IEEE P802.1Qcz
Proposed Project for Congestion Isolation

IETF 101 – London

ICCRG

Paul Congdon

paul.congdon@tallac.com

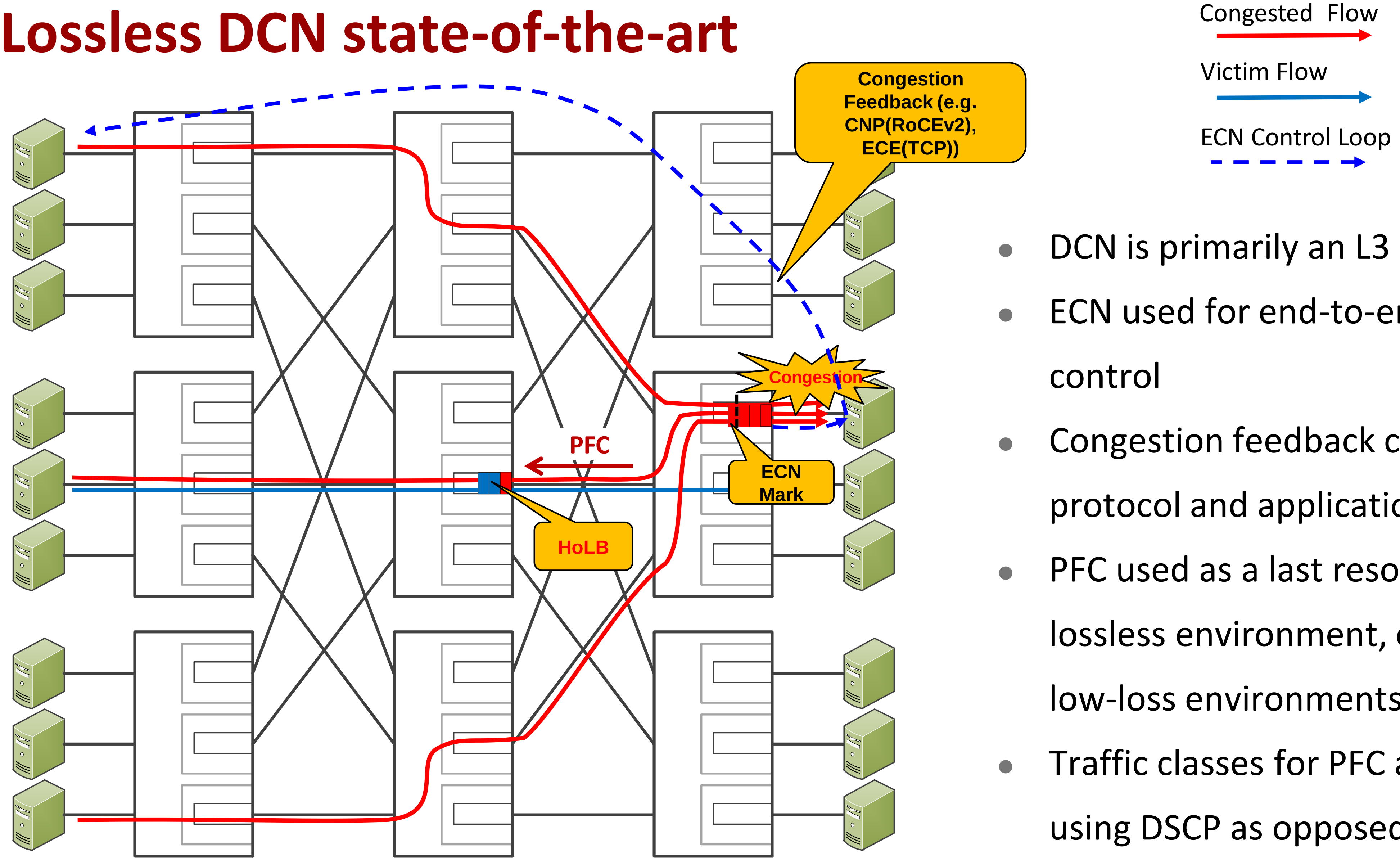
Project Background – P802.1Qcz

- Project Initiation
  - November 2017 - Agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “Congestion Isolation”
  - Motivation discussed in draft report of “802 Network Enhancements For the Next Decade”
    - https://mentor.ieee.org/802.1/dcn/18/1-18-0007-02-ICne-draft-report-lossless-data-center-networks.pdf
- Project Status
  - March 2018 - Approval pending further review, wider exposure and additional simulation analysis.
  - July 2018 – Expected project creation date
- So what is Congestion Isolation?

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 and distributed storage.
  - Bridges will:
    - individually identify flows creating congestion
    - adjust transmission selection (aka egress packet scheduling) for those flows
    - signal congested flow information to the upstream peer.
  - Reduce head-of-line blocking for uncongested flows sharing a traffic class in lossless networks.
  - Intended to be used with higher layer protocols that utilize end-to-end congestion control.

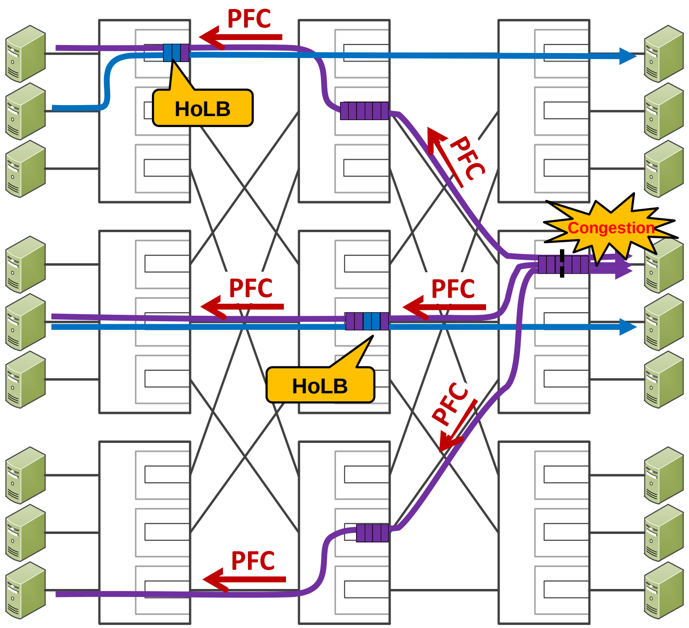
Lossless DCN state-of-the-art



- DCN is primarily an L3 CLOS network
- ECN 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

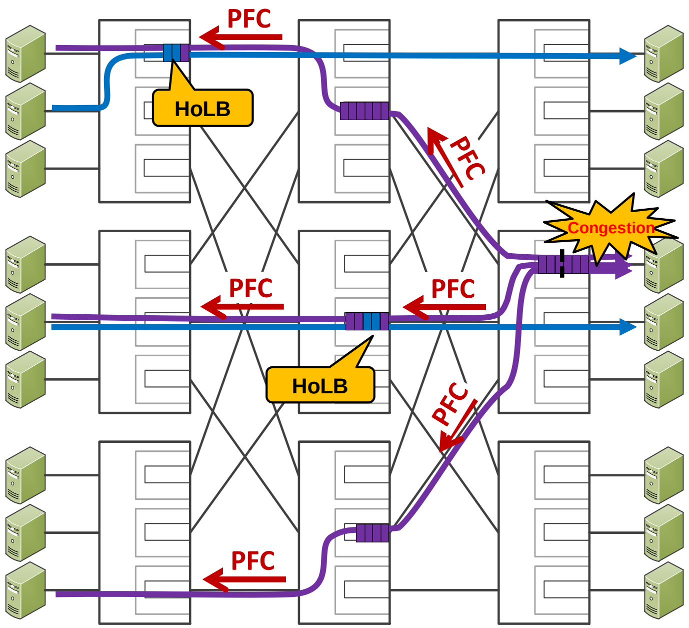


Concerns with over-use

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

Existing 802.1 Congestion Management Tools

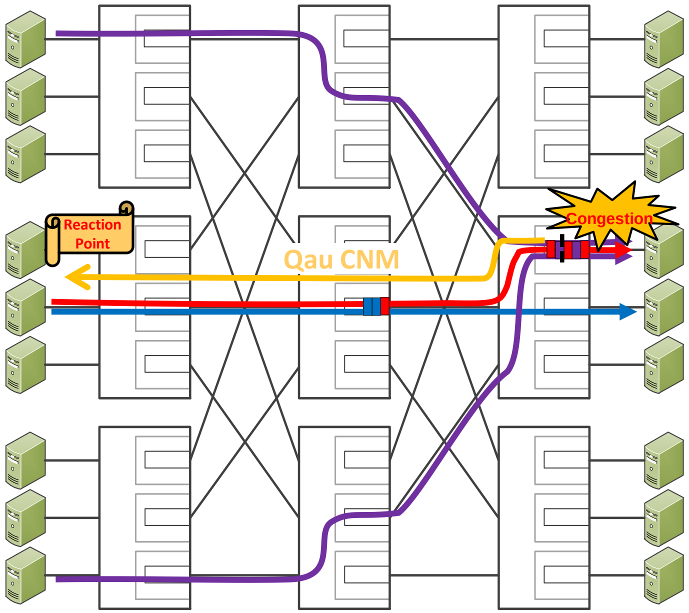
802.1Qbb - Priority-based Flow Control



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



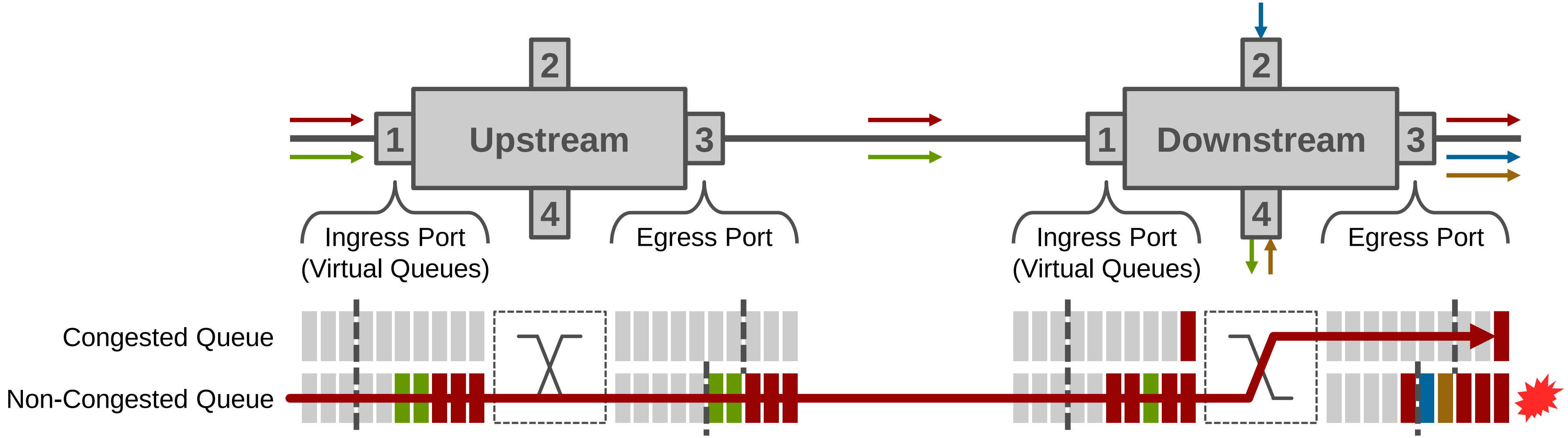
Concerns with deployment

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

P802.1Qcz – Congestion Isolation - Goals

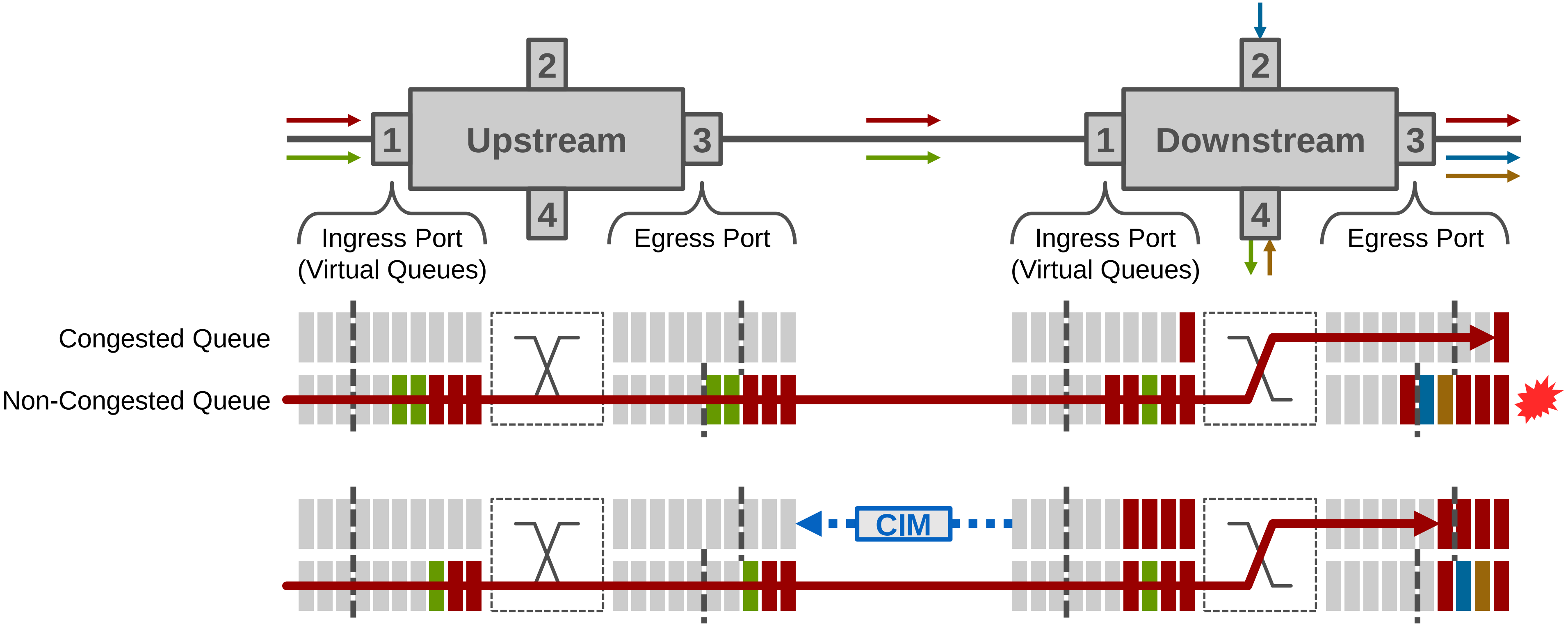
- Work in conjunction with higher-layer end-to-end congestion control (ECN, etc)
- Support larger, faster Ethernet based ***data centers*** (Low-Latency, High-Throughput)
- Support lossless transfers
- Improve performance of TCP and UDP based flows
- Reduce pressure on switch buffer growth
- Reduce the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

Congestion Isolation



1. Identify the flow causing congestion and isolate locally

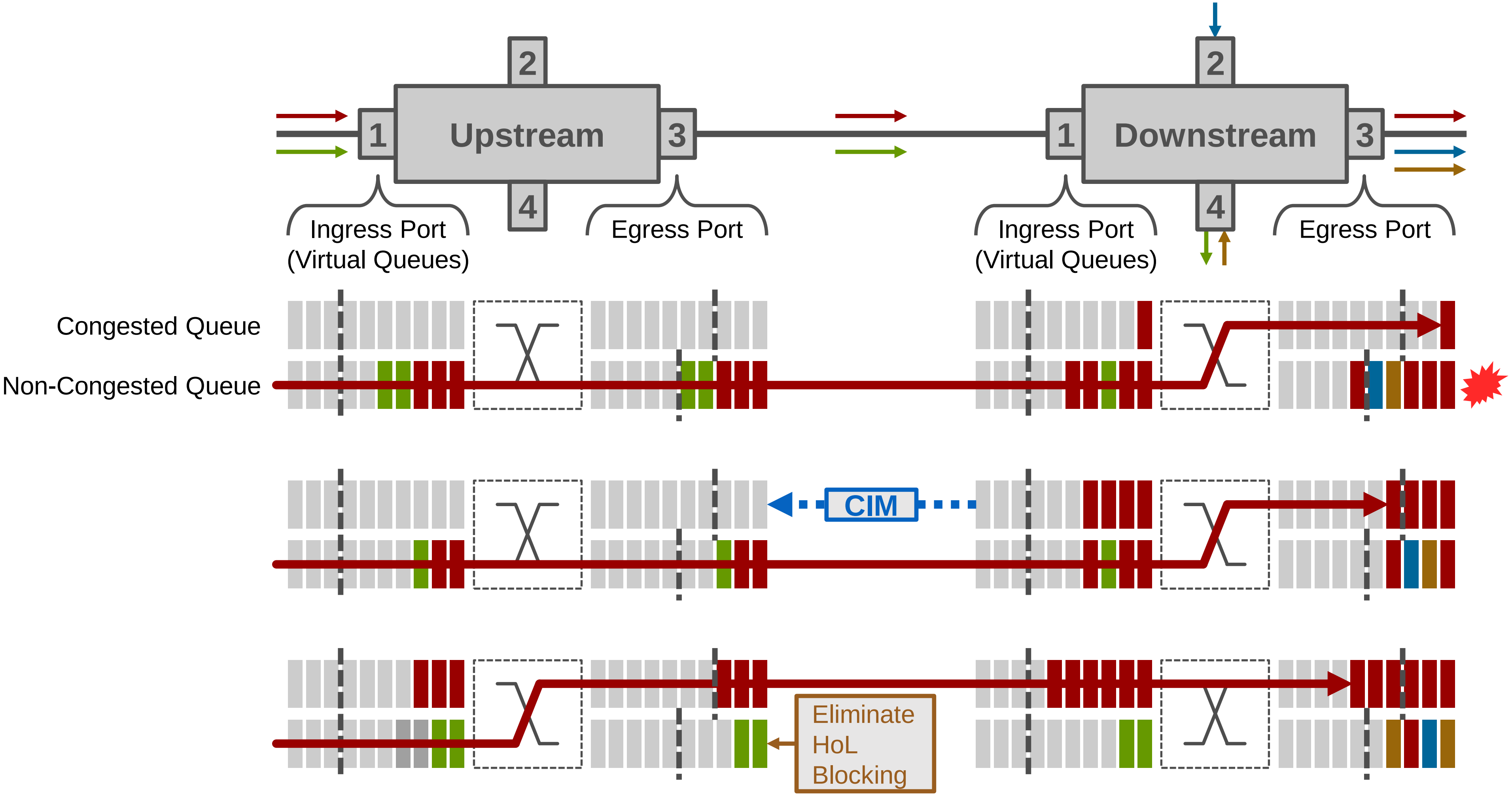
Congestion Isolation



1. Identify the flow causing congestion and isolate locally

2. Signal to neighbor when congested queue fills

Congestion Isolation

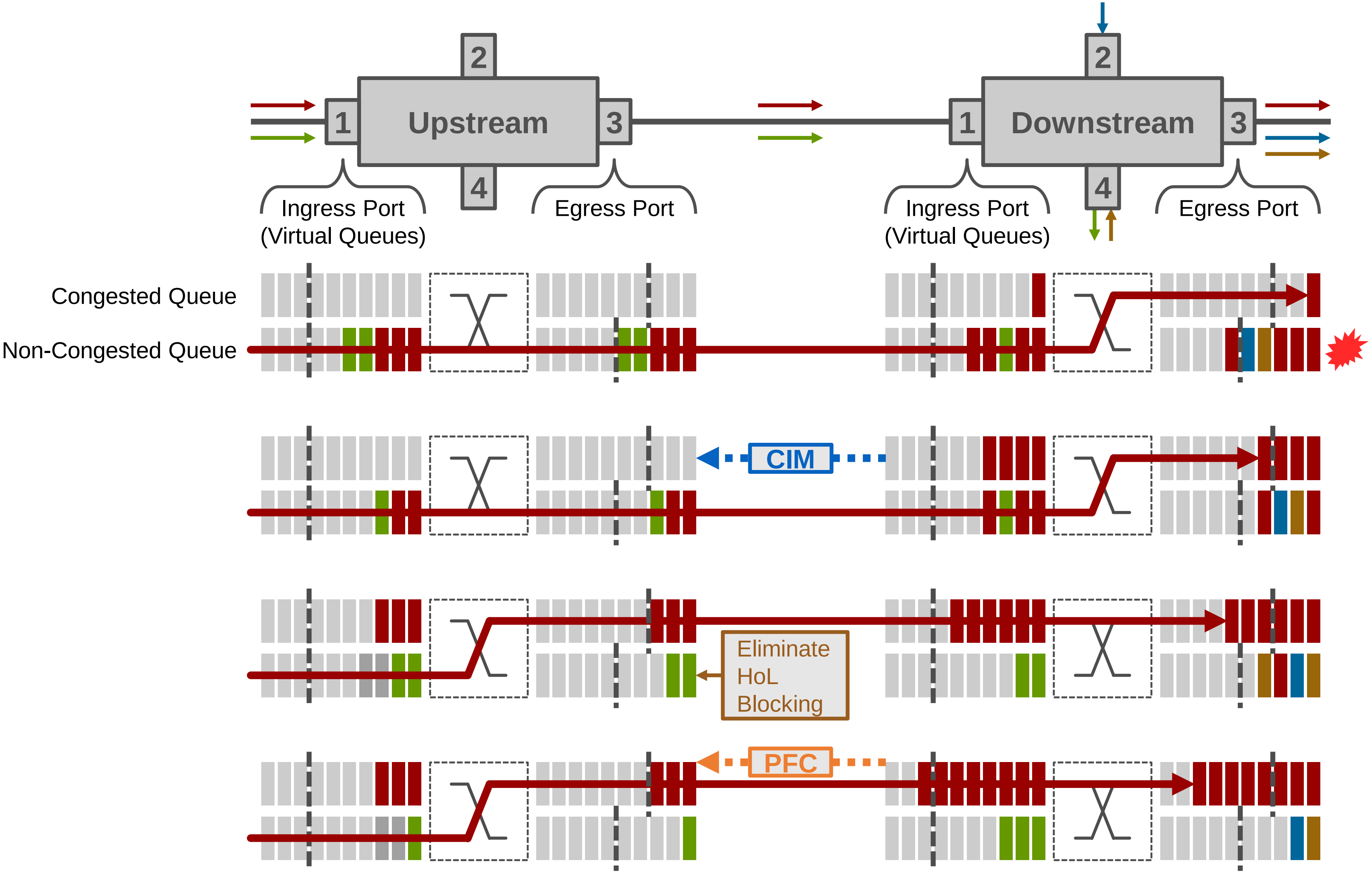


1. Identify the flow causing congestion and isolate locally

2. Signal to neighbor when congested queue fills

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

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

Early Simulation Results

- Environment
  - 2 Tier CLOS: 1152 servers, 72 switches, 100GbE interface, 200ns of link latency
  - RoCEv2 data-mining workload with persistent incast and mixed many-to-many flows
- Preliminary Results
  - Lossless environment with PFC – Reduction in Flow Completion Times
    - 63% (Mice), 23% (Elephants), 38% (Average)
  - Lossless with PFC – Reduction in Pause Frame Counts
    - 84% (switch model dependent)
  - Lossy environment without PFC – Reduction in Overall Packet Loss Rate
    - 66%
- Details available at:
  - http://www.ieee802.org/1/files/public/docs2017/new-dcb-shen-congestion-isolation-simulation-1117-v00.pdf
  - http://www.ieee802.org/1/files/public/docs2018/new-dcb-shen-congestion-isolation-simulation-0118-v01.pdf
  - http://www.ieee802.org/1/files/public/docs2018/cz-shen-congestion-isolation-simulation-0318-v01.pdf

Next Steps

- Continued Technical review with 802.1 Working Group and others (IETF?)
- Additional simulation analysis desired
  - Alternative switch memory architectures
  - Interaction with other CC algorithms (e.g. BBR, other rate or time-based schemes)
- Motion to start standardization in July 2018

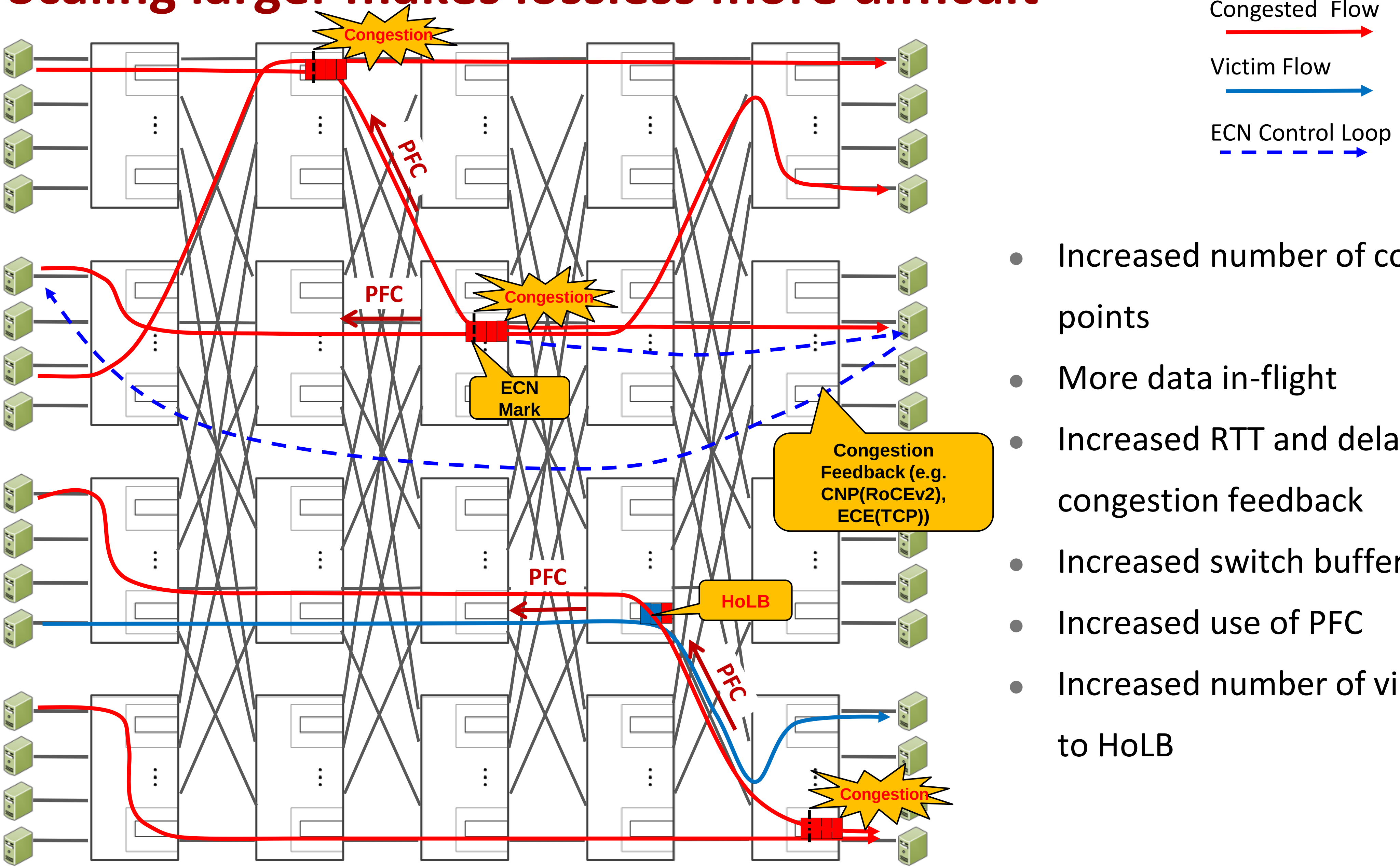
- How can IETF help/participate?
  - Discuss within existing IEEE 802 / IETF interworking relationship (https://www.ietf.org/blog/working-ieee-802/)
  - Provide review comments and feedback to me – paul.congdon@tallac.com
  - Participate and/or review 802 Industry Connections draft report on Next Generation Data Centers (https://1.ieee802.org/802-nend/)

Backup

Important assertions about CI

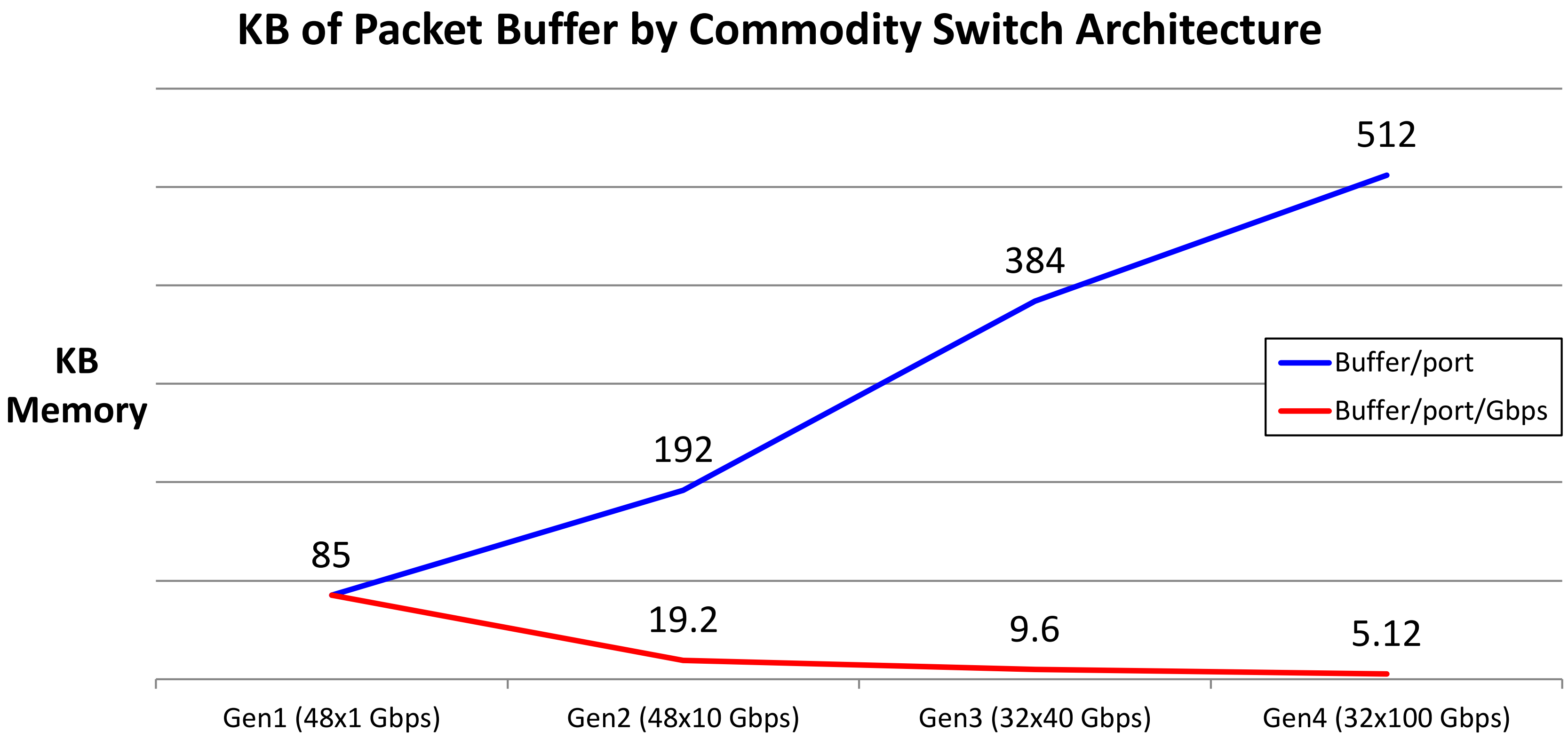
- There are various degrees of conformity that can be specified and agreed upon
  - If lossless operation is NOT a requirement, CI works without enabling PFC
  - CI can perform local isolation only, without signaling
  - CI can coordinate isolation with upstream neighbors – best performance
- CI is designed 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 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

Scaling larger makes lossless more difficult



- Increased number of congestion points
- More data in-flight
- Increased RTT and delay for congestion feedback
- Increased switch buffer requirements
- Increased use of PFC
- Increased number of victim flows due to HoLB

Switch buffer growth is not keeping up



Commodity Shallow Buffer Switches in DCNs are desirable:

- Low Latency
- Low Cost

However, packet loss can create performance issues:

- Source: Broadcom, “White Paper: Buffer Requirements for Datacenter Network Switches”, DNFAMILY-WP1101, August 25, 2015

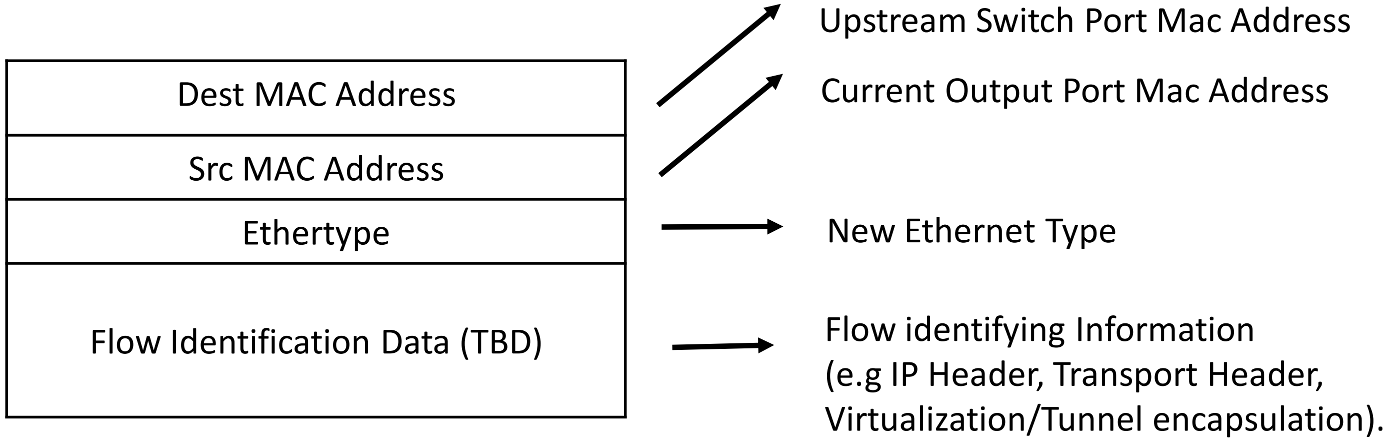
Source: “Congestion Control for High-speed Extremely Shallow-buffered Datacenter Networks”. In Proceedings of APNet’17, Hong Kong, China, August 03-04, 2017, https://doi.org/10.1145/3106989.3107003

Congestion Isolation Packet

- 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



Dest MAC Address → Upstream Switch Port Mac Address

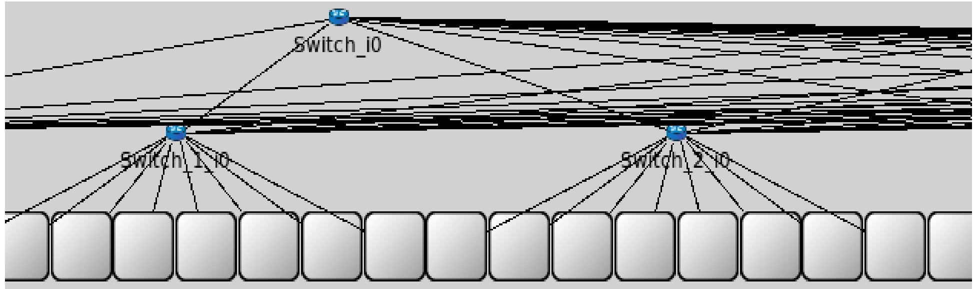
Src MAC Address → Current Output Port Mac Address

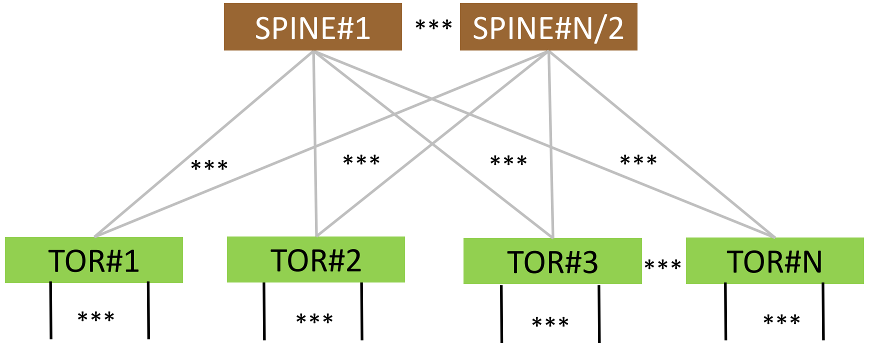
Ethertype → New Ethernet Type

Flow Identification Data (TBD) → Flow identifying Information (e.g IP Header, Transport Header, Virtualization/Tunnel encapsulation).

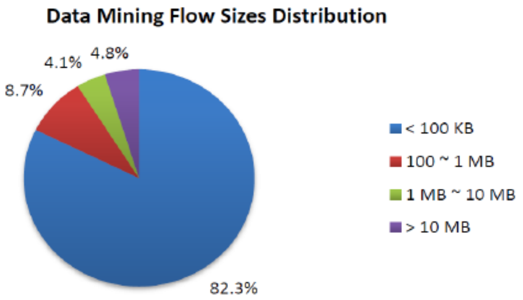
Simulation Highlights

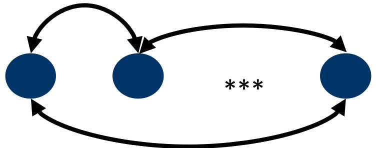
- Complete presentations on simulations are available on 802.1 public repository:
  - http://www.ieee802.org/1/files/public/docs2017/new-dcb-shen-congestion-isolation-simulation-1117-v00.pdf
  - http://www.ieee802.org/1/files/public/docs2018/new-dcb-shen-congestion-isolation-simulation-0118-v01.pdf
  - http://www.ieee802.org/1/files/public/docs2018/cz-shen-congestion-isolation-simulation-0318-v01.pdf
- Set-up – OMNET++



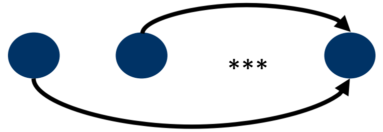


- 2 Tier CLOS: 1152 servers, 72 switches, 100GbE interface, 200ns of link latency (about 40 meters)
- Traffic Patterns:
  - Model data mining application with flow size distributions
  - 50 clusters of 21 servers for many to many traffic
  - 4 sets of 20:1 permanent many to one incast traffic





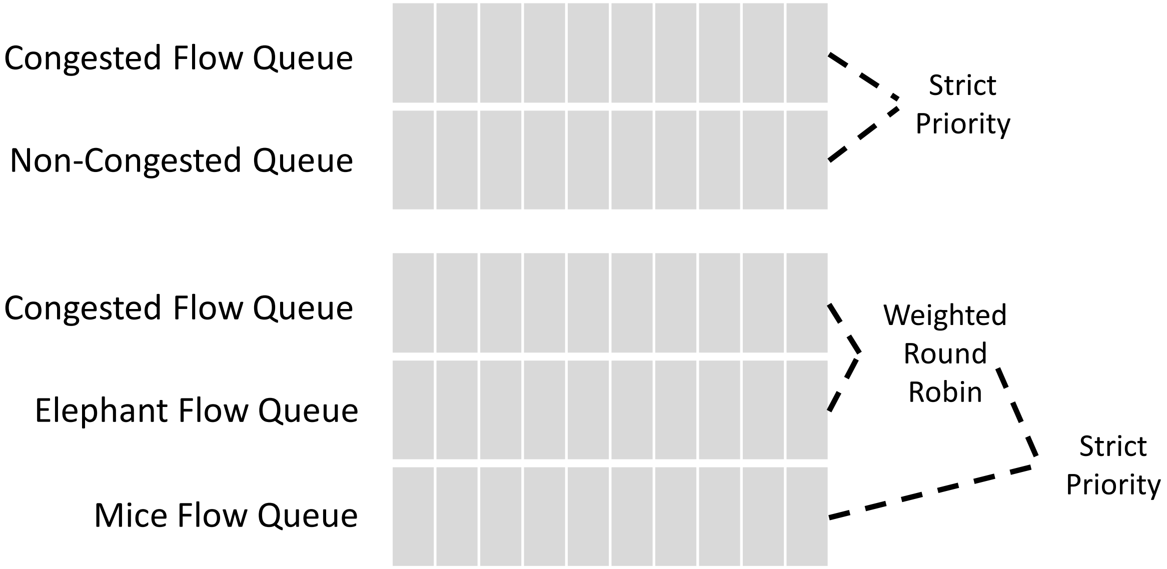
Many to many traffic



Many to one incast traffic

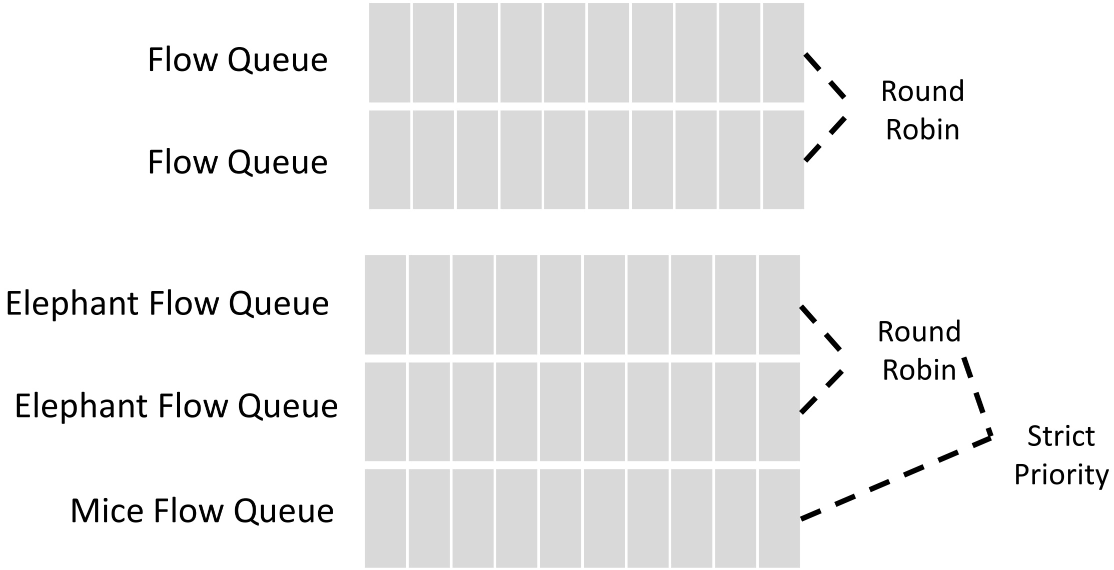
Queue Models Used

With Congestion Isolation (ECN + PFC + CI)



- Congested flows are dynamically isolated based on congestion.
- ECN is marked once a packet is isolated.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB;
  - CI: Low 10 KB, High 300 KB, Max Probability 1%.

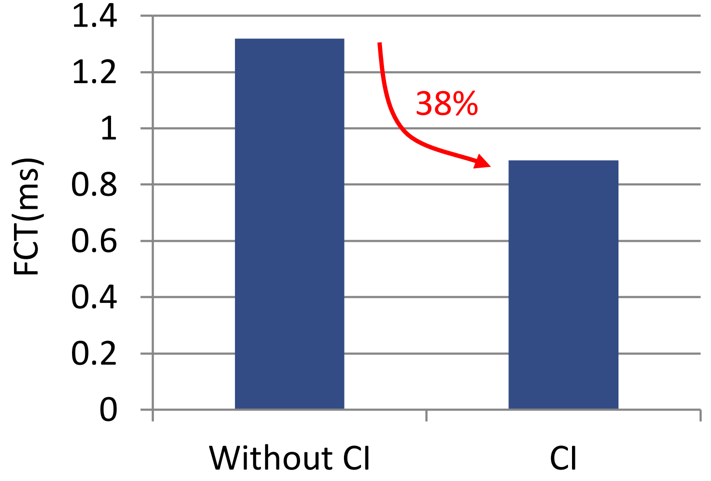
Without Congestion Isolation (ECN + PFC)



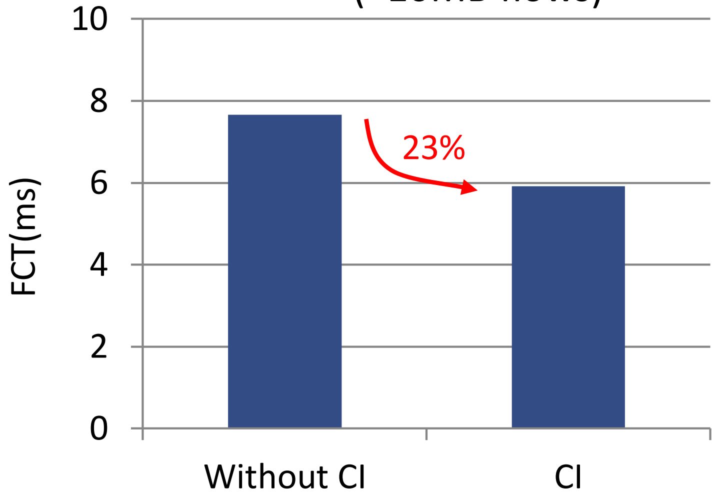
- Flows are mapped to one of the same queues by hash of destination IP.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB;
  - ECN: Low 10 KB, High 300 KB, Max Probability 1%.

FCT Comparison – Lossless Scenario (with PFC)

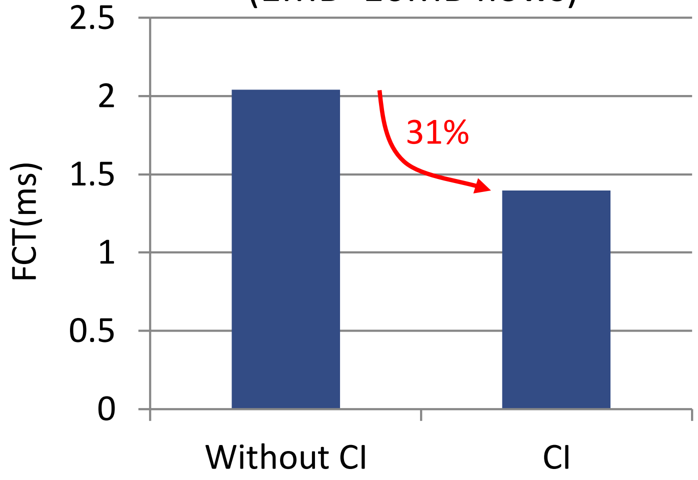
Average flow completion time (all flows)



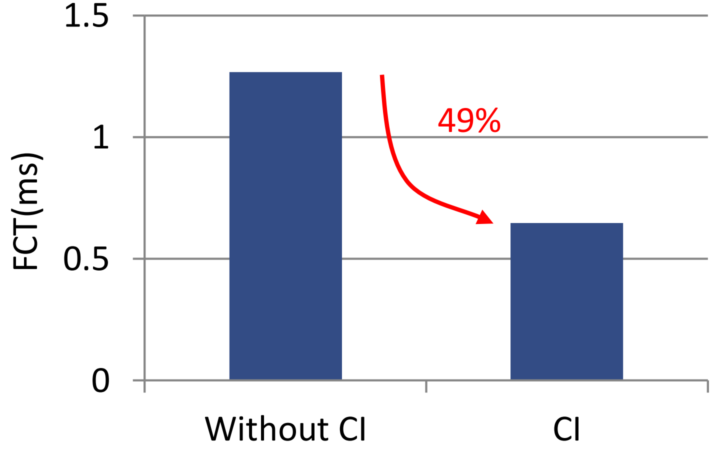
Average flow completion time (>10MB flows)



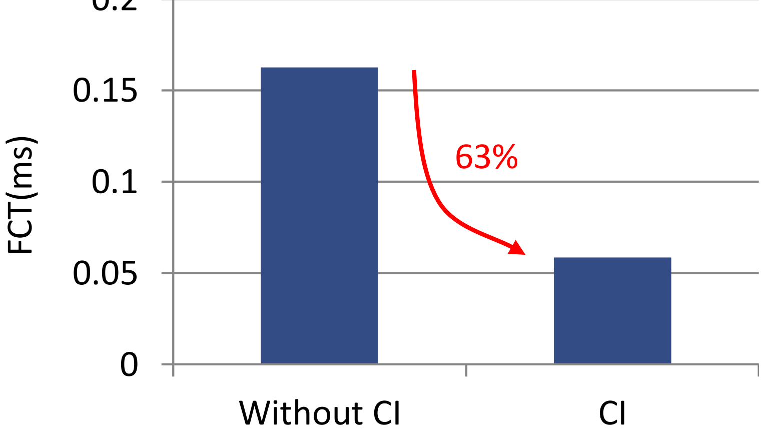
Average flow completion time (1MB~10MB flows)



Average flow completion time (100KB~1MB flows)

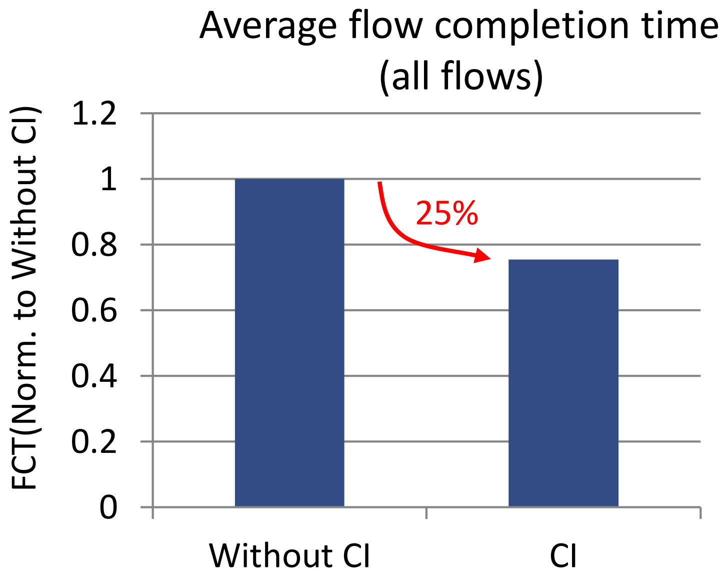


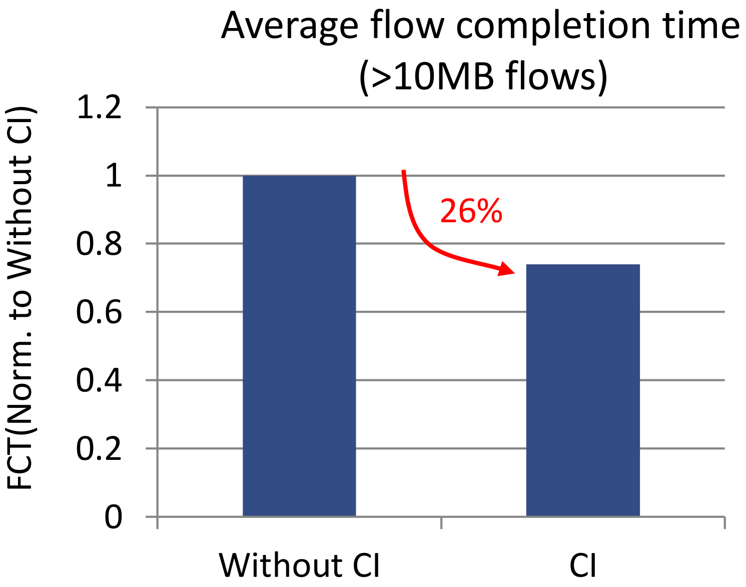
Average flow completion time (<100KB flows)

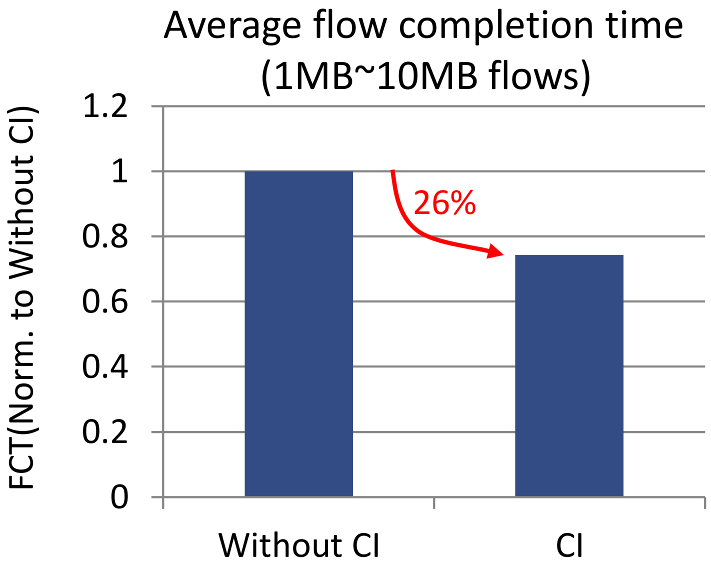


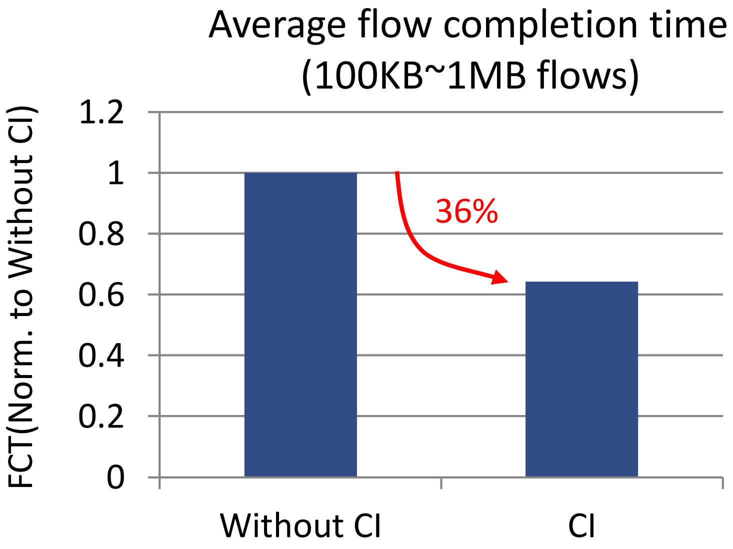
- The mice benefit the most.

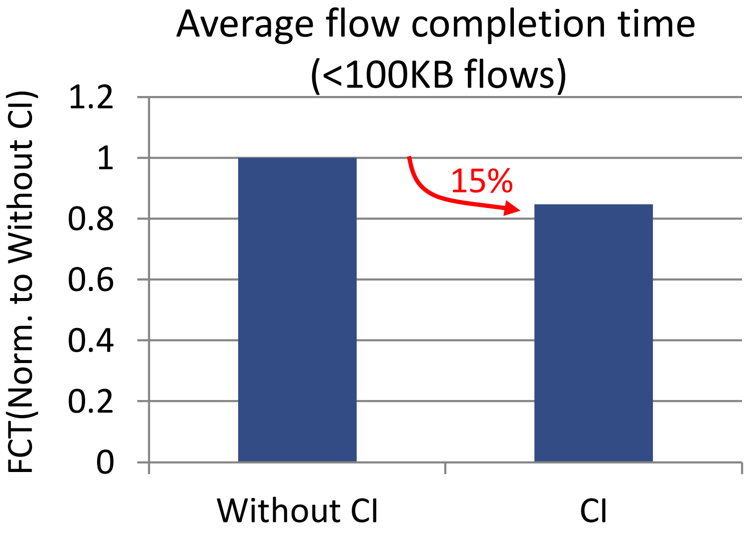
FTC With Mice/Elephant separation (3 Queue Model)


Average flow completion time (all flows)


Average flow completion time (>10MB flows)

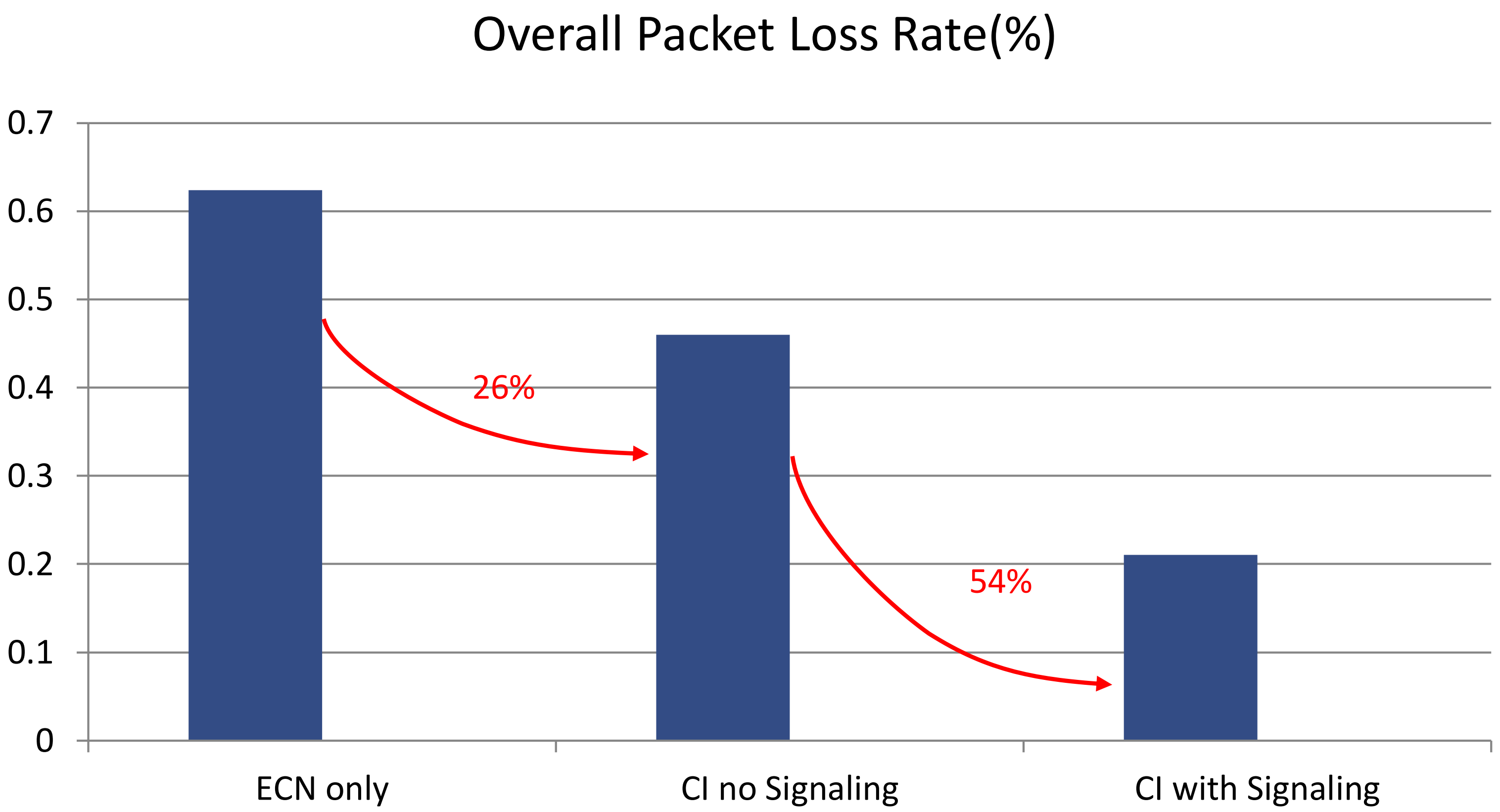

Average flow completion time (1MB~10MB flows)


Average flow completion time (100KB~1MB flows)


Average flow completion time (<100KB flows)

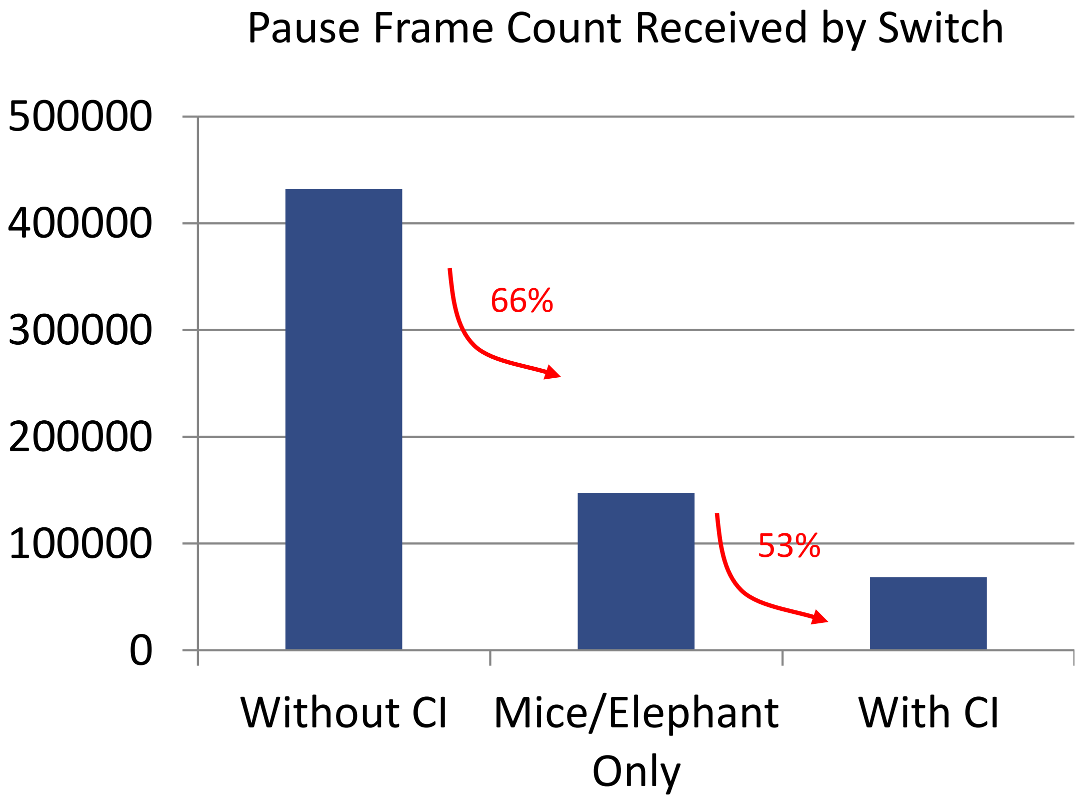
- With 3 queue model both “without CI” and “CI” have mice prioritization mechanism.
- The performance of the mice is not improved as much.

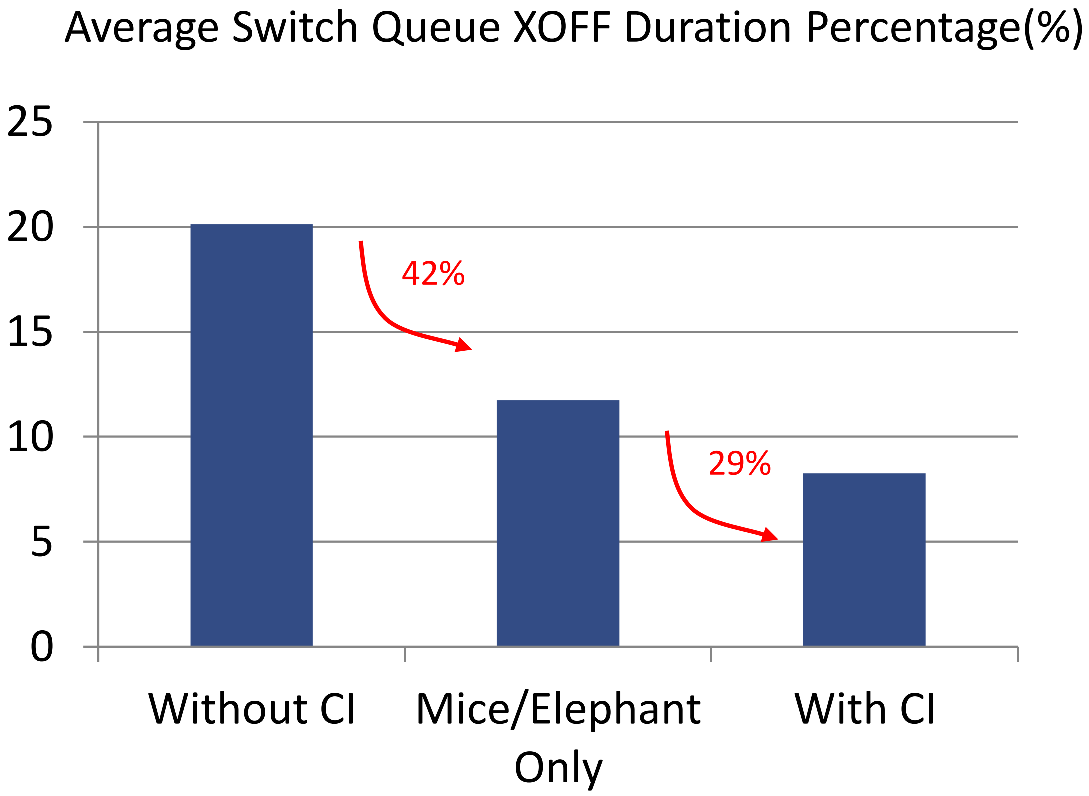
Lossy Scenario (No PFC)



- CI reduces packet loss rate, which means it also reduces packet retransmission and improves performance.

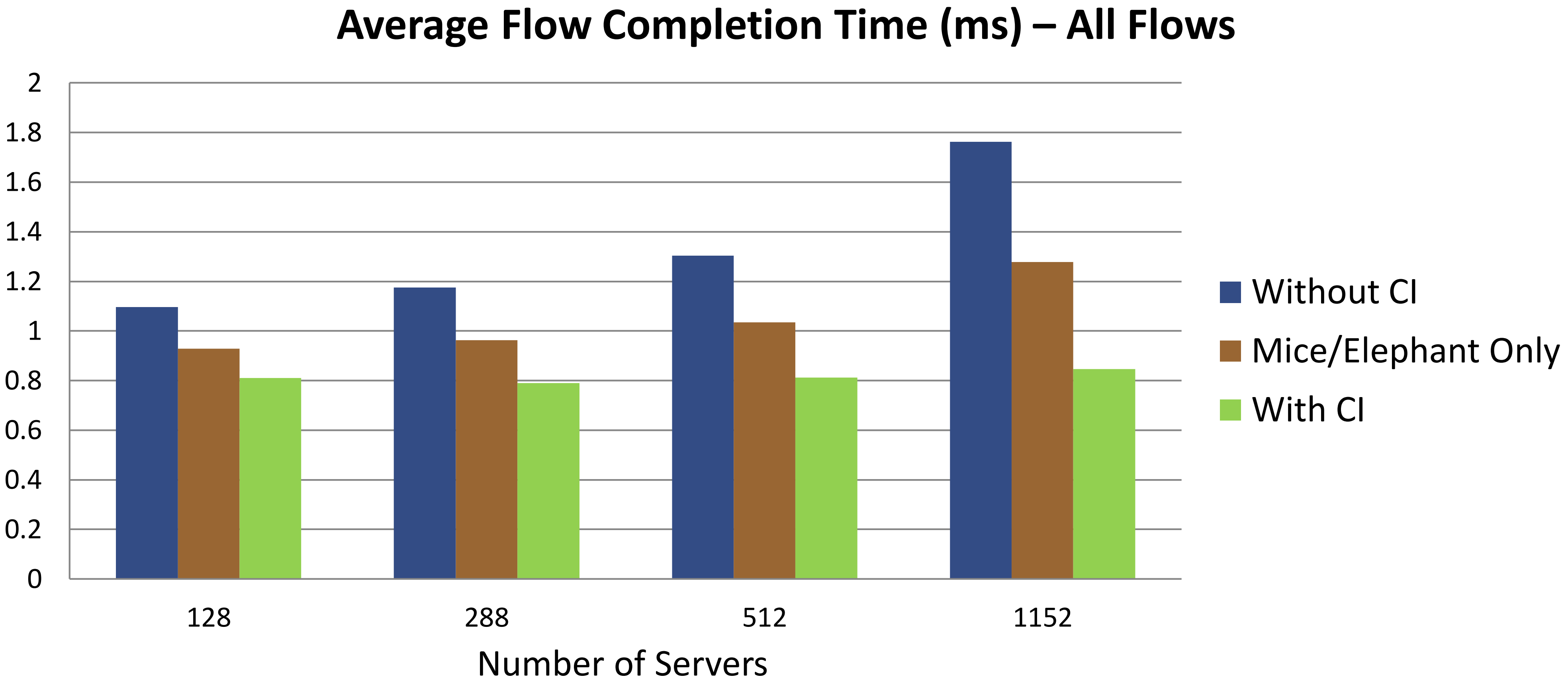
Lossless Scenario - Reducing the Impact of PFC





- CI reduces Pause frame count and XOFF duration.
- XOFF duration is less significant than Pause frame count, because usually pause for low priority queue takes longer time to resume than high priority queue.

Scaling Comparison



- Adding CI allows the data center size to scale.

Summary

- Current data center design will be challenged to support the needs of large scale, low-latency, lossless networks.
- Congestion Isolation provides the following benefits:
  - Supports lossless as well as 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 and feedback
  - Motion to approve project in July 2018