Configuring Priority Flow Control

This chapter contains the following sections:

About Priority Flow Control

Priority flow control (PFC; 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 functions on a per class-of-service (CoS) basis.

When a buffer threshold is exceeded due to congestion, PFC sends a pause frame that indicates which CoS value needs to be paused. A PFC pause frame contains a 2-octet timer value for each CoS that indicates the length of time that the traffic needs to be paused. The unit of time for the timer is specified in pause quanta. A quanta is the time that is required for transmitting 512 bits at the speed of the port. The range is from 0 to 65535. A pause frame with a pause quanta of 0 indicates a resume frame to restart the paused traffic.

By default, PFC is in the auto mode. However, no particular traffic class is enabled for pause.

Note Only certain classes of service of traffic can be flow controlled while other classes are allowed to operate normally.

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 that is not forwarded when received by the peer. When the congestion is mitigated, PFC can request the peer to restart transmitting frames.

Note RDMA over Converged Ethernet (RoCE) v1 and v2 protocols are supported on Cisco Nexus 3000 Series switches.

Guidelines and Limitations for Priority Flow Control

PFC has the following configuration guidelines and limitations:

- If PFC is enabled on a port or a port channel, it does not cause a port flap.
- Ensure that ports or port channels have enough resources before enabling PFC on them.
- PFC configuration enables PFC in both the send (Tx) and receive (Rx) direction.
- Only an exact match of the no-drop CoS is considered as a successful negotiation of PFC by the Data Center Bridging Exchange Protocol (DCBXP).
- Configuration time quanta of the pause frames is not supported.
- The configuration does not support pausing selected streams that are mapped to a particular traffic-class queue. All flows that are mapped to the class are treated as no-drop. It blocks out scheduling for the entire queue, which pauses traffic for all the streams in the queue. To achieve lossless service for a no-drop class, we highly recommend that you have only the no-drop class traffic on the queue.
- For VLAN-tagged packets, priority is always assigned based on the 802.1p field in the VLAN tag and takes precedence over the assigned internal priority(qos-group). DSCP or IP access-list classification cannot be performed on VLAN-tagged frames
- When a no-drop class is classified based on 802.1p CoS x and assigned an internal priority value (qos-group) of y, we recommend that you use the internal priority value x to classify traffic on 802.1p CoS only, and not on any other field. For x, the packet priority assigned is x if the classification is not based on CoS, which results in packets of the internal priority that is x and y to map to the same priority x.
- The PFC feature supports up to three no-drop classes of any MTU size. However, there is a limit on the number of PFC-enabled interfaces based on the following factors:
  - MTU size of the no-drop class
  - Number of 10G and 40G ports
  - Pause buffer size configuration in the input queuing policies
- Interface QoS policy takes precedence over the system policy. PFC priority derivation also occurs in the same order.
- Ensure that you apply the same interface-level QoS policy on all PFC-enabled interfaces for both ingress and egress.

Caution Irrespective of the PFC configuration, we recommend that you stop traffic before you apply or remove the queuing policy that has strict priority levels at the interface level or the system level.

- To achieve end-to-end lossless service over the network, we recommend that you enable PFC on each interface through which the no-drop class traffic flows #(Tx/Rx).
- To achieve lossless service for a no-drop class, it is recommended that you maintain only no-drop class traffic on the egress queue.
- We recommend that you change the PFC configuration when there is no traffic; otherwise, packets already in the memory management unit (MMU) of the system might not get the expected treatment.
- The buffers required for PFC are best allocated automatically. However, you can change the buffer thresholds by configuring input queuing policies.
- For no-drop classes classified based on DSCP/IP access-lists (non CoS based classifications), we highly recommend that you use the same qos-group value as the match CoS value.
- Do not enable WRED on a no-drop class because it results in egress queue drops.
- When you configure a port from the 40 Gigabit Ethernet mode to the 10 Gigabit Ethernet mode or from the 10 Gigabit Ethernet mode to the 40 Gigabit Ethernet mode, the affected ports will be administratively unavailable and PFC will be disabled on these ports. To make these ports available, use the **no shut** command. After the ports are available, PFC will become enabled on them.
- The **no lldp tlv-select dcbxp** command is enhanced so that the PFC is disabled for interfaces on both sides of back-to-back switches.
- When the 'hardware qos pfc mc-drop' configuration is enabled on Cisco Nexus 3064 switches, and if the no-drop and drop qos-groups are mapped to the same queue, then flapping the link or removing/adding PFC configurations will result in the drop-multicast feature not working correctly. This is due to the hardware limitation of having only four multicast queues in Cisco Nexus 3064 switches. To avoid this issue, do one of the following:
  - Specify both of the qos-groups that correspond to a single multicast queue as no-drop.
  - Change the mapping of the multicast queue to qos-groups using the **wrr-queue qos-group-map <queue-no> <qos-groups that are no-drop>** command.
- Beginning with Cisco NX-OS Release 7.0(3)I7(4), when PFC is received on a lossy priority group (non-configured), the event is recorded in the syslog for subsequent analysis.

  Cisco Nexus N3000 and N3100 series switches report BCM_UNEXPECTED_PFC_FRAMES syslog whenever PFC frames are received with unexpected CoS. The syslog contains the approximate count of the unexpected PFC frames received for a particular CoS in the last two seconds. The packets per second (PPS) metric can be derived by dividing this number by two.

Default Settings for Priority Flow Control

The following table lists the default setting for PFC.

Table 1: Default PFC Setting

Parameter	Default
PFC	Auto

Enabling Priority Flow Control on a Traffic Class

You can enable PFC on a particular traffic class:

SUMMARY STEPS

1. switch# **configure terminal**
2. switch(config)# **class-map type qos** *class-name*
3. switch(config-cmap-qos)# **match cos** *cos-value*
4. switch(config-cmap-qos)# **exit**
5. switch(config)# **policy-map type qos** *policy-name*
6. switch(config-pmap-qos)# **class type qos** *class-name*
7. switch(config-pmap-c-qos)# **set qos-group** *qos-group-value*
8. switch(config-pmap-c-qos)# **exit**
9. switch(config)# **interface** *type slot/port*
10. switch(config-if)# **service-policy type qos input** *policy-name*
11. switch(config-if)# **exit**
12. switch(config)# **class-map type network-qos** *class-name*
13. switch(config-cmap-nq)# **match qos-group** *qos-group-value*
14. switch(config-cmap-nq)# **exit**
15. switch(config)# **policy-map type network-qos** *policy-name*
16. switch(config-pmap-nqos)# **class type network-qos** *class-name*
17. switch(config-pmap-c-nq)# **pause no-drop**
18. switch(config-pmap-c-nq)# **exit**
19. switch(config)# **system qos**
20. switch(config-sys-qos)# **service-policy type network-qos** *policy-name*

DETAILED STEPS

	Command or Action	Purpose
Step 1	switch# **configure terminal**	Enters global configuration mode.
Step 2	switch(config)# **class-map type qos** *class-name*	Creates a named object that represents a class of traffic. Class-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 3	switch(config-cmap-qos)# **match cos** *cos-value*	Specifies the CoS value to match for classifying packets into this class. You can configure a CoS value in the range of 0 to 7.

	Command or Action	Purpose
Step 4	switch(config-cmap-qos)# **exit**	Exits class-map mode and enters global configuration mode.
Step 5	switch(config)# **policy-map type qos** *policy-name*	Creates a named object that represents a set of policies that are to be applied to a set of traffic classes. Policy-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 6	switch(config-pmap-qos)# **class type qos** *class-name*	Associates a class map with the policy map, and enters configuration mode for the specified system class. **Note** The associated class map must be the same type as the policy map type.
Step 7	switch(config-pmap-c-qos)# **set qos-group** *qos-group-value*	Configures one or more qos-group values to match on for classification of traffic into this class map. There is no default value.
Step 8	switch(config-pmap-c-qos)# **exit**	Exits policy-map mode and enters global configuration mode.
Step 9	switch(config)# **interface** *type slot/port*	Enters the configuration mode for the specified interface.
Step 10	switch(config-if)# **service-policy type qos input** *policy-name*	Applies the policy map of type qos to the specific interface.
Step 11	switch(config-if)# **exit**	Exits interface configuration mode and enters global configuration mode.
Step 12	switch(config)# **class-map type network-qos** *class-name*	Creates a named object that represents a class of traffic. Class-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 13	switch(config-cmap-nq)# **match qos-group** *qos-group-value*	Configures the traffic class by matching packets based on a list of QoS group values. Values can range from 0 to 7. QoS group 0 is equivalent to class-default.
Step 14	switch(config-cmap-nq)# **exit**	Exits class-map mode and enters global configuration mode.
Step 15	switch(config)# **policy-map type network-qos** *policy-name*	Creates a named object that represents a set of policies that are to be applied to a set of traffic classes. Policy-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 16	switch(config-pmap-nqos)# **class type network-qos** *class-name*	Associates a class map with the policy map, and enters configuration mode for the specified system class. **Note** The associated class map must be the same type as the policy map type.

	Command or Action	Purpose
Step 17	switch(config-pmap-c-nq)# **pause no-drop**	Configures a no-drop class.
Step 18	switch(config-pmap-c-nq)# **exit**	Exits policy-map mode and enters global configuration mode.
Step 19	switch(config)# **system qos**	Enters system class configuration mode.
Step 20	switch(config-sys-qos)# **service-policy type network-qos** *policy-name*	Applies the policy map of type network-qos at the system level or to the specific interface.

Example

This example shows how to enable PFC on a traffic class .

```
switch# configure terminal
switch(config)# class-map type qos c1
switch(config-cmap-qos)# match cos 3
switch(config-cmap-qos)# exit
switch(config)# policy-map type qos p1
switch(config-pmap-qos)# class type qos c1
switch(config-pmap-c-qos)# set qos-group 3
switch(config-pmap-c-qos)# exit
switch(config)# interface ethernet 1/1
switch(config-if)# service-policy type qos input p1
switch(config-if)# exit
switch(config)# class-map type network-qos c1
switch(config-cmap-nq)# match qos-group 3
switch(config-cmap-nq)# exit
switch(config)# policy-map type network-qos p1
switch(config-pmap-nqos)# class type network-qos c1
switch(config-pmap-nqos-c)# pause no-drop
switch(config-pmap-nqos-c)# exit
switch(config)# system qos
switch(config-sys-qos)# service-policy type network-qos p1
```

Configuring Priority Flow Control

You can configure PFC on a per-port basis to enable the no-drop behavior for the CoS as defined by the active network qos policy. PFC can be configured in one of these three modes:

- auto—Enables the no-drop CoS values to be advertised by the DCBXP and negotiated with the peer. A successful negotiation enables PFC on the no-drop CoS. Any failures because of a mismatch in the capability of peers causes the PFC not to be enabled.
- on—Enables PFC on the local port regardless of the capability of the peers.
- off—Disables PFC on the local port.

SUMMARY STEPS

1. **configure terminal**
2. **interface ethernet** [*slot/port-number*]

3. **priority-flow-control mode** {**auto** | **off** | **on**} **priority-flow-control mode {auto |off | on}**
4. **exit**
5. **show interface priority-flow-control**

DETAILED STEPS

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**interface ethernet** [*slot/port-number*]	Enter the interface mode on the specified interface.
Step 3	**priority-flow-control mode** {**auto** | **off** | **on**} **priority-flow-control mode {auto |off | on}**	Sets the PFC to the auto, off, or on mode. By default, PFC mode is set to auto on all ports.
Step 4	**exit**	Exits interface configuration mode.
Step 5	**show interface priority-flow-control**	Displays the status of PFC on all interfaces.

Reserving mmu-buffer for PFC

Complete the following steps to reserve mmu-buffer for PFC.

SUMMARY STEPS

1. **configure terminal**l
2. **hardware profile pfc mmu buffer-reservation ?**

DETAILED STEPS

	Command or Action	Purpose
Step 1	**configure terminal**l	Enters global configuration mode.
Step 2	**hardware profile pfc mmu buffer-reservation ?** **Example:** switch(config)# hardware profile pfc mmu buffer-reservation ?	Reserves the mmu-buffer for PFC. <0-100> Percentage of shared pool buffers to be reserved

Configuring a Priority Flow Control Watchdog Interval

You can configure a PFC watchdog interval to detect whether packets in a no-drop queue are being drained within a specified time period. When the time period is exceeded, all incoming and outgoing packets are dropped on interfaces that match the PFC queue that is not being drained. This feature is supported beginning with Cisco NX-OS Release 6.0(3)U6(9) and only for Cisco Nexus 3000 Series switches.

Starting with Cisco NX-OS Release 7.0(3)I6(1), ingress packets will be dropped for matching to shutdown PFC queue or qos-group of a physical port, and the show command displays the status.

SUMMARY STEPS

1. **configure terminal**
2. **priority-flow-control auto-restore multiplier** *value*
3. **priority-flow-control fixed-restore multiplier** *value*
4. **priority-flow-control watch-dog-interval** {**on** | **off**}
5. **priority-flow-control watch-dog interval** *value*
6. **priority-flow-control watch-dog shutdown-multiplier** *multiplier*
7. (Optional) **sh queuing pfc-queue [interface] [ethernet] [detail]**
8. (Optional) **clear queuing pfc-queue [interface] [ethernet] [intf-name]**
9. (Optional) **priority-flow-control recover interface [ethernet] [intf-name] [qos-group <0-7>]**

DETAILED STEPS

	Command or Action	Purpose

Step 1	**configure terminal**
	Example:
	`switch# configure terminal`
	`switch(config)#`
		Enters global configuration mode.

Step 2	**priority-flow-control auto-restore multiplier** *value*	Configures a value for the PFC auto-restore multiplier.
Step 3	**priority-flow-control fixed-restore multiplier** *value*	Configures a value for the PFC fixed-restore multiplier.

Step 4	**priority-flow-control watch-dog-interval** {**on**	**off**}
		Example:
		`switch(config)# priority-flow-control watch-dog-interval on`
			Globally enables or disables the PFC watchdog interval for all interfaces.
			Note You can use this same command in interface configuration mode to enable or disable the PFC watchdog interval for a specific interface.
			See the following example of the command configured at an interface with a specific shutdown multiplier value (NX-OS 7.0(3)I7(4) and later releases):
			`switch(config)# int e1/36`
			`switch(config-if)# priority-flow-control watch-dog-interval on interface-multiplier 10`
			Note Range of values for interface-multiplier is 1 - 10.

Step 5	**priority-flow-control watch-dog interval** *value* **Example:** `switch(config)# priority-flow-control watch-dog interval 200`	Specifies the watchdog interval value. The range is from 100 to 1000 milliseconds.
Step 6	**priority-flow-control watch-dog shutdown-multiplier** *multiplier* **Example:**	Specifies when to declare the PFC queue as stuck. The range is from 1 to 10, and the default value is 1.

	Command or Action	Purpose
	`switch(config)# priority-flow-control watch-dog shutdown-multiplier 5`	**Note** When the PFC queue is declared as stuck, a syslog entry is created to record the conditions of the PFC queue. (NX-OS 7.0(3)I7(4) and later releases)

Step 7	(Optional) **sh queuing pfc-queue [interface] [ethernet] [detail]**
	Example:
	`switch(config)# sh queuing pfc-queue interface ethernet 1/1 detail`
		Displays the PFCWD statistics. Starting with Cisco NX-OS Release 7.0(3)I6(1), you can use the detail option to account for Ingress drops.
		`| QOS GROUP 2 [Active] PFC [YES] PFC-COS [1]`
		`+----------+`
		`| | Stats |`
		`+----------+`
		`| Shutdown| 0|`
		`| Restored| 0|`
		`| Total bytes drained| 0|`
		`| Total pkts dropped| 0|`
		`| Aggregate pkts dropped| 0|`
		`| Ingress Pkts dropped| 0|`
		`| Aggregate Ingress Pkts dropped| 0|`
		`Global watch-dog interval [Enabled]`
		`+----------+`
		`+----------+`
		`Global PFC watchdog configuration details`
		`PFC watch-dog poll interval : 200 ms`
		`PFC watch-dog shutdown multiplier : 5`
		`PFC watch-dog auto-restore multiplier : 10`
		`PFC watch-dog fixed-restore multiplier : 0`
		`PFC watchdog internal-interface multiplier : 0`

Step 8	(Optional) **clear queuing pfc-queue [interface] [ethernet] [intf-name]** **Example:** `switch(config)# clear queuing pfc-queue interface ethernet 1/1`	Clears the environment variable PFCWD statistics.
Step 9	(Optional) **priority-flow-control recover interface [ethernet] [intf-name] [qos-group <0-7>]**	Recovers the interface manually.

	Command or Action	Purpose
	Example: `switch# priority-flow-control recover interface ethernet 1/1 qos-group 3`	

Verifying the Priority Flow Control Configuration

To display the PFC configuration, perform the following task:

SUMMARY STEPS

1. **configure terminal**
2. **show interface priority-flow-control**
3. (Optional) **show interface priority-flow-control detail**

DETAILED STEPS

	Command or Action	Purpose
Step 1	**configure terminal**	Enter global configuration mode.
Step 2	**show interface priority-flow-control**	Displays the status of PFC on all interfaces.
Step 3	(Optional) **show interface priority-flow-control detail**	Displays the status of PFC for each priority level for each interface.

Monitoring PFC Frame Counter Statistics

You can monitor the Tx and Rx counters for PFC-enabled devices either at an interface level or at a per priority (CoS) level for each interface.

SUMMARY STEPS

1. switch# **show int priority-flow-control** [**detail**]

DETAILED STEPS

	Command or Action	Purpose
Step 1	switch# **show int priority-flow-control** [**detail**]	

Example

This example shows how to display PFC frame counter statistics for each priority level for each interface:

```
switch# show int priority-flow-control detail
```

```
Ethernet1/1/1:
    Admin Mode: On
    Oper Mode: On
    VL bitmap: (14)
    Total Rx PFC Frames: 0
    Total Tx PFC Frames: 0
    ------------------------------------------------------------------------------
----------------------------------------
        |  Priority0  |  Priority1  |  Priority2  |  Priority3  |  Priority4  |
 Priority5  |  Priority6  |  Priority7  |
    ------------------------------------------------------------------------------
----------------------------------------
    Rx  |0            |0            |0            |0            |0            |0
            |0            |0
    ------------------------------------------------------------------------------
----------------------------------------
    Tx  |0            |0            |0            |0            |0            |0
            |0            |0
Ethernet1/1/2:
    Admin Mode: Auto
    Oper Mode: Off
    VL bitmap:
    Total Rx PFC Frames: 0
    Total Tx PFC Frames: 0
    ------------------------------------------------------------------------------
----------------------------------------
        |  Priority0  |  Priority1  |  Priority2  |  Priority3  |  Priority4  |
 Priority5  |  Priority6  |  Priority7  |
    ------------------------------------------------------------------------------
----------------------------------------
    Rx  |0            |0            |0            |0            |0            |0
            |0            |0
    ------------------------------------------------------------------------------
----------------------------------------
    Tx  |0            |0            |0            |0            |0            |0
            |0            |0
```

This example shows how to display PFC frame counter statistics for each interface:

```
switch# show int priority-flow-control
============================================================
Port               Mode Oper(VL bmap)  RxPPP       TxPPP
============================================================

Ethernet1/1/1      On   On  (14)      0           0
Ethernet1/1/2      Auto Off           0           0
Ethernet1/1/3      Auto On  (14)      0           0
Ethernet1/15       Auto On  (14)      0           0
Ethernet1/15       Auto On  (14)      0           0
Ethernet1/15       Auto On  (14)      0           0
Ethernet1/15       Auto On  (14)      0           0
Ethernet1/25       Auto On  (14)      0           0
Ethernet1/32       On   On  (14)      0           0
switch#
```

Configuration Examples for Priority Flow Control

The following example shows how to configure PFC.

```
switch# configure terminal
switch(config)# interface ethernet 5/5
switch(config-if)# priority-flow-control mode on
```