End-to-End QoS Implementation and Operation with Nexus

Nemanja Kamenica
Technical Marketing Engineer
BRKDCN-3223
Cisco Webex App

Questions?
Use Cisco Webex App to chat with the speaker after the session

How
1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 17, 2022.

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-3223
Abstract

• This session will provide a deep understanding of the End-to-End Quality of Service (QoS) Implementation and Operation in a Nexus Data Center network. We will cover the basics of QoS and Queuing and then take a look at the related best practices from a Nexus 9000 perspective. A strong focus will be on the platform architectures as they relate to QoS and Queuing. The session will close with a step-by-step case study of QoS deployment in a Nexus 9000 data center environment.
Session Objectives

• Provide a refresh of QoS Basics
• Understand QOS implementation on Nexus Operating System
• Provide a detailed understanding of QoS on Nexus Nexus 9000 platforms
• Learn how to configure QoS on Nexus 9000 devices through real-world configuration examples
Session Non-Objectives

- Data Centre QoS Methodology
- Nexus hardware architecture deep-dive
- Application Centric Infrastructure (ACI) QOS
Agenda

• Introduction
• QoS Basics
• QoS Implementation on Nexus
• Nexus 9000 Cloud Scale QoS
• Real World Configuration Examples
• Conclusion
Congestion Happens Everyday!
Why QoS in the Data Centre?

Assign Color to Traffic

Manage Congestion

Maximize Throughput

Maximize Throughput and Manage Congestion!
Can Traffic Control help ... ... or confuse

... or hurt
Agenda

• Introduction
• QoS Basics
• QoS Implementation on Nexus
• Nexus 9000 Cloud Scale QoS
• Real World Configuration Examples
• Conclusion
The QoS Toolset

Classification
- Identify and split traffic into different classes

Policing
- Discard misbehaving traffic to maintain network integrity

Marking
- Mark traffic according to behavior and business policies

Queuing and Scheduling
- Prioritize, protect and isolate traffic based on markings

Shaping
- Control bursts and conform traffic
Classification and Marking

Two sides of a coin

- Classification – Identify and separate traffic in classes
  - Identify traffic
    - ACLs
    - CoS
    - DSCP
    - IP PREC
  - Marking – Mark traffic with QoS priority value
- Marking Traffic
  - With new priority value (i.e. CoS or DSCP)
  - Changing Like to Like (i.e. CoS to CoS)
  - Like to Unlike (i.e. DSCP to CoS)

Unclassified:
All traffic treated the same

 Classified:
Traffic treated per class

VoIP/Lync
Email/HTTP
SAP
Backup

VoIP
Email
HTTP
vmotion
mgmt
lync
Backup

Unclassified:
All traffic treated the same
Policing – Limit Misbehaving Traffic

- Policing – Protecting other classes by dropping traffic in misbehaving class
- Single rate Two Color Policer
  - Conform Action (permit)
  - Exceed Action (drop)
- Two rate Three Color Policer
  - Conform Action (permit)
  - Exceed Action (markdown)
  - Violate Action (drop)
Buffering – Why do we need it?

- Buffering – Storing data packets in memory
- Many to One Conversations
  - Client to Server
  - Server to Storage
  - Aggregation Points
- Speed Mismatch
  - Client to WAN to Server
Queueing

- Traffic in buffer is divided logically in the queues
- Queueing provide dedicated buffer for packets of different priority
- Traffic separation allows multiple traffic classes to be mapped to same or different queue
- Traffic in a queue can be treated differently from other queues
Scheduling

- Scheduling – defines order of transmission of traffic out the queues
- Different types of queue are served differently
  - Strict Priority Queue – always serviced first
  - Normal Queues – served only after priority queue is empty
- Normal queues can have different algorithms
Common Scheduling Algorithms

- **Round Robin (RR)**
  - Simple and *Easy to implement*
  - Starvation-free

- **Weighted Round Robin (WRR)**
  - Serves $n$ packets per non-empty queue
  - Assumes a *mean packet size*

- **Deficit Weighted Round Robin**
  - Variable sized packets
  - Uses a deficit counter

- **Shaped Round Robin**
  - More *even distributed ordering*
  - Weighted interleaving of flows
Shaping

- Shaping – Smooth out traffic peaks, microburst, with preserving all traffic
- Usually in egress direction to limit traffic toward ISP
Congestion Avoidance Tools

- Tail Drop (TD)
  - Drop packets at tail of the queue
  - Single threshold per queue
Congestion Avoidance Tools

• Tail Drop (TD)
  • Drop packets at tail of the queue
  • Single threshold per queue
Congestion Avoidance Tools

- Tail Drop (TD)
  - Drop packets at **tail of the queue**
  - **Single threshold** per queue

- Weighted Random Early Drop (WRED)
  - One or more thresholds per queue
  - Threshold associated with priority
  - Buffer usage below threshold no affect
Congestion Avoidance Tools

• Tail Drop (TD)
  • Drop packets at tail of the queue
  • Single threshold per queue

• Weighted Random Early Drop (WRED)
  • One or more thresholds per queue
  • Threshold associated with priority
  • Buffer usage below threshold no affect
  • Buffer usage over min threshold = random drops
Congestion Avoidance Tools

- **Tail Drop (TD)**
  - Drop packets at **tail of the queue**
  - **Single threshold** per queue

- **Weighted Random Early Drop (WRED)**
  - One or more thresholds per queue
  - Threshold associated with priority
  - Buffer usage below threshold no affect
  - Buffer usage over **min threshold** = random drops
  - Buffer usage over **max threshold** = all traffic drop
Agenda

- Introduction
- QoS Basics
- QoS Implementation on Nexus
- Nexus 9000 Cloud Scale QoS
- Real World Configuration Examples
- Conclusion
Nexus uses Modular QOS CLI (MQC)

3 Block Construct

Class-Map

What Traffic do we care about?
- DSCP
- CoS
- IP Precedence
- ACLs

Policy-Map

What actions do I take on the classes?
- Policing
- Marking
- Scheduling
- Queueing

Service-Policy

Where do I apply this policy?
- System Wide
- VLAN
- Interface
- Port-channels
Three Different Types

**Class-map**
- Type QoS
  - CoS
  - DSCP
  - PREC
  - ACLs
- Type Queuing
  - qos-group

**Policy-map**
- Type QoS
  - Classification
  - Marking
  - Policing
- Type Queuing
  - Scheduling
  - Shaping

**Service-policy**
- Type QoS
  - Interface
  - Port-channel
  - VLAN
- Type Queuing
  - Interface
  - System-qos

**Type Network-QoS**
- qos-group
- MTU
- Non-drop
- System-qos
Type Network-QoS Policy

- Define global queuing and scheduling parameters for all interfaces in switch
  - Identify drop/no-drop classes, MTU and WRED/TD, etc.
- One Network-QoS policy per system, applies to all ports
- Assumption is Network-QoS policy defined/applied consistently network-wide
System Based Policy Attachment

- System based QoS Policy gets globally applied to a system (to all interfaces)
- System based QoS Policy is configured in System QoS
- Type Queueing can be attached to the system level
- Type Network-QoS is mandatory to be attached to the system level

```
Nexus(config)# system qos
Nexus(config-sys-qos)# service-policy type network-qos myPolicy
```
VLAN Based QoS Policy Attachment

- VLAN based QoS Policy is configured in VLAN Database
- No SVI (aka L3 VLAN Interface) required

Nexus(config)# vlan configuration <vlan-id>
Nexus(config-vlan)# service-policy type qos input myPolicy
Interface based Type QOS Policy attachment

- Interface based type qos Policy takes precedence over VLAN
- Can also be attached to port-channel and applies to all member-ports

Nexus(config)# interface ethernet 1/1
Nexus(config-if)# service-policy type qos input myPolicy
Interface based Type Queuing Policy attachment

- Type Queuing has to be attached to a physical interface or system level

- Queuing Policy can be attached to port-channel and all member ports

```
Nexus(config)# interface ethernet 1/1
Nexus(config-if)# service-policy type queueing output myPolicy
```
Buffer types – Head of Line Blocking

What is the Problem?

Ingress Module

Egress Module

Ingress Module

Egress Module

Ingress Module

Egress Module

backpressure for all Traffic

buffer

congestion
Virtual Output Queuing

Ingress Module

Egress Module

Crossbar Fabric

VoQ Buffer

buffer

Ingress Module

Egress Module

Ingress Module

Egress Module

congestion
Output Queuing
Buffering on Nexus Models

**VOQ**
Virtual Output Queue

**N ports**
Input buffer for every egress port

**NxN buffer size**

**Output Queue Buffer**

**N ports**
Shared buffer for N egress ports

**N buffer size**

- **Nexus 9500-R**
- **Nexus 3600-R**

- **Nexus 9000**
- **Nexus 3000**
4 Class Queuing Model

- Matches most Service-Provider offerings
- **Ready for No-Drop** traffic like FCoE
- One Class left to place traffic above or below Best-Effort traffic priority
  - Special Application which is drop sensitive (above Best-Effort - Critical)
  - Non-Critical Bandwidth intensive application (below Best-Effort - Scavenger)

<table>
<thead>
<tr>
<th>Class</th>
<th>CoS</th>
<th>Queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority</td>
<td>5-7</td>
<td>PQ</td>
</tr>
<tr>
<td>No-Drop</td>
<td>3</td>
<td>Q2</td>
</tr>
<tr>
<td>Better or Worse than Best-Effort</td>
<td>1,2,4</td>
<td>Q1</td>
</tr>
<tr>
<td>Best-Effort</td>
<td>0</td>
<td>Default-Q</td>
</tr>
</tbody>
</table>

© 2022 Cisco and/or its affiliates. All rights reserved. Cisco Public
8 Class Queuing Model

- Matches often a Campus QoS concept
- DSCP to CoS derivation does NOT apply anymore
  - (Topmost 3-Bit mapping from DSCP to CoS)
- No-Drop still with CoS3
- DSCP 24-30 are usable for IP storage traffic (RoCEv2)

<table>
<thead>
<tr>
<th>Class</th>
<th>DSCP</th>
<th>Queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority</td>
<td>CS6 (CS7)</td>
<td>PQ</td>
</tr>
<tr>
<td>Platinum</td>
<td>EF</td>
<td></td>
</tr>
<tr>
<td>Gold</td>
<td>AF41</td>
<td>Q7</td>
</tr>
<tr>
<td>Silver</td>
<td>CS4</td>
<td>Q6</td>
</tr>
<tr>
<td>No-Drop</td>
<td>CoS3</td>
<td>Q5</td>
</tr>
<tr>
<td>Bronze</td>
<td>AF21</td>
<td>Q4</td>
</tr>
<tr>
<td>Management</td>
<td>CS2</td>
<td>Q3</td>
</tr>
<tr>
<td>Scavenger</td>
<td>AF11</td>
<td>Q2</td>
</tr>
<tr>
<td>Bulk Data</td>
<td>CS1</td>
<td>Q1</td>
</tr>
<tr>
<td>Best-Effort</td>
<td>0</td>
<td>Default-Q</td>
</tr>
</tbody>
</table>
To Trust or Not To Trust?

- Data Centre architecture provides a new set of trust boundaries
- Virtual Switch extends the trust boundary into the Hypervisor
- Nexus Switches always trust CoS and DSCP
Data Center
QoS Capabilities
Data Centre Converged Infrastructure

- Enable, sensitive to drop, storage traffic to use Ethernet
- Simplification of the infrastructure by using Ethernet for data and storage traffic
- Data Center QoS capabilities, enabling new transport:
  - PFC – Priority Flow Control
  - ETS – Enhanced Transmission Selection
  - DCBX – Data Center Bridging Exchange
  - ECN – Explicit Congestion Notification
Priority Flow Control

Flow Control Mechanism – 802.1Qbb

• A.k.a "Lossless Ethernet"

• PFC enables Flow Control on a Per-Priority basis

• Therefore, we have the ability to have lossless and lossy priorities at the same time on the same wire

• Allows traffic to operate over a lossless priority independent of other priorities

• Other traffic assigned to other priority will continue to transmit and rely on upper layer protocols for retransmission
Enhanced Transmission Selection (ETS) Bandwidth Management – 802.1Qaz

- Prevents a single traffic class of “hogging” all the bandwidth and starving other classes
- When a given load doesn’t fully utilize its allocated bandwidth, it is available to other classes
- Helps accommodate for classes of a “bursty” nature

---

**Offered Traffic**

- **t1**: 3G/s
- **t2**: 3G/s
- **t3**: 2G/s

**10 GE Link Realized Traffic Utilization**

- **t1**: 3G/s
- **t2**: HPC Traffic 3G/s
- **t3**: 2G/s

- **t1**: 3G/s
- **t2**: Storage Traffic 3G/s
- **t3**: 3G/s

- **t1**: 3G/s
- **t2**: LAN Traffic 4G/s
- **t3**: 5G/s
Data Center Bridging Exchange Protocol

DCBX Overview - 802.1Qaz

- Negotiates Ethernet capability’s PFC, ETS, CoS values between DCB capable peer devices
- Simplifies Management allows for configuration and distribution of parameters from one node to another
- DCBX is LLDP with new TLV fields

Explicit Congestion Notification (ECN)

- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on a IP network.
- In case of congestion, ECN gets transmitting device to reduce transmission rate until congestion clears, without pausing traffic.
- ECN uses 2 LSB of Type of Service field in IP header.

<table>
<thead>
<tr>
<th>ECN</th>
<th>ECN Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x00</td>
<td>Non ECN Capable</td>
</tr>
<tr>
<td>0x10</td>
<td>ECN Capable Transport (0)</td>
</tr>
<tr>
<td>0x01</td>
<td>ECN Capable Transport (1)</td>
</tr>
<tr>
<td>0x11</td>
<td>Congestion Encountered</td>
</tr>
</tbody>
</table>
IP Storage Transports in Data Center

• Converged storage Protocols:
  • Requirement for FCoE and RoCEv1:
    • PFC
    • ETS
  • Requirement for RoCEv2
    • PFC
    • ETS
    • ECN
Overlay QOS
Overlay QoS

MPLS network

- Mapping between IP priorities to EXP on PE router
- Classification is done biased on COS, DSCP, IP precedence or ACL
- DiffServ Tunneling mode provides different QOS behavior in provider network
  - Uniform mode delivers overlay priority
  - Pipe mode extends underlay priority

<table>
<thead>
<tr>
<th>EXP</th>
<th>COS</th>
<th>DSCP</th>
<th>IP pres</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>56</td>
<td>7</td>
</tr>
</tbody>
</table>
Overlay QOS
MPLS – Default Mode

CE1 — PE1 — P1 — P2 — PE2 — CE2

MPLS Core

DSCP 24 EXP 3
DSCP 24 EXP 3
DSCP 24 EXP 3
DSCP 24 EXP 3
DSCP 24 EXP 3
DSCP 24
Overlay QOS
MPLS – Uniform Mode

- CE1
- PE1
- P1
- P2
- PE2
- CE2

DSCP 24
EXP 3
DSCP 24
EXP 3
DSCP 24
EXP 2
DSCP 16

EXP 3
DSCP 24
EXP 2
DSCP 24
EXP 3
DSCP 24
EXP 2
DSCP 24

MPLS Core
Overlay QOS

MPLS – Pipe Mode

CE1  PE1  P1  P2  PE2  CE2

MPLS Core

DSCP 24  EXP 3  DSCP 24  EXP 5  DSCP 24  EXP 5

DSCP 24  EXP 3  DSCP 24  EXP 5  DSCP 24  EXP 3

DSCP 24  EXP 3  DSCP 24  EXP 5  DSCP 24  EXP 3
Overlay QoS

VXLAN EVPN - VXLAN Encapsulation

- Ingress L3 packet, original priority is mapped to outer header priority
- Ingress L2 frame, COS value will be mapped to outer priority
- VLAN header is not preserved in VXLAN tunnel

<table>
<thead>
<tr>
<th>COS</th>
<th>DSCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>56</td>
</tr>
</tbody>
</table>

Original L3 Packet

Original L2 Frame

VXLAN Encap. Packet

VXLAN Encap. Packet
Overlay QoS

VXLAN EVPN – VXLAN Decapsulation

- DSCP value is derived based on a priority mode for L3 traffic:
  - Uniform mode: delivers overlay priority copying outer header to decapsulated frame
  - Pipe mode: extends original priority copying inner header to decapsulated frame
- Marking can be configure on the egress VTEP mark decapsulated traffic with priority (COS, DSCP)
Overlay QoS

VXLAN – Uniform Mode

- DSCP 26
- COS 3
- Host A: 192.168.10.11
- DSCP 26
- VTEP
- Spine

- DSCP 26
- COS 0
- Host B: 192.168.10.15
- DSCP 26
- VTEP
- Spine

- DSCP 26
- COS 3
- Host C: 192.168.11.20
- DSCP 26
- VTEP
- Spine

VXLAN EVPN

http
Overlay QoS
VXLAN – Pipe Mode

- DSCP 26 to VTEP
- COS 3
- http

Host A: 192.168.10.11
Host C: 192.168.11.20
Host B: 192.168.10.15
Agenda

• Introduction
• QoS Basics
• QoS Implementation on Nexus
• Nexus 9000 Cloud Scale QoS
• Real World Configuration Examples
• Conclusion
Nexus 9000 Overview

- Modular and Fixed chassis
- Optimized for high density 10G/25G/40G/50G/100G/200G/400G
- Standalone and ACI Mode
- Cisco Silicon – Cloud Scale
  - Advanced QoS capabilities
Nexus 9000 - Cloud Scale

**LS25600GX2A**
- 25.6T chip – 8 slices of 8 x 400G each
- 9300-GX2A, 9408 TORs

**LS12800GX2B**
- 12.8T chip – 4 slices of 8 x 400G each
- 9300-GX2B TORs

**LS6400GX**
- 6.4T chip – 4 slices of 4 x 400G each
- 9300-GX TORs
Nexus 9000 - Cloud Scale

**LS3600FX2**
- 3.6T chip – 2 slices of 18 x 100G with MACSEC
- 9300-FX2 TORs

**LS1800FX/FX3**
- 1.8T chip – 1 slice of 18 x 100G with MACSEC
- X9700-FX modular line cards; 9300-FX/FX3 TORs
What Is a “Slice”?  

- Self-contained forwarding complex controlling subset of ports on single ASIC
- Separated into Ingress and Egress functions
- Ingress of each slice connected to egress of all slices
- Slice interconnect provides non-blocking any-to-any interconnection between slices
Cisco Nexus 9000 – Cloud Scale QoS Features

- Classification based on:
  - ACL
  - DSCP, CoS, and IP Precedence
- Marking traffic with:
  - DSCP
  - CoS
  - IP Precedence
- Policing:
  - 1R2C and 2R3C
  - Ingress and Egress
- Buffering/Queueing:
  - Shared egress buffer; 8 Egress Queues
- Scheduling:
  - Strict Priority Queuing and DWRR
- Shaping:
  - Egress per queue shaper
- Congestion Avoidance:
  - Tail Drop
  - WRED with ECN
Buffering

- Cloud Scale platforms implement shared-memory egress buffered architecture
- Each ASIC slice has dedicated buffer – only ports on that slice can use that buffer
- Dynamic Buffer Protection adjusts max thresholds based on class and buffer occupancy
- Intelligent buffer options maximize buffer efficiency
Queuing and Scheduling

- 8 qos-groups per output port
- Egress queuing policy defines priority and weights
- Dedicated classes for CPU traffic and SPAN traffic
Intelligent Buffering
Innovative Buffer Management for Cloud Scale switches

• **Dynamic Buffer Protection (DBP)** – Controls buffer allocation for congested queues in shared-memory architecture

• **Approximate Fair Drop (AFD)** – Maintains buffer headroom per queue to maximize burst absorption

• **Dynamic Packet Prioritization (DPP)** – Prioritizes short-lived flows to expedite flow setup and completion

Miercom Report: Speeding Applications in Data Centre Networks
Dynamic Buffer Protection (DBP)

- Prevents any output queue from consuming more than its fair share of buffer in shared-memory architecture
- Defines dynamic max threshold for each queue
  - If queue length exceeds threshold, packet is discarded
  - Otherwise packet is admitted to queue and scheduled for transmission
- Threshold calculated by multiplying free memory by configurable, per-queue Alpha (α) value (weight)
  - Alpha controls how aggressively DBP maintains free buffer pages during congestion events
Alpha Parameter Examples

Default Alpha on Cloud Scale switches

<table>
<thead>
<tr>
<th>Buffer per queue (MB)</th>
<th>Free buffer (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha (α) = 0.5</td>
<td></td>
</tr>
<tr>
<td>Buffer per queue == ½ free buffer</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Buffer per queue (MB)</th>
<th>Free buffer (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha (α) = 1</td>
<td></td>
</tr>
<tr>
<td>Buffer per queue == free buffer</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Buffer per queue (MB)</th>
<th>Free buffer (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha (α) = 14</td>
<td></td>
</tr>
<tr>
<td>Buffer per queue == 14 x free buffer</td>
<td></td>
</tr>
</tbody>
</table>
Buffering – Ideal versus Reality

**Ideal buffer state**
- Buffer available for burst absorption
- Buffer consumed by sustained-bandwidth TCP flows

**Actual buffer state**
- Buffer available for burst absorption
- Buffer consumed by sustained-bandwidth TCP flows
- Sustained-bandwidth TCP flows consume all available buffer before backing off

Sustained-bandwidth TCP flows back off before all buffer consumed
Approximate Fair Drop (AFD)

Maintain throughput while minimizing buffer consumption by elephant flows – *keep buffer state as close to the ideal as possible*

- Distinguish elephant flows from other flows
  - Higher-bandwidth elephants – higher AFD drop probability
  - Lower-bandwidth elephants – lower AFD drop probability

- Non-elephants – no AFD

Queue admission

- Ideal depth exceeded?
  - Y
  - AFD set?
    - N
    - Y
    - Buffer

- N
Dynamic Packet Prioritization (DPP)

- Prioritize initial packets of new / short-lived flows
- Up to first 1023 packets of each flow assigned to higher-priority qos-group

<= 1023 packets
  - Drive new, higher priority
  - Track per-flow packet count

> 1023 packets
  - Maintain original priority

Identify unique flows

SP queue
Q-default

© 2022 Cisco and/or its affiliates. All rights reserved. Cisco Public
Configuration – Class-Map Type QoS

• Class-map type qos used to classify traffic based on
  • Access List
  • Priority (CoS, DSCP, IP Precedence)

• Match by singe criteria or match all criteria under class-map:
  • match-all: Traffic need to match all criteria under class map
  • match-any: Traffic needs to match any criteria under class map

```plaintext
class-map type qos match-all/match-any class-q1
match access-group HTTP
match cos 1
match dscp 8
```
Configuration – Policy-Map Type QoS

• Policy-map type qos used to take action on class-map traffic
  • Set new priorities (COS, DSCP, IP Precedence)
  • Set a policer

• The policy-map sets qos-group

```plaintext
policy-map type qos Classification-Marking
  class class-q1
    set cos 1
    police cir 1000 mbps bc 200 ms conform transmit violate drop
    set qos-group 1
```
Qos-Group

• QoS group is used to reference classification for all the types class-maps
  • Class-map type queueing and type network qos have class-maps referencing qos-groups
  • Class-maps are present in system by default, no user interaction required

• Default class-map type queueing for Q1:
  ```
  class-map type queuing match-any c-out-8q-q1
  match qos-group 1
  ```

• Default class-map type network-qos for Q1
  ```
  class-map type network-qos c-8q-nq1
description Default class on qos-group 1
  match qos-group 1
  ```
Configuration – Policy-Map Type Queuing

- Policy-map type queueing define queuing and scheduling options
- Define queue limit – change alpha value
- Define scheduling options, strict priority and weight for DWRR queues
- Default Queueing policy cannot be changed
- User needs to define custom policy
- Shaping defined per queue in queueing policy

```plaintext
policy-map type queuing custom-8q-out-policy
  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
```
Configuration – Policy-Map Type Network-QoS

• Policy-map type network-qos define:
  • Non-drop queue
  • End to end queueing policy (8 queue or 4 queue)
• Default Network-QoS policy cannot be changed
• User needs to define custom policy

```cisco
policy-map type network-qos custom-8q-nq-policy
  class type network-qos c-8q-nq7
    mtu 1500
  class type network-qos c-8q-nq6
    mtu 1500
  class type network-qos c-8q-nq5
    mtu 1500
  class type network-qos c-8q-nq4
    mtu 1500
  class type network-qos c-8q-nq3
    mtu 1500
  class type network-qos c-8q-nq2
    mtu 1500
  class type network-qos c-8q-nq1
    mtu 1500
  class type network-qos c-8q-nq-defualt
    mtu 1500
```
Configuration – Putting it all together

class-map type qos match-any class-q1
match access-group HTTP

policy-map type qos Classification-Marking
  class class-q1
    set cos 1
    set qos-group 1

policy-map type queuing custom-8q-out-policy
  [snip]
  class type queuing c-out-8q-q1
    bandwidth remaining percent 50
  class type queuing c-out-8q-q-default
    bandwidth remaining percent 50

policy-map type network-qos custom-8q-nq-policy
  [snip]
  class type network-qos c-8q-nq1
    mtu 1500
  class type network-qos c-8q-nq-default
    mtu 1500

interface Ethernet 1/1
  service-policy type qos input Classification-Marking

system qos
  service-policy type network-qos custom-8q-nq-policy
  service-policy type queuing output custom-8q-out-policy
Nexus 9000 QoS Golden Rules

- CoS and DSCP are TRUSTED by default
- Use QoS-Groups to tie policies together
- Nexus 9000 Cloud Scale – Egress Buffer
  - Queuing/scheduling policy attached in egress direction
Agenda

- Introduction
- QoS Basics
- QoS Implementation on Nexus
- Nexus 9000 Cloud Scale QoS
- Real World Configuration Examples
- Conclusion
What do we want to achieve?

Company XYZ’s Business Goals

• Make sure no disruption in network services
  • Put control traffic in priority queue

• Video/voice hosting also a business objective
  • Put voice traffic in priority queue
  • Dedicated bandwidth to video traffic

• Flexibility in moving applications across servers
  • Dedicated bandwidth to vmotion/mobility
  • Everything else best-effort
Translating to the language of QoS

<table>
<thead>
<tr>
<th>Application</th>
<th>CoS</th>
<th>DSCP</th>
<th>Queuing (Scheduling)</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Effort</td>
<td>0, 1</td>
<td>0, 8</td>
<td>BW remaining 40%</td>
<td>High Volume / Less Important</td>
</tr>
<tr>
<td>vMotion / Live Migration</td>
<td>2</td>
<td>N/A*</td>
<td>BW remaining 20%</td>
<td>Medium Volume / Important</td>
</tr>
<tr>
<td>Multimedia</td>
<td>3, 4</td>
<td>24, 32</td>
<td>BW remaining 30%</td>
<td>Medium Volume Very Important</td>
</tr>
<tr>
<td>Strict Priority</td>
<td>5</td>
<td>46</td>
<td>BW remaining 10%</td>
<td>Low Volume / Important / Delay Sensitive</td>
</tr>
<tr>
<td>Network Control</td>
<td>6,7</td>
<td>48, 56</td>
<td>Priority Queue</td>
<td>Low Volume / Very Important</td>
</tr>
</tbody>
</table>

* Layer 2 traffic without IP header
Topology

- Nexus 9500
- Nexus 9300
- X9700-EX/FX Line cards
- Host VPC
- Core
Classification, Marking and Trust

- Classification of traffic from server, mark with DSCP/COS
- Classification of traffic based on existing COS/DSCP (trust)
- Classification of traffic from core, mark COS/DSCP, or trust

Type: QoS

Nexus 9500

Core
## Marking Definition

<table>
<thead>
<tr>
<th>Application</th>
<th>CoS</th>
<th>DSCP</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Effort</td>
<td>0, 1</td>
<td>0, 8</td>
<td>High Volume / Less Important</td>
</tr>
<tr>
<td>vMotion / Live Migration</td>
<td>2</td>
<td>N/A*</td>
<td>Medium Volume / Important</td>
</tr>
<tr>
<td>Multimedia</td>
<td>3, 4</td>
<td>24, 32</td>
<td>Medium Volume / Very Important</td>
</tr>
<tr>
<td>Strict Priority</td>
<td>5</td>
<td>46</td>
<td>Low Volume / Important / Delay Sensitive</td>
</tr>
<tr>
<td>Network Control</td>
<td>6, 7</td>
<td>48, 56</td>
<td>Low Volume / Very important</td>
</tr>
</tbody>
</table>
Classification and Marking

Nexus 9300 Leaf (Host Interfaces)

```
ip access-list ACL_QOS_LOWPRIO
    10 permit ...
ip access-list ACL_QOS_VMOTION
    10 permit ...
ip access-list ACL_QOS_MULTIMEDIA
    10 permit ...
!
class-map type qos CM_QOS_LOWPRIO_COS1
    match access-group name ACL_QOS_LOWPRIO
!
class-map type qos CM_QOS_VMOTION_COS2
    match access-group name ACL_QOS_VMOTION
!
class-map type qos CM_QOS_MULTIMEDIA_COS4
    match access-group name ACL_QOS_MULTIMEDIA
!
class-map type qos CM_QOS_STRICTPRIO_COS5
    match cos 5

policy-map type qos PM_QOS_MARK_COS_IN
    class CM_QOS_STRICTPRIO_COS5
        set qos-group 5
        set cos 5
        set dscp 46
    class CM_QOS_MULTIMEDIA_COS4
        set qos-group 4
        set cos 4
        set dscp 32
    class CM_QOS_VMOTION_COS2
        set qos-group 2
        set cos 2
    class CM_QOS_LOWPRIO_COS1
        set qos-group 1
        set cos 1
        set dscp 8
!
interface Ethernet 1/1
    service-policy type qos input PM_QOS_MARK_COS_IN
!
vlan configuration 100
    service-policy input PM_QOS_MARK_COS_IN
```
Classification and Marking
Nexus 9300 Leaf (Uplink Interfaces) and Nexus 9500 (Spine Interfaces)

```
class-map type qos match-any 
  CM_QOS_LOWPRIO_COS1
  match dscp 8
!
class-map type qos match-any 
  CM_QOS_VMOTION_COS2
  match dscp 16
!
class-map type qos match-any 
  CM_QOS_MULTIMEDIA_COS4
  match dscp 32
!
class-map type qos match-any 
  CM_QOS_STRICTPRIO_COS5
  match dscp 46
!

policy-map type qos PM_QOS_MARK_COS_IN
  class CM_QOS_STRICTPRIO_COS5
    set qos-group 5
  class CM_QOS_MULTIMEDIA_COS4
    set qos-group 4
  class CM_QOS_VMOTION_COS2
    set qos-group 2
  class CM_QOS_LOWPRIO_COS1
    set qos-group 1
!
interface Ethernet 1/1
  service-policy type qos input PM_QOS_MARK_COS_IN
```
# Queueing and Scheduling

## Nexus 9300, and 9500

<table>
<thead>
<tr>
<th>Application</th>
<th>CoS</th>
<th>DSCP</th>
<th>Queuing (Scheduling)</th>
<th>Queue limit (Alpha)</th>
<th>Queue</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Effort</td>
<td>1</td>
<td>8</td>
<td>BW percent 30%</td>
<td>Default (9)</td>
<td>qos-group 1</td>
<td>High Volume / Less Important</td>
</tr>
<tr>
<td>vMotion / Live Migration</td>
<td>2,3</td>
<td>16</td>
<td>BW percent 20%</td>
<td>Default (9)</td>
<td>qos-group 2</td>
<td>Medium Volume / Important</td>
</tr>
<tr>
<td>Multimedia</td>
<td>4</td>
<td>24, 32</td>
<td>BW percent 30%</td>
<td>Default (9)</td>
<td>qos-group 4</td>
<td>Medium Volume Very Important</td>
</tr>
<tr>
<td>Strict Priority</td>
<td>5</td>
<td>46</td>
<td>BW percent 10%</td>
<td>Default (9)</td>
<td>qos-group 5</td>
<td>Low Volume / Important / Delay Sensitive</td>
</tr>
<tr>
<td>Network Control</td>
<td>6,7</td>
<td>48, 56</td>
<td>Priority Queue</td>
<td>Default (9)</td>
<td>priority</td>
<td>Low Volume / Very important</td>
</tr>
</tbody>
</table>
Queueing and Scheduling
Nexus 9300, and 9500

- Class-maps type queueing are predefined
- Class-maps referring to qos-groups

```
policy-map type queuing custom-8q-out-policy
  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 10
  class type queuing c-out-8q-q4
    bandwidth remaining percent 30
  class type queuing c-out-8q-q3
    bandwidth remaining percent 0
  class type queuing c-out-8q-q2
    bandwidth remaining percent 20
  class type queuing c-out-8q-q1
    bandwidth remaining percent 30
  class type queuing c-out-8q-q-default
    bandwidth remaining percent 10

  system qos
  service-policy type queuing output custom-8q-out-policy
```
Network-QoS

- Keep default Network-QoS:
  - Default 8 Queue model
  - No configuration for non-drop queue
Agenda

- Introduction
- QoS Basics
- QoS Implementation on Nexus
- Nexus 9000 Cloud Scale QoS
- Real World Configuration Examples
- Conclusion
Why QoS in the Data Centre?

Assign Colour to Traffic

Manage Congestion

Maximise Throughput
With some help of my friends

I would like to thank all the people, who started the QoS journey and contributed to it:

• Lukas Krattiger, Distinguished Technical Marketing Engineer
• Tim Stevenson, Distinguished Technical Marketing Engineer
• Matthias Wessendorf, Principal Technical Marketing Engineer
Technical Session Surveys

• Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!

• Attendees will also earn 100 points in the Cisco Live Game for every survey completed.

• These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.
Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. [www.cisco.com/go/certs](http://www.cisco.com/go/certs)

**Learn**
- **Cisco U.**
  IT learning hub that guides teams and learners toward their goals
- **Cisco Digital Learning**
  Subscription-based product, technology, and certification training
- **Cisco Modeling Labs**
  Network simulation platform for design, testing, and troubleshooting
- **Cisco Learning Network**
  Resource community portal for certifications and learning

**Train**
- **Cisco Training Bootcamps**
  Intensive team & individual automation and technology training programs
- **Cisco Learning Partner Program**
  Authorized training partners supporting Cisco technology and career certifications
- **Cisco Instructor-led and Virtual Instructor-led training**
  Accelerated curriculum of product, technology, and certification courses

**Certify**
- **Cisco Certifications and Specialist Certifications**
  Award-winning certification program empowers students and IT Professionals to advance their technical careers
- **Cisco Guided Study Groups**
  180-day certification prep program with learning and support
- **Cisco Continuing Education Program**
  Recertification training options for Cisco certified individuals

Here at the event? Visit us at The Learning and Certifications lounge at the World of Solutions

---

Pay for Learning with Cisco Learning Credits
(CLCs) are prepaid training vouchers redeemed directly with Cisco.
Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand
Thank you