Re-architecting Congestion Management in Lossless Ethernet

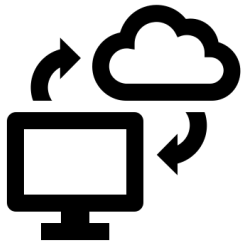
Wenxue Cheng, Kun Qian, Wanchun Jiang(CSU), Tong Zhang, Fengyuan Ren

NNS group @ Department of Computer Science and Technology, Tsinghua University

Data Center Networks

Small RTT

$< 100\mu s$



High Bandwidth

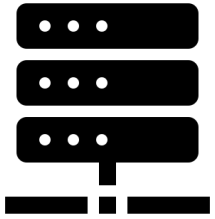
Packets Loss ↑

Shallow Buffer

< 30 MB for ToR

Large Scales

> 10000 machi

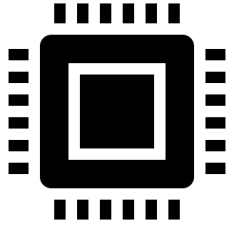


Short Messages

Performance ↓

Special Protocols

RDMA



Data Center Networks

Small RTT

$< 100\mu s$

High Bandwidth

10/40~100/400 Gbps

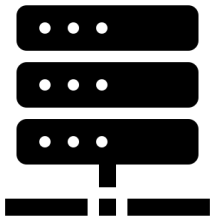
Shallow Buffer

< 30 MB for ToR

Lossless Ethernet

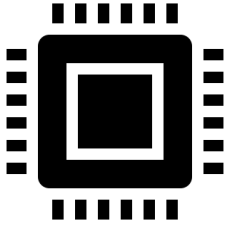
Large

> 10000 mac...

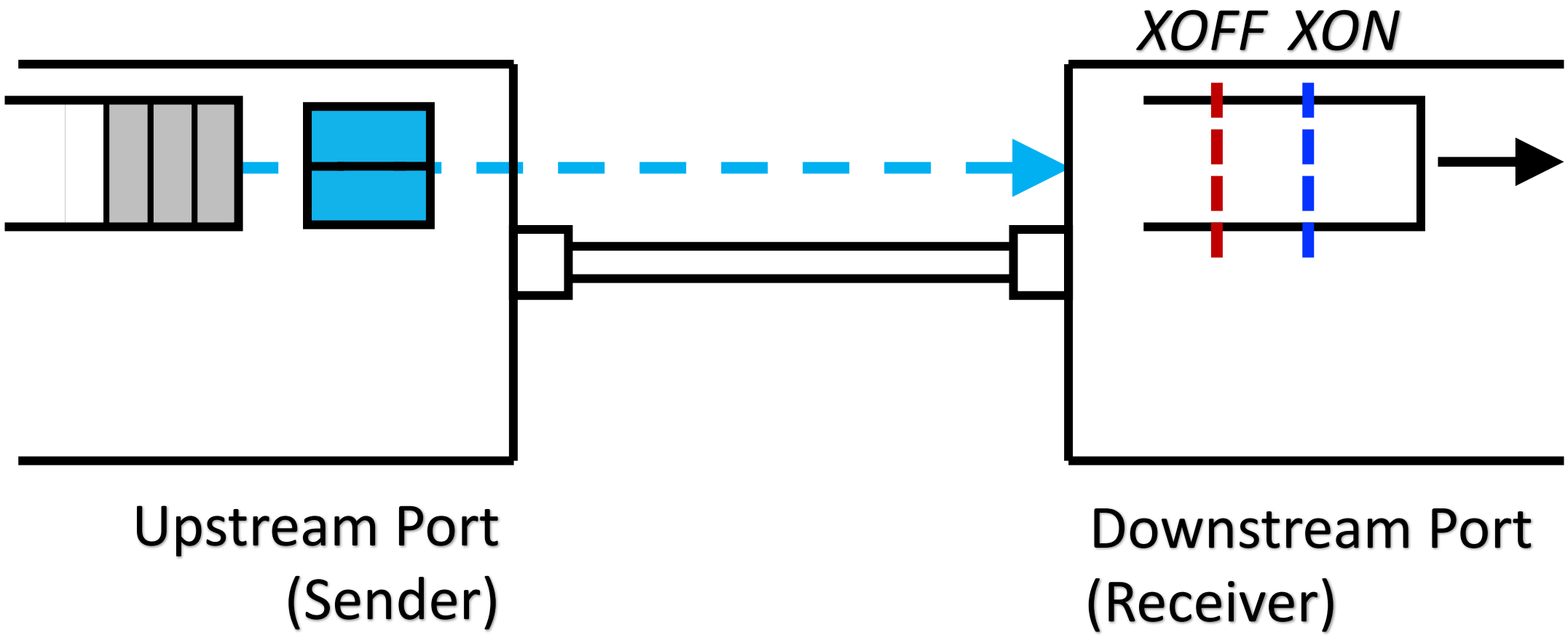


...rotocols

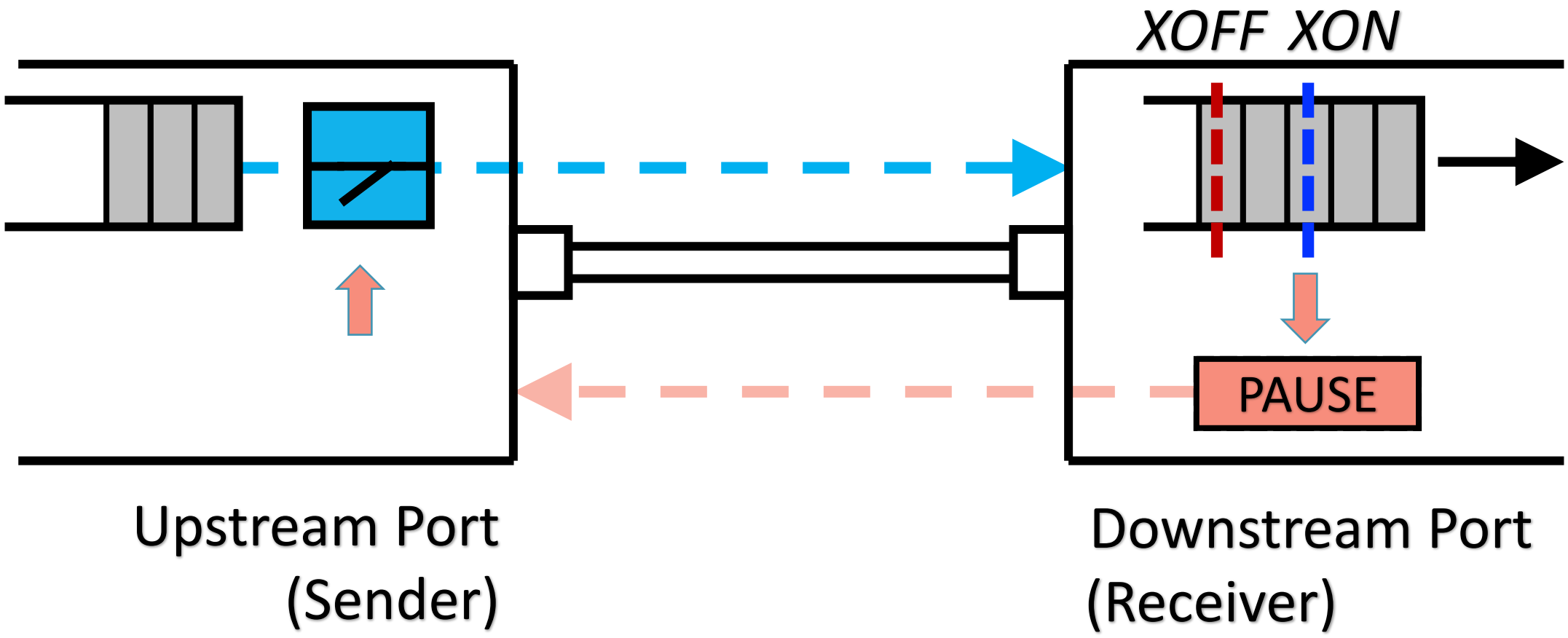
RDMA



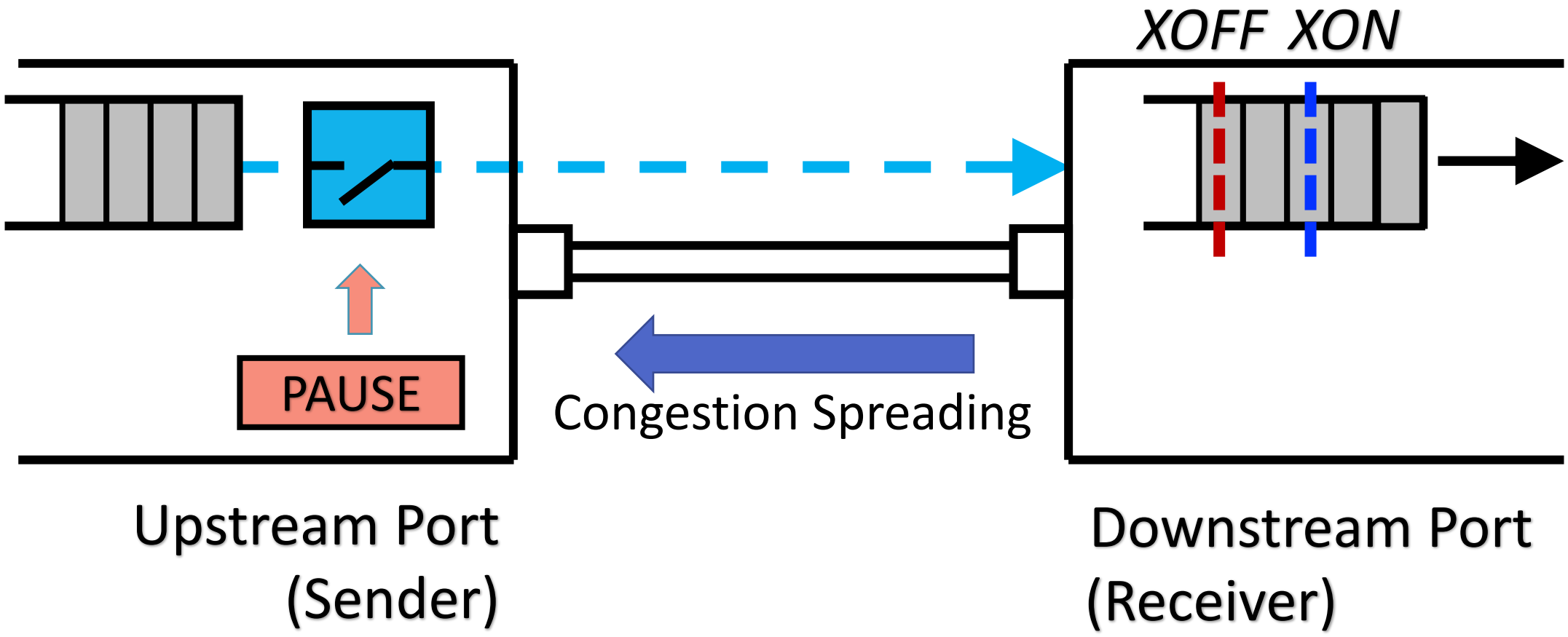
Priority-based Flow Control (PFC)



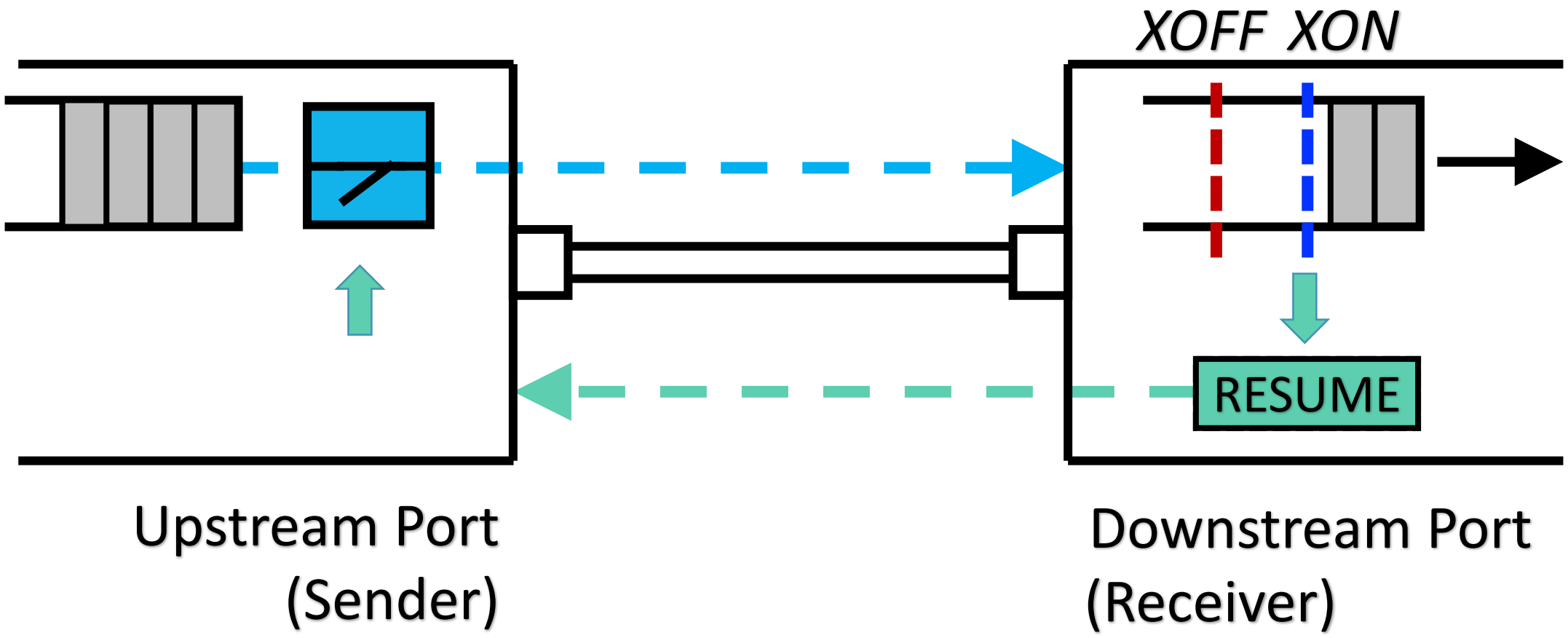
Priority-based Flow Control (PFC)



Priority-based Flow Control (PFC)

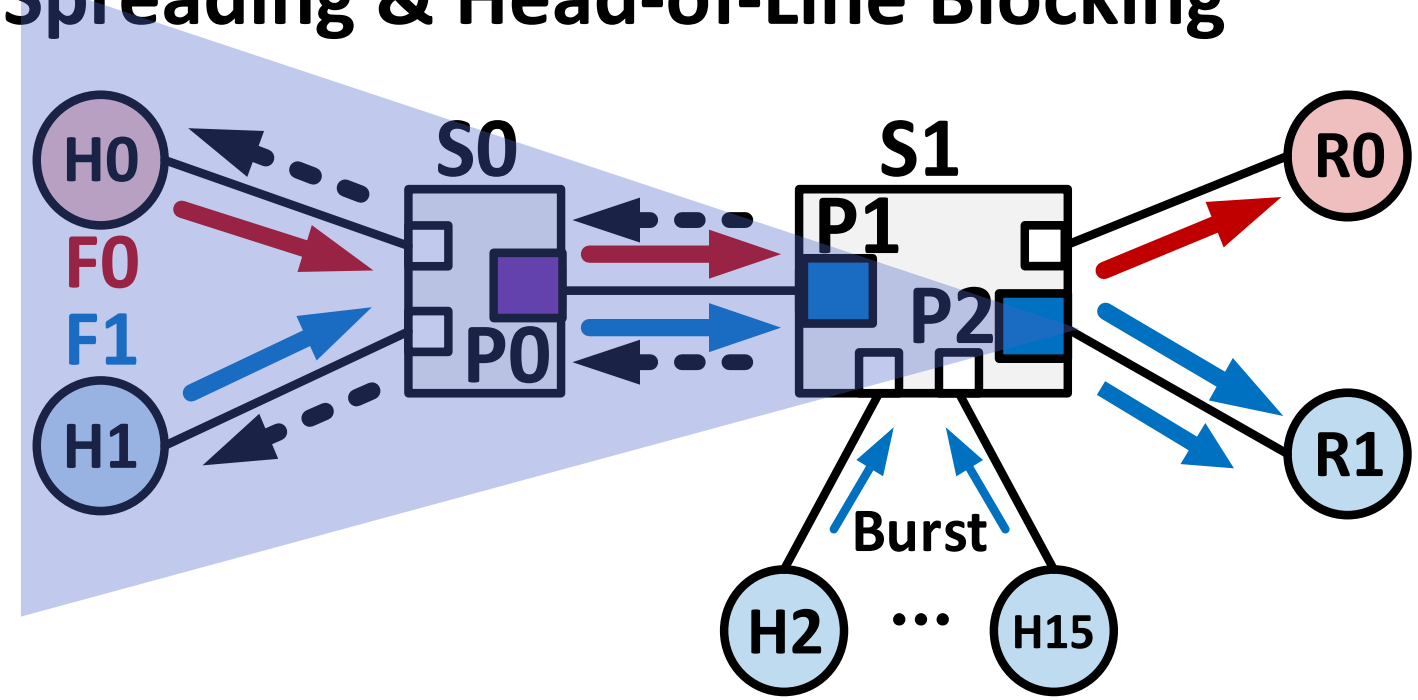


Priority-based Flow Control (PFC)



PFC Issues

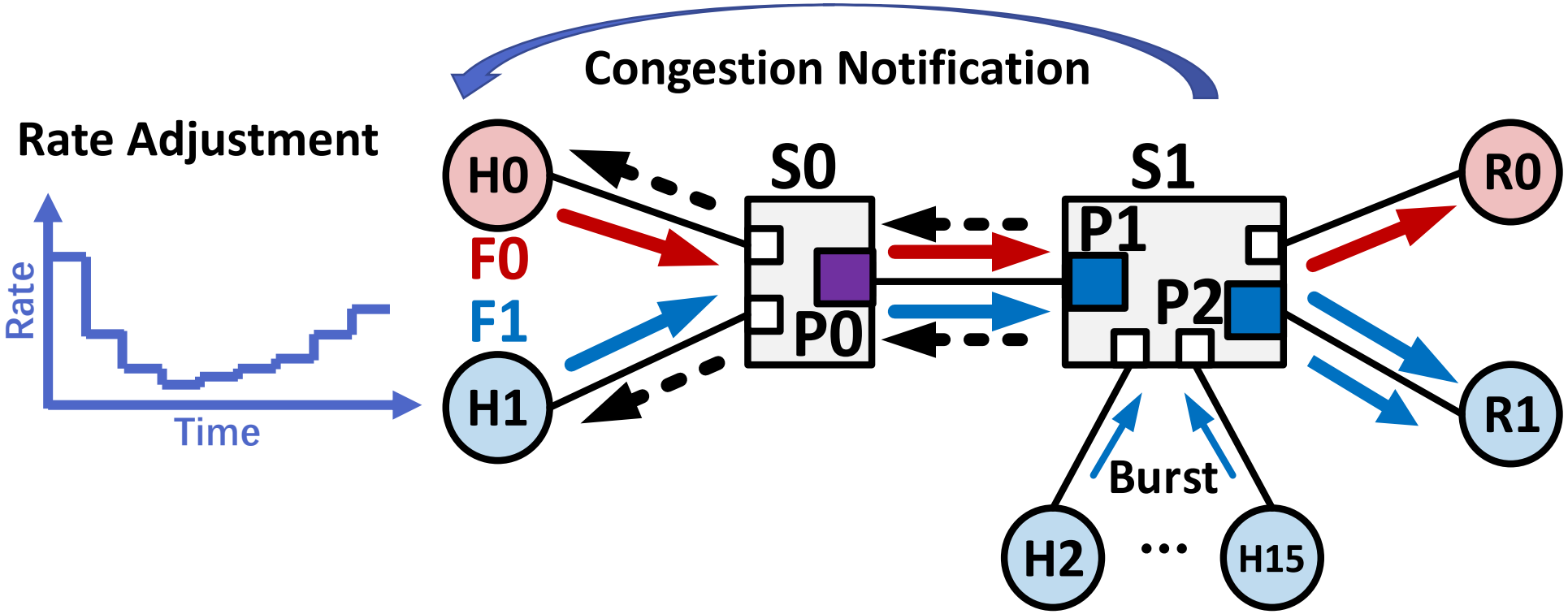
Congestion Spreading & Head-of-Line Blocking



Congestion tree from P2 to H0 and H1.

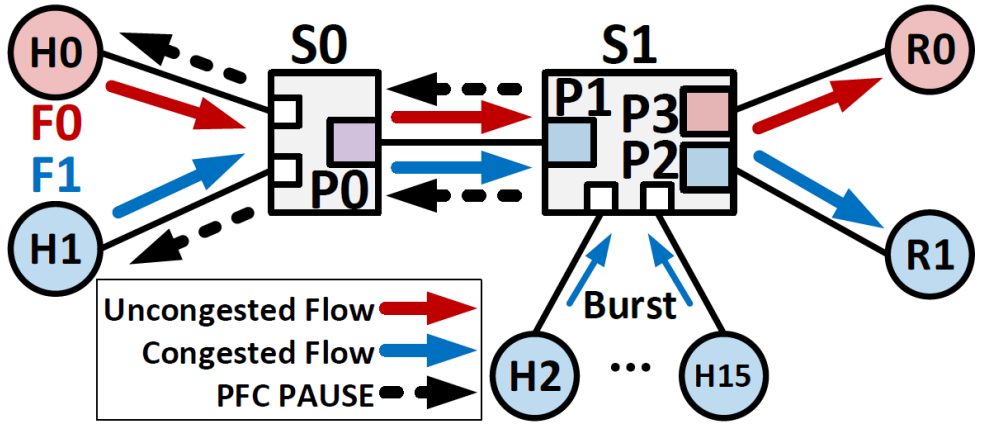
F0 is a victim flow.

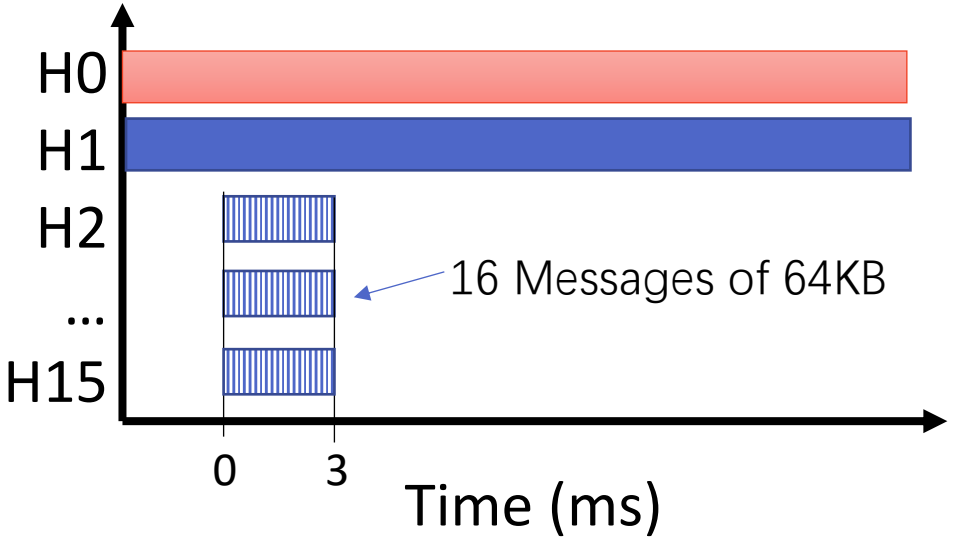
Congestion Control Schemes

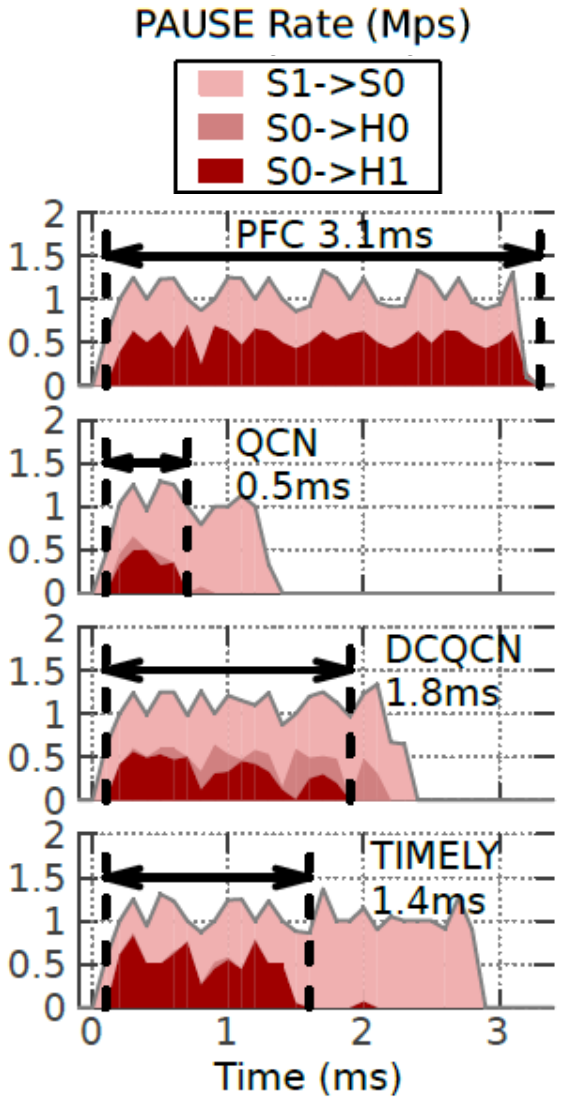


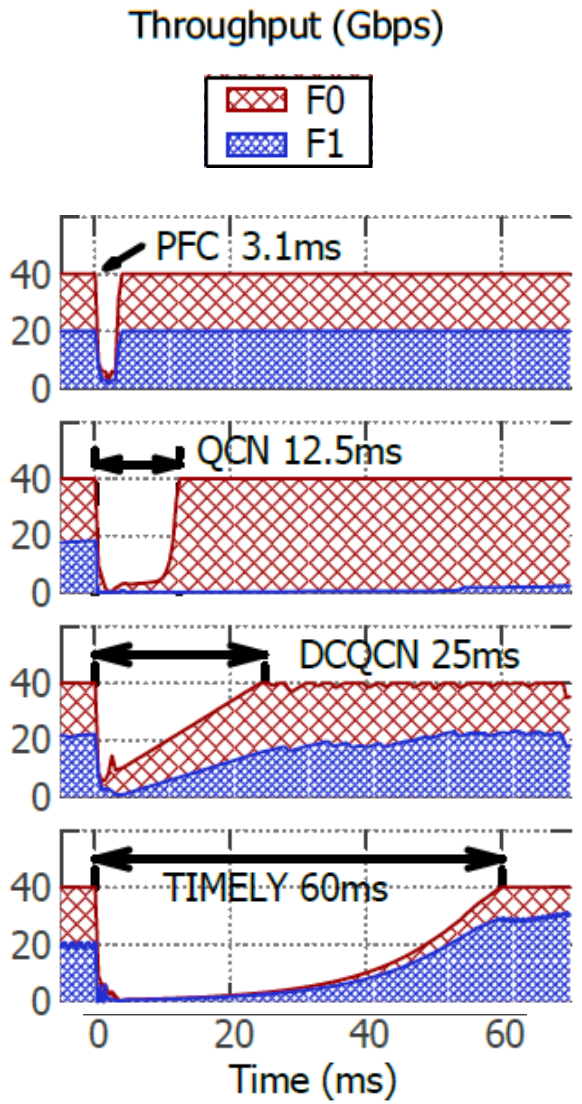
Congestion control schemes are needed
e.g. QCN[IEEE 802.1], DCQCN[RoCEv2] and TIMELY[SIGCOMM 2015].

Experimental Observation

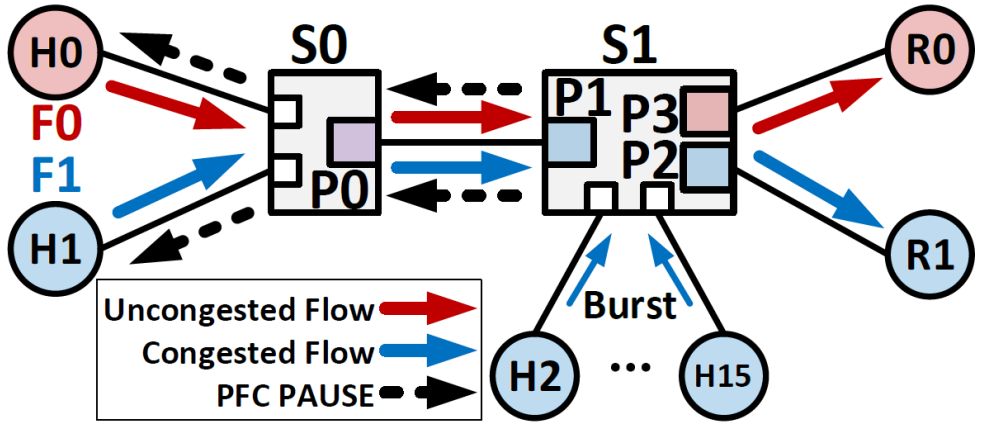






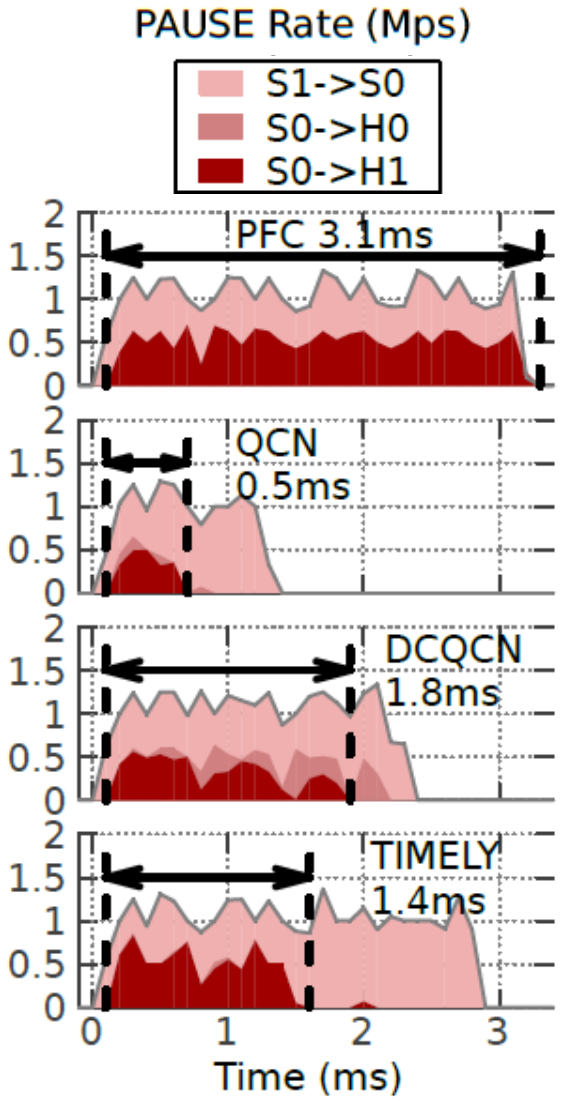


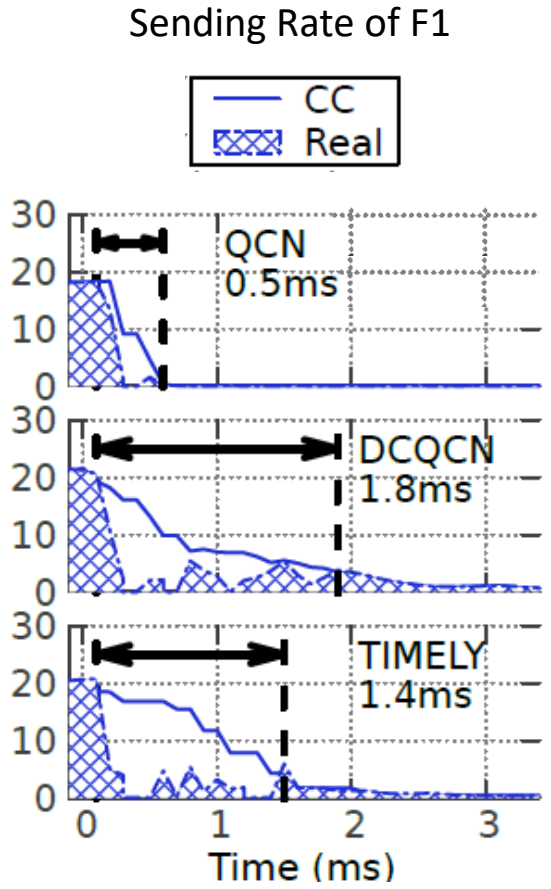
Experimental Observation



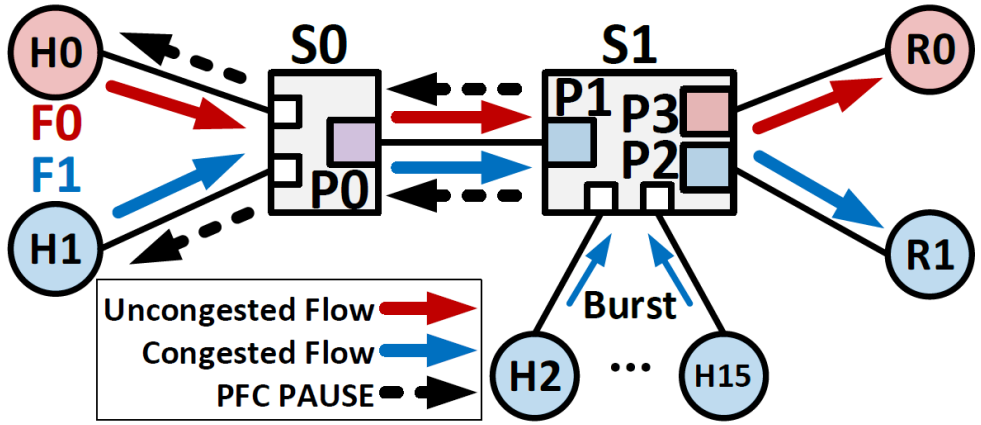
(1) Congestion spreading still exists.

Evolution-based rate decrease is slower than PFC's effect.





Experimental Observation

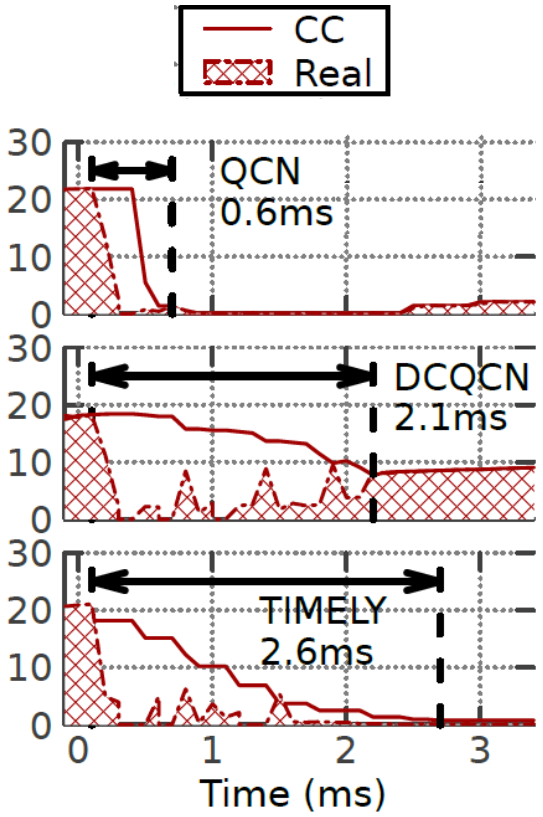


(1) Congestion spreading still exists.

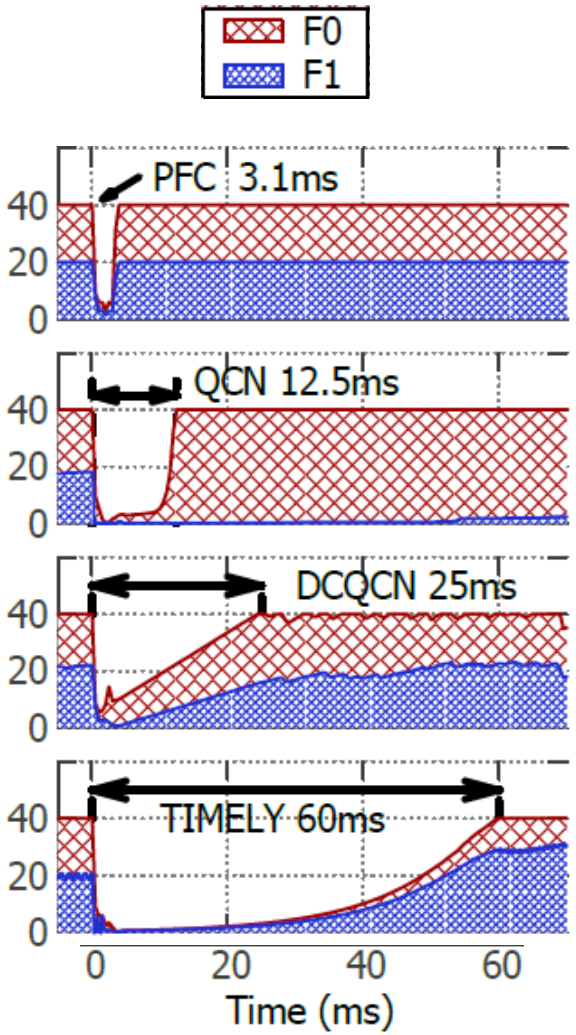
(2) F0 is also victimized by CC.

PFC infects congestion detection of congestion control schemes.

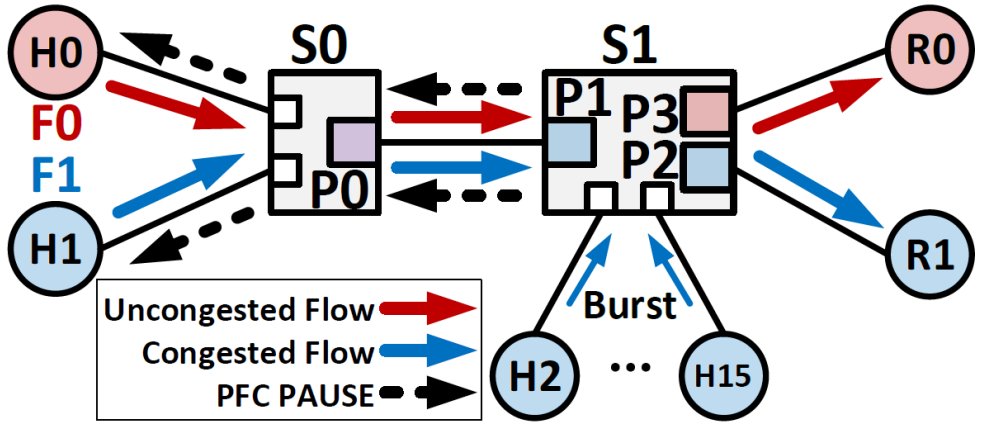
Sending Rate of F0 (Gbps)



Throughput (Gbps)

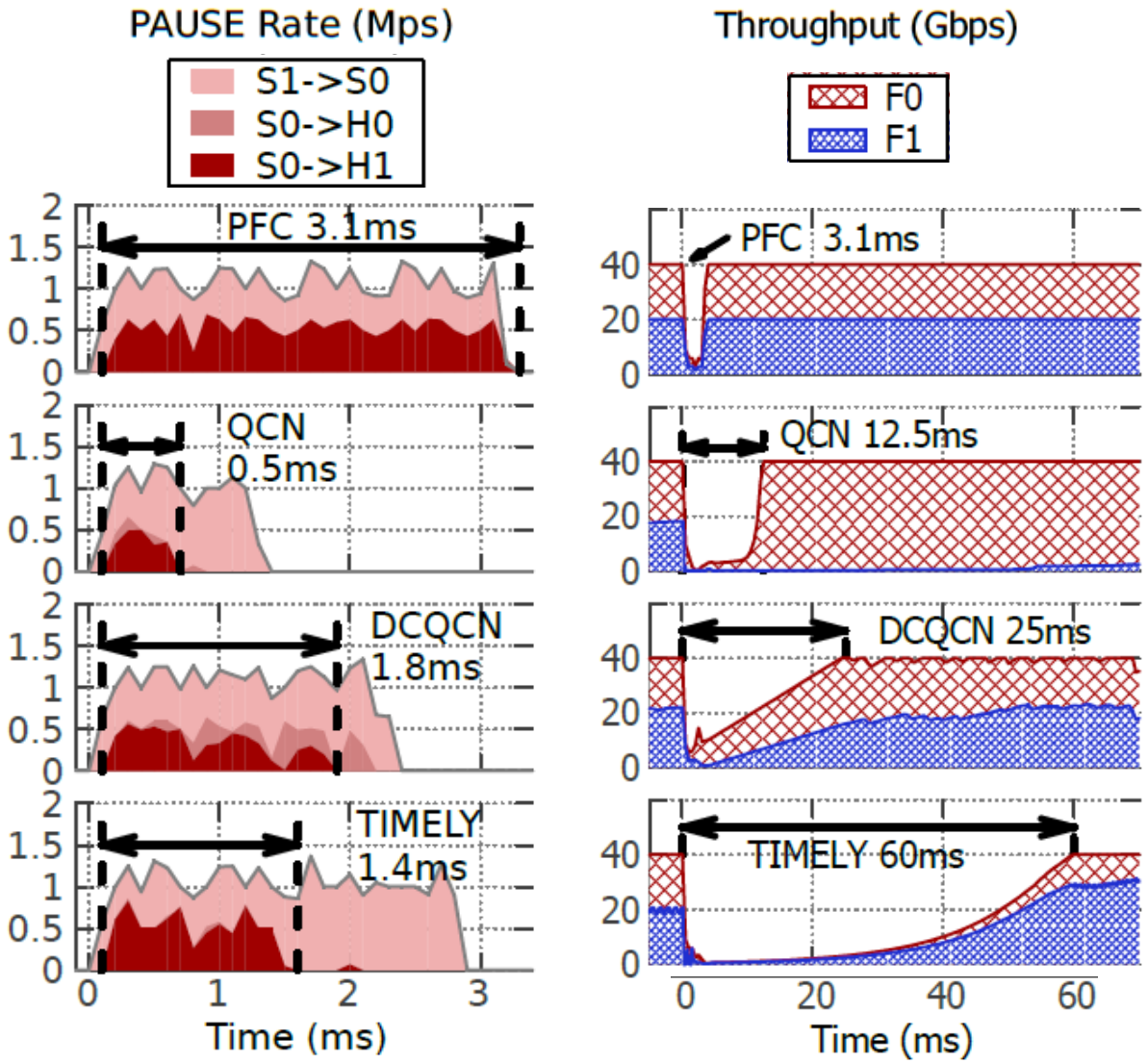


Experimental Observation



(1) Congestion spreading still exists.
(2) F0 is also victimized by CC.
(3) Rate recovery is inadaptable to dynamic network conditions.

Liner rate increase method and tuning parameters.



Basic Idea

Re-architecting Congestion Management

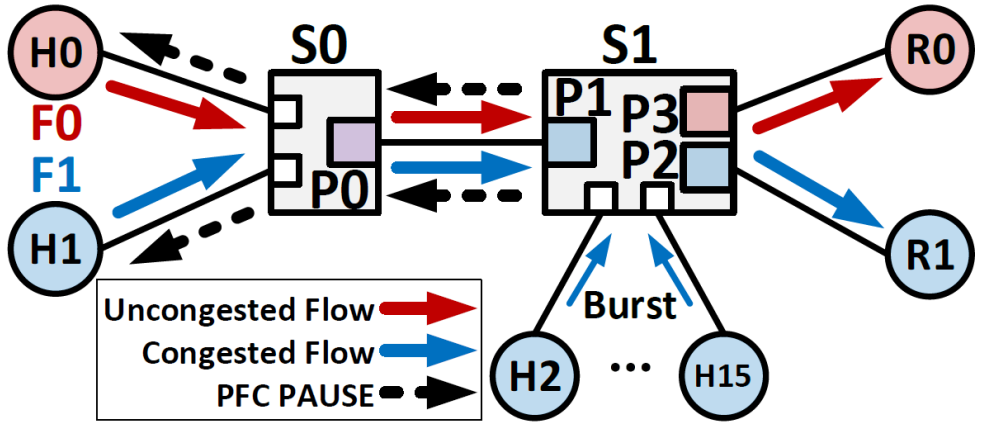
Congestion Detection

- Congestion Flows ⟷ Victim Flows

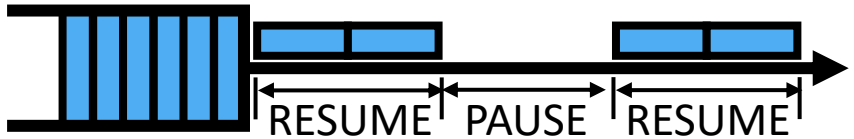
Rate Adjustment

- Fast Rate Decrease
- Automatic Rate Increase

Congestion Detection

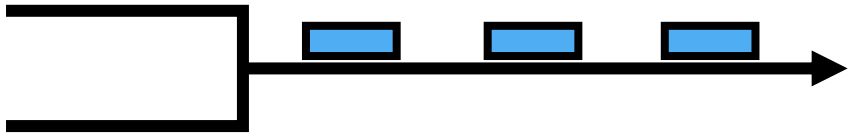


Quasi-Congestion (P0)



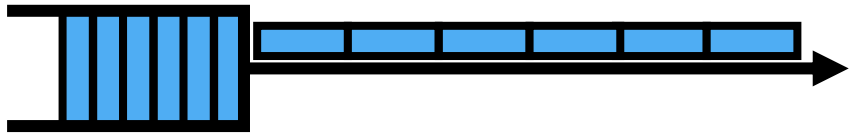
$$\sum R \; ? \; C$$

Non-Congestion (P3)



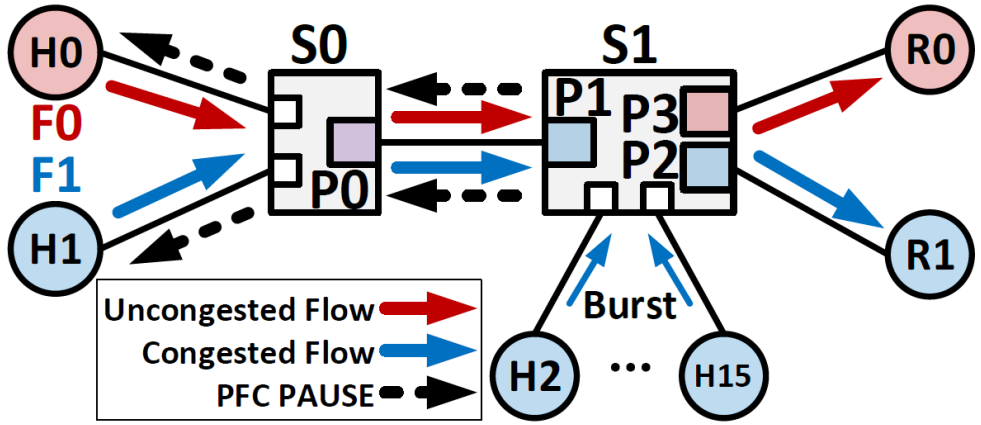
$$\sum R < C$$

Real-Congestion (P2)



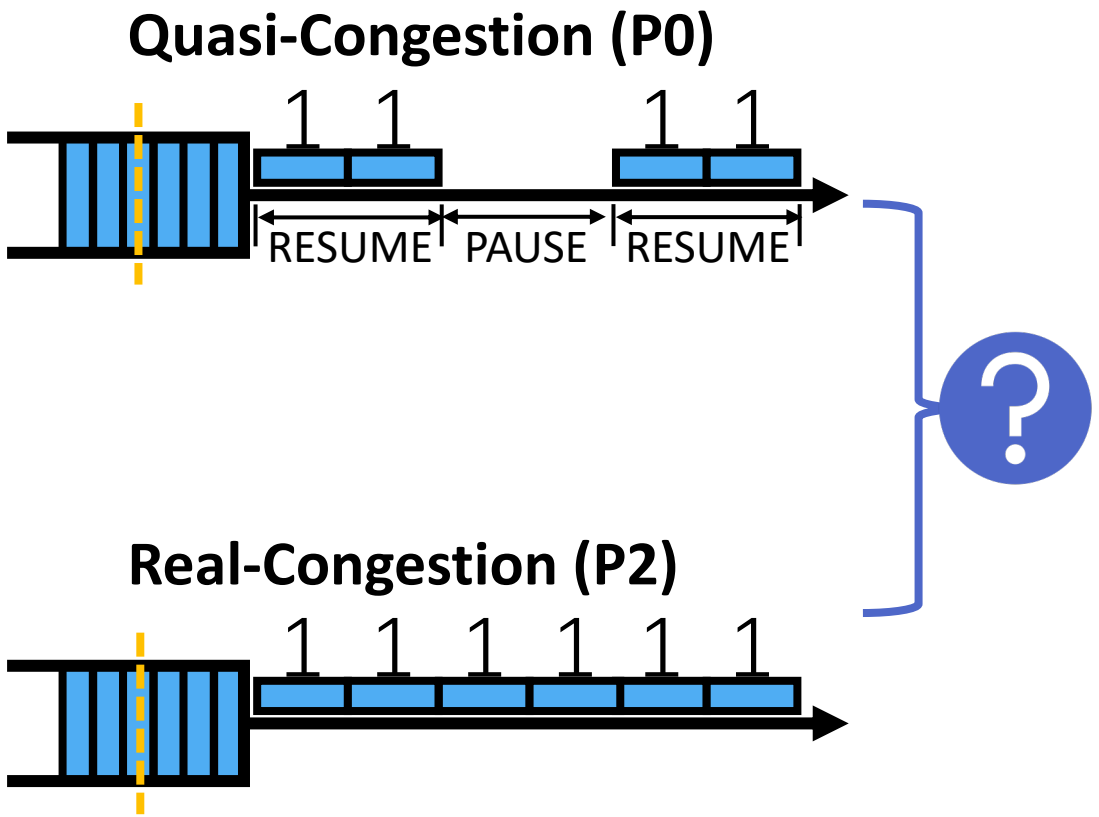
$$\sum R > C$$

Congestion Detection

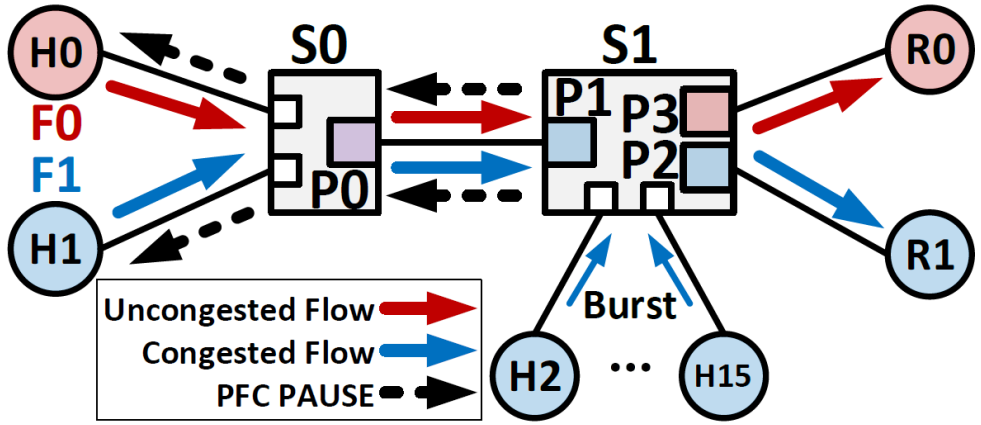


Explicit Congestion Notification (ECN)

- Only based on queue length
- Fail to distinguish quasi-congestion and real-congestion

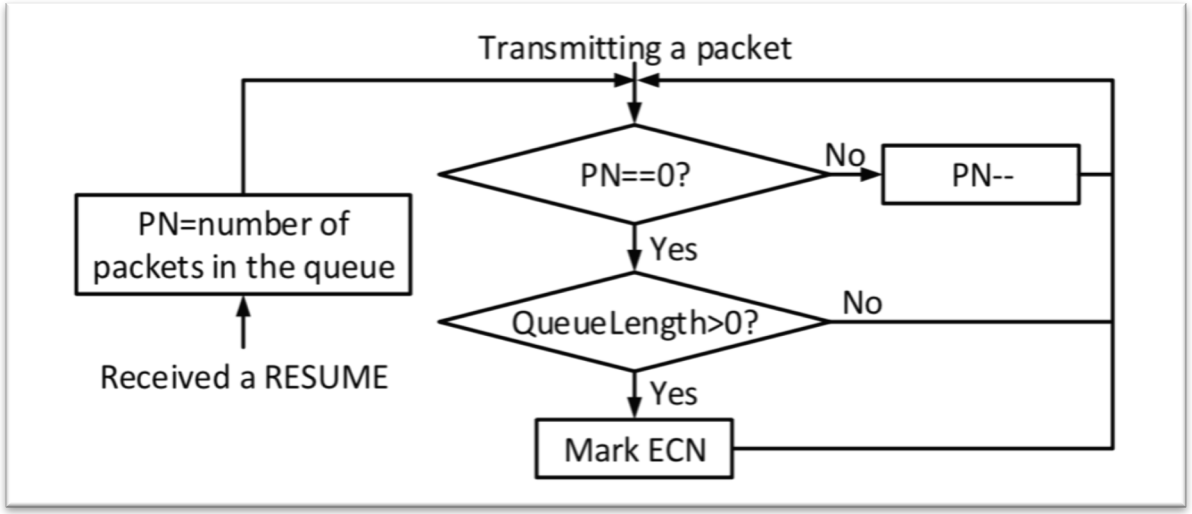


Congestion Detection

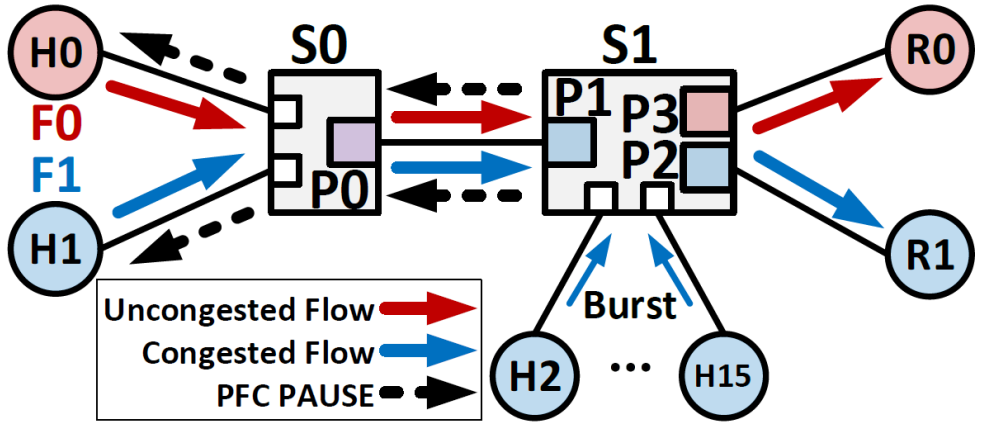


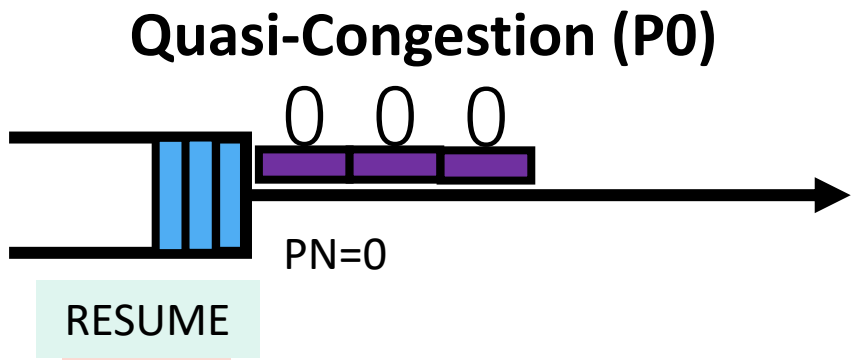
Non-Paused ECN (NP-ECN)

- Don't change ECN for packets that has been paused
- Counter PN: number of packets that has been paused



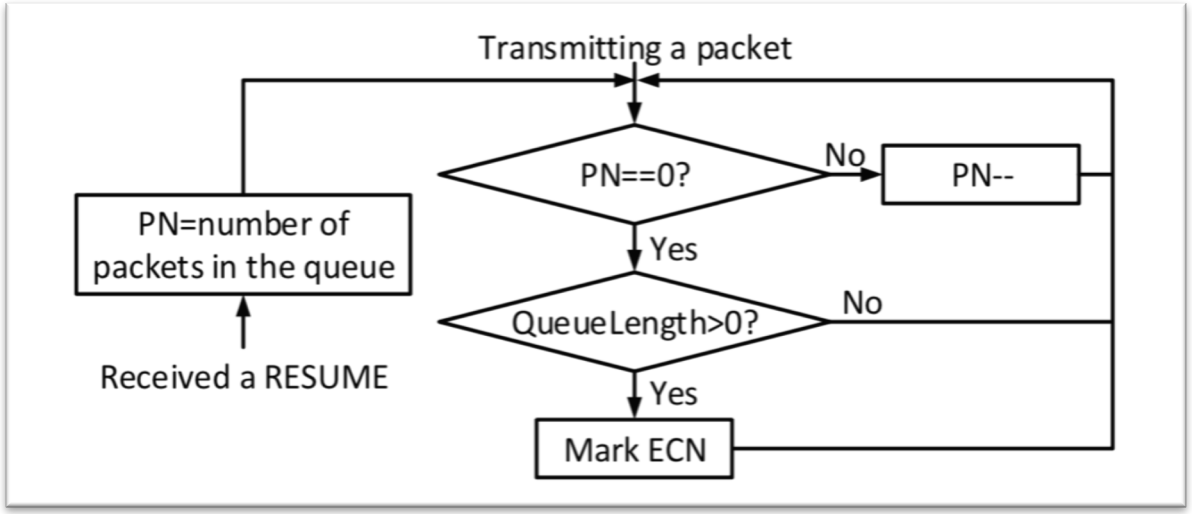
Congestion Detection



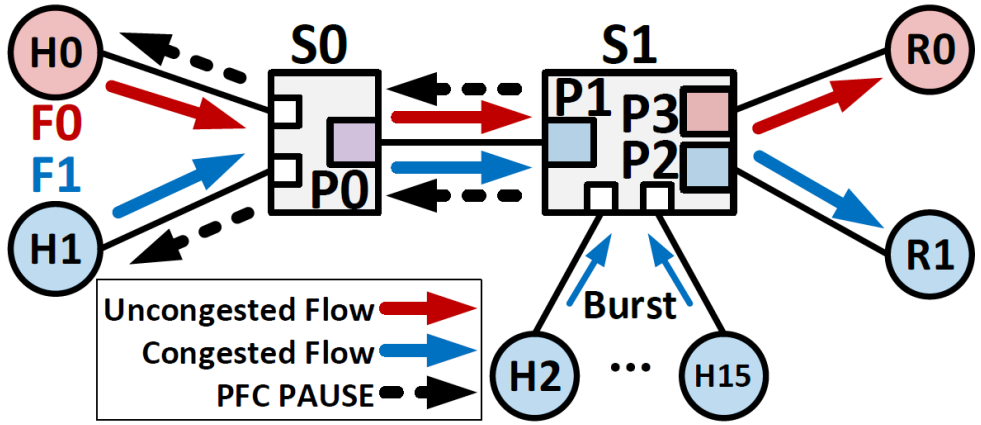


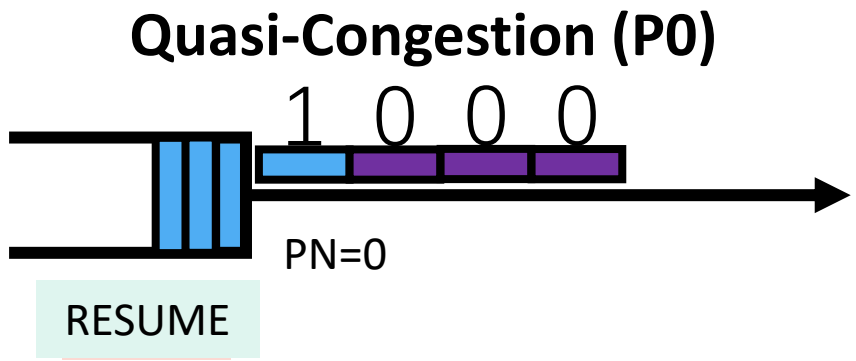
Non-Paused ECN (NP-ECN)

- Don’t change ECN for packets that has been paused
- Counter PN: number of packets that has been paused



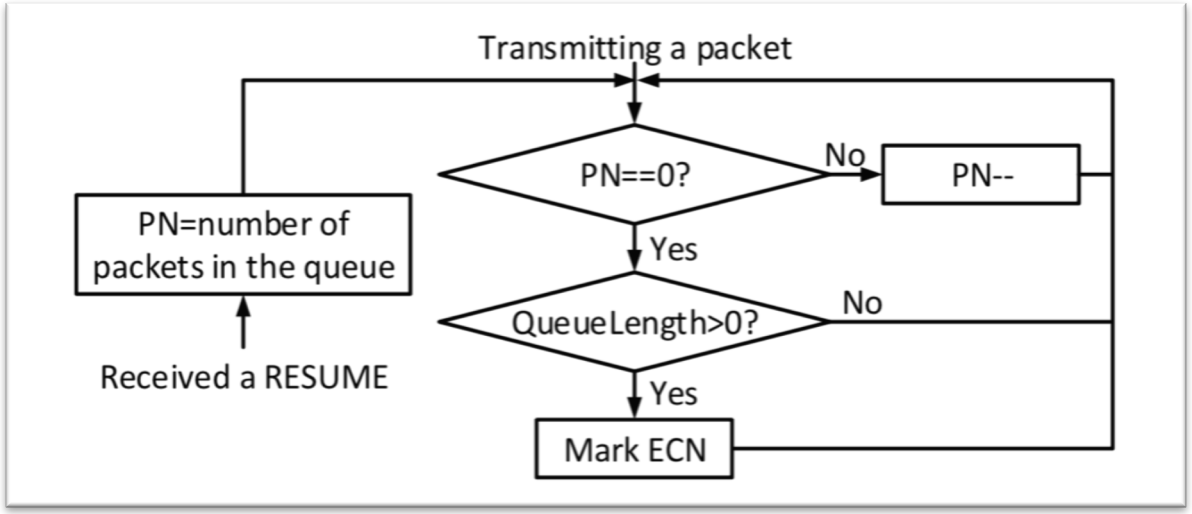
Congestion Detection



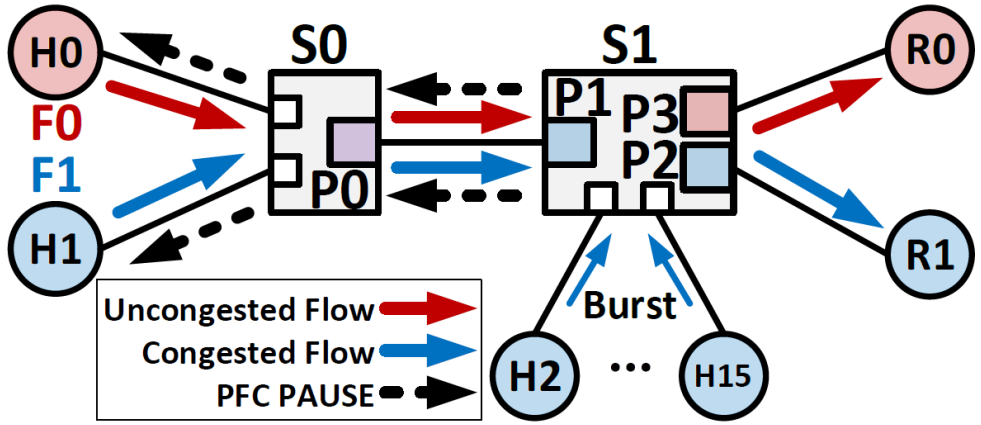


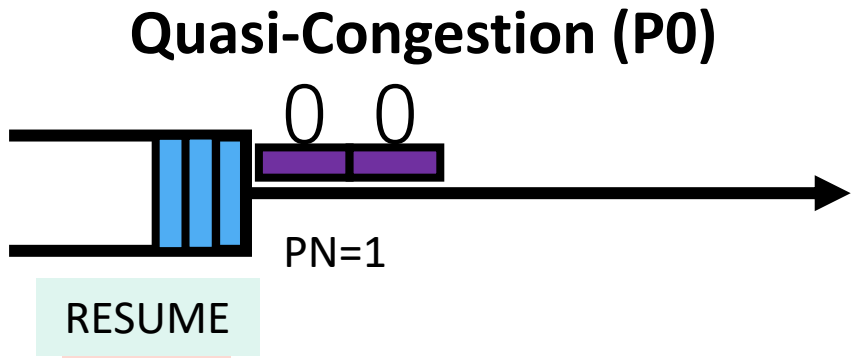
Non-Paused ECN (NP-ECN)

- Don’t change ECN for packets that has been paused
- Counter PN: number of packets that has been paused



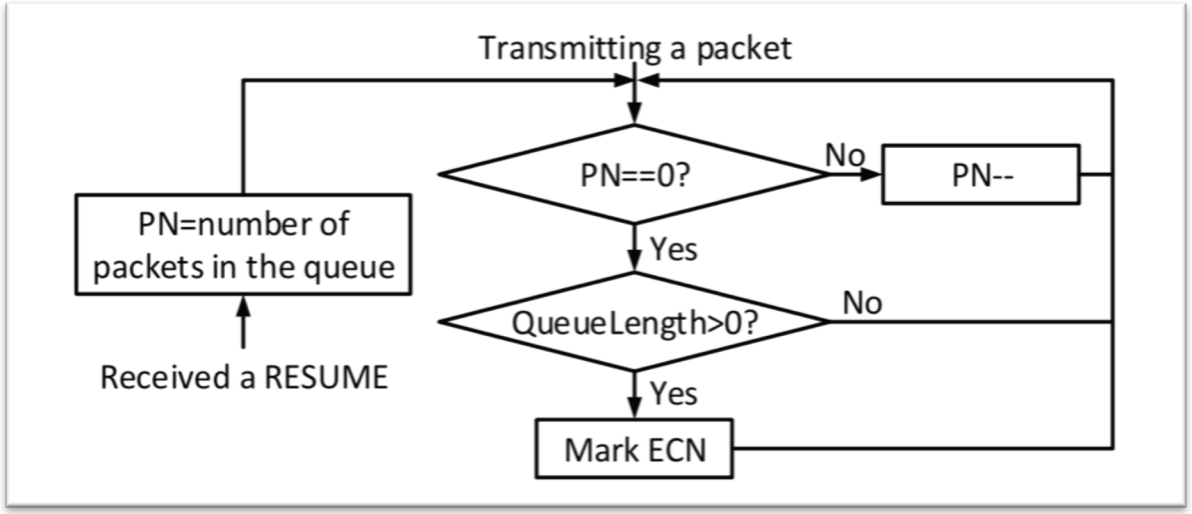
Congestion Detection



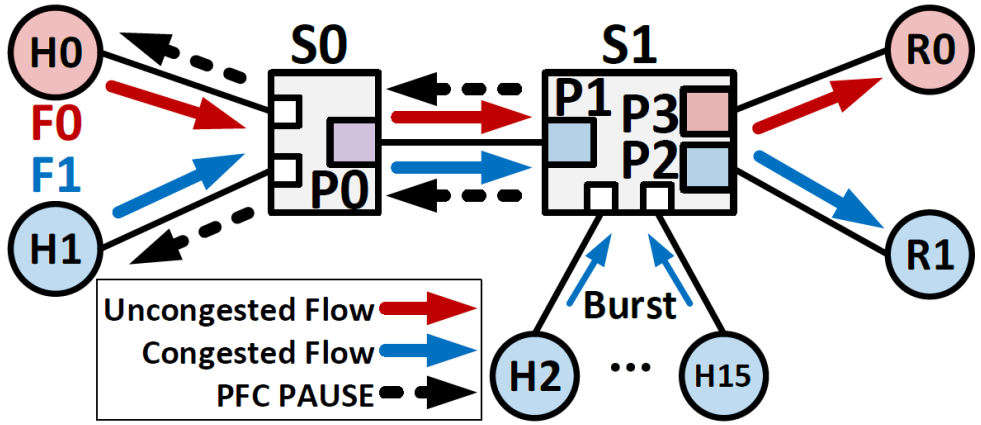


Non-Paused ECN (NP-ECN)

- Don’t change ECN for packets that has been paused
- Counter PN: number of packets that has been paused



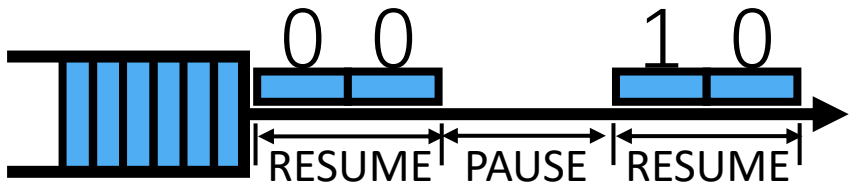
Congestion Detection



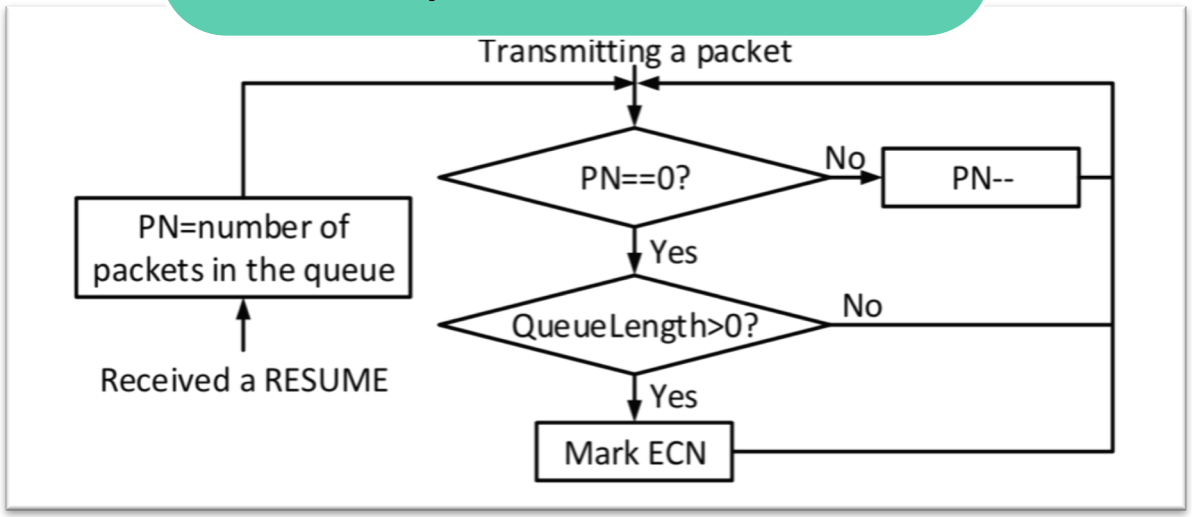
Non-Paused ECN (NP-ECN)

- Don't change ECN for packets that has been paused
- Counter PN: number of packets that has been paused

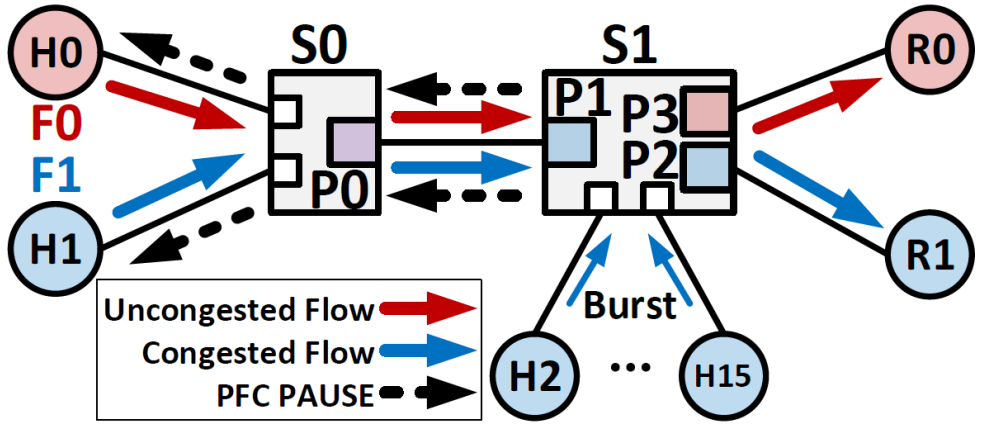
Quasi-Congestion (P0)

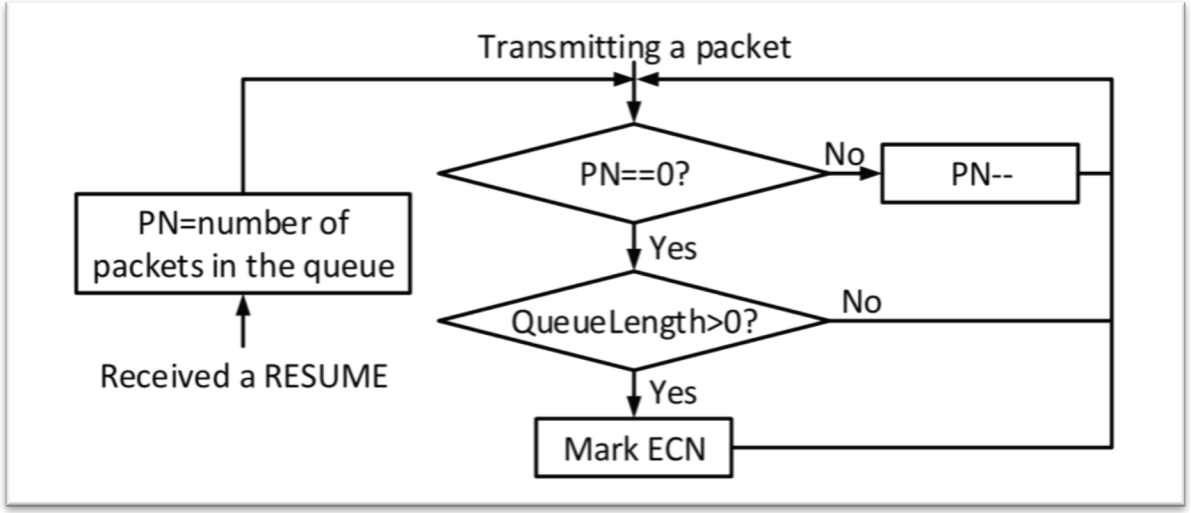


Partially marked with ECN



Congestion Detection

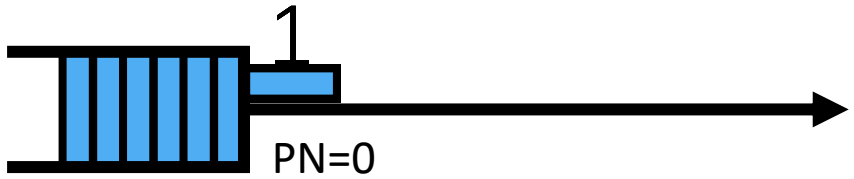




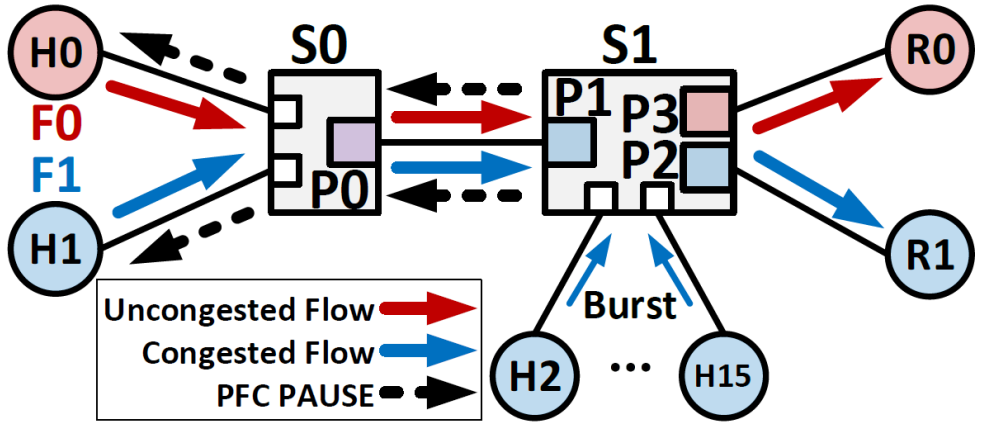
Non-Paused ECN (NP-ECN)

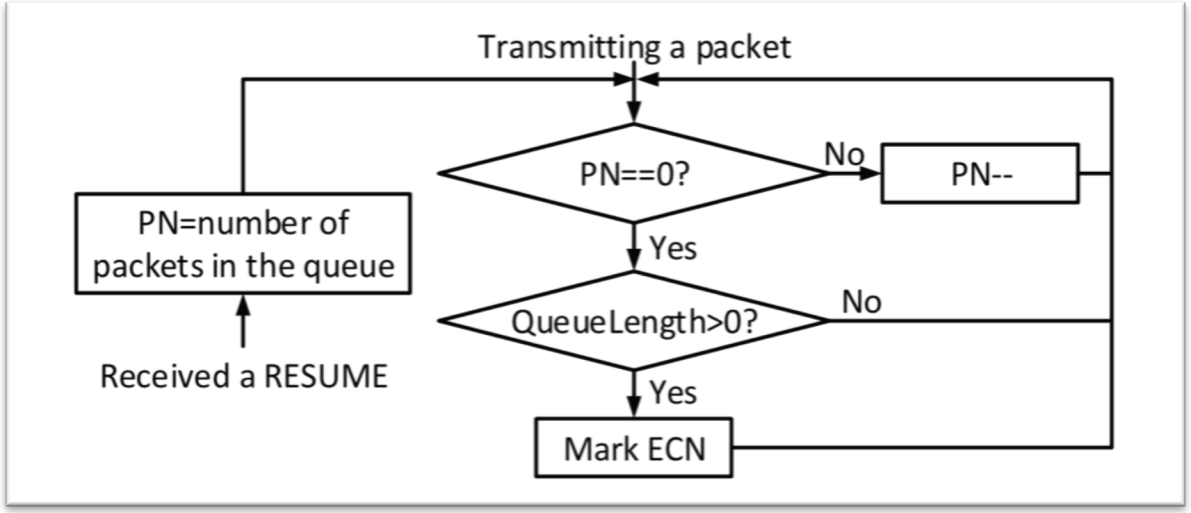
- Don't change ECN for packets that has been paused
- Counter PN: number of packets that has been paused

Real-Congestion (P2)



Congestion Detection

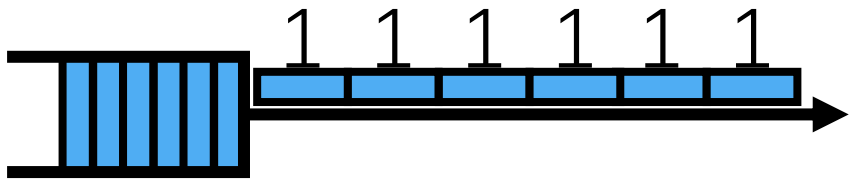




Non-Paused ECN (NP-ECN)

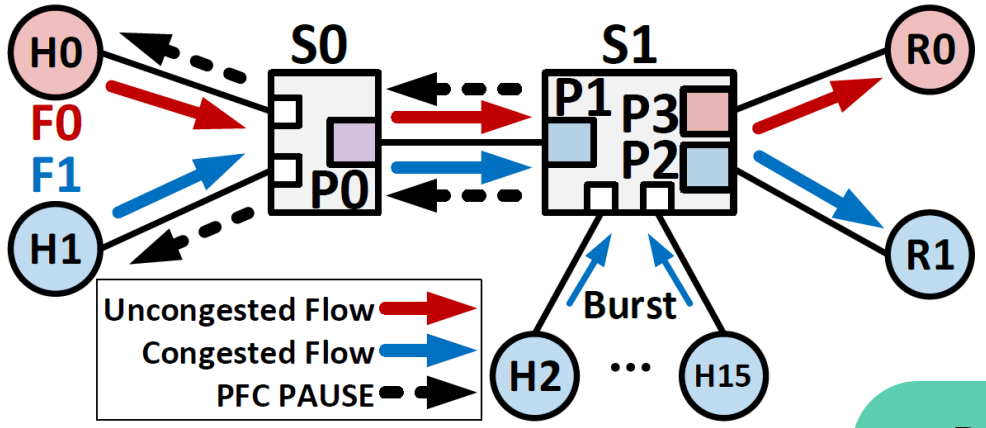
- Don't change ECN for packets that has been paused
- Counter PN: number of packets that has been paused

Real-Congestion (P2)

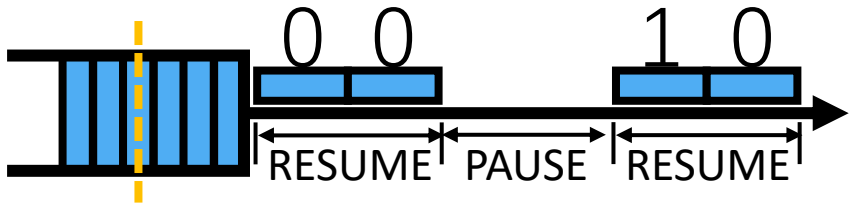


Continuously marked with ECN

Congestion Detection



Quasi-Congestion (P0)

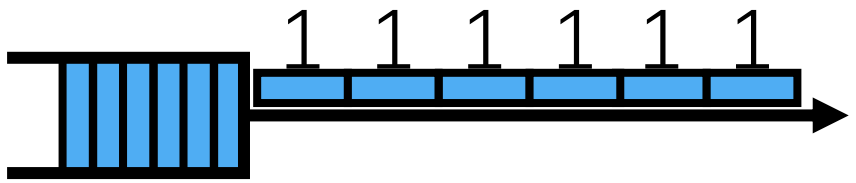


Partially marked with ECN ← Victim Flows

Non-Paused ECN (NP-ECN)

- Don't change ECN for packets that has been paused
- Counter PN: number of packets that has been paused

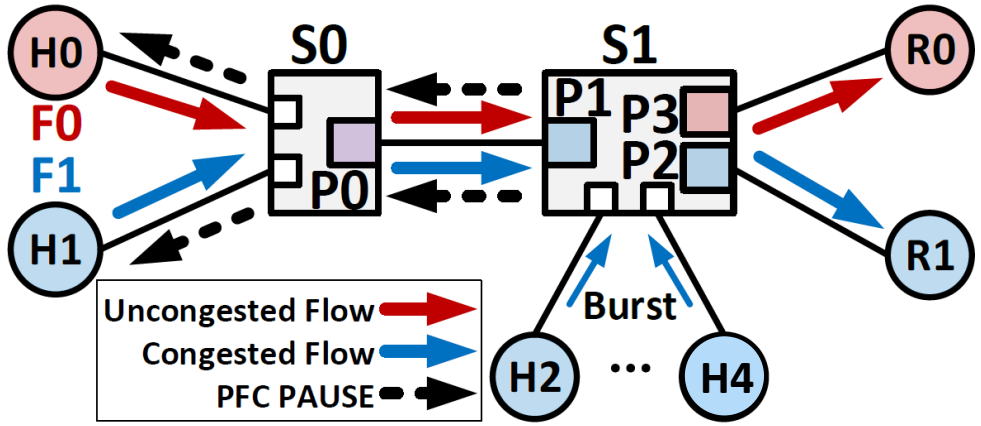
Real-Congestion (P2)



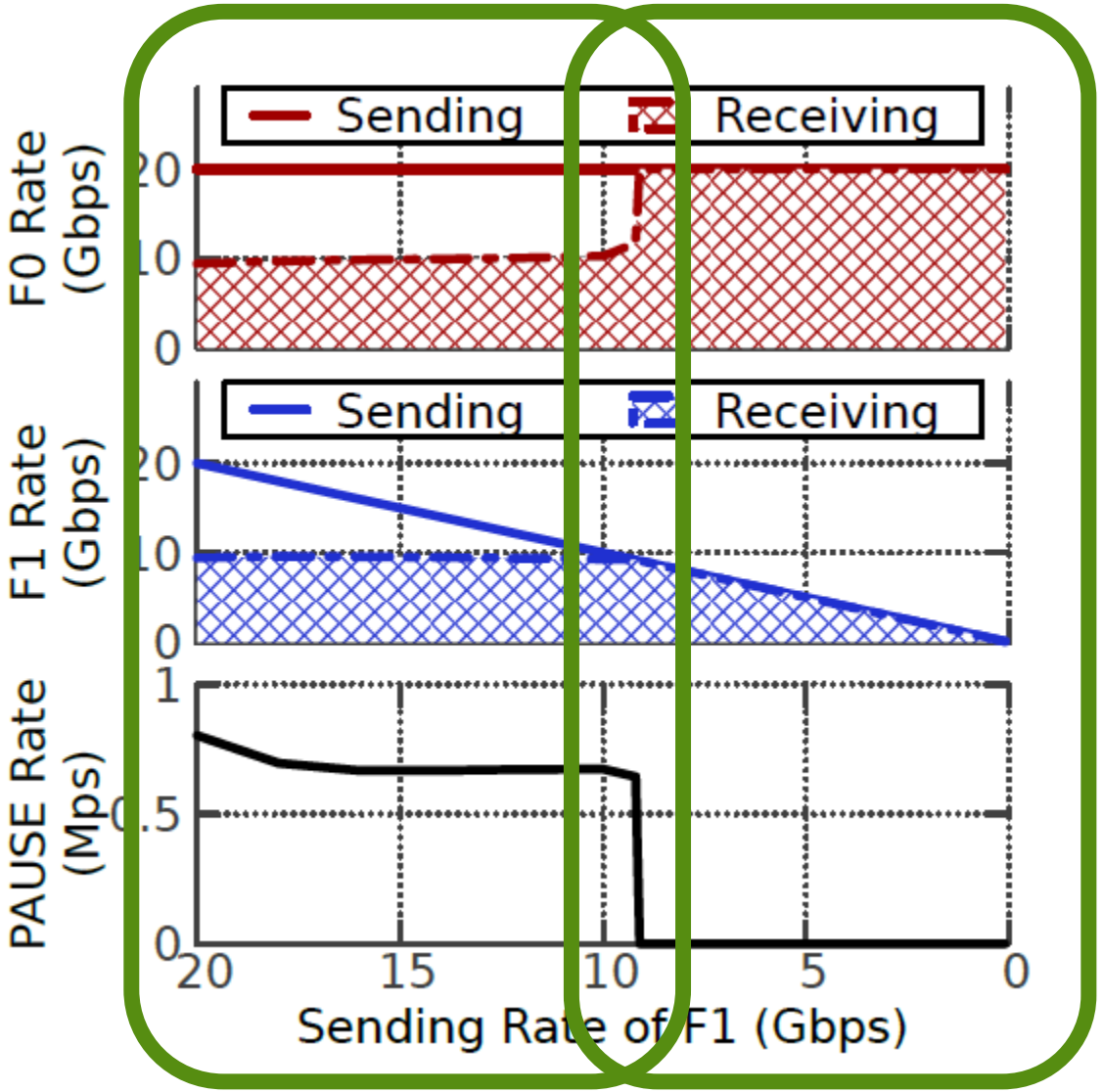
Continuously marked with ECN ← Congested Flows

Rate Adjustment

How to adjust the rates of
- Congested Flows --> target?
- Victim Flows --> no decrease?
- Non-congested Flows



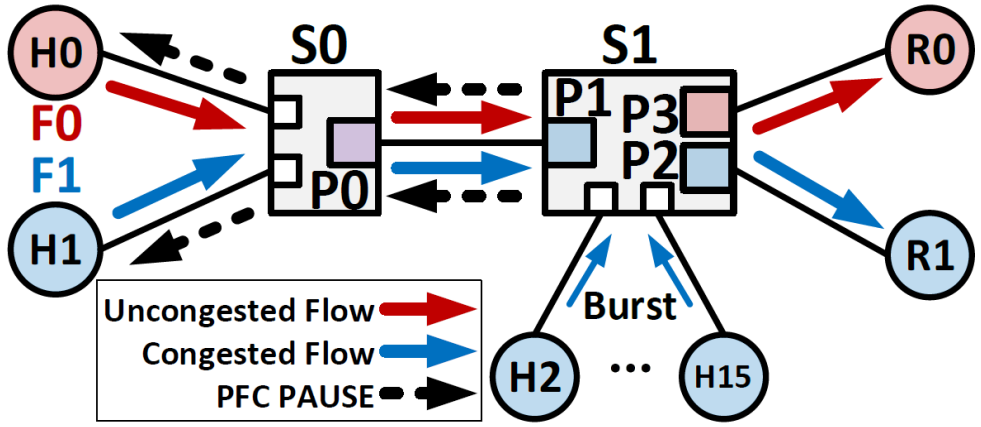
$\text{Burst} = 40Gbps$, $\text{F0} = 20Gbps$,
Reduce F1's rate



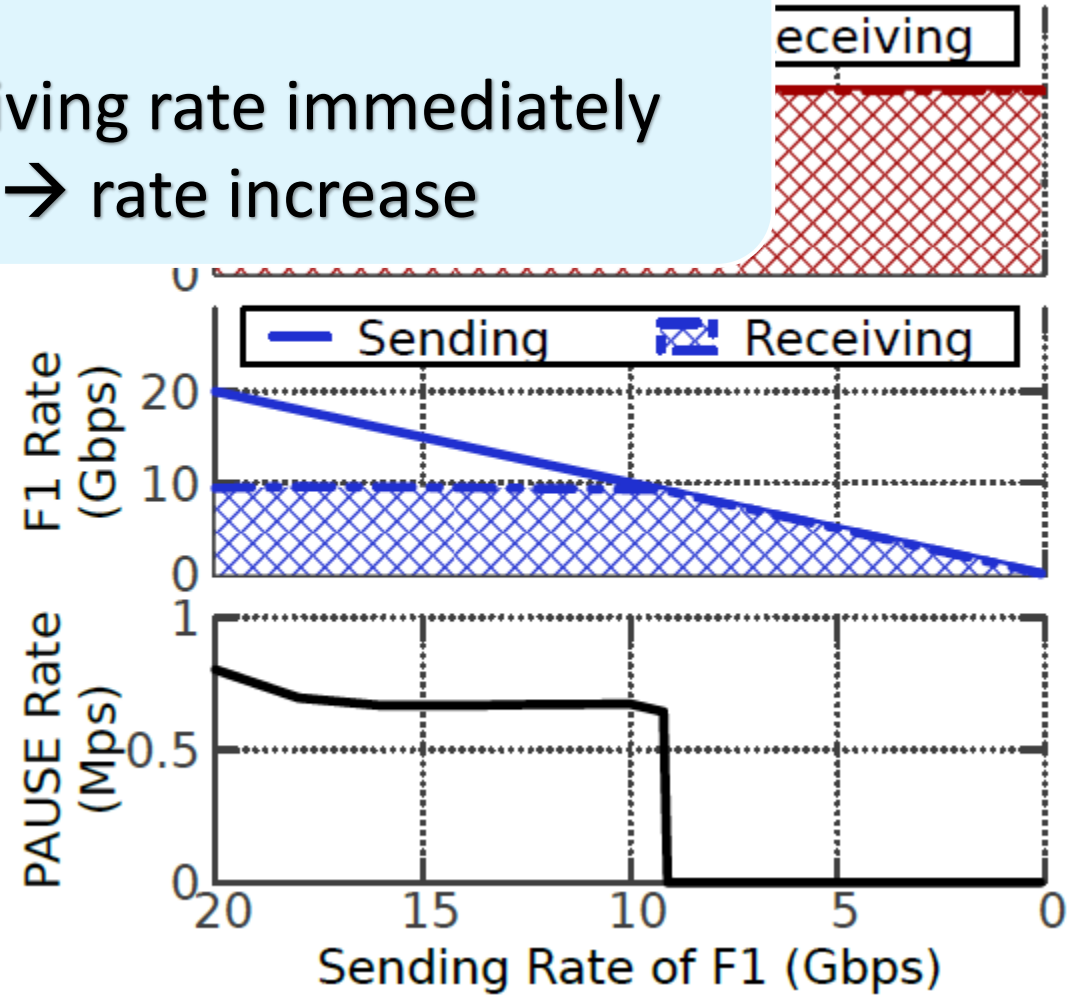
Rate Adjustment

How to adjust the rates of

- Congested Flows → reduce to receiving rate immediately
- Victim Flows & Uncongested Flows → rate increase



$\mathrm{F0} = 20Gbps$, Reduce F1's rate



Rate Adjustment

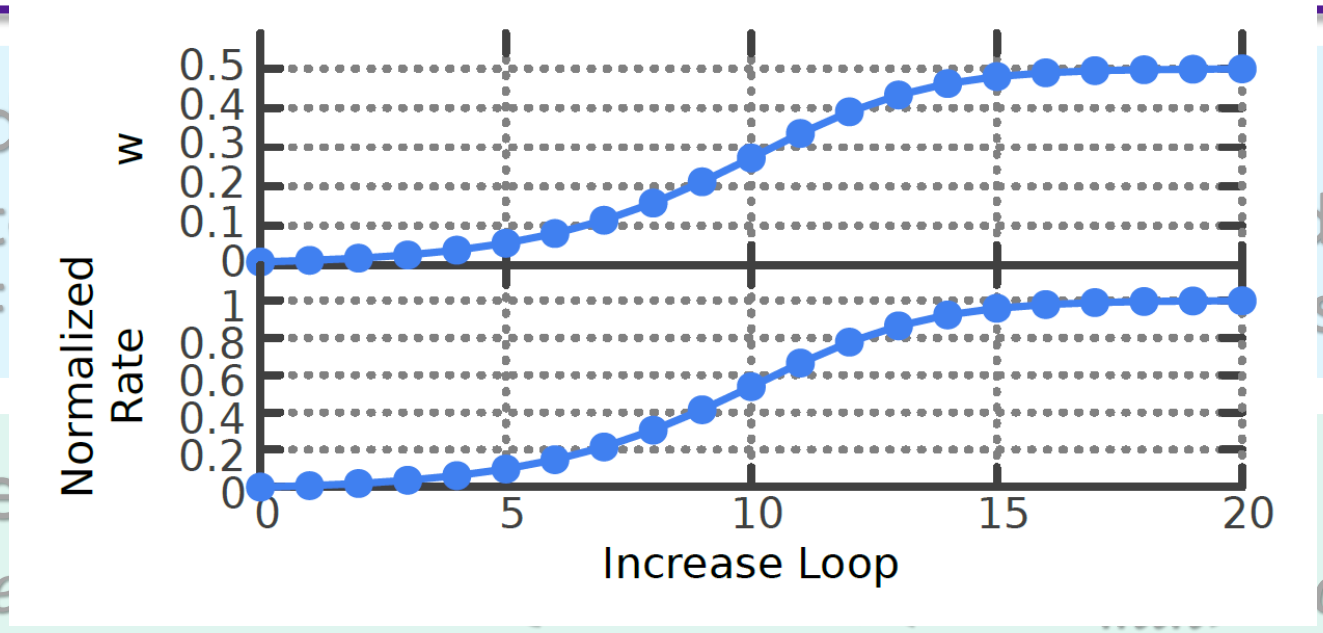
How to adjust the rates of

- Congested Flows → reduce to receiving rate immediately
- Victim Flows & Uncongested Flows → rate increase

Receiver-Driven Rate Decrease

- $sendRate \leftarrow \min\{sendRate, (1 - w_{min})recRate\}$
- No PFC & no serious throughput loss & 1 control loop

Rate Adjustment



How to ac
- Congest
- Victim F

diately
se

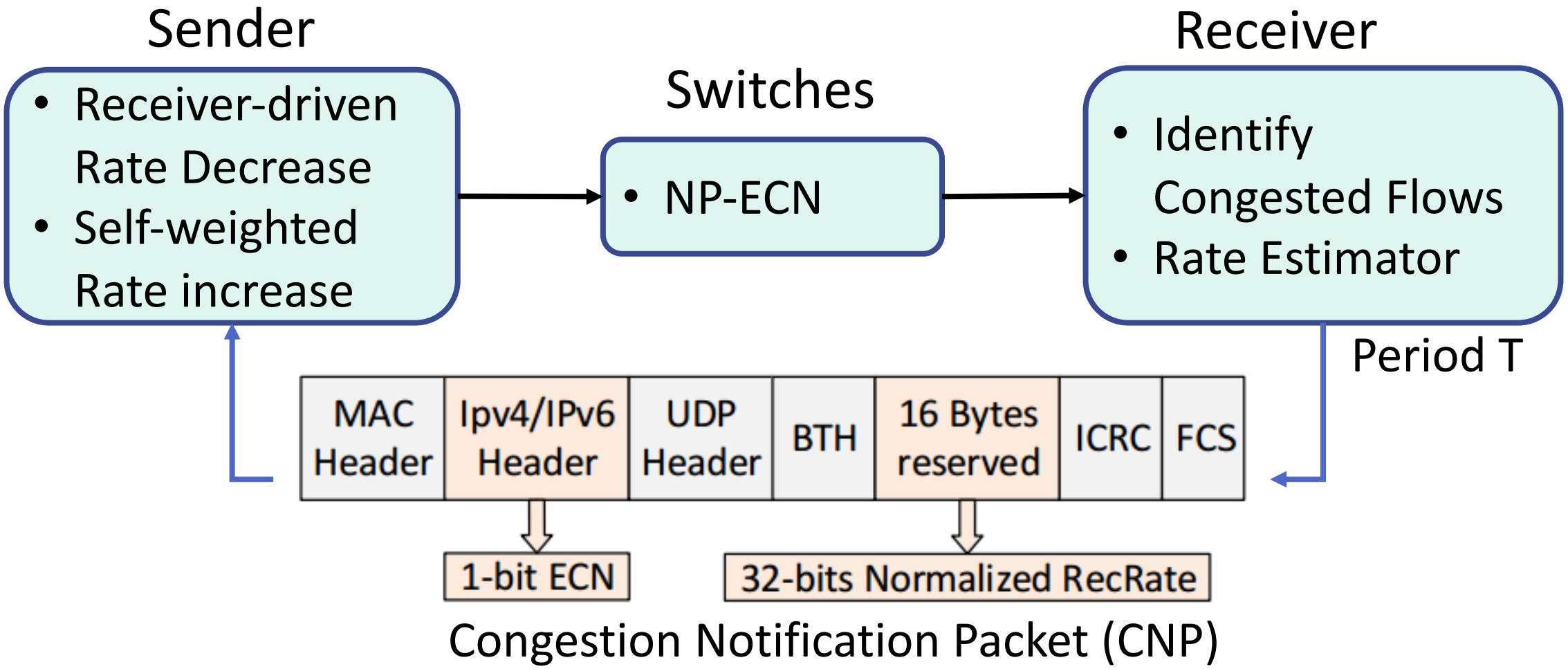
Rece
- se
- No congestion & No PFC triggers in one control loop

cRate}

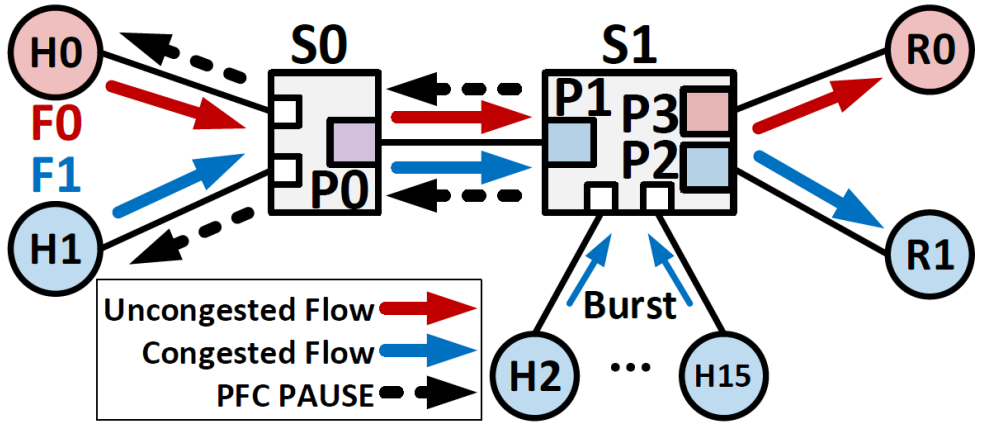
Self-weighted Rate increase

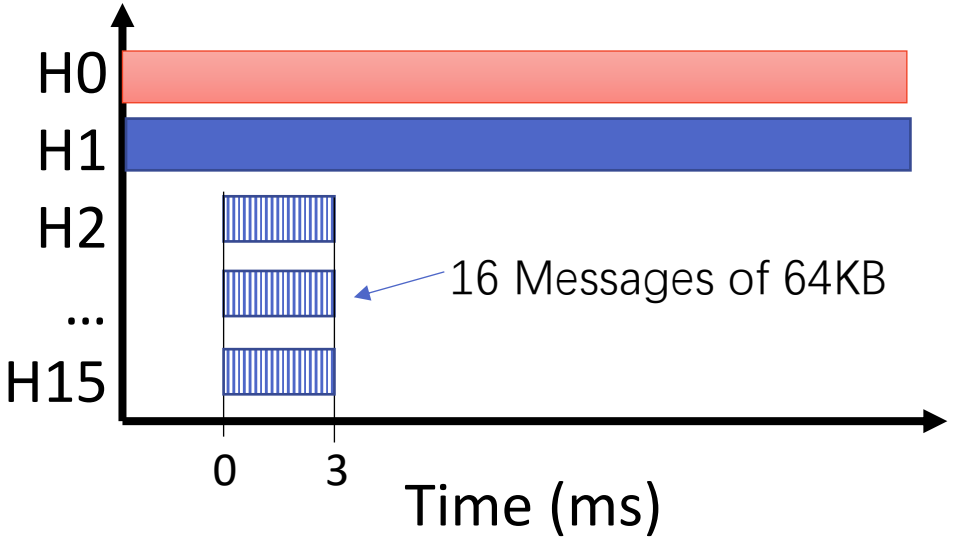
- $\begin{cases} sendRate \leftarrow sendRate(1-w) + MaxRate \cdot w \\ w \leftarrow w(1-w) + w_{max} \cdot w \end{cases}$
- Automatic gentle-to-aggressive

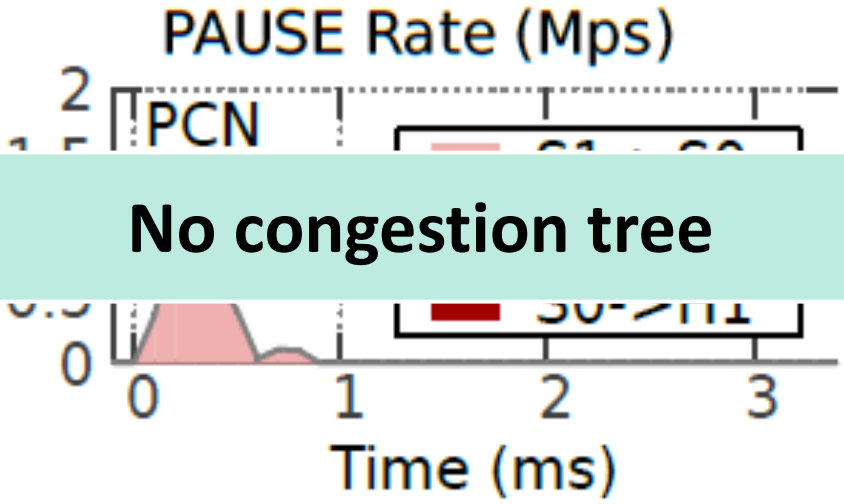
Photonic Congestion Notification (PCN)



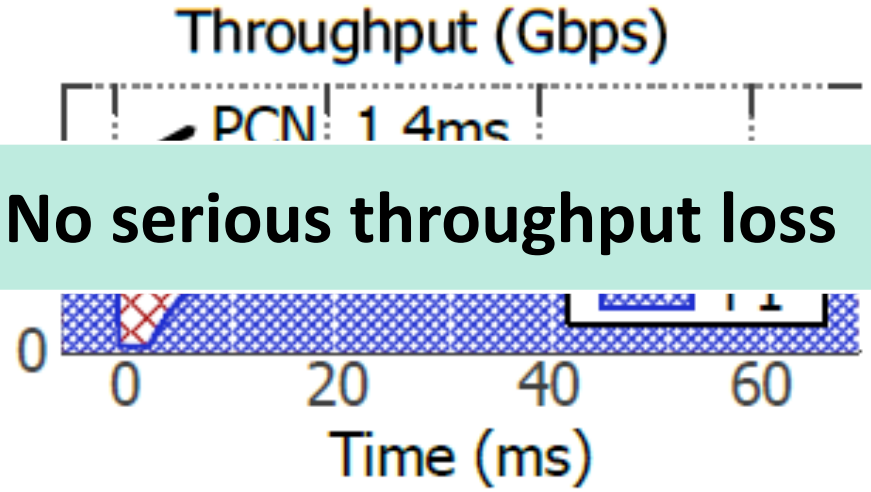
PCN's Benefit





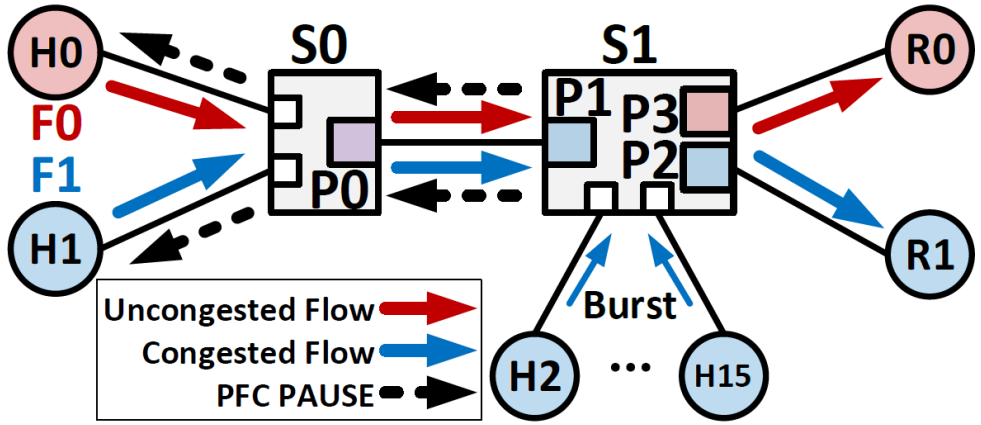


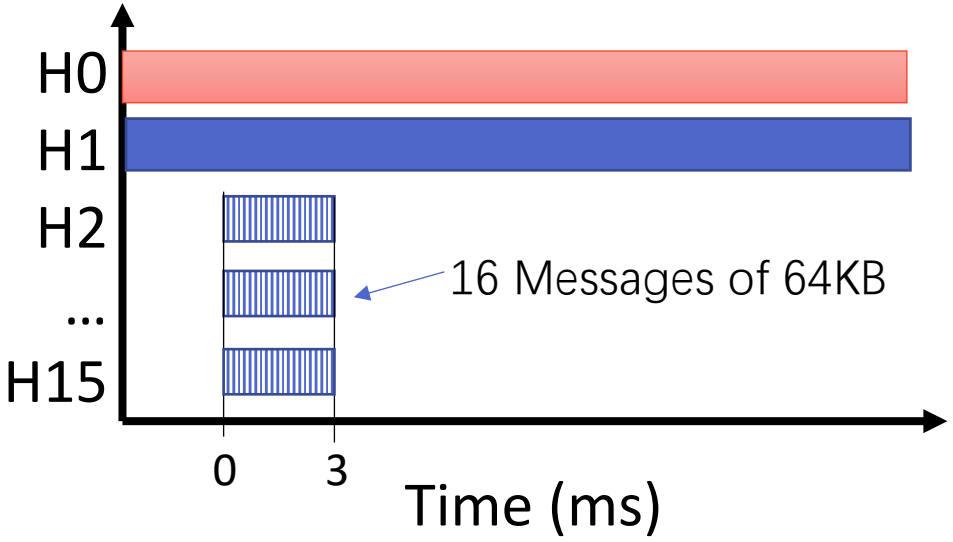
No congestion tree



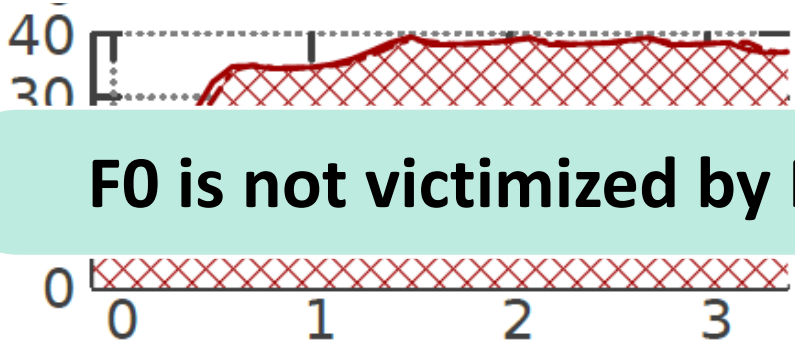
No serious throughput loss

Benefit



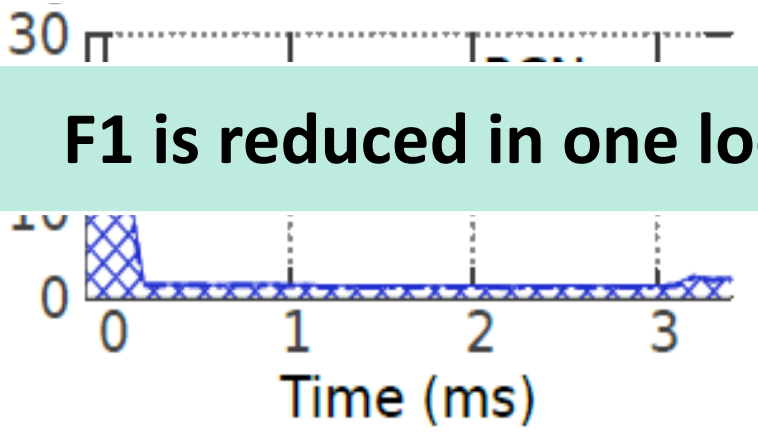


Sending Rate of F0 (Gbps)



F0 is not victimized by PCN

Sending Rate of F1 (Gbps)



F1 is reduced in one loop

Evaluation Setup

Testbed Setup

- Dumbbell topology
- Implementation on DPDK (Intel 82599)
- 4 hosts (PowerEdge R530) connected to single ToR
- 10Gbps

NS-3 Simulation Setup

- Clos topology
- 512 hosts / 32 ToRs / 16 Leafs / 8 Spines
- 10Gbps / 40Gbps

Evaluations

Basic Prosperities

- Convergence
- Fairness
- Stability

Testbed

Workbench

- Burst Tolerance
- Parameter sensitivity
- Realistic Workloads

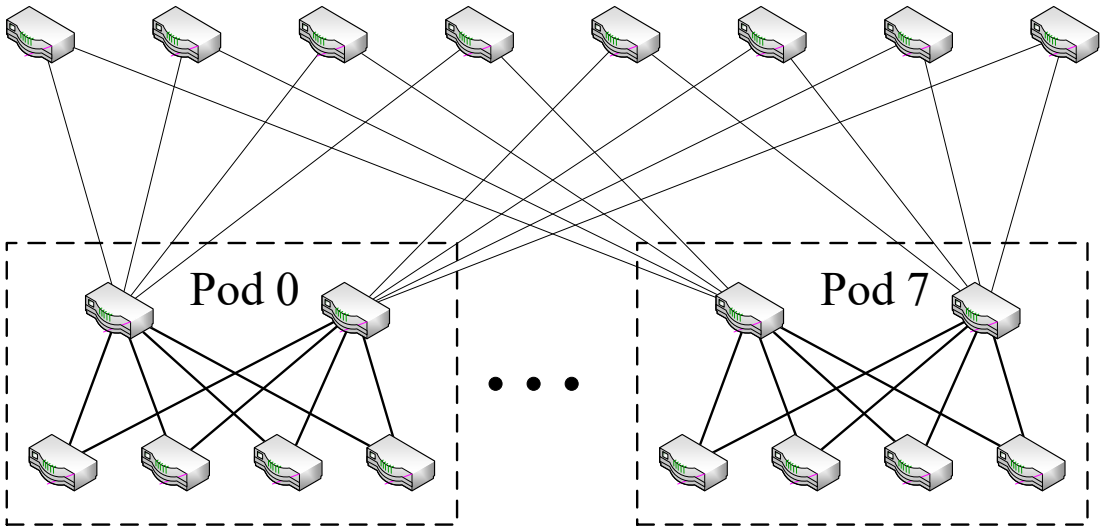
Special Cases

- Flow Scalability
- Adversarial Traffic
- Multiple Bottlenecks
- Multiple Priorities
- Deadlock

NS-3 Simulations

Evaluation: Large-Scale Simulations

Simulation Setup



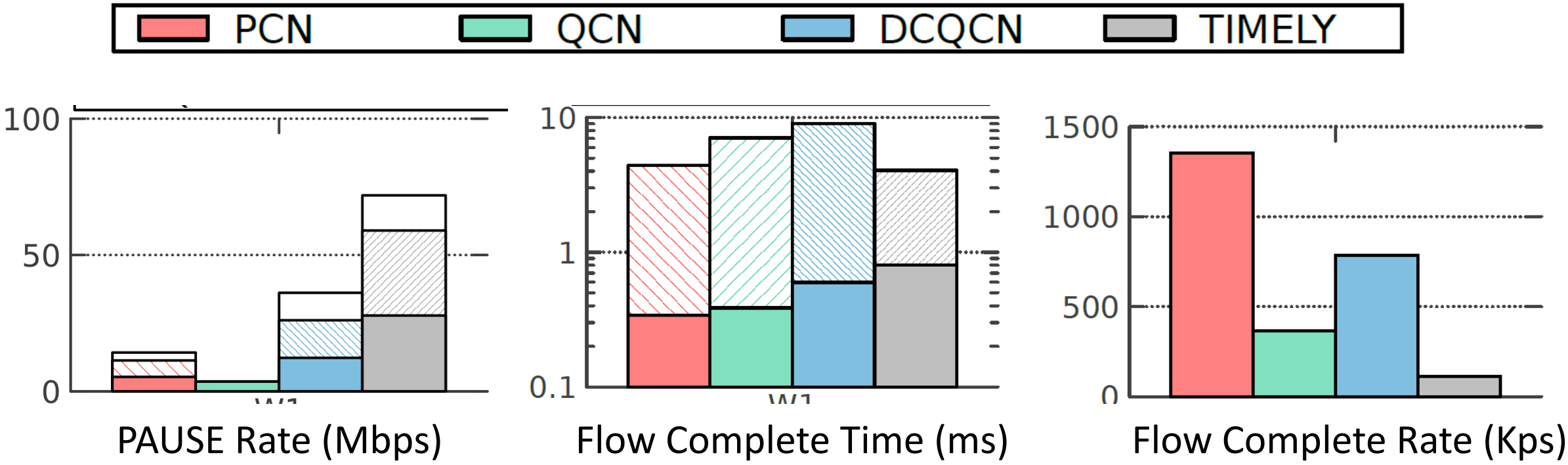
Flow size	% of number		% of traffic	
	W1	W2	W1	W2
0KB-10KB (S)	80.14	70.79	3.08	0.22
10KB-100KB (M)	10.32	16.59	5.89	1.56
100KB-1MB (L)	9.12	3.52	83.8	1.53
1MB- (XL)	0.41	9.1	7.04	96.7

W1: Web-server workload

W2: Hadoop cluster workload

Evaluation: Large-Scale Simulations

Web-server Workload



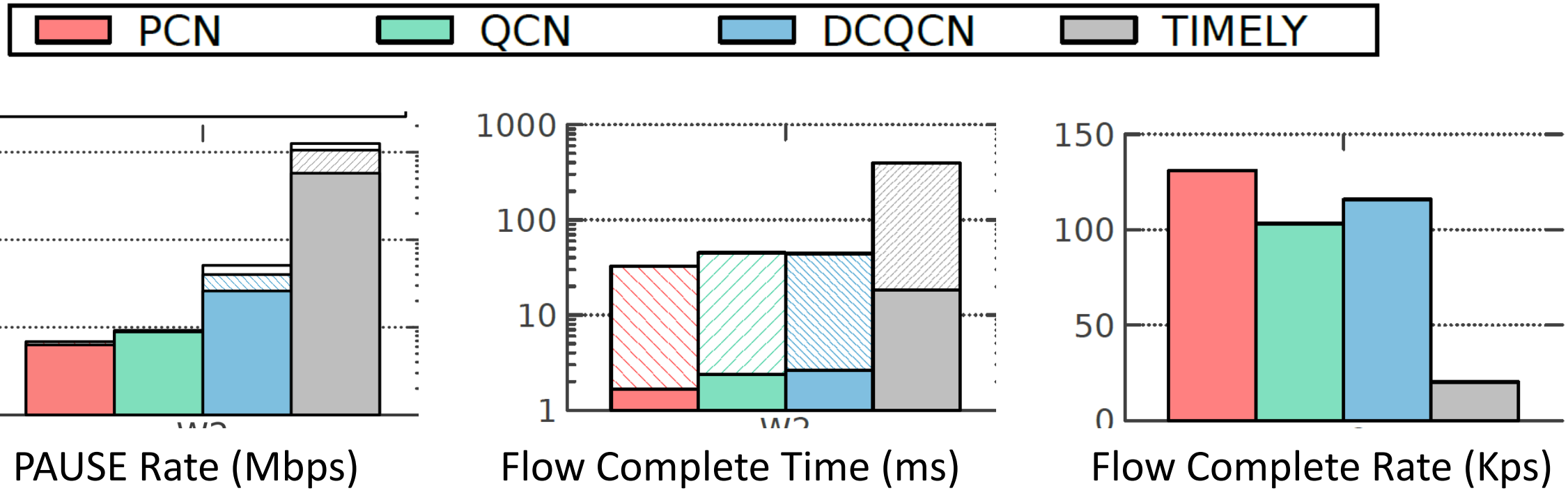
PAUSE Rate (Mbps)

Flow Complete Time (ms)

Flow Complete Rate (Kps)

Evaluation: Large-Scale Simulations

Hadoop Workload



PAUSE Rate (Mbps)

Flow Complete Time (ms)

Flow Complete Rate (Kps)

Conclusion

Re-architecting congestion management

Proposing Photonic Congestion Notification (PCN)

- NP-ECN → victim flows/congested flows
- Receiver-driven rate decrease → no PFC in 1 loop
- Automatic rate increase

Evaluations on testbed and ns-3 simulation show, PCN triggers fewer PFC and achieves lower flow completion time.

Thanks !

pyscwx@126.com
renfy@tsinghua.edu.cn