Configuring VXLAN QoS

This chapter contains the following sections:

Information About VXLAN QoS

VXLAN QoS enables you to provide Quality of Service (QoS) capabilities to traffic that is tunneled in VXLAN.

Traffic in the VXLAN overlay can be assigned to different QoS properties:

- Classification traffic to assign different properties.
- Including traffic marking with different priorities.
- Queuing traffic to enable priority for the protected traffic.
- Policing for misbehaving traffic.
- Shaping for traffic that limits speed per interface.
- Properties traffic sensitive to traffic drops.

Note QoS allows you to classify the network traffic, police and prioritize the traffic flow, and provide congestion avoidance. For more information about QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

This section contains the following topics:

VXLAN QoS Terminology

This section defines VXLAN QoS terminology.

Table 1: VXLAN QoS Terminology

Term	Definition
Frames	Carries traffic at Layer 2. Layer 2 frames carry Layer 3 packets.
Packets	Carries traffic at Layer 3.
VXLAN packet	Carries original frame, encapsulated in VXLAN IP/UDP header.
Original frame	A Layer 2 or Layer 2 frame that carries the Layer 3 packet before encapsulation in a VXLAN header.
Decapsulated frame	A Layer 2 or a Layer 2 frame that carries a Layer 3 packet after the VXLAN header is decapsulated.
Ingress VTEP	The point where traffic is encapsulated in the VXLAN header and enters the VXLAN tunnel.
Egress VTEP	The point where traffic is decapsulated from the VXLAN header and exits the VXLAN tunnel.
Class of Service (CoS)	Refers to the three bits in an 802.1Q header that are used to indicate the priority of the Ethernet frame as it passes through a switched network. The CoS bits in the 802.1Q header are commonly referred to as the 802.1p bits. 802.1Q is discarded prior to frame encapsulation in a VXLAN header, where CoS value is not present in VXLAN tunnel. To maintain QoS when a packet enters the VXLAN tunnel, the type of service (ToS) and CoS values map to each other.
IP precedence	The 3 most significant bits of the ToS byte in the IP header.
Differentiated Services Code Point (DSCP)	The first six bits of the ToS byte in the IP header. DSCP is only present in an IP packet.
Explicit Congestion Notification (ECN)	The last two bits of the ToS byte in the IP header. ECN is only present in an IP packet.
QoS tags	Prioritization values carried in Layer 3 packets and Layer 2 frames. A Layer 2 CoS label can have a value ranging between zero for low priority and seven for high priority. A Layer 3 IP precedence label can have a value ranging between zero for low priority and seven for high priority. IP precedence values are defined by the three most significant bits of the 1-byte ToS byte. A Layer 3 DSCP label can have a value between 0 and 63. DSCP values are defined by the six most significant bits of the 1-byte IP ToS field.

Term	Definition
Classification	The process used for selecting traffic for QoS
Marking	The process of setting: a Layer 2 COS value in a frame, Layer 3 DSCP value in a packet, and Layer 3 ECN value in a packet. Marking is also the process of choosing different values for the CoS, DSCP, ECN field to mark packets so that they have the priority that they require during periods of congestion.
Policing	Limiting bandwidth used by a flow of traffic. Policing can mark or drop traffic.
MQC	The Cisco Modular QoS command line interface (MQC) framework, which is a modular and highly extensible framework for deploying QoS.

VXLAN QoS Features

The following topics describe the VXLAN QoS features that are supported in a VXLAN network:

Trust Boundaries

The trust boundary forms a perimeter on your network. Your network trusts (and does not override) the markings on your switch. The existing ToS values are trusted when received on in the VXLAN fabric.

Classification

You use classification to partition traffic into classes. You classify the traffic based on the port characteristics or the packet header fields that include IP precedence, differentiated services code point (DSCP), Layer 3 to Layer 4 parameters, and the packet length.

The values used to classify traffic are called match criteria. When you define a traffic class, you can specify multiple match criteria, you can choose to not match on a particular criterion, or you can determine the traffic class by matching any or all criteria.

Traffic that fails to match any class is assigned to a default class of traffic called class-default.

Marking

Marking is the setting of QoS information that is related to a packet. Packet marking allows you to partition your network into multiple priority levels or classes of service. You can set the value of a standard QoS field for COS, IP precedence, and DSCP. You can also set the QoS field for internal labels (such as QoS groups) that can be used in subsequent actions. Marking QoS groups is used to identify the traffic type for queuing and scheduling traffic.

Policing

Policing causes traffic that exceeds the configured rate to be discarded or marked down to a higher drop precedence.

Single-rate policers monitor the specified committed information rate (CIR) of traffic. Dual-rate policers monitor both CIR and peak information rate (PIR) of traffic.

Queuing and Scheduling

The queuing and scheduling process allows you to control the queue usage and the bandwidth that is allocated to traffic classes. You can then achieve the desired trade-off between throughput and latency.

You can limit the size of the queues for a particular class of traffic by applying either static or dynamic limits.

You can apply weighted random early detection (WRED) to a class of traffic, which allows packets to be dropped based on the QoS group. The WRED algorithm allows you to perform proactive queue management to avoid traffic congestion.

ECN can be enabled along with WRED on a particular class of traffic to mark the congestion state instead of dropping the packets. ECN marking in the VXLAN tunnel is performed in the outer header, and at the Egress VTEP is copied to decapsulated frame.

Traffic Shaping

You can shape traffic by imposing a maximum data rate on a class of traffic so that excess packets are retained in a queue to smooth (constrain) the output rate. In addition, minimum bandwidth shaping can be configured to provide a minimum guaranteed bandwidth for a class of traffic.

Traffic shaping regulates and smooths out the packet flow by imposing a maximum traffic rate for each port's egress queue. Packets that exceed the threshold are placed in the queue and are transmitted later. Traffic shaping is similar to Traffic Policing, but the packets are not dropped. Because packets are buffered, traffic shaping minimizes packet loss (based on the queue length), which provides better traffic behavior for TCP traffic.

By using traffic shaping, you can control the following:

- Access to available bandwidth.
- Ensure that traffic conforms to the policies established for it.
- Regulate the flow of traffic to avoid congestion that can occur when the egress traffic exceeds the access speed of its remote, target interface.

For example, you can control access to the bandwidth when the policy dictates that the rate of a given interface must not, on average, exceed a certain rate. Despite the access rate exceeding the speed.

Network QoS

The network QoS policy defines the characteristics of each CoS value, which are applicable network wide across switches. With a network QoS policy, you can configure the following:

- Pause behavior—You can decide whether a CoS requires the lossless behavior which is provided by using a priority flow control (PFC) mechanism that prevents packet loss during congestion) or not. You can configure drop (frames with this CoS value can be dropped) and no drop (frames with this CoS value cannot be dropped). For the drop and no drop configuration, you must also enable PFC per port. For more information about PFC, see "Configuring Priority Flow Control".

Pause behavior can be achieved in the VXLAN tunnel for a specific queue-group.

VXLAN Priority Tunneling

In the VXLAN tunnel, DSCP values in the outer header are used to provide QoS transparency in end-to-end of the tunnel. The outer header DSCP value is derived from the DSCP value with Layer 3 packets or the CoS value for Layer 2 frames. At the VXLAN tunnel egress point, the priority of the decapsulated traffic is chosen based on the mode. For more information, see Decapsulated Packet Priority Selection, on page 8.

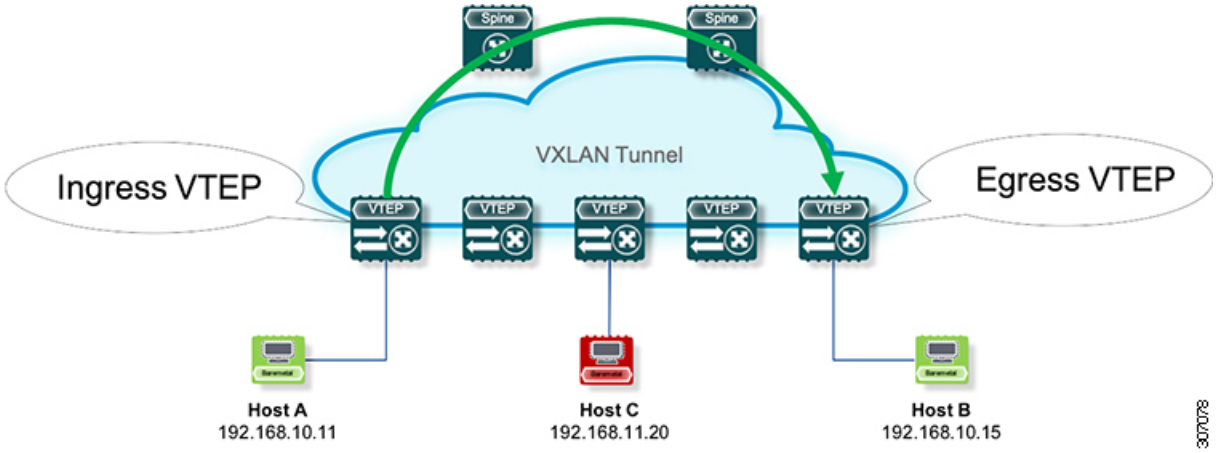
MQC CLI

All available QoS features for VXLAN QoS are managed from the modular QoS command-line interface (CLI). The Modular QoS CLI (MQC) allows you to define traffic classes (class maps), create and configure traffic policies (policy maps), and perform actions that are defined in the policy maps to interface (service policy).

VXLAN QoS Topology and Roles

This section describes the roles of network devices in implementing VXLAN QoS.

Figure 1: VXLAN Network



The network is bidirectional, but in the previous image, traffic is moving left to right.

In the VXLAN network, points of interest are ingress VTEPs where the original traffic is encapsulated in a VXLAN header. Spines are transporting hops that connect ingress and egress VTEPs. An egress VTEP is the point where VXLAN encapsulated traffic is decapsulated and egresses the VTEP as classical Ethernet traffic.

Note Ingress and egress VTEPs are the boundary between the VXLAN tunnel and the IP network.

This section contains the following topics:

Ingress VTEP and Encapsulation in the VXLAN Tunnel

At the ingress VTEP, the VTEP processes packets as follows:

Procedure

Step 1 Layer 2 or Layer 3 traffic enters the edge of the VXLAN network.

Step 2 The switch receives the traffic from the input interface and uses the 802.1p bits or the DSCP value to perform any classification, marking, and policing. It also derives the outer DSCP value in the VXLAN header. For classification of incoming IP packets, the input service policy can also use access control lists (ACLs).

Step 3 For each incoming packet, the switch performs a lookup of the IP address to determine the next hop.

Step 4 The packet is encapsulated in the VXLAN header. The encapsulated packet's VXLAN header is assigned a DSCP value that is based on QoS rules.

Step 5 The switch forwards the encapsulated packets to the appropriate output interface for processing.

Step 6 The encapsulated packets, marked by the DSCP value, are sent to the VXLAN tunnel output interface.

Transport Through the VXLAN Tunnel

In the transport through a VXLAN tunnel, the switch processes the VXLAN packets as follows:

Procedure

Step 1 The VXLAN encapsulated packets are received on an input interface of a transport switch. The switch uses the outer header to perform classification, marking, and policing.

Step 2 The switch performs a lookup on the IP address in the outer header to determine the next hop.

Step 3 The switch forwards the encapsulated packets to the appropriate output interface for processing.

Step 4 VXLAN sends encapsulated packets through the output interface.

Egress VTEP and Decapsulation of the VXLAN Tunnel

At the egress VTEP boundary of the VXLAN tunnel, the VTEP processes packets as follows:

Procedure

Step 1 Packets encapsulated in VXLAN are received at the NVE interface of an egress VTEP, where the switch uses the inner header DSCP value to perform classification, marking, and policing.

Step 2 The switch removes the VXLAN header from the packet, and does a lookup that is based on the decapsulated packet's headers.

Step 3 The switch forwards the decapsulated packets to the appropriate output interface for processing.

Step 4 Before the packet is sent out, a DSCP value is assigned to a Layer 3 packet based on the decapsulation priority or based on marking Layer 2 frames.

Step 5 The decapsulated packets are sent through the outgoing interface to the IP network.

Classification at the Ingress VTEP, Spine, and Egress VTEP

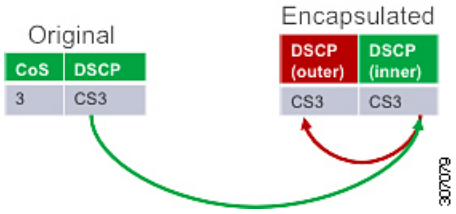
This section includes the following topics:

IP to VXLAN

At the ingress VTEP, the ingress point of the VXLAN tunnel, traffic is encapsulated in the VXLAN header. Traffic on an ingress VTEP is classified based on the priority in the original header. Classification can be performed by matching the CoS, DSCP, and IP precedence values or by matching traffic with the ACL based on the original frame data.

When traffic is encapsulated in the VXLAN, the Layer 3 packet's DSCP value is copied from the original header to the outer header of the VXLAN encapsulated packet. This behavior is illustrated in the following figure:

Figure 2: Copy of Priority from Layer-3 Packet to VXLAN Outer Header



For Layer 2 frames without the IP header, the DSCP value of the outer header is derived from the CoS-to-DSCP mapping present in the hardware illustrated in Default Settings for VXLAN QoS, on page 10. In this way, the original QoS attributes are preserved in the VXLAN tunnel. This behavior is illustrated in the following figure:

Figure 3: Copy of Priority from Layer-2 Frame to VXLAN Outer Header

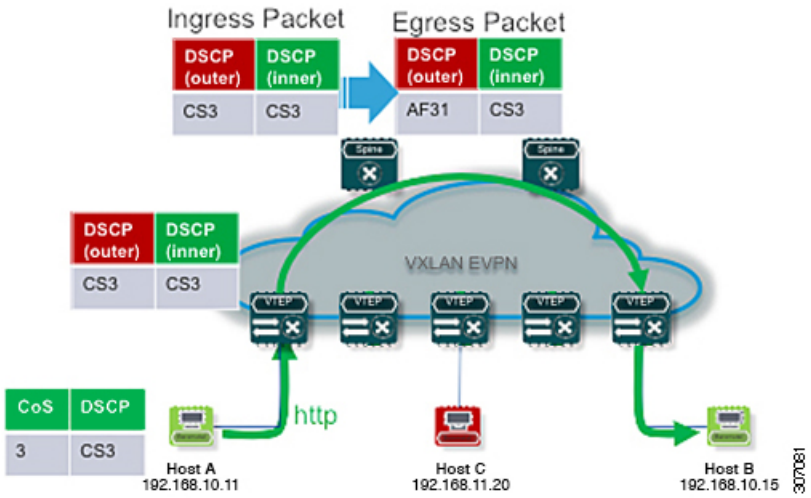
A Layer 2 frame, does not have a DSCP value present because the IP header is not present in the frame. After a Layer 2 frame is encapsulated, the original CoS value is not preserved in the VXLAN tunnel.

Inside the VXLAN Tunnel

Inside the VXLAN tunnel, traffic classification is based on the outer header DSCP value. Classification can be done matching the DCSP value or using ACLs for classification.

If VXLAN encapsulated traffic is crossing the trust boundary, marking can be changed in the packet to match QoS behavior in the tunnel. Marking can be performed inside of the VXLAN tunnel, where a new DSCP value is applied only on the outer header. The new DSCP value can influence different QoS behaviors inside the VXLAN tunnel. The original DSCP value is preserved in the inner header.

Figure 4: Marking Inside of the VXLAN Tunnel



VXLAN to IP

Classification at the egress VTEP is performed for traffic leaving the VXLAN tunnel. For classification at the egress VTEP, the inner header values are used. The inner DSCP value is used for priority-based classification. Classification can be performed using ACLs.

Classification is performed on the NVE interface for all VXLAN tunneled traffic.

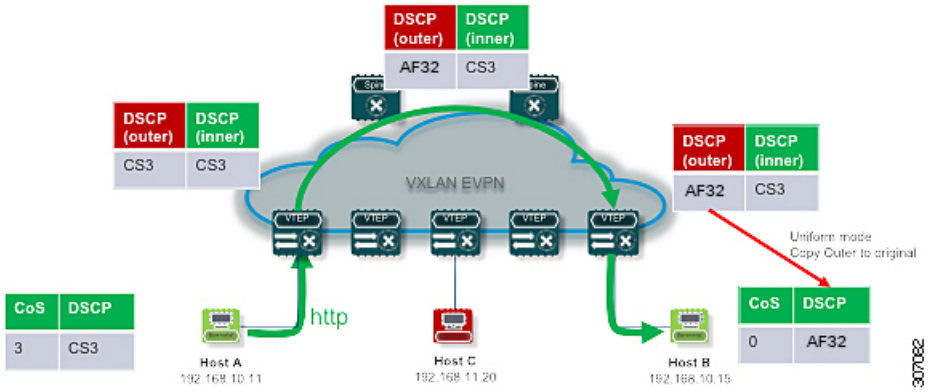
Marking and policing can be performed on the NVE interface for tunneled traffic. If marking is configured, newly marked values are present in the decapsulated packet. Because the original CoS value is not preserved in the encapsulated packet, marking can be performed for decapsulated packets for any devices that expect an 802.1p field for QoS in the rest of the network.

Decapsulated Packet Priority Selection

At the egress VTEP, the VXLAN header is removed from the packet and the decapsulated packet egresses the switch with the DSCP value. The switch assigns the DSCP value of the decapsulated packet based on two modes:

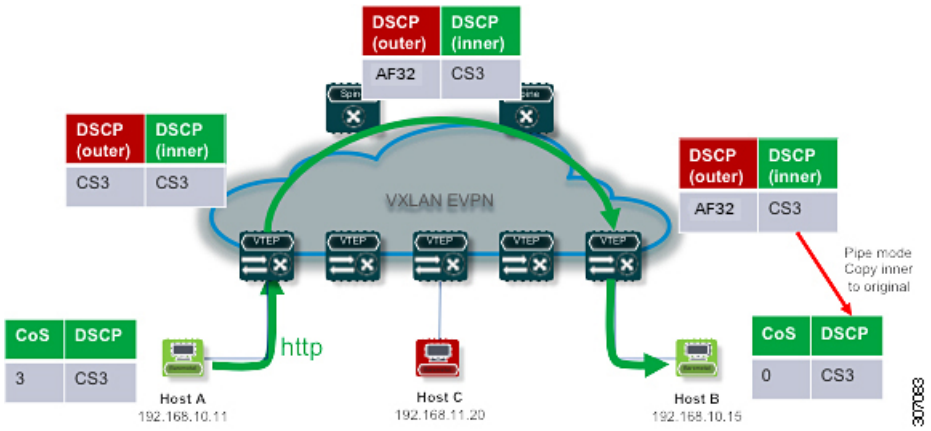
- Uniform mode – the DSCP value from the outer header of the VXLAN packet is copied to the decapsulated packet. Any change of the DSCP value in the VXLAN tunnel is preserved and present in the decapsulated packet. Uniform mode is the default mode of decapsulated packet priority selection.

Figure 5: Uniform Mode Outer DSCP Value is Copied to Decapsulated Packet DSCP Value for a Layer-3 Packet



- Pipe mode – the original DSCP value is preserved at the VXLAN tunnel end. At the egress VTEP, the system copies the inner DSCP value to the decapsulated packet DSCP value. In this way, the original DSCP value is preserved at the end of the VXLAN tunnel.

Figure 6: Pipe Mode Inner DSCP Value is Copied to Decapsulated Packet DSCP Value for Layer-3 Packet



Guidelines and Limitations for VXLAN QoS

Note QoS policy must be configured end-to-end for this feature to work as designed.

VXLAN QoS has the following configuration guidelines and limitations:

- Beginning with Cisco NX-OS Release 9.2(3), support is added for VXLAN QoS.
- VXLAN QoS is not supported on Cisco Nexus 9200 platform switches, Cisco Nexus 9300 platform switches with 9400, 9500, or 9600 line cards.
- This feature is supported in the EVPN fabric.

- The original IEEE 802.1Q header is not preserved in the VXLAN tunnel. The CoS value is not present in the inner header of the VXLAN encapsulated packet.
- Statistics (counters) are present for the NVE interface.
- Entering the **policy-map type qos** command in the output direction for egress policing is not supported in the ingress VTEP.
- If in a vPC, configure the change of the decapsulated packet priority selection on both peers.
- The service policy on an NVE interface can attach only in the input direction.
- If DSCP marking is present on the NVE interface, traffic to the BUD node preserves marking in the inner and outer headers. If a marking action is configured on the NVE interface, BUM traffic is marked with a new DSCP value on Cisco Nexus 9300-EX platform switches and the Cisco Nexus 9364C switch.
- A classification policy applied to an NVE interface, applies only on VXLAN encapsulated traffic. For all other traffic, the classification policy must be applied on the incoming interface.
- To mark the decapsulated packet with a CoS value, a marking policy must be attached to the NVE interface to mark the CoS value to packets where the VLAN header is present.
- In RX series line cards, the default mode is pipe for VXLAN decapsulation (inner packet DSCP not modified based on outer IP header DSCP value). This is a difference in behavior from other line cards types. If RX series line cards and other line cards are used in the same network, the **qos-mode pipe** command can be used in switches where non-RX line cards are present in order to have the same behavior. For details of the configuration command, see Configuring Type QoS on the Egress VTEP, on page 11.
- The following limitations apply to the VXLAN QoS policies when using a Border Gateway (BGW) Spine:
  - If QoS policies are needed for intra-site BUM traffic for VNI with multicast underlay, and that multicast underlay group is also owned by a VNI defined on the BGW Spine, then the QoS policy must be applied to the NVE interface. QoS policies applied to fabric interfaces will not modify these flows since the NVE interface acts as an incoming interface.
  - If QoS policies are needed for intra-site BUM traffic for VNI with multicast underlay, and that multicast group is not owned by a VNI defined on the BGW Spine, then the QoS policy must be applied to a fabric interface. QoS policies applied to the NVE interface will not modify these flows since the NVE is not considered an incoming interface.
  - If the NVE interface of the BGW Spine owns a multicast group used for BUM traffic within the local fabric, QoS policies cannot be applied to both the fabric interfaces and NVE interface to differentiate treatment of intra-site and inter-site flows for that multicast group.

Default Settings for VXLAN QoS

The following table lists the default CoS-to-DSCP mapping in the ingress VTEP for Layer 2 frames:

Table 2: Default CoS-to-DSCP Mapping

CoS of Original Layer 2 Frame	DSCP of Outer VXLAN Header
0	0

CoS of Original Layer 2 Frame	DSCP of Outer VXLAN Header
1	8
2	16
3	26
4	32
5	46
6	48
7	56

Configuring VXLAN QoS

Configuration of VXLAN QoS is done using the MQC model. The same configuration that is used for the QoS configuration applies to VXLAN QoS. For more information about configuring QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

VXLAN QoS introduces a new service-policy attachment point which is NVE – Network Virtual Interface. At the egress VTEP, the NVE interface is the point where traffic is decapsulated. To account for all VXLAN traffic, the service policy must be attached to an NVE interface.

The next section describes the configuration of the classification at the egress VTEP, and **service-policy type qos** attachment to an NVE interface.

Configuring Type QoS on the Egress VTEP

Configuration of VXLAN QoS is done by using the MQC model. The same configuration is used for QoS configuration for VXLAN QoS. For more information about configuring QoS, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 9.2(x).

VXLAN QoS introduces a new service-policy attachment point which is the Network Virtual Interface (NVE). At the egress VTEP, the NVE interface points where traffic is decapsulated. To account for all VXLAN traffic, the service policy must be attached to an NVE interface.

This procedure describes the configuration of classification at the egress VTEP, and **service-policy type qos** attachment to an NVE interface.

Procedure

	Command or Action	Purpose
Step 1	**configure terminal** **Example:** `switch# configure terminal`	Enters global configuration mode.
Step 2	**[no] class-map [type [qos]] | [match-all] | [match-any]** *class-map-name*	Creates or accesses the class map *class--map-name* and enters **class-map** mode.

	Command or Action	Purpose
	Example: `switch(config)# class-map type qos class1`	The *class--map-name* argument can contain alphabetic, hyphen, or underscore characters, and can be up to 40 characters. (match-any is the default when the **no** option is selected and multiple match statements are entered.)
Step 3	**[no] match [access-group | cos | dscp | precedence]** *{name | 0-7 | 0-63 | 0-7}* **Example:** `switch(config-cmap-qos)# match dscp 26`	Configures the traffic class by matching packets based on access-list, **cos** value, **dscp** values, or IP **precedence** value
Step 4	**[no] policy-map type qos** *policy-map-name* **Example:** `switch(config-cmap-qos)# policy-map type qos policy`	Creates or accesses the policy map that is named *policy-map-name* and then enters policy-map mode. The policy-map name can contain alphabetic, hyphen, or underscore characters, is case sensitive, and can be up to 40 characters.
Step 5	**[no] class** *class-name* **Example:** `switch(config-pmap-qos)# class class1`	Creates a reference to class-name and enters policy-map class configuration mode. The class is added to the end of the policy map unless insert-before is used to specify the class to insert before. Use the class-default keyword to select all traffic that is not currently matched by classes in the policy map.
Step 6	**[no] set qos-group** *qos-group-value* **Example:** `switch(config-pmap-c-qos)# set qos-group 1`	Sets the QoS group value to *qos-group-value*. The value can range from 1 through 126. The **qos-group** is referenced in type queuing and type network-qos as matching criteria.
Step 7	**exit** **Example:** `switch(config-pmap-c-qos)# exit`	Exits class-map mode.
Step 8	**[no] interface nve** *nve-interface-number* **Example:** `switch(config)# interface nve 1`	Enters interface mode to configure the NVE interface.
Step 9	**[no] service-policy type qos input** *policy-map-name* **Example:** `switch(config-if-nve)# service-policy type qos input policy`	Adds a service-policy *policy-map-name* to the interface in the input direction. You can attach only one input policy to an NVE interface.
Step 10	(Optional) **[no] qos-mode [pipe]** **Example:**	Selecting decapsulated packet priority selection and using pipe mode. Entering the **no** form of

	Command or Action	Purpose
	`switch(config-if-nve)# qos-mode pipe`	this command negates pipe mode and defaults to uniform mode.

Verifying the VXLAN QoS Configuration

Table 3: VXLAN QoS Verification Commands

Command	Purpose
show class map	Displays information about all configured class maps.
show policy-map	Displays information about all configured policy maps.
show running ipqos	Displays configured QoS configuration on the switch.

VXLAN QoS Configuration Examples

Ingress VTEP Classification and Marking

This example shows how to configure the **class-map type qos** command for classification matching traffic with an ACL. Enter the **policy-map type qos** command to put traffic in qos-group 1 and set the DSCP value. Enter the **service-policy type qos** command to attach to the ingress interface in the input direction to classify traffic matching the ACL.

```
access-list ACL_QOS_DSCP_CS3 permit ip any any eq 80

class-map type qos CM_QOS_DSCP_CS3
 match access-group name ACL_QOS_DSCP_CS3

policy-map type qos PM_QOS_MARKING
 class CM_QOS_DSCP_CS3
 set qos-group 1
 set dscp 24

interface ethernet1/1
 service-policy type qos input PM_QOS_MARKING
```

Transit Switch – Spine Classification

This example shows how to configure the **class-map type qos** command for classification matching DSCP 24 set on the ingress VTEP. Enter the **policy-map type qos** command to put traffic in qos-group 1. Enter the **service-policy type qos** command to attach to the ingress interface in the input direction to classify traffic matching criteria.

```
class-map type qos CM_QOS_DSCP_CS3
 match dscp 24

policy-map type qos PM_QOS_CLASS
 class CM_QOS_DSCP_CS3
```

```
 set qos-group 1

interface Ethernet 1/1
 service-policy type qos input PM_QOS_CLASS
```

Egress VTEP Classification and Marking

This example shows how to configure the **class-map type qos** command for classification matching traffic by DSCP value. Enter the **policy-map type qos** to place traffic in qos-group 1 and mark CoS value in outgoing frames. The **service-policy type qos** command is applied to the NVE interface in the input direction to classify traffic coming out of the VXLAN tunnel.

```
class-map type qos CM_QOS_DSCP_CS3
 match dscp 24

policy-map type qos PM_QOS_MARKING
 class CM_QOS_DSCP_CS3
 set qos-group 1
 set cos 3

interface nve 1
 service-policy type qos input PM_QOS_MARKING
```

Queuing

This example shows how to configure the **policy-map type queueing** command for traffic in qos-group 1. Assigning 50% of the available bandwidth to q1 mapped to qos-group 1 and attaching policy in the output direction to all ports using the **system qos** command.

```
policy-map type queuing PM_QUEUING
class type queuing c-out-8q-q7
      priority level 1
    class type queuing c-out-8q-q6
      bandwidth remaining percent 0
    class type queuing c-out-8q-q5
      bandwidth remaining percent 0
    class type queuing c-out-8q-q4
      bandwidth remaining percent 0
    class type queuing c-out-8q-q3
      bandwidth remaining percent 0
    class type queuing c-out-8q-q2
      bandwidth remaining percent 0
    class type queuing c-out-8q-q1
      bandwidth remaining percent 50
    class type queuing c-out-8q-q-default
      bandwidth remaining percent 50

system qos
 service-policy type queueing output PM_QUEUING
```