Requirement Discussion of Flow-Based Flow Control(FFC)

Nongda Hu

Yolanda Yu

hunongda@huawei.com

yolanda.yu@huawei.com

IEEE 802.1 DCB, Stuttgart, May 2017

Agenda

- Key enablers for the lossless Ethernet network
- Some Issues with PFC
- Proposal to create a per-flow flow control
- Summary

The Enabler of Lossless Ethernet

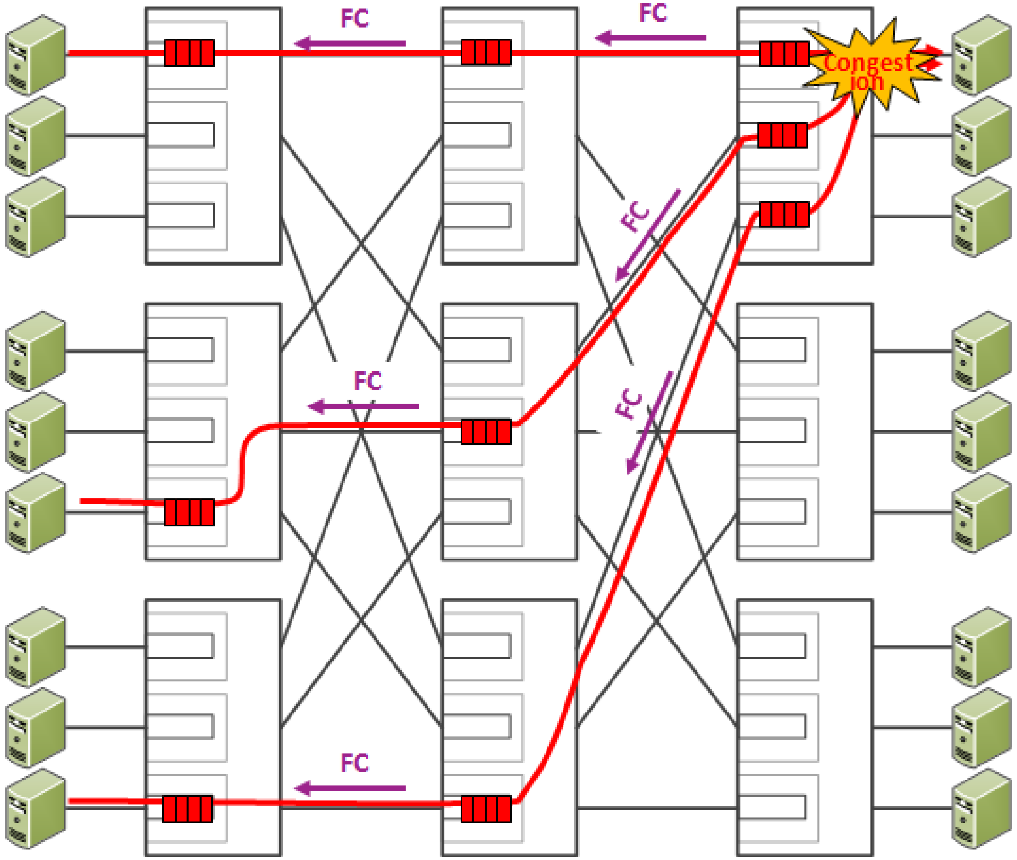
- NVMe over Fabric (NOF)
  - Represent the requirement of high performance storage and resource pooling.
  - NOF need extremely low latency. Can’t accept the latency caused by retransmission.
  - Need lossless Ethernet.

- RDMA technology is more and more used in modern data center
  - The underlying networks for RoCE and RoCEv2 should be configured as lossless.
  - The requirement for an underlying lossless network is aimed at preventing RoCE, RoCEv2 packet drops as a result of contention in the fabric.
  - Currently, IEEE 802.1 Qbb PFC (per-priority link-layer flow-control) is used.

- The convergence of HPC, Storage, LAN
  - Need lossless and low latency Ethernet.

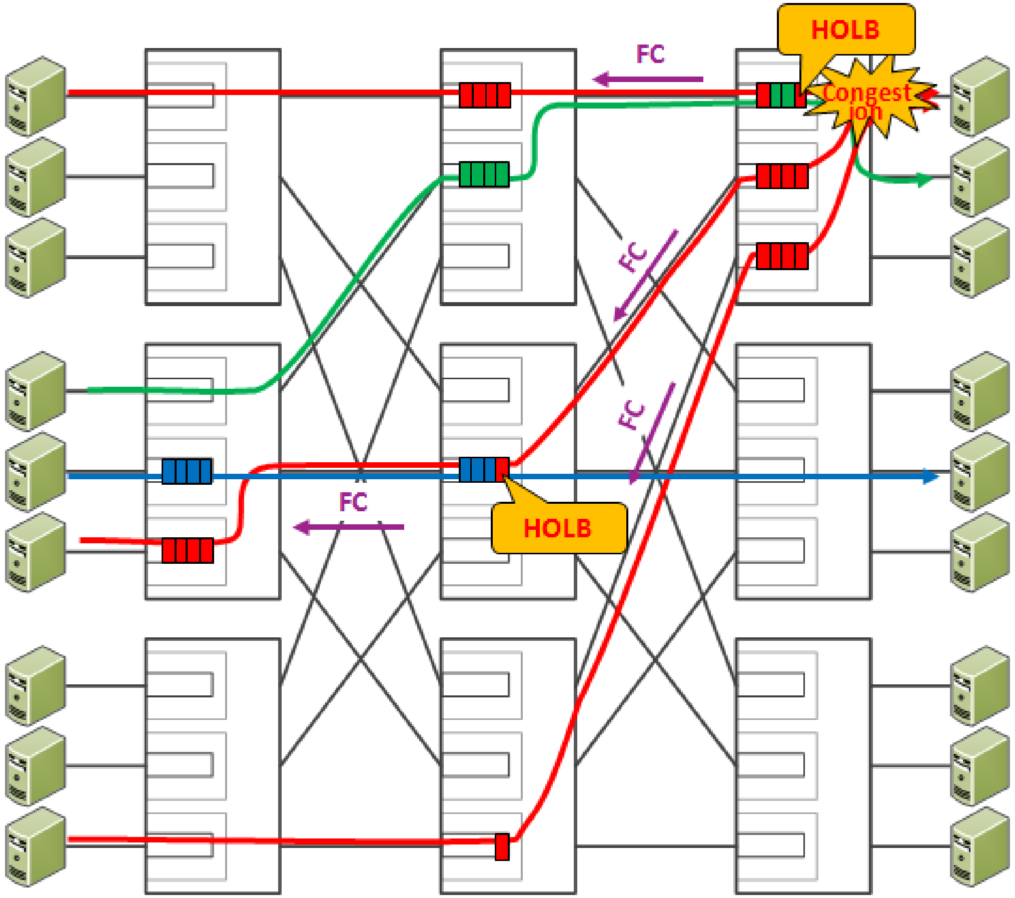
The issues of Lossless Ethernet

- **Congestion Propagation**



Congestion can spread progressively through the network, building up the congestion tree.

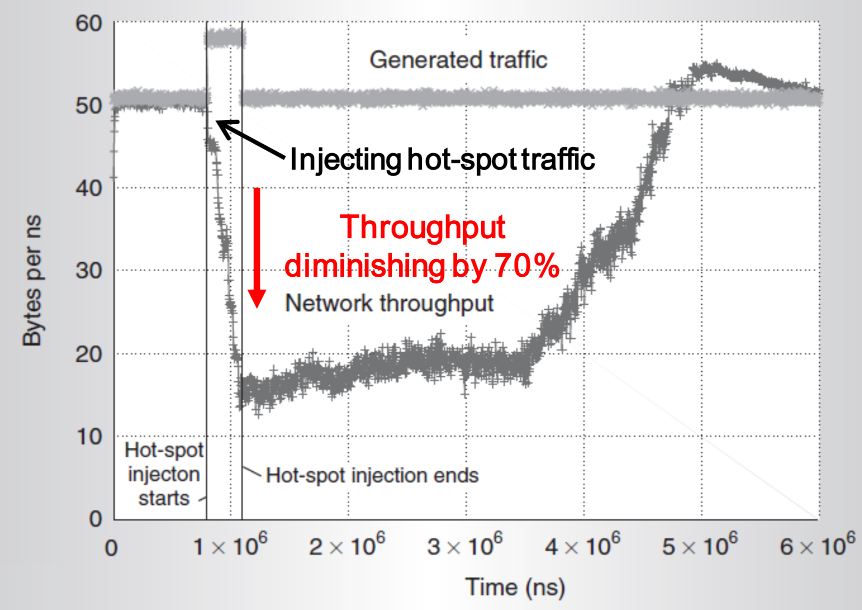
- **HOL-Blocking**



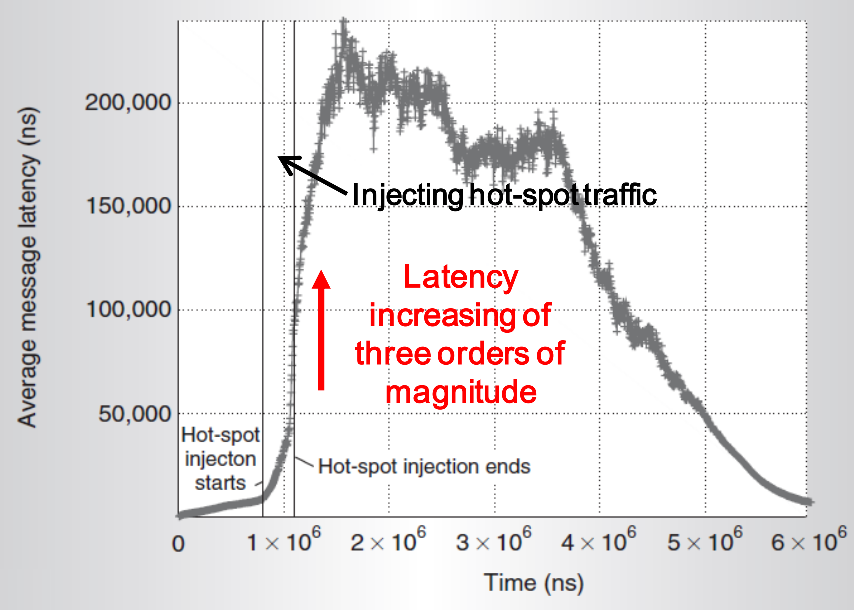
Congested flow can prevent uncongested flow in the same queue, resulting in HOL blocking. The impact of HOL blocking on network performance can be very serious: Network throughput can degrade and packet latency can increase dramatically.

The issues of Lossless Ethernet

- Issues of Lossless network have been well studied in academic community.
- The impact of HOL blocking on network performance can be very serious.
- As show in paper (Pedro J. Garcia et al, IEEE Micro 2006)[1]:



Network Throughput and Generated Traffic



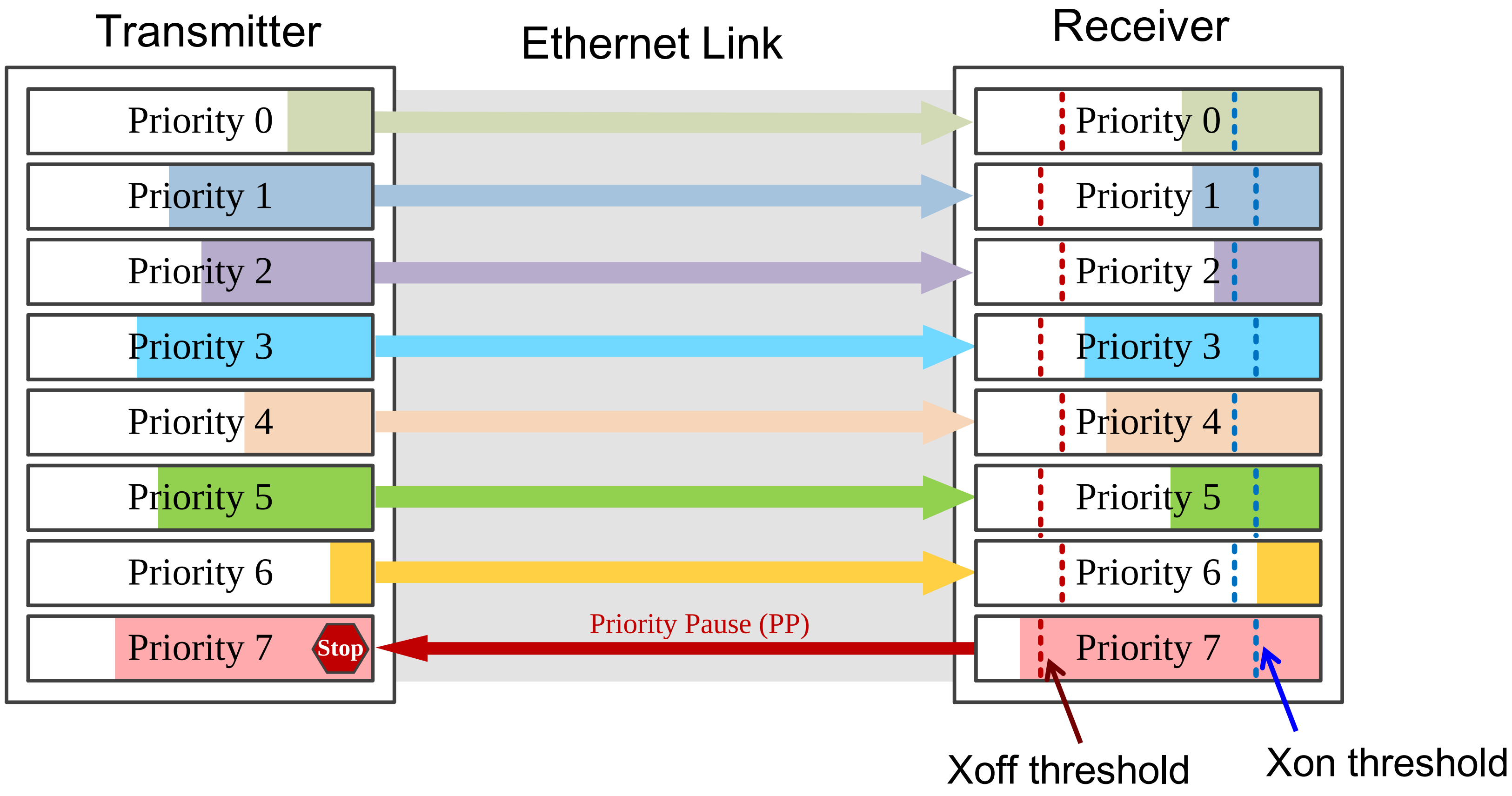
Average Packet Latency

Network Performance Degrades Dramatically after Congestion Appears

[1] Garcia, Pedro Javier, et al. "Efficient, scalable congestion management for interconnection networks." *IEEE Micro* 26.5 (2006): 52-66.

Current PFC Implementation

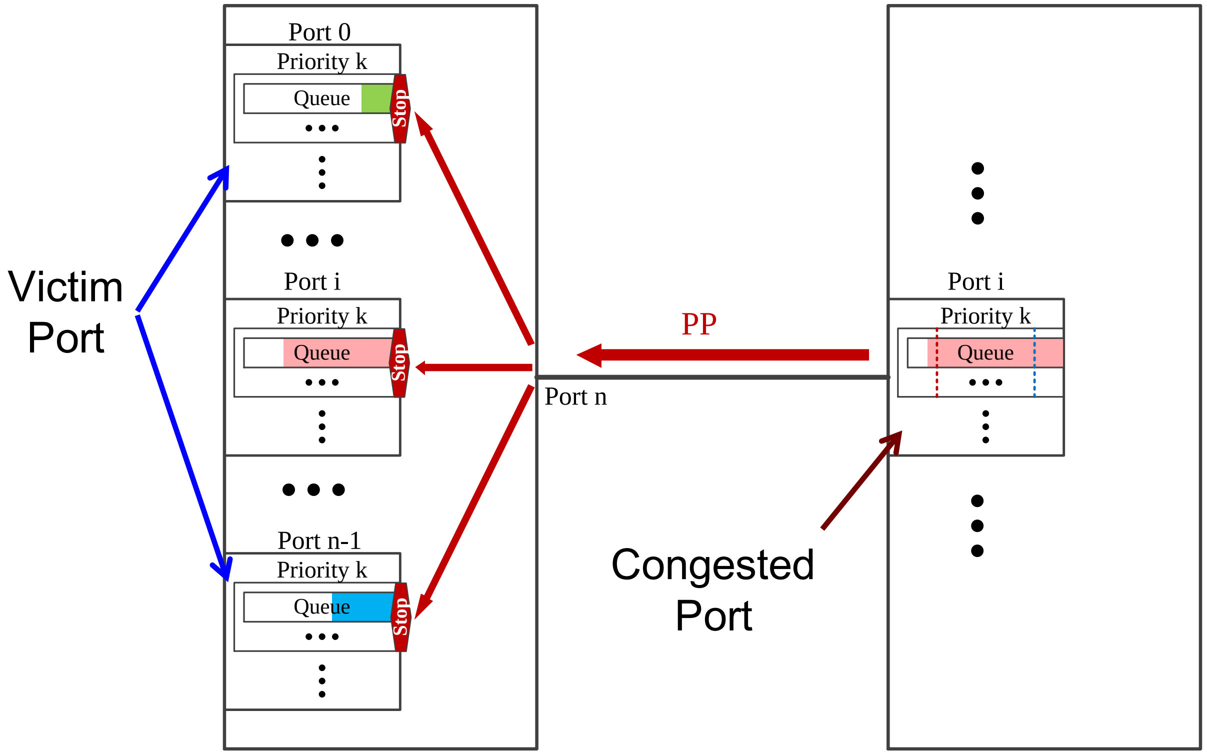
- IEEE 802.1Qbb PFC implement per-priority flow control, supporting eight priorities.
- Flow control of each priority can be enabled independently with individual Xoff/Xon thresholds.



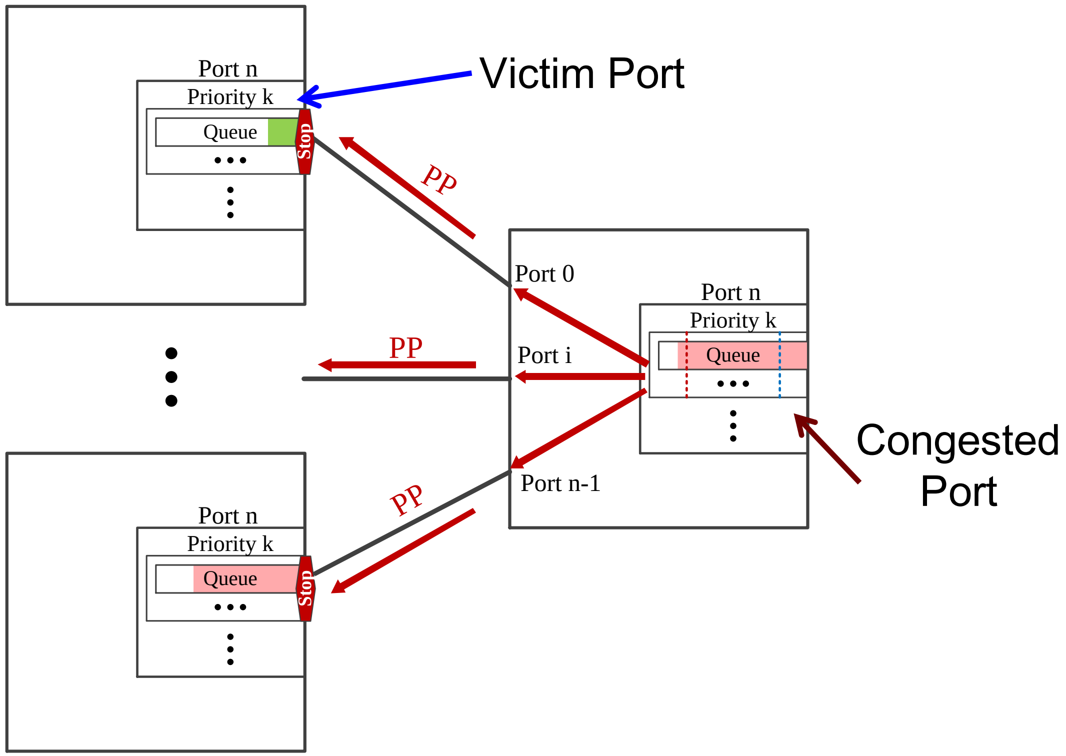
Issues of PFC

- Even more, one congested port in downstream side may block several ports in upstream side, because PFC is static (i.e. one priority k to one or several priority k).

Input Queue Model



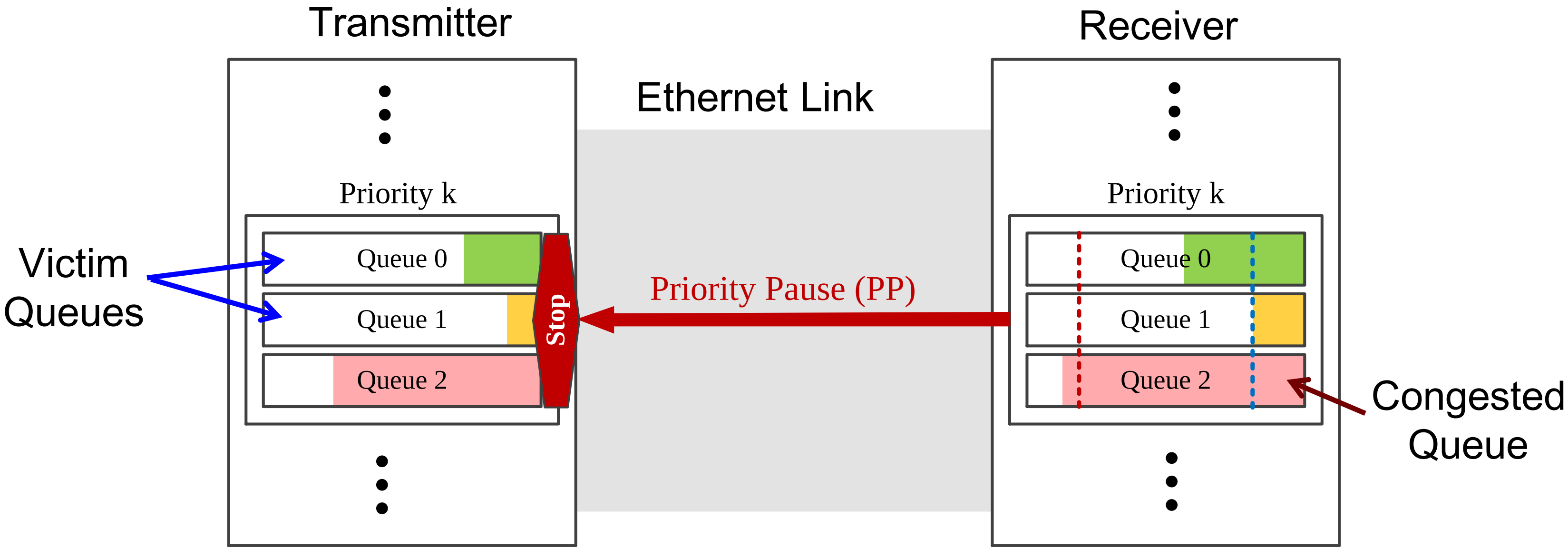
Output Queue Model



Ideally, only the port contributing to congestion should be blocked.

Issues of PFC

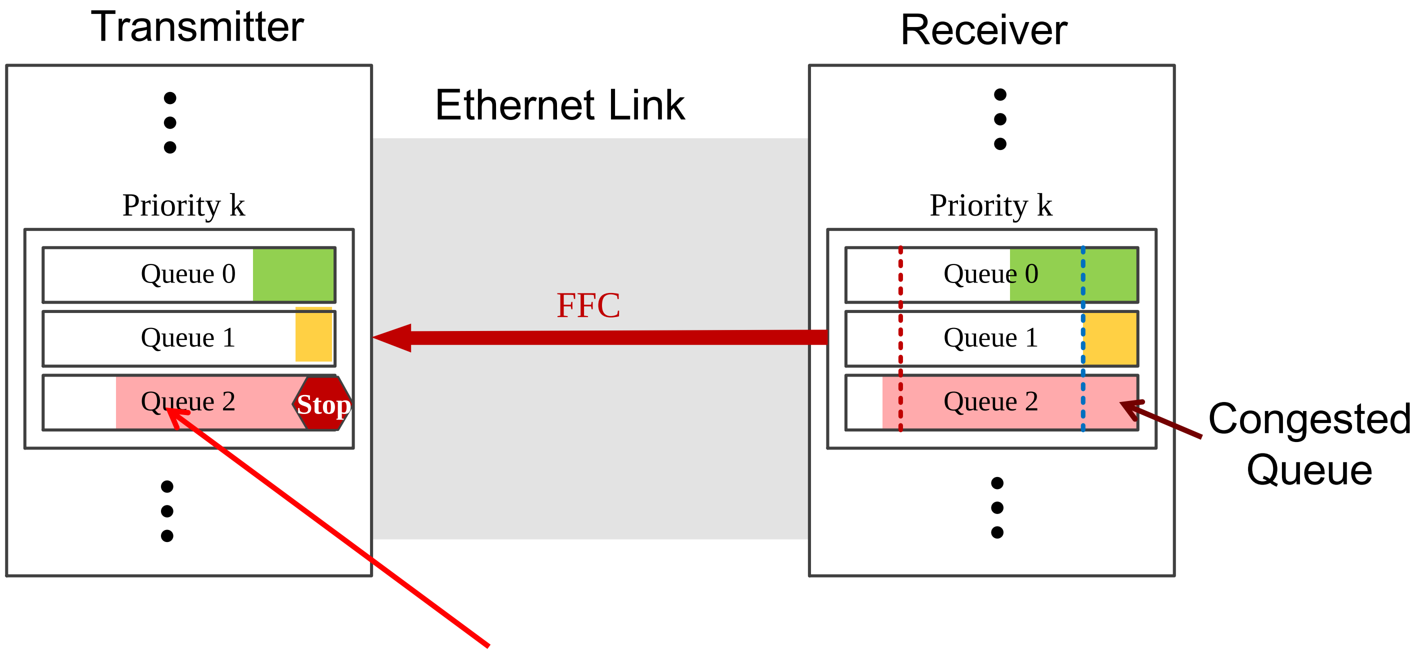
- Since there are hundreds or even thousands of applications in datacenter, a lot of applications have the same priority.
- Traffics of applications in datacenter are very dynamic and unpredictable, and they may affect each other.
- So PFC still suffers congestion propagation and HOL-blocking within each priority.
- Mismatch between coarse grained flow control (eight priorities in PFC) and rich queues in a port may result in victim queues.



Ideally, only the queue contributing to congestion should be blocked.

Proposal

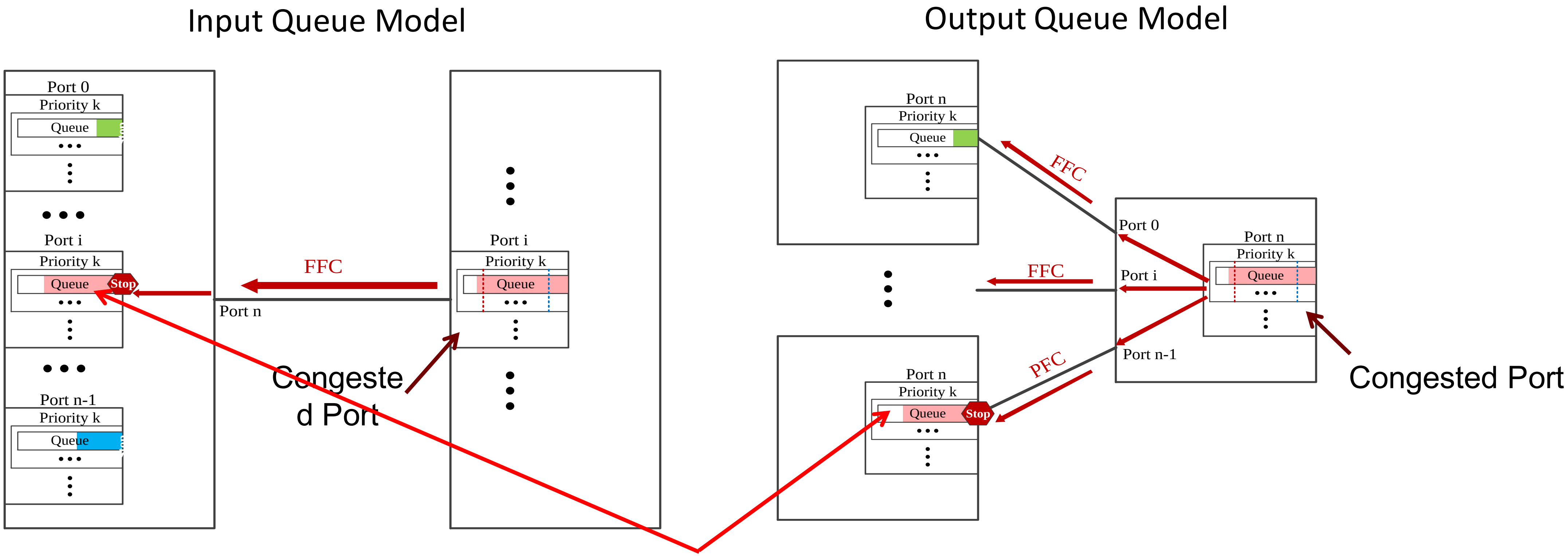
- We propose Flow-based Flow control (FFC):
  - FFC frame carrying flow information to indicate which flows to be paused.
  - On the downstream side, when queue occupancy reaches a threshold, a FFC frame indicating flows entering this queue is generate and sent upward.
  - On the upstream side, when a FFC frame is received, the flow information is parsed from the FFC frame and used to determine which queue to be paused.



Only the queue contributing to congestion is blocked.

Proposal

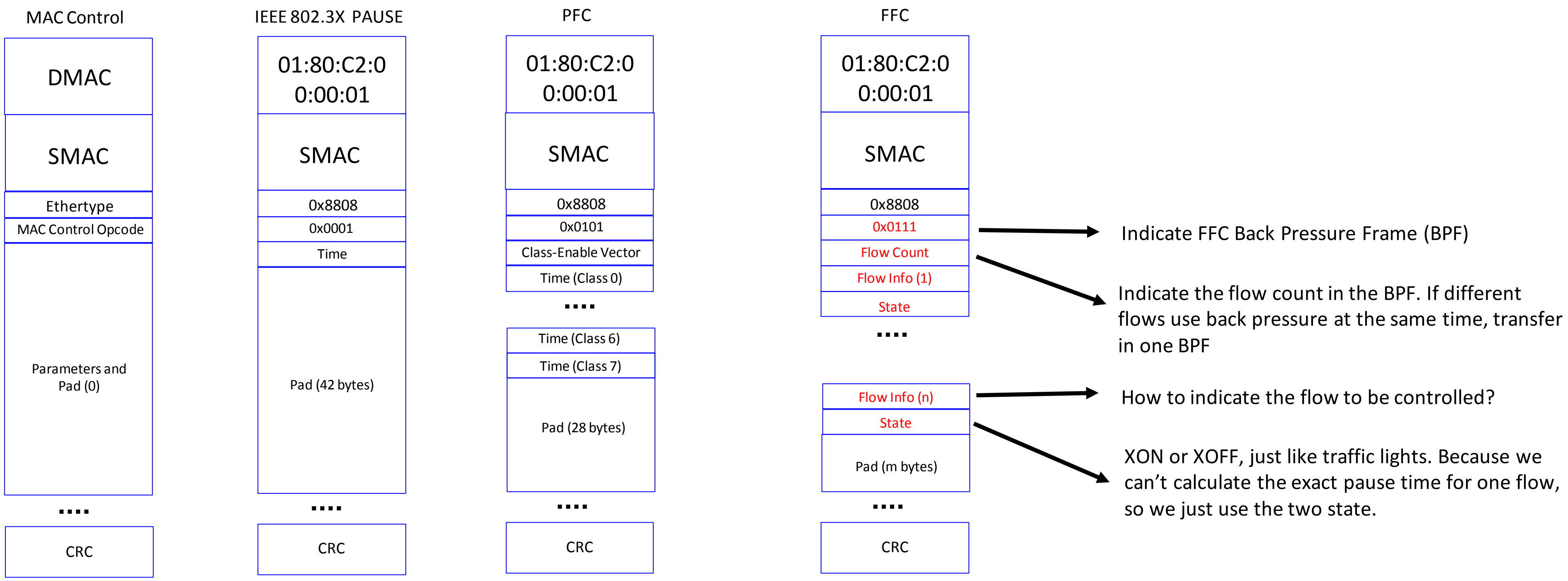
- As a result, two kind of dynamic abilities are acquired:
  - ➢ The granularity of flow control is dynamic: as the number of queue increases, finer grained flow control can be supported.
  - ➢ The queues to be blocked are dynamically determined: only the upstream queues including the flows aiming to the downstream congested queue (i.e. the flows whose information is carried in the FFC frame) will be blocked. => no victim queues.



Only the queue contributing to congestion is blocked.

Proposal

- Create FFC message frames carrying the necessary information for upstream to address the flow control per flow
  - Define the message format, identify the congested flow explicitly
  - Action: Xoff/Xon
- Requirement for switch chips
  - One or more aggregated flows will be mapped into the queues per requirement (PFC map flows to the queue per priority).



MAC Control

- DMAC
- SMAC
- Ethertype
- MAC Control Opcode
- Parameters and Pad (0)
- ....
- CRC

IEEE 802.3X PAUSE

- 01:80:C2:00:00:01
- SMAC
- 0x8808
- 0x0001
- Time
- Pad (42 bytes)
- ....
- CRC

PFC

- 01:80:C2:00:00:01
- SMAC
- 0x8808
- 0x0101
- Class-Enable Vector
- Time (Class 0)
- ....
- Time (Class 6)
- Time (Class 7)
- Pad (28 bytes)
- ....
- CRC

FFC

- 01:80:C2:00:00:01
- SMAC
- 0x8808
- 0x0111
- Flow Count
- Flow Info (1)
- State
- ....
- Flow Info (n)
- State
- Pad (m bytes)
- ....
- CRC

Indicate FFC Back Pressure Frame (BPF)

Indicate the flow count in the BPF. If different flows use back pressure at the same time, transfer in one BPF

How to indicate the flow to be controlled?

XON or XOFF, just like traffic lights. Because we can’t calculate the exact pause time for one flow, so we just use the two state.

Summary

- PFC is a coarse-grained flow control method and may suffer serious HOL blocking which will degrade the network performance dramatically.

- A fine-grained flow control mechanism should be considered in order to resolve the mismatch between the queue and the service priority.

- Flow-based flow control(FFC) can provide fine-grained and dynamic congestion management.

- Need to consider how to mapping the aggregated flows to the queues to acquire the explicit flow control.

- Need to consider the structure of FFC message frame and the inherited relationship with PFC PP(Priority Pause) frame.

Thank you

www.huawei.com