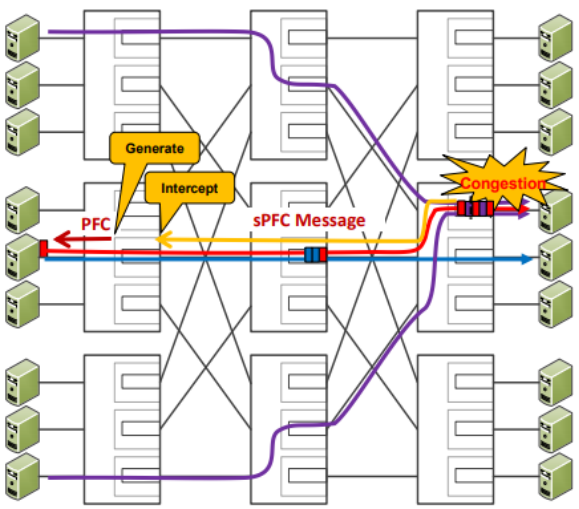
Consideration of sPFC/SFC issues when leveraging Qcz

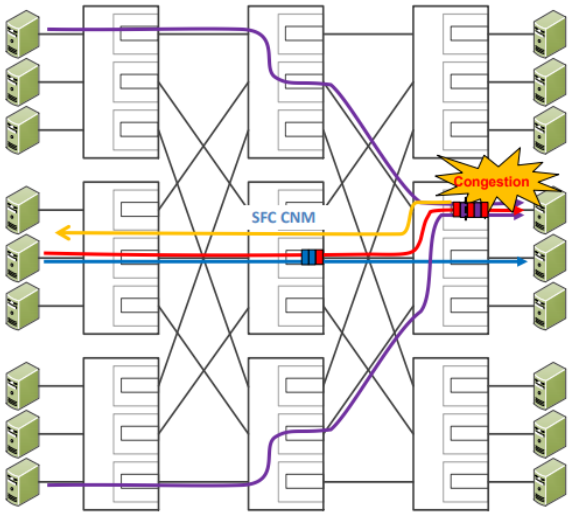
Lily Lv, Jinfeng Yan, Peiying Tao (Huawei)

Jeongkeun “JK” Lee (Intel)

Background

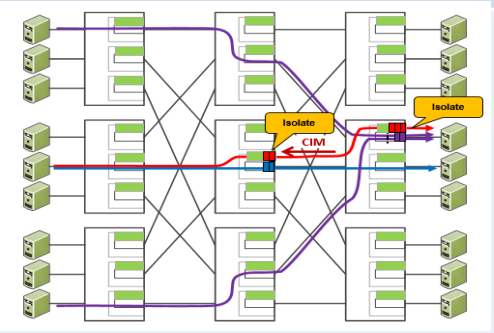
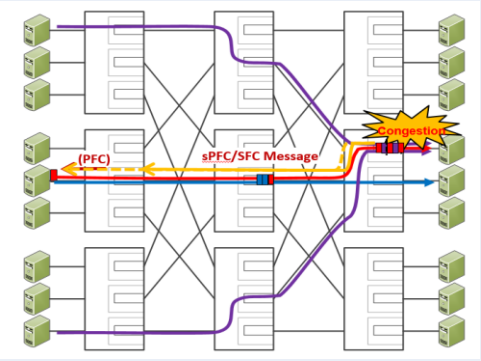
- sPFC/SFC has been discussed in Nendica and IETF
  - Latest contribution: https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne-leveraging-qcz-for-spfc-sfc.pdf
- Main concept of sPFC/SFC
  - sPFC = remote generation of PFC at the source ToR
  - SFC = pause the source at the flow level





- A few issues were raised in the latest contribution.
  - The CI Peer Table configures the UDP port to be used for L3 CIM. This is obtained from upstream neighbor through LLDP
    - Issue: ability to determine UDP port for distant L3 CIM receiver. Better to have well known UDP port used by all systems.
  - Qcz CIM security can use MACSec because it is hop-by-hop. How to secure edge-to-edge sPFC messages?
  - Should SFC message include Qau 'quantized' parameters?
  - When combining with Congestion Isolation, how to identify the source priority to pause (congesting queue or non-congesting queue)?
- This presentation intends to address above issues.

Root Cause Of The Issues

Comparison between CIM message and sPFC/SFC message		
CIM message	**Hop by hop message**	**Analysis**
	• Both L2 and L3 CIM messages are point to point, from downstream congestion point to upstream neighbor • Upstream egress is directly connected to downstream ingress. Frames do not change priority over the wire.	• The message can be protected by MACsec. • LLDP can be used to negotiate parameters in CIM message, like L3 UDP port. • Downstream ingress can determine upstream egress priority/queue, to properly assert PFC
sPFC/SFC message	**Edge to edge message**	**Analysis**
	• sPFC/SFC message crosses the network, from congestion point to source • There may be multiple hops between the congestion point and source, and frame priority may change in between.	• MACsec in not intended for end-to-end security. • Lacking a method to negotiate parameters between multi hop distance switches. LLDP does not help in this case. • Congestion point has no information about congested flow queue status at the source.

Issue 1:

No global well known UDP port has been assigned by IETF. Qcz uses locally assigned UDP port for L3 CIM. The CI Peer Table configures UDP port to be used for L3 CIM. This is obtained through LLDP

- Issue: ability to determine UDP port for distant L3 CIM receiver. Better to have well known UDP port used by all systems.

Explanation/Solution:

- sPFC/SFC is intended to be used in a data center network which is a closed environment within a single administrative domain.
- It is possible to configure the DCN, setting a UDP port number for sPFC/SFC messages by administration.
- Alternatively, request IETF to assign a dedicated UDP port number to sPFC/SFC message.

Issue 2:

Qcz CIM security can use MACSec because it is hop-by-hop. How to secure edge-to-edge sPFC messages?

Explanation/Solution:

- Refer to IETF 112 ICCRG presentation.

  - How is the protocol secured? concerns of spoofing the control messages
    - For a single-domain data center of trusted switching devices
    - Signaling between switches (for SPFC) ~= LLDP or BGP
      - Note) BGP encryption may stop a man-in-the-middle attack; but doesn't solve the problem of a malicious or poorly implemented router
    - SFC signaling to sender transport ~= ECN marking
    - ACL at domain boundaries can block signal pkts coming from NIC/host/outside

- For IPsec supported environment, Layer 3 sPFC/SFC message can also be protected by IPsec.

Issue 3:

Should SFC message include Qau ‘quantized’ parameters?

Explanation/Solution:

- Qau specifies ‘quantized’ parameter Fb. CNM message carries Fb to host as input of rate calculation.

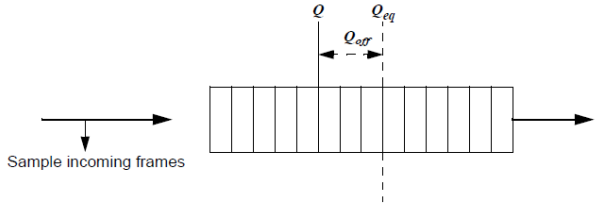


Figure 30-1—Congestion detection in QCN CP

Let Q denote the instantaneous queue size and Q_{old} denote the queue size when the last feedback message was generated. Let $Q_{off} = Q - Q_{eq}$ and $Q_{\delta} = Q - Q_{old}$.

Then F_b is given by the formula

$$F_b = -(Q_{off} + wQ_{\delta})$$

(From 802.1Q -2018 30.2.1 CP algorithm)

- For sPFC, Fb is not needed.
  - sPFC message is used to trigger PFC frame generation on source TOR switch. It does not need Fb, instead, pause time is useful.
- SFC could be part of the standardization. The proposal is to use the same principle as sPFC.
  - The SFC message is sent to host. Host interprets the message as if a PFC frame is received but against each flow/connection to pause. In this case, Fb is not needed.

Issue 4:

When combined with Congestion Isolation, how to identify the source priority to pause (congesting queue or non-congesting queue)?

Explanation/Solution:

- sPFC/SFC message includes information to identify the flow which should be paused, as well as pause time.
- Because of the provided flow information in the sPFC/SFC message, the source knows which queue(priority) needs to be paused.
- PFC can be generated to the source accordingly.

sPFC/SFC interaction with congestion controls

- Theoretically, L3-QCN is a possible way to mitigate congestion. But it is a congestion control approach with a completely different algorithm. sPFC/SFC is flow control.
  - L3-QCN is used to reduce transmission rate at the sender, while sPFC/SFC is to temporarily stop the transmission.
  - The thresholds set on the congestion point for congestion control and flow control are different.
  - Like other congestion control mechanisms, e.g. DCQCN, L3-QCN can be combined with SFC.
  - Hence, including both Fb and pause time in the same message will confuse the source.
- sPFC/SFC as a flow control handles *transient* congestions, while leaving *equilibrium* behavior to existing congestion controls
  - The benefit of modularizing congestion mgmt. into *transience* and *equilibrium* behaviors and tacking the former by instant flow control is well presented at *On-Ramp @ NSDI'21. On-Ramp* measures one-way e2e delay at the end-hosts; the receiver signals it back to the sender for flow control.
  - sPFC/SFC provides the same benefit but with 1-2 orders of magnitude faster control loop.
    - sPFC control loop < NIC-to-NIC congestion-free HW RTT << On-Ramp e2e signal and reaction between host software stack.

Conclusion

- sPFC/SFC is intended for single-domain data center network. Raised issues could be resolved under this pre-condition.
- Propose to use same principle for sPFC and SFC, that is a new flow control mechanism leveraging existing PFC to mitigate in-cast congestion.
- sPFC/SFC can be combined with CI, and other popular congestion control mechanism, such as DCQCN