Leveraging Qcz for Source PFC and/or Source Flow Control

Paul Congdon

802.1 November Plenary, electronic

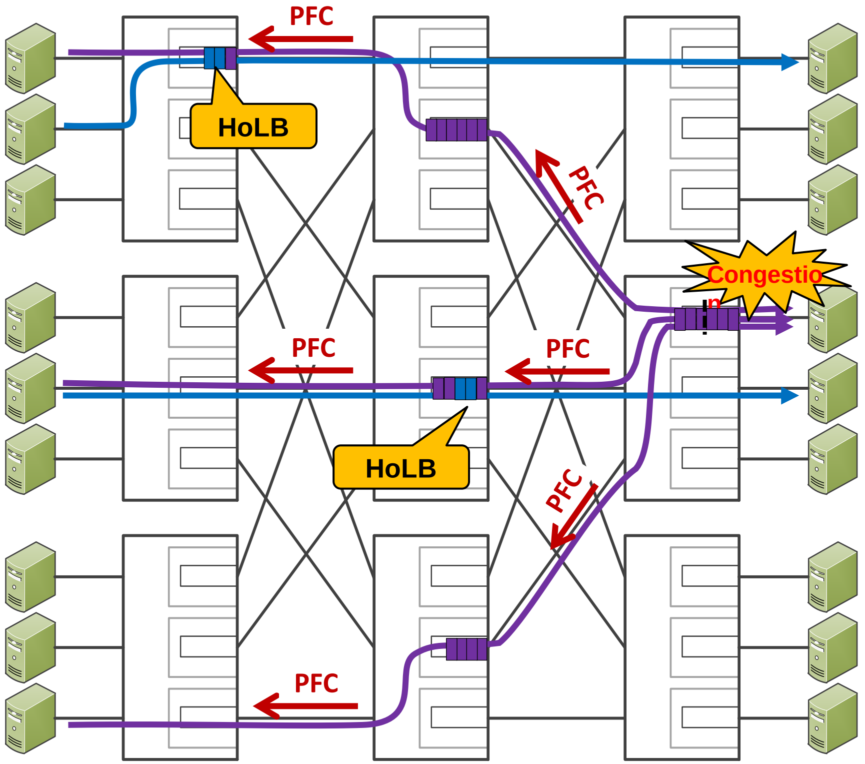
November 11, 2021

Outline

- Existing 802.1 Data Center Congestion Control
- Future 802.1 Data Center Congestion Control
- sPFC vs SFC
- Leveraging Qcz
- Issues to consider
- Next Steps
- History/Background

Existing 802.1 Congestion Management Tools

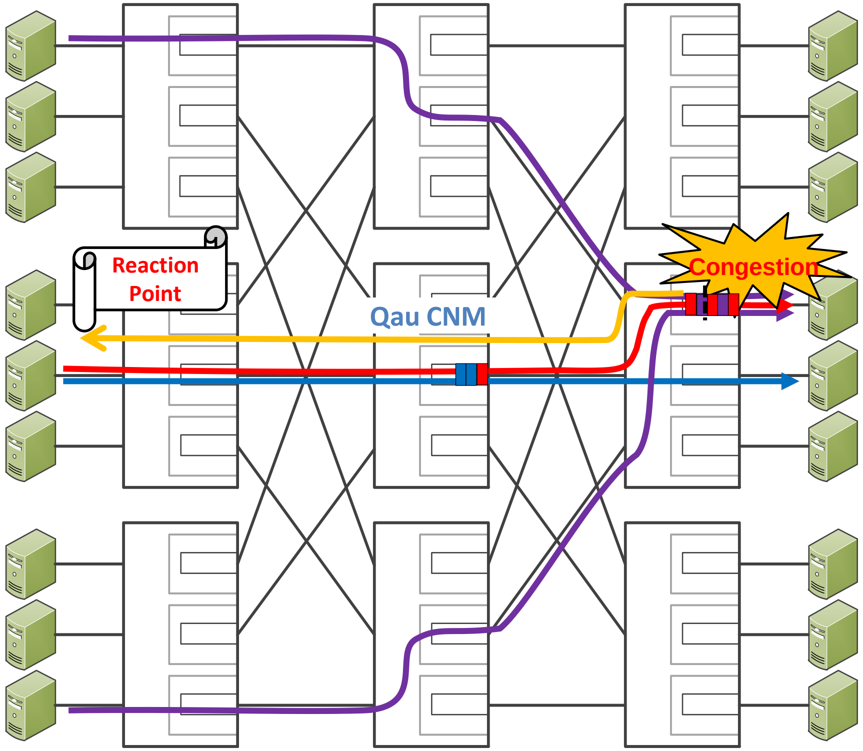
802.1Qbb - Priority-based Flow Control



Concerns with over-use

- Head-of-Line blocking
- Congestion spreading
- Buffer Bloat, increasing latency
- Increased jitter reducing throughput
- Deadlocks with some implementations

802.1Qau - Congestion Notification



Concerns with deployment

- Layer-2 end-to-end congestion control
- NIC based rate-limiters (Reaction Points)
- Designed for non-IP based protocols
  - FCoE
  - RoCE – v1

Future 802.1 Congestion Management Tools

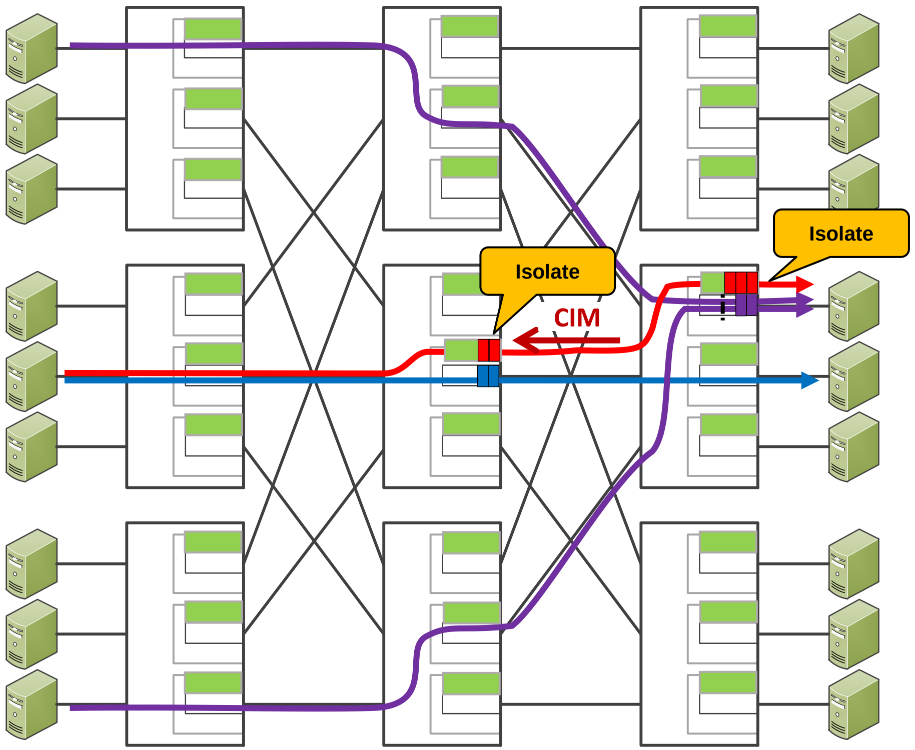
Congested Flow

Congested Queue 🟩

Victim Flow

Normal Queue □

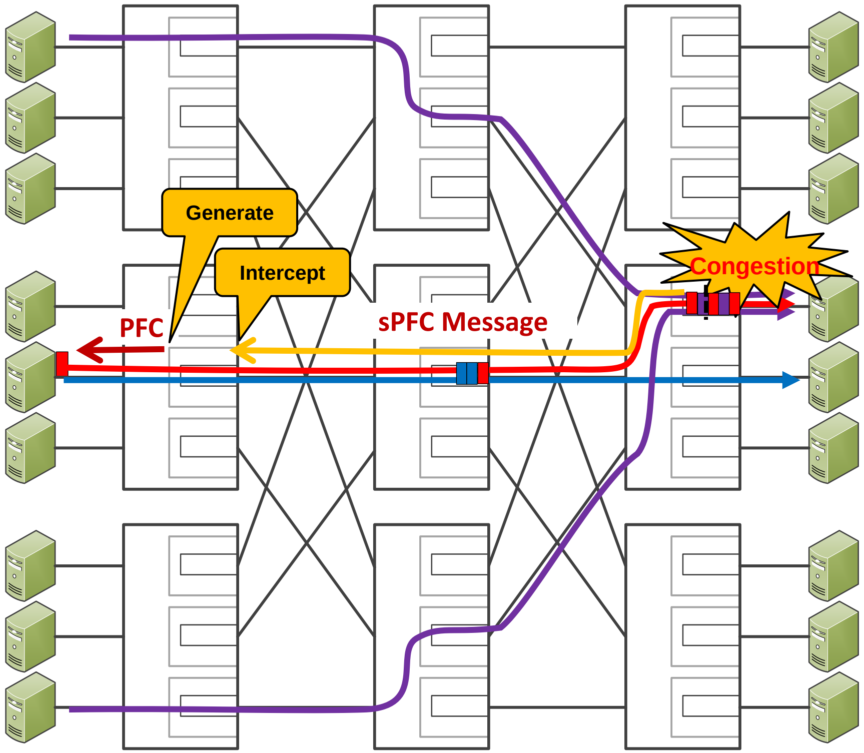
P802.1Qcz - Congestion Isolation



Implementation details

- Congesting flows are isolated locally first
- As queues continue to congest, CIM is generated and sent to upstream bridge/router
- CIM can be L2 or L3 message to support L3 networks (common deployment model).

Source PFC



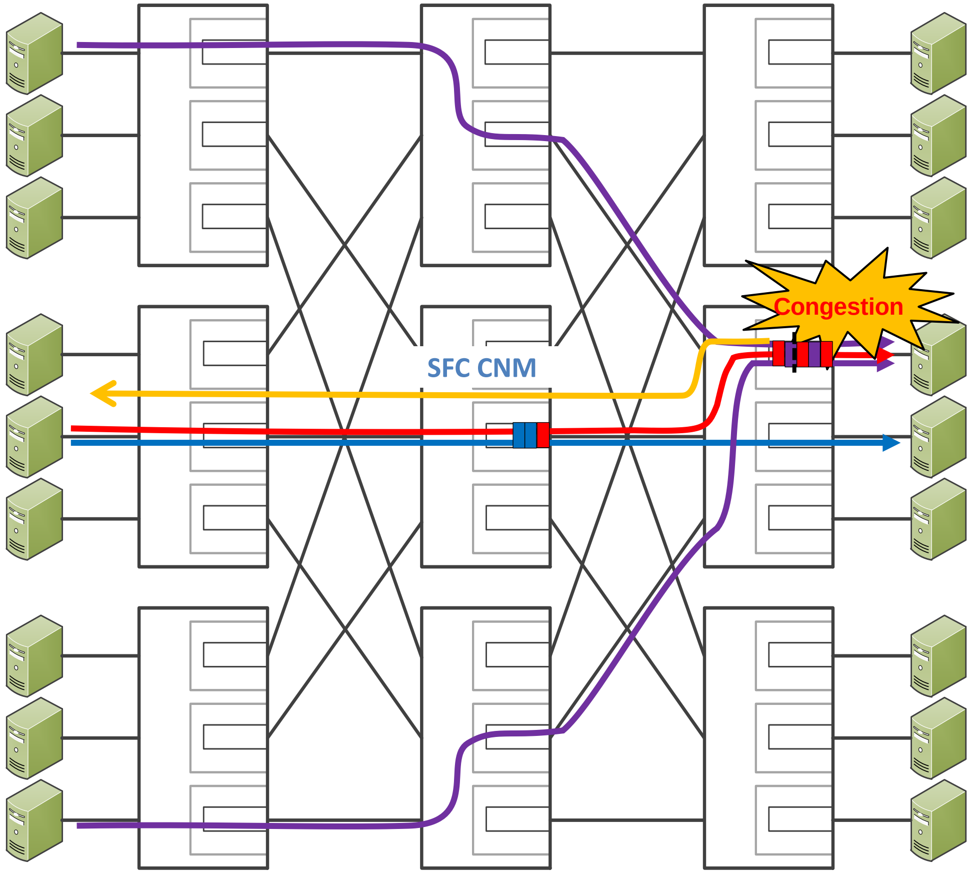
Details

- Can be combined with Congestion Isolation
- If congestion persists, Edge-to-edge signaling using L3 message
- Existing PFC generated at last hop
- NOTE: signaling message could pass to end-station directly if supported.

Source PFC vs Source Flow Control

- sPFC = remote generation of PFC at the source ToR
- SFC = pause at the flow level
- sPFC signaling message direct to end-point
- Basically, a L3 version of 802.1Qau (L3-QCN)
- NOTE: RoCEv2 DCQCN is a L3 adoption of QCN, using the ECN end-to-end congestion control loop

Source Flow Control



What is needed in sPFC/SFC signaling messages?

- Source and destination IP addresses of the data pkt
  - SRC IP for reverse forwarding
  - (Optional) DST IP for caching pause time per dst IP at sender ToR
  - simply swap src IP <-> dst IP from the data pkt into the signal packet; or need to 'learn' sender-ToR
  - DSCP and/or PCP, as needed to identify the PFC priority @ sender NIC
  - Pause time duration **<=** minimal drain time to reach the target queue level
  - (Optional) congestion locator such as congested switch/port/queue IDs

- Additional information for true 'source' flow control (SFC)
  - More tuples of the data pkt, e.g., L4 ports, to identify the sender flow/connection
  - (Note) L4 congestion control becoming part of NIC HW

Levering Qcz Congestion Isolation Message (CIM)

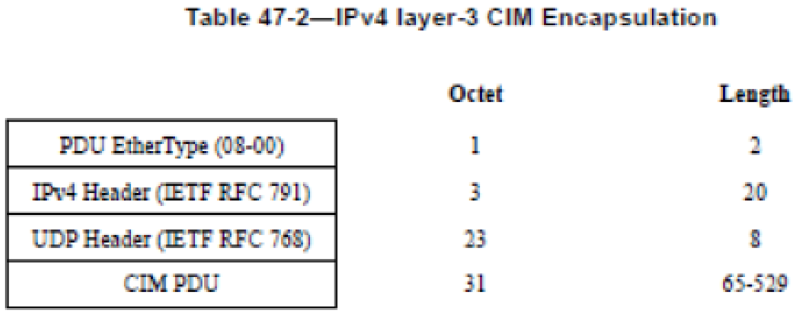


Table 47-2—IPv4 layer-3 CIM Encapsulation

	Octet	Length
PDU EtherType (08-00)	1	2
IPv4 Header (IETF RFC 791)	3	20
UDP Header (IETF RFC 768)	23	8
CIM PDU	31	65-529

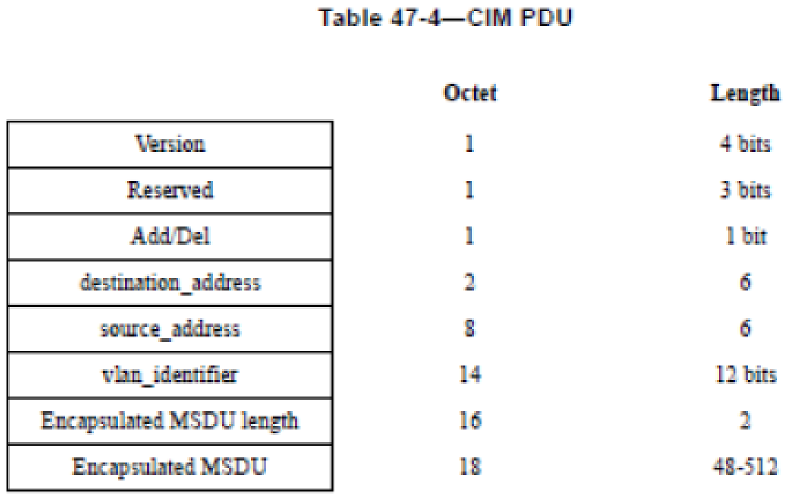


Table 47-4—CIM PDU

	Octet	Length
Version	1	4 bits
Reserved	1	3 bits
Add/Del	1	1 bit
destination_address	2	6
source_address	8	6
vlan_identifier	14	12 bits
Encapsulated MSDU length	16	2
Encapsulated MSDU	18	48-512

- Qcz CIM has Layer-2 and Layer-3 formats
- The CIM PDU contains enough of the payload to identify the offending flow
- Carrying the needed information:
  - Src / Dest IP addresses
  - DSCP
  - Additional tuples of the data pkt
- What’s missing?
  - Pause time
  - Simplified format of above information (i.e not MSDU)
  - Selection of CIM Destination IP (NOT previous hop)

Leveraging the Qcz reference architecture

- Believe it or not, these figures are similar…

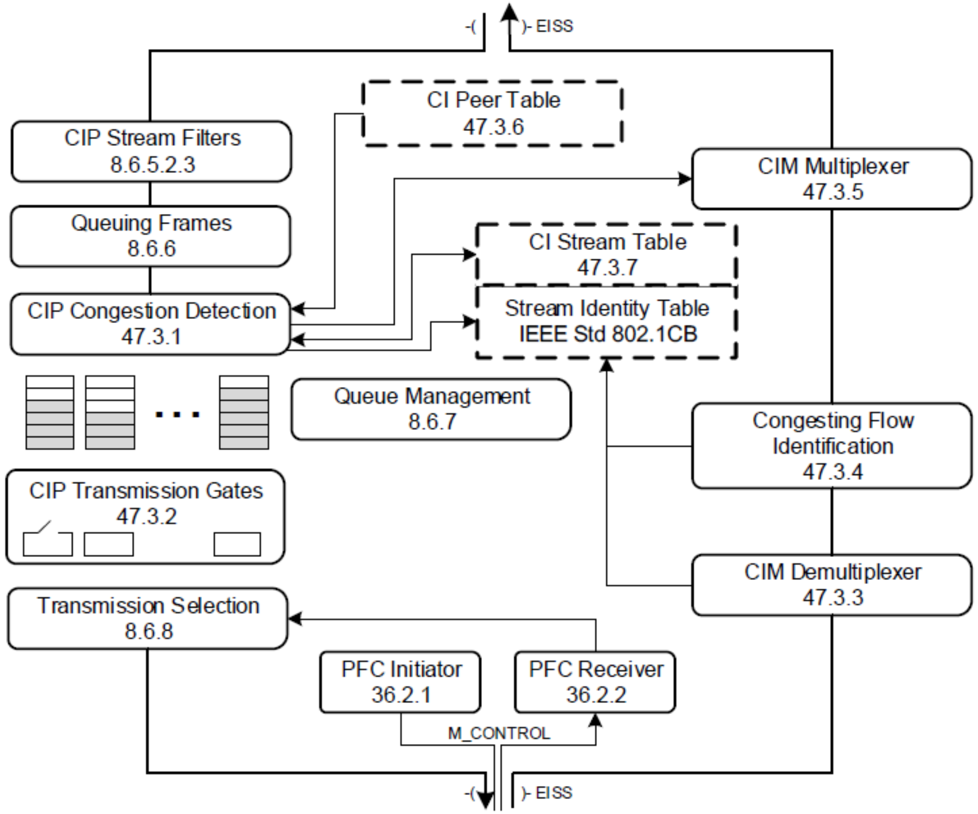
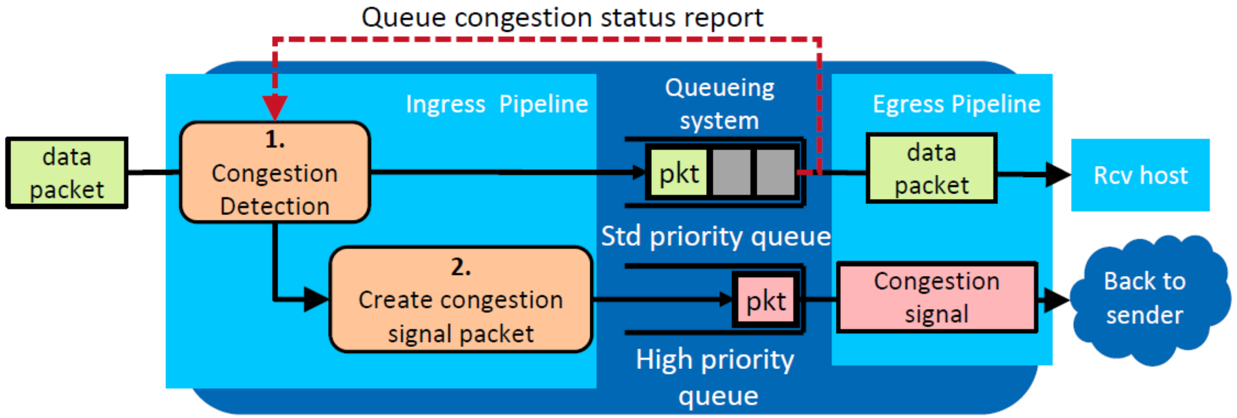


Figure 47-2—Congestion Isolation reference diagram



- Above figure is from https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00
- Congestion detection above (1) is similar to 47.3.1, but perhaps with different thresholds
- Creating signaling packet above (2) is similar to input to CIM Multiplexer 47.3.5, but with different parameters to CIM creation (e.g. Dest IP address)
- CI Peer Table 47.3.6 is used to identify upstream bridge/router – not needed by sPFC – address is in frame.
- CI Stream Table 47.3.7 could be used by Source Flow Control mode, but not needed for sPFC
- CIM Demultiplexer 47.3.3 could be used to intercept sPFC messages?

Issues to consider

- CI Peer Table also 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.
- 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)?
- Others…

Next steps

- Ongoing technical discussions
- Analysis of impact on 802.1Q for an amendment
- Continue to work towards authorization for PAR & CSD development at March 2022 Plenary

History and background material

- Public presentations of the concept and data at P4 Workshops (Apr'20, May'21) and Open Fabrics Alliance (Mar'21)
  - https://opennetworking.org/wp-content/uploads/2020/04/JK-Lee-Slide-Deck.pdf (slide 12)
  - https://www.openfabrics.org/wp-content/uploads/2021-workshop-presentations/503_Lee_flatten.pdf
  - https://opennetworking.org/wp-content/uploads/2021/05/2021-P4-WS-JK-Lee-Slides.pdf (slide 14)
- Previous Nendica presentations
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf - 9/16/2021
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0061-00-ICne-source-remote-pfc-test.pdf – 10/14/2021
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0067-00-ICne-source-remote-pfc-status-update.pdf – 11/04/2021
- IETF Awareness
  - Topic raised at IEEE 802 / IETF Coordination call – 10/25/2021
  - https://datatracker.ietf.org/meeting/112/materials/slides-112-iccrg-source-priority-flow-control-in-data-centers-00 - 11/08/2021