Configure Priority Flow Control

Priority Flow Control Overview

Table 1: Feature History Table

Feature Name	Release Information	Feature Description
Priority Flow Control on Cisco 8808 and Cisco 8812 Modular Chassis Line Cards	Release 7.5.3	Priority Flow Control is now supported on the following line card in the buffer-internal mode: • 88-LC0-34H14FH The feature is supported in the buffer-internal and buffer-extended modes on: • 88-LC0-36FH Apart from the buffer-external mode, support for this feature now extends to the buffer-internal mode on the following line cards: • 88-LC0-36FH-M • 8800-LC-48H
Shortlink Priority Flow Control	Release 7.3.3	This feature and the hw-module profile priority-flow-control command are supported on 88-LC0-36FH line card.

Feature Name	Release Information	Feature Description
Priority Flow Control Support on Cisco 8800 36x400 GbE QSFP56-DD Line Cards (88-LC0-36FH-M)	Release 7.3.15	This feature and the hw-module profile priority-flow-control command are supported on 88-LC0-36FH-M and 8800-LC-48H line cards. All previous functionalities and benefits of this feature are available on these line cards. However, the buffer-internal mode is not supported. In addition, to use the buffer-extended mode on these line cards, you are required to configure the performance capacity or headroom values. This configuration requirement ensures that you can better provision and balance workloads to achieve lossless behavior, which in turn ensures efficient use of bandwidth and resources.
Priority Flow Control	Release 7.3.1	This feature and the hw-module profile priority-flow-control command are not supported.

Priority-based Flow Control (IEEE 802.1Qbb), which is also referred to as Class-based Flow Control (CBFC) or Per Priority Pause (PPP), is a mechanism that prevents frame loss that is due to congestion. PFC is similar to 802.x Flow Control (pause frames) or link-level flow control (LFC). However, PFC functions on a per class-of-service (CoS) basis.

During congestion, PFC sends a pause frame to indicate the CoS value to pause. A PFC pause frame contains a 2-octet timer value for each CoS that indicates the length of time to pause the traffic. The unit of time for the timer is specified in pause quanta. A quanta is the time required for transmitting 512 bits at the speed of the port. The range is from 0 through 65535 quanta.

PFC asks the peer to stop sending frames of a particular CoS value by sending a pause frame to a well-known multicast address. This pause frame is a one-hop frame and isn't forwarded when received by the peer. When the congestion mitigates, the router stops sending the PFC frames to the upstream node.

You can configure PFC for each line card using the **hw-module profile priority-flow-control** command in one of two modes:

- buffer-internal
- buffer-extended

Note PFC threshold configurations are deprecated in pause command. Use the hw-module profile priority-flow-control command to configure PFC threshold configurations.

Tip You can programmatically retrieve the operational state of the PFC configuration using `Cisco-IOS-XR-ofa-npu-pfc-oper.yang` Cisco IOS XR native data model. To get started with using data models, see the *Programmability Configuration Guide for Cisco 8000 Series Routers*.

Related Topics

buffer-internal mode

Use this mode if PFC-enabled devices aren't more than 1 km apart.

You can set values for pause-threshold, headroom (both related to PFC), and ECN for the traffic class using the **hw-module profile priority-flow-control** command in this mode. The buffer-internal configuration applies to all ports that the line card hosts, which mean that you can configure a set of these values per line card.

The existing queue limit and ECN configuration in the queueing policy attached to the interface has no impact in this mode.

The effective queue limit for this mode = pause-threshold + headroom (in bytes)

Restrictions and Guidelines

The following restrictions and guidelines apply while configuring the PFC threshold values using the buffer-internal mode.

- The PFC feature isn't supported on fixed chassis systems.

- Ensure that there's no breakout configured on a chassis that has the PFC configured. Configuring PFC and breakout on the same chassis may lead to unexpected behavior, including traffic loss.

- The feature isn't supported on bundle and non-bundle sub-interface queues.

- The feature is supported on 40GbE, 100 GbE, and 400 GbE interfaces.

- The feature isn't supported in the 4xVOQ queueing mode.

- The feature isn't supported when sharing of VOQ counters is configured.

buffer-extended mode

Use this mode for PFC-enabled devices with long-haul connections.

You can set the value for pause-threshold using the **hw-module profile priority-flow-control** command in this mode. You must, however, configure the queuing policy attached to the interface to set the ECN and queuing limits. The buffer-extended configuration applies to all ports that the line card hosts, which mean that you can configure a set of these values per line card.

Configuration Guidelines

- Important points while configuring the buffer-extended mode on 88-LC0-36FH-M line cards:

  - Apart from pause-threshold, you must also configure values for headroom.

  - The headroom value range is from 4 through 75000.

  - Specify pause-threshold and headroom values in units of kilobytes (KB) or megabytes (MB).

- Important points while configuring the buffer-extended mode on 8800-LC-48H line cards:

  - Configure values only for **pause-threshold**. Don't configure headroom values.

  - Configure **pause-threshold** in units of milliseconds (ms) or microseconds.

  - Don't use units of kilobytes (KB) or megabytes (MB) units, even though the CLI displays them as options. Only use units of milliseconds (ms) or microseconds.

(Also see)

Important Considerations

- If you configure PFC values in the buffer-internal mode, then the ECN value for the line card is derived from the buffer-internal configuration. If you configure PFC values in the buffer-extended mode, then the ECN value is derived from the policy map. (For details on the ECN feature, see Explicit Congestion Notification.)

- The buffer-internal and buffer-extended modes can't coexist on the same line card.

- If you add or remove traffic-class actions on a line card, you must reload the line card.

- When using the buffer-internal mode, you can change values of the following parameters without having to reload the line card. However, if you add a new traffic class and configure these values for the first time on that traffic class, you must reload the line card for the values to come into effect.

  - pause-threshold

  - headroom

  - ECN

- If you add or remove ECN configuration using the **hw-module profile priority-flow-control** command, you must reload the line card for the ECN changes to take effect.

- The PFC threshold value ranges for the buffer-internal mode are as follows.

Threshold	Configured (bytes)
pause (min)	307200

Threshold	Configured (bytes)
pause (max)	422400
headroom (min)	345600
headroom (max)	537600
ecn (min)	153600
ecn (max)	403200

- For a traffic-class, the ECN value must always be lesser than the configured pause-threshold value.

- The combined configured values for pause-threshold and headroom must not exceed 844800 bytes. Else, the configuration is rejected.

- The pause-threshold value range for buffer-extended mode is from 2 milliseconds (ms) through 25 ms and from 2000 microseconds through 25000 microseconds.

Hardware Support for Priority Flow Control

The table lists the PIDs that support PFC per release and the PFC mode in which the support is available.

Table 2: PFC Hardware Support Matrix

Release	PID	PFC Mode
Release 7.3.15	• 88-LC0-36FH-M • 88-LC0-36FH	buffer-extended
Release 7.0.11	8800-LC-48H	buffer-internal

Configure Priority Flow Control

You can configure PFC to enable the no-drop behavior for the CoS as defined by the active network QoS policy.

Configuration Example

You must accomplish the following to complete the PFC configuration:

1. Enable PFC at the interface level.

2. Configure ingress classification policy.

3. Attach the PFC policy to the interface.

4. Configure PFC threshold values using either the buffer-internal or buffer-extended mode.

```
Router# configure
```

```
Router(config)# priority-flow-control mode on
/*Configure ingress classification policy*/
Router(config)# class-map match-any prec7
Router(config-cmap)# match precedence
Router(config)# class-map match-any tc7
/*Ingress policy attach*/
Router(config-if)# service-policy input QOS_marking
/*Egress policy attach*/
Router(config-if)# service-policy output qos_queuing
Router(config-pmap-c)# exit
Router(config-pmap)# exit
Router(config)#show controllers npu priority-flow-control location <loc>
```

Running Configuration

```
*Interface Level*
interface HundredGigE0/0/0/0
   priority-flow-control mode on

*Ingress:*
class-map match-any prec7

 match precedence 7

 end-class-map

!

class-map match-any prec6

 match precedence 6

 end-class-map

!

class-map match-any prec5

 match precedence 5

 end-class-map

!

class-map match-any prec4

 match precedence 4

 end-class-map

!

class-map match-any prec3
match precedence 3
end-class-map
!
class-map match-any prec2
match precedence 2
end-class-map
!
class-map match-any prec1
match precedence 1
end-class-map
!
```

```
!
policy-map QOS_MARKING
class prec7
  set traffic-class 7
  set qos-group 7
!
class prec6
  set traffic-class 6
  set qos-group 6
!
class prec5
  set traffic-class 5
  set qos-group 5
!
class prec4
  set traffic-class 4
  set qos-group 4
!
class prec3
  set traffic-class 3
  set qos-group 3
!
class prec2
  set traffic-class 2
  set qos-group 2
!
class prec1
  set traffic-class 1
  set qos-group 1
!
class class-default
  set traffic-class 0
  set qos-group 0
!

*Egress:*
class-map match-any tc7
 match traffic-class 7
 end-class-map
!
class-map match-any tc6
 match traffic-class 6
 end-class-map
!
class-map match-any tc5
 match traffic-class 5
 end-class-map

!

class-map match-any tc4

 match traffic-class 4

 end-class-map

!

class-map match-any tc3

 match traffic-class 3

 end-class-map
```

```
!

class-map match-any tc2
match traffic-class 2
end-class-map
!
class-map match-any tc1
match traffic-class 1
end-class-map
!
policy-map QOS_QUEUING
class tc7
  priority level 1
  shape average percent 10
!
class tc6
  bandwidth remaining ratio 1
  queue-limit 100 ms
!
class tc5
  bandwidth remaining ratio 20
  queue-limit 100 ms
!
class tc4
  bandwidth remaining ratio 20
  random-detect ecn
  random-detect 6144 bytes 100 mbytes
!
class tc3
  bandwidth remaining ratio 20
  random-detect ecn
  random-detect 6144 bytes 100 mbytes
!
class tc2
  bandwidth remaining ratio 5
  queue-limit 100 ms
!
class tc1
  bandwidth remaining ratio 5
  queue-limit 100 ms
!
class class-default
  bandwidth remaining ratio 20
  queue-limit 100 ms
!
[buffer-extended]

hw-module profile priority-flow-control location 0/0/CPU0
 buffer-extended traffic-class 3 pause-threshold 10 ms
 buffer-extended traffic-class 4 pause-threshold 10 ms
!

[buffer-internal]

hw-module profile priority-flow-control location 0/1/CPU0
 buffer-internal traffic-class 3 pause-threshold 403200 bytes headroom 441600 bytes ecn
224640 bytes
 buffer-internal traffic-class 4 pause-threshold 403200 bytes headroom 441600 bytes ecn
224640 bytes
```

Verification

```
Router#sh controllers hundredGigE0/0/0/22 priority-flow-control
Priority flow control information for interface HundredGigE0/0/0/22:
Priority Flow Control:
Total Rx PFC Frames : 0
Total Tx PFC Frames : 313866
Rx Data Frames Dropped: 0
CoS Status Rx Frames
--- ------ ----------
0   on     0
1   on     0
2   on     0
3   on     0
4   on     0
5   on     0
6   on     0
7   on     0


/*[buffer-internal]*/
Router#show controllers hundredGigE 0/9/0/24 priority-flow-control


Priority flow control information for interface HundredGigE0/9/0/24:

Priority Flow Control:
Total Rx PFC Frames : 0
Total Tx PFC Frames : 313866
Rx Data Frames Dropped: 0
CoS Status Rx Frames
--- ------ ----------
0   on     0
1   on     0
2   on     0
3   on     0
4   on     0
5   on     0
6   on     0
7   on     0
…


/*[buffer-internal, tc3 & tc4 configured. TC4 doesn't have ECN]*/

Router#show controllers npu priority-flow-control location <loc>
Location Id:                     0/1/CPU0
PFC:                             Enabled
PFC-Mode:                        buffer-internal
TC    Pause          Headroom        ECN
-------------------------------------------------------
3     86800 bytes    120000 bytes    76800 bytes
4     86800 bytes    120000 bytes    Not-configured

/*[buffer-extended PFC, tc3 & tc4 configured]*/

Router#show controllers npu priority-flow-control location <loc>
Location Id:                     0/1/CPU0
PFC:                             Enabled
PFC-Mode:                        buffer-extended
TC    Pause
-----------
3     5000 us
4     10000 us

/*[No PFC]*/
```

```
Router#show controllers npu priority-flow-control location <loc>
Location Id:                    0/1/CPU0
PFC:                            Disabled
```

Related Topics

- Priority Flow Control Overview, on page 1

Related Commands   hw-module profile priority-flow-control location

Priority Flow Control Watchdog Overview

PFC Watchdog is a mechanism to identify any PFC storms (queue-stuck condition) in the network. It also prevents the PFC from propagating on the network and running in a loop. You can configure a PFC watchdog interval to detect whether packets in a no-drop queue are drained within a specified time period. When the time period is exceeded, all outgoing packets are dropped on interfaces that match the PFC queue that is not being drained.

This requires monitoring PFC receiving on each port and detecting ports seeing an unusual number of sustained pause frames. Once detected, the watchdog module can enforce several actions on such ports, which include generating a syslog message for network management systems, shutting down the queue, and autorestoring the queue (after the PFC storm stops).

Here's how the PFC Watchdog works:

1. The Watchdog module monitors the PFC-enabled queues to determine the reception of an unusual amount of PFC pause frames in a given interval (Watchdog interval.)

2. Your hardware notifies the Watchdog module when too many PFC frames are received and traffic on the corresponding queues is halted for a time interval.

3. On receiving such notifications, the Watchdog module starts the shutdown timer and moves the queue state to wait-to-shutdown state.

4. At regular intervals during the shutdown interval, the queue is checked for PFC frames and if the traffic in the queue is stuck. If the traffic isn't stuck because the queue didn't receive any PFC frames, the queue moves back to the monitored state.

5. If the traffic is stuck for a longer time and the shutdown-timer expires, the queue switches to a drop state and the PFC Watchdog begins to drop all packets.

6. At regular intervals, the Watchdog checks the queue for PFC frames and whether the traffic in the queue is still stuck. If traffic is stuck in the queue as PFC packets keep arriving, the queue remains in the drop or shutdown state.

7. When the traffic's no longer stuck, the autorestore timer starts. At regular intervals, the module checks if the queue is stuck because of PFC frames.

8. If the queue receives PFC frames during the last autorestore interval, the auto-restore timer is reset upon expiry.

9. If the queue receives no PFC frames during the last autorestore interval, the Watchdog module restores the queue, and traffic resumes.

Related Topics

• Priority Flow Control Overview, on page 1

Configure a Priority Flow Control Watchdog Interval

You can configure PFC Watchdog parameters (Watchdog interval, shutdown multiplier, auto-restore multiplier) at the global or interface levels. Note that:

• When global Watchdog mode is disabled or off, Watchdog is disabled on all interfaces. This condition is regardless of the interface level Watchdog mode settings.

• When global Watchdog mode is enabled or on, the interface level Watchdog mode configuration settings override the global Watchdog mode values.

• When you configure interface level Watchdog attributes such as interval, shutdown multiplier, and auto-restore multiplier, they override the global Watchdog attributes.

Note Configuring the PFC mode and its policies is a prerequisite for PFC Watchdog.

Configuration Example

You can configure the Watchdog at the global or at the interface level.

Note Watchdog is enabled by default, with system default values of:

Watchdog interval = 100 ms

Shutdown multiplier = 1

Auto-restart multiplier = 10

```
P/0/RP0/CPU0:ios#show controllers hundredGigE 0/2/0/0 priority-flow-control

Priority flow control information for interface HundredGigE0/2/0/0:

Priority flow control watchdog configuration:
(D) : Default value
U : Unconfigured
-------------------------------------------------------------------------------
Configuration Item       Global   Interface  Effective
-------------------------------------------------------------------------------
PFC watchdog state :        U        U         Enabled(D)
Poll interval :             U        U         100(D)
Shutdown multiplier :       U        U         1(D)
Auto-restore multiplier :   U        U         10(D)
RP/0/RP0/CPU0:ios#config
RP/0/RP0/CPU0:ios(config)#priority-flow-control watchdog mode off
RP/0/RP0/CPU0:ios(config)#commit

RP/0/RP0/CPU0:ios(config)#do show controllers hundredGigE 0/2/0/0 priority-flo$
```

```
Priority flow control information for interface HundredGigE0/2/0/0:


Priority flow control watchdog configuration:
(D) : Default value
U : Unconfigured
--------------------------------------------------------------------------------
Configuration Item          Global   Interface  Effective
--------------------------------------------------------------------------------
PFC watchdog state :        Disabled  U       Disabled
Poll interval :               U         U       100(D)
Shutdown multiplier :         U         U       1(D)
Auto-restore multiplier :     U         U       10(D)

RP/0/RP0/CPU0:ios(config)#interface hundredGigE 0/2/0/0 priority-flow-control $
RP/0/RP0/CPU0:ios(config)#commit

RP/0/RP0/CPU0:ios(config)#do show controllers hundredGigE 0/2/0/0 priority-flo$

Priority flow control information for interface HundredGigE0/2/0/0:


Priority flow control watchdog configuration:
(D) : Default value
U : Unconfigured
--------------------------------------------------------------------------------
Configuration Item          Global   Interface   Effective
--------------------------------------------------------------------------------
PFC watchdog state :        Disabled  Enabled    Disabled
Poll interval :               U          U       100(D)
Shutdown multiplier :         U          U       1(D)
Auto-restore multiplier :     U          U       10(D)
RP/0/RP0/CPU0:ios(config)#interface hundredGigE 0/2/0/1 priority-flow-control $
RP/0/RP0/CPU0:ios(config)#commit

RP/0/RP0/CPU0:ios(config)#do show controllers hundredGigE 0/2/0/1 priority-flo$

Priority flow control information for interface HundredGigE0/2/0/1:


Priority flow control watchdog configuration:
(D) : Default value
U: Unconfigured
--------------------------------------------------------------------------------
Configuration Item          Global   Interface  Effective
--------------------------------------------------------------------------------
PFC watchdog state :        Enabled  Disabled   Disabled
Poll interval :               U         U       100(D)
Shutdown multiplier :         U         U       1(D)
Auto-restore multiplier :     U         U       10(D)
```

Related Topics