
TOMORROW starts here.
cisco
Cisco live!

QoS Implementation and Operation in a Nexus 7000 environment

BRKDCT-3346

Lukas Krattiger, Consulting Systems Engineer

Lukk@cisco.com

Cisco live!

Abstract

This session will provide a deep understanding of the Quality of Service (QoS) Implementation and Operation in Nexus 7000 data center network environments.

We will cover the “Nuts and Bolts” on the configuration level and take a look at the related best practices from a Nexus 7000 perspective.

A strong focus on the latest additions of the M- and F-Series I/O modules (M2 & F2/F2e) and the hardware architecture on QoS and Queuing will be discussed.

In order to provide a full picture of a Nexus data center network environment, some Nexus 6000/5000/2000 specific architecture and implementation practices will be covered as well.

The session will close with a step-by-step walk through case study where the introduction of QoS in a Nexus 7000/5000/2000 data center environment is covered.

The case study focuses on the hardware and configuration requirements of the different components for a successful QoS data center network deployment.

The session is designed for Network Engineers and Architects involved in current and future Data Centre Designs, Implementation and Operation.

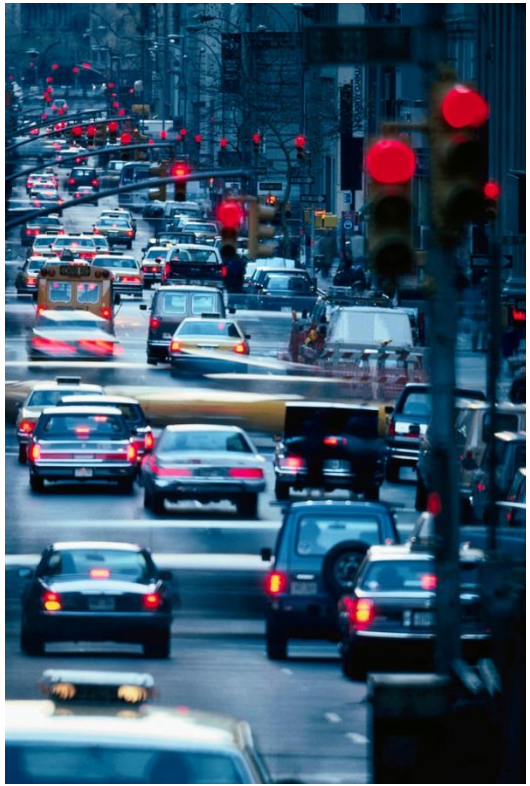
A brief understanding of QoS, Operation and Architecture of data center networks with Nexus platforms is assumed.

Congestion Happens Every Day

"Active traffic management is method of increasing peak capacity and smoothing traffic flows on busy major highways"

Wikipedia: Active Traffic Management

Traffic Management can help …



… or confuse

Session Objectives

- Providing a deep understanding of Implementing and Operating Nexus 7000 Quality of Service (QoS)
- Nexus 7000 Hardware Architecture Deep-Dive for M2 & F2/F2e on QoS/Queuing
- “Nuts and Bolts” on the configuration level and related best practices
- Bridging the gap to implement QoS in a Nexus 7000/6000/5000/2000 Data Center; including architecture info and implementation practices

Session Non-Objectives

- This session will not cover
- Planning, Preparing and Designing of a QoS Strategy (Methodology)
- QoS Application Classification or Marking Strategies
- Policing and Shaping
- Nexus 7000/6000/5000/2000 Hardware Architecture Deep-Dive
- Fiber Channel over Ethernet (FCoE)

Related Sessions

BRKRST-2509	Mastering Data Center QoS
BRKARC-3470	Cisco Nexus 7000 Switch Architecture
BRKARC-2013	Cisco Nexus 3548 Switch Architecture
BRKARC-3452	Cisco Nexus 5000/5500 and 2000 Switch Architecture
BRKARC-3453	Nexus 6000 – Architecture of the next-generation Switch
BRKCRS-2501	Campus QoS Design Simplified
BRKDCT-1044	FCoE of the IP Network Engineer

Recordings available at http://www.ciscolive365.com

QoS Implementation and Operation in a Nexus 7000 environment

BRKDCT-3346

Lukas Krattiger, Consulting Systems Engineer

Lukk@cisco.com

Cisco live!

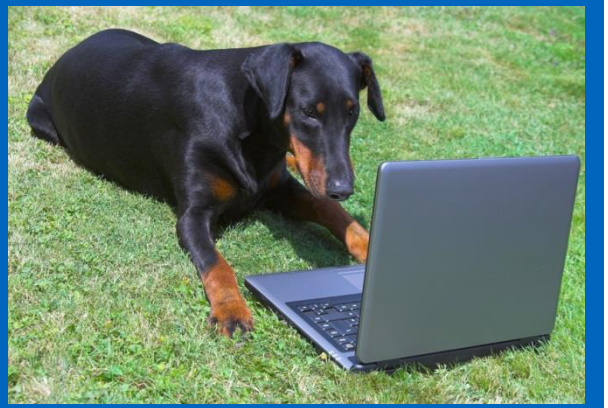
Switzerland

Why QoS in the Data Center?

Assign Color to Traffic

Manage Congestion

Protect Network Resources



maximize throughput, manage congestion!

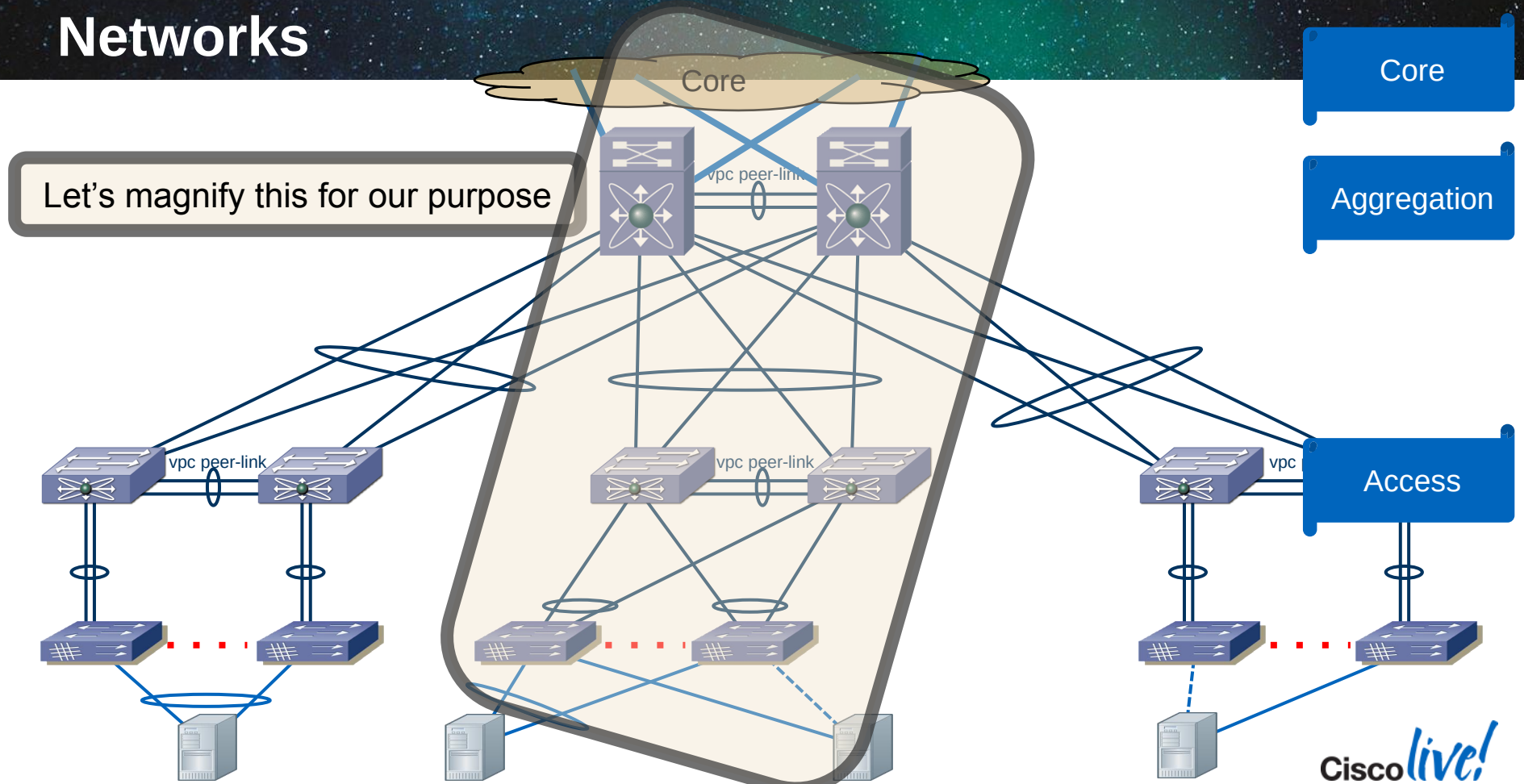
Agenda

- Overview & Introduction
- Nexus 7000 – Configuration & Operation
  - Classification & Marking
  - Queuing
- Case Study – Wrap Up



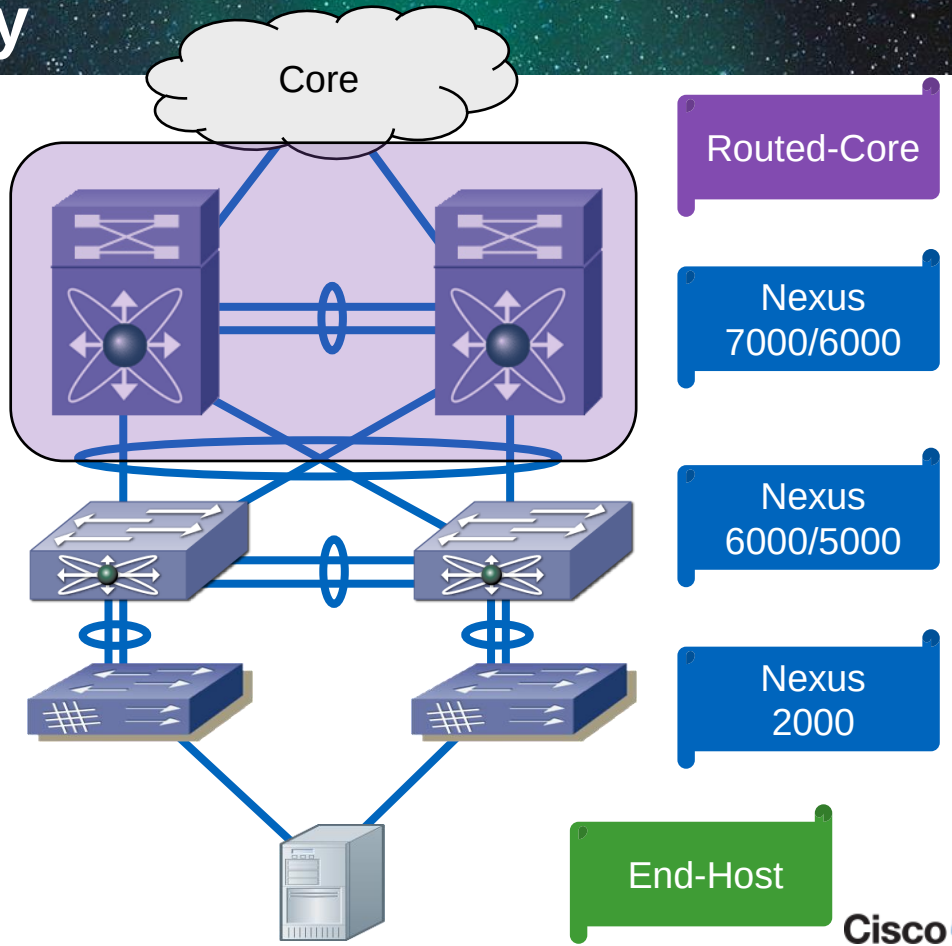

Overview & Introduction

Data Center Networks



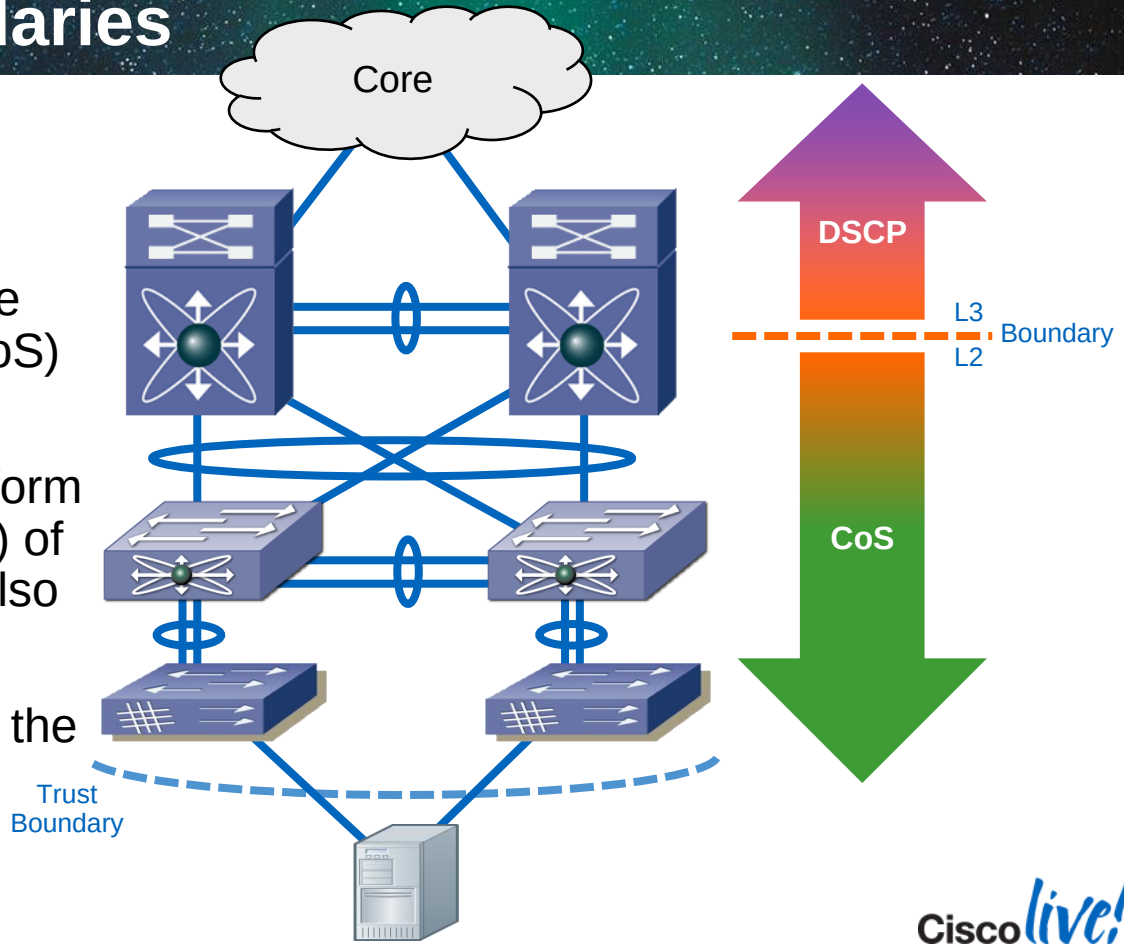
Case Study - Topology

- This is the Topology we will reference to
- Nexus 6000/7000 taking care on Layer-2/Layer-3 boundary in the Aggregation
- Nexus 6000 or 5000 as Layer-2 Access with Nexus 2000 extending the Fabric
- End-Host connected to Nexus 2000 with Ethernet only (Active/Standby)

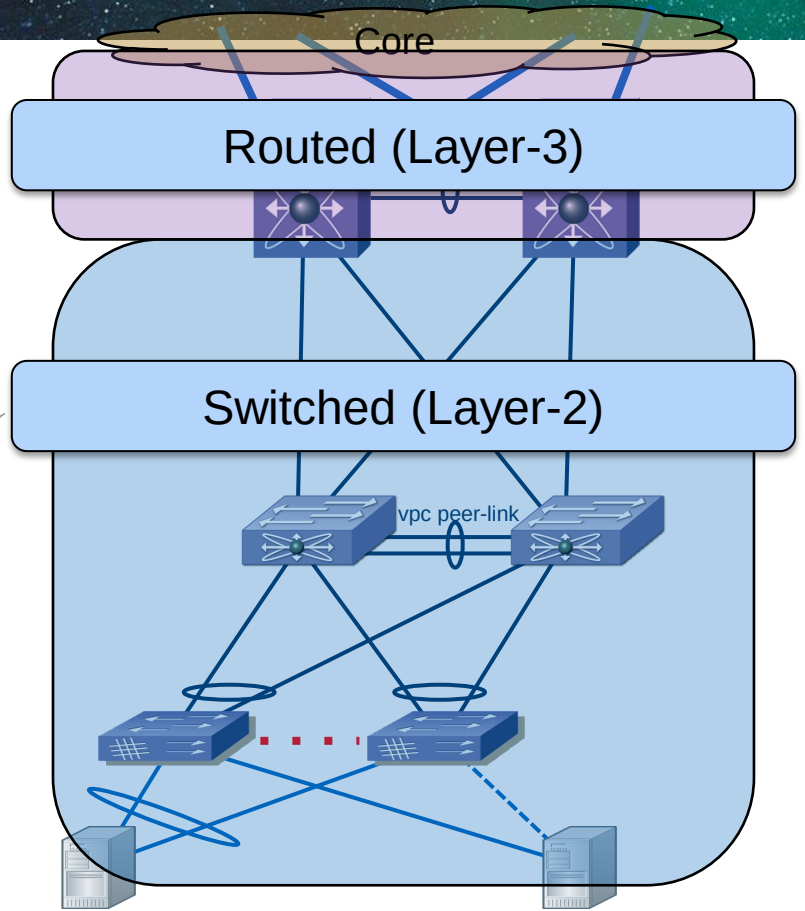


Case Study - Boundaries

- End-Host sends un-marked traffic
- Nexus 5000 and 2000 will be limited to a Layer-2 QoS (CoS) Solution.
- Nexus 7000 required to perform the Layer-3 marking (DSCP) of the Enterprise classes but also Layer-2 (CoS) capable.
- Nexus 6000, the new kid on the block, and a Layer-2/3 (CoS/DSCP) device



Data Center Networks



TOS or EXP gets relevant

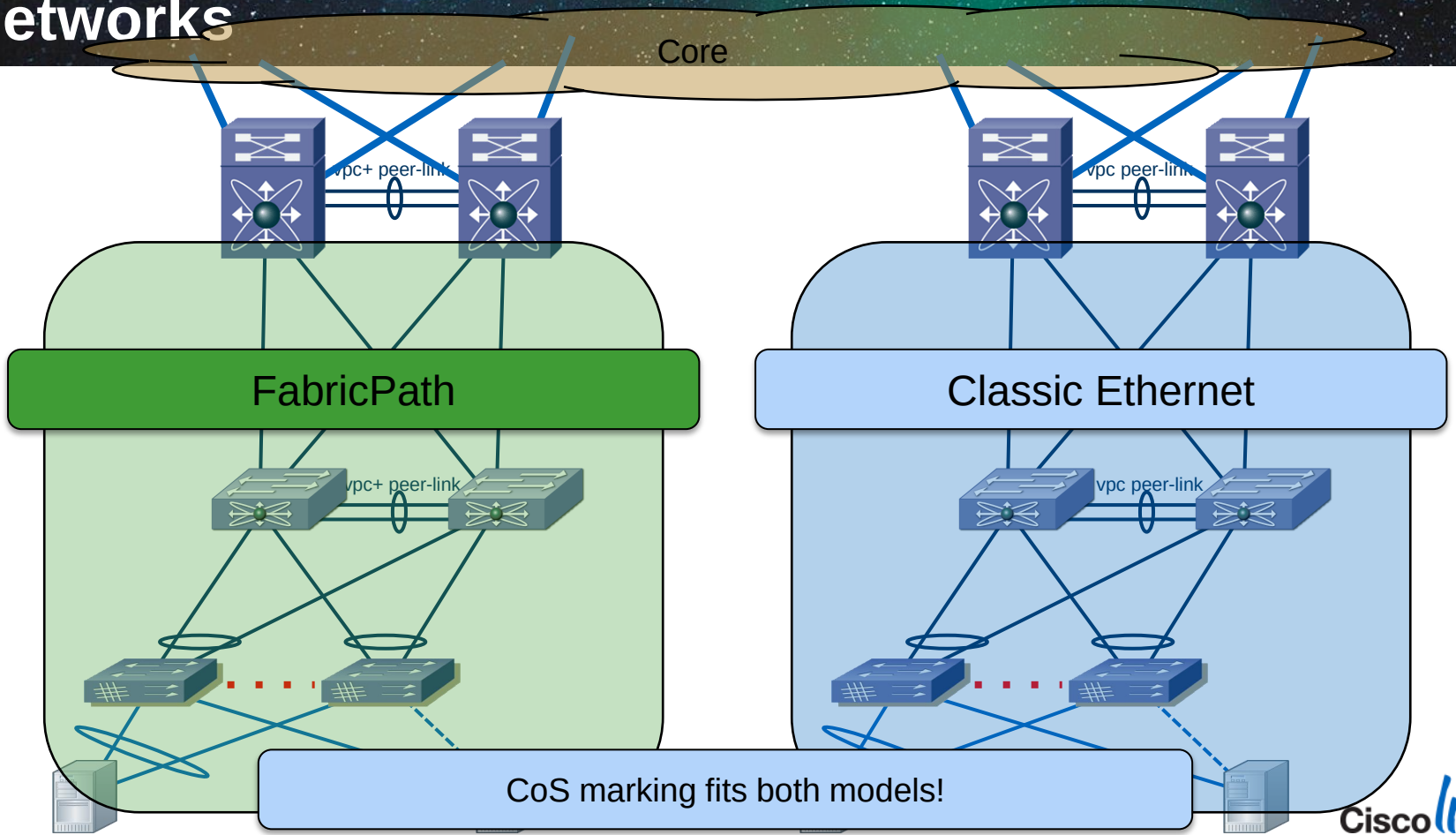
COS is most relevant

Nexus 7000 / 6000

Nexus 6000 / 5000

Nexus 2000

Data Center Networks



CoS or DSCP, this is the question!

- Where to use DSCP and Where to use COS from Design level ... Core/Access/Distribute …
- Data Center Pods are mostly Layer-2
- There is non-IP based Traffic to be considered (FCoE)
- DSCP to CoS / CoS to DSCP mapping is “automatic” at Layer-3 boundary (top most 3bit)

BRKRST-2509
Mastering Data Center QoS

DSCP to CoS / CoS to DSCP – Mapping Tables

```
N7k# show table-map | grep -a 2 cos-dscp-map
  Table-map cos-dscp-map
    default copy
```

```
N7k# show table-map | grep -a 2 dscp-cos-map
  Table-map dscp-cos-map
    default copy
```

```
N7k# show system internal ipqos global-defaults | grep -a 12 cos-dscp-map
table-map: cos-dscp-map (len: 12)
        default copy
        Bit array:
        Values set:

             0  8 16 24 32 40 48 56
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
            -- -- -- -- -- -- -- --
```

```
N7k# show system internal ipqos global-defaults | grep -a 12 dscp-cos-map
table-map: dscp-cos-map (len: 12)
        default copy
        Bit array:
        Values set:

             0  0  0  0  0  0  0  0
             1  1  1  1  1  1  1  1
             2  2  2  2  2  2  2  2
             3  3  3  3  3  3  3  3
             4  4  4  4  4  4  4  4
             5  5  5  5  5  5  5  5
             6  6  6  6  6  6  6  6
             7  7  7  7  7  7  7  7
```

Note: Output taken from Nexus 7000

How are we doing the DSCP to CoS rewrite?

RFC2474

DSCP Class	DSCP (bin)	DSCP (dec)	CoS (bin)	CoS (dec)
BE	000000	0	000	0
CS1 / AF11 / AF12 / AF13	001000-110	8 / 10 / 12 / 14	001	1
CS2 / AF21 / AF22 / AF23	010000-110	16 - 23	010	2
CS3 / AF31 / AF32 / AF33	011000-110	24 / 26 / 28/ 30	011	3
CS4 / AF41 / AF42 / AF43	100000-110	32 / 34 / 36 / 38	100	4
CS5 / EF	101000…110	40 / 46	101	5
CS6	110000	48	110	6
CS7	111000	56	111	7

Default Mapping: 3 most significant bits of DSCP gets mapped to CoS bits

FYI

Case Study - Application Marking Definition

Example

Application	DSCP Differentiated Services Codepoint	PHB Per Hop Behavior	CoS Class of Service
Best Effort High Volume / Less Important	0	BE	0
Low Priority / Scavanger Not Important	8	CS1	1
vMotion / Live Migration Medium Volume / Important	-*		2
Reserved (FCoE) High Volume / Very Important / No Drop	-*		3
Multimedia Medium Volume / Very Important	34	AF41	4
Strict Priority Low Volume / Important / Delay Sensitive	46	EF	5
Network Control Low Volume / Very important	48/56	CS6/CS7	6/7

*Layer-2 only or Non-IP Traffic; will only exist within Data Center Layer-2 Fabric

FYI

Nexus QoS

MQC

- Modular QoS CLI

Classes

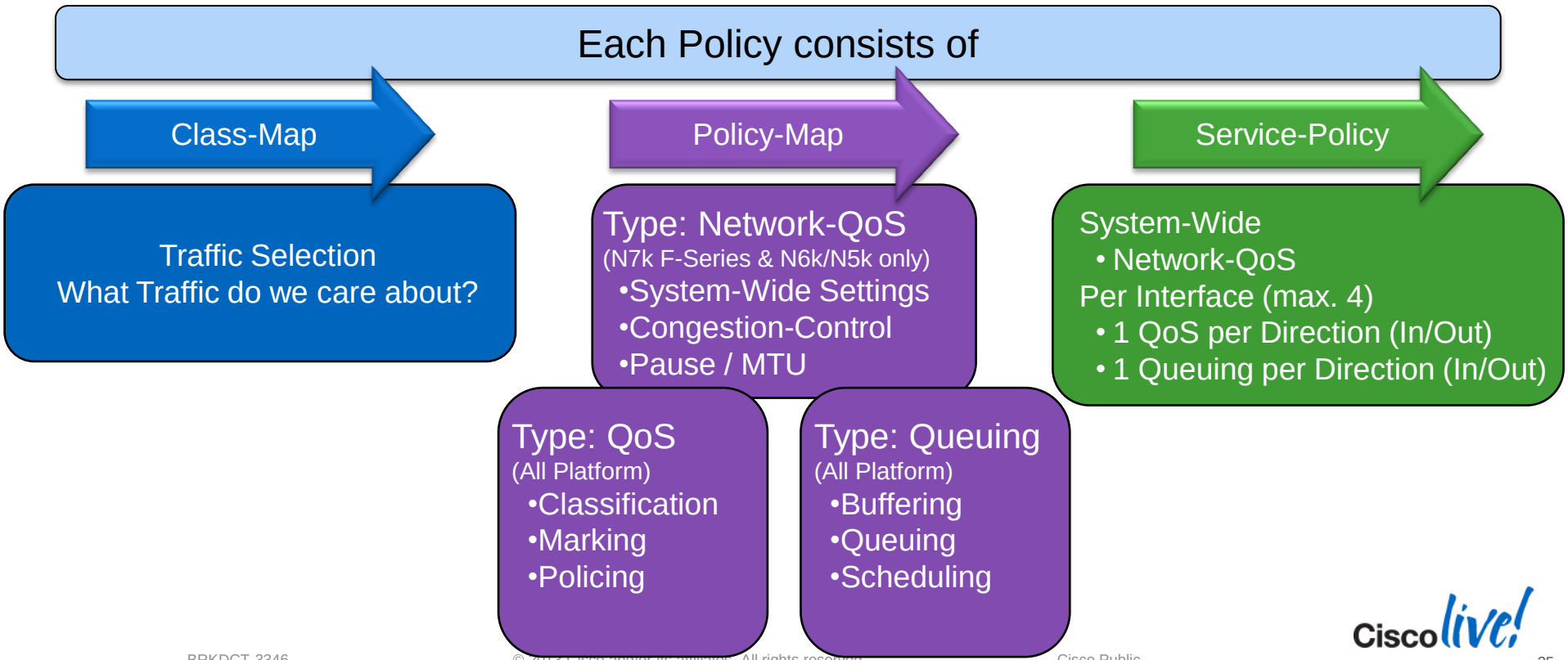
- Nexus 1000v: 64 Classes *(8 predefined)*
- Nexus 3000: 8 Classes *(QoS-Groups)*
- Nexus 5000: 6 Classes *(QoS-Groups)*
- Nexus 6000: 6 Classes *(QoS-Groups)*
- Nexus 7000: 2 to 8 Classes *(Queues)*

Policies

- Type: Network-QOS
- Type: Queuing
- Type: QOS

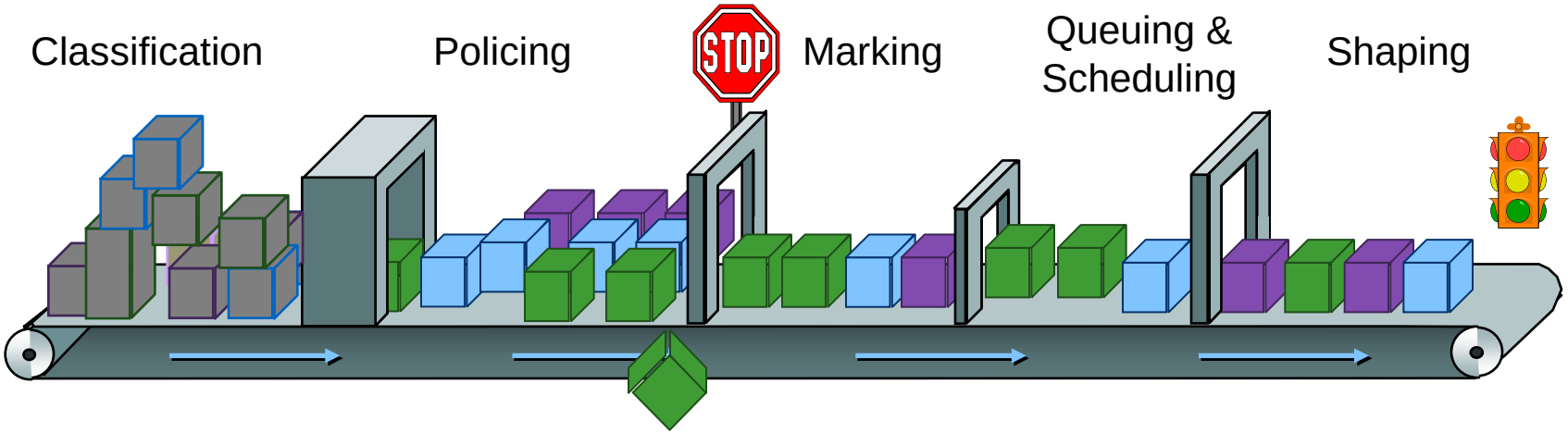
Modular QoS CLI (MQC)

Rules of Engagement



Nexus 7000 – Configuration & Operation

The QoS Toolset



Identify and Split Traffic into Different Classes

Discard Misbehaving Traffic to Maintain Network Integrity

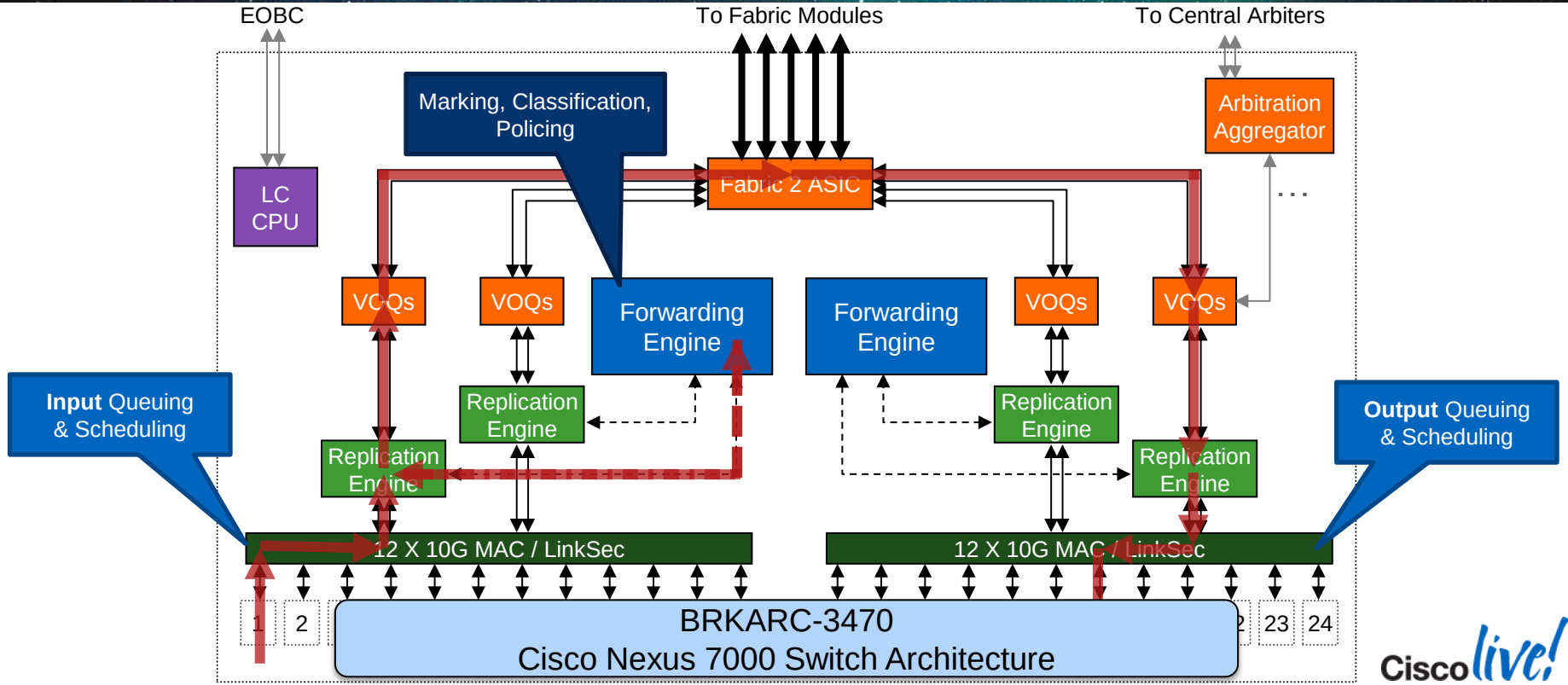
Mark Traffic According to Behavior and Business Policies

Prioritize, Protect and Isolate Traffic Based on Markings

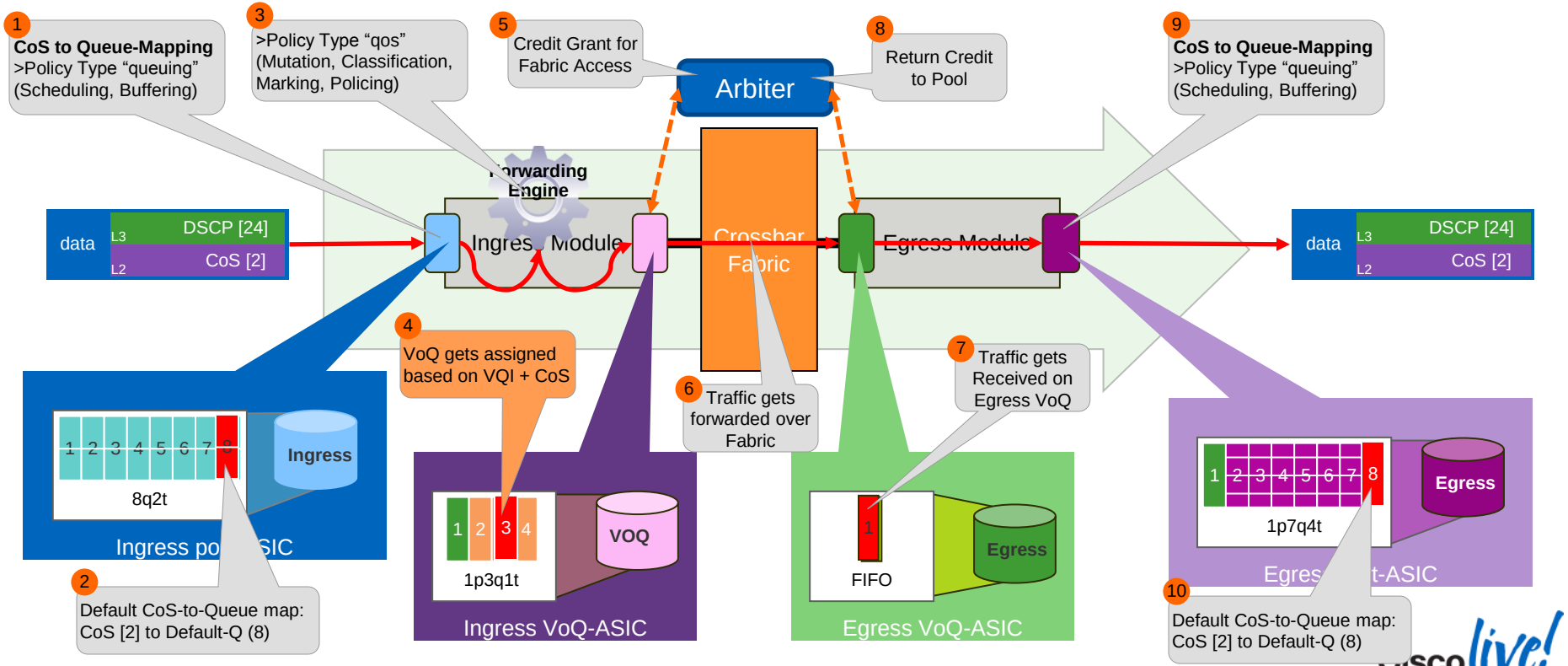
Control Bursts and Conform Traffic

Mapping the QoS Feature Set

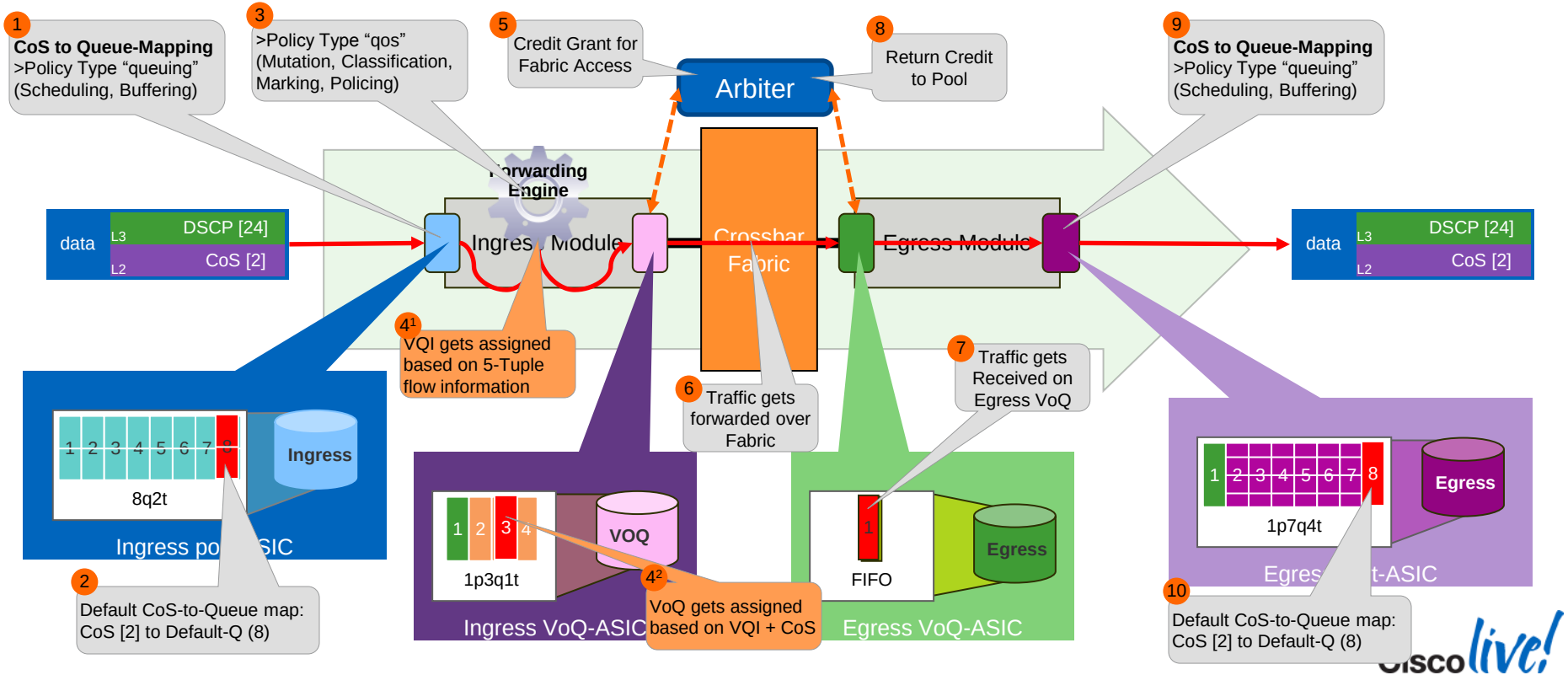
M2 I/O Module Architecture



M2 I/O QoS Packet-Flow (10G)

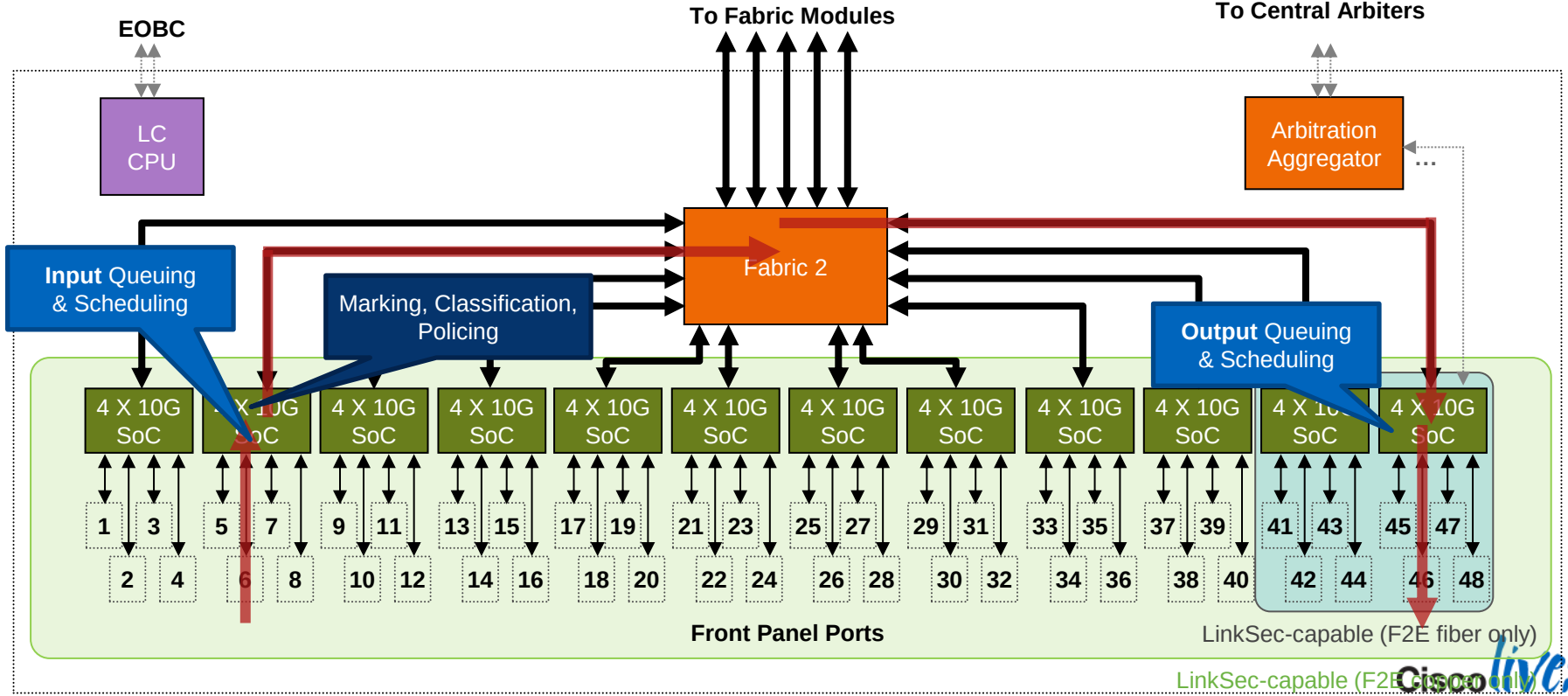


M2 I/O QoS Packet-Flow (40G/100G)



Mapping the QoS Feature Set

F2/F2E I/O Module Architecture



Agenda

- Overview & Introduction
- Nexus 7000 – Configuration & Operation
  - Classification & Marking
  - Queuing
- Case Study – Wrap Up

FYI

Application Classification & Marking - Example

Case Study

Application	DSCP Differentiated Services Codepoint	PHB Per Hop Behavior	CoS Class of Service
Best Effort	0	BE	0
Low Priority / Scavanger	8	CS1	1
vMotion / Live Migration	-*		2
Reserved (FCoE)	-*		3
Multimedia	34	AF41	4
Strict Priority	46	EF	5
Network Control	48/56	CS6/CS7	6/7

*Layer-2 only or Non-IP Traffic; will only exist within Data Center Layer-2 Fabric

FYI

Marking & FCoE Considerations (1)

- F-Series I/O Module is required
- The used template (policy "type network-qos") needs at least one no-drop class, for example the 7e-Template
- If you use a Policy "type qos" which has a "set dscp" action (untrust the port), CoS will be modified based on the topmost 3 bits (unchangeable default dscp to cos mapping)
- If "type qos" Policy with "set dscp" action for DSCP 24-31 is present, DSCP values 24-31 (Binary 011xxx) will automatically map into CoS 3 for egress (unchangeable default dscp to cos mapping)

FYI

Marking & FCoE Considerations (2)

- It is recommended to use symmetrical CoS and DSCP values (example CoS 3 / DSCP 24-31 - Binary 011xxx) in environments where No-Drop queues are used
- It is NOT recommended to use asymmetric CoS and DSCP values (example CoS 2 / DSCP 24-31) in environments where No-Drop queues are used because of potential application of DSCP drive CoS rule
- It is NOT recommended to mix IP traffic and non-IP traffic (FCoE) in one queue; example FCoE traffic has to stay in separate queue alone to be guaranteed

Classification & Marking

The Nexus 7000 Map to Marked Packets

Type: QoS

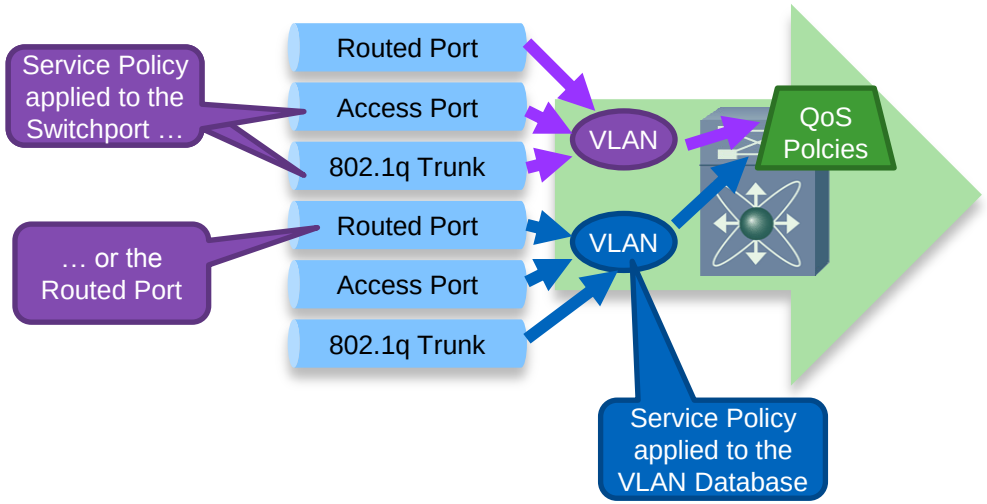
1
```
ip access-list ACL_QOS_LOWPRIO
  10 permit tcp any any eq www
  20 permit tcp any eq www any
```

2
```
class-map type qos match-any CM_QOS_LOWPRIO_COS1
  match access-group name ACL_QOS_LOWPRIO
```

3
```
policy-map type qos PM_QOS_MARK_COS_IN
  class CM_QOS_LOWPRIO_COS1
    set cos 1
```

4
```
interface Ethernet1/1
  service-policy type qos input PM_QOS_MARK_COS_IN
```

Interface or VLAN based QoS



- Interface based QoS Policy takes precedence over VLAN based QoS Policy
- 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 input myPolicy
```

FYI

Classification & Marking – Case Study Config

Type: QoS

```
ip access-list ACL_QOS_LOWPRIO
  10 permit …
ip access-list ACL_QOS_VMOTION
  10 permit …
ip access-list ACL_QOS_MULTIMEDIA
  10 permit …
ip access-list ACL_QOS_SCRICTPRIO
  10 permit …
!
class-map type qos match-any CM_QOS_LOWPRIO_COS1
  match access-group name ACL_QOS_LOWPRIO
!
class-map type qos match-any CM_QOS_VMOTION_COS2
  match access-group name ACL_QOS_VMOTION
!
class-map type qos match-any CM_QOS_MULTIMEDIA_COS4
  match access-group name ACL_QOS_MULTIMEDIA
!
class-map type qos match-any CM_QOS_STRICTPRIO_COS5
  match access-group name ACL_QOS_STRICTPRIO
```

```
policy-map type qos PM_QOS_MARK_COS_IN
  class CM_QOS_STRICTPRIO_COS5
    set cos 5
  class CM_QOS_MULTIMEDIA_COS4
    set cos 4
  class CM_QOS_VMOTION_COS2
    set cos 2
  class CM_QOS_LOWPRIO_COS1
    set cos 1
!
interface Ethernet1/1
  service-policy type qos input PM_QOS_MARK_COS_IN
!
vlan configuration 100
  service-policy input PM_QOS_MARK_COS_IN
```

BRKRST-2509
Mastering Data Center QoS

FYI

Verify Classification & Marking Configuration

Payload-VDC

```
N7k# show policy-map type qos PM_QOS_MARK_COS_IN
…
policy-map type qos PM_QOS_MARK_COS_IN
    class  CM_QOS_VMOTION_COS2
      set cos 2
    class  CM_QOS_LOWPRIO_COS1
      set cos 1
```

shows a specific QoS policy-map configuration

```
N7k# show class-map type qos
…
    class-map type qos match-any CM_QOS_LOWPRIO_COS1
      match access-group name ACL_QOS_LOWPRIO

    class-map type qos match-any CM_QOS_VMOTION_COS2
      match access-group name ACL_QOS_VMOTION
```

shows all class-map of type QoS

```
N7k# show access-lists ACL_QOS_LOWPRIO
…
IP access list ACL_QOS_LOWPRIO
        10 permit ip 1.0.111.11/32 1.0.222.22/32
```

showing a Access-List

FYI

What QOS policies are applied?

Payload-VDC

```
N7k# show policy-map interface brief
…

 Interface/VLAN [Status]:INP QOS       OUT QOS       INP QUE       OUT QUE
================================================================================
…
Ethernet1/1     [Active]:PM_QOS_MARK_C               default-in-po PM_QUEUE_10G-
```

QoS policies

Queuing policies

```
N7k# show policy-map interface ethernet 1/1 type qos
…
  Service-policy (qos) input:   PM_QOS_MARK_COS_IN
    SNMP Policy Index:  285213078
…
    Class-map (qos):   CM_QOS_LOWPRIO_COS1 (match-any)
      5473825 packets
      Match: access-group ACL_QOS_LOWPRIO
      set cos 1
```

#1 command for QOS monitoring

Shows packets matching each class & policer stats

Note: Policer Stats no shown as no Policer applied to Class-Map

FYI

QoS Configuration Limits

Parameter	Verified Limit NX-OS 5.2 – 6.1
Numbers of Class-Maps per Policy[1]	4096
Number of matches in Class-Map	1024
Number of Policers	16’000

[1]ACLs can be used to match packets up to 128 Class-Maps

http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html

FYI

What is QOS HW resources utilization?

Payload-VDC

```
N7k# show hardware capacity
…
         ACL Hardware Resource Utilization (Mod 1)
         ------------------------------------------
                           Used    Free    Percent
                                           Utilization
-------------------------------------------------------
Tcam 0, Bank 0            10      32758   0.03
Tcam 0, Bank 1            3       32765   0.01
Tcam 1, Bank 0            12      32756   0.04
Tcam 1, Bank 1            267     32501   0.81

LOU                        4       100     3.84
…
Protocol CAM               4       3       57.14
Mac Etype/Proto CAM        9       5       64.28

Non L4op labels, Tcam 0    1       6142    0.01
Non L4op labels, Tcam 1    1       6142    0.01
L4 op labels, Tcam 0       0       2047    0.00
L4 op labels, Tcam 1       2       2045    0.09
…
```

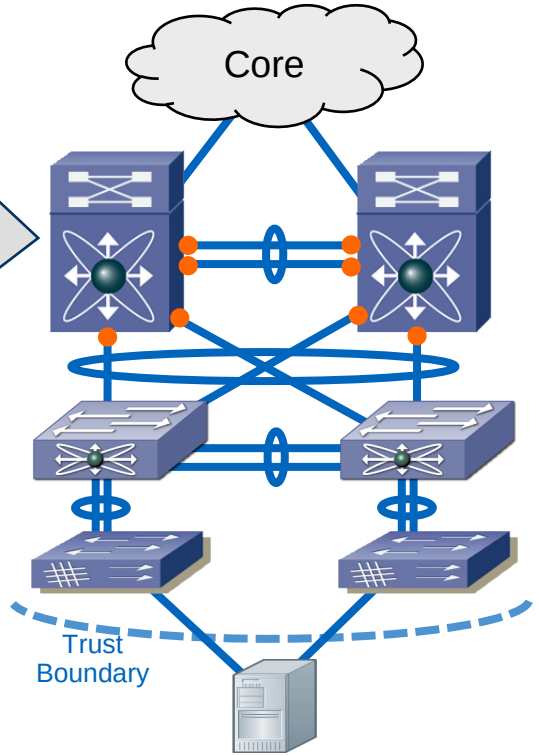
Classification & policing entries consume TCAM space

Note: No Policer resources shown. Would be in section “QoS Resource Utilization” of same command

Classification & Marking (CoS based)

Type: QoS

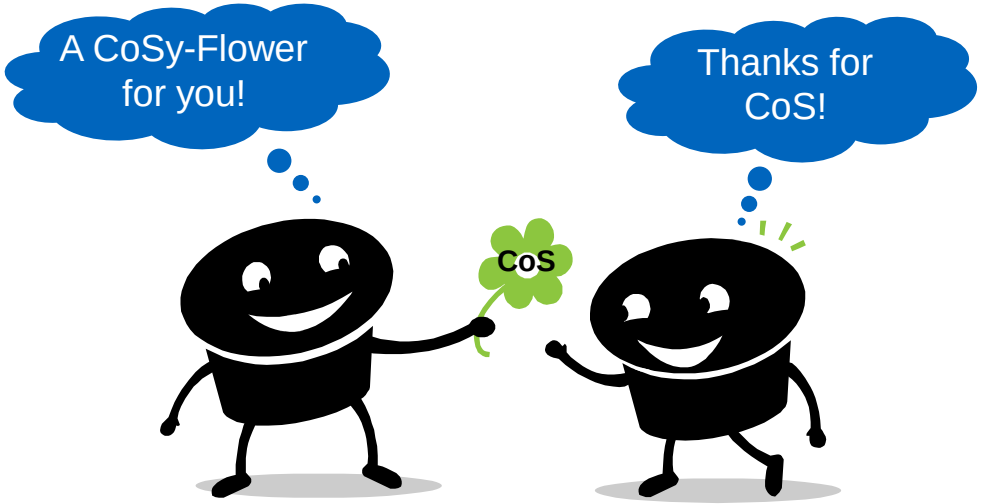
```
class-map type qos match-any CM_QOS_LOWPRIO_COS1
  match cos 1
class-map type qos match-any CM_QOS_VMOTION_COS2
  match cos 2
class-map type qos match-any CM_QOS_FCOE_COS3
  match cos 3
class-map type qos match-any CM_QOS_MULTIMEDIA_COS4
  match cos 4
class-map type qos match-any CM_QOS_STRICTPRIO_COS5
  match cos 5
!
policy-map type qos PM_QOS_MARK_COS_IN
  class CM_QOS_STRICTPRIO_COS5
    set cos 5
  class CM_QOS_MULTIMEDIA_COS4
    set cos 4
  class CM_QOS_FCOE_COS3
    set cos 3
  class CM_QOS_VMOTION_COS2
    set cos 2
 class CM_QOS_LOWPRIO_COS1
    set cos 1
```



Is this really necessary?

Rules of Engagement

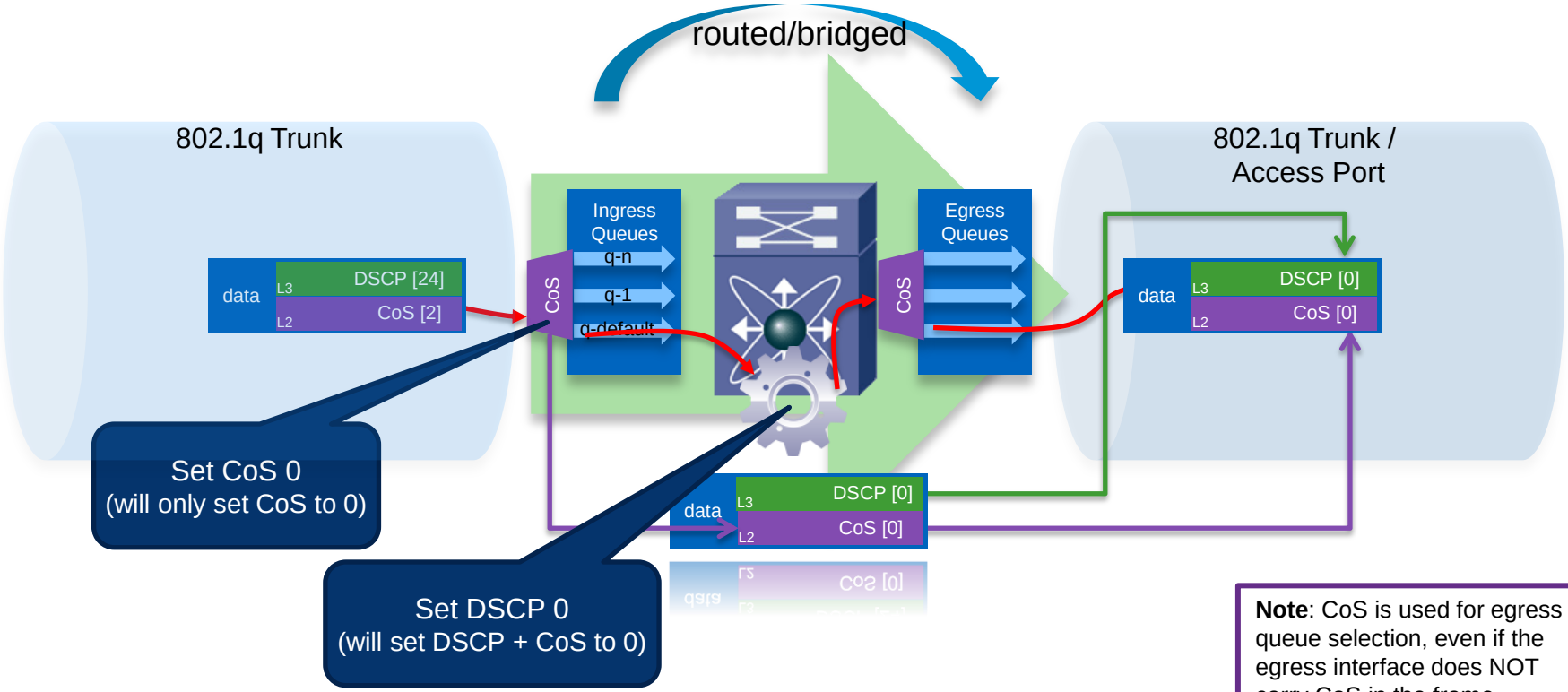
- QoS is enabled by default and cannot be disabled
- Ports “**TRUST**” markings by default
- Lowest common denominator queuing policy applied to all interfaces by default
- Default Queuing and QoS policies are applied to all physical interfaces across all VDCs
- Default policy cannot be modified



Do NOT Trust CoS or DSCP

Changing The Default Trust

Type: QoS



Note: CoS is used for egress queue selection, even if the egress interface does NOT carry CoS in the frame

FYI

Type: QoS Queuing

Do NOT Trust CoS or DSCP

Changing The Default Trust (M-Series I/O Module)

- You can make an interface completely untrusted (CoS & DSCP)
  - CoS affecting bridged traffic
  - DSCP affecting routed traffic
- You need two Policies
  - A "type queuing" policy to set the CoS to 0
  - A "type qos" policy to set the DSCP to 0
- Set DSCP will set the CoS value for Bridged traffic as well.

```
policy-map type queuing Reset-CoS
  class type queuing 8q2t-in-q-default
    set cos 0
    bandwidth percent 100
    queue-limit percent 100
!
policy-map type qos Reset-DSCP
  class class-default
    set dscp 0
!
! Tie to an interface:
interface Ethernet1/1
  service-policy type queuing input Reset-CoS
  service-policy type qos input Reset-DSCP
```

This is how you would configure the Trust Boundary within a Nexus 7000

FYI

Type: QoS Queuing

Do NOT Trust CoS or DSCP

Changing The Default Trust (F-Series I/O Module)

```
qos copy policy-map type queuing default-4q-8e-in-policy prefix UNTRUSTED-
!
policy-map type queuing untrusted-4q-8e-in
  class type queuing 2q4t-8e-in-q1
    queue-limit percent 1
  class type queuing 2q4t-8e-in-q-default
   queue-limit percent 99
   set cos 0
!
policy-map type qos UNTRUSTED
  class class-default
    set dscp 0
!
! Tie to an interface:
interface Ethernet1/1
  service-policy type queuing input untrusted-4q-8e-in
  service-policy type qos input UNTRUSTED
```

FYI

Type: Queuing

Do NOT Trust EXP

Changing The Default Trust (M-Series I/O Module)

- You can make an interface NOT trusting MPLS Experimental
  - EXP affecting MPLS forwarded traffic
- You need one Policy
  - A Type: “queuing" policy and NO set action for MPLS

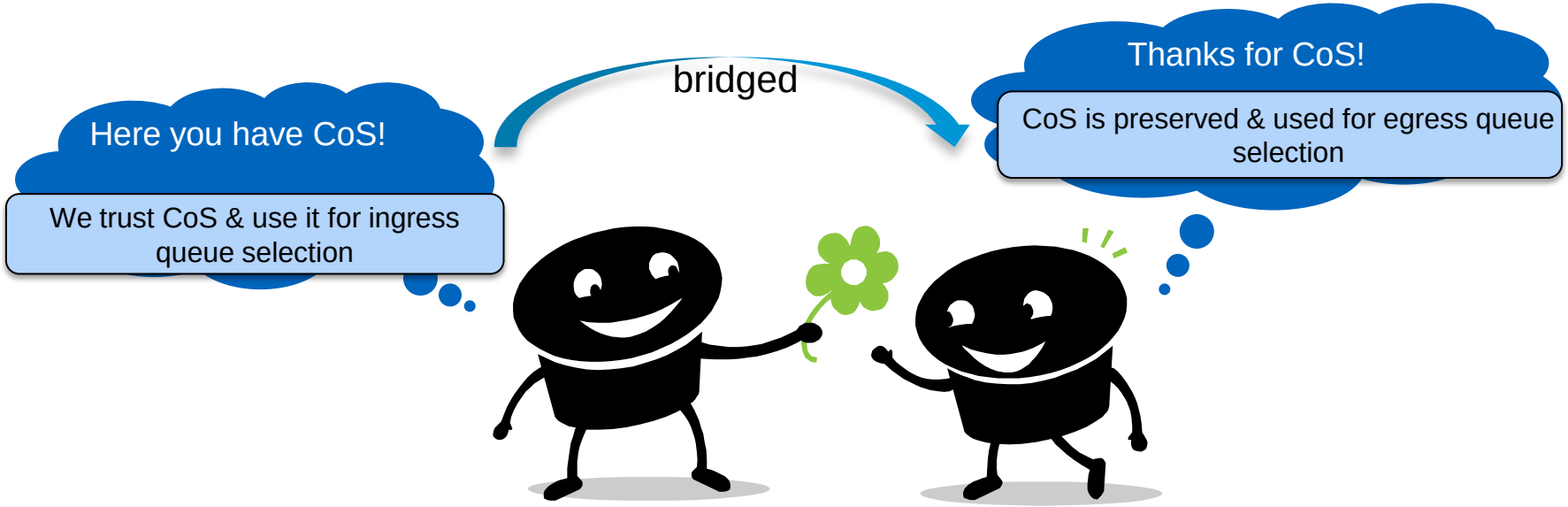
```
policy-map type queuing Reset-EXP
  class type queuing 8q2t-in-q-default
    bandwidth percent 100
    queue-limit percent 100
!
! Tie to an interface:
interface Ethernet1/1
  service-policy type queuing input Reset-EXP
```

This is how you would configure the Trust Boundary within a Nexus 7000

Rules of Engagement

Default Rules

Type: QoS



M-Series I/O Module as per NX-OS 6.1(3):
Packets marked with DSCP but no CoS get treated as CoS 0 thus all traffic uses same ingress queue.

FYI

Default Rules Summary

Type: QoS

Bridged Traffic

- If **CoS** and **DSCP** is present
  - CoS is used for ingress queue selection
  - CoS is preserved
  - DSCP is unmodified
  - CoS is used for egress queue selection
- If **only DSCP** is present
  - No CoS gets treated as CoS 0 on ingress
  - CoS 0 is used for ingress and egress queue selection
  - DSCP is unmodified

Routed Traffic

- If **CoS** and **DSCP** is present
  - CoS is used for ingress queue selection
  - DSCP is preserved and rewrites CoS (top most 3bit)
  - CoS is used for egress queue selection
- If **only DSCP** is present
  - No CoS gets treated as CoS 0 on ingress
  - DSCP is preserved and rewrites CoS (top most 3bit)
  - CoS (derived from DSCP) drives egress queue selection

FYI

Impact of "set DSCP" action in a policy "type qos"

- As soon as you use a Policy "type qos" on a interface, which has a "set dscp" action (untrust the port) the following does apply.
- DSCP will be set as defined in the "set dscp" action
- The "set dscp" action will also drive the CoS value as per the topmost 3 bits (unchangeable default dscp to cos mapping)
- This is true for bridged and routed traffic

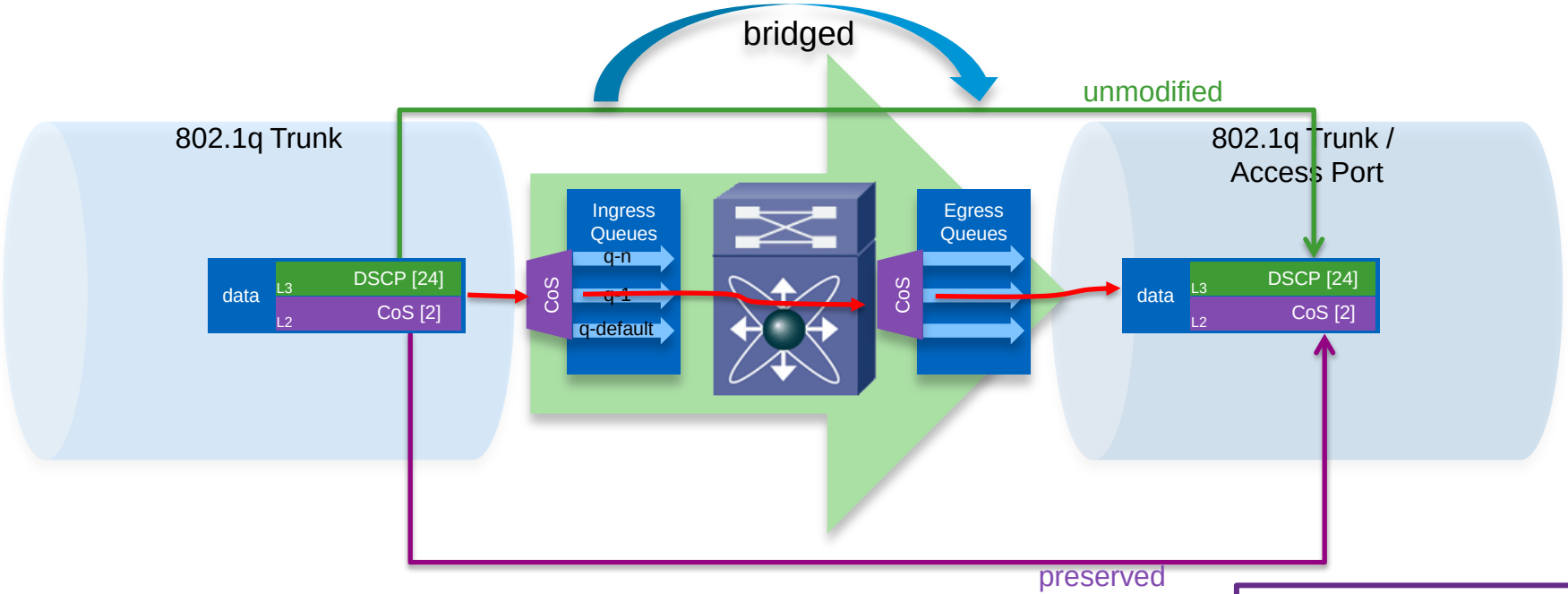
- Note
- The M2 I/O Module behaves differently as per NX-OS 6.1(4). "Set dscp" action does NOT drive the Cos value

FYI

Type: QoS

Bridged Traffic with CoS & DSCP Value

Default Rule #1



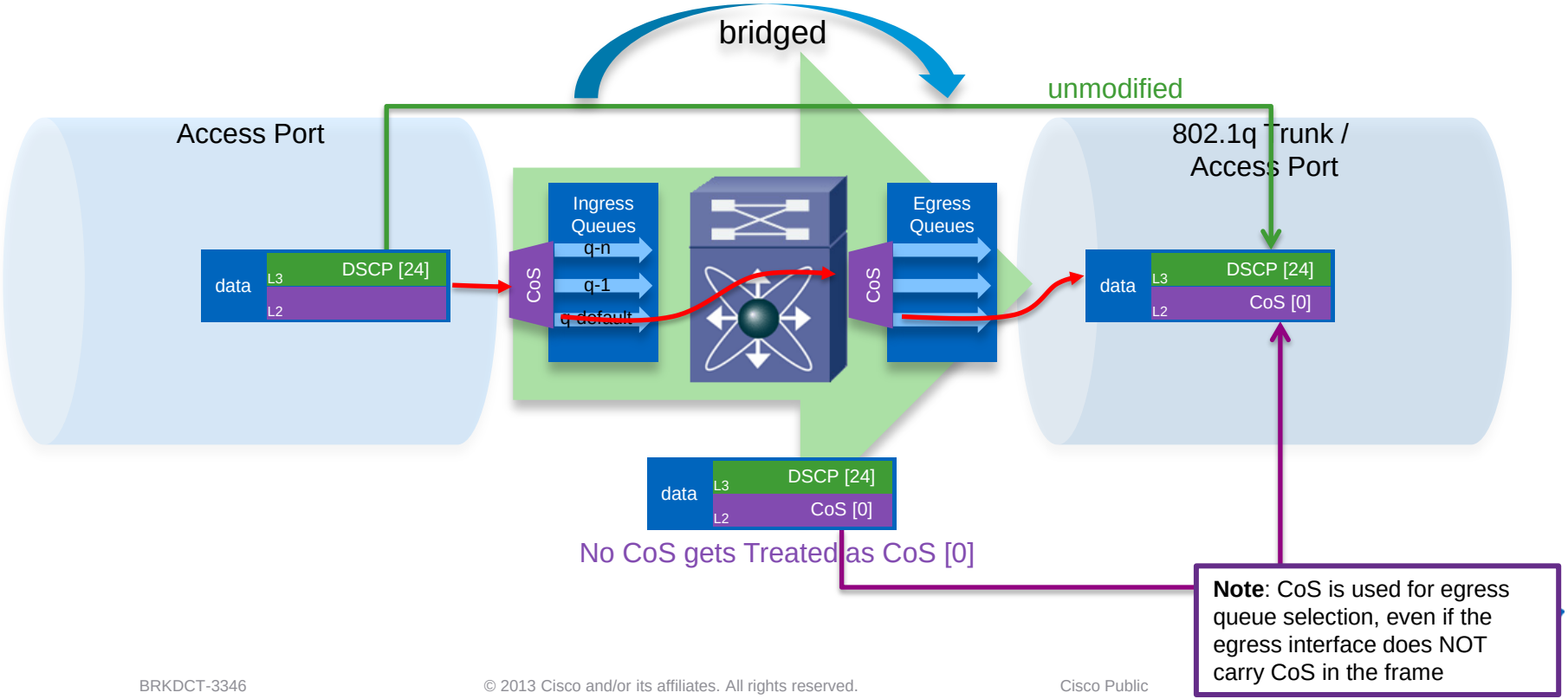
Note: CoS is used for egress queue selection, even if the egress interface does NOT carry CoS in the frame

FYI

Type: QoS

Bridged Traffic with NO CoS but DSCP Value

Default Rule #2

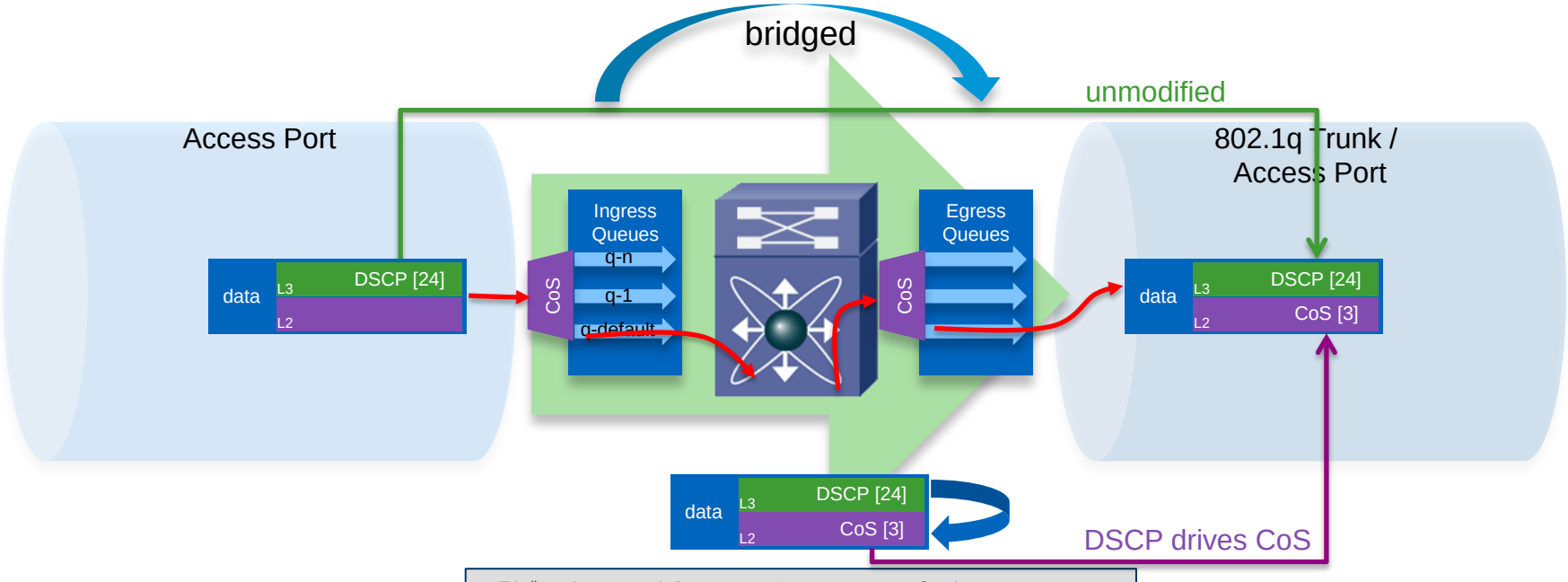


FYI

Type: QoS

Trust Bridged DSCP Traffic

Changing Default Rule #2




```
N7k# show table-map | grep -a 2 dscp-cos-map
  Table-map dscp-cos-map
    default copy
```

Note: CoS is used for egress queue selection, even if the egress interface does NOT carry CoS in the frame

FYI

Type: QoS

Trust Bridged DSCP Traffic

Changing Default Rule #2

- DSCP to drive CoS also for Bridged traffic
- Queuing is based on CoS

```
table-map Trust-DSCP
  default copy
!
policy-map type qos Trust-DSCP
  class class-default
    set dscp dscp table Trust-DSCP
!
! Tie to a VLAN:
vlan configuration <vlan-id>
  service-policy type qos input Trust-DSCP
!
! Tie to an interface:
interface EthernetX/Y
  service-policy type qos input Trust-DSCP
```

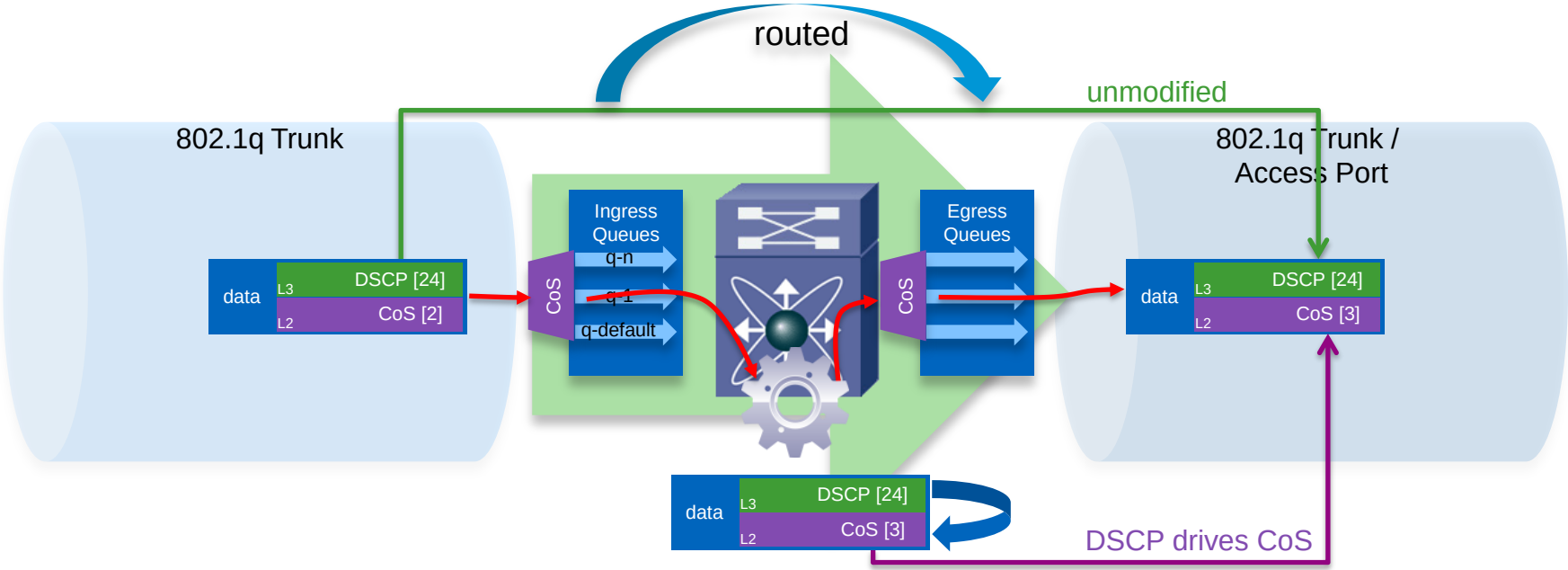
Note: set dscp will also result in DSCP-to-COS rewrite aka consistent QoS information
There is a exception with M2 I/O-Module where DSCP on bridged-frames is preserved

FYI

Type: QoS

Routed Traffic with CoS & DSCP Value

Default Rule #3




```
N7k# show table-map | grep -a 2 dscp-cos-map
  Table-map dscp-cos-map
    default copy
```

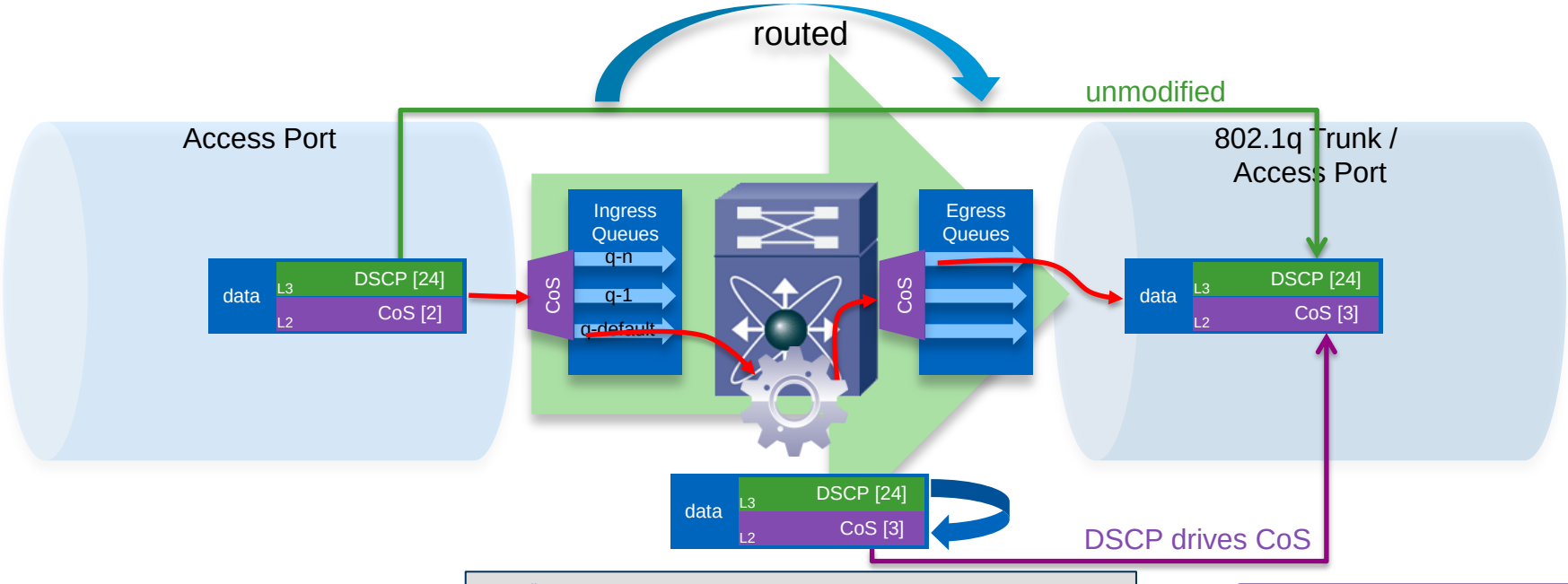
Note: CoS is used for egress queue selection, even if the egress interface does NOT carry CoS in the frame

FYI

Type: QoS

Routed Traffic with NO CoS & DSCP Value

Default Rule #4




```
N7k# show table-map | grep -a 2 dscp-cos-map
  Table-map dscp-cos-map
    default copy
```

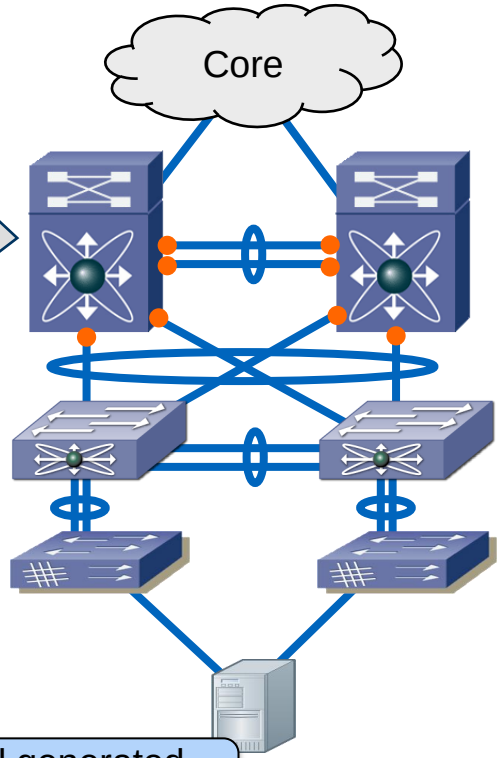
Note: CoS is used for egress queue selection, even if the egress interface does NOT carry CoS in the frame

FYI

Classification & Marking (Re-Marking)

Type: QoS

```
class-map type qos match-any CM_QOS_LOWPRIO_COS1
  match cos 1
class-map type qos match-any CM_QOS_VMOTION_COS2
  match cos 2
class-map type qos match-any CM_QOS_MULTIMEDIA_COS4
  match cos 4
!
policy-map type qos PM_QOS_MARK_COS_IN
 class CM_QOS_MULTIMEDIA_COS4
    set dscp 38
 class CM_QOS_VMOTION_COS2
    set dscp 8
 class CM_QOS_LOWPRIO_COS1
    set dscp 16
```

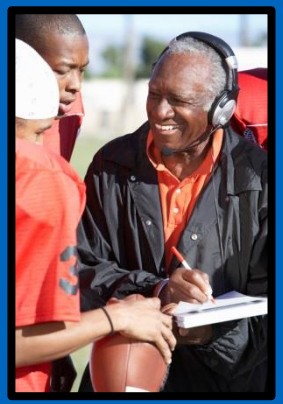


Note: Do NOT use class “Class-Default” for re-marking, local generated traffic could get altered.

Agenda

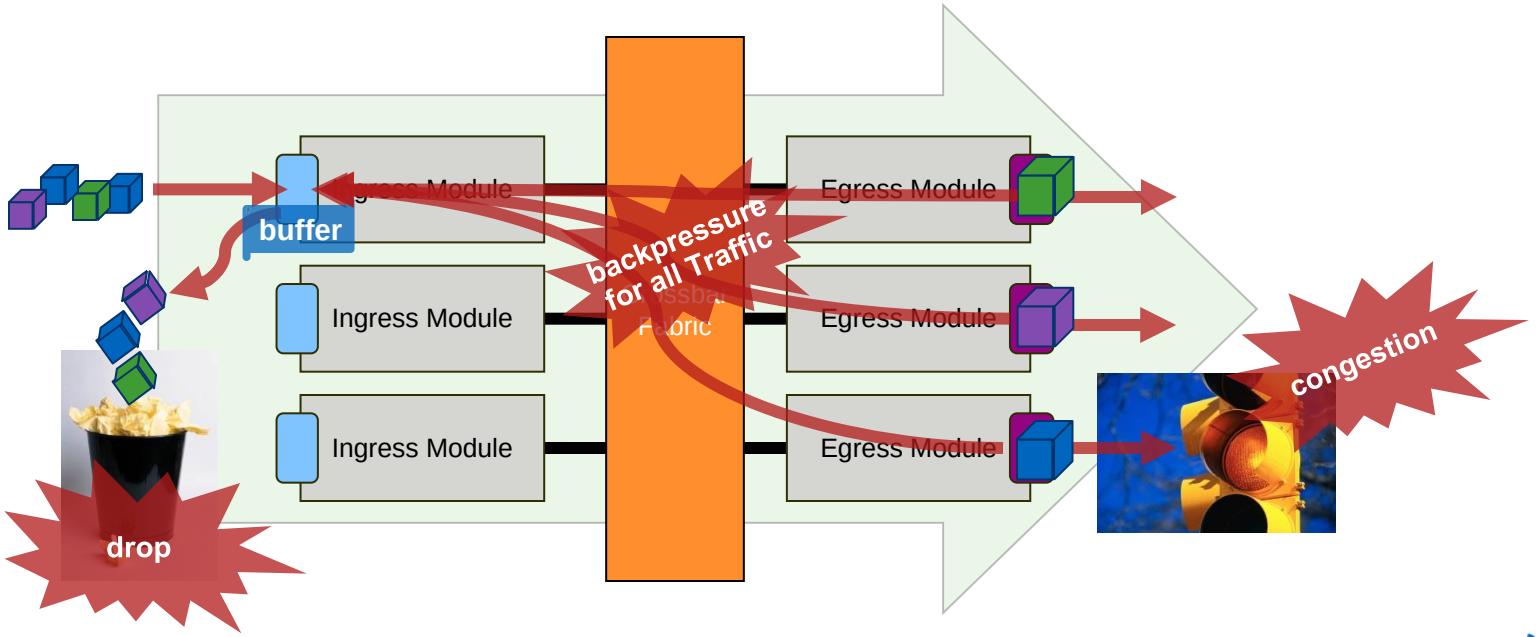
- Overview & Introduction
- Nexus 7000 – Configuration & Operation
  - Classification & Marking
  - Queuing & Buffering
- Case Study – Wrap Up

Head of Line Blocking



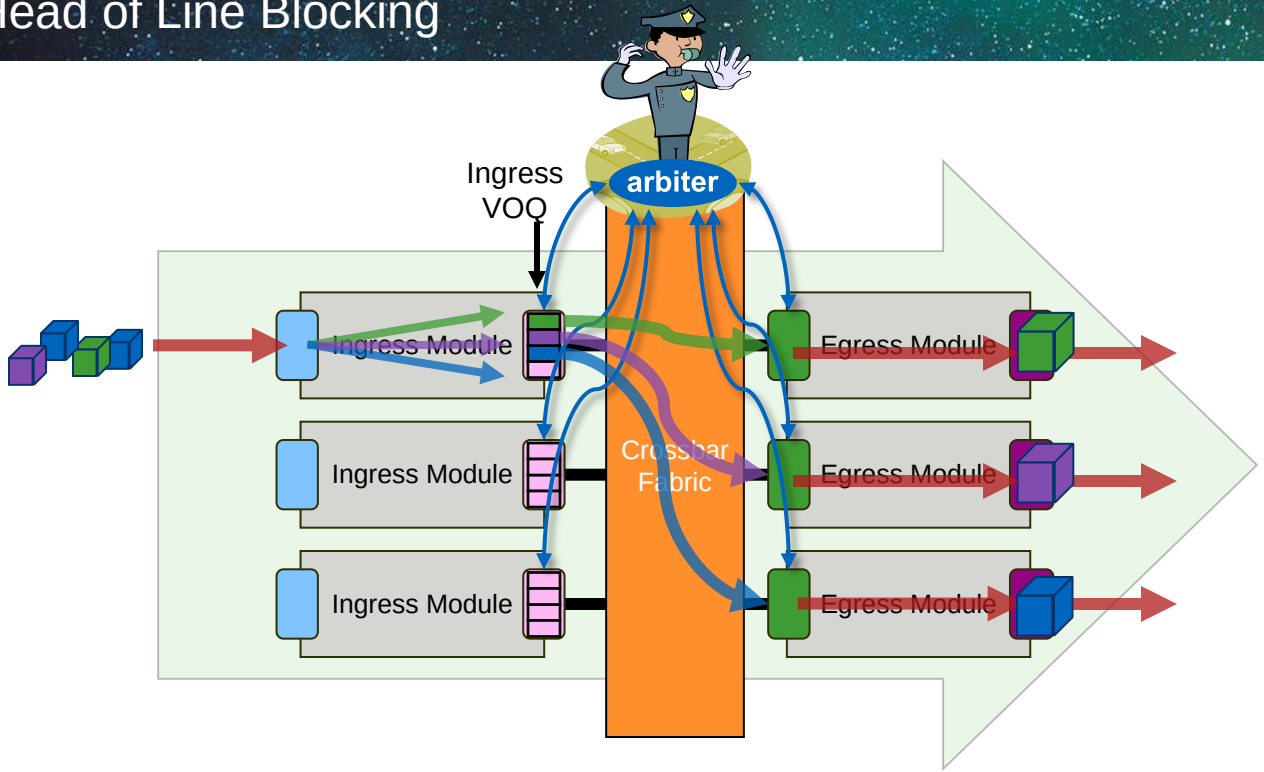
Head of Line Blocking

What is the Problem?



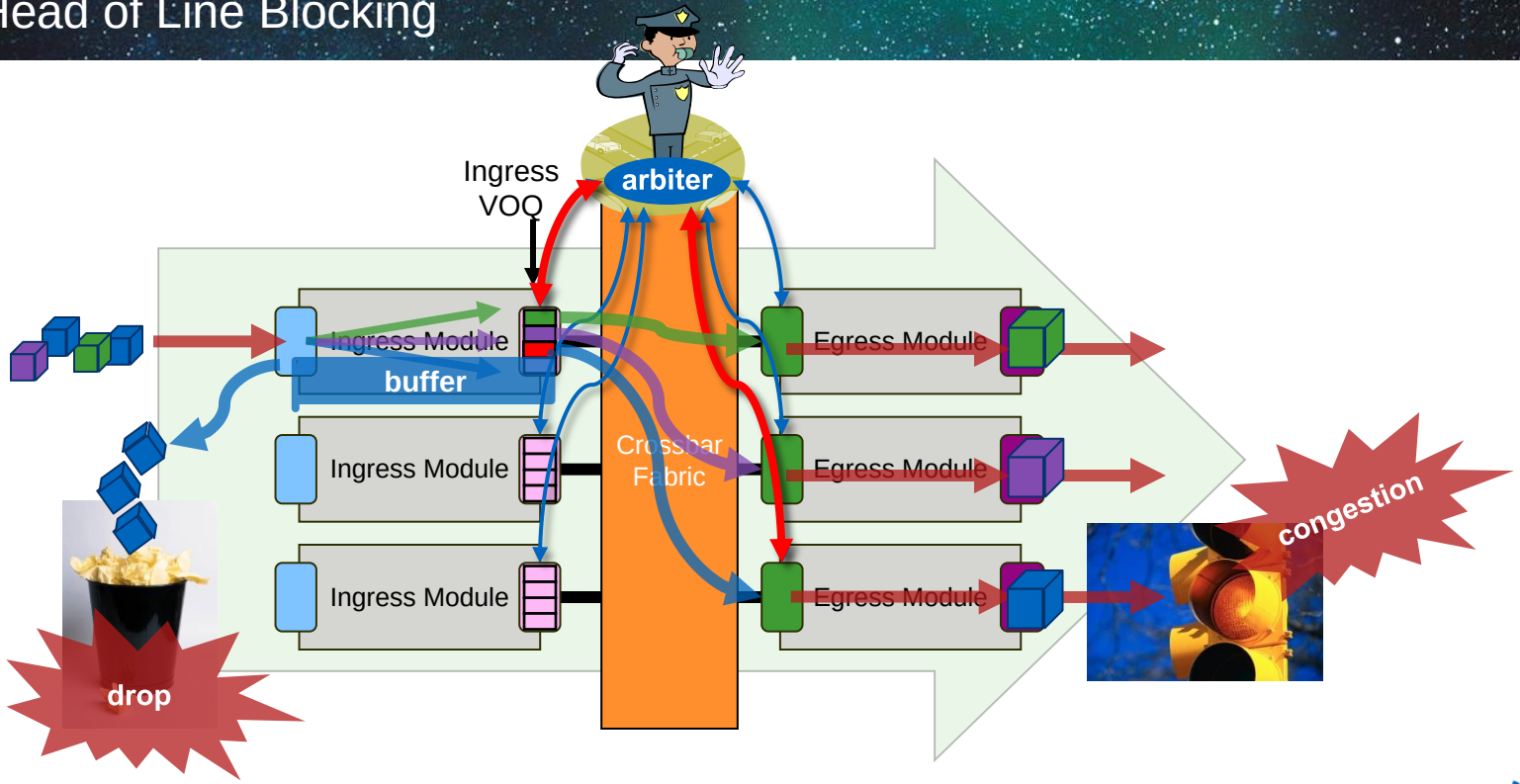
Virtual Output Queues

Avoid Head of Line Blocking



Virtual Output Queues

Avoid Head of Line Blocking



FYI

Benefits of VoQ and Central Arbitration

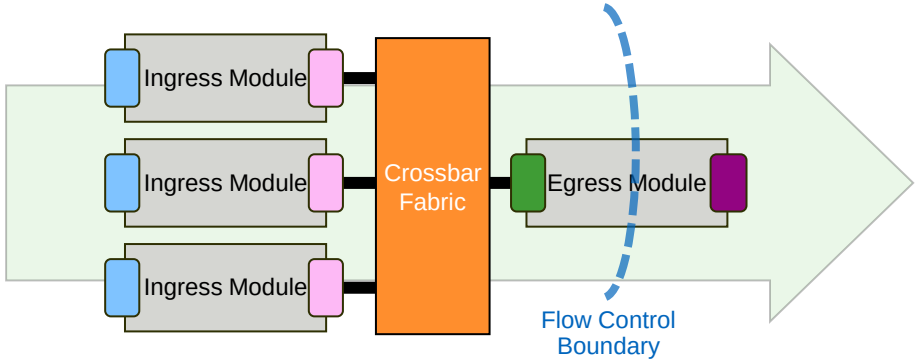
- Ensures **priority traffic takes precedence** over best-effort traffic across fabric
  - Four levels of priority for each VOQ destination
- Ensures **fair access to bandwidth** for multiple ingress ports transmitting to one egress port
  - Central arbiter ensures all traffic sources get appropriate access to fabric bandwidth, even with traffic sources on different modules
- **Prevents** congested egress ports from blocking ingress traffic destined to other ports
  - Mitigates head-of-line blocking by providing independent queues for individual destinations across the fabric
- Enables **lossless service for some traffic classes** across the fabric
  - Can provide strict priority and backpressure (blocking instead of dropping) for certain traffic classes, such as FCoE traffic

Two different Animals

Two different Buffering Models

M-Series vs. F-Series I/O Module

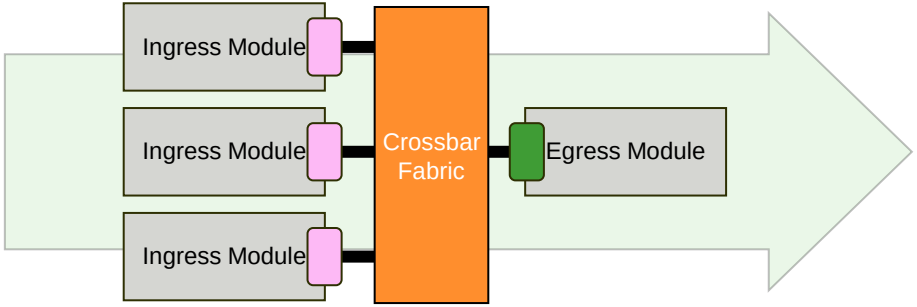
M-Series



Hybrid Model

- combining ingress VOQ-buffered architecture with egress port-buffered architecture
- Similar to Catalyst 6500

F-Series



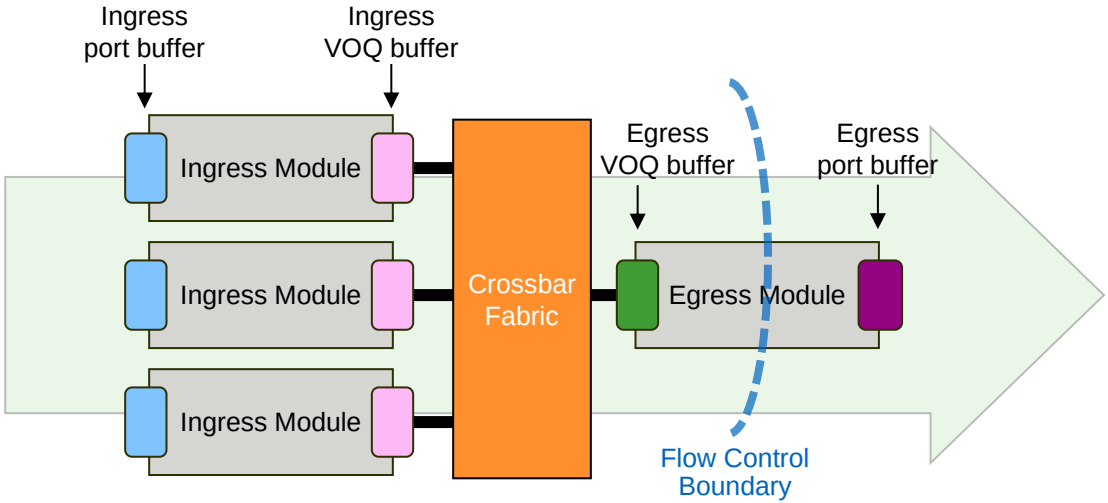
Ingress Model

- pure ingress VOQ-buffered architecture
- Similar to Nexus 5000/6000

M-Series Queuing & Buffering

Hybrid Ingress/Egress Buffering Model

Two different Buffering Models (M2)

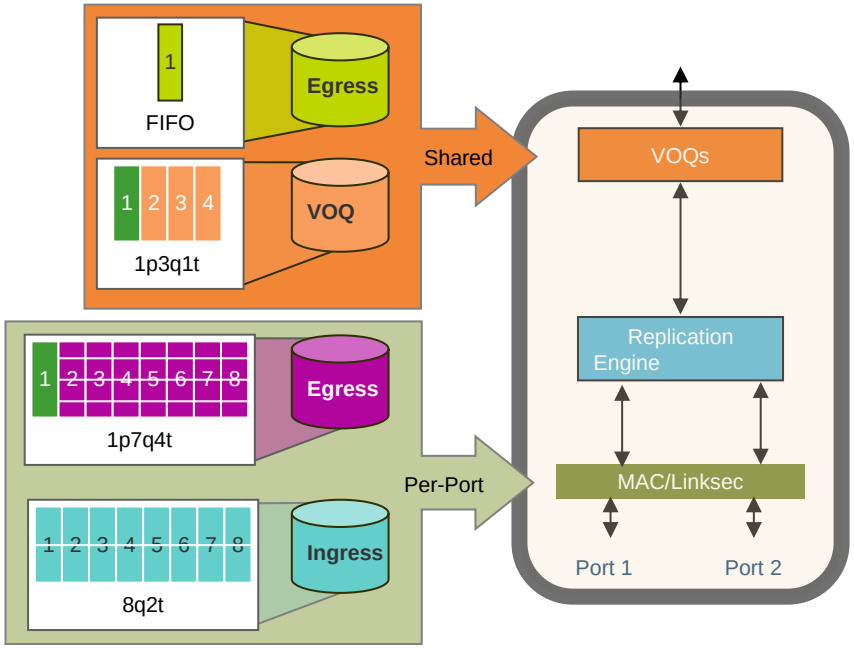


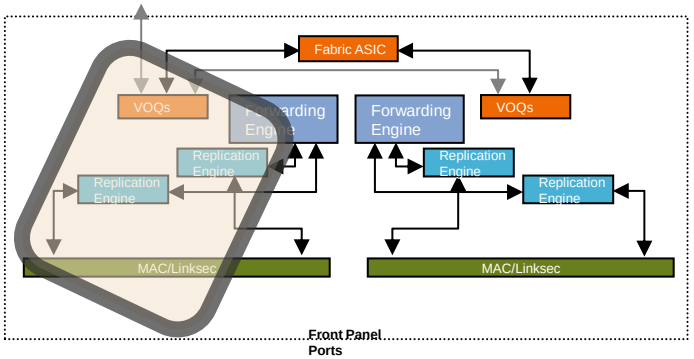
- Ingress port buffer – Manages congestion of ingress forwarding/replication engines, and congestion toward egress destinations (VQIs)
- Ingress VOQ buffer – Manages congestion toward egress destinations (VQIs)
- Egress VOQ buffer – Receives frames from fabric; also buffers multi-destination frames
- Egress port buffer – Manages congestion at egress interface

FYI

M2 I/O Module Buffering

Hybrid ingress/egress buffered architecture on M2 I/O module





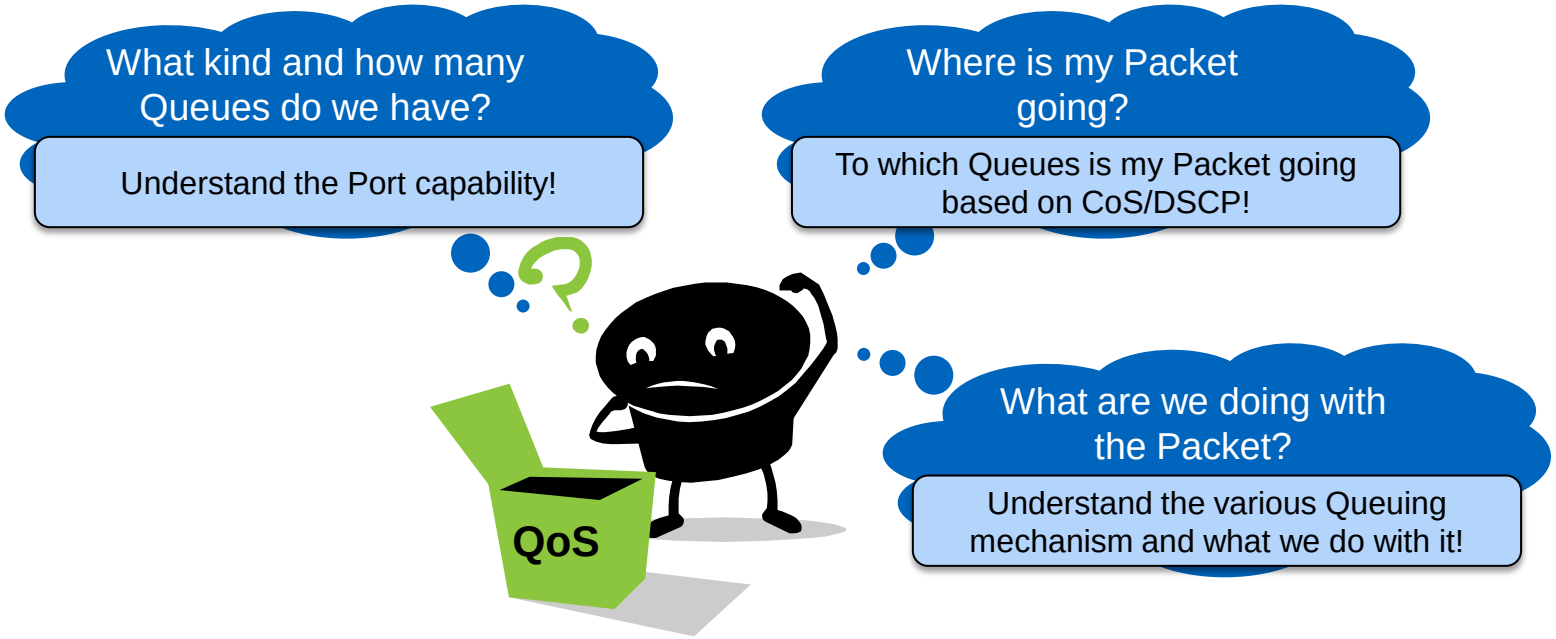
FYI

Queuing and Buffering Details

Nexus 7000 M-Series I/O Module

Module Type	Mode	Port Queuing Structure	Per-Port Buffer Capacity	VOQ Buffer Capacity
M2 24-port 10G	10G	8q2t / 1p7q4t	4.5MB / 4.5MB	64MB per Module
M2 6-port 40G	40G		18MB / 18MB	
	10G (Breakout cable)*		4.5MB / 4.5MB	
M2 2-port 100G	100G		54MB / 54MB	
	40G		18MB / 18MB	
	10G (Breakout cable)*		4.5MB / 4.5MB	
M1 32-port 10G	10G (Dedicated)		1MB per Port + 65MB / 80MB	32MB per Module
	10G (Shared)		1MB per Port Group + 65MB / 80MB	
M1 8-port 10G	10G		92MB / 80MB	
M1 48-port 1G	SFP	2q4t / 1p3q4t	7.56MB / 6.15MB	16MB per Module
M1 48-port 1G	RJ-45			

*roadmap item



Queuing Structures

What kind and how many queues do we have?

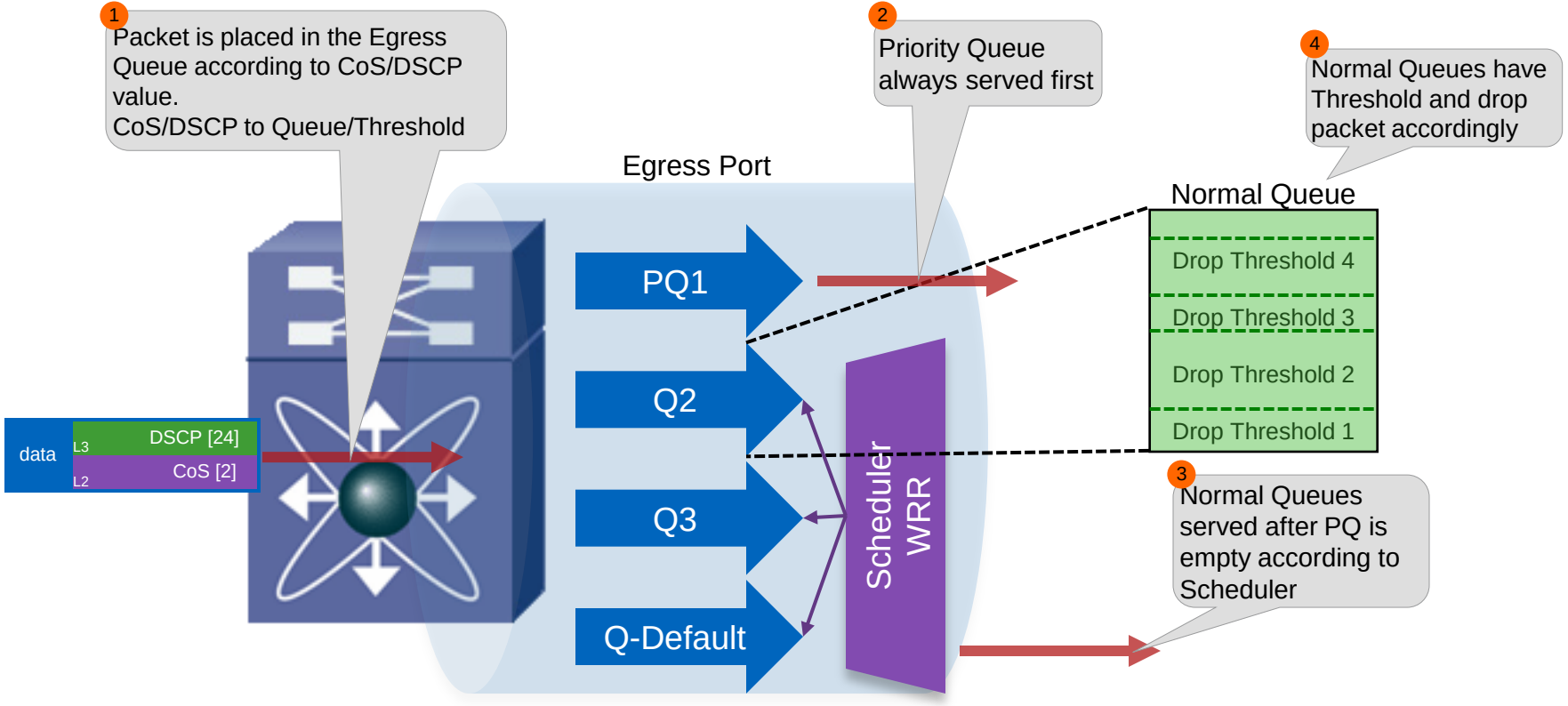
	M1-Series 1G	M1- & M2-Series 10G / 40G / 100G	F2-Series 1G/10G
Ingress	2q4t	8q2t	4q1t / 2q1t
Egress	1p3q4t	1p7q4t	1p3q1t, 2p2q1t, 3p1q1t

***N*p (1p)** Number of **Strict Priority Queue**, this Queue is always served first (NOT part of Scheduling)

***N*q (3q)** Number **Scheduled Queues**, these Queues are served when NO packet exists in the Priority Queue (part of Scheduling; WRR or RR)

***N*t (4t)** Number of **Drop Threshold** applied to the Normal Queues. Watermark for WRED or TailDrop

How Queuing and Scheduling works – 1p3q4t



FYI

Queuing and Scheduling Considerations

- Queuing Structure is I/O Module dependent
- Queuing will be configured with policy-map type queueing and does NOT interfere with policy-map type qos
- Each Queue has a queue-size in Byte per Port-ASIC (Hardware Queue)
- Queues can be tuned by Configuration queue-limit

- How to display Interface Queuing-Capabilities

```
N7k# show interface ethernet 1/1 capabilities | i Model|QOS
  Model:                  N7K-M224XP-23L
  QOS scheduling:         rx-(8q2t),tx-(1p7q4t)
```

Ingress/Egress Queuing Class-Maps

Admin- / Default-VDC only

Type: Queuing

- class-map type queuing – Configures COS to Queue mapping
- Queuing class-map names are static, based on port-type and queue (Predefined)

1G Egress Queue Structure

1G Ingress Queue Structure

```
N7k-ADMIN(config)# class-map type queuing match-any
  1p3q4t-out-pq1                1p7q4t-out-q-default   1p7q4t-out-q6
  1p3q4t-out-q-default          1p7q4t-out-q2          1p7q4t-out-q7
  1p3q4t-out-q2                 1p7q4t-out-q3          2q4t-in-q-default
  1p3q4t-out-q3                 1p7q4t-out-q4          2q4t-in-q1
  1p7q4t-out-pq1                1p7q4t-out-q5          8q2t-in-q-default
N7k-ADMIN(config)# class-map type queuing match-any 1p7q4t-out-pq1
N7k-ADMIN(config-cmap-que)# match cos 7
N7k-ADMIN(config-cmap-que)#
```

```
  8q2t-in-q1    8q2t-in-q6
  8q2t-in-q2    8q2t-in-q7
  8q2t-in-q3
  8q2t-in-q4
  8q2t-in-q5
```

10G/40G/100G Ingress Queue Structure

10G/40G/100G Egress Queue Structure

Admin-/Default-VDC only

FYI

Type: Queuing

CoS to Queue Mapping Configuration

Defaults (Admin- / Default-VDC)

```
class-map type queuing match-any 2q4t-in-q1
  match cos 5-7
class-map type queuing match-any 2q4t-in-q-default
  match cos 0-4
!
class-map type queuing match-any 8q2t-in-q1
  match cos 5-7
class-map type queuing match-any 8q2t-in-q-default
  match cos 0-4
!
class-map type queuing match-any 1p3q4t-out-pq1
  match cos 5-7
class-map type queuing match-any 1p3q4t-out-q-default
  match cos 0-4
!
class-map type queuing match-any 1p7q4t-out-pq1
  match cos 5-7
class-map type queuing match-any 1p7q4t-out-q-default
  match cos 0-4
```

Class-Map Type Queuing and Default Queuing

Type: Queuing

- Predefined policy-map type queuing exist for input and output in each payload VDC.
- They are applied to all ports where there is no user defined policy-map attached
- default policy-maps cannot be modified

```
N7k(config)# policy-map type queuing default-out-policy
ERROR: Changes to default policy-map(s) not permitted
```

```
N7k# show policy-map type queuing


  Type queuing policy-maps
  ========================

  policy-map type queuing default-in-policy
    class type queuing in-q1
      queue-limit percent 50
      bandwidth percent 80
    class type queuing in-q-default
      queue-limit percent 50
      bandwidth percent 20

 policy-map type queuing default-out-policy
    class type queuing out-pq1
      priority level 1
      queue-limit percent 16
    class type queuing out-q2
      queue-limit percent 1
    class type queuing out-q3
      queue-limit percent 1
    class type queuing out-q-default
      queue-limit percent 82
      bandwidth remaining percent 25
```

Policy-Map Type Queuing

Type: Queuing

- policy-map type queuing – Define per-queue behavior such as queue size, WRED, shaping

```
N7k(config)# policy-map type queuing pri-q
N7k(config-pmap-que)# class type queuing 1p7q4t-out-pq1
N7k(config-pmap-c-que)#
bandwidth       no              queue-limit     set
exit            priority        random-detect   shape
N7k(config-pmap-c-que)#
```

- priority – defines queue as the priority queue
- bandwidth – defines WRR weights for each queue
- shape – defines SRR weights for each queue
- queue-limit – defines queue size and defines tail-drop thresholds
- random-detect – sets WRED thresholds for each queue

Note that some “sanity” checks are only performed when you attempt to tie the policy to an interface

FYI

Policy-Map Type Queuing: Capabilities

Type: Queuing

Module Type	Priority Queue (priority)	WRR Scheduler (bandwidth)	SRR – Shaped Round Robin (shape)	Buffer (queue-limit)	Congestion Management	
					WRED Threshold (random-detect)	Taildrop Threshold (queue-limit cos)
M2 24-port 10G	✓[1]	✓	✓[1]	✓	✓	✓
M2 6-port 40G	✓[1]	✓	✓[1]	✓	✓	✓
M2 2-port 100G	✓[1]	✓	✓[1]	✓	✓	✓
M1 32-port 10G	✓[1]	✓	■[2]	■[3]	✓	✓
M1 8-port 10G	✓[1]	✓	■[2]	✓	✓	✓
M1 48-port 1G	✓[1]	✓	✗	✓	✗	✓
M1 48-port 1G	✓[1]	✓	✗	✓	✗	✓

[1] egress Direction only
[2] only without Priority Queue
[3] not configurable

Default Policy-Map gotchas (1)

Type: Queuing

- Default policy-map type queuing only allocate buffer to Priority Queue and Queue default!

```
policy-map type queuing default-out-policy
    class type queuing out-pq1
      priority level 1
      queue-limit percent 16
    class type queuing out-q2
      queue-limit percent 1
    class type queuing out-q3
      queue-limit percent 1
    class type queuing out-q-default
      queue-limit percent 82
      bandwidth remaining percent 25
```

CoS 5-7 by Default

Very Low (1%) buffer allocated to Q2 & Q3

CoS 0-4 by Default

- M 1G I/O Modules have only 4 Queues (1p3q4t)
- M 10G/40G/100G I/O Modules have 8 Queues (1p7Q4t) hence using same default-out-policy BUT Queues 4-7 are having NO Buffer at all

Changing CoS to Queue-Map (Q2, Q3, Q4 or above) does NOT assign necessary buffers and will lead to packet drops !!
Do NOT forget to assign Buffers (Queue-Limit) to the Queues you use !!

Default Policy-Map gotchas (2)

Type: Queuing

- By default, default class-map type queuing only have traffic send to Priority-Queue1 and Default Queue (CoS are assigned to it) !

```
      class-map type queuing match-any 1p7q4t-out-pq1
        Description: Classifier for egress priority queue of type 1p3q4t
        match cos 5-7

      class-map type queuing match-any 1p7q4t-out-q2
        Description: Classifier for egress queue 2 of type 1p7q4t

      class-map type queuing match-any 1p7q4t-out-q3
        Description: Classifier for egress queue 3 of type 1p7q4t

...

      class-map type queuing match-any 1p7q4t-out-q7
        Description: Classifier for egress queue 7 of type 1p7q4t

      class-map type queuing match-any 1p7q4t-out-q-default
        Description: Classifier for egress default queue of type 1p7q4t
```

Modification in those queues (1p7q4t-out-q2) also applies to default-out-policy of Queue2

Allocating CoS to Q2, Q3, Q4 or above means allocating traffic to Queue with very low buffer (1%) if default policy is used !!!

Recommendation:
Modifying CoS to Queue mapping requires to attach the user defined Policy-Map to each Port on each VDC !!

CoS to Queue Mapping - Example

Case Study – M2 I/O Module

Application	CoS	Queuing (Scheduling)	Queue-Limit (Buffer)	Queue (8q2t / 1p7q4t)	Character
Best Effort	0	BW remaining 40%	40%	8q2t-in-q-default / 1p7q4t-out-q-default	High Volume / Less Important
Low Priority / Scavanger	1	BW remaining 10%	10%	8q2t-in-q7 / 1p7q4t-out-q7	Not Important
vMotion / Live Migration	2	BW remaining 20%	10%	8q2t-in-q6 / 1p7q4t-out-q6	Medium Volume / Important
Multimedia	4	BW remaining 30%	29%	8q2t-in-q2 / 1p7q4t-out-q2	Medium Volume Very Important
Strict Priority	5	-	10%	8q2t-in-q1 / 1p7q4t-out-pq1	Low Volume / Important / Delay Sensitive
Network Control	6/7				Low Volume / Very important

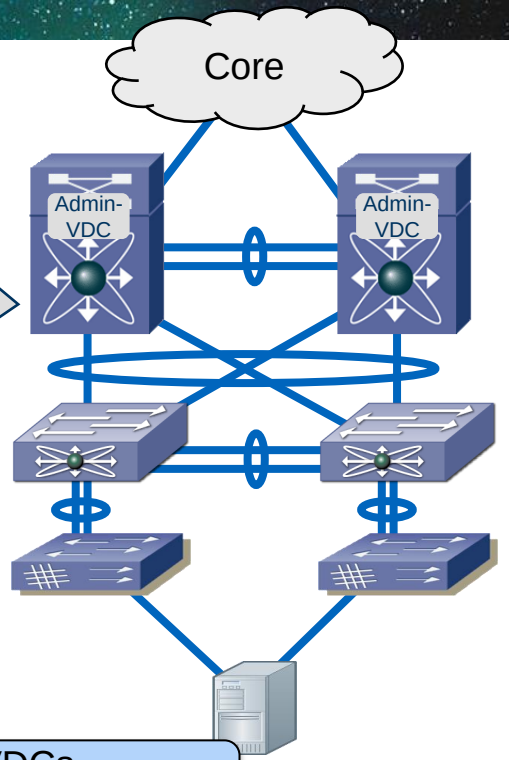
What else need to be done here after changing the cos to queue mapping?

CoS to Queue Mapping Configuration

Case Study (Admin- / Default-VDC)

Type: Queuing

```
class-map type queuing match-any 8q2t-in-q1
  match cos 5-7
class-map type queuing match-any 8q2t-in-q2
  match cos 4
class-map type queuing match-any 8q2t-in-q6
  match cos 2
class-map type queuing match-any 8q2t-in-q7
  match cos 1
class-map type queuing match-any 8q2t-in-q-default
  match cos 0
!
class-map type queuing match-any 1p7q4t-out-pq1
  match cos 5-7
class-map type queuing match-any 1p7q4t-out-q2
  match cos 4
class-map type queuing match-any 1p7q4t-out-q6
  match cos 2
class-map type queuing match-any 1p7q4t-out-q7
  match cos 1
class-map type
  match cos 0
```



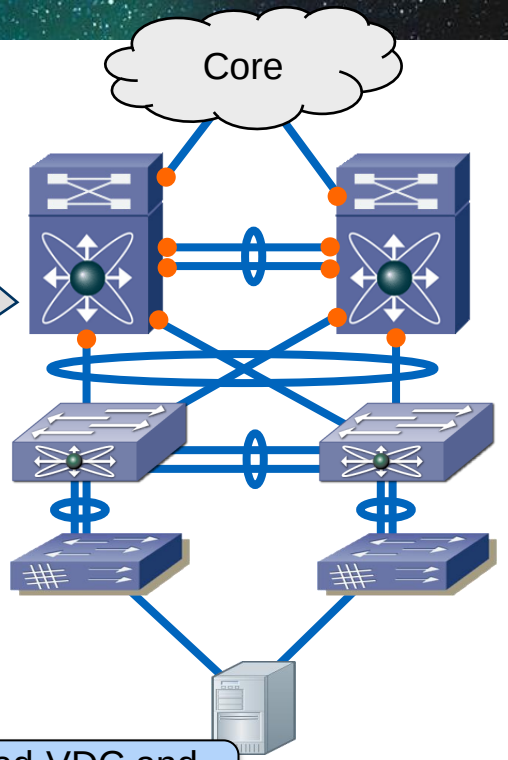
Changes apply to ALL ports of specified type in ALL VDCs
Changes are traffic disruptive for ports of specified type

Queuing Configuration (Egress)

Case Study (Payload-VDC)

Type: Queuing

```
policy-map type queuing PM_QUEUE_10G-40G-100G_OUT
  class type queuing 1p7q4t-out-pq1
    priority level 1
    queue-limit percent 10
  class type queuing 1p7q4t-out-q2
    queue-limit percent 29
    bandwidth remaining percent 30
  class type queuing 1p7q4t-out-q6
    queue-limit percent 10
    bandwidth remaining percent 20
  class type queuing 1p7q4t-out-q7
    queue-limit percent 10
    bandwidth remaining percent 10
  class type queuing 1p7q4t-out-q-default
    queue-limit percent 40
    bandwidth remaining percent 40
!
interface Ethernet1/1
  service-policy type queuing output PM_QUEUE_10G-40G-100G_OUT
```



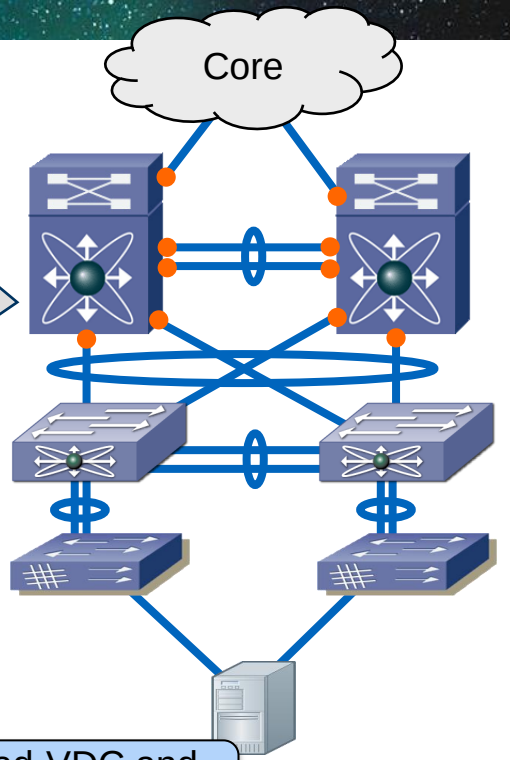
All Policy-Map and Service-Policy are done in relevant Payload-VDC and only affect the interface to which they get applied

Queuing Configuration (Ingress)

Case Study (Payload-VDC)

Type: Queuing

```
policy-map type queuing PM_QUEUE_10G-40G-100G_IN
  class type queuing 8q2t-in-q1
    queue-limit percent 10
    bandwidth remaining percent 10
  class type queuing 8q2t-in-q2
    queue-limit percent 20
    bandwidth remaining percent 30
  class type queuing 8q2t-in-q6
    queue-limit percent 10
    bandwidth remaining percent 10
  class type queuing 8q2t-in-q7
    queue-limit percent 10
    bandwidth remaining percent 10
  class type queuing 8q2t-in-q-default
    queue-limit percent 40
    bandwidth remaining percent 40
!
interface Ethernet1/1
  service-policy type queuing input PM_QUEUE_10G-40G-100G_IN
```



All Policy-Map and Service-Policy are done in relevant Payload-VDC and only affect the interface to which they get applied

FYI

Verify the CoS to Queue mapping (COS2Q)

Payload-VDC

```
N7k# show system internal qos queuing config interface ethernet 1/1 | begin COS2Q
COS2Q Config
Direction: egress
COS 0 Queue: 1p7q4t-out-q-default
COS 1 Queue: 1p7q4t-out-q7
COS 2 Queue: 1p7q4t-out-q6
COS 3 Queue: 1p7q4t-out-q-default
COS 4 Queue: 1p7q4t-out-q2
COS 5 Queue: 1p7q4t-out-pq1
COS 6 Queue: 1p7q4t-out-pq1
COS 7 Queue: 1p7q4t-out-pq1

COS2Q Config
Direction: ingress
COS 0 Queue: 8q2t-in-q-default
COS 1 Queue: 8q2t-in-q7
COS 2 Queue: 8q2t-in-q6
COS 3 Queue: 8q2t-in-q-default
COS 4 Queue: 8q2t-in-q2
COS 5 Queue: 8q2t-in-q1
COS 6 Queue: 8q2t-in-q1
COS 7 Queue: 8q2t-in-q1
…
```

FYI

Monitoring Queuing

Payload-VDC

```
N7k# show policy-map interface ethernet 1/1 type queuing | begin output
  Service-policy (queuing) output:   PM_QUEUE_10G-40G-100G_OUT
    policy statistics status:   enabled

    Class-map (queuing):   1p7q4t-out-pq1 (match-any)
      priority level 1
      queue-limit percent 10
      queue dropped pkts : 0
    Class-map (queuing):   1p7q4t-out-q2 (match-any)
      queue-limit percent 29
      bandwidth remaining percent 30
      queue dropped pkts : 0
…
    Class-map (queuing):   1p7q4t-out-q7 (match-any)
      queue-limit percent 10
      bandwidth remaining percent 10
      queue dropped pkts : 5325325151
    Class-map (queuing):   1p7q4t-out-q-default (match
      queue-limit percent 40
      bandwidth remaining percent 40
      queue dropped pkts : 4235425
```

#1 command for queuing monitoring

Shows queue configuration & drops

Note: these drops appear in 'show interface' output as output errors

FYI

How Queuing is set up in Hardware (1)

Payload-VDC

```
N7k# show queuing interface ethernet 1/1
  Interface Ethernet1/1 TX Queuing strategy: Weighted Round-Robin
  Port QoS is enabled
    Queuing Mode in TX direction: mode-cos
    Transmit queues [type = 1p7q4t]
    Queue Id                        Scheduling   Num of thresholds
    ---------------------------------------------------------------
       1p7q4t-out-q-default         WRR              04
       1p7q4t-out-q2                WRR              04
       1p7q4t-out-q3                WRR              04
       1p7q4t-out-q4                WRR              04
       1p7q4t-out-q5                WRR              04
       1p7q4t-out-q6                WRR              04
       1p7q4t-out-q7                WRR              04
       1p7q4t-out-pq1               Priority         04
…
```

#1 command for queuing monitoring

FYI

How Queuing is set up in Hardware (2)

Payload-VDC

```
...
   Configured WRR
      WRR bandwidth ratios:   40[1p7q4t-out-q-default] 30[1p7q4t-out-q2] 0[1p7q4t-out-q3]
 0[1p7q4t-out-q4] 0[1p7q4t-out-q5] 20[1p7q4t-out-q6] 10[1p7q4t-out-q7]
    WRR configuration read from HW
      WRR bandwidth ratios:   40[1p7q4t-out-q-default] 30[1p7q4t-out-q2] 0[1p7q4t-out-q3]
 0[1p7q4t-out-q4] 0[1p7q4t-out-q5] 20[1p7q4t-out-q6] 10[1p7q4t-out-q7]
     Configured queue-limit ratios
      queue-limit ratios:     40[1p7q4t-out-q-default] 29[1p7q4t-out-q2] *1[1p7q4t-out-q3]
 *1[1p7q4t-out-q4] *1[1p7q4t-out-q5] 10[1p7q4t-out-q6] 10[1p7q4t-out-q7] 10[1p7q4t-out-pq1]
    * means unused queue with mandatory minimum queue-limit
    queue-limit ratios configuration read from HW
      queue-limit ratios:     40[1p7q4t-out-q-default] 29[1p7q4t-out-q2] *1[1p7q4t-out-q3]
 *1[1p7q4t-out-q4] *1[1p7q4t-out-q5] 10[1p7q4t-out-q6] 10[1p7q4t-out-q7] 10[1p7q4t-out-pq1]
...
```

* means unused queue with mandatory minimum queue-limit

FYI

What is the real queue length?

Payload-VDC

```
N7k# show system internal qos queuing config interface ethernet 1/1 | beg HW | beg "Queue Size"
Queue Size Config
interface: Ethernet1/1
Queue: 1p7q4t-out-pq1
Queue Size: 0x1
Size: absolute 0x80000
Size: percent 10
Direction: egress
…
Queue Size Config
interface: Ethernet1/1
Queue: 1p7q4t-out-q7
Queue Size: 0x7
Size: absolute 0x80000
Size: percent 10
Direction: egress

Queue Size Config
interface: Ethernet1/1
Queue: 1p7q4t-out-q-default
Queue Size: 0x8
Size: absolute
Size: percent
Direction: egr
```

524kByte – Queuelimit 10%

524kByte – Queuelimit 10%

2Mbyte – Queuelimit 40%

Hardware may demands some queue space even for unused queues

FYI

Locating Queuing drop in Hardware

Payload-VDC

```
N7k# show policy-map interface e1/1 output | i dropped
      queue dropped pkts : 0
      queue dropped pkts : 0
      queue dropped pkts : 0
      queue dropped pkts : 38235841
```

Drops from port ASICs (packaged information)

```
N7k# show hardware internal statistics module 1 device mac qos port 1 description
Hardware statistics on module 03:
|-------------------------------------------------------------------------|
| Device:Garuda                  Role:MACSECURITY            Mod: 1       |
| Last cleared @ Sun Mar 24 11:45:07 2013
|-------------------------------------------------------------------------|
Instance:0
ID   Name                                          Value              Ports
--   ----                                          -----              -----
5144 igr ibc:  total pkts enqueued in q0           0000000000003532   1 -
5145 igr ibc: bytes enqueued in q0                 0000000005121400   1 -
5336 egr ebc:  total pkts enqueued in q0           0000000000003532   1 -
5337 egr ebc: bytes enqueued in q0                 0000000005050760   1 -
28744 [sw] igr ibc:  total pkts enqueued in q0      0000000000003532   1 -
28745 [sw] igr ibc: bytes enqueued in q0            0000000005121400   1 -
```

Detailed stats from port ASICs (source of information)

Note: The detailed ASIC output does NOT show drops here. Watch for “taildrop” or “drop” In Name

Viewing the per Queue detail Statistics

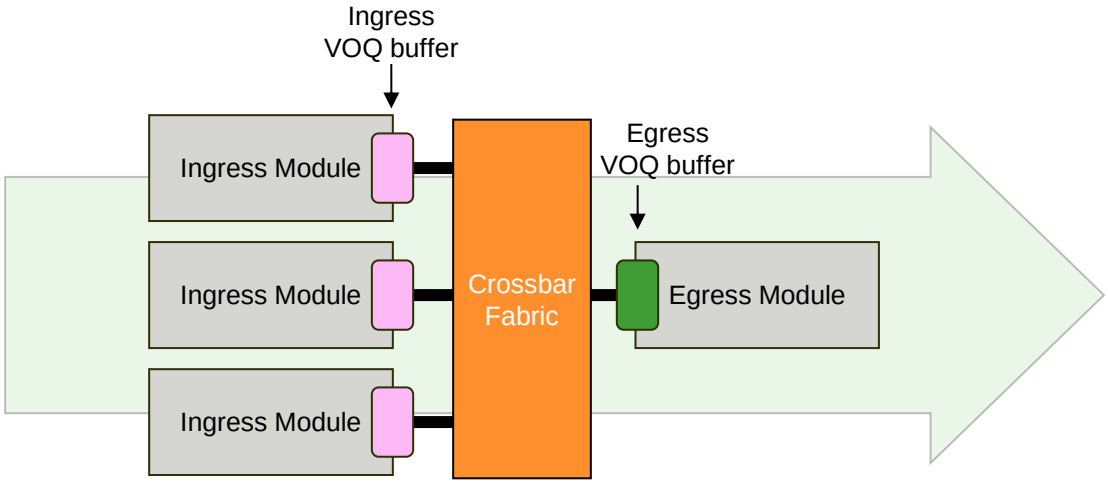
```
N7k# show system internal qos queuing stats interface e1/1
Interface Ethernet1/1 statistics
Receive queues
----------------------------------------
    Queue 8q2t-in-q1
        Total bytes                 0
        Total packets               0
        Current depth in bytes      0
        Min pg drops                NA
        No desc drops               NA
        WRED drops                  NA
        Taildrop drops              NA
        Dropped total               0

Transmit queues
----------------------------------------
    Queue 1p7q4t-out-pq1
        Total bytes                 144
        Total packets               2
        Current depth in bytes      0
        Min pg drops                NA
        No desc drops               NA
        WRED drops                  NA
        Taildrop drops              NA
        Dropped total               0
```

Shows detailed output per different counters

F-Series Queuing & Buffering

Ingress Buffering Model

Two different Buffering Models (F2)



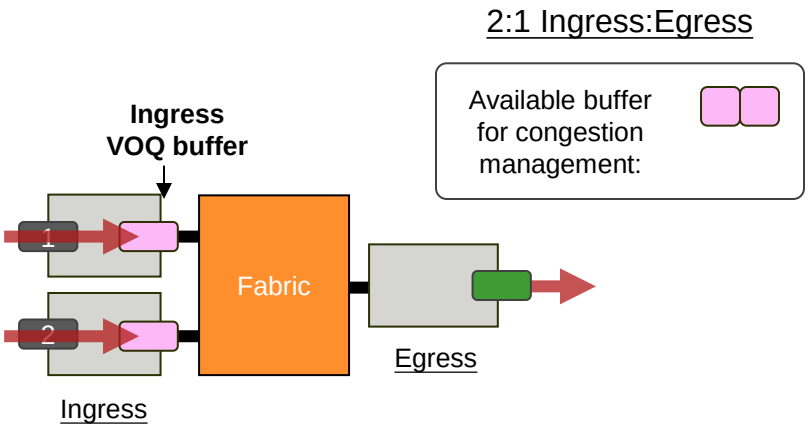
- Ingress VOQ buffer – Manages congestion toward egress destinations (VQIs)
- Egress VOQ buffer – Receives frames from fabric; also buffers multi-destination frames

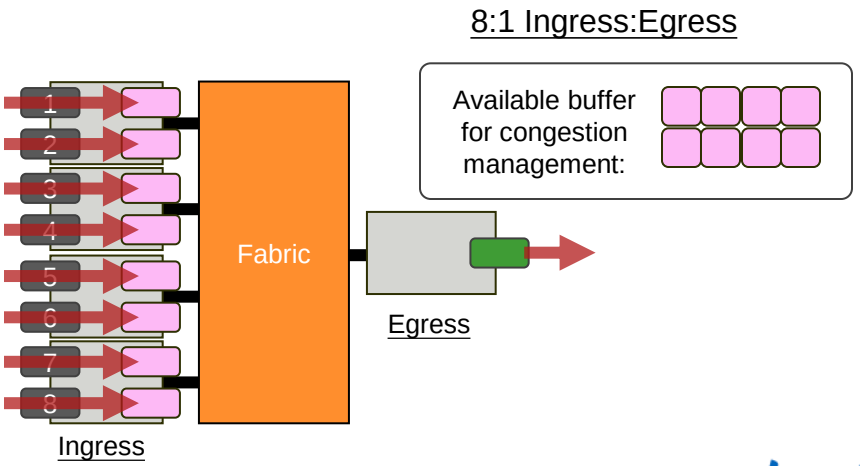
FYI

Distributed Ingress Buffering

Two different Buffering Models (F2)

- Ingress-buffered architecture implements large, distributed buffer pool to absorb congestion
- Absorbs congestion at all ingress ports contributing to congestion, leveraging all per-port ingress buffer
- Excess traffic does not consume fabric bandwidth, only to be dropped at egress port
- Multicast/Broadcast/Unknown Unicast is not arbitrated and will be dropped on Egress





FYI

Queuing and Buffering Details

Nexus 7000 F2-Series I/O Module

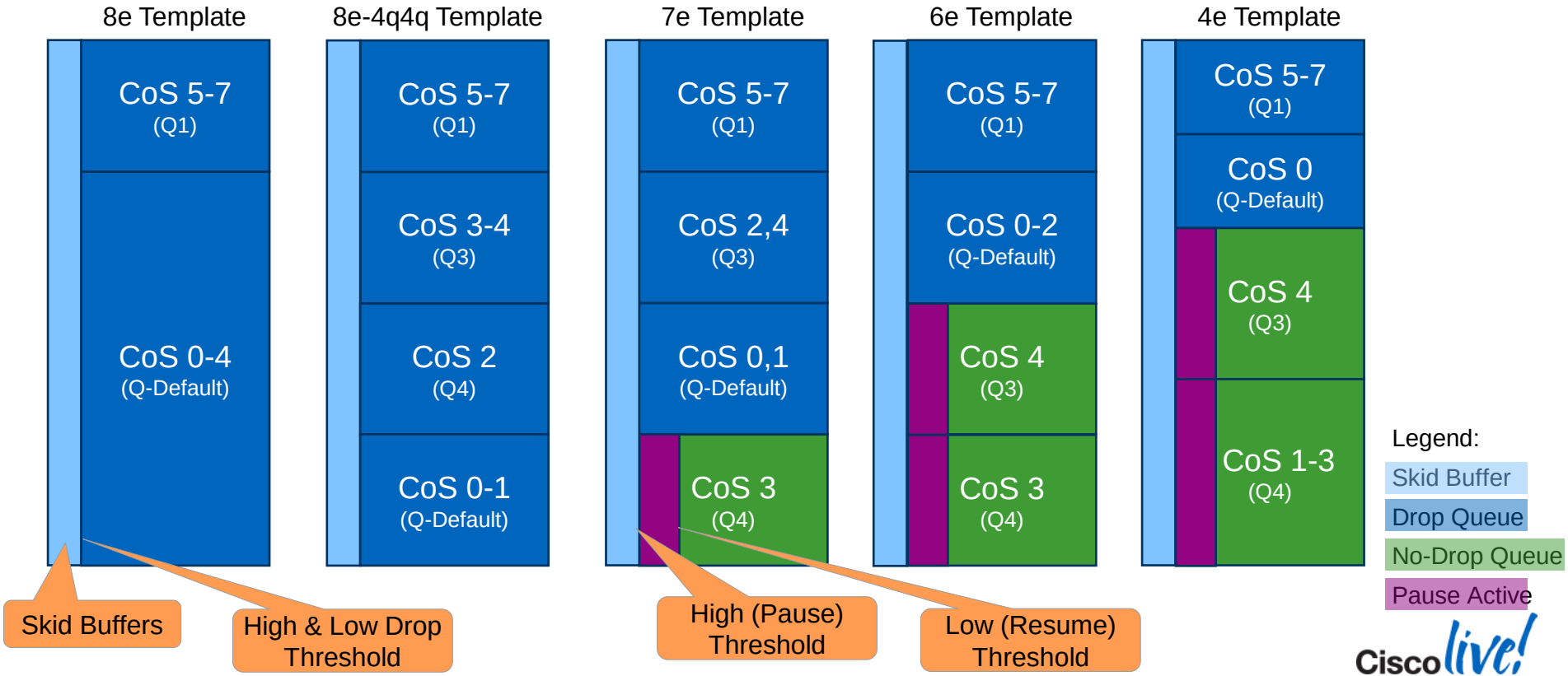
Module Type	Mode	Port Queuing Structure	Per-Port Buffer Capacity	Fabric Queuing Structure	VOQ Buffer Capacity
F2 48-port 10G	10G / 1G	Configurable template-based queuing modes: 4q1t and 2q1t / 1p3q1t, 2p2q1t, and 3p1q1t	Same as VoQ Buffer Capacity	Same as Port Queuing Structure	72MB per I/O Module, 6MB per SoC, 1.5MB per Port
F2E 48-port 10G	10G / 1G SFP				
F2E 48-port 10G	10G / 1G RJ45				

FYI

Ingress Queuing: M-Series vs. F-Series

M I/O Modules ingress queuing	F I/O Modules ingress queuing
Default policy is “default-in-policy”	One default policy per template “default-4q-8e-in-policy”, “default-4q-7e-in-policy”, “default-4q-6e-in-policy”, “default-4q-4e-in-policy”
Default policy attached to each port and port-channel	Default policy attached to “system qos” target.
Fixed number of ingress queues for 1G (4q2t) and 10G (8q2t)	Ingress queues created based on the template type. (2q4t, 4q4t)
WRED can be enabled per ingress port in the queuing policy	WRED not supported per port. It’s a per CoS setting in the network-qos policy
Only one queue can be strict priority	Depending on the template up to of 2 queues can be strict priority – one per drop class
No hierarchical policy support.	Hierarchical policy supported

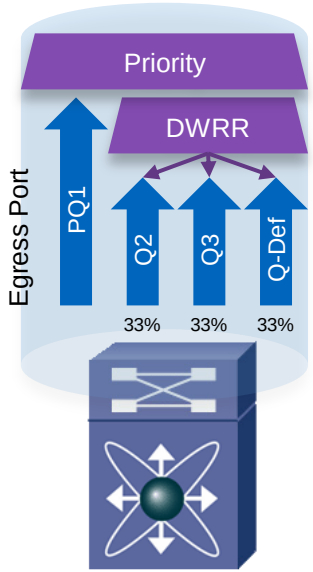
Ingress Queuing – Logical View



Egress Queuing – Logical View

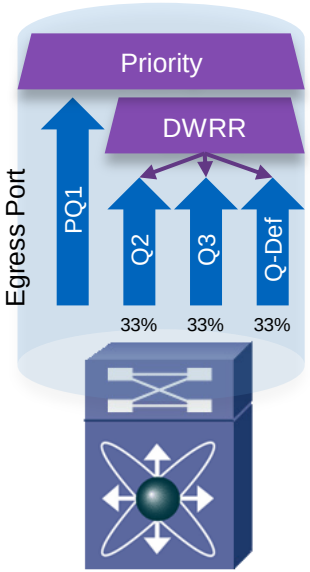
default-4q-8e-out-policy

PQ1	Q2	Q3	Q-Def.
(5,6,7)	(3,4)	(2)	(0,1)



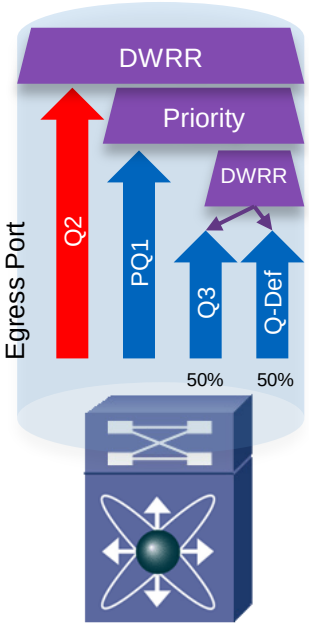
default-4q4q-8e-out-policy

PQ1	Q2	Q3	Q-Def.
(5,6,7)	(3,4)	(2)	(0,1)



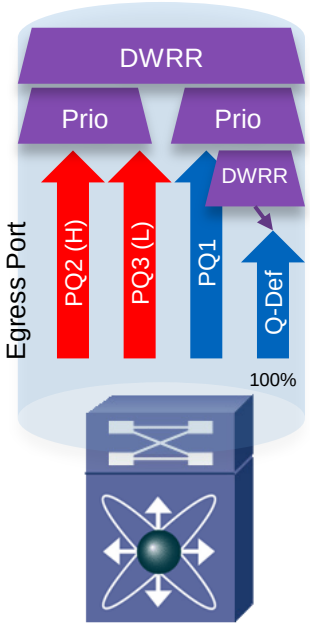
default-4q-7e-out-policy

Q2	PQ1	Q3	Q-Def.
(3)	(5,6,7)	(2,4)	(0,1)



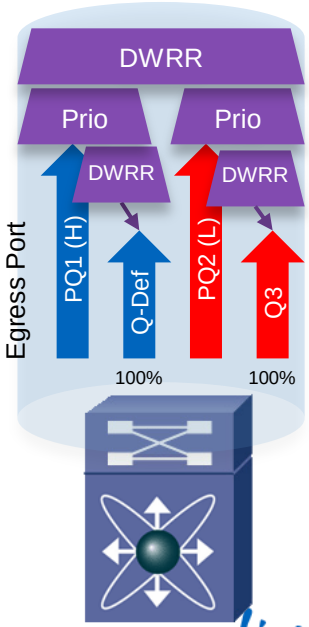
default-4q-6e-out-policy

PQ2.	PQ3	PQ1	Q-Def.
(4)	(3)	(5,6,7)	(0-2)



default-4q-4e-out-policy

PQ1	Q-Def.	PQ2	Q3
(5,6,7)	(0)	(4)	(1,2,3)



red indicates no-drop

QoS Policy Templates

Default network-qos policies

- For ease of configuration and manageability a fixed set of predefined network-qos policies (templates) are provided.
- User can configure policies, but has to conform to one of the templates
- Always attached to “system qos” target in Admin-/Default-VDC

Template	Drop CoS	Priority	No Drop CoS	Priority
default-nq-8e-policy	0,1,2,3,4,5,6,7	5,6,7	-	-
default-nq-8e-4q4q-policy (not supported with F1)	0,1,2,3,4,5,6,7	5,6,7	-	-
default-nq-7e-policy	0,1,2,4,5,6,7	5,6,7	3	-
default-nq-6e-policy	0,1,2,5,6,7	5,6,7	3,4	4
default-nq-4e-policy	0,5,6,7	5,6,7	1,2,3,4	4

Ingress/Egress Queuing Class-Maps

Admin- / Default-VDC only (F)

Type: Queuing

- class-map type queuing – Configures COS/DSCP to Queue mapping
- Queuing class-map names are static, based on port-type and queue template (default-nq-8e-policy)

F2 Egress Queue Structure

```
N7k-ADMIN(config)# class-map type queuing match-any
  1p3q1t-8e-out-pq1                1p3q4t-out-q2                1p7q4t-out-q4                2q4t-in-q-default            8q2t-in-q4
  1p3q1t-8e-out-q-default          1p3q4t-out-q3                1p7q4t-out-q5                2q4t-in-q1                   8q2t-in-q5
  1p3q1t-8e-out-q2                 1p7q4t-out-pq1               1p7q4t-out-q6                8q2t-in-q-default            8q2t-in-q6
  1p3q1t-8e-out-q3                 1p7q4t-out-q-default         1p7q4t-out-q7                8q2t-in-q1                   8q2t-in-q7
  1p3q4t-out-pq1                   1p7q4t-out-q2                2q4t-8e-in-q-default         8q2t-in-q2
  1p3q4t-out-q-default             1p7q4t-out-q3                2q4t-8e-in-q1                8q2t-in-q3
N7k-ADMIN(config)# class-map type queuing match-any 1p3q1t-8e-out-pq1
N7k-ADMIN(config-cmap-que)# match cos 7
N7k-ADMIN(config-cmap-que)#
```

Admin-/Default-VDC only

F2 Ingress Queue Structure

Note: Queue-Template is included

Network-QoS Configuration

Type: Network-QoS

- Copy (cloning) the 8e template and edit required parameters

```
N7k-ADMIN# qos copy policy-map type network-qos default-nq-8e-policy prefix my-
N7k-ADMIN(config)# policy-map type network-qos my-nq-8e
N7k-ADMIN(config-pmap-nqos)# class type network-qos c-nq-8e
N7k-ADMIN(config-pmap-nqos-c)# mtu 9216
```

- Activate the newly created policy by attaching it to system target.

```
N7k-ADMIN(config)# system qos
N7k-ADMIN(config-pmap-nqos)# service policy type network-qos my-default-nq-8e
```

- All the 8 CoS values have to be matched in the policy either by an explicit class-map or the class-default.

Note: Changing network-qos policy is a disruptive operation and can cause traffic drops on any/all ports across VDCs.

Network-QoS and Default Queuing (global)

Type: Network-QoS

- default-nq-8e-policy is default network-qos policy and attached to system qos in Admin-/Default-VDC
- The system queuing policy applied by default can be overridden on a per port basis.

```
N7k# show policy-map type queuing | beg default 4q-8e

policy-map type queuing default-4q-8e-in-policy
    class type queuing 2q4t-8e-in-q1
      queue-limit percent 10
      bandwidth percent 50
    class type queuing 2q4t-8e-in-q-default
      queue-limit percent 90
      bandwidth percent 50

policy-map type queuing default-4q-8e-out-policy
    class type queuing 1p3q1t-8e-out-pq1
      priority level 1
    class type queuing 1p3q1t-8e-out-q2
      bandwidth remaining percent 33
    class type queuing 1p3q1t-8e-out-q3
      bandwidth remaining percent 33
    class type queuing 1p3q1t-8e-out-q-default
      bandwidth remaining percent 33
```

Note: show policy-map system does display similar output

Policy-Map Type Queuing: Capabilities

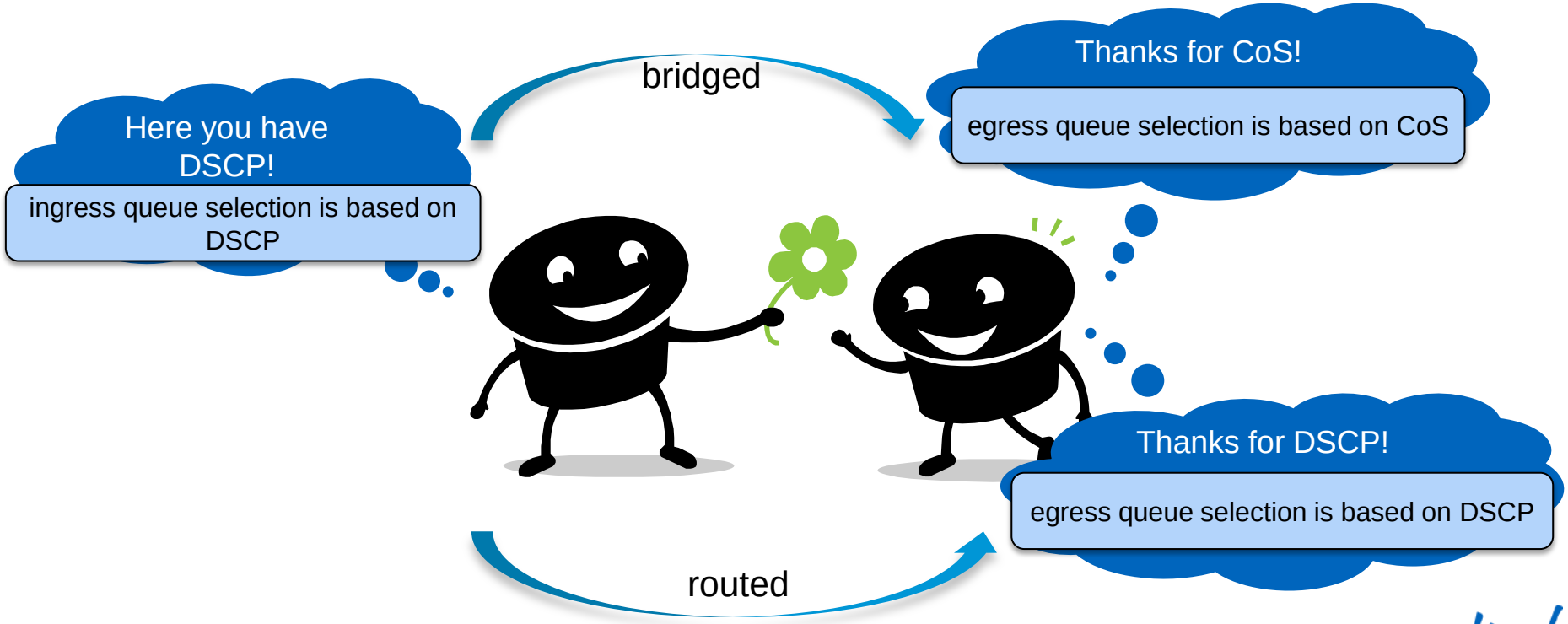
Type: Queuing

Module Type	Priority Queue (priority)	DWRR Scheduler (bandwidth)	Shaper (shape)	Buffer (queue-limit)	Congestion Management (network-qos)	
					WRED Thresholds	Taildrop Thresholds
F2 48-port 10G (SFP)	✓[1]	✓[1]	✓	✓[2]	✗	✗
F2E 48-port 10G (SFP)	✓[1]	✓[1]	✓	✓[2]	✓[4,5]	✓[4,5]
F2E 48-port 10G (RJ-45)	✓[1]	✓[1]	✓	✓[2]	✓[4,5]	✓[4,5]

[1] egress Direction only
[2] ingress Direction only
[3] only on individual Queue
[4] mesh Optimized
[5] burst Optimized

DSCP to Queue Mapping

Default Rule on F2 I/O Module for bridged and routed traffic



CoS to Queue Mapping - Example

Case Study – F2 I/O Module

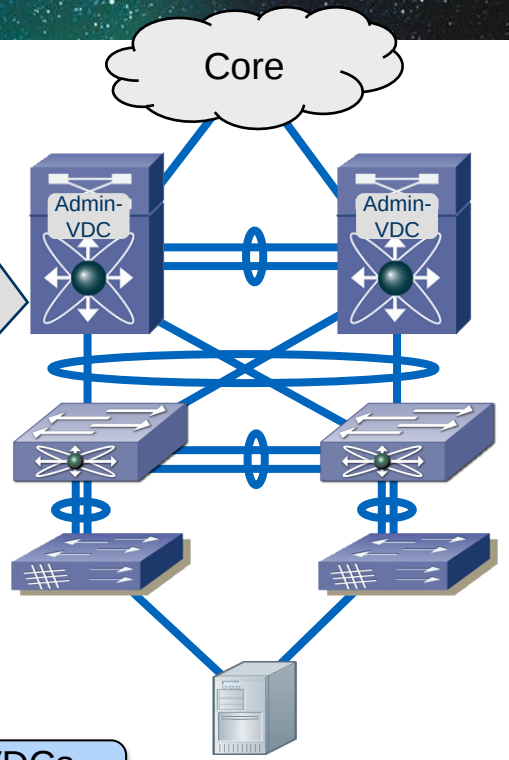
Application	CoS	Queuing (Scheduling)	Queue-Limit (Buffer)	Queue (default-nq-8e-in-policy)	Character
Best Effort	0	BW remaining 50%	50%	2q4t-8e-in-q-default / 1p3q1t-8e-out-q-default	High Volume / Less Important
Low Priority / Scavanger	1				Not Important
vMotion / Live Migration	2	BW remaining 20%	10%	2q4t-8e-in-q-default / 1p3q1t-8e-out-q3	Medium Volume / Important
Multimedia	4	BW remaining 30%	29%	2q4t-8e-in-q-default / 1p3q1t-8e-out-q2	Medium Volume Very Important
Strict Priority	5	-	10%	2q4t-8e-in-q1 / 1p3q1t-8e-out-pq1	Low Volume / Important / Delay Sensitive
Network Control	6/7				Low Volume / Very important

CoS to Queue Mapping Configuration

Case Study (Admin- / Default-VDC)

Type: Queuing

```
qos copy policy-map type network-qos default-nq-8e-policy prefix QOS
!
class-map type queuing match-any 1p3q1t-8e-out-pq1
 match cos 5-7
class-map type queuing match-any 1p3q1t-8e-out-q2
 match cos 4
class-map type queuing match-any 1p3q1t-8e-out-q3
 match cos 2
class-map type queuing match-any 1p3q1t-8e-out-q-default
 match cos 0-1
!
class-map type queuing match-any 2q4t-8e-in-q1
 match cos 5-7
class-map type queuing match-any 2q4t-8e-in-q-default
 match cos 0,1,2,4
!
system qos
 service-policy type network-qos QOSdefault-nq-8e
```



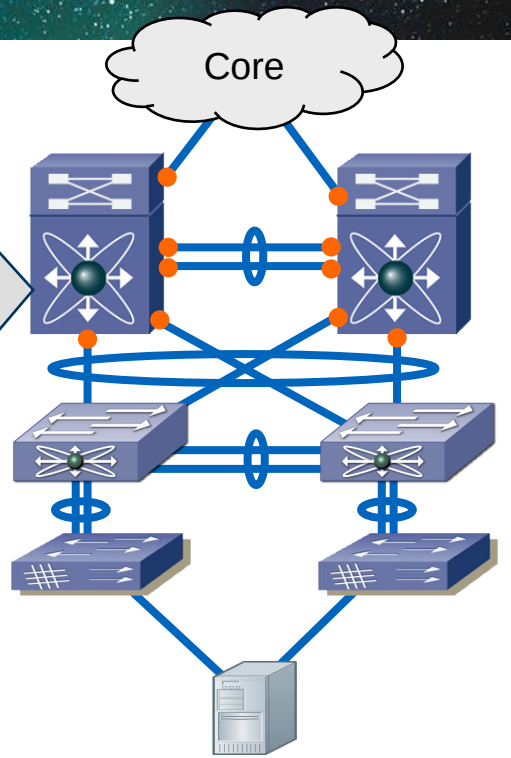
Changes apply to ALL ports of specified type in ALL VDCs
Changes are traffic disruptive for ports of specified type

Queuing Configuration (Ingress/Egress)

Case Study (Payload-VDC)

Type: Queuing

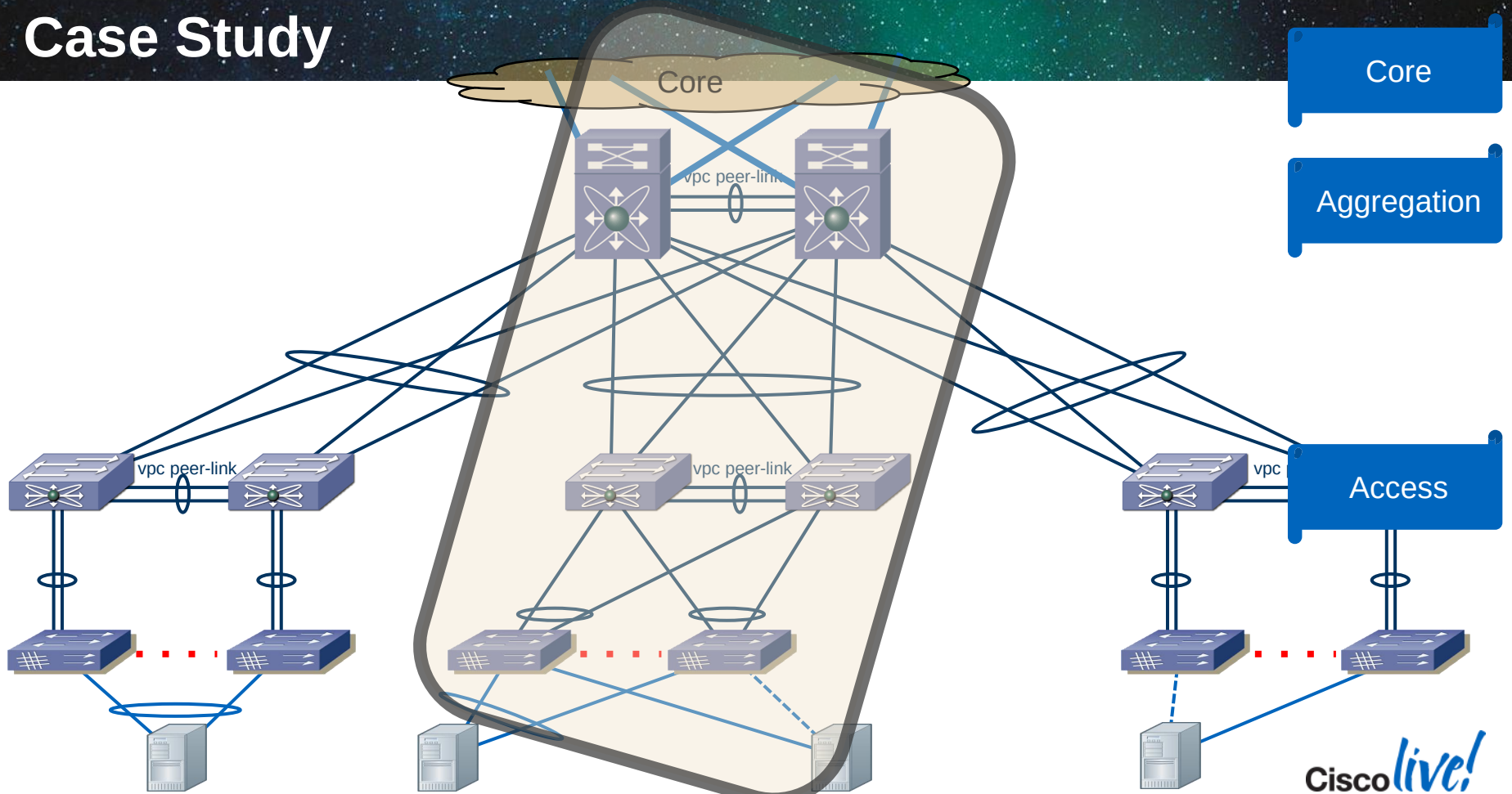
```
qos copy policy-map type queuing default-4q-8e-in-policy prefix QOS
qos copy policy-map type queuing default-4q-8e-out-policy prefix QOS
!
policy-map type queuing QOS4q-8e-in-policy
  class type queuing 2q4t-8e-in-q1
    queue-limit percent 10
  class type queuing 2q4t-8e-in-q-default
    queue-limit percent 90
policy-map type queuing QOS4q-8e-out-policy
  class type queuing 1p3q1t-8e-out-pq1
    priority level 1
  class type queuing 1p3q1t-8e-out-q2
    bandwidth remaining percent 30
  class type queuing 1p3q1t-8e-out-q3
    bandwidth remaining percent 20
  class type queuing 1p3q1t-8e-out-q-default
    bandwidth remaining percent 50
!
interface Ethernet1/1
  service-policy type queuing input QOS4q-8e-in-policy
  service-policy type queuing output QOS4q-8e-out-policy
```



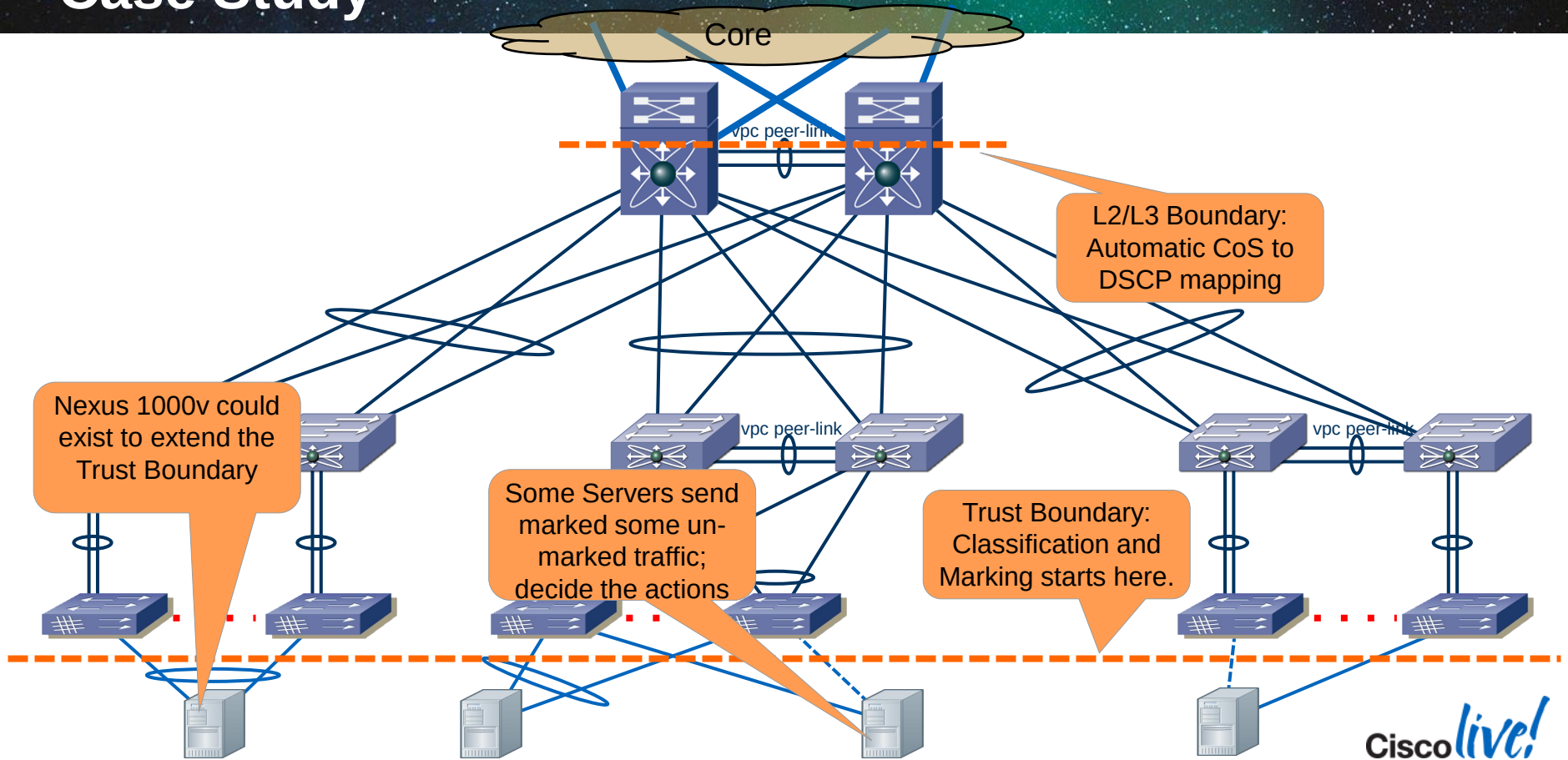



Case Study – Wrap Up

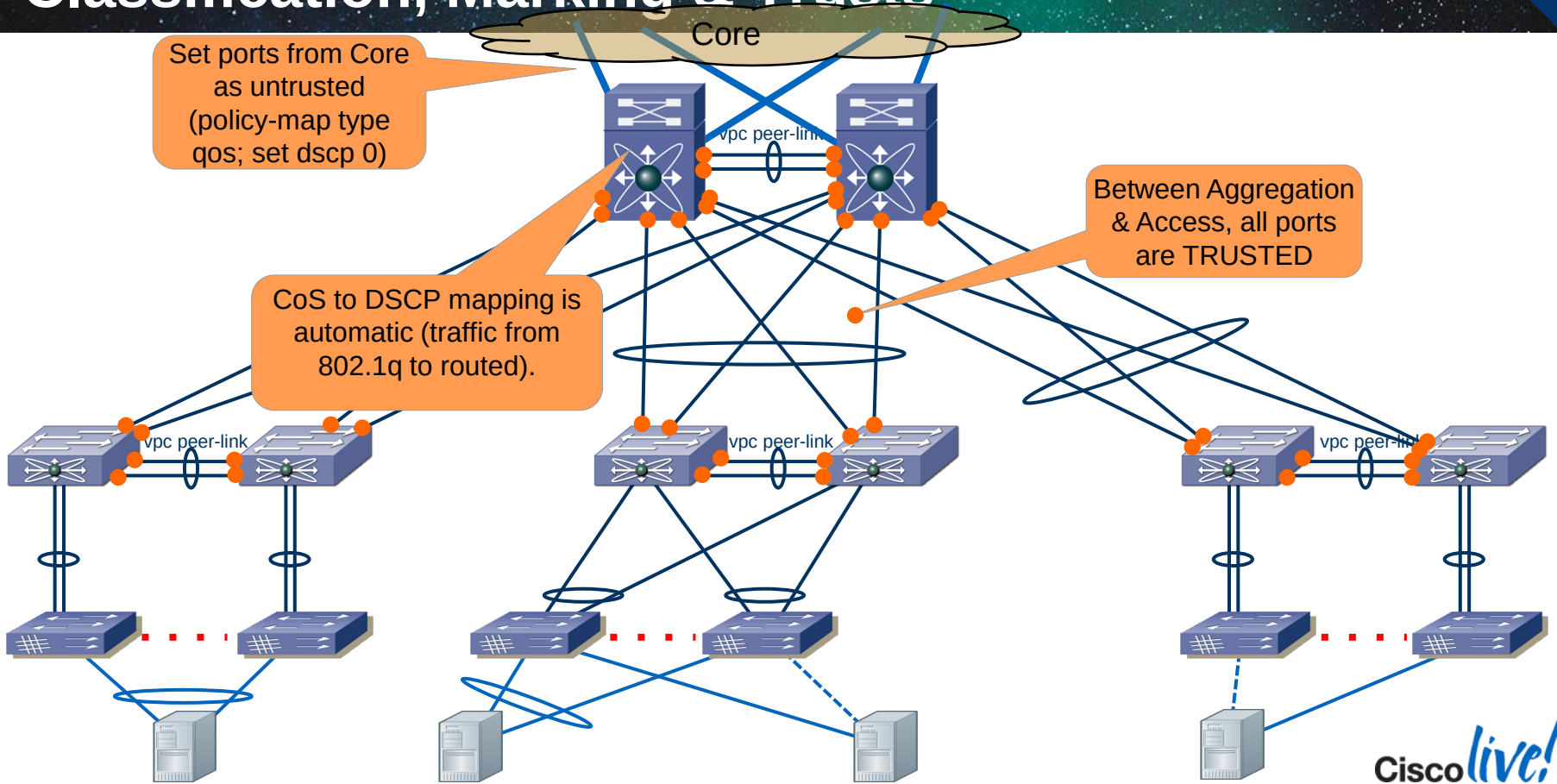
Case Study



Case Study

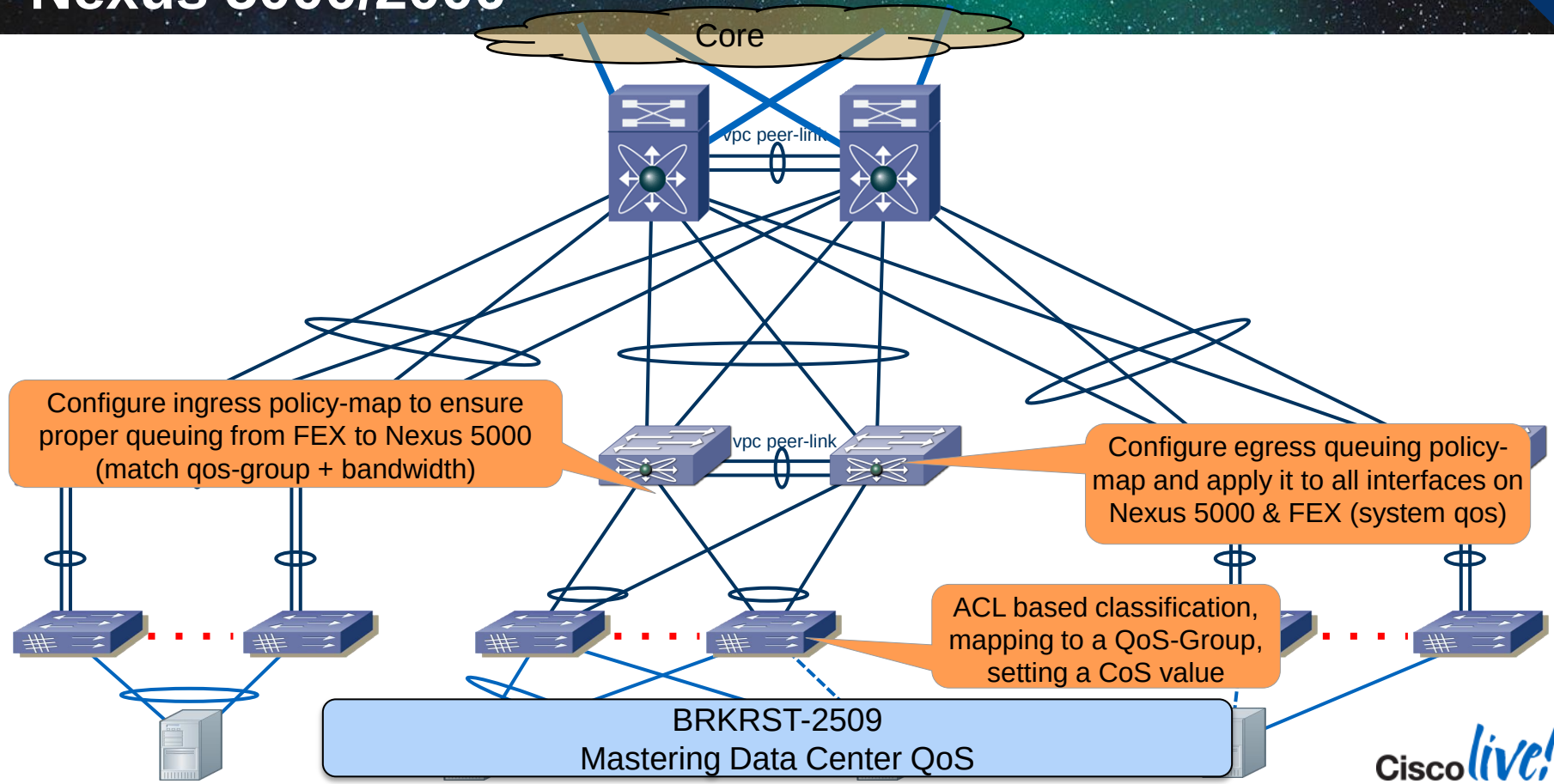


Classification, Marking & Trusts

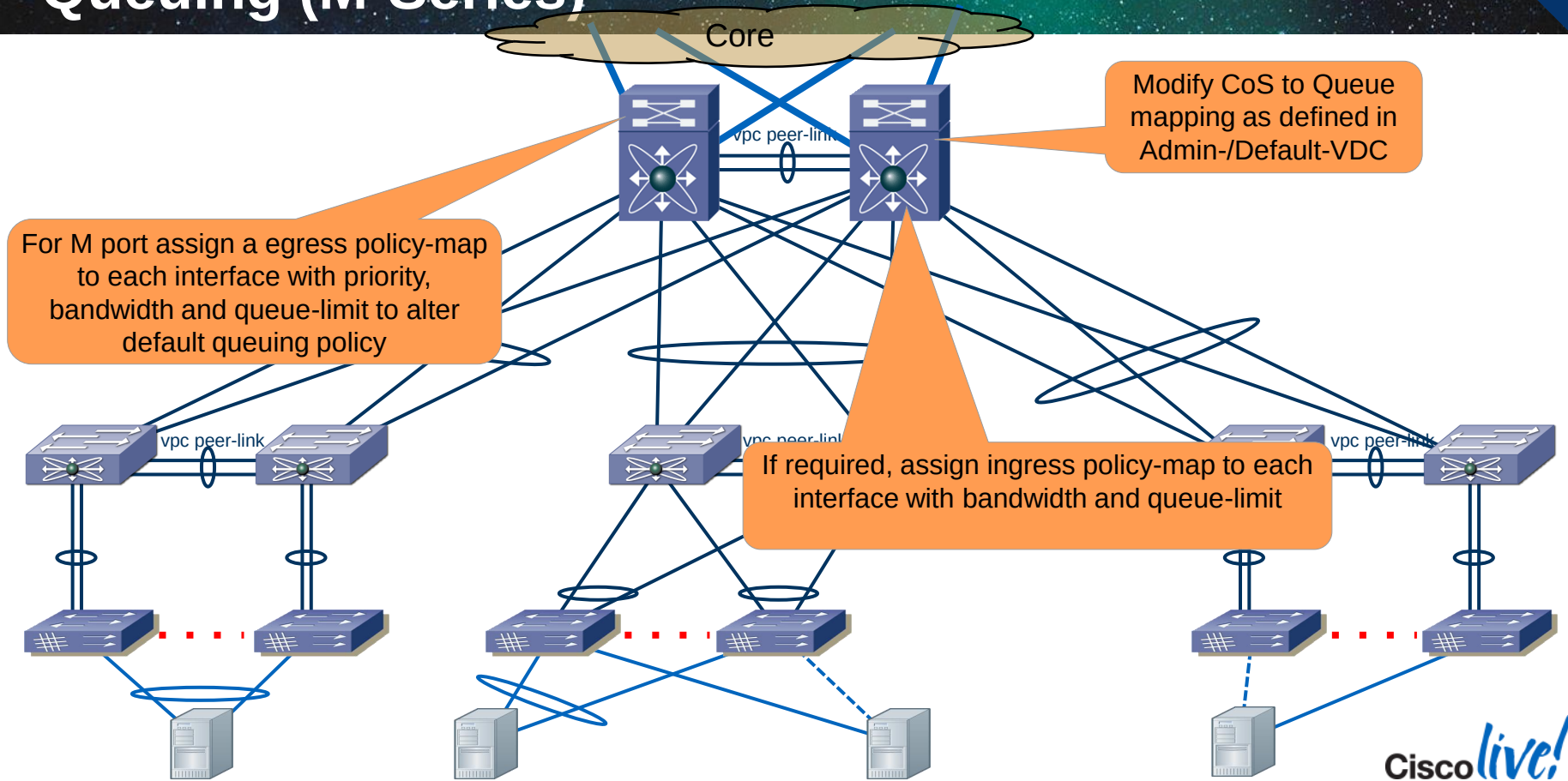


Nexus 5000/2000

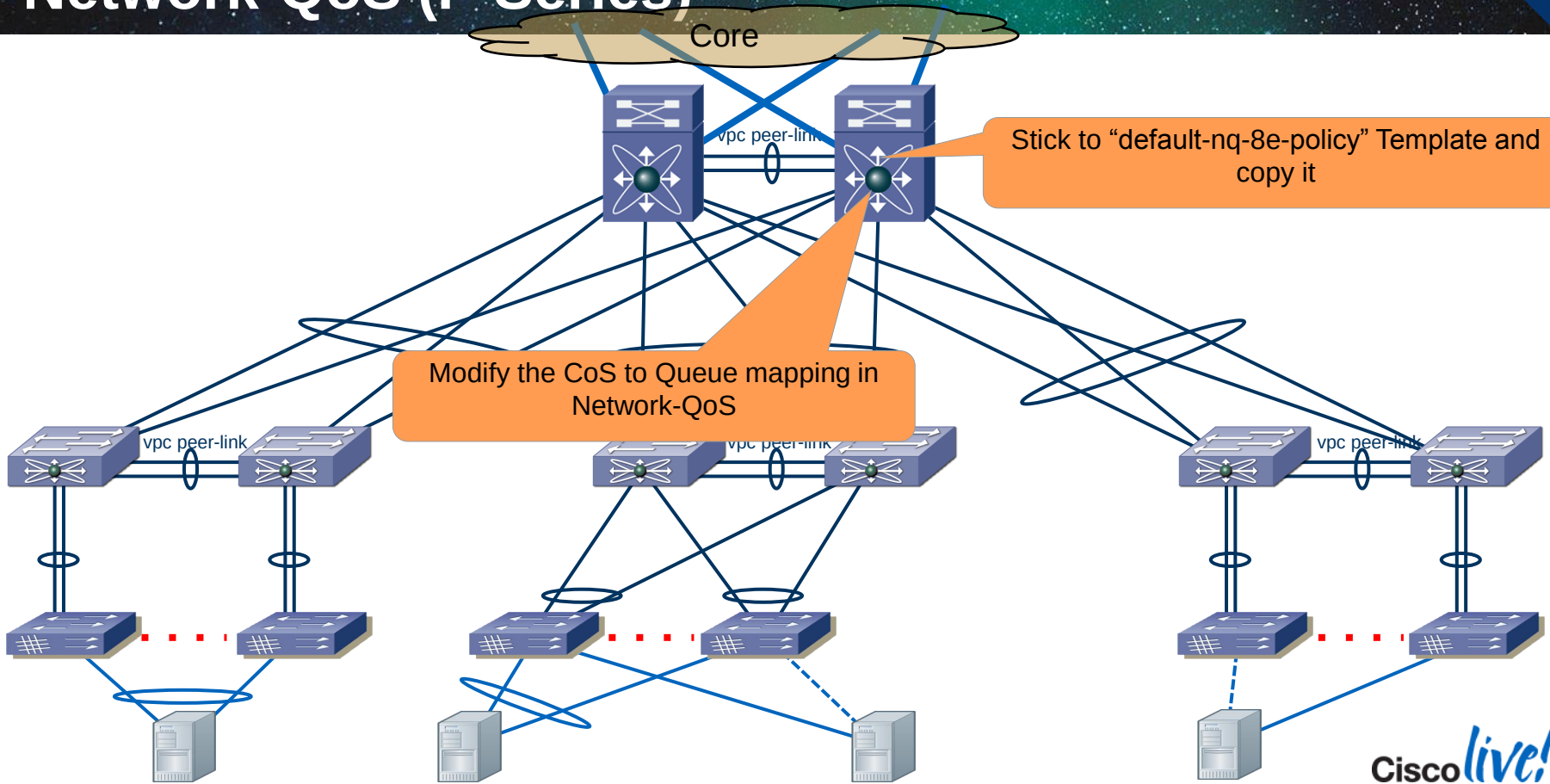
Type: Queuing



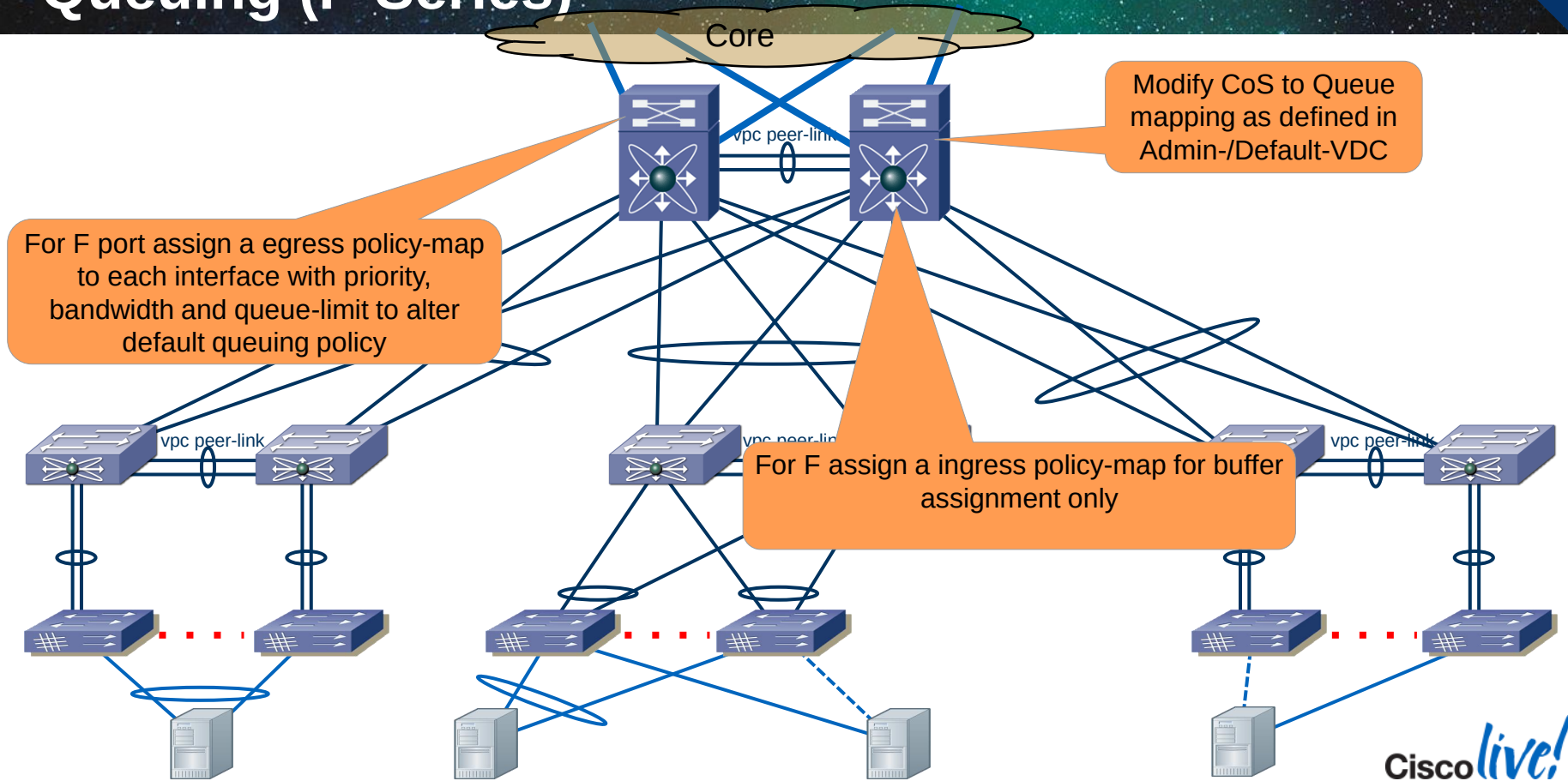
Queuing (M-Series)



Network-QoS (F-Series)



Queuing (F-Series)

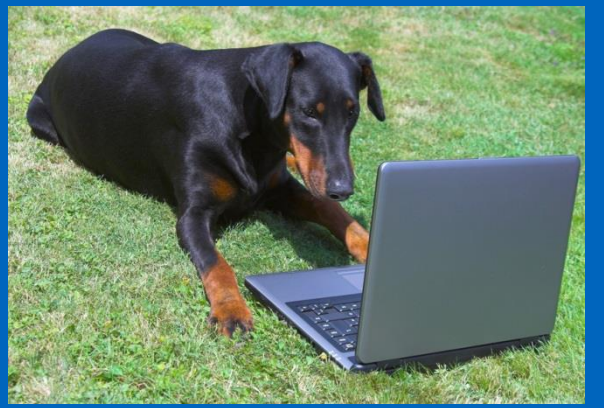


Why QoS in the Data Center?

Assign Color to Traffic

Manage Congestion

Protect Network Resources



maximize throughput, manage congestion!

Where could QoS help YOU

to MAXIMIZE your Throughput

in the Data Center?”

Complete Your Online Session Evaluation

- Give us your feedback and you could win fabulous prizes. Winners announced daily.
- Receive 20 Cisco Daily Challenge points for each session evaluation you complete.
- Complete your session evaluation online now through either the mobile app or internet kiosk stations.

Maximize your Cisco Live experience with your free Cisco Live 365 account. Download session PDFs, view sessions on-demand and participate in live activities throughout the year. Click the Enter Cisco Live 365 button in your Cisco Live portal to log in.

Note: This slide is now a Layout choice


CISCO
TM




Appendix

Agenda

- Nexus 5000/6000 – Configuration & Operation
  - Classification & Marking
  - Queuing & Buffering

Application Classification & Marking - Example

Case Study

Application	DSCP Differentiated Services Codepoint	PHB Per Hop Behavior	CoS Class of Service
Best Effort	0	BE	0
Low Priority / Scavanger	8	CS1	1
vMotion / Live Migration	-*		2
Reserved (FCoE)	-*		3
Multimedia	34	AF41	4
Strict Priority	46	EF	5
Network Control	48	CS6	6

*Layer-2 only or Non-IP Traffic; will only exist within Data Center Layer-2 Fabric

Classification & Marking (1)

The Nexus 5000/6000 Map to Marked Packets

Type: Network-QoS QoS

1
```
ip access-list ACL_QOS_LOWPRIO
  10 permit tcp any any eq www
  20 permit tcp any eq www any
```

2
```
class-map type qos match-any CM_QOS_LOWPRIO_COS1
  match access-group name ACL_QOS_LOWPRIO
```

3
```
policy-map type qos PM_QOS_MARK_COS_IN
  class CM_QOS_LOWPRIO_COS1
    set qos-group 2
```

4
```
system qos
  service-policy type qos input PM_QOS_MARK_COS_IN
```

Note: Point of Policy attachament could also be VLAN or Interface

Classification & Marking (2)

The Nexus 5000/6000 Map to Marked Packets

5
```
class-map type network-qos CM_N-QOS_MATCH_QG2_COS1
  match qos-group 2
```

6
```
policy-map type network-qos PM_N-QOS_SYSTEM
  class type network-qos CM_N-QOS_MATCH_QG2_COS1
    set cos 1
```

7
```
system qos
  service-policy type network-qos PM_N-QOS_SYSTEM
```

FYI

Classification & Marking – Case Study Config (1)

Type: Network-QoS QoS

```
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 match-any CM_QOS_LOWPRIO_COS1
  match access-group name ACL_QOS_LOWPRIO
!
class-map type qos match-any CM_QOS_VMOTION_COS2
  match access-group name ACL_QOS_VMOTION
!
class-map type qos match-any CM_QOS_MULTIMEDIA_COS4
  match access-group name ACL_QOS_MULTIMEDIA
!
class-map type qos match-any 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
  class CM_QOS_MULTIMEDIA_COS4
    set qos-group 4
  class CM_QOS_VMOTION_COS2
    set qos-group 3
  class CM_QOS_LOWPRIO_COS1
    set qos-group 2
!
system qos
  service-policy type qos input PM_QOS_MARK_COS_IN
```

The Nexus 5000/6000 is not the Trust Boundary, class-map type qos matches can also be a CoS value

Note: QoS-Group has no direct relation with CoS value

FYI

Classification & Marking – Case Study Config (2)

Type: Network-QoS QoS

```
class-map type network-qos CM_N-QOS_MATCH_QG2_COS1
  match qos-group 2
class-map type network-qos CM_N-QOS_MATCH_QG3_COS2
  match qos-group 3
class-map type network-qos CM_N-QOS_MATCH_QG4_COS4
  match qos-group 4
class-map type network-qos CM_N-QOS_MATCH_QG5_COS5
  match qos-group 5
```

```
policy-map type network-qos PM_N-QOS_SYSTEM
  class type network-qos CM_N-QOS_MATCH_QG2_COS1
    set cos 1
  class type network-qos CM_N-QOS_MATCH_QG3_COS2
    set cos 2
  class type network-qos CM_N-QOS_MATCH_QG4_COS4
    set cos 4
  class type network-qos CM_N-QOS_MATCH_QG5_COS5
    set cos 5
    queue-limit 20480 bytes
!
system qos
  service-policy type network-qos PM_N-QOS_SYSTEM
```

QoS-Group # is mapping between Slide 1 & Slide 2

FYI

Verify Classification & Marking Configuration

Payload-VDC

```
N5k# show policy-map type qos PM_QOS_MARK_COS_IN
…
policy-map type qos PM_QOS_MARK_COS_IN
    class  CM_QOS_LOWPRIO_COS1
      set cos 1
…
N5k# show class-map type qos
…
    class-map type qos match-any CM_QOS_LOWPRIO_COS1
      match access-group name ACL_QOS_LOWPRIO
…
N5k# show access-lists ACL_QOS_LOWPRIO

IP access list ACL_QOS_LOWPRIO
        10 permit ip 1.0.111.11/32 1.0.222.22/32
…
N5k# show policy-map type network-qos
…
N5k# show policy-map type network-qos
…
```

FYI

What QOS policies are applied?

Payload-VDC

```
N5k# show policy-map interface brief

 Interface/VLAN [Status]:INP QOS       OUT QOS       INP QUE       OUT QUE
================================================================================
…
Ethernet1/1     [Active]:PM_QOS_MARK_C                default-in-po default-out-p
…
N5k# show policy-map interface ethernet 1/1 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

Ethernet1/1

  Service-policy (qos) input:   PM_QOS_MARK_COS_IN
    policy statistics status:   disabled

    Class-map (qos):   CM_QOS_LOWPRIO_COS1 (match-any)
      Match: access-group ACL_QOS_LOWPRIO
      set qos-group 2
…
```

QoS policies

Queuing policies

#1 command for QOS monitoring

Shows packets matching each class & policer stats

FYI

What is QOS HW resources utilization?

Payload-VDC

```
N5k#  show platform afm info tcam 0 region qos
qos tcam TCAM configuration for asic id 0:
[ vacl tcam]: range     0 - 2047
[ifacl tcam]: range  2048 - 3199
[  qos tcam]: range  3200 - 3647 *
[rbacl tcam]: range  3648 - 3775
[ span tcam]: range  3776 - 3839
[  sup tcam]: range  3840 - 3967

    TCAM [qos tcam]: [v:1, size:448, start:3200 end:3647]
    In use tcam entries: 11
        3200-3206,3644-3647

N5k#
```

ACL classification rules share 4K TCAM space with other features, such as PACL, VACL

448 TCAM entries allocated for QoS classification

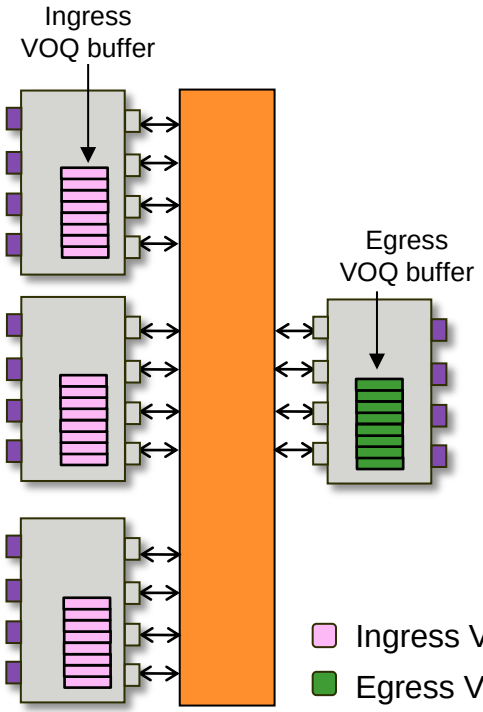
Agenda

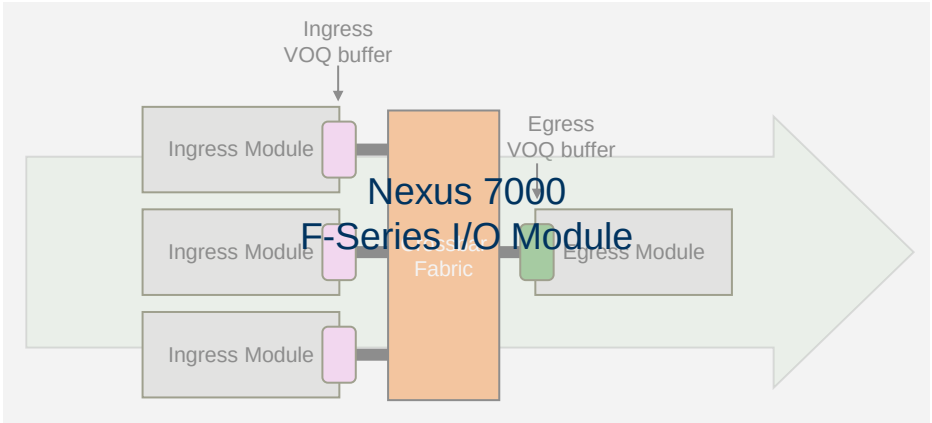
- Nexus 5000/6000 – Configuration & Operation
  - Classification & Marking
  - Queuing & Buffering

FYI

Ingress Buffering Model

Nexus 5000/6000 compared to Nexus 7000 F-Series I/O Module





- Ingress VOQ buffer – Manages congestion toward egress destinations (VQIs)
- Egress VOQ buffer – Receives frames from fabric; also buffers multi-destination frames

FYI

Key Concepts – Common Points

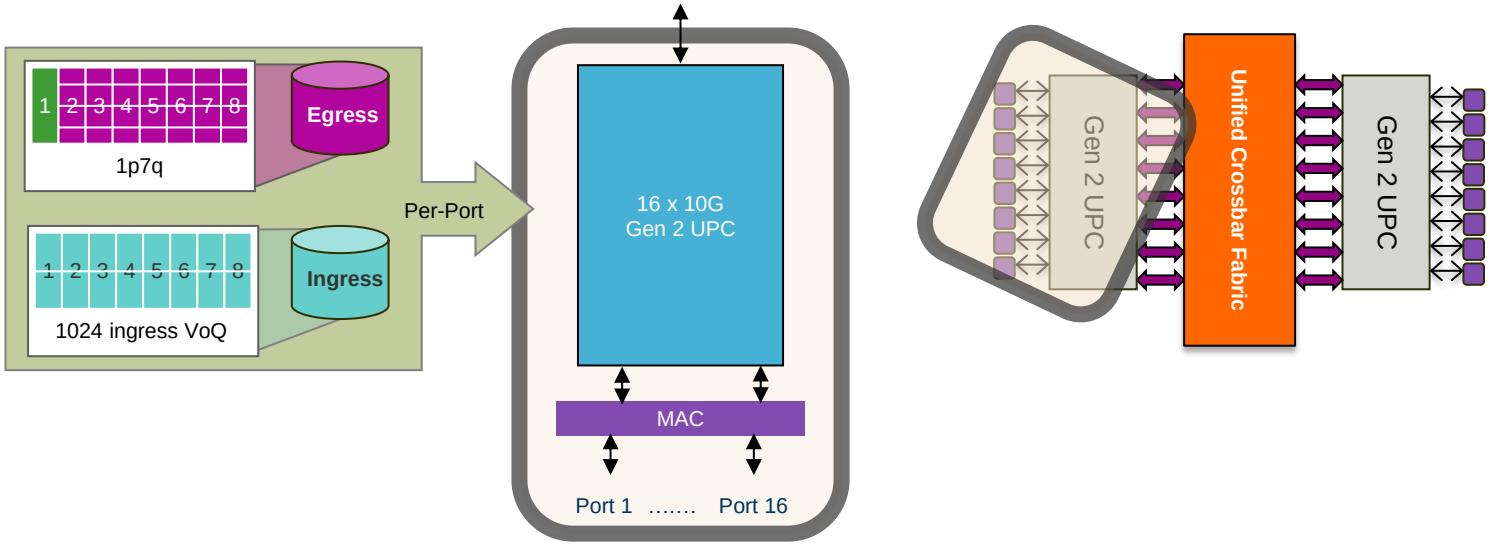
Nexus 7000 compared to Nexus 5000/6000 QoS

- Nexus 5000/6000 & Nexus 7000 F-Series I/O Modules are sharing the Ingress Buffer Model
- Ingress buffering and queuing (as defined by ingress queuing policy) occurs at VOQ of each ingress port
  - Ingress VOQ buffers are primary congestion-management point for arbitrated traffic
- Egress scheduling (as defined by egress queuing policy) enforced by egress port
  - Egress scheduling dictates manner in which egress port bandwidth made available at ingress
  - Per-port, per-priority grants from arbiter control which ingress frames reach egress port

FYI

Nexus 5000/6000 Buffering

Ingress buffered architecture on Nexus 5000/6000



Note: We are showing the Nexus 5000 Diagram. Nexus 6000 is similar from a concept but has significant differences in detail.

FYI

Queuing and Buffering Details

Nexus 5000 / 6000

Module Type	Mode	Port Queuing Structure	Per-Port Buffer Capacity	Fabric Queuing Structure	VOQ Buffer Capacity
Nexus 5500	10G / 1G	6q1t / 1p6q0t	128kb	-	512kb
Nexus 6000	40G / 10G		9MB	-	16MB

FYI

Buffering and Scheduling Considerations

- Nexus 5000 & 6000 are using two stages to achieve Buffering & Scheduling
- Buffering is achieved via policy-map type network-qos where queue-limit can be configured
- Scheduling is configured with policy-map type queuing where priority & bandwidth for DWRR is defined
  - Queuing Policy can be attached in both direction, ingress and egress
  - Egress controls how the traffic is leaving the Switch
  - Ingress controls how a attached DCBX capable Device is sending
- Queuing will be configured with policy-map type queueing and does NOT interfere with policy-map type qos

```
N5k# show interface ethernet 1/1 capabilities | i Model|QOS
  Model:                N5K-C5548UP-SUP
  QOS scheduling:       rx-(6q1t),tx-(1p6q0t)
```

FYI

Type: Queuing

Class-Map Type Queuing and Default Queuing

- Predefined policy-map type queuing exist for input and output on all ports.
- default-policy & FCoE-default-policy created; only Default-Policy is attached.
- They can NOT be modified

```
N5k(config)# policy-map type queuing default-in-policy
ERROR: Changes to default policy-map(s) not permitted
```

```
N5k# show policy-map type queuing


  Type queuing policy-maps
  ========================

  policy-map type queuing default-in-policy
    class type queuing class-default
      bandwidth percent 100
  policy-map type queuing default-out-policy
    class type queuing class-default
      bandwidth percent 100
  policy-map type queuing fcoe-default-in-policy
    class type queuing class-fcoe
      bandwidth percent 50
    class type queuing class-default
      bandwidth percent 50
  policy-map type queuing fcoe-default-out-policy
    class type queuing class-fcoe
      bandwidth percent 50
    class type queuing class-default
      bandwidth percent 50
```

CoS to Queue Mapping - Example

Case Study – Nexus 5000/6000

Application	CoS	Queuing (Scheduling)	Queue-Limit (Buffer)	Queue (6q1t / 1p6q0t)	Character
Best Effort	0	BW percent 30%	remaining (226kByte)	qos-group 0 (default)	High Volume / Less Important
Low Priority / Scavanger	1	BW percent 10%	22kByte	qos-group 2	Not Important
vMotion / Live Migration	2	BW percent 20%	41kByte	qos-group 3	Medium Volume / Important
Multimedia	4	BW percent 30%	41kByte	qos-group 4	Medium Volume Very Important
Strict Priority	5	BW percent 10%	20kByte (min)	qos-group5 / priority	Low Volume / Important / Delay Sensitive
Network Control	6				Low Volume / Very important

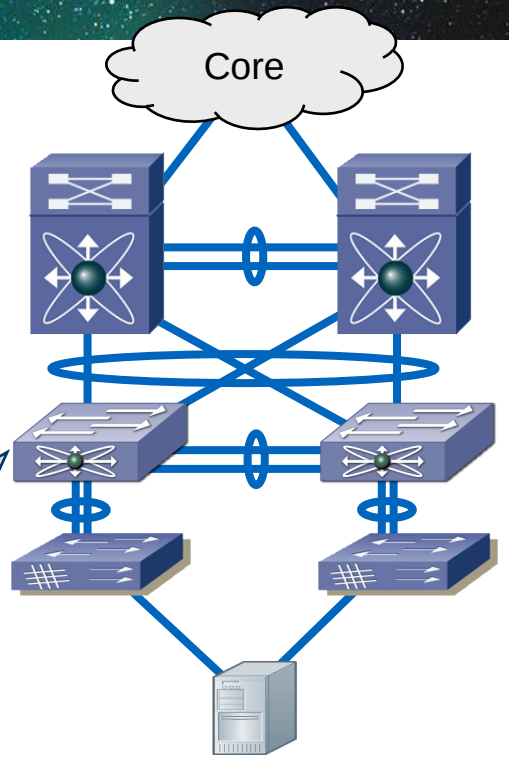
QoS-Group 1 is “reserved” for FCoE

Queue-Limit (Buffer) Configuration

Case Study

Type: Network-QoS

```
policy-map type network-qos PM_N-QOS_SYSTEM
  class type network-qos CM_N-QOS_MATCH_QG2_COS1
    set cos 1
  class type network-qos CM_N-QOS_MATCH_QG3_COS2
    set cos 2
    queue-limit 40960 bytes
  class type network-qos CM_N-QOS_MATCH_QG4_COS4
    set cos 4
    queue-limit 40960 bytes
  class type network-qos CM_N-QOS_MATCH_QG5_COS5
    set cos 5
    queue-limit 20480 bytes
  class type network-qos class-default
    multicast-optimize
!
system qos
  service-policy type network-qos PM_N-QOS_SYSTEM
```

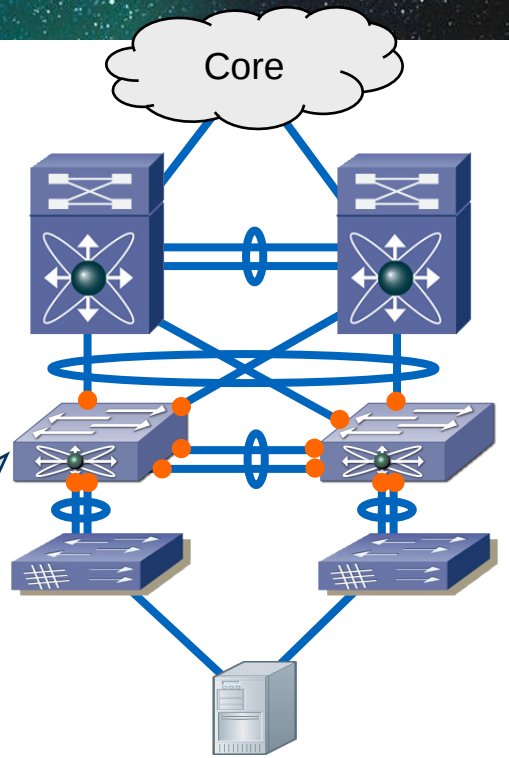


Queuing Configuration (Egress)

Case Study

Type: Queuing

```
class-map type queuing CM_Q_MATCH_QG2_COS1
  match qos-group 2
class-map type queuing CM_Q_MATCH_QG3_COS2
  match qos-group 3
class-map type queuing CM_Q_MATCH_QG4_COS4
  match qos-group 4
class-map type queuing CM_Q_MATCH_QG5_COS5
  match qos-group 5
!
policy-map type queuing PM_QUEUING_SYSTEM_OUT
  class type queuing CM_Q_MATCH_QG2_COS1
    bandwidth percent 10
  class type queuing CM_Q_MATCH_QG3_COS2
    bandwidth percent 20
  class type queuing CM_Q_MATCH_QG4_COS4
    bandwidth percent 30
  class type queuing CM_Q_MATCH_QG5_COS5
    priority
    bandwidth percent 10
  class type queuing class-default
    bandwidth percent 30
```



FYI

Monitoring Queuing

```
N5k# show queuing interface ethernet 1/1
Ethernet1/1 queuing information:
  TX Queuing
    qos-group  sched-type  oper-bandwidth
        0       WRR             30
        2       WRR             10
        3       WRR             20
        4       WRR             30
        5       priority        10
…
Ethernet1/1 queuing information:
  TX Queuing
…
  RX Queuing
    qos-group 0
    q-size: 225920, HW MTU: 1500 (1500 configured)
    drop-type: drop, xon: 0, xoff: 225920
    Statistics:
        Pkts received over the port             : 4
        Ucast pkts sent to the cross-bar        : 2
        Mcast pkts sent to the cross-bar        : 2
        Ucast pkts received from the cross-bar  : 1423468014
        Pkts sent to the port                   : 992929364
        Pkts discarded on ingress               : 0
        Per-priority-pause status               : Rx (Inactive), Tx (Inactive)
```

#1 command for queuing monitoring

Shows queue configuration & drops

Note: these drops appear in 'show interface' output as output errors

FYI

Verify CoS to QoS-Group Mapping

```
N5K_111# show platform afm info map-tbls

E2Q mapping table
------------------
Ext-cos: 0      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 1      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 2      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 3      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 4      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 5      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 6      -  Qos-Group: 0      -   SetByUser: 0
Ext-cos: 7      -  Qos-Group: 0      -   SetByUser: 0

Q2C mapping table [Flags:: 0x20=is_fcoe 0x10=pfc-cos 0x08=COS_based 0x04=ACL_based 0x02=setcos
0x01=Valid]
------------------
Qos-Group  -  Class  - cos-rw - BaseClass - flags
0          -   3     -   0    -  3         -  0x 1
1          -   3     -   0    -  0         -  0x 0
2          -   4     -   1    -  4         -  0x 3
3          -   5     -   2    -  5         -  0x 3
4          -   6     -   4    -  6         -  0x 3
5          -   7     -   5    -  7         -  0x 3
```

FYI

Verify CoS to QoS-Group Mapping

```
N5K_111# sh platform software qd info global
System info:
cells 1364, sup_hi_cos 7, sup_hi_ext_to_int 7, drop_cos_map 0xff
cos2class-id[7-0]      :  3 3 3 3  3 3 3 3
pfc-cos2class-bmap[7-0]: f8f8f8f8 f8f8f8f8
cls|flag|pause|d_qsize|mtuT|mtuS|icells|ecells|mc|cosM|grpM
  0|0x00|  0  |      0|   0|1500|     0|     0| 0|0x00|0x00
  1|0x07|  0  |      0|   0|9216|   116|    62| 0|0x00|0x00
  2|0x07|  0  |      0|   0|9216|   174|    62| 0|0x80|0x00
  3|0x05|  0  |      0|   0|1500|  1412|     0| 1|0xff|0x01
  4|0x09|  0  |      0|   0|1500|   142|     0| 0|0x00|0x04
  5|0x09|  0  |  40960|   0|1500|   256|     0| 0|0x00|0x08
  6|0x09|  0  |  40960|   0|1500|   256|     0| 0|0x00|0x10
  7|0x09|  0  |  20480|   0|1500|   128|     0| 0|0x00|0x20

Scheduling info
          actual                    configured
cls|priIn|priOut|bwIn|Out   qgrp|priIn|priOut|bwIn|Out
  0|  0  |   0  |  0 |  0     0 |  0  |   0  | 30 | 30
  1|  0  |   0  |  0 |  0     1 |  0  |   0  |  0 |  0
  2|  1  |   1  |  0 |  0     2 |  0  |   0  | 10 | 10
  3|  0  |   0  | 30 | 30     3 |  0  |   0  | 20 | 20
  4|  0  |   0  | 10 | 10     4 |  0  |   0  | 30 | 30
  5|  0  |   0  | 20 | 20     5 |  1  |   1  | 10 | 10
  6|  0  |   0  | 30 | 30
  7|  1  |   1  | 10 | 10
```


CISCO
TM