IEEE P802.1Qcz
Congestion Isolation

Update for Pittsburgh Interim

May 21, 2018

Paul Congdon

paul.congdon@tallac.com

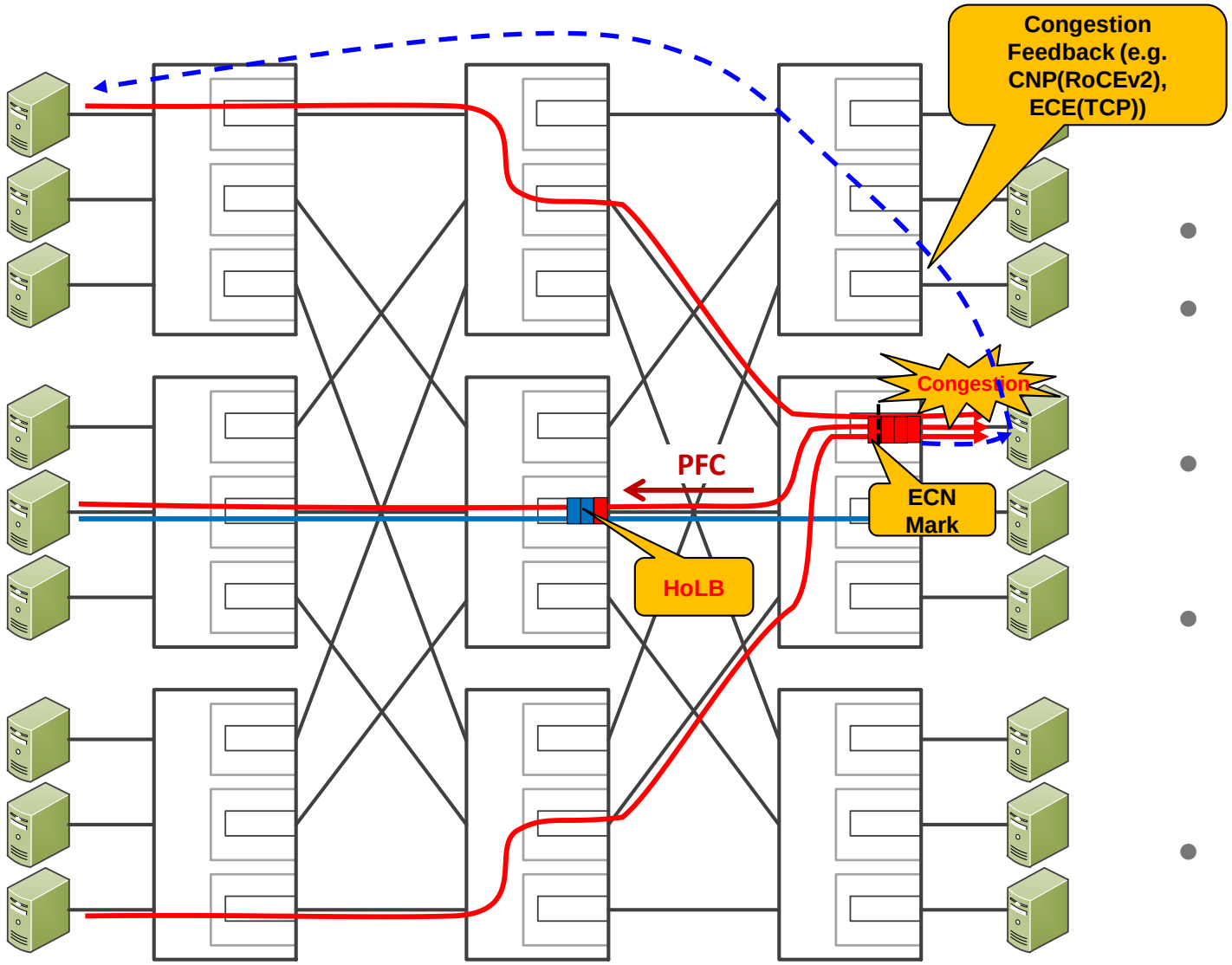
Status Update – P802.1Qcz

- PAR and CSD Status
  - The PAR and CSD were pre-circulated before March and all comments from other 802 WGs were resolved. The latest versions are available here:
    - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-congestion-isolation-PAR-0318-v1.pdf
    - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-congestion-isolation-CSD-0318-v01.pdf
  - March motion to forward PAR and CSD to Nescom narrowly failed – Many abstains
  - March motion to authorize May interim to pre-circulate PAR and CSD a second time passed
- Progress since March
  - Project introduced and discussed at London IETF-101 – tsvwg, iccrg, hotrfc
  - Technical detail review on TSN conference call – April 16th
  - Design team expansion and re-engagement
  - New simulation models from published papers
- Plans for May Interim
  - Additional technical review and TSN WG exposure. Objective: convert abstain votes to yes/no in July
  - Responses to Analysis presented in March
  - Additional simulation results
  - Additional detailed discussion around motivation, design team, scope of Q-changes

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



Congested Flow

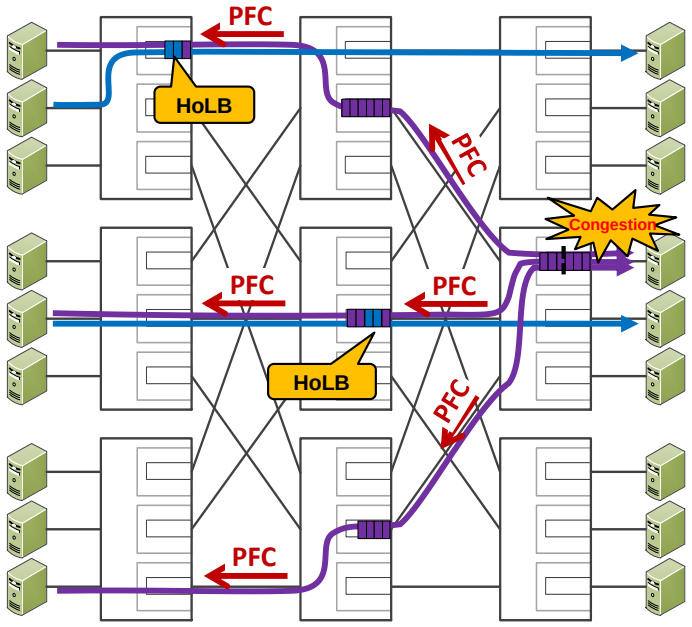
Victim Flow

ECN Control Loop

- 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

Existing 802.1 Congestion Management Tools

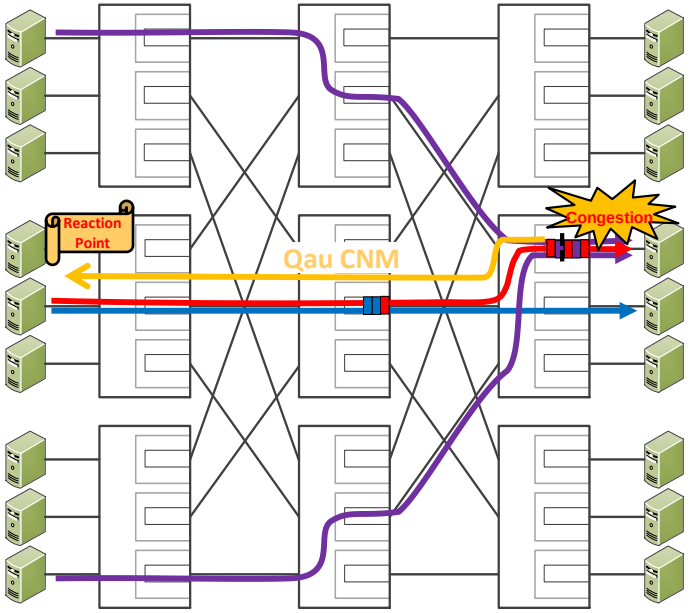
802.1Qbb - Priority-based Flow Control



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

Qcz simplifications over Qau

- No congestion domains to discover or defend against
- CI is hop-by-hop, so no issue within the PBB domain
- No new reaction points

P802.1Qcz – Congestion Isolation - Goals

- Work in conjunction with higher-layer end-to-end congestion control (ECN, etc)
- Support larger, faster data centers (Low-Latency, High-Throughput)
- Support lossless transfers
- Improve performance of TCP and UDP based flows
- Reduce pressure on switch buffer growth
- Reduce the frequency of relying on PFC for a lossless environment
- Eliminate or significantly reduce HOLB

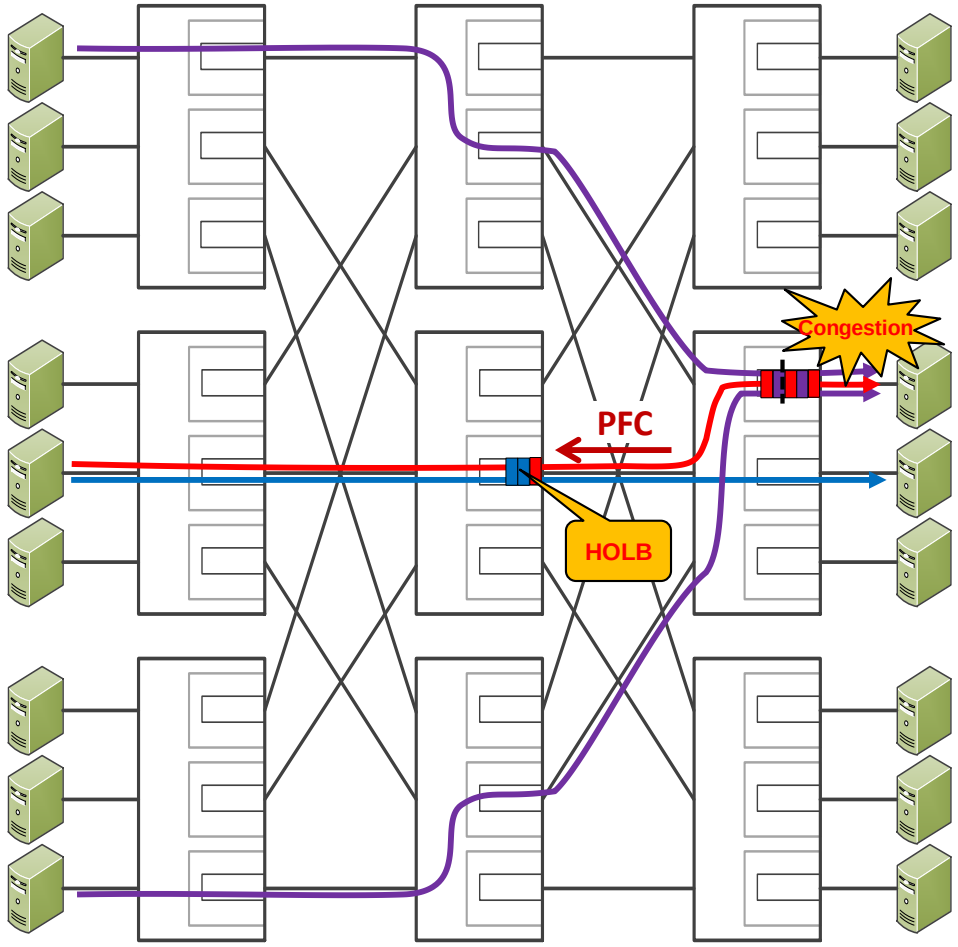
Important assertions about Qcz

- There are various degrees of conformity that can be specified and agreed upon
  - If lossless operation is NOT a requirement, CI works without enabling PFC
  - CI can perform local isolation only, without signaling
  - CI can coordinate isolation with upstream neighbors – best performance
- CI is designed to support higher layer end-to-end congestion control
  - CI is NOT an improvement on PFC
  - CI is NOT an improvement on QCN (Congestion Notification)
  - Congestion isolation reacts immediately and allows end-to-end congestion control to be less aggressive
  - Congestion isolation provides necessary time for the end-to-end congestion control loop.
- To create a fully lossless network, PFC is needed as a last resort
  - CI has been shown to reduce both the number of pause frames and duration of pause

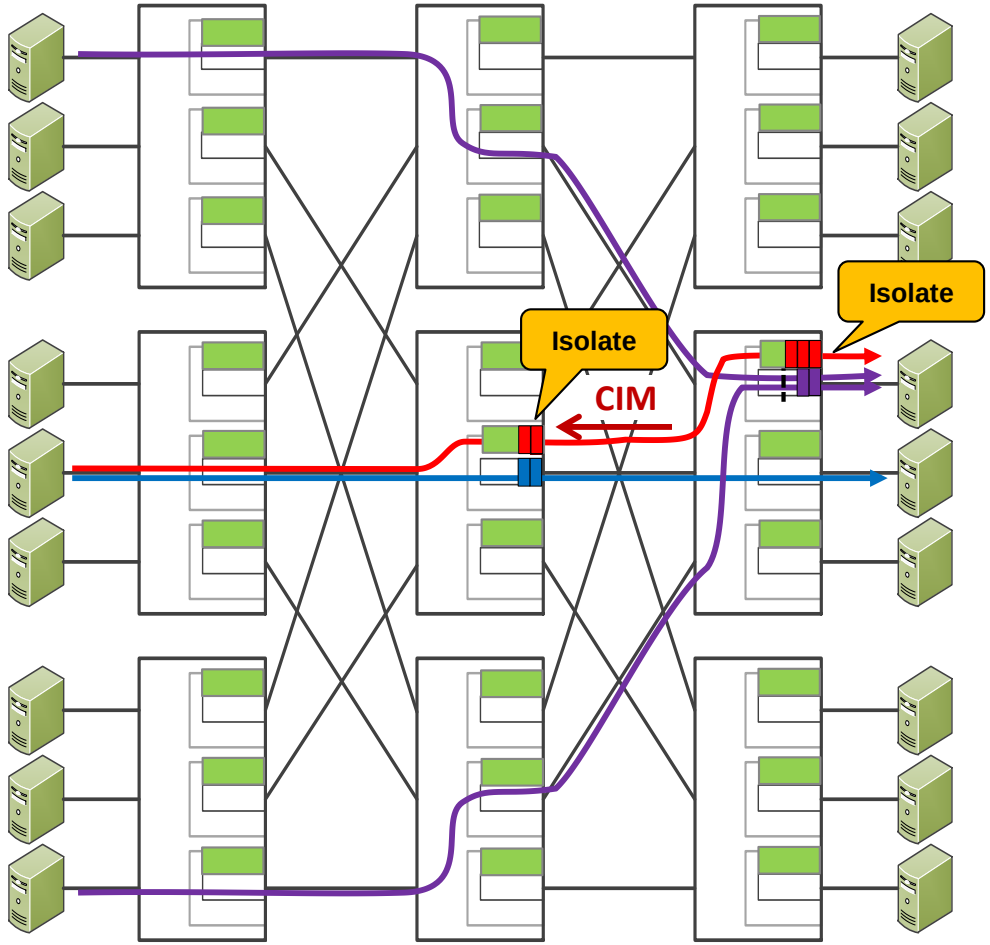
Isolate the congestion to mitigate HOLB

Congested Flow
Congested Queue
Victim Flow
Normal Queue

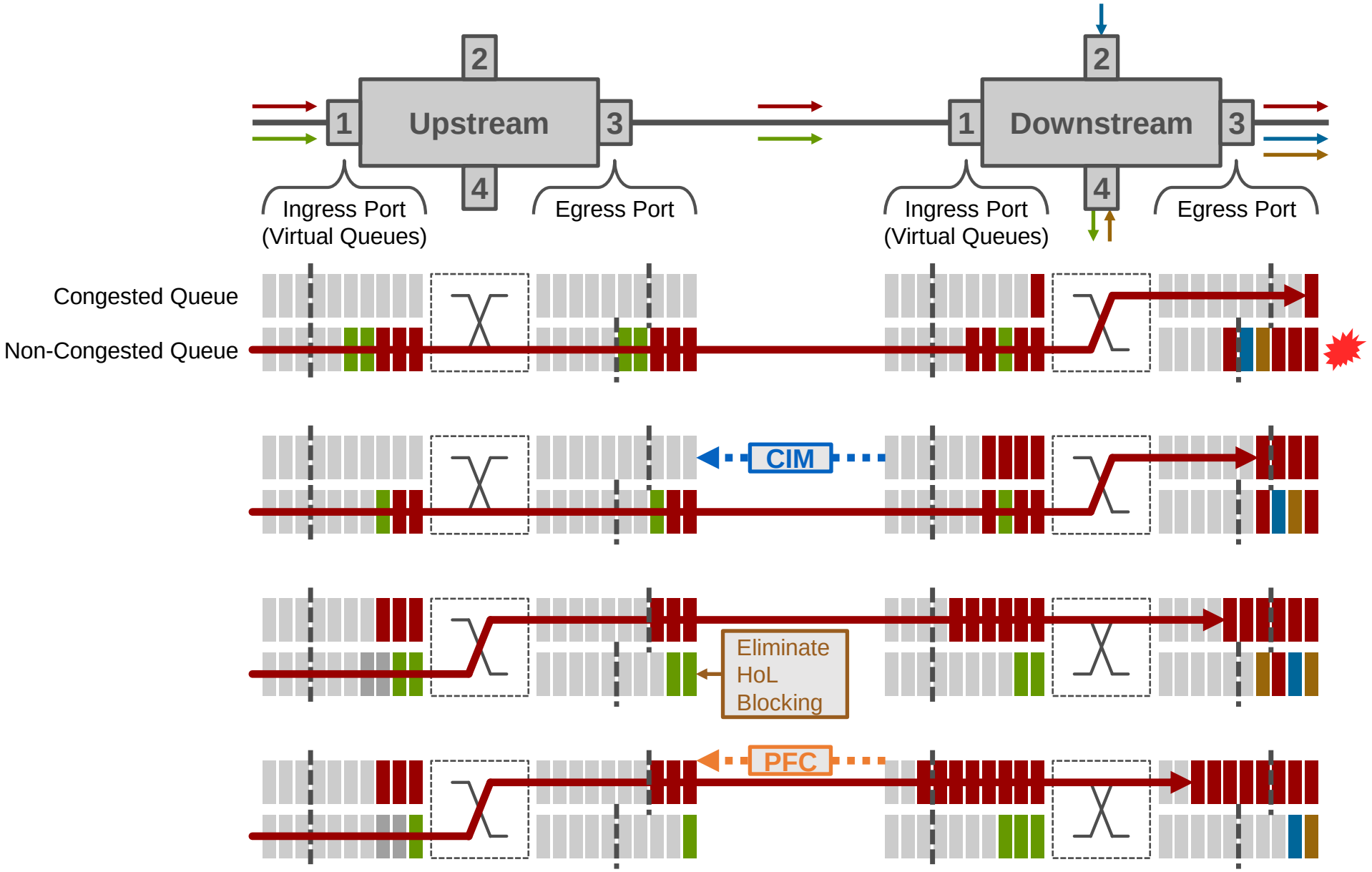
Today – Without Congestion Isolation



Congestion Isolation



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

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
  - Continued Technical review within 802.1 Working Group during May Interim
  - Pre-circulation of PAR and CSD before July 2018
  - Motion to start standardization in July 2018

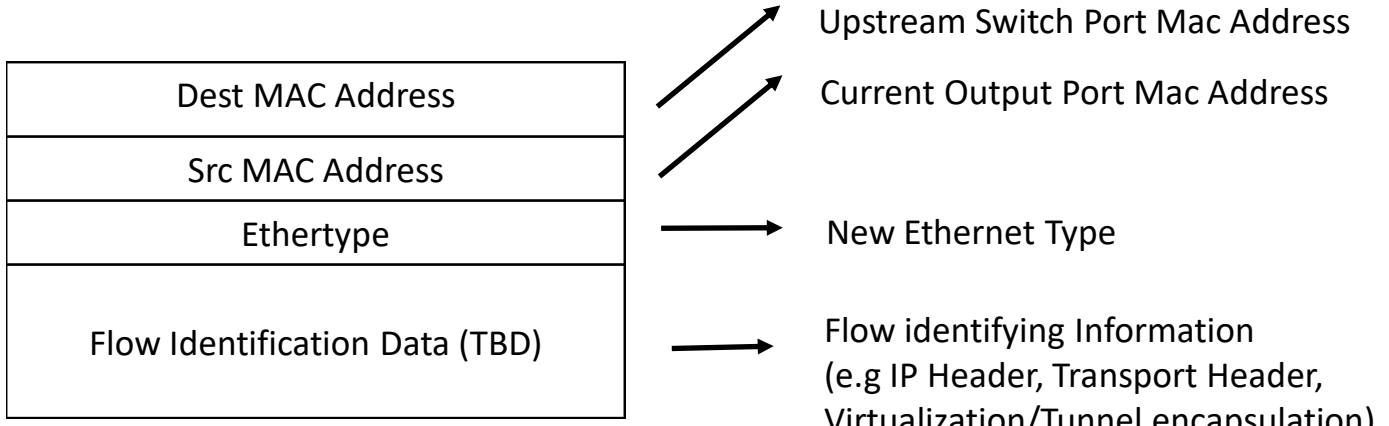
Backup

Congestion Isolation Message

- Objectives/Requirements:
  - Provide upstream neighbor with an indication that a flow has been isolated
  - Provide upstream neighbor with flow identification information
  - No adverse effects of single packet loss
  - Low overhead

- **NOTE:** Consider re-using 802.1Qau CNM format, but use upstream switch as DA MAC?

Format of Congestion Isolation Packet



Capability Discover via LLDP

- Objectives/Requirements:
  - Peer bridges must know that each is capability of Congestion Isolation
  - Bridges should agree on the traffic class used for the Congested Flow Queue
  - Bridges should agree on the traffic classes that will monitored for congestion
  - It may be helpful to inform the upstream switch of the inactivity timeout - if used.

Format of LLDP TLV

TLV Type	TLV info length	802.1 OUI	subtype	Congested Queue	Monitored Queues	Inactivity Timeout

Proposed Reference Diagram – work in progress