redundancy for the UCSM. The data plane for the UCS operates in active/active mode.
- The UCS FIs are configured in End-host (EH) mode.
- Three UCS 5108 chassis are connected to a pair of UCS 6248 FIs.
- Each chassis has four server links to each FI.
- The uplinks on the FIs are bundled into port-channels to upstream switches with disjoint Layer 2 (L2) networks. Refer to UCS Uplinks Configuration for more details.
- The FIs connect to two Nexus 5000 switches with Fibre Channel (FC) links for access to SAN storage. Refer to Storage Implementation Overview for more details.
- The UCSM configuration make use of updating templates to ensure consistent configuration and updates across all blade servers.
- The BIOS policy for each blade server is optimized for ESXi hypervisor application. Refer to UCS and VMware documentation for more details.
- The blade servers are configured to boot via vSphere Auto Deploy, and boot disk (SAN or local) is not required on each blade server. Refer to Hypervisor vSphere ESXi Implementation for more details.
- Networking for each blade server (ESXi hypervisor) is managed by the Nexus 1000V. Refer to Nexus 1000V Series Switches for more details.
- Each blade server is configured with four Virtual Network Interface Cards (vNICs) for access to the disjoint upstream L2 networks for redundancy. On the UCSM, fabric failover for each vNIC is not required/enabled.

- Each blade server is configured with two Virtual Host Bus Adapters (vHBAs) for access to SAN storage via SAN-A and SAN-B for storage multipathing. Refer to Storage Implementation Overview for more details.

1. Disjoint L2 upstream network configuration on the UCSM requires EH mode on the UCS FIs.
2. Some configuration changes will either cause server reboot or service disruption. Multiple templates of the same type should be used to prevent any single change to cause service disruption to all blade servers.

Figure 2-2 shows the service-profile configuration for one of the blade servers on the UCSM. Updating service-profile templates, updating vNIC templates, and updating vHBA templates are used to ensure that the configuration across multiple blade servers are consistent and up to date. Server pools are configured, and each service-profile is associated with its respective server pool. Each server pool has blade servers from two or more chassis for redundancy purposes.

Figure 2-2 Service-Profile Configuration



service-profile associated with server pool

vNICs for management network; bound to template; Fabric failover disabled

vNICs for tenant data network; bound to template; Fabric failover disabled

service-profile bound to SP template

vHBA for SAN-A; bound to vHBA template

vHBA for SAN-B; bound to vHBA template

802.1Q trunk of management VLANs; VLANs added/pinned to uplink port-channel98

802.1Q trunk of management VLANs; VLANs added/pinned to uplink port-channel99

802.1Q trunk of tenants' data VLANs; VLANs added/pinned to uplink port-channel88

802.1Q trunk of tenants' data VLANs; VLANs added/pinned to uplink port-channel89

UCS Uplinks Configuration

The UCS FIs connect to two upstream networks:

- DC management network, for management/administration access to network devices and compute hypervisors
- Network for tenants' data traffic

The two networks are not connected to each other. Access to both networks by all blades servers is necessary for proper operations of VMDC architecture. The upstream disjoint L2 networks' capability is configured to allow access to both networks.

Take note of the following considerations when configuring the disjoint L2 network on the UCSM:

- The UCSM must be configured in EH mode.
- In a High Availability (HA) UCSM cluster, symmetrical configuration on both fabric A and fabric B is recommended, and both FIs should be configured with the same set of VLANs.
- UCSM verifies the VLANs' configuration, and the VLANs used for the disjoint L2 networks must be configured and assigned to an uplink Ethernet port or uplink Ethernet port channel.

- UCSM does not support overlapping VLANs in disjoint L2 networks. Ensure that each VLAN only connects to one upstream disjoint L2 network.
- A vNIC (VMNIC in the vSphere ESXi hypervisor or physical NIC in the bare metal server) can only communicate with one disjoint L2 network. If a server needs to communicate with multiple disjoint L2 networks, configure a vNIC for each of those networks.
- Do not configure any vNICs with a default VLAN (VLAN ID 1).

By default, all VLANs are trunked to all available uplinks to maintain backward compatibility, however, this default behavior would cause a data traffic black hole when connecting the FIs to disjoint upstream L2 networks. VLANs must be explicitly assigned to the appropriate uplink(s) to ensure proper network operations. On the UCSM, VLANs are assigned to specific uplinks using the

LAN Uplinks Manager. Refer to UCSM documentation for more details about LAN Uplinks Manager usage.

Figure 2-3 UCS Disjoint Upstream L2 Network Layout

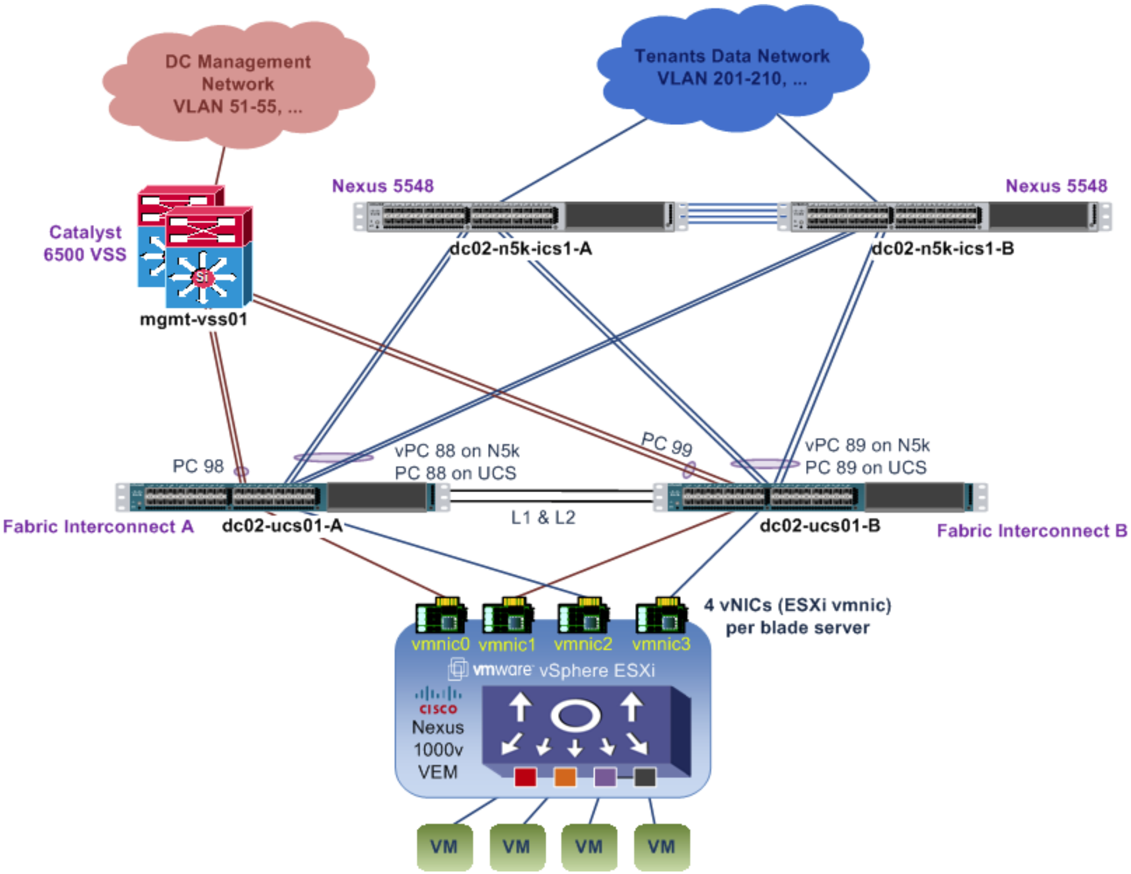


Figure 2-3 shows the disjoint L2 networks setup for this implementation. Each FI has two port-channel uplinks to two different upstream L2 networks. Each upstream L2 network has a completely non-overlapped set of VLANs. Table 2-3 shows the VLANs to uplink port-channel mapping.

Table 2-3 LANs to Uplink Port-channel Mapping

Uplink	Server vNIC	VLANs Assigned	Remarks
Port-channel 98	vmnic0 (Fabric A)	51-55	Management VLANs

Table 2-3 LANs to Uplink Port-channel Mapping (continued)

Port-channel 99	vmnic1 (Fabric B)	119-121	
Port-channel 88	vmnic2 (Fabric A)	201-210, 301-310,	VLANs for Gold,
		401-410, 1601-1610	Silver, Bronze, etc.
Port-channel 89	vmnic3 (Fabric B)		tenants
		501-520,6 01-620,	
		701-720 801-820	
		1801-1860,1990	
		2001-2010	

On the Nexus 5548 upstream switches to the tenants' data network, a vPC is configured to provide redundancy for the UCS compute stack. The configuration of one of the Nexus 5548 devices is shown below; the other Nexus 5548 has similar configuration.

```
interface port-channel88
  description vPC to dc02-ucs01-a
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  spanning-tree port type edge trunk
  vpc 88
interface port-channel89
  description vPC to dc02-ucs01-b
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  spanning-tree port type edge trunk
  vpc 89

interface Ethernet2/1
  description to UCS FI-A
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  channel-group 88 mode active
interface Ethernet2/2
  description to UCS FI-A
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  channel-group 88 mode active
interface Ethernet2/3
  description to UCS FI-B
  shutdown
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  channel-group 89 mode active
interface Ethernet2/4
  description to UCS FI-B
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,
1990,2001-2010
  channel-group 89 mode active
```

The upstream to the management network is the Catalyst 6500 Virtual Switch System (VSS). A vPC is not required/supported on the VSS, as the VSS uses Multi-Chassis EtherChannel (MEC). The member links for each port-channel consist of switchports from two different chassis. The configuration of the Catalyst 6500 VSS is shown below.

```
interface Port-channel98
 description dc02-ucs01-a
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
interface Port-channel99
 description dc02-ucs01-b
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
end

interface TenGigabitEthernet1/4/7
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
 channel-group 98 mode active
interface TenGigabitEthernet1/4/8
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
 channel-group 99 mode active
interface TenGigabitEthernet2/4/7
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
 channel-group 98 mode active
interface TenGigabitEthernet2/4/8
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 51-55,119-121
 switchport mode trunk
 channel-group 99 mode active
```

Note UCS FI uses Link Aggregation Control Protocol (LACP) as the port-channel aggregation protocol. The opposing upstream switches must be configured with LACP **active** mode.

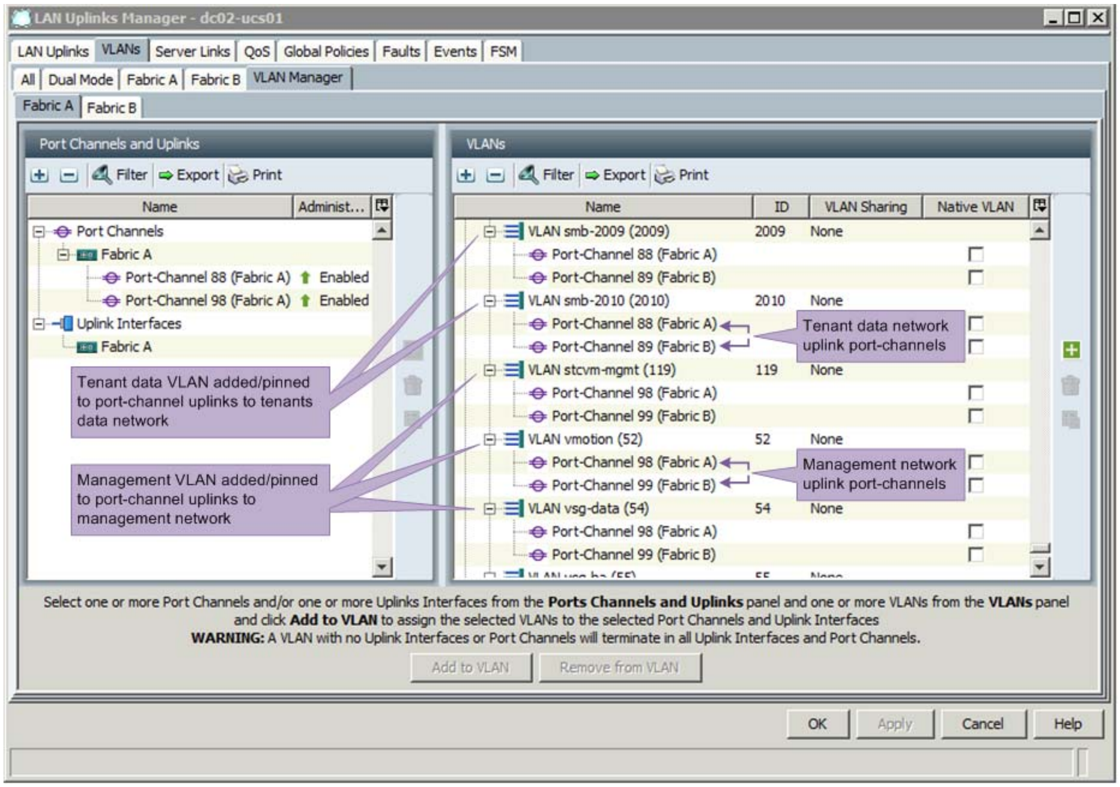
In this implementation, each blade server needs to communicate with both the management network and the tenants' data network. Since each vNIC (VMNIC in the ESXi hypervisor) can only communicate with one disjoint L2 network, at least two vNICs are required; for redundancy, four vNICs (two for management network, two for tenants' data network) are deployed per blade server, as shown in the figure and table above. The need for four vNICs on each half-width B-Series blade server mandates use of the Cisco Virtual Interface Card. The following adapters are acceptable:

- Cisco UCS Virtual Interface Card 1280

- Cisco UCS Virtual Interface Card 1240
- Cisco UCS M81KR Virtual Interface Card

Figure 2-4 shows the disjoint upstream L2 networks configuration in the UCSM VLAN Manager. On UCSM, global VLANs are configured for HA, and each VLAN is added/pinned to the respective port-channel uplinks on fabric A and fabric B in accordance to which upstream L2 network the VLAN belongs to. The VLANs for both upstream L2 networks do not overlap.

Figure 2-4 Adding/Pinning VLANs to Upstream Port-channels


LAN Uplinks Manager - dc02-ucs01
LAN Uplinks | VLANs | Server Links | QoS | Global Policies | Faults | Events | FSM
All | Dual Mode | Fabric A | Fabric B | VLAN Manager
Fabric A | Fabric B
Port Channels and Uplinks
Filter | Export | Print
VLANs
Filter | Export | Print
Tenant data VLAN added/pinned to port-channel uplinks to tenants data network
Management VLAN added/pinned to port-channel uplinks to management network
Tenant data network uplink port-channels
Management network uplink port-channels
Add to VLAN | Remove from VLAN
OK | Apply | Cancel | Help

Name	Administ...
Port Channels	
Fabric A	
Port-Channel 88 (Fabric A)	Enabled
Port-Channel 98 (Fabric A)	Enabled
Uplink Interfaces	
Fabric A	

Name	ID	VLAN Sharing	Native VLAN
VLAN smb-2009 (2009)	2009	None	
Port-Channel 88 (Fabric A)			☐
Port-Channel 89 (Fabric B)			☐
VLAN smb-2010 (2010)	2010	None	
Port-Channel 88 (Fabric A)			☐
Port-Channel 89 (Fabric B)			☐
VLAN stcvm-mgmt (119)	119	None	
Port-Channel 98 (Fabric A)			☐
Port-Channel 99 (Fabric B)			☐
VLAN vmotion (52)	52	None	
Port-Channel 98 (Fabric A)			☐
Port-Channel 99 (Fabric B)			☐
VLAN vsg-data (54)	54	None	
Port-Channel 98 (Fabric A)			☐
Port-Channel 99 (Fabric B)			☐

Select one or more Port Channels and/or one or more Uplinks Interfaces from the **Ports Channels and Uplinks** panel and one or more VLANs from the **VLANs** panel and click **Add to VLAN** to assign the selected VLANs to the selected Port Channels and Uplink Interfaces
WARNING: A VLAN with no Uplink Interfaces or Port Channels will terminate in all Uplink Interfaces and Port Channels.

Note 1.UCSM implicitly assigns default VLAN 1 to all uplink ports and port-channels. Do not configure any vNICs with default VLAN 1. It is advisable not to use VLAN 1 for carrying any user data traffic.

2. UCSM reserved some VLANs for internal system use, and these reserved VLANs should not be used to carry any user management and data traffic.

Storage Implementation Overview

Storage is part of the ICS stack. In VMDC 2.3 implementation, the NetApp Filer FAS6040 is used to provide the storage needs of the solution. The FAS6040 is based on a unified storage architecture and provides Storage Area Network (SAN) and Network-Attached Storage (NAS) capabilities on a single platform. In this solution, a common infrastructure is used to provide both SAN and NAS capabilities. The Nexus 5548 ICS switch provides LAN capabilities for NAS connectivity, and also is the FC switch that connects server blades and storage to provide SAN capability.

For details on storage best practices, refer to the NetApp FlexPod Solutions Guide, which provides an overview of FlexPod.

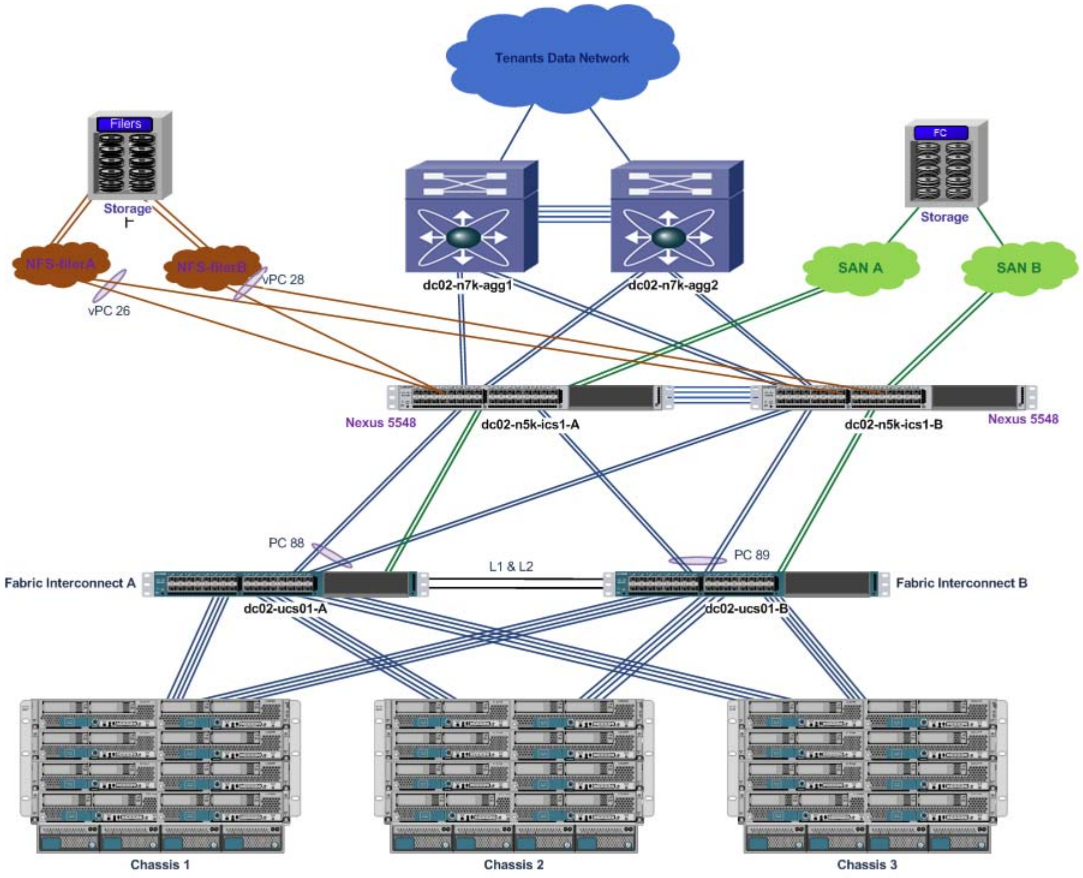
This section presents the following topics:

- SAN Implementation Overview, page 2-9
- NAS Implementation Overview, page 2-15
- NetApp FAS6040 Configuration Overview, page 2-25

SAN Implementation Overview

Figure 2-5 shows an overview of the SAN infrastructure. This section explains the Fibre Channel over Ethernet (FCoE) connection from servers to the FI and Fibre Channel (FC) connectivity to carry SAN traffic from the FI to the Nexus 5000 (storage switch) to NetApp Filers FAS6040.

Figure 2-5 Storage Infrastructure

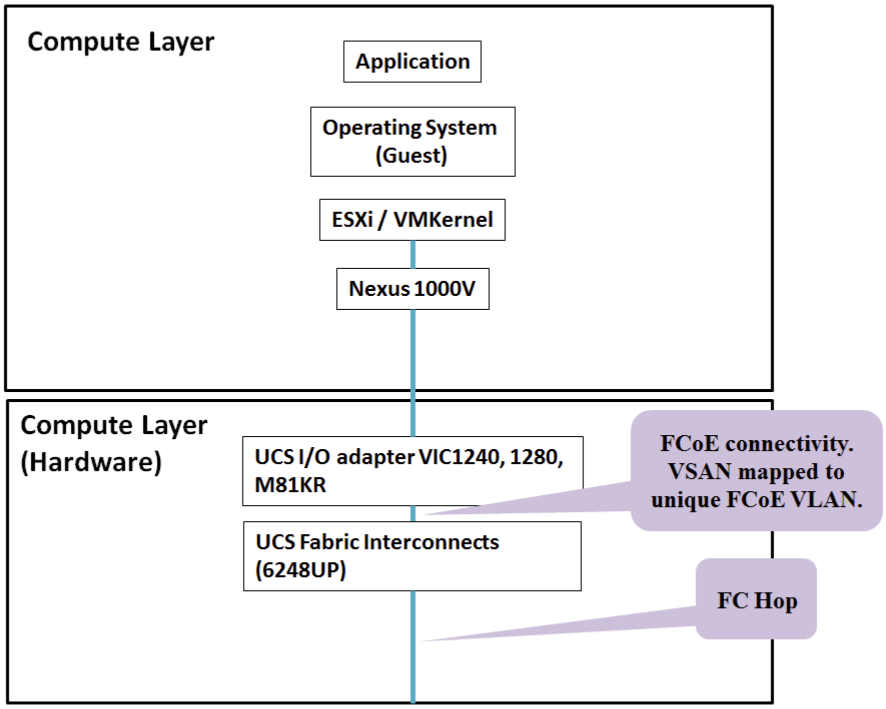


The following is an end-to-end flow diagram from application (user VM) to storage using SAN infrastructure. The compute, network, and storage portions of the flow are shown separately.

Compute

Figure 2-6 shows how the different components of the Compute layer are stacked up and the traffic that flows between them.

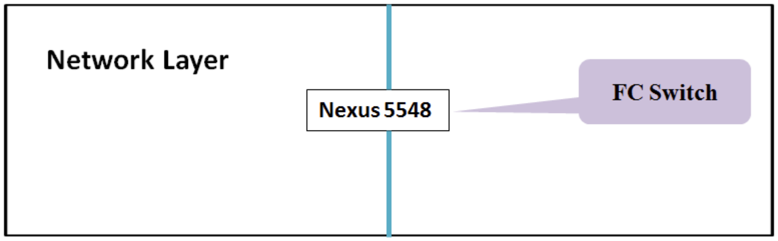
Figure 2-6 Compute Flow


Compute Layer
Application
Operating System (Guest)
ESXi / VMKernel
Nexus 1000V
Compute Layer (Hardware)
UCS I/O adapter VIC1240, 1280, M81KR
UCS Fabric Interconnects (6248UP)
FCoE connectivity. VSAN mapped to unique FCoE VLAN.
FC Hop

Network

Figure 2-7 shows the L2 switch, which is a Nexus 5500 series switch. Both the Server (Host) and Storage (NetApp Filer FAS6040) are connected to this switch, which is a part of the ICS stack.

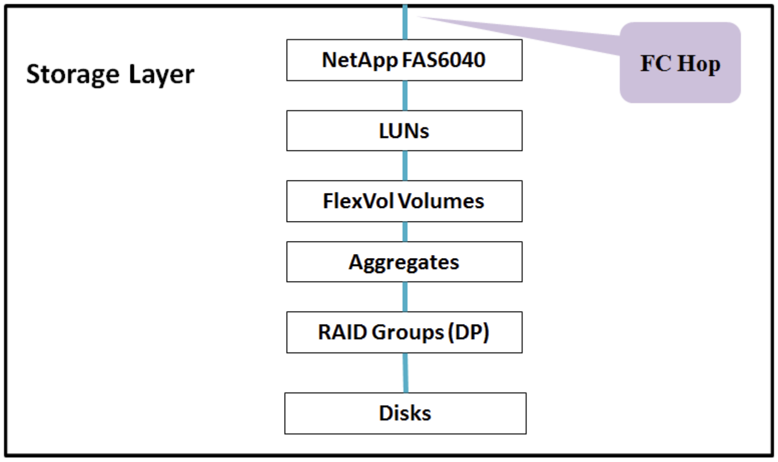
Figure 2-7 Network Flow



Storage

The NetApp FAS6040 provides the SAN storage, which is shown in Figure 2-8.

Figure 2-8 Storage Flow



This section presents the following topics:

- FCoE in UCS Fabric Interconnect, page 2-11
- SAN from Nexus 5500 to Storage, page 2-13
- NetApp FAS6040, page 2-14

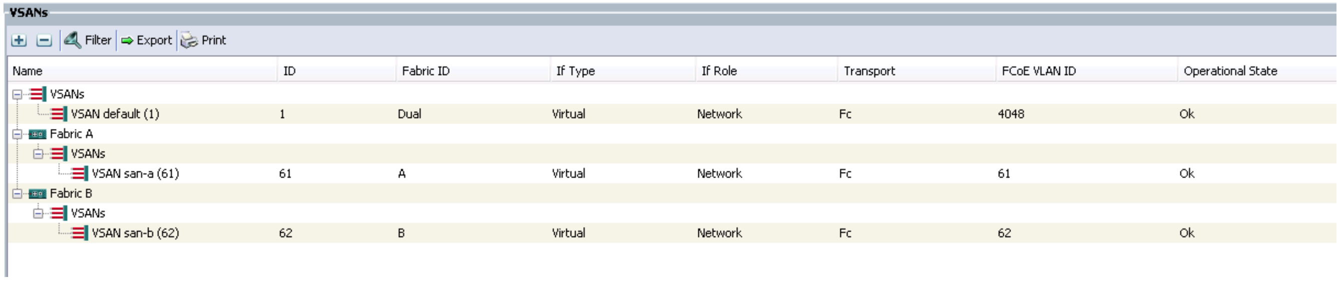
FCoE in UCS Fabric Interconnect

The salient features of FC configuration in the DC are as follows:

- Each blade server has two vHBAs that provide server to storage SAN connectivity. Virtual SANs (VSANs) are used to provide scalability, availability, and security, and allows multiple VSANs to share the common SAN infrastructure.
- Multiple vHBAs per blade server provide HBA redundancy at the server.
- Storage traffic from server blades to FIs is FCoE. Each VSAN is mapped to a unique VLAN that carries storage traffic from server to FI.
- FC traffic is mapped to a no-packet drop class using the system Quality of Service (Qos) policy. This assures that FC traffic will not be dropped during congestion.
- Storage traffic from FIs to the Nexus 5548 ICS switch is sent as FC traffic using FC port-channels. Each port-channel has two links, and thus, provides link-level redundancy.
- Each FI is mapped to one VSAN. In this case, FI-A carries all VSAN61 traffic and FI-B carries all VSAN62 traffic.
- The Nexus 5548 ICS switch is configured in N-port Identifier Virtualization (NPIV) mode. SAN storage traffic from FI-A is sent to the Nexus 5548 ICS switch-A; likewise, all of the SAN traffic from FI-B is sent to the Nexus 5548 ICS switch-B.
- The port-channel-trunk feature is enabled on the Nexus 5000 to enable port-channel configuration on FC interfaces connected to FIs.

Figure 2-9 configuration shows the list of VSANs in the SAN infrastructure: VSAN61, VSAN62, and the mapping of each of those VSANs to a unique FCoE VLAN ID (61,62). This mapping is required because both SAN and LAN traffic is carried using the same FCoE links between server and FIs. VSAN61 is transported on Fabric-A, and VSAN62 is transported on Fabric-B. Even though they share the common infrastructure, the traffic that flows on them is strictly isolated.

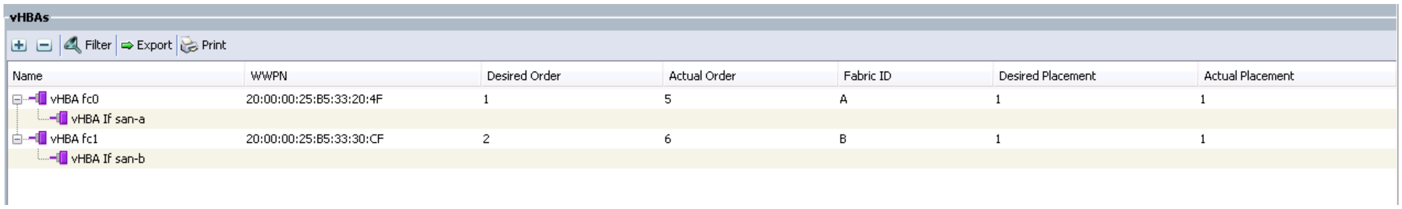
Figure 2-9 VSANs in the SAN Infrastructure



Name	ID	Fabric ID	If Type	If Role	Transport	FCoE VLAN ID	Operational State
VSANs							
VSAN default (1)	1	Dual	Virtual	Network	Fc	4048	Ok
Fabric A							
VSANs							
VSAN san-a (61)	61	A	Virtual	Network	Fc	61	Ok
Fabric B							
VSANs							
VSAN san-b (62)	62	B	Virtual	Network	Fc	62	Ok

Figure 2-10 configuration shows the vHBA configuration on each server blade. vHBAs are configured using service profiles generated from service-profile templates. There are two vHBA adapters configured per server blade. As shown, vHBA0 traffic is sent on san-a, and vHBA1 traffic is sent on san-b. Each vHBA is placed on a unique, isolated SAN network.

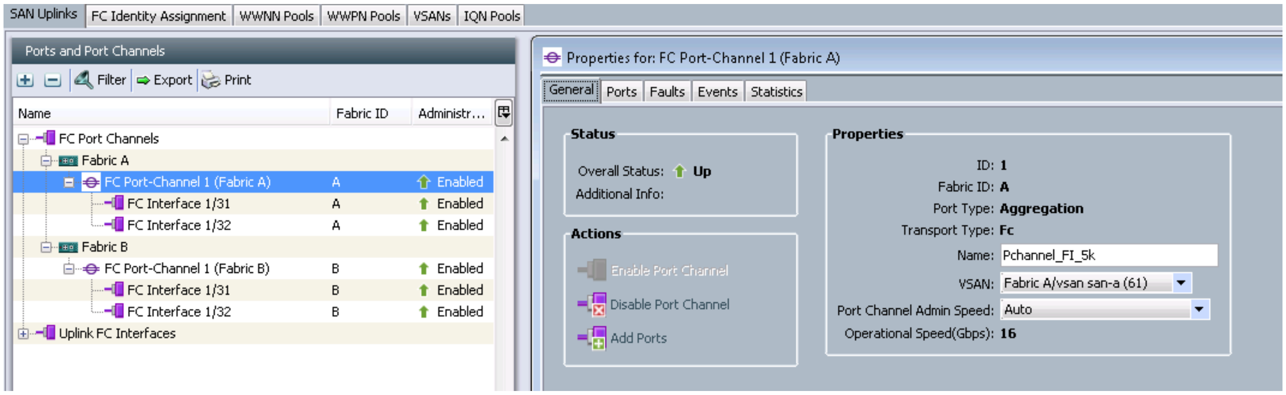
Figure 2-10 vHBA Configuration on Each Server Blade



Name	WWPN	Desired Order	Actual Order	Fabric ID	Desired Placement	Actual Placement
vHBA fc0	20:00:00:25:B5:33:20:4F	1	5	A	1	1
vHBA If san-a						
vHBA fc1	20:00:00:25:B5:33:30:CF	2	6	B	1	1
vHBA If san-b						

Figure 2-11 configuration shows the port-channel configuration of the FC between FI and the Nexus 5548 ICS switch for SAN traffic. Port-channel(1) is shown expanded in the right column. Portchannel(1) carries FC SAN traffic that flows on VSAN61 from FI-A to the Nexus 5548 ICS switch-A. Similarly, port-channel(2) carries VSAN62 traffic from FI-B to the Nexus 5548 ICS switch-B.

Figure 2-11 Port-channel Configuration of FC Between FI and Nexus 5548 ICS switch for SAN Traffic



Name	Fabric ID	Administr...
FC Port Channels		
Fabric A		
FC Port-Channel 1 (Fabric A)	A	Enabled
FC Interface 1/31	A	Enabled
FC Interface 1/32	A	Enabled
Fabric B		
FC Port-Channel 1 (Fabric B)	B	Enabled
FC Interface 1/31	B	Enabled
FC Interface 1/32	B	Enabled
Uplink FC Interfaces		

Overall Status: Up
Additional Info:

ID: 1
Fabric ID: A
Port Type: Aggregation
Transport Type: Fc
Name: Pchannel_FI_5k
VSAN: Fabric A/vsan san-a (61)
Port Channel Admin Speed: Auto
Operational Speed(Gbps): 16

SAN from Nexus 5500 to Storage

The following are the salient points of the SAN configuration:

- Nexus 5000-A carries SAN-A(VSAN61) traffic from FIs to NetApp filer-A and filer-B. Similarly, Nexus 5000-B carries SAN-B(VSAN62) traffic.
- FC links between the Nexus 5548 ICS switch and FIs are configured as the SAN port-channel.
- Each Nexus is connected to both filer-A and filer-B for filer-level redundancy.
- The Nexus is configured in NPIV mode. FC ports connected to FIs or the NetApp Filer are configured as F ports. The Nexus 5548 ICS switch is configured to be the FC switch. The following configuration needs to be enabled on the FC switch, the Nexus 5548:

```
feature npiv
feature fport-channel-trunk
```

- Soft zoning (using World Wide Port Name (WWPN) names) is configured on the Nexus to allow servers with specific identity (WWPN) to communicate only with NetApp filers. Each filer connection has its own WWPN name. The configuration below shows the zoning configuration for one server blade per VSAN (in this case, SAN-B). As mentioned before, vHBA1 of any server blade is placed in the SAN-B infrastructure and vHBA0 is placed in SAN-A.

The zoning configuration is shown below.

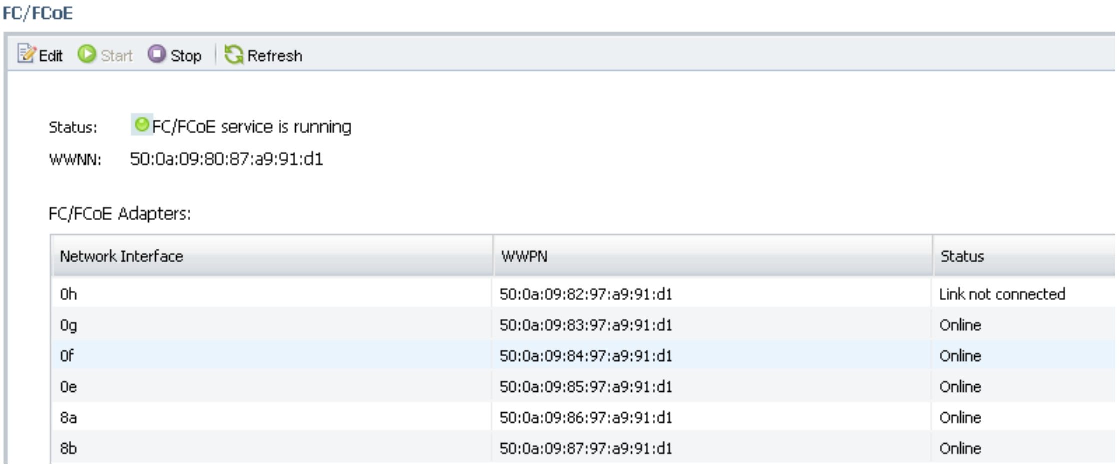
```
zone name dc02-c03-esxi08-hba1 vsan 62
pwwn 20:00:00:25:b5:33:30:9e \[dc02-c03-esxi08-hba1\] pwwn 50:0a:09:87:97:a9:91:d1
\[netapp-filera\] pwwn 50:0a:09:87:87:a9:91:d1 \[netapp-filerb\]
```

- As you can infer from the above configuration, "single initiator zoning" has been implemented. Each zone contains only one host server vHBA and can contain multiple storage array targets in the same zone.
- The FC interface on the Nexus 5548 ICS switch is used to connect to the NetApp FAS6040 for FC connectivity. Below is the interface configuration.

```
interface fc2/16
switchport mode F
switchport description to Netapp FAS6040-B no shutdown
```

- The WWPN of vHBAs is obtained from the UCSM (shown in the previous section). The WWPN of NetApp filers is fetched using the NetApp OnCommand System Manager GUI. Figure 2-12 is a screenshot that details the WWPN for each of the ports connected on Filer-a. An **Online** status implies that the FC link is up, whereas, an **Offline** status implies that the FC link is down.

Figure 2-12 VWPN for Ports Connected on Filer-a



Status: FC/FCoE service is running
WWNN: 50:0a:09:80:87:a9:91:d1

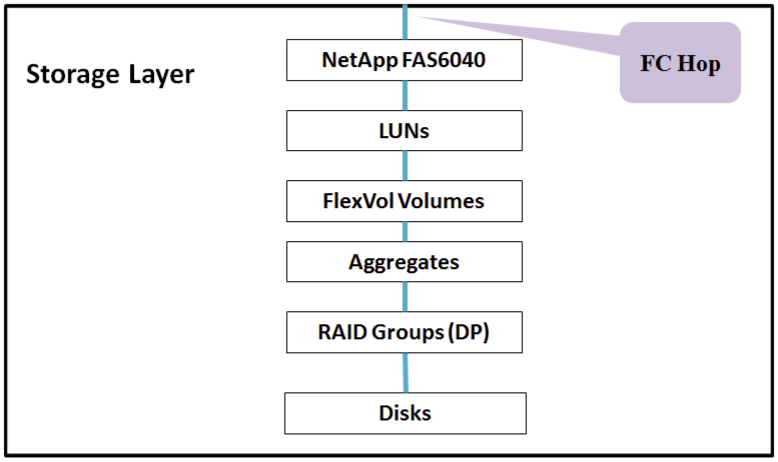
FC/FCoE Adapters:

Network Interface	WWPN	Status
0h	50:0a:09:82:97:a9:91:d1	Link not connected
0g	50:0a:09:83:97:a9:91:d1	Online
0f	50:0a:09:84:97:a9:91:d1	Online
0e	50:0a:09:85:97:a9:91:d1	Online
8a	50:0a:09:86:97:a9:91:d1	Online
8b	50:0a:09:87:97:a9:91:d1	Online

NetApp FAS6040

Figure 2-13 shows the end-to-end logical components within the NetApp FAS6040.

Figure 2-13 End-to-end Logical Components within the NetApp FAS6040



Starting from the bottom up, raw disks are formatted with a RAID group (in this case, RAID6) and grouped to form an Aggregate. From the Aggregate, volumes are carved out. From the volumes, LUNs are created. LUNs appear as logical disks to server hosts during a LUN scan.

The following are the salient features of the SAN configuration on the NetApp FAS6040:

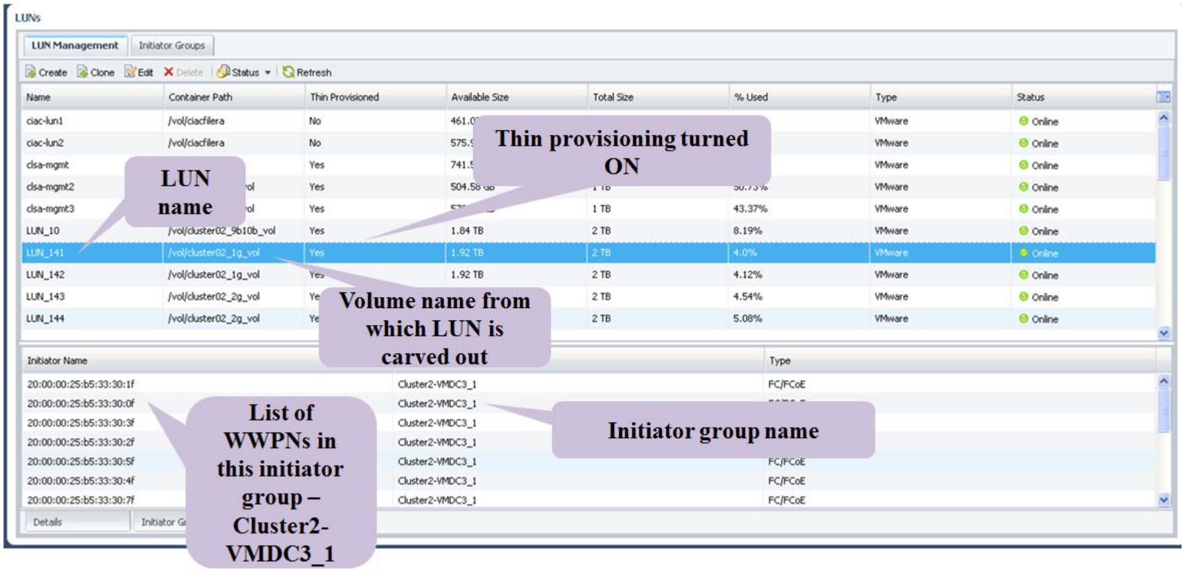
- Initiator groups are used to grant the server hosts access to LUNs.
- To avoid problems with Virtual Machine File System (VMFS) resignaturing, it is recommended that within a SAN environment, all ESXi hosts refer to the same LUN using the same LUN ID.
- In this implementation, a single initiator group is used for all of the VMware ESXi hosts in the SAN environment and then mapped LUNs to LUN IDs using that single initiator group, however, depending on the deployment and management needs, multiple initiator groups can also be used.

The following steps show an overview of the SAN storage configuration:

1. Configure Aggregates on the NetApp filer from physical disks and configure RAID-DP.
2. Configure volumes from Aggregates. Volumes are Thin Provisioned.

Figure 2-14 shows the configuration of LUNs from a previously defined volume.

Figure 2-14 NetApp LUNs


LUNs
LUN Management
Initiator Groups
Create
Clone
Edit
Delete
Status
Refresh
Thin provisioning turned ON
LUN name
Volume name from which LUN is carved out
List of WWPNs in this initiator group – Cluster2-VMDC3_1
Initiator group name
Details
Initiator G

Name	Container Path	Thin Provisioned	Available Size	Total Size	% Used	Type	Status
ciac-lun1	/vol/ciacfilera	No	461.0	[illegible]	[illegible]	VMware	Online
ciac-lun2	/vol/ciacfilera	No	575.9	[illegible]	[illegible]	VMware	Online
clsa-mgmt	[illegible]	Yes	741.5	[illegible]	[illegible]	VMware	Online
clsa-mgmt2	[illegible]	Yes	504.58 GB	1 TB	50.73%	VMware	Online
clsa-mgmt3	[illegible]	Yes	[illegible]	1 TB	43.37%	VMware	Online
LUN_10	/vol/cluster02_9b10b_vol	Yes	1.84 TB	2 TB	8.19%	VMware	Online
LUN_141	/vol/cluster02_1g_vol	Yes	1.92 TB	2 TB	4.0%	VMware	Online
LUN_142	/vol/cluster02_1g_vol	Yes	1.92 TB	2 TB	4.12%	VMware	Online
LUN_143	/vol/cluster02_2g_vol	Ye	[illegible]	2 TB	4.54%	VMware	Online
LUN_144	/vol/cluster02_2g_vol	Ye	[illegible]	2 TB	5.08%	VMware	Online

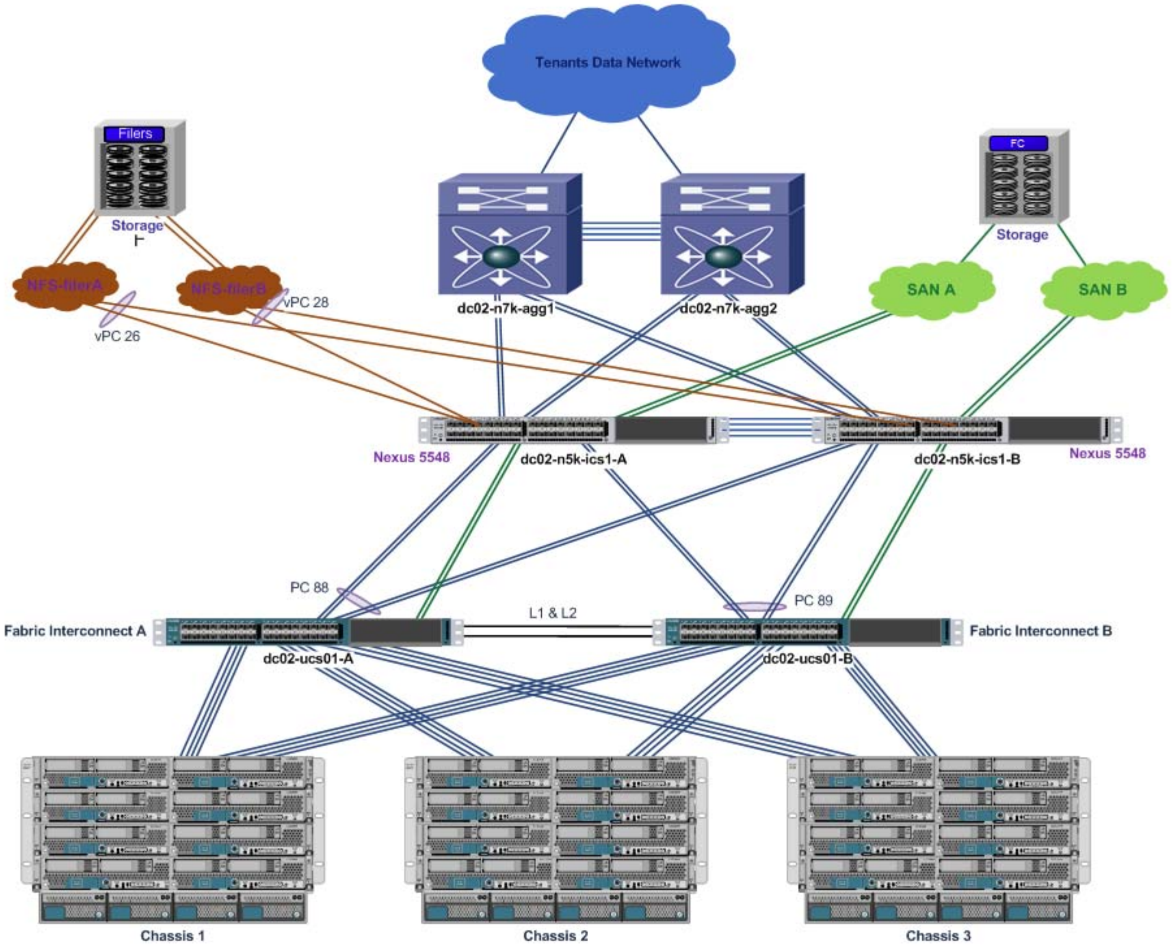
Initiator Name		Type
20:00:00:25:b5:33:30:1f	Cluster2-VMDC3_1	FC/FCoE
20:00:00:25:b5:33:30:0f	Cluster2-VMDC3_1	[illegible]
20:00:00:25:b5:33:30:3f	Cluster2-VMDC3_1	[illegible]
20:00:00:25:b5:33:30:2f	Cluster2-VMDC3_1	[illegible]
20:00:00:25:b5:33:30:5f	Cluster2-VMDC3_1	FC/FCoE
20:00:00:25:b5:33:30:4f	Cluster2-VMDC3_1	FC/FCoE
20:00:00:25:b5:33:30:7f	Cluster2-VMDC3_1	FC/FCoE

The LUN provides block-level storage to the server. The operating system (in this case, ESXi) is provided with a unique list of LUNs based on the server adapter WWPN. Each LUN is configured with a LUNID, that is commonly referred to as the Host LUNID (the ID that the host will use to access a particular LUN).

NAS Implementation Overview

Figure 2-15 provides an overview of the end-to-end storage infrastructure.

Figure 2-15 End-to-End Storage Infrastructure

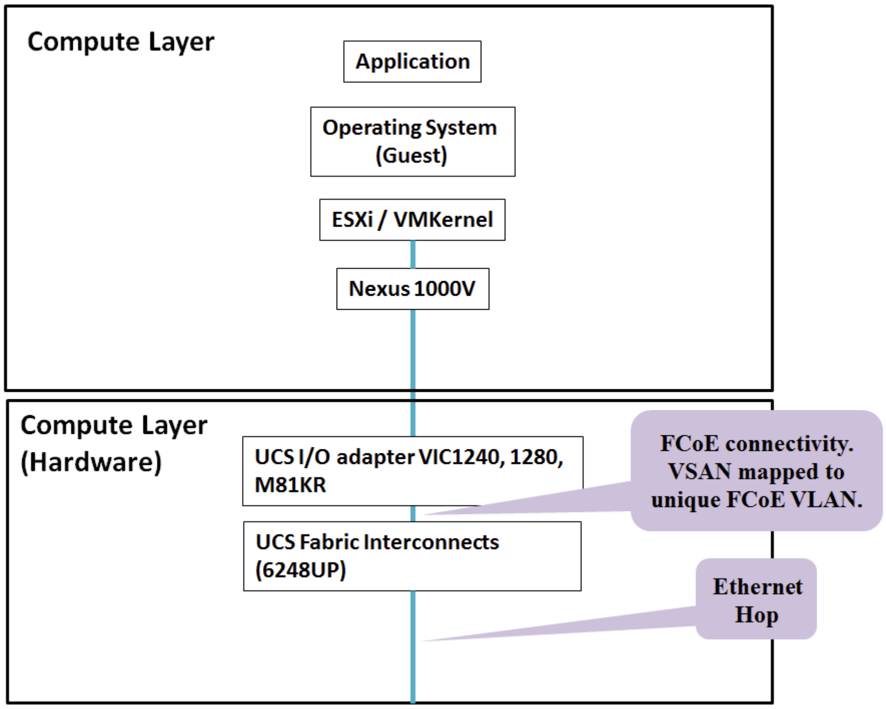


The following is an end-to-end flow diagram from application (user VM) to storage using NFS infrastructure. The compute, network, and storage portions of the flow are shown separately.

Compute

Figure 2-16 shows how the different components of the compute layer are stacked up and the traffic that flows between them.

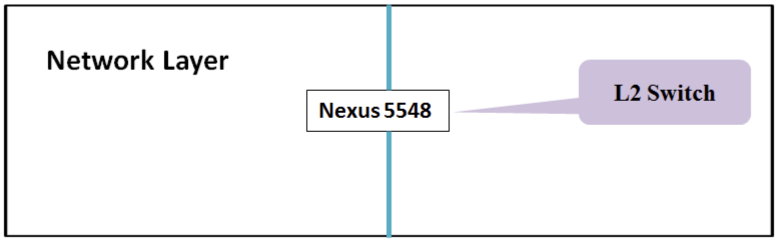
Figure 2-16 Compute Flow


Compute Layer
Application
Operating System (Guest)
ESXi / VMKernel
Nexus 1000V
Compute Layer (Hardware)
UCS I/O adapter VIC1240, 1280, M81KR
UCS Fabric Interconnects (6248UP)
FCoE connectivity. VSAN mapped to unique FCoE VLAN.
Ethernet Hop

Network

Figure 2-17 shows the L2 switch, which is a Nexus 5500 series switch. Both the Server (Host) and Storage (NetApp Filer FAS6040) are connected to this switch, which is a part of the ICS stack.

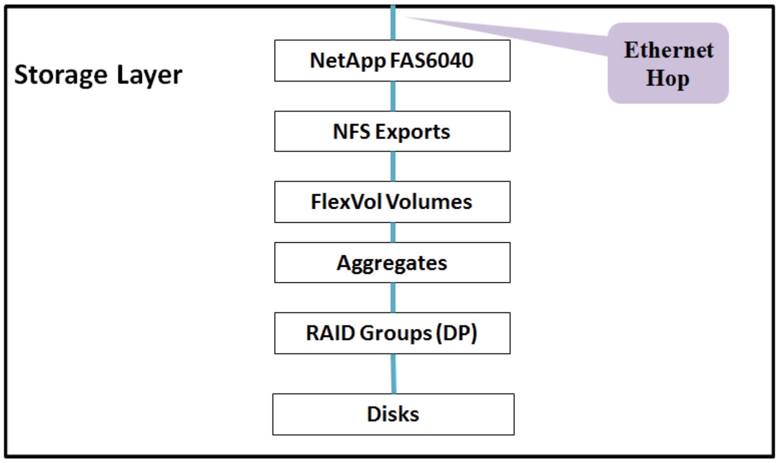
Figure 2-17 Network Flow



Storage

The storage filers used to provide NFS are the NetApp FAS6040, as shown in Figure 2-18.

Figure 2-18 Storage Flow



This section presents the following topics:

- NFS from Fabric Interconnect to Nexus 5500, page 2-18
- NFS from Nexus 5500 to FAS6040, page 2-22
- NetApp FAS6040—NFS, page 2-22

NFS from Fabric Interconnect to Nexus 5500

The salient features of the Network File System (NFS) configuration in the DC are as follows:

- NFS traffic is isolated from other LAN traffic by using a separate VLAN.
- Configure the NFS VLAN on the Nexus 1000V, UCSM, and Nexus 5548 ICS switches. In this case, VLAN 1990 is chosen to be the NFS VLAN.
- It is important to pin the NFS VLAN only on all of the uplinks in the data network (data uplinks are - Po88 and Po89). Refer to the end-to-end storage infrastructure diagram here). This configuration is done both on the UCSM and Nexus 1000V. This implies that the NFS VLAN must be blocked on the management network port-channels/uplinks.

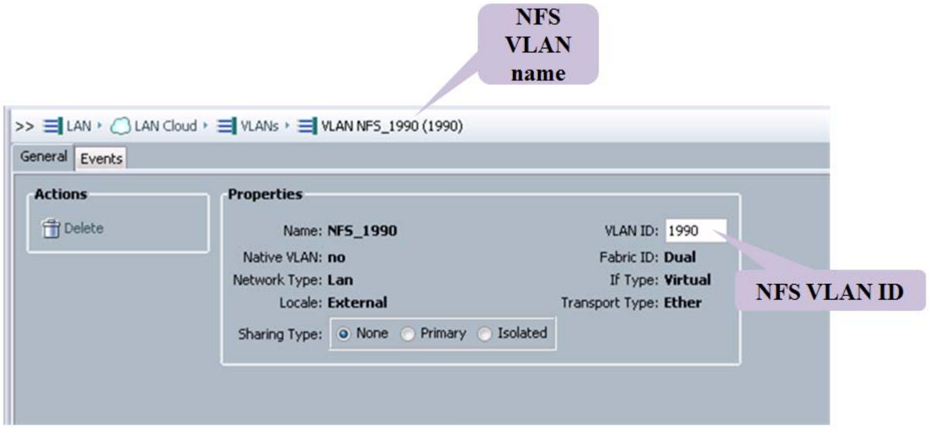
Below is an overview of the steps involved in NFS configuration on the UCS, FIs, and Nexus 5548 ICS switch.

1. To enable NFS connectivity, the ESX server requires a special connection type, referred to as a VMKernel port.
2. Create a new VMKernel port on the ESXi host and allow the NFS VLAN. Configure an IP (in the NFS subnet) for the VMKernel port for every server blade that needs NFS connectivity.

3. Configure an NFS VLAN port-group profile in Nexus 1000V.
4. Allow the VLAN on the LAN port-channel between the UCS and Nexus 5548 ICS switch.
5. Mount the NFS filer on the VMware ESXi Host (configure the mount after finishing the configuration on NetApp filers).

Below are snapshot diagrams that explain the details of the steps listed above. Figure 2-19 shows the configuration of the NFS VLAN in the UCSM.

Figure 2-19 NFS VLAN in the UCSM



>> LAN ▸ LAN Cloud ▸ VLANs ▸ VLAN NFS_1990 (1990)

General | Events

Actions

Delete

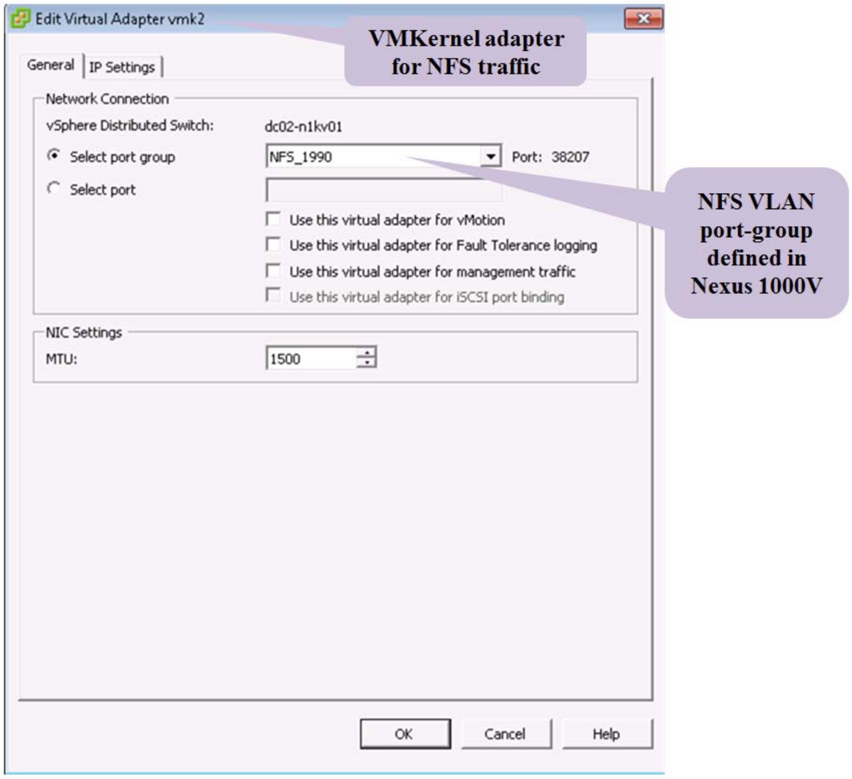
Properties

Name: NFS_1990
Native VLAN: no
Network Type: Lan
Locale: External
Sharing Type: ● None ○ Primary ○ Isolated

VLAN ID: 1990
Fabric ID: Dual
If Type: Virtual
Transport Type: Ether

Figure 2-20 shows the configuration of the VMKernel interface, which is used by the server to establish NFS connectivity on the LAN.

Figure 2-20 VMKernel Interface



Network Connection

vSphere Distributed Switch: dc02-n1kv01

● Select port group NFS_1990 Port: 38207

○ Select port

□ Use this virtual adapter for vMotion

□ Use this virtual adapter for Fault Tolerance logging

□ Use this virtual adapter for management traffic

□ Use this virtual adapter for iSCSI port binding

NIC Settings

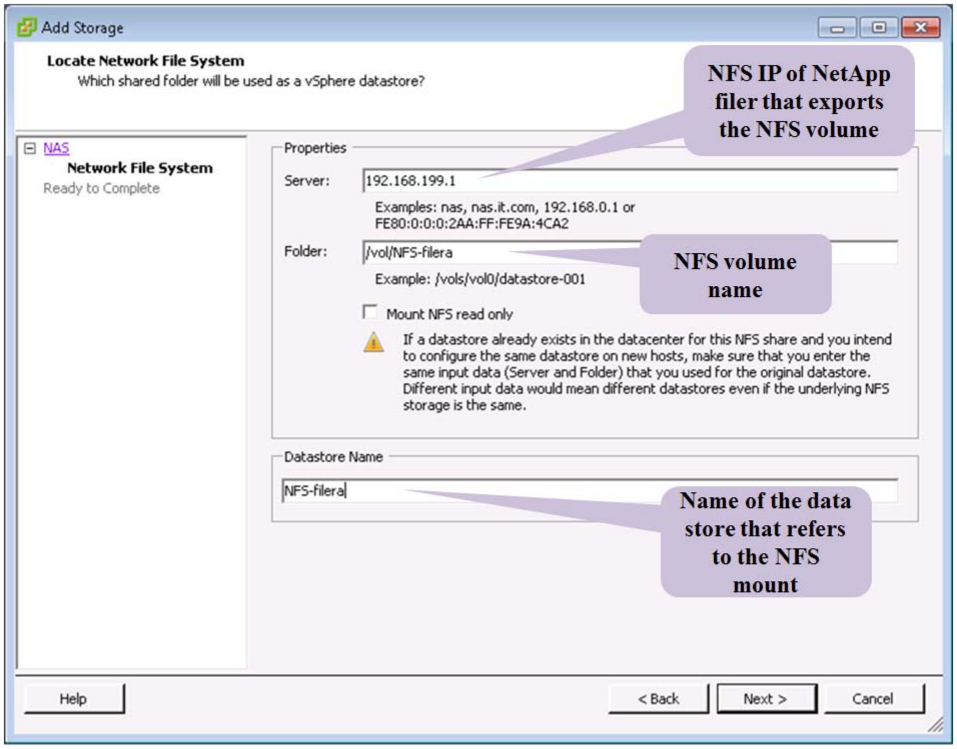
MTU: 1500

VMKernel adapter for NFS traffic

NFS VLAN port-group defined in Nexus 1000V

Assuming, that storage for NFS is set up (refer to NetApp FAS6040 - NFS for more details), Figure 2-21 provides an overview of NFS mount in the VMware vCenter.

Figure 2-21 NFS Mount in VMware vCenter



Locate Network File System

Which shared folder will be used as a vSphere datastore?

Properties

Server: 192.168.199.1

Examples: nas, nas.it.com, 192.168.0.1 or FE80:0:0:0:2AA:FF:FE9A:4CA2

Folder: /vol/NFS-filera

Example: /vols/vol0/datastore-001

□ Mount NFS read only

⚠ If a datastore already exists in the datacenter for this NFS share and you intend to configure the same datastore on new hosts, make sure that you enter the same input data (Server and Folder) that you used for the original datastore. Different input data would mean different datastores even if the underlying NFS storage is the same.

Datastore Name

NFS-filera

NFS IP of NetApp filer that exports the NFS volume

NFS volume name

Name of the data store that refers to the NFS mount

Below is the configuration of the NFS VLAN in the Nexus 1000V.

```
port-profile NFS_1990
 type: Vethernet
 description:
 status: enabled
 max-ports: 32
 min-ports: 1
 inherit:
 config attributes:
  switchport mode access
  switchport access vlan 1990
  service-policy input nfs
  no shutdown
 evaluated config attributes:
  switchport mode access
  switchport access vlan 1990
  service-policy input nfs
  no shutdown
 assigned interfaces:
  Vethernet1413
 port-group: NFS_1990
 system vlans: none
 capability l3control: no
 capability iscsi-multipath: no
 capability vxlan: no
 capability l3-vservice: no
 port-profile role: none
 port-binding: ephemeral
```

NFS from Nexus 5500 to FAS6040

This section provides an overview of the configuration required to establish NFS connectivity between the Nexus 5548 ICS switch and NetApp filers FAS6040. The salient features of NFS configuration are as follows:

- In this implementation, NFS filers are in the same subnet as ESXi hosts (NFS clients). NFS traffic is carried in the same end-to-end VLAN (L2) between server blades and filers. In this implementation, we have NFS ESXi clients and storage in the same subnet as it minimizes latency (eliminates routing overhead) and reduces the number of hops.
- NFS storage traffic between FIs and the Nexus 5500 is carried on the Ethernet port-channel, and traffic isolation is achieved using a separate VLAN (VLAN 1990), which is the same as the FCoE VLAN used for carrying NFS traffic from the UCS to the Nexus 5548 ICS switch.
- NFS traffic between the Nexus 5500 and NetApp filers is sent on 10G links as Ethernet traffic. There is one Ethernet port-channel for every filer, which is connected to both the Nexus 5548 ICS switchs (as a vPC).

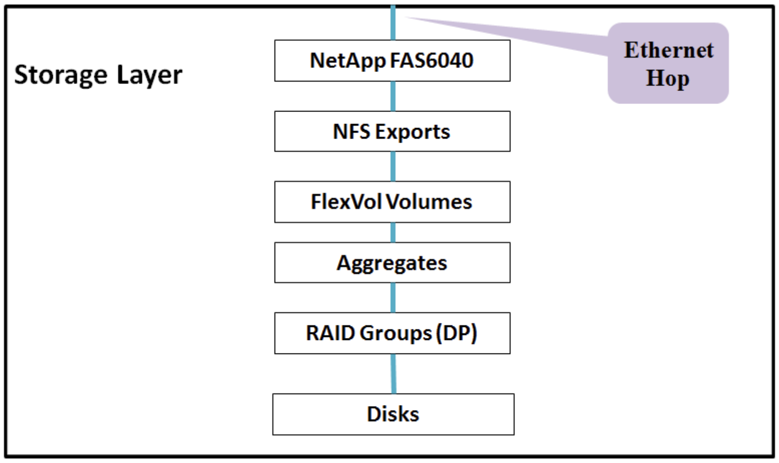
The following configuration shows the Nexus 5548 ICS switch configuration of the vPC, which is connected to the Ethernet interface on the NetApp filer:

```
interface port-channel28
  description vPC to netapp -B
  switchport mode trunk
  switchport trunk allowed vlan 1990
  service-policy type queuing input vmdc-nas-in-policy
  service-policy type queuing output vmdc-nas-out-policy
  vpc 28
```

NetApp FAS6040—NFS

Figure 2-22 shows the end-to-end logical components within the NetApp FAS6040.

Figure 2-22 End-to-end Components within the NetApp FAS6040



Starting from the bottom up, raw disks are formatted with a RAID group (in this case, RAID6) and grouped to form an Aggregate. From the Aggregate, volumes are carved out. Volumes are exported using the NFS protocol, and security rules are also applied on the volume (to selectively filter the clients accessing the NFS share and also assign authorization levels).

The following are the salient features of the NAS configuration on the NetApp FAS6040:

- NFS volume that is exported must be given a security style, Unix or Windows depending on whether NFS or CIFS protocol is used.
- NFS uplinks to the Nexus 5548 ICS switch are bundled as a port-channel for link redundancy. A multi-mode VIF port-channel is created and uses the LACP protocol for efficient port-channel management.

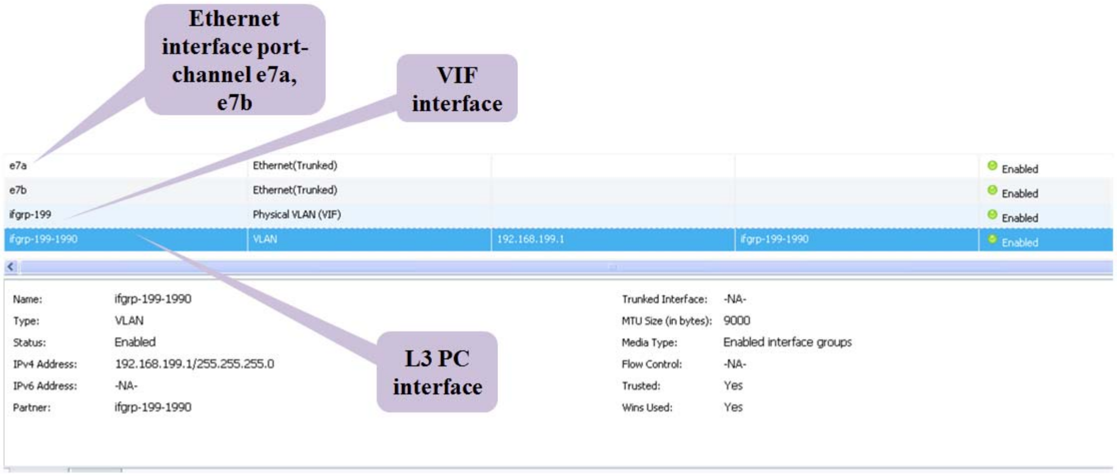
The following steps show an overview of the NFS storage configuration:

1. Configure Aggregates on the NetApp filer from physical disks and configure RAID-DP.
2. Configure volumes from Aggregates.
3. Start the NFS server using the **nfs on** CLI command.
4. Configure the multi-mode VIF port-channel (uplinks) connected to the Nexus 5548 ICS switch. Below is the status of the VIF port-channel configured on one of the filers.

```
s6040a-vmdc> vif status
ifgrp: command "vif" is deprecated in favor of command "ifgrp"
default: transmit 'IP Load balancing', Ifgrp Type 'multi_mode', fail 'log'
ifgrp-199: 2 links, transmit 'IP+port Load balancing', Ifgrp Type 'lacp' fail
'default'
         Ifgrp Status   Up      Addr_set
        up:
        e7b: state up, since 03Mar2013 22:11:38 (7+12:45:17)
                mediatype: auto-10g_sr-fd-up
                flags: enabled
                active aggr, aggr port: e7a
                input packets 1546529, input bytes 1164594806
                input lacp packets 80171, output lacp packets 80603
                output packets 876806, output bytes 139597180
                up indications 213, broken indications 140
                drops (if) 0, drops (link) 0
                indication: up at 03Mar2013 22:11:38
                        consecutive 0, transitions 353
        e7a: state up, since 06Feb2013 15:27:07 (32+19:29:48)
                mediatype: auto-10g_sr-fd-up
                flags: enabled
                active aggr, aggr port: e7a
                input packets 2745009, input bytes 2221709322
                input lacp packets 100463, output lacp packets 100476
                output packets 2261379, output bytes 697836908
                up indications 9, broken indications 5
                drops (if) 0, drops (link) 0
                indication: up at 06Feb2013 15:27:07
                        consecutive 0, transitions 14
```

A virtual network interface (VIF) is a mechanism that supports aggregation of network interfaces into one logical interface unit. Once created, a VIF is indistinguishable from a physical network interface. VIFs are used to provide fault tolerance of the network connection, and in some cases, higher throughput to the storage device. Figure 2-23 shows the port-channel and VLAN formed using Ethernet interfaces on filers that provide NFS connectivity.

Figure 2-23 NFS Port-channel



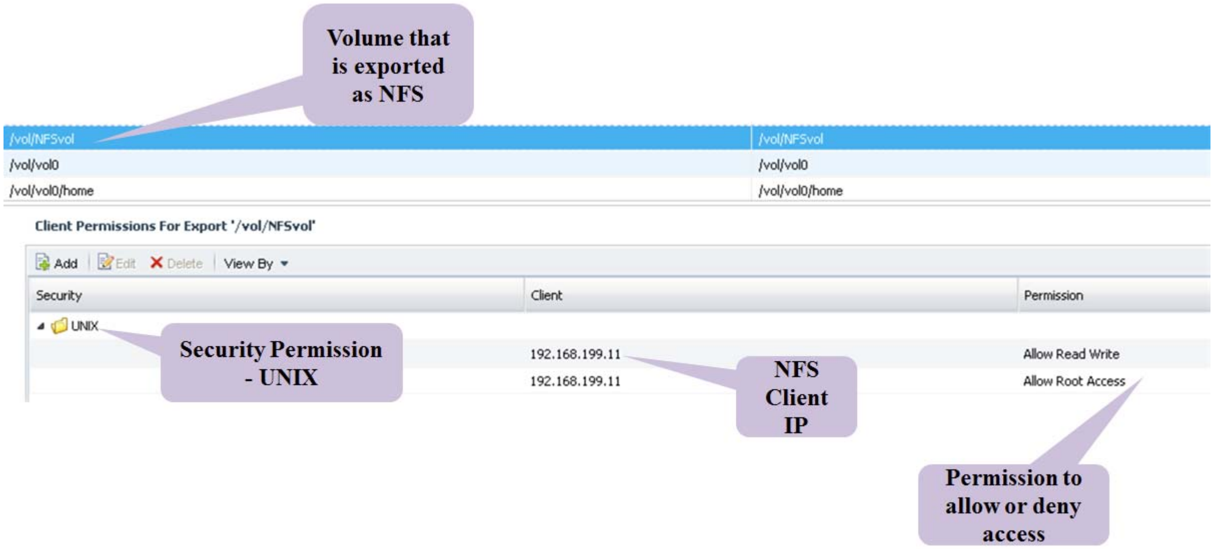
e7a	Ethernet(Trunked)			Enabled
e7b	Ethernet(Trunked)			Enabled
ifgrp-199	Physical VLAN (VIF)			Enabled
ifgrp-199-1990	VLAN	192.168.199.1	ifgrp-199-1990	Enabled

Name:	ifgrp-199-1990	Trunked Interface:	-NA-
Type:	VLAN	MTU Size (in bytes):	9000
Status:	Enabled	Media Type:	Enabled interface groups
IPv4 Address:	192.168.199.1/255.255.255.0	Flow Control:	-NA-
IPv6 Address:	-NA-	Trusted:	Yes
Partner:	ifgrp-199-1990	Wins Used:	Yes

Note that the MTU configured on the NFS interface (shown in Figure 2-23) is 9000. It is mandatory to configure **system jumbomtu 9216** on the ICS Cisco Nexus 5500 series switches to avoid MTU mismatch errors. Jumbo MTU size refers to the MTU size for L2 interfaces.

Select a volume for export using NFS. Figure 2-24 shows a volume exported as an NFS volume. Use exportfs to export the volume using NFS. The security permissions on this volume are Unix style, and also have the rules to allow/deny NFS client requests (from servers) based on the NFS IP address.

Figure 2-24 NFS Volume


Volume that is exported as NFS
Add
Edit
Delete
View By
Security Permission - UNIX
NFS Client IP
Permission to allow or deny access

/vol/NFSvol	/vol/NFSvol
/vol/vol0	/vol/vol0
/vol/vol0/home	/vol/vol0/home

Client Permissions For Export '/vol/NFSvol'

Security	Client	Permission
UNIX		
	192.168.199.11	Allow Read Write
	192.168.199.11	Allow Root Access

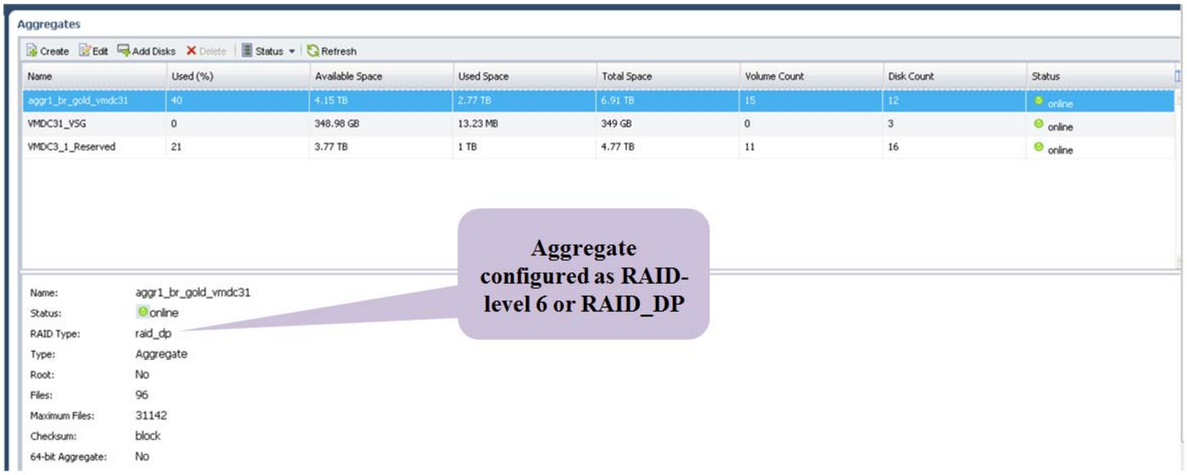
NetApp FAS6040 Configuration Overview

VMDC 2.3 supports SAN or NAS storage options depending on the overall DC requirements. To ensure HA of storage, two FAS6040 filers are configured to operate in seven-mode cluster. They both run the same version of the ONTAP 8.1.1 operating system software. Each NetApp Fabric-Attached Storage (FAS) controller shares a NetApp Unified Storage Architecture based on the Data ONTAP 8G operating system and uses an integrated suite of application-aware manageability software. This efficiently consolidates SAN, NAS, primary storage, and secondary storage on a single platform while allowing concurrent support for block and file protocols by using Ethernet and FC interfaces. These interfaces include FCoE, Network File System (NFS), Common Internet File System protocol (CIFS), and Internet Small Computer System Interface (iSCSI).

The data disks are configured as a Redundant Array of Independent Disks (RAID) group with RAID-level Double Parity (RAID-DP) (NetApp high-performance RAID 6), which offers superior data protection with little or no performance loss. Data deduplication is enabled on volumes to increase storage efficiency by saving storage space. Thin Provisioning is enabled on all of the Logical Unit Numbers (LUNs) to maximize storage capacity utilization efficiency.

Figure 2-25 shows the configuration of Aggregates from data disks and configuration of RAID-DP.

Figure 2-25 NetApp Aggregate



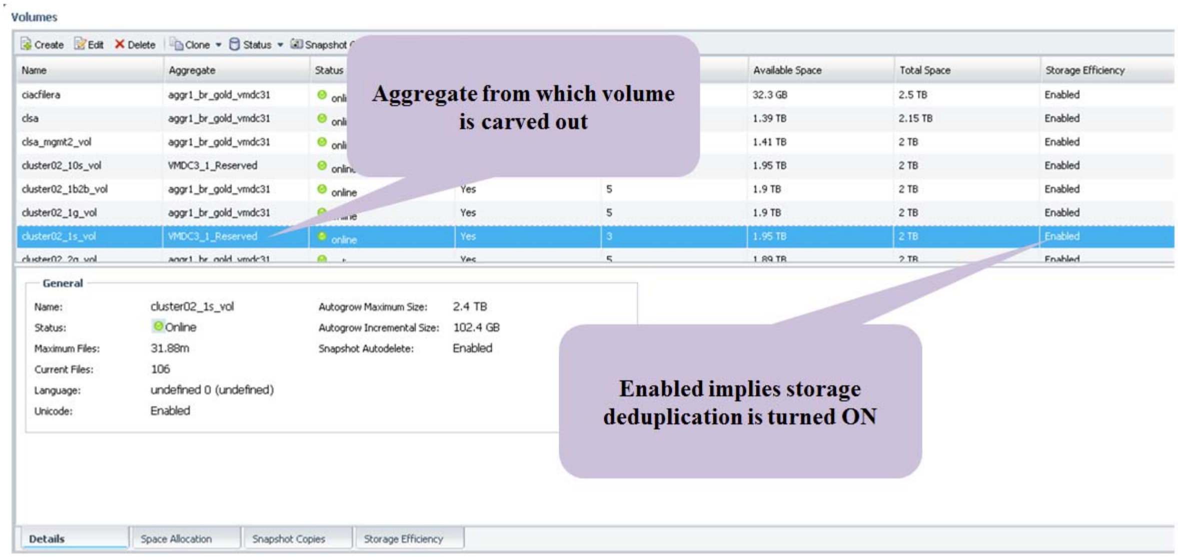
Name	Used (%)	Available Space	Used Space	Total Space	Volume Count	Disk Count	Status
aggr1_br_gold_vmdc31	40	4.15 TB	2.77 TB	6.91 TB	15	12	online
VMDC31_VSG	0	348.98 GB	13.23 MB	349 GB	0	3	online
VMDC3_1_Reserved	21	3.77 TB	1 TB	4.77 TB	11	16	online

Name:	aggr1_br_gold_vmdc31
Status:	online
RAID Type:	raid_dp
Type:	Aggregate
Root:	No
Files:	96
Maximum Files:	31142
Checksum:	block
64-bit Aggregate:	No

Aggregate configured as RAID-level 6 or RAID_DP

Figure 2-26 shows the configuration of volumes from a previously configured Aggregate.

Figure 2-26 NetApp Volumes



Name	Aggregate	Status			Available Space	Total Space	Storage Efficiency
ciacfilera	aggr1_br_gold_vmdc31	online			32.3 GB	2.5 TB	Enabled
clsa	aggr1_br_gold_vmdc31	online			1.39 TB	2.15 TB	Enabled
clsa_mgmt2_vol	aggr1_br_gold_vmdc31	online			1.41 TB	2 TB	Enabled
cluster02_10s_vol	VMDC3_1_Reserved	online			1.95 TB	2 TB	Enabled
cluster02_1b2b_vol	aggr1_br_gold_vmdc31	online	Yes	5	1.9 TB	2 TB	Enabled
cluster02_1g_vol	aggr1_br_gold_vmdc31	online	Yes	5	1.9 TB	2 TB	Enabled
cluster02_1s_vol	VMDC3_1_Reserved	online	Yes	3	1.95 TB	2 TB	Enabled
cluster02_2g_vol	aggr1_br_gold_vmdc31	[illegible]	Yes	5	1.89 TB	2 TB	Enabled

General

Name:	cluster02_1s_vol	Autogrow Maximum Size:	2.4 TB
Status:	Online	Autogrow Incremental Size:	102.4 GB
Maximum Files:	31.88m	Snapshot Autodelete:	Enabled
Current Files:	106		
Language:	undefined 0 (undefined)		
Unicode:	Enabled		

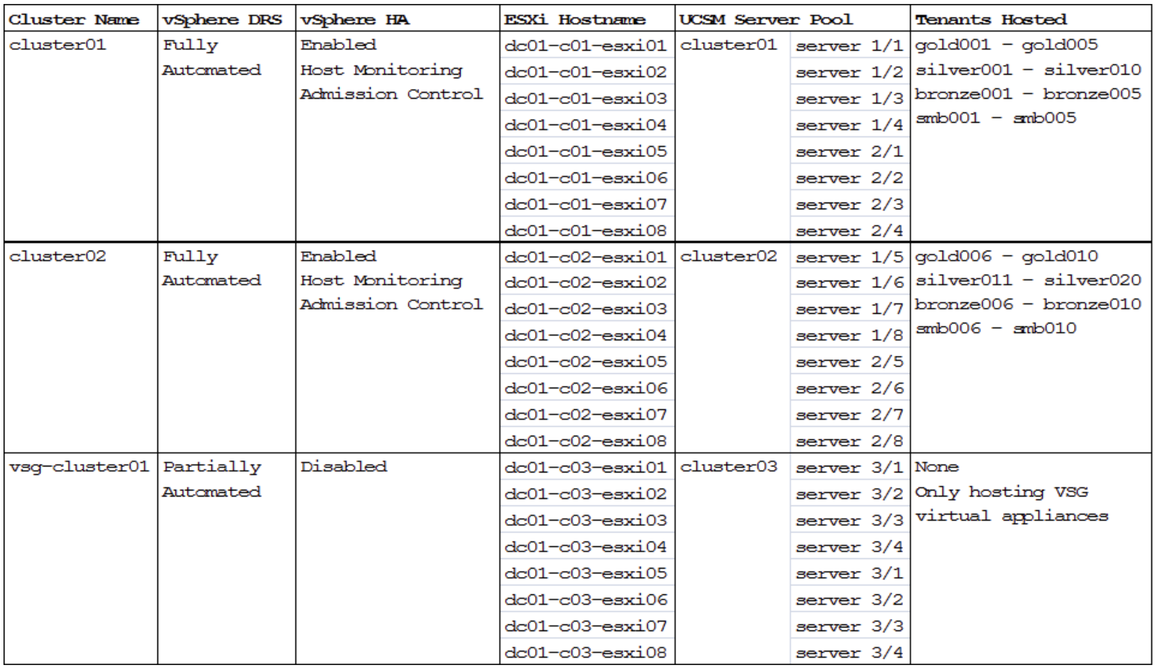
Aggregate from which volume is carved out

Enabled implies storage deduplication is turned ON

Hypervisor vSphere ESXi Implementation

The vSphere ESXi hypervisor is the workhorse for the compute infrastructure, providing the compute resources for hosting the VMs. Figure 2-27 shows the vSphere clusters, ESXi hosts, blade server assignments, and the tenants' distribution.

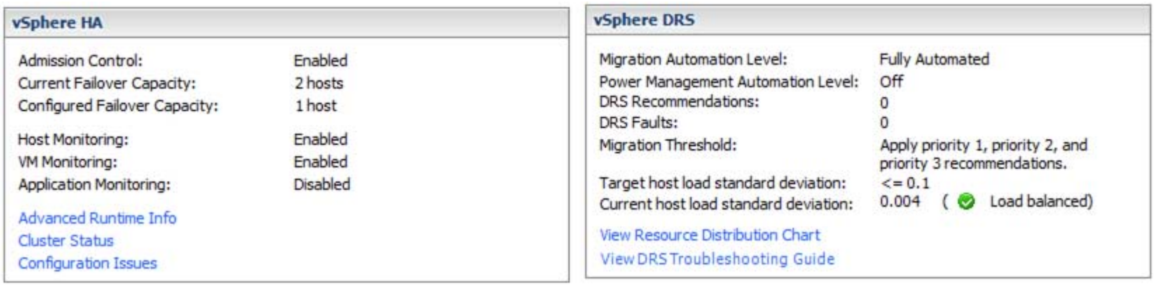
Figure 2-27 vSphere Clusters, ESXi Hosts, Blade Server Assignments, Tenants' Distribution



Cluster Name	vSphere DRS	vSphere HA	ESXi Hostname	UCSM Server Pool		Tenants Hosted
cluster01	Fully Automated	Enabled Host Monitoring Admission Control	dc01-c01-esxi01	cluster01	server 1/1	gold001 - gold005 silver001 - silver010 bronze001 - bronze005 smb001 - smb005
			dc01-c01-esxi02		server 1/2	
			dc01-c01-esxi03		server 1/3	
			dc01-c01-esxi04		server 1/4	
			dc01-c01-esxi05		server 2/1	
			dc01-c01-esxi06		server 2/2	
			dc01-c01-esxi07		server 2/3	
			dc01-c01-esxi08		server 2/4	
cluster02	Fully Automated	Enabled Host Monitoring Admission Control	dc01-c02-esxi01	cluster02	server 1/5	gold006 - gold010 silver011 - silver020 bronze006 - bronze010 smb006 - smb010
			dc01-c02-esxi02		server 1/6	
			dc01-c02-esxi03		server 1/7	
			dc01-c02-esxi04		server 1/8	
			dc01-c02-esxi05		server 2/5	
			dc01-c02-esxi06		server 2/6	
			dc01-c02-esxi07		server 2/7	
			dc01-c02-esxi08		server 2/8	
vsg-cluster01	Partially Automated	Disabled	dc01-c03-esxi01	cluster03	server 3/1	None Only hosting VSG virtual appliances
			dc01-c03-esxi02		server 3/2	
			dc01-c03-esxi03		server 3/3	
			dc01-c03-esxi04		server 3/4	
			dc01-c03-esxi05		server 3/1	
			dc01-c03-esxi06		server 3/2	
			dc01-c03-esxi07		server 3/3	
			dc01-c03-esxi08		server 3/4	

ESXi hosts in cluster01and cluster02 are used to host tenants' VMs. vSphere Distributed Resource Scheduling (DRS) is enabled to provide efficient load balancing of the computing workload across ESXi hosts in the cluster. vSphere HA is enabled to provide HA to the entire virtualized environment. Figure 2-28 shows the vSphere HA and DRS settings for cluster01 and cluster02.

Figure 2-28 vSphere HA and DRS for cluster01 and cluster02



vSphere HA

Admission Control: Enabled
Current Failover Capacity: 2 hosts
Configured Failover Capacity: 1 host

Host Monitoring: Enabled
VM Monitoring: Enabled
Application Monitoring: Disabled

Advanced Runtime Info
Cluster Status
Configuration Issues

vSphere DRS

Migration Automation Level: Fully Automated
Power Management Automation Level: Off
DRS Recommendations: 0
DRS Faults: 0
Migration Threshold: Apply priority 1, priority 2, and priority 3 recommendations.

Target host load standard deviation: <= 0.1
Current host load standard deviation: 0.004 (✓ Load balanced)

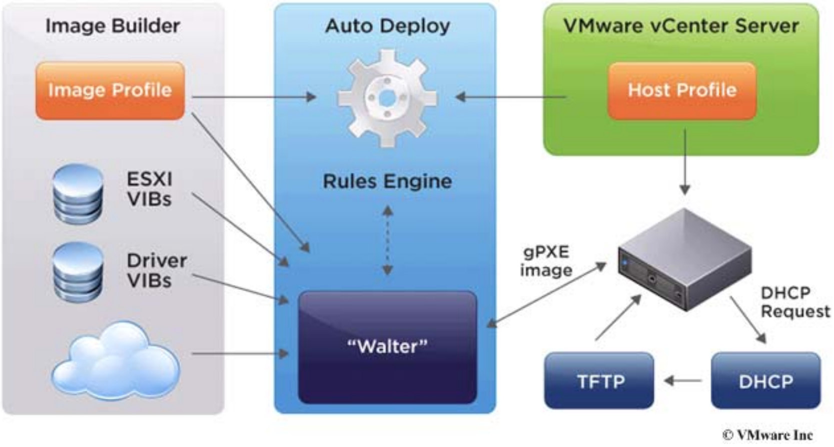
View Resource Distribution Chart
View DRS Troubleshooting Guide

ESXi hosts in the vsg-cluster01 are dedicated for hosting primary and secondary VSG virtual appliances. The VSG implements its own HA scheme, as such vSphere HA is not supported/required. The VSG does not support live vMotion, and vSphere DRS is set to partially automated for initial virtual appliance power on placement only.

vSphere Auto Deploy

Beginning with version 5.0, vSphere provides the Auto Deploy option for the deployment of ESXi hosts. Auto Deploy uses a PXE boot infrastructure in conjunction with vSphere host profiles and an image builder to provision and customize the ESXi host. No state is stored on the ESXi host itself, instead, the Auto Deploy server manages state information for each ESXi host. Figure 2-29 shows the Auto Deploy architecture.

Figure 2-29 vSphere Auto Deploy Architecture



Note vSphere Auto Deploy is used for this implementation for deploying ESXi software to the blade servers. Alternatively, the ESXi hosts can also be configured to boot from SAN. Installing ESXi software on a locally attached disk is not recommended, as this breaks the stateless computing capabilities of UCS.

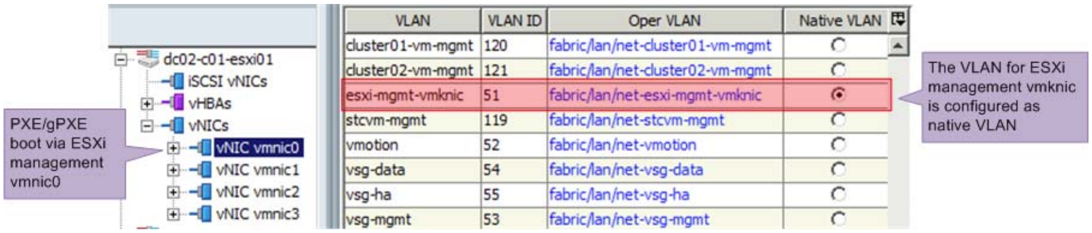
The following links provide more details about Auto Deploy and its installation and configuration:

- http://kb.vmware.com/kb/2005131
- http://blogs.vmware.com/vsphere/2012/01/understanding-the-auto-deploy-components.html
- http://pubs.vmware.com/vsphere-50/topic/com.vmware.vsphere.install.doc_50/GUIDCAB84194-3D8E-45F0-ABF9-0277710C8F98.html
- http://kb.vmware.com/kb/2000988
- http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/whitepaper_c11-701953.html

Take the following into consideration when setting up Auto Deploy:

- PXE support on the ESXi host network adapter is required.
- Auto Deploy makes use of PXE chainloading (http://etherboot.org/wiki/pxechaining) to load the gPXE bootloader to boot the ESXi host via HTTP (instead of TFTP; TFTP is used only to load the small gPXE bootloader). As such, upon booting, the ESXi host will request for IP information from the DHCP server twice with the same MAC address (first request from PXE, second request from gPXE). The DHCP server used must support such usage (most Linux DHCP servers do).
- On the DHCP server, configure static binding of the MAC address to the IP address, ensuring that the ESXi host always gets the same IP address.
- On newly deployed server hardware, make sure the BIOS clock is set correctly. If the BIOS clock is outdated by more than 60 days, the ESXi deployed on the server will not be able to join the vCenter.
- The PXE/gPXE bootloader does not support 802.1Q tagging of DHCP frames. Configure the VLAN where the ESXi management vmk interface resides as the native VLAN, as shown in Figure 2-30.

Figure 2-30 Configure Native VLAN for UCS Blade Server vNIC



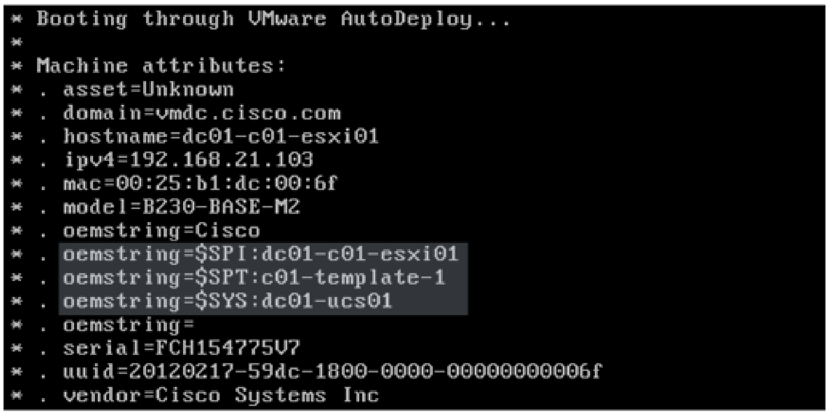
VLAN	VLAN ID	Oper VLAN	Native VLAN
cluster01-vm-mgmt	120	fabric/lan/net-cluster01-vm-mgmt	○
cluster02-vm-mgmt	121	fabric/lan/net-cluster02-vm-mgmt	○
esxi-mgmt-vmknic	51	fabric/lan/net-esxi-mgmt-vmknic	●
stcvm-mgmt	119	fabric/lan/net-stcvm-mgmt	○
vmotion	52	fabric/lan/net-vmotion	○
vsg-data	54	fabric/lan/net-vsg-data	○
vsg-ha	55	fabric/lan/net-vsg-ha	○
vsg-mgmt	53	fabric/lan/net-vsg-mgmt	○

PXE/gPXE boot via ESXi management vmnic0

The VLAN for ESXi management vmknic is configured as native VLAN

- Auto Deploy makes use of DNS. Configure both forward and reverse DNS resolution for the ESXi hostname on the DNS server. See http://blogs.vmware.com/vsphere/2012/11/auto-deployadding-host-to-vcenter-using-ip.html for more information.
- Include up-to-date drivers' VIBs for the hardware on the server (ENIC, FNIC, etc.) to the Auto Deploy image profile used. Make sure to include the correct version of the drivers in accordance with the UCSM and vSphere versions.
- If the ESXi hosts are part of the vSphere HA cluster, include the vmware-fdm VIB to the Auto Deploy image profile used. The vmware-fdm package can be retrieved from the software depot published by the vCenter server at the following URL: http://<VC-Address>/vSphere-HA-depot
- For the UCS blade server with the Cisco VIC adapter (Cisco UCS VIC 1280, Cisco UCS VIC 1240, Cisco UCS M81KR VIC, etc.), the ESXi host boot time will be much longer than those with other adapters. See http://tools.cisco.com/Support/BugToolKit/search/getBugDetails.do?method=fetchBugDetails&bugId=CSCtu17983 for more details.
- The standard ESXi software package provisions the ESXi host with VMware Tools binaries. If network boot time or memory and storage overhead is a concern, Auto Deploy can provision the ESXi host without VMware Tools binaries. Refer to http://kb.vmware.com/kb/2004018 for details.
- The UCS sets some OEM-specific strings in the SMBIOS to ease the configuration of Auto Deploy rules. The following oemstring is available (see Figure 2-31):
  - **$SPI—**Service-Profile Instance. The name of the service profile assigned to that specific blade server.
  - **$SPT—**Service-Profile Template. The name of the service profile template used to create that specific service profile.
  - **$SYS—**System. The name of the UCSM system that manages the blade server.

Figure 2-31 Cisco UCS oemstring in Auto Deploy



```
* Booting through VMware AutoDeploy...
*
* Machine attributes:
* . asset=Unknown
* . domain=vmdc.cisco.com
* . hostname=dc01-c01-esxi01
* . ipv4=192.168.21.103
* . mac=00:25:b1:dc:00:6f
* . model=B230-BASE-M2
* . oemstring=Cisco
* . oemstring=$SPI:dc01-c01-esxi01
* . oemstring=$SPT:c01-template-1
* . oemstring=$SYS:dc01-ucs01
* . oemstring=
* . serial=FCH154775V7
* . uuid=20120217-59dc-1800-0000-00000000006f
* . vendor=Cisco Systems Inc
```

- If the vCenter server, Auto Deploy server, DHCP server, or TFTP server are unavailable when the ESXi host boots up, the ESXi host will not be able to complete the boot and deployment process, rendering it unusable. Refer to https://blogs.vmware.com/techpubs/2012/03/highlyavailable-auto-deploy-infrastructure.html for recommendations on setting up highly available Auto Deploy infrastructure.
- The Auto Deploy server is an HTTP server at its core. Simultaneously booting large numbers of ESXi hosts places a significant load on the Auto Deploy server. VMware recommends using existing web server scaling technologies to help distribute the load. Refer to http://communities.vmware.com/groups/vsphere-autodeploy/blog/2012/04/20/scaling-out-autodeployusing-a-reverse-caching-proxy, which describes one way of scaling Auto Deploy with reverse caching proxy.
- In ESXi version 5.0, the ESXi Network Dump Collector feature is supported only with Standard vSwitches and cannot be used on a VMkernel network interface connected to a vSphere Distributed Switch or Cisco Nexus 1000V Switch. See http://kb.vmware.com/kb/2000781 for more details.
- vCenter creates an associated scheduled task to check host-profile compliance when a new host profile is created. The default properties (run frequency, start time, etc.) for the scheduled task might not be suitable, make changes as appropriate.
- If a host profile that is saved from a reference host with the local SCSI-3 device, applying the host profile to another ESXi host will cause compliance failure. See http://kb.vmware.com/ kb/2002488 for more details.
- A newly created host profile, or a host profile that has been updated from a reference host, would overwrite some manually entered configuration with defaults. Make sure to edit the host profile after it has been created or updated from the reference host. The following settings are known to be overwritten:
  - Networking configuration > Host virtual NIC > *name of vmknic* > Determine how the MAC address for vmknic should be decided.
  - Security configuration > Administrator password.
  - If the Enable/Disable Profile Configuration dialog box has been edited because of local SCSI-3 device described above, the changes will be overwritten as well.

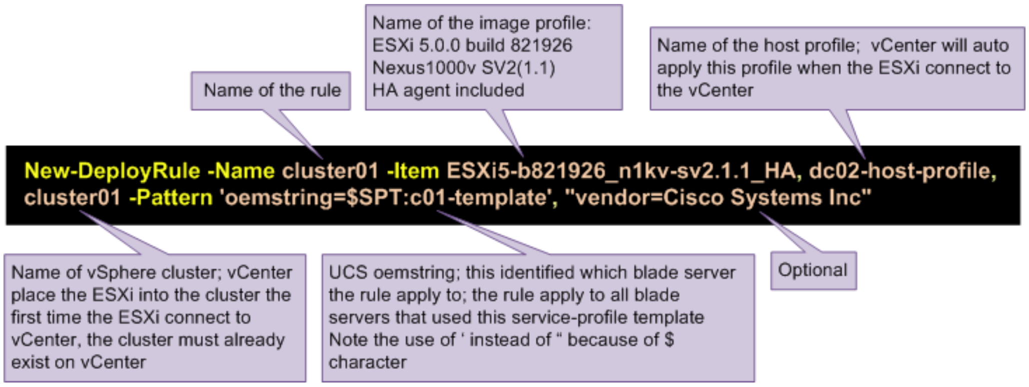
The following configuration shows the vSphere PowerCLI script used for Auto Deploy in this implementation:

```
# add ESXi packages
# download the offline image from http://www.vmware.com/patchmgr/download.portal
Add-EsxSoftwareDepot C:\temp\ESXi500-201209001.zip
# add HA package
Add-EsxSoftwareDepot http://192.168.13.14/vSphere-HA-depot/index.xml
# add Nexus1000v VEM package
# download the Nexus1000v image from VSM, http://<vsm-ip> Add-EsxSoftwareDepot
c:\temp\cisco-vem-v150-4.2.1.2.1.1.0-3.0.1.zip
# add enic driver for VIC adapter
Add-EsxSoftwareDepot c:\temp\enic_driver_2.1.2.22-offline_bundle-564611.zip
# add fnic driver for VIC adapter
Add-EsxSoftwareDepot c:\temp\fnic_driver_1.5.0.8-offline_bundle-758653.zip
# view the software depot Get-EsxSoftwareChannel
# remove all softweare depot
#Remove-EsxSoftwareDepot $DefaultSoftwareDepots
# view the image profile
Get-EsxImageProfile | select name
# view the available software packages Get-EsxSoftwarePackage
# clone a new image profile from existing profile, image profile with VMware Tools is
used
```

```
New-EsxImageProfile -CloneProfile ESXi-5.0.0-20120904001-standard -Name ESXi5-
b821926_n1kv-sv2.1.1_HA -Vendor vmdc
# add vmware HA package to the image profile
Add-EsxSoftwarePackage -ImageProfile ESXi5-b821926_n1kv-sv2.1.1_HA -SoftwarePackage
vmware-fdm
# add Nexus1000v VEM package to the image profile
Add-EsxSoftwarePackage -ImageProfile ESXi5-b821926_n1kv-sv2.1.1_HA -SoftwarePackage
cisco-vem-v150-esx
# add cisco enic driver package to the image profile
Add-EsxSoftwarePackage -ImageProfile ESXi5-b821926_n1kv-sv2.1.1_HA -SoftwarePackage
net-enic
# add cisco fnic driver package to the image profile
Add-EsxSoftwarePackage -ImageProfile ESXi5-b821926_n1kv-sv2.1.1_HA -SoftwarePackage
scsi-fnic
# create new deploy rule for ESXi host in cluster01
New-DeployRule -Name cluster01 -Item ESXi5-b821926_n1kv-sv2.1.1_HA, dc02-host-profile,
cluster01 -Pattern 'oemstring=$SPT:c01-template', "vendor=Cisco Systems Inc"
Add-DeployRule cluster01
# create new deploy rule for ESXi host in cluster02
New-DeployRule -Name cluster02 -Item ESXi5-b821926_n1kv-sv2.1.1_HA, dc02-host-profile,
cluster02 -Pattern 'oemstring=$SPT:c02-template', "vendor=Cisco Systems Inc"
Add-DeployRule cluster02
# create new deploy rule for ESXi host in cluster02
New-DeployRule -Name vsg-cluster01 -Item ESXi5-b821926_n1kv-sv2.1.1_HA,
dc02-hostprofile, vsg-cluster01 -Pattern 'oemstring=$SPT:c03-template', "vendor=Cisco
Systems Inc"
Add-DeployRule vsg-cluster01
```

Figure 2-32 shows the parameters of the vSphere PowerCLI **New-DeployRule** command.

Figure 2-32 vSphere PowerCLI New-DeployRule Command



```
New-DeployRule -Name cluster01 -Item ESXi5-b821926_n1kv-sv2.1.1_HA, dc02-host-profile,
cluster01 -Pattern 'oemstring=$SPT:c01-template', "vendor=Cisco Systems Inc"
```

Name of the rule

Name of the image profile:
ESXi 5.0.0 build 821926
Nexus1000v SV2(1.1)
HA agent included

Name of the host profile; vCenter will auto apply this profile when the ESXi connect to the vCenter

Name of vSphere cluster; vCenter place the ESXi into the cluster the first time the ESXi connect to vCenter, the cluster must already exist on vCenter

UCS oemstring; this identified which blade server the rule apply to; the rule apply to all blade servers that used this service-profile template Note the use of ' instead of " because of $ character

Optional

Nexus 1000V Series Switches

The Nexus 1000V Series Switches provide a comprehensive and extensible architectural platform for VM and cloud networking. In this implementation, all networking needs of the VMs are provided by Nexus 1000V Series Switches.

Nexus 1010 Virtual Services Appliance

The Nexus 1000V Virtual Supervisor Module (VSM) HA-pair is hosted on an HA-pair of the Nexus 1010 VPN Services Adapter (VSA). The Nexus 1010 is configured with network uplink topology type 3. The following configuration shows the relevant Nexus 1010 configuration:

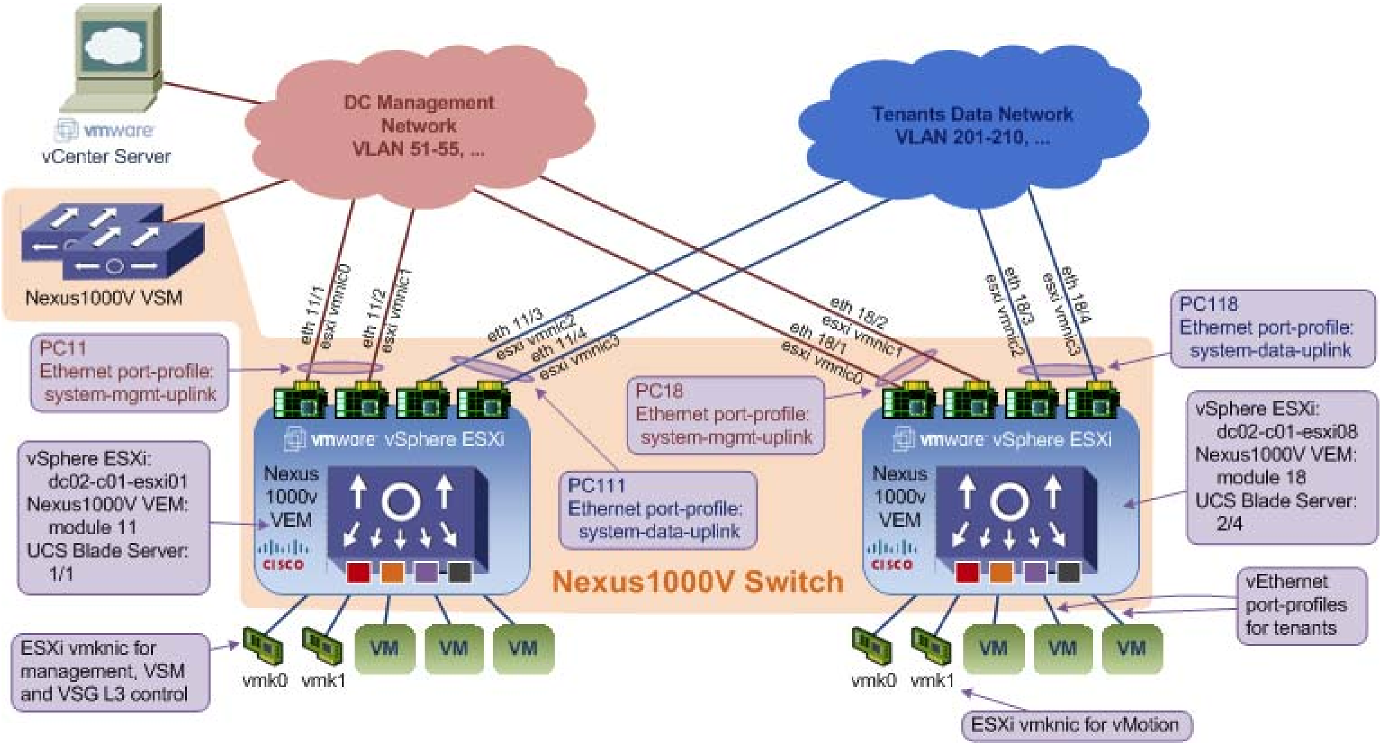
```
vlan 50
  name mgmt
network-uplink type 3
interface GigabitEthernet1
interface GigabitEthernet2
interface GigabitEthernet3
interface GigabitEthernet4
interface GigabitEthernet5
interface GigabitEthernet6
interface PortChannel1
interface PortChannel2
virtual-service-blade dc02-n1kv01
  virtual-service-blade-type name VSM-1.2
  description dc02-n1kv01
  interface control vlan 50
  interface packet vlan 50
  ramsize 2048
  disksize 3
  numcpu 1
  cookie 1098716425
  no shutdown primary
  no shutdown secondary
interface VsbEthernet1/1
interface VsbEthernet1/2
interface VsbEthernet1/3

svs-domain
  domain id 2381
  control vlan 50
  management vlan 50
  svs mode L2
```

The Nexus 1000V VSM is configured in L3 SVS mode. All three interfaces of the VSM are configured in the same VLAN (VLAN ID 50).

Nexus 1000V Distributed Virtual Switch

Figure 2-33 depicts the Nexus 1000V deployment for this implementation. As shown in the figure, the Nexus 1000V Distributed Virtual Switch (DVS) is made up of the VSM HA-pair and the VEM/ESXi that the VSM control. For simplicity, the UCS FIs, UCS chassis, the upstream Nexus 5000 (tenants' data network), and Catalyst 6500 Virtual Switch System (VSS) (management network) switches are not shown in the diagram; only the components relevant to Nexus 1000V configuration are shown. Only two of the 24 ESXi/VEMs are shown in the diagram, and the other ESXi/VEM has similar configuration.

Figure 2-33 Nexus 1000V Network Layout


vmware
vCenter Server
DC Management Network VLAN 51-55, ...
Tenants Data Network VLAN 201-210, ...
Nexus1000V VSM
PC11 Ethernet port-profile: system-mgmt-uplink
eth 11/1 esxi vmnic0
eth 11/2 esxi vmnic1
eth 11/3 esxi vmnic2
eth 11/4 esxi vmnic3
eth 18/2 esxi vmnic1
eth 18/1 esxi vmnic0
eth 18/3 esxi vmnic2
eth 18/4 esxi vmnic3
PC118 Ethernet port-profile: system-data-uplink
vSphere ESXi: dc02-c01-esxi01 Nexus1000V VEM: module 11 UCS Blade Server: 1/1
vmware vSphere ESXi
Nexus 1000v VEM
cisco
PC18 Ethernet port-profile: system-mgmt-uplink
PC111 Ethernet port-profile: system-data-uplink
Nexus1000V Switch
vmware vSphere ESXi
Nexus 1000v VEM
cisco
vSphere ESXi: dc02-c01-esxi08 Nexus1000V VEM: module 18 UCS Blade Server: 2/4
vEthernet port-profiles for tenants
ESXi vmknic for management, VSM and VSG L3 control
vmk0
vmk1
VM
VM
VM
vmk0
vmk1
VM
VM
VM
ESXi vmknic for vMotion

The Nexus 1000V VSM is configured in L3 SVS mode. In L3 SVS mode, VSM encapsulates the control and packet frames into User Datagram Protocol (UDP) packets. The VSM uses its mgmt0 interface to communicate with the VEMs. The VEMs are located in a different IP subnet from the VSM mgmt0 interface. On each VEM, the vmk0 vmkernel interface is used to communicate with the VSM. The following configuration shows the VSM svs-domain configuration:

```
svs-domain
  domain id 2288
  svs mode L3 interface mgmt0
```

Note Make sure that the SVS domain IDs for the Nexus 1010 VSA and Nexus 1000V VSM are unique. A domain ID is a parameter that is used to identify a VSM and VEM as related to one another.

The UCS is configured with disjoint upstream L2 networks; each ESXi/VEM host is configured with four NICs (also referred to as the ESXi VM Network Interface Card (VMNIC) or UCS vNIC), two

NICs for the management network (for UCS fabric A - fabric B redundancy), and two NICs for the tenants' data network (for UCS fabric A - fabric B redundancy). On the Nexus 1000V, two Ethernet uplink port profiles are configured. The following configuration shows the Ethernet port-profile configuration:

```
port-profile type ethernet system-mgmt-uplink
  vmware port-group
  port-binding static
  switchport mode trunk
  switchport trunk native vlan 51
  switchport trunk allowed vlan 51-55,119-121
  channel-group auto mode on mac-pinning
  no shutdown
  system vlan 51-55,119
  state enabled
```

```
port-profile type ethernet system-data-uplink
  vmware port-group
  port-binding static
  switchport mode trunk
  switchport trunk allowed vlan 201-210,301-310,401-410,501-520,601-620
  switchport trunk allowed vlan add 701-720,801-820,1601-1610,1801-1860
  switchport trunk allowed vlan add 1990,2001-2010
  channel-group auto mode on mac-pinning
  no shutdown
  state enabled
```

When the ESXi host is added to the Nexus 1000V DVS, the vmnic0 and vmnic1 interfaces are attached to the **system-mgmt-uplink** Ethernet uplink port profile, while the vmnic2 and vmnic3 interfaces are attached to the **system-data-uplink** Ethernet uplink port profile. For each VEM/ESXi added to the Nexus 1000V, the Nexus 1000V binds vmnic0 and vnmic1 into one MAC-pinning mode port-channel (to the management upstream network), while vmnic2 and vmnic3 are bound into another mac-pinning mode port-channel (to the tenants' data upstream network).

Note 1. The list of allowed VLANs configured on the two uplink Ethernet port profiles must not overlap. Defining two uplinks to carry the same VLAN is an unsupported configuration.

2. The allowed VLANs list on each of the Ethernet port profiles should match what has been configured on the UCSM vNIC.

In this implementation, the vmknic ESXi kernel interfaces (vmk0 and vmk1) are also managed by the Nexus 1000V. The following shows the configuration used for the ESXi management and vMotion vmkernel interfaces respectively:

```
port-profile type vethernet esxi-mgmt-vmknic
  capability l3control
  vmware port-group
  port-binding static
  switchport mode access
  switchport access vlan 51
  pinning id 0
  no shutdown
  capability l3-vn-service
  system vlan 51
  max-ports 64
  state enabled
port-profile type vethernet vmotion
  vmware port-group
  port-binding static
  switchport mode access
  switchport access vlan 52
  pinning id 0
  no shutdown
  system vlan 52
  max-ports 64
  state enabled
```

Note the following:

- The ESXi vmk0 interface is configured as the management interface and is attached to the **esximgmt-vmknic** vEthernet port profile.
- The ESXi vmk1 interface is configured as the vMotion interface and is attached to the **vmotion** vEthernet port profile

- Both port profiles are configured as the system port profile with the **system vlan** command. The VLANs in the vEthernet port profiles also have to be configured as the system VLAN in the Ethernet uplink port profile.
- The **esxi-mgmt-vmknic** port profile is configured with **capability l3control**, as the vmk0 interface is used for L3 control of the Nexus 1000V.
- The **esxi-mgmt-vmknic** port profile is also configured with **capability l3-vn-service**. L3 control of the VSG also uses the vmk0 interface.

Table 2-4 lists the port profiles configured for the tenants' VMs hosted on the compute infrastructure. For each tenant service class, only the port profiles for the first tenant are shown in the table, and the rest of the tenants for the same service class have similar configuration.

Table 2-4 Port Profiles Configured for the Tenant's VMs

Tenant	Port Profile	VLAN	Port Profile Configuration	Remark

Gold001 - Tier 1	gold001-v0201	201	`port-profile type vethernet gold-profile`
			`  switchport mode access`
			`  pinning id 2`
			`  no shutdown`
			`  state enabled`
			`port-profile type vethernet gold001-v0201`
			` vmware port-group`
			` inherit port-profile gold-profile`
			` switchport access vlan 201`
			` state enabled`
				The configuration shows the parent port profile for Gold tenants.
				All Gold tenants inherit the
				Gold parent port profile. Presentation/Web tier VMs in the Private Zone.

Gold001 - Tier 2	gold001-v0301	301	`port-profile type vethernet gold001-v0301`
			` vmware port-group`
			` inherit port-profile gold-profile`
			` switchport access vlan 301`
			` state enabled`
				Logic/ Application tier VMs in the Private Zone.

Gold001 - Tier 3	gold001-v0401	401	`port-profile type vethernet gold001-v0401`
			` vmware port-group`
			` inherit port-profile gold-profile`
			` switchport access vlan 401`
			` state enabled`
				Data/Database tier VMs in the Private Zone.

Gold001 - DMZ	gold001-v1601	1601	`port-profile type vethernet gold001-v1601`
			`  vmware port-group`
			`  inherit port-profile gold-profile`
			`  switchport access vlan 1601`
			`  state enabled`
				DMZ Zone VMs.

Table 2-4 Port Profiles Configured for the Tenant's VMs (continued)

Other Gold tenants	Tier 1 Tier 2 Tier 3 DMZ	202-299 302-399 402-499 1602-1699	. . .	...

Silver001 - Tier 1	silver001-v0501	501	`port-profile type vethernet silver-profile`
			` switchport mode access`
			` pinning id 3`
			` no shutdown`
			` state enabled`
			`port-profile type vethernet silver001-v0501`
			` vmware port-group`
			` inherit port-profile silver-profile`
			` switchport access vlan 501`
			` state enabled`
				The configuration shows the parent port profile for Silver tenants. All Silver tenants inherit the
				Silver parent port profile. Presentation/Web tier VMs.

Silver001 - Tier 2	silver001-v0601	601	`port-profile type vethernet silver001-v0601`
			` vmware port-group`
			` inherit port-profile silver-profile`
			` switchport access vlan 601`
			` state enabled`
				Logic/
				Application tier VMs.

Silver001 - Tier 3	silver001-v0701	701	`port-profile type vethernet silver001-v0701`
			` vmware port-group`
			` inherit port-profile silver-profile`
			` switchport access vlan 701`
			` state enabled`
				Data/Database tier VMs.

Other Silver tenants	Tier 1, 2, 3	502-599 602-699 702-799	. . .	...

Table 2-4 Port Profiles Configured for the Tenant's VMs (continued)

Bronze001	bronze001-v0801	801	`port-profile type vethernet bronze-profile`
			`switchport mode access`
			`service-policy input bronze`
			`pinning id 3`
			`no shutdown`
			`state enabled`
			`port-profile type vethernet bronze001-v0801`
			`vmware port-group`
			`inherit port-profile bronze-profile`
			`switchport access vlan 801`
			`state enabled`
				The configuration shows the parent port profile for Bronze tenants. All Bronze tenants inherit the Bronze parent port profile. All Bronze VMs in a single tier.

Other Bronze tenants	bronze	802-999	`...`	...

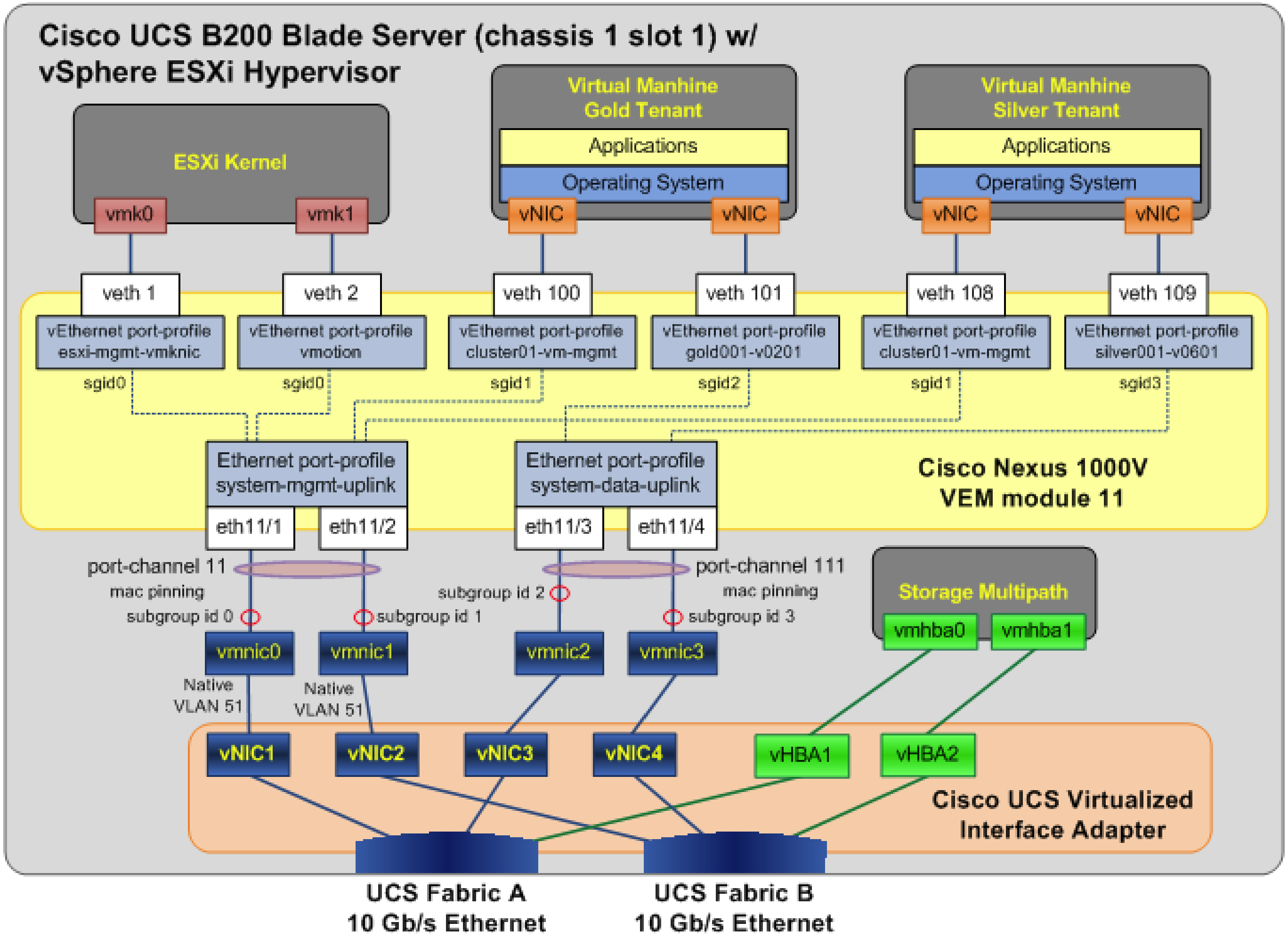
Copper/SMB001	smb001-v2001	2001	`port-profile type vethernet smb-profile`
			`switchport mode access`
			`pinning id 3`
			`no shutdown`
			`state enabled`
			`port-profile type vethernet smb001-v2001`
			`vmware port-group`
			`inherit port-profile smb-profile`
			`switchport access vlan 2001`
			`state enabled`
				The configuration shows the parent port profile for Copper/SMB tenants. All Copper/SMB tenants inherit the Copper/ SMB parent port profile. All Copper/SMB VMs in a single tier.

Other Copper/ SMB tenants	Copper/SMB	2002-2099	`...`	...

The configuration of the vEthernet port profile for tenants makes use of port-profile inheritance. Port-profile inheritance eases configuration and administration of setup with lots of port profiles. When properly deployed, inheritance enforces consist configuration across port profiles of similar nature. In this implementation, the tenants consist of Gold, Silver, Bronze, and SMB service classes. Tenants in the same service class have the same network requirements. Port-profile inheritance is used to ensure that each tenant in the same service class has the same network treatments. Tenants for Gold and Silver service classes are assigned with multiple port profiles, allowing their VMs to be placed in multiple VLANs.

Figure 2-34 shows the Nexus 1000V VEM and its related UCS configuration details for one of the blade servers. The diagram depicts the subgroup ID pinning configured for the various port profiles.

Figure 2-34 Nexus 1000V Configuration Details



2.5 Compute and Storage Best Practices and Caveats

UCS Best Practices

- When using UCSM configuration templates, be aware that some configuration changes will either cause server reboot or service disruption. Multiple templates of the same type should be used to prevent any single change to cause service disruption to all blade servers.
- When configuring server pools, select servers from multiple chassis to avoid single chassis failure bringing down all servers in the pool.
- Disable fabric failover for all vNICs configured for the blade servers, and let the Nexus 1000V manage the vNIC failure.
- UCSM does not support overlapping VLANs in disjoint L2 networks. Ensure that each VLAN only connects to one upstream disjoint L2 network.
- UCS FI uses LACP as the port-channel aggregation protocol. The opposing upstream switches must be configured with LACP active mode.

- A vNIC (VMNIC in the vSphere ESXi hypervisor or physical NIC in the bare metal server) can only communicate with one disjoint L2 network. If a server needs to communicate with multiple disjoint L2 networks, configure a vNIC for each of those networks.
- UCSM implicitly assigns default VLAN 1 to all uplink ports and port-channels. Do not configure any vNICs with default VLAN 1. It is advisable not to use VLAN 1 for carrying any user data traffic.

Storage Best Practices

- If using NetApp OnCommand System Manager 2.0 to configure storage filers, it is recommended to configure the following using the command line:
  - Configuring VIF and VLAN interfaces for NFS port-channel.
  - Configure security style (Unix or Windows) permissions when a volume is exported as NFS.
- To take advantage of Thin Provisioning, it is recommended to configure Thin Provisioning on both volumes/LUNs in storage and in VMFS.
- Configure Asymmetric Logical Unit Access (ALUA) on the filers for asymmetric logical unit access of LUNs.
- Enable storage deduplication on volumes to improve storage efficiency.
- Nexus 5000 is the storage switch in this design. It is mandatory to enable NPIV mode on the Nexus 5000, and also configure soft zoning (enables server mobility) that uses WWPNs.

vSphere ESXi Best Practices

- vSphere Auto Deploy makes use of PXE and gPXE. The PXE/gPXE bootloader does not support 802.1Q tagging of DHCP frames. Configure the VLAN where the ESXi management vmk interface resides as the native VLAN.
- vSphere Auto Deploy makes use of DNS. Configure both forward and reverse DNS resolution for the ESXi hostname on the DNS server.
- When using vSphere Auto Deploy, make sure that the vCenter server, Auto Deploy server, DHCP server, and TFTP server are made highly available.

vSphere ESXi Caveats

- For the UCS blade server with the Cisco VIC adapter (Cisco UCS VIC 1280, Cisco UCS VIC 1240, Cisco UCS M81KR VIC, etc.), the ESXi host boot time will be much longer than those with other adapters. See CSCtu17983 for more details.
- In ESXi version 5.0, the ESXi Network Dump Collector feature is supported only with Standard vSwitches and cannot be used on a VMkernel network interface connected to a vSphere Distributed Switch or Nexus 1000V Switch. See VMware Knowledge Base for more details.

Nexus 1000V Series Switches Best Practices

- Make sure that the SVS domain IDs for the Nexus 1010 VSA and the Nexus 1000V VSM are unique.
- Configure port profiles for management and vMotion vmknic as **system vlan**.
- Make use of port-profile inheritance to enforce consistent configuration and ease of management.

CHAPTER 3

Layer 2 Implementation

In the Virtualized Multiservice Data Center (VMDC) 2.3 solution, the goal is to minimize the use of Spanning Tree Protocol (STP) convergence and loop detection by the use of Virtual Port Channel (vPC) technology on the Nexus 7000. While STP is still running for protection as a backup, the logical topology is without loops and mostly edge ports, and the only non-edge ports are the ones to the Integrated Compute and Storage (ICS) switches (Nexus 5000), which are connected with back-to-back vPC. This is explained in more detail in the Layer 2 at Nexus 7004 Aggregation section.

The Nexus 7000 based DC-Aggregation switches form the heart of the Layer 2 (L2) network design and implement the L2/Layer 3 (L3) boundary. All services appliances and ICS stacks attach to the Aggregation layer using vPCs. Integrated compute and storage includes a switching layer that aggregates compute attachments and connects to the DC Aggregation layer. In VMDC 2.3, the ICS layer includes the Nexus 5500 series switches. Within the Compute layer, this solution uses Unified Computing System (UCS) 6248 Fabric Interconnects (FIs) and B-series blades, and there is a virtualized switching layer implemented with the Nexus 1000V. These aspects are covered in detail in the Compute and Storage Implementation chapter.

The L2 implementation details are split into the following major topics:

- Layer 2 at Integrated Compute and Storage, page 3-1
- Layer 2 Implementation at ICS Nexus 5500, page 3-2
- Layer 2 at Nexus 7004 Aggregation, page 3-4
- Connecting Service Appliances to Aggregation, page 3-9
- Port-Channel Load-Balancing, page 3-15
- Layer 2 Best Practices and Caveats, page 3-18

Layer 2 at Integrated Compute and Storage

This section presents the following topics:

- Nexus 1000V to Fabric Interconnect, page 3-2
- Fabric Interconnect to Nexus 5500, page 3-2

Nexus 1000V to Fabric Interconnect

The Nexus 1000V provides the virtualized switch for all of the tenant VMs. The Nexus 1000V is a virtualized L2 switch, and supports standard switch features, but is applied to virtual environments. Refer to Nexus 1000V Series Switches for more details.

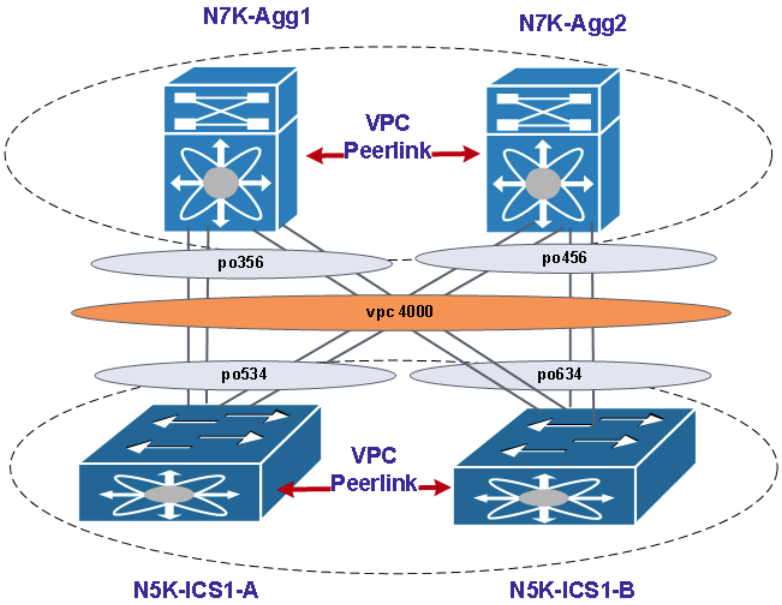
Fabric Interconnect to Nexus 5500

The pair of UCS 6248 FIs connect to the pair of Nexus 5500s using a vPC on the Nexus 5500 end. Refer to UCS Uplinks Configuration for more details.

Layer 2 Implementation at ICS Nexus 5500

The Nexus 7004 is used in the Aggregation layer and uses vPC technology to provide loop-free topologies. The Nexus 5548 is used in the Access layer and is connected to the Aggregation layer using back-to-back vPC. Figure 3-1 shows the entire vPC topology that is used in the Aggregation and Access layers.

Figure 3-1 vPC Topology in the Aggregation and Access Layers



The main difference between a vPC configuration and a non-vPC configuration is in the forwarding behavior of the vPC peer link and the Bridge Protocol Data Unit (BPDU) forwarding behavior of vPC member ports only.

A vPC deployment has two main spanning-tree modifications that matter:

- vPC imposes the rule that the peer link should never be blocking because this link carries important traffic such as the Cisco Fabric Services over Ethernet (CFSoE) protocol. The peer link is always forwarding.
- For vPC ports only, the operational primary switch generates and processes BPDUs. The operational secondary switch forwards BPDUs to the primary switch.

The advantages of Multiple Spanning Tree (MST) over Rapid Per-VLAN Spanning Tree Plus (PVST +) are as follows:

- MST is an IEEE standard.
- MST is more resource efficient. In particular, the number of BPDUs transmitted by MST does not depend on the number of VLANs, as Rapid PVST+ does.
- MST decouples the creation of VLANs from the definition for forwarding the topology.
- MST simplifies the deployment of stretched L2 networks, because of its ability to define regions.

For all these reasons, it is advisable for many vPC deployments to migrate to an MST-based topology.

Rapid PVST+ offers slightly better flexibility for load balancing VLANs on a typically V-shape spanning-tree topology. With the adoption of vPC, this benefit is marginal because topologies are becoming intrinsically loop free, at which point the use of per-VLAN load balancing compared to per-instance load balancing is irrelevant (with vPC, all links are forwarding in any case).

In our implementation, we have used two instances in MST. MST0 is reserved and is used by the system for BPDU processing. Within each MST region, MST maintains multiple spanning-tree instances. Instance 0 is a special instance for a region, known as the IST. The IST is the only spanning-tree instance that sends and receives BPDUs. MST 1 has all of the VLAN instances (1-4094) mapped to it. Since the per-VLAN benefits are marginal compared to per-instance load balancing, we prefer to use a single MST instance (MST1).

The following configuration details the MST and vPC configuration used on the Nexus 7004 (Aggregation) switch:

```
spanning-tree mode mst
spanning-tree mst 0-1 priority 0
spanning-tree mst configuration
  name dc2
  instance 1 vlan 1-4094

interface port-channel456
  description PC-to-N5K-VPC
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1-1120,1601-1610,1801-1860,2001-2250
  switchport trunk allowed vlan add 3001-3250
  spanning-tree port type network
  service-policy type qos input ingress-qos-policy
  service-policy type queuing output vmdc23-8e-4q4q-out
  vpc 4000
```

The following details the MST and vPC configuration used on the Nexus 5548 (ICS) switch:

```
interface port-channel534
  description vPC to N7K-Aggs
  switchport mode trunk
  spanning-tree port type network
  speed 10000
  vpc 4000

spanning-tree mode mst
spanning-tree mst 0 priority 4096
spanning-tree mst configuration
  name dc2
  instance 1 vlan 1-4094
```

The salient features of the connection between ICS and aggregation are as follows:

- Pre-provision all VLAN instances on MST and then create them later as needed.

- The operational secondary switch cannot process BPDUs and it forwards them to the operational primary when they are received.
- Unlike SPT, vPC can have two root ports. The port on the secondary root that connects to primary root (vPC peer link) is a root port.
- Type-1 inconsistencies must be resolved for a vPC to be formed. Associate the root and secondary root role at the Aggregation layer. It is preferred to match the vPC primary and secondary roles with the root and secondary root.
- You do not need to use more than one instance for vPC VLANs.
- Make sure to configure regions during the deployment phase.
- If you make changes to the VLAN-to-instance mapping when vPC is already configured, remember to make changes on both the primary and secondary vPC peers to avoid a Type-1 global inconsistency.
- Use the **dual-active exclude interface-vlan** command to avoid isolating non-vPC VLAN traffic when the peer link is lost.
- From a scalability perspective, it is recommended to use MST instead of Rapid PVST+.

For more details on STP guidelines for Cisco NX-OS Software and vPC, refer to Chapter 4: Spanning Tree Design Guidelines for Cisco NX-OS Software and Virtual Port-channels. This document explains the best practices and presents the argument for using MST versus Rapid PVST+ as STP.

Layer 2 at Nexus 7004 Aggregation

The Aggregation layer, which is sometimes referred to as the Distribution layer, aggregates connections and traffic flows from multiple Access layer devices to provide connectivity to the MPLS PE routers. In this solution, a pair of Nexus 7004 switches are used as the Aggregation layer devices. The Nexus 7004 is a four-slot switch. In a compact form factor, this switch has the same NX-OS operational features of other Cisco Nexus 7000 Series Switches. The Nexus 7004 offers high availability, high performance, and great scalability. This switch has two dedicated supervisor slots and two I/O module slots. This switch supports Sup2 and Sup2E, and it does not require fabric modules. This switch is only 7 Rack Units (RU) and is designed with side-to-rear airflow.

vPC and Spanning Tree

In this solution, a back-to-back vPC is used between the Nexus 7004 Aggregation layer and ICS Nexus 5500 series switches. The logical view of vPC is that two switches look like one to the other side, and hence both sides see the other as one switch and one port-channel of 8 links connecting to it. This eliminates any loops, and the vPC rules prevent any packet from being looped back. STP is still run in the background to prevent any accidental vPC failure or for non-vPC ports connected together. The Nexus 7004 STP bridge priority is kept higher to elect the Nexus 7004 pair as the root bridge and have all the ICS switches as non-root bridges. All services appliances are connected to the Nexus 7004 using vPC as well. These are connected as edge ports.

To prevent any L2 spanning-tree domain connection with the management L2 domain, the connection from the management network is directly to the UCS and uses disjoint VLANs on the UCS to connect only the management VLANs on these ports facing the management network. One pair of ports is connected to the Nexus 7004 Aggregation layer switches to transport the Application Control Engine (ACE) management VLANs back to the management network. These are connected as a vPC, but also as an access switchport, and hence are edge ports. See ACE 4710 to Nexus 7004 for detailed information about ACE management.

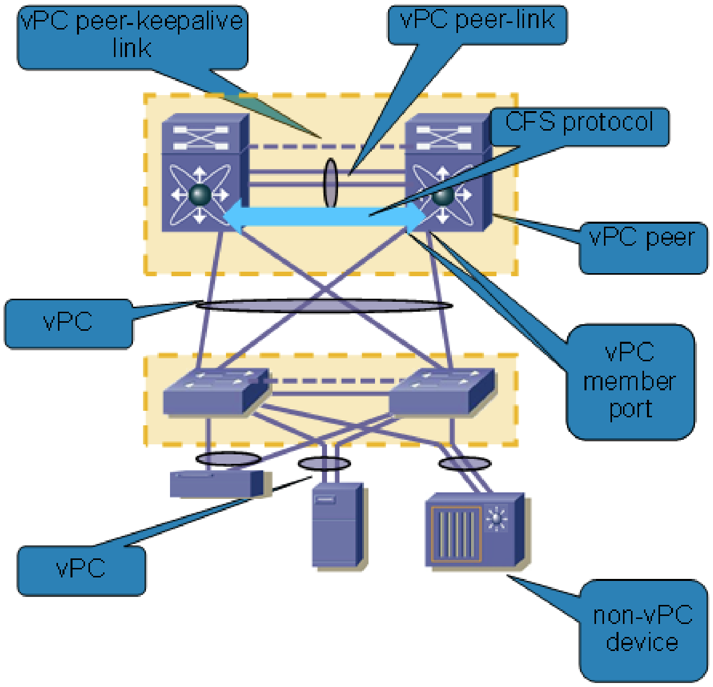
The details of the vPC configuration are discussed below.

See VPC Best Practices Design Guide for additional information.

vPC

A vPC is a logical entity formed by L2 port-channels distributed across two physical switches to the far-end attached device (Figure 3-2).

Figure 3-2 vPC Terminology



The following components make up a vPC:

- **vPC peer.** A vPC switch, one of a pair.
- **vPC member port.** One of a set of ports (port-channels) that form a vPC.
- **vPC.** The combined port-channel between the vPC peers and the downstream device.
- **vPC peer-link.** The link used to synchronize the state between vPC peer devices; must be 10GbE.
- **vPC peer-keepalive link.** The keepalive link between vPC peer devices, i.e., backup to the vPC peer-link.

Refer to Configuring vPCs for a detailed vPC configuration guide. Below is the vPC configuration on the Nexus 7000 switch.

```
feature vpc

vpc domain 998
  peer-switch
  role priority 30000
  peer-keepalive destination 192.168.50.21
  delay restore 120
  peer-gateway
  auto-recovery
  delay restore interface-vlan 100
  ip arp synchronize
```

```
interface port-channel34
  vpc peer-link                                <=========================vPC peer link

interface port-channel35
  vpc 35                                       <=========================vPC link 35
port-channel 35 for ASA

interface port-channel111
  vpc 111                                     <=========================vPC link 111
port-channel 111 for ACE mgmt

interface port-channel356
  vpc 4000                                    <=========================vPC link 4000
port-channel 356 to the N5K
```

Below are useful commands for configuring vPCs.

```
show vpc brief
show vpc role
show vpc peer-keepalive
show vpc statistics
show vpc consistency-parameters
```

Note 1. **vPC peer-keepalive link implementation.** The peer-keepalive link between the vPC peers is used to transmit periodic, configurable keepalive messages. L3 connectivity between the peer devices is needed to transmit these messages. In this solution, management VRF and management ports are used.

2. **peer switch.** See the Spanning Tree Protocol Interoperability with vPC section below.

3. **delay-restore.** This feature will delay the vPC coming back up until after the peer adjacency forms and the VLAN interfaces are back up. This feature avoids packet drops when the routing tables may not be converged before the vPC is once again passing traffic.

4. **arp sync.** This feature addresses table synchronization across vPC peers using the reliable transport mechanism of the CFSoE protocol. Enabling IP Address Resolution Protocol (ARP) synchronize can get faster convergence of address tables between the vPC peers. This convergence is designed to overcome the delay involved in ARP table restoration for IPv4 when the peer link port-channel flaps or when a vPC peer comes back online.

5. **auto-recovery.** This feature enables the Nexus 7000 Series device to restore vPC services when its peer fails to come online by using the **auto-recovery** command. On reload, if the peer link is down and three consecutive peer-keepalive messages are lost, the secondary device assumes the primary STP role and the primary LACP role. The software reinitializes the vPCs, bringing up its local ports. Because there are no peers, the consistency check is bypassed for the local vPC ports. The device elects itself to be the STP primary regardless of its role priority, and also acts as the master for LACP port roles.

6. **peer gateway.** This feature enables vPC peer devices to act as the gateway for packets that are destined to the vPC peer device's MAC address.

7. **peer link redundancy.** For the peer link, it is better to use two or more links from different line cards to provide redundancy.

8. **role priority.** There are two defined vPC roles, primary and secondary. The vPC role defines which of the two vPC peer devices processes BPDUs and responds to ARP.

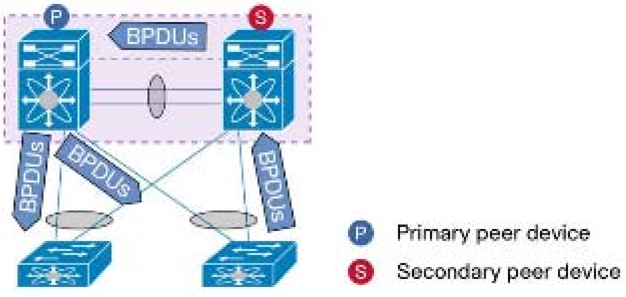
Spanning Tree Protocol Interoperability with vPC

The vPC maintains dual-active control planes, and STP still runs on both switches.

For vPC ports, only the vPC primary switch runs the STP topology for those vPC ports. In other words, STP for vPCs is controlled by the vPC primary peer device, and only this device generates then sends out BPDUs on STP designated ports. This happens irrespectively of where the designated STP root is located. STP on the secondary vPC switch must be enabled, but it does not dictate the vPC member port state. The vPC secondary peer device proxies any received STP BPDU messages from access switches toward the primary vPC peer device.

Both vPC member ports on both peer devices always share the same STP port state (FWD state in a steady network). Port-state changes are communicated to the secondary via Cisco Fabric Service (CFS) messages through peer link. Peer link should never be blocked. As the vPC domain is usually the STP root for all VLANs in the domain, the rootID value is equal to the bridgeID of the primary peer device or secondary peer device. Configuring aggregation on vPC peer devices as the STP root primary and STP root secondary is recommenced. It is also recommended to configure the STP root on the vPC primary device and configure the STP secondary root on the vPC secondary device (Figure 3-3).

Figure 3-3 vPC and STP BPDUs

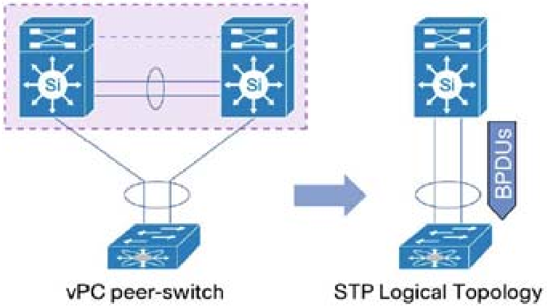


Peer-switch Feature

The vPC peer-switch feature address performance concerns around STP convergence. This feature allows a pair of vPC peer devices to appear as a single STP root in the L2 topology (they have the same bridge ID). This feature eliminates the need to pin the STP root to the vPC primary switch and improves vPC convergence if the vPC primary switch fails. When the vPC peer switch is activated, it is mandatory that both peer devices have the exact same spanning tree configuration, and more precisely, the same STP priority for all vPC VLANs.

To avoid loops, the vPC peer link is excluded from the STP computation. In vPC peer switch mode, STP BPDUs are sent from both vPC peer devices to avoid issues related to STP BPDU timeout on the downstream switches, which can cause traffic disruption. This feature can be used with the pure-peer switch topology, in which the devices all belong to the vPC (Figure 3-4).

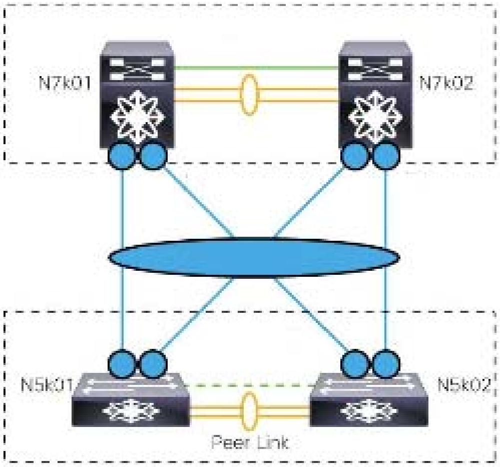
Figure 3-4 Peer-switch



Spanning Tree Implementation in this Solution

The ASA and ACE do not support spanning tree and configuring the edge trunk port on the Nexus 7000. A pair of Nexus 5000s in vPC mode connects to the pair of Nexus devices in vPC mode in this solution, and this is often referred to as "double-sided vPC" (Figure 3-5).

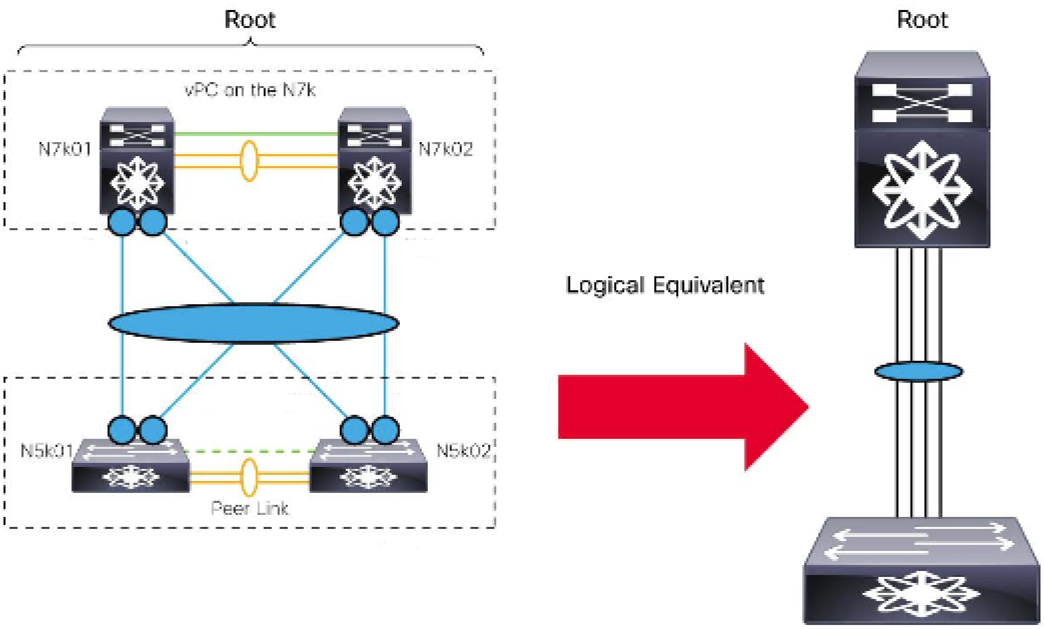
Figure 3-5 Double-sided vPC



Double-sided vPC simplifies the spanning-tree design, provides a higher resilient architecture, and provides more bandwidth from the Access to Aggregation layer as no ports are blocked.

With the peer-switch feature, the Nexus 7000 switches are placed as the root of the spanning tree in this solution. Figure 3-6 shows the spanning-tree view of the double-sided vPC.

Figure 3-6 Spanning-Tree View of Double-sided vPC



See Layer 2 Implementation at ICS Nexus 5500 for more information about Nexus 7000 to Nexus 5000 connections.

Connecting Service Appliances to Aggregation

In this solution, the Application Control Engine (ACE) 4710 is used to provide the load-balancing service, and the Adaptive Security Appliance (ASA) is used to provide firewall and VPN services.

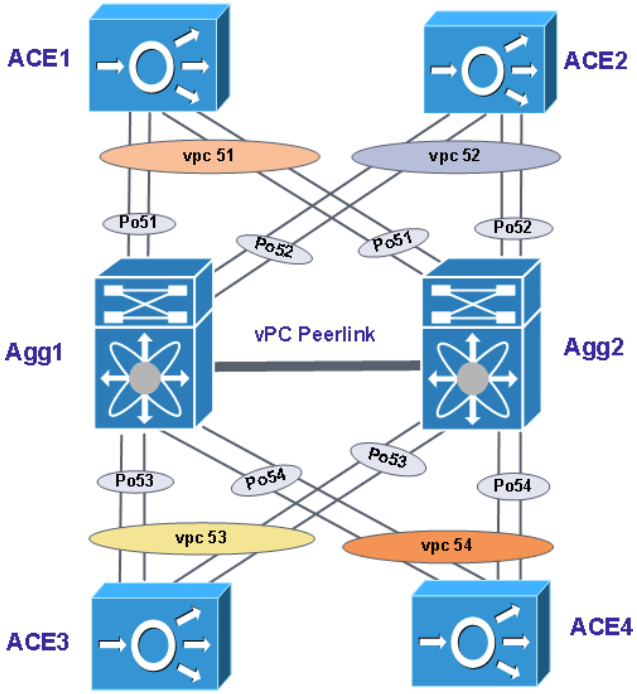
This section presents the following topics:

- ACE 4710 to Nexus 7004, page 3-9
- ASA 5500/5800 to Nexus 7004, page 3-13

ACE 4710 to Nexus 7004

The ACE 4710 has four Gigabit Ethernet interfaces. In this solution, a pair of ACE 4710 devices is used to form active/active redundancy. A vPC is used to connect a pair of ACEs to the Nexus 7004 devices. Figure 3-7 shows the physical connection.

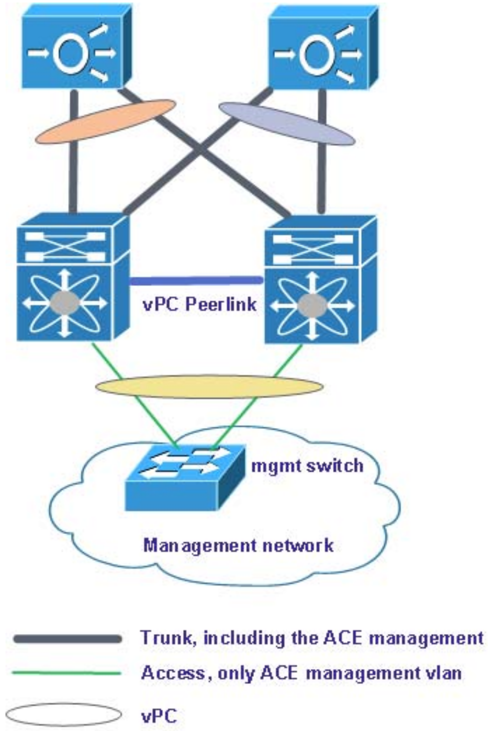
Figure 3-7 Physical Connection



In this diagram, ACE1 and ACE2 are the redundancy pair for the Gold tenants, and ACE3 and ACE4 are the redundancy pair for the Silver tenants. The ACE is not running spanning tree, and there is no loop in this topology, so in the Nexus 7000, the vPC to the ACE is configured as an edge trunk. One pair of ports is connected to the Nexus 7004 Aggregation layer switches to transport the ACE management VLANs back to the management network. These are connected as a vPC, but also as an access switchport, and hence are edge ports. The management switch uses an L3 interface for this ACE management connection to prevent possible spanning-tree loops in the management network.

In this implementation, we are using all four GigabitEthernet ports on the ACE 4710 in a port-channel to connect to the Nexus 7004 aggregation nodes in the DC. This is to enable the full capacity of the ACE 4710 to be available for customer traffic, however, this requires the Fault-Tolerant (FT) VLAN and Management VLAN to also be trunked over this port-channel. For management connectivity, particular attention has to be given and steps taken to avoid merging two different L2 domains and spanning trees. Alternative options are to dedicate one physical interface for Management access, one physical interface for FT traffic, tie them back-to-back between the ACE 4710 pair, and use the other two interfaces available on the ACE 4710 for data traffic, which provides for 2 Gbps of inbound and 2 Gbps of outbound traffic (Figure 3-8). The throughput limit for the ACE 4710 is 4 Gbps. Refer to the Cisco ACE 4710 Application Control Engine Data Sheet for more details.

Figure 3-8 ACE Management Connection


vPC Peerlink
mgmt switch
Management network
Trunk, including the ACE management
Access, only ACE management vlan
vPC

Below is the related configuration on the AGG1 device.

```
interface port-channel51
  description connection to ACE1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 60,201-210,301-310,401-410,1601-1610
  switchport trunk allowed vlan add 1998
  spanning-tree port type edge trunk
  vpc 51

interface port-channel52
  description connection to ACE2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 60,201-210,301-310,401-410,1601-1610
  switchport trunk allowed vlan add 1998
  spanning-tree port type edge trunk
  vpc 52

interface port-channel53
  description connection to ACE3
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 60,501-520,601-620,701-720,1998
  spanning-tree port type edge trunk
  vpc 53

interface port-channel54
  description connection to ACE4
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 60,501-520,601-620,701-720,1998
  spanning-tree port type edge trunk
  vpc 54

interface port-channel111
  description this is for ACE mgmt
  switchport
  switchport access vlan 60
  spanning-tree port type edge
  speed 1000
  vpc 111
```

Below is the related configuration on the ACE1 device.

```
interface port-channel 1
  ft-port vlan 1998
  switchport trunk allowed vlan
60,201-210,301-310,401-410,501-520,601-620,701-720,1601-1610
  port-channel load-balance src-dst-port
  no shutdown

interface gigabitEthernet 1/1
  speed 1000M
  duplex FULL
  qos trust cos
  channel-group 1
  no shutdown
interface gigabitEthernet 1/2
  speed 1000M
  duplex FULL
  qos trust cos
  channel-group 1
  no shutdown
interface gigabitEthernet 1/3
  speed 1000M
  duplex FULL
  qos trust cos
  channel-group 1
  no shutdown
interface gigabitEthernet 1/4
  speed 1000M
  duplex FULL
  qos trust cos
  channel-group 1
  no shutdown
```

Below are useful commands for the Nexus 7000 and the ACE.

Nexus 7000

```
show vpc
show port-channel summary
```

ACE

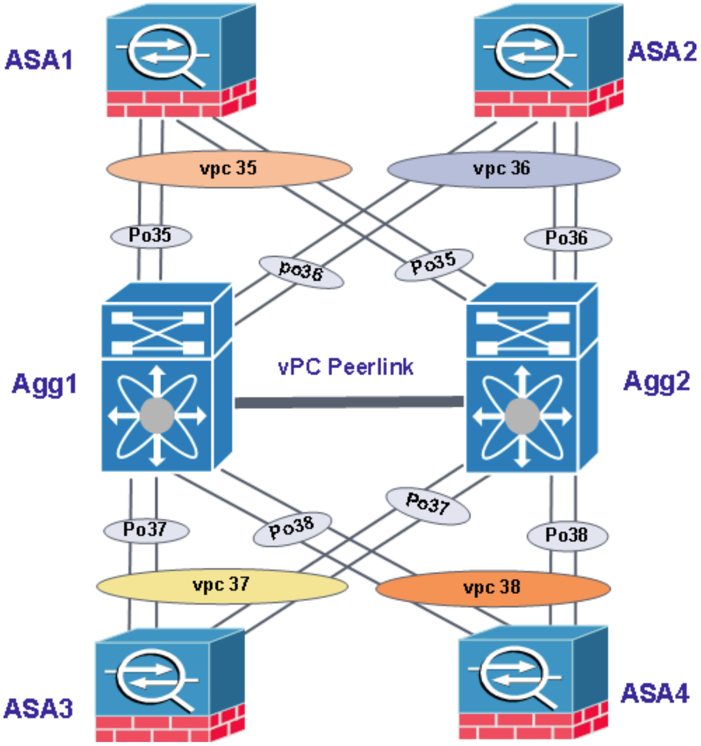
```
show interface
show interface port-channel
```

ASA 5500/5800 to Nexus 7004

The ASA 5585 is used as the per-tenant firewall, and the ASA 5555 is used as the VPN server for the IPsec and Cisco AnyConnect clients. In this solution, active/active is used to provide redundancy for the firewall purpose, and active/standby is used to provide the VPN server redundancy. To connect to the ASA in the pair, separate port-channels are created in each Nexus 7000. vPC links are used in the Nexus 7000 to connect to the ASA devices. From the view of the ASA, it is using one single port-channel to connect to the Nexus 7000 routers.

Figure 3-9 shows the physical connection.

Figure 3-9 Physical Connection



In this diagram, ASA1 and ASA2 are the redundancy pair for the firewall service, and ASA3 and ASA4 are the redundancy pair for the VPN service. The ASA is not running spanning tree, and there is no loop in this topology, so in the Nexus 7000, the vPC to the ASA is configured as an edge trunk.

Below is the related configuration on the Nexus 7000 Agg1 device.

```
interface port-channel35
  description PC-to-FW1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1201-1210,1301-1310,1401-1410,1501-1510
  switchport trunk allowed vlan add 1701-1710,2000,3001-4000
  spanning-tree port type edge trunk
  no lacp graceful-convergence
  vpc 35

interface port-channel36
  description PC-to-FW2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1201-1210,1301-1310,1401-1410,1501-1510
  switchport trunk allowed vlan add 1701-1710,2000,3001-4000
  spanning-tree port type edge trunk
  no lacp graceful-convergence
  vpc 36

interface port-channel37
  description PC-to-VPN1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1701-1710,2000
  spanning-tree port type edge trunk
  no lacp graceful-convergence
  vpc 37

interface port-channel38
  description PC-to-VPN2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1701-1710,2000
  spanning-tree port type edge trunk
  no lacp graceful-convergence
  vpc 38
```

Note By default, LACP graceful convergence is enabled. In this solution, we disable it to support LACP interoperability with devices where the graceful failover defaults may delay the time taken for a disabled port to be brought down or cause traffic from the peer to be lost.

Below is the related configuration on the ASA1 device.

```
interface TenGigabitEthernet0/6
 channel-group 1 mode active

!
interface TenGigabitEthernet0/7
 channel-group 1 mode active

!
interface TenGigabitEthernet0/8
 channel-group 1 mode active

!
interface TenGigabitEthernet0/9
 channel-group 1 mode active


interface Port-channel1
 port-channel load-balance vlan-src-dst-ip-port
```

Below are useful commands for the Nexus 7000 and ASA.

Nexus 7000

```
show port-channel summary show vpc
```

ASA

```
show interface port-channel
```

Port-Channel Load-Balancing

The port-channel (EtherChannel) is a port-link-aggregation technology. It allows grouping of several physical Ethernet links to create one logical Ethernet link for the purpose of providing FT and load-balancing links between switches, routers, and other devices. To load balance the traffic, hash schemes are used to select a port member of a bundle that is used for forwarding, and usually they make this decision based on fixed field values of either L2, L3, or Layer 4 (L4) headers, or Boolean operation on fixed field values on two or three protocol headers. To determine which fields to use, traffic analysis should be done to determine the best hash scheme. Load-balancing options differ across different platforms. The following sections discuss Nexus 7000, Nexus 5000, ASA, ACE, and FI port-channel load-balancing techniques.

Nexus 7000 Load-balancing Optimization

The NX-OS software load balances traffic across all operational interfaces in a port-channel by hashing the addresses in the frame to a numerical value that selects one of the links in the channel. The fields that can be used to hash are MAC addresses, IP addresses, or L4 port numbers. It can use either source or destination addresses or ports or both source and destination addresses or ports. Load-balancing mode can be configured to apply to all port-channels that are configured on the entire device or on specified modules. The per-module configuration takes precedence over the load-balancing configuration for the entire device. The default load-balancing method for L3 interfaces is source and destination IP address. The default load-balancing method for L2 interfaces is source and destination MAC address.

In this solution, the Nexus 7000 is configured as follows:

```
port-channel load-balance src-dst ip-l4port-vlan
```

Below are useful commands for the Nexus 7000.

```
dc02-n7k-agg1# sh port-chan load-balance
 System config:
  Non-IP: src-dst mac
  IP: src-dst ip-l4port-vlan rotate 0
Port Channel Load-Balancing Configuration for all modules:
Module 3:
  Non-IP: src-dst mac
  IP: src-dst ip-l4port-vlan rotate 0
Module 4:
  Non-IP: src-dst mac
  IP: src-dst ip-l4port-vlan rotate 0
```

Nexus 5000 Load-balancing Optimization

In the Nexus 5000 switches, NX-OS load balances traffic across all operational interfaces in a port-channel by reducing part of the binary pattern formed from the addresses in the frame to a numerical value that selects one of the links in the channel. Port-channels provide load balancing by default.

The basic configuration uses the following criteria to select the link:

- For an L2 frame, it uses the source and destination MAC addresses.
- For an L3 frame, it uses the source and destination MAC addresses and the source and destination IP addresses.
- For an L4 frame, it uses the source and destination MAC addresses and the source and destination IP addresses.

The switch can be configured to use one of the following methods to load balance across the port-channel:

- Destination MAC address

- Source MAC address
- Source and destination MAC address
- Destination IP address
- Source IP address
- Source and destination IP address
- Destination TCP/UDP port number
- Source TCP/UDP port number
- Source and destination TCP/UDP port number

Traffic analysis needed to be carried out to determine which fields to use. In this solution, the Nexus 5000 is configured as follows:

```
port-channel load-balance ethernet source-dest-port
```

Below are useful commands for the Nexus 5000.

```
dc02-n5k-ics1-A# sh port-channel load-balance

Port Channel Load-Balancing Configuration:
System: source-dest-port

Port Channel Load-Balancing Addresses Used Per-Protocol:
Non-IP: source-dest-mac
IP: source-dest-port source-dest-ip source-dest-mac
```

ASA Load-balancing Optimization

In the ASA, an 802.3ad EtherChannel is a logical interface (called a port-channel interface) consisting of a bundle of individual Ethernet links (a channel group) to increase the bandwidth for a single network. A port-channel interface is used in the same way as a physical interface when interface-related features are configured. The EtherChannel aggregates the traffic across all available active interfaces in the channel. The port is selected using a proprietary hash algorithm, based on source or destination MAC addresses, IP addresses, TCP and UDP port numbers, and VLAN numbers.

In the ASA, load balancing is configured in the port-channel interface, not in the global device. The default load-balancing method is the source and destination IP address. The following methods can be configured for load balancing:

```
dc02-asa-fw1(config-if)# port-channel load-balance ?

interface mode commands/options:
  dst-ip                Dst IP Addr
  dst-ip-port           Dst IP Addr and TCP/UDP Port
  dst-mac               Dst Mac Addr
  dst-port              Dst TCP/UDP Port
  src-dst-ip            Src XOR Dst IP Addr
  src-dst-ip-port       Src XOR Dst IP Addr and TCP/UDP Port
  src-dst-mac           Src XOR Dst Mac Addr
  src-dst-port          Src XOR Dst TCP/UDP Port
  src-ip                Src IP Addr
  src-ip-port           Src IP Addr and TCP/UDP Port
  src-mac               Src Mac Addr
  src-port              Src TCP/UDP Port
  vlan-dst-ip           Vlan, Dst IP Addr
  vlan-dst-ip-port      Vlan, Dst IP Addr and TCP/UDP Port
  vlan-only             Vlan
  vlan-src-dst-ip       Vlan, Src XOR Dst IP Addr
  vlan-src-dst-ip-port  Vlan, Src XOR Dst IP Addr and TCP/UDP Port
  vlan-src-ip           Vlan, Src IP Addr
  vlan-src-ip-port      Vlan, Src IP Addr and TCP/UDP Port
```

To determine which fields to use, traffic analysis should be done to determine the best hash scheme. In this solution, the ASA is configured as follows:

```
interface Port-channel1
 port-channel load-balance vlan-src-dst-ip-port
```

Below are useful **show** commands.

```
dc02-asa-fw1# sh port-channel 1 load-balance
EtherChannel Load-Balancing Configuration:
        vlan-src-dst-ip-port

EtherChannel Load-Balancing Addresses UsedPer-Protocol:
Non-IP: Source XOR Destination MAC address
  IPv4: Vlan ID and Source XOR Destination IP address and TCP/UDP (layer-4)port number
  IPv6: Vlan ID and Source XOR Destination IP address and TCP/UDP (layer-4)port number
```

ACE Load-balancing Optimization

An EtherChannel bundles individual L2 Ethernet physical ports into a single, logical link that provides the aggregate bandwidth of up to four physical links on the ACE appliance. The EtherChannel provides full-duplex bandwidth up to 4000 Mbps between the ACE appliance and another switch. In the ACE, the load-balance policy (frame distribution) can be based on a MAC address (L2), an IP address (L3), or a port number (L4). Load balancing is configured in the interface level (not the global device level).

The options are as follows:

```
dc02-ace-1/Admin(config-if)# port-channel load-balance ?
  dst-ip        Dst IP Addr
  dst-mac       Dst Mac Addr
  dst-port      Dst TCP/UDP Port
  src-dst-ip    Src XOR Dst IP Addr
  src-dst-mac   Src XOR Dst Mac Addr
  src-dst-port  Src XOR Dst TCP/UDP Port
  src-ip        Src IP Addr
  src-mac       Src Mac Addr
  src-port      Src TCP/UDP Port
```

Traffic should be analyzed to determine the best hash scheme. In this solution, the ACE is configured as follows:

```
interface port-channel 1
  port-channel load-balance src-dst-port
```

Below are useful commands for configuring the ACE.

```
dc02-ace-1/Admin# sh interface port-channel 1

PortChannel 1:
----------------------------
Description:
mode: Trunk
native vlan: 0
status: (UP), load-balance scheme: src-dst-port
```

UCS FI Load-balancing Optimization

Load balancing in the UCS Fabric FI is not required/configurable. The UCSM is configured in End-host (EH) mode. In this mode, server VIFs are dynamically pinned to the uplinks by default. The UCSM also allows static pinning with pin-group configuration. The uplinks to upstream networks on the UCS FI are configured as a port-channel. The UCSM does not have configuration option to change the port-channel load-balancing option.

Nexus 1000V Load-balancing Optimization

The Ethernet uplinks of each ESXi/VEM are configured as a MAC-pinning mode port-channel, and no port-channel load-balancing configuration is required for this kind of port-channel. In the default configuration, vEth interfaces are dynamically pinned to the individual member link of the port-channel. Static pinning can be used to better control the traffic flow. The following configuration pins the management/control traffic and Gold tenant traffic to fabric A, while traffic from other tenants is pinned to fabric B:

```
port-profile type vethernet esxi-mgmt-vmknic
  pinning id 0
port-profile type vethernet vmotion
  pinning id 0

port-profile type vethernet vsg-data
  pinning id 0
port-profile type vethernet vsg-mgmt
  pinning id 0
port-profile type vethernet vsg-ha
  pinning id 0

port-profile type vethernet gold-profile
  pinning id 2
port-profile type vethernet silver-profile
 pinning id 3
port-profile type vethernet bronze-profile
 pinning id 3
port-profile type vethernet smb-profile
 pinning id 3
```

Layer 2 Best Practices and Caveats

vPC Best Practices

- A vPC peer link is recommended to use ports from different modules to provide bandwidth and redundancy.
- "ip arp synchronize," "peer-gateway," and "auto-recovery" should be configured in the vPC configuration.
- LACP should be used if possible
- It is recommended to disable LACP graceful convergence when the other end of port-channel neighbors are non NX-OS devices.
- Pre-provision all VLANs on MST and then create them as needed.
- On the Aggregation layer, create a root or a secondary root device as usual. Design the network to match the primary and secondary roles with the spanning-tree primary and secondary switches.
- If making changes to the VLAN-to-instance mapping when the vPC is already configured, remember to make changes on both the primary and secondary vPC peers to avoid a Type-1 global inconsistency.

CHAPTER 4

Layer 3 Implementation

This chapter contains the following major topics:

End-to-End Routing

In the multiple tenants' Data Center (DC) environment, the tenants must be separated. In order for the clients to access the resources in the DC, the clients must have route reachability to the DC. This solution uses Virtual Routing and Forwarding (VRF)-Lite technology to separate tenants and Border Gateway Protocol (BGP) and static routes as the routing protocols.

This section presents the following topics:

VRF-Lite in the Data Center

VRF is a key element in the DC that allows multiple instances of a routing table to coexist within the same router at the same time. Routing instances are independent, providing a separated environment for each customer. The same or overlapping IP addresses can be used without conflicting with each other.

Each VRF instance has its own:

- IP routing table
- Derived forwarding table
- Set of interfaces
- Set of routing protocols and routing peers that inject information into the VRF

VRF-Lite is a feature that equals to VRF without the need to run Multiprotocol Label Switching (MPLS). VRF-Lite uses input interfaces to distinguish routes for different customers and forms virtual packet forwarding tables by associating one or more Layer 3 (L3) interfaces with each VRF instance.

The VRF interface can be physical, such as Ethernet ports, or logical, such as a subinterface or VLAN Switched Virtual Interface (SVI). An end-to-end VRF-Lite instance supports network virtualization and provides total separation between customer networks. Communication between customers is not possible within the cloud and backbone network.

In the ASA and ACE, a similar concept "context" can be created. In the ASA, a single security appliance can be partitioned into multiple virtual devices, known as "security contexts." Each context is an independent device, with its own security policy, interfaces, and administrators. Multiple contexts are similar to having multiple stand-alone devices. In the ACE, a virtual environment, called a "virtual context," can be created using ACE virtualization. A single ACE appears as multiple virtual devices, and each is configured and managed independently. A virtual context allows for closely and efficiently managing system resources, ACE users, and the services that are provided to customers.

In this solution, every tenant has its own VRF instance in the ASR 1000, a VRF-Lite instance in the Nexus 7000, and its own contexts in the ASA and ACE.

Tenant Load Distribution

Redundancy and load balancing are a must in the DC design. From the network topology, we see there are redundant devices and links in every layer. To best use redundancy, traffic is divided into "left" and "right" to achieve load balance for both southbound and northbound traffic (Figure 4-1 and Figure 4-2). In a normal situation, the "left" traffic will use the left half of the topology (ASR1K-1, Nexus7k-Agg1), and the "right" traffic will use the right half of the topology (ASR1K-2, Nexus7K-Agg2). If the link or node fails, the traffic will converge to other nodes and links. If the "left" traffic reaches the right half of the topology by some reason, the first choice is going through the cross link and returning to the left half of the topology. As BGP is the routing protocol for every tenant, community is a natural choice to transmit the load-balancing information. After receiving the routes with community information, the ASR 1000 and Nexus 7000 use local preference to prefer the routes.

Figure 4-1 Topology Divided to Left and Right, Northbound Traffic

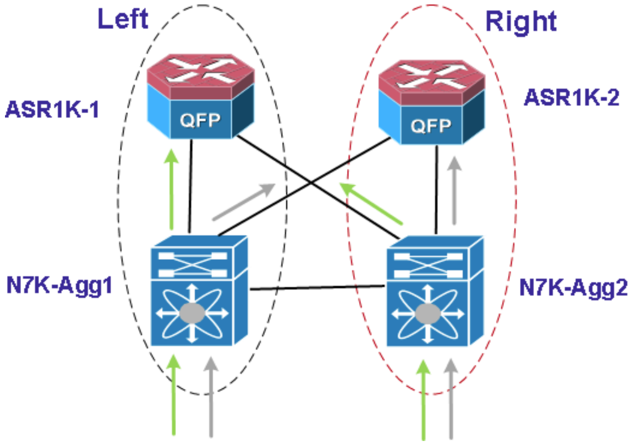
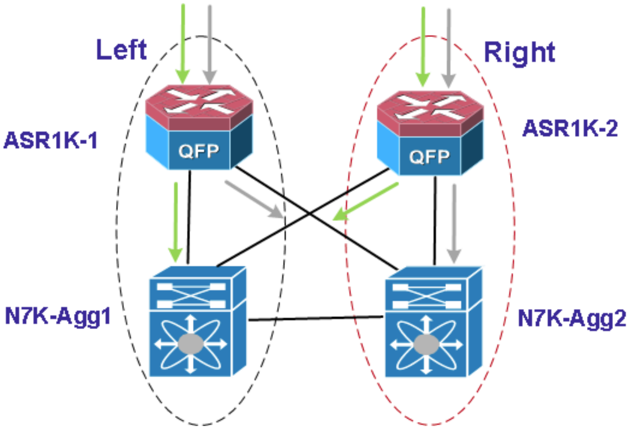


Figure 4-2 Topology Divided to Left and Right, Southbound Traffic



Implementation in the Nexus 7000

The Nexus 7000 runs BGP with the local PE and advertises the server subnets to BGP (redistribute to BGP from connect or static). To achieve load balance, a route-map is used to add community during the redistribution. In the Nexus 7000 Agg1, community 31:31 is attached to the "left" traffic, and 31:32 is attached to the "right" traffic routes. In the Nexus 7000 Agg2, community 32:31 is attached to the "left" traffic, and 32:32 is attached to the "right" traffic routes.

Implementation in the ASR 1000 (Local PE)

The ASR 1000 injects a default route to BGP and advertises the route to the Nexus 7000 for every tenant. To achieve load balance, ASR1K-1 adds community 21:21, and ASR1K-2 adds community 22:22 when the route is redistributed to BGP.

Overview of the flow

For the north to south traffic, when the ASR 1000 routers receive the routes of the server via BGP, it will use the community to set the local preference. For example, if the PE1 receives "left" traffic routes with community 31:31 (from Agg1), it will set local preference 10000, the same route, but with community 32:31 (received from Agg2), and it will set preference 5000. The PE1 will choose Agg1 as the next hop.

For the south to north traffic, the Nexus 7000 routers receive the default route via BGP from both the ASR 1000 PE1 and PE2. For the "left" traffic, the Agg1 sets the local preference higher for the routes learned from PE1 and sets PE1 as the next hop. For the "right" traffic, the Agg2 sets the local preference higher for the routes learned from PE2 and sets PE2 as the next hop.

Configuration Examples and Useful Commands

Below are configuration examples using tenant bronze_1.

Nexus 7000 a1

```
router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self
 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
    neighbor 10.3.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM permit 10
  match ip address prefix-list SERVER-NET
  set community 31:31

route-map SERVER-NET-SET-COMM permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.2.0.0/16

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

Nexus 7000 a2

```
router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM-PATH2
    neighbor 10.3.2.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.4.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.34.3
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community
```

```
route-map SERVER-NET-SET-COMM-PATH2 permit 10
  match ip address prefix-list SERVER-NET
  set community 32:31

route-map SERVER-NET-SET-COMM-PATH2 permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.0.0.0/8 le 24

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

ASR 1000 PE1

```
router bgp 109

 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both

 address-family ipv4 vrf customer_bronze1
  neighbor 10.3.1.2 remote-as 65501
  neighbor 10.3.1.2 activate
  neighbor 10.3.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.3.4.2 remote-as 65501
  neighbor 10.3.4.2 activate
  neighbor 10.3.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
 !
route-map DC2_PATH_PREFERENCE permit 10
 match community PREFER-N7K1
 set local-preference 10000
!
route-map DC2_PATH_PREFERENCE permit 20
 match community PREFER-N7K2
 set local-preference 1000
!
route-map DC2_PATH_PREFERENCE permit 30
 match community BACKUP
 set local-preference 5000
!
route-map DC2_PATH_PREFERENCE permit 40

route-map default permit 10
 match ip address prefix-list default
 set community 21:21

route-map default-condition permit 10
 match ip address prefix-list default-condition
 set community 21:21

ip prefix-list default seq 5 permit 0.0.0.0/0
ip prefix-list default-condition seq 5 permit 169.0.0.0/8

ip community-list standard PREFER-N7K1 permit 31:31
ip community-list standard PREFER-N7K2 permit 32:32
ip community-list standard BACKUP permit 32:31
```

ASR 1000 PE2

```
router bgp 109

 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both

 address-family ipv4 vrf customer_bronze1
  neighbor 10.3.2.2 remote-as 65501
  neighbor 10.3.2.2 activate
  neighbor 10.3.2.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.3.3.2 remote-as 65501
  neighbor 10.3.3.2 activate
  neighbor 10.3.3.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family

route-map DC2_PATH_PREFERENCE permit 10
 match community PREFER-N7K1
 set local-preference 1000
!
route-map DC2_PATH_PREFERENCE permit 20
 match community PREFER-N7K2
 set local-preference 10000
!
route-map DC2_PATH_PREFERENCE permit 30
 match community BACKUP
 set local-preference 5000
!
route-map DC2_PATH_PREFERENCE permit 40
!
route-map default permit 10
 match ip address prefix-list default
 set community 22:22

route-map default-condition permit 10
 match ip address prefix-list default-condition
 set community 22:22

ip prefix-list default seq 5 permit 0.0.0.0/0
ip prefix-list default-condition seq 5 permit 169.0.0.0/8

ip community-list standard PREFER-N7K1 permit 31:31
ip community-list standard PREFER-N7K2 permit 32:32
ip community-list standard BACKUP permit 31:32
```

Below are useful commands for the ASR 1000 and Nexus 7000. ASR 1000:

```
sh ip bgp vpnv4 vrf customer_bronze1 x.x.x.x
sh ip route vrf customer_bronze1
show route-map
sh ip community-list
```

Nexus 7000

```
sh ip bgp vrf customer_bronze1 0.0.0.0
sh ip bgp vrf customer_bronze1 x.x.x.x
sh ip route vrf customer_bronze1
show route-map
show ip community-list
```

ASR 1000 PE Implementation Overview

In order for the ASR 1000 to receive routes from its client networks across the Service Provider cloud, it must peer with the client PE router and receive VPNv4 prefixes from these peers. Also, in order to have Internet reachability, it must peer with the required IPv4 Internet routers. To achieve Service Provider client and Internet client reachability from the DC, the ASR 1000 conditionally injects a default route into the appropriate routing table, since it has all route prefixes in its routing table. Using this approach allows for efficient use of the routing table of devices/routers in the DC network.

To receive VPNv4 prefixes from the client PE routers, the ASR 1000 must run MP-iBGP with these routers. This requires running MPLS on the ASR 1000 and enabling Label Distribution Protocol (LDP) on relevant interfaces. In this solution, LDP is enabled on two ASR 1000 interfaces, one to the core and the other on the L3 interface that connects it to the other ASR 1000. The core interfaces on the ASR 1000 are 10G.

The ASR 1000 also has an Internet Protocol version 4 (IPv4) External Border Gateway Protocol (eBGP) neighborship with the Nexus 7004 aggregation routers. Using this peering, it advertises a default route into these routers and receives server specific network routes from them. Each tenant has a sub-interface in the VRF specific to the tenant and runs tenant specific eBGP sessions in these VRF instances between the ASR 1000 and Nexus 7004 aggregation routers.

ASR 1000 Core Routing Configuration

The core configuration on the ASR 1000 routers involves Open Shortest Path First (OSPF) and MPLS configuration, as well as the Multiprotocol Internal Border Gateway Protocol (MP-iBGP) configuration required to receive VPNv4 prefixes from client PE routers. Routing optimizations are also included for faster convergence. This includes Nonstop Forwarding (NSF) and Nonstop Routing (NSR) for OSPF, graceful-restart for MPLS, and BGP PIC Core and Edge and BGP graceful restart.

MPLS and OSPF Configuration

```
mpls ldp graceful-restart

router ospf 1
 nsr
 nsf
!
dc02-asr1k-pe1#sh run int te0/0/0
Building configuration...

dc02-asr1k-pe1#sh run int te0/0/0
Building configuration...

Current configuration : 298 bytes
!
interface TenGigabitEthernet0/0/0
 description uplink-to-core
 ip address 10.4.21.1 255.255.255.0
 ip ospf 1 area 0
 load-interval 30
 carrier-delay up 30
 plim qos input map mpls exp  5  queue strict-priority
 mpls ip
 cdp enable
 service-policy input wan-in
 service-policy output wan-out
end
```

```
dc02-asr1k-pe1#sh run int te0/1/0
Building configuration...

Current configuration : 313 bytes
!
interface TenGigabitEthernet0/1/0
 description connect to pe2
 ip address 10.21.22.1 255.255.255.0
 ip ospf network point-to-point
 ip ospf 1 area 0
 carrier-delay up 30
 plim qos input map mpls exp  5  queue strict-priority
 mpls ip
 cdp enable
 service-policy input wan-in
 service-policy output wan-out
end
```

Additional Routing Optimization Configuration

```
ip routing protocol purge interface
cef table output-chain build favor convergence-speed
cef table output-chain build indirection recursive-prefix non-recursive-prefix
cef table output-chain build inplace-modify load-sharing
```

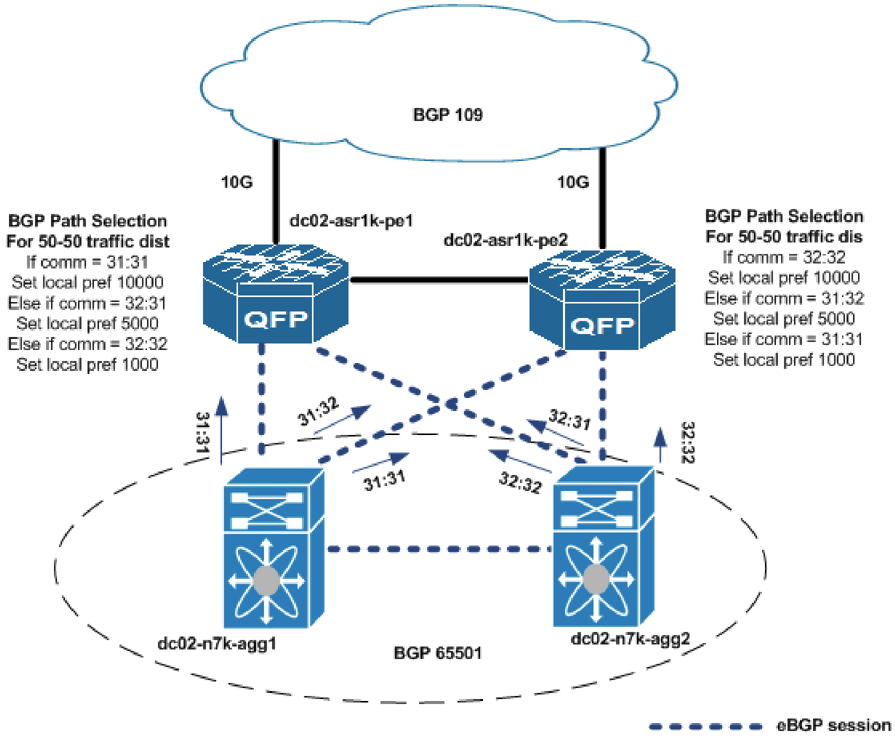
ASR 1000 DC Routing Configuration

BGP is used for DC routing between the ASR 1000 PE and the Nexus 7000 aggregation routers. The DC prefixes and server addresses are advertised by the Nexus 7000 aggregation routers, while a default route is conditionally advertised for each tenant configured on the aggregation routers. DC addresses advertised include server private and public prefixes, and the ASA public addresses used for VPN termination and the Copper tenants.

To ensure effective distribution of traffic to each of the Nexus 7000 routers, a BGP load-balancing scheme is configured. This scheme ensures 50-50 traffic distribution for all configured DC tenants by using the community value advertised by the aggregation routers to determine the preferred path. BGP path selection is based on the local preference set based on these received community values. See Tenant Load Distribution for a complete understanding of the BGP scheme used to forward traffic to the aggregation Nexus 7000 switches. BGP community values and BGP local preferences are used to determine a secondary path to be used if the primary path used by BGP fails. Using this scheme, both ASR 1000 PEs will forward traffic to the aggregation routers if the primary paths used to send tenant traffic fails. BGP PIC Edge optimization is configured to achieve faster convergence when the BGP paths fails. Both the primary and secondary BGP paths will be installed in the routing table with the secondary installed as a repair-path.

Figure 4-3 and Figure 4-4 show a diagrammatic overview of the BGP scheme used on the ASR 1000 for routing and overview of the secondary paths used by BGP to forward traffic if the primary paths fails.

Figure 4-3 ASR 1000 BGP DC Routing Overview

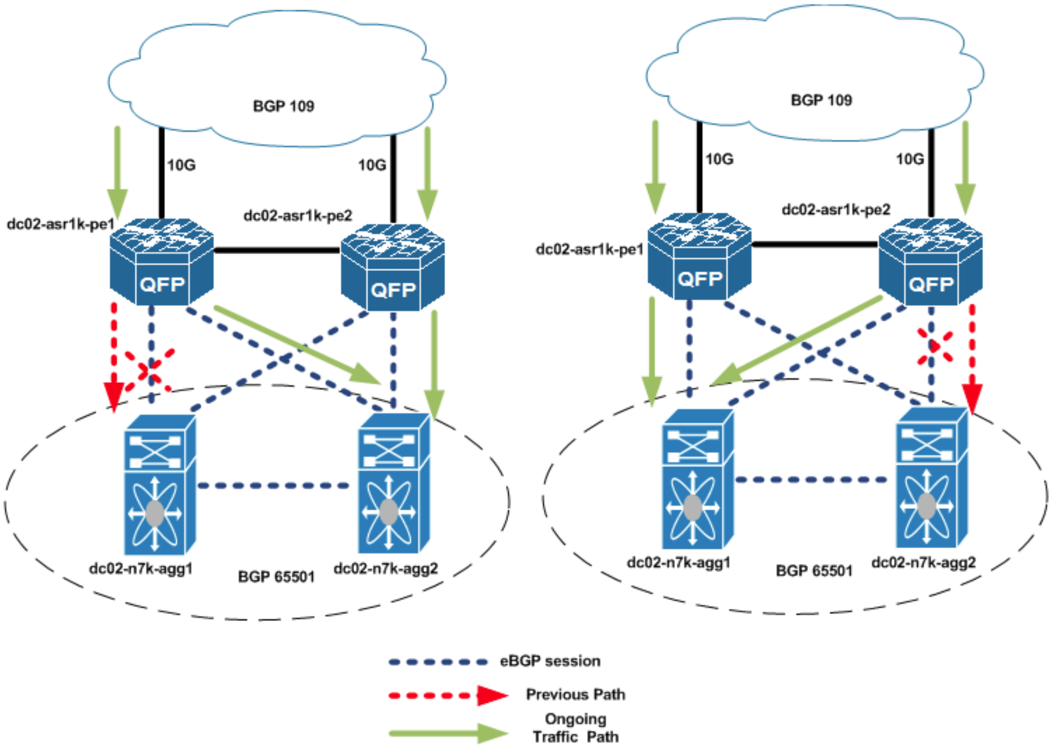


ASR 1000 BGP Routing Configuration For Sample Tenant

```
router bgp 109
 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both
 exit-peer-policy
 !
address-family vpnv4
  bgp additional-paths install
  bgp recursion host
!
address-family ipv4 vrf customer_gold2
  neighbor 10.1.1.2 remote-as 65501
  neighbor 10.1.1.2 activate
  neighbor 10.1.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.1.4.2 remote-as 65501
  neighbor 10.1.4.2 activate
  neighbor 10.1.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
!

dc02-asr1k-pe1#sh ip bgp vpnv4 vrf customer_gold1  11.1.0.0
BGP routing table entry for 21:1:11.1.0.0/16, version 347789
Paths: (2 available, best #2, table customer_gold1)
  Additional-path-install
  Advertised to update-groups:
     998
  Refresh Epoch 1
  65501
```

```
    10.1.4.2 from 10.1.4.2 (10.1.5.3)
      Origin incomplete, metric 0, localpref 5000, valid, external, backup/repair
      Community: 32:31
      Extended Community: RT:21:1 , recursive-via-connected
      mpls labels in/out 2362/nolabel
      rx pathid: 0, tx pathid: 0
  Refresh Epoch 1
  65501
    10.1.1.2 from 10.1.1.2 (10.1.5.2)
      Origin incomplete, metric 0, localpref 10000, valid, external, best
      Community: 31:31
      Extended Community: RT:21:1 , recursive-via-connected
      mpls labels in/out 2362/nolabel
      rx pathid: 0, tx pathid: 0x0
dc02-asr1k-pe1#


dc02-asr1k-pe1#sh ip route vrf customer_gold1 repair-paths 11.1.0.0

Routing Table: customer_gold1
Routing entry for 11.1.0.0/16
  Known via "bgp 109", distance 20, metric 0
  Tag 65501, type external
  Last update from 10.1.1.2 1d17h ago
  Routing Descriptor Blocks:
  * 10.1.1.2, from 10.1.1.2, 1d17h ago, recursive-via-conn
      Route metric is 0, traffic share count is 1
      AS Hops 1
      Route tag 65501
      MPLS label: none
      MPLS Flags: NSF
    [RPR]10.1.4.2, from 10.1.4.2, 1d17h ago, recursive-via-conn
      Route metric is 0, traffic share count is 1
      AS Hops 1
      Route tag 65501
      MPLS label: none
      MPLS Flags: NSF
```

Figure 4-4 ASR 1000 BGP Routing with Failure of Primary Path



With failure of the primary BGP path for a tenant, traffic will be rerouted to the repair-path/secondary path associated with the tenant prefix to ensure 50-50 traffic distribution on both ASR 1000 PEs for all configured DC tenants. In Figure 4-4, if the primary path for a tenant, dc02-asr1k-pe1->dc02-n7k-agg1 or dc02-asr1k-pe2->dc02-n7k-agg2, fails, then based on the BGP routing configuration, traffic will be routed on the dc02-asr1k-pe1->dc02-n7k-agg2 or dc02-asr1k-pe2->dc02-n7k-agg1 paths respectively.

Layer 3 Implementation on the Nexus 7004 Aggregation

In this solution, a pair of Nexus 7004 switches are placed in the Aggregation layer. This section presents the following topics:

- VRF-Lite, page 4-11
- BGP, page 4-13
- HSRP, page 4-16

VRF-Lite

Cisco NX-OS supports VRF instances. Multiple VRF instances can be configured in a Nexus 7000 switch. Each VRF contains a separate address space with unicast and multicast route tables for IPv4 and IPv6 and makes routing decisions independent of any other VRF instance. Interfaces and route protocols can be assigned to a VRF to create virtual L3 networks. An interface exists in only one VRF instance.

Each Nexus 7K router has a default VRF instance and a management VRF instance. The management VRF instance is for management purposes only, and only the mgmt0 interface can be in the management VRF instance. All L3 interfaces exist in the default VRF instance until they are assigned to another VRF instance. The default VRF instance uses the default routing context and is similar to the global routing table concept in Cisco IOS.

Below is an example of creating a VRF instance and assigning a VRF membership to the interfaces for tenant customer_silver1.

```
vrf context customer_silver1

interface Vlan501
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.1.2/24
  no ipv6 redirects
  no ip arp gratuitous hsrp duplicate

interface Vlan601
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.2.2/24
  no ipv6 redirects

interface Vlan701
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.3.2/24
  no ipv6 redirects

interface Vlan1821
  vrf member customer_silver1
  no ip redirects
  ip address 113.3.1.2/24
  no ipv6 redirects

interface port-channel343.501
  vrf member customer_silver1
  ip address 10.2.34.3/24

interface Ethernet3/9.501
  vrf member customer_silver1
  ip address 10.2.1.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.501
  vrf member customer_silver1
  ip address 10.2.3.2/24
  no ip arp gratuitous hsrp duplicate
```

Routing protocols can be associated with one or more VRF instances. In this solution, BGP is used as the routing protocol. For example, below is the routing configuration for tenant customer_silver1 on the Nexus 7000 Agg1 device.

```
router bgp 65501
  vrf customer_silver1
    graceful-restart-helper
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
      additional-paths send
      additional-paths receive
    neighbor 10.2.1.1
```

```
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.2.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.2.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community
```

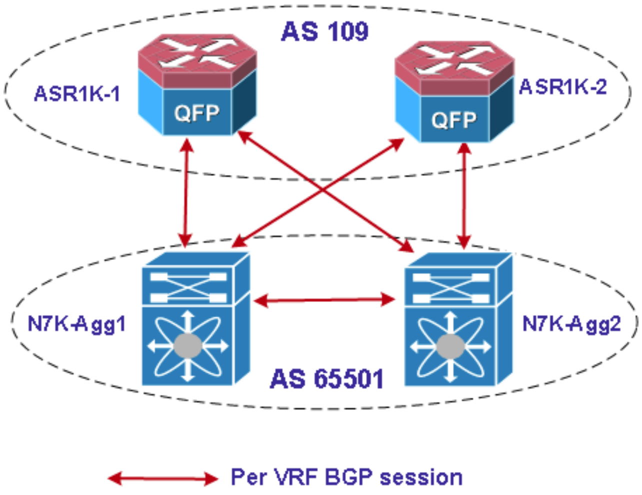
Below are useful commands.

```
show vrf XXX
show vrf XXX interface
show run vrf XXX
show ip route vrf XXX
show ip bgp vrf XXX
```

BGP

BGP is the routing protocol used in the tenants to convey routing information. Direct and/or static routes are distributed to BGP from the Nexus 7000 routers and are advertised out to the ASR 1000 routers. The Nexus 7000 routers also learn the default route from the ASR 1000 routers through BGP. Figure 4-5 shows the BGP sessions per tenant.

Figure 4-5 BGP Sessions Per Tenant



For example, for the tenant customer_bronze1, the Nexus 7000 Agg1 builds an eBGP session with ASR1k-1 using the e3/9.801 subinterface, and an eBGP session with ASR1k-2 using the e4/9.801 subinterface. To provide redundancy, there is also an iBGP session between the Nexus 7000 Agg1 and Agg2 using the po343.801 subinterface. In the iBGP session, we use the next-hop-self option. The Nexus 7000 Agg2 builds an eBGP session with ASR1k-2 using the e4/9.801 subinterface, and an eBGP session with ASR1k-1 using the e3/9.801 subinterface. As discussed in the Tenant Load Distribution section, the community is needed to convey the load-balancing information. Community sends the information to the eBGP peers (ASR 1000 routers).

Below are the related configurations of the Nexus 7000 routers for tenant customer_bronze1.

Nexus Agg

```
interface port-channel343.801
  vrf member customer_bronze1
  ip address 10.3.34.3/24

interface Ethernet3/9.801
  vrf member customer_bronze1
  ip address 10.3.1.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.801
  vrf member customer_bronze1
  ip address 10.3.3.2/24
  no ip arp gratuitous hsrp duplicate

router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
    neighbor 10.3.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM permit 10
  match ip address prefix-list SERVER-NET
  set community 31:31

route-map SERVER-NET-SET-COMM permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.2.0.0/16

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

Nexus Agg2

```
interface port-channel434.801
```

```
  vrf member customer_bronze1
  ip address 10.3.34.4/24

interface Ethernet3/9.801
  vrf member customer_bronze1
  ip address 10.3.4.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.801
  vrf member customer_bronze1
  ip address 10.3.2.2/24
  no ip arp gratuitous hsrp duplicate

router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM-PATH2
    neighbor 10.3.2.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.4.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.34.3
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM-PATH2 permit 10
  match ip address prefix-list SERVER-NET
  set community 32:31

route-map SERVER-NET-SET-COMM-PATH2 permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.0.0.0/8 le 24

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

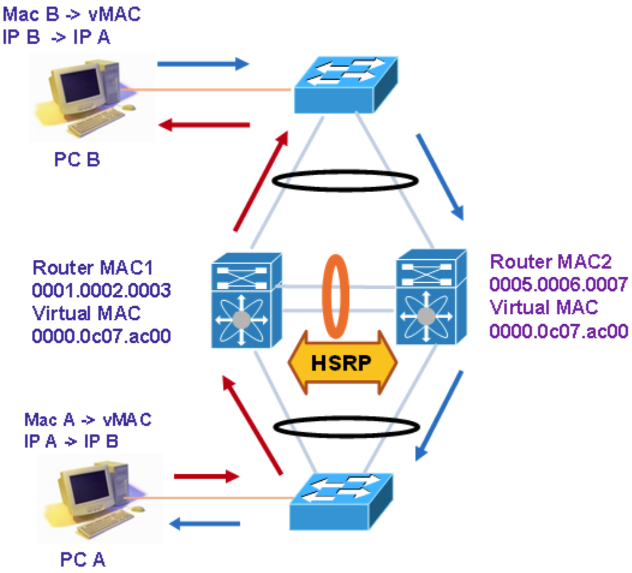
Below are useful **show** commands.

```
show ip bgp vrf XXX summ
show ip bgp vrf XXX neighbors
show ip bgp x.x.x.x vrf XXX
```

HSRP

Hot Standby Router Protocol (HSRP) is a First-Hop Redundancy Protocol (FHRP) that allows a transparent failover of the first-hop IP router. When HSRP is configured on a network segment, a virtual MAC address and a virtual IP address are provided for the HSRP group. HSRP will select one router in the group to be the active router. The active router receives and routes packets destined for the virtual MAC address of the group. In the Nexus 7000, HSRP interoperates with vPCs and behaves slightly different (Figure 4-6). Both active HSRP routers and the standby HSRP router will forward the traffic sent to it. In this solution, we put each Nexus 7000 as active (with high priority) for half of all the HSRP groups. It does not make a difference in the data plane, as the traffic is determined by the load-balance algorithm of the downstream devices (ACE, ASA, or Nexus 5000).

Figure 4-6 HSRP Behavior in a vPC Environment



vPC Peer Gateway and HSRP

Some third-party devices can ignore the HSRP virtual MAC address and instead use the source MAC address of an HSRP router. In a vPC environment, the packets using this source MAC address may be sent across the vPC peer link, causing a potential dropped packet. Configure the vPC peer gateway to enable the HSRP routers to directly handle packets sent to the local vPC peer MAC address and the remote vPC peer MAC address, as well as the HSRP virtual MAC address.

Below is the vPC configuration for the Nexus 7000 router.

```
vpc domain 998
  peer-switch
  role priority 30000
  peer-keepalive destination 192.168.50.21
  delay restore 120
  peer-gateway                    <====================
  auto-recovery
  delay restore interface-vlan 100
  ip arp synchronize
```

HSRP is used in the Nexus 7000 as the gateway of the server, ASA, and ACE. Below is the sample configuration of the server gateway for tenant customer_bronze1.

Agg1

```
interface Vlan801
  no shutdown
  vrf member customer_bronze1
  no ip redirects
  ip address 11.3.1.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 801
    preempt
    priority 150
    ip 11.3.1.1
```

Agg2

```
interface Vlan801
  no shutdown
  vrf member customer_bronze1
  no ip redirects
  ip address 11.3.1.3/24
  no ipv6 redirects
  hsrp version 2
  hsrp 801
    preempt
    priority 120
    ip 11.3.1.1
```

Below is the sample configuration of the ASA outside interface gateway for tenant customer_gold1.

Agg1

```
interface Vlan1301
  no shutdown
  vrf member customer_gold1_pub
  no ip redirects
  ip address 10.1.5.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1301
    preempt
    priority 150
    ip 10.1.5.1
```

Agg2

```
interface Vlan1301
  no shutdown
  vrf member customer_gold1_pub
  no ip redirects
  ip address 10.1.5.3/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1301
    preempt
    priority 120
    ip 10.1.5.1
```

Below is the sample configuration of the ASA inside interface gateway for tenant customer_gold1.

Agg1

```
interface Vlan1201
  no shutdown
  vrf member customer_gold1_priv
```

```
      no ip redirects
      ip address 10.1.6.2/24
      no ipv6 redirects
      hsrp version 2
      hsrp 1201
        preempt
        priority 150
        ip 10.1.6.1
```

Agg2

```
    interface Vlan1201
      no shutdown
      vrf member customer_gold1_priv
      no ip redirects
      ip address 10.1.6.3/24
      no ipv6 redirects
      hsrp version 2
      hsrp 1201
        preempt
        priority 120
        ip 10.1.6.1
```

The ACE is running in one-arm mode, and the interface is in the same subnet as the server, so the ACE gateway is the same as the server gateway.

Below are useful commands.

```
    show hsrp brief
    show hsrp group
```

Services Layer Routing Configuration

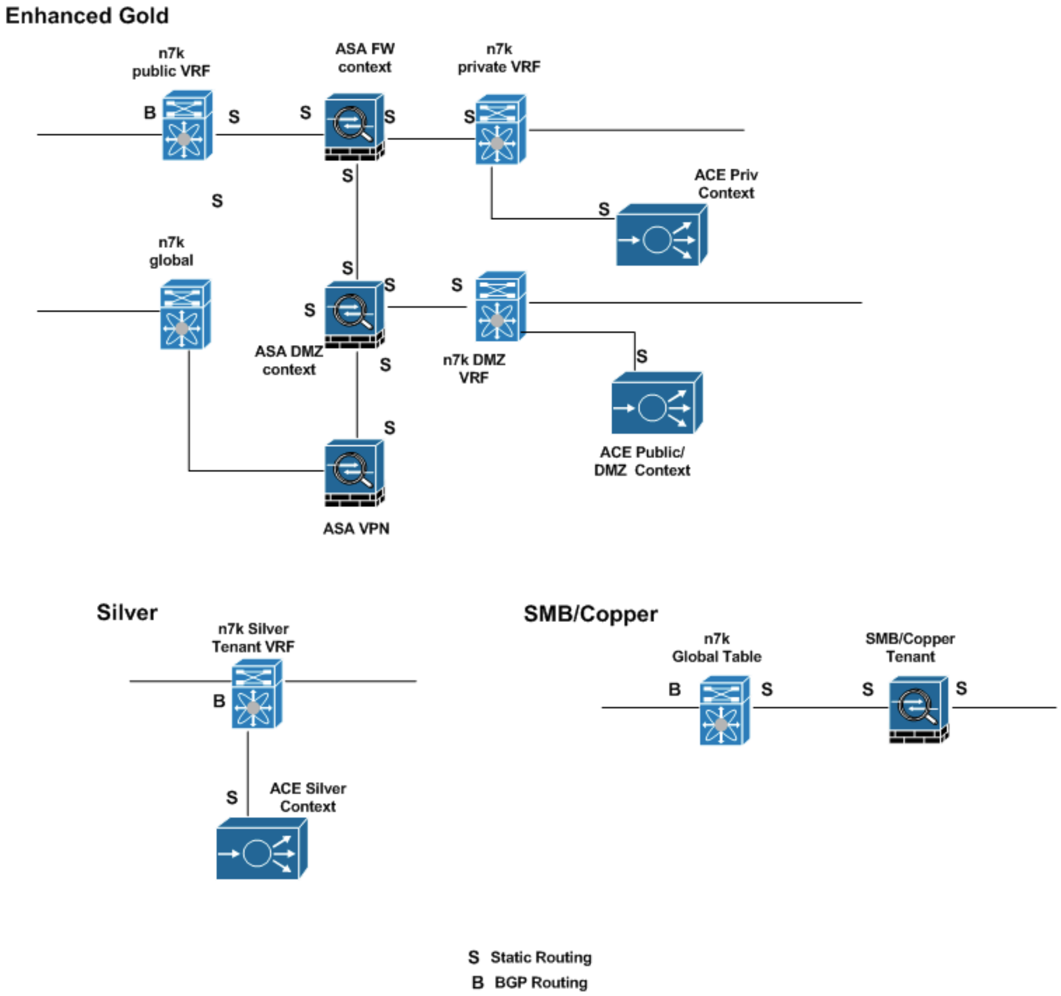
This section provides details on how end-to-end routing is achieved through the Services layer.

All Service Provider client-server Gold tenant traffic received on the Nexus 7000 aggregation routers is forwarded to the tenant ASA firewall context. Also, this traffic might be forwarded to the ACE appliances if the traffic is to be load balanced or is return traffic to load-balanced traffic. Internet client-server traffic will either be sent to the ASA VPN or the ASA firewall, depending on the type of security services associated with that traffic. All Service Provider Silver client-server traffic received on the Nexus 7000 aggregation routers will either be forwarded to the ACE appliance if application services (application load balancing, SSL overload, etc.) need to be provided, or will be forwarded to the Compute layer. Internet services are not provided to Silver tenants. All Service Provider Bronze client-server traffic received will be forwarded to the Compute layer. No Internet/security/application services are provided to these tenants. All Internet SMB/Copper tenants received on the Nexus 7000 routers will be forwarded to the ASA firewall context associated with this tenant.

This section also provides the routing configuration required for end-to-end traffic routing through the ASA firewall, ASA VPN, and ACE appliances. For end-to-end reachability through the Services layer, the appropriate static routes are configured on both the ASA and on the Nexus 7000 tenant VRF instance. Since the VIP and ACE client NAT pools are in the same subnet as the servers, the Nexus 7000 aggregation routes L2 forward packets destined to the ACE appliances.

An overview of Services layer routing is provided in Figure 4-7.

Figure 4-7 Services Layer Overview



This section presents the following topics:

- ASA Firewall Context Routing Configuration, page 4-19
- ASA VPN Routing Configuration, page 4-21
- ACE Routing Configuration, page 4-21

ASA Firewall Context Routing Configuration

For end-to-end routing, the ASA firewall should be able to provide routing for the following traffic:

1. **Tenant Service Provider client traffic that is destined to the Service Provider server private network.** This include Service Provider server and VIP private addresses. To achieve this, static routes are configured on the tenant firewall context to provide reachability to the Service Provider client and server network.
2. **Tenant Service Provider client traffic that is destined to the DMZ network.** This includes the DMZ server and VIP private addresses. To achieve this, Service Provider client addresses are NAT'd before traffic is sent to the DMZ firewall context. The firewall DMZ context has static routes that provide reachability to the Service Provider client NAT network and DMZ server network.

3. **Tenant Internet client traffic that is destined to the DMZ network.** To achieve this, static routes are configured on the DMZ firewall context that provides reachability to the Internet and DMZ server network.
4. **Tenant Internet VPN client (SSL or IPsec) traffic destined to the DMZ and private network.** Static routes are configured on the ASA VPN to provide reachability to the DMZ and private networks. Since VPN client addresses are in the same subnet with the VPN interface in the DMZ context associated with a tenant, static routes providing reachability to VPN client addresses are not required in the DMZ context.

Sample routing configurations for a Gold tenant's private and public ASA contexts are shown below:

Expanded Gold Tenant Private Context Routing Configuration

```
dc02-asa-fw1/customer-gold1-dmz# changeto c customer-gold1
dc02-asa-fw1/customer-gold1# sh route

!snip
dc02-asa-fw1/customer-gold1-dmz# changeto c customer-gold1
dc02-asa-fw1/customer-gold1# sh route

Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, ia - IS-IS inter area
       * - candidate default, U - per-user static route, o - ODR
       P - periodic downloaded static route

Gateway of last resort is 10.1.5.1 to network 0.0.0.0

S    111.0.0.0 255.0.0.0 [1/0] via 10.1.6.1, inside
C    10.1.8.0 255.255.255.0 is directly connected, dmz
C    10.1.6.0 255.255.255.0 is directly connected, inside
C    10.1.5.0 255.255.255.0 is directly connected, outside
S    11.0.0.0 255.0.0.0 [1/0] via 10.1.6.1, inside   # route to private server network
S    11.1.4.0 255.255.255.0 [1/0] via 10.1.8.11, dmz # route to DMZ server network
S    11.255.0.0 255.255.0.0 [1/0] via 10.1.8.11, dmz # route to VPN client networks
C    192.168.50.0 255.255.255.0 is directly connected, mgmt
S*   0.0.0.0 0.0.0.0 [1/0] via 10.1.5.1, outside # default route to private client 
networks
S    192.168.0.0 255.255.0.0 [1/0] via 192.168.50.1, mgmt
dc02-asa-fw1/customer-gold1#
dc02-asa-fw1/customer-gold1# changeto c customer-gold1-dmz
dc02-asa-fw1/customer-gold1-dmz# sh route

Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, ia - IS-IS inter area
       * - candidate default, U - per-user static route, o - ODR
       P - periodic downloaded static route

Gateway of last resort is 100.200.1.1 to network 0.0.0.0

S    51.0.0.0 255.0.0.0 [1/0] via 10.1.8.21, inside  # route to private client 
networks
C    100.200.1.0 255.255.255.0 is directly connected, internet
C    10.1.8.0 255.255.255.0 is directly connected, inside
C    10.1.7.0 255.255.255.0 is directly connected, dmz
```

```
S    11.1.1.0 255.255.255.0 [1/0] via 10.1.8.21, inside # route to private server
network
S    11.1.4.0 255.255.255.0 [1/0] via 10.1.7.1, dmz # route to dmz server network
C    11.255.1.0 255.255.255.0 is directly connected, vpn # interface to VPN network
C    192.168.50.0 255.255.255.0 is directly connected, mgmt
S*   0.0.0.0 0.0.0.0 [1/0] via 100.200.1.1, internet # default route to internet
S    192.168.0.0 255.255.0.0 [1/0] via 192.168.50.1, mgmt
dc02-asa-fw1/customer-gold1-dmz#
```

ASA VPN Routing Configuration

As mentioned in the previous section, the ASA VPN should have static routes that provide reachability to the DMZ and private network. The next hop for these routes must be associated with a VLAN interface, and this VLAN interface must be configured under the tunnel group used to establish the tunnel. See ASA IPsec VPN Configuration and ASA SSL VPN Configuration for details on how to associate a VLAN ID with a VPN tunnel group. For Internet reachability, the ASA has a default route that points to the Internet SVI interface on the Nexus 7000.

Sample ASA VPN Routing Configuration

```
route internet 0.0.0.0 0.0.0.0 100.200.1.1 1
route dmz1 11.1.0.0 255.255.0.0 11.255.1.251 1
route dmz2 11.1.0.0 255.255.0.0 11.255.2.251 2
route dmz3 11.1.0.0 255.255.0.0 11.255.3.251 3
route dmz4 11.1.0.0 255.255.0.0 11.255.4.251 4
route dmz5 11.1.0.0 255.255.0.0 11.255.5.251 5
route dmz6 11.1.0.0 255.255.0.0 11.255.6.251 6
route dmz7 11.1.0.0 255.255.0.0 11.255.7.251 7
route dmz8 11.1.0.0 255.255.0.0 11.255.8.251 8
route dmz9 11.1.0.0 255.255.0.0 11.255.9.251 9
route dmz10 11.1.0.0 255.255.0.0 11.255.10.251 10
route management 192.168.0.0 255.255.0.0 192.168.50.1 1
dc02-asa5555-1#
```

ACE Routing Configuration

Each ACE tenant context is configured with a default route that points to the HSRP VIP of the web VLAN interface on the Nexus 7000 switches. In this implementation, the web interface is used by the ACE to forward traffic to the Service Provider L3VPN client networks, and the ACE web VIP addresses are in the same subnet with the server and the VLAN interface on the Nexus 7000 aggregation switches. This eliminates the need to have static routes from the Nexus 7000 switches to the ACE, however, if required, separate subnets can be used for VIP addresses, and static routes would be needed on the Nexus 7004 VRF instances for the tenant pointing to the ACE interface address.

Layer 3 Best Practices and Caveats

Best Practices

1. To accelerate L3 convergence, spread the L3 ports on different SoCs on the F2 module. This is due to the fact that on the F2 module, each port is mapped to a VRF instance and then the FIB for that VRF is downloaded. If an SoC has all ports as L2, then during reload and possibly other conditions, when the ports come up, FIB download is delayed until the SVI to VRF mapping is done, and hence FIB download happens after the port comes up and L2 convergence and mapping of VLANs to that

port is complete. In VMDC 2.3 implementation, the L3 ports to the DC PEs and the VPC peer links were spread across five SoCs per module to get the benefit of FIB download immediately on reload. Refer to Cisco Nexus 7000 F2-Series 48-Port 1 and 10 Gigabit Ethernet Module Data Sheet for more information about F2 card and SoCs. Also, see CSCue67104 below.

2. To reduce traffic loss after system reload, delay the time that it takes for VLAN interface and vPCs to come online. By default, VLAN interfaces are brought online 10 seconds after the peer link is up, and vPCs are brought online 30 seconds after the VLAN interfaces are brought up. Based on scale characteristics of this validation, we delay VLAN interfaces and vPCs from coming online by 90 seconds each.
3. The ACE 4710 appliances do not support LACP, and hence their port-channels to the Nexus 7000 switches are static with mode on. We expect to see some traffic loss when the system comes online after a reload. To protect against this loss, carrier delays can be configured on the ACE GigabitEthernet interfaces to prevent this interface from coming online. Using this scheme will introduce a carrier-delay time during a vPC shut/no shut test or similar negative event.
4. Carrier delay can be configured on the ASR 1000 interfaces to the Nexus 7000 aggregation routers to delay the L3 interface from coming up. This ensures that these L3 interfaces are brought up at a time when the Nexus 7000 routers are ready to successfully set up and establish BGP sessions. In this validation, the carrier delay on the ASR 1000 PE was set to the maximum of 60 seconds.
5. By default, the ACE 4710 appliance will renew ARP entries for a configured host every 300 seconds. We increase the ARP rates to 1440 seconds to reduce the possibility of the ACE ARP request being lost as the system comes online after a reload.
6. To get better convergence performance, use BGP policy to divert traffic away from the Nexus 7004 aggregation switch under certain conditions such as VPC peer link fail or secondary shutdown. This is because the FIB programming on the F2 card is slower, leading to additional packet losses of up to 10 seconds in the scale validated, and this can be higher with a high-programmed prefix count. BGP configuration on the ASR 1000 and Nexus 7000 aggregation routers is set up so that the ASR 1000 reroutes traffic to an alternate path if the vPC peer link fails and shuts down the VPC secondary. This eliminates up to 10 seconds of traffic loss that occurs due to the F2 FIB programming delay. If the peer link fails, expect up to 13 seconds of traffic convergence, which is due to up to 8 seconds being required for the VLAN interface to go down, and due to up to 5 seconds being required for the BGP and RIB update on the Nexus 7000 aggregation routers. The causes of this convergence delay in FIB programming is under investigation. See CSCue59878 below. For overall vPC convergence, there are a few enhancements targeted for the next NX-OS software release 6.2.
7. BGP PIC, BGP graceful restart, and other routing optimization should be enabled on the ASR 1000 PE devices for faster convergence. BGP PIC and graceful restart are enabled by default on the Nexus 7000 aggregation routers.

Caveats

1. CSCud23607 was an HSRP programming issue seen if the MAC address table size limits are reached. This is fixed in NX-OS 6.1.3. Prior to NX-OS 6.1.3 , the workaround was to manually flap the affected HSRP interfaces.
2. CSCue59878 was filed to investigate the FIB programming delay after routing convergence during a vPC shut test or similar scenarios. This issue is under investigation. The reason for delay is due to the FIB programming mechanism used for the F2 module. The module has to program TCAM for all 12 SoCs, and as the number of prefixes gets higher, it takes additional time to calculate and program each of the SoCs. The workarounds are to reduce the number of SoCs used, i.e., less number of ports and to reduce the number of prefixes per SoC (by mapping specific VRF instances (ports) to SoCs so that the total prefix is less per SoC). If convergence times need to be quicker, and with a larger number of prefixes, consider using M2 or M1 series modules.

3. CSCue67104 was filled to investigate convergence delays due to packet losses after system reload of the nexus 7000 aggregation router. These losses are seen as FIB losses when the vPC port-channels are brought up and can last 10 or more seconds. This issue was closed as this is expected. On F2 modules, which have an SoC design, each SoC needs to map all of its ports into VRF instances, and then download the FIB. When all of the ports on an SoC are L2 only, the L2 ports need to come up and the SVIs need to be mapped to VRF instances before downloading the FIB for those VRF instances. This takes additional time after the port comes up (see CSCue59878 above, F2 FIB convergence is slow). To work around this issue, have a mix of both L2 and L3 ports on the same SoC. The L3 ports being on the SoC will cause all FIBs for the VRF instances on the L3 port to be downloaded as soon as the module comes up. In VMDC 2.3, all VRF instances used are allowed on the L3 port, so all FIBs will be downloaded to any SoC that has L3 ports. Since there are two L3 uplinks and four L3 peer links for iBGP per box, this provides one L3 port for uplink and two iBGP ports for peer per module. These ports should be spread on three different SoCs. Additionally, we can also spread the vPC peer link ports in different SoCs. Since there are four ports in the vPC peer link, two ports from each module, this covers two more SoCs. This helps with the reload case, as the vPC peer link will come online first and have SVIs mapped to it followed by FIB download, before the actual vPC port-channels come up, however, this will not help in the module restore case, as the vPC peer link port SoCs and FIB download will still be delayed. Additional L3 ports can help, if they are configured on any additional SoCs used. The goal with this workaround is to have all SoC FIBs programmed by the time the vPC port-channels come online.
4. CSCuc51879 is an issue seen during RP failover either due to RPSO or In-Service System Upgrade (ISSU). This is an issue related to traffic loss seen during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration.

The following performance fixes are expected in the 6.2 release of NX-OS. These fixes are expected to help with convergence.

- CSCtn37522: Delay in L2 port-channels going down
- CSCud82316: VPC Convergence optimization
- CSCuc50888: High convergence after F2 module OIR

CHAPTER 5

Services Implementation

Please refer to the Service Tiers section for the different services offered to different types of tenants, Gold, Silver, Bronze, and Copper. This section discusses the implementation on the services nodes, the Application Control Engine (ACE) 4710 for Server Load Balancing (SLB), the Adaptive Security Appliance (ASA) 5585-based perimeter firewall, the ASA 5555-x-based VPN access, and the Virtual Security Gateway (VSG) for the virtualized compute firewall.

ACE 4710 Appliance

This section presents the following topics:

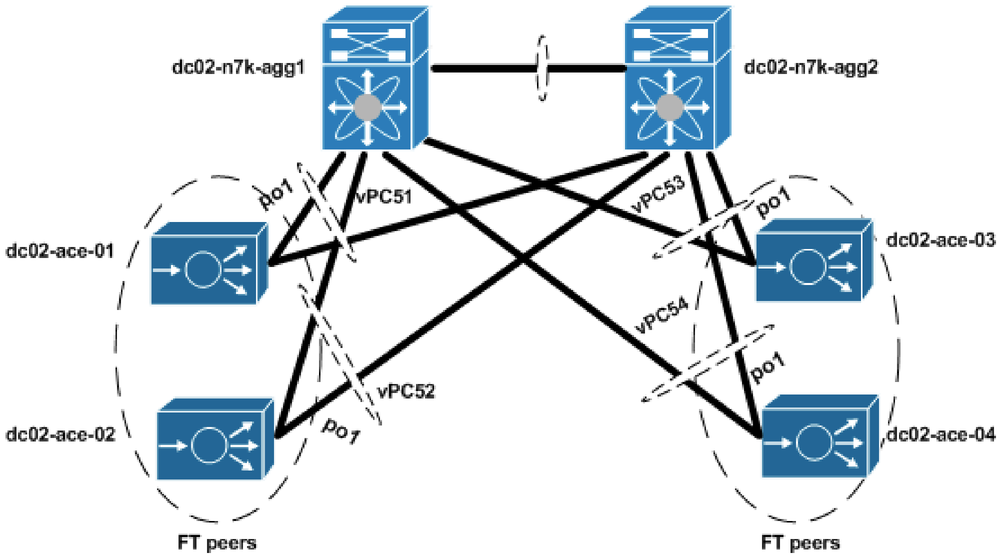
ACE Redundancy Configuration

The ACE appliances used in this solution are configured in active/active redundancy mode, i.e., a pair of ACE appliances forming a Fault-Tolerant (FT) peer will forward traffic for different contexts. To provide redundancy, a FT VLAN is configured on both ACE appliances. This FT VLAN is used by the ACE appliances to send state information, replication data redundancy protocol packets, heartbeat packets, and configuration synchronization packets. For the ACE appliances, the FT VLAN should be trunked using the **ft-port vlan <vlan-id>** command. This identifies a FT VLAN on a trunk port and ensures that proper QoS treatment is applied to packets on that VLAN. Proper QoS treatment ensures that FT messages between peers are not dropped in the network. The Nexus 7000 devices are also configured to ensure that proper QoS treatment is given to FT packets.

In this solution, the ACE appliance FT packets are switched through the Layer 2 (L2) network. The ACE FT peer connects to the L2 network through port-channels with all four Gigabit Ethernet interfaces as members. The endpoints of these port-channels are vPCs on a pair of Nexus 7000 switches. Based on L2 forwarding rules on the Nexus 7000 switches, redundancy packets of the ACE received by a Nexus 7000 switch will be switched to the vPC that connects to the other ACE

FT peer. It will be rare for the redundancy packets to transverse any other link, especially the Nexus 7000 vPC peer links. Figure 5-1 shows the L2 connection between the ACE appliances and the Nexus 7000 switches. A single management VLAN is used to manage the four ACE appliances used in this solution. To avoid MAC collision, each of the ACE appliances are assigned a different shared-vlan hostid to ensure that they derive their MAC addresses from a different MAC address pool. The ACE modules and appliances are built with 16 pools of MAC addresses.

Figure 5-1 ACE-Nexus 7000 Topology



Below is a sample configuration required for ACE redundancy.

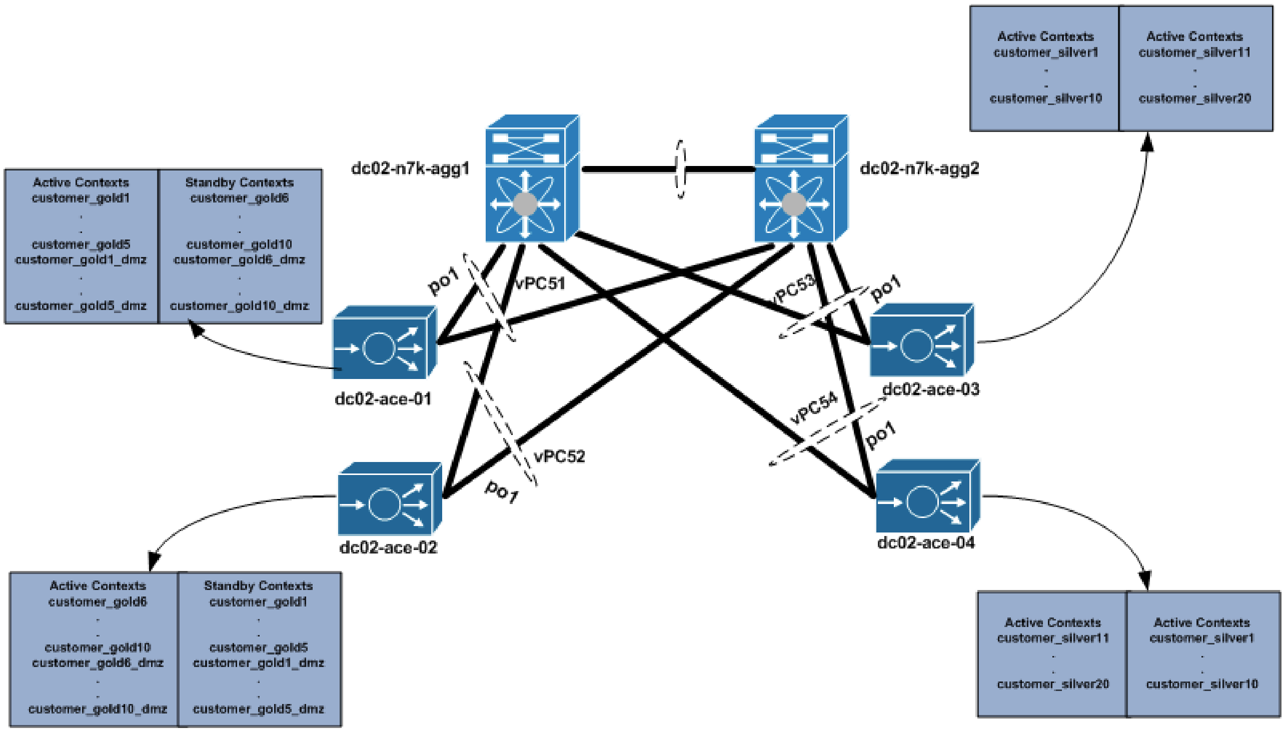
```
shared-vlan-hostid 3
peer shared-vlan-hostid 4

context customer_silver1
  allocate-interface vlan 60
  allocate-interface vlan 501
  allocate-interface vlan 601
  allocate-interface vlan 701
  member silver-sticky-class

ft peer 1
  heartbeat interval 1000
  heartbeat count 10
  ft-interface vlan 1998
ft group 31
  peer 1
  priority 255
  peer priority 120
  associate-context customer_silver1
  inservice

interface port-channel 1
  ft-port vlan 1998
  switchport trunk allowed vlan
60,201-210,301-310,401-410,501-520,601-620,701-720,1601-1610
  port-channel load-balance src-dst-port
```

ACE Context Distribution

A total of 40 contexts are used in this validation. In this solution, each Gold tenant network configuration is validated with two ACE contexts. One ACE context is used to load balance tenant Private, or PVT, server resources, and the other ACE context is used to load balance Demilitarized Zone (DMZ) server resources. A single ACE context is assigned to each Silver tenant. Two ACE 4710 appliance are used for Gold tenants, while two other ACE 4710s are used for Silver tenants. The ACEAP-04-LIC licenses are installed in each appliance. This license type allows the system to reach the context and throughput limits of each appliance. The number of ACE appliances can be increased to accommodate more Gold/Silver tenants.

Figure 5-2 shows the ACE context distribution used in this solution. For effective load balancing and redundancy, half of each tenant type is active in each of the ACE 4710s assigned to that tenant type, while the other half will be standby on the other ACE 4710 appliance.

Figure 5-2 ACE Context Distribution



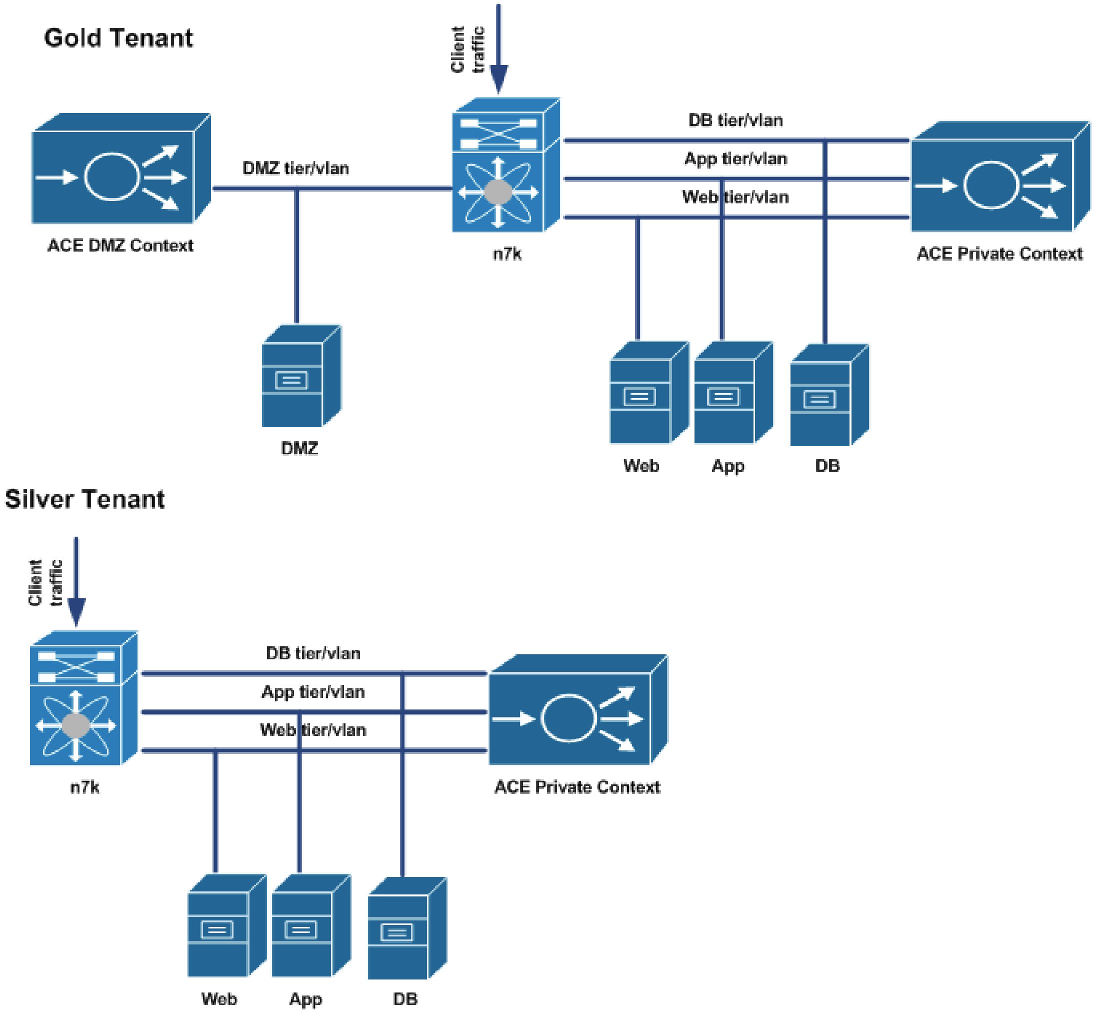
ACE SLB Configuration

The ACE contexts for the tenants are configured in a one-arm mode. Using this mode, the ACE data interfaces are in the same VLAN with that of the servers. In this implementation, the ACE VIP addresses are also chosen to be in the same server subnet. This eliminates the need to have additional static routes on the Nexus 7000 aggregation switches. Each ACE context is configured with a default route pointing to the VLAN interface of the Nexus 7000. Client addresses are translated to addresses in the same server subnet to ensure load-balanced return traffic goes to the ACE context. Each ACE context is configured to load balance Layer 4 (L4) - Layer 7 (L7) traffic. L4 load-balancing policies are configured for User Datagram Protocol (UDP) traffic, while L7 load-balancing policies are configured for Hypertext Transfer Protocol (HTTP) traffic.

In this implementation, each Gold and Silver ACE private context is configured to load balance web client traffic, web tier traffic directed to the app tier VIP, and app tier traffic directed to the Database (DB) tier VIP address. An ACE DMZ context is configured for each Gold tenant to load balance traffic destined to the DMZ servers.

Figure 5-3 shows the ACE SLB topology for Gold and Silver tenants.

Figure 5-3 Figure 5-3.Gold and Silver ACE Context



The following sections show the ACE SLB configurations.

Rserver Configuration

The rserver configuration is used to associate the real server IP address to an object name. This object name is the rserver name and will be used to define members of a server farm. A sample configuration is shown below.

Sample Rserver Configuration

```
rserver host web-server1
  ip address 11.1.1.11
  inservice
rserver host web-server2
  ip address 11.1.1.12
  inservice
rserver host web-server3
  ip address 11.1.1.13
  inservice
```

Server Farm Configuration

The server farm is a set of real servers providing the same application service. Client traffic is load balanced among the real servers in a server farm using a predictor algorithm. By default, the predictor used is round robin and this is used in this solution. The server farm configuration also provide a convenient way to take real servers offline or bring real servers online. Real servers information is added into the server farm configuration using their associated rserver names. In addition, probes can be applied to a server farm to ensure that the VIP addresses are taken offline if there is no real server available to handle requests. A sample configuration is shown below.

```
serverfarm host app-serverfarm
  rserver app-server1
    inservice
  rserver app-server2
    inservice
  rserver app-server3
    inservice
serverfarm host db-serverfarm
  rserver db-server1
    inservice
  rserver db-server2
    inservice
  rserver db-server3
    inservice
serverfarm host udp-serverfarm
  rserver udp-host
    inservice
serverfarm host udp-serverfarm:30000
  rserver udp-host:30000
    inservice
serverfarm host web-serverfarm
  rserver web-server1
  rserver web-server2
  rserver web-server3
  rserver web-spirent
    inservice
```

Class-Maps

Management, L4, and L7 class-maps are used in the configuration. The Management class-map defines management related traffic that is allowed to the ACE contexts. L4 class-maps are used to define the L4 ports that are used as match criteria for client traffic that will be load balanced. Typically, UDP and Transmission Control Protocol (TCP) ports are used as match criteria. L7 class-maps are used to define the L7 header values that will be used as match criteria for load balancing. In this implementation, HTTP URL values are used to define criteria. A sample configuration used in this solution is shown below.

Sample Management Class-map

```
class-map type management match-any management-traffic
  2 match protocol ssh any
  3 match protocol http any
  4 match protocol https any
  5 match protocol icmp any
  6 match protocol telnet any
  7 match protocol snmp source-address 192.168.0.0 255.255.0.0
```

Sample L4 Class-map

```
class-map match-all udp-vip
  2 match virtual-address 11.1.1.111 udp eq 69
class-map match-all udp-vip:30000
  2 match virtual-address 11.1.1.111 udp eq 30000
class-map match-all web->app-vip
  2 match virtual-address 11.1.2.111 tcp eq www
class-map match-all web-vip
  2 match virtual-address 11.1.1.111 tcp eq www
class-map match-all app->db-vip
  2 match virtual-address 11.1.3.111 tcp eq www
```

Sample L7 Class-map

```
class-map type http loadbalance match-any cm-app-subnet
  2 match source-address 11.1.2.0 255.255.255.0
class-map type http loadbalance match-any cm-http
  2 match http url /.*.txt
  3 match http url /.*.html
class-map type http loadbalance match-any cm-web-subnet
  2 match source-address 11.1.1.0 255.255.255.0
class-map type http loadbalance match-all cm-app->db
  2 match class-map cm-http
  3 match class-map cm-app-subnet
class-map type http loadbalance match-all cm-web->app
  2 match class-map cm-http
  3 match class-map cm-web-subnet
```

NAT Configuration

Either Source NAT with PAT (SNAT) or Policy Based Routing (PBR) are used to implement the one-arm ACE topology. In this solution, we use SNAT with PAT to implement the one-arm ACE configuration. This involves the ACE translating the client source address to an address in a pool to ensure that client return traffic from the server farm is received by the ACE appliance. We use a server farm subnet address range to define the NAT pool, and this eliminates the need to have static routes on the Nexus 7000 switches. The server receiving the client traffic will have an ARP entry that it receives from the ACE context. We use PAT to ensure that we do not quickly deplete the pool when client requests are received. The NAT pool is defined on the interface, and this NAT pool is associated with an L4 policy-map. A sample configuration used in this solution is shown below.

Sample NAT configuration

```
interface vlan 201
  description web tier
  ip address 11.1.1.22 255.255.255.0
  alias 11.1.1.21 255.255.255.0
  peer ip address 11.1.1.23 255.255.255.0
  access-group input web-acl
  nat-pool 1 11.1.1.24 11.1.1.30 netmask 255.255.255.0 pat
  nat-pool 11 11.1.1.41 11.1.1.41 netmask 255.255.255.255
  nat-pool 12 11.1.1.42 11.1.1.42 netmask 255.255.255.255
  service-policy input lb-policy
  no shutdown
interface vlan 301
  description app tier
  ip address 11.1.2.22 255.255.255.0
  alias 11.1.2.21 255.255.255.0
  peer ip address 11.1.2.23 255.255.255.0
  access-group input app-acl
  nat-pool 2 11.1.2.24 11.1.2.30 netmask 255.255.255.0 pat
  service-policy input web->app-lb
  no shutdown
interface vlan 401
  description db tier
  ip address 11.1.3.22 255.255.255.0
  alias 11.1.3.21 255.255.255.0
  peer ip address 11.1.3.23 255.255.255.0
  access-group input db-acl
  nat-pool 3 11.1.3.24 11.1.3.30 netmask 255.255.255.0 pat
  service-policy input app->db-lb
  no shutdown
```

Policy-Maps

Management, L4, and L7 policy-maps are used in the configuration. The Management policy-map defines the action that will be taken if there is a match in the management class-map. The L4 load-balance policy combines the L4 class-map with an L7 load-balance policy. This L4 load-balance policy defines what traffic should be load balanced and what load-balance policy should be applied to this traffic. The load-balance policy applied to matched traffic is defined by the L7 load-balance policy. This policy defines L7 match criteria for received traffic and the server farm that handles L7 traffic. L4 polices are applied to data interfaces on the ACE context using the service-policy. The sample configurations used in this solution are shown below.

Sample Management Configuration

```
policy-map type management first-match management-traffic
  class management-traffic
    permit
```

Sample L7 Policy-map

```
policy-map type loadbalance first-match app->db-lb-policy
  class cm-app->db
    sticky-serverfarm customer_gold1-app->db
policy-map type loadbalance first-match udp-lb-policy
  class class-default
    serverfarm udp-serverfarm
policy-map type loadbalance first-match udp-lb-policy:30000
  class class-default
    serverfarm udp-serverfarm:30000
policy-map type loadbalance first-match web->app-lb-policy
  class cm-web->app
```

```
    sticky-serverfarm customer_gold1-web->app
policy-map type loadbalance first-match web-lb-policy
  class cm-http
    sticky-serverfarm customer_gold1-http
```

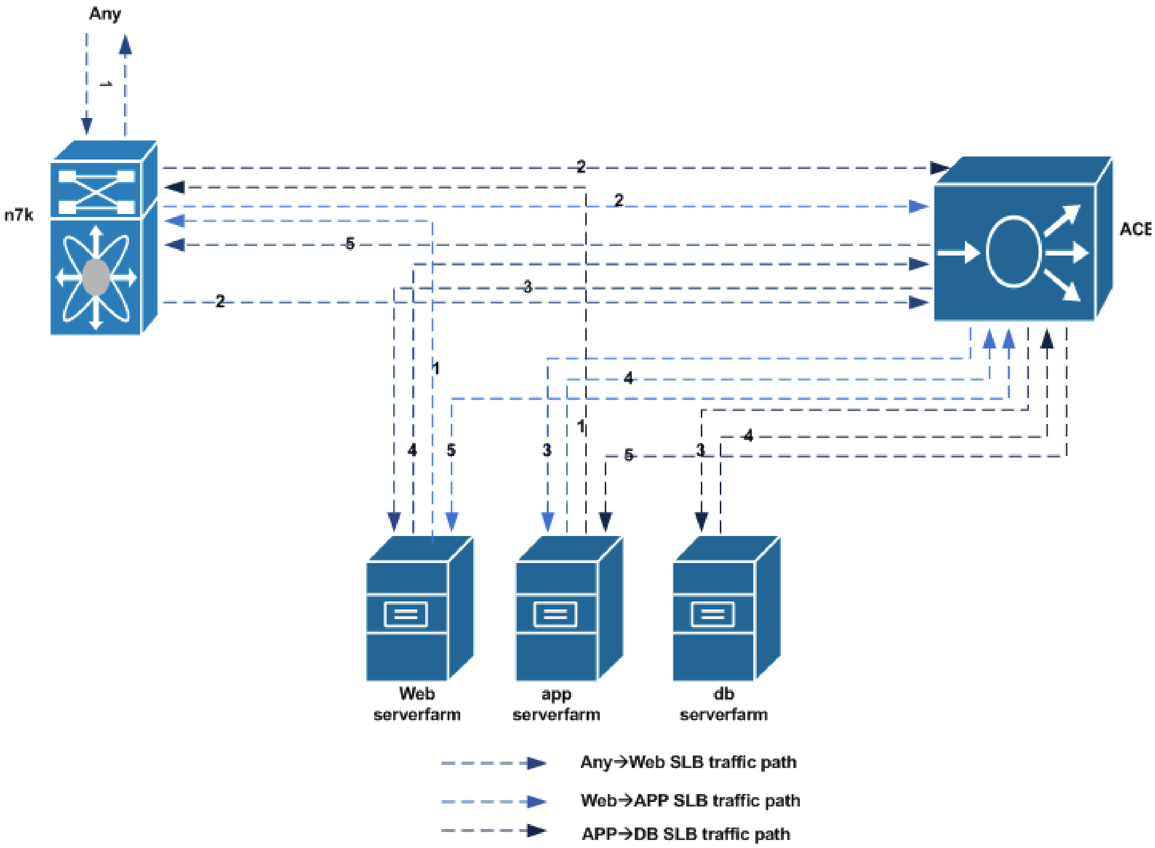
Sample L4 Policy-map

```
policy-map multi-match app->db-lb
  class app->db-vip
    loadbalance vip inservice
    loadbalance policy app->db-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 3 vlan 401
policy-map multi-match lb-policy
  class web-vip
    loadbalance vip inservice
    loadbalance policy web-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 1 vlan 201
    connection advanced-options tcp_pm
  class udp-vip
    loadbalance vip inservice
    loadbalance policy udp-lb-policy
    loadbalance vip icmp-reply
    nat dynamic 11 vlan 201
    connection advanced-options udp_pm
  class udp-vip:30000
    loadbalance vip inservice
    loadbalance policy udp-lb-policy:30000
    loadbalance vip icmp-reply active
    nat dynamic 12 vlan 201
    connection advanced-options udp_pm
policy-map multi-match web->app-lb
  class web->app-vip
    loadbalance vip inservice
    loadbalance policy web->app-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 2 vlan 301
```

ACE SLB Traffic Path Overview

Figure 5-3 shows the ACE SLB traffic path overview.

Figure 5-4 ACE SLB Traffic Path Overview



AnyWeb VIP Traffic Path

1. Traffic destined to the web VIP is received from Gold and Silver tenant client networks.
2. Traffic destined to the web VIP will be forwarded by the Nexus 7000 aggregation routers (on the same web VLAN) to the ACE context.
3. At the ACE context, client traffic is load balanced to the server in the server farm that will handle the client request.
4. The return traffic from the web server will be forwarded to the ACE context.
5. The ACE context will forward the client to its gateway, which is an HSRP VIP address on the Nexus 7000 aggregation router.

Web—APP VIP Traffic Path

1. Traffic destined to the app VIP from a web server, will be forwarded to the web server gateway which is the nexus 7000 aggregation router
2. the nexus 7000 aggregation router will forward this traffic to the ACE context. Both nexus 7000 aggregation router and the ACE Context app interface are in the same VLAN
3. At the ACE context, web server traffic is load balanced to a server in the app server farm.
4. The return traffic from the app server will be sent to the ACE context. This is due to the web server address being translated to a pool on the ACE on the same server subnet as the app servers.

5. The ACE context will send it to the web server that originated the traffic. Return traffic does not reach the nexus 7000 aggregation router.

APP—DB VIP Traffic Path

1. Traffic destined to the DB VIP address from an app server will be forwarded to the app server gateway, which is the Nexus 7000 aggregation router.
2. The Nexus 7000 aggregation router will forward this traffic to the ACE context. Both the Nexus 7000 aggregation router and the ACE context DB interface are in the same VLAN.
3. At the ACE context, app server traffic is load balanced to a server in the DB server farm.
4. The return traffic from the DB server will be sent to the ACE context. This is due to the app server address being translated to a pool on the ACE on the same server subnet as the DB servers.
5. The ACE context will send this traffic to the app server that originated the traffic. Return traffic does not reach the Nexus 7000 aggregation router.

Table 5-1 and Table 5-2 provide an overview of Gold and Silver client traffic that will be load balanced. Note that L7 class-maps are used to match allowed HTTP URL and client source address.

Table 5-1 Allowed SLB Traffic for Gold Tenant

Traffic Origination	Destination	Operation	Restriction Mode
Any	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL
Any	DMZ VIP (DMZ context)	Load balance to DMZ server farm	L7 class-map based on HTTP URL
Any	App VIP	Reset/Drop Connection	L7 class-map based on HTTP URL and Source IP
Any	DB VIP	Reset/Drop Connection	L7 class-map based on HTTP URL and Source IP
Web tier	App VIP	Load balance to app server farm	L7 class-map based on HTTP URL and Source IP
Web tier	DB VIP	Reset	L7 class-map based on HTTP URL and Source IP
App tier	DB VIP	Load balance to DB server farm	L7 class-map based on HTTP URL and Source IP
App tier	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL
DB tier	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL

Table 5-2 Allowed SLB Traffic for Silver Tenant

Traffic Origination	Destination	Operation	Restriction Mode
Any	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL
Any	App VIP	Reset/Drop Connection	L7 class-map based on HTTP URL and Source IP
Any	DB VIP	Reset/Drop Connection	L7 class-map based on HTTP URL and Source IP

Table 5-2 Allowed SLB Traffic for Silver Tenant (continued)

Web tier	App VIP	Load balance to app server farm	L7 class-map based on HTTP URL and Source IP
Web tier	DB VIP	Reset	L7 class-map based on HTTP URL and Source IP
App tier	DB VIP	Load balance to DB server farm	L7 class-map based on HTTP URL and Source IP
App tier	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL
DB tier	Web VIP	Load balance to web server farm	L7 class-map based on HTTP URL

Refer to Associating a Layer 7 SLB Policy Map with a Layer 3 and Layer 4 SLB Policy Map for additional information on ACE SLB configuration.

ASA Perimeter Firewall

This section presents the following topics:

- ASA Firewall Redundancy, page 5-11
- Gold Tenant ASA Firewall Configuration, page 5-13
- Copper Firewall Details, page 5-16

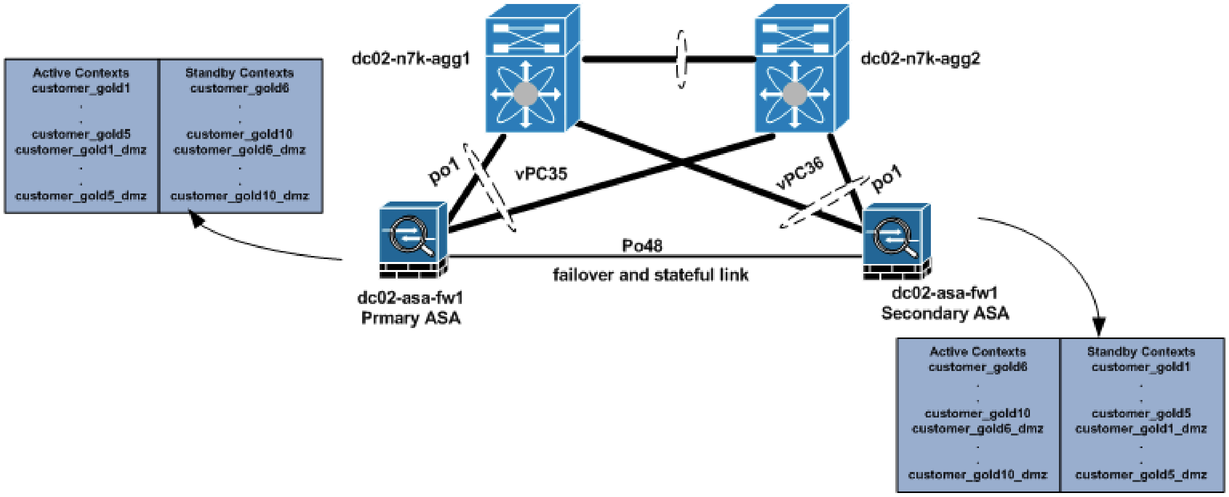
ASA Firewall Redundancy

In this implementation, two ASA firewalls are used to provide Gold tenants with security services, e.g., inspection, ACL, NAT, etc., and these firewalls are configured in active/active redundancy mode to maximize their capability. Separate, dedicated interfaces are used for failover and stateful failover interfaces between the ASA firewall. For more information on how to set up ASA firewall redundancy, refer to the following links:

- Configuring Active/Active Redundancy
- VMDC 2.2 Implementation Guide

ASA port-channels are used to connect to the Nexus 7000 aggregation switch vPC. The data VLANs used for communication through the ASA are trunked on these interfaces. To protect against single vPC failures, interfaces allocated to firewall tenants should be monitored. This ensures that if a failure occurs, the failure policy condition also occurs. For effective load balancing and redundancy, the tenants' contexts are distributed among the two ASA firewalls used in this solution (Table 5-2).

Figure 5-5 ASA Firewall Setup



Below is the ASA sample firewall failover configuration.

```
failover
failover lan unit primary
failover lan interface FL Port-channel48.4093
failover polltime unit 5 holdtime 15
failover replication http
failover link SL Port-channel48.4094
failover interface ip FL 9.9.9.1 255.255.255.0 standby 9.9.9.2
failover interface ip SL 9.9.10.1 255.255.255.0 standby 9.9.10.2
failover group 1
  preempt
  replication http
  polltime interface 1 holdtime 5
  interface-policy 50%
failover group 2
  secondary
  preempt
  replication http
  polltime interface 1 holdtime 5
  interface-policy 50%

ASA Failover Configuration Required on context
----------------------------------------------
dc02-asa-fw1/customer-gold1# sh run monitor-interface
monitor-interface inside
monitor-interface outside
monitor-interface dmz
```

Two failover groups on the ASA are used to distribute active contexts among the primary ASA and the secondary ASA. By default, failover group 1 is assigned to the primary ASA. To have active contexts on the secondary ASA, failover group 2 is assigned to the secondary ASA. To distribute contexts on both ASA devices, half of all configured Gold contexts are assigned to failover group 1, and the other half are assigned to failover group 2. A sample configuration for two Gold tenants is shown below.

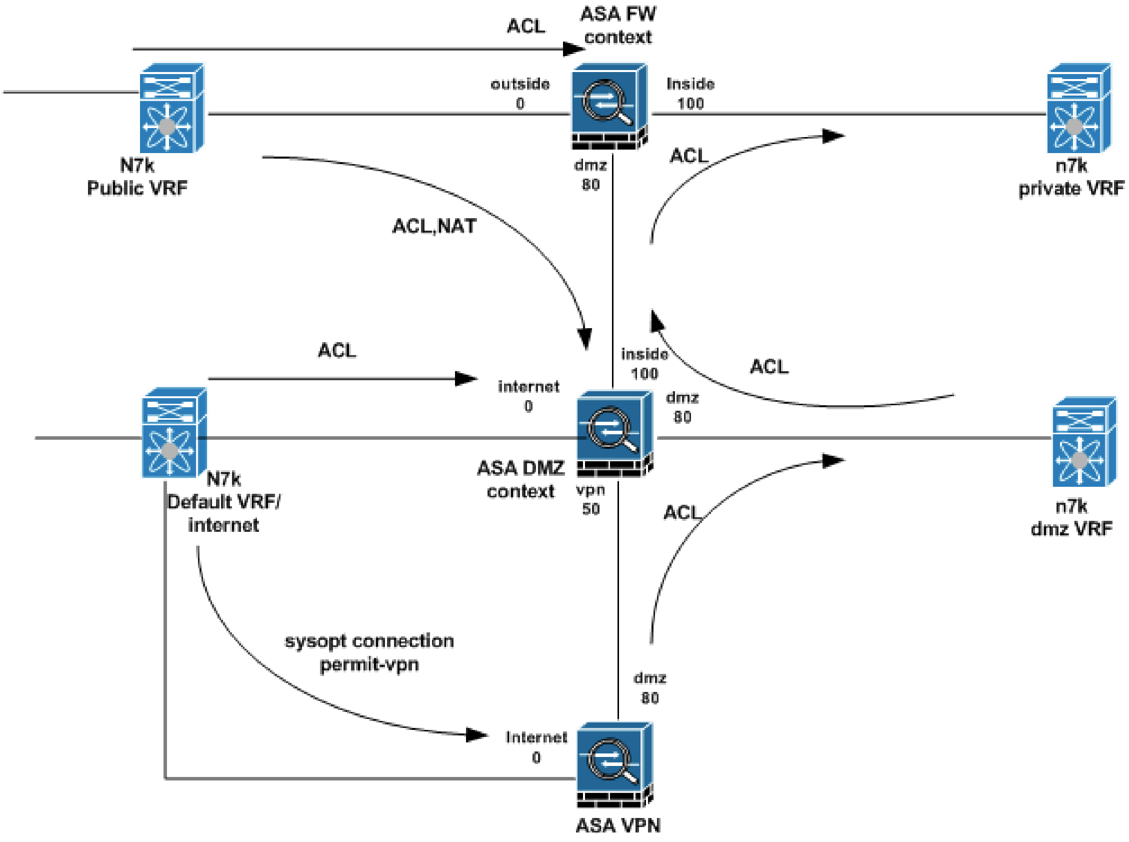
ASA Context Configuration

```
dc02-asa-fw1# sh run context customer-gold1
context customer-gold1
  allocate-interface Management0/0
  allocate-interface Port-channel1.1201
  allocate-interface Port-channel1.1301
  allocate-interface Port-channel1.1401
  config-url disk0:/vmdc3.1/customer-gold1
  join-failover-group 1
!

dc02-asa-fw1# sh run context customer-gold6
context customer-gold6
  allocate-interface Management0/0
  allocate-interface Port-channel1.1206
  allocate-interface Port-channel1.1306
  allocate-interface Port-channel1.1406
  config-url disk0:/vmdc3.1/customer-gold6
  join-failover-group 2
!
```

Gold Tenant ASA Firewall Configuration

Figure 5-6 provides an overview of a typical ASA firewall configuration for a Gold tenant in this implementation.

Figure 5-6 Gold Tenant ASA Firewall Configuration Overview



Routing Configuration

Refer to ASA Firewall Context Routing Configuration for the configuration required to route through the ASA. To route between the Private and DMZ contexts for a tenant, the **mac-address auto prefix <16-bit prefix>** command is configured to ensure that all active interfaces on the ASA have a unique MAC address. This configuration is required for inter-context routing on the ASA.

ACL Configuration

ACLs are configured on the Gold tenant ASA firewall context outside interfaces to allow permitted protocol traffic from Service Provider client networks to be forwarded to the inside and DMZ networks. In this implementation, an object group is used to simplify the ACL configuration. Two object-group types are used in the ACL configuration. The network-object group is used to identify the networks to be allowed, while the service-object group is used to identify the UDP and TCP ports that are allowed for these networks. Also, ACLs are applied on the context DMZ interfaces to identify the DMZ server traffic that should be allowed to the private server networks. ACLs are also configured in the DMZ firewall contexts for the tenants. These ACLs control allowed traffic from Internet to DMZ networks.

A sample configuration of a configured object group and ACL is shown below.

```
dc02-asa-fw1/customer-gold1# sh run object-group
object-group network SP-CLIENTS-NETWORK
 network-object 40.1.0.0 255.255.0.0
 network-object 10.1.0.0 255.255.0.0
 network-object 131.0.0.0 255.0.0.0
object-group service SP-CLIENTS-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ftp
 port-object eq ssh
 port-object eq domain
object-group service SP-CLIENTS-PROTOCOLS-UDP udp
 port-object eq tftp
 port-object eq domain
 port-object range 10000 30000
object-group network DMZ-VPN-NETWORK
 network-object 11.1.4.0 255.255.255.0
 network-object 11.255.0.0 255.255.0.0
object-group service DMZ-VPN-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ssh
 port-object eq ftp
object-group service DMZ-VPN-PROTOCOLS-UDP udp
 port-object eq tftp
 port-object eq domain
 port-object range 10000 30000
dc02-asa-fw1/customer-gold1# sh run access-list
access-list DMZ-VPN extended permit tcp object-group DMZ-VPN-NETWORK any object-group
DMZ-VPN-PROTOCOLS-TCP
access-list DMZ-VPN extended permit udp object-group DMZ-VPN-NETWORK any object-group
DMZ-VPN-PROTOCOLS-UDP
access-list DMZ-VPN extended permit icmp object-group DMZ-VPN-NETWORK any
access-list OUTSIDE extended permit tcp object-group SP-CLIENTS-NETWORK any
object-group SP-CLIENTS-PROTOCOLS-TCP
access-list OUTSIDE extended permit udp object-group SP-CLIENTS-NETWORK any
object-group SP-CLIENTS-PROTOCOLS-UDP
access-list OUTSIDE extended permit icmp object-group SP-CLIENTS-NETWORK any
dc02-asa-fw1/customer-gold1# sh run access-group
access-group OUTSIDE in interface outside
access-group DMZ-VPN in interface dmz
dc02-asa-fw1/customer-gold1#
```

NAT Configuration

Due to the use of static default routes on tenant contexts, dynamic NAT is configured on the private tenant contexts. To enable the DMZ context, know how to forward return traffic for Service Provider clients from the DMZ networks. This dynamic NAT translates the source IP addresses of Service Provider clients whose traffic is destined to DMZ server network. Static NAT configuration is also added in the DMZ context to translate IP addresses of DMZ resources to global addresses. Traffic sent to these resources from the Internet must be destined to their global IP addresses. Network objects are used to identify addresses to be translated and the pool or public IP to use during translation. A sample NAT configuration is shown below.

Dynamic NAT Configuration

```
dc02-asa-fw1/customer-gold1#
dc02-asa-fw1/customer-gold1# sh run object
object network SP-CLIENTS-POOL
 range 51.1.1.1 51.1.1.254
object network SP-CLIENTS->DMZ
 range 0.0.0.0 255.255.255.255
dc02-asa-fw1/customer-gold1# sh run nat
!
```

Static NAT configuration

```
dc02-asa-fw1/customer-gold1# changeto c customer-gold1-dmz
dc02-asa-fw1/customer-gold1-dmz# sh run object
object network SERVER1
 host 11.1.4.11
object network SERVER3
 host 11.1.4.13
object network SERVER2
 host 11.1.4.12
object network WEB-VIP
 host 11.1.4.111
object network t1
object network SERVER8
 host 11.1.4.100
object network SERVER7
 host 11.1.4.151
dc02-asa-fw1/customer-gold1-dmz# sh run nat
!
object network SERVER1
 nat (dmz,internet) static 100.200.2.24
object network SERVER3
 nat (dmz,internet) static 100.200.2.25
object network SERVER2
 nat (dmz,internet) static 100.200.2.26
object network WEB-VIP
 nat (dmz,internet) static 100.200.2.1
object network SERVER8
 nat (dmz,internet) static 100.200.2.31
object network SERVER7
 nat (dmz,internet) static 100.200.2.30
dc02-asa-fw1/customer-gold1-dmz#
```

Application Inspection

The ASA context default inspection policy is used in this implementation. By default, this inspection policy is implicitly applied to all active interfaces configured on an ASA context.

Copper Firewall Details

The Copper service is used for Internet users to access servers in the DC. Those servers can have public or private IP addresses. For the private IP address servers, NAT is needed for the access to these servers, and the IP addresses can be overlapped for different tenants.

This section presents the following topics:

- Shared Firewall Setup, page 5-16
- NAT Setup, page 5-19

Shared Firewall Setup

The Copper tenants' traffic comes from the Internet, and all of the Copper tenants share the same ASA context. This is a use case for a tenant to have VMs in the Service Provider public cloud offering and the VMs are accessed over the Internet. The IP addressing on the tenant VMs can be from routable public IP addressing space or can be private addressing. In the public addressing, these are reachable directly from the Internet, and each tenant would have a subnet for VMs, that is part of the Service Provider's block. In the private addressing scenario, the tenant VMs use a private subnet, and NAT is done on the ASA to translate to a set of public IP addresses. With a private addressing model, overlapping subnets can be used by tenant VMs, however, the public addresses on the outside need to be unique, and NAT translations need to be used to reach the correct inside interface.

Additionally, the inside interface on the ASA connecting to the inside per-tenant VRF instance has a private subnet, however, these subnets cannot be overlapping as only one context is used on the ASA. If overlapping addresses on connected interfaces are required, then different contexts on the ASA need to be used.

This section presents the following topics:

- Public Address Tenants, page 5-16
- Private Address Tenants, page 5-17

Public Address Tenants

For the tenant servers with the public address, NAT is not needed. Static routes are defined in the ASA SMB context to select the egress interface and the next hop. In the Nexus 7000, tenant VRF instances are created to separate the tenants.

From the ASA View

For the north to south traffic, the next hop is the HSRP addresses of the individual tenant VRF instance of the Nexus 7000. For the south to north traffic, the next hop is the HSRP address of a global VLAN, and it uses the global routing table of the Nexus 7000.

From the Nexus 7000 Global Routing View

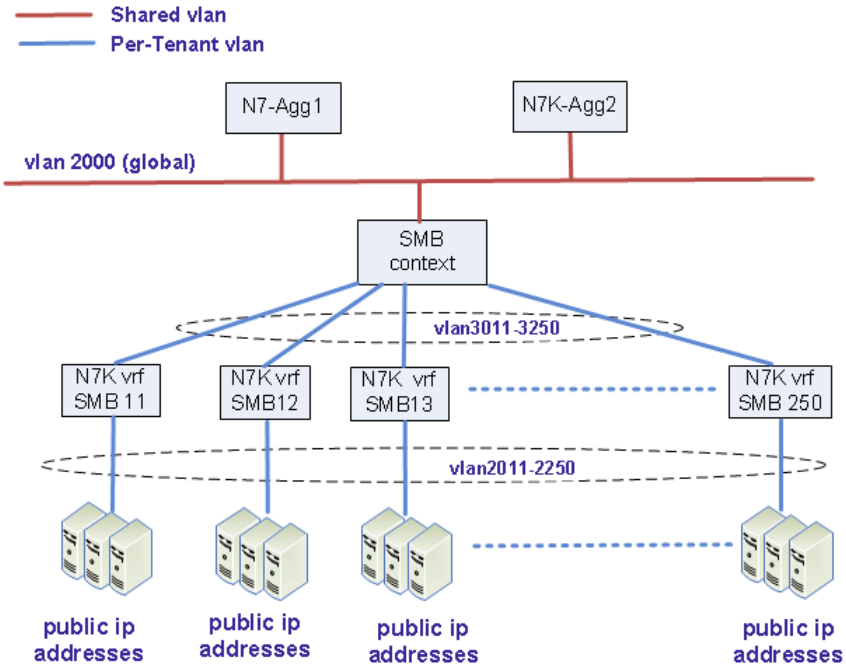
For the north to south traffic, the next hop is the shared ASA context outside interface. For the south to north traffic, the Nexus 7000 will use its routing table to route the traffic to the PE routers.

From the Nexus 7000 Per-Tenant VRF View

For the north to south traffic, the next hop is the server. For the south to north traffic, the next hop is the shared ASA context per-tenant-inside interface

Figure 5-7 shows the public address tenants.

Figure 5-7 Public Address Tenants



Tenants 11 to 250 are public address tenants.

Below are examples of the SMB configuration for SMB tenant 11 and tenant 12.

```
interface Port-channel1.2000
 nameif outside
 security-level 0
 ip address 100.200.1.61 255.255.255.0 standby 100.200.1.62

interface Port-channel1.3011
 nameif smb11
 security-level 100
 ip address 10.9.11.61 255.255.255.0 standby 10.9.11.62
!
interface Port-channel1.3012
 nameif smb12
 security-level 100
 ip address 10.9.12.61 255.255.255.0 standby 10.9.12.62

route smb11 100.201.11.0 255.255.255.0 10.9.11.1 1
route smb12 100.201.12.0 255.255.255.0 10.9.12.1 1
```

Below are useful commands.

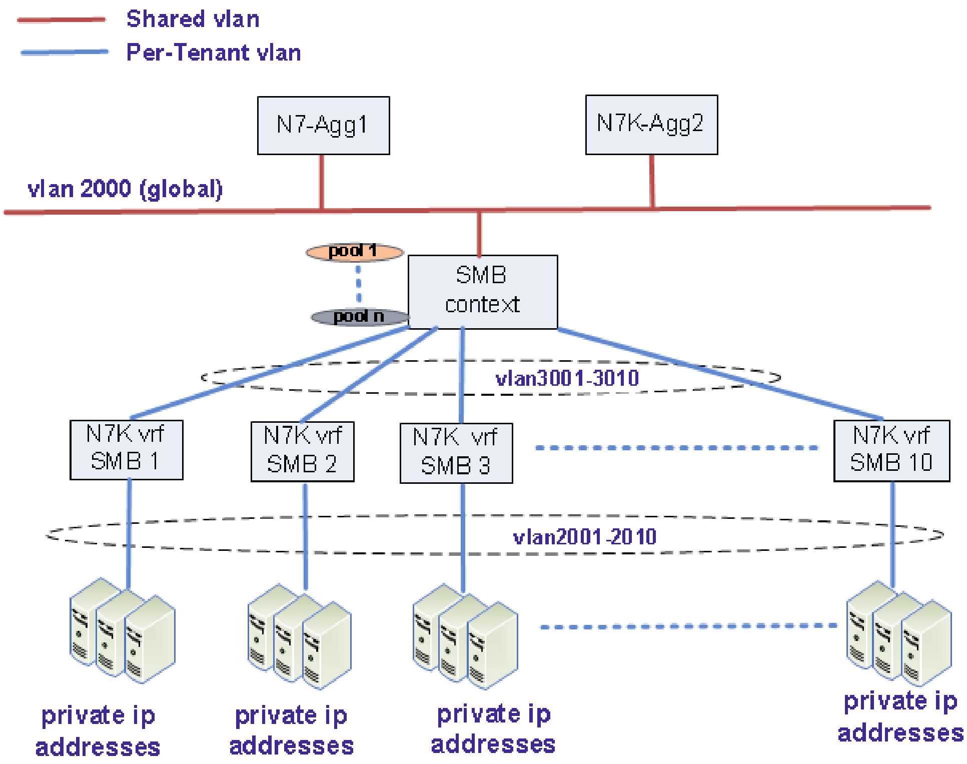
```
show route
show conn
show interface
show interface ip brief
```

Private Address Tenants

Refer to Figure 5-8 for the topology of private address tenants.

Tenants 1 to 10 shown in the diagram are private address tenants. In the Nexus 7000, tenant VRF instances are created to separate the tenants. The IP addresses for tenant VMs can be reused in different tenants - allows for overlapping IP addresses for tenant VM subnets for different tenants. The connected interface on the ASA itself however has to be uniquely addressed, and is in private space as well as no need to be able to route to it.

Figure 5-8 Private Address Tenants



The servers of these tenants have private IP addresses in the DC and have public IP addresses (mapped IP addresses) in the ASA SMB context. For these servers, NAT and static routes are needed for access to these servers from the Internet as well as these servers initiating traffic to the Internet. The server IP addresses can be overlapped for different tenants, but it is not recommended.

From the ASA View

For the north to south traffic, static NAT is used to direct the traffic to the egress interface. Static routes are used to find the next hop. The next hop to the real server is the HSRP address of the individual Nexus 7000 tenant. The return traffic will be NAT'd to the public IP address in the SMB context and points to the HSRP address of the Nexus 7000 VLAN 2000 (global routing table) as the next hop.

For the traffic initiated by the servers (other than the configured static NAT servers) in the DC, the per-tenant dynamic pool in the ASA SMB context is used for NAT. The next hop is the HSRP address of the Nexus 7000 VLAN 2000. For the returned traffic, NAT will decide the egress interface, and static routes are used to find the next hop.

From the Nexus 7000 Global Routing View

For the north to south traffic, the next hop is the shared ASA context outside interface. For the south to north traffic, the Nexus 7000 will use its routing table to route the traffic to the PE routers.

From the Nexus 7000 Per-Tenant VRF View

For the north to south traffic, the next hop is the server. For the south to north traffic, the next hop is the shared ASA context per-tenant-inside interface.

See NAT Setup for detailed NAT setup information.

NAT Setup

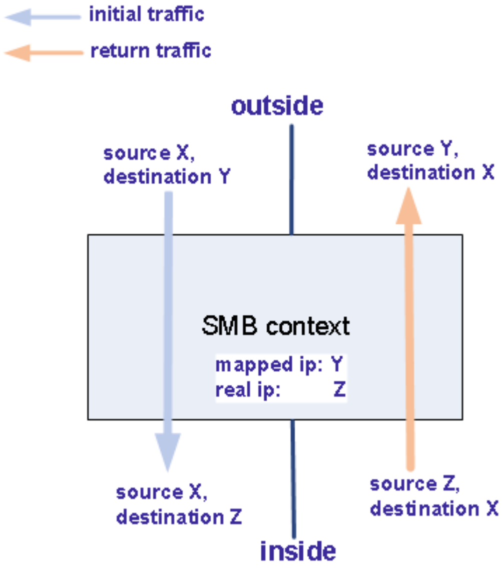
This section presents the following topics:

- Static NAT, page 5-19
- Dynamic NAT, page 5-21
- Interconnecting Tenants, page 5-22

Static NAT

Static NAT creates a fixed translation of a real address to a mapped address. Because the mapped address is the same for each consecutive connection, static NAT allows bidirectional connection initiation, both to and from the server (Figure 5-9).

Figure 5-9 Static NAT



Refer to Information About Static NAT with Port Address Translation for detailed information about static NAT.

In this solution, static NAT is used for access to the private IP address servers from the Internet. For tenant 1, the related configuration on the ASA is shown below.

```
interface Port-channel1.2000
 nameif outside
 security-level 0
 ip address 100.200.1.61 255.255.255.0 standby 100.200.1.62
!
interface Port-channel1.3001
 nameif smb1
 security-level 100
 ip address 10.9.1.61 255.255.255.0 standby 10.9.1.62
```

```
object network smb-1-server1
 host 11.4.1.11
object network smb-1-server2
 host 11.4.1.12
object network smb-1-server3
 host 11.4.1.13
object network smb-1-server21
 host 11.4.1.21
object network smb-1-server22
 host 11.4.1.22
object network smb-1-server23
 host 11.4.1.23
object network smb-1-server24
 host 11.4.1.24
object network smb-1-server25
 host 11.4.1.25
object network smb-1-server26
 host 11.4.1.26
object network smb-1-server27
 host 11.4.1.27
object network smb-1-server28
 host 11.4.1.28
object network smb-1-server29
 host 11.4.1.29
object network smb-1-server30
 host 11.4.1.30

object network smb-1-server1
 nat (smb1,outside) static 100.201.1.11
object network smb-1-server2
 nat (smb1,outside) static 100.201.1.12
object network smb-1-server3
 nat (smb1,outside) static 100.201.1.13
object network smb-1-server21
 nat (smb1,outside) static 100.201.1.21
object network smb-1-server22
 nat (smb1,outside) static 100.201.1.22
object network smb-1-server23
 nat (smb1,outside) static 100.201.1.23
object network smb-1-server24
 nat (smb1,outside) static 100.201.1.24
object network smb-1-server25
 nat (smb1,outside) static 100.201.1.25
object network smb-1-server26
 nat (smb1,outside) static 100.201.1.26
object network smb-1-server27
 nat (smb1,outside) static 100.201.1.27
object network smb-1-server28
 nat (smb1,outside) static 100.201.1.28
object network smb-1-server29
 nat (smb1,outside) static 100.201.1.29
object network smb-1-server30
 nat (smb1,outside) static 100.201.1.30

route outside 0.0.0.0 0.0.0.0 100.200.1.1 1
route smb1 11.4.1.0 255.255.255.0 10.9.1.1 1
```

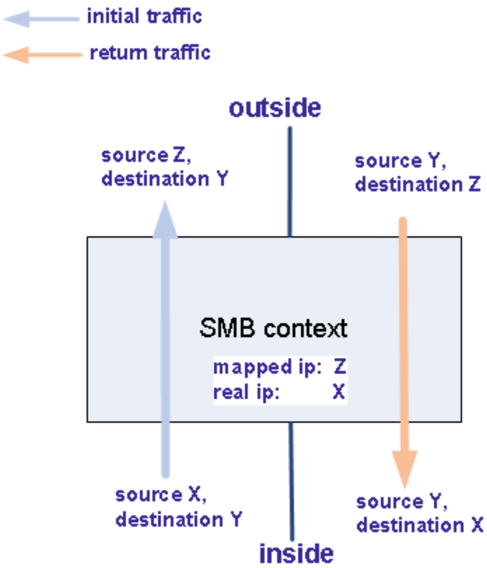
Below are useful commands.

```
show xlate
show route
show conn
```

Dynamic NAT

Dynamic NAT translates a group of real addresses to a pool of mapped addresses that are routable on the destination network. The translation is created only when the real host initiates the connection (Figure 5-10).

Figure 5-10 Dynamic NAT



Refer to Dynamic NAT for a detailed dynamic NAT explanation and configuration guide.

In this solution, dynamic NAT is used for the traffic initiated from the servers (other than the servers already configured as static NAT) in the DC.

Below is the configuration for the tenant SMB1.

```
interface Port-channel1.2000
 nameif outside
 security-level 0
 ip address 100.200.1.61 255.255.255.0 standby 100.200.1.62
!
interface Port-channel1.3001
 nameif smb1
 security-level 100
 ip address 10.9.1.61 255.255.255.0 standby 10.9.1.62

object network smb1-mapped
 range 100.201.1.1 100.201.1.10

object network smb1-real
 subnet 11.4.1.0 255.255.255.0

object network smb1-real
 nat (smb1,outside) dynamic smb1-mapped
```

Below are useful commands.

```
show xlate
show route
show conn
```

Interconnecting Tenants

By default, interfaces on the same security level cannot communicate with each other. To enable the communication between the same security level, the **same-security-traffic permit inter-interface** command must be configured. With the **same-security-traffic permit inter-inerface** command enabled, complicated ACLs are placed between tenants for the security. In this implementation, tenants are not allowed to talk to each other inside the ASA, and the **same-security-traffic permit inter-interface** command is disabled.

In this implementation, we have private IP address servers and public IP address servers, and we want to block the inter-tenants communications to the public IP addresses servers.

Table 5-3 shows if the inter-tenants communications is possible in this implementation.

Table 5-3 Inter-tenants Communication

	Destination Public	Destination Private
Initiated from Public	No	Yes
Initiated from Private	No	Yes

In general, if the packet is initiated by private or public, but the destination is a public IP address server, the inter-tenant communication will fail. If the packet is initiated by private or public, but the destination is a private IP address server, the inter-tenant communication will succeed. For those inter-tenants, communication is possible. The inter-tenants' traffic is sent to the Nexus 7000 by the ASA, and the Nexus 7000 sends the traffic back to the ASA. We do not recommend inter-tenants communications due to the security considerations and recommend using public IP addresses in all Copper tenants.

These scenarios are discussed in detail below.

For Scenario 1, Public to Public (Fail)

When the packet reaches the ASA from the public IP address server, the ASA will use the routing table to find the egress interface, and the ASA will find the destination SMB tenant interface as the egress interface. As **same-security-traffic permit inter-interface** is disabled, the packet will be dropped.

For Scenario 2, Public to Private (Pass)

The destination of this kind of traffic is a mapped public IP address. When the ASA receives the packet, it will use the routing table to find the egress interface. There is no entry in the routing table for the mapped IP address, and the ASA will use the default route to send this traffic to the Nexus 7000 HSRP address. The Nexus 7000 routers have a static route for this subnet and point to the ASA outside interface as the next hop. When the ASA receives the packet from the outside interface, it will use static NAT to find the egress interface and the routing table to find the next hop. The return traffic will use NAT to direct the traffic to the outside interface of the ASA. After reaching the Nexus 7000, the

Nexus 7000 will send the packet back to the ASA, and the ASA will use the routing table to find the egress interface and the next hop.

For Scenario 3, Private to Public (Fail)

As the destination is a public IP address server, the ASA will use static routes to find the egress, and the egress interface is the SMB tenant inside interface. Since **same-security-traffic permit inter-interface** is disabled, the packet will be dropped.

For Scenario 4, Private to Private (Pass)

If the traffic is initiated by the private IP address servers, (does not matter if it is configured as the static NAT server or other servers), when the packet reaches the ASA, the ASA will look up the routing table, as the destination is a mapped address. The ASA will use the default routes to direct the traffic to the egress interface. The packet will go to the Nexus 7000 and come back as Nexus 7000 routers have static routes pointing to the ASA outside interface as the next hop. When the ASA receives the packet, it will use static NAT to NAT and direct the packet to the egress interface, which is the SMB tenant inside interface. The return traffic will be static NAT'd and directed to the outside interface of the ASA. The traffic will go to the Nexus 7000 and then return. The ASA will use NAT to NAT and direct the packet to the egress interface of the SMB tenant.

ASA VPN Configuration

The ASA 5555-X appliance is used to provide VPN services to the Gold tenants. This appliance is a 4 Gbps firewall device capable of up to 1 Gbps of VPN throughput. Up to 5000 VPN peers can be configured on this appliance. In this implementation, a mix of IPsec and SSL VPN tunnels are established on this appliance.

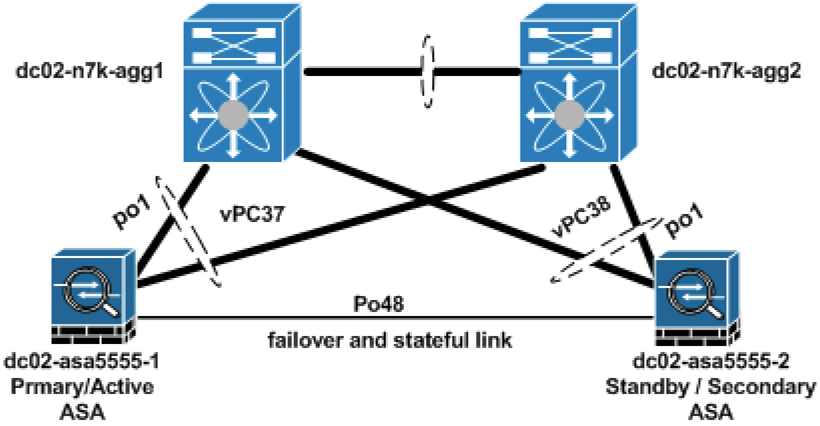
This section presents the following topics:

- ASA VPN Redundancy, page 5-23
- ASA VPN Multitenancy, page 5-24
- ASA IPsec VPN Configuration, page 5-24
- ASA SSL VPN Configuration, page 5-25

ASA VPN Redundancy

The ASA 5555-X providing VPN services and acting as the VPN gateway must be configured in single-context mode and can only be configured in active/standby redundancy mode. In active/standby redundancy mode, only the active ASA is responsible for all VPN gateway functionality. The active ASA terminates all established SSL VPN and IPsec remote access VPN tunnels and maps the client traffic to the appropriate outbound VLANs. Since the active ASA responds to traffic directed to the active IP address, all traffic is directed to the active ASA. All established tunnels in the VPN session database are replicated from the active ASA to the standby ASA. This ensures that the standby ASA can take over the responsibility of the active, should it fail. All state information is sent through the stateful link (Figure 5-11).

Figure 5-11 ASA VPN Redundancy



As with the other ASA used to provide security services to the Gold tenant, the ASA 5555-x are port-channels that are connected to the vPC on a pair of aggregation Nexus 7000 switches.

ASA VPN Multitenancy

The ASA 5555-X providing VPN services for the Gold service class runs in single-context mode. The ASA configured in multiple-context mode cannot be used to provide VPN services. Due to this limitation, the virtualization capability of the ASA cannot be used to differentiate between Gold tenants in this DC environment. To differentiate between these tenants, two VPN configuration options are used in the ASA. These options are group-policy VLAN ID and tunnel-group configuration.

Group-policy VLAN ID

The group policy is a set of attributes that define how users use a connection through a tunnel after tunnel establishment. The group-policy VLAN ID determines which outbound VLAN the decrypted traffic should be placed on. For every tenant, since all clear text traffic is sent to the DMZ vFW, the VLAN group must correspond to the VPN-outside interface of the tenant DMZ vFW.

Tunnel Group

A tunnel group consists of a set of records that determines tunnel connection policies. These records identify the servers to which the tunnel user is authenticated, as well as the accounting servers, if any, to which connection information is sent. They also identify a default group policy for the connection, and they contain protocol-specific connection parameters. Tunnel groups include a small number of attributes that pertain to creating the tunnel itself. Tunnel groups include a pointer to a group policy that defines user-oriented attributes.

ASA IPsec VPN Configuration

IPsec VPN services are provided using IPsec framework, which comprises a suite of protocols that provide confidentiality, authentication, integrity, and advanced security services. Confidentiality is provided using cryptographic algorithms and ciphers. Authentication is provided using hashing algorithms that generate message authentication codes. IPsec uses the following protocols to perform various functions:

- Authentication Headers (AH) provide integrity, authentication, and protection against reply attacks.

- Encapsulating Standard Protocol (ESP) provides confidentiality, authentication, integrity, and anti-replay.
- Security Association (SA) provides the bundle of algorithms and parameters (such as keys) that are used to encrypt and authenticate a particular flow in one direction
- Internet Security Association and Key Management Protocol (ISAKMP) is the negotiation protocol that lets two hosts agree on how to build an IPsec SA. A sample IPsec VPN configuration is shown below.

Cisco ASA platform has a rich set of features used for secure access. More details can be found about new features released in latest release at this link:Cisco ASA Series , Release Notes , 9.0(x).

Sample IPsec VPN Configuration

```
dc02-asa5555-1# sh run crypto
crypto ipsec ikev1 transform-set ipsec-tz esp-3des esp-md5-hmac
crypto ipsec security-association pmtu-aging infinite
crypto dynamic-map ipsec-cm 1 set ikev1 transform-set ipsec-tz
crypto dynamic-map ipsec-cm 1 set security-association lifetime seconds 7200
crypto map ipsec-cm 1 ipsec-isakmp dynamic ipsec-cm
crypto map ipsec-cm interface internet
crypto ca trustpool policy
crypto ikev1 enable internet
crypto ikev1 policy 1
 authentication pre-share
 encryption 3des
 hash md5
 group 2
 lifetime 3600
dc02-asa5555-1# sh run tunnel-group customer_gold1-ipsec
tunnel-group customer_gold1-ipsec type remote-access
tunnel-group customer_gold1-ipsec general-attributes
 address-pool customer_gold1
 authentication-server-group (internet) LOCAL
 authorization-server-group (internet) LOCAL
tunnel-group customer_gold1-ipsec ipsec-attributes
 ikev1 pre-shared-key *****
dc02-asa5555-1# sh run group-policy customer_gold1-ipsec
group-policy customer_gold1-ipsec internal
group-policy customer_gold1-ipsec attributes
 vpn-simultaneous-logins 200
 vpn-tunnel-protocol ikev1
 group-lock value customer_gold1-ipsec
 split-tunnel-policy tunnelspecified
 split-tunnel-network-list value customer_gold1
 vlan 1701
```

ASA SSL VPN Configuration

The ASA 5555-X provides two types of SSL VPN tunnels, with the major difference between the two being what the client uses to establish these tunnels. The ASA provide clientless SSL VPN/WebVPN tunnel services, which enable users to use their supported web browser to set up and establish SSL/ TLS tunnels to the ASA endpoint. Typically, these users gain access to web resources. The ASA also provides AnyConnect SSL VPN tunnels, which enable clients to gain access to full network resources. These clients establish SSL or IPsec tunnels to the ASA using the Cisco AnyConnect Secure Mobility client. In both cases, users are required to authenticate to the respective tunnel groups. If this authentication is successful for the clientless SSL VPN user, that user will be presented with the

Cisco Secure Desktop (CSD) portal. If an AnyConnect user authenticates using a web browser, the AnyConnect Secure Mobility Client installer is pushed from the ASA that matches the user's OS. The AnyConnect SSL VPN tunnel is successfully established if users authenticate with a current version of the AnyConnect Client on the ASA. A sample configuration required to set up SSL VPN tunnels is shown below.

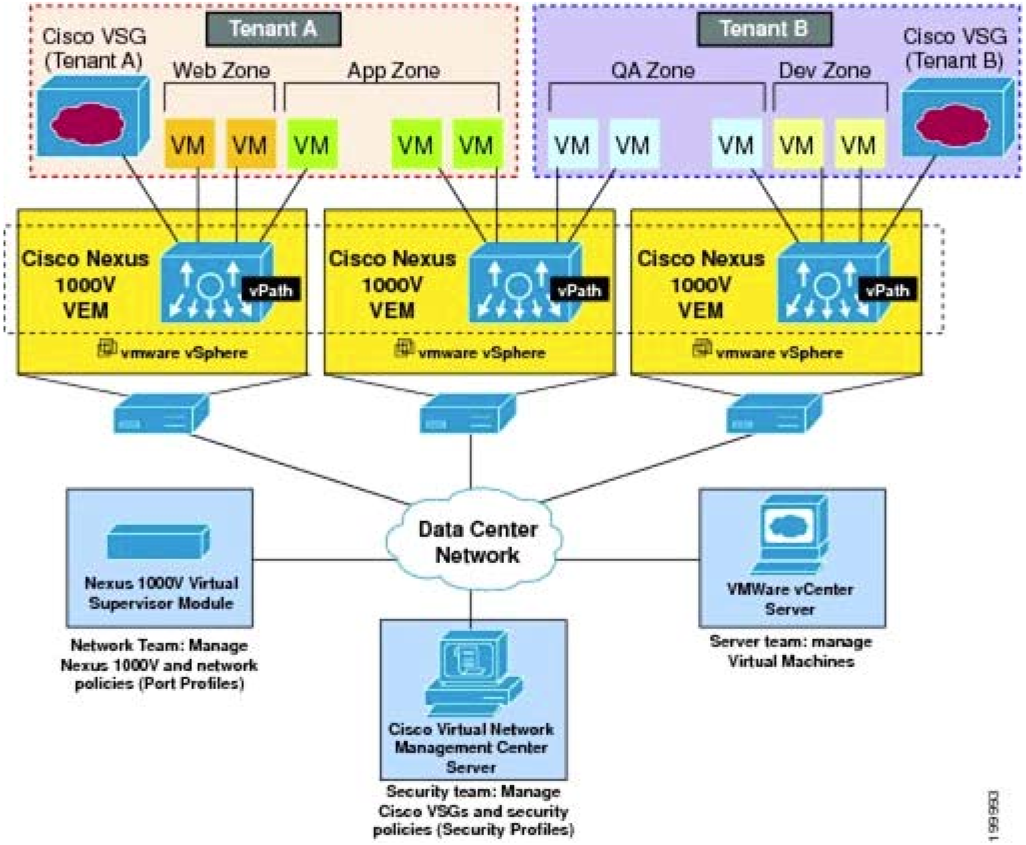
Sample ASA SSL VPN Configuration

```
dc02-asa5555-1# sh run webvpn
webvpn
 enable internet
 no anyconnect-essentials
 csd image disk0:/csd_3.6.6210-k9.pkg
 anyconnect image disk0:/anyconnect-win-3.1.01065-k9.pkg 1
 anyconnect profiles anyconnect-profile disk0:/RDP.xml
 anyconnect enable
 tunnel-group-preference group-url
dc02-asa5555-1#
dc02-asa5555-1#
dc02-asa5555-1# sh run tunnel-group customer_gold1-ssl
tunnel-group customer_gold1-ssl type remote-access
tunnel-group customer_gold1-ssl general-attributes
 address-pool customer_gold1
 authentication-server-group (internet) LOCAL
 authorization-server-group (internet) LOCAL
tunnel-group customer_gold1-ssl webvpn-attributes
 group-url https://100.200.1.51/customer_gold1 enable
dc02-asa5555-1#
dc02-asa5555-1# sh run group-policy customer_gold1-ssl
group-policy customer_gold1-ssl internal
group-policy customer_gold1-ssl attributes
 vpn-simultaneous-logins 200
 vpn-tunnel-protocol ssl-client ssl-clientless
 group-lock value customer_gold1-ssl
 split-tunnel-policy tunnelspecified
 split-tunnel-network-list value customer_gold1
 vlan 1701
 webvpn
  anyconnect profiles value anyconnect-profile type user
dc02-asa5555-1#
dc02-asa5555-1# sh run user ssl1
username ssl1 password JSKNK4oromgGd3D9 encrypted
username ssl1 attributes
 vpn-group-policy customer_gold1-ssl
```

Compute Firewall

In this implementation, the VSG acts as the compute firewall. The VSG is a virtual firewall appliance that provides trusted access to virtual DC and cloud environments with dynamic policy-driven operation, mobility-transparent enforcement, and scale-out deployment for dense multitenancy. The VSG operates with the Nexus 1000V DVS in the VMware vSphere hypervisor. The VSG leverages the virtual network service data path (vPath) that is embedded in the Nexus 1000V VEM. Figure 5-12 shows a typical VSG deployment topology.

Figure 5-12 VSG Deployment Topology


Cisco VSG (Tenant A)
Tenant A
Web Zone
App Zone
VM
Tenant B
QA Zone
Dev Zone
Cisco VSG (Tenant B)
Cisco Nexus 1000V VEM
vPath
vmware vSphere
Data Center Network
Nexus 1000V Virtual Supervisor Module
Network Team: Manage Nexus 1000V and network policies (Port Profiles)
VMWare vCenter Server
Server team: manage Virtual Machines
Cisco Virtual Network Management Center Server
Security team: Manage Cisco VSGs and security policies (Security Profiles)
199993

This section presents the following topics:

- VSG Deployment, page 5-27
- Organization Structure, page 5-33
- Security Policies, page 5-35
- ASA SSL VPN Configuration, page 5-25

VSG Deployment

This section presents the following topics:

- ESXi Service Cluster, page 5-27
- Virtual Network Management Center, page 5-29
- Virtual Security Gateway, page 5-31

ESXi Service Cluster

The VSG virtual appliances are hosted on ESXi hosts. For this implementation, the VSG appliances are hosted on a cluster of ESXi hosts dedicated for hosting virtual services appliances. Table 5-4 shows the vSphere cluster, ESXi hosts, and blade server assignment for ESXi hosts dedicated for hosting VSG virtual appliances.

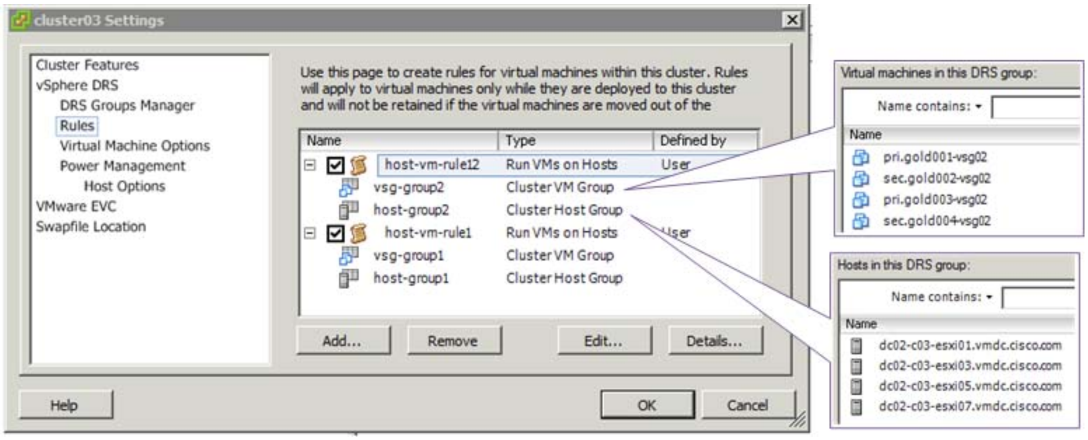
The VSG comes with High Availability (HA). It is not recommended to use the vSphere HA feature, FT feature, or vSphere Distributed Resource Scheduling (DRS) feature for the VSG. For this implementation, vSphere HA is disabled, and vSphere DRS is set to partially automated for initial virtual appliance power on placement only.

Table 5-4 vSphere Cluster, ESXi Hosts, and Blade Server Assignment for ESXi Hosts

vSphere cluster	vSphere DRS	vSphere HA	ESXi Host	UCSM Server Pool
vsg-cluster01	Partially automated	Disabled	dc02-c03- esxi01	cluster03 -
			dc02-c03- esxi02	server 3/1
			dc02-c03- esxi03	server 3/2
			dc02-c03- esxi04	server 3/3
			dc02-c03- esxi05	server 3/4
			dc02-c03- esxi06	server 3/5
			dc02-c03- esxi07	server 3/6
			dc02-c03- esxi08	server 3/7
				server 3/8

For this implementation, each VSG is an HA-pair of primary and secondary nodes. For HA, the primary and secondary nodes of each VSG should not be hosted on the same ESXi host. vSphere DRS groups are configured for this purpose. Figure 5-13 shows the vSphere DRS groups' configuration.

Figure 5-13 vSphere DRS Groups and Rules for VSG Virtual Appliances



Use this page to create rules for virtual machines within this cluster. Rules will apply to virtual machines only while they are deployed to this cluster and will not be retained if the virtual machines are moved out of the

Name	Type	Defined by
⊟ ☑ host-vm-rule12	Run VMs on Hosts	User
vsg-group2	Cluster VM Group	
host-group2	Cluster Host Group	
⊟ ☑ host-vm-rule1	Run VMs on Hosts	User
vsg-group1	Cluster VM Group	
host-group1	Cluster Host Group	

Virtual machines in this DRS group:

Name
pri.gold001-vsg02
sec.gold002-vsg02
pri.gold003-vsg02
sec.gold004-vsg02

Hosts in this DRS group:

Name
dc02-c03-esxi01.vmdc.cisco.com
dc02-c03-esxi03.vmdc.cisco.com
dc02-c03-esxi05.vmdc.cisco.com
dc02-c03-esxi07.vmdc.cisco.com

Two host DRS groups are configured, with each host group containing half of the ESXi hosts in the cluster. Two VMs' DRS groups are also configured; for each VSG HA-pair, the primary node is placed in one VMs, group, while the secondary node is placed in the other group. Each VMs, group is configured to include both primary and secondary nodes (from different VSG HA-pair) to ensure even load on the ESXi hosts.

Four SAN data stores are made available for hosting the VSG virtual appliances, and all ESXi hosts in the cluster have access to all four data stores. The primary and secondary nodes of each VSG should not be placed on the same data store.

Each VSG node (be it the active or standby node of the VSG HA-pair) reserves CPU and memory resources on the ESXi host. On vSphere, a reservation specifies the guaranteed minimum allocation of resources for a VM. The ESXi host performs reservation admission control. If a VM has a reservation defined, the ESXi host must have at least that much resource unreserved (not just unused, but unreserved), or else it will refuse to power on the VM. The ESXi host guarantees the availability of the reserved resources even when the physical server is heavily loaded.

Note As of VSG version 4.2(1)VSG1(4.1), the resources reservation of each VSG node is as follows:

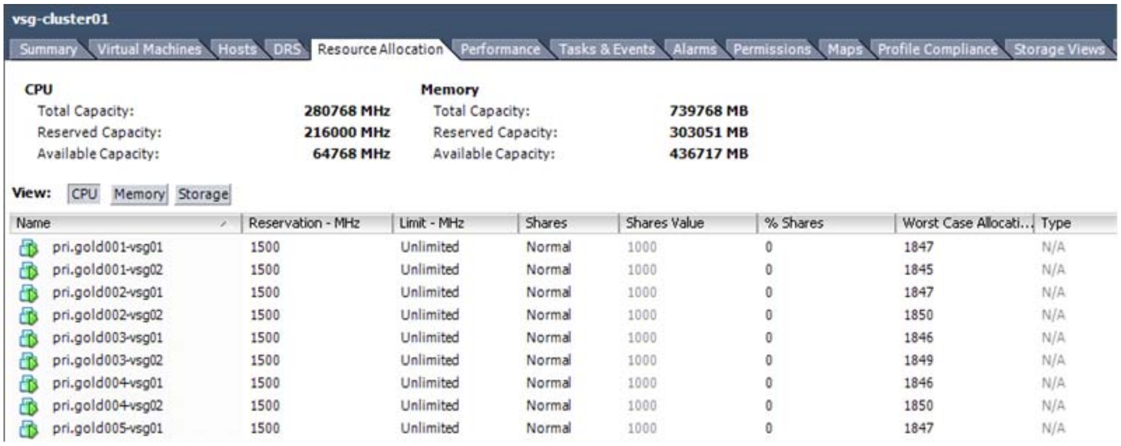
- CPU - 1500 MHz per vCPU
- Memory - 2048 MB

Resources reservation placed a limit on the number of powered on VSG nodes that each ESXi host can support. The number of ESXi hosts required for hosting VSG appliances depends on the following:

- CPU (number of CPU, number of cores per CPU) and memory (amount of DRAM installed) of the ESXi hosts
- Number of VSG appliances deployed

Figure 5-14 shows the resource allocation of the **vsg-cluster01** cluster. This figure shows the total, reserved, and available capacities of CPU and memory for all ESXi hosts in the cluster.

Figure 5-14 vSphere Resource Allocation of a Cluster



vsg-cluster01

CPU		Memory	
Total Capacity:	280768 MHz	Total Capacity:	739768 MB
Reserved Capacity:	216000 MHz	Reserved Capacity:	303051 MB
Available Capacity:	64768 MHz	Available Capacity:	436717 MB

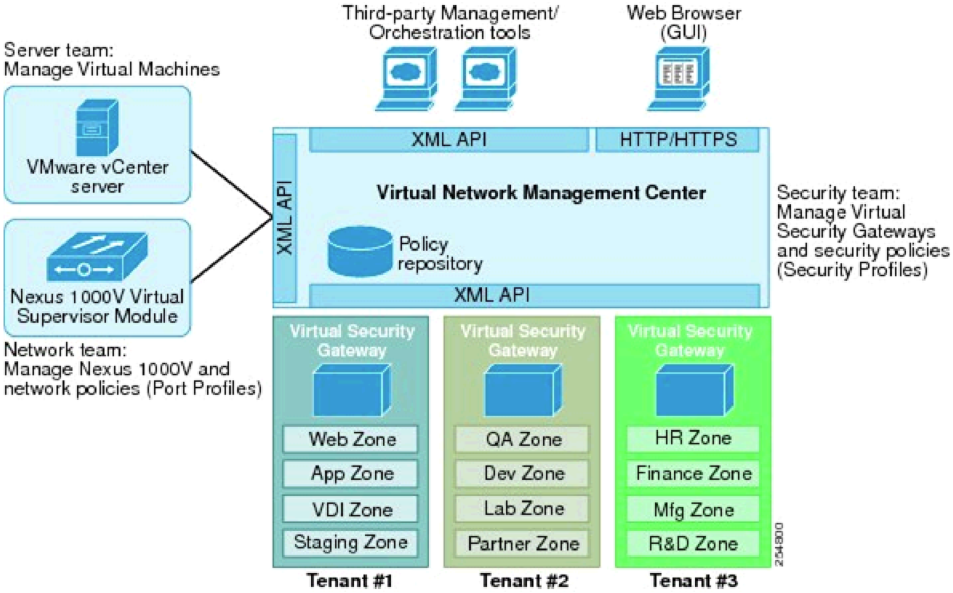
Name	Reservation - MHz	Limit - MHz	Shares	Shares Value	% Shares	Worst Case Allocati...	Type
pri.gold001-vsg01	1500	Unlimited	Normal	1000	0	1847	N/A
pri.gold001-vsg02	1500	Unlimited	Normal	1000	0	1845	N/A
pri.gold002-vsg01	1500	Unlimited	Normal	1000	0	1847	N/A
pri.gold002-vsg02	1500	Unlimited	Normal	1000	0	1850	N/A
pri.gold003-vsg01	1500	Unlimited	Normal	1000	0	1846	N/A
pri.gold003-vsg02	1500	Unlimited	Normal	1000	0	1849	N/A
pri.gold004-vsg01	1500	Unlimited	Normal	1000	0	1846	N/A
pri.gold004-vsg02	1500	Unlimited	Normal	1000	0	1850	N/A
pri.gold005-vsg01	1500	Unlimited	Normal	1000	0	1847	N/A

Virtual Network Management Center

The Cisco Virtual Network Management Center (VNMC) virtual appliance is a custom VM based on Red Hat Enterprise Linux (RHEL). The VNMC provides centralized device and security policy management of the VSG for the Nexus 1000V Series switch. The VNMC manages the VSG appliances that are deployed throughout the DC from a centralized location. For this implementation, the VNMC virtual appliance is deployed on a management vSphere cluster that is not part of the

tenants' compute infrastructure. VNMC does not provide its own HA capability, relying on vSphere HA instead. For proper operations, the VNMC must have management access to vSphere vCenter, Nexus 1000V VSM, and all VSG appliances under its administration. The VNMC registers to vSphere vCenter as a plug-in, and the VSM and VSG appliances register to VNMC via the policy agent. Figure 5-15 shows the VNMC components.

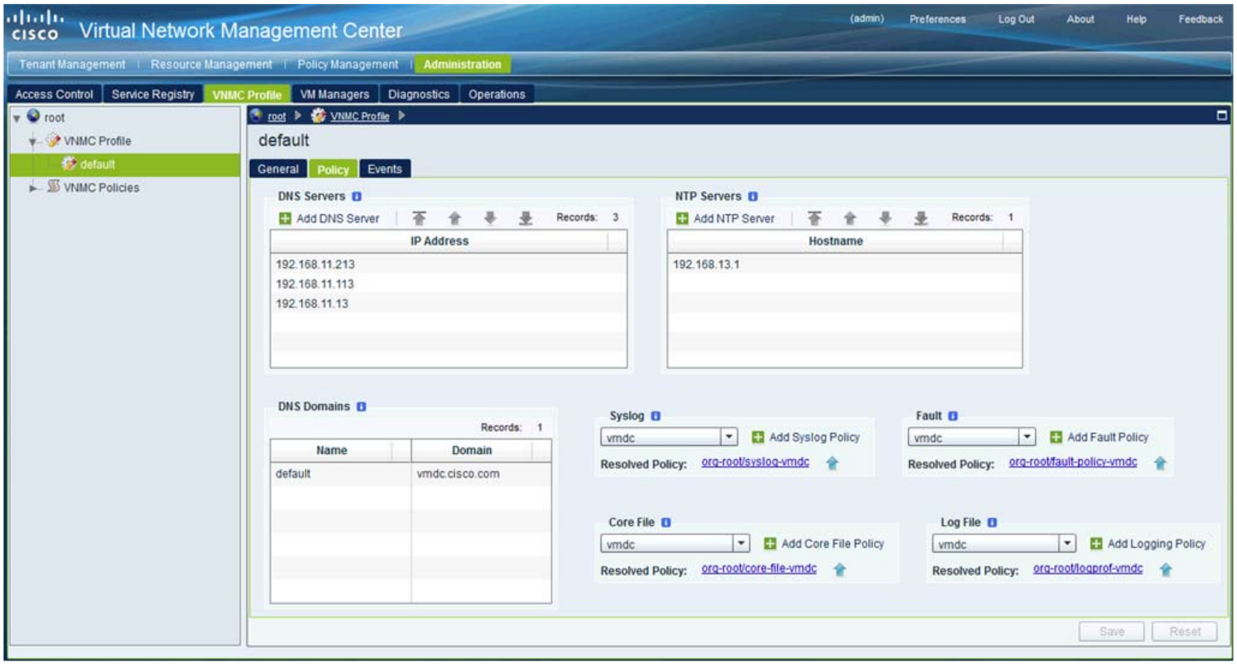
Figure 5-15 VNMC Components


Third-party Management/ Orchestration tools
Web Browser (GUI)
Server team: Manage Virtual Machines
VMware vCenter server
Nexus 1000V Virtual Supervisor Module
Network team: Manage Nexus 1000V and network policies (Port Profiles)
XML API
HTTP/HTTPS
XML API
Virtual Network Management Center
Policy repository
XML API
Security team: Manage Virtual Security Gateways and security policies (Security Profiles)
Virtual Security Gateway
Web Zone
App Zone
VDI Zone
Staging Zone
Tenant #1
Virtual Security Gateway
QA Zone
Dev Zone
Lab Zone
Partner Zone
Tenant #2
Virtual Security Gateway
HR Zone
Finance Zone
Mfg Zone
R&D Zone
Tenant #3
254800

The device specific settings for the VNMC virtual appliance are configured via the VNMC Profile. Figure 5-16 shows a typical VNMC Profile. The following settings are configurable:

- **Time zone**
- **DNS Servers**
- **DNS Domain Name**
- **NTP Servers**—The clock on both VNMC and VSG appliances **must** be in-sync.
- **Syslog Server**
- **Core File Policy**—TFTP server for VNMC to upload any process core file for later analysis.
- **Fault Policy**—Determine if and for how long cleared faults on the VNMC should be kept.
- **Logging Policy**—Logging settings for logging to the local log file on the VNMC.

Figure 5-16 VNMC Profile



default

DNS Servers — Records: 3

IP Address
192.168.11.213
192.168.11.113
192.168.11.13

NTP Servers — Records: 1

Hostname
192.168.13.1

DNS Domains — Records: 1

Name	Domain
default	vmdc.cisco.com

Syslog: vmdc
Resolved Policy: org-root/syslog-vmdc

Fault: vmdc
Resolved Policy: org-root/fault-policy-vmdc

Core File: vmdc
Resolved Policy: org-root/core-file-vmdc

Log File: vmdc
Resolved Policy: org-root/logprof-vmdc

Virtual Security Gateway

The VSG virtual appliance is a VM based on NX-OS. The VSG is a virtual firewall appliance that provides trusted access to virtual DC and cloud environments with dynamic policy-driven operation, mobility-transparent enforcement, and scale-out deployment for dense multitenancy.

For this implementation, the VSG virtual appliances are deployed on the same vSphere and Nexus 1000V infrastructure as the protected VMs. A dedicated set of ESXi hosts in a cluster are used to host the VSG virtual appliances.

Each VSG has three network interfaces:

- Data/Service interface (1st vNIC)
- Management interface (2nd vNIC)
- HA interface (3rd vNIC)

The Data/Service and HA interfaces require their own dedicated VLANs and port profiles. The Management interface can use the existing management VLAN and port profile, however, for this implementation, a dedicated VSG management VLAN and port profile is used.

VSG appliances for each tenant can be provisioned with one VLAN for Data/Service interfaces (for all the VSG appliances of the tenants) and one VLAN for HA interfaces (for all the VSG appliances of the tenants), however, in a deployment with many tenants, such a deployment option would use up many VLAN resources. For this implementation, one shared Data/Service VLAN and one shared HA VLAN are used for all VSG appliances for all tenants. Table 5-5 shows the configuration for the interfaces.

Table 5-5 VSG Interface Configuration

Interface	VLAN ID	Port Profile

Data/Service Interface	54	port-profile type vethernet vsg-data
		  vmware port-group
		  port-binding static
		  switchport mode access
		  switchport access vlan 54
		  service-policy input vsg-data
		  pinning id 0
		  no shutdown
		  system vlan 54
		  max-ports 192
		  state enabled

Management Interface	53	port-profile type vethernet vsg-mgmt
		  vmware port-group
		  port-binding static
		  switchport mode access
		  switchport access vlan 53
		  service-policy input vsg-mgmt
		  pinning id 0
		  no shutdown
		  system vlan 53
		  max-ports 192

HA Interface	55	port-profile type vethernet vsg-ha
		  vmware port-group
		  port-binding static
		  switchport mode access
		  switchport access vlan 55
		  service-policy input vsg-ha
		  pinning id 0
		  no shutdown
		  system vlan 55
		  max-ports 192

Note 1. Licenses for VSG appliances are based on the number of CPUs (regardless of the number of cores on the CPU) of the ESXi hosts hosting the appliances. Each VSG virtual appliance has three vNICs. In an HA setup with a VGS HA-pair, the pair would consume six vNICs.

2. In large scale implementation with a larger number of VGS appliances, the VSG will consume a large number of vEth interfaces on the Nexus 1000V. In such deployments, it would be better to deploy a separate Nexus 1000V dedicated for the VSG virtual appliances. Deploying an additional Nexus 1000V should not require more licenses than when VSG appliances shared the same Nexus 1000V of the tenants' VMs.

Figure 5-17 VSG Network Topology

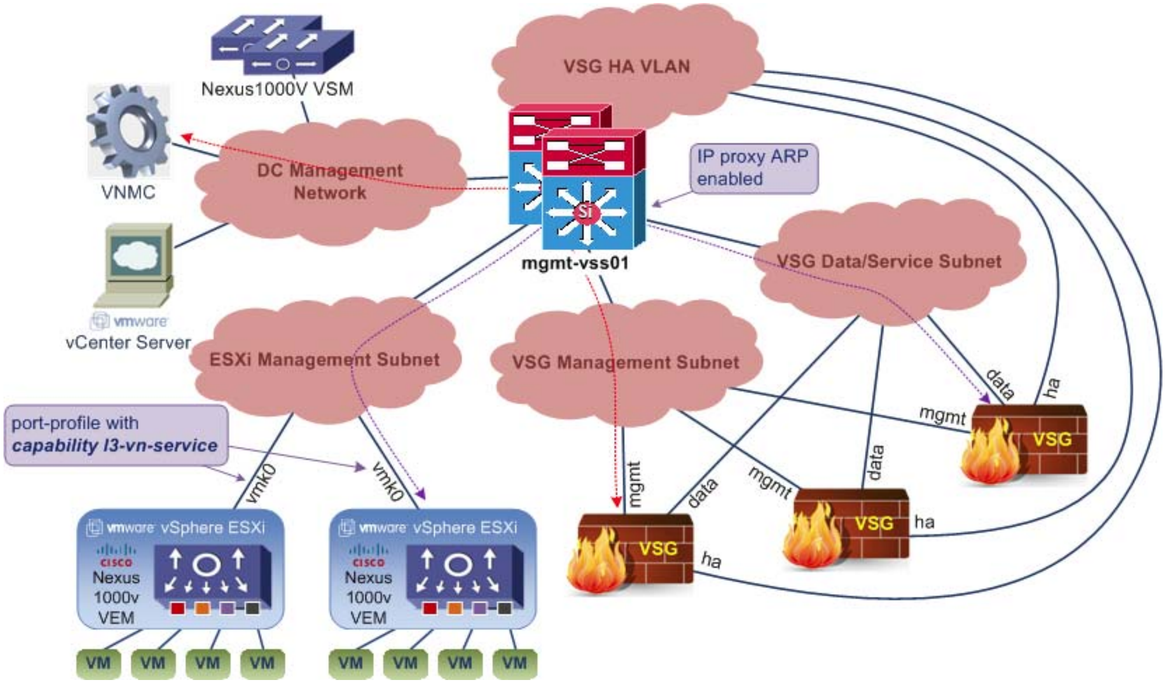


Figure 5-17 shows the network topology for the VSG deployment for this implementation. The VSG appliances are configured to communicate with the vPaths/VEMs in L3 mode. Each vPath/VEM uses the vmk0 interface to communicate with the VGS appliances, via the IP network. The Nexus 1000V port profile that the vmk0 interfaces attach to must be configured with **capability l3-vn-service**.

VSG uses the Management interface to communicate with VNMC, and it uses the Data/Service interface to communicate with the vPath/VEM. On the router interface facing the VSG appliances Data/Service interfaces, IP-proxy Address Resolution Protocol (ARP) must be enabled in order for the VSG appliances to be able to ARP for the IP addresses of the vmk0 interfaces of ESXi hosts.

Organization Structure

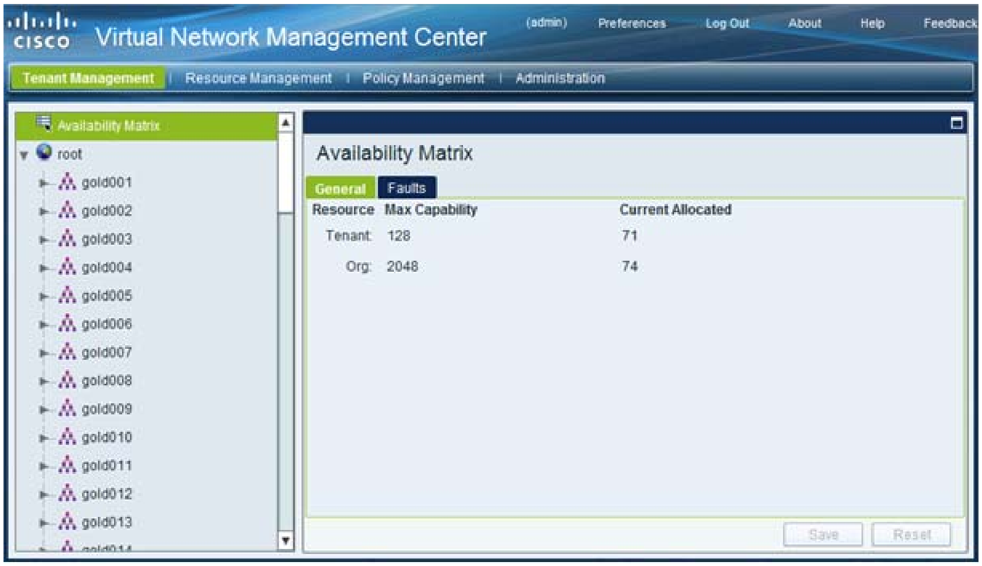
The VNMC provides the ability to support multitenant environments. A multitenant environment enables the division of large physical infrastructures into logical entities/organizations. Multitenancy allows logical isolation between organizations without providing a dedicated physical infrastructure for each organization. The VNMC administrator can assign unique resources to each tenant through the related organization in the multitenant environment. VNMC provides a strict organizational hierarchy as follows:

1. Root
2. Tenant
3. Virtual Data Center
4. Application
5. Tier

VMDC Tenants' Organization Structure

For each VMDC tenant, a flat, one-level organizational hierarchy is created for the tenant under the Root organization, and VSG compute firewalls are assigned at the tenant org level. Figure 5-18 shows the tenants created for this implementation.

Figure 5-18 VMDC Tenants



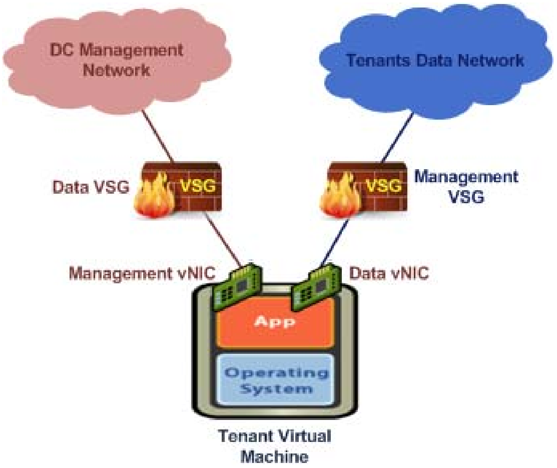
Resource	Max Capability	Current Allocated
Tenant:	128	71
Org:	2048	74

Note Compute firewall(s) should be added at the tenant level or below, and not at the Root org level.

Management Organization Structure

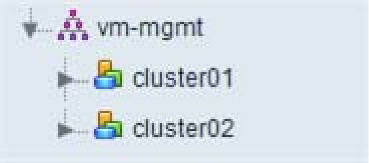
For this implementation, each protected VM has two vNICs, the data vNIC for the tenant data traffic and the management vNIC for back-end management traffic of the VMs. The data vNIC of each tenant VM would be protected by the dedicated VSG assigned to the tenant, while the management vNIC of all tenants' VMs are organized into one **vm-mgmt** tenant with its own VSG resources. Figure 5-19 illustrates the VSG appliances' deployment for a typical tenant VM.

Figure 5-19 Tenant Virtual Machine and VSG Appliances



The **vm-mgmt** tenant is organized into two virtual DCs organizational hierarchy. The organizational hierarchy is along the vSphere cluster, and each sub-org is assigned with one VSG to protect the management vNIC of the VMs in that cluster. Figure 5-20 shows the **vm-mgmt** organizational hierarchy.

Figure 5-20 Management Tenant



vm-mgmt
cluster01
cluster02

Security Policies

This section presents the following topics:

- Tenants Security Policies, page 5-35
- Management Security Policies, page 5-44

Tenants Security Policies

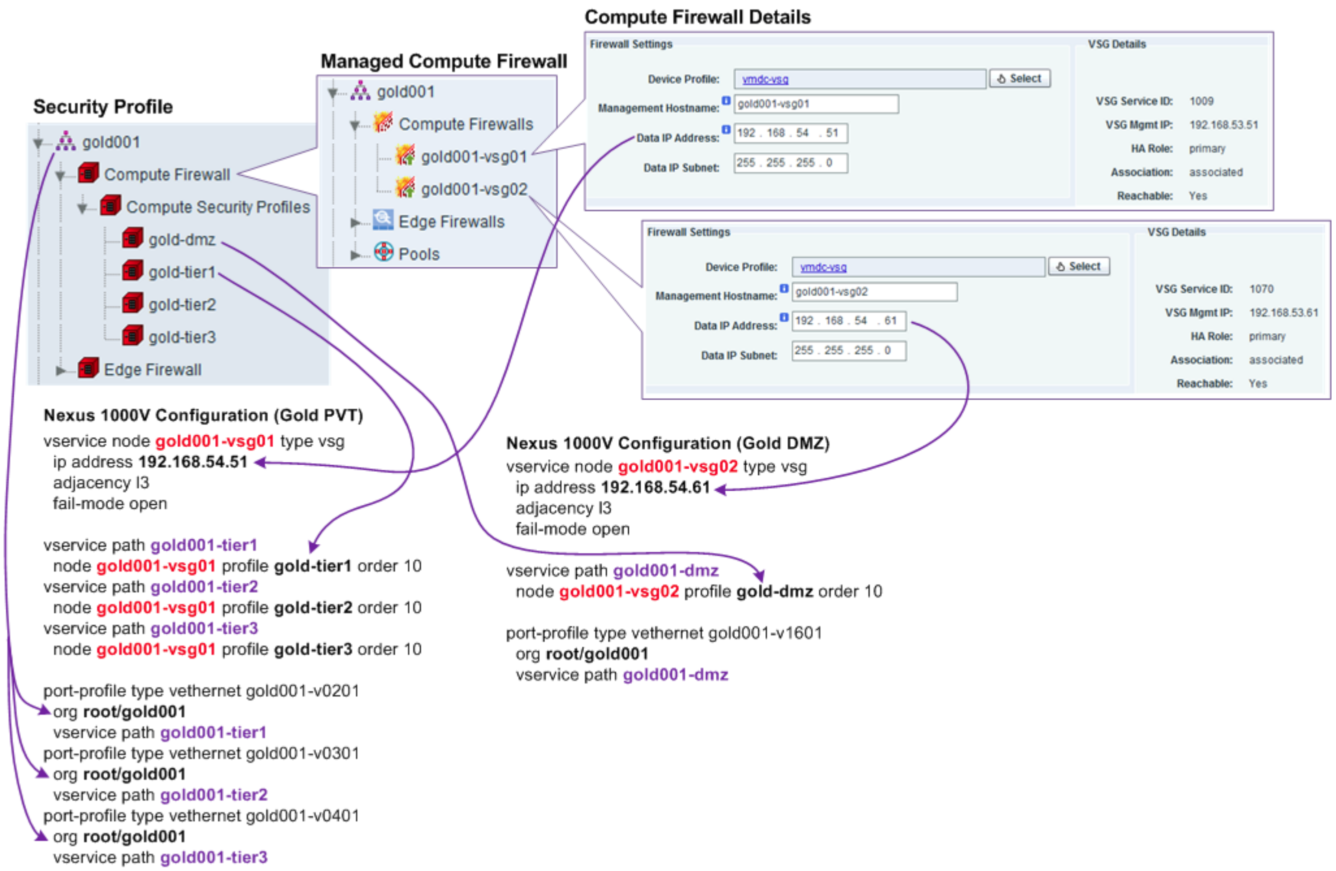
This section presents the following topics:

- Expanded Gold Tenant, page 5-36
- Private Zone, page 5-37
- Demilitarized Zone, page 5-38
- Bronze Tenant, page 5-41
- Copper/SMB Tenant, page 5-43

Expanded Gold Tenant

Each VMDC Expanded Gold tenant is provisioned with two security zones, a Private (PVT) zone and a DMZ. The Gold tenant PVT zone is allotted with three VLANs, and the related three Nexus 1000V port profiles, allowing the Gold tenant to construct a three-tier application architecture (presentation/ web tier, logic/application tier, data/database tier) with three IP subnets on the PVT zone. The Gold tenant DMZ is allotted with only one VLAN and one related Nexus 1000V port profile. All VMs for the Gold tenant DMZ belong to one IP subnet. Each Gold tenant is assigned with two VSG appliances, one for the PVT zone and DMZ respectively. Three security profiles are configured for the Gold tenant PVT zone, one for each of the three Nexus 1000V port profiles. The Gold tenant DMZ has one security profile for its single Nexus 1000V port profile. Figure 5-21 shows the security profiles, VSG appliances' assignment, and the related Nexus 1000V configuration for one of the Gold tenants.

Figure 5-21 Expanded Gold Tenant Security Profiles



Device Profile: vmdc-vsg
Management Hostname: gold001-vsg01
Data IP Address: 192 . 168 . 54 . 51
Data IP Subnet: 255 . 255 . 255 . 0
VSG Service ID: 1009
VSG Mgmt IP: 192.168.53.51
HA Role: primary
Association: associated
Reachable: Yes

Device Profile: vmdc-vsg
Management Hostname: gold001-vsg02
Data IP Address: 192 . 168 . 54 . 61
Data IP Subnet: 255 . 255 . 255 . 0
VSG Service ID: 1070
VSG Mgmt IP: 192.168.53.61
HA Role: primary
Association: associated
Reachable: Yes

Nexus 1000V Configuration (Gold PVT)

```
vservice node gold001-vsg01 type vsg
  ip address 192.168.54.51
  adjacency l3
  fail-mode open

vservice path gold001-tier1
  node gold001-vsg01 profile gold-tier1 order 10
vservice path gold001-tier2
  node gold001-vsg01 profile gold-tier2 order 10
vservice path gold001-tier3
  node gold001-vsg01 profile gold-tier3 order 10

port-profile type vethernet gold001-v0201
  org root/gold001
  vservice path gold001-tier1
port-profile type vethernet gold001-v0301
  org root/gold001
  vservice path gold001-tier2
port-profile type vethernet gold001-v0401
  org root/gold001
  vservice path gold001-tier3
```

Nexus 1000V Configuration (Gold DMZ)

```
vservice node gold001-vsg02 type vsg
  ip address 192.168.54.61
  adjacency l3
  fail-mode open

vservice path gold001-dmz
  node gold001-vsg02 profile gold-dmz order 10

port-profile type vethernet gold001-v1601
  org root/gold001
  vservice path gold001-dmz
```

Note For the test implementation, the VMs deployed are not provisioned with any actual applications set forth in the three-tier application architecture. Instead, each VM is provisioned with HTTP, FTP, and TFTP servers, which are used to approximate the three-tier data traffic flows.

For the three-tier policies using three VLANs, FTP/TFTP policies are not tested. The VSG FTP/ TFTP protocol inspect on the VSG fails to open the pinhole required for the data connection when the source and destination vNICs are under the same VSG protection, but on different VLANs. See CSCud39323 for more details.

This section presents the following topics:

- Private Zone, page 5-37
- Demilitarized Zone, page 5-38

Private Zone

Gold Tier 1 (Presentation/Web Tier) Security Policies

Table 5-6 lists the security policies configuration for the **gold-tier1** security profile.

Table 5-6 Security Policies for the gold-tier1 Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 1 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 1 VMs. This rule is optional, since the rule below already covered it; shown here for clarity.
Any	Tier 1 VMs Port 80	TCP	Permit	Allow HTTP to tier 1
Tier 1 VMs	Tier 2 VIP addresses Port 80	TCP	Permit	Allow HTTP, tier 1 to tier 2 load-balanced VIP addresses
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Gold Tier 2 (Logic/Application Tier) Security Policies

Table 5-7 lists the security policies configuration for the **gold-tier2** security profile.

Table 5-7 Security Policies for the gold-tier2 Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 2 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 2 VMs
Tier 2 SLB SNAT Pool	Tier 2 VMs Port 80	TCP	Permit	Allow HTTP, tier 1 to tier 2
Tier 2 VMs	Tier 3 VIP addresses Port 80	TCP	Permit	Allow HTTP, tier 2 to tier 3 load-balanced VIP addresses
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Gold Tier 3 (Data/Database Tier) Security Policies

Table 5-8 lists the security policies configuration for the **gold-tier3** security profile.

Table 5-8 Security Policies for the gold-tier3 Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 3 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 3 VMs
Tier 3 SLB SNAT Pool	Tier 3 VMs Port 80	TCP	Permit	Allow HTTP, tier 2 to tier 3
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Note 1. All conditions are and conditions, unless specified otherwise.

2. For each tier, a server load balancer is deployed. The server load balancer load balances traffic to the VM. Clients do not see the real IP addresses of the VMs, they only see the virtual IP addresses.

3. The server load balancer NATs the source IP address of load-balanced traffic going to the VMs to IP addresses in the SLB SNAT pool. Each tier has its own SNAT pool.

4. SLB Probe, Tier 1 VMs, Tier 1 VIP addresses, Tier 2 SLB SNAT Pool, etc, are conditions configured as a VNMC object group. The conditions in the object group have or semantics.

Demilitarized Zone

Table 5-9 lists the security policies configuration for the **gold-dmz** security profile.

Table 5-9 Security Policies Configuration for the gold-dmz Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	DMZ VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to DMZ VMs. This rule is optional, since the rule below already covered it; shown here for clarity.
Any	DMZ VMs Port 80	TCP	Permit	Allow HTTP to DMZ VMs
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Note

1. All conditions are AND conditions, unless specified otherwise.

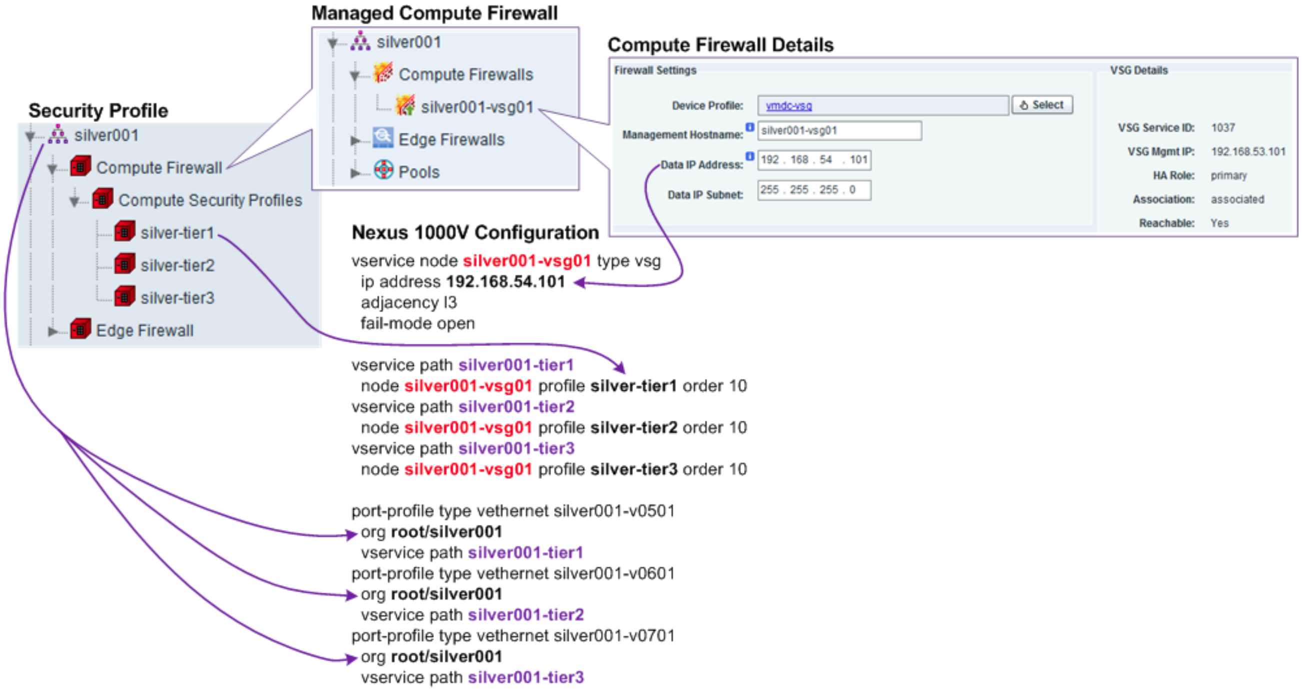
2. A server load balancer is deployed for the DMZ VMs. The server load balancer load balances traffic to the VM. Clients do not see the real IP addresses of the VMs, they only see the virtual IP addresses.

3. The server load balancer NATs the source IP address of load-balanced traffic going to the VMs to IP addresses in the SLB SNAT pool. Each tier has its own SNAT pool.

4. SLB Probe, DMZ VMs, etc, are conditions configured as a VNMC object group. The conditions in the object group have OR semantics.

Each VMDC Silver tenant is allotted with three VLANs, and the related three Nexus 1000V port profiles, allowing the Silver tenant to construct a three-tier application architecture (presentation/web tier, logic/application tier, data/database tier) with three IP subnets. Each Silver tenant is assigned with one VSG. Three security profiles are configured for the Silver tenant, one for each of the three Nexus 1000V port profiles. Figure 5-22 shows the security profiles, VSG assignment, and the related Nexus 1000V configuration for one of the Silver tenants.

Figure 5-22 Silver Tenant Security Profiles



Device Profile: vmdc-vsg
Management Hostname: silver001-vsg01
Data IP Address: 192 . 168 . 54 . 101
Data IP Subnet: 255 . 255 . 255 . 0

VSG Service ID: 1037
VSG Mgmt IP: 192.168.53.101
HA Role: primary
Association: associated
Reachable: Yes

```
vservice node silver001-vsg01 type vsg
 ip address 192.168.54.101
 adjacency l3
 fail-mode open

vservice path silver001-tier1
 node silver001-vsg01 profile silver-tier1 order 10
vservice path silver001-tier2
 node silver001-vsg01 profile silver-tier2 order 10
vservice path silver001-tier3
 node silver001-vsg01 profile silver-tier3 order 10

port-profile type vethernet silver001-v0501
 org root/silver001
 vservice path silver001-tier1
port-profile type vethernet silver001-v0601
 org root/silver001
 vservice path silver001-tier2
port-profile type vethernet silver001-v0701
 org root/silver001
 vservice path silver001-tier3
```

1. For the test implementation, the VMs deployed are not provisioned with any actual applications set forth in the three-tier application architecture. Instead, each VM is provisioned with HTTP, FTP, and TFTP servers, which are used to approximate the three-tier data traffic flows.
2. For the three-tier policies using three VLANs, FTP/TFTP policies are not tested. The VSG FTP/TFTP protocol inspect on the VSG fails to open the pinhole required for data connection when the source and destination vNICs are under the same VSG protection, but on different VLANs. See CSCud39323 for more details.

Silver Tier 1 (Presentation/Web Tier) Security Policies

Table 5-10 lists the security policies configuration for the **silver-tier1** security profile.

Table 5-10 Security Policies Configuration for the silver-tier1 Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 1 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 1 VMs. This rule is optional, since the rule below already covered it; shown here for clarity.
Any	Tier 1 VMs Port 80	TCP	Permit	Allow HTTP to tier 1
Tier 1 VMs	Tier 2 VIP addresses Port 80	TCP	Permit	Allow HTTP, tier 1 to tier 2 load-balanced VIP addresses
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Silver Tier 2 (Logic/Application Tier) Security Policies

Table 5-11 lists the security policies configuration for the **silver-tier2** security profile.

Table 5-11 Security Policies Configuration for the silver-tier2 Security Profile

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 2 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 2 VMs
Tier 2 SLB SNAT Pool	Tier 2 VMs Port 80	TCP	Permit	Allow HTTP, tier 1 to tier 2
Tier 2 VMs	Tier 3 VIP addresses Port 80	TCP	Permit	Allow HTTP, tier 2 to tier 3 load-balanced VIP addresses
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Silver Tier 3 (Data/Database Tier) Security Policies

Table 5-12 lists the security policies configuration for the **silver-tier3** security profile.

Table 5-12 Security Policies Configuration for the silver-tier3 Security Profile

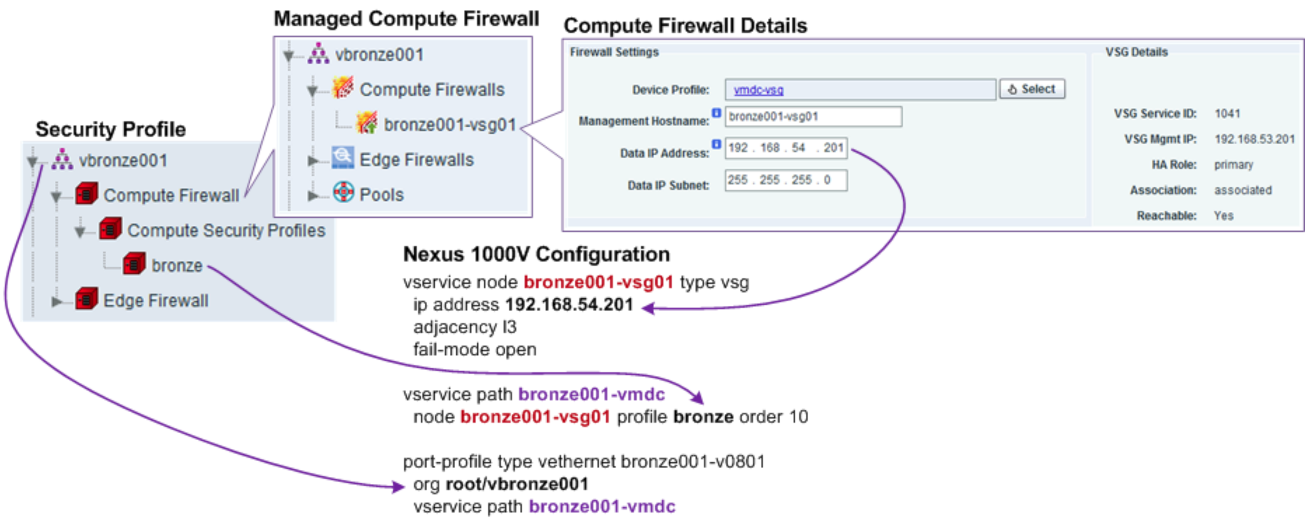
Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
SLB Probe	Tier 3 VMs Port 80	TCP	Permit	Allow HTTP probe from server load balancer to tier 3 VMs
Tier 3 SLB SNAT Pool	Tier 3 VMs Port 80	TCP	Permit	Allow HTTP, tier 2 to tier 3
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

1. All conditions are and conditions, unless specified otherwise.
2. For each tier, a server load balancer is deployed. The server load balancer load balances traffic to the VM. Clients do not see the real IP addresses of the VMs, they only see the virtual IP addresses.
3. The server load balancer NATs the source IP address of load-balanced traffic going to the VMs to IP addresses in the SLB SNAT pool. Each tier has its own SNAT pool.
4. **SLB Probe**, **Tier 1 VMs**, **Tier 1 VIP addresses**, **Tier 2 SLB SNAT Pool**, etc, are conditions configured as a VNMC object group. The conditions in the object group have or semantics.

Bronze Tenant

Each VMDC Bronze tenant is allotted with one VLAN and one related Nexus 1000V port profile. All VMs for the Bronze tenant belong to one IP subnet. Each Bronze tenant is assigned with one VSG. One security profile is configured for the Bronze tenant, with one Nexus 1000V port profile. Figure 5-23 shows the security profiles, VSG assignment, and the related Nexus 1000V configuration for one of the Bronze tenants.

Figure 5-23 Bronze Tenant Security Profile



Device Profile: vmdc-vsg
Management Hostname: bronze001-vsg01
Data IP Address: 192 . 168 . 54 . 201
Data IP Subnet: 255 . 255 . 255 . 0

VSG Service ID: 1041
VSG Mgmt IP: 192.168.53.201
HA Role: primary
Association: associated
Reachable: Yes

Nexus 1000V Configuration

```
vservice node bronze001-vsg01 type vsg
  ip address 192.168.54.201
  adjacency l3
  fail-mode open

vservice path bronze001-vmdc
  node bronze001-vsg01 profile bronze order 10

port-profile type vethernet bronze001-v0801
  org root/vbronze001
  vservice path bronze001-vmdc
```

Note

Note For the test implementation, the VMs deployed are not provisioned with any actual applications set forth in the three-tier application architecture. Instead, each VM is provisioned with HTTP, FTP, and TFTP servers, which are used to approximate the three-tier data traffic flows.

Bronze Flat Security Policies

Table 5-13 shows the security policies for one of the Bronze tenants; other Bronze tenants have similar settings.

Table 5-13 Security Policies for a Bronze Tenant

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
Any	Bronze VMs Port 80	TCP	Permit	Allow HTTP to Bronze VM
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Bronze Three-tier Security Policies

Table 5-14 show that each Bronze tenant is only assigned one VLAN and one Nexus 1000V port profile, and thus, only one security profile is assigned. With vZone and object groups, the VNMC provides the flexibility to configure three-tier separation using only one VLAN (one IP subnet, one security profile, one port profile).

Table 5-14 Bronze Three-tier Security Policies

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
Any	Port 22 or port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
Any	Tier 1 VMs Port 80	TCP	Permit	Allow HTTP to tier 1
Tier 1 VMs	Tier 2 VMs Port 80	TCP	Permit	Allow HTTP, tier 1 to tier 2
Tier 2 VMs	Tier 3 VMs Port 80	TCP	Permit	Allow HTTP, tier 2 to tier 3
Any	Any	Any	Drop, log	Deny everything else, log the denied requests

Note

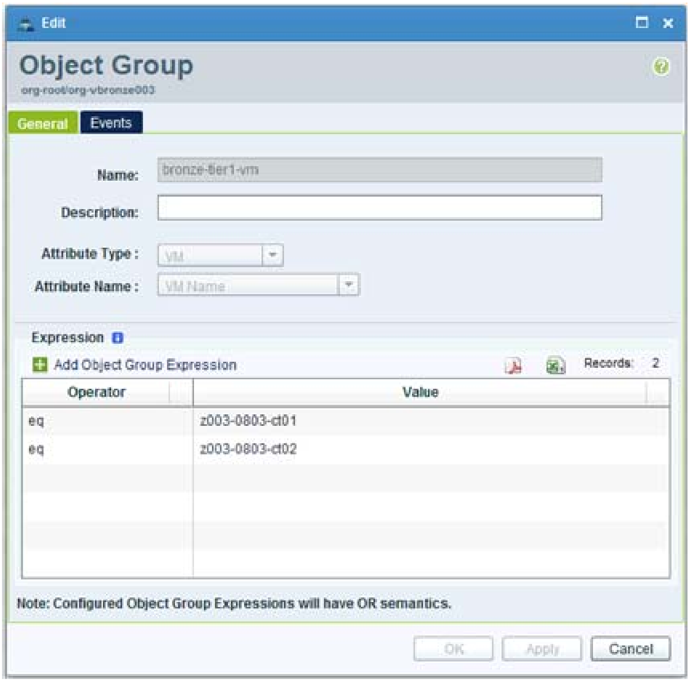
1. It is possible/supported to bind one security profile to multiple port profiles, but one port profile can only bind to one security profile at a time.

2. All conditions are and conditions, unless specified otherwise.

3. The Bronze VMs, Tier 1 VMs, etc, are conditions configured as a VNMC object group.

4. Tier 1 VM, Tier 2 VM, and Tier 3 VM are configured using object groups. The object-group filter expression matches against the VM name attribute. Figure 5-24 shows the object-group configuration for Tier 1 VM. The conditions in the object group have or semantics.

Figure 5-24 Object-group Configuration



Object Group

org-root/org-vbronze003

Name: bronze-tier1-vm

Description:

Attribute Type : VM

Attribute Name : VM Name

Expression

Records: 2

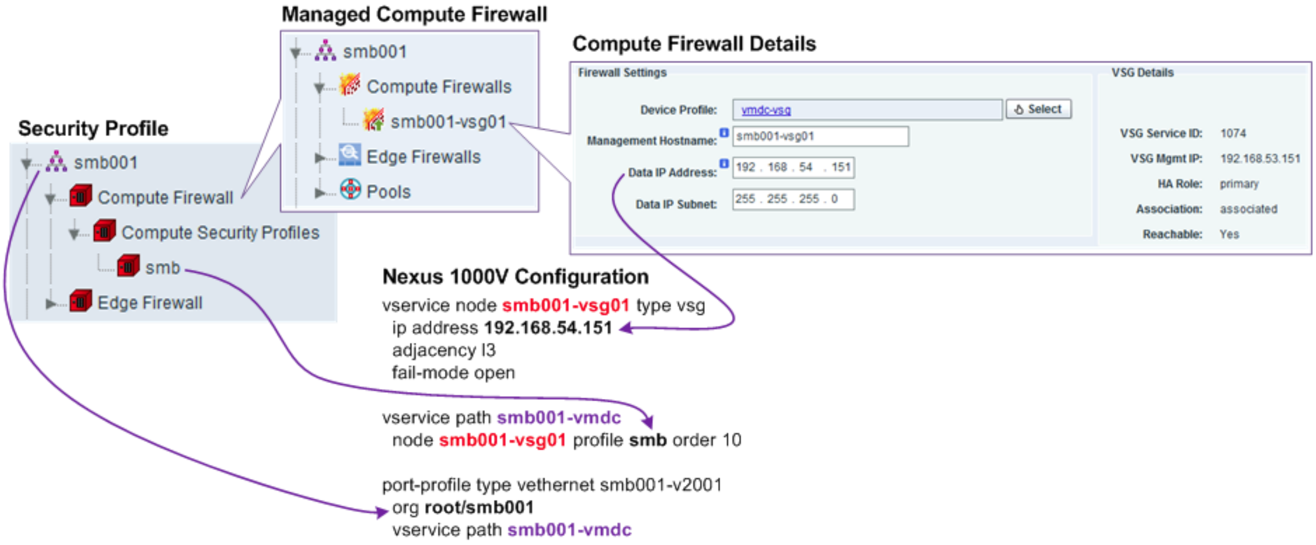
Operator	Value
eq	z003-0803-ct01
eq	z003-0803-ct02

Note: Configured Object Group Expressions will have OR semantics.

Copper/SMB Tenant

Each VMDC Copper/SMB tenant is allotted with one VLAN and one related Nexus 1000V port profile. All VMs for the Copper/SMB tenant belong to one IP subnet. Each Copper/SMB tenant is assigned with one VSG. One security profile is configured for the Copper/SMB tenant, with one Nexus 1000V port profile. Figure 5-25 shows the security profiles, VSG assignment, and the related Nexus 1000V configuration for one of the Copper/SMB tenants.

Figure 5-25 Copper/SMB Tenant Security Profile



Device Profile: vmdc-vsg
Management Hostname: smb001-vsg01
Data IP Address: 192 . 168 . 54 . 151
Data IP Subnet: 255 . 255 . 255 . 0

VSG Service ID: 1074
VSG Mgmt IP: 192.168.53.151
HA Role: primary
Association: associated
Reachable: Yes

```
vservice node smb001-vsg01 type vsg
 ip address 192.168.54.151
 adjacency l3
 fail-mode open

vservice path smb001-vmdc
 node smb001-vsg01 profile smb order 10

port-profile type vethernet smb001-v2001
 org root/smb001
 vservice path smb001-vmdc
```

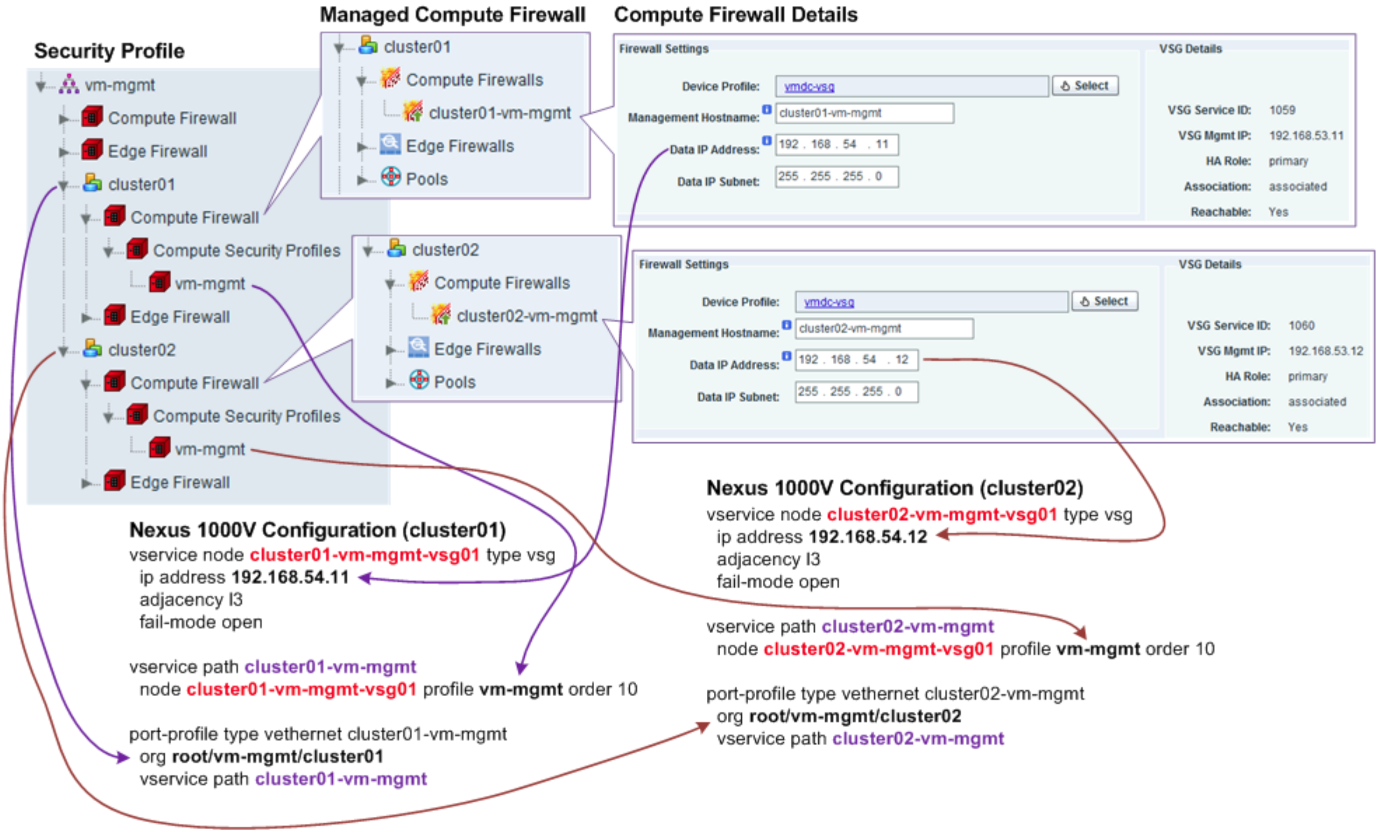
The security policies for the Copper/SMB tenant are similar to those of Bronze tenant. See the Bronze Tenant section for details.

Management Security Policies

The **vm-mgmt** tenant consists of all of the back-end vNICs of all VMs of all tenants. The back-end vNICs are used for management of VMs by the Service Provider. The **vm-mgmt** tenant is sub-divided into two virtual DCs, one virtual DC for each cluster. Each cluster is allotted with one VLAN and one related Nexus 1000V port profile, and all vNICs for the cluster belong to one IP subnet (Table 5-15).

Each cluster is assigned with one VSG. One security profile is configured for each cluster, along with one Nexus 1000V port profile. Figure 5-26 shows the security profiles, VSG assignment, and the related Nexus 1000V configuration for one of the **vm-mgmt** tenants.

Figure 5-26 vm-mgmt Tenant Security Profile



Device Profile: vmdc-vsg
Management Hostname: cluster01-vm-mgmt
Data IP Address: 192 . 168 . 54 . 11
Data IP Subnet: 255 . 255 . 255 . 0

VSG Service ID: 1059
VSG Mgmt IP: 192.168.53.11
HA Role: primary
Association: associated
Reachable: Yes

Device Profile: vmdc-vsg
Management Hostname: cluster02-vm-mgmt
Data IP Address: 192 . 168 . 54 . 12
Data IP Subnet: 255 . 255 . 255 . 0

VSG Service ID: 1060
VSG Mgmt IP: 192.168.53.12
HA Role: primary
Association: associated
Reachable: Yes

Nexus 1000V Configuration (cluster01)

```
vservice node cluster01-vm-mgmt-vsg01 type vsg
  ip address 192.168.54.11
  adjacency l3
  fail-mode open

vservice path cluster01-vm-mgmt
  node cluster01-vm-mgmt-vsg01 profile vm-mgmt order 10

port-profile type vethernet cluster01-vm-mgmt
  org root/vm-mgmt/cluster01
  vservice path cluster01-vm-mgmt
```

Nexus 1000V Configuration (cluster02)

```
vservice node cluster02-vm-mgmt-vsg01 type vsg
  ip address 192.168.54.12
  adjacency l3
  fail-mode open

vservice path cluster02-vm-mgmt
  node cluster02-vm-mgmt-vsg01 profile vm-mgmt order 10

port-profile type vethernet cluster02-vm-mgmt
  org root/vm-mgmt/cluster02
  vservice path cluster02-vm-mgmt
```

Table 5-15 vm-mgmt Tenant Security Profile Configuration

Source Condition	Destination Condition	Protocol	Action	Description
Any	Any	ICMP	Permit	Allow ping
VM mgmt vNIC	VM mgmt vNIC	Any	Drop, log	Deny tenant VMs to talk to each other via mgmt network, log the denied requests
Any	Port 67 or Port 68	UDP	Permit	Allow DHCP
Any	port 22 or Port 3389	TCP	Permit	Allow remote access
Any	Port 53	UDP	Permit	Allow DNS
Any	Port 123	UDP	Permit	Allow NTP
SNMP Managers	VM mgmt vNIC Port 161	UDP	Permit	Allow SNMP
VM mgmt vNIC	SNMP Managers Port 162	UDP	Permit	Allow SNMP Traps
Any	VM mgmt vNIC Port 21	TCP	Permit, inspect	Allow FTP to VM mgmt vNIC
Any	VM mgmt vNIC Port 69	UDP	Permit, inspect	Allow TFTP to VM mgmt vNIC
Any	VM mgmt vNIC Port 80	TCP	Permit	Allow HTTP to VM mgmt vNIC

Table 5-15 vm-mgmt Tenant Security Profile Configuration (continued)

Any	NFS Servers Port 111	Any	Permit	Allow Sun RPC, required for NFS
Any	NFS Servers Port 2049	Any	Permit	Allow NFS
Any	Any	Any	Drop, log	deny everything else, log the denied requests

Note 1. All conditions are and conditions, unless specified otherwise.

2. The IP address of the management vNIC is assigned via DHCP.

3. VM management vNIC, SNMP Managers, NFS Servers, etc. are conditions configured as a VNMC object group, using the IP subnet or IP address list as the filter expression. The conditions in the object group havee or semantics.

Services Best Practices and Caveats

ASA Firewall Appliance Best Practices

- The ASA FT and stateful links should be dedicated interfaces between the primary and secondary ASA.
- Failover interface policies should be configured to ensure that the security context fails over to the standby ASA if monitored interfaces are down.
- Configure an appropriate port-channel load-balancing scheme on the ASA to ensure that all port-channel interfaces are used to forward traffic out of the ASA.

Copper Implementation Best Practices

- Configure all Copper tenants' servers with either public or private IP addresses, not a mix of both types. If both types are needed, use seperate ASA context for all public addressed tenants and a separate context for all private addressed tenants.
- Private IP addresses for servers can be overlapped for different tenants, and requires the use of NAT with separate public IP addresses per tenant for outside.

Compute Firewall Best Practices

- The VSG does not support vSphere HA and DRS. On clusters dedicated for hosting VSG virtual appliances, disable vSphere HA and DRS. On clusters hosting both VSG virtual appliances and other VMs, disable HA and DRS for the VSG virtual appliances.
- For a VSG HA-pair, the primary and secondary nodes should be hosted on the same ESXi host. Use vSphere anti-affinity rules or DRS groups to ensure this.
- Each VSG virtual appliance (be it active or standby node) reserves CPU and memory resources from the ESXi host. Make sure the ESXi host has enough unreserved CPU and memory resources, otherwise, the VSG virtual appliance will not power on.
- Make sure that the clocks on the VNMC, VSGs, and Nexus 1000V are synchronized. The VSGs and Nexus 1000V will not be able to register to the VNMC if the clocks are too out of sync.

- Enable IP proxy ARP on the router interface(s) on the subnet/VLAN facing the VSG data interfaces.
- On the VNMC, compute firewalls should be added at the tenant level or below, and not at the Root org level.
- For the tenant, the DMZ should have its own VSG compute firewall, separate from the firewall used on the PVT zone.
- When configuring security policies/rules on the VNMC, the attributes used for filtering conditions should be preferred in the following order:
  - Network attributes, most prefer, providing highest performance for VSG – VM Attributes
  - vZone, lest prefer, lowest VSG performance **Compute Firewall Caveats**
- The VSG FTP/TFTP protocol inspect on the VSG fails to open the pinhole required for data connection when the source and destination vNICs are under the same VSG protection, but on different VLANs. See CSCud39323 for more details.

ACE 4710 Appliance Best Practices

- The FT VLAN should be configured using the **ft-port vlan <vlan-id>** command to ensure that FT packets have the right QoS labels. This ensures that proper treatment is given to ACE FT packets in the network.
- Configure an appropriate port-channel load-balancing scheme to ensure that all port-channel interfaces are used to forward traffic out of the ACE appliance.
- To avoid MAC collision among operational ACE appliances on the same VLAN, use an appropriate shared-vlan host-id <1-16> to ensure that each ACE appliance has a unique MAC address on a shared VLAN.

CHAPTER 6

QoS Implementation

Quality of Service (QoS) is implemented to support differentiated service level agreements with tenants. Edge policies enforce the contractual limits agreed upon, and rate limits depending on class of traffic and tenant type. Inside the Data Center (DC), policies provide appropriate bandwidth and queuing treatment for different service classes. This chapter discusses QoS implementation for the Virtualized Multiservice Data Center (VMDC) 2.3 solution.

This chapter contains the following major topics:

- End-to-End QoS, page 6-1
- Nexus 1000V QoS, page 6-8
- UCS QoS, page 6-15
- Nexus 7000 QoS, page 6-17
- Nexus 5000 QoS, page 6-24
- ASA QoS, page 6-30
- ACE QoS, page 6-30
- ASR 1000 QoS, page 6-31
- QoS Best Practices and Caveats, page 6-46

End-to-End QoS

This section discusses the key implementation goals for implementing QoS in VMDC 2.3. This section presents the following topics:

- QoS Domains and Trust Boundaries, page 6-1
- QoS Transparency, page 6-2
- QoS Traffic Service Classes, page 6-5

QoS Domains and Trust Boundaries

There are three QoS domains to be considered for end-to-end QoS:

1. The **end tenant network** (for example, Enterprise customer for a Service Provider-hosted Infrastructure as a Service (IaaS)) is in its own QoS domain and implements policies, as it wishes, independently from the DC and WAN network. This topic is not covered in this implementation.

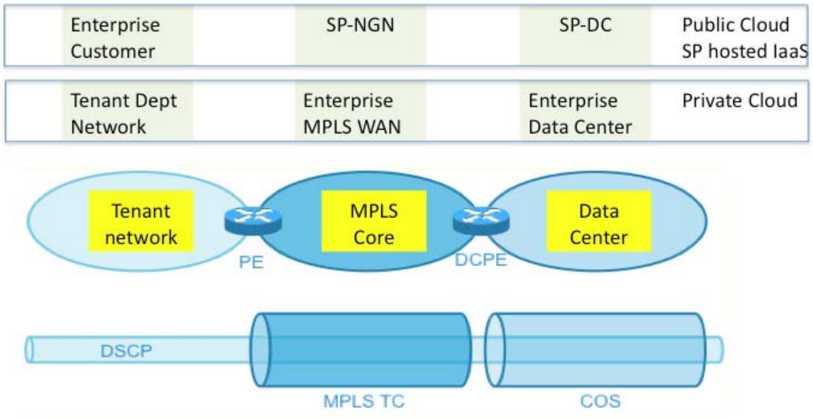
2. The **MPLS-Core network** (for example, Service Provider Next Generation Network (SP-NGN) or an Enterprise-wide MPLS-Core WAN) implements a QoS that supports the different offered services for WAN transport. The end tenant customer traffic is mapped to one of the WAN/ SP-NGN service classes based on the contractual Service Level Agreement (SLA) between the tenant and the Enterprise-WAN or SP-NGN Service Provider.
3. There is another boundary between the Enterprise-WAN/SP-NGN and the DC. Inside the DC, another QoS domain exists to support DC service offerings. The tenant customer's traffic is mapped into one of the DC classes of service to implement the contractual SLA.

The remote Provider Edge (PE) is the boundary between the tenant network and the provider network, i.e., WAN/SP-NGN, and will classify and mark traffic incoming into the WAN/SP-NGN from the tenant. This is also the enforcement point for the traffic entering the WAN/SP-NGN, and hence traffic is treated to enforce the contractual agreement and support agreed upon service level agreements by policing/rate-limiting and mark down. Traffic that is allowed is marked with a WAN/SP-NGN class marking so that the rest of the Enterprise/SP-NGN QoS domain and the DC QoS domain can trust this marking and use it to classify and provide appropriate treatment.

The Data Center Provider Edge (DC-PE) is the boundary between the WAN/SP-NGN and the DC. While the WAN/SP-NGN and DC can also be two independent Service Providers/Operators, in this implementation guide, they are assumed to be one. For the ingress direction from WAN/NGN to the DC, the DC-PE trusts the WAN/NGN markings and classifies traffic into similar classes within the DC. The meaning of the markings in the MPLS network that use the MPLS Traffic Class (MPLSTC) field are kept consistent with the dot1p Class of Service (CoS) markings used within the DC. In the egress direction, i.e., from the DC to the MPLS network, the DC-PE implements tenant aggregate policy enforcement, as well as mapping between the DC classes to the WAN/NGN classes.

Figure 6-1 shows the end-to-end QoS domains.

Figure 6-1 End-to-End QoS Domains



QoS Transparency

QoS transparency is an important requirement for many customers, where end customer/tenant traffic transits networks managed and operated by external entities such as a central Enterprise WAN or an external entity such as an SP-NGN-based network service. In these cases, after transiting these external networks, the tenant traffic should be received with tenant QoS markings unchanged. Also, it is important that the tenant network be able to use QoS classes and markings independently and flexibly to meet its needs without any restrictions when traffic transits externally managed networks such as an SP-NGN or an Enterprise WAN that is managed outside of the tenant network. The QoS markings are done using IP/Differentiated Services Code Point (DSCP) bits or IP/Precedence bits. When this traffic crosses into the WAN/SP-NGN, these markings are preserved as the traffic transits the WAN/SP-NGN and also crosses the DC QoS domains.

Within the WAN/SP-NGN provider QoS domains, the provider organization has its own classes of service and QoS policies, and outer header markings are used to implement the provider organizations' QoS. The tenant network's DSCP markings are left intact and not used by the WAN/DC provider. Inside the provider QoS domains in the WAN/SP-NGN and DC, MPLS-TC and dot1p CoS bits are used to mark provider classes of service. These provider classes are entirely different and independent from the tenant's QoS classes.

To support QoS transparency, customer traffic containing IP/DSCP bits needs to be preserved without change as traffic transits the WAN/SP-NGN MPLS-Core network and DC nodes. For this phase of VMDC, the QoS transparency desired could not be achieved for load-balanced traffic as the ACE load balancer does not yet preserve dot1p CoS, however, for all other traffic, it is possible to achieve QoS transparency. If the ACE-based Server Load Balancing (SLB) services are used, then the DC network needs to use IP/DSCP and remark DSCP to map to a provider marking. Due to this, QoS transparency cannot be maintained, i.e., customer assigned IP/DSCP markings are overwritten by the provider. The tenant network edge has to reclassify traffic coming in from the provider and remark the traffic in the customer edge routers to map to tenant classes and markings. The enhancement request CSCtt19577 has been filed for the implementation of the dot1p CoS preservation on the ACE.

On the Nexus 1000V, the DSCP values are set along with dot1p CoS. The class selector values mapping to dot1p CoS bits are used so that mapping between DSCP and dot1p is straight forward, and default mappings can be used on most hardware platforms where "trust dscp" is configured. Refer to QoS Traffic Service Classes for more information. The Nexus 1000V does the marking in the south to north direction, i.e., DC to tenant.

For the tenant to DC direction, the remote PE where the tenant traffic enters the WAN/SP-NGN needs to mark the IP/DSCP, as well as the MPLS-TC for traffic entering into the MPLS-Core network. Based on the contractual agreement, in-contract traffic is marked with MPLS-TC and IP/DSCP, and out-of-contract traffic is marked with a different marking. The DSCP values used correspond to the eight class selector values, and the corresponding values from the three bits are mapped to MPLS-TC. Refer to QoS Traffic Service Classes for more information.

Trust Boundaries

For traffic from the tenant network to the MPLS network, the remote PE implements the trust boundary. Traffic coming in from the tenant network is untrusted, and classification and policing is done to accept traffic based on the contractual SLA at a per-customer/tenant level. Based on this classification, traffic is mapped to a provider's class and marked with appropriate MPLS-TC and IP/ DSCP. The nodes in the MPLS network will use MPLS-TC to provide the desired treatment to the whole aggregate service class. On arrival at the DC-PE, which is the MPLS network and DC network boundary, mapping between MPLS network classes and DC classes is done. In this implementation, these are owned by the same entity, and hence the MPLS network (WAN/SP-NGN) class is mapped to a DC class, i.e, it is trusted, and MPLS-TC markings are used for classification and mapped to DC classes. For traffic from the DC to the tenant, the DC implements a trust boundary at the Nexus 1000V where the VM is connected. VMs are considered untrusted at the virtualized Switch layer. At this layer, the traffic is classified based on the contractual SLA and marked to an SP-DC class using dot1p CoS bits.

There is an additional per-tenant enforcement point at the DC-PE. As traffic from all of the VMs towards the tenant transits the DC-PE, the aggregate per-tenant SLA is enforced. The need for aggregate per-tenant SLA is needed as WAN bandwidth can be expensive and managed per tenant. In the case of an Enterprise using a Service Provider's IaaS offering, the tenant is charged for bandwidth consumed when injecting traffic into the MPLS-Core, referred to as the "point to hose model." Traffic is usually policed and within contract traffic is allowed into the appropriate class for transport across the MPLS-Core. The excess is either dropped or marked down into lower classes of traffic to use available bandwidth and will be dropped first during congestion.

Figure 6-2 shows the enforcement point for customer to DC traffic and the trust boundary.

Figure 6-2 Customer to DC Traffic Trust Boundary

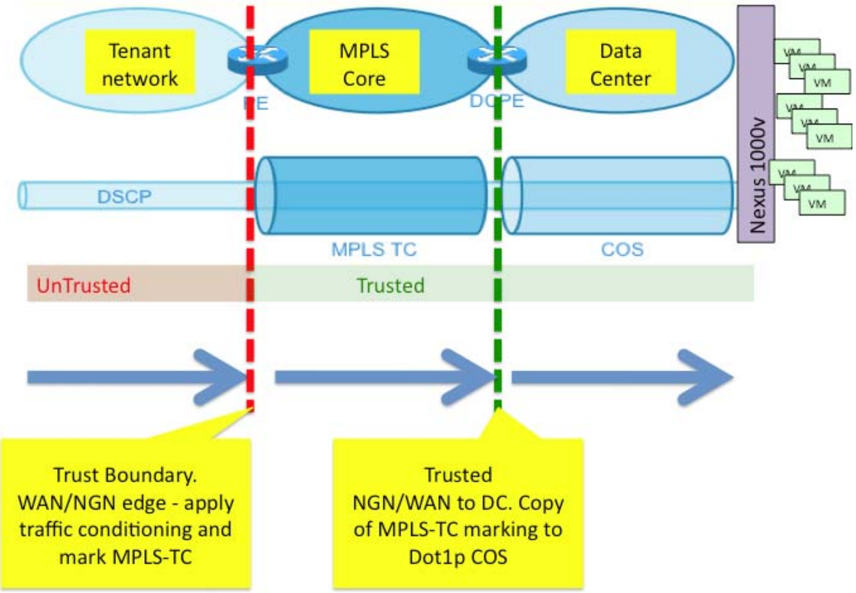
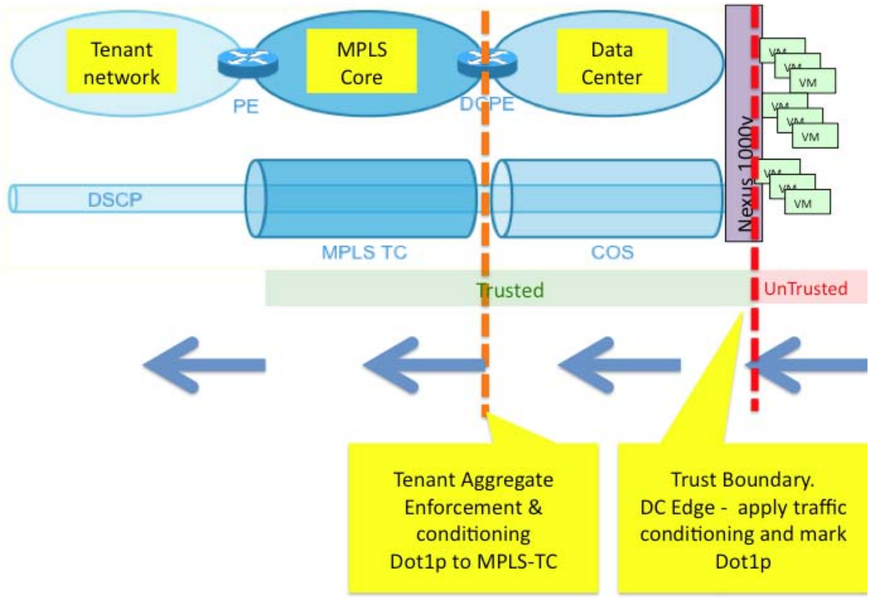

Tenant network
MPLS Core
Data Center
PE
DC PE
Nexus 1000v
VM
VM
VM
DSCP
MPLS TC
COS
UnTrusted
Trusted
Trust Boundary. WAN/NGN edge - apply traffic conditioning and mark MPLS-TC
Trusted NGN/WAN to DC. Copy of MPLS-TC marking to Dot1p COS

Figure 6-3 shows the trust boundaries for the DC to customer traffic.

Figure 6-3 DC to Customer Traffic Trust Boundary


Tenant network
PE
MPLS Core
DC PE
Data Center
Nexus 1000v
VM
DSCP
MPLS TC
COS
Trusted
UnTrusted
Tenant Aggregate Enforcement & conditioning Dot1p to MPLS-TC
Trust Boundary. DC Edge - apply traffic conditioning and mark Dot1p

QoS Traffic Service Classes

Table 6-1 lists the service classes that are implemented in the DC domain and the MPLS-Core network domain to support the different tenant traffic types.

Table 6-1 Traffic Service Classes

QoS Traffic Class	Data Center dot1p CoS Marking	MPLS-Core MPLS- TC Marking	Provider Marked IP/DSCP	Treatment
Management	7	7	cs7	BW reserved
Network Control	6	6	cs6	BW reserved
VoIP	5	5	cs5	Priority/Low Latency
Video	4 [1]	4	cs4	Not used in this phase
Call Control	3 [2]	3	cs3	BW reserved
Premium Data	2,1	2,1	cs2,cs1	BW reserved WRED
Standard Class	0	0	cs0	BW reserved WRED

Note 1. CoS3 is used for Fibre Channel over Ethernet (FCoE) traffic inside the Unified Computing System (UCS), however, this traffic is separated from the UCS Fabric Interconnect (FI) onwards and uses a dedicated native Fibre Channel (FC).

2. CoS4 is used for Network File System (NFS) traffic, and this traffic is seen in the Access layer (Nexus 5000) and then separated into dedicated links to NFS storage.

Tenant Type Mapping to QoS Traffic Classes

The DC IaaS offering uses four different tenant types with differing service levels and is expected to be priced differently, offering a range of options from premium services to standard lower priced services. These differentiated service levels are also mapped to a set of different DC and MPLS network-QoS service classes for traffic treatment.

In terms of end tenant offerings, the following traffic classes are offered:

1. **Low Latency Switched traffic—**For real time apps such as Voice over IP (VoIP).
2. **Call Signaling class—**Bandwidth (BW) guaranteed for signaling for VoIP and other multimedia.
3. **BW Guaranteed Data Class—**Premium data class.
4. **Best effort Data Class—**Standard data class.

To make the offerings simple, these traffic classes are bundled and mapped to tenant types.

Table 6-2 shows that Gold tenants can send traffic with dscp=ef/cs5, and the DC/MPLS network will classify it as VoIP in their domain and provide low latency guarantee by switching it in the priority queue. Call control is also allowed for Gold tenants. All other traffic from Gold tenants is treated as premium data QoS service class traffic. For Silver tenants, all traffic is treated as premium data, and there is no VoIP or Call Signaling class offered. For Bronze and Copper tenants, all traffic is treated as standard data class.

Table 6-2 Tenant Type Mapping to QoS Traffic Service Classes

Data Center and WAN Traffic class	Customer Marking	Gold	Silver	Bronze	Copper
VoIP (Low Latency)	IP/dscp=ef or cs5	x			
Call Control (BW Guaranteed)	IP/dscp=cs3	x			
Premium Data (BW Guaranteed)	Any	x	x		
Standard Data (Avail BW)	Any			x	x

Sample Contractual SLA Implemented

Table 6-3 lists contractual SLA for Gold/Silver and Bronze tenant types in validating implementation.

Table 6-3 Sample Contractual SLA Implemented

Traffic Type	Enforcement Point	Gold	Silver	Bronze/Copper
VoIP	Per-vNIC (VM to DC)	50 Mbps[1,3]	-	-
VoIP	Tenant total (DC to NGN)	100 Mbps	-	-
Call Control	Per-vNIC (VM to DC)	-	-	-
Call Control	Tenant total (DC to NGN)	10 Mbps [1]	-	-
Data	Per-vNIC (VM to DC)	250 Mbps[2,3] Excess also allowed but marked down	62.5 Mbps[2] Excess also allowed but marked down	500 Mbps[1] Strict limit
Data DC-> NGN	Tenant total CIR/ PIR	500 Mbps/3 Gbps[2]	250 Mbps/2 Gbps[2]	0/1 Gbps[1]
Data NGN->DC	Tenant total CIR/PIR	500 Mbps/3 Gbps	250 Mbps/2 Gbps	100 Mbps[4] /1 Gbps

1. One Rate 2 Colors (1R2C) policing is done to drop all exceed/violate.
2. Dual Rate Three Color (2R3C) policing is done. All exceed traffic is marked down and violate traffic is dropped. On the Nexus 1000V implementing per-vNIC limits, violate traffic is not dropped, i.e, out-of-contract traffic is not limited for Gold/Silver, and there is no Peak Information Rate (PIR) enforcement at the per-vNIC level.
3. In this validation, the per-vNIC rate-limits were set to four times oversubscription of the the per-tenant limit to allow for east-west and chatty VM traffic. For example, per-vNIC limit = (per-tenant aggregate limit x 4)/ #virtual machines.
4. Bronze tenants were configured with 100 Mbps bandwidth reservation. Alternatively, instead of bandwidth provisioning, "bandwidth remaining percent" only can be provisioned to provide Bronze tenants with no bandwidth guarantee to differentiate to a lower tier of service.
5. Copper tenants are policed at aggregate of all Copper tenants put together.

Per-tenant QoS Enforcement Points

To avoid operational complexity, the number of points of implementation of per-tenant QoS is kept small and is at the edges at these points:

1. DC-PE in SP-DC
2. Nexus 1000V virtualized switch where the VMs connect
3. Remote PE in SP-NGN where the customer connects

VMDC Infrastructure QoS

Within the DC, the following classes are to be supported. The Edge devices (Remote-PE, DC-PE, and Nexus 1000V) will mark traffic and the rest of the infrastructure trusts those markings and implements this QoS. There are no tenant specific differences in the QoS configuration needed in the DC-Core, DC-Agg, and Services layers, i.e., all of the tenants have the same QoS configuration. Table 6-4 lists the sample infrastructure egress QoS policy.

Table 6-4 Sample Infrastructure Egress QoS Policy

QoS Service Class	Data Center dot1p CoS Marking	Treatment Desired	Nexus 7000	Nexus 5000	UCS
Management	7		pq	qg5-pq	Default
Network Control	6		pq	qg5-pq	Platinum(cos=6
VoIP	5	Priority	pq[1]	qg5-pq	Gold (cos=5)
NFS	4	Infra		qg4 (10%)	Silver (cos=4)
Call Control	3	BW Guarantee	q2 (10%)	qg3 (10%)	Shared w[4] Fiber (cos=3)
Premium Data in-contract	2	BW Guarantee	q3[2] (70%)	qg2[3] (60%)	Bronze (cos=2)
Premium Data out-of-contract	1	Drop if need to	Default included	qg1 (0%)	Default
Standard Class	0	Best effort	Default (20%)	Default (5%)	Default

[1] On the Nexus 7000, all priority queue traffic is treated with strict priority. See the Nexus 7000 QoS section for more details. On F2 cards, only four queues are available, as compared to M1 or M2 cards where eight queues are available. VMDC 2.3 uses F2 cards to reduce cost.

[2] The F2 card has a single threshold per queue, and hence differentiated Weighted Random Early Detection (WRED) cannot be implemented.

[3] There is no WRED on the Nexus 5000.

[4] On the UCS, call signaling and FCoE traffic share the same queue. This is a no-drop queue. Some customers may prefer not to put this traffic in the same queue on the UCS. In that case, a different marking needs to be used for call signaling in the DC.

Nexus 1000V QoS

The QoS feature enables network administrators to provide differentiated services to traffic flowing through the network infrastructure. QoS performs such functions as classification and marking of network traffic, admission control, policing and prioritization of traffic flows, congestion management and avoidance, and so forth. The Nexus 1000V supports classification, marking, policing, and queuing functions. In this implementation of compute virtualization deployment with the VMware vSphere, UCS, Nexus 1000V and VSG, the following are the general classifications of traffic flowing through the network infrastructure:

- Nexus 1000V management, control, and packet traffic
- vSphere management and vMotion traffic
- VSG traffic
- Storage traffic
- Tenants VMs' data traffic

For this implementation, the traffic is marked with 802.1Q CoS bits; use of DSCP marking is required due to the ACE 4710 load balancer not preserving CoS. Table 6-5 shows the respective markings used in this implementation.

Table 6-5 Nexus 1000V QoS Markings

Traffic Class	Sub-Clsss	CoS Marking
Nexus 1000V management, control and packet traffic	N/A	6
vSphere	ESXi management	6
	vMotion	6
VSG traffic	Management	6
	Data/Service HA	6 6
Storage traffic	FCoE	3
	NFS	4
Gold tenants data traffic	Priority (policed at 50 Mbps, drop excess)	5
		2 1
	Within Contract (within 250	
	Mbps) Excess (above 250	
	Mbps)	
Silver tenants data traffic	Within Contract (within 62.5	2
	Mbps)	
		1
	Excess (above 62.5 Mbps)	
Bronze tenants data traffic	Policed at 500 Mbps, drop excess	0
Copper/SMB tenants data	Policed at 500 Mbps, drop	0
traffic	excess	

Nexus 1000V Management, Control, and Packet Traffic

For this implementation, the Nexus 1000V is configured with L3 SVS mode, and control and packet VLANs are not used. On each VEM/ESXi, the vmk0 management interface is used for communication between the VSM and VEM. See below for the QoS configuration for the ESXi vmk0 management interfaces.

vSphere Management and vMotion Traffic

Both the Nexus 1000V VSM and vCenter need access to the ESXi vmk0 management interface for monitoring, control, and configuration. If the vSphere HA feature is enabled, the ESXi hosts also exchange HA state information among themselves using the vmk0 interfaces. Control and Management traffic usually has low bandwidth requirements, but it should be treated as high-priority traffic. The following shows the QoS classification and marking configuration:

```
ip access-list erspan-traffic
  10 permit gre any any

class-map type qos match-all erspan-traffic
  match access-group name erspan-traffic
```

```
policy-map type qos esxi-mgmt
  class erspan-traffic
    set cos 0
    set dscp 0
  class class-default
    set cos 6
    set dscp cs6
policy-map type qos vmotion
  class class-default
    set cos 6
    set dscp cs6

port-profile type vethernet esxi-mgmt-vmknic
  capability l3control
  capability l3-vn-service
  service-policy input esxi-mgmt
port-profile type vethernet vmotion
  service-policy input vmotion
```

Note The IP access-list to match ERSPAN traffic shows the permit statement as:

10 permit gre any any

When configuring the permit statement, replace the **gre** keyword with **47**. The Nexus 1000V

CLI only shows the **gre** keyword with the show commands, but the protocol number is used during configuration. GRE is assigned protocol number 47.

On each ESXi host, the vmk0 management interface is attached to the **esxi-mgmt-vmknic** port profile, and the port profile is configured with **capability l3control** and **capability l3-vn-service** to allow the VEM/ESXi to use the vmk0 interface for communication with VSM and VSG respectively.

The addition of **capability l3control** configuration also allows the ESXi vmk0 interface to be used as the Encapsulated Remote Switched Port Analyzer (ERSPAN) source interface. ERSPAN traffic is voluminous, bursty, and usually not very important, and thus, does not require priority treatment.

On each ESXi host, the vmk1 interface is configured for vMotion, and the vmk1 interface is attached to the **vmotion** port profile. Depending on the vSphere configuration, vMotion activities could be bandwidth intensive, so configure bandwidth reservation to guarantee minimum bandwidth for vMotion traffic. Rate limiting configuration could be used to limit the bandwidth usage of vMotion if the network is always congested. Note, however that, policing vMotion traffic to too low bandwidth could cause excessive drops, which would cause vMotion to fail.

VSG Traffic

Each VSG virtual appliance has three interfaces:

- **Management**. Each VSG virtual appliance registers to the VNMC over the VSG management interface. The VNMC deploys security policies to the VSG over the VSG management interface. The communication between the VSG and VNMC takes place over an SSL connection on TCP port 443.
- **Data/Service**. The VSG receives traffic on the data interface from VEM/vPath for policy evaluation when protection is enabled on a port profile. The VSG then transmits the policy evaluation results to the VEM/vPath via the data interface. The VSGs are configured in L3 adjacency mode, and the packet exchange between VSG and VEM/vPath is encapsulated as IP packet.
- **HA**. When the VSG is deployed in high availability mode, the active and standby VSG nodes exchange heartbeats over the HA interface. These heartbeats are carried in L2 frames.

VSG traffic usually has low bandwidth requirements, but it should be treated as high-priority traffic. Packet drops on this traffic could lead to the VSG not operating properly. The following configuration shows the QoS classification and marking configuration for VSG traffic:

```
ip access-list vsg-to-vnmc
  10 permit ip any 192.168.13.16/32

class-map type qos match-all vsg-mgmt
  match access-group name vsg-to-vnmc

policy-map type qos vsg-mgmt
  class vsg-mgmt
    set cos 6
    set dscp cs6
  class class-default
    set cos 0
    set dscp 0
policy-map type qos vsg-data
  class class-default
    set cos 6
    set dscp cs6
policy-map type qos vsg-ha
  class class-default
    set cos cs6

port-profile type vethernet vsg-mgmt
  service-policy input vsg-mgmt
port-profile type vethernet vsg-data
  service-policy input vsg-data
port-profile type vethernet vsg-ha
  service-policy input vsg-ha
```

On each VSG virtual appliance, the management interface is attached to the **vsg-mgmt** port profile, the data/service interface is attached to the **vsg-data** port profile, and the HA interface is attached to the **vsg-ha** port profile. In addition to communication with VNMC, the VSG also uses the management interface for other purposes such as ssh, syslog, tftp, etc. Only traffic to the VNMC needs to be treated as high-priority traffic.

Storage Traffic

Storage traffic is the traffic generated when servers make access to remote storage, such as Network Attached Storage (NAS), Microsoft Common Internet File System (CIFS), SAN disks, etc. Servers use NFS, NetBIOS over TCP/IP, iSCSI, or FCoE protocols to access the remote storage. Storage traffic is lossless and receives priority over other traffic.

For this implementation, each ESXi host has access to NFS file storage and SAN disks for storing VMs data (vmdk disk, configuration files, etc).

FCoE transports storage traffic to access SAN disks. FCoE operates over a Data Center Ethernet (DCE) enhanced network. FCoE packets do not pass through the Nexus 1000V.

For each ESXi host, the vmk2 interface is configured for mounting NFS file storage, and the vmk2 interface is attached to the **NFS_1990** port profile. The following configuration shows the QoS classification and marking configuration for NFS storage traffic:

```
policy-map type qos nfs
  class class-default
    set cos 4

port-profile type vethernet NFS_1990
  service-policy input nfs
```

Tenants' Virtual Machines Data Traffic

VM data traffic is a generalization of all traffic transmitted or received by user VMs hosted on the virtual compute infrastructure. In the compute virtualization environment, this is the bulk of the traffic in the network. The QoS treatment for these traffic flows depends on the usage and applications running on the VMs. For this implementation, VMs are generally categorized into four classes of service, Gold, Silver, Bronze, and Copper/SMB.

Gold Tenant Virtual Machine

Data traffic from Gold tenant VMs is classified and treated as follows:

- **Priority traffic**—The application within the Gold VMs marks the packets with DSCP=EF to signal to the network that the packets require priority treatment. The Nexus 1000V marks packets meeting this criteria with CoS=5, and polices the data rate to 50 Mbps for each Gold VM so as not to starve out all other traffic classes.
- **Within contract traffic**—All other traffic from each Gold VM up to the limit of 250 Mbps is considered within contract. The Nexus 1000V marks and transmits packets meeting this criteria with CoS=2.
- **Excess traffic**—All other traffic from each Gold VM in excess of 250 Mbps is considered to belong in this criteria. The Nexus 1000V marks and transmits packets meeting this criteria with CoS=1.

Silver Tenant Virtual Machine

Traffic from the Silver tenant VMs is classified and treated as follows:

- **Within contract traffic**—All traffic from each Silver VM up to the limit of 62.5 Mbps is considered within contract. The Nexus 1000V marks and transmits packets meeting this criteria with CoS=2.
- **Excess traffic**—All traffic from each Silver VM in excess of 62.5 Mbps is considered to belong in this criteria. The Nexus 1000V marks and transmits packets meeting this criteria with CoS=1.

Bronze Tenant Virtual Machine

Traffic from the Bronze tenant VMs is classified and treated as follows:

- Each Bronze VM is limited to 500 Mbps of bandwidth, and all excess traffic is dropped. The Nexus 1000V marks and polices all packets within the bandwidth allotted with CoS=0.

Copper/SMB Tenant Virtual Machine

Traffic from the Copper/SMB tenant VMs is classified and treated the same as the Bronze tenant as follows:

- Each Copper/SMB VM is limited to 500 Mbps of bandwidth, and all excess traffic is dropped. The Nexus 1000V marks and polices all packets within the bandwidth allotted with CoS=0.

QoS Configuration for Tenants' Data Traffic

The following configuration shows the QoS classification, marking, and policing configuration. The QoS policy for vEthernet interfaces of VMs are applied when packets ingress to the Nexus 1000V.

```
class-map type qos match-all gold-ef
  match dscp ef
class-map type qos match-all gold-excess
  match qos-group 88
  match cos 0
class-map type qos match-all silver-excess
  match qos-group 89
  match cos 0

policy-map type qos gold
```

```
  class gold-ef
    police cir 50 mbps bc 200 ms conform set-cos-transmit 5 violate drop
    set cos 5
    set dscp cs5
<!--- Required as ACE does not preserve CoS, see note below. --->
  class class-default
    police cir 250 mbps bc 200 ms conform set-cos-transmit 2 violate set dscp dscp
table pir-markdown-map
    set qos-group 88
    set dscp cs2
<!--- Required as ACE does not preserve CoS, see note below. --->
policy-map type qos silver
  class class-default
    police cir 62500 kbps bc 200 ms conform set-cos-transmit 2 violate set dscp dscp
table pir-markdown-map
    set qos-group 89
    set dscp cs2
<!--- Required as ACE does not preserve CoS, see note below. --->
policy-map type qos bronze
  class class-default
    police cir 500 mbps bc 200 ms conform transmit violate drop
    set cos 0
    set dscp 0
<!--- Required as ACE does not preserve CoS, see note below. --->
policy-map type qos esxi-egress-remark
  class gold-excess
    set cos 1
    set dscp cs1
<!--- Required as ACE does not preserve CoS, see note below. --->
  class silver-excess
    set cos 1
    set dscp cs1
<!--- Required as ACE does not preserve CoS, see note below. --->

#---- parent port-profiles
port-profile type vethernet gold-profile
  service-policy input gold
port-profile type vethernet silver-profile
  service-policy input silver
port-profile type vethernet bronze-profile
  service-policy input bronze
port-profile type vethernet smb-profile
  service-policy input bronze

#---- port-profiles for the tenants
port-profile type vethernet gold001-v0201
  inherit port-profile gold-profile
port-profile type vethernet silver001-v0501
  inherit port-profile silver-profile
port-profile type vethernet bronze001-v0801
  inherit port-profile bronze-profile
port-profile type vethernet smb001-v2001
  inherit port-profile smb-profile

port-profile type ethernet system-data-uplink
  service-policy output esxi-egress-remark
```

Current QoS implementation uses trust CoS for Service Provider QoS marking, however, due to the ACE not supporting preservation of dot1p CoS, the DSCP also needs to be set at the Nexus 1000V layer. The values chosen for DSCP are such that the three higher order bits map directly to the dot1p CoS by default in the Nexus 7000 switches.

The QoS policies are attached to the parent port profiles for the Gold, Silver, and Bronze (also Copper/SMB) service classes. Port profiles for the individual tenant inherit the parent port profile configuration of their respective class. The hierarchical port profiles setup enforces consistent configuration.

As of version 4.2(1)SV2(1.1), the Nexus 1000V does not support 1R2C configuration when the two colors marking uses IEEE 802.1p CoS. This version only supports the two colors marking using DSCP. See CSCtr57528 for more information. The QoS configuration above shows the workaround for the 1R2C configuration for Gold and Silver tenants using CoS:

- By default, all incoming packets from the VM vEth interface have CoS=0 (the port profiles for vEth interfaces are all configured as access switchport). For Gold and Silver VMs, a policer is attached to the incoming vEth interface to mark all packets that conform to the configured rate with CoS=2, and all excess traffic is transmitted with a do nothing **pir-markdown-map** map.
- In addition, upon incoming, all packets are also tagged with a specific QoS-group value (88 and 89 in this case). For this implementation, only one QoS-group value is required, but the configuration shows two QoS-group values for clarity).
- On the egress side on the Ethernet uplink, the **esxi-egress-remark** QoS policy is attached to remark any packet that meets the following criteria with CoS=1:
  - CoS=0, AND
  - the specific QoS-group (88 and 89)

The Nexus 1000V configuration for 1R2C mandates the use of the **pir-markdown-map** for DSCP mutation for the violate action that is not dropped. The **pir-markdown-map** must be used even when no DSCP mutation is required. On the Nexus 1000V, the **pir-markdown-map** configuration is not shown as part of the running-config. For this implementation, DSCP mutation is not required, so make sure to change the **pir-markdown-map** to the following:

```
dc02-n1kv01# sh table-map pir-markdown-map

  Table-map pir-markdown-map
    default copy
```

QoS Queuing

The Nexus 1000V supports Class-Based Weighted Fair Queuing (CBWFQ) for congestion management. CBWFQ is a network queuing technique that provides support for user-defined traffic classes. The traffic classes are defined based on criteria such as protocols, IEEE 802.1p CoS values, etc. Packets satisfying the match criteria for a class constitute the traffic for that class. A queue is reserved for each class, and traffic belonging to a class is directed to the queue for that class with its own reserved bandwidth.

Use the following guidelines and limitations when configuring CBWFQ on the Nexus 1000V:

- Queuing is only supported on ESX/ESXi hosts version 4.1.0 and above.
- A queuing policy can only be applied on an uplink Ethernet interface in the egress (outbound) direction
- Only one queuing policy per VEM, and the policy can be applied on one physical interface or port-channel.
- For port-channel interfaces, queuing bandwidth applies on the member ports.
- Different VEMs can have different queuing policies (by assigning the VEM uplinks to different port profiles).
- The total number of traffic classes supported for a queuing policy is 16.
- 6.3 UCS QoS

- The total bandwidth allocated to each traffic class in a queuing policy should add up to 100%.

The following configuration shows the QoS queuing configuration on the Nexus 1000V. The QoS queuing policy is attached to the **esxi-egress-queuing** Ethernet port profile in the egress direction.

```
class-map type queuing match-all queuing-cos0
  match cos 0
class-map type queuing match-all queuing-cos1
  match cos 1
class-map type queuing match-all queuing-cos2
  match cos 2
class-map type queuing match-all queuing-cos4
  match cos 4
class-map type queuing match-all queuing-cos5
  match cos 5
class-map type queuing match-any mgmt-n-control
  match protocol n1k_control
  match protocol n1k_packet
  match protocol n1k_mgmt
  match protocol vmw_mgmt
  match protocol vmw_vmotion
  match cos 6

policy-map type queuing esxi-egress-queuing
  class type queuing queuing-cos5
    bandwidth percent 10
  class type queuing queuing-cos4
    bandwidth percent 10
  class type queuing queuing-cos2
    bandwidth percent 60
  class type queuing queuing-cos1
    bandwidth percent 5
  class type queuing queuing-cos0
    bandwidth percent 5
  class type queuing mgmt-n-control
    bandwidth percent 10

port-profile type ethernet system-data-uplink
  service-policy type queuing output esxi-egress-queuing
```

UCS QoS

UCS uses DCE to handle all traffic inside a Cisco UCS instance. The UCS unified fabric unifies LAN and SAN traffic on a single Ethernet transport for all blade servers within a UCS instance. SAN traffic is supported by the FCoE protocol, which encapsulates FC frames in Ethernet frames. The Ethernet pipe on the UCS unified fabric is divided into eight virtual lanes. Two virtual lanes are reserved for the internal system and management traffic, and the other six virtual lanes are user configurable. On the UCSM, the QoS system classes determine how the unified fabric bandwidth in these six virtual lanes is allocated across the entire UCS instance.

QoS System Class

The QoS system class defines the overall bandwidth allocation for each traffic class on the system. Each system class reserves a specific segment of the bandwidth for a specific type of traffic. This provides a level of traffic management, even in an oversubscribed system. UCSM provides six user configurable QoS system classes. In this implementation of compute virtualization deployment with VMware vSphere, UCS, Nexus 1000V and VSG, the following are the general classifications of traffic flowing through the network infrastructure:

- Nexus 1000V management, control and packet traffic
- vSphere management and vMotion traffic
- VSG traffic
- Storage traffic
- Tenants' VMs data traffic

For more details of the traffic classes, refer to Nexus 1000V QoS.

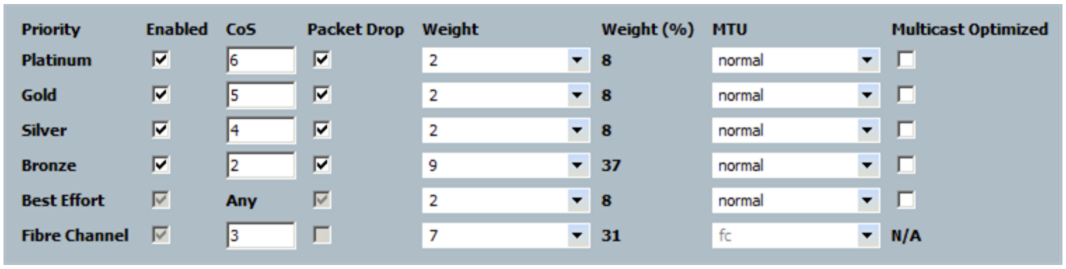
Table 6-6 shows the traffic classes to UCS QoS system classes' mapping.

Table 6-6 Mapping Traffic Classes to UCS QoS System Classes

UCS QoS System Class	Traffic Type	CoS Marking	Assured Bandwidth
Platinum	Nexus 1000V Management, Control and Packet VMware Management and vMotion VSG management, HA and Data	6	7%
Gold	DSCP=EF traffic from Gold VMs	5	7%
Silver	NFS Storage	4	7%
Bronze	In contract traffic from Gold and Silver VMs	2	42%
Best Effort	CoS=1, out of contract traffic from Gold and Silver VMs CoS=0, traffic from Bronze and Copper/ SMB VMs	0, 1	7%
Fiber Channel	FCoE Storage	3	30%

Figure 6-4 shows the QoS system class configuration on the UCSM.

Figure 6-4 UCS QoS System Class Configuration



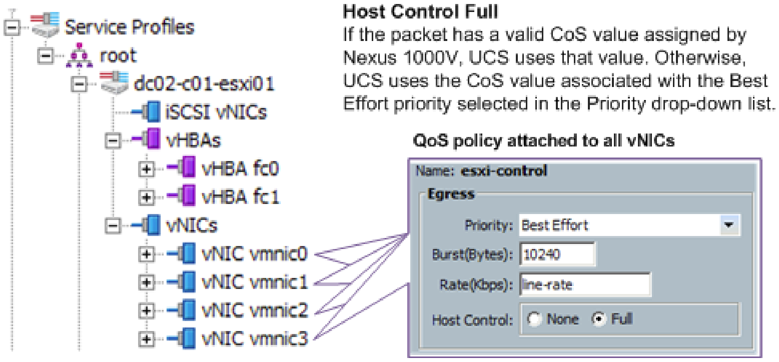
Priority	Enabled	CoS	Packet Drop	Weight	Weight (%)	MTU	Multicast Optimized
Platinum	☑	6	☑	2	8	normal	☐
Gold	☑	5	☑	2	8	normal	☐
Silver	☑	4	☑	2	8	normal	☐
Bronze	☑	2	☑	9	37	normal	☐
Best Effort	☑	Any	☑	2	8	normal	☐
Fibre Channel	☑	3	☐	7	31	fc	N/A

The QoS system class configuration uses weight (value range 1 - 10) to determine the bandwidth allocation for each class. The system then calculates the bandwidth percentage for each class based on the individual class weight divided by the sum of all weights; getting the correct weight for each class to meet the desired percentage for each class requires some trials and errors. The final results might not exactly match the desired design.

QoS Policy

The QoS policy determines the QoS treatment for the outgoing traffic from a vNIC of the UCS blade server. For UCS servers deployed with the Nexus 1000V, it is highly recommended to do the CoS marking at the Nexus 1000V level. On the UCSM, a QoS policy with the **Host Control Full** setting is attached to all vNICs on the service profile (logical blade server). The policy allows UCS to preserve the CoS markings assigned by the Nexus 1000V. If the egress packet has a valid CoS value assigned by the host (i.e., marked by Nexus 1000V QoS policies), UCS uses that value. Otherwise, UCS uses the CoS value associated with the **Best Effort** priority selected in the Priority drop-down list. Figure 6-5 shows the QoS policy configuration.

Figure 6-5 UCS QoS Policy for vNICs



Host Control Full
If the packet has a valid CoS value assigned by Nexus 1000V, UCS uses that value. Otherwise, UCS uses the CoS value associated with the Best Effort priority selected in the Priority drop-down list.

QoS policy attached to all vNICs

Name: **esxi-control**

Egress

Priority: Best Effort
Burst(Bytes): 10240
Rate(Kbps): line-rate
Host Control: ○ None ● Full

Nexus 7000 QoS

This section discusses Nexus 7000 QoS at the DC-Agg layer.

- Nexus 7000 QoS Policy Implementation, page 6-17
- Nexus 7000 Queuing Policy Implementation, page 6-19

Nexus 7000 QoS Policy Implementation

The Nexus 7000 is used in the DC-Agg layer. The QoS requirement at this layer is to support infrastructure QoS and will be kept tenant agnostic. On the Nexus 7000 platform, QoS is implemented in two parts, the QoS policy configuration and the queuing policy configuration.

In this section, the QoS policy implementation details are discussed. The main reason why QoS policy is required is to preserve dot1p CoS for all traffic.

By default, for routed traffic, the Nexus 7000 preserves DSCP, i.e., will use DSCP to create outbound dot1p CoS from IP/precedence, however, to support QoS transparency, this default behavior is not desirable. Service Provider classes are marked using dot1p CoS bits at the trust boundaries of the Nexus 1000V switch and DC-PE, and this is not to be overridden at the Nexus 7000 based DC-Agg layers. This default behavior is overridden by configuring a QoS policy.

Ingress Policy Details

The ingress policy classifies each class based on dot1p CoS bits, and these traffic markings are trusted. Traffic coming into the Nexus 7004 Agg is expected to be marked with provider class markings in the dot1p header. In the north to south direction, the ASR 1000 PE does that, and in the south to north direction, the Nexus 1000V virtual switch does that. All devices need to preserve dot1p CoS bits as traffic transits through them in the DC, except for ACE 4710 SLB.

Since the ACE 4710 does not currently transit traffic, depending on the load-balancing policy, the system may lose the CoS markings, and DC providers have to use DSCP markings at the edges, i.e., the Nexus 1000V needs to mark for south to north. For north to south, the expectation is that the remote PE will mark appropriate markings for DSCP as well as MPLS-TC. In terms of the QoS facing the ACE 4710, the QoS policy classifies based on DSCP and marks dot1p CoS. This allows the rest of the DC network to use dot1p CoS, and modify changes only in the edges when the support for dot1p CoS marking is available on any of the services appliances that do not currently support CoS bits preservation.

Traffic is marked so that output dot1p is set based on input dot1p CoS instead of DSCP. This ingress QoS policy is applied to all L2 switch trunks at the port-channel level on the DC-Agg, facing the Nexus 5000 Integrated Compute and Storage (ICS) switch, vPC peer link, and other port-channels used to connect to the ASA and ACE appliances. The ingress QoS policy is also applied to all L3 subinterfaces on the DC-Agg facing the DC-PE.

Refer to Configuring Queuing and Scheduling on F-Series I/O Modules for QoS configuration on the Nexus 7000 series.

```
!!! Classmaps facing all ports other than ACE 4710 SLB appliances

class-map type qos match-all mgmt
      match cos 7
    class-map type qos match-all network-control
      match cos 6
    class-map type qos match-all voip-cos
      match cos 5
    class-map type qos match-all call-control-cos
      match cos 3
    class-map type qos match-all premium-data-cos
      match cos 2
    class-map type qos match-all premium-data-cos-out
      match cos 1

!!! Policy map facing all L2 ports and L3 subinterfaces, except ports facing ACE4710

Type qos policy-maps
  ====================

  policy-map type qos ingress-qos-policy
    class  voip-cos
      set cos 5
    class  premium-data-cos
      set cos 2
    class  call-control-cos
      set cos 3
    class  premium-data-cos-out
      set cos 1
    class  mgmt
      set cos 7
    class  network-control
      set cos 6
    class  class-default
      set cos 0

===

!!! Classmaps used for ACE facing policy-maps

    class-map type qos match-all dscp-mgmt
      match dscp 56
    class-map type qos match-all dscp-network-control
      match dscp 48
    class-map type qos match-all dscp-voip-cos
```

```
        match dscp 40
      class-map type qos match-all dscp-call-control-cos
        match dscp 24
      class-map type qos match-all dscp-premium-data-cos
        match dscp 16
      class-map type qos match-all dscp-premium-data-cos-out
        match dscp 8

!!!! Policy map facing the ACE4710s

    Type qos policy-maps
    ====================

    policy-map type qos dscp-ingress-qos-policy
      class  dscp-voip-cos
        set cos 5
      class  dscp-premium-data-cos
        set cos 2
      class  dscp-call-control-cos
        set cos 3
      class  dscp-premium-data-cos-out
        set cos 1
      class  dscp-mgmt
        set cos 7
      class  dscp-network-control
        set cos 6
      class  class-default
        set cos 0
```

Egress QoS Policy

The egress QoS policy is **not used** in VMDC 2.3.

In VMDC 2.2, an egress policy was used to police VoIP traffic. This is not a hard requirement as edges do police the amount of traffic injected into the DC, so this is additional mostly for protection against unforeseen errors. The configuration used had traffic classified based on the QoS-group marked in the ingress policy. A 1R2C policer was applied to drop all violate traffic on a per-tenant basis to a fixed max value, which each tenant was not expected to exceed. The egress QoS policy is applied only on the L3 subinterfaces on the DC-Agg layer towards the DC-PE.

When implementing on F2 cards, however, implementing the same configuration is quite operationally intensive. The F2 card uses System on Chip (SoC) architecture, and hence there are 12 SoCs that implement policing on the ingress ports for traffic mapping to egress. This can cause the policing rate to be in effect 12x of the desired rate. To avoid this, specific ingress ports mapping to the same SoCs could be used, however, this might be operationally difficult. Also, set/match QoS-group functionality is not supported on F-series modules. Given that egress policing is not really needed, this configuration has been removed from VMDC 2.3. If this configuration is needed, M1 or M2-based designs should be considered.

Nexus 7000 Queuing Policy Implementation

Network-QoS Configuration

On the F2 cards on the Nexus 7000, there is a concept of network-QoS, which defines the no-drop vs. tail-drop behavior and the Maximum Transmission Unit (MTU). This is mainly used to allocate CoS markings used for FCoE traffic and provide no-drop treatment, as well as support different MTU sizes.

Refer to Network-QoS Policy Configuration for an explanation of how to configure network-QoS. There are built-in defaults for network-QoS, which automatically configures the number of queues and CoS markings reserved for FCoE traffic and Ethernet traffic. The configuration below shows the available default network-QoS templates.

For VMDC 2.3, the default-nq-8e-4q4q-policy is used, which provides for four ingress queues and four egress queues, and all eight CoS markings are tail-drop classes and sets the MTU to 1500. This is configured at the system level under the system QoS global configuration. If required, this network-QoS configuration can be copied using the **qos copy** command. This configuration can also be customized, for example, changing the MTU to support jumbo frames.

Refer to Configuring Queuing and Scheduling on F-Series I/O Modules for more information on QoS and queuing on F-series cards.

```
dc02-n7k-agg2# conf t
Enter configuration commands, one per line.  End with CNTL/Z.
dc02-n7k-agg2(config)# system qos
dc02-n7k-agg2(config-sys-qos)# service-policy type network-qos ?
  default-nq-4e-policy       Default 4-ethernet policy (4-drop 4-nodrop CoS)
  default-nq-6e-policy       Default 6-ethernet policy (6-drop 2-nodrop CoS)
  default-nq-7e-policy       Default 7-ethernet policy (7-drop 1-nodrop CoS)
  default-nq-8e-4q4q-policy  Default 8-ethernet policy 4 queues for ingress and 4 
queues for egress (8-drop CoS)
  default-nq-8e-policy       Default 8-ethernet policy 2 queues for ingress and 4 
queues for egress (8-drop CoS)

dc02-n7k-agg2(config-sys-qos)# service-policy type network-qos 
default-nq-8e-4q4q-policy
dc02-n7k-agg2(config-sys-qos)#

 %IPQOSMGR-2-QOSMGR_NETWORK_QOS_POLICY_CHANGE: Policy default-nq-8e-4q4q-policy is now 
active

dc02-n7k-agg2(config-sys-qos)#

#### Network-qos

dc02-n7k-agg1# show class-map type network-qos c-nq-8e-4q4q

  Type network-qos class-maps
  ===========================
  class-map type network-qos match-any c-nq-8e-4q4q
      Description: 8E-4q4q Drop CoS map
    match cos 0-7



dc02-n7k-agg1# show policy-map type network-qos default-nq-8e-4q4q-policy

  Type network-qos policy-maps
  ============================
  policy-map type network-qos default-nq-8e-4q4q-policy template 8e-4q4q
    class type network-qos c-nq-8e-4q4q
      congestion-control tail-drop
      mtu 1500
```

Egress Queuing Policy Details

The queues and the queuing class-maps are fixed when the network-QoS template is selected. For the network-QoS template used in VMDC 2.3, which is default-nq-8e-4q4q-policy, the ingress and egress queuing classes and the default in and out queuing policies are as follows:

```
class-map type queuing match-any 4q1t-8e-4q4q-in-q1
```

```
      Description: Classifier for Ingress queue 1 of type 4q1t-8e-4q4q
      match cos 5-7

    class-map type queuing match-any 4q1t-8e-4q4q-in-q-default
      Description: Classifier for Ingress queue 2 of type 4q1t-8e-4q4q
      match cos 0-1

    class-map type queuing match-any 4q1t-8e-4q4q-in-q3
      Description: Classifier for Ingress queue 3 of type 4q1t-8e-4q4q
      match cos 3-4

    class-map type queuing match-any 4q1t-8e-4q4q-in-q4
      Description: Classifier for Ingress queue 4 of type 4q1t-8e-4q4q
      match cos 2


 !!!! For ingress queuing, use defaults for vmdc23 for classes and policy
dc02-n7k-agg1# show policy-map type queuing default-8e-4q4q-in-policy


  Type queuing policy-maps
  ========================

  policy-map type queuing default-8e-4q4q-in-policy
    class type queuing 4q1t-8e-4q4q-in-q1
      queue-limit percent 10
      bandwidth percent 25
    class type queuing 4q1t-8e-4q4q-in-q-default
      queue-limit percent 30
      bandwidth percent 25
    class type queuing 4q1t-8e-4q4q-in-q3
      queue-limit percent 30
      bandwidth percent 25
    class type queuing 4q1t-8e-4q4q-in-q4
      queue-limit percent 30
      bandwidth percent 25


!!!!! Output Queueing classes, use default classes but policy should be changed

  class-map type queuing match-any 1p3q1t-8e-4q4q-out-pq1
      Description: Classifier for Egress Priority queue 1 of type 1p3q1t-8e-4q4q
      match cos 5-7

    class-map type queuing match-any 1p3q1t-8e-4q4q-out-q2
      Description: Classifier for Egress queue 2 of type 1p3q1t-8e-4q4q
      match cos 3-4

    class-map type queuing match-any 1p3q1t-8e-4q4q-out-q3
      Description: Classifier for Egress queue 3 of type 1p3q1t-8e-4q4q
      match cos 2

    class-map type queuing match-any 1p3q1t-8e-4q4q-out-q-default
      Description: Classifier for Egress queue 4 of type 1p3q1t-8e-4q4q
      match cos 0-1



!!! Default output 8e-4q Queuing policy

dc02-n7k-agg1# show policy-map type queuing default-8e-4q4q-out-policy


  Type queuing policy-maps
  ========================

  policy-map type queuing default-8e-4q4q-out-policy
    class type queuing 1p3q1t-8e-4q4q-out-pq1
      priority level 1
    class type queuing 1p3q1t-8e-4q4q-out-q2
      bandwidth remaining percent 33
    class type queuing 1p3q1t-8e-4q4q-out-q3
      bandwidth remaining percent 33
    class type queuing 1p3q1t-8e-4q4q-out-q-default
      bandwidth remaining percent 33
dc02-n7k-agg1#
```

Refer to Configuring Queuing and Scheduling on F-Series I/O Modules for more information on the queuing configuration on the F2 card.

Table 6-7 provides the queuing values used in VMDC 2.3, which are left mapped to the default class-maps.

Table 6-7 Queuing Values

Class Map Queue Name	CoS Values	Comment
1p3q1t-8e-4q4q-out-pq1	5,6,7	VoIP, Network Control, Network Management
1p3q1t-8e-4q4q-out-q2	3,4	Call Signaling
1p3q1t-8e-4q4q-out-q3	2	Gold & Silver Data in-contract
1p3q1t-8e-4q4q-out-q-default	0,1	Bronze, Copper as well as Gold/Silver out-of-contract

Note 1. Default CoS-queue mapping can be modified only in the default vDC.

2. If CoS-queue mapping is modified, then make sure to configure a queuing policy-map and allocate sufficient bandwidth to the respective queues. This queuing policy should be applied on all interfaces in all vDCs to prevent unexpected traffic blackholing.

3. In VMDC 2.3, default CoS-queue mappings are used.

An egress queuing policy is configured with specific bandwidth allocation, as shown in Table 6-8, and applied to all physical and port-channel interfaces. The bandwidth weights and queue-limits are based on the infrastructure QoS, as detailed in the End-to-End QoS section. Queue-limits are kept proportional to bandwidth weights, and the remaining bandwidth is calculated after assuming 15% traffic from VoIP/priority queue.

Table 6-8 Egress Queuing Policy

Class Map Queue Name	Traffic Description	Assured Bandwidth	CoS Values
1p3q1t-8e-4q4q-out-q3	Gold and Silver Data	70%	2

Table 6-8 Egress Queuing Policy (continued)

1p3q1t-8e-4q4q-out- pq1	Gold VoIP	Priority	5,6,7
1p3q1t-8e-4q4q-out-q2	Call Control	10%	3,4
1p3q1t-8e-4q4q-out-q- default	Bronze, Copper, and Out-of-contract Gold/ Silver	20%	0,1

The following sample code shows a queuing policy configuration on the Nexus 7000 DC-Agg:

```
!!! Copy queuing config from default out

 dc02-n7k-agg1(config)# qos copy policy-map type queuing default-8e-4q4q-out-policy 
prefix vmdc23

!!! VMDC23 output queuing policy

dc02-n7k-agg1# show policy-map type queuing vmdc23-8e-4q4q-out

  Type queuing policy-maps
  ========================

  policy-map type queuing vmdc23-8e-4q4q-out
    class type queuing 1p3q1t-8e-4q4q-out-pq1
      priority level 1
    class type queuing 1p3q1t-8e-4q4q-out-q2
      bandwidth remaining percent 10
    class type queuing 1p3q1t-8e-4q4q-out-q3
      bandwidth remaining percent 70
    class type queuing 1p3q1t-8e-4q4q-out-q-default
      bandwidth remaining percent 20
```

Ingress Queuing Policy Details

For this implementation, the default ingress queuing policy was used.

Attaching Queuing Policy

The queuing policies are attached to the physical Ethernet interfaces or to port-channel interfaces. Queuing policies are attached at the parent-interface level, and cannot be attached at the subinterface level for L3 Ethernet or port-channels.

```
interface port-channel356
  description PC-to-N5K-VPC
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 1-1120,1601-1610,1801-1860,2001-2250
  switchport trunk allowed vlan add 3001-3250
  spanning-tree port type network
  logging event port link-status
  service-policy type qos input ingress-qos-policy
  service-policy type queuing output vmdc23-8e-4q4q-out
  vpc 4000
dc02-n7k-agg1# show policy-map interface po356 type queuing

Global statistics status :   enabled

port-channel356

  Service-policy (queuing) input:   default-8e-4q4q-in-policy
    SNMP Policy Index:  302013613
    Class-map (queuing):   4q1t-8e-4q4q-in-q1 (match-any)
      queue-limit percent 10
      bandwidth percent 25
      queue dropped pkts : 0

    Class-map (queuing):   4q1t-8e-4q4q-in-q-default (match-any)
      queue-limit percent 30
      bandwidth percent 25
      queue dropped pkts : 0

    Class-map (queuing):   4q1t-8e-4q4q-in-q3 (match-any)
      queue-limit percent 30
      bandwidth percent 25
      queue dropped pkts : 0

    Class-map (queuing):   4q1t-8e-4q4q-in-q4 (match-any)
      queue-limit percent 30
      bandwidth percent 25
      queue dropped pkts : 0

  Service-policy (queuing) output:   vmdc23-8e-4q4q-out
    SNMP Policy Index:  302015884

    Class-map (queuing):   1p3q1t-8e-4q4q-out-pq1 (match-any)
      priority level 1
      queue dropped pkts : 0

    Class-map (queuing):   1p3q1t-8e-4q4q-out-q2 (match-any)
      bandwidth remaining percent 10
      queue dropped pkts : 0

    Class-map (queuing):   1p3q1t-8e-4q4q-out-q3 (match-any)
      bandwidth remaining percent 70
      queue dropped pkts : 0

    Class-map (queuing):   1p3q1t-8e-4q4q-out-q-default (match-any)
      bandwidth remaining percent 20
      queue dropped pkts : 0
```

Nexus 5000 QoS

This section discusses Nexus 5000 QoS at the ICS switch layer. In VMDC 2.3, the Nexus 5000 is used as part of the ICS stacks to aggregate multiple UCS Fabric Interconnects (FIs). The Nexus 5000 is used purely in L2 mode as a switch. The following sections detail the QoS implementation for VMDC 2.3 on the Nexus 5000 ICS switch.

- Nexus 5000 QoS Policy Implementation, page 6-25
- Nexus 5000 Network-QoS Policy Implementation, page 6-26
- Nexus 5000 Queuing Policy Implementation, page 6-28

Nexus 5000 QoS Policy Implementation

The Nexus 5000 uses QoS policy to match on CoS and mark a QoS-group for further treatment and queuing within the switch. The following configuration is used to implement VMDC 2.3 QoS, and the goal of this configuration is to map all traffic into six classes. CoS5, 6, and 7 are mapped to one class called vmdc-priority, and the rest of the CoS are mapped one-to-one. The Nexus 5000 series can support up to five classes of user traffic, besides class-default, which gives a total of six classes of traffic.

On the Nexus 5000, the QoS policy is used only on ingress to map incoming traffic into internal markings based in the QoS-group, and further treatment and queuing uses QoS-group.

```
class-map type qos match-any class-default
      match any

    class-map type qos match-any class-vmdc-p1
      match cos 1

    class-map type qos match-any class-vmdc-p2
      match cos 2

    class-map type qos match-any class-vmdc-p3
      match cos 3

    class-map type qos match-any class-vmdc-p4
      match cos 4

    class-map type qos match-any class-all-flood
      match all flood

    class-map type qos match-any class-ip-multicast
      match ip multicast

    class-map type qos match-any class-vmdc-priority
      match cos 5-7

c02-n5k-ics1-A# show policy-map vmdc-qos-policy

  Type qos policy-maps
  ====================

  policy-map type qos vmdc-qos-policy
    class type qos class-vmdc-priority
      set qos-group 5
    class type qos class-vmdc-p2
      set qos-group 2
    class type qos class-vmdc-p3
      set qos-group 3
    class type qos class-vmdc-p4
      set qos-group 4
    class type qos class-vmdc-p1
      set qos-group 1
    class type qos class-default
      set qos-group 0
dc02-n5k-ics1-A#

conf t
system qos
  service-policy type qos input vmdc-qos-policy


### Show commands
c02-n5k-ics1-A# show policy-map interface po534 type qos
```

```
Global statistics status :   enabled

 NOTE: Type qos policy-map configured on VLAN will take precedence
       over system-qos policy-map for traffic on the VLAN

port-channel534

  Service-policy (qos) input:   vmdc-qos-policy
    policy statistics status:   disabled

    Class-map (qos):   class-vmdc-priority (match-any)
      Match: cos 5-7
      set qos-group 5

    Class-map (qos):   class-vmdc-p2 (match-any)
      Match: cos 2
      set qos-group 2

    Class-map (qos):   class-vmdc-p3 (match-any)
      Match: cos 3
      set qos-group 3

    Class-map (qos):   class-vmdc-p4 (match-any)
      Match: cos 4
      set qos-group 4

    Class-map (qos):   class-vmdc-p1 (match-any)
      Match: cos 1
      set qos-group 1

    Class-map (qos):   class-default (match-any)
      Match: any
      set qos-group 0

dc02-n5k-ics1-A#
```

Nexus 5000 Network-QoS Policy Implementation

The goal of the network-QoS policy on the Nexus 5000 is to enable different treatment in terms of drop or no-drop as well as MTU. This is primarily to support the different treatment for FCoE and Ethernet traffic. In VMDC 2.3, there is no FCoE traffic on Ethernet links, and dedicated FC links are used for FC/SAN connectivity. The network-QoS configuration maps all classes of traffic to tail-drop behavior, i.e., not a no-drop class. The MTU is set to 1500B as this is the most common used MTU with Ethernet frames. If desired, jumbo frame support can be enabled at this level. Refer to Cisco Nexus 5000 Series NX-OS Quality of Service Configuration Guide, Release 5.1(3)N1(1) for more information.

```
Type network-qos class-maps
  ==============================

    class-map type network-qos class-default
      match qos-group 0

    class-map type network-qos class-vmdc-p1
      match qos-group 1

    class-map type network-qos class-vmdc-p2
      match qos-group 2

    class-map type network-qos class-vmdc-p3
      match qos-group 3
    class-map type network-qos class-vmdc-p4
      match qos-group 4

    class-map type network-qos class-ethernet
      match qos-group 5

    !!!! The following class exist by default.
    class-map type network-qos class-all-flood
      match qos-group 2

    !!!! The following class exist by default.
    class-map type network-qos class-ip-multicast
      match qos-group 2

    class-map type network-qos class-vmdc-priority
      match qos-group 5


policy-map type network-qos vmdc-nofc-nq-policy
    class type network-qos class-vmdc-priority

      mtu 1500
    class type network-qos class-vmdc-p2

      mtu 1500
    class type network-qos class-vmdc-p3

      mtu 1500
    class type network-qos class-vmdc-p4

      mtu 1500
    class type network-qos class-vmdc-p1

      mtu 1500
    class type network-qos class-default

      mtu 1500

!!! Network QoS policy is attached to the system qos while configuring

system qos
  service-policy type network-qos vmdc-nofc-nq-policy


dc02-n5k-ics1-A# show policy-map system  type network-qos


  Type network-qos policy-maps
  ===============================

  policy-map type network-qos vmdc-nofc-nq-policy
    class type network-qos class-vmdc-priority
      match qos-group 5

      mtu 1500
    class type network-qos class-vmdc-p2
      match qos-group 2

      mtu 1500
    class type network-qos class-vmdc-p3
      match qos-group 3

      mtu 1500
```

```
class type network-qos class-vmdc-p4
  match qos-group 4

  mtu 1500
class type network-qos class-vmdc-p1
  match qos-group 1

  mtu 1500
class type network-qos class-default
  match qos-group 0

  mtu 1500
```

Nexus 5000 Queuing Policy Implementation

The queuing policy applied on the Nexus 5000 provides differentiated treatment to the following traffic types:

1. Priority queuing for CoS5, 6, and 7 traffic, which maps to VoIP and real-time tenant services, network-control, and network-management traffic.
2. Bandwidth guarantee for CoS2 traffic for bandwidth guaranteed tenant traffic and the premium data class for Gold and Silver in-contract traffic.
3. Bandwidth guarantee for CoS3 for call-signaling services.
4. NFS traffic with a bandwidth guarantee.
5. Best effort for standard data class traffic for Bronze and Copper tenants; a small amount is reserved to avoid starving this class (5%).
6. Left-over bandwidth for Gold/Silver out-of-contract bandwidth.

The QoS policy classifies traffic based on dot1p CoS and marks each packet with an internal marking called QoS-group. The queuing policy uses QoS-group markings to provide appropriate queuing behavior. See the configuration snippets below for the implementation.

The queuing policy is applied under system QoS to apply as the default policy for all ports. This makes the configuration simple, however, on the ports facing storage, since no other traffic other than NFS traffic is expected, a different ingress and egress policy is applied, giving 90% of bandwidth to NFS class traffic. A small amount of bandwidth is reserved for CoS6 and CoS7, which is the VMDCPQ class in case any network control and management marked traffic is used towards storage. Note that the VMDC-PQ class is not given priority switching in this direction, as no VoIP or real-time application is expected in this segment, and only CoS6 and CoS7 are expected.

For overriding the system QoS queuing policy on NAS facing interfaces, configuring this specific queuing policy for ingress and egress under the port-channels facing NAS is required.

```
dc02-n5k-ics1-A# show class-map type queuing

  Type queuing class-maps
  =======================

    dc02-n5k-ics1-A# show class-map type queuing

  Type queuing class-maps
  =======================

    class-map type queuing vmdc-p1
      match qos-group 1

    class-map type queuing vmdc-p2
```

```
      match qos-group 2

    class-map type queuing vmdc-p3
      match qos-group 3

    class-map type queuing vmdc-p4
      match qos-group 4

    class-map type queuing vmdc-pq
      match qos-group 5

!!! Exists by default, not used in vmdc2.3
    class-map type queuing class-fcoe
      match qos-group 1

    class-map type queuing class-default
      match qos-group 0

!!! Exists by default, not used in vmdc2.3
    class-map type queuing class-all-flood
      match qos-group 2

!!! Exists by default, not used in vmdc2.3
    class-map type queuing class-ip-multicast
      match qos-group 2

!!! These queuing policy are applied at system level so all interfaces get this policy
!!! NAS facing interfaces have a different queuing policy as shown later below.

dc02-n5k-ics1-A# show policy-map type queuing
policy-map type queuing vmdc-ethernet-in-policy
    class type queuing vmdc-pq
      priority
    class type queuing vmdc-p2
      bandwidth percent 70
    class type queuing vmdc-p3
      bandwidth percent 10
    class type queuing vmdc-p4
      bandwidth percent 10
    class type queuing vmdc-p1
      bandwidth percent 0
    class type queuing class-default
      bandwidth percent 10
  policy-map type queuing vmdc-ethernet-out-policy
    class type queuing vmdc-pq
      priority
    class type queuing vmdc-p2
      bandwidth percent 70
    class type queuing vmdc-p3
      bandwidth percent 10
    class type queuing vmdc-p4
      bandwidth percent 10
    class type queuing vmdc-p1
      bandwidth percent 0
    class type queuing class-default
      bandwidth percent 10

!!! Facing NAS, COS4 is given most of bw, and some reservation for COS5,6,7 as well as
COS0 traffic
!!! These policies are applied to the interface facing NAS storage and overrides
system qos config

policy-map type queuing vmdc-nas-out-policy
    class type queuing vmdc-p4
```

```
      bandwidth percent 90
    class type queuing vmdc-pq
      bandwidth percent 5
    class type queuing class-default
      bandwidth percent 5

 policy-map type queuing vmdc-nas-in-policy
    class type queuing vmdc-p4
      bandwidth percent 90
    class type queuing vmdc-pq
      bandwidth percent 5
    class type queuing class-default
      bandwidth percent 5

 dc02-n5k-ics1-A# show run | b system
system qos
  service-policy type qos input vmdc-qos-policy
  service-policy type queuing input vmdc-ethernet-in-policy
  service-policy type queuing output vmdc-ethernet-out-policy
  service-policy type network-qos vmdc-nofc-nq-policy
 dc02-n5k-ics1-A# show run int port-ch26

!Command: show running-config interface port-channel26
!Time: Wed Mar  6 11:44:07 2013

version 5.2(1)N1(2)

interface port-channel26
  description vPC to netapp -A
  switchport mode trunk
  switchport trunk allowed vlan 1990
  service-policy type queuing input vmdc-nas-in-policy
  service-policy type queuing output vmdc-nas-out-policy
  vpc 26

dc02-n5k-ics1-A#
```

ASA QoS

On the ASA firewall, there is no specific configuration required for QoS. By default, the ASA preserves IP/DSCP, and also the dot1p CoS bits for traffic transiting through the firewall.

ACE QoS

In VMDC 2.3, ACE 4710 appliances are used to implement SLB.

On the ACE 4710, for traffic transiting through it, the IP/DSCP is preserved, i.e., copied, however, dot1p CoS bits are not preserved for L7 load-balanced traffic, but are preserved for L4 load-balanced traffic. Any L7 load-balancing traffic will not preserve the CoS that was marked on it as it transits the ACE 4710 and will then be treated as standard data class (best-effort Per-Hop Behavior (PHB)) in the rest of the network.

For traffic that is load balanced, in the north to south direction, i.e., from outside the DC into the DC, the SLB transit happens close to the server VM, and hence does not create a problem. Return traffic going through the ACE 4710 SLB (with L7 LB config), however, will get its CoS marking reset to 0, and hence DSCP also needs to be marked as shown in the Nexus 1000V QoS section. Due to this reason, QoS transparency cannot be achieved.

For deployments that use the ACE 4710 and other appliances that do not support dot1p CoS bits preservation, edge marking based on DSCP is used for the Nexus 1000V. Additionally, for the ports facing the ACE 4710, on the Nexus 7004 where they are attached, the policy classifies based on DSCP.

ASR 1000 QoS

The ASR 1000 router is used as a DC-PE router and sits on the boundary of the DC cloud and MPLS cloud. The ASR 1000 router provides hardware-based QoS packet-processing functionality. QoS features are enabled through the Modular QoS Command-Line Interface (MQC) feature. The MQC is a Command Line Interface (CLI) that allows users to create traffic polices and attach these polices to interfaces. The QoS requirements of a DC-PE are to support the classes of service used by the Service Provider and to enforce per-tenant service level agreements.

- ASR 1000 DC-PE WAN Egress QoS, page 6-31
- ASR 1000 DC-PE WAN Ingress QoS, page 6-33
- ASR 1000 DC-PE DC Egress HQoS, page 6-35
- ASR 1000 DC-PE DC Ingress QoS, page 6-43

ASR 1000 DC-PE QoS Implementation

The ASR 1000 router is used as a DC-PE. The DC-PE is the demarcation between the MPLS cloud (for example, SP-NGN network) and the DC cloud, and implements and maps services and associated QoS between the WAN/SP-NGN QoS domain and the DC QoS domain. The QoS implementation supports a per-tenant SLA, which is a concatenation of the WAN/SP-NGN SLA and DC SLA. The ASR 1000 DC-PE router enforces the edge service level agreements for both domains.

The ASR 1000 router serves as an MPLS PE router. The Internet Engineering Task Force (IETF) has defined three MPLS QoS models to tunnel the DiffServ information, the pipe model, short pipe model, and uniform model.

1. In the **pipe model**, the EXP bit can be copied from the IP precedence or set through configuration on the ingress Label Switching Router (LSR). On a P router, EXP bits are propagated from incoming label to outgoing label. On the egress LSR, the forwarding treatment of the packet is based on the MPLS EXP, and EXP bits are not propagated to the IP precedence.
2. The **short pipe** model is similar to the pipe model with one difference. On the egress LSR, the forwarding treatment of the packet is based on the IP precedence, and EXP information is not propagated to the IP precedence.
3. In the **uniform** model, the MPLS EXP information must be derived from the IP precedence on the ingress LSR. On a P router, the EXP bits are propagated from incoming label to outgoing label. On the egress LSR, the MPLS EXP information must be propagated to the IP precedence.

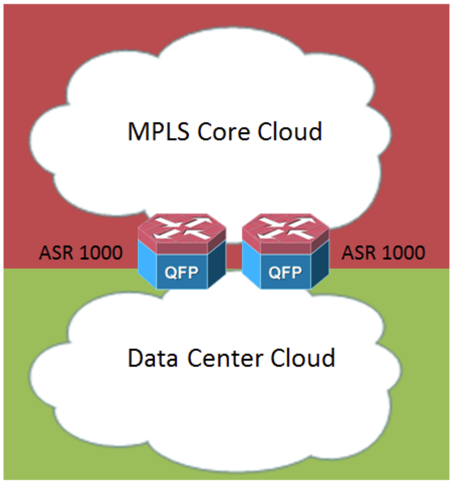
In this solution, the pipe model is used. All markings are based on Service Provider classification and use outer header QoS markings to support RFC3270, and the IP/DSCP or precedence is left untouched to support QoS transparency.

Refer to Quality of Service (QoS) for ASR 1000 QoS documentation.

ASR 1000 DC-PE WAN Egress QoS

The ASR 1000 DC-PE serves as the demarcation point between the IP/MPLS cloud and the DC cloud. It is also the boundary of QoS for the SP-DC QoS domain and the SP-NGN QoS domain (Figure 6-6).

Figure 6-6 Demarcation Point Between MPLS Cloud and Data Center Cloud



The following treatment is applied:

1. **Classification—**Use MPLS-TC for MPLS traffic. Packets are marked with `set mpls exp imposition <mpls-tc>` on the ingress from the DC interface, which is used to classify on egress.
2. **Marking—**No further marking is required. MPLS-TC is already set based on the DC ingress policies.
3. **Priority Class Traffic—**VoIP traffic uses the priority queue and is strictly policed.
4. All other classes get a bandwidth guarantee using bandwidth percent statement.
5. The WAN uplinks are 10GE Ethernet links. In VMDC 2.3, port-channels are not used on the ASR 1000. For VMDC 2.3 sizing, the bandwidth of one 10GE interface uplink to the MPLS-Core is expected. For customers that want to do more than 10GE, multiple links may be used, however, port-channels are not recommended as the ASR 1000 does not support QoS configuration on Gigabit EtherChannel bundles for flat-interface level policies.
6. Congestion avoidance using WRED on premium data class and standard data class (class-default).
7. For the premium data class, out-of-contract traffic is dropped before in-contract traffic during congestion when WRED kicks in.

Class-maps

```
class-map match-any cmap-premdata-dscp-exp
 match mpls experimental topmost 1  2
 match dscp cs1  cs2

class-map match-any cmap-callctrl-dscp-exp
 match mpls experimental topmost 3
 match dscp cs3
! Future use of video
class-map match-any cmap-video-dscp-exp
 match mpls experimental topmost 4
 match dscp cs4
class-map match-any cmap-voip-dscp-exp
 match dscp cs5
 match mpls experimental topmost 5
class-map match-any cmap-ctrl-dscp-exp
 match dscp cs6
 match mpls experimental topmost 6
```

```
class-map match-any cmap-mgmt-dscp-exp
 match dscp cs7
 match mpls experimental topmost 7
```

Policy-map

```
policy-map wan-out
 class cmap-voip-dscp-exp
  police rate percent 15
  priority level 1
 class cmap-premdata-dscp-exp
  bandwidth percent 60
  random-detect discard-class-based
 class cmap-callctrl-dscp-exp
  bandwidth percent 2
 class cmap-ctrl-dscp-exp
  bandwidth percent 4
 class cmap-mgmt-dscp-exp
  bandwidth percent 4
 class class-default
  random-detect
  bandwidth percent 15
```

ASR 1000 DC-PE WAN Ingress QoS

The following treatment is applied:

1. **Classification**
    a. **Ingress MPLS Traffic** Based on MPLS-TC (formerly known as EXP) bits in the MPLS label. The SP-NGN uses up to eight classes of traffic, but in this phase, only seven are used (the video service class is reserved for future use). The remote PE router classifies and marks with MPLS-TC. This is the far-end PE router at the edge of the SP-NGN where the customer's edge router connects.
    b. **Ingress IP Traffic** Based on IP/DSCP bits or IP/precedence bits. The SP-NGN uses eight classes of traffic. All tenant traffic arrives via MPLS in this phase except for Internet-based traffic for SSL/IPsec VPN access, which will arrive in the Internet class.
2. **Marking—** QoS-group marking is set one-to-one to each of the Service Provider classes. This is to support the pipe model to do egress into SP-DC using ingress MPLS-TC-based classification. On the DC facing egress interface, the QoS-group is used to classify, and corresponding CoS markings are added for the DC to use for classification.
3. The VoIP traffic class is marked as priority to enable all traffic in this class to use priority queuing.
4. The WAN uplinks are 10GE Ethernet links. In VMDC 2.3, port-channels are not used on the ASR 1000. For VMDC 2.3 sizing, the bandwidth of one 10GE interface uplink to the MPLS-Core is expected. For customers that want to do more than 10GE, multiple links may be used, however port-channels are not recommended as the ASR 1000 does not support QoS configuration on Gigabit EtherChannel bundles for flat interface level policies.

Class-maps

```
class-map match-any cmap-premdata-out-dscp-exp
 match mpls experimental topmost 1
 match dscp cs1

class-map match-any cmap-premdata-in-dscp-exp
 match mpls experimental topmost 2
 match dscp cs2
```

```
class-map match-any cmap-callctrl-dscp-exp
 match mpls experimental topmost 3
 match dscp cs3
class-map match-any cmap-video-dscp-exp
 match mpls experimental topmost 4
 match dscp cs4
class-map match-any cmap-voip-dscp-exp
 match dscp cs5
 match mpls experimental topmost 5

class-map match-any cmap-ctrl-dscp-exp
 match dscp cs6
 match mpls experimental topmost 6

class-map match-any cmap-mgmt-dscp-exp
 match dscp cs7
 match mpls experimental topmost 7
!
```

Policy-map

```
Policy Map wan-in
    Class cmap-mgmt-dscp-exp
      set qos-group 7
    Class cmap-ctrl-dscp-exp
      set qos-group 6
    Class cmap-voip-dscp-exp
      set qos-group 5
    Class cmap-video-dscp-exp
      set qos-group 4
    Class cmap-callctrl-dscp-exp
      set qos-group 3
    Class cmap-premdata-in-dscp-exp
      set qos-group 2
      set discard-class 2
    Class cmap-premdata-out-dscp-exp
      set qos-group 1
      set discard-class 1
    Class class-default
      set qos-group 0
```

The policy is attached to the core facing uplink. Additionally, MPLS-TC=5 traffic is mapped to use the high-priority queue internally from the ASR 1000 SIP to the ASR 1000 ESP. See the following configuration and **show** command used to verify this policy.

```
dc02-asr1k-pe2#sh run int ten0/0/0
Building configuration...

Current configuration : 298 bytes
!
interface TenGigabitEthernet0/0/0
 description uplink-to-core
 ip address 10.5.22.1 255.255.255.0
 ip ospf 1 area 0
 load-interval 30
 carrier-delay up 30
 plim qos input map mpls exp  5  queue strict-priority
 mpls ip
 cdp enable
 service-policy input wan-in
 service-policy output wan-out
end
dc02-asr1k-pe2#show platform hardware interface TenGigabitEthernet0/0/0 plim qos input
map
```

```
Interface TenGigabitEthernet0/0/0
    Low Latency Queue(High Priority):
        IP PREC, 6, 7
        IPv6 TC, 46
        MPLS EXP, 5, 6, 7
```

ASR 1000 DC-PE DC Egress HQoS

The ASR 1000 DC-PE implements Hierarchical QoS (HQoS) to treat tenant traffic going into the DC. It is a policy enforcement boundary for implementing per-tenant SLA. The primary purpose of QoS in this direction is to differentiate and support agreed upon service level agreements. Traffic coming from the MPLS-Core has already been classified and marked with MPLS-TC, and the DC-PE will map this to QoS-group markings. The egress into DC will trust these QoS-group markings and treat traffic for each class appropriately to ensure SLA under congestion.

Traffic in this direction is going into the DC and might be arriving from different remote sites. The total aggregate traffic getting into the DC might be higher than the commit rate, and thus, cause congestion. For example, in the test setup for VMDC 2.3, each of the ASR 1000 PEs has a 10GE connection to the MPLS-Core, as well as a 10 GE connection between them. Each ASR 1000 has a connection to the two DC-Agg switches. The Nexus 7000s, with all tenants preferring one of the DCAgg for load balancing, however, during failure conditions, for example if the other ASR 1000 DC-PE loses its links to the Nexus 7000, its possible that this traffic also arrives and congests the link between the first DC-PE and DC-AGG layer. The key requirement here is that each tenant receives the guaranteed bandwidth and latency as contractually agreed in the SLA under conditions of congestion.

The following treatment is applied:

1. Policies are tenant specific and configured on the tenant subinterfaces. The number of different policies needed maps to the number of tenant types supported - in this implementation four classes of tenants are supported - Gold, Silver, Bronze and Copper, and hence four types of polices are needed. Copper tenants access the DC over a shared context such as Internet, so the Copper traffic and Gold Demilitarized Zone (DMZ)-bound traffic is treated by the Internet to DC policy.
2. **Classification.** QoS-group is expected to be set on packets in the egress direction into the DC. See the WAN ingress policy to see how this is marked, based on MPLS-TC.
3. **Marking.** Packets are marked with dot1p CoS settings by configuring **set cos** under each class, which maps the Service Provider class markings. Please note that the premium data class traffic has both in-contract and out-contract traffic in the same class, and WRED is applied to drop out-of-contract traffic first. Yet, as traffic egresses into the DC, there is just one marking (CoS=2) for all premium data class traffic. This is because it is not possible to mark two different CoS values for traffic in the same class, which it needs to do for WRED.
4. The parent class is shaped to the PIR agreed upon for each tenant type. Gold is allowed to burst up to 3 Gbps, Silver up to 2 Gbps, and Bronze up to 1 Gbps. Internet traffic is also shaped to 3 Gbps.
5. Excess bandwidth is shared using a bandwidth remaining ratio. The weights are 300 for Internet, four for each Gold tenant, two for each Silver tenant, and one for each Bronze tenant.
6. In this implementation, port-channels are not used, as the DC-PE has connectivity only of 10GE to each DC-Agg Nexus 7000, as well as only 10GE connectivity to the MPLS-Core. Also, the ASR 1000 does not currently support flow-based QoS service policy configuration for ingress QoS.
7. The ASR 1000 DC-PE is connected to two different Nexus 7000 DC-Agg routers towards the DC, however, by controlling BGP routing, only one route is installed, i.e., there is no equal cost multiple path. Even though the exact same policy is repeated on both links, there is no duplication of policed rates or CIR rates.

Class-maps

```
class-map match-all cmap-callctrl-qg
 match qos-group 3

class-map match-all cmap-voip-qg
 match qos-group 5

class-map match-any cmap-premdata-qg
 match qos-group 2
 match qos-group 1
```

Policy-maps

Each tenant type has a separate policy-map to treat traffic appropriately.

Gold Policy-map

1. The Gold child policy allows three classes, VoIP, call control, and premium data.
2. VoIP is priority queued and strictly policed.
3. Call control is given a bandwidth guarantee.
4. Premium data receives a bandwidth guarantee.
5. WRED is used for congestion avoidance for the premium data class.
6. Out-of-contract traffic is dropped before in-contract traffic is dropped when WRED kicks in.

```
policy-map gold-out-parent
 class class-default
  shape peak 3000000000
  bandwidth remaining ratio 4
   service-policy gold-out-child
!


policy-map gold-out-child
 class cmap-voip-qg
  priority level 1
  police rate 100000000
  set cos 5
 class cmap-premdata-qg
  bandwidth 500000
  queue-limit 100 ms
  random-detect discard-class-based
  random-detect discard-class 1 40 ms 80 ms
  random-detect discard-class 2 80 ms 100 ms
  set cos 2
 class cmap-callctrl-qg
  bandwidth 1000
  set cos 3
 class class-default
  set cos 0
  random-detect
!
dc02-asr1k-pe2#sh run int ten0/2/0.201
Building configuration...

Current configuration : 331 bytes
!
interface TenGigabitEthernet0/2/0.201
 encapsulation dot1Q 201
 vrf forwarding customer_gold1
 ip address 10.1.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input gold-in
 service-policy output gold-out-parent
end

dc02-asr1k-pe2#

dc02-asr1k-pe2#show policy-map interface ten0/2/0.201 output
 TenGigabitEthernet0/2/0.201

  Service-policy output: gold-out-parent

    Class-map: class-default (match-any)
      56857 packets, 3878343 bytes
      30 second offered rate 0000 bps, drop rate 0000 bps
      Match: any
      Queueing
      queue limit 12499 packets
      (queue depth/total drops/no-buffer drops) 0/0/0
      (pkts output/bytes output) 56857/3878343
      shape (peak) cir 3000000000, bc 12000000, be 12000000
      target shape rate 1705032704
      bandwidth remaining ratio 4

      Service-policy : gold-out-child

        queue stats for all priority classes:
          Queueing
          priority level 1
          queue limit 512 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0

        Class-map: cmap-voip-qg (match-all)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 5
          Priority: Strict, b/w exceed drops: 0

          Priority Level: 1
          police:
              rate 100000000 bps, burst 3125000 bytes
            conformed 0 packets, 0 bytes; actions:
              transmit
            exceeded 0 packets, 0 bytes; actions:
              drop
            conformed 0000 bps, exceeded 0000 bps
          QoS Set
            cos 5
              Packets marked 0

        Class-map: cmap-premdata-qg (match-any)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 2
            0 packets, 0 bytes
            30 second rate 0 bps
          Match: qos-group 1
            0 packets, 0 bytes
            30 second rate 0 bps
          Queueing
          queue limit 100 ms/ 6250000 bytes
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0
          bandwidth 500000 kbps

            Exp-weight-constant: 9 (1/512)
            Mean queue depth: 0 ms/ 2464 bytes
           discard-class       Transmitted         Random drop      Tail drop
Minimum        Maximum     Mark
                 pkts/bytes           pkts/bytes      pkts/bytes         thresh
thresh     prob
                                                                         ms/bytes
ms/bytes
            0               0/0               0/0              0/0
25/1562500      50/3125000   1/10
            1               0/0               0/0              0/0
40/2500000      80/5000000   1/10
            2               0/0               0/0              0/0
80/5000000     100/6250000   1/10
            3               0/0               0/0              0/0
34/2148437      50/3125000   1/10
            4               0/0               0/0              0/0
37/2343750      50/3125000   1/10
            5               0/0               0/0              0/0
40/2539062      50/3125000   1/10
            6               0/0               0/0              0/0
43/2734375      50/3125000   1/10
            7               0/0               0/0              0/0
46/2929687      50/3125000   1/10
          QoS Set
            cos 2
              Packets marked 0

        Class-map: cmap-callctrl-qg (match-all)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 3
          Queueing
          queue limit 64 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0
          bandwidth 1000 kbps
          QoS Set
            cos 3
              Packets marked 0

        Class-map: class-default (match-any)
          56857 packets, 3878343 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: any

          queue limit 12499 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 56857/3878343
          QoS Set
            cos 0
              Packets marked 56857
            Exp-weight-constant: 4 (1/16)
            Mean queue depth: 1 packets
            class       Transmitted         Random drop      Tail drop
Minimum        Maximum     Mark
                 pkts/bytes           pkts/bytes      pkts/bytes         thresh
thresh     prob
```

```
           0          1825/116704        0/0            0/0             3124
6249  1/10
           1             0/0              0/0            0/0             3514
6249  1/10
           2             0/0              0/0            0/0             3905
6249  1/10
           3             0/0              0/0            0/0             4295
6249  1/10
           4             0/0              0/0            0/0             4686
6249  1/10
           5             0/0              0/0            0/0             5076
6249  1/10
           6         55032/3761639        0/0            0/0             5467
6249  1/10
           7             0/0              0/0            0/0             5857
6249  1/10
dc02-asr1k-pe2#
```

Silver Policy-map

1. The Silver child policy allows just one class, premium data.
2. Premium data receives a bandwidth guarantee.
3. WRED is used for congestion avoidance for the premium data class.
4. Out-of-contract traffic is dropped before in-contract traffic is dropped when WRED kicks in.

```
policy-map silver-out-parent
 class class-default
  shape peak 2000000000
  bandwidth remaining ratio 2
   service-policy silver-out-child
!
policy-map silver-out-child
 class cmap-premdata-qg
  bandwidth 250000
  queue-limit 100 ms
  random-detect discard-class-based
  random-detect discard-class 1 40 ms 80 ms
  random-detect discard-class 2 80 ms 100 ms
  set cos 2
 class class-default
!
dc02-asr1k-pe2#sh run int ten0/2/0.501
Building configuration...

Current configuration : 337 bytes
!
interface TenGigabitEthernet0/2/0.501
 encapsulation dot1Q 501
 vrf forwarding customer_silver1
 ip address 10.2.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input silver-in
 service-policy output silver-out-parent
end

dc02-asr1k-pe2#
```

Bronze Policy-map

1. The Bronze class uses class-default (standard data class), and has WRED configured for congestion avoidance.
2. Optionally, the Bronze class can be bandwidth remaining value only, so that there is no bandwidth reservation, and only the bandwidth remaining can be allotted to Bronze tenants to support the no reservation model. A bandwidth reservation of 100 Mbps per Bronze tenant was configured in this solution.

```
policy-map bronze-out-parent
 class class-default
  shape peak 1000000000
  bandwidth remaining ratio 1
   service-policy bronze-out-child
!

policy-map bronze-out-child
 class class-default
  queue-limit 100 ms
  bandwidth 100000
  random-detect
  set cos 0
!

dc02-asr1k-pe2#show run int ten0/2/0.801
Building configuration...

Current configuration : 337 bytes
!
interface TenGigabitEthernet0/2/0.801
 encapsulation dot1Q 801
 vrf forwarding customer_bronze1
 ip address 10.3.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input bronze-in
 service-policy output bronze-out-parent
end
```

Internet Policy-map

1. Internet to DC traffic uses a policy that looks very similar to the Gold policy. This is because Gold DMZ traffic comes in from the Internet, and hence all Gold traffic classes are configured. Also, Copper traffic comes in and has to be supported in this policy as a standard data class.
2. The total of all Internet to DC traffic is shaped to 3 Gbps in the following configs, but it is really optional depending on the deployment scenario as this shaping is not per tenant level. In most deployment cases, it would be better to allow incoming traffic to consume any available bandwidth.
3. The rate-limiting for VoIP is not done at a tenant level, however, for example, a total of 100 Mbps of VoIP class traffic is allowed from the Internet to DC, but no specific per-tenant limits are configured and enforced. Enforcing per-tenant limits can be done using class-maps based on the ACL to identify traffic bound to tenant destination addresses (not shown in this document).
4. For the total of all Internet to Gold DMZ of all tenants put together in a premium data class, a bandwidth guarantee of 500 Mbps is configured. For all of the Copper tenants put together, the total bandwidth guarantee is 100 Mbps.

```
policy-map internet-out-parent
 class class-default
  shape peak 3000000000
  bandwidth remaining ratio 300
```

```
   service-policy internet-out-child
 policy-map internet-out-child
 class cmap-voip-qg
  priority level 1
  set cos 5
  police rate 100000000
 class cmap-premdata-qg
  bandwidth 500000
  queue-limit 100 ms
  random-detect discard-class-based
  random-detect discard-class 1 40 ms 80 ms
  random-detect discard-class 2 80 ms 100 ms
  set cos 2
 class cmap-callctrl-qg
  bandwidth 1000
  set cos 3
 class class-default
  set cos 0
  bandwidth 100000
c02-asr1k-pe2#

dc02-asr1k-pe2#show run int ten0/2/0.2000
Building configuration...

Current configuration : 277 bytes
!
interface TenGigabitEthernet0/2/0.2000
 encapsulation dot1Q 2000
 ip address 100.200.0.9 255.255.255.252
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 cdp enable
 service-policy input internet-in
 service-policy output internet-out-parent
end

dc02-asr1k-pe2#show policy-map int ten0/2/0.2000 output
 TenGigabitEthernet0/2/0.2000

  Service-policy output: internet-out-parent

    Class-map: class-default (match-any)
      423658096 packets, 633710961578 bytes
      30 second offered rate 0000 bps, drop rate 0000 bps
      Match: any
      Queueing
      queue limit 12499 packets
      (queue depth/total drops/no-buffer drops) 0/0/0
      (pkts output/bytes output) 423658096/633710961578
      shape (peak) cir 3000000000, bc 12000000, be 12000000
      target shape rate 1705032704
      bandwidth remaining ratio 300

      Service-policy : internet-out-child

        queue stats for all priority classes:
          Queueing
          priority level 1
          queue limit 512 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0

        Class-map: cmap-voip-qg (match-all)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 5
          Priority: Strict, b/w exceed drops: 0

          Priority Level: 1
          QoS Set
            cos 5
              Packets marked 0
          police:
              rate 100000000 bps, burst 3125000 bytes
            conformed 0 packets, 0 bytes; actions:
              transmit
            exceeded 0 packets, 0 bytes; actions:
              drop
            conformed 0000 bps, exceeded 0000 bps

        Class-map: cmap-premdata-qg (match-any)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 2
            0 packets, 0 bytes
            30 second rate 0 bps
          Match: qos-group 1
            0 packets, 0 bytes
            30 second rate 0 bps
          Queueing
          queue limit 100 ms/ 6250000 bytes
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0
          bandwidth 500000 kbps

            Exp-weight-constant: 9 (1/512)
            Mean queue depth: 0 ms/ 0 bytes
            discard-class       Transmitted         Random drop      Tail drop
Minimum        Maximum     Mark
                 pkts/bytes           pkts/bytes      pkts/bytes         thresh
thresh     prob
                                                                         ms/bytes
ms/bytes
            0               0/0               0/0              0/0
25/1562500      50/3125000   1/10
            1               0/0               0/0              0/0
40/2500000      80/5000000   1/10
            2               0/0               0/0              0/0
80/5000000     100/6250000   1/10
            3               0/0               0/0              0/0
34/2148437      50/3125000   1/10
            4               0/0               0/0              0/0
37/2343750      50/3125000   1/10
            5               0/0               0/0              0/0
40/2539062      50/3125000   1/10
            6               0/0               0/0              0/0
43/2734375      50/3125000   1/10
            7               0/0               0/0              0/0
46/2929687      50/3125000   1/10
          QoS Set
            cos 2
              Packets marked 0

        Class-map: cmap-callctrl-qg (match-all)
          0 packets, 0 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: qos-group 3
          Queueing
          queue limit 64 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 0/0
          bandwidth 1000 kbps
          QoS Set
            cos 3
              Packets marked 0

        Class-map: class-default (match-any)
          423658096 packets, 633710961578 bytes
          30 second offered rate 0000 bps, drop rate 0000 bps
          Match: any
          Queueing
          queue limit 416 packets
          (queue depth/total drops/no-buffer drops) 0/0/0
          (pkts output/bytes output) 423658096/633710961578
          QoS Set
            cos 0
              Packets marked 423658096
          bandwidth 100000 kbps
dc02-asr1k-pe2#
```

ASR 1000 DC-PE DC Ingress QoS

For traffic from the DC to the MPLS network ingressing into the DC-PE, the following treatment is applied:

1. CoS dot1p bits are used for classification. Eight classes can be supported, and seven are used in this phase. It is expected that this traffic from the DC to the MPLS network is already marked with correct CoS values and the DC-PE can trust CoS. The Nexus 1000V virtual switch sets the correct CoS for any traffic originating from tenant VMs, however, for traffic going through a load balancer such as the ACE 4710 tested in VMDC 2.3, CoS is not preserved when it transits the SLB. This traffic comes into the DC-PE with CoS0 and receives classified into class-default and is treated with best-effort PHB.
2. Tenant service level agreements are enforced on the ASR 1000 DC-PE. Each tenant can send contractually agreed upon bandwidth per class of traffic into the WAN/ NGN network, and policing is applied to enforce this limit as the DC-PE is the boundary between DC and WAN/SPNGN.
3. Three different tenant types receive different service level agreements applied to the subinterfaces for the specific tenant types. Additionally, a policy for the Internet subinterface is required. The Internet subinterface will see traffic from all Copper tenants, as well as all Gold tenants for their DMZ VMs. Also, Gold tenants have VPN remote access over the Internet.
4. For VoIP, which is treated as priority traffic, strict policing is applied, and exceed/violate traffic is dropped.
5. For the premium data class, conditional marking is done to indicate that in-contract and out-ofcontract traffic are marked down.
6. Bronze tenants receive the standard data class, which uses class-default.
7. Traffic bound to the MPLS network is marked with appropriate MPLS-TC markings.
8. In this implementation, port-channels are not used, as the DC-PE has connectivity only of 10GE to each DC-Agg Nexus 7000 as well as only 10GE connectivity to the MPLS-Core. Also, the

ASR 1000 does not currently support flow-based QoS service policy configuration for ingress QoS.

To configure ingress QoS, complete the following steps:

Step 1 Configure classification settings using the following configuration example commands:

```
class-map match-all cmap-voip-cos
 match cos  5
class-map match-all cmap-ctrl-cos
 match cos  6
class-map match-all cmap-callctrl-cos
 match cos  3
class-map match-all cmap-mgmt-cos
 match cos  7
class-map match-any cmap-premdata-cos
 match cos  1
 match cos  2
```

Step 2 The policy-maps for Gold include three classes of traffic, VoIP, call control, and data. Gold customers receive the premium data class, and the CIR=500 Mbps with a PIR of 3 Gbps. VoIP and call control have strict policing limits with 1R2C.

```
policy-map gold-in
 class cmap-voip-cos
  police rate 1500000000
  set mpls experimental imposition 5
 class cmap-callctrl-cos
  police rate 10000000
  set mpls experimental imposition 3
 class cmap-premdata-cos
  police rate 500000000  peak-rate 3000000000
   conform-action set-mpls-exp-imposition-transmit 2
   conform-action set-discard-class-transmit 2
   exceed-action set-mpls-exp-imposition-transmit 1
   exceed-action set-discard-class-transmit 1
   violate-action drop
 class class-default
  set mpls experimental imposition 0
```

Step 3 The policy-map for Silver includes a single data class with a CIR of 250 Mbps with a PIR of 2 Gbps of premium data class traffic allowed.

```
policy-map silver-in
 class cmap-premdata-cos
  police rate 250000000  peak-rate 2000000000
   conform-action set-mpls-exp-imposition-transmit 2
   conform-action set-discard-class-transmit 2
   exceed-action set-mpls-exp-imposition-transmit 1
   exceed-action set-discard-class-transmit 1
   violate-action drop
 class class-default
  set mpls experimental imposition 0
!
```

Step 4 The policy-map for Bronze includes rate-limiting to a max of 100 Mbps of standard data class.

```
policy-map bronze-in
 class class-default
  set mpls experimental imposition 0
  police rate 100000000
!
```

Step 5 The policy-map for Internet includes 100 Mbps of VoIP/real time and 500 Mbps guaranteed of premium data with up to 3 Gbps peak rate (marked down above 500 Mbps). There is no reservation or rate-limiting for standard class where Copper tenant traffic will be classified. This traffic receives best-effort treatment and uses any available bandwidth.

```
policy-map internet-in
 class cmap-callctrl-cos
  police rate 10000000
  set mpls experimental imposition 3
 class cmap-voip-cos
  police rate 100000000
  set mpls experimental imposition 5
 class cmap-premdata-cos
  police rate 500000000  peak-rate 3000000000
   conform-action set-mpls-exp-imposition-transmit 2
   conform-action set-discard-class-transmit 2
   exceed-action set-mpls-exp-imposition-transmit 1
   exceed-action set-discard-class-transmit 1
   violate-action drop
 class class-default
  set mpls experimental imposition 0
!
```

Step 6 The tenant specific policy-map is applied on all of the tenants' subinterfaces connecting to the Nexus 7000 DC-Agg. For example:

```
! Example for Gold tenant
dc02-asr1k-pe2#show run int ten0/2/0.201
Building configuration...

Current configuration : 331 bytes
!
interface TenGigabitEthernet0/2/0.201
 encapsulation dot1Q 201
 vrf forwarding customer_gold1
 ip address 10.1.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input gold-in
 service-policy output gold-out-parent
end

dc02-asr1k-pe2#show run int ten0/2/0.501
Building configuration...

Current configuration : 337 bytes
!
interface TenGigabitEthernet0/2/0.501
 encapsulation dot1Q 501
 vrf forwarding customer_silver1
 ip address 10.2.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input silver-in
 service-policy output silver-out-parent
end

dc02-asr1k-pe2#show run int ten0/2/0.801
Building configuration...

Current configuration : 305 bytes
!
interface TenGigabitEthernet0/2/0.801
 encapsulation dot1Q 801
 vrf forwarding customer_bronze1
 ip address 10.3.3.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input bronze-in
 service-policy output bronze-out-parent
end

dc02-asr1k-pe2#show run int ten0/2/0.2000
Building configuration...

Current configuration : 201 bytes
!
interface TenGigabitEthernet0/2/0.2000
 encapsulation dot1Q 2000
 ip address 100.200.0.9 255.255.255.252
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 cdp enable
 plim qos input map cos5 queue strict-priority
 service-policy input internet-in
 service-policy output internet-out-parent
end
dc02-asr1k-pe2#
```

On the ASR 1000, the SIP to ESP connection has two queues, high priority and low Priority. To ensure low latency for CoS5 traffic, these packets with CoS5 need to also be mapped to use the high priority queue. By default, the ASR 1000 only does it for CoS6 and CoS7. See the following show command output that shows the mapping after configuring the following under the subinterface:

```
conf t
interface TenGigabitEthernet0/2/0.201
  plim qos input map cos5 queue strict-priority
!

dc02-asr1k-pe2#show platform hardware interface TenGigabitEthernet0/2/0.201 plim qos
input map
Interface TenGigabitEthernet0/2/0.201
    Low Latency Queue(High Priority):
        dot1Q COS, 5, 6, 7
```

QoS Best Practices and Caveats

Nexus 1000V QoS Best Practices

- Configure QoS policies to classify, mark, police, and prioritize traffic flows. Different traffic types should have different network treatment.

UCS QoS Best Practices

- Reserve bandwidth for each traffic type using QoS system class. Each type of traffic should have a guaranteed minimum bandwidth.
- For UCS servers deployed with the Nexus 1000V, it is highly recommended to do the CoS marking at the Nexus 1000V level. Configure UCS QoS policy with **Host Control Full** and attach the policy to all vNICs of UCS servers.

ACE QoS Best Practices and Caveats Caveats

- **CSCtt19577:** need ACE to preserve L7 traffic dot1p CoS

  - QoS transparency requires that DSCP not be touched, and that only CoS be used to support DC QoS in the VMDC system. The tenant uses DSCP for their markings, and the DC operator can use independent QoS markings by using dot1P CoS bits. To support this, both DSCP and dot1p CoS need to be preserved as packets transit the ACE, however, the ACE does not currently support CoS preservation for L7 traffic. This enhancement requests support for CoS preservation and DSCP preservation for all scenarios including L7 traffic.

Nexus 7000 QoS Best Practices and Caveats Best Practices

- The Nexus 7000 series uses four fabric queues across modules, and CoS values are mapped to these four queues statically, i.e., they cannot be changed. The priority queue for CoS5,6, and 7 is switched with strict priority, and the other three queues are switched with equal weights. The F2 cards used in VMDC 2.3 use the 8e-4q4q model, which class-maps that map to the CoS values in the same way as the fabric queues. This is particularly important as the F2 card uses buffers in the ingress card, and back pressure from the egress interface congestion is mapped to ingress queues. Packets are dropped at ingress when such congestion happens. It is important to use the 8e-4q4q model to track each class separately. This model is supported from NX-OS release 6.1.3 onwards.

Caveats

- **CSCue55938:** duplicating policy-maps for egress queuing.
  - Attaching two queuing policies for the same direction under a port is allowed under some conditions.
- **CSCud46159:** all interfaces in the module are gone after reboot
  - When a scaled up configuration with many interfaces is configured with the same policy-map that includes egress policing, upon reload, the Nexus 7004 aggregation switch loses its configuration of all interfaces. This workaround is to configure multiple policy-maps with the same policy and divide the total number of subinterfaces into three or four groups and attaching a different policy-map to each group.
- **CSCud26031:** F2: aclqos crash on configuring QoS policy on subinterfaces
  - ACLQOS crash is observed when attaching a service policy that includes egress policing on a large number of subinterfaces. The workaround is to use different policy-maps (with the same underlying policy) so that the number of subinterfaces using the same policy-map is reduced.
- **CSCud26041:** F2: scale QoS configs by not allocating policer stats when no policing
  - Qos per class stats use hardware resources that are shared with policers. On the F-series card, this is restricted to a small amount, i.e., currently 1024, which is the total of all classes in policies multiplied by attachments. For example, with an eight-class policy, only 128 attachments can be done on 128 subinterfaces on the same SoC. This bug requests disabling default per-class statistics collection and providing proper error messaging to indicate the actual issue. Statistics are enabled by default, and hence the workaround is to add **no-stats** to the service policy attachments.

Nexus 5000 QoS Best Practices and Caveats Best Practices

- Use all six classes of traffic for the Ethernet class if no FCoE traffic is expected.
- Account for NFS traffic at this layer of the DC, and provide a separate class and queuing to provide a BW guarantee.

Caveats

- **CSCue88052:** Consistency between Nexus 5000 and Nexus 7000 QoS config

  - Nexus 5500 Series switches currently have different semantics of similar sounding QoS configuration items, and this bug tracks specifically the fact that the Nexus 5500 allows the configuration of bandwidth percent for a class in a policy-map where priority is configured. Also, the bandwidth percent semantics in a policy-map that has priority class is actually called "bandwidth remaining." This is confusing and not consistent with the Nexus 7000 semantics, which have checks in place to prevent priority and bandwidth percent configuration for the same class in a policy-map.

ASR 1000 QoS Best Practices and Caveats **Best Practices**

- QoS on port-channel interfaces is not supported. For the MPLS-Core facing interfaces, port-channels are not recommended, as the VMDC 2.3 QoS policies cannot be implemented.
- QoS on port-channel subinterfaces have restrictions. For example, ingress QoS cannot be done in flow-based mode, and egress QoS requires a QoS configuration on the member links. The recommendation for VMDC 2.3 is to use multiple links between the DC-PE and DC-AGG if more than 10GE is required.
- NetFlow on the ASR 1000 series with custom NetFlow records can impact the switching performance. The recommendation is to use default NetFlow record formats. While this is not exactly a QoS best practice, this can impact QoS due to dropping of packets earlier than expected due to switching performance rather than actual link congestion.
- Mapping of priority traffic based on CoS and MPLS-TC to the high-priority queue between SIP and the ESP is required to provide priority traffic low latency treatment.
- Calculation of bandwidth requirement for both normal and failure cases should be accounted for, as the ASR 1000 is a centralized switching platform and all traffic is funneled and switched at the ESP. In this design, a SIP-40 is used with 4x10GE shared port adapters, and with ESP-40, which can handle 40 Gbps of switching. This provides 10 Gbps of traffic from north-south, and 10 Gbps of traffic from south-north, for a total of 20 Gbps for normal conditions. Different failure scenarios will not cause any oversubscription at the ESP-40.

Caveats

- **CSCud51708:** wrong calc for bytes w ms based queue-limit config after random-detect
  - If the queue-limit is configured in milliseconds after configuring random-detect, the bytes calculation is wrong for the specified number of milliseconds in the queue-limit. The workaround is to first configure the queue-limit in milliseconds and then configure random-detect.

CHAPTER 7

Resiliency and High Availability

The Virtualized Multiservice Data Center (VMDC) 2.3 design provides a number of High Availability (HA) features and is a highly resilient network. The following sections provide an overview of network resiliency and also summarize the validation results of convergence around service impacting failures as tested in the lab configuration.

This section presents the following topics:

Resiliency Against Link and Node Failure

HA has different aspects that are implemented at different layers in the network. The VMDC2.3 design does not have any single point of failure, and the service impacting failures are minimized by ensuring quick convergence around the failing link or node. In terms of converging around the failing link or node, this may be required as part of planned maintenance or an unplanned failure event. Planned events are most commonly done to upgrading software on various nodes in the Data Center (DC) and other maintenance reasons on power plants, and to address facilities issues.

In VMDC 2.3, the network portion has dual paths, with two nodes supporting each path, in an active/active configuration with load balancing of traffic achieved by using Border Gateway Protocol (BGP). During maintenance events on one node, wherein the node is taken down, the traffic and services can continue to be provided using the other path, however, there could be local congestion during such events, as one node going down would cause all traffic to use the other path. For example, when the Provider Edger (PE) node is down, all traffic uses the surviving PE and WAN link, which causes the bandwidth available for the entire DC to be reduced to half. This can be avoided by using dual-redundant route processors and the Encapsulating Standard Protocol (ESP) on the ASR 1006 DC-PE, and by using dual supervisors on the Nexus 7004 DC-Agg routers, which is our recommendation. In addition to the benefit of being able to perform In Service Software Upgrade (ISSU), any unexpected failure of the supervisors when configured with redundancy will cause automatic switchover to the the redundant RP/supervisor, and forwarding is minimally impacted. Similarly, it is highly recommended to deploy other services appliances and compute infrastructure, as well as the Nexus 1000V Virtual Supervisor Module (VSM) and Virtual Security Gateway (VSG) in a HA configuration with a pair of devices to support failover. Additionally, for redundancy on the link level on the Nexus 7004, two modules are used, and port-channels with members from both modules are used to provide service continuously for planned or unplanned events on each module.

Table 7-1 lists the redundancy model for the ASR 1006 and Nexus 7004.

Table 7-1 Redundancy Model for ASR 1006 and Nexus 7004

Event Type	Planned/Unplanned	Redundancy	Mechanism	Service Impact
Software upgrades	Planned	Not redundant	Routing convergence	Yes, convergence event during link/ node shut
Software upgrades	Planned	Redundant	HA/SSO	Minimum, zero packet loss in most conditions *
Software or hardware error	Unplanned	Not redundant	Routing convergence	Yes, convergence event
Software or hardware error	Unplanned	Redundant	HA/SSO	Minimum, zero packet loss in most conditions *

Note The ASR 1000 is impacted by CSCuc51879. This issue causes packet drops during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration and is still under investigation as of this publication.

For other nodes used in the VMDC 2.3-based DC, Table 7-2 lists the redundancy model to support not having a single point of failure.

Table 7-2 Redundancy Model for Services and Compute

Node Type	Role/Services	HA
ASA 5585	Firewall	FT using active/standby pair
ASA 5555-X	VPN	FT using active/standby pair
ACE 4710	SLB	FT using active/standby pair
Nexus 1010	Virtual Service Blades	Paired Nexus 1010 in active/ standby
Nexus 1000V VSM	Virtual Supervisor Module	Paired VSM in active/standby
UCS Fabric Interconnect 6248	Fabric Interconnect	Pair of FI/6248 devices in active/standby for management, active/active for data
ICS Switch Nexus 5000	ICS Access switch	Pair in virtual port-channel, no stateful sync
Compute Cluster	Compute	VMware HA/DRS cluster
FC Storage/NAS Storage	Storage	NetApp dual controllers
VSG	Compute Firewall	Pair of active/standby, statefully synced
VNMC	Compute Firewall Management	Use VMware HA

Convergence Test Results

Table 7-3 and Table 7-4 detail convergence results for ASR 1006 DC-PE, and Nexus 7004 aggregation switch convergence events.

1. For the network test scenarios, traffic was sent using traffic tools (IXIA, Spirent TestCenter) for all tenants north to south.

2. A convergence event was triggered for all tenants north to south.
3. MAC scale was set to 13,000 - 14,000 MAC addresses on the Nexus 7004 devices.
4. Additional traffic was sent between east/west tenants for the first 50 tenants.
5. The worst case impacted flow amongst all the flows is reported. It should be noted that all flows are not impacted due to alternate paths, which are not impacted during tests.

Table 7-3 ASR 1006 DC-PE Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Node fail	2.7-5.3 sec	0.3 sec		
2	Node recovery	zpl	zpl		
3	Link fail ASR 1000 to Nexus 7004 Agg	2.502 sec	4.628 sec		
4	Link restore ASR 1000 to Nexus 7004 Agg	zpl	zpl		
5	ASR 1000 RP switchover CLI	0 sec	0.5 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879[1]
6	ASR 1000 RP switchover - RP module pull	0	2.6-5 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879
7	ASR 1000 ISSU	15 sec	23 sec	With scaled up configuration on the ASR 1000 PE, some losses were seen instead of zpl	CSCuc51879
8	ASR 1000 ESP module pull	zpl	zpl		
9	ASR 1000 ESP switchover via CLI	zpl	zpl		

[1] The fix for this issue is still under investigation at the time of this publication.

Table 7-4 Nexus 7004 Aggregation Switch Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Nexus 7004 AGG module fail	3.6 sec	3.12 sec		2 3
2	Nexus 7004 AGG module restore	9.9 sec	10.9 sec		2 6
3	Nexus 7004 AGG node fail	1-2 sec	1-2 sec		
4	Nexus 7004 AGG node recovery	5 sec	8 sec	See Layer 3 Best Practices and Caveatsfor more information. Additional steps are needed to workaround the issue. 4 5	1 2 4 5

Table 7-4 Nexus 7004 Aggregation Switch Convergence Events (continued)

5	Nexus 7004 AGG vPC peer link fail	13 sec	2 sec	See Layer 3 Best Practices and Caveats for more information. BGP convergence will move traffic off the Nexus 7004 path. Future fixes to help with convergence.	1 2 5
6	Nexus 7004 AGG vPC peer link restore	3.5 sec	6.3 sec		2
7	Nexus 7004 AGG link to ICS Nexus 5548 SW fail	0.2 sec	0.2 sec	Fiber pull	
8	Nexus 7004 AGG link to ICS Nexus 5548 SW restore	0.2 sec	0.2 sec	Fiber restore	
9	Nexus 7004 AGG supervisor fail - module pull	zpl	zpl		
10	Nexus 7004 AGG supervisor switchover - CLI	zpl	zpl		
11	Nexus 7004 ISSU	zpl	zpl		

Note The following issues are being fixed in the Nexus 7000, but are not available in the release tested. These fixes are currently planned for release in the 6.2-based release of NX-OS for the Nexus 7000.

- [1]CSCtn37522: Delay in L2 port-channels going down
- [2]CSCud82316: vPC Convergence optimization
- [3]CSCuc50888: High convergence after F2 module OIR The following issue is under investigation by the engineering team:
- [4]CSCue59878: Layer 3 convergence delays with F2 module - this is under investigation. With scale tested, the additional delay is between 10-17s in the vPC shut case. The workaround used is to divert the traffic away from N7K Agg as control plane (BGP) does converge quicker and traffic bypasses the N7K agg. Also use the least amount of port groups possible to reduce the number of programming events. Alternatively, consider using M1/M2 modules for higher scale of prefixes and better convergence.

The following issues are closed without any fix, as this is the best convergence time with the F2 module after the workarounds are applied:

- [5]CSCue67104: Layer 3 convergence delays during node recovery (reload) of N7k Agg. The workaround is to use L3 ports in every port group to download FIB to each port-group.
- [6]CSCue82194:High Unicast convergence seen with F2E Module Restore

Table 7-5, Table 7-6, Table 7-7, and Table 7-8 detail convergence results for Nexus 5500 Series ICS switch, ACE 4710 and Nexus 7004, ASA 5585, and other convergence events.

Table 7-5 Nexus 5500 Series ICS Switch Convergence Events

	Event	N-S	S-N	Comments	Issues
1	Nexus 5548 ICS SW vPC peer link fail	1.7 sec	2.5 sec		
2	Nexus 5548 ICS SW vPC peer link restore	7.14 sec	9.35 sec		
3	Nexus 5548 ICS SW Node fail	1.7 sec	0.36 sec		
4	Nexus 5548 ICS SW Node Recovery	9.8 sec	8.5 sec		

Table 7-6 ACE 4710 and Nexus 7004 Convergence Events

	Event	N-S	S-N	Comment	Issues
1	ACE failover with CLI	0.072 sec	0.072 sec		
2	ACE node fail	9.3 sec	9.3 sec	ACE FT configured 10 sec	
3	ACE node recovery	5.7 sec	3.1 sec		
4	ACE Single link fail	0.5 sec	0.5 sec		
5	ACE Single link restore	0.05 sec	0.05 sec		
6	ACE Dual links to same Nexus 7000 fail	2.5 sec	2.5 sec		
7	ACE Dual link to same Nexus 7000 restore	2.5 sec	1.2 sec		
8	ACE Port-ch fail	13 sec	13 sec	ACE FT configured 10 sec	
9	ACE port-ch restore	10 sec	10 sec	ACE FT configured 10 sec	

Table 7-7 ASA 5585 Convergence Events

	Event	N-S	S-N	Comment	Issues
1	ASA FT link fail	zpl	zpl	Failover is disabled	
2	ASA FT link restore	zpl	zpl	Failover is disabled	
3	ASA reload	4-6 sec	4-6 sec	Failover pull time/hold time 1/3 sec	
4	ASA recovery	0.2-3 sec	0.2-3 sec	Default preemption	

Table 7-8 Other Convergence Events

	Events	Traffic Impact	Comments	Issues
1	UCS FI 6248 fail	2.20648 sec		
2	UCS FI 6248 restore	0.6909 sec		
3	UCS FT switchover	2.2954 sec when FT failed, 0.3893 sec when FT restored.	No packet drop if only the control/ management plane of the FT switchover	

Table 7-8 Other Convergence Events (continued)

4	Ethernet link fail between 6248 and 5500	Minimal impact seen.	VM running script continuously accessing a file (read/write) on NFS data store continues after a momentary stop	5
5	Ethernet link Restore between 6248 and 5500	No impact seen.	VM running script continuously accessing a file (read/write) on NFS data store sees no impact	
6	FC link fail between 6248 and 5500	Minimal impact seen.	VM running script continuously accessing a file (read/write) on FC data store continues after a momentary stop	
7	FC link restore between 6248 and 5500	No impact seen.	VM running script continuously accessing a file (read/write) on FC data store sees no impact	
8	FC link fail between 5500 and NetApp	No impact seen.	VM is up and running and able to browse data store	
9	FC link restore between 5500 and NetApp	No impact seen.	VM is up and running and able to browse data store	
10	Link fail between FT and TOM	3.9674 sec		
11	Link restore between FT and TOM	0.2098 sec		
12	Fail IOM	1.12765 sec		
13	Restore IOM	0.25845 sec		
14	Fail Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
15	Restore Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
16	CLI switchover Nexus 1010	0 sec	Nexus 1010 not in the forwarding path of data traffic.	
17	Nexus 1000V VSM switchover	0 sec	VSM not in the forwarding path of data traffic.	
18	Nexus 1000V VSM failure	0 sec	VSM not in the forwarding path of data traffic.	
19	Nexus 1000V VSM restore	0 sec	VSM not in the forwarding path of data traffic. The failed node booted up the the standby node.	
20	Nexus 1000V VSG failure	0 sec for established flow. 7.98 sec for new connection setup.		
21	Nexus 1000V VSG restore	0 sec	The failed node booted up the the standby node.	

Table 7-8 Other Convergence Events (continued)

22	UCS blade fail	A few minutes.	All VMs on the failed blade will fail. vSphere HA will restart the VM on another blade, guest OS boot up takes a few minutes.	
23	UCS blade restore	0 sec	No impact, unless vSphere DRS vMotions the VMs to the restored blade because of the load on other blades. vMotion would cause a packet drop of 1-2 sec. This depends on which VMs are moved and the resource usage.	

Authors

- Sunil Cherukuri
- Krishnan Thirukonda
- Chigozie Asiabaka
- Qingyan Cui
- Boo Kheng Khoo
- Padmanaba Kesav Babu Rajendran

APPENDIX A

Best Practices and Caveats

This appendix provides the best practices and caveats for implementing the VMDC 2.3 solution.

Compute and Storage Best Practices and Caveats

UCS Best Practices

- When using UCSM configuration templates, be aware that some configuration changes will either cause server reboot or service disruption. Multiple templates of the same type should be used to prevent any single change to cause service disruption to all blade servers.
- When configuring server pools, select servers from multiple chassis to avoid single chassis failure bringing down all servers in the pool.
- Disable fabric failover for all vNICs configured for the blade servers, and let the Nexus 1000V manage the vNIC failure.
- UCSM does not support overlapping VLANs in disjoint L2 networks. Ensure that each VLAN only connects to one upstream disjoint L2 network.
- UCS FI uses LACP as the port-channel aggregation protocol. The opposing upstream switches must be configured with LACP active mode.
- A vNIC (VMNIC in the vSphere ESXi hypervisor or physical NIC in the bare metal server) can only communicate with one disjoint L2 network. If a server needs to communicate with multiple disjoint L2 networks, configure a vNIC for each of those networks.
- UCSM implicitly assigns default VLAN 1 to all uplink ports and port-channels. Do not configure any vNICs with default VLAN 1. It is advisable not to use VLAN 1 for carrying any user data traffic.

Storage Best Practices

- If using NetApp OnCommand System Manager 2.0 to configure storage filers, it is recommended to configure the following using the command line:
  - Configuring VIF and VLAN interfaces for NFS port-channel.
  - Configure security style (Unix or Windows) permissions when a volume is exported as NFS.
- To take advantage of Thin Provisioning, it is recommended to configure Thin Provisioning on both volumes/LUNs in storage and in VMFS.
- Configure Asymmetric Logical Unit Access (ALUA) on the filers for asymmetric logical unit access of LUNs.
- Enable storage deduplication on volumes to improve storage efficiency.

- Nexus 5000 is the storage switch in this design. It is mandatory to enable NPIV mode on the Nexus 5000, and also configure soft zoning (enables server mobility) that uses WWPNs.

vSphere ESXi Best Practices

- vSphere Auto Deploy makes use of PXE and gPXE. The PXE/gPXE bootloader does not support 802.1Q tagging of DHCP frames. Configure the VLAN where the ESXi management vmk interface resides as the native VLAN.
- vSphere Auto Deploy makes use of DNS. Configure both forward and reverse DNS resolution for the ESXi hostname on the DNS server.
- When using vSphere Auto Deploy, make sure that the vCenter server, Auto Deploy server, DHCP server, and TFTP server are made highly available.

vSphere ESXi Caveats

- For the UCS blade server with the Cisco VIC adapter (Cisco UCS VIC 1280, Cisco UCS VIC 1240, Cisco UCS M81KR VIC, etc.), the ESXi host boot time will be much longer than those with other adapters. See CSCtu17983 for more details.
- In ESXi version 5.0, the ESXi Network Dump Collector feature is supported only with Standard vSwitches and cannot be used on a VMkernel network interface connected to a vSphere Distributed Switch or Nexus 1000V Switch. See VMware Knowlesge Base for more details.

Nexus 1000V Series Switches Best Practices

- Make sure that the SVS domain IDs for the Nexus 1010 VSA and the Nexus 1000V VSM are unique.
- Configure port profiles for management and vMotion vmknic as **system vlan**.
- Make use of port-profile inheritance to enforce consistent configuration and ease of management.

Layer 2 Best Practices and Caveats

vPC Best Practices

- A vPC peer link is recommended to use ports from different modules to provide bandwidth and redundancy.
- "ip arp synchronize," "peer-gateway," and "auto-recovery" should be configured in the vPC configuration.
- LACP should be used if possible
- It is recommended to disable LACP graceful convergence when the other end of port-channel neighbors are non NX-OS devices.
- Pre-provision all VLANs on MST and then create them as needed.
- On the Aggregation layer, create a root or a secondary root device as usual. Design the network to match the primary and secondary roles with the spanning-tree primary and secondary switches.
- If making changes to the VLAN-to-instance mapping when the vPC is already configured, remember to make changes on both the primary and secondary vPC peers to avoid a Type-1 global inconsistency.

Layer 3 Best Practices and Caveats

Best Practices

1. To accelerate L3 convergence, spread the L3 ports on different SoCs on the F2 module. This is due to the fact that on the F2 module, each port is mapped to a VRF instance and then the FIB for that VRF is downloaded. If an SoC has all ports as L2, then during reload and possibly other conditions, when the ports come up, FIB download is delayed until the SVI to VRF mapping is done, and hence FIB download happens after the port comes up and L2 convergence and mapping of VLANs to that port is complete. In VMDC 2.3 implementation, the L3 ports to the DC PEs and the VPC peer links were spread across five SoCs per module to get the benefit of FIB download immediately on reload. Refer to Cisco Nexus 7000 F2-Series 48-Port 1 and 10 Gigabit Ethernet Module Data Sheet for more information about F2 card and SoCs. Also, see CSCue67104 below.
2. To reduce traffic loss after system reload, delay the time that it takes for VLAN interface and vPCs to come online. By default, VLAN interfaces are brought online 10 seconds after the peer link is up, and vPCs are brought online 30 seconds after the VLAN interfaces are brought up. Based on scale characteristics of this validation, we delay VLAN interfaces and vPCs from coming online by 90 seconds each.
3. The ACE 4710 appliances do not support LACP, and hence their port-channels to the Nexus 7000 switches are static with mode on. We expect to see some traffic loss when the system comes online after a reload. To protect against this loss, carrier delays can be configured on the ACE GigabitEthernet interfaces to prevent this interface from coming online. Using this scheme will introduce a carrier-delay time during a vPC shut/no shut test or similar negative event.
4. Carrier delay can be configured on the ASR 1000 interfaces to the Nexus 7000 aggregation routers to delay the L3 interface from coming up. This ensures that these L3 interfaces are brought up at a time when the Nexus 7000 routers are ready to successfully set up and establish BGP sessions. In this validation, the carrier delay on the ASR 1000 PE was set to the maximum of 60 seconds.
5. By default, the ACE 4710 appliance will renew ARP entries for a configured host every 300 seconds. We increase the ARP rates to 1440 seconds to reduce the possibility of the ACE ARP request being lost as the system comes online after a reload.
6. To get better convergence performance, use BGP policy to divert traffic away from the Nexus 7004 aggregation switch under certain conditions such as VPC peer link fail or secondary shutdown. This is because the FIB programming on the F2 card is slower, leading to additional packet losses of up to 10 seconds in the scale validated, and this can be higher with a high-programmed prefix count. BGP configuration on the ASR 1000 and Nexus 7000 aggregation routers is set up so that the ASR 1000 reroutes traffic to an alternate path if the vPC peer link fails and shuts down the VPC secondary. This eliminates up to 10 seconds of traffic loss that occurs due to the F2 FIB programming delay. If the peer link fails, expect up to 13 seconds of traffic convergence, which is due to up to 8 seconds being required for the VLAN interface to go down, and due to up to 5 seconds being required for the BGP and RIB update on the Nexus 7000 aggregation routers. The causes of this convergence delay in FIB programming is under investigation. See CSCue59878 below. For overall vPC convergence, there are a few enhancements targeted for the next NX-OS software release 6.2.
7. BGP PIC, BGP graceful restart, and other routing optimization should be enabled on the ASR 1000 PE devices for faster convergence. BGP PIC and graceful restart are enabled by default on the Nexus 7000 aggregation routers.

Caveats

1. CSCud23607 was an HSRP programming issue seen if the MAC address table size limits are reached. This is fixed in NX-OS 6.1.3. Prior to NX-OS 6.1.3 , the workaround was to manually flap the affected HSRP interfaces.
2. CSCue59878 was filed to investigate the FIB programming delay after routing convergence during a vPC shut test or similar scenarios. This issue is under investigation. The reason for delay is due to the FIB programming mechanism used for the F2 module. The module has to program TCAM for all 12 SoCs, and as the number of prefixes gets higher, it takes additional time to calculate and program each of the SoCs. The workarounds are to reduce the number of SoCs used, i.e., less number of ports and to reduce the number of prefixes per SoC (by mapping specific VRF instances (ports) to SoCs so that the total prefix is less per SoC). If convergence times need to be quicker, and with a larger number of prefixes, consider using M2 or M1 series modules.
3. CSCue67104 was filled to investigate convergence delays due to packet losses after system reload of the nexus 7000 aggregation router. These losses are seen as FIB losses when the vPC port-channels are brought up and can last 10 or more seconds. This issue was closed as this is expected. On F2 modules, which have an SoC design, each SoC needs to map all of its ports into VRF instances, and then download the FIB. When all of the ports on an SoC are L2 only, the L2 ports need to come up and the SVIs need to be mapped to VRF instances before downloading the FIB for those VRF instances. This takes additional time after the port comes up (see CSCue59878 above, F2 FIB convergence is slow). To work around this issue, have a mix of both L2 and L3 ports on the same SoC. The L3 ports being on the SoC will cause all FIBs for the VRF instances on the L3 port to be downloaded as soon as the module comes up. In VMDC 2.3, all VRF instances used are allowed on the L3 port, so all FIBs will be downloaded to any SoC that has L3 ports. Since there are two L3 uplinks and four L3 peer links for iBGP per box, this provides one L3 port for uplink and two iBGP ports for peer per module. These ports should be spread on three different SoCs. Additionally, we can also spread the vPC peer link ports in different SoCs. Since there are four ports in the vPC peer link, two ports from each module, this covers two more SoCs. This helps with the reload case, as the vPC peer link will come online first and have SVIs mapped to it followed by FIB download, before the actual vPC port-channels come up, however, this will not help in the module restore case, as the vPC peer link port SoCs and FIB download will still be delayed. Additional L3 ports can help, if they are configured on any additional SoCs used. The goal with this workaround is to have all SoC FIBs programmed by the time the vPC port-channels come online.
4. CSCuc51879 is an issue seen during RP failover either due to RPSO or In-Service System Upgrade (ISSU). This is an issue related to traffic loss seen during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration.
5. The following performance fixes are expected in the 6.2 release of NX-OS. These fixes are expected to help with convergence.
   a. CSCtn37522: Delay in L2 port-channels going down
   b. CSCud82316: VPC Convergence optimization
   c. CSCuc50888: High convergence after F2 module OIR

Services Best Practices and Caveats

ASA Firewall Appliance Best Practices

- The ASA FT and stateful links should be dedicated interfaces between the primary and secondary ASA.

- Failover interface policies should be configured to ensure that the security context fails over to the standby ASA if monitored interfaces are down.
- Configure an appropriate port-channel load-balancing scheme on the ASA to ensure that all port-channel interfaces are used to forward traffic out of the ASA.

Copper Implementation Best Practices

- Configure all Copper tenants' servers with either public or private IP addresses, not a mix of both types. If both types are needed, use seperate ASA context for all public addressed tenants and a separate context for all private addressed tenants.
- Private IP addresses for servers can be overlapped for different tenants, and requires the use of NAT with separate public IP addresses per tenant for outside.

Compute Firewall Best Practices

- The VSG does not support vSphere HA and DRS. On clusters dedicated for hosting VSG virtual appliances, disable vSphere HA and DRS. On clusters hosting both VSG virtual appliances and other VMs, disable HA and DRS for the VSG virtual appliances.
- For a VSG HA-pair, the primary and secondary nodes should be hosted on the same ESXi host. Use vSphere anti-affinity rules or DRS groups to ensure this.
- Each VSG virtual appliance (be it active or standby node) reserves CPU and memory resources from the ESXi host. Make sure the ESXi host has enough unreserved CPU and memory resources, otherwise, the VSG virtual appliance will not power on.
- Make sure that the clocks on the VNMC, VSGs, and Nexus 1000V are synchronized. The VSGs and Nexus 1000V will not be able to register to the VNMC if the clocks are too out of sync.
- Enable IP proxy ARP on the router interface(s) on the subnet/VLAN facing the VSG data interfaces.
- On the VNMC, compute firewalls should be added at the tenant level or below, and not at the Root org level.
- For the tenant, the DMZ should have its own VSG compute firewall, separate from the firewall used on the PVT zone.
- When configuring security policies/rules on the VNMC, the attributes used for filtering conditions should be preferred in the following order:
  - Network attributes, most prefer, providing highest performance for VSG
  - VM Attributes
  - vZone, lest prefer, lowest VSG performance

Compute Firewall Caveats

- The VSG FTP/TFTP protocol inspect on the VSG fails to open the pinhole required for data connection when the source and destination vNICs are under the same VSG protection, but on different VLANs. See CSCud39323 for more details.

ACE 4710 Appliance Best Practices

- The FT VLAN should be configured using the **ft-port vlan <vlan-id>** command to ensure that FT packets have the right QoS labels. This ensures that proper treatment is given to ACE FT packets in the network.
- Configure an appropriate port-channel load-balancing scheme to ensure that all port-channel interfaces are used to forward traffic out of the ACE appliance.

- To avoid MAC collision among operational ACE appliances on the same VLAN, use an appropriate shared-vlan host-id <1-16> to ensure that each ACE appliance has a unique MAC address on a shared VLAN.

QoS Best Practices and Caveats

Nexus 1000V QoS Best Practices

- Configure QoS policies to classify, mark, police, and prioritize traffic flows. Different traffic types should have different network treatment.

UCS QoS Best Practices

- Reserve bandwidth for each traffic type using QoS system class. Each type of traffic should have a guaranteed minimum bandwidth.
- For UCS servers deployed with the Nexus 1000V, it is highly recommended to do the CoS marking at the Nexus 1000V level. Configure UCS QoS policy with **Host Control Full** and attach the policy to all vNICs of UCS servers.

ACE QoS Best Practices and Caveats Caveats

- **CSCtt19577:** need ACE to preserve L7 traffic dot1p CoS
  - QoS transparency requires that DSCP not be touched, and that only CoS be used to support DC QoS in the VMDC system. The tenant uses DSCP for their markings, and the DC operator can use independent QoS markings by using dot1P CoS bits. To support this, both DSCP and dot1p CoS need to be preserved as packets transit the ACE, however, the ACE does not currently support CoS preservation for L7 traffic. This enhancement requests support for CoS preservation and DSCP preservation for all scenarios including L7 traffic.

Nexus 7000 QoS Best Practices and Caveats Best Practices

- The Nexus 7000 series uses four fabric queues across modules, and CoS values are mapped to these four queues statically, i.e., they cannot be changed. The priority queue for CoS5,6, and 7 is switched with strict priority, and the other three queues are switched with equal weights. The F2 cards used in VMDC 2.3 use the 8e-4q4q model, which class-maps that map to the CoS values in the same way as the fabric queues. This is particularly important as the F2 card uses buffers in the ingress card, and back pressure from the egress interface congestion is mapped to ingress queues. Packets are dropped at ingress when such congestion happens. It is important to use the 8e-4q4q model to track each class separately. This model is supported from NX-OS release 6.1.3 onwards.

Caveats

- **CSCue55938:** duplicating policy-maps for egress queuing.
  - Attaching two queuing policies for the same direction under a port is allowed under some conditions.
- **CSCud46159:** all interfaces in the module are gone after reboot
  - When a scaled up configuration with many interfaces is configured with the same policy-map that includes egress policing, upon reload, the Nexus 7004 aggregation switch loses its configuration of all interfaces. This workaround is to configure multiple policy-maps with the same policy and divide the total number of subinterfaces into three or four groups and attaching a different policy-map to each group.
- **CSCud26031:** F2: aclqos crash on configuring QoS policy on subinterfaces

  - ACLQOS crash is observed when attaching a service policy that includes egress policing on a large number of subinterfaces. The workaround is to use different policy-maps (with the same underlying policy) so that the number of subinterfaces using the same policy-map is reduced.
- **CSCud26041:** F2: scale QoS configs by not allocating policer stats when no policing
  - Qos per class stats use hardware resources that are shared with policers. On the F-series card, this is restricted to a small amount, i.e., currently 1024, which is the total of all classes in policies multiplied by attachments. For example, with an eight-class policy, only 128 attachments can be done on 128 subinterfaces on the same SoC. This bug requests disabling default per-class statistics collection and providing proper error messaging to indicate the actual issue. Statistics are enabled by default, and hence the workaround is to add **no-stats** to the service policy attachments.

Nexus 5000 QoS Best Practices and Caveats Best Practices

- Use all six classes of traffic for the Ethernet class if no FCoE traffic is expected.
- Account for NFS traffic at this layer of the DC, and provide a separate class and queuing to provide a BW guarantee.

Caveats

- **CSCue88052:** Consistency between Nexus 5000 and Nexus 7000 QoS config
  - Nexus 5500 Series switches currently have different semantics of similar sounding QoS configuration items, and this bug tracks specifically the fact that the Nexus 5500 allows the configuration of bandwidth percent for a class in a policy-map where priority is configured. Also, the bandwidth percent semantics in a policy-map that has priority class is actually called "bandwidth remaining." This is confusing and not consistent with the Nexus 7000 semantics, which have checks in place to prevent priority and bandwidth percent configuration for the same class in a policy-map.

ASR 1000 QoS Best Practices and Caveats Best Practices

- QoS on port-channel interfaces is not supported. For the MPLS-Core facing interfaces, port-channels are not recommended, as the VMDC 2.3 QoS policies cannot be implemented.
- QoS on port-channel subinterfaces have restrictions. For example, ingress QoS cannot be done in flow-based mode, and egress QoS requires a QoS configuration on the member links. The recommendation for VMDC 2.3 is to use multiple links between the DC-PE and DC-AGG if more than 10GE is required.
- NetFlow on the ASR 1000 series with custom NetFlow records can impact the switching performance. The recommendation is to use default NetFlow record formats. While this is not exactly a QoS best practice, this can impact QoS due to dropping of packets earlier than expected due to switching performance rather than actual link congestion.
- Mapping of priority traffic based on CoS and MPLS-TC to the high-priority queue between SIP and the ESP is required to provide priority traffic low latency treatment.
- Calculation of bandwidth requirement for both normal and failure cases should be accounted for, as the ASR 1000 is a centralized switching platform and all traffic is funneled and switched at the ESP. In this design, a SIP-40 is used with 4x10GE shared port adapters, and with ESP-40, which can handle 40 Gbps of switching. This provides 10 Gbps of traffic from north-south, and 10 Gbps of traffic from south-north, for a total of 20 Gbps for normal conditions. Different failure scenarios will not cause any oversubscription at the ESP-40.

Caveats

- **CSCud51708:** wrong calc for bytes w ms based queue-limit config after random-detect
  - If the queue-limit is configured in milliseconds after configuring random-detect, the bytes calculation is wrong for the specified number of milliseconds in the queue-limit. The workaround is to first configure the queue-limit in milliseconds and then configure random-detect.

APPENDIX B

Related Documentation

The Virtualized Multiservice Data Center (VMDC) design recommends that general Cisco Data Center (DC) design best practices be followed as the foundation for Infrastructure as a Service (IaaS) deployments. The companion documents listed in this appendix provide guidance on such a foundation.

Cisco Related Documentation

The following Cisco Validated Design (CVD) companion documents provide guidance on VMDC design and implementation.

- VMDC 2.0 Solution Overview
- VMDC 2.0 Solution White Paper
- VMDC 2.1 Design Guide
- VMDC 2.1 Implementation Guide
- VMDC 2.2 Design Guide
- VMDC 2.2 Implementation Guide
- VMDC 2.2 EoMPLS DCI for Hybrid Cloud with vCloud Director
- VMDC 2.3 Design Guide
- VMDC 3.0 Design Guide
- VMDC 3.0 Implementation Guide
- Previous VMDC System Releases
- VMDC based Cloud Ready Infrastructure kit
- Data Center Designs: Data Center Interconnect
- VMDC Hybrid Cloud with vCloud Director Design and Implementation Guide
- Data Center Design—IP Network Infrastructure
- Data Center Service Patterns
- Data Center Interconnect
- Security and Virtualization in the Data Center
- Vblock Infrastructure Solutions

Cloud Enablement Services from Cisco Advanced Services and partners can help customers realize the full business value of their IT investments faster. Backed by our networking and security expertise, an architectural approach, and a broad ecosystem of partners, these intelligent services enable customers to build a secure, agile, and highly automated cloud infrastructure.

- Cisco Advanced Services Cloud Enablement

Third-Party Related Documentation

The following links provide information on BMC CLM and VCE Vblock:

- BMC Cloud Lifecycle Management
- Cisco BMC Sales Engagement
- Vblock Infrastructure Solutions

The following links provide information on the EMC Symmetrix VMAX:

- VMAX - Enterprise Storage, Virtual Data Center
- Symmetrix Family—Enterprise Storage, Virtualization

The following links provide information on the NetApp FAS6080:

- FAS6000 Series Enterprise Storage Systems
- FAS6000 Series Technical Specifications

APPENDIX C

Configuration Templates

This appendix provides the configurations per tenant type.

- Configuration Template for Gold Service Class, page C-1
- Configuration Template for Silver Service Class, page C-16
- Configuration Template for Bronze Service Class, page C-24
- Configuration Template for Copper Service Class, page C-26
- Configuration Template for Nexus 5548 ICS switch, page C-31

Configuration Template for Gold Service Class

This section presents the configuration templates for the Gold service class.

- Aggregation Nexus 7000 Gold Configuration, page C-1
- ASA Gold Configurations, page C-3
- ACE Gold Configuration, page C-10
- ASR 1000 PE Gold Tenant Configuration, page C-14
- Nexus 1000V Gold Configuration, page C-15

Aggregation Nexus 7000 Gold Configuration

This section provides an aggregation Nexus 7000 Gold configuration.

```
vrf context customer_gold1_priv
  ip route 0.0.0.0/0 10.1.6.11

vrf context customer_gold1_pub
  ip route 11.1.0.0/16 10.1.5.11

interface Vlan201
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_gold1_priv
  no ip redirects
  ip address 11.1.1.2/24
  no ipv6 redirects
  no ip arp gratuitous hsrp duplicate
  hsrp version 2
  hsrp 201
    preempt
    priority 150
    ip 11.1.1.1

interface Vlan301
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_gold1_priv
  no ip redirects
  ip address 11.1.2.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 301
    preempt
    priority 150
    ip 11.1.2.1

interface Vlan401
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_gold1_priv
  no ip redirects
  ip address 11.1.3.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 401
    preempt
    priority 150
    ip 11.1.3.1

interface Vlan1201
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_gold1_priv
  no ip redirects
  ip address 10.1.6.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1201
    preempt
    priority 150
    ip 10.1.6.1

interface Vlan1301
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_gold1_pub
  no ip redirects
  ip address 10.1.5.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1301
    preempt
    priority 150
    ip 10.1.5.1

interface port-channel343.201
  vrf member customer_gold1_pub
  ip address 10.1.34.3/24

interface Ethernet3/9.201
  vrf member customer_gold1_pub
  ip address 10.1.1.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.201
  vrf member customer_gold1_pub
  ip address 10.1.3.2/24
  no ip arp gratuitous hsrp duplicate

router bgp 65501
  vrf customer_gold1_pub
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute static route-map SET-COMM
      additional-paths send
      additional-paths receive
    neighbor 10.1.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.1.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.1.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community
        next-hop-self
```

ASA Gold Configurations

This section provided templates for ASA Gold configurations.

- ASA Gold Tenant Perimeter Firewall Configuration, page C-3
- ASA Gold Tenant DMZ Firewall Configuration, page C-7
- ASA Gold Tenant SSL and IPSec VPN Configuration, page C-9

ASA Gold Tenant Perimeter Firewall Configuration

This section provides an ASA Gold tenant perimeter firewall configuration.

```
dc02-asa-fw1/admin# changeto c customer-gold1
dc02-asa-fw1/customer-gold1# sh run
: Saved
:
ASA Version 9.0(1) <context>
!
terminal width 511
hostname customer-gold1
enable password 8Ry2YjIyt7RRXU24 encrypted
xlate per-session deny tcp any4 any4
xlate per-session deny tcp any4 any6
xlate per-session deny tcp any6 any4
xlate per-session deny tcp any6 any6
xlate per-session deny udp any4 any4 eq domain
xlate per-session deny udp any4 any6 eq domain
xlate per-session deny udp any6 any4 eq domain
xlate per-session deny udp any6 any6 eq domain
passwd 2KFQnbNIdI.2KYOU encrypted
```

```
names
!
interface Management0/0
 management-only
 nameif mgmt
 security-level 100
 ip address 192.168.50.201 255.255.255.0 standby 192.168.50.202
!
interface Port-channel1.1201
dc02-asa-fw1/customer-gold1# ter
dc02-asa-fw1/customer-gold1# terminal ?

  monitor  Syslog monitor
  no       Turn off syslogging to this terminal
  pager    Control page length for pagination. The page length set here is not saved
to configuration.
dc02-asa-fw1/customer-gold1# terminal pa
dc02-asa-fw1/customer-gold1# terminal pager ?

  <0-2147483647>  Pager lines, 0 means no page-limit
  lines           The number following this keyword determines the number of lines in
a page before ---more--- prompt appears, default is 24
dc02-asa-fw1/customer-gold1# terminal pager 0
dc02-asa-fw1/customer-gold1# sh run
: Saved
:
ASA Version 9.0(1) <context>
!
terminal width 511
hostname customer-gold1
enable password 8Ry2YjIyt7RRXU24 encrypted
xlate per-session deny tcp any4 any4
xlate per-session deny tcp any4 any6
xlate per-session deny tcp any6 any4
xlate per-session deny tcp any6 any6
xlate per-session deny udp any4 any4 eq domain
xlate per-session deny udp any4 any6 eq domain
xlate per-session deny udp any6 any4 eq domain
xlate per-session deny udp any6 any6 eq domain
passwd 2KFQnbNIdI.2KYOU encrypted
names
!
interface Management0/0
 management-only
 nameif mgmt
 security-level 100
 ip address 192.168.50.201 255.255.255.0 standby 192.168.50.202
!
interface Port-channel1.1201
 nameif inside
 security-level 100
 ip address 10.1.6.11 255.255.255.0 standby 10.1.6.12
!
interface Port-channel1.1301
 nameif outside
 security-level 0
 ip address 10.1.5.11 255.255.255.0 standby 10.1.5.12
!
interface Port-channel1.1401
 nameif dmz
 security-level 80
 ip address 10.1.8.21 255.255.255.0 standby 10.1.8.22
!
object network SP-CLIENTS-POOL
```

```
 range 51.1.1.1 51.1.1.254
object network SP-CLIENTS->DMZ
 range 0.0.0.0 255.255.255.255
object network test1
 range 51.1.2.1 51.1.2.254
object-group network SP-CLIENTS-NETWORK
 network-object 40.1.0.0 255.255.0.0
 network-object 10.1.0.0 255.255.0.0
 network-object 131.0.0.0 255.0.0.0
 network-object 51.1.2.0 255.255.255.0
object-group service SP-CLIENTS-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ftp
 port-object eq ssh
 port-object eq domain
object-group service SP-CLIENTS-PROTOCOLS-UDP udp
 port-object eq tftp
 port-object eq domain
 port-object range 10000 30000
object-group network DMZ-VPN-NETWORK
 network-object 11.1.4.0 255.255.255.0
 network-object 11.255.0.0 255.255.0.0
object-group service DMZ-VPN-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ssh
 port-object eq ftp
object-group service DMZ-VPN-PROTOCOLS-UDP udp
 port-object eq tftp
 port-object eq domain
 port-object range 10000 30000
access-list DMZ-VPN extended permit tcp object-group DMZ-VPN-NETWORK any object-group
DMZ-VPN-PROTOCOLS-TCP
access-list DMZ-VPN extended permit udp object-group DMZ-VPN-NETWORK any object-group
DMZ-VPN-PROTOCOLS-UDP
access-list DMZ-VPN extended permit icmp object-group DMZ-VPN-NETWORK any
access-list OUTSIDE extended permit tcp object-group SP-CLIENTS-NETWORK any
object-group SP-CLIENTS-PROTOCOLS-TCP
access-list OUTSIDE extended permit udp object-group SP-CLIENTS-NETWORK any
object-group SP-CLIENTS-PROTOCOLS-UDP
access-list OUTSIDE extended permit icmp object-group SP-CLIENTS-NETWORK any
pager lines 24
logging enable
logging timestamp
logging standby
logging monitor debugging
logging buffered debugging
logging trap errors
logging asdm informational
logging facility 17
logging device-id context-name
logging host mgmt 192.168.11.100
no logging message 713167
no logging message 713123
no logging message 313001
no logging message 725001
no logging message 725002
no logging message 710005
no logging message 113009
no logging message 302015
no logging message 302014
no logging message 302013
no logging message 602303
no logging message 609001
no logging message 715007
no logging message 302016
mtu mgmt 1500
mtu inside 1500
mtu outside 1500
mtu dmz 1500
monitor-interface inside
monitor-interface outside
monitor-interface dmz
icmp unreachable rate-limit 1 burst-size 1
no asdm history enable
arp timeout 14400
!
object network SP-CLIENTS->DMZ
 nat (outside,dmz) dynamic SP-CLIENTS-POOL
object network test1
 nat (outside,inside) dynamic test1
access-group OUTSIDE in interface outside
access-group DMZ-VPN in interface dmz
route outside 0.0.0.0 0.0.0.0 10.1.5.1 1
route inside 11.0.0.0 255.0.0.0 10.1.6.1 1
route dmz 11.1.4.0 255.255.255.0 10.1.8.11 1
route dmz 11.255.0.0 255.255.0.0 10.1.8.11 1
route inside 111.0.0.0 255.0.0.0 10.1.6.1 1
route mgmt 192.168.0.0 255.255.0.0 192.168.50.1 1
timeout xlate 3:00:00
timeout pat-xlate 0:00:30
timeout conn 1:00:00 half-closed 0:10:00 udp 0:02:00 icmp 0:00:02
timeout sunrpc 0:10:00 h323 0:05:00 h225 1:00:00 mgcp 0:05:00 mgcp-pat 0:05:00
timeout sip 0:30:00 sip_media 0:02:00 sip-invite 0:03:00 sip-disconnect 0:02:00
timeout sip-provisional-media 0:02:00 uauth 0:05:00 absolute
timeout tcp-proxy-reassembly 0:01:00
timeout floating-conn 0:00:00
user-identity default-domain LOCAL
snmp-server host mgmt 192.168.11.12 community ***** version 2c
no snmp-server location
no snmp-server contact
crypto ipsec security-association pmtu-aging infinite
telnet timeout 5
ssh timeout 5
no threat-detection statistics tcp-intercept
!
class-map inspection_default
 match default-inspection-traffic
!
!
policy-map type inspect dns preset_dns_map
 parameters
  message-length maximum client auto
  message-length maximum 512
policy-map global_policy
 class inspection_default
  inspect ftp
  inspect h323 h225
  inspect h323 ras
  inspect ip-options
  inspect netbios
  inspect rsh
  inspect rtsp
  inspect skinny
  inspect esmtp
  inspect sqlnet
  inspect sunrpc
  inspect tftp
  inspect sip
  inspect xdmcp
  inspect dns preset_dns_map
!
Cryptochecksum:d41d8cd98f00b204e9800998ecf8427e
: end
dc02-asa-fw1/customer-gold1
```

ASA Gold Tenant DMZ Firewall Configuration

This section provides an ASA Gold tenantDMZ firewall configuration.

```
dc02-asa-fw1/customer-gold1# changeto c customer-gold1-dmz
dc02-asa-fw1/customer-gold1-dmz# ter
dc02-asa-fw1/customer-gold1-dmz# terminal p 0
dc02-asa-fw1/customer-gold1-dmz# sh run
: Saved
:
ASA Version 9.0(1) <context>
!
terminal width 511
hostname customer-gold1-dmz
enable password 8Ry2YjIyt7RRXU24 encrypted
xlate per-session deny tcp any4 any4
xlate per-session deny tcp any4 any6
xlate per-session deny tcp any6 any4
xlate per-session deny tcp any6 any6
xlate per-session deny udp any4 any4 eq domain
xlate per-session deny udp any4 any6 eq domain
xlate per-session deny udp any6 any4 eq domain
xlate per-session deny udp any6 any6 eq domain
xlate per-session deny tcp any4 any4
xlate per-session deny tcp any4 any6
xlate per-session deny tcp any6 any4
xlate per-session deny tcp any6 any6
xlate per-session deny udp any4 any4 eq domain
xlate per-session deny udp any4 any6 eq domain
xlate per-session deny udp any6 any4 eq domain
xlate per-session deny udp any6 any6 eq domain
passwd 2KFQnbNIdI.2KYOU encrypted
names
!
interface Management0/0
 management-only
 nameif mgmt
 security-level 100
 ip address 192.168.50.221 255.255.255.0 standby 192.168.50.222
!
interface Port-channel1.1401
 nameif inside
 security-level 100
 ip address 10.1.8.11 255.255.255.0 standby 10.1.8.12
!
interface Port-channel1.1501
 nameif dmz
 security-level 80
 ip address 10.1.7.11 255.255.255.0 standby 10.1.7.22
!
interface Port-channel1.1701
 nameif vpn
 security-level 50
 ip address 11.255.1.251 255.255.255.0 standby 11.255.1.252
```

```
!
interface Port-channel1.2000
 nameif internet
 security-level 0
 ip address 100.200.1.11 255.255.255.0 standby 100.200.1.12
!
object network SERVER1
 host 11.1.4.11
object network SERVER3
 host 11.1.4.13
object network SERVER2
 host 11.1.4.12
object network WEB-VIP
 host 11.1.4.111
object network t1
object network SERVER8
 host 11.1.4.100
object network SERVER7
 host 11.1.4.151
object-group service INTERNET-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ssh
object-group service VPN-PROTOCOLS-TCP tcp
 port-object eq www
 port-object eq https
 port-object eq ssh
object-group service INTERNET-PROTOCOLS-UDP udp
 port-object eq tftp
 port-object range 10000 30000
access-list INTERNET extended permit tcp any any object-group INTERNET-PROTOCOLS-TCP
access-list INTERNET extended permit icmp any any
access-list INTERNET extended permit udp any any object-group INTERNET-PROTOCOLS-UDP
access-list VPN extended permit tcp any any object-group INTERNET-PROTOCOLS-TCP
access-list VPN extended permit icmp any any
access-list DMZ extended permit ip any any
pager lines 24
logging enable
logging timestamp
logging standby
logging monitor debugging
logging buffered debugging
logging trap errors
logging asdm informational
logging facility 17
logging device-id context-name
logging host mgmt 192.168.11.100
no logging message 713167
no logging message 713123
no logging message 313001
no logging message 725001
no logging message 725002
no logging message 710005
no logging message 113009
no logging message 302015
no logging message 302014
no logging message 302013
no logging message 602303
no logging message 609001
no logging message 715007
no logging message 302016
mtu mgmt 1500
```

ASA Gold Tenant SSL and IPSec VPN Configuration

This section provides an ASA Gold tenant SSL and IPSec VPN configuration.

```
rypto ipsec ikev1 transform-set ipsec-tz esp-3des esp-md5-hmac
crypto ipsec security-association pmtu-aging infinite
crypto dynamic-map ipsec-cm 1 set ikev1 transform-set ipsec-tz
crypto dynamic-map ipsec-cm 1 set security-association lifetime seconds 7200
crypto map ipsec-cm 1 ipsec-isakmp dynamic ipsec-cm
crypto map ipsec-cm interface internet
crypto ca trustpool policy
crypto ikev1 enable internet
crypto ikev1 policy 1
 authentication pre-share
 encryption 3des
 hash md5
 group 2
 lifetime 3600
tunnel-group customer_gold1-ipsec type remote-access
tunnel-group customer_gold1-ipsec general-attributes
 address-pool customer_gold1
 authentication-server-group (internet) LOCAL
 authorization-server-group (internet) LOCAL
tunnel-group customer_gold1-ipsec ipsec-attributes
 ikev1 pre-shared-key *****
group-policy customer_gold1-ipsec internal
group-policy customer_gold1-ipsec attributes
 vpn-simultaneous-logins 200
 vpn-tunnel-protocol ikev1
 group-lock value customer_gold1-ipsec
 split-tunnel-policy tunnelspecified
 split-tunnel-network-list value customer_gold1
 vlan 1701
username ipsec1 password S8ZObXJyIluJKbJX encrypted
username ipsec1 attributes
 vpn-group-policy customer_gold1-ipsec
webvpn
 enable internet
 no anyconnect-essentials
 csd image disk0:/csd_3.6.6210-k9.pkg
 anyconnect image disk0:/anyconnect-win-3.1.01065-k9.pkg 1
 anyconnect profiles anyconnect-profile disk0:/RDP.xml
 anyconnect enable
 tunnel-group-preference group-url
tunnel-group customer_gold1-ssl type remote-access
tunnel-group customer_gold1-ssl general-attributes
 address-pool customer_gold1
 authentication-server-group (internet) LOCAL
 authorization-server-group (internet) LOCAL
tunnel-group customer_gold1-ssl webvpn-attributes
 group-url https://100.200.1.51/customer_gold1 enable
dc02-asa5555-1# sh run group-policy customer_gold1-ssl
group-policy customer_gold1-ssl internal
group-policy customer_gold1-ssl attributes
 vpn-simultaneous-logins 200
 vpn-tunnel-protocol ssl-client ssl-clientless
 group-lock value customer_gold1-ssl
 split-tunnel-policy tunnelspecified
 split-tunnel-network-list value customer_gold1
 vlan 1701
 webvpn
  anyconnect profiles value anyconnect-profile type user
dc02-asa5555-1# sh run username ssl1
```

```
username ssl1 password JSKNK4oromgGd3D9 encrypted
username ssl1 attributes
 vpn-group-policy customer_gold1-ssl
dc02-asa5555-1#
```

ACE Gold Configuration

This section provides an ACE Gold configuration.

```
dc02-ace-1/Admin# changeto customer_gold1
dc02-ace-1/customer_gold1# terminal length 0
dc02-ace-1/customer_gold1# sh run
Generating configuration....

logging enable
logging standby
logging timestamp
logging trap 6
logging buffered 7
logging monitor 6
logging facility 17
logging device-id context-name
logging host 192.168.11.100 udp/514
no logging message 251008
no logging message 302022
no logging message 302023
no logging message 302024
no logging message 302025
no logging message 106023

arp interval 1440

access-list app-acl line 8 extended permit ip any any
access-list db-acl line 8 extended permit ip any any
access-list t1 line 8 extended permit tcp 11.1.1.0 255.255.255.0 11.1.2.0
255.255.255.0
access-list web-acl line 8 extended deny udp 11.0.0.0 255.0.0.0 eq tftp any
access-list web-acl line 16 extended deny udp 11.0.0.0 255.0.0.0 eq 30000 any
access-list web-acl line 24 extended permit ip any any

probe ftp ftp-probe
  interval 2
  faildetect 5
  passdetect interval 2
  passdetect count 5
  receive 1
  expect status 200 400
  connection term forced
probe http http-probe
  interval 2
  faildetect 5
  passdetect interval 2
  passdetect count 5
  receive 1
  expect status 200 400
  connection term forced

rserver host app-server1
  ip address 11.1.2.11
  inservice
rserver host app-server2
  ip address 11.1.2.12
  inservice
rserver host app-server3
  ip address 11.1.2.13
  inservice
rserver host db-server1
  ip address 11.1.3.11
  inservice
rserver host db-server2
  ip address 11.1.3.12
  inservice
rserver host db-server3
  ip address 11.1.3.13
  inservice
rserver host udp-host
  ip address 11.1.1.100
  inservice
rserver host udp-host:30000
  ip address 11.1.1.101
  inservice
rserver host web-server1
  ip address 11.1.1.11
  inservice
rserver host web-server2
  ip address 11.1.1.12
  inservice
rserver host web-server3
  ip address 11.1.1.13
  inservice
rserver host web-spirent
  ip address 11.1.1.151
  inservice

serverfarm host app-serverfarm
  rserver app-server1
    inservice
  rserver app-server2
    inservice
  rserver app-server3
    inservice
serverfarm host db-serverfarm
  rserver db-server1
    inservice
  rserver db-server2
    inservice
  rserver db-server3
    inservice
serverfarm host udp-serverfarm
  rserver udp-host
    inservice
serverfarm host udp-serverfarm:30000
  rserver udp-host:30000
    inservice
serverfarm host web-serverfarm
  rserver web-server1
    inservice
  rserver web-server2
  rserver web-server3
  rserver web-spirent
    inservice

parameter-map type connection tcp_pm
  set tcp wan-optimization rtt 0
parameter-map type connection udp_pm
  set timeout inactivity 300
```

```
sticky http-cookie customer_gold1-http-cookie customer_gold1-http
  cookie insert browser-expire
  serverfarm web-serverfarm
  timeout 10
  replicate sticky
sticky http-cookie customer_gold1-web-app-cookie customer_gold1-web->app
  cookie insert browser-expire
  serverfarm app-serverfarm
  timeout 10
  replicate sticky
sticky ip-netmask 255.255.255.255 address both customer_gold1-app->db
  serverfarm db-serverfarm
  timeout 10
  replicate sticky

class-map type http loadbalance match-any cm-app-subnet
  2 match source-address 11.1.2.0 255.255.255.0
class-map type http loadbalance match-any cm-http
  2 match http url /.*.txt
  3 match http url /.*.html
class-map type http loadbalance match-any cm-web-subnet
  2 match source-address 11.1.1.0 255.255.255.0

class-map match-all app->db-vip
  2 match virtual-address 11.1.3.111 tcp eq www
class-map type http loadbalance match-all cm-app->db
  2 match class-map cm-http
  3 match class-map cm-app-subnet
class-map type http loadbalance match-all cm-web->app
  2 match class-map cm-http
  3 match class-map cm-web-subnet
class-map type management match-any management-traffic
  2 match protocol ssh any
  3 match protocol http any
  4 match protocol https any
  5 match protocol icmp any
  6 match protocol telnet any
  7 match protocol snmp source-address 192.168.0.0 255.255.0.0
class-map match-all udp-vip
  2 match virtual-address 11.1.1.111 udp eq 69
class-map match-all udp-vip:30000
  2 match virtual-address 11.1.1.111 udp eq 30000
class-map match-all web->app-vip
  2 match virtual-address 11.1.2.111 tcp eq www
class-map match-all web-vip
  2 match virtual-address 11.1.1.111 tcp eq www

policy-map type management first-match management-traffic
  class management-traffic
    permit

policy-map type loadbalance first-match app->db-lb-policy
  class cm-app->db
    sticky-serverfarm customer_gold1-app->db
policy-map type loadbalance first-match udp-lb-policy
  class class-default
    serverfarm udp-serverfarm
policy-map type loadbalance first-match udp-lb-policy:30000
  class class-default
    serverfarm udp-serverfarm:30000
policy-map type loadbalance first-match web->app-lb-policy
  class cm-web->app
    sticky-serverfarm customer_gold1-web->app
```

```
policy-map type loadbalance first-match web-lb-policy
  class cm-http
    sticky-serverfarm customer_gold1-http

policy-map multi-match app->db-lb
  class app->db-vip
    loadbalance vip inservice
    loadbalance policy app->db-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 3 vlan 401
policy-map multi-match lb-policy
  class web-vip
    loadbalance vip inservice
    loadbalance policy web-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 1 vlan 201
    connection advanced-options tcp_pm
  class udp-vip
    loadbalance vip inservice
    loadbalance policy udp-lb-policy
    loadbalance vip icmp-reply
    nat dynamic 11 vlan 201
    connection advanced-options udp_pm
  class udp-vip:30000
    loadbalance vip inservice
    loadbalance policy udp-lb-policy:30000
    loadbalance vip icmp-reply active
    nat dynamic 12 vlan 201
    connection advanced-options udp_pm
policy-map multi-match web->app-lb
  class web->app-vip
    loadbalance vip inservice
    loadbalance policy web->app-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 2 vlan 301

service-policy input management-traffic

interface vlan 60
  description mgmt
  ip address 192.168.60.21 255.255.255.0
  peer ip address 192.168.60.22 255.255.255.0
  no shutdown
interface vlan 201
  description web tier
  ip address 11.1.1.22 255.255.255.0
  alias 11.1.1.21 255.255.255.0
  peer ip address 11.1.1.23 255.255.255.0
  access-group input web-acl
  nat-pool 1 11.1.1.24 11.1.1.30 netmask 255.255.255.0 pat
  nat-pool 11 11.1.1.41 11.1.1.41 netmask 255.255.255.255
  nat-pool 12 11.1.1.42 11.1.1.42 netmask 255.255.255.255
  service-policy input lb-policy
  no shutdown
interface vlan 301
  description app tier
  ip address 11.1.2.22 255.255.255.0
  alias 11.1.2.21 255.255.255.0
  peer ip address 11.1.2.23 255.255.255.0
  access-group input app-acl
  nat-pool 2 11.1.2.24 11.1.2.30 netmask 255.255.255.0 pat
  service-policy input web->app-lb
  no shutdown
interface vlan 401
```

```
  description db tier
  ip address 11.1.3.22 255.255.255.0
  alias 11.1.3.21 255.255.255.0
  peer ip address 11.1.3.23 255.255.255.0
  access-group input db-acl
  nat-pool 3 11.1.3.24 11.1.3.30 netmask 255.255.255.0 pat
  service-policy input app->db-lb
  no shutdown
interface vlan 501
  no normalization

ft track host 1

ip route 0.0.0.0 0.0.0.0 11.1.1.1
ip route 192.168.0.0 255.255.0.0 192.168.60.1

snmp-server community public group Network-Monitor

snmp-server host 192.168.11.39 traps version 2c public

snmp-server host 192.168.11.41 traps version 2c public

snmp-server trap-source vlan 60

snmp-server enable traps rate-limit bandwidth
snmp-server enable traps slb serverfarm
snmp-server enable traps slb vserver
snmp-server enable traps slb real
snmp-server enable traps syslog
snmp-server enable traps snmp authentication
snmp-server enable traps snmp linkup
snmp-server enable traps snmp linkdown
username admin password 5 $1$d0VCV53d$J1bjlQoaSO8xhAoYReeh90  role Admin domain
default-domain

dc02-ace-1/customer_gold1#
```

ASR 1000 PE Gold Tenant Configuration

This section provides an ASR 1000 PE Gold tenant configuration.

```
dc02-asr1k-pe1#sh run vrf customer_gold1
Building configuration...

Current configuration : 1386 bytes
vrf definition customer_gold1
 rd 21:1
 route-target export 21:1
 route-target import 31:1
 !
 address-family ipv4
 exit-address-family
!
!
interface TenGigabitEthernet0/2/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/2/0.201
 encapsulation dot1Q 201
 vrf forwarding customer_gold1
 ip address 10.1.1.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input gold-in
 service-policy output gold-out-parent
!
interface TenGigabitEthernet0/3/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/3/0.201
 encapsulation dot1Q 201
 vrf forwarding customer_gold1
 ip address 10.1.4.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input gold-in
 service-policy output gold-out-parent
!
router bgp 109
 !
 address-family ipv4 vrf customer_gold1
  neighbor 10.1.1.2 remote-as 65501
  neighbor 10.1.1.2 activate
  neighbor 10.1.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.1.4.2 remote-as 65501
  neighbor 10.1.4.2 activate
  neighbor 10.1.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
!
ip route vrf customer_gold1 169.0.0.0 255.0.0.0 Null0 track 1
end
```

Nexus 1000V Gold Configuration

This section provides a Nexus 1000V Gold configuration.

```
#---- one time config
class-map type qos match-all gold-ef
  match dscp 46
policy-map type qos gold
  class gold-ef
    set cos 5
    police cir 50 mbps bc 200 ms conform set-cos-transmit 5 violate drop
    set dscp 40
  class class-default
    police cir 250 mbps bc 200 ms conform set-cos-transmit 2 violate set dscp dscp 
table pir-markdown-map
    set qos-group 88
    set dscp 16

port-profile type vethernet gold-profile
  switchport mode access
  service-policy input gold
  pinning id 2
  no shutdown
  state enabled
```

```
#--- once for each tenant
vlan 201,301,401,1601

vservice node gold001-vsg01 type vsg
  ip address 192.168.54.51
  adjacency l3
  fail-mode open
vservice node gold001-vsg02 type vsg
  ip address 192.168.54.61
  adjacency l3
  fail-mode open

vservice path gold001-tier1
  node gold001-vsg01 profile gold-tier1 order 10
vservice path gold001-tier2
  node gold001-vsg01 profile gold-tier2 order 10
vservice path gold001-tier3
  node gold001-vsg01 profile gold-tier3 order 10
vservice path gold001-dmz
  node gold001-vsg02 profile gold-dmz order 10

port-profile type ethernet system-data-uplink
  switchport trunk allowed vlan add 201,301,401,1601

port-profile type vethernet gold001-v0201
  vmware port-group
  inherit port-profile gold-profile
  switchport access vlan 201
  state enabled
  org root/gold001
  vservice path gold001-tier1
port-profile type vethernet gold001-v0301
  vmware port-group
  inherit port-profile gold-profile
  switchport access vlan 301
  state enabled
  org root/gold001
  vservice path gold001-tier2
port-profile type vethernet gold001-v0401
  vmware port-group
  inherit port-profile gold-profile
  switchport access vlan 401
  state enabled
  org root/gold001
  vservice path gold001-tier3
port-profile type vethernet gold001-v1601
  vmware port-group
  inherit port-profile gold-profile
  switchport access vlan 1601
  state enabled
  org root/gold001
  vservice path gold001-dmz
```

Configuration Template for Silver Service Class

This section presents the configuration templates for the Silver service class.

- Aggregation Nexus 7000 Silver Configuration, page C-17

- ACE Silver Tenant Configuration, page C-18
- ASR 1000 PE Silver Tenant Configuration, page C-22
- Nexus 1000V Silver Configuration, page C-23

Aggregation Nexus 7000 Silver Configuration

```
vrf context customer_silver1

interface Vlan501
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.1.2/24
  no ipv6 redirects
  no ip arp gratuitous hsrp duplicate
  hsrp version 2
  hsrp 501
    preempt
    priority 150
    ip 11.2.1.1

interface Vlan601
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.2.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 601
    preempt
    priority 150
    ip 11.2.2.1

interface Vlan701
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.3.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 701
    preempt
    priority 150
    ip 11.2.3.1

interface port-channel343.501
  encapsulation dot1q 501
  service-policy type qos input ingress-qos-policy
  vrf member customer_silver1
  ip address 10.2.34.3/24
  no shutdown

interface Ethernet3/9.501
  encapsulation dot1q 501
  service-policy type qos input ingress-qos-policy
  vrf member customer_silver1
  ip address 10.2.1.2/24
  no ip arp gratuitous hsrp duplicate
  no shutdown

interface Ethernet4/9.501
  encapsulation dot1q 501
  service-policy type qos input ingress-qos-policy no-stats
  vrf member customer_silver1
  ip address 10.2.3.2/24
  no ip arp gratuitous hsrp duplicate
  no shutdown

router bgp 65501
  vrf customer_silver1
    graceful-restart-helper
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
      additional-paths send
      additional-paths receive
    neighbor 10.2.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.2.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.2.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community
```

ACE Silver Tenant Configuration

This section provides an ACE Silver tenant configuration.

```
dc02-ace-3/customer_silver1# sh run
Generating configuration....

logging enable
logging standby
logging timestamp
logging trap 6
logging facility 17
logging device-id context-name
logging host 192.168.11.100 udp/514
no logging message 106023

arp interval 1440

access-list app-acl line 8 extended permit ip any any
access-list capture-list line 8 extended permit ip any any
access-list db-acl line 8 extended permit ip any any
access-list web-acl line 8 extended deny udp 11.0.0.0 255.0.0.0 eq tftp any
access-list web-acl line 16 extended deny udp 11.0.0.0 255.0.0.0 eq 30000 any
access-list web-acl line 24 extended permit ip any any

probe ftp ftp-probe
  interval 2
  faildetect 5
  passdetect interval 2
  passdetect count 5
  receive 1
  expect status 200 400
  connection term forced
probe http http-probe
  interval 2
  faildetect 5
  passdetect interval 2
  passdetect count 5
  receive 1
  expect status 200 400
  connection term forced

rserver host app-server1
  ip address 11.2.2.11
  inservice
rserver host app-server2
  ip address 11.2.2.12
  inservice
rserver host app-server3
  ip address 11.2.2.13
  inservice
rserver host db-server1
  ip address 11.2.3.11
  inservice
rserver host db-server2
  ip address 11.2.3.12
  inservice
rserver host db-server3
  ip address 11.2.3.13
  inservice
rserver host udp-host
  ip address 11.2.1.100
  inservice
rserver host web-server1
  ip address 11.2.1.11
  inservice
rserver host web-server2
  ip address 11.2.1.12
  inservice
rserver host web-server3
  ip address 11.2.1.13
  inservice
rserver host web-spirent
  ip address 11.2.1.151
  inservice

serverfarm host app-serverfarm
  rserver app-server1
    inservice
  rserver app-server2
    inservice
  rserver app-server3
    inservice
serverfarm host db-serverfarm
  rserver db-server1
    inservice
  rserver db-server2
    inservice
  rserver db-server3
    inservice
serverfarm host udp-serverfarm
  rserver udp-host
    inservice
```

```
serverfarm host web-serverfarm
  rserver web-server1
  rserver web-server2
  rserver web-server3
  rserver web-spirent
    inservice

parameter-map type connection tcp_pm
  set tcp wan-optimization rtt 0
parameter-map type connection udp_pm
  set timeout inactivity 300

sticky http-cookie customer_gold1-http-cookie customer_gold1-http
  cookie insert browser-expire
  serverfarm web-serverfarm
  timeout 10
  replicate sticky
sticky http-cookie customer_gold1-web-app-cookie customer_gold1-web->app
  cookie insert browser-expire
  serverfarm app-serverfarm
  timeout 10
  replicate sticky
sticky ip-netmask 255.255.255.255 address both customer_gold1-app->db
  serverfarm db-serverfarm
  timeout 10
  replicate sticky

class-map type http loadbalance match-any cm-app-subnet
  2 match source-address 11.2.2.0 255.255.255.0
class-map type http loadbalance match-any cm-http
  2 match http url /.*.txt
  3 match http url /.*.html
class-map type http loadbalance match-any cm-web-subnet
  2 match source-address 11.2.1.0 255.255.255.0

class-map match-all app->db-vip
  2 match virtual-address 11.2.3.111 tcp eq www
class-map type http loadbalance match-all cm-app->db
  2 match class-map cm-http
  3 match class-map cm-app-subnet
class-map type http loadbalance match-all cm-web->app
  2 match class-map cm-http
  3 match class-map cm-web-subnet
class-map type management match-any management-traffic
  2 match protocol ssh any
  3 match protocol http any
  4 match protocol https any
  5 match protocol icmp any
  6 match protocol telnet any
  7 match protocol snmp source-address 192.168.0.0 255.255.0.0
class-map match-all udp-vip
  2 match virtual-address 11.2.1.111 udp eq 69
class-map match-all web->app-vip
  2 match virtual-address 11.2.2.111 tcp eq www
class-map match-all web-vip
  2 match virtual-address 11.2.1.111 tcp eq www

policy-map type management first-match management-traffic
  class management-traffic
    permit

policy-map type loadbalance first-match app->db-lb-policy
  class cm-app->db
    sticky-serverfarm customer_gold1-app->db
```

```
policy-map type loadbalance first-match udp-lb-policy
  class class-default
    serverfarm udp-serverfarm
policy-map type loadbalance first-match web->app-lb-policy
  class cm-web->app
    sticky-serverfarm customer_gold1-web->app
policy-map type loadbalance first-match web-lb-policy
  class cm-http
    sticky-serverfarm customer_gold1-http

policy-map multi-match app->db-lb
  class app->db-vip
    loadbalance vip inservice
    loadbalance policy app->db-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 3 vlan 701
  class udp-vip
policy-map multi-match lb-policy
  class web-vip
    loadbalance vip inservice
    loadbalance policy web-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 1 vlan 501
    connection advanced-options tcp_pm
  class udp-vip
    loadbalance vip inservice
    loadbalance policy udp-lb-policy
    loadbalance vip icmp-reply
    nat dynamic 11 vlan 501
    connection advanced-options udp_pm
policy-map multi-match web->app-lb
  class web->app-vip
    loadbalance vip inservice
    loadbalance policy web->app-lb-policy
    loadbalance vip icmp-reply active
    nat dynamic 2 vlan 601

service-policy input management-traffic

interface vlan 60
  description mgmt
  ip address 192.168.60.61 255.255.255.0
  peer ip address 192.168.60.62 255.255.255.0
  no shutdown
interface vlan 501
  description web tier
  ip address 11.2.1.22 255.255.255.0
  alias 11.2.1.21 255.255.255.0
  peer ip address 11.2.1.23 255.255.255.0
  access-group input web-acl
  nat-pool 1 11.2.1.24 11.2.1.30 netmask 255.255.255.0 pat
  nat-pool 11 11.2.1.41 11.2.1.41 netmask 255.255.255.0
  service-policy input lb-policy
  no shutdown
interface vlan 601
  description app tier
  ip address 11.2.2.22 255.255.255.0
  alias 11.2.2.21 255.255.255.0
  peer ip address 11.2.2.23 255.255.255.0
  access-group input app-acl
  nat-pool 2 11.2.2.24 11.2.2.30 netmask 255.255.255.0 pat
  service-policy input web->app-lb
  no shutdown
interface vlan 701
```

```
  description db tier
  ip address 11.2.3.22 255.255.255.0
  alias 11.2.3.21 255.255.255.0
  peer ip address 11.2.3.23 255.255.255.0
  access-group input db-acl
  nat-pool 3 11.2.3.24 11.2.3.30 netmask 255.255.255.0 pat
  service-policy input app->db-lb
  no shutdown

ip route 0.0.0.0 0.0.0.0 11.2.1.1
ip route 192.168.0.0 255.255.0.0 192.168.60.1

dc02-ace-3/customer_silver1
```

ASR 1000 PE Silver Tenant Configuration

This section provides an ASR 1000 PE CE Silver tenant configuration.

```
vrf definition customer_silver1
 rd 22:1
 route-target export 22:1
 route-target import 32:1
 !
 address-family ipv4
 exit-address-family
!
!
interface TenGigabitEthernet0/2/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/2/0.501
 encapsulation dot1Q 501
 vrf forwarding customer_silver1
 ip address 10.2.1.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input silver-in
 service-policy output silver-out-parent
!
interface TenGigabitEthernet0/3/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/3/0.501
 encapsulation dot1Q 501
 vrf forwarding customer_silver1
 ip address 10.2.4.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy input silver-in
 service-policy output silver-out-parent
!
router bgp 109
 !
 address-family ipv4 vrf customer_silver1
```

```
  import path selection all
  import path limit 10
  bgp advertise-best-external
  neighbor 10.2.1.2 remote-as 65501
  neighbor 10.2.1.2 activate
  neighbor 10.2.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.2.4.2 remote-as 65501
  neighbor 10.2.4.2 activate
  neighbor 10.2.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
!
ip route vrf customer_silver1 169.0.0.0 255.0.0.0 Null0 track 1
end
```

Nexus 1000V Silver Configuration

This section provides a Nexus 1000v Silver configuration.

```
#---- one time config
policy-map type qos silver
  class class-default
    set qos-group 89
    police cir 62500 kbps bc 200 ms conform set-cos-transmit 2 violate set dscp dscp
table pir-markdown-map
    set dscp 16

port-profile type vethernet silver-profile
  switchport mode access
  service-policy input silver
  pinning id 3
  no shutdown
  state enabled

#--- once for each tenant
vlan 501,601,701

vservice node silver001-vsg01 type vsg
  ip address 192.168.54.101
  adjacency l3
  fail-mode open

vservice path silver001-tier1
  node silver001-vsg01 profile silver-tier1 order 10
vservice path silver001-tier2
  node silver001-vsg01 profile silver-tier2 order 10
vservice path silver001-tier3
  node silver001-vsg01 profile silver-tier3 order 10

port-profile type ethernet system-data-uplink
  switchport trunk allowed vlan add 501,601,701

port-profile type vethernet silver001-v0501
  vmware port-group
  inherit port-profile silver-profile
  switchport access vlan 501
  state enabled
  org root/silver001
  vservice path silver001-tier1
port-profile type vethernet silver001-v0601
  vmware port-group
  inherit port-profile silver-profile
  switchport access vlan 601
  state enabled
  org root/silver001
  vservice path silver001-tier2
port-profile type vethernet silver001-v0701
  vmware port-group
  inherit port-profile silver-profile
  switchport access vlan 701
  state enabled
  org root/silver001
  vservice path silver001-tier3
```

Configuration Template for Bronze Service Class

This section presents the configuration templates for the Bronze service class.

- Aggregation Nexus 7000 Bronze Configuration, page C-24
- ASR 1000 PE Bronze Configuration, page C-25
- Nexus 1000V Bronze Configuration, page C-26

Aggregation Nexus 7000 Bronze Configuration

This section provides an aggregation Nexus 7000 Bronze configuration.

```
interface Vlan801
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_bronze1
  no ip redirects
  ip address 11.3.1.2/24
  no ipv6 redirects
  no ip arp gratuitous hsrp duplicate
  hsrp version 2
  hsrp 801
    preempt
    priority 150
    ip 11.3.1.1

interface port-channel343.801
  encapsulation dot1q 801
  service-policy type qos input ingress-qos-policy no-stats
  vrf member customer_bronze1
  ip address 10.3.34.3/24
  no shutdown

interface Ethernet3/9.801
  encapsulation dot1q 801
  service-policy type qos input ingress-qos-policy no-stats
  vrf member customer_bronze1
  ip address 10.3.1.2/24
  no ip arp gratuitous hsrp duplicate
  no shutdown

interface Ethernet4/9.801
  encapsulation dot1q 801
  service-policy type qos input ingress-qos-policy no-stats
  vrf member customer_bronze1
  ip address 10.3.3.2/24
  no ip arp gratuitous hsrp duplicate
  no shutdown

vrf context customer_bronze1
router bgp 65501
  vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
      nexthop trigger-delay critical 100 non-critical 300
    neighbor 10.3.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.34.4
      remote-as 65501
      address-family ipv4 unicast
```

ASR 1000 PE Bronze Configuration

This section provides an ASR 1000 PE Bronze configuration.

```
vrf definition customer_bronze1
 rd 23:1
 route-target export 23:1
 route-target import 33:1
 !
 address-family ipv4
 exit-address-family
!
!
interface TenGigabitEthernet0/2/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/2/0.801
 encapsulation dot1Q 801
 vrf forwarding customer_bronze1
 ip address 10.3.1.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy output bronze-out-parent
!
interface TenGigabitEthernet0/3/0
 no ip address
 load-interval 30
 carrier-delay up 60
 cdp enable
!
interface TenGigabitEthernet0/3/0.801
 encapsulation dot1Q 801
 vrf forwarding customer_bronze1
 ip address 10.3.4.1 255.255.255.0
 ip flow monitor input_monitor input
 ip flow monitor output_monitor output
 plim qos input map cos  5  queue strict-priority
 service-policy output bronze-out-parent
!
router bgp 109
 !
 address-family ipv4 vrf customer_bronze1
  neighbor 10.3.1.2 remote-as 65501
  neighbor 10.3.1.2 activate
  neighbor 10.3.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.3.4.2 remote-as 65501
  neighbor 10.3.4.2 activate
  neighbor 10.3.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
!
ip route vrf customer_bronze1 169.0.0.0 255.0.0.0 Null0 track 1
```

Nexus 1000V Bronze Configuration

This section provides a Nexus 1000V Bronze configuration.

```
#---- one time config
policy-map type qos bronze
  class class-default
    set cos 0
    police cir 500 mbps bc 200 ms conform transmit violate drop
    set dscp 0

port-profile type vethernet bronze-profile
  switchport mode access
  service-policy input bronze
  pinning id 3
  no shutdown
  state enabled

#--- once for each tenant
vlan 801

vservice node bronze001-vsg01 type vsg
  ip address 192.168.54.201
  adjacency l3
  fail-mode open

vservice path bronze001-vmdc
  node bronze001-vsg01 profile bronze order 10

port-profile type ethernet system-data-uplink
  switchport trunk allowed vlan add 801

port-profile type vethernet bronze001-v0801
  vmware port-group
  inherit port-profile bronze-profile
  switchport access vlan 801
  state enabled
  org root/vbronze001
  vservice path bronze001-vmdc
```

Configuration Template for Copper Service Class

This section presents the configuration templates for the Copper service class.

- Aggregation Nexus 7000 Copper Configuration, page C-27
- ASA Copper Configuration, page C-28
- ASR 1000 PE Copper Configuration, page C-30
- Nexus 1000V Copper Configuration, page C-30

Aggregation Nexus 7000 Copper Configuration

This section provides an aggregation Nexus 7000 Copper configuration.

```
ip route 100.201.1.0/24 100.200.1.61 tag 1111

interface Vlan2000
  no shutdown
  no ip redirects
  ip address 100.200.1.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 2000
    preempt
    priority 110
    ip 100.200.1.1

interface Vlan2001
  description test for snmptrap
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_smb1
  no ip redirects
  ip address 11.4.1.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 2001
    preempt
    priority 110
    ip 11.4.1.1

interface Vlan3001
  no shutdown
  ip flow monitor fm_vmdc23 input sampler sp_vmdc23
  vrf member customer_smb1
  no ip redirects
  ip address 10.9.1.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 3001
    preempt
    priority 110
    ip 10.9.1.1

interface port-channel343
  service-policy type qos input ingress-qos-policy
  service-policy type queuing output vmdc23-8e-4q4q-out
  ip address 100.200.0.17/30

interface Ethernet3/9.2000
  description PC-to-PE1
  encapsulation dot1q 2000
  service-policy type qos input ingress-qos-policy
  ip address 100.200.0.2/30
  no shutdown
```

```
interface Ethernet4/9.2000
  encapsulation dot1q 2000
  service-policy type qos input ingress-qos-policy no-stats
  ip address 100.200.0.10/30
  no shutdown

router bgp 65501

  address-family ipv4 unicast
    redistribute direct route-map DC2-INTERNET-SUBNET
    redistribute static route-map SERVICED-PREFIXES-SET-COMM
    nexthop trigger-delay critical 100 non-critical 300

  neighbor 100.200.0.1
    remote-as 109
    address-family ipv4 unicast
      send-community
      weight 60000
      next-hop-self
  neighbor 100.200.0.9
    remote-as 109
    address-family ipv4 unicast
      send-community
      next-hop-self
  neighbor 100.200.0.18 remote-as 65501
    address-family ipv4 unicast
      route-map filter-100.200.1.61 out
      next-hop-self
```

ASA Copper Configuration

This section provides an ASA Copper configuration.

```
interface Port-channel1.2000
 nameif outside
 security-level 0
 ip address 100.200.1.61 255.255.255.0 standby 100.200.1.62
!
interface Port-channel1.3001
 nameif smb1
 security-level 100
 ip address 10.9.1.61 255.255.255.0 standby 10.9.1.62
!

object network smb1-mapped
 range 100.201.1.1 100.201.1.10
object network smb1-real
 subnet 11.4.1.0 255.255.255.0
object network smb-1-server1
 host 11.4.1.11
object network smb-1-server2
 host 11.4.1.12
object network smb-1-server3
 host 11.4.1.13

object network smb-1-server21
 host 11.4.1.21
object network smb-1-server22
 host 11.4.1.22
object network smb-1-server23
 host 11.4.1.23
```

```
object network smb-1-server24
 host 11.4.1.24
object network smb-1-server25
 host 11.4.1.25
object network smb-1-server26
 host 11.4.1.26
object network smb-1-server27
 host 11.4.1.27
object network smb-1-server28
 host 11.4.1.28
object network smb-1-server29
 host 11.4.1.29
object network smb-1-server30
 host 11.4.1.30

mtu smb1 1500

monitor-interface outside
icmp unreachable rate-limit 1 burst-size 1
no asdm history enable
arp timeout 14400
!
object network smb1-real
 nat (smb1,outside) dynamic smb1-mapped
object network smb-1-server1
 nat (smb1,outside) static 100.201.1.11
object network smb-1-server2
 nat (smb1,outside) static 100.201.1.12
object network smb-1-server3
 nat (smb1,outside) static 100.201.1.13

object network smb-1-server21
 nat (smb1,outside) static 100.201.1.21
object network smb-1-server22
 nat (smb1,outside) static 100.201.1.22
object network smb-1-server23
 nat (smb1,outside) static 100.201.1.23
object network smb-1-server24
 nat (smb1,outside) static 100.201.1.24
object network smb-1-server25
 nat (smb1,outside) static 100.201.1.25
object network smb-1-server26
 nat (smb1,outside) static 100.201.1.26
object network smb-1-server27
 nat (smb1,outside) static 100.201.1.27
object network smb-1-server28
 nat (smb1,outside) static 100.201.1.28
object network smb-1-server29
 nat (smb1,outside) static 100.201.1.29
object network smb-1-server30
 nat (smb1,outside) static 100.201.1.30

route outside 0.0.0.0 0.0.0.0 100.200.1.1 1
route smb1 11.4.1.0 255.255.255.0 10.9.1.1 1
```

Note The configuration above is for the private ip address server tenants, if public ip address server tenants, remove all the nat configurations.

ASR 1000 PE Copper Configuration

This section provides an ASR 1000 PE Copper configuration.

```
interface TenGigabitEthernet0/2/0.2000
 encapsulation dot1Q 2000
 ip address 100.200.0.1 255.255.255.252
 cdp enable
 service-policy input internet-in
 service-policy output internet-out-parent

interface TenGigabitEthernet0/3/0.2000
 encapsulation dot1Q 2000
 ip address 100.200.0.13 255.255.255.252
 cdp enable
 service-policy input internet-in
 service-policy output internet-out-parent

router bgp 109
 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both
 exit-peer-policy
 !
 bgp log-neighbor-changes
 bgp graceful-restart restart-time 120
 bgp graceful-restart stalepath-time 360
 bgp graceful-restart

 neighbor 100.200.0.2 remote-as 65501
 neighbor 100.200.0.14 remote-as 65501
 !
 address-family ipv4
  bgp additional-paths install
  redistribute connected
  redistribute static

  neighbor 100.200.0.2 activate
  neighbor 100.200.0.2 route-map DC2_INT_PREFER in
  neighbor 100.200.0.14 activate
  neighbor 100.200.0.14 route-map DC2_INT_PREFER in
  maximum-paths 2
  maximum-paths ibgp 2
 exit-address-family
```

Nexus 1000V Copper Configuration

This section provides a Nexus 1000V Copper configuration.

```
#---- one time config
policy-map type qos bronze
  class class-default
    set cos 0
    police cir 500 mbps bc 200 ms conform transmit violate drop
    set dscp 0

port-profile type vethernet smb-profile
  switchport mode access
  service-policy input bronze
```

```
    pinning id 3
    no shutdown
    state enabled

  #--- once for each tenant
  vlan 2001

  vservice node smb001-vsg01 type vsg
    ip address 192.168.54.151
    adjacency l3
    fail-mode open

  vservice path smb001-vmdc
    node smb001-vsg01 profile smb order 10

  port-profile type ethernet system-data-uplink
    switchport trunk allowed vlan add 2001

  port-profile type vethernet smb001-v2001
    vmware port-group
    inherit port-profile smb-profile
    switchport access vlan 2001
    state enabled
    org root/smb001
    vservice path smb001-vmdc
```

Configuration Template for Nexus 5548 ICS switch

The Nexus 5000 ICS switch does not have any per-tenant configurations, other than the VLANs to be allowed. The data VLANs used by tenants can be added on the Nexus 5000 ICS switch, but this should be planned and configured in advance for different ranges needed. Further modifications and updates can be done as tenants are added and deleted as required.

LAN Configuration

The following configuration shows the port-channel configuration between the Nexus 5548 ICS switch and Nexus 7004 Aggregation switches:

```
#Portchannel between dc02-n5k-ics1 and dc02-n7k-agg1

interface port-channel534
  description vPC to N7K-Aggs
  switchport mode trunk
  spanning-tree port type network
  speed 10000
  vpc 4000
```

The following configuration shows the port-channel configuration between the Nexus 5548 ICS switch and the UCS Fabric Interconnect 6248. There are two port-channels, 88 and 89, that carry all LAN data traffic in the data network.

```
interface port-channel88
  description vPC to dc02-ucs01-a
  switchport mode trunk
  switchport trunk allowed vlan 
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,1990,2001-
2010
  spanning-tree port type edge trunk
  vpc 88
```

```
interface port-channel89
  description vPC to dc02-ucs01-b
  switchport mode trunk
  switchport trunk allowed vlan
201-210,301-310,401-410,501-520,601-620,701-720,801-820,1601-1610,1801-1860,1990,2001-
2010
  spanning-tree port type edge trunk
  vpc 89
```

Only the VLANs that carry the LAN data traffic for all the tenants (Gold, Silver, Bronze, and Copper) are allowed on the port-channels going to the UCS. A list of all data VLANs is obtained from the above configuration.

The following configuration shows the port-channel between the Nexus 5548 ICS switch and NetApp Filers 6040. This port-channel carries only the NFS traffic, and hence only the NFS VLAN (1990) is allowed on the port-channel. There are two port-channels with one going to each of the filers (Filer-A and Filer-B).

```
interface port-channel26
  description vPC to netapp -A
  switchport mode trunk
  switchport trunk allowed vlan 1990
  service-policy type queuing input vmdc-nas-in-policy
  service-policy type queuing output vmdc-nas-out-policy
  vpc 26

interface port-channel28
  description vPC to netapp -B
  switchport mode trunk
  switchport trunk allowed vlan 1990
  service-policy type queuing input vmdc-nas-in-policy
  service-policy type queuing output vmdc-nas-out-policy
  vpc 28
```

SAN Configuration

The following configuration shows the port-channel between Nexus 5548 ICS switch and NetApp Filers and/or UCS Fabric Interconnect for FCP connectivity:

```
interface san-port-channel 2
  channel mode active
  switchport mode F
  switchport description Port-channel UCS FI/Filers & 5k ICS switch

interface fc2/10
  switchport mode F
  switchport description to_UCS_fi
  channel-group 2 force
  no shutdown

interface fc2/12
  switchport mode F
  switchport description to netapp filer
  no shutdown
```

APPENDIX D

Glossary

Numbers

1R2C	One Rate 2 Colors
2R3C	Dual Rate Three Color

A

ACE	Application Control Engine (Cisco)
ACL	Access Control List
AH	Authentication Header
ALUA	Asymmetric Logical Unit Access
ARP	Address Resolution Protocol
ASA	Adaptive Security Appliance (Cisco)
ASIC	Application-Specific Integrated Circuit
ASR	Aggregation Services Router (Cisco)

B

BECN	Backward Explicit Congestion Notification (Frame Relay QoS)
BGP	Border Gateway Protocol
BPDU	Bridge Protocol Data Unit
BW	Bandwidth

C

CAPEX	Capital Expense
CBWFQ	Class-based Weighted Fair Queuing (QoS)
CDP	Cisco Discovery Protocol (Cisco)
CFS	Cisco Fabric Service (Cisco)
CFSoE	Cisco Fabric Service over Ethernet
CIFS	Common Internet File System
CIR	Committed Information Rate
CLI	Command Line Interface
CNA	Converged Network Adapter

CoS	Class of Service
CPU	Central Processing Unit
CSD	Cisco Secure Desktop (Cisco)

D

DB	Database
DC	Data Center
DCBX	Data Center Bridging Exchange (IEEE 802.1AB)
DCE	Data Center Ethernet
DC-PE	Data Center Provider Edge
DFC	Distributed Forwarding Card (Cisco)
DMZ	Demilitarized Zone
DRS	Distributed Resource Scheduling
DSCP	Differentiated Services Code Point
DSN	Data Center Service Node
DVS	Distributed Virtual Switch
DWRR	Deficit Weighted Round Robin

E

eBGP	External Border Gateway Protocol
EH	End-host (mode)
EoMPLS	Ethernet over MPLS
ERSPAN	Encapsulated Remote Switched Port Analyzer
ESP	Encapsulating Standard Protocol
ETS	Enhanced Transmission Selection

F

FAS	Fabric-Attached Storage
FC	Fibre Channel
FC-GS	Fibre Channel Generic Services
FCoE	Fibre Channel over Ethernet
FHRP	First-Hop Redundancy Protocol
FI	Fabric Interconnect
FIB	Forwarding Information Base
FSPF	Fabric Shortest Path First
FT	Fault-Tolerant
FTP	File Transfer Protocol
FWSM	Firewall Services Module (Cisco)

G

GE	Gigabit Ethernet
GUI	Graphical User Interface

H

HA	High Availability
HBA	Host Bus Adapter
HQoS	Hierarchical QoS
HSRP	Hot Standby Router Protocol
HTTP	Hypertext Transfer Protocol

I

IaaS	Infrastructure as a Service
iBGP	Internal Border Gateway Protocol
ICMP	Internet Control Message Protocol
ICS	Integrated Compute and Storage
IETF	Internet Engineering Task Force
IKE	Internet Key Exchange
IOPS	Input/Output Operations Per Second
IPsec	IP security
IPv4	Internet Protocol version 4
ISAKMP	Internet Security Association and Key Management Protocol
iSCSI	Internet Small Computer System Interface
ISSU	In-Service System Upgrade

L

L2	Layer 2
L3	Layer 3
L4	Layer 4
L7	Layer 7
LACP	Link Aggregation Control Protocol
LAMP	Linux, Apache, MySQL, PHP, Perl, and Python
LAN	Local Area Network
LDP	Label Distribution Protocol
LSR	Label Switching Router
LUN	Logical Unit Number

M

MAC	Media Access Control
MDS	Multilayer Director Switch

MEC	Multi-Chassis EtherChannel
MP-iBGP	Multiprotocol Internal Border Gateway Protocol
MPLS	Multiprotocol Label Switching
MPLS-TC	MPLS Traffic Class
MQC	Modular QoS CLI (Cisco)
MST	Multiple Spanning Tree
MSTI	Multiple Spanning Tree Instance
MTU	Maximum Transmission Unit

N

NAS	Network-Attached Storage
NAT	Network Address Translation
NFS	Network File System
NGN	Next Generation Network
NIC	Network Interface Card
NPIV	n-port Identifier Virtualization
NSF	Nonstop Forwarding (Cisco)
NSR	Nonstop Routing

O

OPEX	Operating Expense
OS	Operating System
OSPF	Open Shortest Path First

P

PB	PetaByte
PBR	Policy Based Routing (Cisco)
PE	Provider Edge
PFC	Priority Flow Control
PHB	Per-Hop Behavior
PIR	Peak Information Rate
Pod	Point of Delivery. A basic infrastructure module that is a physical, repeatable construct with predictable infrastructure characteristics and deterministic functions. A pod identifies a modular unit of data center components and enables customers to add network, compute, and storage resources incrementally.
PQ	Priority Queuing
PVST+	Per-VLAN Spanning Tree Plus
PVT	Private

Q

QoS	Quality of Service

R

RAID	Redundant Array of Independent Disks
RAID-DP	RAID-level Double Parity (NetApp)
RHEL	Red Hat Enterprise Linux
RHI	Route Health Injection
RU	Rack Unit

S

SA	Security Association
SADB	Security Association Database
SAN	Storage Area Network
SATA	Serial AT Attachment
SCSI	Small Computer System Interface
SLA	Service Level Agreement
SLB	Server Load Balancing
SMB	Small/Medium Business
SNMP	Simple Network Management Protocol
SPAN	Switched Port Analyzer
SP-DC	Service Provider Data Center
SPI	Security Parameter Index
SP-NGN	Service Provider Next Generation Network
SoC	System on Chip
SSL	Secure Sockets Layer
STP	Spanning Tree Protocol
SVI	Switched Virtual Interface

T

TCAM	Ternary Content Addressable Memory
TCP	Transmission Control Protocol
TDEV	Thin Device
TFTP	Trivial File Transfer Protocol
TLS	Transport Layer Security
TTM	Time to Market

U

UCS	Unified Computing System (Cisco)
UCSM	Unified Computing System Manager (Cisco)
UDP	User Datagram Protocol

V

VCE	Virtual Computing Environment
vCPU	Virtual CPU
vDS	vNetwork Distributed Switch
VEM	Virtual Ethernet Module
vFW	Virtual Firewall
vHBA	Virtual Host Bus Adapter
VIC	Virtual Interface Card
VIF	Virtual Interface
VIP	Virtual IP
VLAN	Virtual LAN
VM	Virtual Machine
VMAC	Virtual MAC
VMDC	Virtualized Multiservice Data Center (Cisco)
VMFS	Virtual Machine File System
vMotion	Virtual Motion
VMNIC	VM Network Interface Card
vNIC	Virtual Network Interface Card
VNMC	Virtual Network Management Center (Cisco)
VoIP	Voice over IP
VOQ	Virtual Output Queue
vPC	Virtual Port-Channel
VQI	Virtual Queue Index
VRF	Virtual Routing and Forwarding
VSA	VPN Services Adapter
VSAN	Virtual SAN
VSG	Virtual Security Gateway (Cisco)
VSL	Virtual Switch Link
vSLB	Virtual Server Load Balancer
VSM	Virtual Supervisor Module
VSS	Virtual Switch System (Cisco)

W

WAMP	Windows Apache, MySQL, PHP, Perl, and Python
WRED	Weighted Random Early Detection
WWPN	World Wide Port Name