Source Flow Control Overview

Paul Congdon

802.1 January Interim, electronic

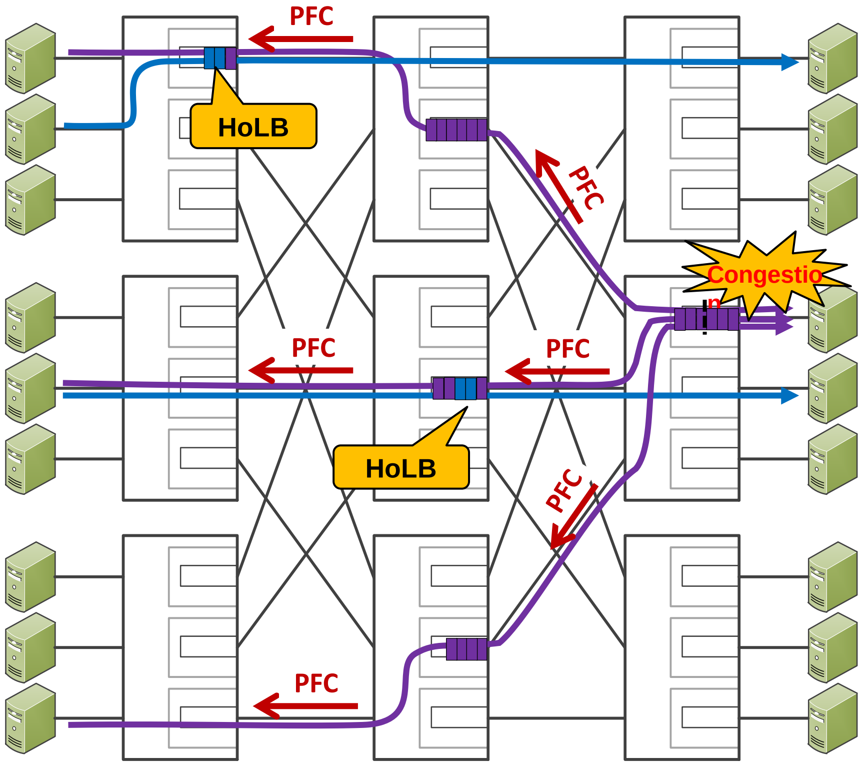
January 17, 2022

Outline

- Existing 802.1 Data Center Congestion Control
- Future 802.1 Data Center Congestion Control
- Leveraging Qcz signaling
- Next Steps
- History/Background

Existing 802.1 Congestion Management Tools

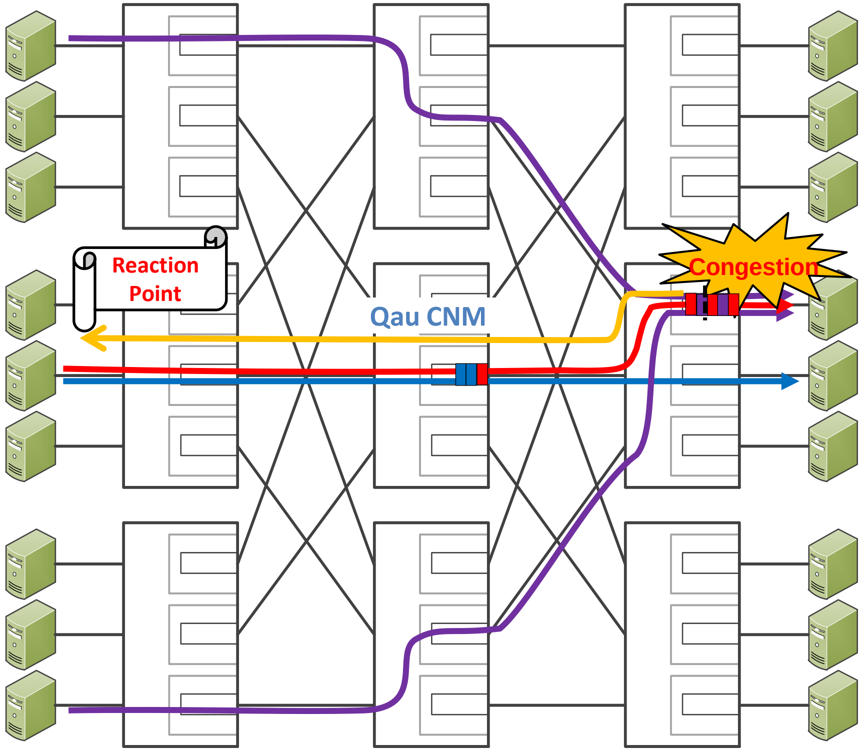
802.1Qbb - Priority-based Flow Control



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

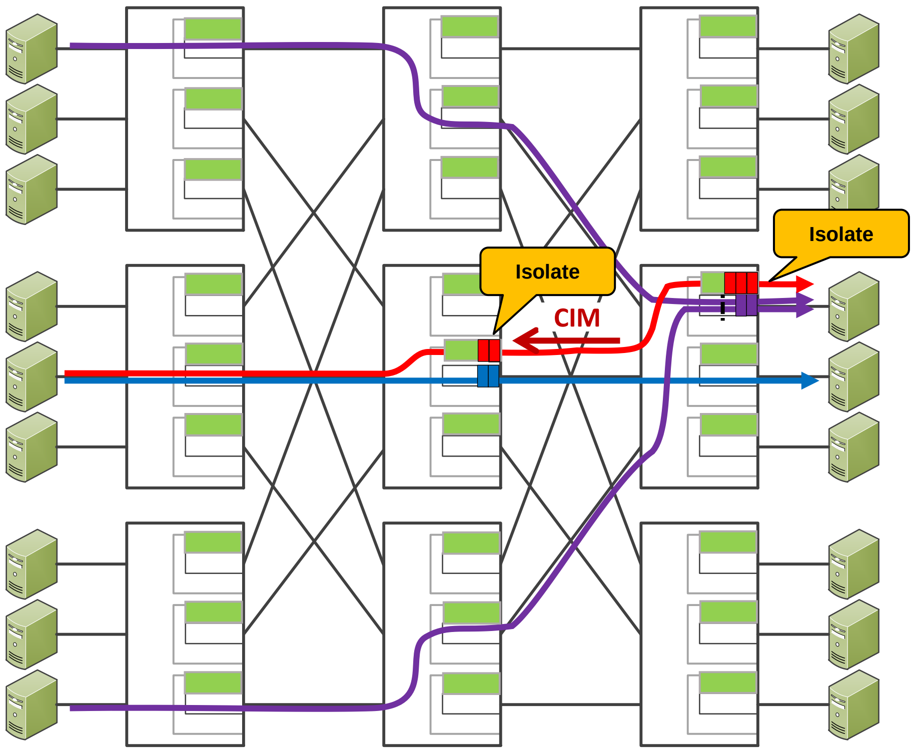
Congested Flow →

Congested Queue 🟩

Victim Flow →

Normal Queue □

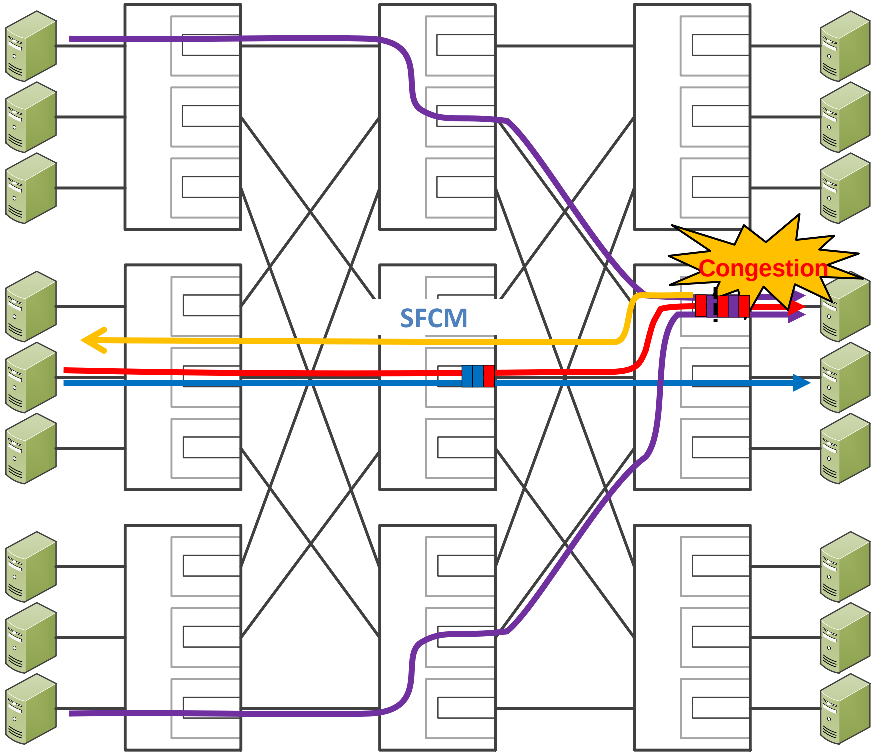
P802.1Qcz - Congestion Isolation



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 Flow Control



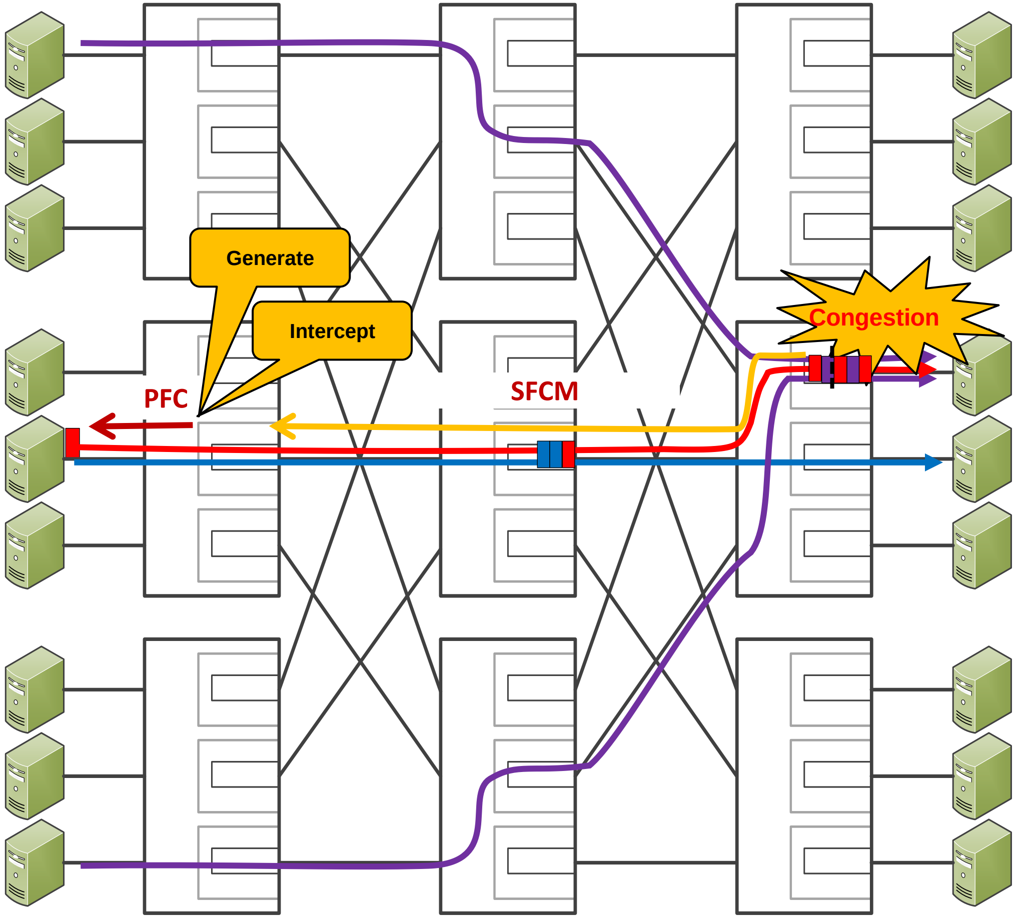
Details

- Can be combined with Congestion Isolation
- If congestion persists, Edge-to-Source signaling using L3 message
- Somewhat like a L3 version of 802.1Qau (L3-QCN), but no Reaction Point (RP) rate controller defined – instead, this is Flow Control
- Optional source Top-of-Rack switch involvement (see next slide)

Top-of-Rack Source Flow Control (proxy)

- Important use case for early deployment.
- ToR intercepts SFCM to non-supporting host using an egress stream_filter matching SFCM UDP port number
- ToR generates traditional PFC frame from SFCM

Source Flow Control (ToR Proxy)



What is needed in the SFCM signaling message?

- 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 source ToR
    - simply swap src IP <-> dst IP from the data pkt into the signal packet
- DSCP and/or PCP, as needed to identify the PFC priority @ source NIC
- Pause time duration **<=** minimal drain time to reach the target queue level
- (Optional) congestion locator such as Topology Recognition level to identify ‘incast’ congestion verses ‘in-network’ congestion.

Levering Qcz Congestion Isolation Message (CIM)

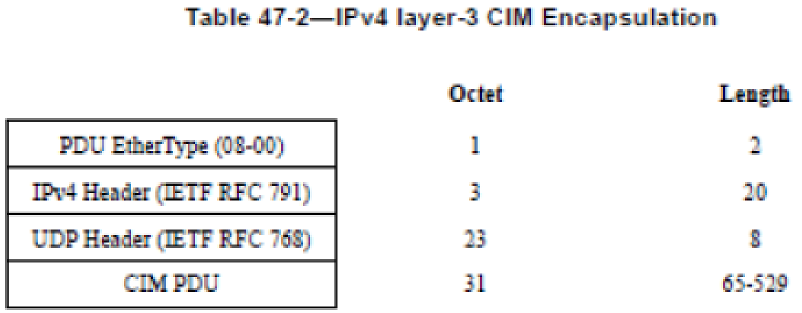


Table 47-2—IPv4 layer-3 CIM Encapsulation

	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

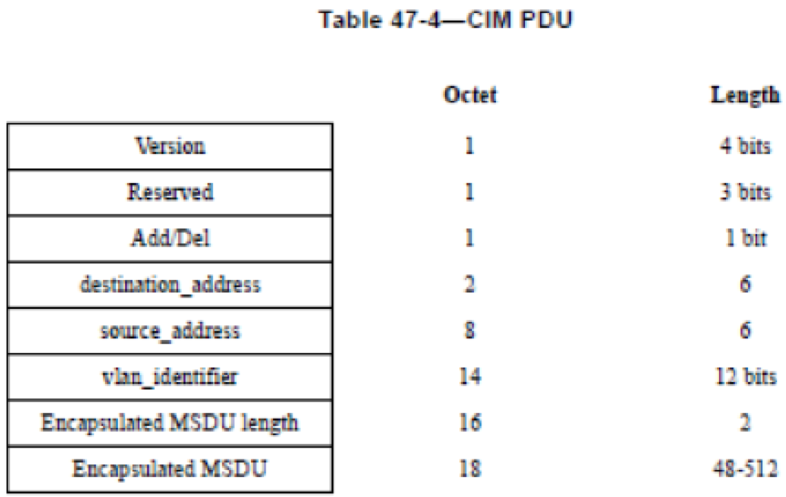


Table 47-4—CIM PDU

	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)

Next steps

- Ongoing design team discussion and analysis – new participants welcome
- Ongoing technical discussions in Nendica
- 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
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0077-00-ICne-consideration-of-spfc-sfc-issues-when-leveraging-qcz.pdf - 12/16/2021
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0079-00-ICne-spfc-sfc-next-steps.pdf - 12/23/2021
- 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