IEEE P802.1Qcz
Congestion Isolation

TSN Conference Call

April 16, 2018

Paul Congdon

paul.congdon@tallac.com

Project Background – P802.1Qcz

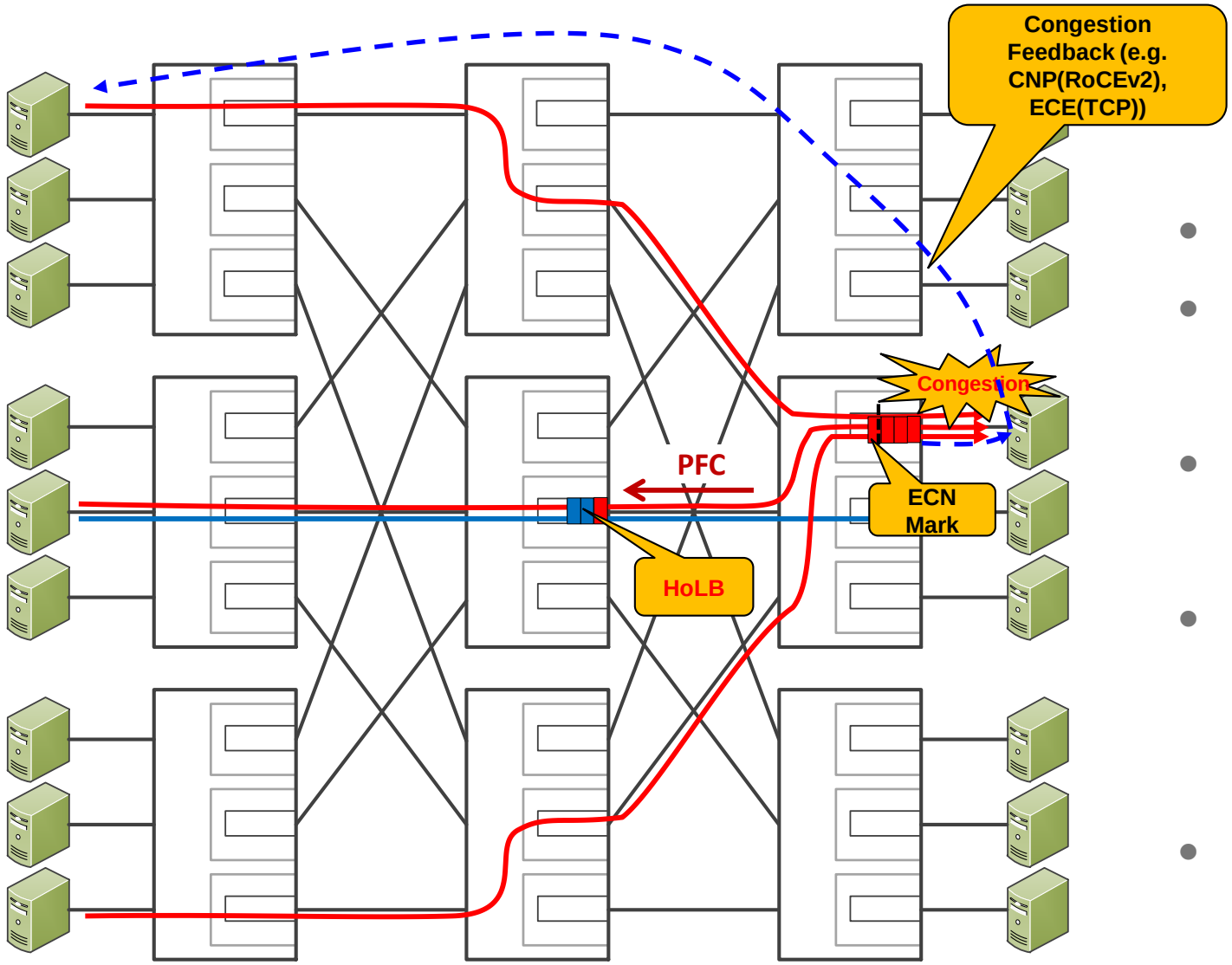
- Project Initiation
  - November 2017 – IEEE 802.1 agreed to develop a Project Authorization Request (PAR) and Criteria for Standards Development (CSD) to amend IEEE 802.1Q with “Congestion Isolation”
  - Motivation discussed in draft report of “802 Network Enhancements For the Next Decade”
    - https://mentor.ieee.org/802.1/dcn/18/1-18-0007-02-ICne-draft-report-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, 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 the upstream peer.
  - 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.

Lossless DCN state-of-the-art



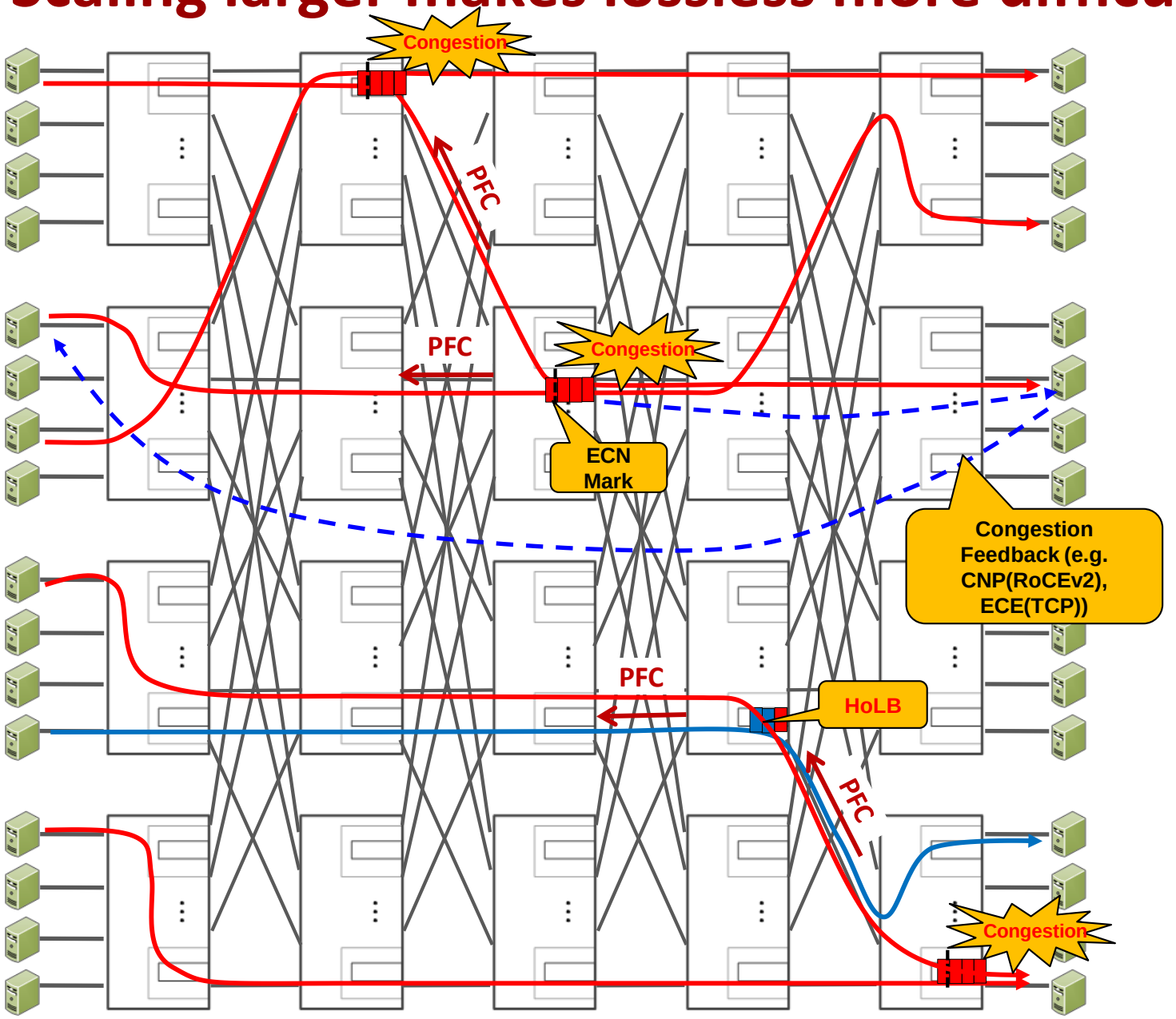
Congested Flow

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

Scaling larger makes lossless more difficult



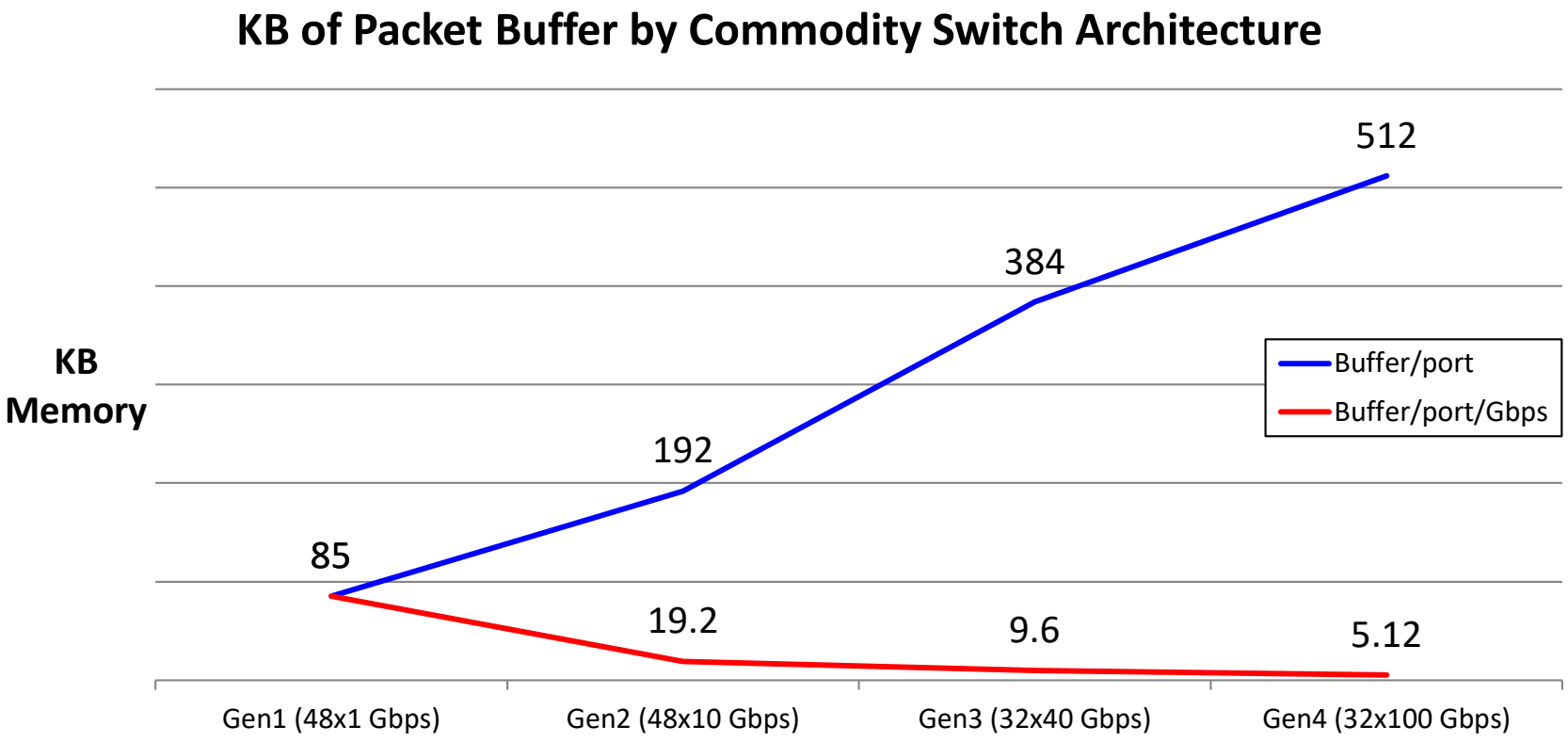
Congested Flow

Victim Flow

ECN Control Loop

- 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

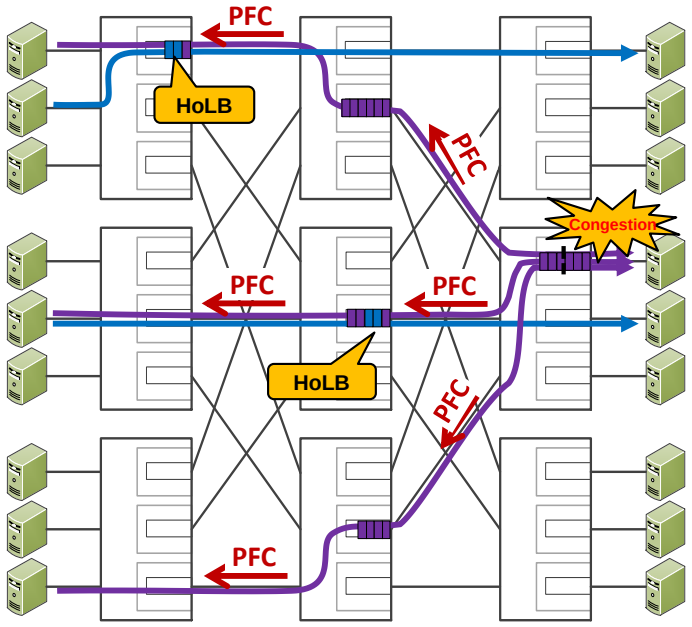
However, packet loss can create performance issues:

- 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

Existing 802.1 Congestion Management Tools

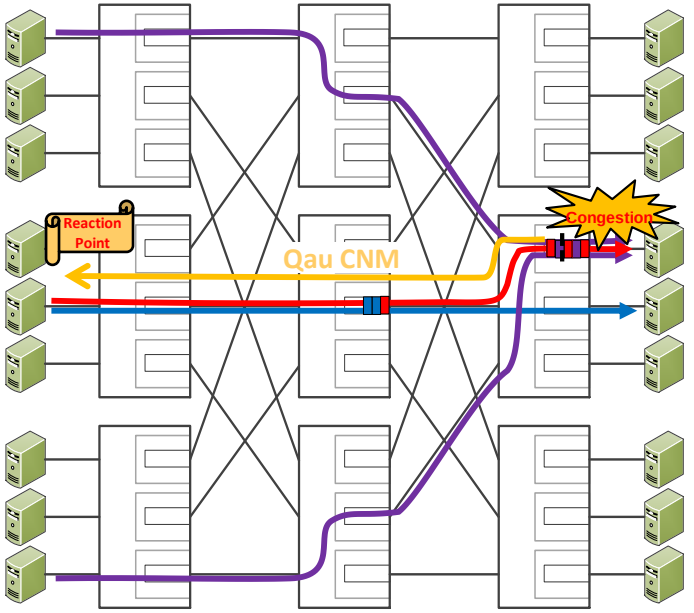
802.1Qbb - Priority-based Flow Control



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 caused by over-use of PFC

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 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

Congestion Isolation

Definition: An approach to isolate flows causing congestion and signal upstream to isolate the same flows to avoid head-of-line blocking.

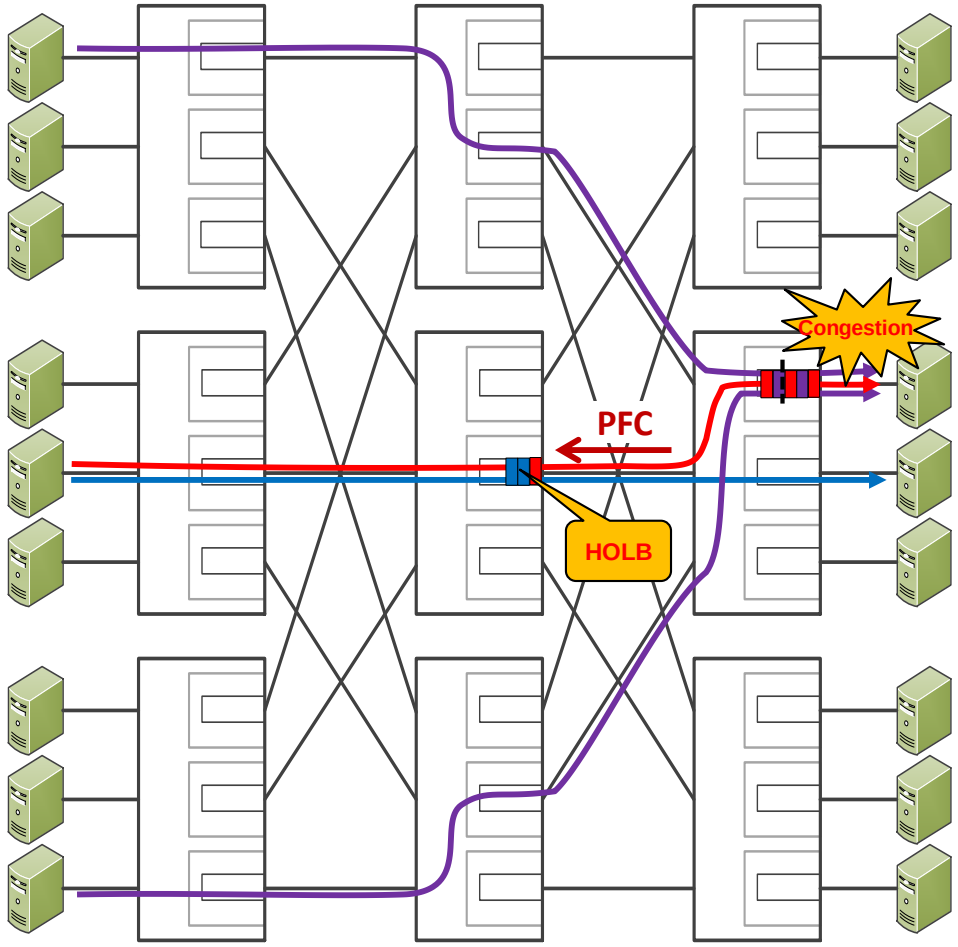
The approach involves:

1. Identifying the flows creating congestion (e.g. perhaps already done for 802.1Qau and/or ECN)
2. Using implementation specific approaches to dynamically adjust the traffic class of offending flows without packet re-ordering
3. Signaling upstream indications via a Congestion Isolation Message (CIM)

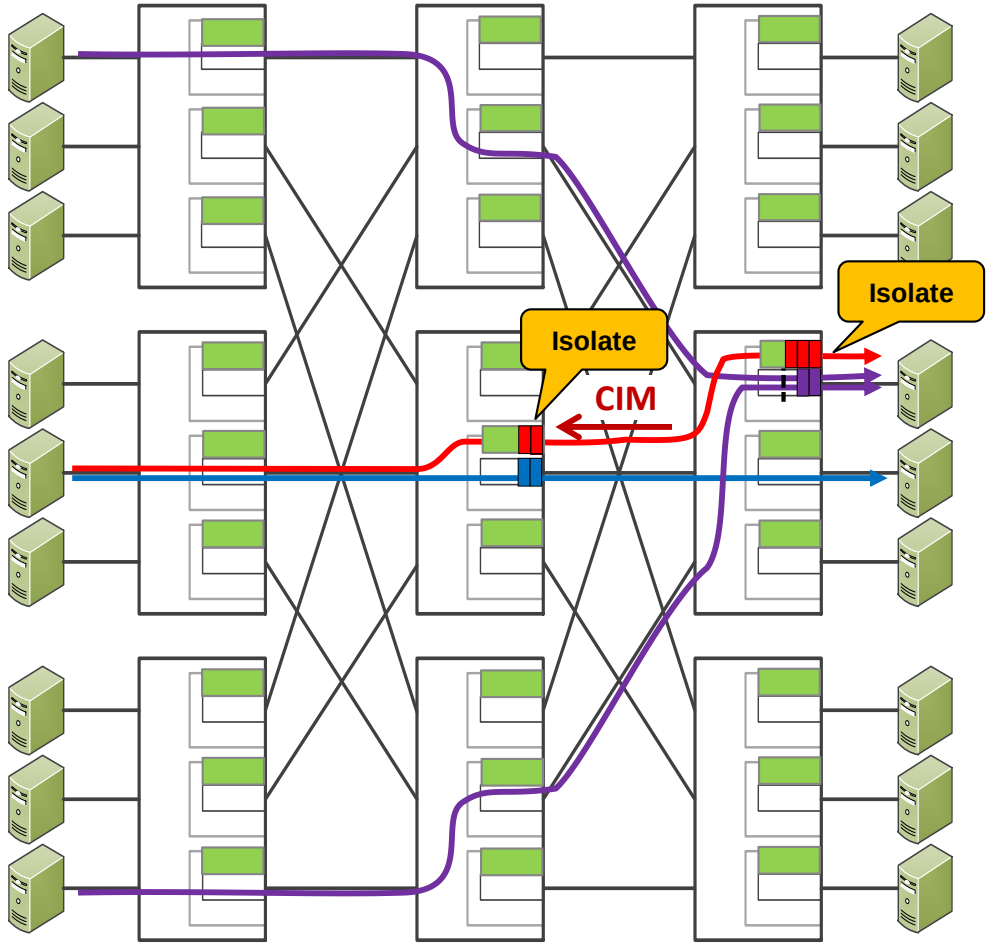
Isolate the congestion to mitigate HOLB

Congested Flow
Congested Queue
Victim Flow
Normal Queue

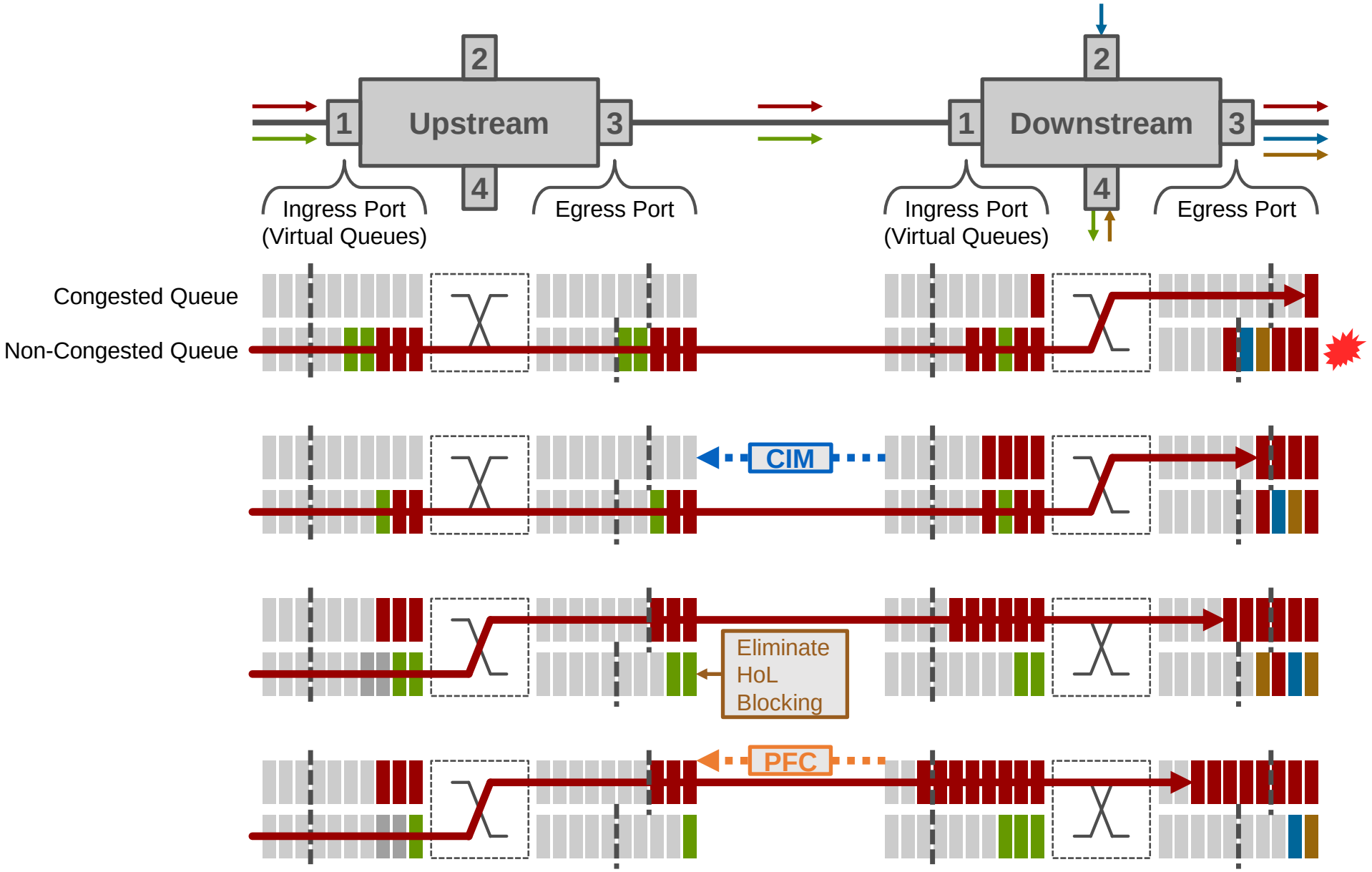
Today – Without Congestion Isolation



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

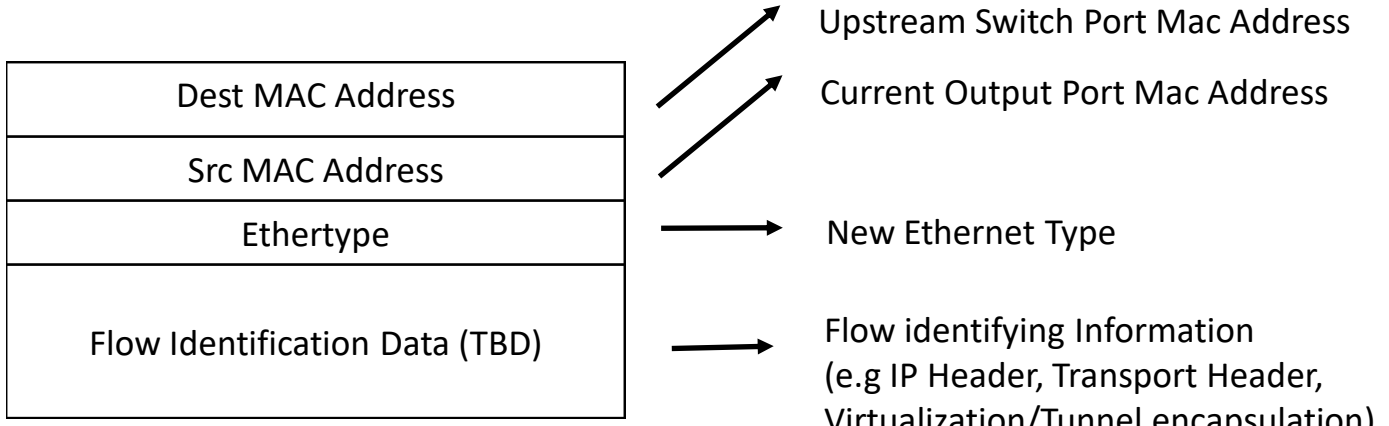
4. Last Resort! If congested queue continues to fill, invoke PFC for lossless

Congestion Isolation Message

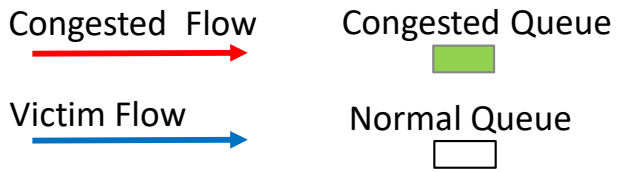
- Objectives/Requirements:
  - Provide upstream neighbor with an indication that a flow has been isolated
  - Provide upstream neighbor with flow identification information
  - No adverse effects of single packet loss
  - Low overhead

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

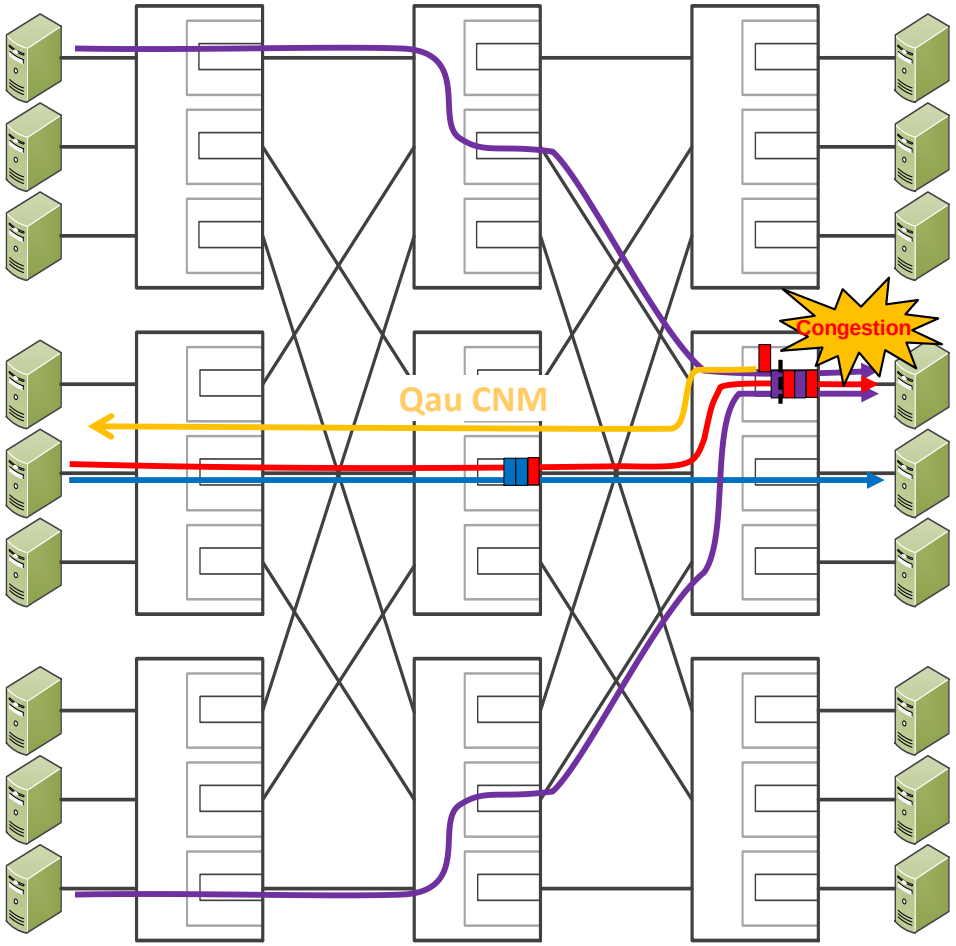
Format of Congestion Isolation Packet



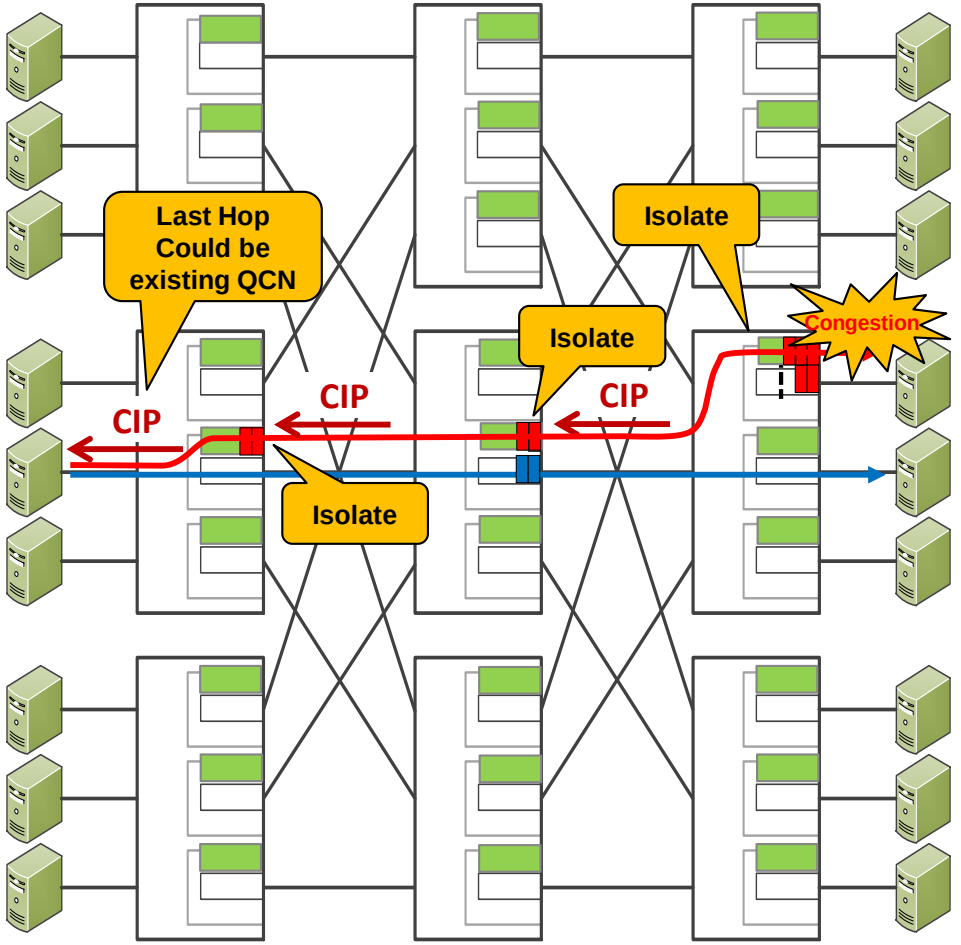
Leverage existing CNM message?



End-to-End
QCN Congestion Notification Message



Hop-by-Hop
Congestion Isolation



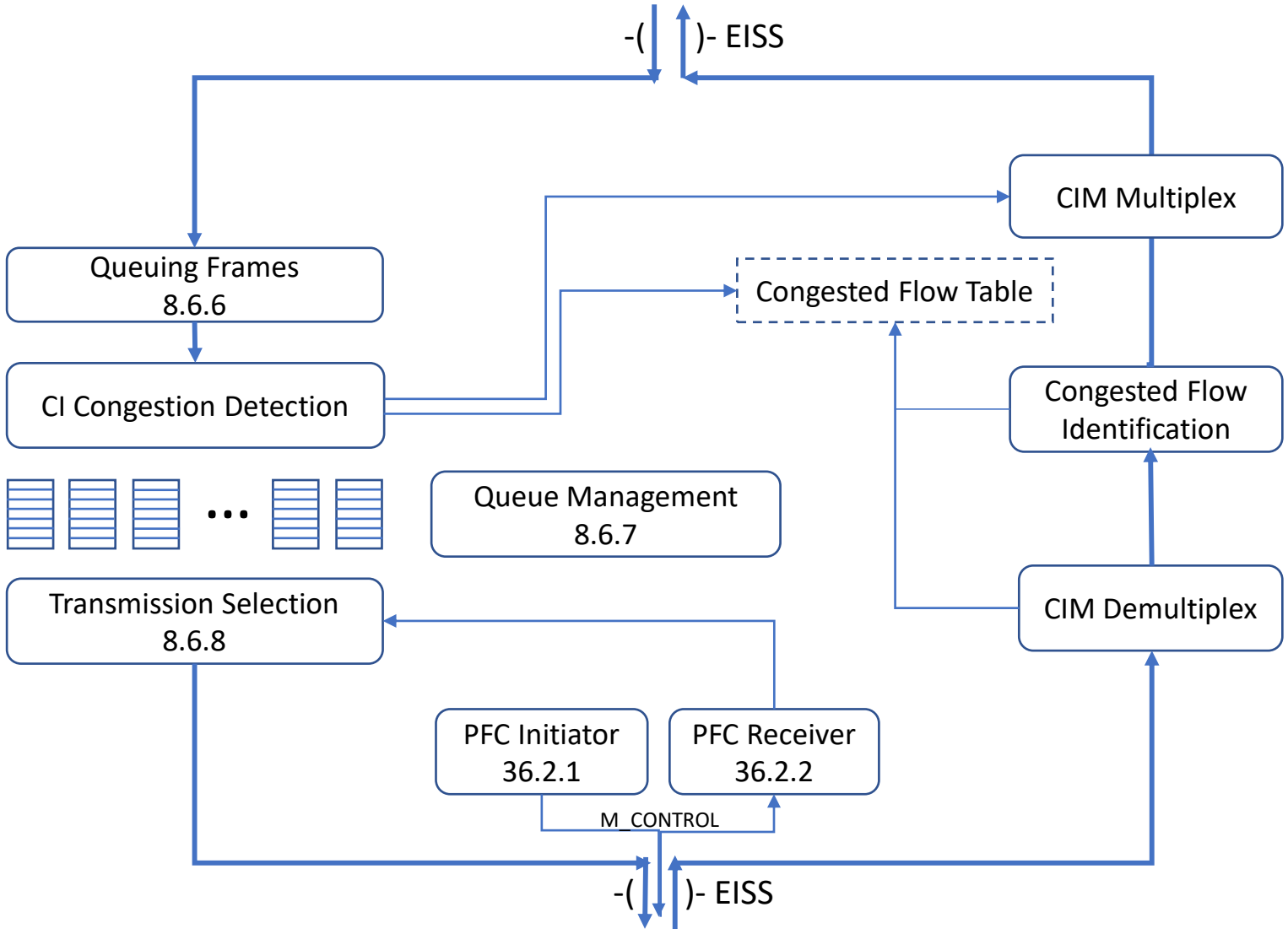
Capability Discover via LLDP

- 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
  - Helpful to inform the upstream switch of the inactivity timeout used downstream so it may use a larger timeout to avoid early ageing.

Format of LLDP TLV

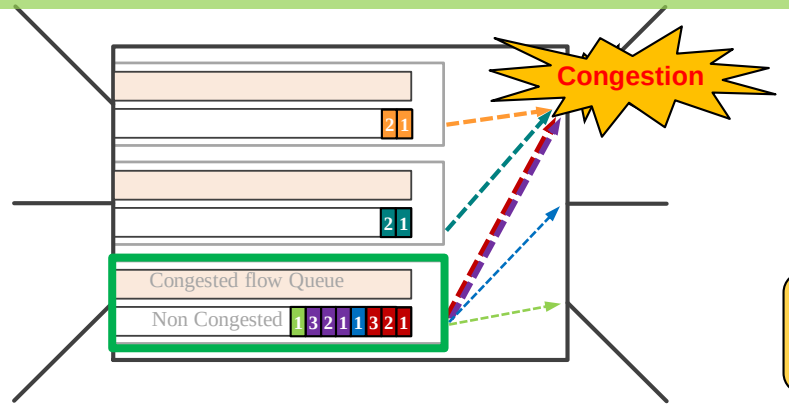
TLV Type	TLV info length	802.1 OUI	subtype	Congested Queue	Monitored Queues	Inactivity Timeout

Proposed Reference Diagram – work in progress



Handling the potential out-of-order problem

An instance：Red flow and purple flow are judged as congested flow and are moved to congested flow queue successively.



cong queue mark counter1 | Non-cong queue mark counter2

0 0

Congested flow Queue

Non Congested

Judged as congested flow， moved to congested flow queue

1 ≤ 1

S1. one mark enqueue respectively for two queues
S2. mark counter increases by 1
S3. the packet of red flow put into congested flow queue

1 ≤ 1

When T gets to the head of congested queue, it should wait until counter1 > counter2

1 ≤ 1

Judged as congested flow， moved to congested flow queue

2 ≤ 2

S1. one mark enqueue respectively for two queues
S2. mark counter increases by 1
S3. the packet of purple flow put into congested flow queue

2 > 1

Dequeue the mark, and decrease the counter by 1

1 1

counter1 > counter2, start to schedule congested flow queue, mark output, then decrease the counter

1 > 0

If the next T gets to the head of congested flow queue, stop schedule, wait until counter1 > counter2 again

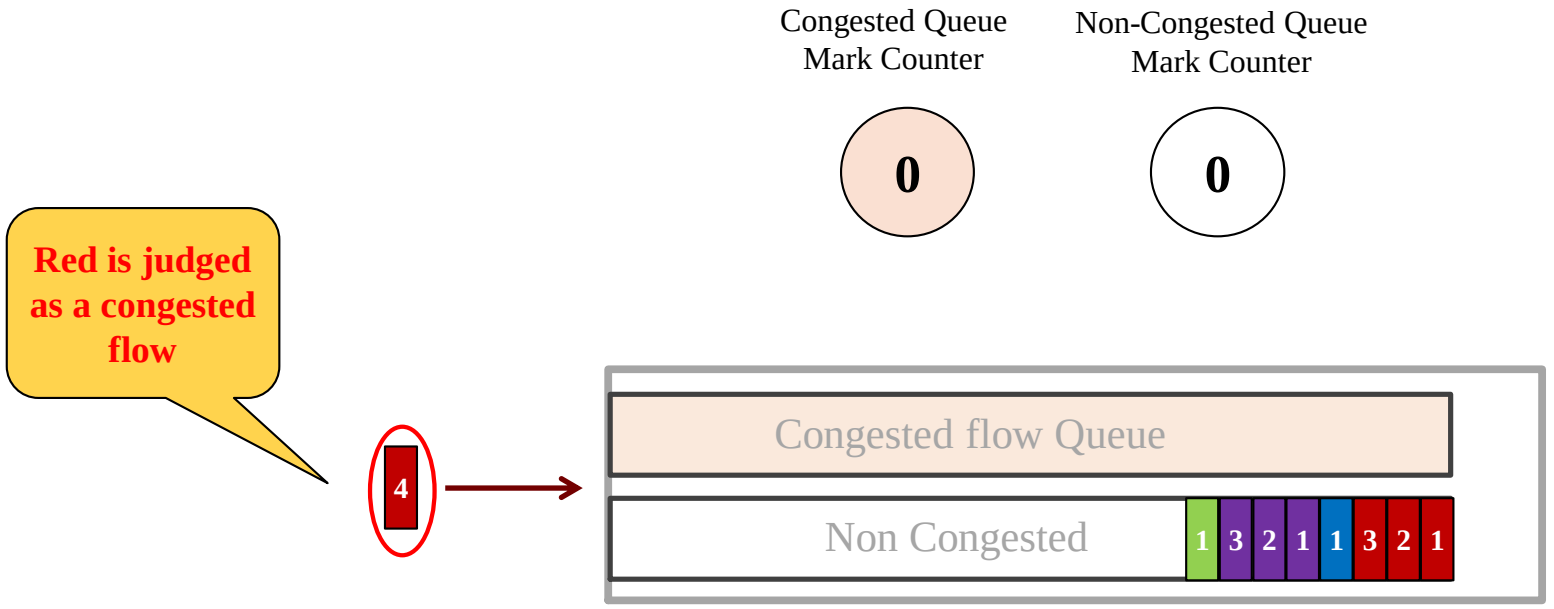
0 0

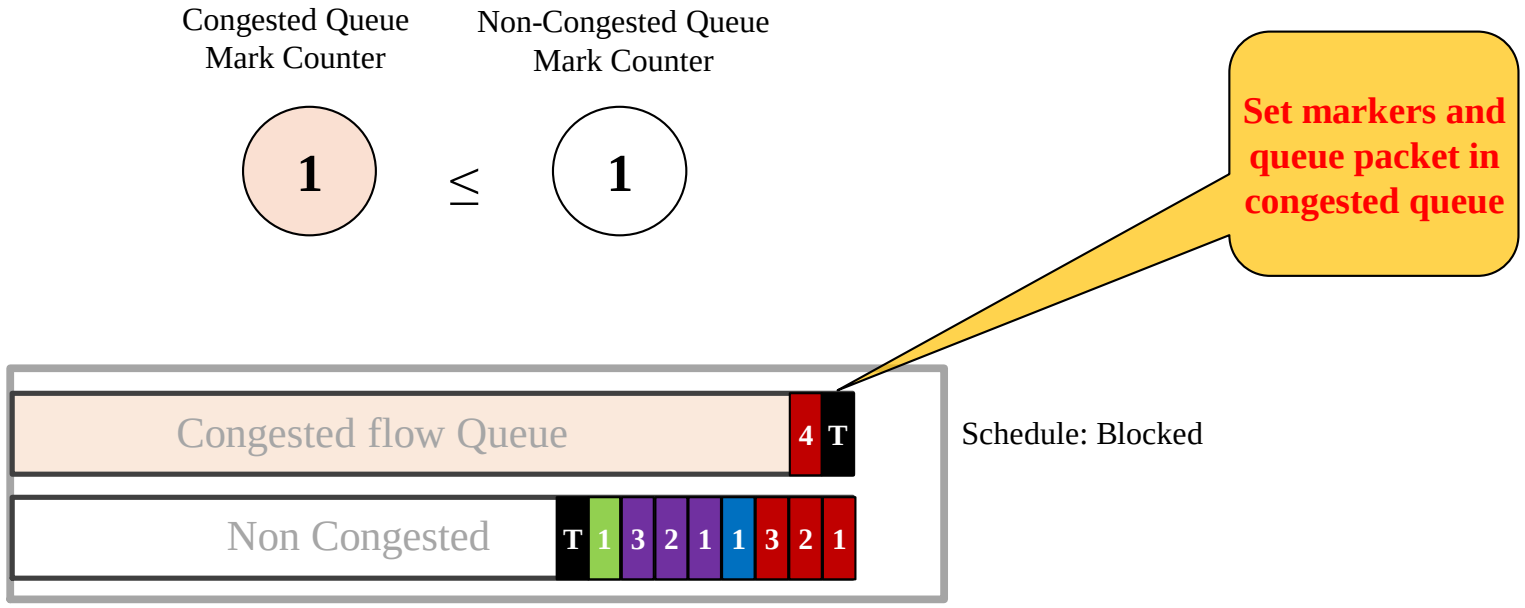
Start to schedule congested flow queue

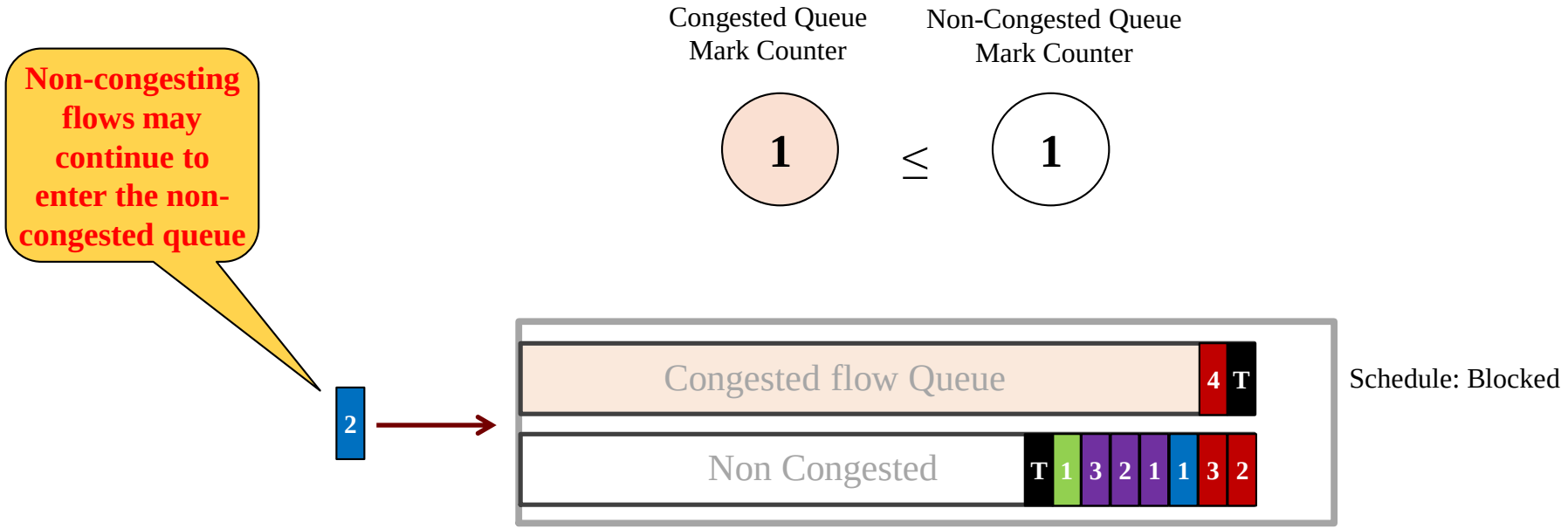
0 0

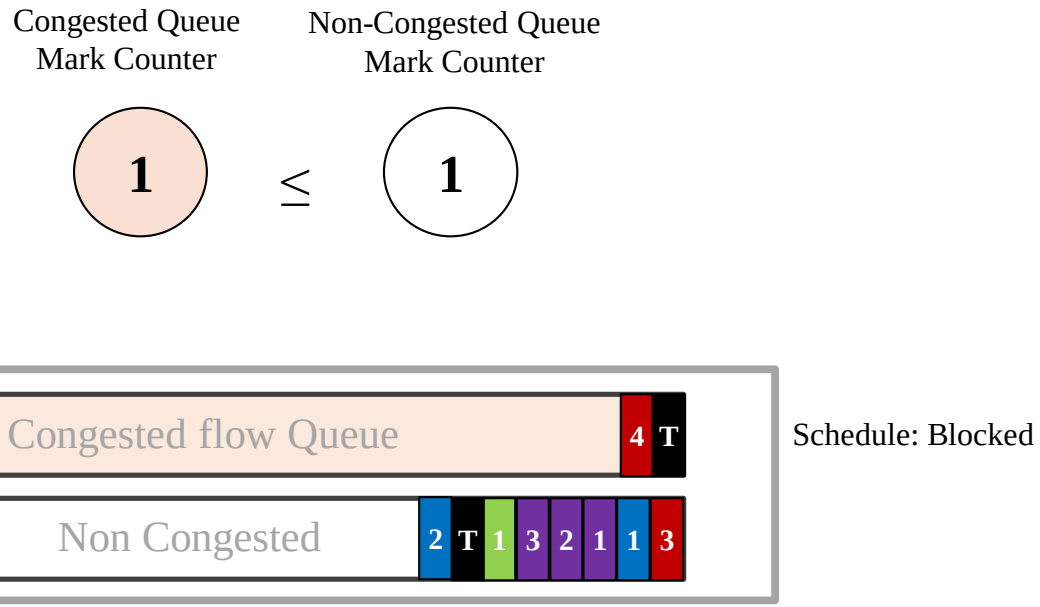
Output normally

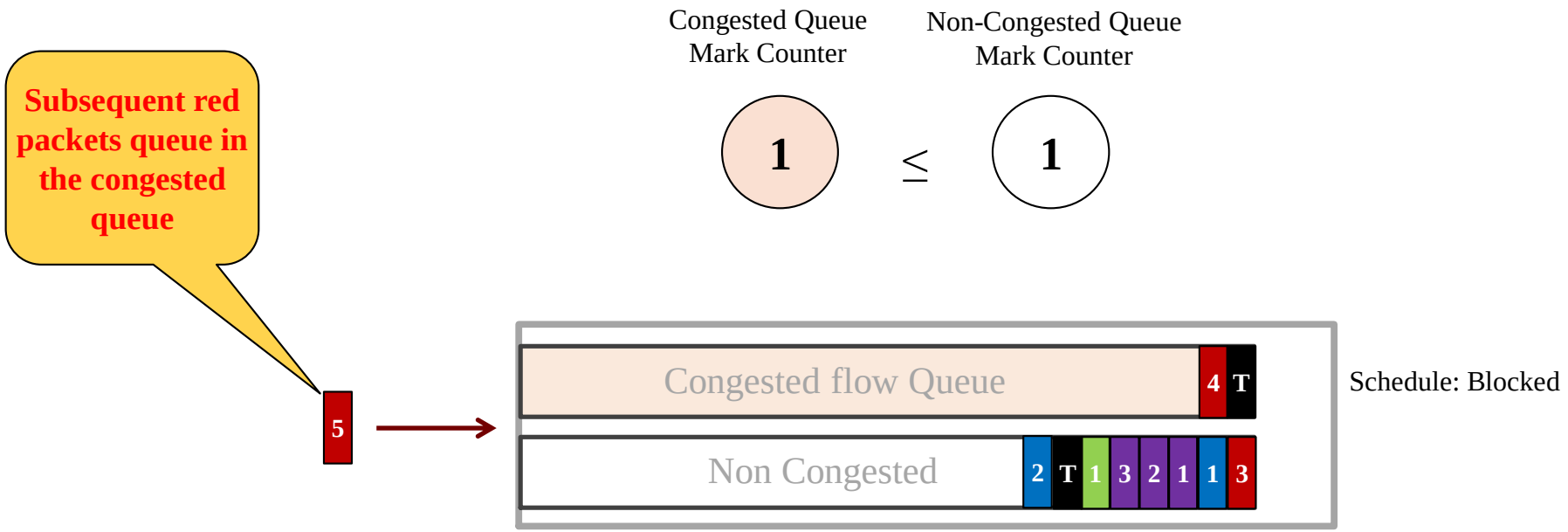
Handling the potential out-of-order problem

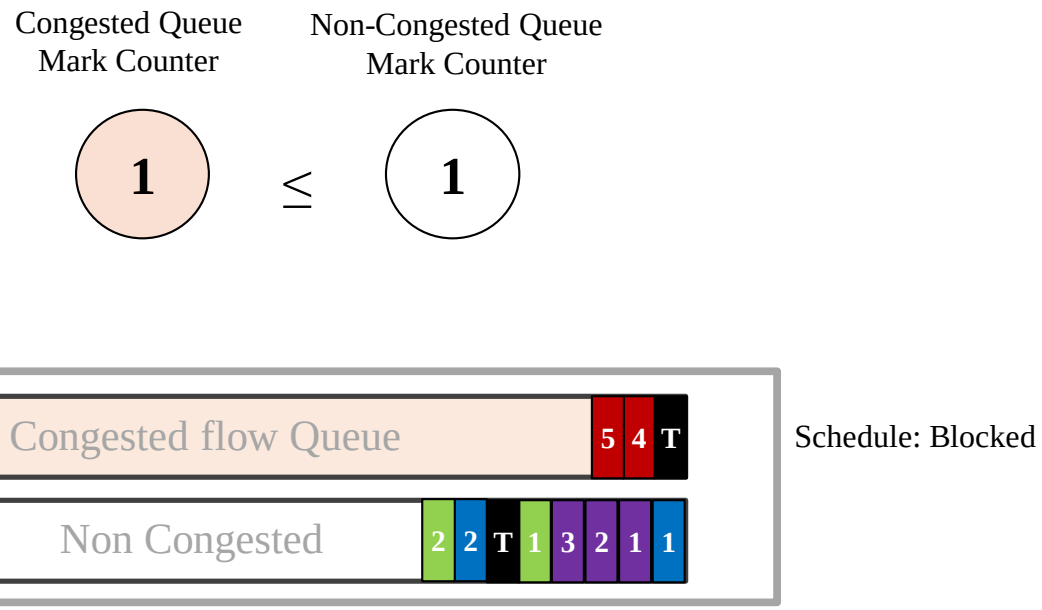


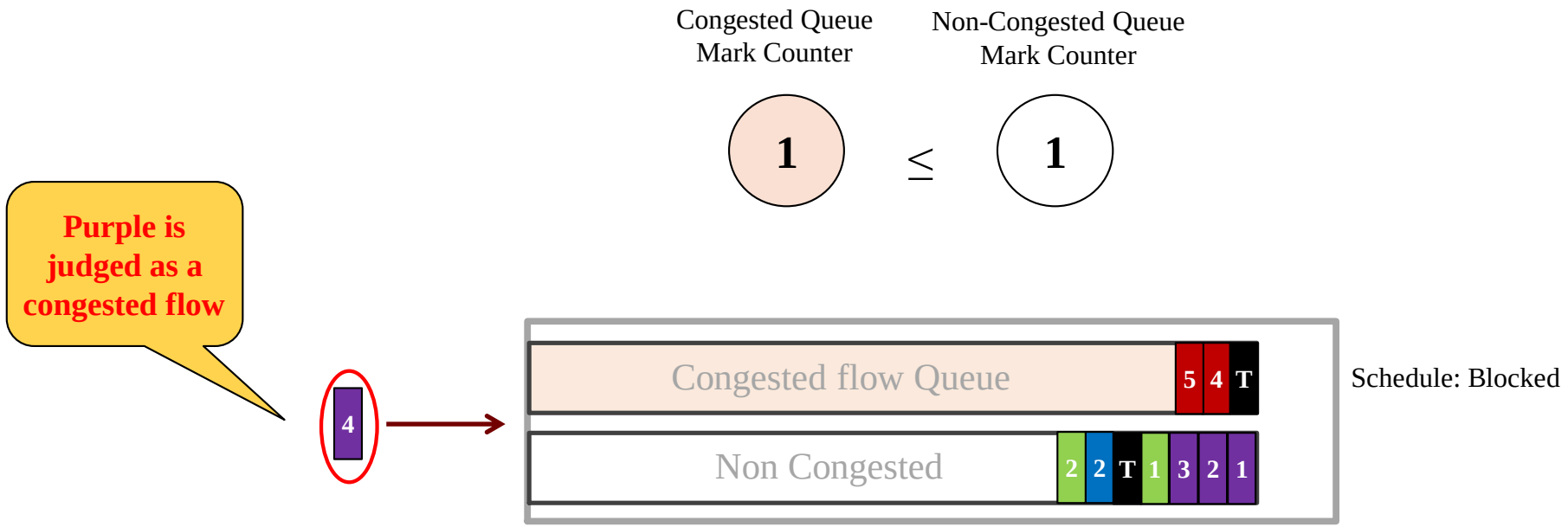

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Set markers and queue packet in congested queue
Congested flow Queue
4
T
Schedule: Blocked
Non Congested
T
1
3
2
1
1
3
2
1

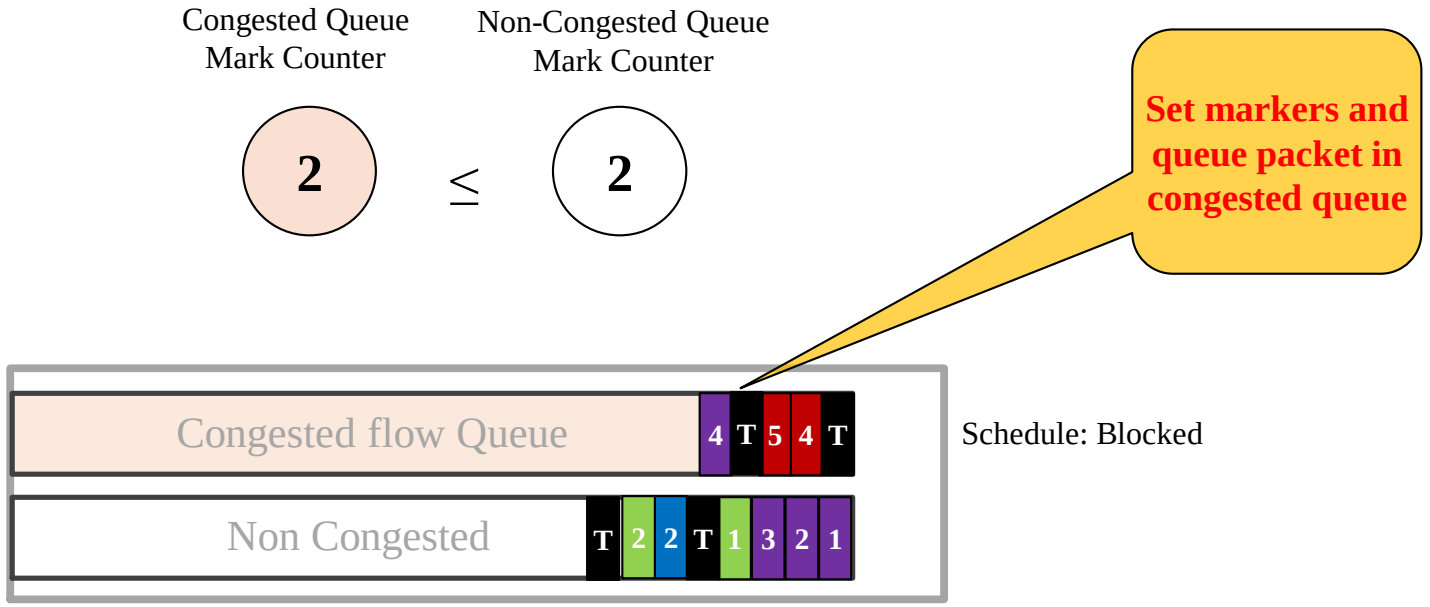

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Non-congesting flows may continue to enter the non-congested queue
2
Congested flow Queue
4
T
Non Congested
T
1
3
2
1
1
3
2
Schedule: Blocked

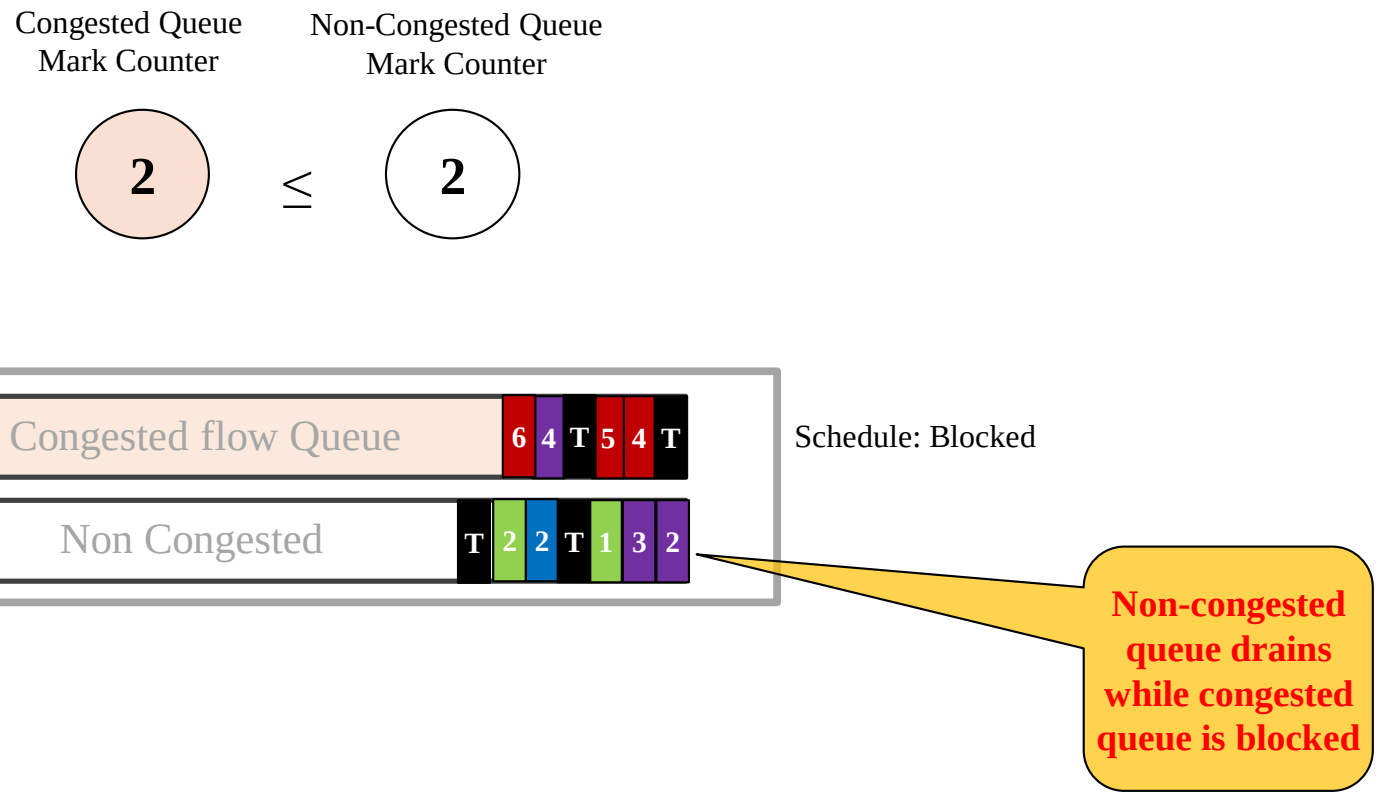

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Congested flow Queue
4
T
Non Congested
2
T
1
3
2
1
1
3
Schedule: Blocked

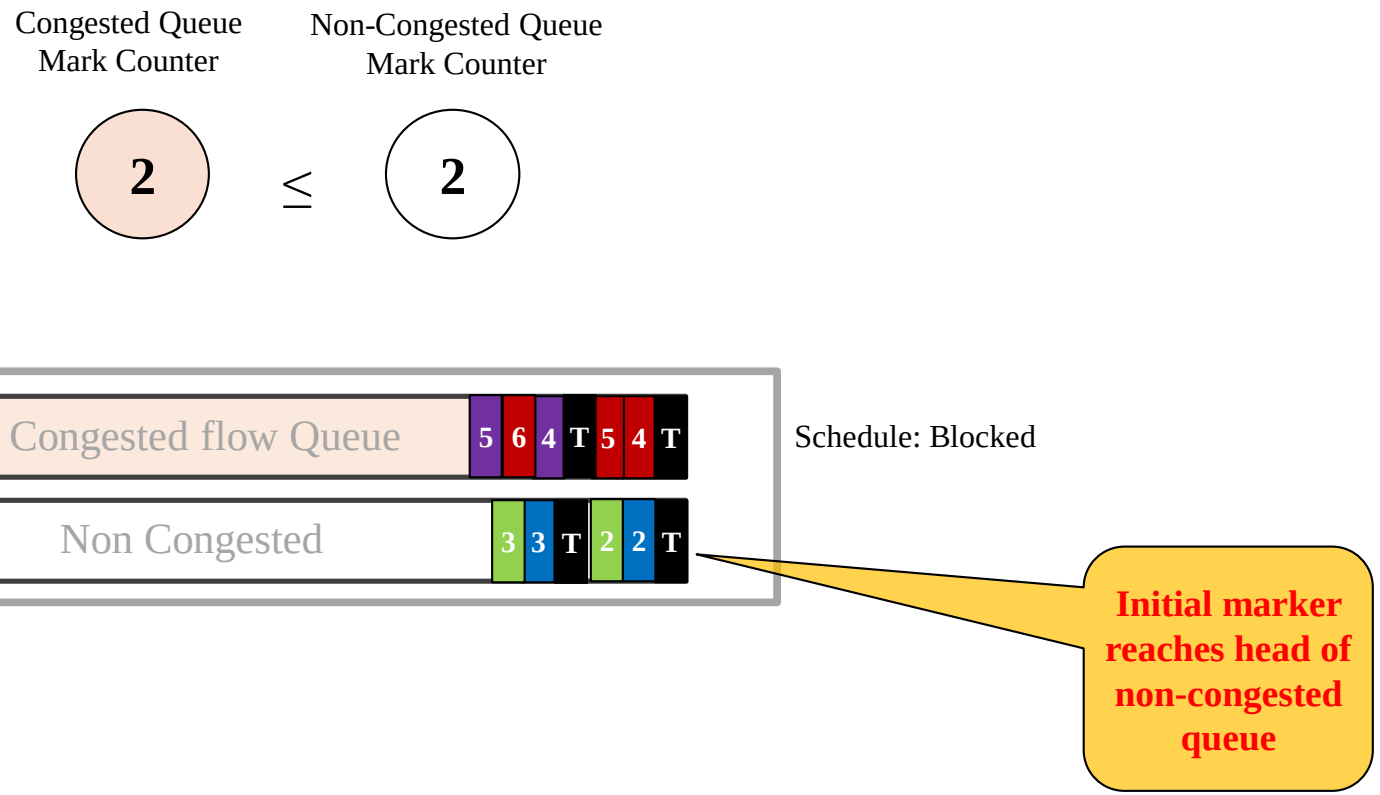

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Subsequent red packets queue in the congested queue
5
Congested flow Queue
4
T
Non Congested
2
T
1
3
2
1
1
3
Schedule: Blocked

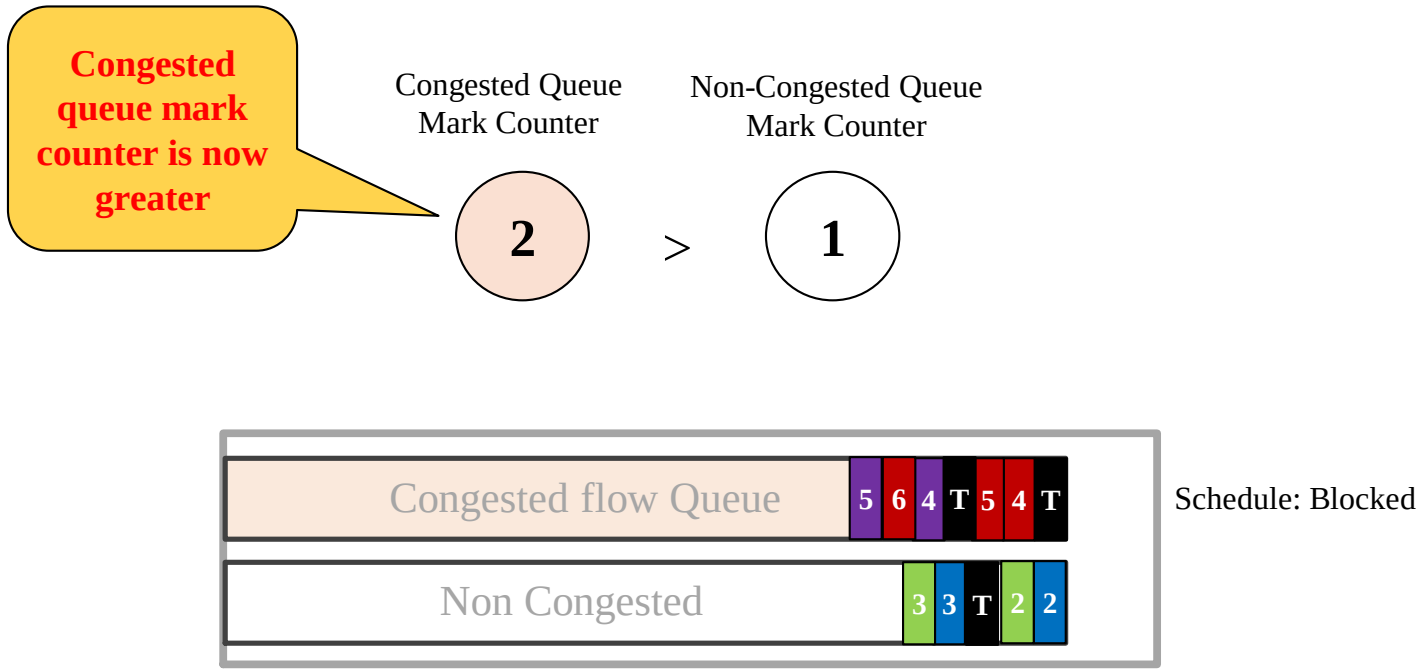

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Congested flow Queue
5
4
T
Non Congested
2
2
T
1
3
2
1
1
Schedule: Blocked

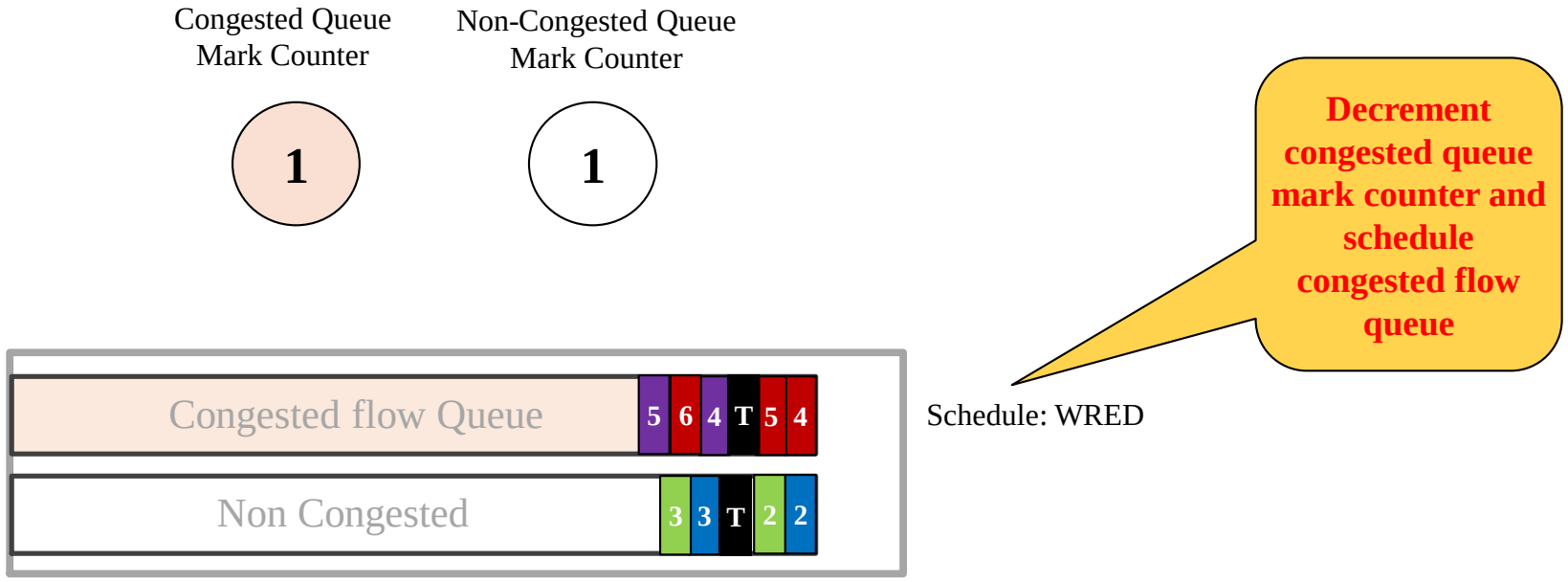

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Purple is judged as a congested flow
4
Congested flow Queue
5 4 T
Non Congested
2 2 T 1 3 2 1
Schedule: Blocked

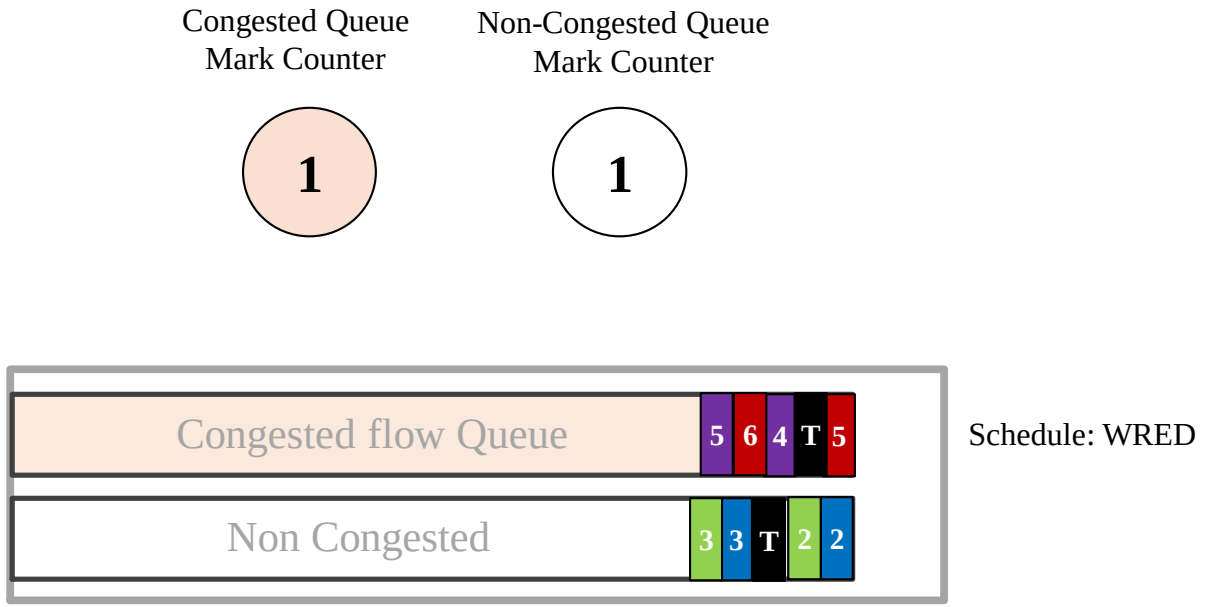

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
2
≤
2
Set markers and queue packet in congested queue
Congested flow Queue
4 T 5 4 T
Non Congested
T 2 2 T 1 3 2 1
Schedule: Blocked

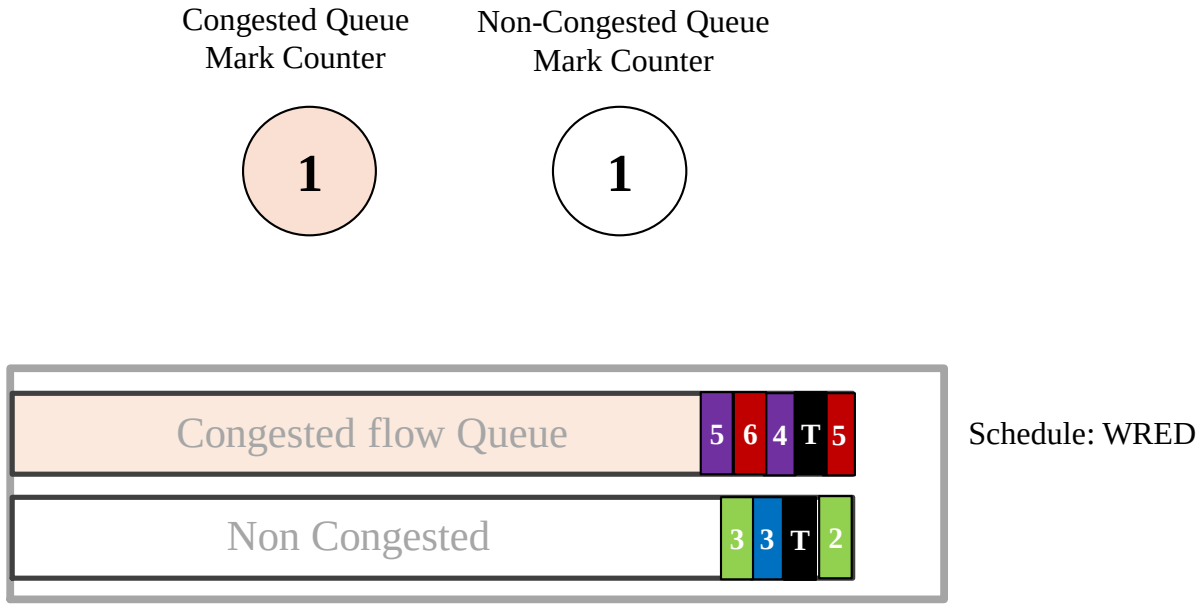

Congested Queue Mark Counter
Non-Congested Queue Mark Counter
2
≤
2
Congested flow Queue
6 4 T 5 4 T
Non Congested
T 2 2 T 1 3 2
Schedule: Blocked
Non-congested queue drains while congested queue is blocked

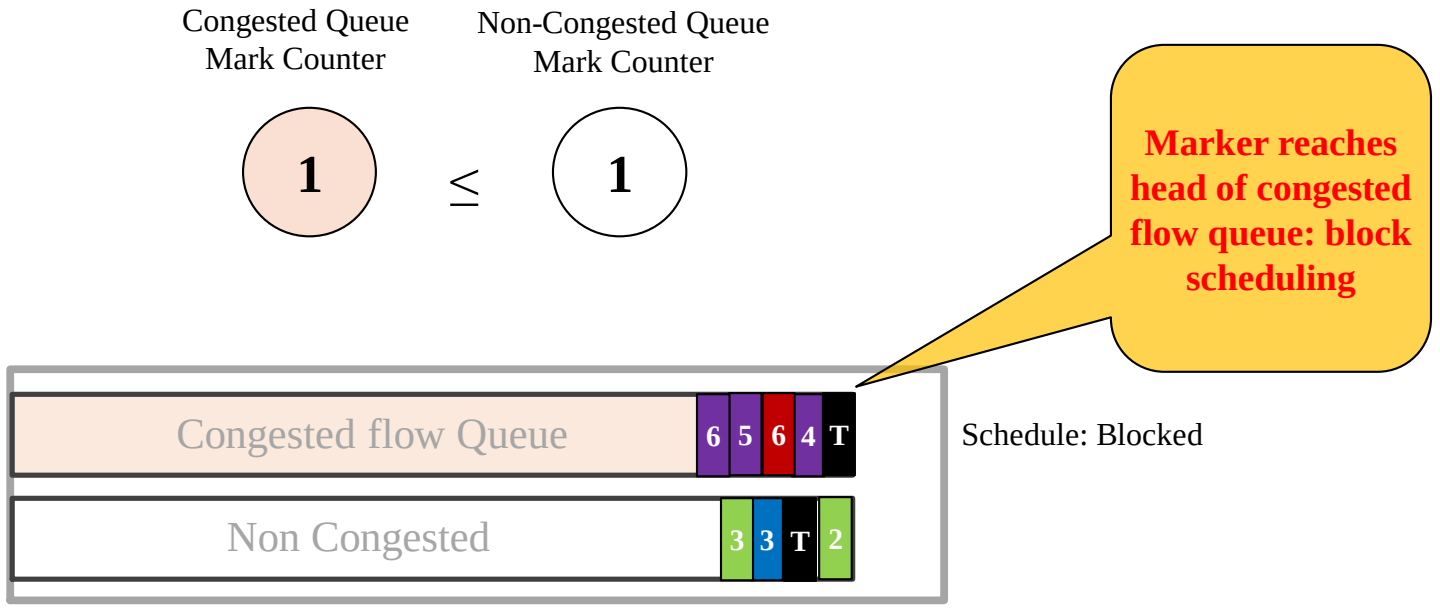

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
2
≤
2
Congested flow Queue
5 6 4 T 5 4 T
Non Congested
3 3 T 2 2 T
Schedule: Blocked
Initial marker reaches head of non-congested queue

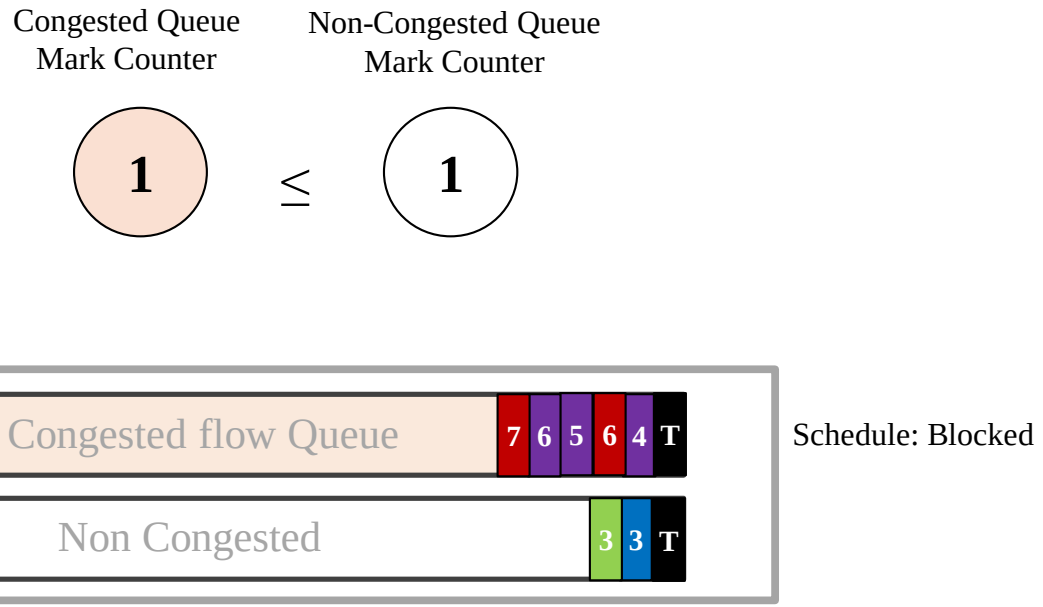

Congested queue mark counter is now greater
Congested Queue Mark Counter
Non-Congested Queue Mark Counter
2
>
1
Congested flow Queue
5 6 4 T 5 4 T
Non Congested
3 3 T 2 2
Schedule: Blocked

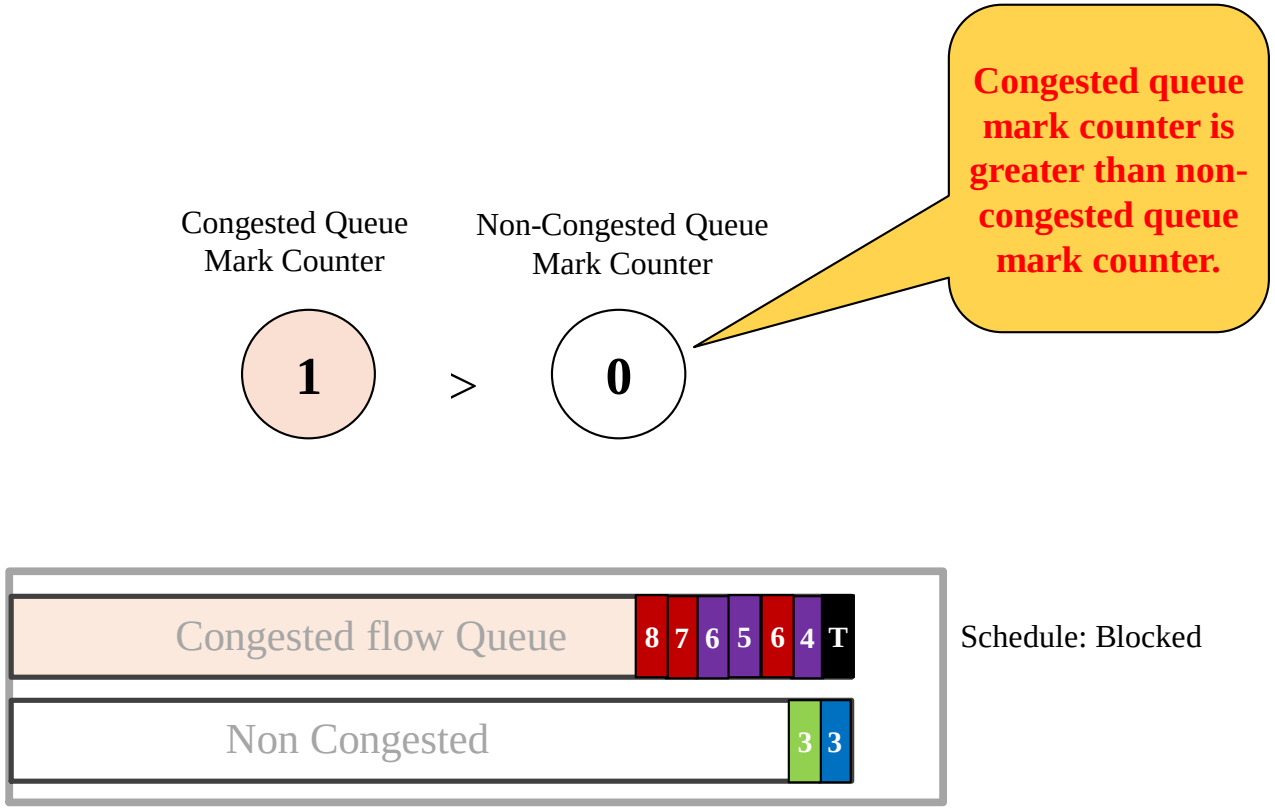

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
1
Congested flow Queue
5 6 4 T 5 4
Non Congested
3 3 T 2 2
Schedule: WRED
Decrement congested queue mark counter and schedule congested flow queue

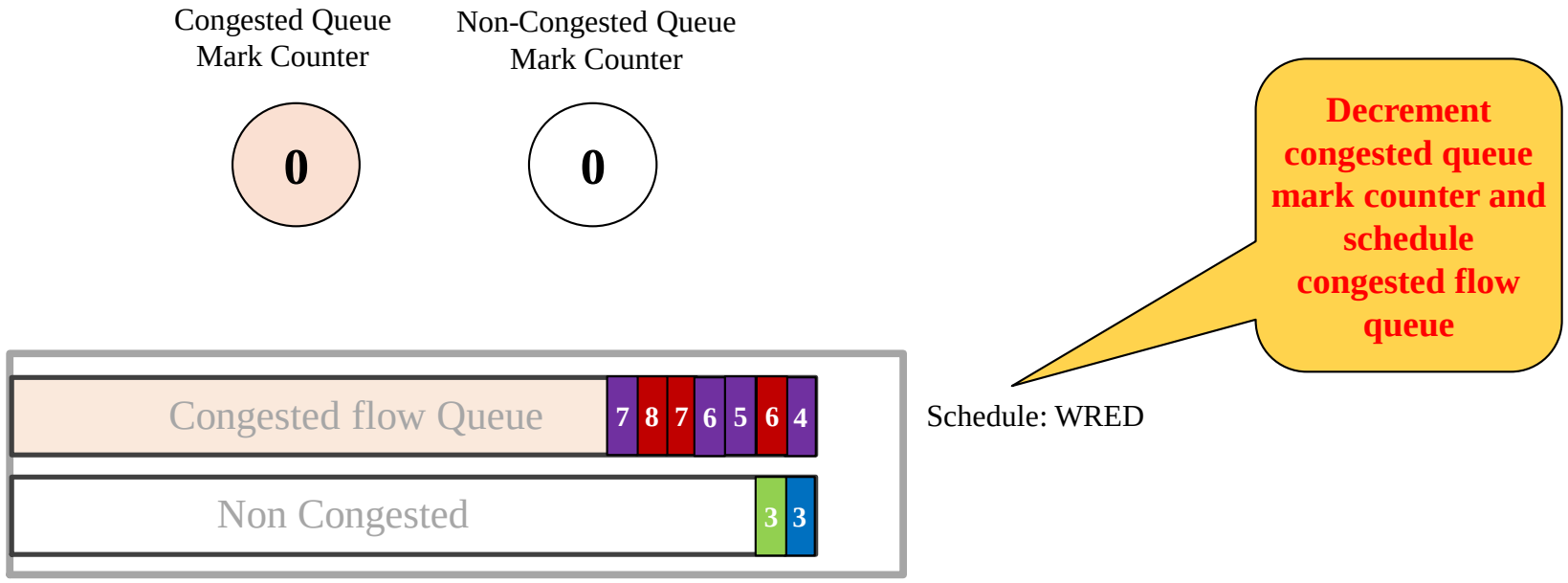

Congested Queue
Mark Counter
1
Non-Congested Queue
Mark Counter
1
Congested flow Queue
5 6 4 T 5
Non Congested
3 3 T 2 2
Schedule: WRED

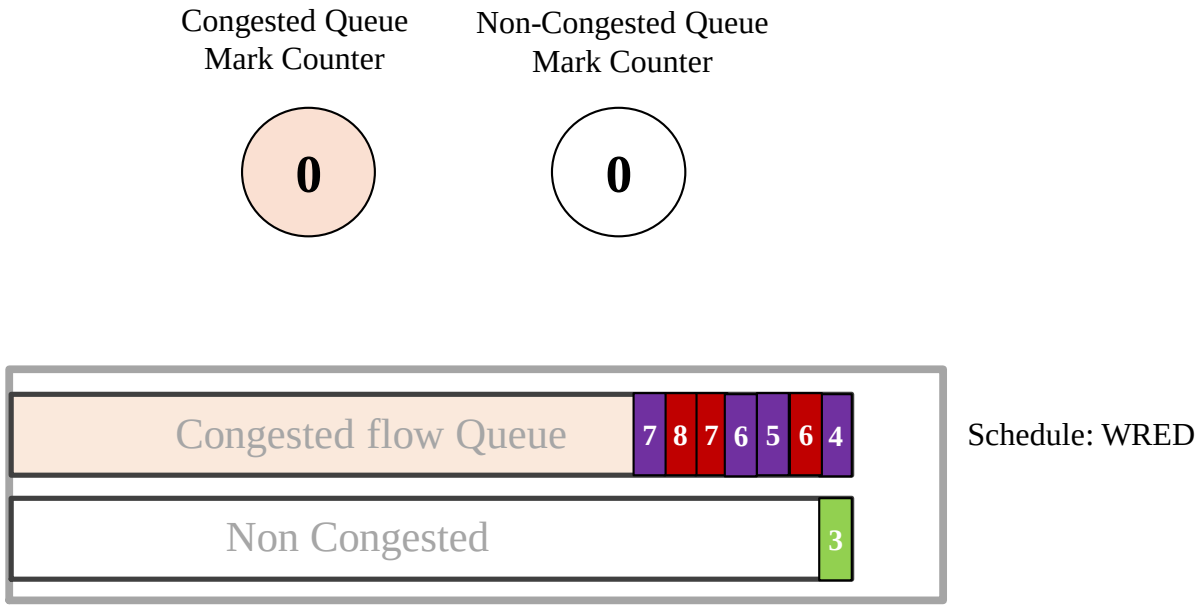

Congested Queue
Mark Counter
1
Non-Congested Queue
Mark Counter
1
Congested flow Queue
5
6
4
T
5
Non Congested
3
3
T
2
Schedule: WRED


Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Marker reaches head of congested flow queue: block scheduling
Congested flow Queue
6 5 6 4 T
Non Congested
3 3 T 2
Schedule: Blocked


Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
1
≤
1
Congested flow Queue
7
6
5
6
4
T
Non Congested
3
3
T
Schedule: Blocked

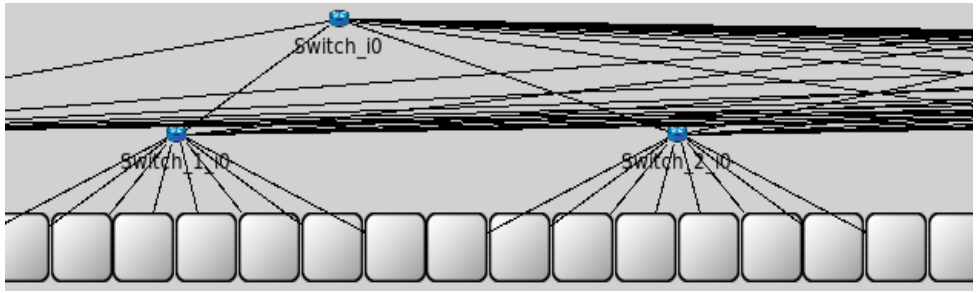

Congested Queue Mark Counter
Non-Congested Queue Mark Counter
1
>
0
Congested queue mark counter is greater than non-congested queue mark counter.
Congested flow Queue
8 7 6 5 6 4 T
Non Congested
3 3
Schedule: Blocked

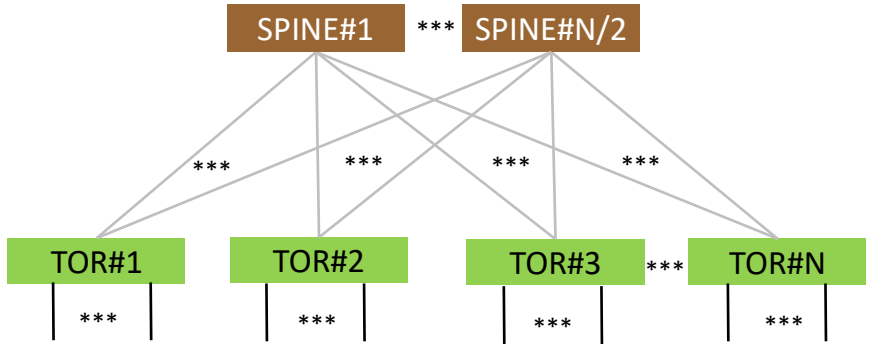

Congested Queue
Mark Counter
Non-Congested Queue
Mark Counter
0
0
Congested flow Queue
7 8 7 6 5 6 4
Non Congested
3 3
Schedule: WRED
Decrement congested queue mark counter and schedule congested flow queue


Congested Queue
Mark Counter
0
Non-Congested Queue
Mark Counter
0
Congested flow Queue
7
8
7
6
5
6
4
Non Congested
3
Schedule: WRED

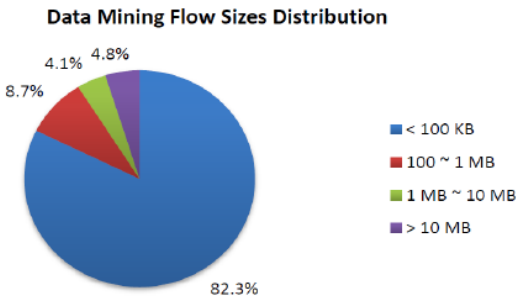
Simulation Highlights

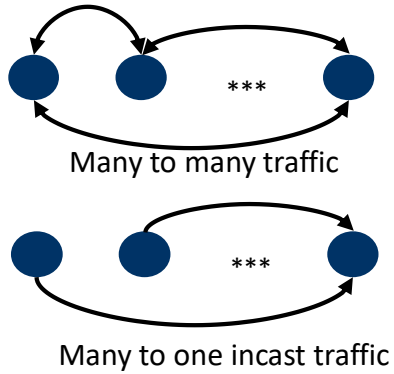
- Complete presentations on simulations are available on 802.1 public repository:
  - http://www.ieee802.org/1/files/public/docs2017/new-dcb-shen-congestion-isolation-simulation-1117-v00.pdf
  - http://www.ieee802.org/1/files/public/docs2018/new-dcb-shen-congestion-isolation-simulation-0118-v01.pdf
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





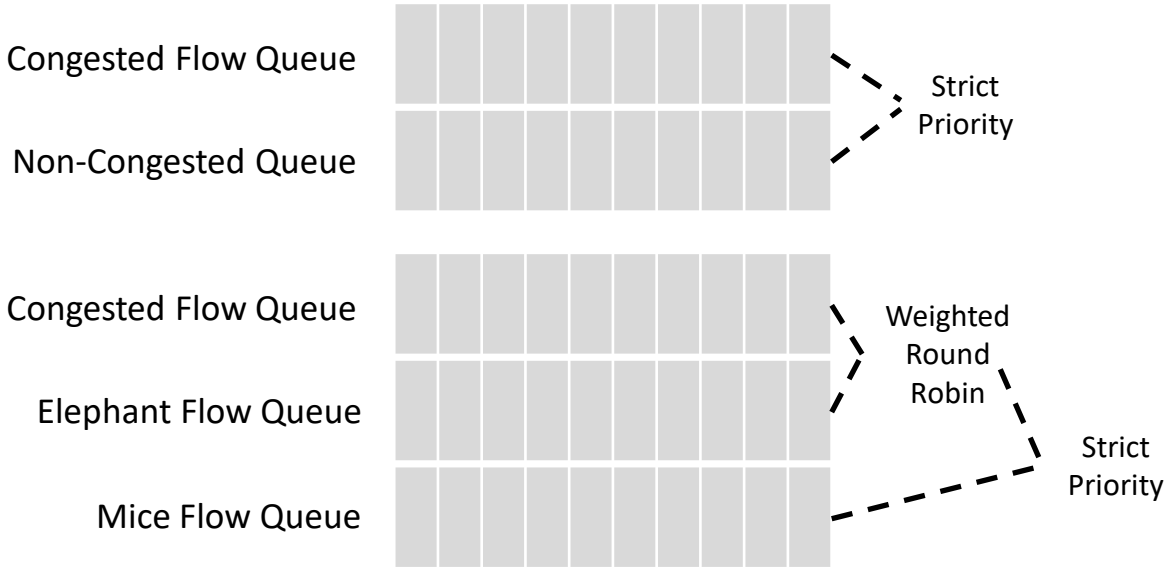
- 2 Tier CLOS: 1152 servers, 72 switches, 100GbE interface, 200ns of link latency (about 40 meters)
- Traffic Patterns:
  - Model data mining application with flow size distributions
  - 50 clusters of 21 servers for many to many traffic
  - 4 sets of 20:1 permanent many to one incast traffic





Queue Models Used

With Congestion Isolation (ECN + PFC + CI)

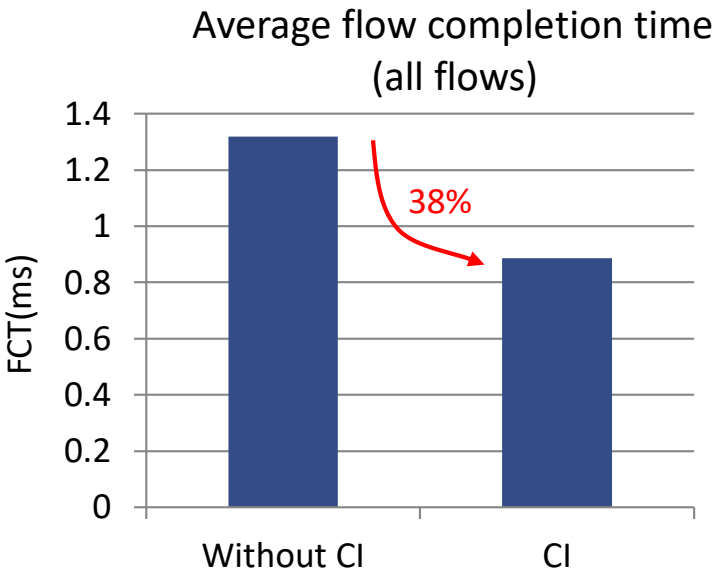


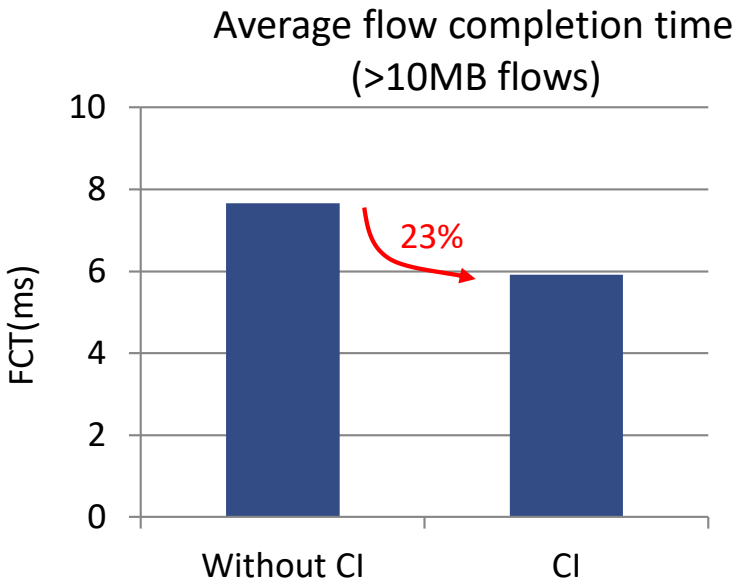
- Congested flows are dynamically isolated based on congestion.
- ECN is marked once a packet is isolated.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB;
  - CI: Low 10 KB, High 300 KB, Max Probability 1%.

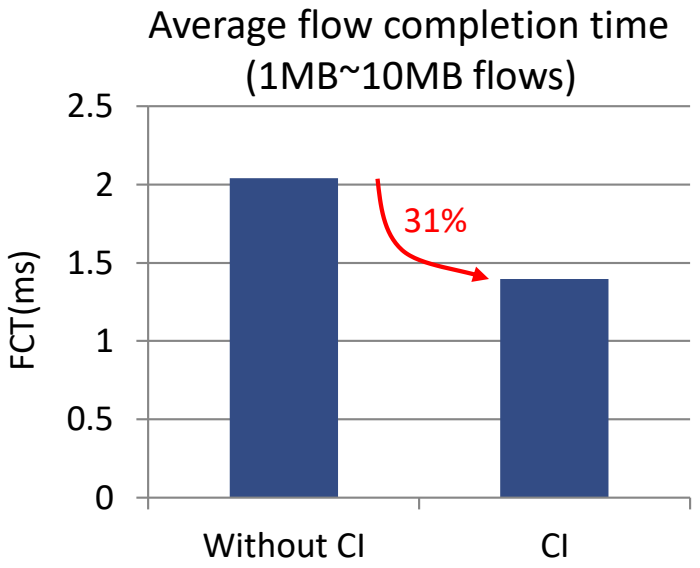
Without Congestion Isolation (ECN + PFC)

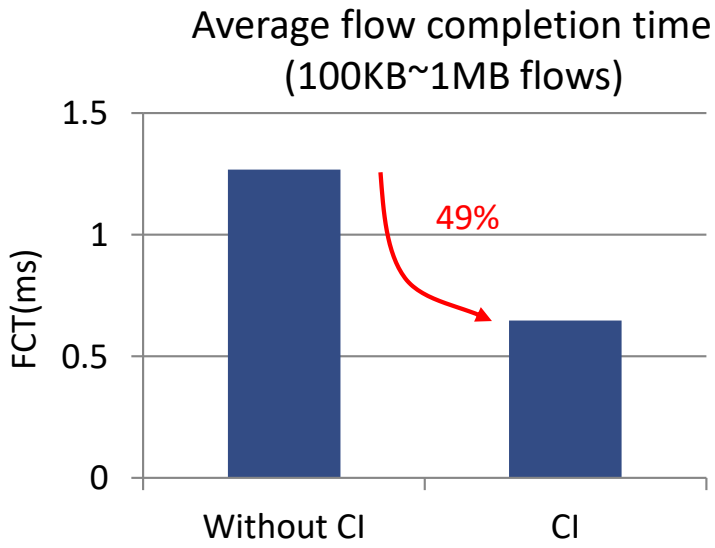
- Flows are mapped to one of the same queues by hash of destination IP.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB;
  - ECN: Low 10 KB, High 300 KB, Max Probability 1%.

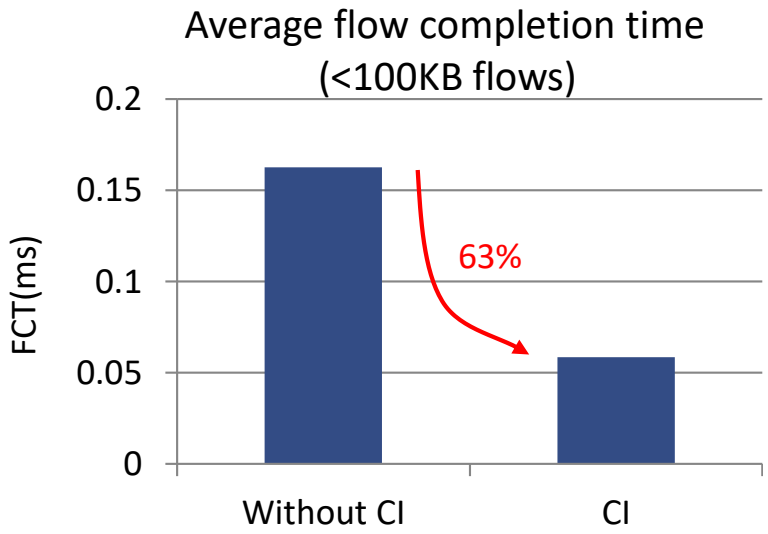
FCT Comparison – Lossless Scenario (with PFC)





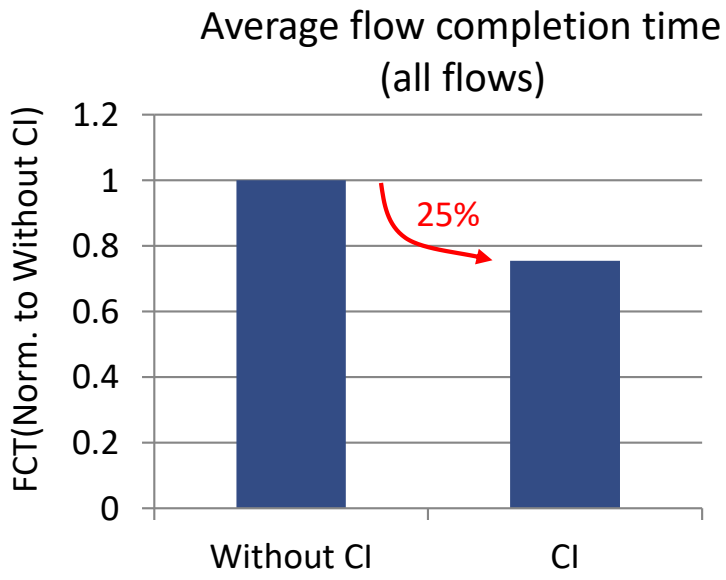


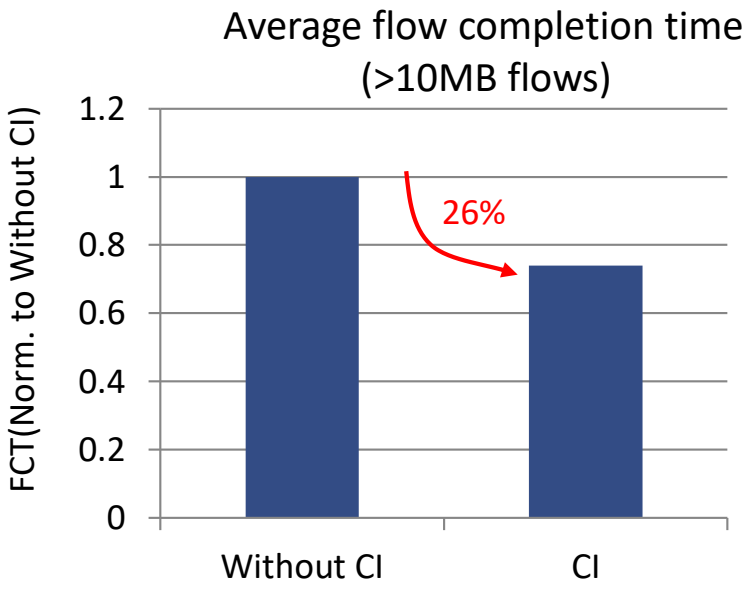


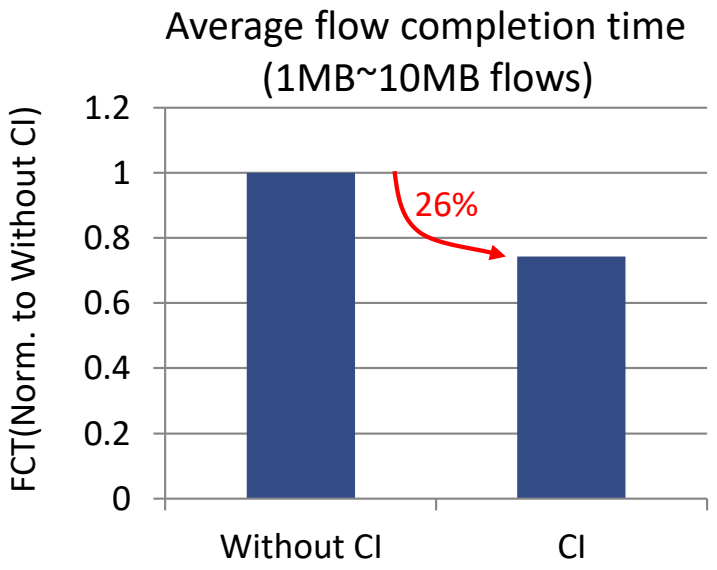


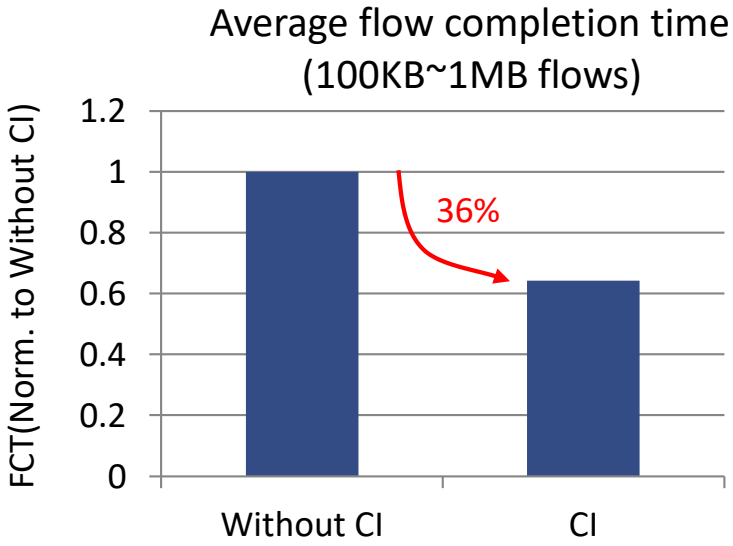
- The mice benefit the most.

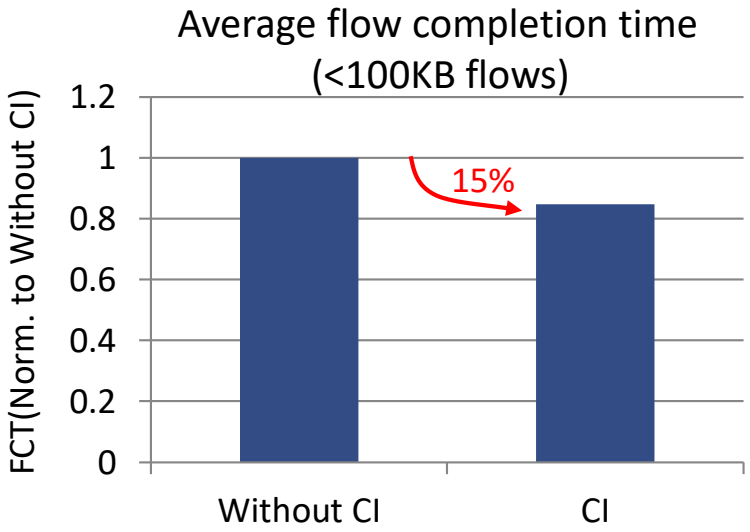
FTC With Mice/Elephant separation (3 Queue Model)





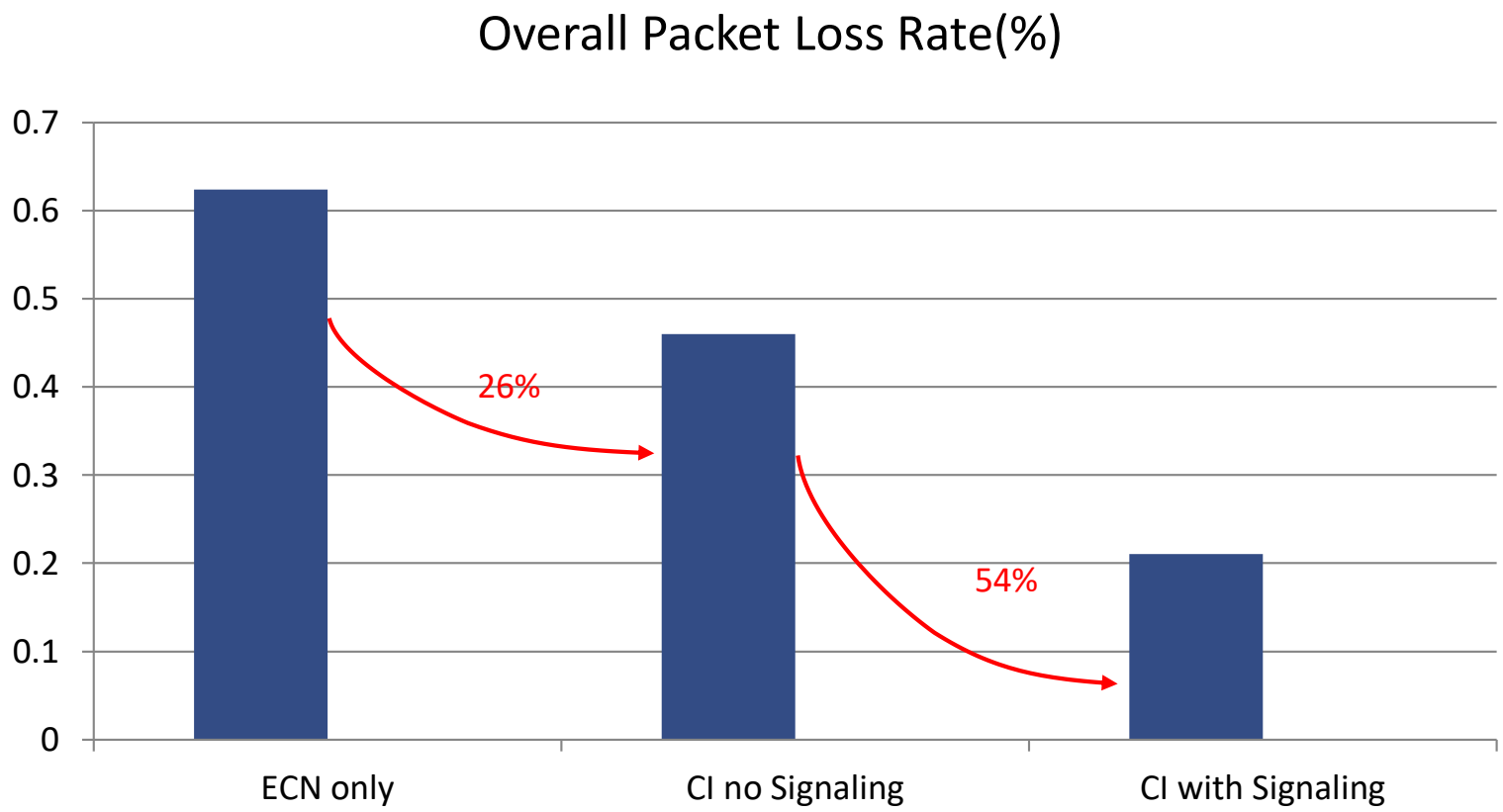






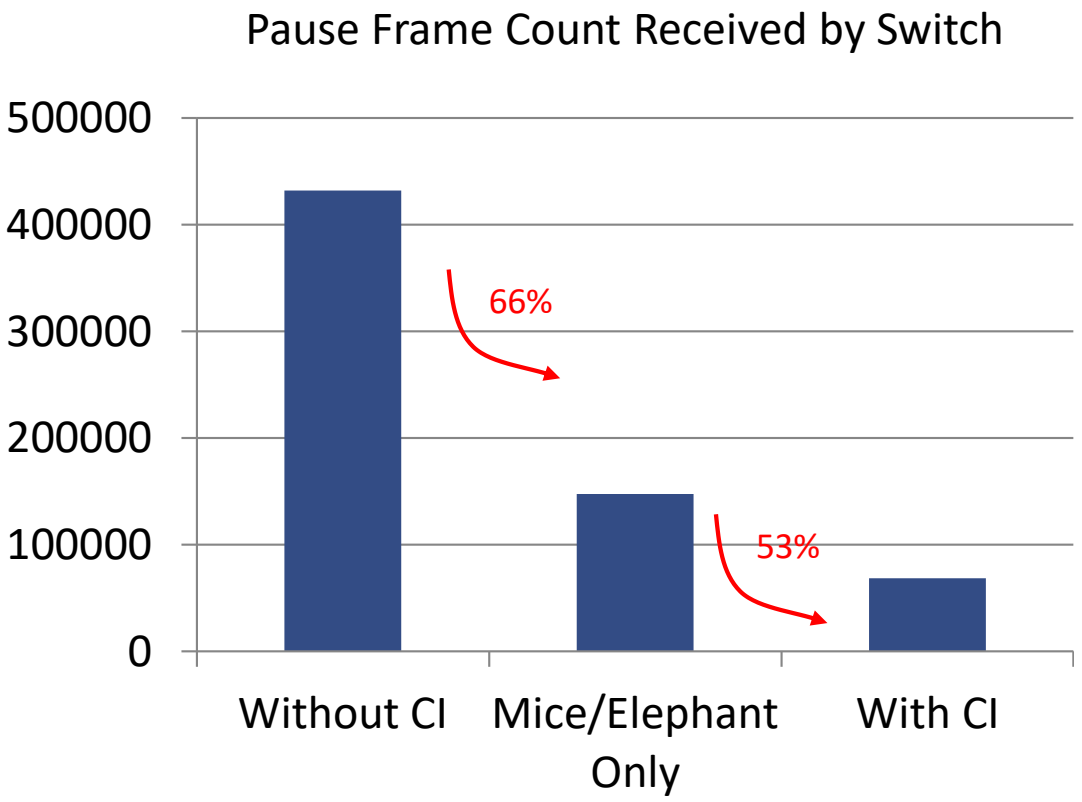
- With 3 queue model both “without CI” and “CI” have mice prioritization mechanism.
- The performance of the mice is not improved as much.

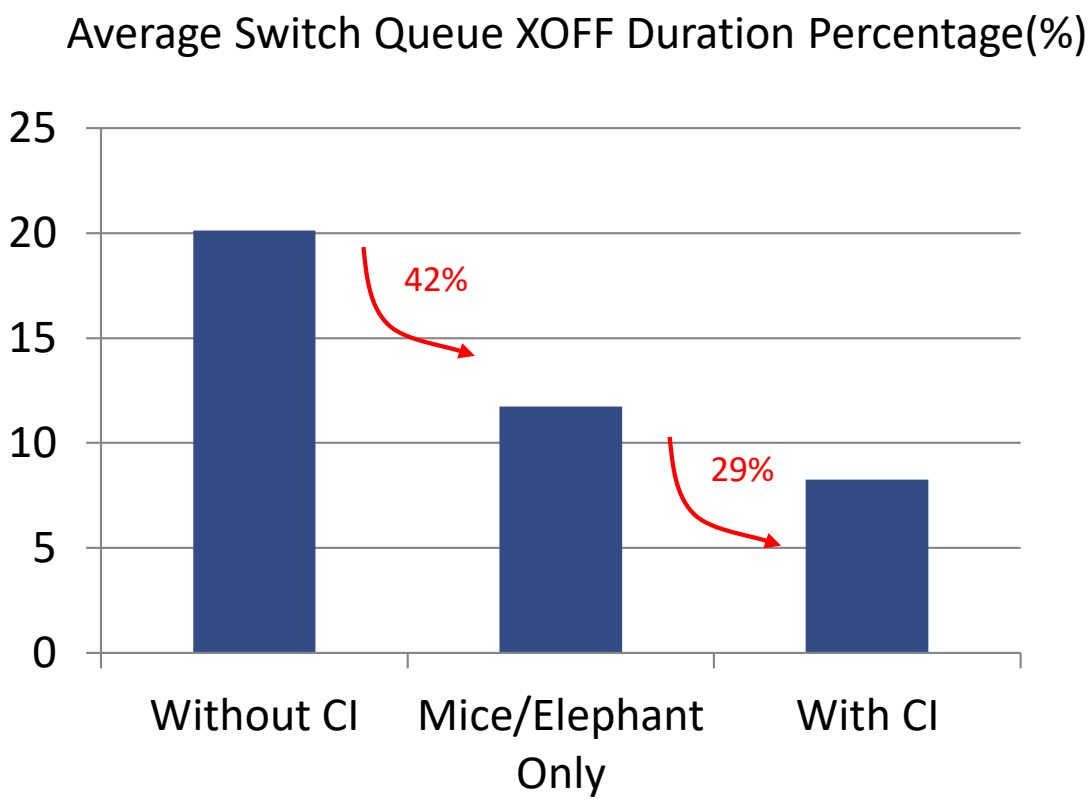
Lossy Scenario (No PFC)



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

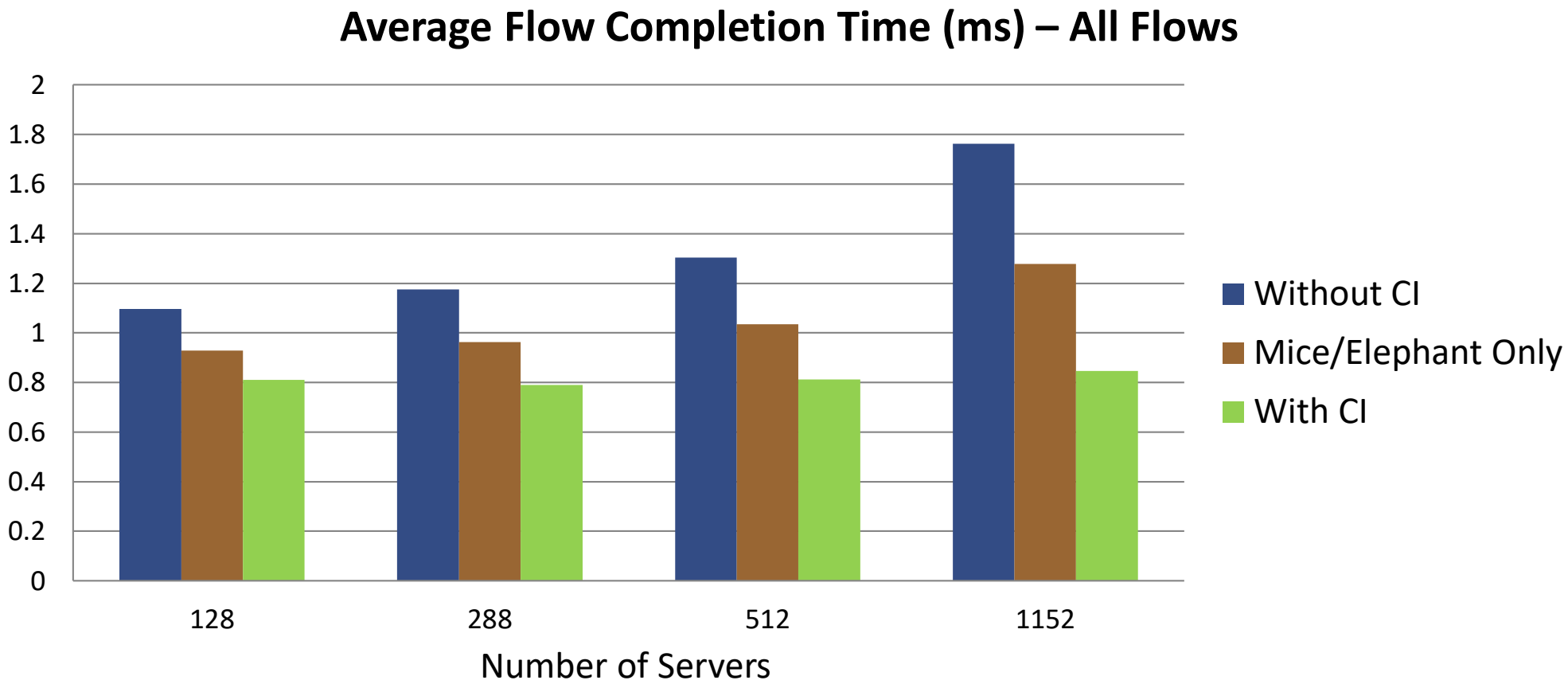
Lossless Scenario - Reducing the Impact of PFC





- 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.

Issues raised by cz-tabatabaee-CIAnalysis-0318-v01.pdf

- Congested flow detection
  - Short-lived flows are not stopped effectively
  - Unfairness across flows
  - Poor tail latency and FCT
  - Lag in detecting flows that cause congestion
- Congested to uncongested transition using inactivity timeout is not sufficient
  - Packet reordering due to interaction with PFC
  - HOL blocking for flows that their rate is controlled by an ECN based congestion management + traffic pacer
- Increased buffer requirements
  - Increases burst absorption buffer requirement
- Congested flow packets can be in noncongested queue when PFC is triggered
  - Increases headroom buffer requirement

Response: Congested Flow Detection

CI is not expecting to specify how to detect congested flows. CI uses existing tried-and-true techniques for detecting congested flows. Implementations for ECN marking and/or existing Qau Congestion Notification can be used. Other approaches are possible and within the scope of the project.

- Claim: Short-lived flows are not stopped effectively
  - Local isolation occurs immediately after detection, regardless of flow-size. Local isolation reacts quicker than ECN. Upstream isolation occurs after reception of CIM from downstream switch and is in place well before congestion effectively propagates to upstream switch
- Unfairness across flows
  - Unsubstantiated claim. Results show that FCT improves significantly. Elephant flows are the primary cause of congestion and are able to adjust rate using end-to-end congestion control
- Poor tail latency and FCT
  - Unsubstantiated claim. Results show that FCT improves significantly. Tail latency needs measurement.
- Lag in detecting flows that cause congestion
  - Local isolation occurs immediately after detection, reacting much faster than ECN.

Response: Congested to uncongested transition

Agreed that a simply inactivity timer is not sufficient, however the impact of remaining in the congested queue is not detrimental. Once congestion has subsided, the scheduling of the congested queue is round-robin and the flow is not penalized.

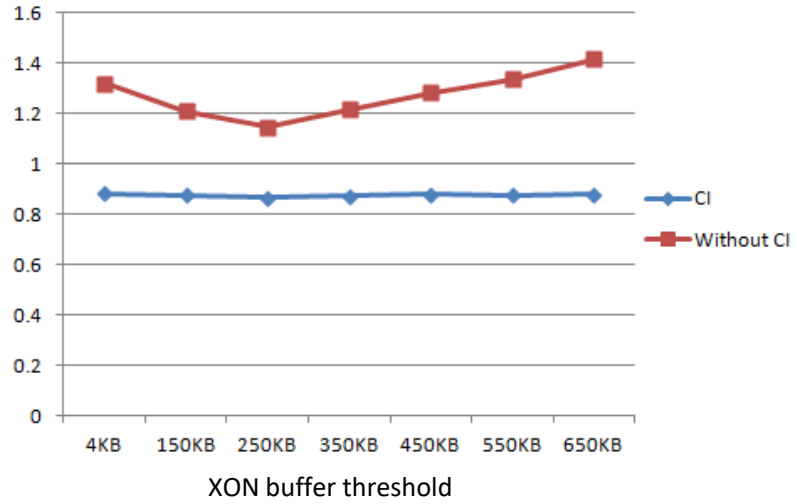
- Claim: Packet reordering due to interaction with PFC
  - This assumes PFC-XOFF is enabled longer than the inactivity timeout and an active flow is removed
  - Once the congested flow queue is empty, the entire congested flow table can be flushed.
- Claim: HOL blocking for flows that their rate is controlled by an ECN based congestion management + traffic pacer
  - More clarity needed. I do not understand the assertion.

Response: Increased buffer requirements

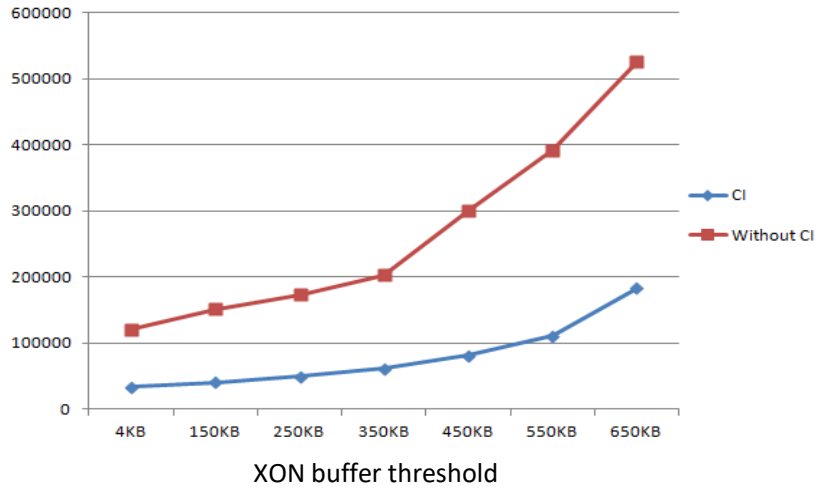
Further simulation is desired, but initial results indicate that overall memory utilization is lower when CI is enabled.

- Claim: Increases burst absorption buffer requirement
  - Simulation results from cz-shen-congestion-isolation-simulation-0318-v01.pdf show that varying the burst absorption buffer has little impact on performance when CI is enabled.

Average flow completion time(ms)
(all flows)



Pause Frame Count Received by Servers



Response: Congested flow packets can be in noncongested queue when PFC is triggered

CI is using two traffic classes. This situation is only an issue for lossless mode. PFC sent on the congested queue will not stop the non-congested queue if congested packets are buffered in that class upstream. Additional thresholds are required to avoid packet loss.

- Claim: Increases headroom buffer requirement
  - True, but in data center environments, the amount of headroom required for 100M links is not significant.

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