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

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

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.

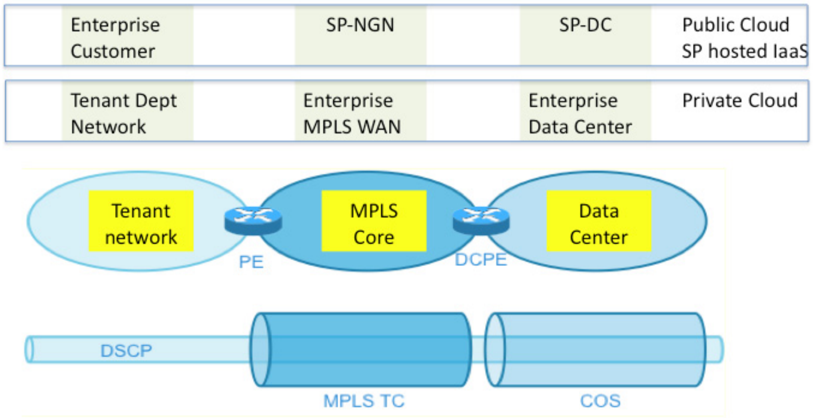
2. The **MPLS-Core network** (for example, Service Provider Next Generation Network (SP-NGN) or an Enterprise-wide MPLS-Core WAN) implements a QoS that supports the different offered services for WAN transport. The end tenant customer traffic is mapped to one of the WAN/ SP-NGN service classes based 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

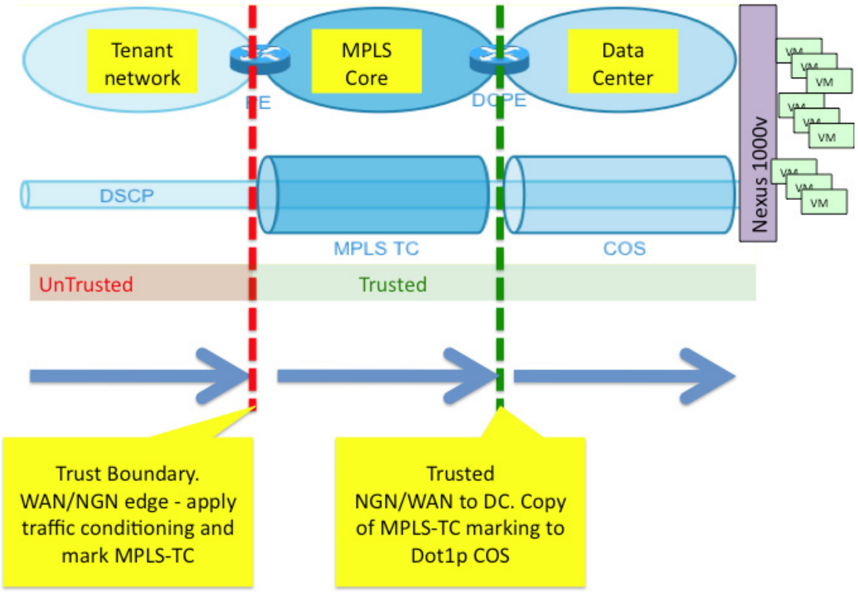
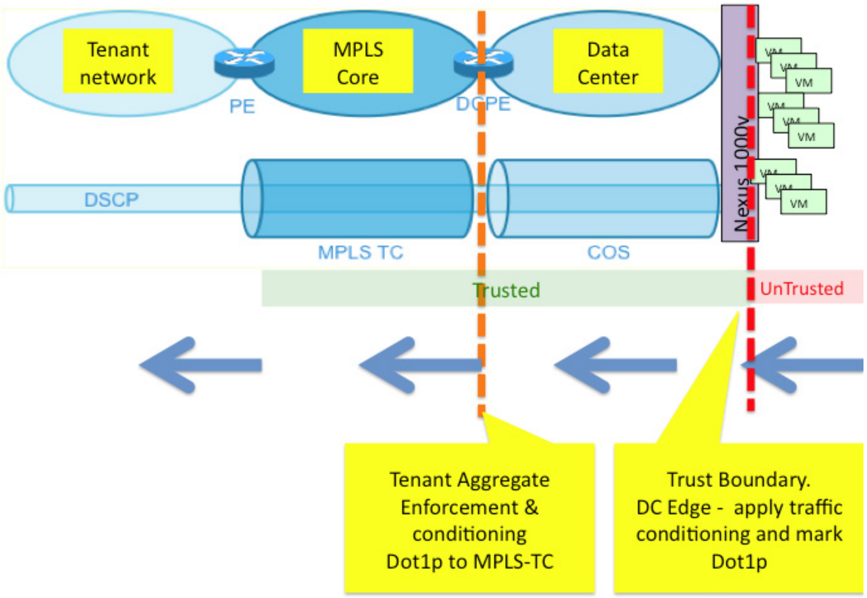

Tenant network
MPLS Core
Data Center
PE
DC PE
Nexus 1000v
VM
VM
VM
VM
VM
VM
VM
VM
VM
DSCP
MPLS TC
COS
UnTrusted
Trusted
Trust Boundary. WAN/NGN edge - apply traffic conditioning and mark MPLS-TC
Trusted NGN/WAN to DC. Copy of MPLS-TC marking to Dot1p COS

Figure 6-3 shows the trust boundaries for the DC to customer traffic.

Figure 6-3 DC to Customer Traffic Trust Boundary


Tenant network
MPLS Core
Data Center
PE
DCPE
VM
VM
VM
VM
VM
VM
VM
VM
VM
Nexus 1000v
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.

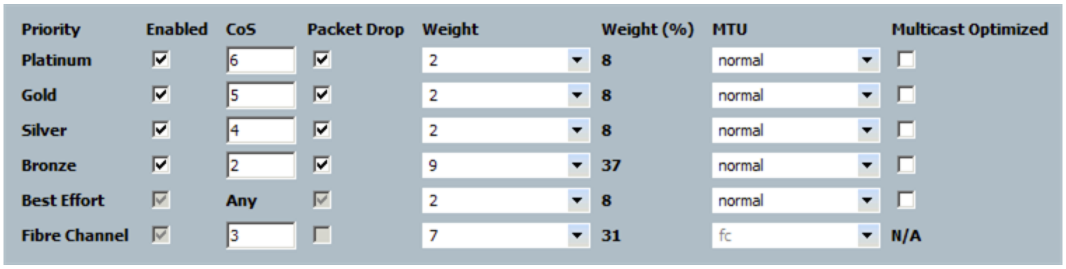
Table 6-6 shows the traffic classes to UCS QoS system classes' mapping.

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



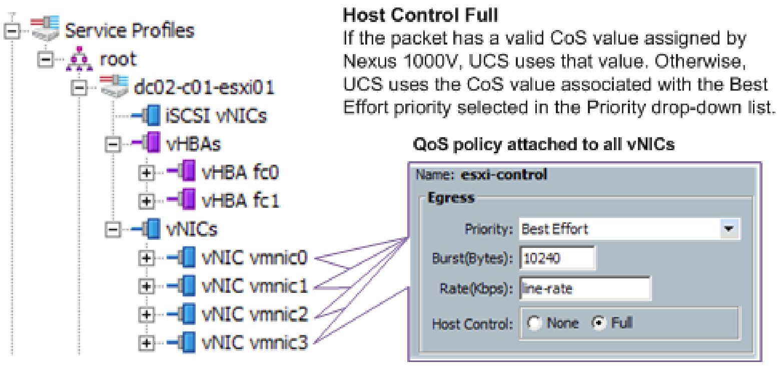
Priority	Enabled	CoS	Packet Drop	Weight	Weight (%)	MTU	Multicast Optimized
Platinum	☑	6	☑	2	8	normal	☐
Gold	☑	5	☑	2	8	normal	☐
Silver	☑	4	☑	2	8	normal	☐
Bronze	☑	2	☑	9	37	normal	☐
Best Effort	☑	Any	☑	2	8	normal	☐
Fibre Channel	☑	3	☐	7	31	fc	N/A

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.

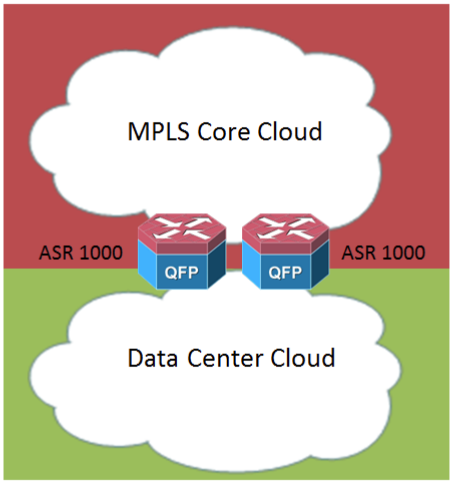
In this solution, the pipe model is used. All markings are based on Service Provider classification and use outer header QoS markings to support RFC3270, and the IP/DSCP or precedence is left untouched to support QoS transparency.

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.