Simulation Analysis of Congestion Isolation (CI)

Kevin Shen

kevin.shenli@huawei.com

Sam Sun

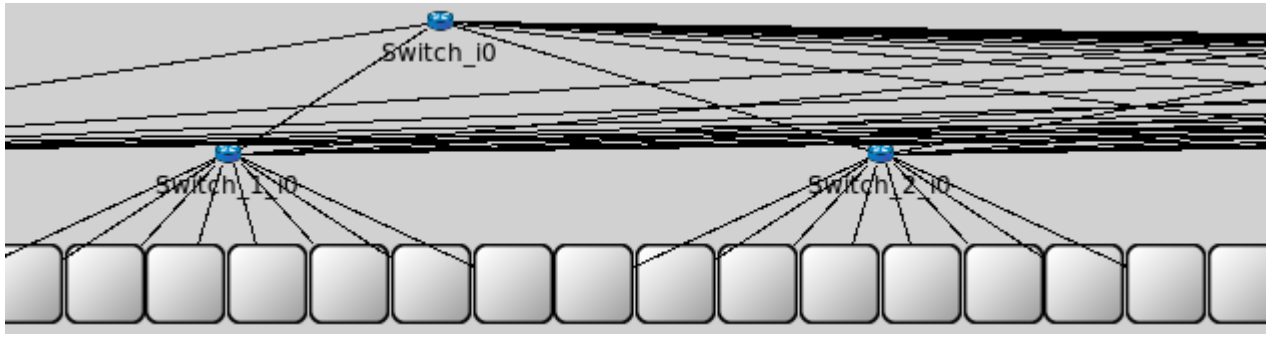
sam.sunwenhao@huawei.com

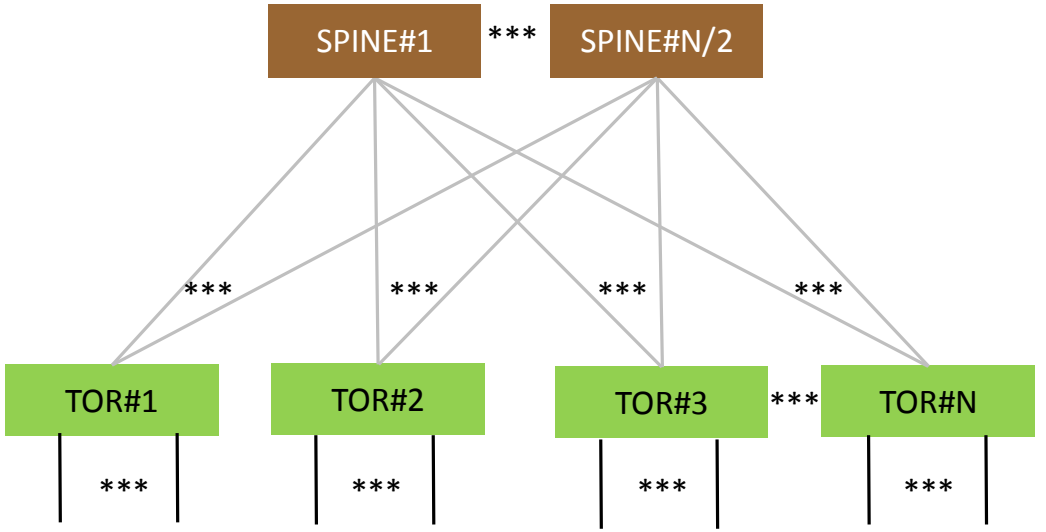
IEEE 802.1 DCB, Geneva Switzerland, January 2018

Objectives of the Analysis

- **Two queue model** (congested and non-congested queues; no mice prioritization)
  - How CI performs when only using 2 queues without mice/elephant separation.
- **Memory sensitivity**
  - How CI performs when modifying queue buffer size and threshold.
- **Including static switch latency**
  - How CI performs when adding additional static latency.
- **Queue depth**
  - Compare queue depths with and without CI.
- **Lossy scenario (no PFC)**
  - How CI performs without PFC enabled.

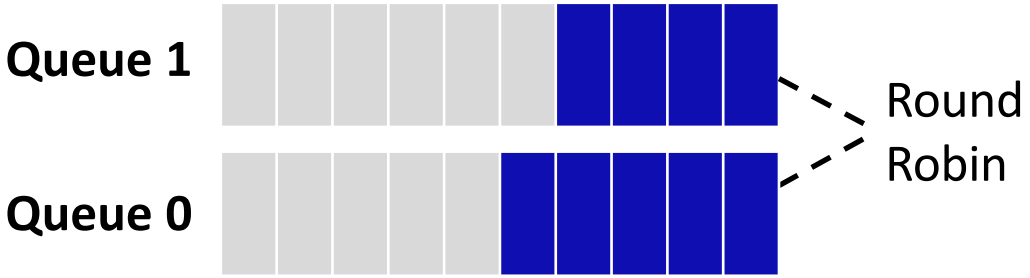
Simulation Set-up





- **Platform:** OMNET++
- **2 Tier CLOS:** 100G interface with 200ns of link latency (about 40 meters)
- **Scale:** 128 ~ 1152 servers, 24 ~ 72 switches
- **Traffic Pattern:** Data Mining Application with 82% of mice

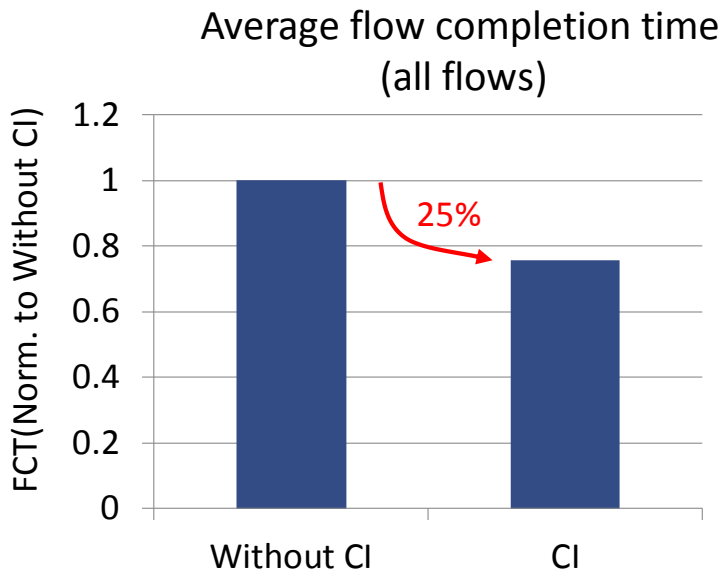
Compared Solutions

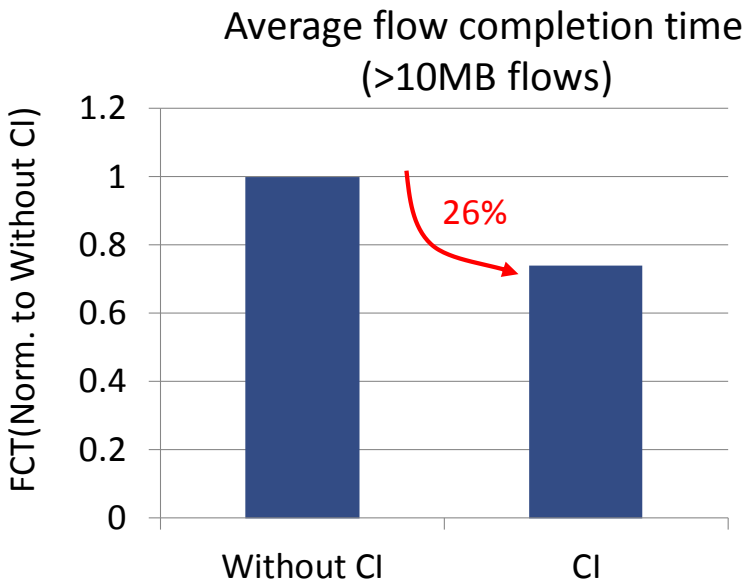


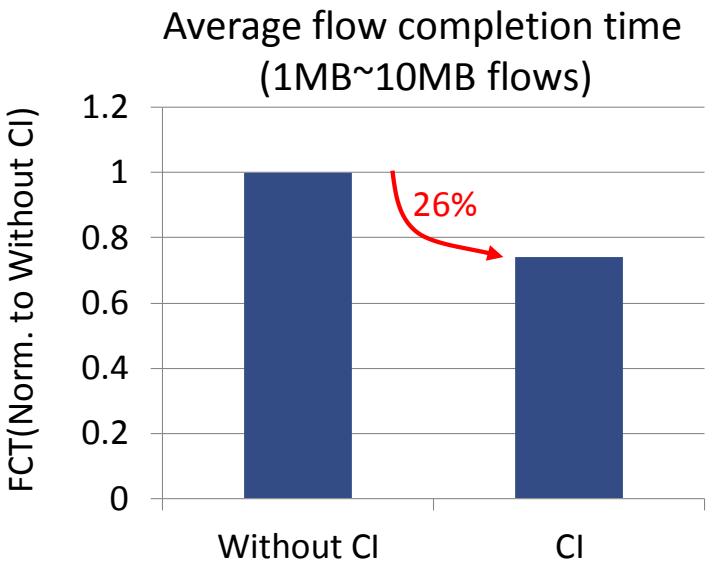
- Solution "**without CI**" means PFC + ECN without CI.
- Flows are mapped to one of the two queues by hash of destination IP.
- PFC and ECN are enabled on both queues.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB, XON 4 KB;
  - ECN: Low 10 KB, High 300 KB, Max Probability 1%.

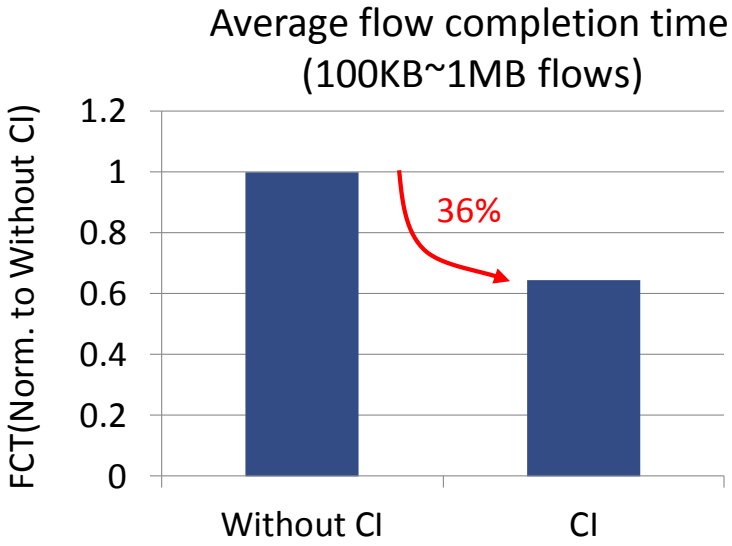
- Solution "**CI**" means PFC + ECN with CI.
- Flows go through the non-congested queue by default, and congested flows are dynamically isolated to the congested queue based on congestion.
- ECN is marked once a packet is isolated.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB, XON 4 KB;
  - CI: Low 10 KB, High 300 KB, Max Probability 1%.

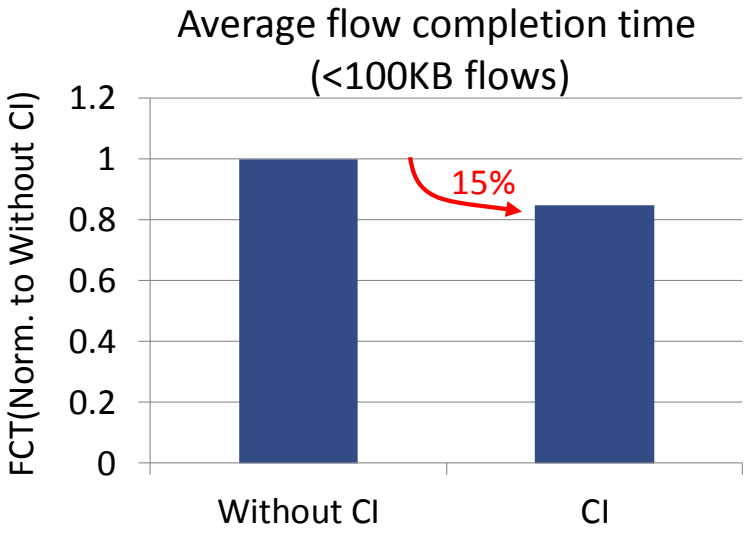
Review: Previous Data With 3 Queue Model





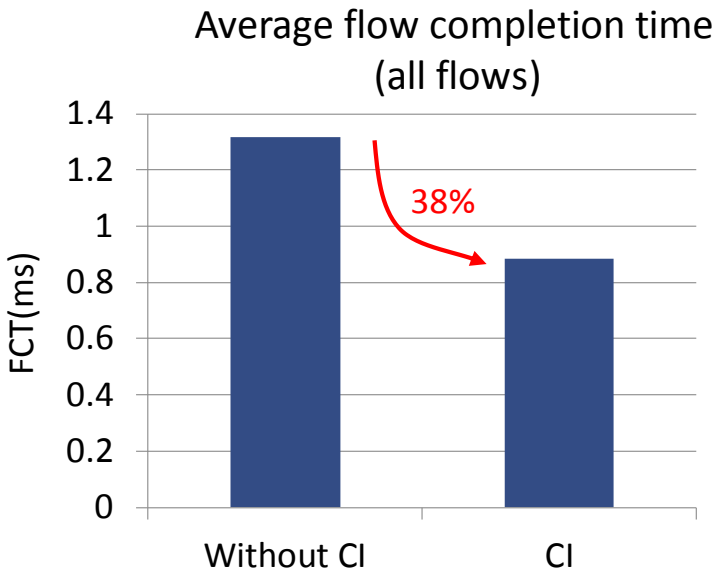


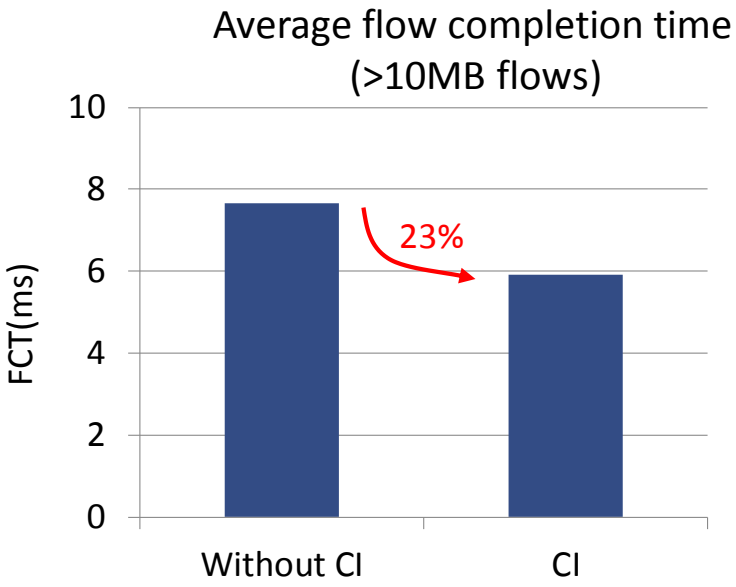


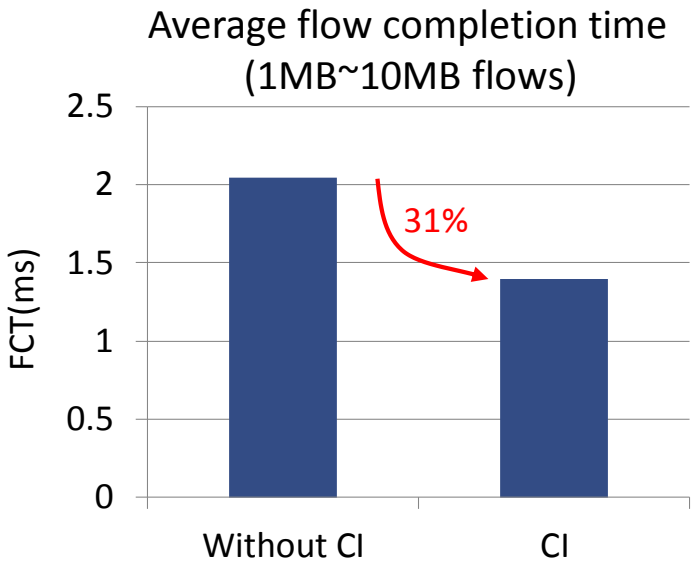


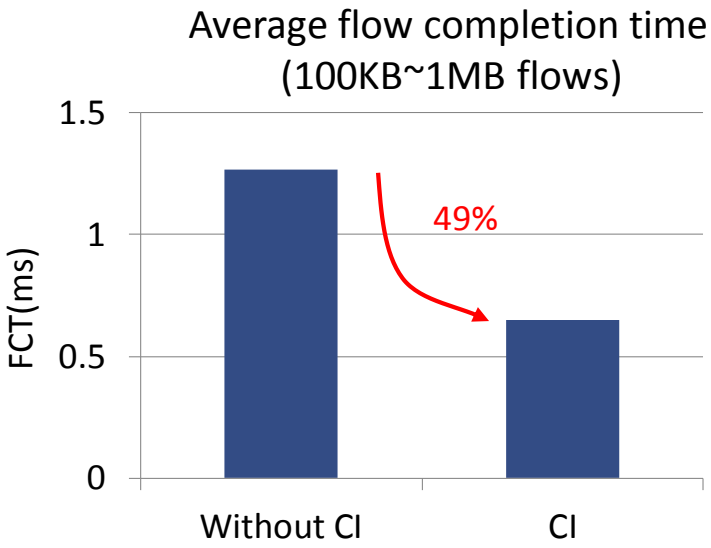
- In previous simulation with 3 queue model, Solution "without CI" and "CI" both have mice prioritization mechanism.
- The performance of the mice is not improved obviously.

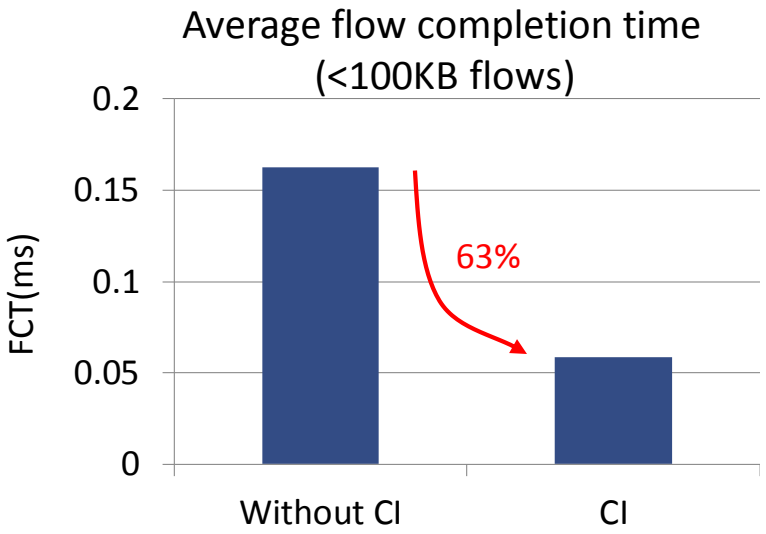
2 Queue Model





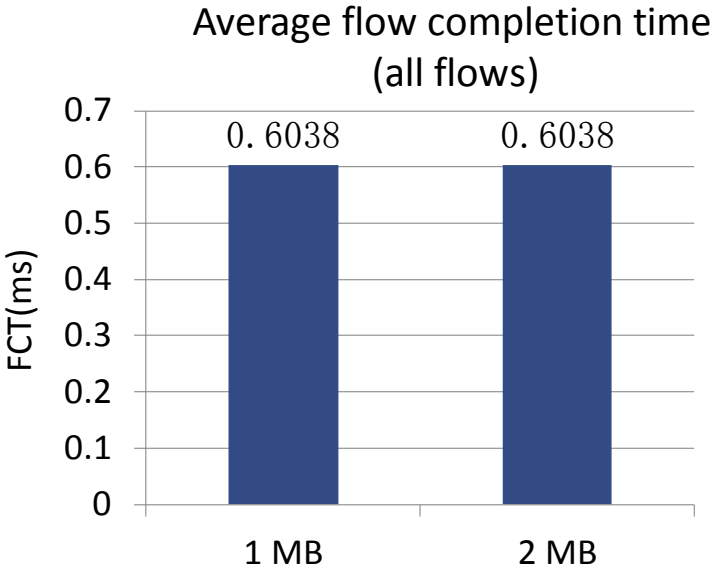


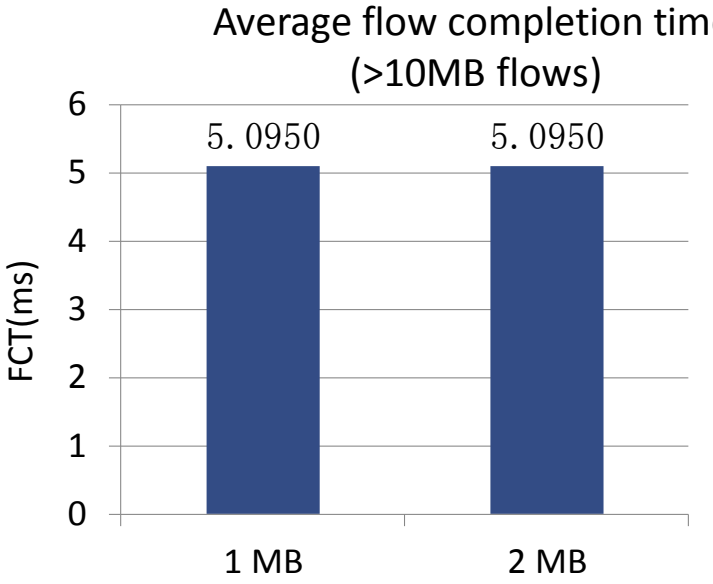


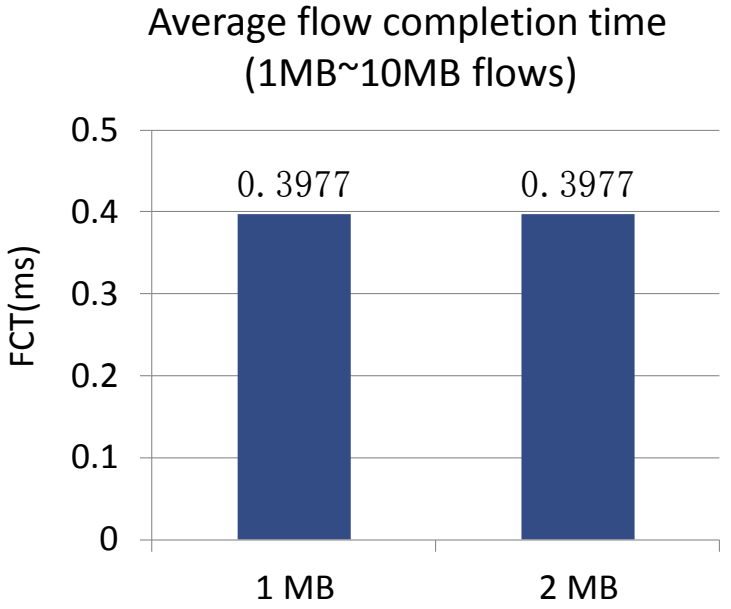


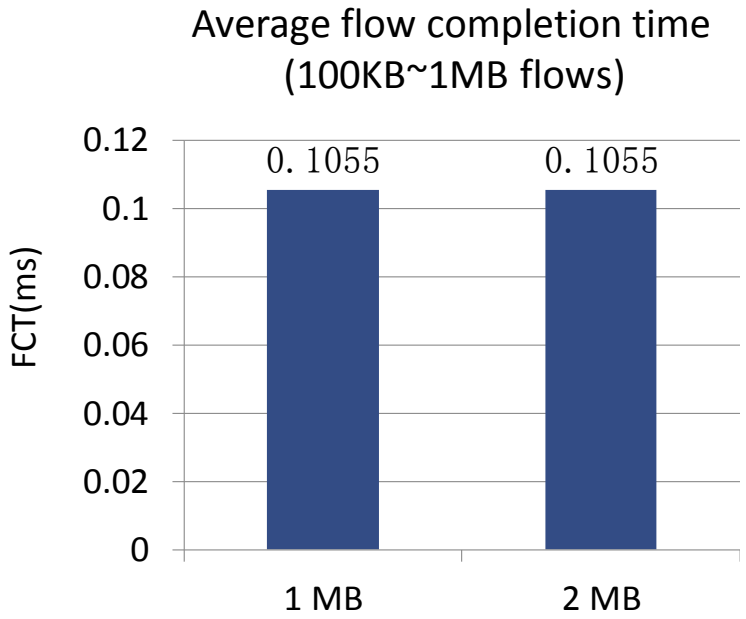
- In 2 queue model without mice prioritization, CI performs even better.
- The mice benefit the most.

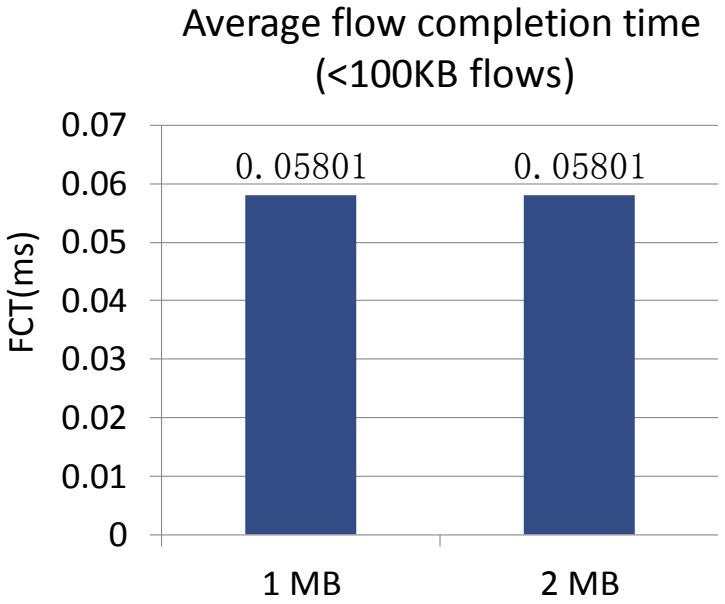
Memory Sensitivity





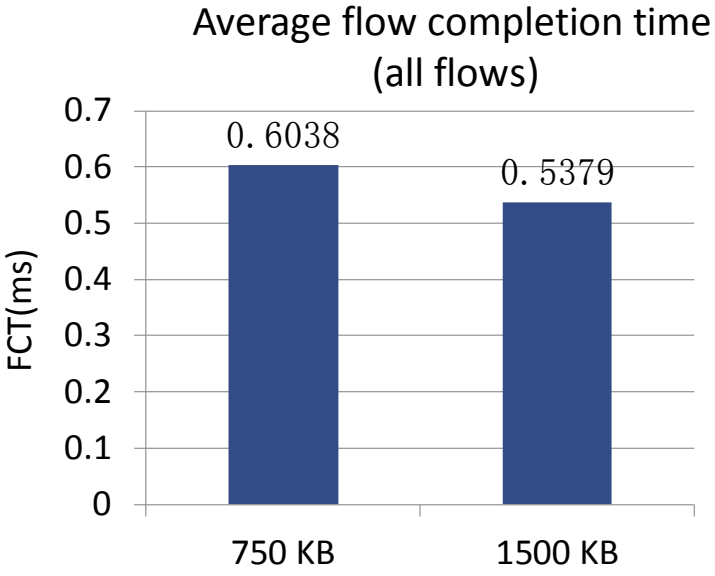


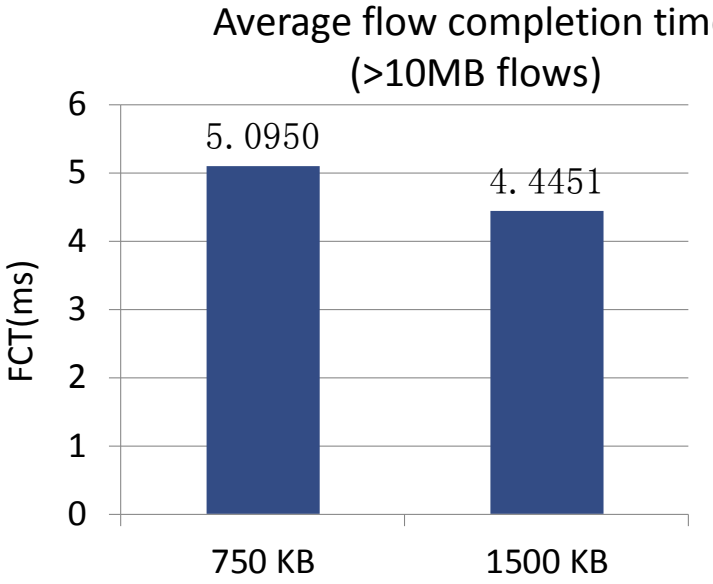


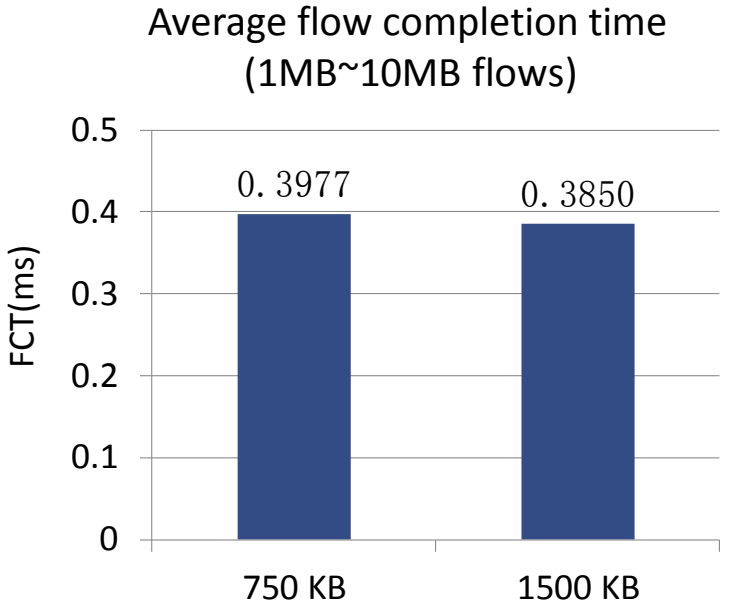


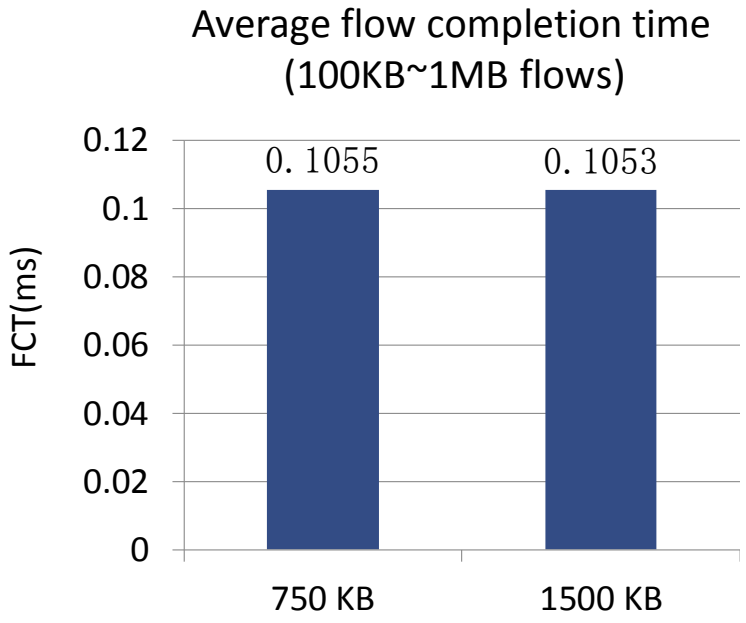
- PFC threshold: XOFF 750KB, XON 4KB.
- CI threshold: Low 10KB, High 300KB, Max Probability 1%.
- Keep the PFC and CI threshold unchanged, just enlarge queue size from 1MB to 2MB, performance does not change at all.

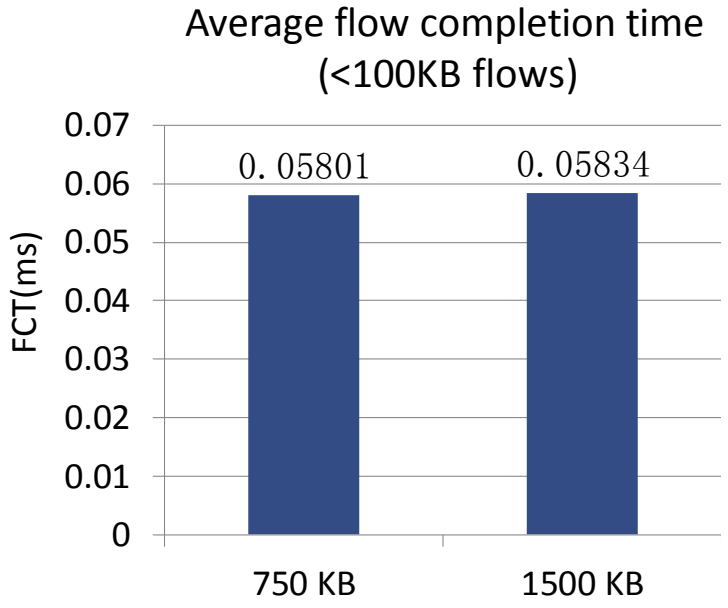
Memory Sensitivity





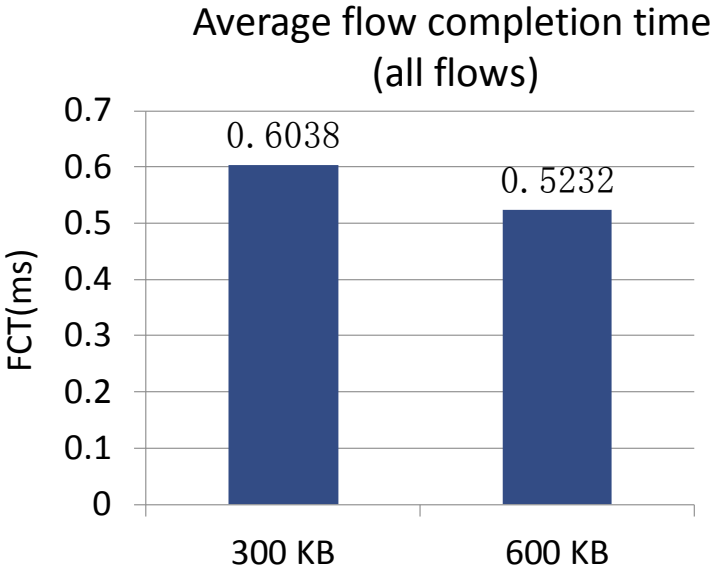


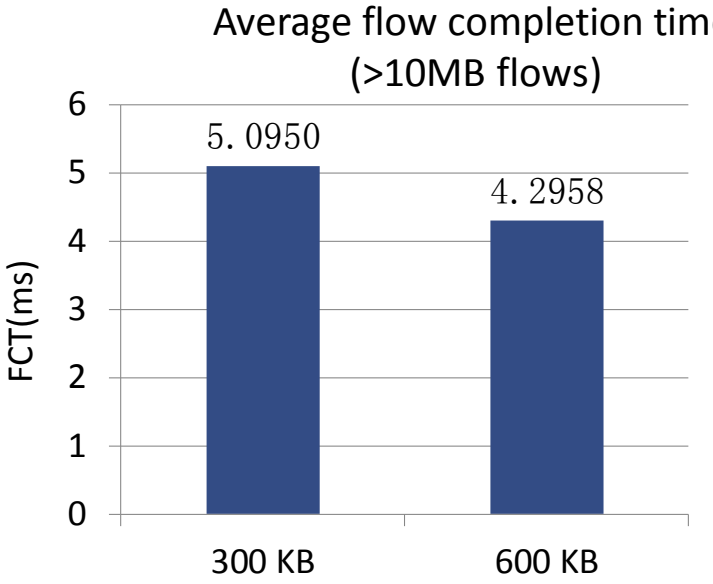


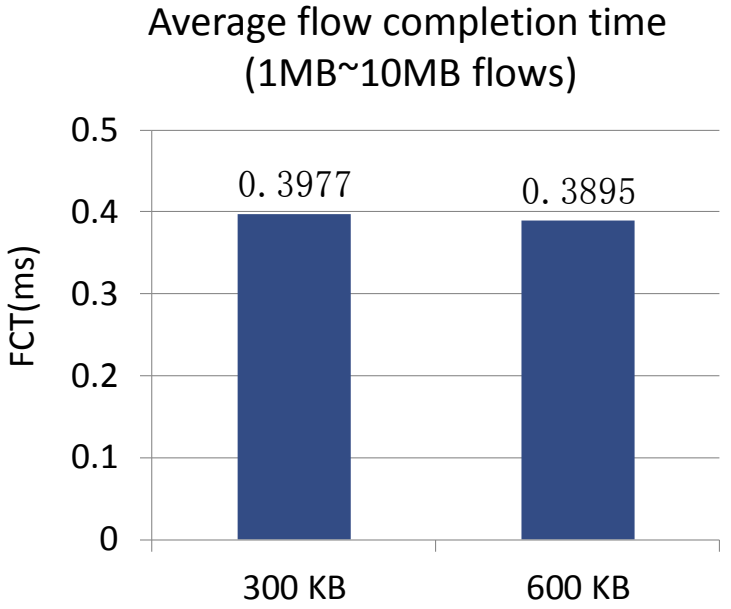


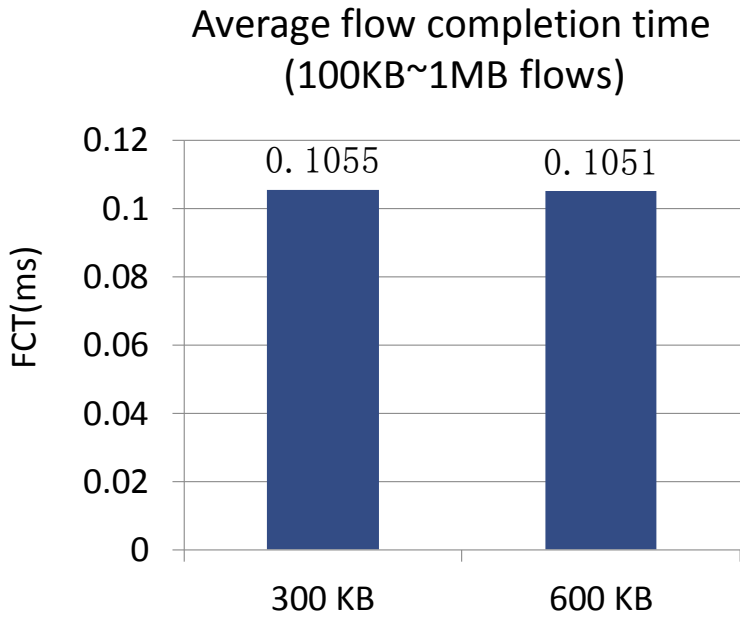
- Only change PFC XOFF threshold from 750KB to 1500KB, large flows are affected more than small flows.
- Performance improvement is achieved because fewer pause and CNP frames are triggered under 1500KB.

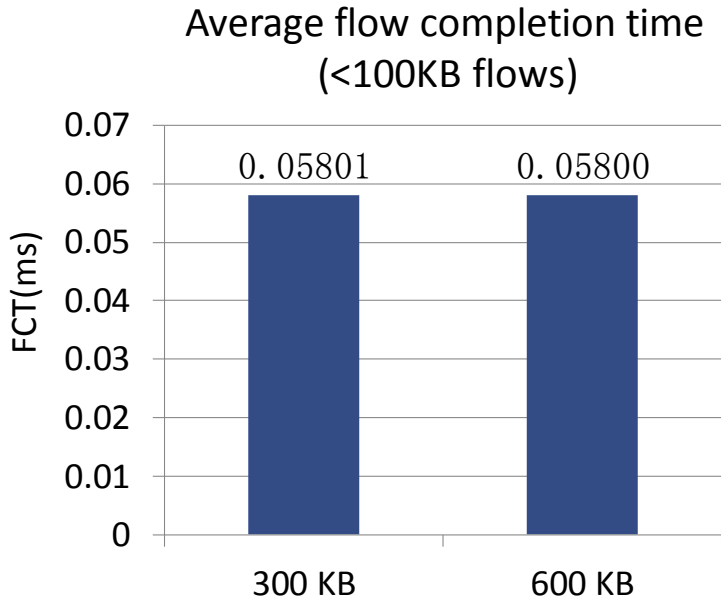
Memory Sensitivity





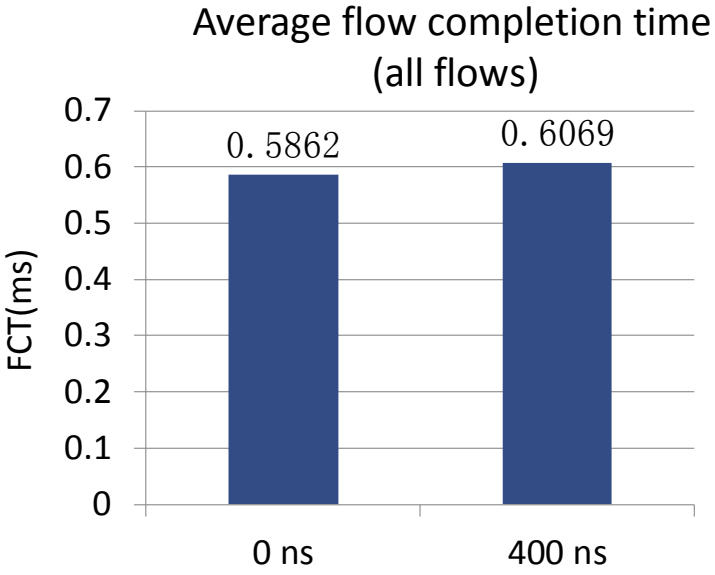


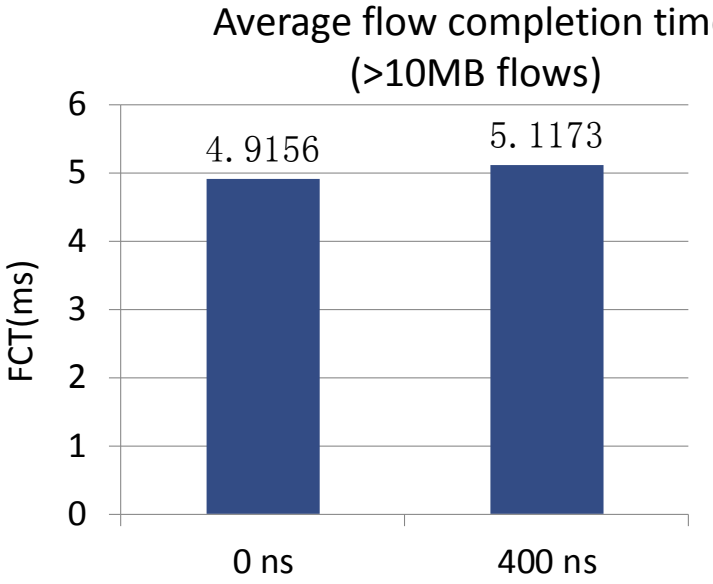


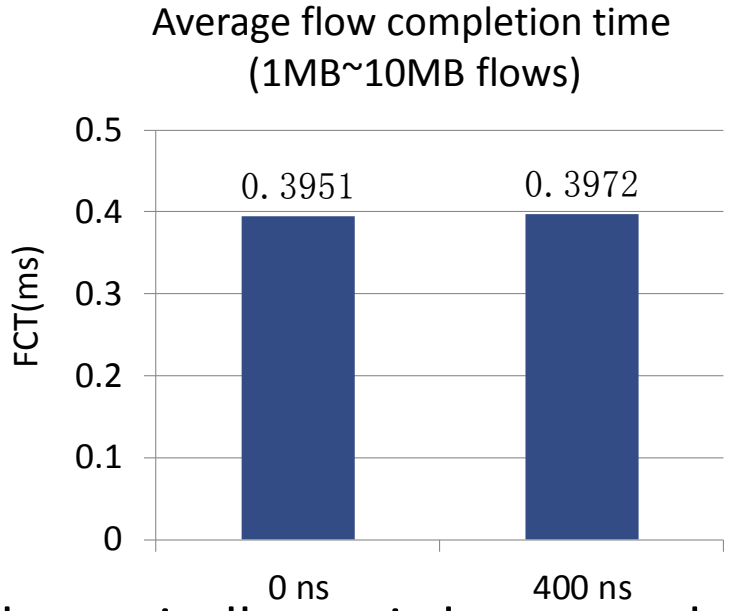


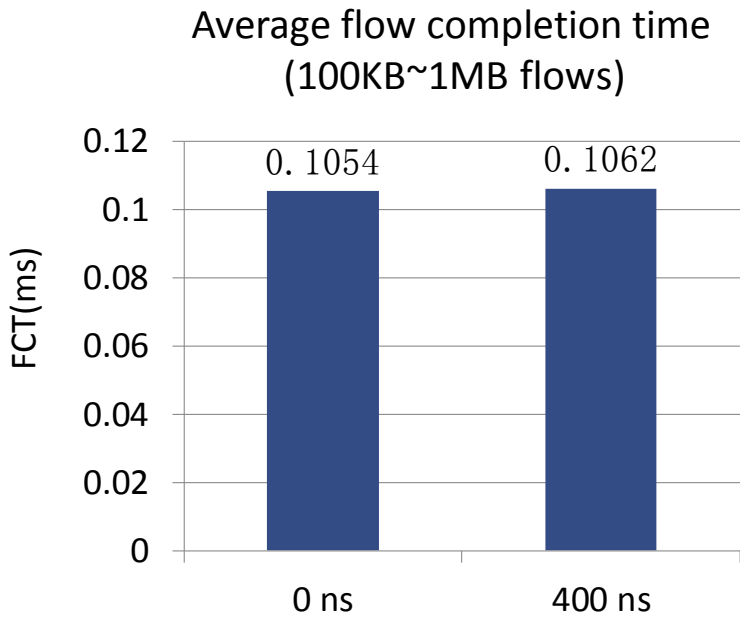
- Only change the CI high threshold from 300KB to 600KB, still large flows are affected more than small flows.
- Performance improvement is achieved because fewer pause and CNP frames are triggered under 600KB.

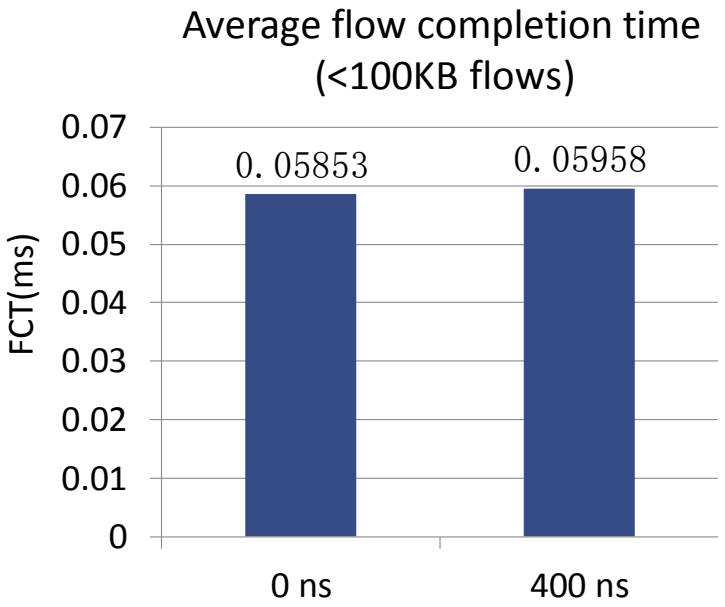
Including Static Switch Latency





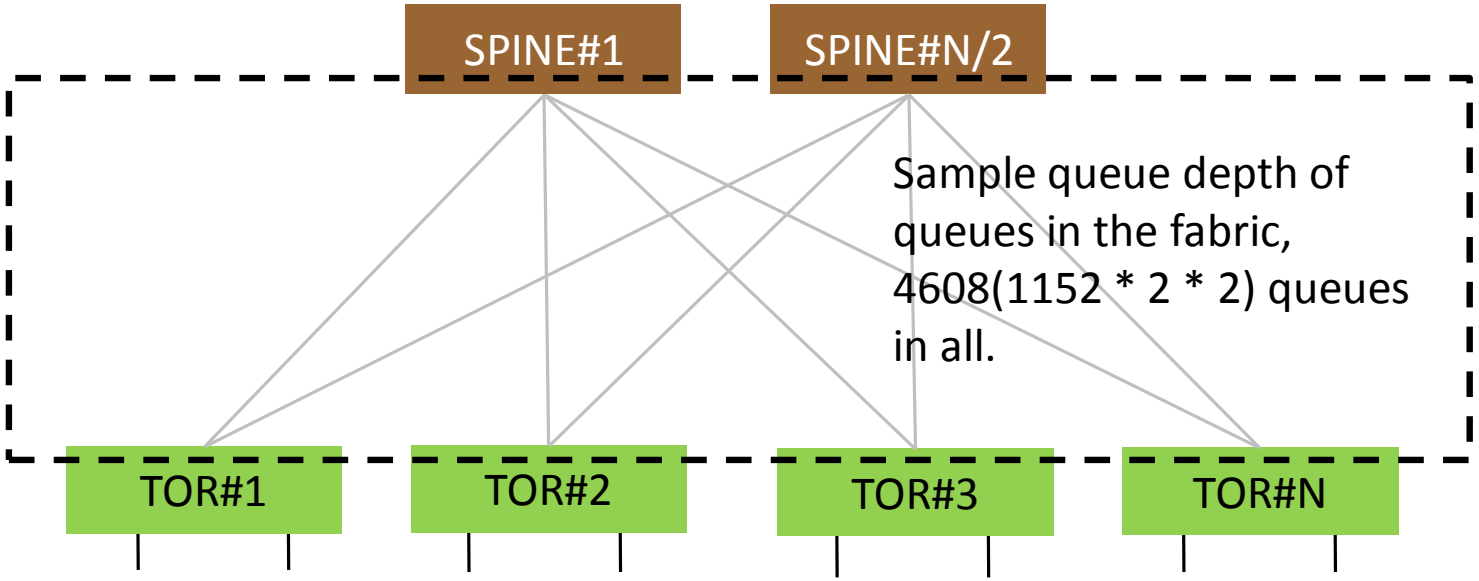






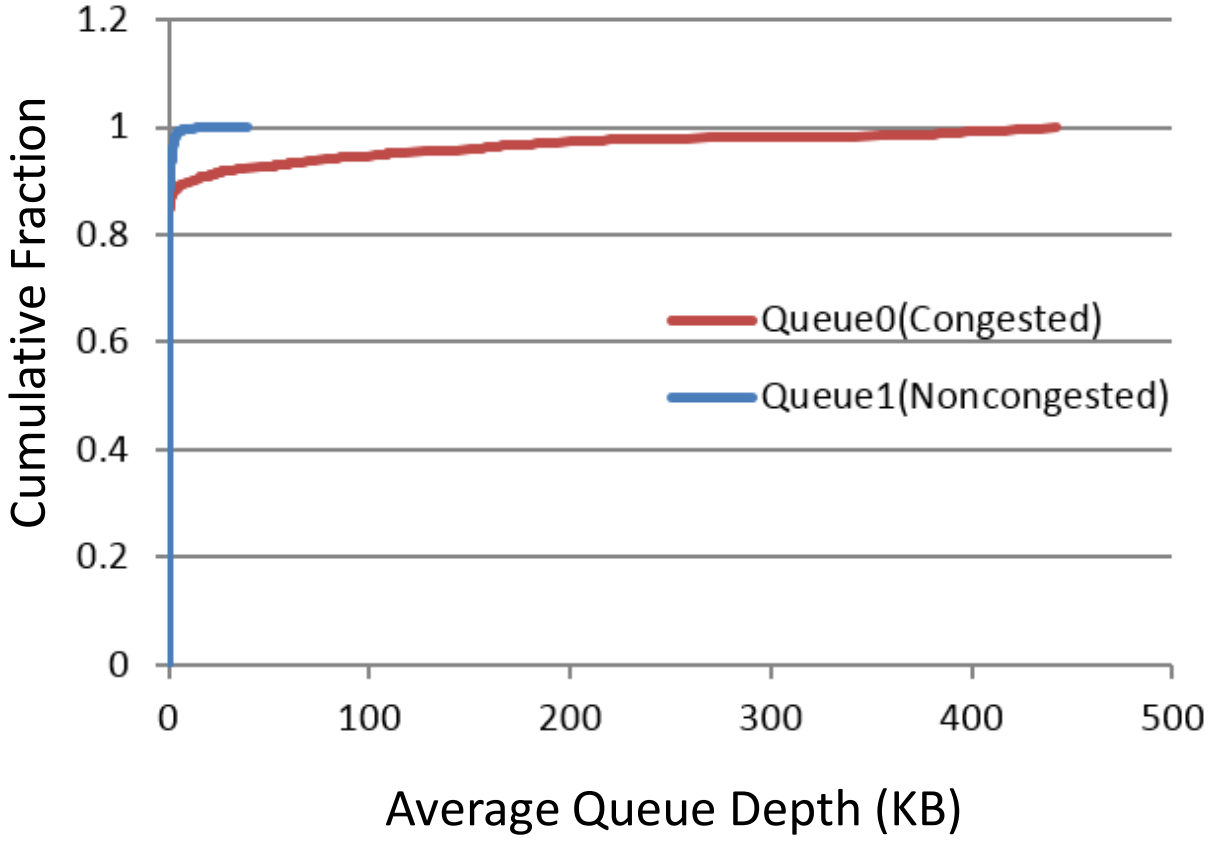
- Theoretically, static latency only increases the FCT with the latency value.
- For small flows, the increment is close to 1.2us (400ns * 3hop).
- For the elephant, the increment is 200us, which is much more than 1.2us, mainly because of more pause frames.

Queue Depth Comparison

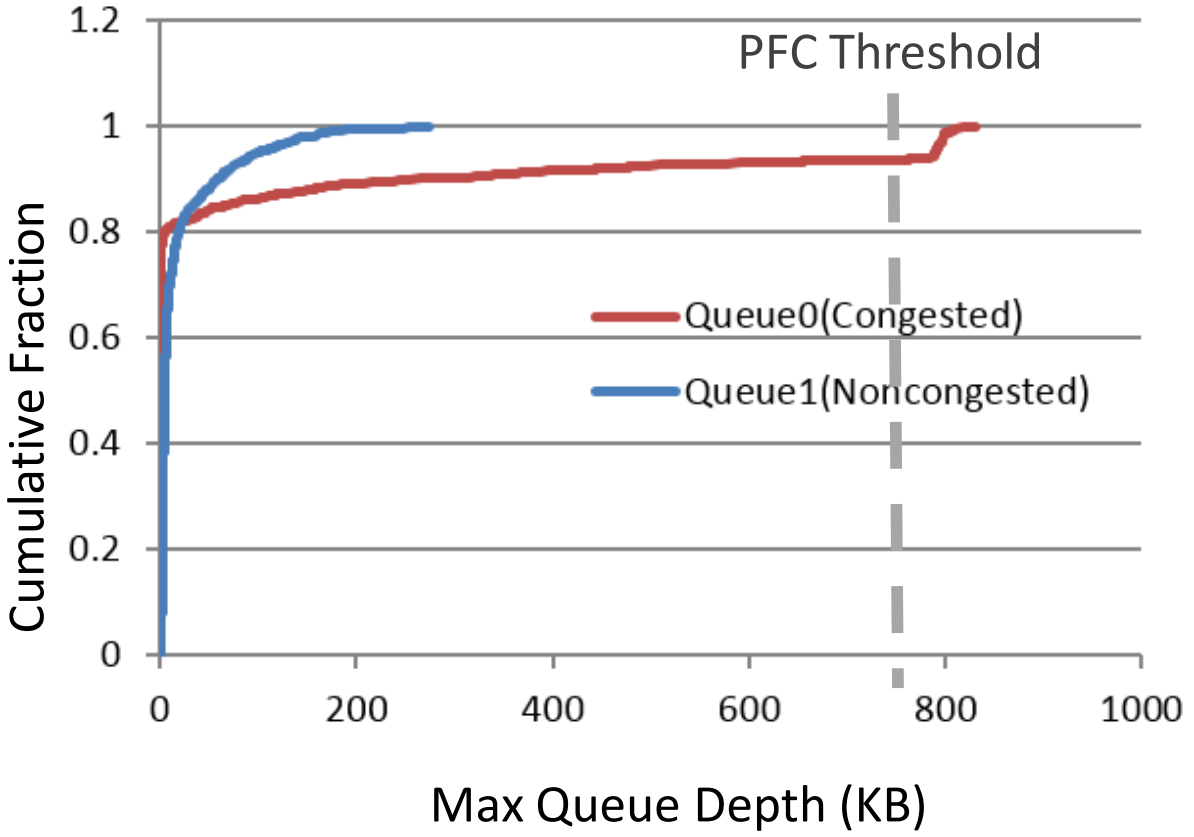


- In this comparison, we have *N* equal to 48, which means 1152 servers, 48 TORs and 24 SPINEs.
- Sample the queue depth in the fabric periodically; record the number of sample times, cumulative queue depth and max queue depth.
- Queue setting:
  - Queue size: 1 MB;
  - PFC threshold: XOFF 750 KB, XON 4 KB;
  - CI threshold: Low 10 KB, High 300 KB, Max Probability 1%.

CI: Queue Depth

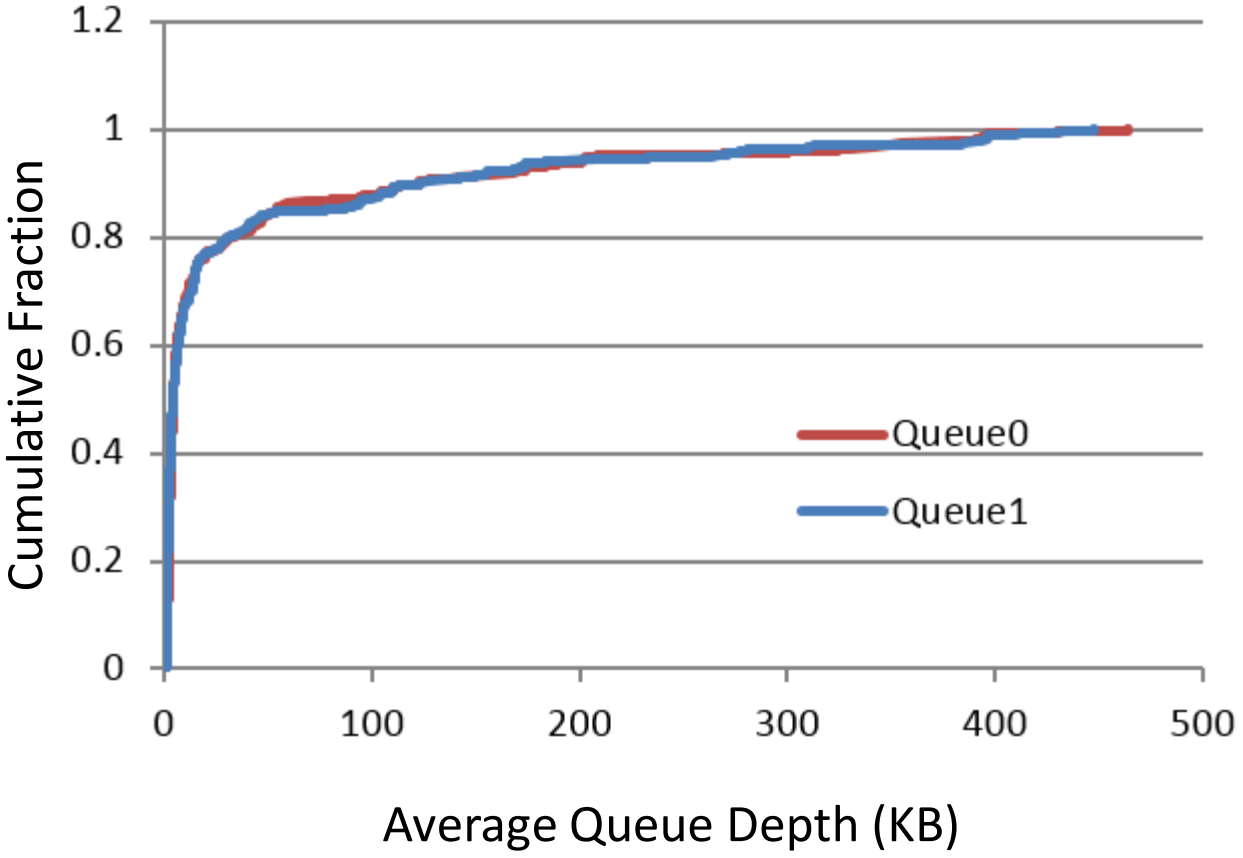


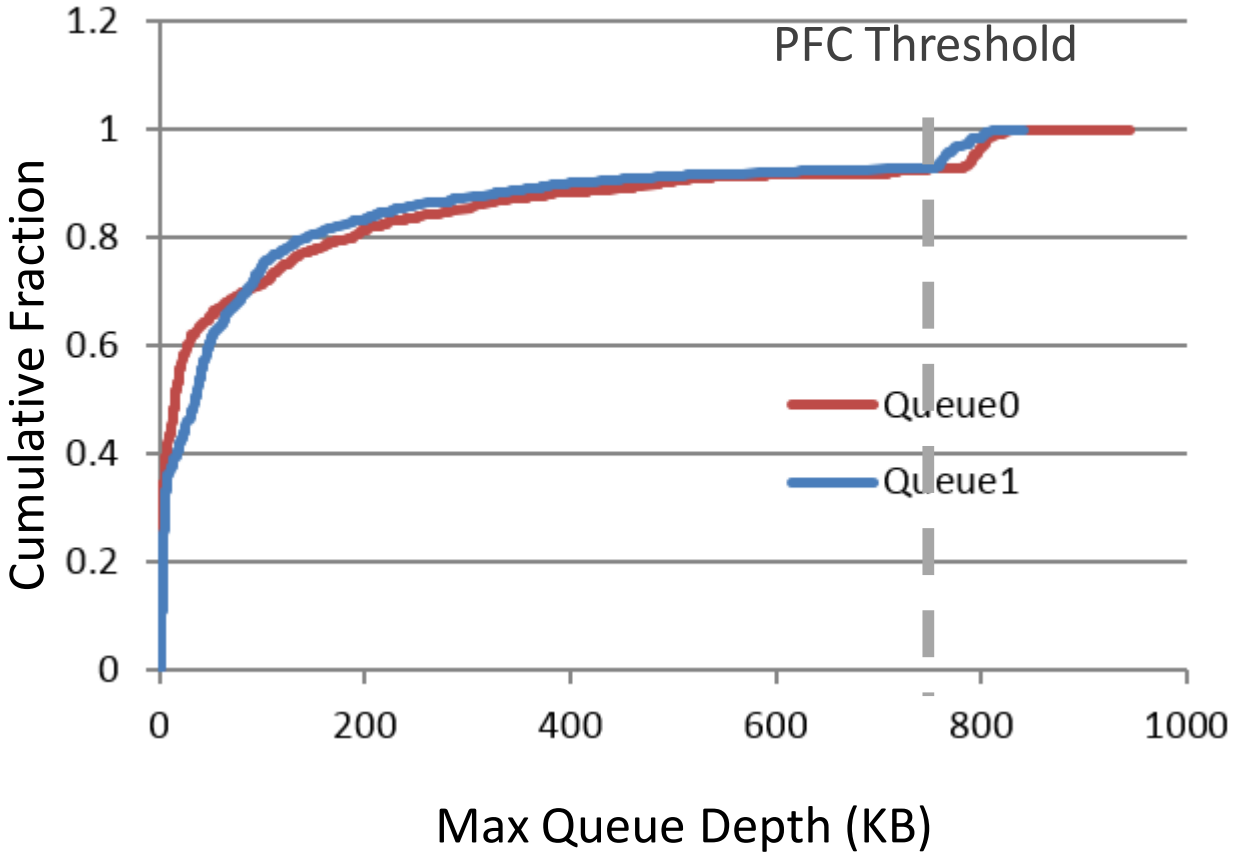
- CI makes the average queue depth of non-congested queue quite low.



- The max queue depth of non-congested queue never exceeds the PFC threshold due to the immediate isolation of congested flows.

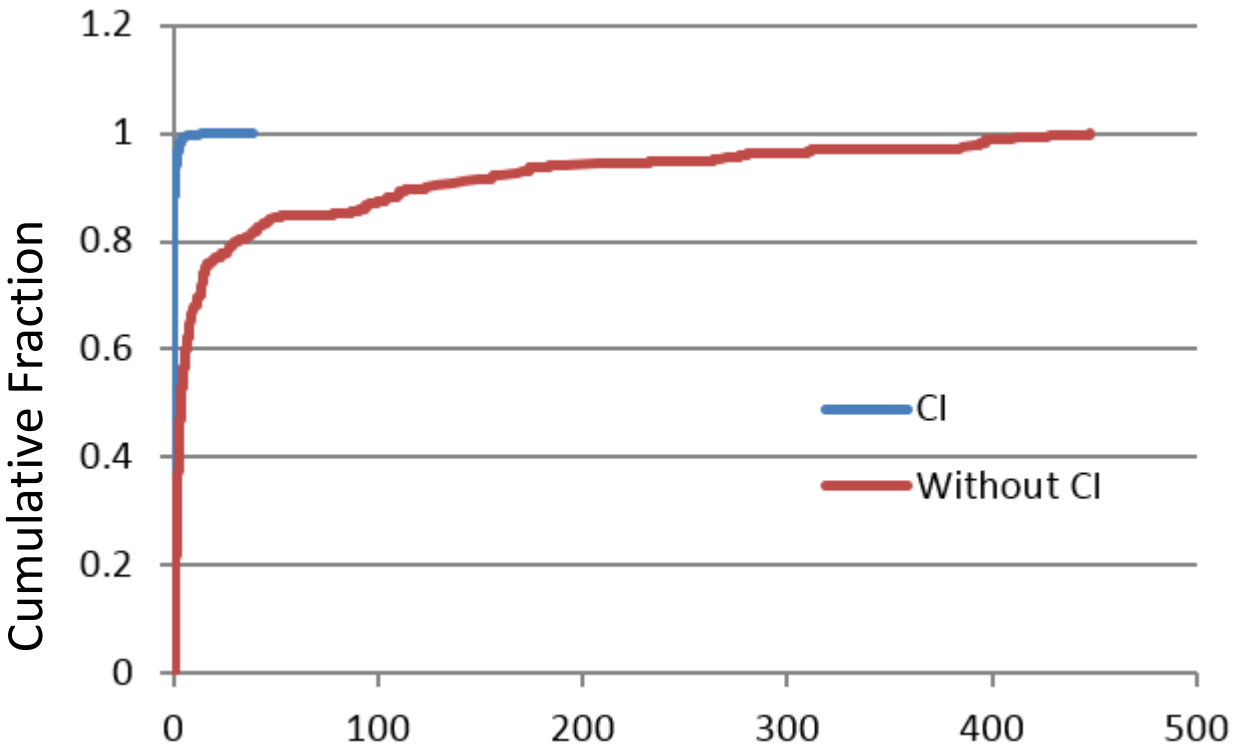
Without CI: Queue Depth



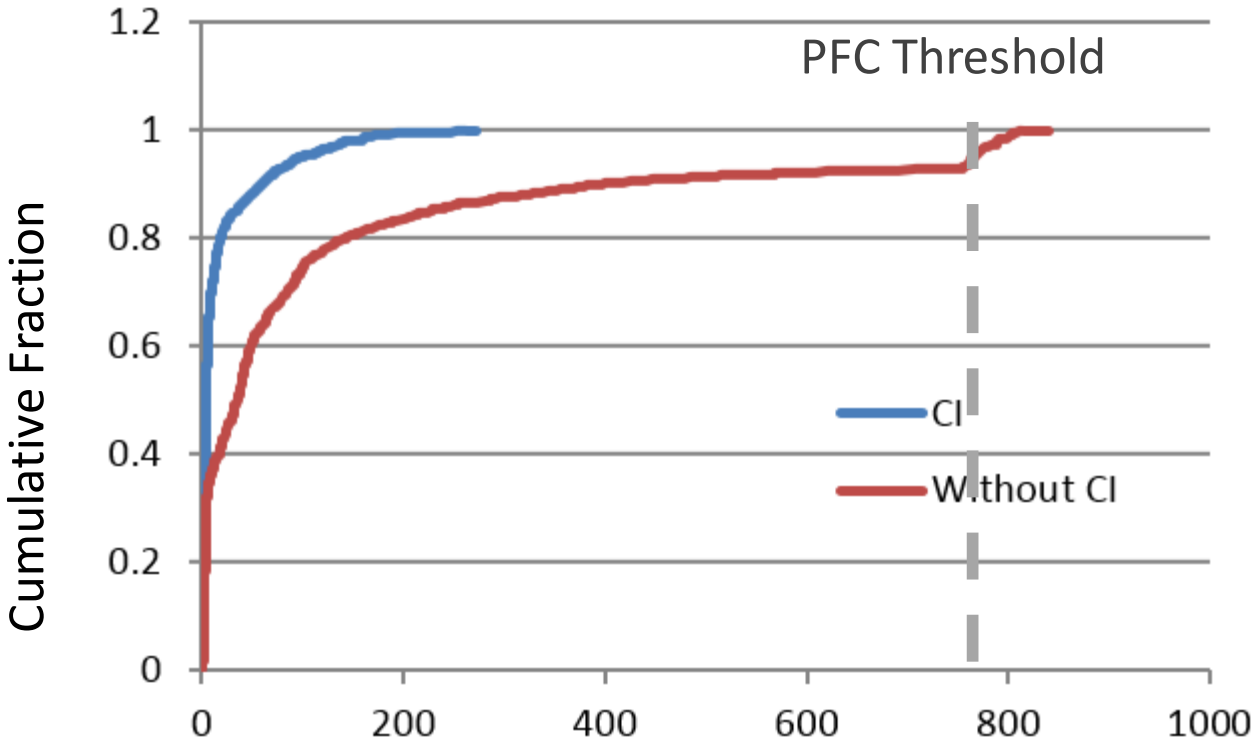


- The queue depth distribution of two queues is similar as expected.

Queue Depth Comparison



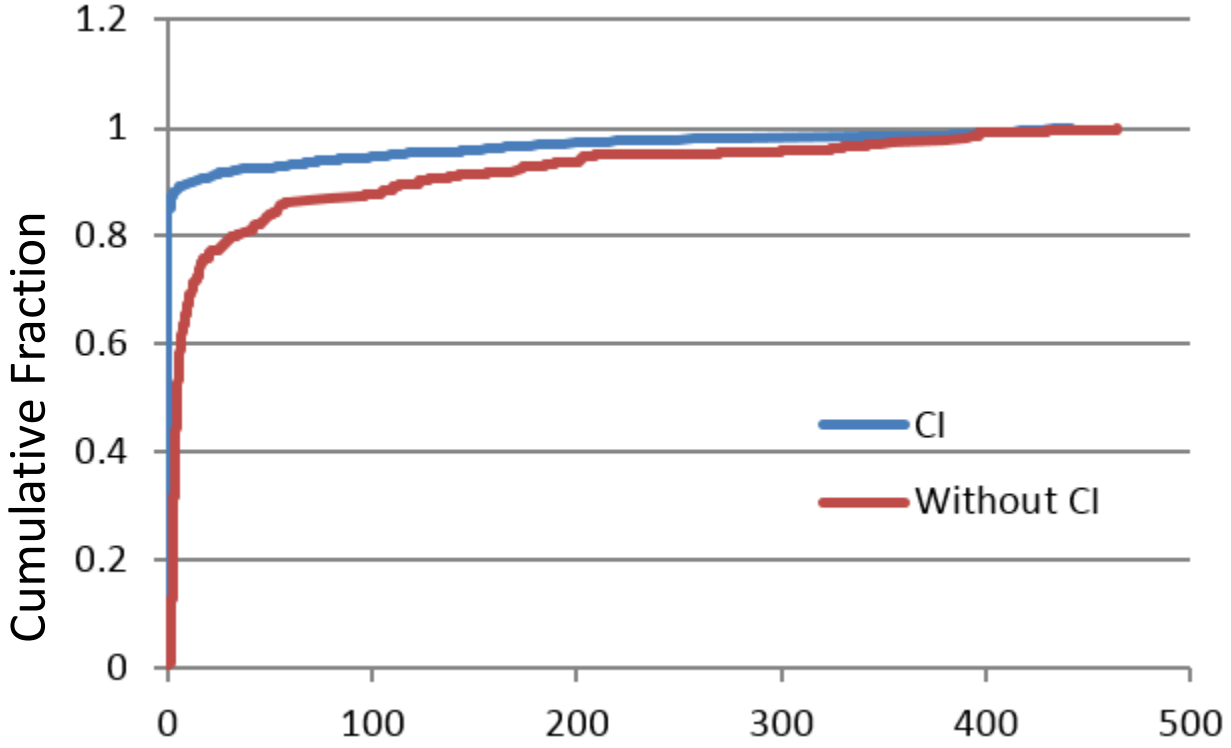
Queue 1 (Non-congested) Average Queue Depth (KB)



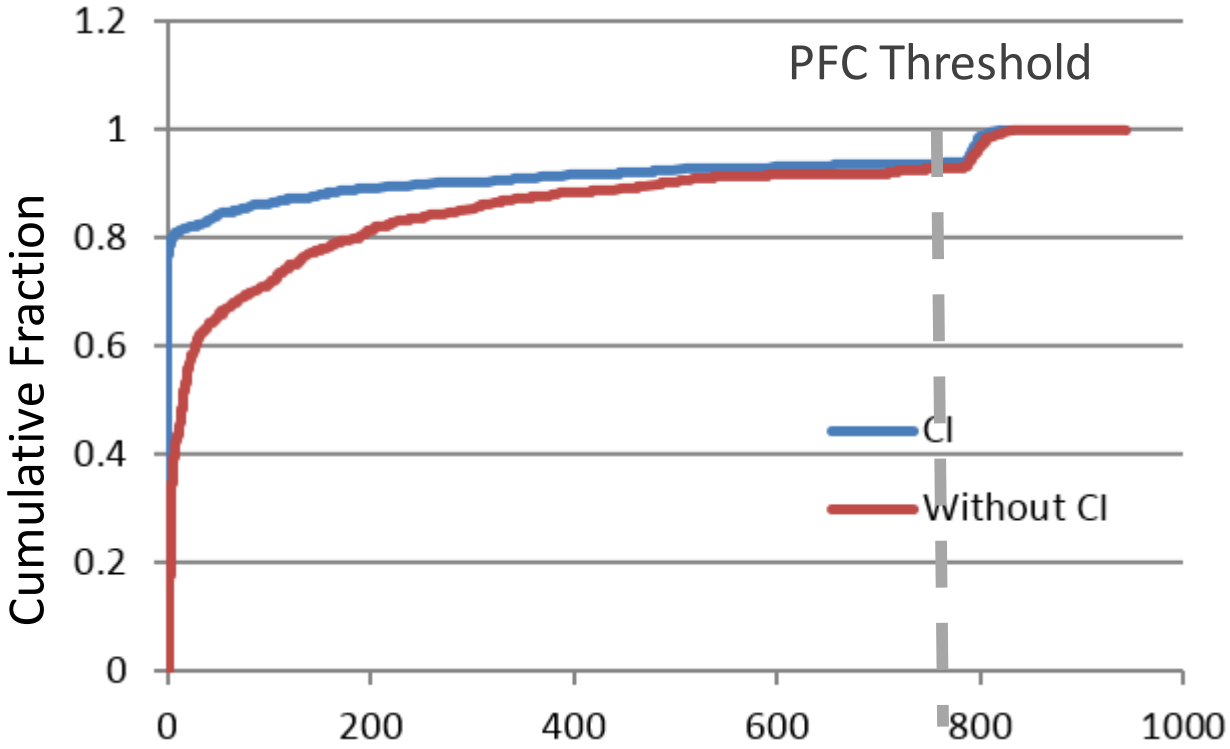
Queue 1 (Non-congested) Maximum Queue Depth (KB)

- For Queue 1 (non-congested), CI maintains more shallow queue depths as compared without CI.
- With CI, HOLB never occur in Queue 1 because PFC XOFF threshold never be exceeded.

Queue Depth Comparison



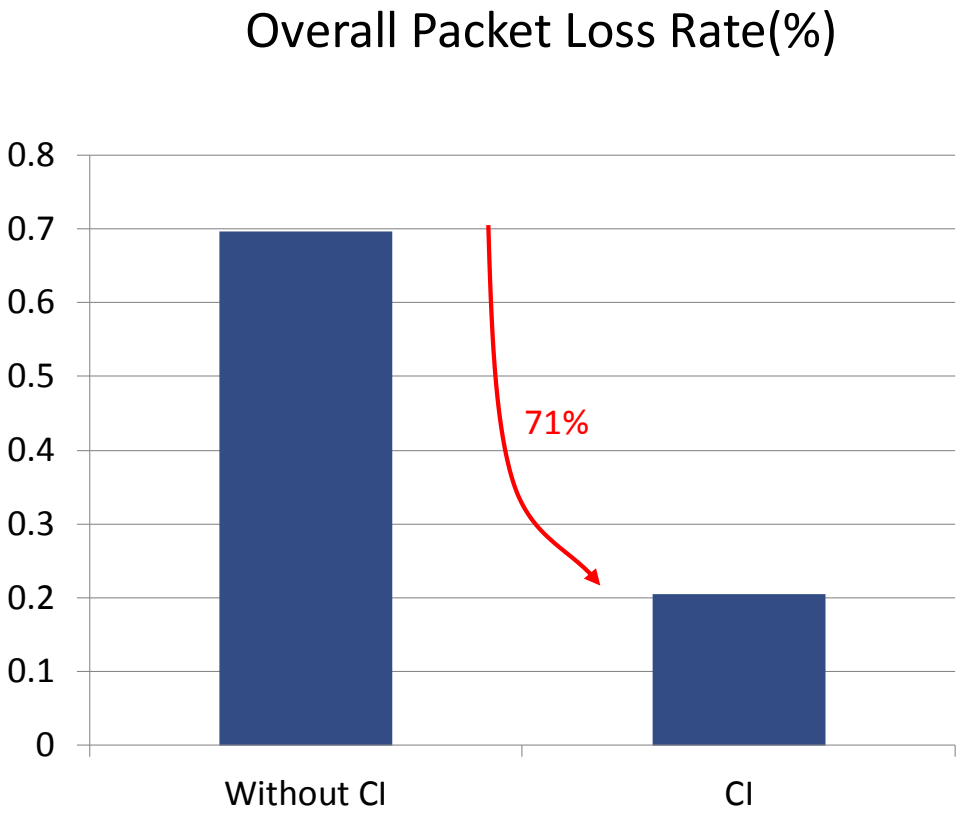
Queue 0 (Congested) Average Queue Depth (KB)



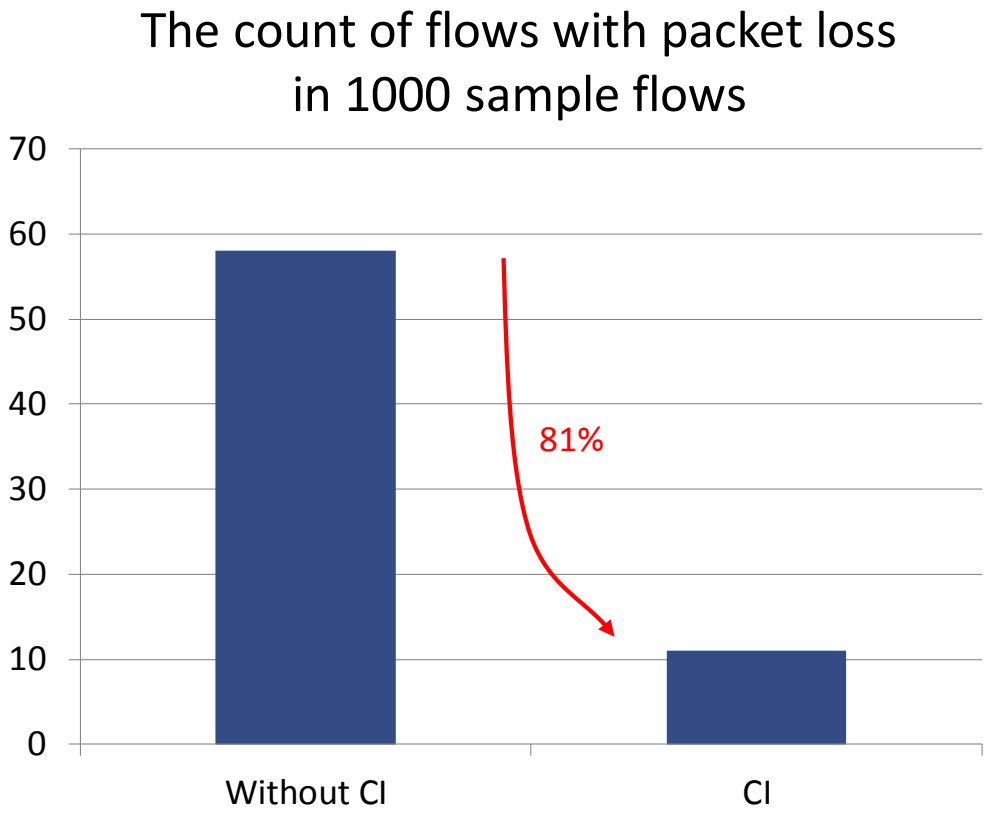
Queue 0 (Congested) Maximum Queue Depth (KB)

- For queue 0(congested), fewer queues across the fabric suffer from congestion because with CI fewer flows are in the congested queue.
- HOLB is limited to the congested queues holding congested flows.

Lossy Scenario (No PFC)



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



- CI reduces the number of flows experiencing packet loss.
- Only packets from congested flows are dropped. Non-congested queue never fills.

Summary

- **Two queue model** (congested and non-congested queues; no mice prioritization)
  - CI achieves even better performance; especially for the mice.
- **Memory sensitivity**
  - Threshold setting seems critical for the congested flows, but not for the non-congested flows.
- **Including static switch latency**
  - Static latency influence the result very little, so the analysis result will not alter.
- **Queue depth**
  - CI can keep most queues low depth.
- **Lossy scenario (no PFC)**
  - CI improves performance by reducing overall packet loss and flows experiencing packet loss.

Questions?