Congestion Management

Traffic congestion management

Congestion management is a QoS traffic-management model that

- defines how traffic is organized into queues when congestion occurs on an interface
- defines how classified packets are mapped to those queues, and
- defines how queues are scheduled for transmission based on priority or policy rules.

Key features for congestion management

- **Low Latency Queueing with Strict Priority Queueing**: Ensures time-sensitive traffic (such as voice or real-time video) is placed in a strict-priority queue so it is serviced first, minimizing latency and jitter while other traffic is still managed fairly.
- **Traffic Shaping**: Smooths traffic bursts by buffering and pacing packets so that the egress rate matches downstream link speeds or contracted rates, helping prevent congestion caused by data-rate mismatches.
- **Traffic Policing**: Enforces a maximum traffic rate on an interface using token-bucket algorithms and drops (or marks) packets that exceed the configured rate, helping protect bandwidth and maintain service quality across classes of traffic.

The table lists the key differences among Low Latency Queuing (with Strict Priority), Traffic Shaping, and Traffic Policing to help you understand how each feature manages traffic under congestion.

Table 1: Key differences among low latency queuing (with strict priority), traffic shaping, and traffic policing

Aspect	LLQ with strict priority queuing	Traffic shaping	Traffic policing
Primary goal	Prioritize time-sensitive traffic to minimize latency and jitter.	Match traffic output rate to downstream speed or contracted profile.	Enforce a maximum allowed traffic rate using token-bucket metering.

Aspect	LLQ with strict priority queuing	Traffic shaping	Traffic policing
Behavior with excess traffic	Strict priority can cause lower-priority queues to not get serviced if priority traffic is heavy.	Buffers excess packets and sends them later. It *smooths* bursts.	Does not buffer. Packets that exceed rate may be marked or dropped.
Traffic direction	Egress	Egress only (explicit).	Ingress (egress policing is not supported).
Impact on other traffic	May starve lower-priority queues unless an optional shaper is used.	Prevents downstream bottlenecks by smoothing bursts; does not drop traffic by design.	Can drop packets based on exceed/violate actions; partitions traffic into service classes.
Handling of traffic bursts	Priority traffic bypasses others; does not inherently smooth bursts or buffer.	Buffers bursts and releases later to match rate profile.	Bursts may trigger exceed/violate events; uses burst parameters.

Low-latency queuing with strict priority queuing

Low-latency queuing with strict priority queuing is a QoS congestion management feature that

- integrates strict priority queuing into the Class-Based Weighted Fair Queuing (CBWFQ) model
- ensures delay-sensitive traffic is always dequeued ahead of other traffic classes, and
- minimizes latency and jitter for critical applications such as voice and real-time video.

CBWFQ is a QoS scheduling model that assigns bandwidth to traffic classes based on configured weights and ensures each class receives its allocated portion of service during congestion.

Priority Queuing is a scheduling method that dequeues packets from the highest-priority queue before any other queue, allowing delay-sensitive traffic to pre-empt lower-priority classes; in strict priority mode, lower-priority queues may not be serviced if high-priority traffic is continuously present.

Table 2: Feature History Table

Feature Name	Release Information	Feature Description
Priority Propagation on Egress Queues	Release 25.3.1	Introduced in this release on: Centralized Systems (8400 [ASIC: K100])(select variants only*) * This feature is now supported on Cisco 8404-SYS-D routers.

Feature Name	Release Information	Feature Description
Priority Propagation on Egress Queues	Release 25.1.1	Introduced in this release on: Fixed Systems (8700 [ASIC: K100], 8010 [ASIC: A100])(select variants only*) This feature is supported on: • 8712-MOD-M • 8011-4G24Y4H-I
Low Latency Queueing with Strict Priority Queueing	Release 25.1.1	Introduced in this release on: Fixed Systems (8010 [ASIC: A100])(select variants only*) *This feature is supported on Cisco 8011-4G24Y4H-I routers.
Low Latency Queueing with Strict Priority Queueing	Release 24.4.1	Introduced in this release on: Fixed Systems (8700) (select variants only*) This feature ensures that time-sensitive traffic receives dedicated priority, ensuring minimal latency and jitter while congestion for other types of traffic is managed fairly, guaranteeing high performance for critical applications and efficient overall traffic management. *This functionality is now supported on Cisco 8712-MOD-M routers.

Feature Name	Release Information	Feature Description
Priority Propagation on Egress Queues	Release 24.3.1	Introduced in this release on: Fixed Systems (8200 [ASIC: P100], 8700 [ASIC: P100])(select variants only*); Modular Systems (8800 [LC ASIC: P100])(select variants only*) You can now ensure that high-priority traffic is consistently prioritized across multiple sub-interfaces on the same network port, enhancing the performance of critical applications. This is achieved by enabling support for egress queuing policy maps that allow high-priority P1 traffic (on class TC7) to take precedence over non-P1 traffic across sub-interfaces. This feature ensures that critical traffic, such as voice or real-time video, is always given the highest priority, regardless of the sub-interface it is associated with, thereby maintaining optimal network performance and reducing latency for essential services. *This feature is supported on: • 8212-48FH-M • 8711-32FH-M • 88-LC1-36EH • 88-LC1-12TH24FH-E • 88-LC1-52Y8H-EM

Feature Name	Release Information	Feature Description
Low Latency Queueing with Strict Priority Queueing	Release 24.3.1	Introduced in this release on: Modular Systems (8800 [LC ASIC: P100]) (select variants only*), Fixed Systems (8200) (select variants only*), Fixed Systems (8700 (P100, K100)) (select variants only*) This feature ensures that time-sensitive traffic receives dedicated priority, ensuring minimal latency and jitter while congestion for other types of traffic is managed fairly, guaranteeing high performance for critical applications and efficient overall traffic management. *This feature is supported on: • 88-LC1-12TH24FH-E • 88-LC1-52Y8H-EM • 8212-48FH-M • 8711-32FH-M

Feature Name	Release Information	Feature Description
Low Latency Queueing with Strict Priority Queueing	Release 24.2.11	Introduced in this release on: Modular Systems (8800 [LC ASIC: P100])(select variants only*) With this feature, time-sensitive traffic receives dedicated priority, ensuring minimal latency and jitter while congestion for other types of traffic is managed fairly, guaranteeing high performance for critical applications and efficient overall traffic management. You can also configure low latency queueing with strict priority queueing at the subinterface level, ensuring that the strict priority policy applies only to traffic on that specific subinterface, not the entire physical interface. This configuration is useful in scenarios where multiple services share the same physical link but have different QoS requirements. *This feature is supported on 88-LC1-36EH.

Guidelines for configuring low-latency queuing with strict priority queuing

Priority and scheduling behavior

- Only priority levels 1 to 7 are supported, where 1 is the highest and 7 is the lowest priority; the default CoSQ 0 has the lowest priority among all classes.
- Egress policing is not supported, and in strict priority mode, heavy priority traffic can prevent lower-priority queues from being serviced.
- The platform enforces a strict class hierarchy, where higher traffic classes take precedence over lower ones. This hierarchy can be redefined through an egress queuing policy map when configuring LLQ with strict priority.
- When operating in strict-priority mode, egress scheduling follows the default class order TC7 >TC6 > TC5 > TC4 >TC3 > TC2 > TC1 > TC0, unless a queuing policy map overrides this behavior.

Shaping and rate control

- You can configure shape average and queue-limit commands along with the priority command.
- An optional shaper can be configured to cap the maximum traffic rate and prevent starvation of lower-priority traffic classes.

Interface-level configuration

LLQ with strict priority queuing can also be applied at the subinterface level, allowing strict priority behavior to affect only the specific subinterface rather than the entire physical interface.

Configure low-latency queuing with strict priority queuing

Before you begin

Ensure that you have an existing QoS class map that classifies the intended traffic.

Follow these steps to configure LLQ with strict priority on an egress interface.

Procedure

Step 1 Create or modify a policy-map that applies LLQ with strict priority.

This policy-map defines the class of traffic that receives strict priority handling.

Example:

```
Router# configure
Router(config)# policy-map test-priority-1
Router(config-pmap)# class qos1
Router(config-pmap-c)# priority level 7
Router(config-pmap-c)# exit
Router(config-pmap)# exit
```

The policy-map now includes a strict priority class with the specified priority level.

Step 2 (Optional) Shape the strict priority class to a specific bit rate.

Shaping prevents lower-priority queues from starving during high volumes of strict priority traffic.

Example:

```
Router(config-pmap-c)# shape average percent 40
```

Step 3 Apply the policy-map to the egress interface.

Attaching the policy-map makes strict priority queuing active for egress traffic on the specified interface.

Example:

```
Router(config)# interface HundredGigE 0/6/0/18
Router(config-if)# service-policy output test-priority-1
Router(config-if)# no shutdown
Router(config-if)# commit
```

The LLQ strict priority policy is now active on the selected output interface.

Step 4 Verify the strict priority behavior on the interface.

Confirm that the strict priority class (for example, HP7) is visible in the queue statistics.

Example:

```
Router# show qos interface HundredGigE 0/6/0/18 output
```

```
NOTE:- Configured values are displayed within parentheses
Interface HundredGigE0/6/0/18 ifh 0x3000220  -- output policy
NPU Id:                        3
Total number of classes:       3
Interface Bandwidth:           100000000 kbps
VOQ Base:                      11176
VOQ Stats Handle:              0x88550ea0
Accounting Type:               Layer1 (Include Layer 1 encapsulation and above)
------------------------------------------------------------------------------
Level1 Class (HP7)                       =   qos1
Egressq Queue ID                         =   11177 (HP7 queue)
TailDrop Threshold                       =   125304832 bytes / 10 ms (default)
WRED not configured for this class

Level1 Class (HP6)                       =   qos-2
Egressq Queue ID                         =   11178 (HP6 queue)
TailDrop Threshold                       =   125304832 bytes / 10 ms (default)
WRED not configured for this class

Level1 Class (class-default)
Egressq Queue ID                         =   11176 (Default LP queue)
Queue Max. BW.                           =   101803495 kbps (default)
Queue Min. BW.                           =   0 kbps (default)
Inverse Weight / Weight                  =   1 (BWR not configured)
TailDrop Threshold                       =   1253376 bytes / 10 ms (default)
WRED not configured for this class
```

The interface shows the strict priority class and confirms that LLQ with strict priority is functioning as expected.