# Understanding CoS Buffer Configuration

Each QFX3500 and QFX3600 switch has 9 MB of Packet Forwarding Engine (PFE) wide common packet buffer memory that is used to store packets on interface queues. Each QFX5100 and EX4600 switch has 12 MB of PFE wide common packet buffer memory. The buffer memory has separate ingress and egress accounting to make accept, drop, or pause decisions. Because the switch has a single pool of memory with separate ingress and egress accounting, the full amount of buffer memory is available from both the ingress and the egress perspective. Packets are accounted for as they enter and leave the switch, but there is no concept of a packet arriving at an ingress buffer and then being moved to an egress buffer.

The buffers are divided into two pools from both an ingress and an egress perspective:

1. *Shared buffers* are a global memory pool that the switch allocates dynamically to ports as needed, so the buffers are shared among the switch ports.
2. *Dedicated buffers* are a memory pool divided equally among the switch ports. Each port receives a minimum guaranteed amount of buffer space, dedicated to each port, not shared among ports.

**Note:** Lossless traffic is traffic on which you enable priority-based flow control (PFC) to ensure lossless transport. Lossless traffic does not refer to best-effort traffic on a link enabled for Ethernet PAUSE (IEEE 802.3x).

The switch reserves nonconfigurable buffer space to ensure that ports and queues receive a minimum memory allocation. You can configure how the system uses the rest of the buffer space to optimize the allocation for your mix of network traffic. You can configure the percentage of available buffer space used as shared buffer space versus dedicated buffer space. You can also configure how shared buffer space is allocated to different types of traffic. You can optimize the buffer settings for the traffic on your network.

The default buffer configuration is designed for networks that have a balance of best-effort and lossless traffic.

The default class-of-service configuration provides two lossless forwarding classes (`fcoe` and `no-loss`), a best-effort unicast forwarding class, a network control traffic forwarding class, and one multidestination (multicast, broadcast, and destination lookup fail) forwarding class. Each default forwarding class maps to a different default output queue. The default configuration allocates the buffers in a manner that supports a moderate amount of lossless traffic while still providing the ability to absorb bursts in best-effort traffic transmission.

Changing the buffer settings changes the abilities of the buffers to absorb traffic bursts and handle lossless traffic. For example, networks with mostly best-effort traffic require allocating most of the shared buffer space to best-effort buffers. This provides deep, flexible buffers that can absorb traffic bursts with minimal packet loss, at the expense of buffer availability for lossless traffic.

Conversely, networks with mostly lossless traffic require allocating most of the shared buffer space to lossless headroom buffers. This prevents packet loss on lossless flows at the expense of absorbing bursty best-effort traffic efficiently.

**Caution:** Changing the buffer configuration is a disruptive event. Traffic stops on *all* ports until buffer reprogramming is complete.

This topic describes the buffer architecture and settings:

## Buffer Pools

From both an ingress and an egress perspective, the PFE buffer is split into two main pools, a shared buffer pool and a dedicated buffer pool that ensures a minimum allocation to each port. You can configure the amount of buffer space allocated to each of the two pools. A portion of the buffer space is reserved so that there is always a minimum amount of shared and dedicated buffer space available to each port.

- Shared buffer pool—A global memory space that all of the ports on the switch share dynamically as they need buffers. The shared buffer pool is further partitioned into buffers for best-effort unicast, best-effort multidestination (broadcast, multicast, and destination lookup fail), and PFC (lossless) traffic types. You can allocate global shared memory space to buffer partitions to better support different mixes of network traffic. The larger the shared buffer pool, the better the switch can absorb traffic bursts because more shared memory is available for the traffic.
- Dedicated buffer pool—A reserved global memory space allocated equally to each port. The switch reserves a minimum dedicated buffer pool that is not user-configurable. You can divide the dedicated buffer allocation for a port among the port queues on a per-port, per-queue basis. (For example, this enables you to dedicate more buffer space to queues that transport lossless traffic.)

  A larger dedicated buffer pool means a larger amount of dedicated buffer space for each port, so congestion on one port is less likely to affect traffic on another port because the traffic does not need to use as much shared buffer space. However, the larger the dedicated buffer pool, the less bursty traffic the switch can handle because there is less dynamic shared buffer memory.

You can configure the way the available unreserved portion of the buffer space is allocated to the global shared buffer pool and to the dedicated shared buffer pool by configuring the ingress and egress shared buffer percentages.

By default, 100 percent of the available unreserved buffer space is allocated to the shared buffer pool. If you change the percentage of space allocated to the shared buffer, the available buffer space that is not allocated to the shared buffer is allocated to the dedicated buffer. For example, if you configure the ingress shared buffer pool as 80 percent, the remaining 20 percent of the available buffer space is allocated to the dedicated buffer pool and divided equally across the ports.

**Note:** When 100 percent of the available (user-configurable) buffers are allocated to the shared buffer pool, the switch still reserves a minimum dedicated buffer pool.

You can separately configure ingress and egress shared buffer pool allocations. You can also partition the ingress and egress shared buffer pool to allocate percentages of the shared buffer pool to specific types of traffic. If you do not use the default configuration or one of the recommended configurations, pay particular attention to the ingress configuration of the lossless and lossless headroom buffers (these buffers handle PFC pause during periods of congestion) and to the egress configuration of the best-effort buffers to handle incast congestion (multiple synchronized sources sending data to the same receiver in parallel).

In addition to the shared buffer pool and the dedicated buffer pool, there is also a small ingress global headroom buffer pool that is reserved and is not configurable.

When contention for buffer space occurs, the switch uses an internal algorithm to ensure that the buffer pools are distributed fairly among competing flows. When traffic for a given flow exceeds the amount of dedicated port buffer reserved for that flow, the flow begins to consume memory from the dynamic shared buffer pool. Competing flows compete for shared buffer memory with other flows that also have exhausted

their dedicated buffers. When there is no congestion, there are no competing flows.



## Buffer Handling of Lossless Flows (PFC) Versus Ethernet PAUSE

When we discuss lossless buffers in the following sections, we mean buffers that handle traffic on which you enable PFC to ensure lossless transport. The lossless buffers are not used for best-effort traffic on a link on which you enable Ethernet PAUSE (IEEE 802.3x). The lossless ingress and egress shared buffers, and the ingress lossless headroom shared buffer, are used only for traffic on which you enable PFC.

**Note:** To support lossless flows, you must configure the appropriate data center bridging capabilities (PFC, DCBX, or ETS) and scheduling properties.

## Shared Buffer Pool and Partitions

The shared buffer pool is a global memory space that all of the ports on the switch share dynamically as they need buffers. The switch uses the shared buffer pool to absorb traffic bursts after the dedicated buffer pool for a port is exhausted.

You can divide both the ingress shared buffer pool and the egress shared buffer pool into three partitions to allocate percentages of each buffer pool to different types of traffic. When you partition the ingress or egress shared buffer pool:

- If you explicitly configure one ingress shared buffer partition, you must explicitly configure all three ingress shared buffer partitions. (You either explicitly configure all three ingress partitions or you use the default setting for all three ingress partitions.)

  If you explicitly configure one egress shared buffer partition, you must explicitly configure all three egress shared buffer partitions. (You either explicitly configure all three egress partitions or you use the default setting for all three egress partitions.)

  The switch returns a commit error if you do not explicitly configure all three partitions when configuring the ingress or egress shared buffer partitions.

- The combined percentages of the three ingress shared buffer partitions must total exactly 100 percent.

  The combined percentages of the three egress shared buffer partitions must total exactly 100 percent.

  When you explicitly configure ingress or egress shared buffer partitions, the switch returns a commit error if the total percentage of the three partitions does not equal 100 percent.

- If you explicitly partition one set of shared buffers, you do not have to explicitly partition the other set of shared buffers. For example, you can explicitly configure the ingress shared buffer partitions and use the default egress shared buffer partitions. However, if you change the buffer partitions for the ingress buffer pool to match the expected types of traffic flows, you would probably also want to change the buffer partitions for the egress buffer pool to match those traffic flows.

You can configure the percentage of available unreserved buffer space allocated to the shared buffer pool. Space that you do not allocate to the shared buffer pool is added to the dedicated buffer pool and divided equally among the ports. The default configuration allocates 100 percent of the unreserved ingress and

egress buffer space to the shared buffers.

Configuring the ingress and egress shared buffer pool partitions enables you to allocate more buffers to the types of traffic your network predominantly carries, and fewer buffers to other traffic.

### Ingress Shared Buffer Pool Partitions

You can configure three ingress buffer pool partitions:

- Lossless buffers—Shared buffer pool for all lossless ingress traffic. The recommended minimum value for lossless buffers is 5 percent.
- Lossless headroom buffers—Shared buffer pool for packets received while a pause is asserted. If PFC is enabled on priorities on a port, when the port sends a pause message to the connected peer, the port uses the headroom buffers to store the packets that arrive between the time the port sends the pause message and the time the last packet arrives after the peer pauses traffic. The minimum value for lossless headroom buffers is 0 (zero) percent. (Lossless headroom buffers are the only buffers for which the recommended value can be less than 5 percent.)
- Lossy buffers—Shared buffer pool for all best-effort ingress traffic (best-effort unicast, multidestination, and strict-high priority traffic). The recommended minimum value for best-effort buffers is 5 percent.

The combined percentage values of the ingress lossless, lossless headroom, and best-effort buffer partitions must total exactly 100 percent. If the buffer percentages total more than 100 percent or less than 100 percent, the switch returns a commit error. If you explicitly configure an ingress shared buffer partition, you must explicitly configure all three ingress buffer partitions, even if the lossless headroom buffer partition has a value of 0 (zero) percent.

### Egress Shared Buffer Pool Partitions

You can configure three egress buffer pool partitions:

- Lossless buffers—Shared buffer pool for all lossless egress queues. The recommended minimum value for lossless buffers is 5 percent.
- Lossy buffers—Shared buffer pool for all best-effort egress queues (best-effort unicast, and strict-high priority queues). The recommended minimum value for best-effort buffers is 5 percent.
- Multicast buffers—Shared buffer pool for all multidestination (multicast, broadcast, and destination lookup fail) egress queues. The recommended minimum value for multicast buffers is 5 percent.

The combined percentage values of the egress lossless, lossy, and multicast buffer partitions must total exactly 100 percent. If the buffer percentages total more than 100 percent or less than 100 percent, the switch returns a commit error. All egress buffer partitions must be explicitly configured and should have a value of at least 5 percent. If you explicitly configure an egress shared buffer partition, you must explicitly configure all three egress buffer partitions, and each partition should have a value of at least 5 percent.

## Dedicated Port Buffer Pool and Buffer Allocation to Queues

The global dedicated buffer pool is memory that is allocated equally to each port, so each port receives a guaranteed minimum amount of buffer space. Dedicated buffers are not shared among ports. Each port receives an equal proportion of the dedicated buffer pool.

The amount of dedicated buffer space is not user-configurable and depends on the percentage of available nonreserved buffers allocated to the shared buffers. (The dedicated buffer space is equal to the minimum reserved port buffers plus the remainder of the available nonreserved buffers that are not allocated to the shared buffer pool.)

When traffic enters and exits the switch, the switch ports use their dedicated buffers to store packets. If the dedicated buffers are not sufficient to handle the traffic, the switch uses shared buffers. The only way to increase the dedicated buffer pool is to decrease the shared buffer pool from its default value of 100 percent of available unreserved buffers.

**Note:** If 100 percent of the available unreserved buffers are allocated to the shared buffer pool, the switch still reserves a minimum dedicated buffer pool.

The larger the shared buffer pool, the better the burst absorption across the ports. The larger the dedicated buffer pool, the larger the amount of dedicated buffer space for each port. The greater the dedicated buffer space, the less likely that congestion on one port can affect traffic on another port, because the traffic does not need to use as much shared buffer space.

### Allocating Dedicated Port Buffers to Queues

You can divide the dedicated buffer allocation for an egress port among the port queues by including the `buffer-size` statement in the scheduler configuration. This enables you to control the egress port dedicated buffer allocation on a per-port, per-queue basis. (For example, this enables you to dedicate more buffer space to queues that transport lossless traffic, or to stop the port from reserving buffers for queues that do not carry traffic.) Egress dedicated port buffer allocation is a hierarchical structure that allocates a global dedicated buffer pool evenly among ports, and then divides the allocation for each port among the port queues.

By default, ports divide their allocation of dedicated buffers among their egress queues in the same proportion as the default scheduler sets the minimum guaranteed transmission rates (the `transmit-rate` option) for traffic. Only the queues included in the default scheduler receive bandwidth and dedicated buffers, in the proportions shown in Table 1:

***Table 1: Default Dedicated Buffer Allocation to Egress Queues (Based on Default Scheduler)***

| Forwarding Class | Queue | Minimum Guaranteed Bandwidth (`transmit-rate`) | Proportion of Reserved Dedicated Port Buffers |
|---|---|---|---|
| best-effort | 0 | 5% | 5% |
| fcoe | 3 | 35% | 35% |
| no-loss | 4 | 35% | 35% |
| network-control | 7 | 5% | 5% |
| mcast | 8 | 20% | 20% |

In the default configuration, no egress queues other than the ones shown in Table 1 receive an allocation of dedicated port buffers.

**Note:** The switch uses hierarchical scheduling to control port and queue bandwidth allocation, as described in Understanding CoS Hierarchical Port Scheduling (ETS) and shown in Example: Configuring CoS Hierarchical Port Scheduling (ETS). For egress queue

buffer size configuration, when you attach a traffic control profile (includes the queue scheduler information) to a port, the dedicated egress buffers on the port are divided among the queues as configured in the scheduler.

If you do not want to use the default allocation of dedicated port buffers to queues, use the `buffer-size` option in the scheduler that is attached to the port to configure the queue allocation. You can configure the dedicated buffer allocation to queues in two ways:

- As a percentage—The queue receives the specified percentage of dedicated port buffers when the queue is mapped to the scheduler and the scheduler is attached to a port.
- As a remainder—After the port services the queues that have an explicit percentage buffer size configuration, the remaining dedicated port buffer space is divided equally among the other queues to which a scheduler is attached. (No default or explicit scheduler for a queue means no dedicated buffer allocation for that queue.) If you configure a scheduler and you do not specify a buffer size as a percentage, *remainder* is the default setting.

**Note:** The total of all of the explicitly configured buffer size percentages for all of the queues on a port cannot exceed 100 percent.

### Configuring Dedicated Port Buffer Allocation to Queues

In a port configuration that includes multiple forwarding class sets, with multiple forwarding classes mapped to multiple schedulers, the allocation of port dedicated buffers to queues depends on the mix of queues with buffer sizes configured as explicit percentages and queues configured with (or defaulted to) the `remainder` option.

The best way to demonstrate how using the percentage and remainder options affects dedicated port buffer allocation to queues is by showing an example of queue buffer allocation, and then showing how the queue buffer allocation changes when you add another forwarding class (queue) to the port.

Table 2 shows an initial configuration that includes four forwarding class sets, the five default forwarding classes (mapped to the five default queues for those forwarding classes), the `buffer-size` option configuration, and the resulting buffer allocation for each queue. Table 3 shows the same configuration after we add another forwarding class (best-effort-2, mapped to queue 1) to the best-effort forwarding class set. Comparing the buffer allocations in each table shows you how adding another queue affects buffer allocation when you use remainders and explicit percentages to configure the buffer allocation for different queues.

***Table 2: Egress Queue Dedicated Buffer Allocation (Example 1)***

| Forwarding Class Set (Priority Group) | Forwarding Class | Queue | Scheduler Buffer Size Configuration | Buffer Allocation per Queue (Percentage) |
|---|---|---|---|---|
| fc-set-be | best-effort | 0 | 10% | 10% |
| fc-set-lossless | fcoe | 3 | 20% | 20% |
| | no-loss | 4 | 40% | 40% |

| | | | | |
|---|---|---|---|---|
| fc-set-strict-high | network-control | 7 | remainder | 15% |
| fc-set-mcast | mcast | 8 | remainder | 15% |

In this first example, 70 percent of the egress port dedicated buffer pool is explicitly allocated to the best-effort, fcoe, and no-loss queues. The remaining 30 percent of the port dedicated buffer pool is split between the two queues that use the `remainder` option (network-control and mcast), so each queue receives 15 percent of the dedicated buffer pool.

Now we add another forwarding class (queue) to the best-effort priority group (fc-set-be) and configure it with a buffer size of *remainder* instead of configuring a specific percentage. Because a third queue now shares the remaining dedicated buffers, the queues that share the remainder receive fewer dedicated buffers, as shown in Table 3. The queues with explicitly configured percentages receive the configured percentage of dedicated buffers.

***Table 3: Egress Queue Dedicated Buffer Allocation with Another Remainder Queue (Example 2)***

| Priority Group (fc-set) | Forwarding Class | Queue | Scheduler Buffer Size Configuration | Buffer Allocation per Queue (Percentage) |
|---|---|---|---|---|
| fc-set-be | best-effort | 0 | 10% | 10% |
| | best-effort-2 | 1 | remainder | 10% |
| fc-set-lossless | fcoe | 3 | 20% | 20% |
| | no-loss | 4 | 40% | 40% |
| fc-set-strict-high | network-control | 7 | remainder | 10% |
| fc-set-mcast | mcast | 8 | remainder | 10% |

The two tables show how the port divides the dedicated buffer space that remains after servicing the queues that have an explicitly configured percentage of dedicated buffer space.

## Trade-off Between Shared Buffer Space and Dedicated Buffer Space

The trade-off between shared buffer space and dedicated buffer space is:

- Shared buffers provide better absorption of traffic bursts because there is a larger pool of dynamic buffers that ports can use as needed to handle the bursts. However, all flows that exhaust their dedicated buffer space compete for the shared buffer pool. A larger shared buffer pool means a smaller dedicated buffer pool, and therefore more competition for the shared buffer pool because more flows exhaust their dedicated buffer allocation. Too much shared buffer space results in

no single flow receiving very much shared buffer space, to maintain fairness when many flows contend for that space.

- Dedicated buffers provide guaranteed buffer space to each port. The larger the dedicated buffer pool, the less likely that congestion on one port affects traffic on another port, because the traffic does not need to use as much shared buffer space. However, less shared buffer space means less ability to dynamically absorb traffic bursts.

For optimal burst absorption, the switch needs enough dedicated buffer space to avoid persistent competition for the shared buffer space. When fewer flows compete for the shared buffers, the flows that need shared buffer space to absorb bursts receive more of the shared buffer because fewer flows exhaust their dedicated buffer space.

The default configuration and all of the recommended configurations allocate 100 percent of the user-configurable memory space to the global shared buffer pool because the amount of space reserved for dedicated buffers provides enough space to avoid persistent competition for dynamic shared buffers. This results in fewer flows competing for the shared buffers, so the competing flows receive more of the buffer space.

### Order of Buffer Consumption

The total buffer pool is divided into ingress and egress shared buffer pools and dedicated buffer pools. When traffic flows through the switch, the buffer space is used in a particular order that depends on the type of traffic.

On ingress, the order of buffer consumption is:

- Best-effort unicast traffic:
  1. Dedicated buffers
  2. Shared buffers
  3. Global headroom buffers (very small)
- Lossless unicast traffic:
  1. Dedicated buffers
  2. Shared buffers
  3. Lossless headroom buffers
  4. Global headroom buffers (very small)
- Multidestination traffic:
  1. Dedicated buffers
  2. Shared buffers
  3. Global headroom buffers (very small)

On egress, the order of buffer consumption is the same for unicast best-effort, lossless unicast, and multidestination traffic:

- Dedicated buffers
- Shared buffers

In all cases on all ports, the switch uses the dedicated buffer pool first and the shared buffer pool only after the dedicated buffer pool for the port or queue is exhausted. This reserves the maximum amount of dynamic shared buffer space to absorb traffic bursts.

## Default Buffer Pool Values

You can view the default or configured ingress and egress buffer pool values in KB units using the `show class-of-service shared-buffer` operational command. You can view the configured shared buffer pool values in percent units using the `show configuration class-of-service shared-buffer` operational command.

This section provides the default total buffer, shared buffer, and dedicated buffer values.



## Total Buffer Pool Size

The total buffer pool is common memory that has separate ingress and egress accounting, so the full buffer pool is available from both the ingress and egress perspective. The total buffer pool consists of the dedicated buffer space and the shared buffer space. The size of the total buffer pool is not user-configurable, but the allocation of buffer space to the dedicated and shared buffer pools is user-configurable.

On QFX3500 and QFX3600 switches, the combined total size of the ingress and egress buffer pools is approximately 9 MB (exactly 9360 KB).

On QFX5100 and EX4600 switches, the combined total size of the ingress and egress buffer pools is approximately 12 MB (exactly 12480 KB).

## Shared Buffer Pool Default Values

The QFX5100 and EX4600 switches have a larger shared buffer pool (12 MB) than QFX3500 and QFX3600 switches (9 MB). However, the allocation of shared buffer space to the individual ingress and egress buffer pools is the same on a percentage basis, even though the absolute values are different. For example, the default ingress lossless buffer is 9 percent of the total shared ingress buffer space on QFX5100, EX4600, QFX3500, and QFX3600 switches, even though the default absolute value of the ingress lossless buffer is 861.05KB on QFX5100 and EX4600 switches, and 648.18KB on QFX3500 and QFX3600 switches.

This section describes the default values in percent and in KB for the shared ingress and shared egress buffers.



### Shared Ingress Buffer Default Values

The QFX5100 and EX4600 switches have a larger shared ingress buffer than the QFX3500 and QFX3600 switches. Table 4 shows the default ingress shared buffer allocation values in KB units for QFX5100 and EX4600 switches.

***Table 4: QFX5100 and EX4600 Switch Default Shared Ingress Buffer Values (KB)***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 9567.19 KB | 861.05 KB | 4305.23 KB | 4400.91 KB |

Table 5 shows the default ingress shared buffer allocation values in KB units for QFX3500 and QFX3600

switches.

***Table 5: QFX3500 and QFX3600 Switch Default Shared Ingress Buffer Values (KB)***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 7202 KB | 648.18 KB | 3240.9 KB | 3312.92 KB |

Table 6 shows the default ingress shared buffer allocation values as percentages for QFX5100, EX4600, QFX3500, and QFX3600 switches. (If you change the default shared buffer allocation, you configure the change as a percentage.)

***Table 6: Default Shared Ingress Buffer Values (Percentage)***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 100% | 9% | 45% | 46% |

### Shared Egress Buffer Default Values

The QFX5100 and EX4600 switches have a larger shared egress buffer than the QFX3500 and QFX3600 switches. Table 7 shows the default egress shared buffer allocation values in KB units for QFX5100 and EX4600 switches.

***Table 7: QFX5100 and EX4600 Switch Default Shared Egress Buffer Values (KB)***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 8736 KB | 4368 KB | 2708.16 KB | 1659.84 KB |

Table 8 shows the default egress shared buffer allocation values in KB units.

***Table 8: QFX3500 and QFX3600 Switch Default Shared Egress Buffer Values (KB)***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 6656 KB | 3328 KB | 2063.36 KB | 1264.64 KB |

Table 9 shows the default egress shared buffer allocation values as percentages.

***Table 9: Default Shared Egress Buffer Values (Percentage)***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 100% | 50% | 31% | 19% |

### Dedicated Buffer Pool Default Values

The system reserves ingress and egress dedicated buffer pools that are divided equally among the switch ports. By default, the system allocates 100 percent of the available unreserved buffer space to the shared buffer pool. If you reduce the percentage of available unreserved buffer space allocated to the shared buffer pool, the remaining unreserved buffer space is added to the dedicated buffer pool allocation. You configure the amount of dedicated buffer pool space by reducing (or increasing) the percentage of buffer space allocated to the shared buffer pool. You do not directly configure the dedicated buffer pool allocation.

The default dedicated buffer pool values for QFX3500 and QFX3600 switches in KB units are:

- Ingress dedicated buffer—2158 KB
- Egress dedicated buffer—2704.0 KB

The default dedicated buffer pool values for QFX5100 switches in KB units are:

- Ingress dedicated buffer—2912.81 KB
- Egress dedicated buffer—3744 KB

## Shared Buffer Configuration Recommendations for Different Network Traffic Scenarios

The way you configure the shared buffer pool depends on the mix of traffic on your network. This section provides shared buffer configuration recommendations for five basic network traffic scenarios:

- Balanced traffic—The network carries a balanced mix of unicast best-effort, lossless, and multicast traffic. (This is the default configuration.)
- Best-effort unicast traffic—The network carries mostly unicast best-effort traffic.
- Best-effort traffic with Ethernet PAUSE (IEEE 802.3X) enabled—The network carries mostly best-effort traffic with Ethernet PAUSE enabled on the links.
- Best-effort multicast traffic—The network carries mostly multicast best-effort traffic.
- Lossless traffic—The network carries mostly lossless traffic (traffic on which PFC is enabled).

**Note:** Lossless traffic is defined as traffic on which you enable PFC to ensure lossless transport. Lossless traffic does not refer to best-effort traffic on a link on which you enable Ethernet PAUSE. Start with the recommended profiles for each network traffic scenario, and adjust them if necessary for your network traffic conditions.

**Caution:** Changing the buffer configuration is a disruptive event. Traffic stops on *all* ports until buffer reprogramming is complete. This includes changing the default configuration to one of the recommended configurations.

Because you configure buffer allocations in percentages, the recommended allocations for each network traffic scenario are valid for all QFX Series switches and the EX4600 switch. Use one of the following recommended shared buffer configurations for your network traffic conditions. Start with a recommended configuration, then make small adjustments to the buffer allocations to fine-tune the buffers if necessary as described in Optimizing Buffer Configuration .

- Balanced Traffic (Default Configuration)
- Best-Effort Unicast Traffic
- Ethernet PAUSE Traffic
- Best-Effort Multicast (Multidestination) Traffic
- Lossless Traffic

## Balanced Traffic (Default Configuration)

The default shared buffer configuration is optimized for networks that carry a balanced mix of best-effort unicast, lossless, and multidestination (multicast, broadcast, and destination lookup fail) traffic. The default class-of-service (CoS) configuration is also optimized for networks that carry a balanced mix of traffic.

We recommend that you use the default shared buffer configuration for networks that carry a balanced mix of traffic, especially if you are using the default CoS settings. Table 10 shows the default ingress shared buffer allocations:

***Table 10: Default Ingress Shared Buffer Configuration***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 100% | 9% | 45% | 46% |

Table 11 shows the default egress shared buffer allocations:

***Table 11: Default Egress Shared Buffer Configuration***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 100% | 50% | 31% | 19% |

## Best-Effort Unicast Traffic

If your network carries mostly best-effort (lossy) unicast traffic, then the default shared buffer configuration allocates too much buffer space to support lossless transport. Instead of wasting those buffers, we recommend that you use the following ingress shared buffer settings (see Table 12) and egress shared buffer settings (see Table 13):

***Table 12: Recommended Ingress Shared Buffer Configuration for Networks with Mostly Best-Effort Unicast Traffic***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 100% | 5% | 0% | 95% |

***Table 13: Recommended Egress Shared Buffer Configuration for Networks with Mostly Best-Effort Unicast Traffic***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 100% | 5% | 75% | 20% |

See Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Unicast Traffic for an example that shows you how to configure the recommended buffer settings shown in Table 12 and Table 13.

## Ethernet PAUSE Traffic

If your network carries mostly best-effort (lossy) traffic *and* enables Ethernet PAUSE on links, then the default shared buffer configuration allocates too much buffer space to the shared ingress buffer (Ethernet PAUSE traffic uses the dedicated buffers instead of shared buffers) and not enough space to the lossless-headroom buffers. We recommend that you use the following ingress shared buffer settings (see Table 14) and egress shared buffer settings (see Table 15):

***Table 14: Recommended Ingress Shared Buffer Configuration for Networks with Mostly Best-Effort Traffic and Ethernet PAUSE Enabled***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 70% | 5% | 80% | 15% |

***Table 15: Recommended Egress Shared Buffer Configuration for Networks with Mostly Best-Effort Traffic and Ethernet PAUSE Enabled***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 100% | 5% | 75% | 20% |

See Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Traffic on Links with Ethernet PAUSE Enabled for an example that shows you how to configure the recommended buffer settings shown in Table 12 and Table 13.

## Best-Effort Multicast (Multidestination) Traffic

If your network carries mostly best-effort (lossy) multicast traffic, then the default shared buffer configuration allocates too much buffer space to support lossless transport. Instead of wasting those buffers, we recommend that you use the following ingress shared buffer settings (see Table 16) and egress shared buffer settings (see Table 17):

***Table 16: Recommended Ingress Shared Buffer Configuration for Networks with Mostly Best -Effort Multicast Traffic***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 100% | 5% | 0% | 95% |

***Table 17: Recommended Egress Shared Buffer Configuration for Networks with Mostly Best-Effort Multicast Traffic***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|
| 100% | 5% | 20% | 75% |

See Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Multicast Traffic for an example that shows you how to configure the recommended buffer settings shown in Table 16 and Table 17.

## Lossless Traffic

If your network carries mostly lossless traffic, then the default shared buffer configuration allocates too much buffer space to support best-effort traffic. Instead of wasting those buffers, we recommend that you use the following ingress shared buffer settings (see Table 18) and egress shared buffer settings (see Table 19):

***Table 18: Recommended Ingress Shared Buffer Configuration for Networks with Mostly Lossless Traffic***

| Total Shared Ingress Buffer | Lossless Buffer | Lossless-Headroom Buffer | Lossy Buffer |
|---|---|---|---|
| 100% | 15% | 80% | 5% |

***Table 19: Recommended Egress Shared Buffer Configuration for Networks with Mostly Lossless Traffic***

| Total Shared Egress Buffer | Lossless Buffer | Lossy Buffer | Multicast Buffer |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 100% | 90% | 5% | 5% |

See Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Lossless Traffic for an example that shows you how to configure the recommended buffer settings shown in Table 18 and Table 19.

## Optimizing Buffer Configuration

Starting from the default configuration or from a recommended buffer configuration, you can further optimize the buffer allocation to best support the mix of traffic on your network. Adjust the settings gradually to fine-tune the shared buffer allocation. Use caution when adjusting the shared buffer configuration, not just when you fine-tune the ingress and egress buffer partitions, but also when you fine-tune the total ingress and egress shared buffer percentage. (Remember that if you allocate less than 100 percent of the available buffers to the shared buffers, the remaining buffers are added to the dedicated buffers). Tuning the buffers incorrectly can cause problems such as ingress port congestion.

**Caution:** Changing the buffer configuration is a disruptive event. Traffic stops on *all* ports until buffer reprogramming is complete.

The relationship between the sizes of the ingress buffer pool and the egress buffer pool affects when and where packets are dropped. The buffer pool sizes include the shared buffers and the dedicated buffers. In general, if there are more ingress buffers than egress buffers, the switch can experience ingress port congestion because egress queues fill before ingress queues can empty.

Use the `show class-of-service shared-buffer` operational command to see the sizes in kilobytes (KB) of the dedicated and shared buffers and of the shared buffer partitions.

For best-effort traffic (unicast and multidestination), the combined ingress lossy shared buffer partition and ingress dedicated buffers must be *less than* the combined egress lossy and multicast shared buffer partitions plus the egress dedicated buffers. This prevents ingress port congestion by ensuring that egress best-effort buffers are deeper than ingress best-effort buffers, and ensures that if packets are dropped, they are dropped at the egress queues. (Packets dropping at the ingress prevents the egress schedulers from working properly.)

For lossless traffic (traffic on which you enable PFC), the combined ingress lossless shared buffer partition and a reasonable portion of the ingress headroom buffer partition, plus the dedicated buffers, must be *less than* the total egress lossless shared buffer partition and dedicated buffers. (A reasonable portion of the ingress headroom buffer is approximately 20 to 25 percent of the buffer space, but this varies depending on how much buffer headroom is required to support the lossless traffic.) When these conditions are met, if there is ingress port congestion, the ingress port congestion triggers PFC on the ingress port to prevent packet loss. If the total lossless ingress buffers exceed the total lossless egress buffers, packets could be dropped at the egress instead of PFC being applied at the ingress to prevent packet loss.

**Note:** If you commit a buffer configuration for which the switch does not have sufficient resources, the switch might log an error instead of returning a commit error. After you commit a buffer configuration, check the syslog messages to ensure that the new buffer configuration did not fail to commit.

If the buffer configuration commits but you receive a syslog message that indicates the configuration cannot be implemented, you can:

- Reconfigure the buffers or reconfigure other parameters (for example, the PFC configuration, which affects the need for lossless headroom buffers and lossless buffers—the more priorities you pause, the more lossless and lossless headroom buffer space you need), then attempt the commit operation again.
- Roll back the switch to the last successful configuration.

If you receive a syslog message that says the buffer configuration cannot be implemented, you must take corrective action. If you do not fix the configuration or roll back to a previous successful configuration, the system behavior is unpredictable.

## General Buffer Configuration Rules and Considerations

Keep the following rules and considerations in mind when you configure the buffers:

- Changing the buffer configuration is a disruptive event. Traffic stops on *all* ports until buffer reprogramming is complete.
- If you configure the ingress or egress shared buffer percentages as less than 100 percent, the remaining percentage of buffer space is added to the dedicated buffer pool.
- The sum of all of the ingress shared buffer partitions must equal 100 percent. Each partition must be configured with a value of at least 5 percent except the lossless headroom buffer, which can have a value of 0 percent.
- The sum of all of the egress shared buffer partitions must equal 100 percent. Each partition must be configured with a value of at least 5 percent.
- Lossless and lossless headroom shared buffers serve traffic on which you enable PFC, and do not serve traffic subject to Ethernet PAUSE.
- The switch uses the dedicated buffer pool first and the shared buffer pool only after the dedicated buffer pool for a port or queue is exhausted.
- Too little dedicated buffer space results in too much competition for shared buffer space.
- Too much dedicated buffer space results in poorer burst absorption because there is less available shared buffer space.
- Always check the syslog messages after you commit a new buffer configuration.
- The optimal buffer configuration for your network depends on the types of traffic on the network. If your network carries less traffic of a certain type (for example, lossless traffic), then you can reduce the size of the buffers allocated to that type of traffic (for example, you can reduce the sizes of the lossless and lossless headroom buffers).

### Related Documentation

QFabric System, QFX Series standalone switches

- Example: Configuring Queue Schedulers

QFX Series standalone switches

- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Unicast Traffic
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Traffic on Links with Ethernet PAUSE Enabled
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Multicast Traffic
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Lossless Traffic
- Configuring Global Ingress and Egress Shared Buffers

Additional Information

- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Unicast Traffic
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Best-Effort Traffic on Links with Ethernet PAUSE Enabled
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Multicast Traffic
- Example: Recommended Configuration of the Shared Buffer Pool for Networks with Mostly Lossless Traffic

- Example: Configuring Queue Schedulers
- Configuring Global Ingress and Egress Shared Buffers

---

Published: 2014-07-23