Enhance PFC

To support quantized flow control

Yuehua Wei (wei.yuehua@zte.com.cn)
Jinghai Yu (yu.jinghai@zte.com.cn)
Hesong Li (li.hesong@zte.com.cn)

IEEE 802.1 Meeting
March 11-15 , 2024
Denver, Colorado

Contents

- Recap of PFC and QCN
- A Rough Proposal

Recap of PFC and QCN

- PFC(Priority-based Flow Control) uses Pause to relieve congestion. When congestion occurs, the corresponding flow of PFC enabled priority on a link will pause while all of the other priorities on the link continue to send frames
  - PFC is standardized as IEEE 802.1Qbb (now incorporated into IEEE 802.1Q)
  - P802.1Qdt specifies automatic configuration of PFC headroom, and MACsec protection of PFC frames. (Ongoing)
- QCN(Quantized Congestion Notification) provides a means for a bridge to notify a source of congestion causing the source to reduce the flow rate. QCN is standardized as IEEE 802.1Qau (now incorporated into IEEE 802.1Q)
  - Provide a way to mitigate congestion spreading from link flow control such as PFC
  - Allow source to identify the flow to apply the rate limit

Why mention PFC and QCN together?

- Data Center Quantized Congestion Notification (DCQCN) is an end-to-end congestion control scheme for RoCEv2 which is widely deployed.
- DCQCN relies on ECN(RFC 3168) . It combines elements of DCTCP and **QCN**
- The idea behind DCQCN is to allow ECN to do flow control by decreasing the transmission rate when congestion starts, thereby minimizing the time **PFC** is triggered, which stops the flow altogether.
- **QCN** and **PFC** are both good L2 tools to deal with congestion.

Let’s recap technical features of PFC and QCN

	PFC	QCN
Goals: （From 802.1Q-2022）	PFC enables **to not discard frames due to congestion** for protocols that require this property	Congestion notification depends on the formation of a cooperating set of systems including VLAN Bridges and end stations to **achieve the reduction in frame loss**
Scope	Hop-by-Hop	End-to-End (CP to RP)
Source Action	Pause	Rate limit
Granularity	Coarse buffer-aware only	Fine buffer-aware + flow-aware
NIC based	No	Yes
Signaling protocol	PFC frame（IEEE MAC-specific Control Protocols with group address 01-80-C2-00-00-01） + DCBX (LLDP)	CNM (CN-TAG) + LLDP Congestion Notification TLV
Extra tag encap&decap	No	Yes. Source tags frames with a CN-Tag
Per priority queuing	Yes	Yes A CNPV consists of one value of the priority parameter such that all of the Bridges’ and end stations’ ports in a Congestion Notification Domain (CND) are configured to assign frames at that value to the same CP and/or an RP

Motivation to enhance PFC

Advantages of PFC

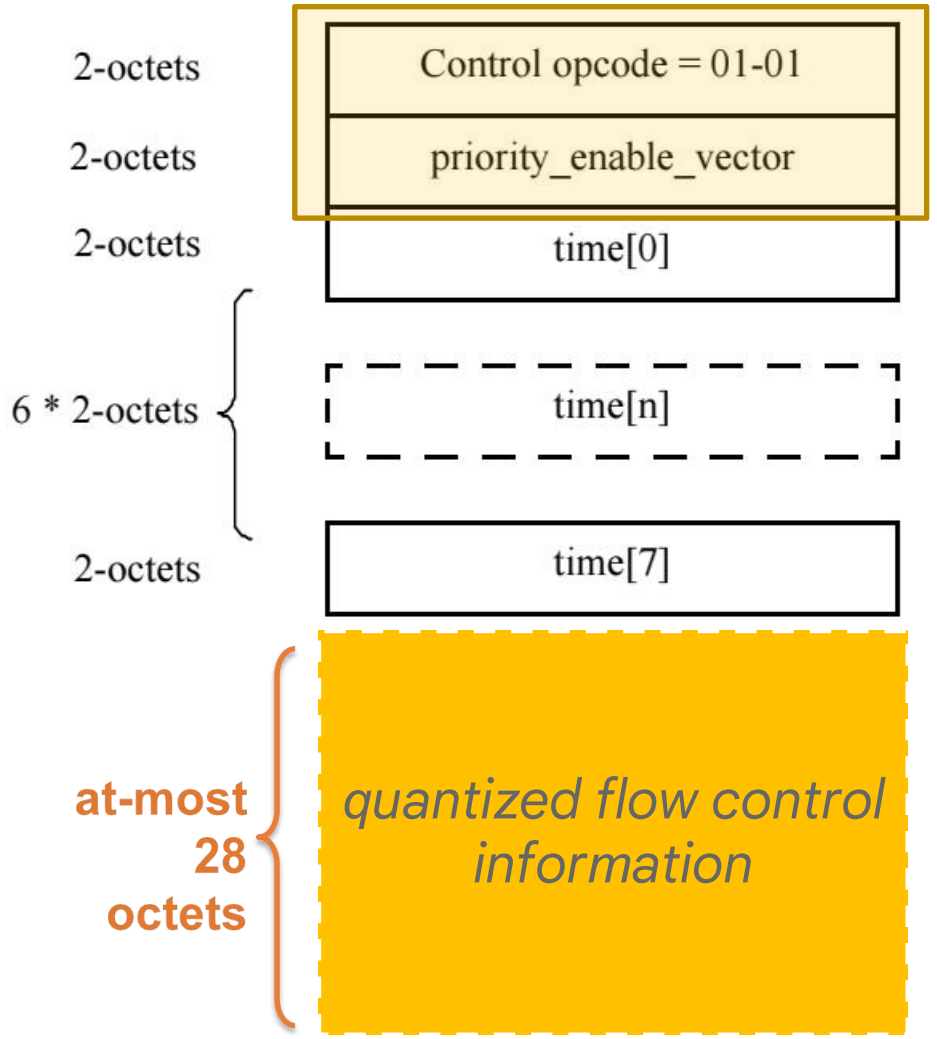
- PFC is relatively a SIMPLE mechanism.
  - It adopts LLDP as signaling protocol which is wildely deployed in L2 networks
  - It needn’t extra tag encapsulation and decapsulation for all the data frames which
    - Saves the frame process cost among the nodes along the path
    - Saves overhead cost of the total bandwidth
- PFC is widely deployed in FCoE and RoCE network
  - Nearly all major vendors support PFC in DC networks

Limitations of PFC

- If PFC pause time interval is not properly set， it may bring frame loss. Upper layer packets retransmission will lead congestion spread.
- Pause will cause preemption mechanism fail
- Pause may bring unfairness, deadlock and victim flow
- Read https://www.ieee802.org/1/files/public/docs2024/dt-seaman-clause-36-proposal-0124-v1.pdf for more detailed analysis

What's the idea and what are we going to get?

- The idea is quite simple: extend PFC to support quantized flow control such as rate limit.
- Then
  - We will have a HBH mechanism with the advantages the previous page talked about
  - PFC may inherit some good ideas from QCN, such as
    - buffer-aware + flow-aware
    - rate adjustment algorithm
  - Some critical flow will not be interrupted to avoid service fail
- PFC enhancements make Ethernet technology more applicable and appealing for data center environments.



Thank You