P802.1Qcz
Congestion Isolation

IEEE 802 / IETF

Workshop on Data Center Networking

Bangkok

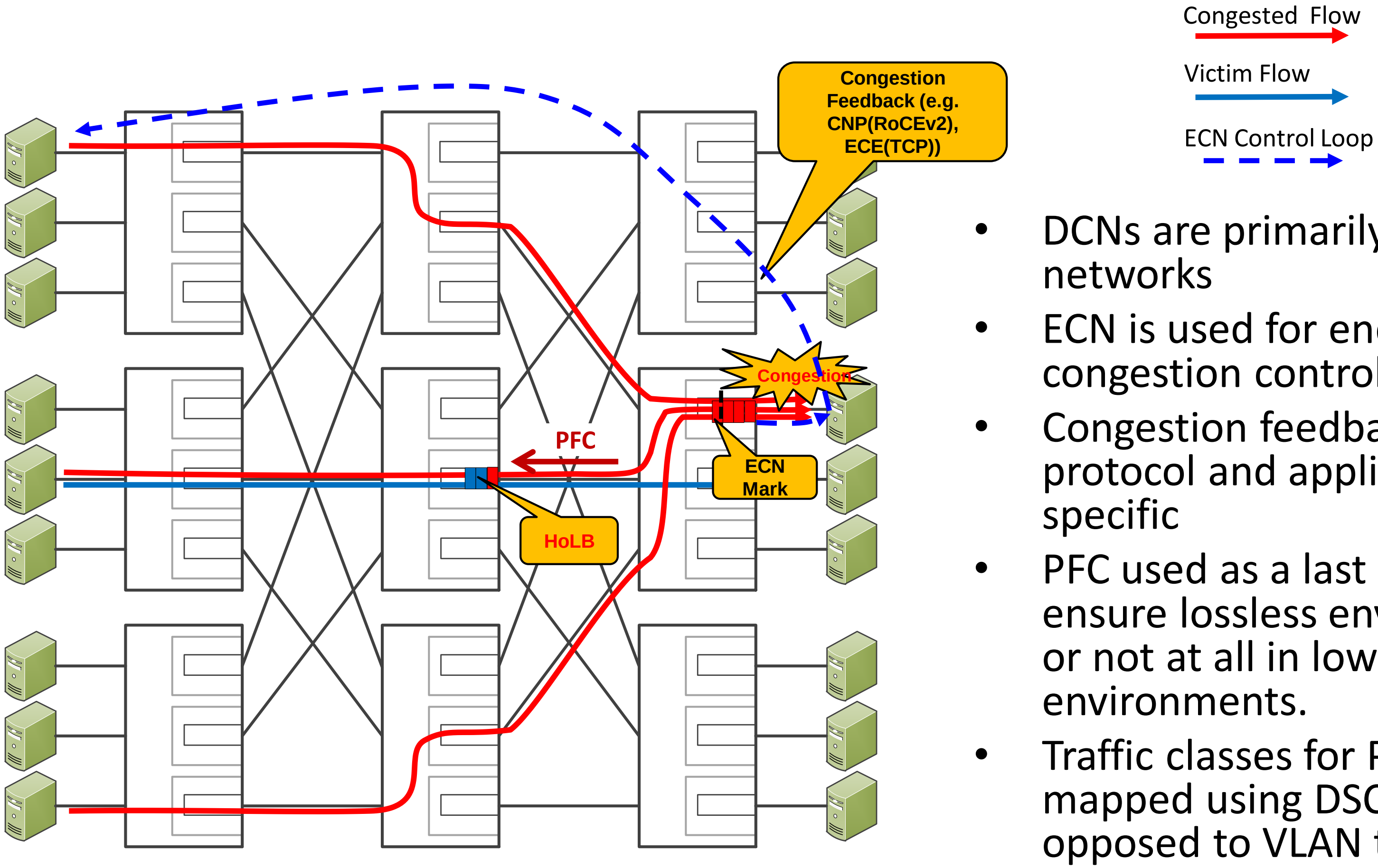
November 2018

Paul Congdon (Huawei/Tallac)

The Case for Low-latency, Lossless, Large-Scale DCNs

- More and more latency-sensitive applications are being deployed in data centers
  - Distributed Storage
  - AI / Deep Learning
  - Cloud HPC
  - High-Frequency Trading
- RDMA is operating at larger scales thanks to RoCEv2
  - Chuanxiong Guo, et. al., Microsoft, ”RDMA over Commodity Ethernet at Scale”, SIGCOMM 2016
  - Y Zhu, H Eran, et. al., Microsoft, Mellanox, “Congestion control for large-scale RDMA deployments”, SIGCOMM 2015
  - Radhika Mittal, et. al., UC Berkeley, Google, “TIMELY: RTT-based Congestion Control for the Datacenter”, SIGCOMM 2015
- The scale of Data Center Networks continues to grow
  - Larger, faster clusters are better than more smaller size clusters
  - Server growth continues at 25% - 30% putting pressure on cluster sizes and networking costs

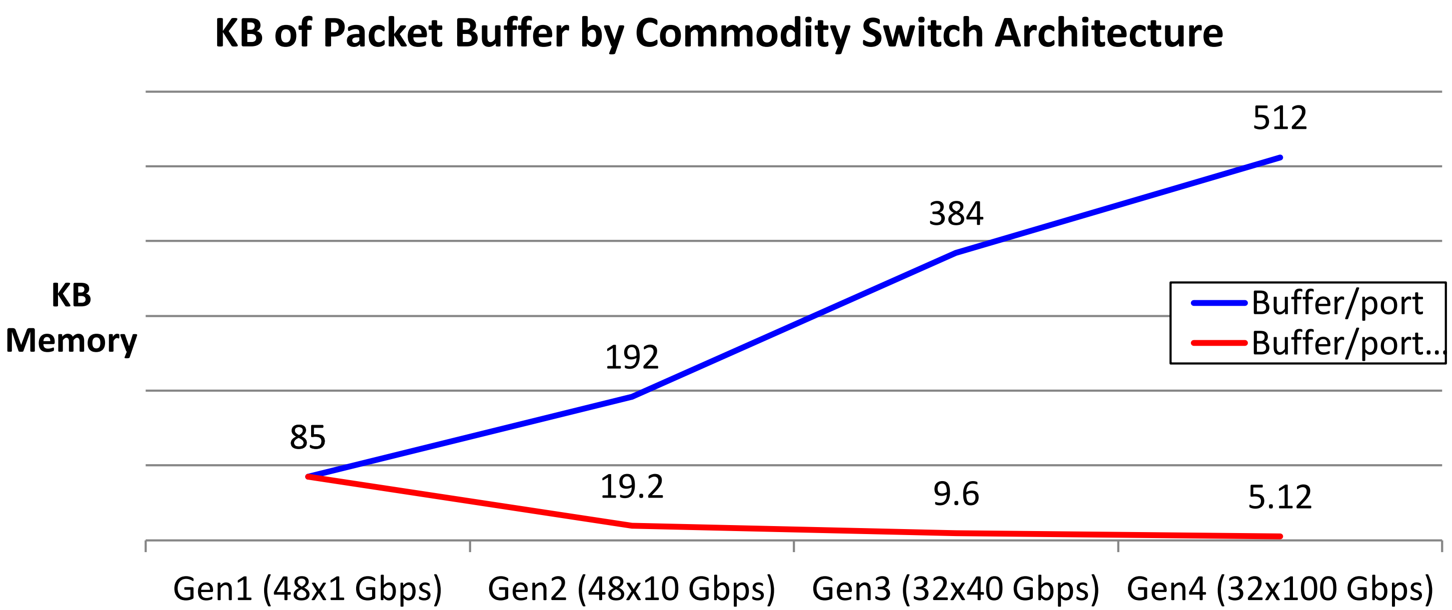
Lossless DCN state-of-the-art



- 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

Challenges going forward

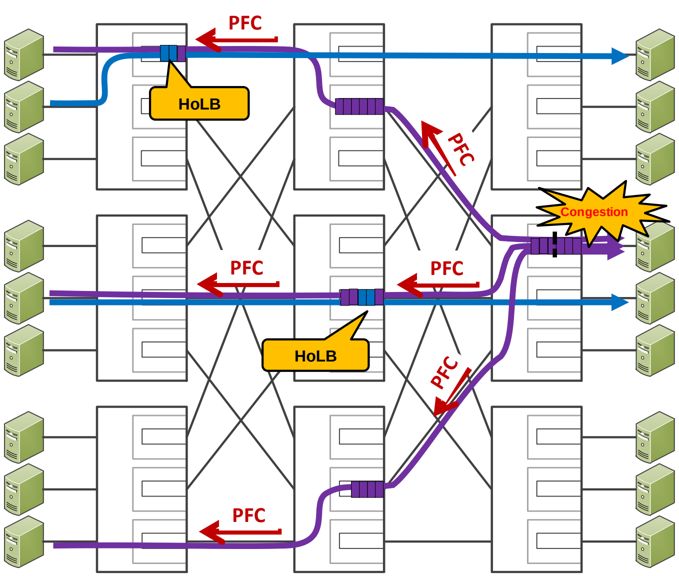
- Scaling the high-performance data center
  - More hops => more congestion points
  - Faster links => more data in flight
- Switch buffer growth is not keeping up



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

Existing 802.1 Congestion Management Tools

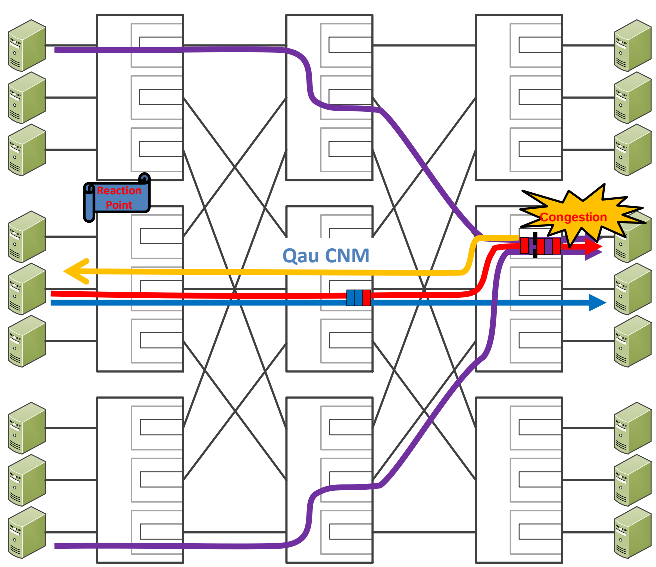
802.1Qbb - Priority-based Flow Control



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

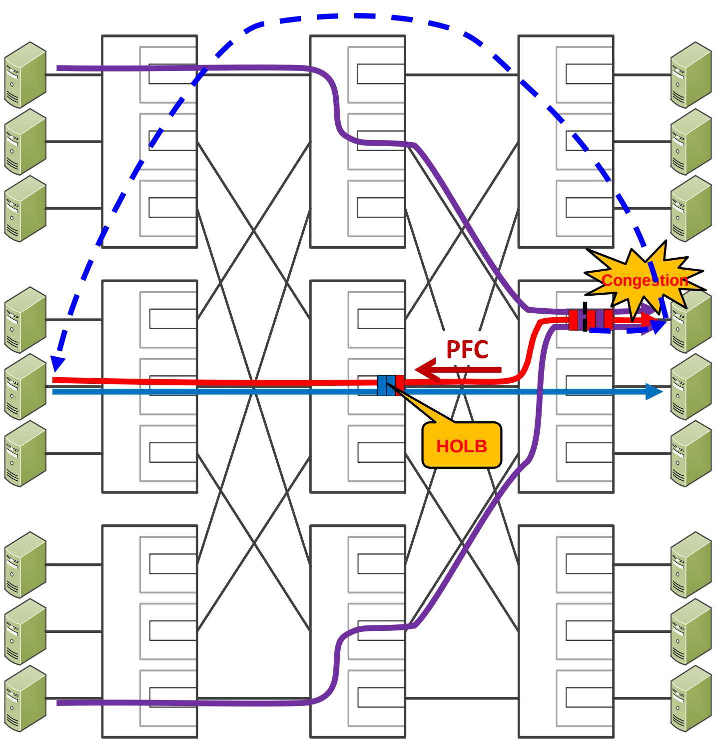
- Layer-2 end-to-end congestion control
- NIC based rate-limiters (Reaction Points)
- Designed for non-IP based protocols
  - FCoE
  - RoCE – v1

Congestion Isolation at a High Level

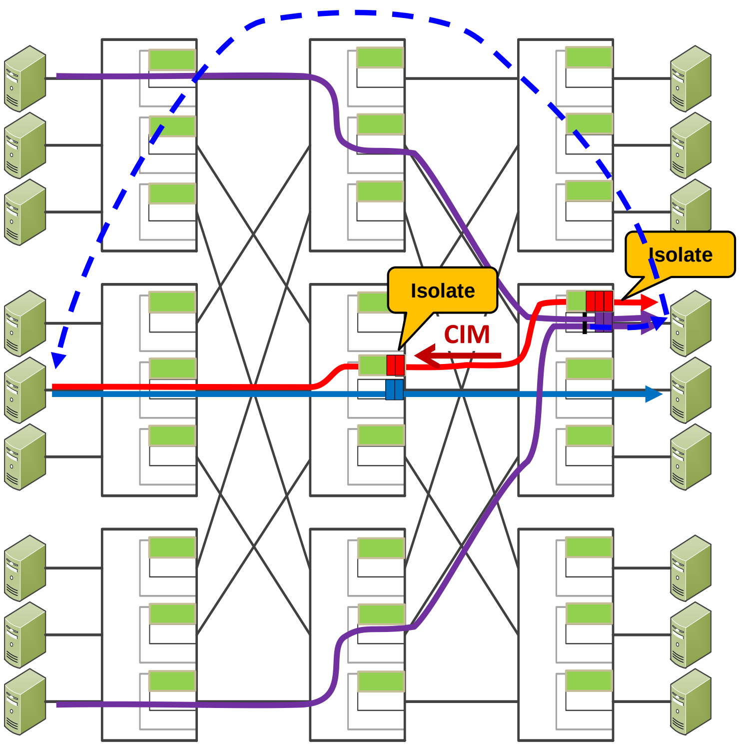
Congested Flow — Congested Queue

Victim Flow — Normal Queue

Today – Without Congestion Isolation



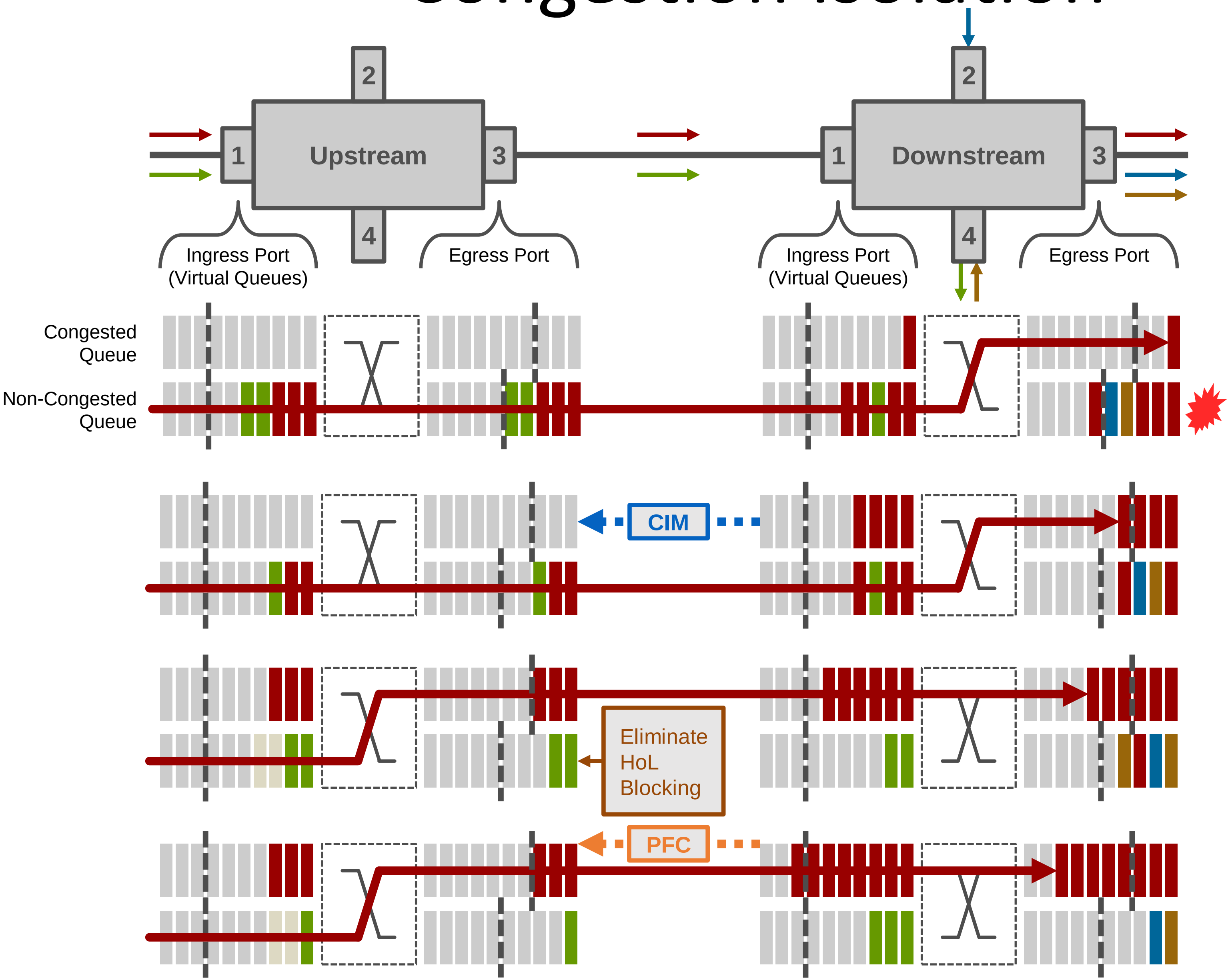
Congestion Isolation



P802.1Qcz – Congestion Isolation - Goals

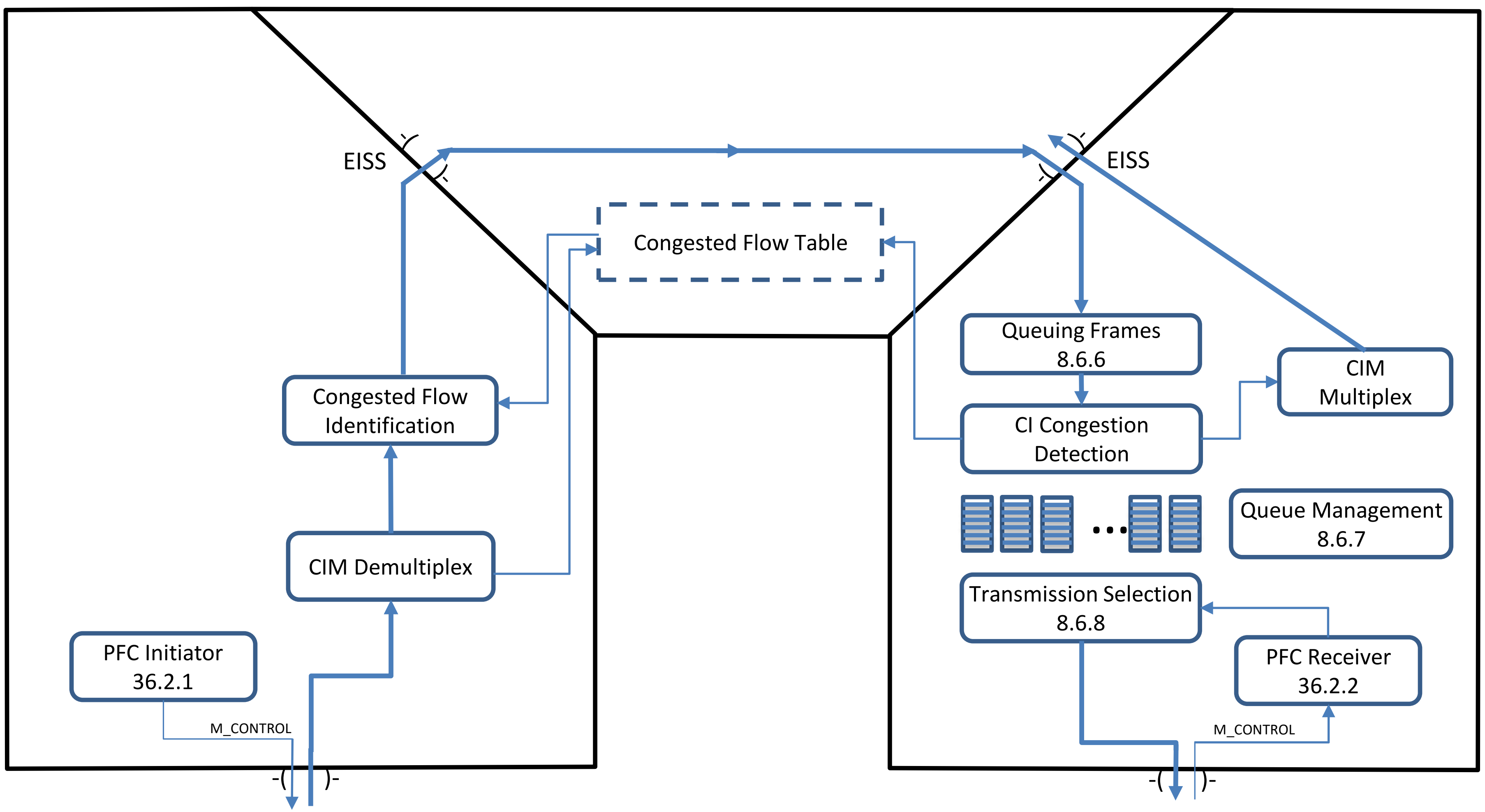
- Work in conjunction with higher-layer end-to-end congestion control (ECN, BBR, etc)
- Support larger, faster data centers (Low-Latency, High-Throughput)
- Support lossless and low-loss environments
- Improve performance of TCP and UDP based flows
- Reduce pressure on switch buffer growth
- Reduce the 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
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

Proposed Reference Diagram

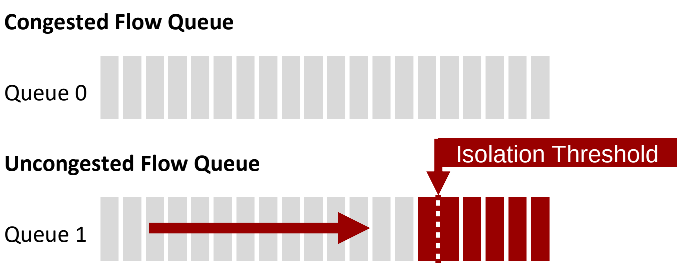


Congestion Isolation Critical Processes

1. Detecting flows causing congestion
2. Creating flows in the congested flow table
3. Signaling congested flow identify to neighbors
4. Isolating congested flows without ordering issues
5. Interaction with PFC generation
6. Detecting when congested flows are no longer congested
7. Signaling congested to non-congested flow transitions to neighbors
8. Un-isolating previously congested flows without ordering issues

Detecting Flows that Cause Congestion

- Expect to not to specify a new mechanism
  - Reference existing Quantized Congestion Notification (QCN) sampling approach (Clause 30.2.1)
  - Reference IETF standards/recommendations?
  - Allow implementation flexibility



Creating flows in the congested flow table

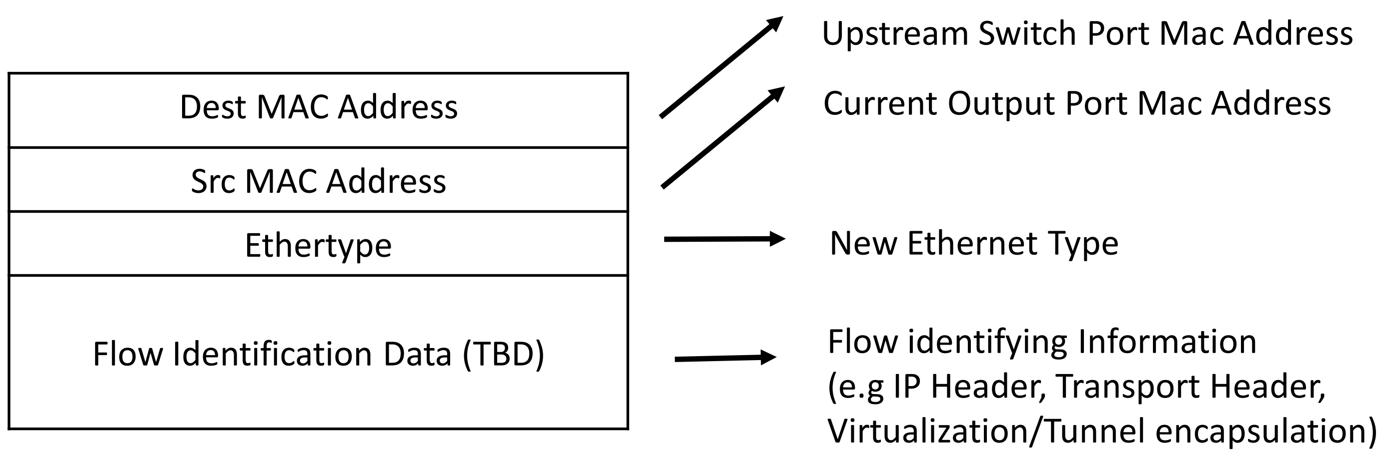
- Should only be required to register congested flows
- A flow is a traditional 5-tuple
- No failure if flow table is full
- May be useful to know how many and when CIMs were generated

Congested Flow Table

Src IP	Dest IP	Protocol	Src Port	Dest Port	Packet Counter	Last Active Time	…
10.136.159.100	10.136.169.100	17	3245	4791	100	0x4a32fa32	…
…	…	…	…	…	…	…	…
10.120.31.21	10.120.34.21	6	2345	80	20	0x4a33231f	…
10.189.32.20	10.189.31.21	6	23	81	1022	0x4a3323f4	…

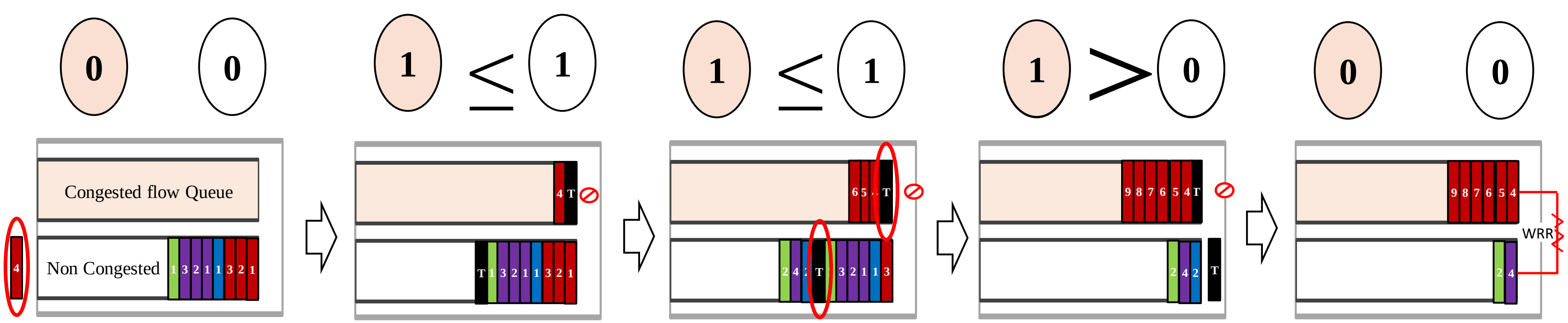
Signaling congested flow identify to neighbors

- Neighbor needs enough information to identify the same flow
  - First n-bytes of frame or explicit packet fields?
  - Leverage QCN format for Congestion Notification Message
- No adverse effects of single packet loss
- No requirement to identify flows within virtualization overlay encapsulation
- Mandatory or optional functionality?



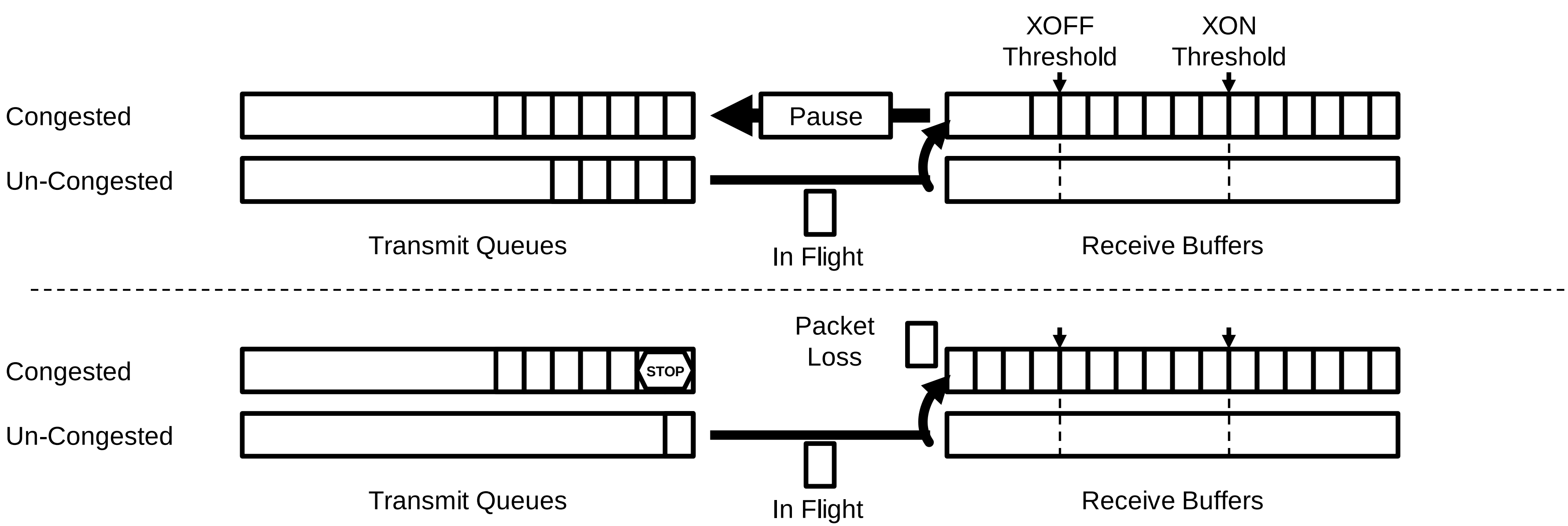
Isolating congested flows without ordering issues

- Conformance will be specified as “externally observable behavior”
- Potential example approach as an informative Annex



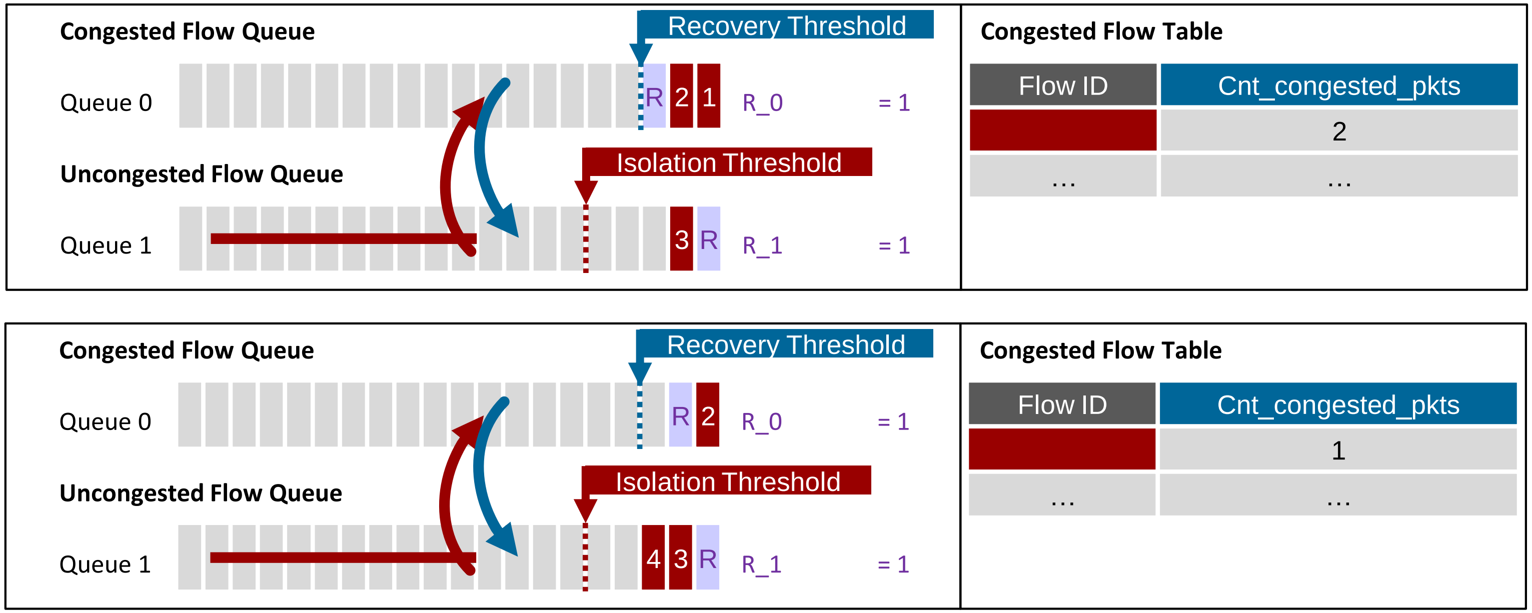
Interaction with PFC generation

- Priority-based Flow Control (PFC) is required for fully ‘lossless’ mode of operation
- The goal is, that if needed, PFC should be issued primarily on the congested traffic class. However…
  - Due to CIM signaling delays, packets may exist in both un-congested and congested upstream traffic classes after isolation.
  - The downstream switch/router needs to ‘know’ which traffic class was used by an upstream switch - Solution is to use Priority tagged or VLAN tagged format



Removing entries from congested flow table (a flow is no longer congested)

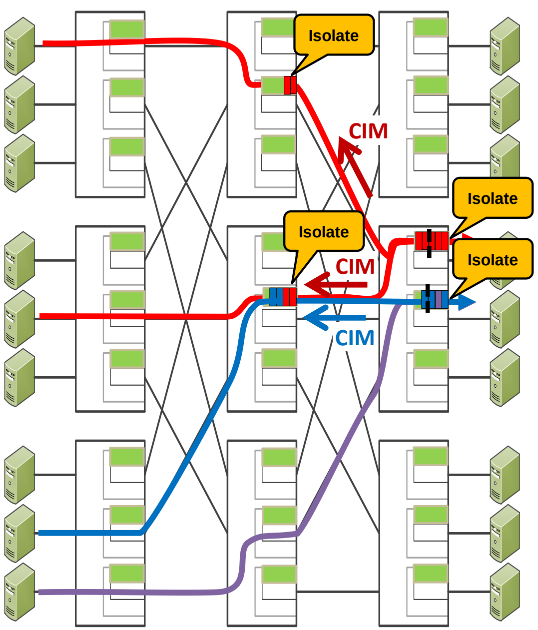
- Mechanisms are needed to free entries in the congested flow table
  - flush all entries egressing a port if the congested queue empties – Obvious
  - Define a recovery threshold in the congested flow queue and maintain order through scheduling
  - Count packets of a flow in the congested flow queue – Ideal, but maybe difficult to implement



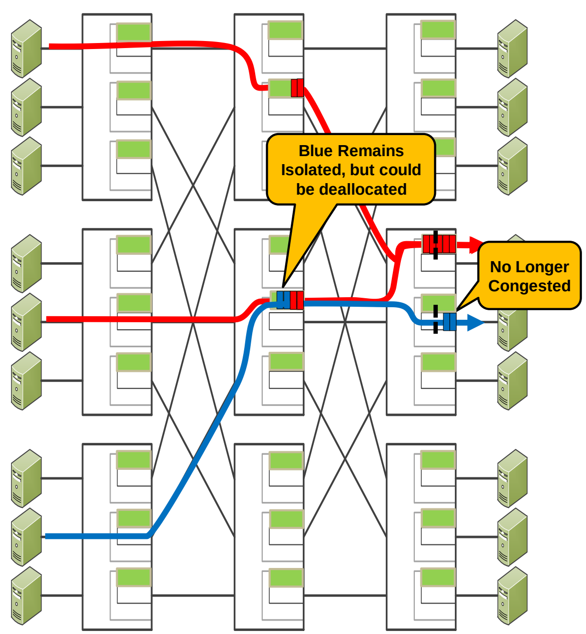
Signaling congested to non-congested transitions to neighbors

- Flows may remain congested upstream longer than necessary
- No natural mechanism to generate multiple messages

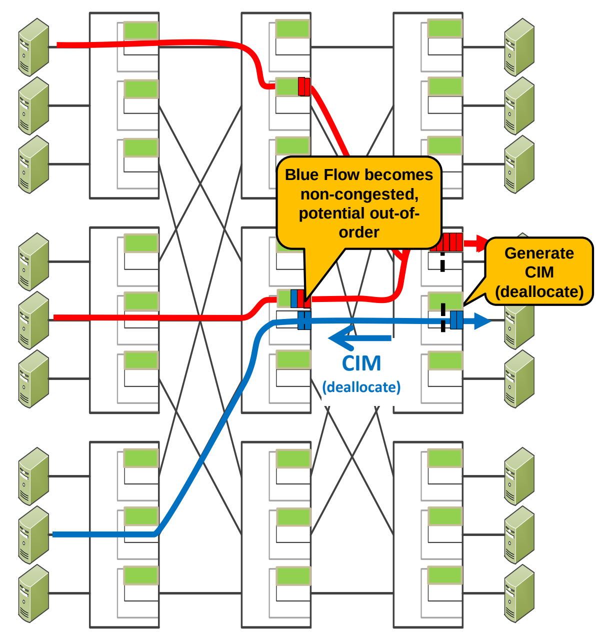
Two congested egress queues



Congested subsides on one

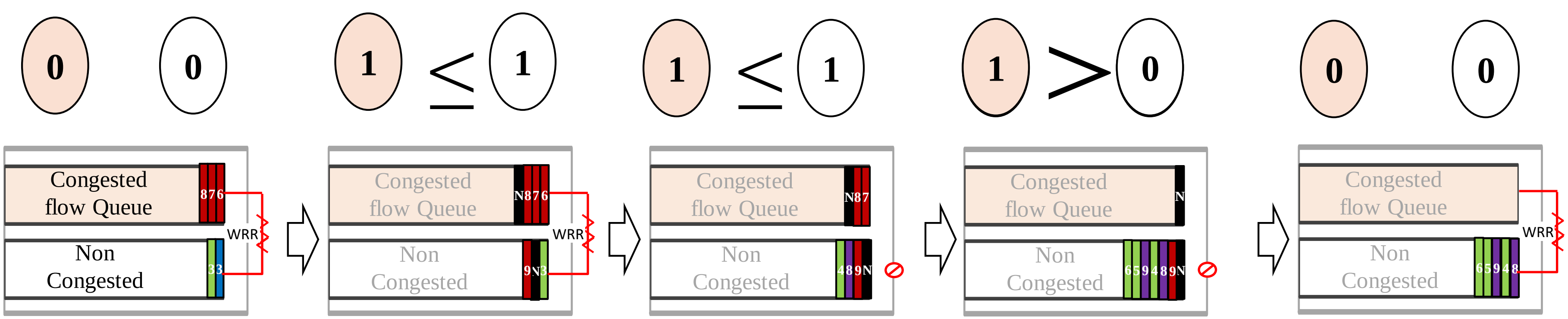


Inform neighbor



Un-isolating previously congested flows without ordering issues

- Specify externally observable behavior, not implementation details
- Similar mechanism when isolating a flow, but in reverse



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

References

- P802.1Qcz project web-page
  - https://1.ieee802.org/tsn/802-1qcz/
- Useful Presentations
  - Objectives Discussion
    - http://www.ieee802.org/1/files/public/docs2018/new-dcb-congdon-ci-objectives-0118-v02.pdf
  - Technical overview of CI
    - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-congestion-isolation-review-0418-v1.pdf
  - Simulation Results
    - http://www.ieee802.org/1/files/public/docs2018/cz-sun-ci-simulation-update-0518-v01.pdf
  - Possible changes to 802.1Q
    - http://www.ieee802.org/1/files/public/docs2018/cz-congdon-ci-Q-changes-0618-v1.pdf