Cisco Nexus 9300 Platform Buffer and Queuing Architecture

White Paper

November 2014

Contents

What You Will Learn

Cisco Nexus® 9300 platform switches are fixed-configuration Cisco Nexus 9000 Series Switches. The platform delivers industry-leading 1, 10, 40 Gigabit Ethernet port density and performance with high energy efficiency in a compact form factor.

Cisco Nexus 9300 platform switches can operate in either traditional Cisco® NX-OS mode or Cisco Application Centric Infrastructure (ACI) mode. When running in Cisco NX-OS mode, the Cisco Nexus 9300 platform uses the comprehensive Cisco NX-OS Software Layer 2 and 3 feature set and extensive programmability capabilities to offer data center solutions with high performance, operation efficiency, and design flexibility. When deployed in Cisco ACI mode, the Cisco Nexus 9300 platform switches function as leaf nodes in the high-speed, fully automated, bipartite Cisco ACI fabric architecture. They provide attachment points for application endpoints and perform policy-based forwarding and enforcement for Cisco ACI tenant applications.

This document discusses the buffer and queuing architecture of the Cisco Nexus 9300 platform in Cisco NX-OS mode. Cisco ACI mode and the leaf node functions are not within the scope of this document.

Buffer Requirements for Data Center Network Access Layer

Although the need for a deep buffer at the data center network aggregation layer has been eliminated by switch platforms such as the Cisco Nexus 9500 platform, which provide nonblocking, low-latency, line-rate performance with high 10 and 40 Gigabit Ethernet port density, sufficient buffering capacity at the network access layer remains a critical network design principle for several reasons:

- Port speed mismatch often occurs on the access switches because of the presence of a variety of host connectivity types. Uplinks normally have higher speeds than host ports. When traffic is moving from a fast port to a slow port, such as from a 10 Gigabit Ethernet port to a 1 Gigabit Ethernet port, additional buffer space is needed to accommodate the port-speed difference.
- The access layer is often designed with an oversubscription ratio between the host ports and uplink ports.
- Applications with in-cast traffic patterns require deeper buffers on the access-switch ports.

Cisco Nexus 9300 Platform Buffer Structure

A Cisco Nexus 9300 platform switch consists of one network forwarding engine (NFE) and one application leaf engine (ALE) or ALE-2. NFE performs most of the network functions when the Cisco Nexus 9300 platform switch runs in Cisco NX-OS mode, and ALE or ALE-2 provides additional buffer space and facilitates advanced network functions such as routing between Virtual Extensible LANs (VXLANs).

ALE or ALE-2 can be found on either the generic extension module (GEM) of Cisco Nexus 9300 platform switches or on the switch baseboard of certain types of Cisco Nexus 9300 platform switches. Table 1 lists the different types of GEMs and the ALE types with which they are equipped. Table 2 lists the different types of ALE application-specific integrated circuits (ASICs) and their supported Cisco Nexus 9300 switch platforms.

Table 1. Cisco Nexus 9300 Platform GEMs

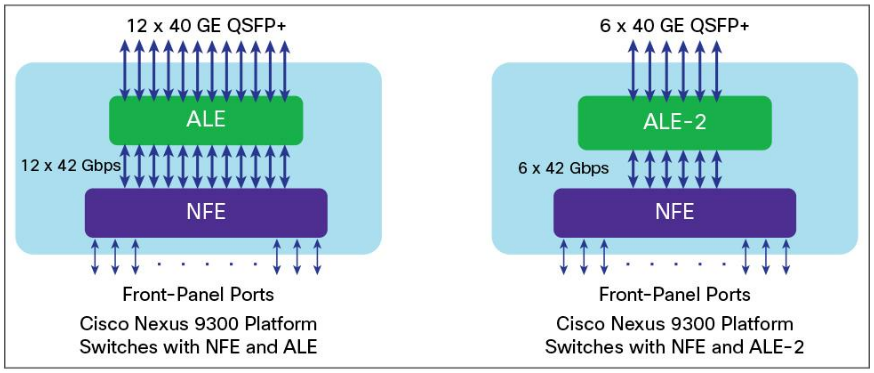
GEM Type	ALE Type	Supported Cisco Nexus 9300 Platform
N9K-M12PQ	ALE	All Cisco Nexus 9300 platform switches that have a GEM slot, including Cisco Nexus 9396PX, 9396TX, 93128PX, and 93128TX Switches
N9K-M6PQ	ALE-2	All Cisco Nexus 9300 platform switches that have a GEM slot, including Cisco Nexus 9396PX, 9396TX, 93128PX, and 93128TX Switches

Table 2. ALE Types and Supported Cisco Nexus 9300 Platform Switches

ALE Type	Buffer Size	Supported Cisco Nexus 9300 Platform
ALE	40 MB	Cisco Nexus 9396PX and 9396TX with an N9K-M12PQ module Cisco Nexus 93128PX and 93128TX with an N9K-M12PQ module
ALE-2	25 MB	Cisco Nexus 9396PX and 9396TX with an N9K-M6PQ module Cisco Nexus 93128PX and 93128TX with an N9K-M6PQ module Cisco Nexus 9372PX and 9372TX Switches (ALE-2 is on the switch baseboard) Cisco Nexus 9332PQ Switch (ALE-2 is on the switch baseboard)

Depending on the ALE type it uses, a Cisco Nexus 9300 platform switch has one of the two internal architectures shown in Figure 1.

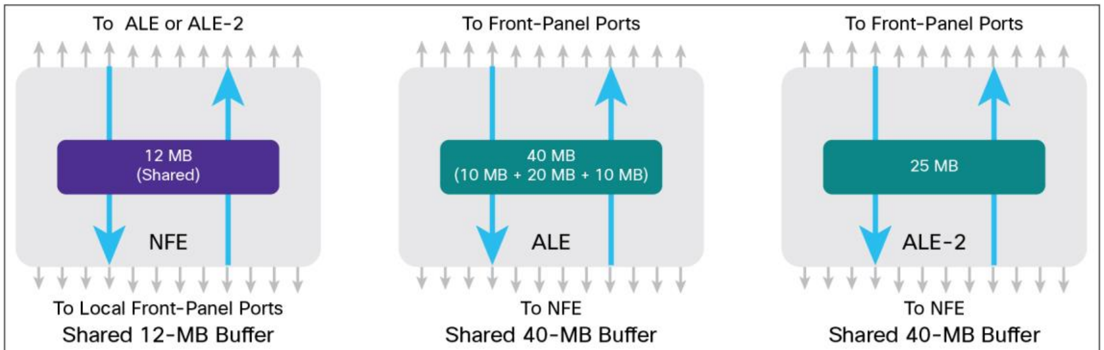
Figure 1. Internal Block Diagrams of Cisco Nexus 9300 Platform Switches



NFE and ALE /ALE-2 provide on-chip buffer space. Figure 2 shows the possible buffer spaces in a Cisco Nexus 9300 platform switch. It includes:

- 12 MB on NFE (shared by all ports on NFE for all traffic)
- 40 MB on ALE (divided into three regions: for traffic from ALE front-panel ports to NFE front-panel ports, for traffic from NFE front-panel ports to ALE front-panel ports, and for hair-pinned traffic between two NFE front-panel ports)
- 25 MB on ALE-2 (shared by all ports on ALE-2 for all traffic)

Figure 2. Buffers in Cisco Nexus 9300 Platform Switches



Buffer Location	Buffer Size
NFE	12 MB
ALE	40 MB
ALE-2	25 MB

Depending on the ALE type, a Cisco Nexus 9300 platform switch can have 52 MB of buffer memory (12 MB on NFE and 40 MB on ALE) or 37 MB of buffer memory (12 MB on NFE and 25 MB on ALE-2).

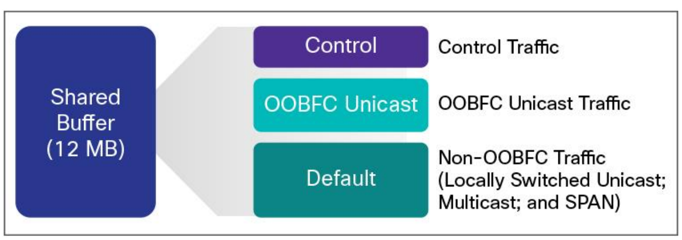
Buffer on Network Forwarding Engine

The 12-MB buffer on NFE is dynamically shared by all ports on NFE. It is divided into three service pools (Figure 3):

- Control service pool
- Out-of-band flow control (OOBFC) unicast service pool
- Default service pool

Control traffic is served with the dedicated buffer resource in the control service pool. The OOBFC unicast service pool serves unicast traffic that has extended output queues on the ALE of the Cisco Nexus 9300 platform switch.

Figure 3. Cisco Nexus 9300 Platform NFE Buffer Service Pools

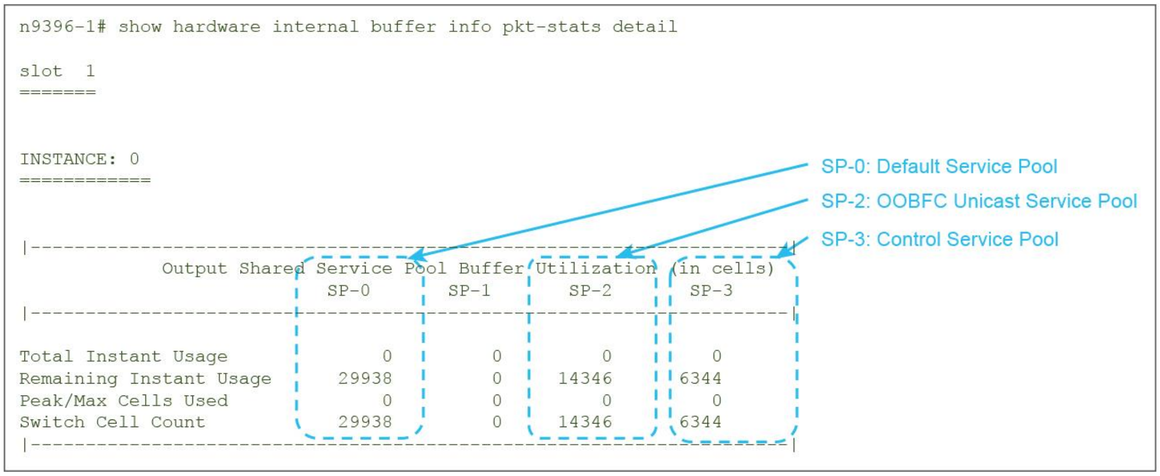


Cisco NX-OS Software for the Cisco Nexus 9300 platform provides command-line interface (CLI) commands for users to dynamically monitor switch buffer configuration and utilization. Figure 4 shows sample output from the monitoring command for the NFE buffer in the switch. In the command output:

- SP-0 is the default service pool
- SP-2 is the OOBFC service pool
- SP-3 is the control service pool

Note that NFE supports up to four buffer service pools. SP-1 is left unused on Cisco Nexus 9300 platform switches.

Figure 4. Cisco Nexus 9300 Platform NFE Buffer Display

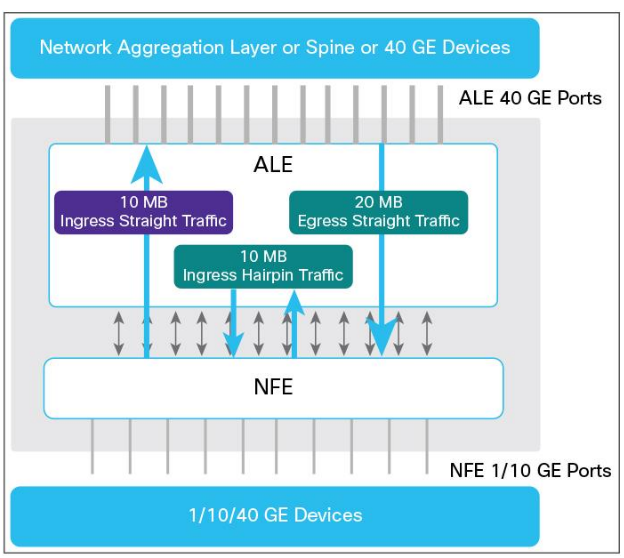


Buffer on Application Leaf Engine

The 40-MB buffer on the ALE consists of three separate regions (Figure 5:

- Buffer for ingress straight traffic (10 MB): The traffic direction is relative to the network. Ingress here means going to the network aggregation layer or spine. So this buffer is for the traffic going out of the ALE 40 Gigabit Ethernet ports of a Cisco Nexus 9300 platform switch.
- Buffer for ingress hairpin traffic (10 MB): This buffer is for traffic travelling between two front-panel ports on NFE. The traffic can be hair-pinned to ALE with the buffer boost feature to take advantage of the additional 10 MB buffer space on ALE.
- Buffer for egress straight traffic (20 MB): Egress means coming from the network and going out to host devices. So this buffer is for traffic coming from ALE 40 Gigabit Ethernet ports and going out on an NFE front-panel port.

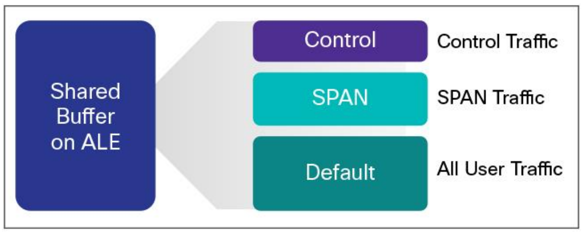
Figure 5. Buffer Regions on ALE



The buffer memory in each of these three regions is dynamically shared by the ports that they serve in the corresponding direction. They are divided into three services pools (Figure 6):

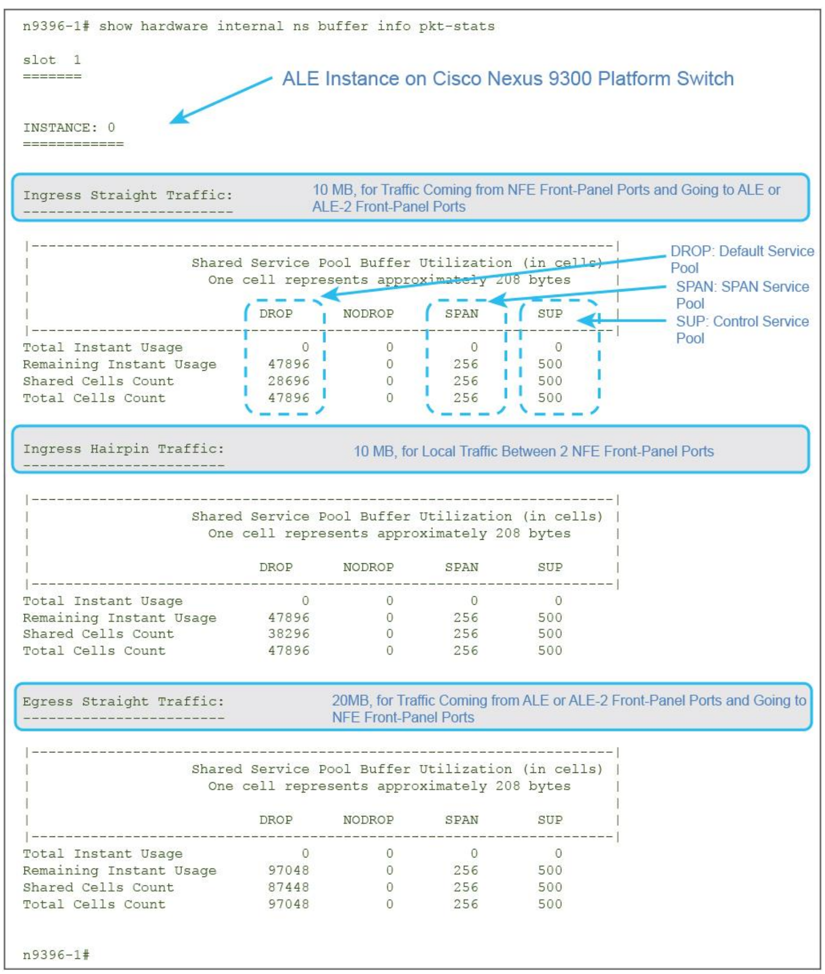
- Control service pool: For all control-plane traffic
- Cisco Switched Port Analyzer (SPAN) service pool: For SPAN traffic
- Default service pool: For all other data traffic

Figure 6. Buffer Service Pools on ALE



Cisco NX-OS for the Cisco Nexus 9000 Series provides CLI commands to display ALE buffer allocation and dynamic utilization. Figure 7 shows sample output from the monitoring command.

Figure 7. ALE Buffer Service Pools Display



```
n9396-1# show hardware internal ns buffer info pkt-stats

slot  1
=======

INSTANCE: 0
============

Ingress Straight Traffic:
-------------------------

|-----------------------------------------------------------------------|
|              Shared Service Pool Buffer Utilization (in cells)        |
|                 One cell represents approximately 208 bytes           |
|                                                                       |
|                         DROP      NODROP      SPAN       SUP          |
|-----------------------------------------------------------------------|
Total Instant Usage          0           0         0         0
Remaining Instant Usage  47896           0       256       500
Shared Cells Count       28696           0       256       500
Total Cells Count        47896           0       256       500

Ingress Hairpin Traffic:
------------------------

|-----------------------------------------------------------------------|
|              Shared Service Pool Buffer Utilization (in cells)        |
|                 One cell represents approximately 208 bytes           |
|                                                                       |
|                         DROP      NODROP      SPAN       SUP          |
|-----------------------------------------------------------------------|
Total Instant Usage          0           0         0         0
Remaining Instant Usage  47896           0       256       500
Shared Cells Count       38296           0       256       500
Total Cells Count        47896           0       256       500

Egress Straight Traffic:
------------------------

|-----------------------------------------------------------------------|
|              Shared Service Pool Buffer Utilization (in cells)        |
|                 One cell represents approximately 208 bytes           |
|                                                                       |
|                         DROP      NODROP      SPAN       SUP          |
|-----------------------------------------------------------------------|
Total Instant Usage          0           0         0         0
Remaining Instant Usage  97048           0       256       500
Shared Cells Count       87448           0       256       500
Total Cells Count        97048           0       256       500

n9396-1#
```

The command output in Figure 7 identifies the three ALE buffer service pool as follows:

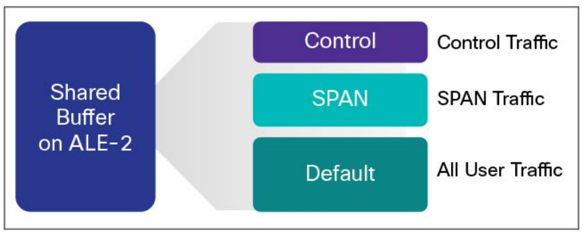
- Drop: Default service pool
- SPAN: SPAN service pool
- SUP: Control service pool

Note that ALE can support four service pools, Drop, Non-drop, SPAN, and SUP. The Non-drop service pool is currently unused on Cisco Nexus 9300 platform. It can be used for Priority Flow Control (PFC) in the future.

Buffer on Application Leaf Engine-2

ALE-2 has 25 MB of buffer memory that is dynamically shared by all the ports on ALE-2 for all traffic. It combines the three regions that ALE has, but keeps the same service pool definitions: Control, SPAN, and Default (Figure 8).

Figure 8. Buffer Service Pools on ALE-2



The ALE-2 service pools are identified in the buffer monitoring command in the same way as the ALE buffer service pools:

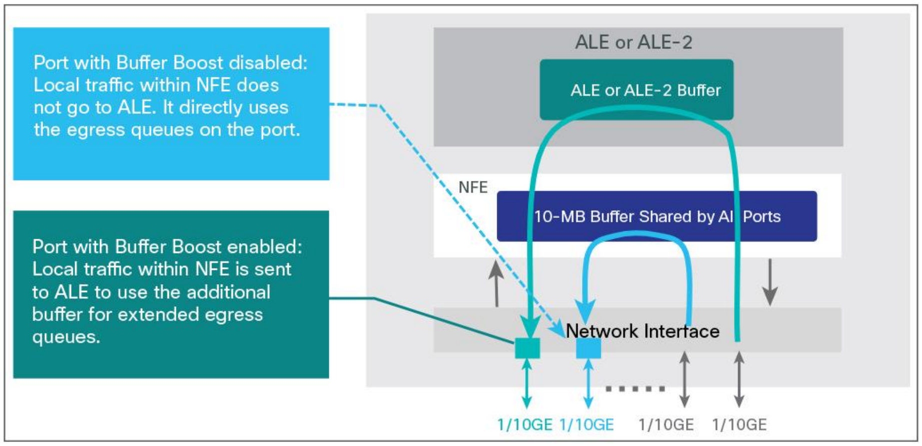
- Drop: Default service pool
- SPAN: SPAN service pool
- SUP: Control service pool

Buffer Boost Feature on Cisco Nexus 9300 Platform

One significant advantage of the Cisco Nexus 9300 platform over other access-switch platforms with the same or similar port density is its larger buffer size. In addition to the 12-MB buffer on NFE, it has the additional 40-MB buffer provided by ALE or 25-MB buffer provided by ALE-2. 10 MB of the 40-MB buffer on ALE is reserved for local traffic between two 1 and 10 Gigabit Ethernet front panel ports on NFE.

Additional buffer space is desirable even for NFE local traffic if the source port has a higher speed than the destination port - for example, from a 10 Gigabit Ethernet port to a 1 Gigabit Ethernet port - or if the local traffic is bursty or in an in-cast pattern. Because NFE performs packet lookup and forwarding, the local traffic between two NFE front-panel ports doesn't need to go to ALE for the forwarding process. However, packets need to be sent to ALE for them to take advantage of the additional ALE buffer. The Buffer Boost feature is introduced for this purpose (Figure 9).

Figure 9. Cisco Nexus 9300 Platform Buffer Boost Feature



When Buffer Boost is enabled on an NFE front-panel port, unicast traffic to this port from another NFE front-panel port will be redirected to ALE or ALE-2 to use the additional buffer space for local traffic. ALE and ALE-2 will hairpin the traffic back for NFE to forward the packets to the egress port. On ALE, a 10-MB buffer space is dedicated for the hairpinned NFE local traffic. On ALE-2, the hairpinned local traffic shares the 25-MB buffer with other traffic. When Buffer Boost is disabled on an NFE front-panel port, NFE will not redirect the traffic from another local port to this port to ALE. Instead, it forwards the traffic directly to this egress port.

Buffer Boost is an egress-port configuration property. It can be enabled or disabled on a per-port basis. It is enabled on all NFE 1 and 10 Gigabit Ethernet front-panel ports by default. Buffer Boost applies only to local unicast traffic. It doesn't change multicast traffic forwarding.

Cisco Nexus 9300 Platform Egress Queues and Extended Output Queues

Cisco Nexus 9300 platform switches use a simple yet efficient class-based egress queuing mechanism to handle link congestion. Cisco Nexus 9300 platform switches use the following types of traffic classes for queuing:

- Control traffic class
- SPAN traffic class
- User traffic classes

The control traffic class and SPAN traffic class are defined internally in the system and are transparent to users. Network control-plane traffic, including traffic for network-control protocols such as Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), and Network Time Protocol (NTP), is classified in the control class.

SPAN traffic, including local SPAN and Encapsulated Remote (ERSPAN) traffic, is categorized in the SPAN class. Control traffic is treated with the highest priority and has reserved buffer resources. SPAN traffic has the lowest priority on a port and uses the remaining bandwidth.

Four user traffic classes are used for egress queuing:

- c-out-q-default: Egress default queue
- c-out-q1: Egress queue 1
- c-out-q2: Egress queue 2
- c-out-q3: Egress queue 3

Users can define and apply traffic classification rules on ingress ports to control the way that traffic is mapped to the four classes. Traffic can be classified based on IP Differentiated Services Code Point (DSCP) or precedence, IEEE 802.1q class of service (CoS), IP access control list (ACL), MAC address ACL, etc. Each class is assigned a quality-of-service (QoS)–group number as its internal identification in the switch system. QoS-group numbers range from 0 through 3.

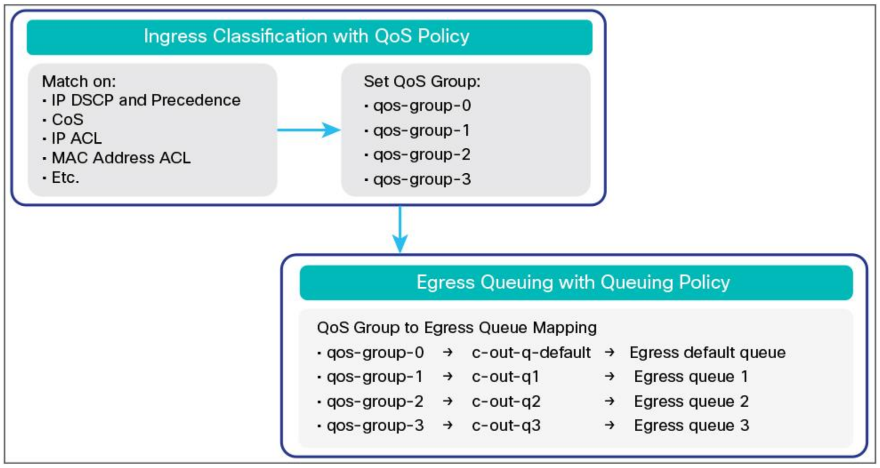
On the egress ports, QoS groups are mapped to the traffic classes as shown here:

- qos-group-0 > c-out-q-default (egress default queue)
- qos-group-1 > c-out-q1 (egress queue 1)
- qos-group-2 > c-out-q2 (egress queue 2)
- qos-group-3 > c-out-q3 (egress queue 3)

A user can define the queuing policies for each class. After it has been classified into a QoS group on the ingress port, traffic will be subject to the egress queuing policies defined for this QoS group on the egress port.

Figure 10 shows the ingress traffic classification and egress queuing process.

Figure 10. Cisco Nexus 9300 Platform QoS Classification and Queuing

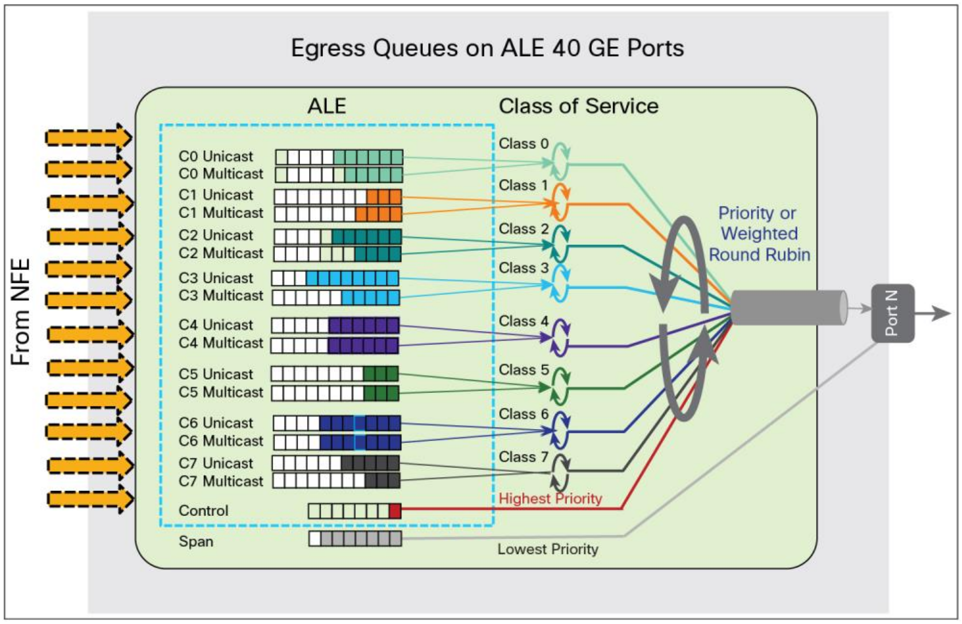


Egress Queues on ALE and ALE-2 40 Gigabit Ethernet Ports

Figure 11 depicts the egress queue structure for the 40 Gigabit Ethernet ports that are provided by ALE or ALE-2. Queues are structured with six traffic classes, including the control traffic class, the SPAN traffic class, and four user-definable classes (internally identified by QoS groups). Within each user-defined class, there is a unicast queue and a multicast queue. Therefore, each ALE 40 Gigabit Ethernet port has the following egress queues:

- One control traffic queue
- One SPAN traffic queue
- Four unicast queues
- Four multicast queues

Figure 11. Output Queues on 40 Gigabit Ethernet Ports on ALE and ALE-2



These egress queues on the 40 Gigabit Ethernet ports consume the 10-MB ingress straight traffic buffer on ALE, or if the ports are on ALE-2, they share the 25-MB buffer with other traffic through ALE-2.

Egress and Extended Egress Queues on NFE Front-Panel Ports

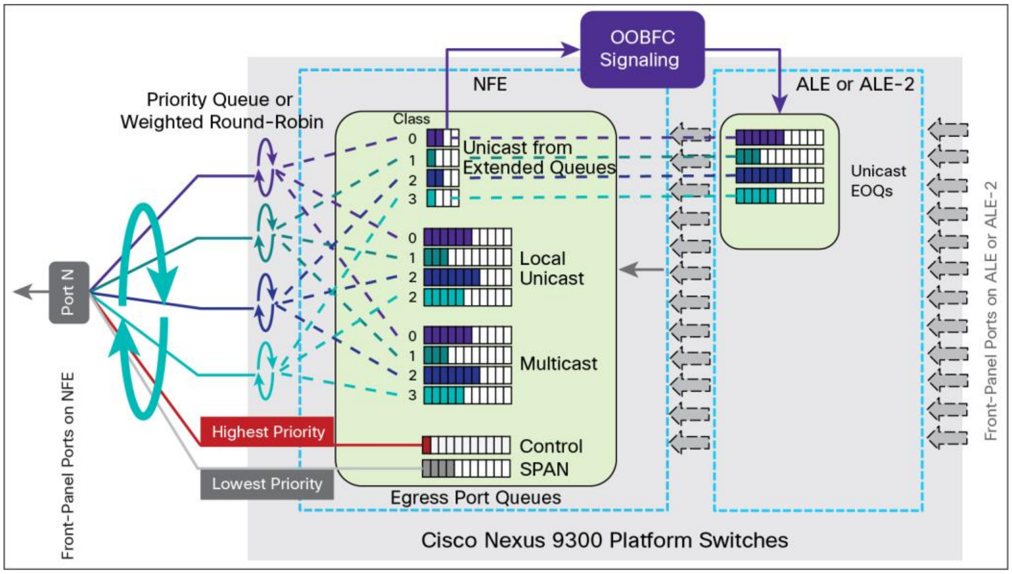
Like 40 Gigabit Ethernet ports on ALE, each front-panel port on NFE has the set of egress queues for control traffic, SPAN traffic, multicast traffic, and unicast traffic. Additionally, each NFE port has four OOBFC unicast queues. Theses queues are for unicast extended egress queues on ALE. As a result, the following queues are seen on each NFE 1 and 10 Gigabit Ethernet egress port:

- One control traffic queue
- One SPAN traffic queue
- Four multicast queues
- Four unicast queues (for local unicast traffic)
- Four OOBFC unicast queues (These queues are for OOBFC-controlled unicast traffic, including hairpinned NFE local unicast traffic and egress straight unicast traffic from ALE 40 Gigabit Ethernet ports to NFE front-panel ports.)

On ALE, there are four corresponding unicast extended output queues (EoQs) for each NFE egress port. NFE uses the OOBFC signaling channel to tell ALE when to stop or when to resume sending traffic to NFE on a per-egress-port and per-unicast-class basis. When ALE is instructed to stop sending traffic to NFE, it queues the packets in the appropriate EoQ using its own buffer. As a result, the egress unicast queues on NFE are extended to the ALE EoQs to use the additional ALE buffer resources. The unicast traffic that can take advantage of OOBFC-signaled EoQ on ALE includes the egress straight traffic travelling from ALE 40 Gigabit Ethernet ports to an NFE front-panel port, and the hairpinned local traffic travelling between two NFE ports.

Figure 12 shows the NFE egress queues and ALE EoQs for an NFE front-panel port of a Cisco Nexus 9300 platform switch.

Figure 12. Cisco Nexus 9300 Platform NFE Front-Panel Port Egress and Extended Egress Queues



Weighted Round-Robin and Priority Queuing on Cisco Nexus 9300 Platform

Cisco Nexus 9300 platform switches use the Weighted Round-Robin (WRR) and Priority Queuing (PQ) mechanisms to manage the egress queues and extended egress queues on NFE and ALE.

These are the default queuing polices for the four-user traffic classes:

- c-out-q3
- c-out-q2
- c-out-q1
- c-out-q-default

```
n9396-1# sh policy-map type queuing default-out-policy


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

  policy-map type queuing default-out-policy
    class type queuing c-out-q3
      priority level 1
    class type queuing c-out-q2
      bandwidth remaining percent 0
    class type queuing c-out-q1
      bandwidth remaining percent 0
    class type queuing c-out-q-default
      bandwidth remaining percent 100
n9396-1#
```

In a WRR queuing policy, bandwidth can be defined as a percentage of the link bandwidth, or as a percentage of the remaining bandwidth.

When you use priority queuing, the other nonpriority queues (WRR queues) can have their bandwidth defined only as a percentage of the remaining bandwidth. Cisco Nexus 9300 platform switches support up to three priority queues. They must start with the class c-out-q3 in the policy-map configuration and move to c-out-q2 and c-out-q1 in sequence.

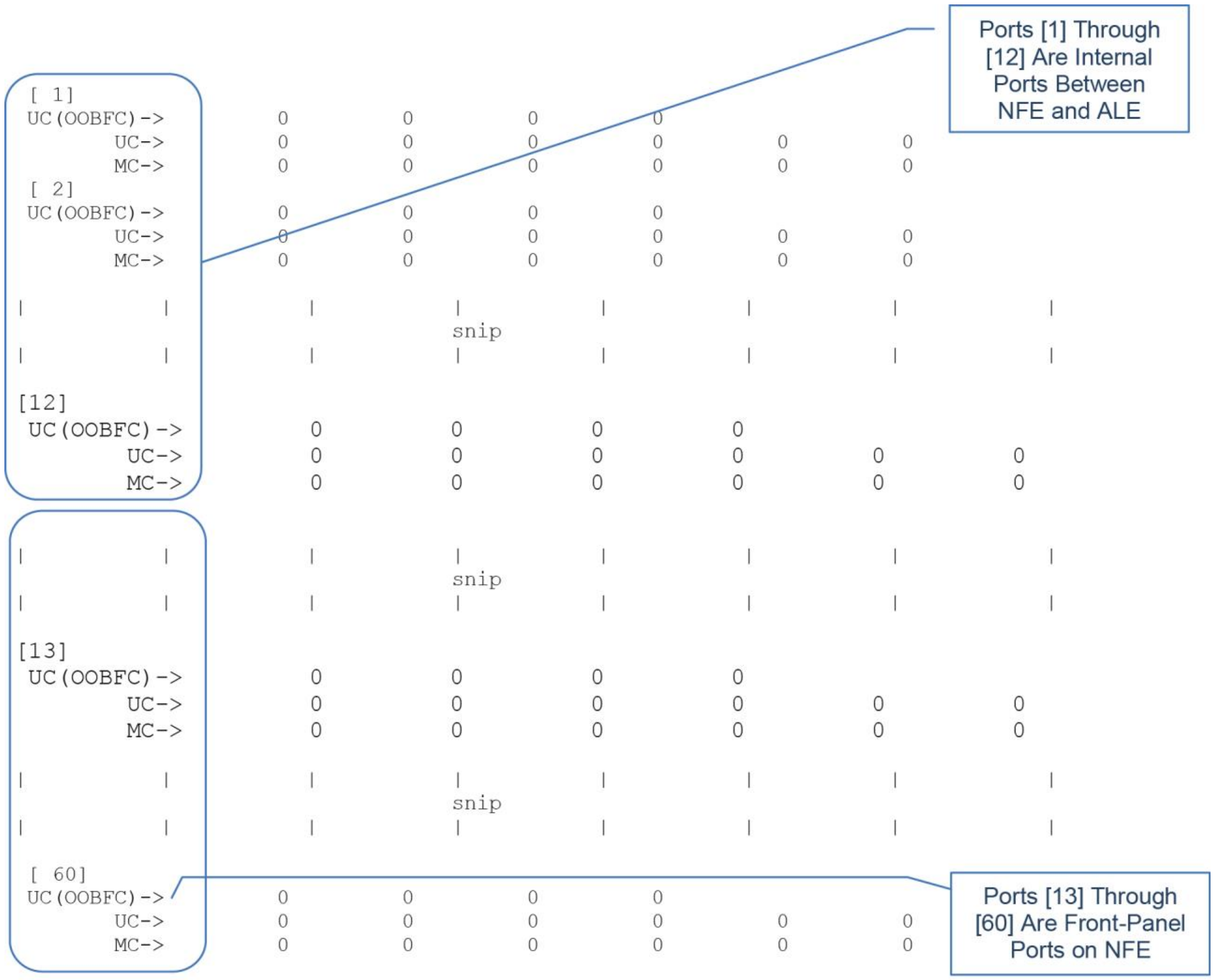
Egress Queue and Extended Egress Queue Monitoring

Buffer and Queue Monitoring on NFE

The following example shows the buffer and queue monitoring results on NFE for a Cisco Nexus 9396PX Switch. The **show hardware internal buffer info pkt-state detail** command shows dynamic buffer statistics for all ports on NFE on a per-traffic-class and per-queue basis. Each port has six classes: Q3, Q2, Q1, Q0, CPU, and SPAN. Classes Q3 through Q0 each have an OOBFC unicast queue, a non-OOBFC unicast queue, and a multicast queue. Classes CPU and SPAN each have a unicast queue and a multicast queue.

```
n9396-1# show hardware internal buffer info pkt-stats  detail

slot  1
=======


INSTANCE: 0
============


|--------------------------------------------------------------------|
             Output Shared Service Pool Buffer Utilization (in cells)
                            SP-0       SP-1       SP-2       SP-3
|--------------------------------------------------------------------|

Total Instant Usage             0          0          0          0
Remaining Instant Usage     29938          0      14346       6344
Peak/Max Cells Used            33          0       1531        163
Switch Cell Count           29938          0      14346       6344
|--------------------------------------------------------------------|

|-------------------------------------------------------------------------|
|              Instant Buffer utilization per queue per port              |
|        Each line displays the number of cells utilized for a given      |
|                     port for each QoS queue                              |
|                One cell represents approximately 208 bytes              |
|-------------+---------+---------+---------+---------+---------+--------+|
|ASIC Port         Q3        Q2        Q1        Q0        CPU      SPAN  |
|-------------+---------+---------+---------+---------+---------+--------+|
```

6 Classes per NFE Port



Note: This command output shows buffer statistics for all active ports on NFE, starting with the internal ports between NFE and ALE or ALE-2 followed by NFE front-panel ports. The preceding example is taken from a Cisco Nexus 9396PX Switch that has 12 internal 40 Gigabit Ethernet ports between NFE and ALE, and 48 1 and 10 Gigabit Ethernet front-panel ports on NFE. Therefore, the command output shows 60 ASIC ports:

- Ports 1 through 12: Internal ports between NFE and ALE
- Ports 13 through 60: Front-panel ports on NFE

A variation of the preceding buffer monitoring command shows the peak buffer utilization value in each queue. Sample output for the high-watermark monitoring is shown here:


```
n9396-1# show hardware internal buffer info pkt-stats peak

slot  1
=======


INSTANCE: 0
============


|--------------------------------------------------------------------|
            Output Shared Service Pool Buffer Utilization (in cells)
                           SP-0       SP-1       SP-2       SP-3
|--------------------------------------------------------------------|

Total Instant Usage             0          0          0          0
Remaining Instant Usage     29938          0      14346       6344
Peak/Max Cells Used            33          0       1531        163
Switch Cell Count           29938          0      14346       6344
|--------------------------------------------------------------------|

|--------------------------------------------------------------------------|
|              Peak Buffer utilization per queue per port                  |
|       Each line displays the number of cells utilized for a given        |
|                    port for each QoS queue                                |
|            One cell represents approximately 208 bytes                   |
|--------------+---------+---------+---------+---------+---------+--------+|
|ASIC Port         Q3        Q2        Q1        Q0        CPU      SPAN   |
|--------------+---------+---------+---------+---------+---------+--------+|

 [ 1]
 UC(OOBFC)->        0         0         0         0
        UC->        0         0         0         3         74         0
        MC->        0         0         0         1          0         0
 [ 2]
 UC(OOBFC)->        0         0         0         0
        UC->        0         0         0         1         74         0
        MC->        0         0         0         1          0         0
 [ 3]
 UC(OOBFC)->        0         0         0         0
        UC->        0         0         0         1         72         0
        MC->        0         0         0         1          1         0
 [ 4]
 UC(OOBFC)->        0         0         0         0
        UC->        0         0         0         3         73         0
        MC->        0         0         0         1          1         0
[20]
 UC(OOBFC)->        0         0         0       224
        UC->        0         0         0         0          8         0
        MC->        0         0         0         0          1         0
```

Buffer and Queue Monitoring on ALE and ALE-2

The **show hardware internal ns buffer info pkt-stats** command monitors buffer utilization and queue statistics for ALE.

```
n9396-1# show hardware internal ns buffer info pkt-stats detail

slot  1
=======
INSTANCE: 0
============


Ingress Straight Traffic:
------------------------

|------------------------------------------------------------------|
|                 Shared Service Pool Buffer Utilization (in cells) |
|                   One cell represents approximately 208 bytes     |
```



```
|                           DROP      NODROP      SPAN      SUP      |
|--------------------------------------------------------------------|
Total Instant Usage           0          0          0          0

Remaining Instant Usage   47896          0        256        500
Shared Cells Count        28696          0        256        500
Total Cells Count         47896          0        256        500

|------------------------------------------------------------------|
|        Instant Buffer utilization per port per pool               |
|    Each line displays number of cells utilized for a given        |
|             port for each policy class                             |
|        One cell represents approximately 208 bytes                |
|-------------+---------+---------+---------+---------+---------+|
|ASIC Port        Q0        Q1        Q2        Q3       SUP      |
|-------------+---------+---------+---------+---------+---------+|

 [MACN0]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN1]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN2]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN3]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN4]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN5]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN6]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN7]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN8]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN9]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN10]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --
 [MACN11]
        UC->      0         0         0         0        --
        MC->      0         0         0         0        --

Ingress Hairpin Traffic:
------------------------

|--------------------------------------------------------------------|
|                  Shared Service Pool Buffer Utilization (in cells) |
|                    One cell represents approximately 208 bytes     |
|                                                                    |
|                           DROP      NODROP      SPAN      SUP      |
|--------------------------------------------------------------------|
Total Instant Usage           0          0          0          0
Remaining Instant Usage   47896          0        256        500
Shared Cells Count        38296          0        256        500
Total Cells Count         47896          0        256        500
```

12 x 40 Gigabit Ethernet Front-Panel Ports on ALE

12 x 40 Gigabit Ethernet ALE Internal Ports to NFE

```
[MACF0]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF1]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF2]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF3]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF4]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF5]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF6]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF7]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF8]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF9]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF10]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --
[MACF11]
        UC->        0        0        0        0        --
        MC->        0        0        0        0        --

|-----------------------------------------------------------|
|         Instant Buffer utilization per EOQ per pool       |
|       Each line displays number of cells utilized for     |
|            a given eoq for each policy class              |
|          One cell represents approximately 208 bytes      |

[EOQ 0 : BCM 13  ]        0        0        0        0
[EOQ 1 : BCM 14  ]        0        0        0        0

|           |           |           |           |           |

|           |           |           |           |           |

|           |           |           |           |           |

[EOQ 46 : BCM 59 ]        0        0        0        0
[EOQ 47 : BCM 60 ]        0        0        0        0
[EOQ 48          ]        0        0        0        0
[EOQ 49          ]        0        0        0        0
|           |           |           |           |           |

|           |           |           |           |           |

|           |           |           |

[EOQ 94          ]        0        0        0        0
[EOQ 95          ]        0        0        0        0
```

Unicast EoQs for Each Front-Panel Egress Port on NFE

Each ALE Can Support EoQs for Up to 96 NFE Front-Panel Egress Ports

Egress Straight Traffic:

```
Egress Straight Traffic:
------------------------

|--------------------------------------------------------------------|
|                   Shared Service Pool Buffer Utilization (in cells) |
|                     One cell represents approximately 208 bytes     |
|                                                                     |
|                          DROP      NODROP      SPAN       SUP       |
|--------------------------------------------------------------------|
Total Instant Usage           0          0          0          0
Remaining Instant Usage   97048          0        256        500
Shared Cells Count        87448          0        256        500
Total Cells Count         97048          0        256        500

 [MACF0]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF1]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF2]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF3]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF4]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF5]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF6]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF7]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF8]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF9]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF10]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --
 [MACF11]
        UC->         0          0          0          0         --
        MC->         0          0          0          0         --

 [EOQ 0 : BCM 13  ]       0          0          0          0
 [EOQ 1 : BCM 14  ]       0          0          0          0
 |          |           |           |           |           |
|           |           |           |           |           |
|           |           |           |           |           |
 [EOQ 46 : BCM 59 ]       0          0          0          0
 [EOQ 47 : BCM 60 ]       0          0          0          0
 [EOQ 48          ]       0          0          0          0
 [EOQ 49          ]       0          0          0          0
|           |           |           |           |           |
|           |           |           |           |           |
|           |           |           |           |           |
 [EOQ 94          ]       0          0          0          0
 [EOQ 95          ]       0          0          0          0
n9396-1#
```

12 x 40 Gigabit Ethernet ALE Internal Ports to NFE

Unicast EoQs for Each NFE Front-Panel Egress Port, Up to 96

Queue Monitoring on Interfaces

```
n9396-1# sh queuing interface e1/1 summary

slot  1
=======


Egress Queuing for Ethernet1/1 [System]
------------------------------------------------------------------
QoS-Group# Bandwidth% PrioLevel                Shape
                                  Min           Max         Units

------------------------------------------------------------------
     3            -          1            -             -      -
     2            0          -            -             -      -
     1            0          -            -             -      -
     0          100          -            -             -      -
+---------------------------------------------------------------------+
|                              QOS GROUP 0                            |
+---------------------------------------------------------------------+
|                 |  Unicast       | OOBFC Unicast  |  Multicast     |
+---------------------------------------------------------------------+
|        Tx Pkts |               0|     5325011301|                0|
|        Tx Byts |               0|  5954391263104|                0|
|   Dropped Pkts |               0|              0|                0|
|   Dropped Byts |               0|              0|                0|
|   Q Depth Byts |               0|              0|                0|
+---------------------------------------------------------------------+
|                              QOS GROUP 1                            |
+---------------------------------------------------------------------+
|                 |  Unicast       | OOBFC Unicast  |  Multicast     |
+---------------------------------------------------------------------+
|        Tx Pkts |               0|              0|                0|
|        Tx Byts |               0|              0|                0|
|   Dropped Pkts |               0|              0|                0|
|   Dropped Byts |               0|              0|                0|
|   Q Depth Byts |               0|              0|                0|
+---------------------------------------------------------------------+
|                              QOS GROUP 2                            |
+---------------------------------------------------------------------+
|                 |  Unicast       | OOBFC Unicast  |  Multicast     |
+---------------------------------------------------------------------+
|        Tx Pkts |               0|              0|                0|
|        Tx Byts |               0|              0|                0|
|   Dropped Pkts |               0|              0|                0|
|   Dropped Byts |               0|              0|                0|
|   Q Depth Byts |               0|              0|                0|
+---------------------------------------------------------------------+
|                              QOS GROUP 3                            |
+---------------------------------------------------------------------+
|                 |  Unicast       | OOBFC Unicast  |  Multicast     |
+---------------------------------------------------------------------+
|        Tx Pkts |               0|              0|                0|
|        Tx Byts |               0|              0|                0|
|   Dropped Pkts |               0|              0|                0|
|   Dropped Byts |               0|              0|                0|
|   Q Depth Byts |               0|              0|                0|
+---------------------------------------------------------------------+
```

```
+----------------------------------------------------------------------+
|                       CONTROL QOS GROUP 4                            |
+----------------------------------------------------------------------+
|                 |  Unicast        | OOBFC Unicast  |  Multicast      |
+----------------------------------------------------------------------+
|         Tx Pkts |             8714|               0|                0|
|         Tx Byts |          1024410|               0|                0|
|    Dropped Pkts |                0|               0|                0|
|    Dropped Byts |                0|               0|                0|
|    Q Depth Byts |                0|               0|                0|
+----------------------------------------------------------------------+
|                          SPAN QOS GROUP 5                            |
+----------------------------------------------------------------------+
|                 |  Unicast        | OOBFC Unicast  |  Multicast      |
+----------------------------------------------------------------------+
|         Tx Pkts |                0|               0|                0|
|         Tx Byts |                0|               0|                0|
|    Dropped Pkts |                0|               0|                0|
|    Dropped Byts |                0|               0|                0|
|    Q Depth Byts |                0|               0|                0|
+----------------------------------------------------------------------+

Port Ingress Statistics
--------------------------------------------------------
Ingress MMU Drop Pkts                                0
Ingress MMU Drop Bytes                               0


Port Egress Statistics
--------------------------------------------------------
WRED Drop Pkts                               0
NS Straight EOQ(qos-group-0) Drop Pkts                  893
NS BufferBoost EOQ(qos-group-0) Drop Pkts                 0

PFC Statistics
--------------------------------------------------------
TxPPP:                     0, RxPPP:                    0
--------------------------------------------------------
 COS QOS Group   TxPause   TxCount        RxPause        RxCount
   0         -  Inactive         0       Inactive              0
   1         -  Inactive         0       Inactive              0
   2         -  Inactive         0       Inactive              0
   3         -  Inactive         0       Inactive              0
   4         -  Inactive         0       Inactive              0
   5         -  Inactive         0       Inactive              0
   6         -  Inactive         0       Inactive              0
   7         -  Inactive         0       Inactive              0
--------------------------------------------------------

n9396-1#
```

Queue Limit Control

The queue limit can be defined on a per-port and per-class basis on Cisco Nexus 9300 platform switches. It provides a mechanism for preventing a given port or a given traffic class from using too much of the buffer resources and causing buffer starvation for other ports or traffic classes. The queue limit can also be used to allocate more buffer space to a given port or traffic class when needed.

Cisco Nexus 9300 platform switches support both static queue limits and dynamic queue limits. A static queue limit specifies the exact number of bytes, kilobytes, or megabytes for a particular traffic class in the queue. A static limit can also be specified as the amount of time in milliseconds or microseconds that packets are allowed to remain in the queue. Static queue limits are helpful when precise buffer and queue control is needed for a particular traffic class on some ports.

A dynamic queue limit provides a flexible and dynamic means of controlling per-port and per-class queue limits. By selecting a dynamic queue-limit factor from the options listed in Table 3, a user can specify the amount of available buffer space a queue can consume per port and per class at any given time.

Table 3. Dynamic Queue-Limit Factors

Dynamic Queue-Limit Factor	Queue Limit as Percentage of Available Buffer Space
Option 0: 1/128	1%
Option 1: 1/64	2%
Option 2: 1/32	3%
Option 3: 1/16	6%
Option 4: 1/8	11%
Option 5: 1/4	20%
Option 6: 1/2	33%
Option 7: 1	50%
Option 8: 2	67%
Option 9: 4	80%
Option 10: 8	89%

A dynamic queue limit provides optimal utilization of the buffer space while preventing a queue from consuming too much of the buffer resources. The default queue-limit setting is option 8, which allows per-class and per-queue use of up to 67% of the available buffer space. If the traffic on a port or for a particular class is anticipated to be bursty, the user can change the queue limit for it to option 9 or 10 to use up to 89% of the available bandwidth.

Burst Profile and Flow Prioritization on ALE and ALE-2

ALE and ALE-2 Burst Profiles

ALE and ALE-2 provide three burst profiles:

- Burst: Burst optimized
- Mesh: Mesh optimized
- Ultra-burst: Ultra-burst optimized

Mesh is the default burst mode. However, if the traffic through a Cisco Nexus 9300 platform switch is known to be bursty, the burst mode is recommended. The following global command can be used to change the burst profile. The command change doesn't request system reboot.

```
n9396-1(config)# hardware qos ns-buffer-profile ?
  burst         Burst optimized
  mesh          Mesh optimized
  ultra-burst  Ultra burst optimized
```

The CLI command **show hardware qos ns-buffer-profile** displays the current burst profile in the switch running configuration:

```
n9396-1# show hardware qos ns-buffer-profile
NS Buffer Profile: Burst optimized
n9396-1#
```


ALE and ALE-2 Flow Prioritization

ALE and ALE-2 have built-in intelligence that can prioritize flows based on their life spans. Given a mixture of long-lived flows and short and bursty flows, ALE and ALE-2 can recognize and prioritize the short flows. In the event of link congestion in which the switch has to drop some packets, ALE and ALE-2 will first drop packets from the long flows while allowing short flows to go through without packet loss.

Figure 13 shows the results of a flow prioritization test on a Cisco Nexus 9396PX Switch. In the test, a constant 10 Gigabit Ethernet flow and a short-lived 10 Gigabit Ethernet burst flow were sent to each of the egress 10 Gigabit Ethernet ports on NFE. The results show that the constant flows experienced packet loss, but the burst flows went through without packet loss.

Figure 13. ALE and ALE-2 Flow Prioritization Demonstration

	Tx Port	Rx Port	Traffic Item	Tx Frames	Rx Frames	Frames Delta	Loss %	Tx Frame Rate	Rx Frame Rate	Tx L1 Rate (bps)	Rx L1 Rate (bps)	Rx Bytes
1	40GE-9396-2/9	10GE-9396-1/1	const-9396	388,470,164	388,469,183	981	0.000	2,349,389.151	2,349,389.651	9,999,000,225...	9,923,821,884...	197,342,3...
2	40GE-9396-2/9	10GE-9396-1/1	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000
3	40GE-9396-2/10	10GE-9396-1/2	const-9396	388,470,185	388,469,842	343	0.000	2,349,388.571	2,349,389.571	9,998,997,757...	9,923,821,548...	197,342,6...
4	40GE-9396-2/10	10GE-9396-1/2	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000
5	40GE-9396-2/11	10GE-9396-1/3	const-9396	388,471,352	388,470,940	412	0.000	2,349,388.707	2,349,388.707	9,998,998,335...	9,923,817,896...	197,343,2...
6	40GE-9396-2/11	10GE-9396-1/3	burst-9396	5,000	5,000	0	0.000	0.000	0.000	0.000	0.000	2,540,000

Many data center applications use long flows for data transport and short flows for state synchronization or for requests. These short flows are more sensitive to packet drops and latency. By prioritizing these short flows over the long-lasting data-transport flows, the ALE and ALE-2 flow prioritization feature can help improve data center application performance.

Conclusion

Cisco Nexus 9300 platform switches are designed to provide high-performance, cost-effective network connectivity and an extensive programmability feature set to support the operating model of the modern data center. The platform's industry-leading 1, 10, and 40 Gigabit Ethernet port density in a compact fixed-configuration form factor enables organizations to migrate the data center network access layer from 1 Gigabit Ethernet to 10 Gigabit Ethernet for host access, and from 10 Gigabit Ethernet to 40 Gigabit Ethernet for uplinks to data center aggregation and spine layers. The extended buffer capacity and enhanced egress queuing architecture on Cisco Nexus 9300 platform switches helps ensure application performance in a diversified and dynamic network environment.

For More Information

For more information, go to: http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-732452-00 11/14

