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 is similar to 802.3x Flow Control (pause frames). However, PFC functions on a per class-of-service (CoS) basis.

During 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.

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 3400-S switches.

Licensing Requirements for Priority Flow Control

The following table shows the licensing requirements for this feature:

Product	License Requirement
Cisco NX-OS	The PFC feature does not a require license. Any feature not included in a license package is bundled with the NX-OS image and is provided at no extra charge to you. For a complete explanation of the Cisco NX-OS licensing scheme, see the Cisco NX-OS Licensing Guide.

Prerequisites for Priority Flow Control

PFC has the following prerequisites:

• You must be familiar with using modular QoS CLI.

• You are logged on to the device.

Guidelines and Limitations for Priority Flow Control

Priority Flow Control (PFC) has the following guidelines and limitations:

• PFC Pause Buffer Threshold Guidelines & Limitations:

• Ingress queuing policy allows you to optimize buffer utilization between the ports, giving certain ports a larger buffer, if needed, so long as the total buffer stays within the limit (60K * 224B). If the total buffer size exceeds the limit, the ingress queuing policy is rejected.

• If the ingress queuing policy is accepted, the switch must be reload for it to take effect.

• Total buffer size per block has to be managed not to exceed the limit of ~60K cells (each cell is 224B). The pause buffer limit and a large number of ports per block can cause the following limitations if the ports are configured to the default headroom

• Long cable lengths, e.g., 300m, may cause lossless queue drops.

• Only a single lossless class is supported (two lossless classes cannot be supported).

• Only one no-drop class is supported.

• Adding the "pause buffer size threshold" configuration is optional for cable lengths that are less than 100 meters and it does need not to be configured.

- PFC is not supported on sub-interfaces.

- For cable lengths greater than 100m, the "pause buffer size threshold" configuration is mandatory and it is required as part of the ingress queuing policy configuration, however, ingress queuing is not supported in this release.

- PFC configuration enables PFC in both the send (Tx) and receive (Rx) direction.

- Configuration time quanta of the pause frames is not supported.

- 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 outgoing packets are dropped on interfaces that match the PFC queue that is not being drained.

- 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 recommend that you have only the no-drop class traffic on the queue.

- You can define the upper limit of any MTU in the system using the **systemjumbomtu** command. The MTU range is from 1500 to 9216 bytes, and the default is 9216 bytes.

- For no-drop class, the supported MTU size is up to 2240 bytes.

- The interface QoS policy takes precedence over the system policy. PFC priority derivation also happens 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 applying or removing a 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).

- 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 may not get the expected treatment.

- We recommend WRED on a no-drop class with ECN enabled.

- ECN functions for TCP/UDP traffic. It does not function for non-IP traffic.

- Classification is supported based on DSCP/IP precedence or priority in the CoS.

- IP/MAC access-list based classification is not supported for no-drop class traffic.

- Only an exact match of the no-drop CoS is considered as a successful negotiation of PFC by the DCBXP.

- The **no lldp tlv-select dcbxp** command is enhanced so that PFC is disabled for interfaces on both sides of back-to-back switches.

Default Settings for Priority Flow Control

Table 1: Default PFC Setting

Parameter	Default
PFC	Auto

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 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.

> **Note** You can use the **priority-flow-control override-interface mode off** command to globally disable PFC on all interfaces regardless of the current interface configuration. This command, which is meant to be used during troubleshooting, allows you to quickly disable PFC without having to disable PFC on each interface.

SUMMARY STEPS

1. **configure terminal**
2. **interface** *type slot/port*
3. **priority-flow-control mode** [**auto** | **off** |**on**]
4. **show interface priority-flow-control**

DETAILED STEPS

	Command or Action	Purpose
| **Step 1** | **configure terminal**<br><br>**Example:**<br>```
switch# configure terminal
switch(config)#
``` | Enters global configuration mode. |
| **Step 2** | **interface** *type slot/port*<br><br>**Example:**<br>```
switch(config)# interface ethernet 2/5
switch(config-if)#
``` | Enters interface mode on the interface specified. |

	Command or Action	Purpose

Step 3	**priority-flow-control mode [auto	off |on]**
		
		Example:
		`switch(config-if)# priority-flow-control mode on`
		`switch(config-if)#`
			Sets PFC to the on mode.

Step 4	**show interface priority-flow-control** **Example:** `switch# show interface priority-flow-control`	(Optional) Displays the status of PFC on all interfaces.

Enabling Priority Flow Control on a Traffic Class

You can enable PFC on a particular traffic class.

SUMMARY STEPS

1. **configure terminal**
2. **class-map type qos** *class-name*
3. **match cos** *cos-value*
4. **exit**
5. **policy-map type qos** *policy-name*
6. **class type qos** *class-name*
7. **set qos-group** *qos-group-value*
8. **exit**
9. **exit**
10. **class-map type network-qos match-any** *class-name*
11. **match qos-group** *qos-group-value*
12. **exit**
13. **class-map type network-qos** *class-name*
14. **match qos-group** *qos-group-value*
15. **exit**
16. **policy-map type network-qos** *policy-name*
17. **class type network-qos** *class-name*
18. **pause pfc-cos** *value*
19. **exit**
20. **exit**
21. **system qos**
22. **service-policy type network-qos** *policy-name*
23. **exit**
24. **interface ethernet** *slot/number*
25. **priority-flow-control mode on** *slot/number*
26. **exit**

DETAILED STEPS

	Command or Action	Purpose

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

Step 2	**class-map type qos** *class-name*
	
	Example:
	`switch(config)# class-map type qos c1`
	`switch(config-cmap-qos)#`
		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	**match cos** *cos-value* **Example:** `switch(config-cmap-qos)# match cos 2`	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.

Step 4	**exit**
	
	Example:
	`switch(config-cmap-qos)# exit`
	`switch(config)#`
		Exits class-map mode and enters global configuration mode.

Step 5	**policy-map type qos** *policy-name*
	
	Example:
	`switch(config)# policy-map type qos p1`
	`switch(config-pmap-qos)#`
		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	**class type qos** *class-name*
	
	Example:
	`switch(config-pmap-qos)# class type qos c1`
	`switch(config-pmap-c-qos)#`
		Associates a class map with the policy map and enters the configuration mode for the specified system class.
		
		Note The associated class map must be the same type as the policy map type.

Step 7	**set qos-group** *qos-group-value* **Example:** `switch(config-pmap-c-qos)# set qos-group 2`	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	**exit**
	
	Example:
	`switch(config-pmap-c-qos)# exit`
	`switch(config-pmap-qos)#`
		Exits the system class configuration mode and enters policy-map mode.

Step 9	**exit**
	
	Example:
	`switch(config-pmap-qos)# exit`
	`switch(config)#`
		Exits policy-map mode and enters global configuration mode.

	Command or Action	**Purpose**

Step 10	**class-map type network-qos match-any** *class-name*
	
	 Example:
	 `switch(config)# class-map type network-qos`
	 `match-any c1`
	 `switch(config-cmap-nqos)#`
		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 11	**match qos-group** *qos-group-value* **Example:** `switch(config-cmap-nqos)# match qos-group 3`	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. **Note** The *qos-group-value* should match the **pause pfc-cos** *value*. See the **pause pfc-cos** command below in this procedure.

Step 12	**exit**
	
	 Example:
	 `switch(config-cmap-nqos)# exit`
	 `switch(config)#`
		Exits class-map mode and enters global configuration mode.

Step 13	**class-map type network-qos** *class-name*
	
	 Example:
	 `switch(config)# class-map type network-qos nw-qos3`
	 `switch(config-cmap-nqos)#`
		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 14	**match qos-group** *qos-group-value* **Example:** `switch(config-cmap-nqos)# match qos-group 3`	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 15	**exit**
	
	 Example:
	 `switch(config-cmap-nqos)# exit`
	 `switch(config)#`
		Exits class-map mode and enters global configuration mode.

Step 16	**policy-map type network-qos** *policy-name*
	
	 Example:
	 `switch(config)# policy-map type network-qos`
	 `pfc-qos`
	 `switch(config-pmap-nqos)#`
		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 17	**class type network-qos** *class-name*
	
	 Example:
	 `switch(config-pmap-nqos)# class type network-qos`
	 ` nw-qos3`
	 `switch(config-pmap-nqos-c)#`
		Associates a class map with the policy map, and enters the configuration mode for the specified system class.
		
		 Note The associated class map must be the same type as the policy map type.

Step 18	**pause pfc-cos** *value* **Example:**	PFC sends a pause frame that indicates which CoS value needs to be paused.

		Command or Action	Purpose
		`switch(config-pmap-nqos-c)# pause pfc-cos 3` `switch(config-pmap-nqos)#`	**Note** The **pause pfc-cos** *value* should match the *qos-group-value* in the **match qos-group** command. See Step Step 11, on page 7.

Step 19	**exit**
	
	Example:
	`switch(config-pmap-nqos-c)# exit`
	`switch(config-pmap-nqos)#`
			Exits configuration mode and enters policy-map mode.

Step 20	**exit**
	
	Example:
	`switch(config-pmap-nqos)# exit`
	`switch(config)#`
			Exits policy-map mode and enters global configuration mode.

Step 21	**system qos**
	
	Example:
	`switch(config)# system qos`
	`switch(config-sys-qos)#`
			Enters system class configuration mode.

Step 22	**service-policy type network-qos** *policy-name*
	
	Example:
	`switch(config-sys-qos)# service-policy type`
	`network-qos pfc-qos`
			Applies the policy map of type network-qos at the system level or to the specific interface.

Step 23	**exit**
	
	Example:
	`switch(config-sys-qos)# exit`
	`switch(config)#`
			Exits policy-map mode and enters global configuration mode.

Step 24	**interface ethernet** *slot/number*
	
	Example:
	`switch(config)# interface ethernet 1/1`
	`switch(config-if)#`
			Enters the ethernet interface configuration mode for the selected slot and chassis number.

Step 25	**priority-flow-control mode on** *slot/number*
	
	Example:
	`switch(config-if)# priority-flow-control mode on`
	`switch(config-if)#`
			Enables the priority flow control policy for the interface.

Step 26	**exit**
	
	Example:
	`switch(config-if)# exit`
	`switch(config)#`
			Exits the ethernet interface mode and enters the global configuration mode.

Configuring Pause Buffer Thresholds

To react to port congestion issues, a buffer space (pause buffer) is available to store incoming packets until the sender reacts to a system-generated pause request and stops sending packets. By default, the buffer size is automatically set by the system and is based on ingress link speed and the link cable length.

The pause buffer feature allows you to override the default buffer size by configuring buffer values in an ingress queuing policy.

Note   Global ingress queuing policies are not supported.

SUMMARY STEPS

1. **configure terminal**
2. **policy-map type queuing** *name*
3. **class type queuing { queue type }**
4. **pause buffer-size** *bytes* **pause-threshold** *bytes* **resume-threshold** *bytes*
5. **exit**
6. **exit**
7. **interface ethernet** *slot* / *number*
8. **service-policy type queuing input** *name*

DETAILED STEPS

	Command or Action	Purpose

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

Step 2	**policy-map type queuing** *name*
	
	Example:
	`switch (config)# policy-map type queuing policy-1`
	`switch(config-pmap-que)#`
		Enters the policy map queue configuration mode.

Step 3	**class type queuing { queue type }**
	
	Example:
	`switch (config-pmap-que)# class type queuing`
	`c-out-8q-q3`
	`switch(config-pmap-c-que)#`
		Enters the policy map class queue configuration mode.

Step 4	**pause buffer-size** *bytes* **pause-threshold** *bytes* **resume-threshold** *bytes*	Configures the pause buffer for the policy.

	Command or Action	**Purpose**
	Example: switch (config-pmap-que)# pause buffer-size 145000 pause-threshold 100000 resume-threshold 12480 switch(config-pmap-c-que)#	**Note** Cisco Nexus 3400-S switches do not support user-defined pause-threshold and resume-threshold as individual settings. Modification of the pause buffer size is supported using the following formula: buffer size = buffer-size - pause-threshold. For example, to set a buffer size of 32000 bytes, you can configure the pause buffer-size as 62000, the pause-threshold size as 30000, and the resume-threshold size as15000. In this configuration, the pause-threshold and resume-threshold are required, but will not be honored by the switch as individual settings.
Step 5	**exit** **Example:** switch (config-pmap-c-que)# exit switch(config)#	Enters the policy map queue configuration mode.
Step 6	**exit** **Example:** switch (config-pmap-que)# exit switch(config)#	Enters the global configuration mode.
Step 7	**interface ethernet** *slot* / *number* **Example:** switch (config-if)# interface ethernet 1/7 switch(config-if)#	Enters the interface configuration mode.
Step 8	**service-policy type queuing input** *name* **Example:** switch (config-if)# service-policy type queuing input policy1 switch(config-if)#	Applies the queuing policy to the interface port.

Verifying the Priority Flow Control Pause Buffer Threshold Configuration

To display the PFC pause buffer threshold configurations on all interfaces, perform the following task:

Command	**Purpose**
show queuing interface	Displays configuration and statistics for all ingress queuing interfaces.

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 outgoing packets are dropped on interfaces that match the PFC queue that is not being drained.

Note   Ingress drops are supported only on the front-panel ports.

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) **priority-flow-control recover interface [ethernet|ii] [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. This command should be configured at global and also at an interface.
			
			See the following example of the command configured at global:
			
			`switch(config)# priority-flow-control`
			`watch-dog-interval on`
			
			See the following example of the command configured at an interface:
			
			`switch(config)# interface ethernet 7/5`
			`switch(config-if)# priority-flow-control`
			`watch-dog-interval on`

	Command or Action	Purpose
		Note You can use this same command in interface configuration mode to enable or disable the PFC watchdog interval for a specific interface.
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:** `switch(config)# priority-flow-control watch-dog shutdown-multiplier 5`	Specifies when to declare the PFC queue as stuck. The range is from 1 to 10, and the default value is 1. **Note** When the PFC queue is declared as stuck, a syslog entry is created to record the conditions of the PFC queue.
Step 7	(Optional) **priority-flow-control recover interface [ethernet|ii] [intf-name] [qos-group <0-7>]** **Example:** `switch# priority-flow-control recover interface ethernet 1/1 qos-group 3`	Recovers the interface manually.

Verifying the Priority Flow Control Configuration

To display the PFC configuration, perform the following task:

Command	Purpose
show interface priority-flow-control [**module** *number*]	Displays the status of PFC on all interfaces or on specific modules.

Configuration Examples for Priority Flow Control

The following example shows how to configure PFC:

```
configure terminal
interface ethernet 5/5
priority-flow-control mode on
```

The following example shows how to enable PFC on a traffic class:

```
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-pmap-qos)# exit
```

```
switch(config)# class-map type network-qos match-any c1
switch(config-cmap-nqos)# match qos-group 3
switch(config-cmap-nqos)# exit
switch(config)# policy-map type network-qos p1
switch(config-pmap-nqos)# class type network-qos c-nq1
switch(config-pmap-nqos-c)# pause pfc-cos 3
switch(config-pmap-nqos-c)# exit
switch(config-pmap-nqos)# exit
switch(config)# system qos
switch(config-sys-qos)# service-policy type network-qos p1
```