Configuring Priority Flow Control

- About Priority Flow Control, on page 1
- Prerequisites for Priority Flow Control, on page 2
- Guidelines and Limitations for Priority Flow Control, on page 2
- Default Settings for Priority Flow Control, on page 4
- Configuring Priority Flow Control, on page 4
- Enabling Priority Flow Control on a Traffic Class, on page 5
- Configuring Pause Buffer Thresholds and Queue Limit Using Ingress Queuing Policy, on page 8
- Verifying the Priority Flow Control Configuration, on page 10
- Configuration Examples for Priority Flow Control, on page 10

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) or link-level flow control (LFC). However, PFC functions on a per class-of-service (CoS) basis.

When a buffer threshold is exceeded due to congestion, LFC sends a pause frame to its peer to pause all data transmission on the link for a specified period of time. When the congestion is mitigated (traffic comes under the configured threshold), a resume frame is generated to restart data transmission on the link.

In contrast, 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 Cisco Nexus 9000 Series switches support the transport of RDMA over Converged Ethernet (RoCE) v1 and v2 protocols.

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

PFC has the following configuration guidelines and limitations:

- PFC is not supported on the Cisco Nexus 9508 switch (NX-OS 7.0(3)F3(3).
- The **show** commands with the **internal** keyword are not supported.
- Adding pause buffer size threshold configuration is optional for cable lengths that are less than 100 meters and it need not be configured.
- Input queuing policy maps cannot have pause buffer and priority/bandwidth together.
- For cable lengths that are greater than 100m, the pause buffer size threshold configuration is mandatory and it is required as part of the QoS policy configuration.
- If PFC is enabled on a port or a port channel, it does not cause a port flap.
- 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. Beginning with Cisco NX-OS Release 7.0(3)I4(2), this feature is supported only for Cisco Nexus 9200 Series switches, Cisco Nexus 93108TC-EX, and 93180YC-EX switches, and Cisco Nexus 9508 switches with the X9732C-EX line cards.
- 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, Cisco recommends that you have only the no-drop class traffic on the queue.
- When a no-drop class is classified based on 802.1p CoS x and assigned a internal priority value (qos-group) of y, Cisco recommends that you use the internal priority value x to classify traffic on 802.1p CoS only, and not on any other field. The packet priority assigned is x if the classification is not based on CoS, which results in packets of internal priority x and y to map to the same priority x.

- The PFC feature supports up to three no-drop classes of any maximum transmission unit (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
- 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.
- 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, Cisco recommends 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, Cisco recommends that you enable PFC on each interface through which the no-drop class traffic flows (Tx/Rx).
- Cisco recommends 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.
- Cisco recommends that you use default buffer sizes for no-drop classes or configure different input queuing policies suitable to 10G and 40G interfaces and the no-drop class MTU size. If the buffer size is specified through the CLI, it allocates the same buffer size for all ports irrespective of the link speed and MTU size. Applying the same pause buffer-size on 10G and 40G interfaces is not supported.
- Do not enable WRED on a no-drop class because it results in egress queue drops.
- Dynamic load balancing cannot be enabled for internal links with PFC. You must disable DLB and enable RTAG7 load-balancing for internal links with the port-channel load-balance internal rtag7 command.
- The dynamic load balancing (DLB) based hashing scheme is enabled by default on all internal links of a linecard. When DLB is enabled, no-drop traffic might experience out-of-order packet delivery when congestion on internal links occurs and PFC is applied. If applications on the system are sensitive to out-of-order delivery, you can adjust for this by disabling DLB at the qos-group level. Disable DLB by using the **set dlb-disable** action in the QoS policy-maps and the **set qos-group** action for no-drop classes.

  In the following example assume that qos-group 1 is a no-drop class. DLB is disabled for this no-drop class by adding the **set dlb-disable** action and the **set qos-group** action.

```
switch(config)# policy-map p1
switch(config-pmap-qos)# class c1
switch(config-pmap-c-qos)# set qos-group 1
switch(config-pmap-c-qos)# set dlb-disable
switch(config-pmap-c-qos)# end
switch# show policy-map p1


  Type qos policy-maps
  ====================
```

```
policy-map type qos p1
  class  c1
    set qos-group 1
    set dlb-disable
```

Note The following Cisco Nexus platform switches do not support the **set-dlb-disable** command:

- Cisco Nexus 9200-series platform switches
- Cisco Nexus 9300-EX/FX/FX2 platform switches
- Cisco Nexus 9500-series platform switches with -EX and -FX line cards

- For VLAN-tagged packets, priority is 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.
- For non VLAN-tagged frames, priority is assigned based on the **set qos-group** action given by the ingress QoS policy. Classification is based on a QoS policy-allowed match condition such as precedence, DSCP, or access-list. You must ensure that the **pfc-cos** value provided in the network-qos policy for this class is the same as the **qos-group** value in this case.

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:

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

DETAILED STEPS

	Command or Action	Purpose

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

Step 2	**interface** *type slot/port*
	Example:
	`switch(config)# interface ethernet 2/5`
	`switch(config-if)#`
		Enters interface mode on the interface specified.

Step 3	**priority-flow-control mode** [| **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 match** { **all** | **any** } *class-name*
3. **match cos** *cos-value*
4. **match dscp** *dscp-value*
5. **exit**
6. **policy-map type qos** *policy-name*
7. **class** *class-name*
8. **set qos-group** *qos-group-value*
9. **exit**
10. **exit**
11. **policy-map type network-qos** *policy-name*
12. **class type network-qos** *class-name*
13. **pause pfc-cos** *value* [**receive**]
14. **exit**
15. **exit**
16. **system qos**
17. **service-policy type network-qos** *policy-name*

18. **exit**
19. **interface ethernet** *slot* / *number*
20. **priority-flow-control mode** { **auto** | **on** | **off** }
21. **service-policy type qos input** *policy-name*
22. **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 match** { **all** | **any** } *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.
		match { **all** | **any** }: Default is **match all** (if multiple matching statements are present all of them must be matched).

Step 3	**match cos** *cos-value*
	Example:
	`switch(config-cmap-qos)# match cos 2`
	`switch(config-cmap-qos)#`
		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	**match dscp** *dscp-value*
	Example:
	`switch(config-cmap-qos)# match dscp 3`
	`switch(config-cmap-qos)#`
		Specifies the DSCP value to match for classifying packets into this class. You can configure a DSCP value in the range of 0 to 63 or the listed values.

Step 5	**exit**
	Example:
	`switch(config-cmap-qos)# exit`
	`switch(config)#`
		Exits class-map mode and enters global configuration mode.

Step 6	**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 7	**class** *class-name*
	Example:
	`switch(config-pmap-qos)# class 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.

	Command or Action	Purpose

Step 8	**set qos-group** *qos-group-value*
	Example:
	`switch(config-pmap-c-qos)# set qos-group 3`
	`switch(config-pmap-c-qos)#`
		Configures one or more qos-group values to match on for classification of traffic into this class map. There is no default value.

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

Step 10	**exit**
	Example:
	`switch(config-pmap-qos)# exit`
	`switch(config)#`
		Exits policy-map mode and enters global configuration mode.

Step 11	**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 12	**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 13	**pause pfc-cos** *value* [**receive**]
	Example:
	`switch(config-pmap-nqos-c)# pause pfc-cos 3 receive`
	`switch(config-pmap-nqos-c)#`
		PFC sends a pause frame that indicates which CoS value needs to be paused. Only PFC receive is enabled for the list of PCF CoS values.
		receive: When this optional keyword is used, PFC only receives and honors pause frames. PFC will never send pause frames. This is known as "Asymmetric PFC".
		Note Although not required, the **pause pfc-cos** *value* should match the *qos-group-value* in the **set qos-group** command. See the **set qos-group** command in steps 8 above.

Step 14	**exit**
	Example:
	`switch(config-pmap-nqos-c)# exit`
	`switch(config-pmap-nqos)#`
		Exits configuration mode and enters policy-map mode.

Step 15	**exit**
	Example:
	`switch(config-pmap-nqos)# exit`
	`switch(config)#`
		Exits policy-map mode and enters global configuration mode.

	Command or Action	Purpose

Step 16	**system qos**
	Example:
	`switch(config)# system qos`
	`switch(config-sys-qos)#`
		Enters system class configuration mode.

Step 17	**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 18	**exit**
	Example:
	`switch(config-sys-qos)# exit`
	`switch(config)#`
		Exits policy-map mode and enters global configuration mode.

Step 19	**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 20	**priority-flow-control mode** { **auto** | **on** | **off** }
	Example:
	`switch(config-if)# priority-flow-control mode on`
	`switch(config-if)#`
		Enables the priority flow control policy for the interface.

Step 21	**service-policy type qos input** *policy-name* **Example:** `switch(config-if)# service-policy type qos input p1`	Adds classification to the interface ensuring that packets matching the previously configured CoS or DSCP values are classified in the correct QoS group.

Step 22	**exit**
	Example:
	`switch(config-if)# exit`
	`switch(config)#`
		Exits the ethernet interface mode and enters the global configuration mode.

Configuring Pause Buffer Thresholds and Queue Limit Using Ingress Queuing Policy

The pause buffer thresholds specified in the network-qos policy are shared by all the ports in the system. However, there are situations where a few ports may need different thresholds (such as long distance connections). An ingress queuing policy can be used for this purpose.

An ingress queuing policy also allows the configuration of the queue-limit to restrict the amount of shared buffer that can be used in addition to the reserved pause buffer by the no-drop class.

Each no-drop class is mapped internally to one of the port's priority-group in the ingress direction. The configured pause buffer thresholds and queue-limit are applied to the priority-group associated with the class.

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

For cable lengths that are greater than 100m, the pause buffer size threshold configuration is mandatory and it is required as part of the QoS policy configuration.

Note About queue limits for 100G enabled devices (such as the Cisco Nexus 9300 platform switch with the N9K-M4PC-CFP2 GEM):

- The maximum dynamic queue-limit alpha value supported by the device might be greater that 8. However 8 is the maximum alpha value supported. Configuring the alpha value to a value greater than 8 is overridden by the maximum alpha value of 8.

  No message is issued when the alpha value is overridden.

- The static queue-limit has a maximum of 20,000 cells. Any value specified greater than the maximum 20,000 cell limit is overridden by the 20,000 cell limit.

  No message is issued when the cell limit is overridden.

SUMMARY STEPS

1. **configure terminal**
2. **policy-map type queuing** *policy-map-name*
3. **class type queuing** *c-in-q1*
4. **pause buffer-size** *buffer-size* **pause threshold** *xoff-size* **resume threshold** *xon-size*
5. **no pause buffer-size** *buffer-size* **pause threshold** *xoff-size* **resume threshold** *xon-size*
6. **queue-limit** *queue size* [**dynamic** *dynamic threshold*]

DETAILED STEPS

	Command or Action	Purpose
Step 1	**configure terminal**	Enters global configuration mode.
Step 2	**policy-map type queuing** *policy-map-name*	Enters policy-map queuing class mode and identifies the policy map assigned to the type queuing policy map.
Step 3	**class type queuing** *c-in-q1*	Attaches the class map of type queuing and then enters policy-map class queuing mode. Class queuing names are listed in the System-Defined Type queuing Class Maps table. **Note** The qos-group associated with the class must be defined as a no-drop class in the network-qos policy applied in the system qos.

	Command or Action	Purpose
Step 4	**pause buffer-size** *buffer-size* **pause threshold** *xoff-size* **resume threshold** *xon-size*	Specifies the buffer threshold settings for pause and resume.
Step 5	**no pause buffer-size** *buffer-size* **pause threshold** *xoff-size* **resume threshold** *xon-size*	Removes the buffer threshold settings for pause and resume.
Step 6	**queue-limit** *queue size* [**dynamic** *dynamic threshold*]	(Optional) Specifies either the static or dynamic shared limit available to the ingress priority-group. The static queue limit defines the fixed size to which the priority-group can grow. The dynamic queue limit allows the priority-group's threshold size to be decided depending on the number of free cells available, in terms of the alpha value. **Note** Cisco Nexus 9200 platform switches only support a class level dynamic threshold configuration with respect to the alpha value. This means that all ports in a class share the same alpha value.

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
```