IEEE P802.1Qcz
Congestion Isolation

Update for San Diego Plenary

July 8, 2018

Paul Congdon

paul.congdon@tallac.com

PAR and CSD Status Update – P802.1Qcz

- Refined PAR and CSD from May Interim has been pre-circulated for July Plenary. The latest versions are available here:
  - http://www.ieee802.org/1/files/public/docs2018/cz-draft-PAR-0518-v02.pdf
  - http://www.ieee802.org/1/files/public/docs2018/cz-draft-CSD-0518-v01.pdf
- All comments from previous pre-circulation where resolved in March.
- Awaiting new comments for July – Response due Wednesday evening

Progress since May

- May Interim presentations:
  - PAR & CSD update
    - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-ci-update-0518-v1.pdf
  - Analysis Response
    - http://www.ieee802.org/1/files/public/docs2018/cz-escuderosahuquillo-CIAnalysis-response-0518-v01.pdf
  - New simulation model
    - http://www.ieee802.org/1/files/public/docs2018/cz-sun-ci-simulation-update-0518-v01.pdf
  - Need for project
    - http://www.ieee802.org/1/files/public/docs2018/cz-gafni-ci-need-0518-v1.pdf
- TSN conference call on June 11th, discussing changes to 802.1Q-2018
  - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-ci-Q-changes-0618-v1.pdf
- Informal design team discussions

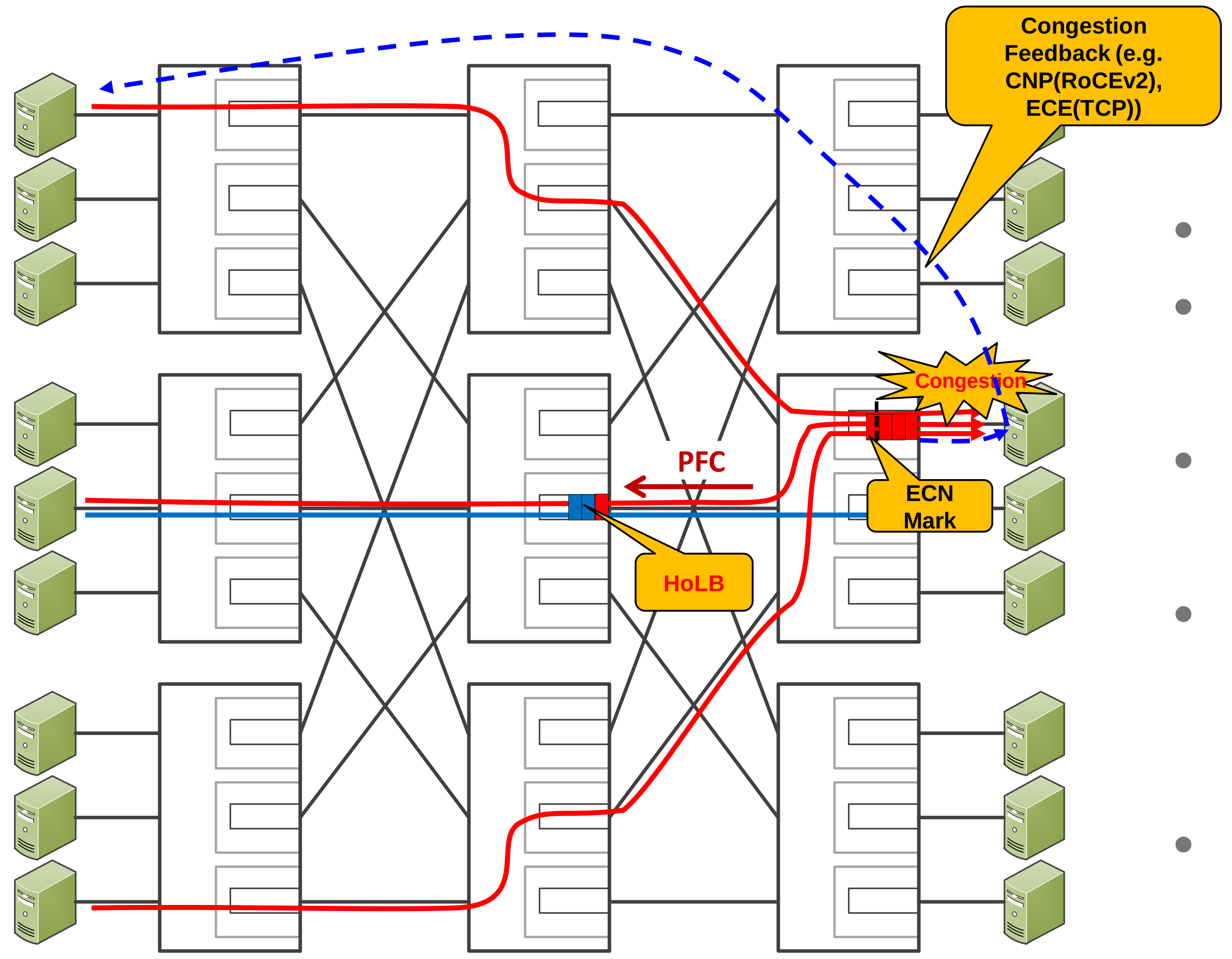
Progress since March Plenary

- Project and Nendica activity introduced and discussed at London IETF-101
  - TSVWG
    - https://datatracker.ietf.org/doc/slides-101-tsvwg-sessb-41-congestion-isolation-in-ieee-8021/
  - ICCRG
    - https://datatracker.ietf.org/doc/slides-101-iccrg-proposed-ieee-8021qcz-work/
  - HOTRFC
    - http://snaggletooth.akam.ai/IETF-101-HotRFC/01-Congdon.pdf
- Technical detail review on TSN conference call – April 16th
  - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-congestion-isolation-review-0418-v1.pdf
- Refined simulation based upon open source models from published papers
  - Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., & Liron, Y., et al. (2015). Congestion Control for Large-Scale RDMA Deployments. ACM SigComm Computer Communication Review, 45(4), 523-536.
  - https://github.com/bobzhuyb/ns3-rdma

P802.1Qcz – Congestion Isolation

- Amendment to IEEE 802.1Q-2014
- Scope
  - Support the isolation of congested data flows within ***data center environments***, such as high-performance computing, distributed storage and central offices re-architected as data centers.
  - Bridges (aka L3 Switches) will:
    - individually identify flows creating congestion
    - adjust transmission selection (i.e egress packet scheduling) for those flows
    - signal congested flow information to peers as needed.
  - Reduces head-of-line blocking for uncongested flows sharing a traffic class.
  - Intended to be used with higher layer protocols that utilize end-to-end congestion control.

DCN state-of-the-art



- DCNs are primarily L3 CLOS networks
- ECN is used for end-to-end congestion control
- Congestion feedback can be protocol and application specific
- PFC used as a last resort to ensure lossless environment, or not at all in low-loss environments.
- Traffic classes for PFC are mapped using DSCP as opposed to VLAN tags

Isolate the congestion to mitigate HOLB

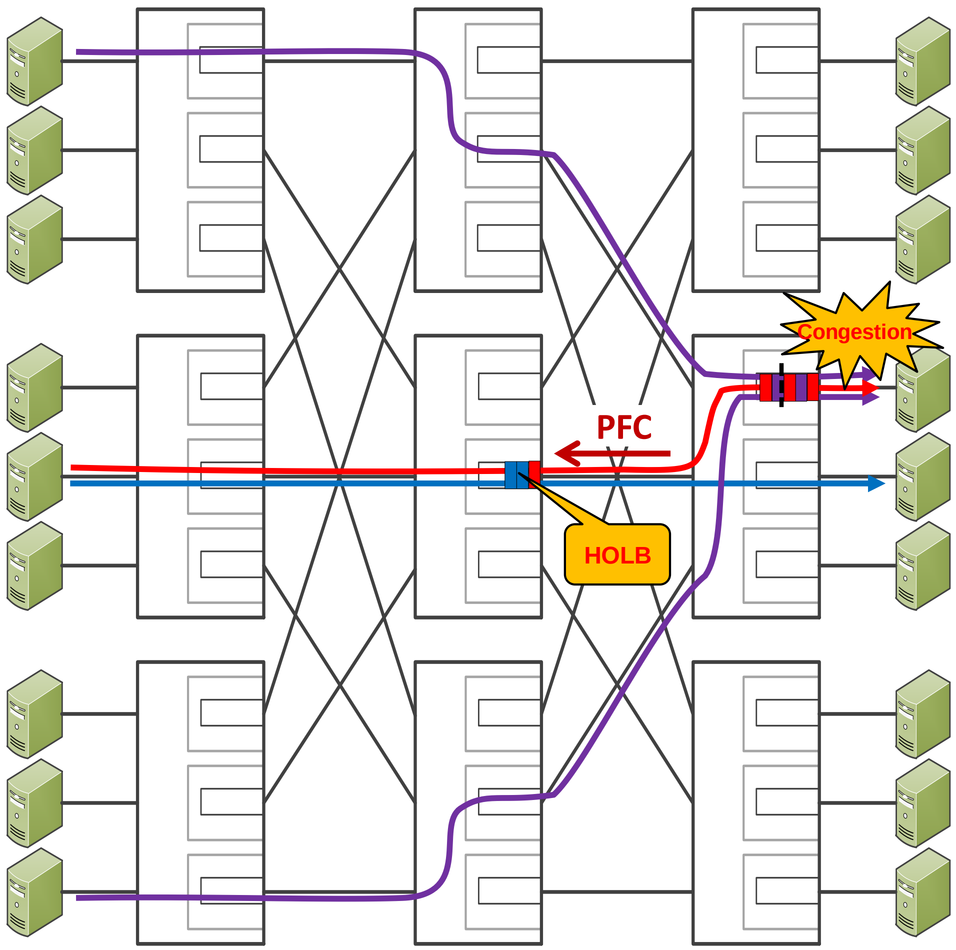
Congested Flow →

Victim Flow →

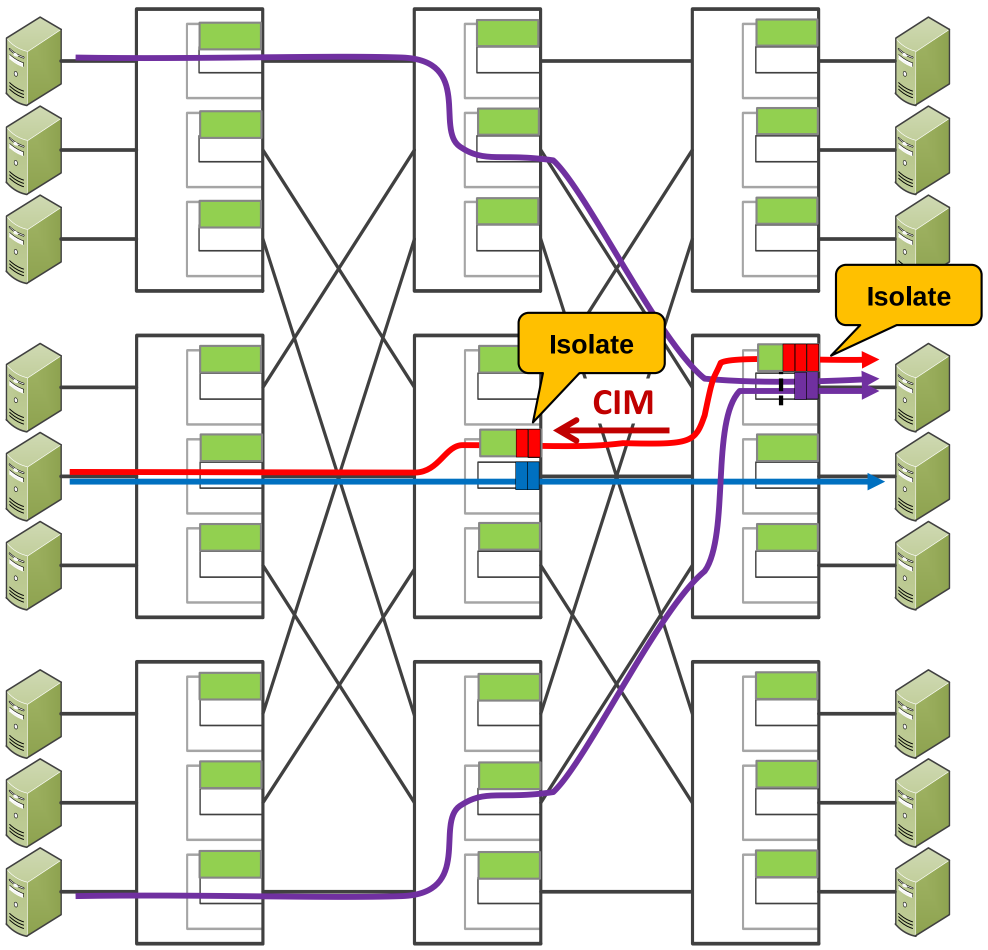
Congested Queue 🟩

Normal Queue □

Today – Without Congestion Isolation



Congestion Isolation

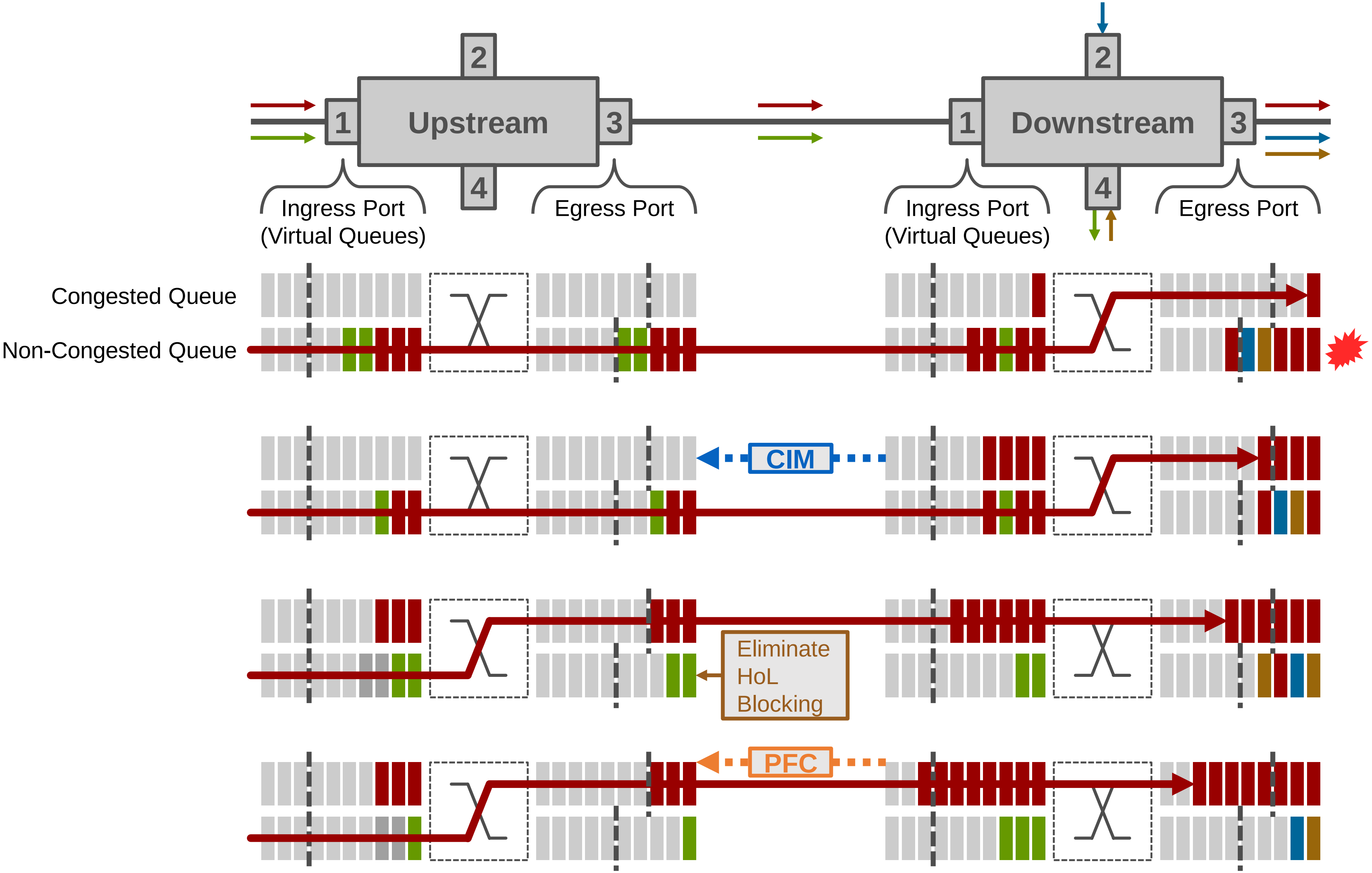


Summary

- Current data center design will be challenged to support the needs of large scale, low-latency, lossless or low-loss networks.
- P802.1Qcz: Congestion Isolation provides the following benefits:
  - Supports lossless and lossy networks to improve low-latency
  - Mitigates Head-of-Line blocking caused by PFC
  - Improves average flow completion times
  - Reduces or eliminates the need for PFC on non-congested flow queues
- Next Steps
  - Respond to comments on pre-circulated PAR and CSD
  - Motion to PAR to Nescom in July 2018

Backup

Congestion Isolation



1. Identify the flow causing congestion and isolate locally

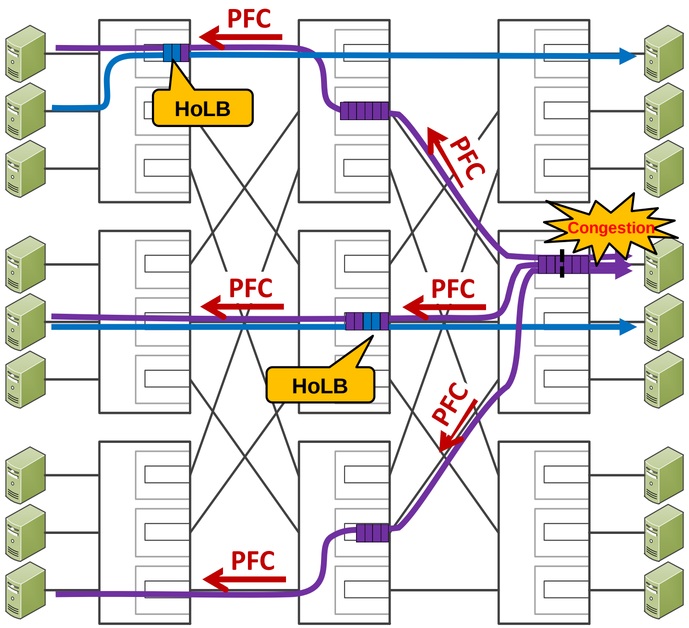
2. Signal to neighbor when congested queue fills

3. Upstream isolates the flow too, eliminating head-of-line blocking

4. Last Resort! If congested queue continues to fill, invoke PFC for lossless

Existing 802.1 Congestion Management Tools

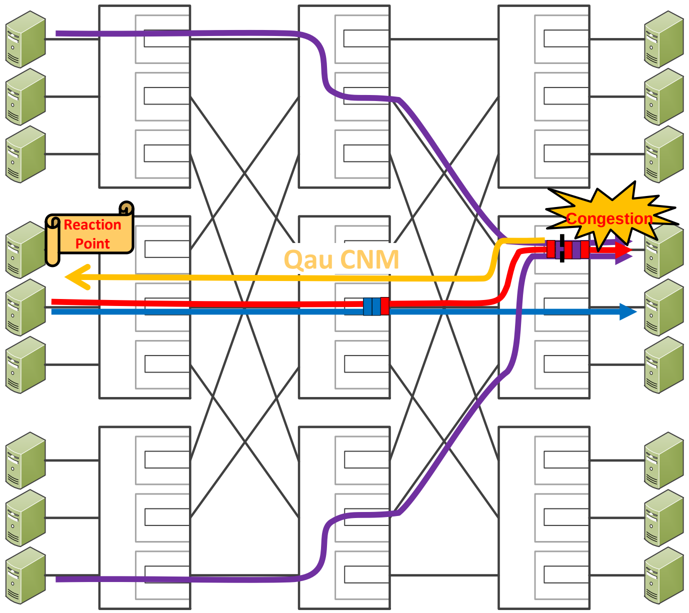
802.1Qbb - Priority-based Flow Control



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