New Simulation Results of Congestion Isolation (CI)

Sam Sun (sam.sunwenhao@huawei.com)

Kevin Shen (kevin.shenli@huawei.com)

IEEE 802.1 Interim, Pittsburgh, May 2018

Contents

- Objectives of This Work
- Simulation Settings
- Simulation Results
- Conclusions

Objectives of This Work

- Simulate the CI mechanism using a ‘published’ shared buffer switch model
  - Understand how CI works with shared buffer switches
- Compare the performance of DCQCN with and without CI
  - Verify that CI (local and immediate congestion management) interoperates with existing congestion management, specifically PFC (Priority-based Flow Control), and end-to-end congestion management mechanisms, e.g., DCQCN (Data Center Quantized Congestion Notification)
- See also previous work in [1][2]

[1] http://www.ieee802.org/1/files/public/docs2018/new-dcb-shen-congestion-isolation-simulation-0118-v01.pdf
[2] http://www.ieee802.org/1/files/public/docs2018/cz-shen-congestion-isolation-simulation-0318-v01.pdf

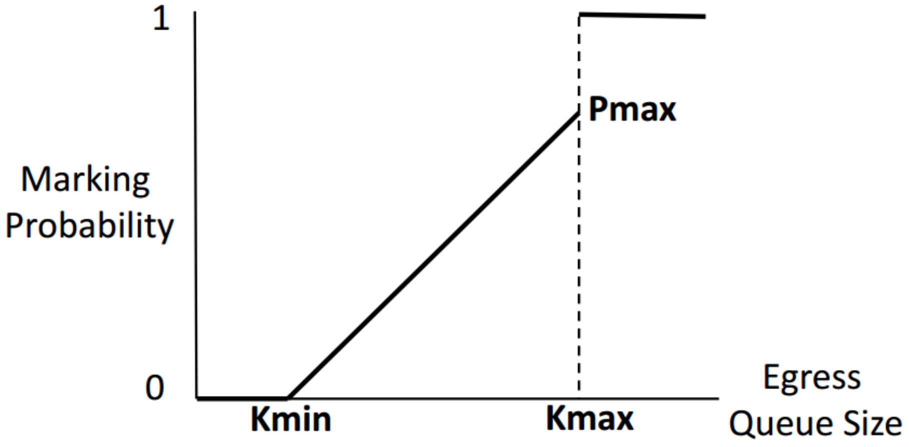
Brief Introduction to DCQCN [3]

- DCQCN is a rate-based end-to-end congestion control scheme for RoCEv2.
- Builds upon concepts from 802.1Qau (QCN), but end-to-end at Layer 3
- Switch: an arriving packet is ECN-marked if the queue length exceeds a threshold.
- Receiver: the NP (Notification Point) algorithm specifies how and when CNPs (Congestion Notification Packets) should be generated.
- Sender: when an RP (Reaction Point) gets a CNP, it executes the RP algorithm and reduces its current rate.

[3] Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., & Liron, Y., et al. (2015). Congestion Control for Large-Scale RDMA Deployments. ACM SigComm Computer Communication Review, 45(4), 523-536.

Simulation Settings

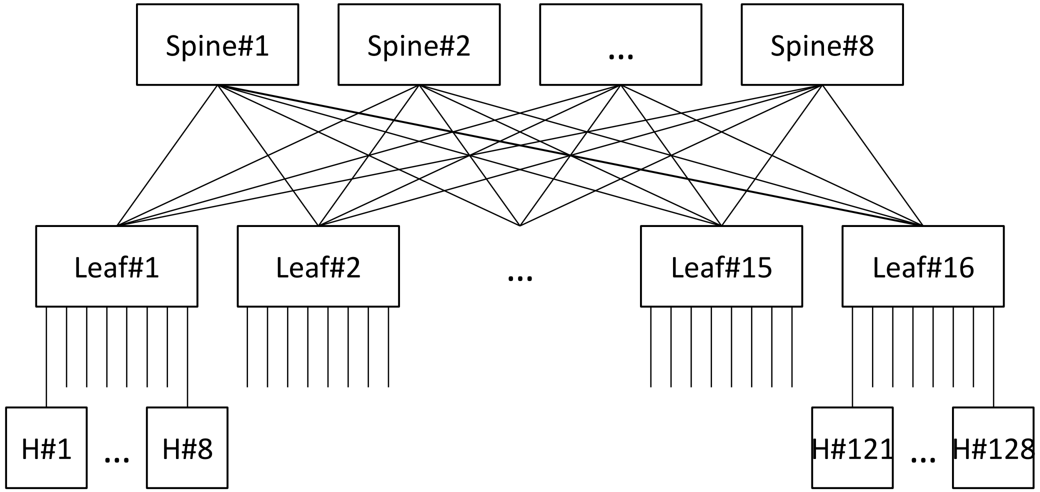
- Model: NS-3 with DCQCN and PFC implemented [4]
  - Buffer: Total = 12 MB, Shared = 8 MB
  - DCQCN Parameters (as described in [3]):
    - K_{max} = 200 KB
    - K_{min} = 5 KB
    - P_{max} = 1%



[4] https://github.com/bobzhuyb/ns3-rdma

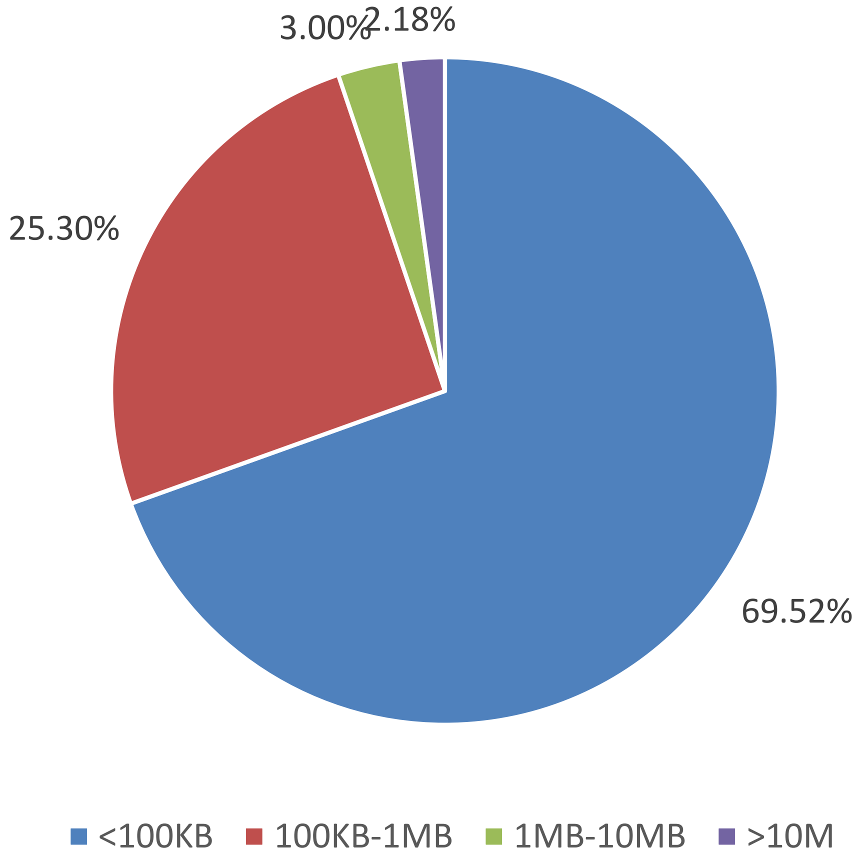
Simulation Settings

Topology:	2-tier Clos
Server #:	128
Switch #:	20
Link Bandwidth:	40 Gbps
Link Delay:	1 µs



Simulation Settings

- Traffic:
  - Weibull Distribution [5]
  - Traffic Characteristics of Distributed Storage in Facebook's Datacenter [6]
  - MTU: 1KB (as in [4])



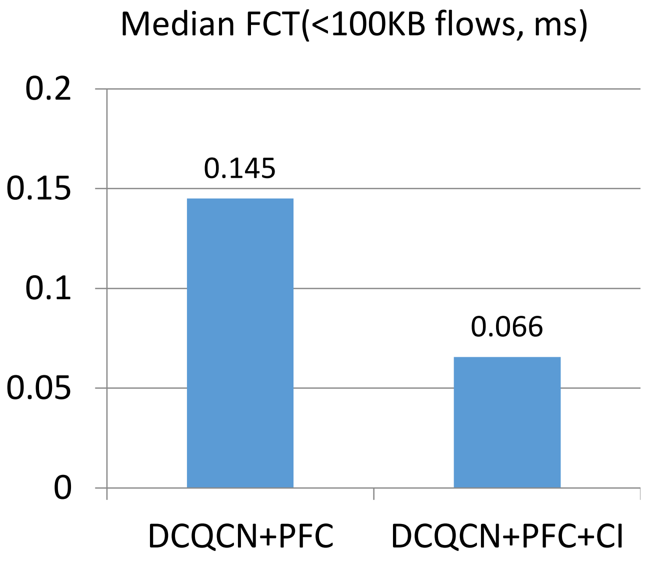
[5] Theophilus Benson, Aditya Akella, and David A. Maltz. 2010. Network traffic characteristics of data centers in the wild. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (IMC '10). ACM, New York, NY, USA, 267-280.

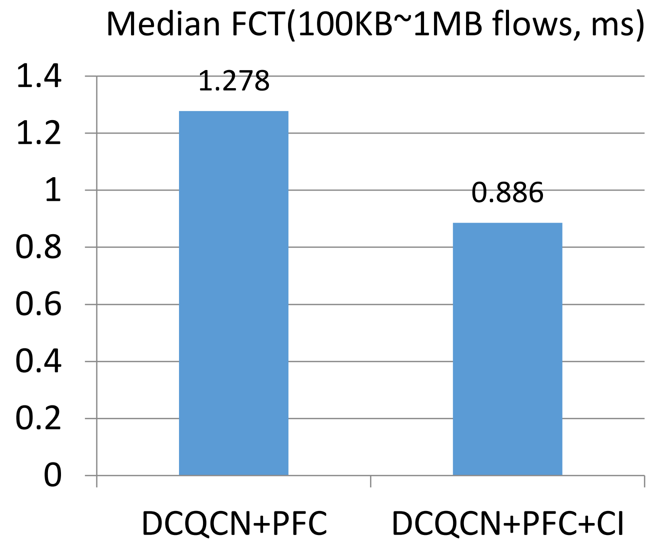
[6] Bagga, J., Bagga, J., Bagga, J., Porter, G., & Snoeren, A. C. (2015). Inside the Social Network's (Datacenter) Network. ACM Conference on Special Interest Group on Data Communication (Vol.45, pp.123-137). ACM.

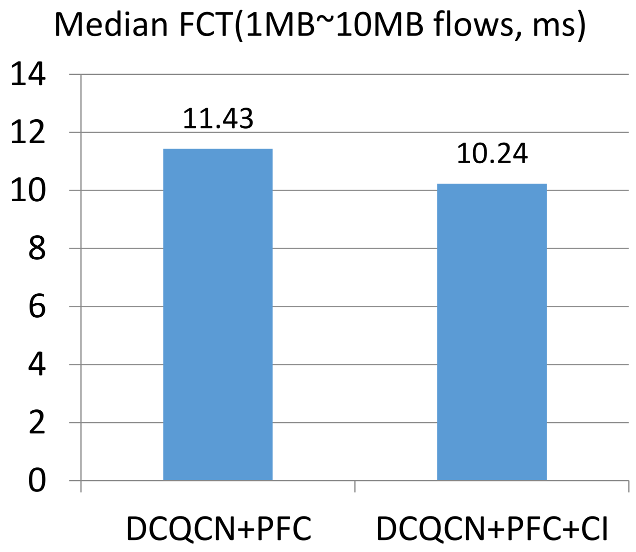
Simulation Settings

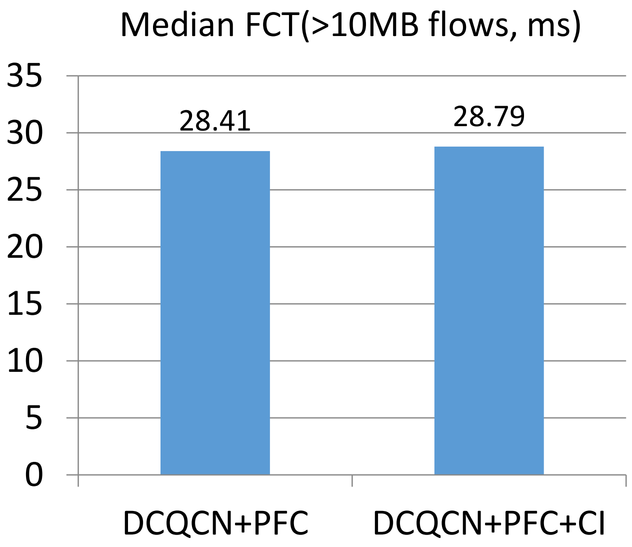
- Compared solutions:
  - DCQCN+PFC: two queues (Queue 0 and Queue 1), round robin.
  - DCQCN+PFC+CI: a congested flow queue (Queue 0), and an uncongested flow queue (Queue 1), strict priority.

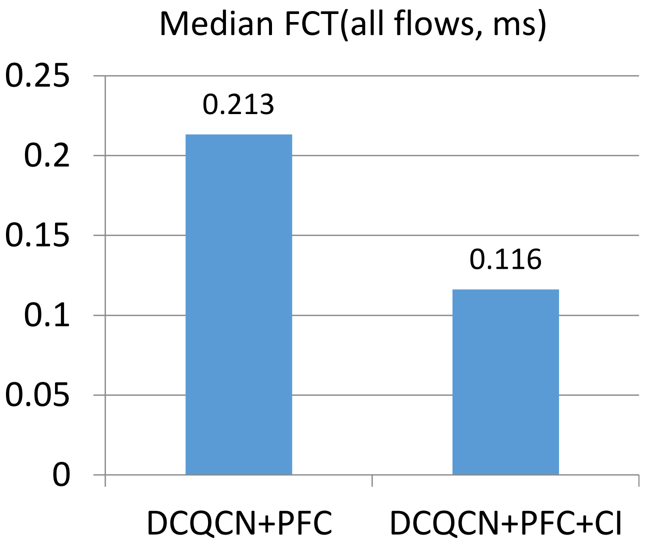
Simulation Results: Flow Completion Time

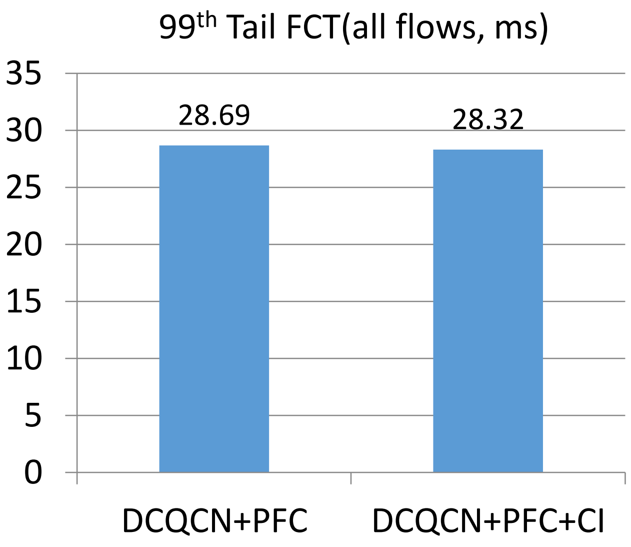

Median FCT(<100KB flows, ms)


Median FCT(100KB~1MB flows, ms)

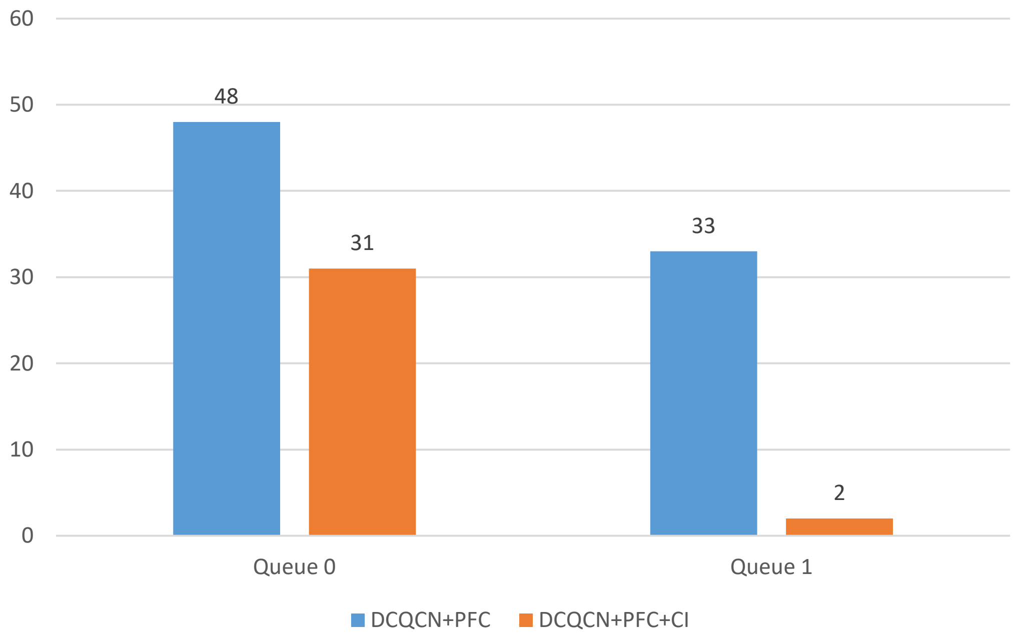

Median FCT(1MB~10MB flows, ms)


Median FCT(>10MB flows, ms)

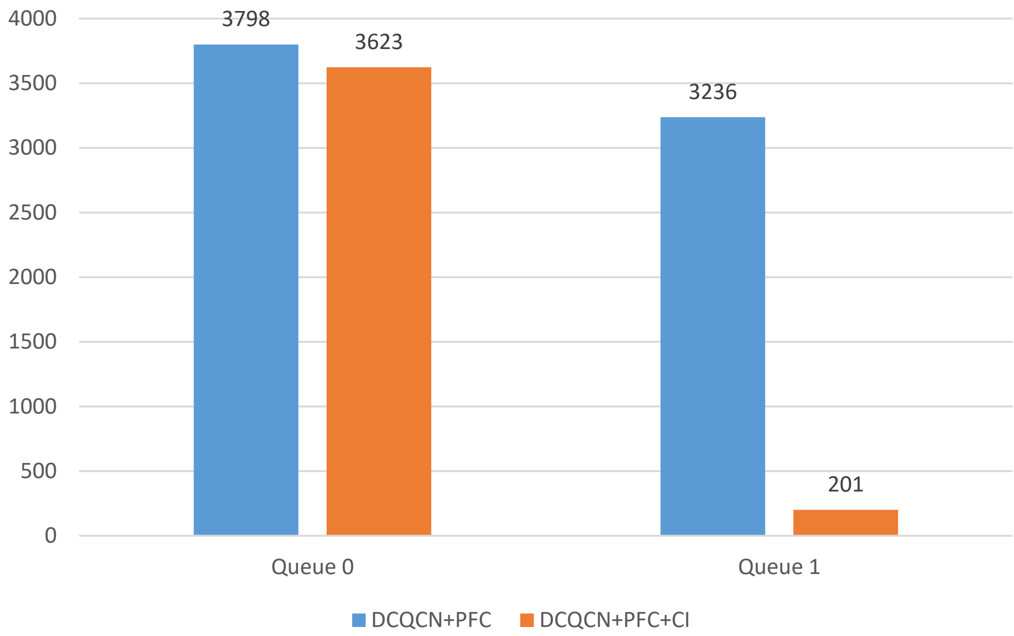

Median FCT(all flows, ms)


99th Tail FCT(all flows, ms)

Simulation Results: Queue Length



Average Queue Length (KBytes)



Maximum Queue Length (KBytes)

Conclusions

- CI works well with shared buffer switches
- When working with DCQCN, CI improves the performance as well

Thank you!

Questions?