Congestion Monitoring and Optimization

- Congestion monitoring and optimization, on page 1

Congestion monitoring and optimization

The table summarizes monitoring and optimization techniques, with brief descriptions and highway traffic analogies to illustrate how each method works in practice.

Technique	Short Description	Highway Traffic Analogy
Sharing of VOQ statistics counters, on page 1	Allows multiple VOQs to share a set of counters to conserve hardware resources while still tracking traffic and drops.	Like using a shared traffic camera to monitor several nearby intersections instead of one camera per intersection.
Traffic class queue high water marks monitoring, on page 3	Records peak queue occupancy and delay for each traffic class to help identify congestion patterns.	Like noting the worst traffic buildup times at each highway exit to plan better traffic control.
VOQ evictions to High Bandwidth Memory , on page 11	Shows which VOQs have been moved from on-chip SMS to high-bandwidth memory due to congestion.	Like tracking which highway lanes have been diverted onto an overflow road during heavy traffic.

Sharing of VOQ statistics counters

Sharing of Virtual Output Queue (VOQ) counters is a congestion optimization technique that

- reduces the counter consumption required to monitor enqueued and dropped packets across VOQs and,
- allows multiple VOQs to share a set of counters, thus conserving hardware counter resources.

Key attributes of sharing VOQ statistics counters

- Each VOQ group that shares counters reports both enqueued and dropped packets (in packets and bytes).
- Sharing options include 1 (no sharing), 2, 4, or 8 VOQs per counter set.

- Sharing VOQ counters enables scalability in large-scale deployments by mitigating device counter capacity limits

Guidelines for configuring VOQ counter sharing

Configure VOQ counter sharing before applying egress policies

To conserve VOQ counter resources and prevent counter exhaustion, configure VOQ counter sharing (using **hw-module profile stats voqs-sharing-counters**) before applying any egress queuing policy.

If you do not configure VOQ counter sharing first, the egress queuing policies may use default or previously set counter-sharing configurations, which could exceed hardware limits or misrepresent queue statistics.

Delete all existing egress queuing policies and reload all nodes (using router**#reload location all**) to apply the new shared counter configuration successfully.

Shared VOQ counter modes

Use the **hw-module profile stats voqs-sharing-counters** command to configure VOQ counter-sharing. The available VOQ counter-sharing modes and their intended use cases are described in this table.

VOQ counter mode	Description	Use case
1	One counter per VOQ	Highest granularity. Use when counters are abundant.
2	Two VOQs share one counter set	This mode provides a good balance between granularity and resource usage.
4	Four VOQs share one counter set	This mode is recommended when conserving counter space.

Configure shared VOQ counters

Use this task to specify how many VOQs share a set of counters, optimizing counter resource usage.

This configuration is hardware-level and affects all VOQs across the device. It must be applied before deploying queuing policies.

Before you begin

- Ensure no egress queuing policy is applied on any interface.
- Prepare to reload all nodes

Procedure

Step 1 Choose one of the available options (1, 2, or 4) based on your needs for granularity versus counter resource conservation.

Example:

```
Router(config)#hw-module profile stats voqs-sharing-counters 2
Router(config)#commit
```

Step 2 Reload all nodes on your router. This step is required for the configuration to take effect.

Example:

```
Router#reload location all
```

This example sets the mode where 2 VOQs share counters.

Step 3 Verify the VOQ statistics.

Example:

```
Router#show controllers npu stats voq ingress interface hundredGigE 0/0/0/16 instance all location
0/RP0/CPU0

Interface Name Interface Handle Location Asic Instance
= = = =
Hu0/0/0/16 f0001b0 0/RP0/CPU0 0
VOQ Base = 10288 Port Speed(kbps) Local Port VOQ Mode
= = =
100000000 local 8 Shared Counter Mode = 2
ReceivedPkts ReceivedBytes DroppedPkts DroppedBytes
TC_{0,1} = 114023724 39908275541 113945980 39881093000
TC_{2,3} = 194969733 68239406550 196612981 68814543350
TC_{4,5} = 139949276 69388697075 139811376 67907466750
TC_{6,7} = 194988538 68242491778 196612926 68814524100
```

- The **HundredGigE 0/0/0/16** interface is actively collecting VOQ statistics.
- The VOQ counter sharing mode is set to 2, meaning two VOQs are sharing a single counter set, as configured.
- The counters are actively capturing received and dropped packets and bytes across various traffic classes, showing traffic flow and congestion events

Traffic class queue high water marks monitoring

Traffic class queue high water marks are metrics that

- indicate the peak congestion levels experienced by traffic class queues on an egress interface and,
- reflect the maximum queue occupancy and maximum queue delay, helping identify potential bottlenecks.
- Queue occupancy is the amount of buffer space utilized by a traffic class at a given moment.
- Queue delay is the duration that packets wait in the queue before transmission.

For example, by using high water mark data, you can determine whether certain traffic classes are consistently congested and adjust QoS policies.

Table 1: Feature History Table

Feature Name	Release Information	Feature Description
Traffic Class Queue High Water Marks Monitoring	Release 25.3.1	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*) *This feature is supported on: • 8711-32FH-M • 8212-48FH-M • 88-LC1-36EH
Traffic Class Queue High Water Marks Monitoring	Release 24.2.11	Introduced in this release on Cisco 8000 Series Routers with Cisco Silicon One Q200 network processors. The Cisco 8608 router is not currently supported. This feature monitors egress interface traffic class queues and records the queue occupancy and queue delay high water marks information for each traffic class. This information includes the virtual output queue that experienced the high water mark and a timestamp indicating when the high water mark was recorded. You can use this data to identify network bottlenecks and prevent traffic congestion. This feature introduces these changes: CLI: • **hw-module profile qos high-water-marks** • **show controllers npu qos high-water-marks** • **clear controller npu qos high-water-marks** YANG Data Models: • cisco-IOS-XR-ofa-npu-qos-oper.yang • cisco-IOS-XR-ofa-npu-qos-act.yang • cisco-IOS-XR-um-8000-hw-module-profile-cfg.yang • cisco-IOS-XR-npu-hw-profile-cfg.yang

High water mark data types

High water mark data types are key metrics that reflect the maximum observed values for queue occupancy and queue delay on an interface's traffic class queue.

The table lists and describes high water mark data types.

Data Type	Description
Queue occupancy high water marks	Maximum queue occupancy observed in kilobytes and percentage of queue size; includes timestamp and Virtual Output Queue (VOQ) identifier.
Queue delay high water marks	Maximum queue delay observed in nanoseconds (ns); includes timestamp, queue occupancy at the time, and VOQ identifier.

Recommendations and requirements for monitoring traffic class queue high water marks

Priority Flow Control compatibility

This feature can co-exist with Priority Flow Control (PFC). When using PFC in **buffer-internal** mode, configure **ecn** and **max-threshold** values.

Bundle interface monitoring

High water marks are monitored per bundle member interface, not the bundle interface itself.

Subinterface contributions

Packets egressing from subinterfaces and bundle subinterfaces contribute to the high water marks on the corresponding parent interface.

Queue occupancy monitoring support

Queue occupancy monitoring is supported only for packets sourced from Q200-based network processors (NPUs).

Queue delay monitoring support

Packets from non-Q200 NPUs are also supported for queue delay monitoring.

Reload requirement

Ensure that you manually reload the chassis or all line cards after enabling or disabling this feature.

VOQ mapping factors

The number of VOQs per traffic class depends on the VOQ mode of the router and the number of slices.

Occupancy percentage basis

Queue occupancy percentages are calculated using the maximum queue-limit, not custom limits.

Configure traffic class queue high water marks monitoring

Use this task to enable monitoring of peak congestion levels for egress traffic class queues.

This task helps identify congestion patterns on supported Cisco 8000 Series Routers by using high water mark data.

Before you begin

- Verify that your router supports Q200-based line cards.
- Plan for a reload of the chassis or all line cards to activate or deactivate the feature.

Use these steps to configure the traffic class queue high water marks monitoring feature.

Procedure

Step 1 Enable the monitoring feature for egress traffic class queues.

Example:

```
Router(config)#hw-module profile qos high-water-marks
Router(config)#commit
```

Step 2 Reload the chassis or all line cards to apply the changes.

Example:

```
Router(admin)#reload location all
```

Step 3 Verify that the feature is enabled

Example:

```
Router#show controllers npu qos high-water-marks interface all
Mon Jun  3 06:02:50.138 UTC
Not supported or not enabled on location 0/0/CPU0
RP/0/RP0/CPU0:ios#
```

The output indicates that the feature is not enabled on location **0/0/CPU0**.

Step 4 (Optional) View the monotonic high water marks for all traffic classes.

Example:

```
Router#show controllers npu qos high-water-marks interface fourHundredGigE 0/0/0/11

Interface Name        =   FH0/0/0/11
Interface Handle      =        0x1F8
System Port Gid       =           96
Asic Instance         =            0

       Queue Occupancy High Water Marks                                    Queue Delay High Water Marks

       Max Occupancy    Queue     Src Sys Port                           Max Queue   Occupancy     Src
 Sys Port
       %    kilobytes   Delay ns  Slot/NPU/Slc/Gid Timestamp             Delay ns    %  kilobytes
Slot/NPU/Slc/Gid Timestamp
-------------------------------------------------------------------------------------------------------------------

TC_0 =   6.00 30965     73728     0/0/2/40         04/08/23 08:39:35   102400     3.00  15482
0/0/1/44       04/05/23 12:22:05
TC_1 =   0.00 0         0         0/0/0/0          -                   0          0.00  0
0/0/0/0        -
TC_2 =  25.00 129024    1114112   0/0/0/48         04/07/23 01:10:23   1179648    15.00 77414
0/0/0/48       04/07/23 21:40:53
```

```
TC_3 =  70.00 361267    8912896   0/1/1/56         04/02/23 08:41:44   8912896    70.00 361267
0/1/1/58       04/02/23 08:41:44
TC_4 =  40.00 206438    2228224   3/0/2/4          04/09/23 06:38:35   2359296    25.00 129024
3/0/2/5        04/04/23 18:30:56
TC_5 =   0.00 0         0         0/0/0/0          -                   0          0.00  0
0/0/0/0        -
TC_6 =  78.00 599       6437184   3/1/0/24         04/10/23 16:35:00   8628192    64.00 492
7/0/2/76       04/10/23 16:35:00
TC_7 =  25.00 129024    139264    3/0/0/14         04/06/23 08:39:41   155648     15.00 77414
0/2/2/66       04/08/23 08:39:41

[ -----------     Occupancy High Water Marks     ---------------    ]  [ --------------   Delay High
 Water Marks  -----------  ]
```

The output displays monotonic high water marks data for all traffic classes on interface **fourHundredGigE 0/0/0/11**, recorded since bootup or after the most recent clear operation.

Tip
Monotonic high water marks are displayed if neither the **monotonic** or **periodic** keyword is used.

Step 5 (Optional) View the monotonic high water marks for a single traffic class.

Example:

```
Router#show controllers npu qos high-water-marks monotonic interface fourHundredGigE 0/0/0/2
traffic-class 5

Interface Name        =    FH0/0/0/2
Interface Handle      =    0xF000120
System Port Gid       =            6
Asic Instance         =            0

       Queue Occupancy High Water Marks                                  Queue Delay High Water Marks


       Max Occupancy    Queue     Src Sys Port                          Max Queue  Occupancy     Src
 Sys Port
       %     kilobytes  Delay ns  Slot/NPU/Slc/Gid Timestamp            Delay ns   %  kilobytes
Slot/NPU/Slc/Gid Timestamp
-------------------------------------------------------------------------------------------------------------------

TC_5 = 40.00 206438     6815744   3/0/0/15         11/11/23 17:43:30   1811939328  25.00 129024
7/1/2/89         11/27/23 11:21:26
[ -----------     Occupancy High Water Marks     ---------------    ]  [ --------------   Delay High
 Water Marks  ----------- ]
```

The output displays monotonically increasing high water marks data for traffic class 5 on interface **fourHundredGigE 0/0/0/2**, recorded since bootup or after the most recent clear operation. .

Step 6 (Optional) View the periodic high water marks for a single traffic class.

Example:

```
Router#show controllers npu qos high-water-marks periodic last 3 interface fourHundredGigE 0/0/0/5
traffic-class 7

Interface Name        =    FH0/0/0/5
Interface Handle      =    0xF000138
System Port Gid       =            9
Asic Instance         =            0

                                          Queue Occupancy High Water Marks           Queue Delay
 High Water Marks
```

```
                                                Max Occupancy   Queue    Src Sys Port      Max Queue
Occupancy       Src Sys Port
       Interval Start      End                  %     kilobytes Delay ns Slot/NPU/Slc/Gid  Delay ns
%     kilobytes Slot/NPU/Slc/Gid
-------------------------------------------------------------------------------------------------------------------------------
TC_7 = 1 12/01/23 17:46:30 12/01/23 17:46:59 50.00 258048    34680274 7/1/2/91           34680274
50.00 258048    7/1/2/91
       2 12/01/23 17:45:58 12/01/23 17:46:30 60.00 309657    52296260 0/2/1/68           61348106
50.00 258048    7/1/2/91
       3 12/01/23 17:45:30 12/01/23 17:45:58 40.00 206438    15290430 0/2/1/68           15290430
40.00 206438    0/2/1/68

                                                [----- Occupancy High Water Marks ----- ]  [ -------
Delay High Water Marks ------  ]
```

The output displays high water marks data for traffic class 7 on interface **fourHundredGigE 0/0/0/5** for the last three periodic collection intervals.

Descriptions of show command output fields for queue high water marks

This section describes the fields in the show command outputs that help you monitor queue high water marks for different traffic classes and scenarios.

The table presents the common fields displayed in the show command output for queue high water marks.

Table 2: Common Fields

Field	Description
Interval Start and End (periodic only)	Displays the periodic collection interval number, and the start and end time of that interval.
TC_Number = (Number range is 0-7)	Identifies the traffic class associated with the high water mark data. For periodic output, the field appears only for the traffic class's first interval.

The table describes queue occupancy fields in the show command output, including maximum occupancy values, delay, source port details, and timestamps.

These fields report the maximum queue occupancy values observed for a traffic class. Use them to determine how full the queue became and the conditions under which that peak occurred.

Table 3: Queue Occupancy Fields

Field	Description
Max Occupancy %	Displays the maximum queue occupancy for this traffic class as a percentage of the total queue size. Because of limited queue quantization thresholds provided by the NPU, this value is an estimate and may differ from the actual maximum occupancy.

Field	Description
Max Occupancy kilobytes	Displays the maximum queue occupancy for this traffic class in kilobytes. The value is calculated assuming that all buffers are fully packed (for example, 384 bytes per Shared Memory System buffer). As a result, the displayed value may be higher than the actual number of kilobytes queued.
Queue Delay ns	Displays the delay, in nanoseconds, when the maximum queue occupancy high water mark occurred.
Src Sys Port Slot/NPU/Slc/Gid	Identifies the slot, NPU, slice, and global identifier (GID) of the virtual output queue where the queue occupancy high water mark occurred. The GID is the global identifier of the source system port whose packet was dequeued when the maximum occupancy was reached. In most cases, all ports on a slice share a virtual output queue. The identified source port may not be the only contributor; other ports sharing the same queue could also have contributed to the burst of packets. **Note** In fair-4 or fair-8 VOQ mode, each source port has its own virtual output queue. In Priority Flow Control (PFC) **buffer-internal** mode, a port may share a virtual output queue with other ports in the same interface group (IFG). Each slice has two IFGs. Use the **show controllers npu voq-usage** command to verify which other ports share a virtual output queue with the identified source port.
Timestamp (monotonic only)	Displays the timestamp when the maximum queue occupancy high water mark was recorded. The timestamp reflects when the NPU reported the data, not when the event occurred. Because the NPU is queried every 30 seconds, the timestamp shows the end of the 30-second window during which the high water mark occurred. For example, a timestamp of 16:56:44 indicates that the event occurred sometime between 16:56:14 and 16:56:44.

The table describes the queue delay fields in the show command output, including maximum delay, occupancy details, source port identifiers, and timestamps.

These fields report the maximum queue delay experienced by a traffic class. Use them to understand the delay impact and the corresponding queue occupancy at the time of the peak delay.

Table 4: Queue Occupancy Fields

Field	Description
Max Queue Delay ns	Displays the maximum delay experienced by this traffic class, in nanoseconds.
Queue Occupancy %	Displays the queue occupancy as a percentage of the total queue size when the maximum queue delay high water mark occurred. Because of limited queue quantization thresholds provided by the NPU, this value is an estimate and may differ from the actual maximum occupancy.
Queue Occupancy kilobytes	Displays the queue occupancy in kilobytes at the time the maximum queue delay high water mark occurred. The value is calculated assuming all buffers are fully packed (for example, 384 bytes per Shared Memory System buffer). As a result, the displayed value may be higher than the actual number of kilobytes queued.
Src Sys Port Slot/NPU/Slc/Gid	Identifies the slot, NPU, slice, and global identifier (GID) of the virtual output queue where the maximum queue delay high water mark occurred. The GID identifies the source system port whose packet was dequeued when the maximum delay was reached. In most cases, all ports on a slice share a virtual output queue. The identified source port may not be the only contributor; other ports sharing the same queue could also have contributed to the burst of packets. **Note** In fair-4 or fair-8 VOQ mode, each source port has its own virtual output queue. In Priority Flow Control **buffer-internal** mode, a port may share a virtual output queue with other ports in the same interface group (IFG). Each slice has two IFGs. Use the **show controllers npu voq-usage** command to verify which other ports share a virtual output queue with the identified source port.
Timestamp (monotonic only)	Displays the timestamp when the maximum delay high water mark was recorded. The timestamp reflects when the NPU reported the data, not when the event occurred. Because the NPU is queried every 30 seconds, the timestamp shows the end of the 30-second window during which the event occurred. For example, a timestamp of 16:56:44 indicates that the high water mark was observed sometime between 16:56:14 and 16:56:44.

VOQ evictions to High Bandwidth Memory

VOQ evictions to High Bandwidth Memory (HBM) are buffer-management optimizations that

- offload virtual output queues (VOQs) from the on-chip Shared Memory System (SMS) to external High Bandwidth Memory (HBM),
- are triggered when local occupancy thresholds are exceeded, and
- help preserve packet buffering capacity during periods of high traffic.

Table 5: Feature History Table

Feature Name	Release Information	Feature Description
ViewVOQs Evicted to HBM	Release 24.2.11	The newly introduced command displays the virtual output queues (VOQs) that are evicted to the High Bandwidth Memory (HBM) and the VOQs' HBM buffer usage details. You can use this information when monitoring and debugging congestion scenarios. This feature introduces the **show controllers npu voq in-extended-memory instance** command. This feature modifies the Cisco-IOS-XR-[illegible] data model. (see GitHub, YANG Data Models Navigator)

How VOQ eviction to HBM works

When a VOQ in the Shared Memory System (SMS) exceeds its congestion threshold, it may be evicted to High Bandwidth Memory (HBM) to avoid packet drops.

Summary

This process dynamically moves VOQs between SMS and HBM based on buffer pressure and traffic conditions. The HBM acts as overflow memory for congestion scenarios.

Workflow

These stages describe how VOQ eviction to HBM works.

1. Normal operation: The NPU places packets into the SMS buffer for every VOQ.
2. Congestion detection: When SMS usage for a VOQ exceeds the congestion threshold, the VOQ is marked for eviction.

3. Eviction to HBM: The VOQ is offloaded to HBM, which serves as extended buffer memory.
4. HBM buffer monitoring: As traffic drains from the HBM-based VOQ, the system tracks usage and adapts the VOQ size dynamically.
5. Reversion to SMS: When HBM pressure reduces and VOQ activity stabilizes, the VOQ is returned to the SMS buffer.

Result

This system-level process helps avoid packet loss by using adaptive buffering and reducing memory contention among active VOQs.

Best practices for viewing VOQ eviction to HBM

Requirement for automation scripts

Do not use the **show controllers npu voq in-extended-memory instance** command in an automation script.

Guidelines for viewing VOQ eviction to HBM

Use these guidelines to correctly interpret Virtual Output Queue (VOQ) eviction behavior under different Priority Flow Control (PFC) modes.

- Priority handling with PFC **buffer-extended** mode: In PFC **buffer-extended** mode, associated VOQs are evicted to HBM on priority, while the remaining VOQs stay in SMS.
- Retention behavior with PFC **buffer-internal mode** mode: In PFC **buffer-internal mode**, associated VOQs are retained in Shared Memory System (SMS), while the remaining VOQs are evicted to High Bandwidth Memory (HBM).
- Eviction policy when PFC is disabled: If PFC isn't enabled on a device, VOQs are evicted to the HBM based on the VOQs' age and buffer usage.

View evicted VOQs to HBM

Use this task to identify congestion events and diagnose buffer pressure by viewing Virtual Output Queues (VOQs) evicted to High Bandwidth Memory (HBM) on your Cisco 8000 router.

This task is useful when analyzing congestion symptoms or buffer-related behavior during traffic surges.

Before you begin

- You must have access to the router in privileged EXEC mode.
- Ensure that the router is running Cisco IOS XR Release 24.2.11 or later.

Procedure

View the VOQs evicted to the HBM for all device instances.

Example:

```
Router#show controllers npu voq in-extended-memory instance all location 0/6/CPU0

* Use this CLI with caution.
```

```
* This should not be integrated with any automation scripts.

Total Entries  0  | Slot 6   | Device  0  | Percent in Evict Voq Buff(s) 0.000000
-------------------------------------------------------------------------------
Egress Interface  | VOQ_Base  | TC  | Slice  |  Buff_Usage |  In_Bytes     |

Total Entries  2  | Slot 6   | Device  1  | Percent in Evict Voq Buff(s) 0.004883
-------------------------------------------------------------------------------
Egress Interface  | VOQ_Base  | TC  | Slice  |  Buff_Usage  |  In_Bytes  |
-------------------------------------------------------------------------------
    FH0/6/0/0           28672       2     0         14722          87837459
    FH0/6/0/3           28728       2     0         1049           6335369

Total Entries  0  | Slot 6   | Device  2  | Percent in Evict Voq Buff(s) 0.000000
-------------------------------------------------------------------------------
   Egress Interface  | VOQ_Base  | TC  | Slice  |  Buff_Usage  |  In_Bytes  |
```

The output displays VOQs that are evicted to the HBM for node **0/6/cpu0** and all instances. In this case, VOQs from device instance **Device 1** are evicted to the HBM.

Output fields for evicted VOQs:

- **Egress Interface**: The egress interface of the virtual output queue.
- **VOQ_Base**: Base VOQ ID.
- **TC**: Traffic Class number.
- **Slice**: Source slice number.
- **Buff_Usage and In_Bytes**: Buffer usage in blocks and in bytes, respectively

---

You can view VOQs evicted to HBM and correlate congestion events with traffic patterns.