Simulation Analysis of Congestion Isolation

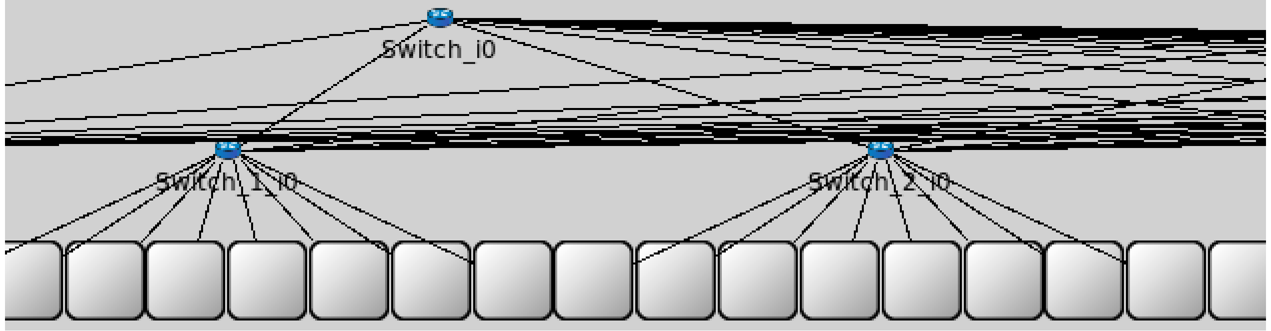
Kevin Shen

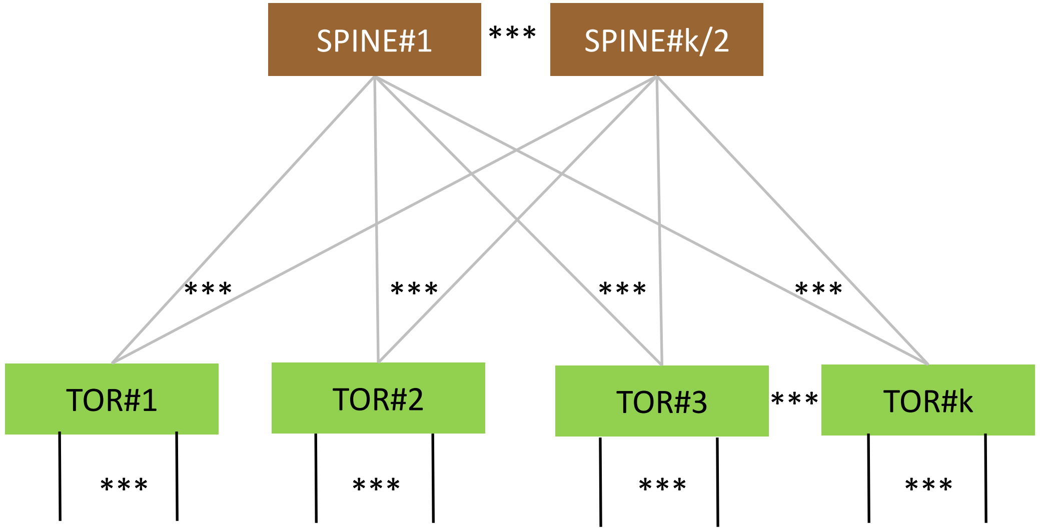
kevin.shenli@huawei.com

IEEE 802.1 DCB

Orlando Florida, November 2017

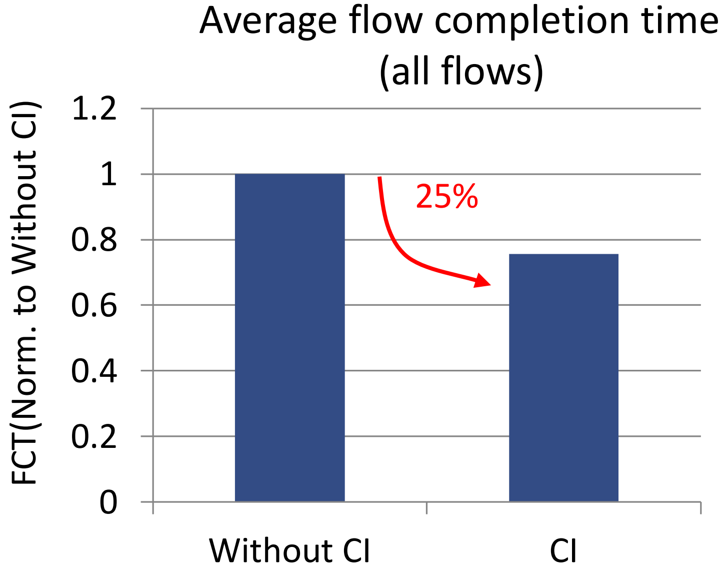
Simulation Set-up

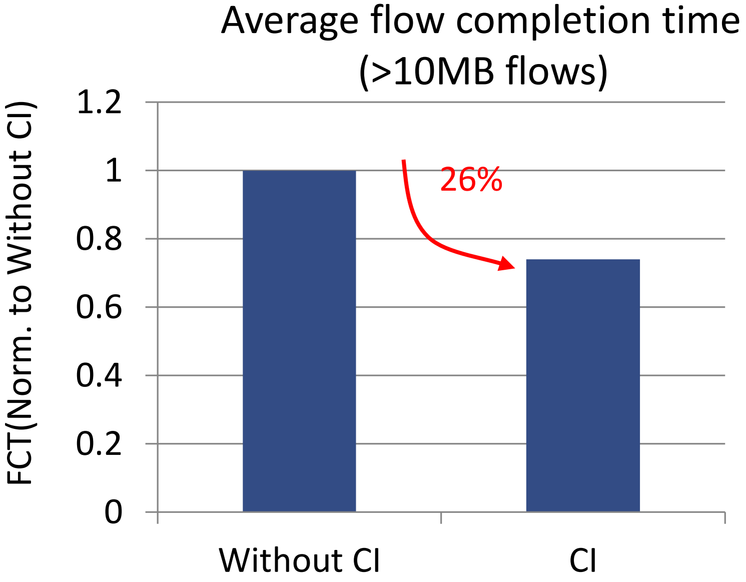


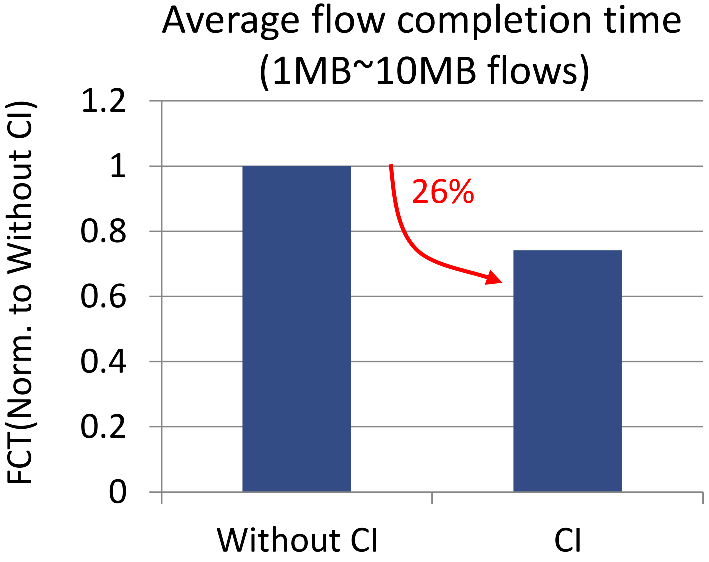


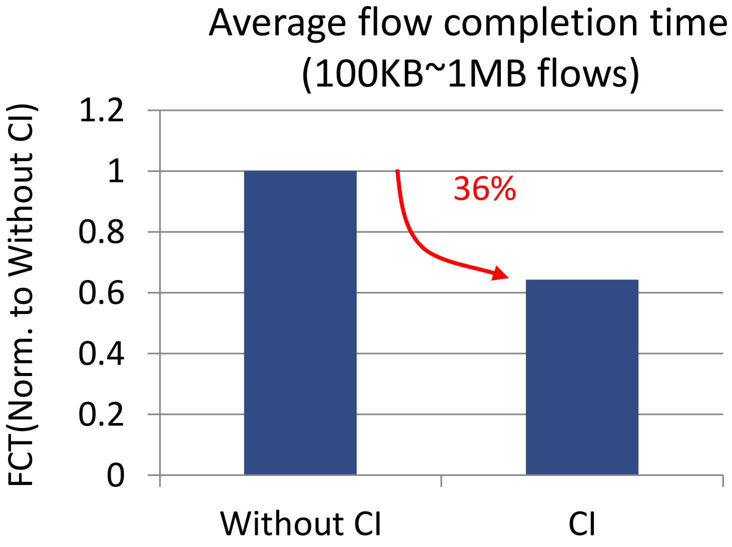
- OMNET++ Platform
- 2 Tier CLOS： 100G interface with 200ns of link latency 200ns(about 40m)
- Scale： 128 ~ 1152 servers, 24 ~ 72 switches
- Traffic Patterns: Data Mining Application, Several regional all to all with some persistent incast

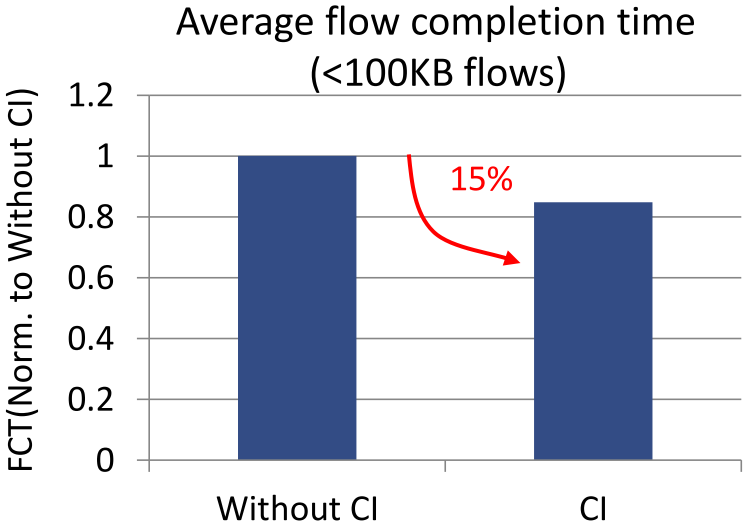
Recall the simulation data





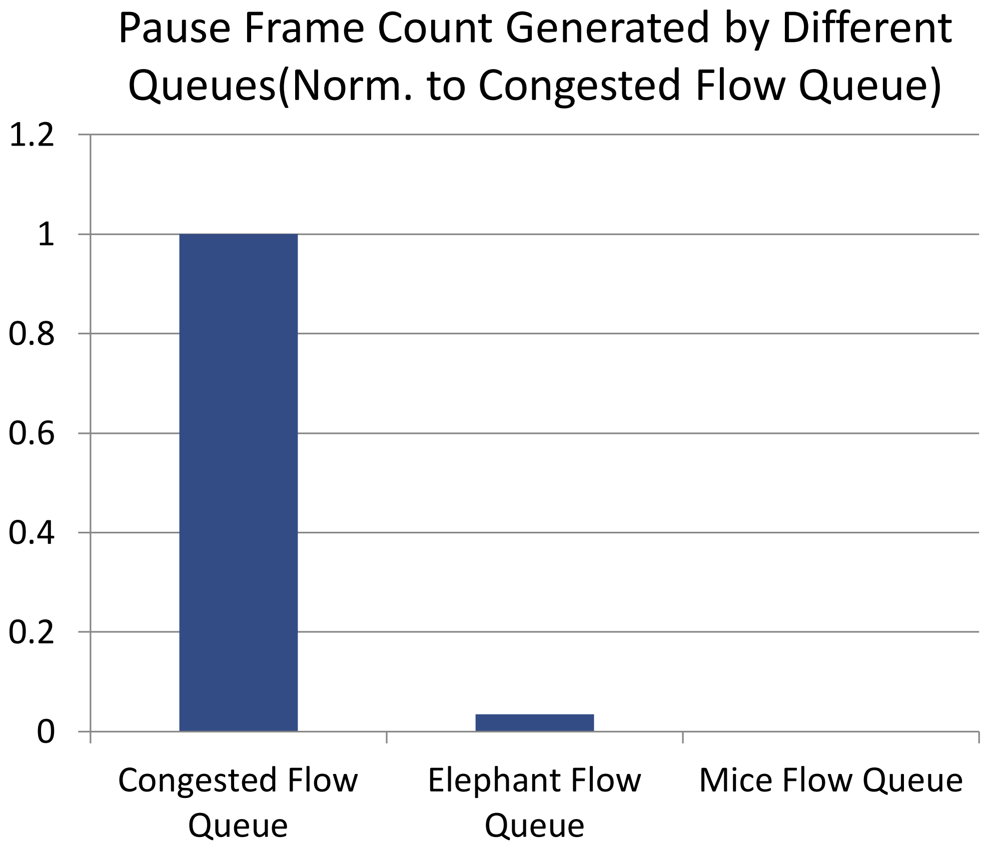


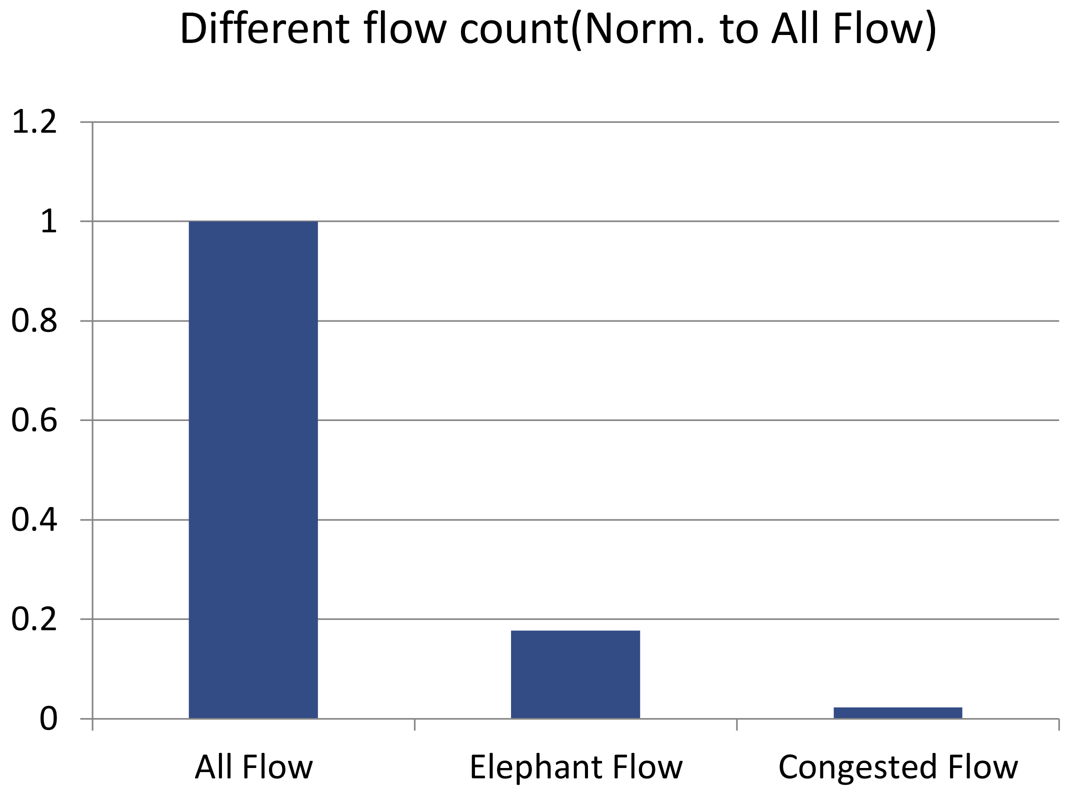




- CI mitigates HOLB, which can improve the performance of all kinds of flows

Recall the simulation data





- 96.6% of the pause frames are generated by congested flow queues.

- The count of isolated flows is quite small. The proportion is 2% for total flows , and 12% for large flows.
- So the HOLB only occurs among the congested flows.

Questions raised in last meeting

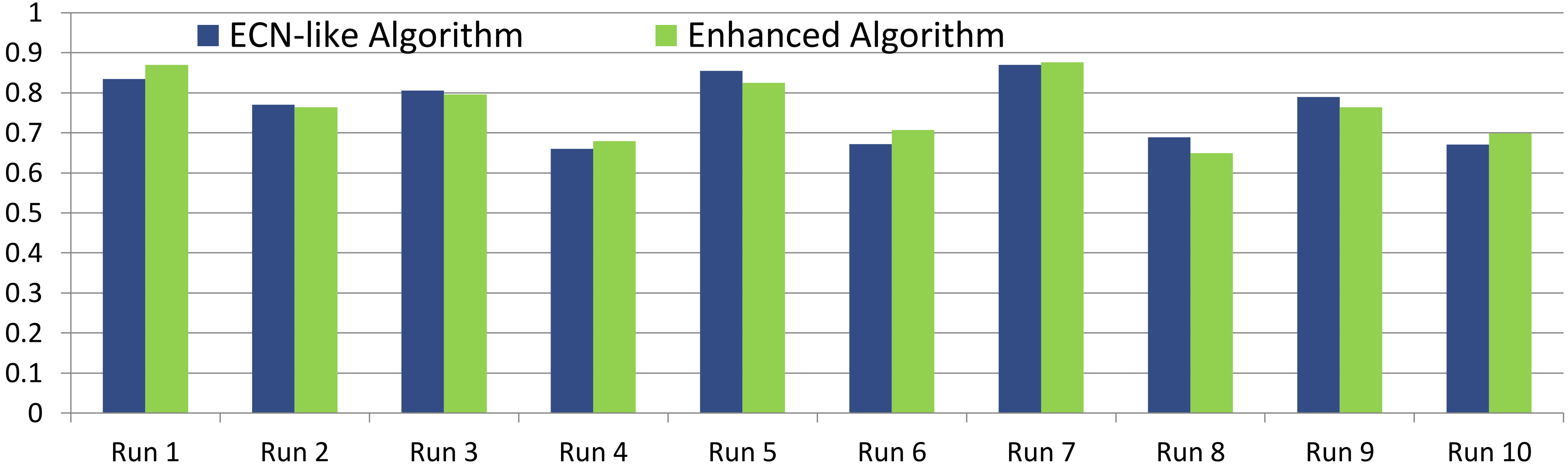
- ECN-like algorithm is too random and may pick out wrong flow. Is there a better way?

- Compared with pause frame count, how about the queue XOFF duration?

A better congested flow selection scheme

- Counters in flow table to count the bytes buffered in the queue for each flow.
- When a packet enqueues, increase the counter by the bytes of the packet. When a packet dequeues, decrease the counter by the bytes of the packet.
- Record several maximum flows in the queue.
- When congested, isolate detected congested flows.

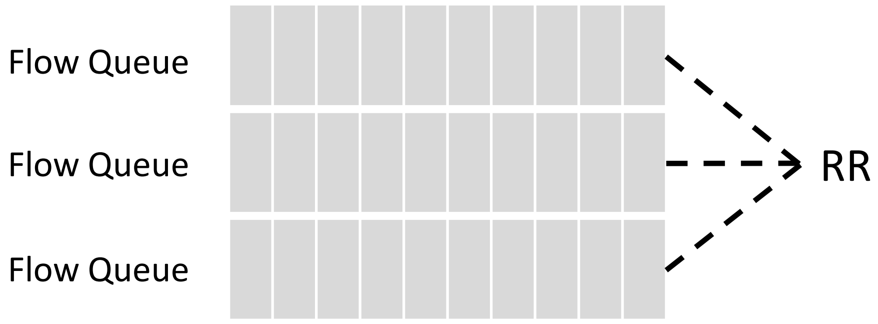
A better congested flow selection scheme


Average flow completion time(ms)
(all flows)

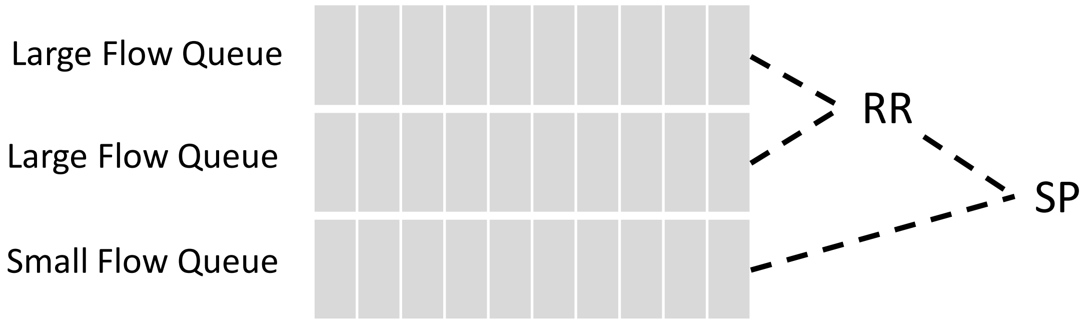
- A sophisticated congested flow selection algorithm brings little help. It’s not so critical.
- Mostly because if CI select a wrong flow, it will select another one.

Compared Solutions

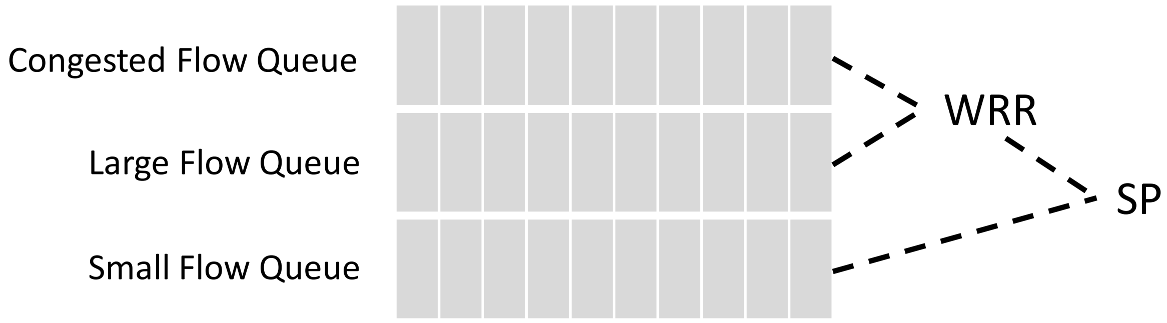
RR: Round Robin SP: Strict Priority WRR: Weighted Round Robin



- Solution 1: PFC + ECN



- Solution 2: PFC + ECN with mice prioritization



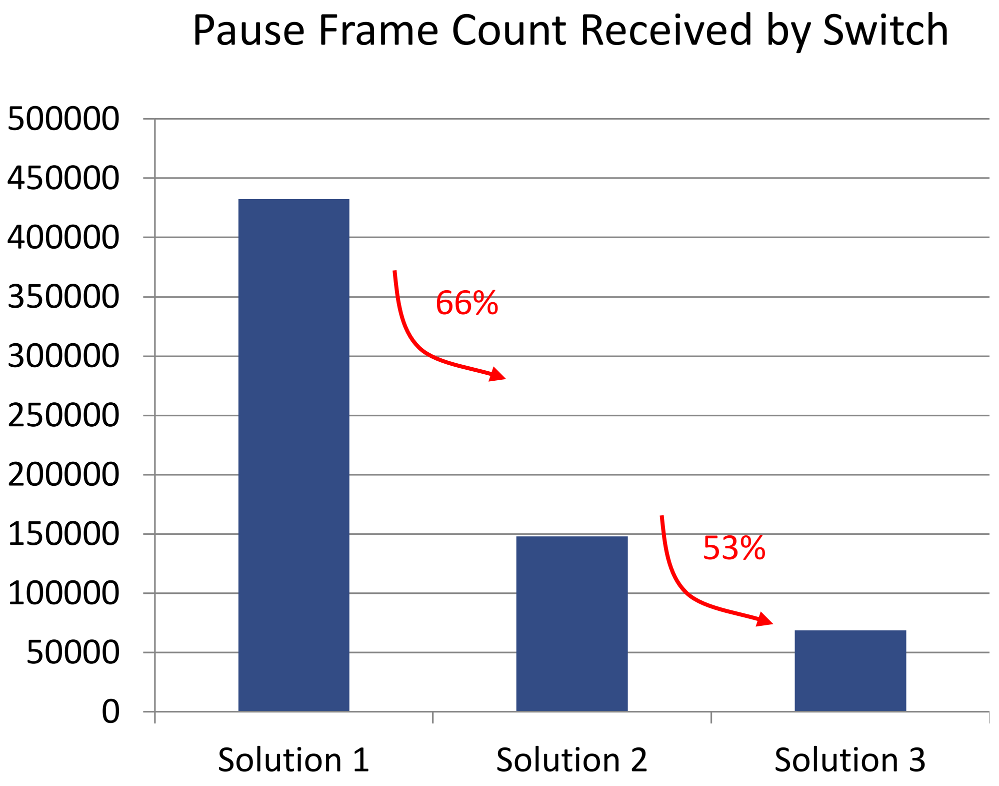
- Solution 3: PFC + ECN with mice prioritization and CI

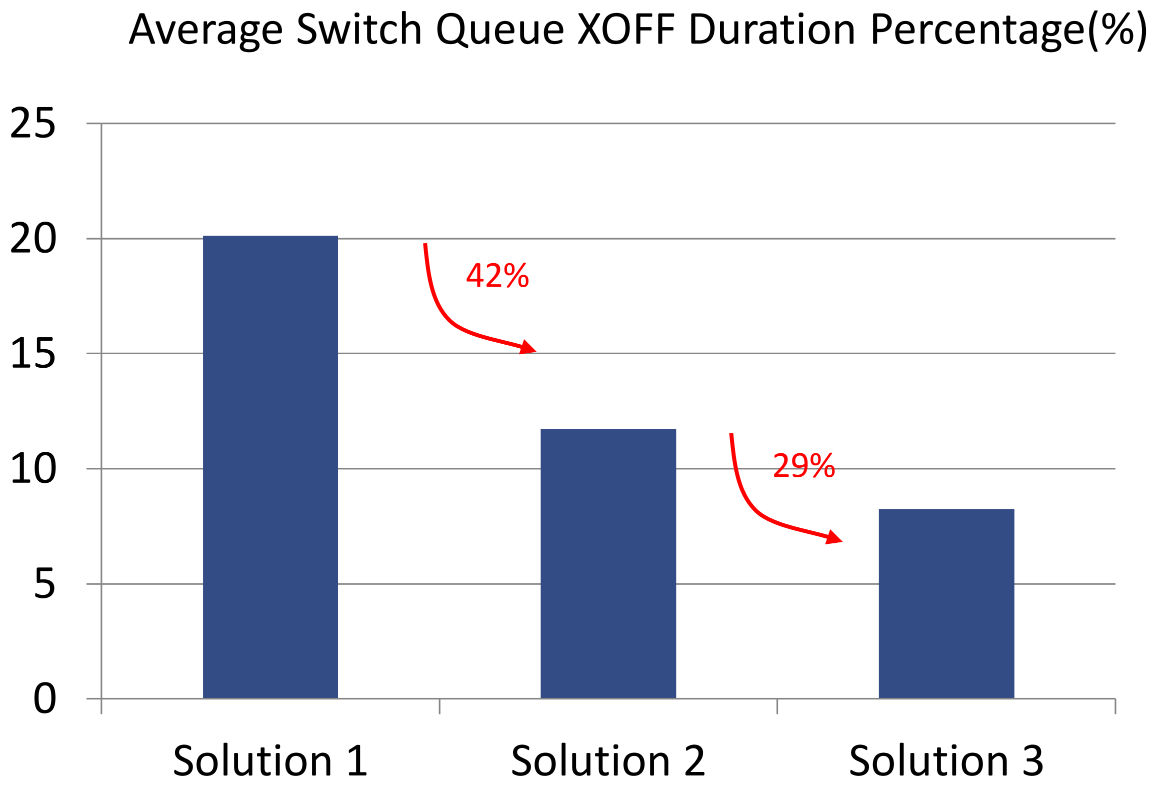
Compared with different metrics:

- FCT(Flow Completion Time)
- Pause Frame Count
- Queue XOFF Duration
- CIP Count

Solution Comparison

- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI

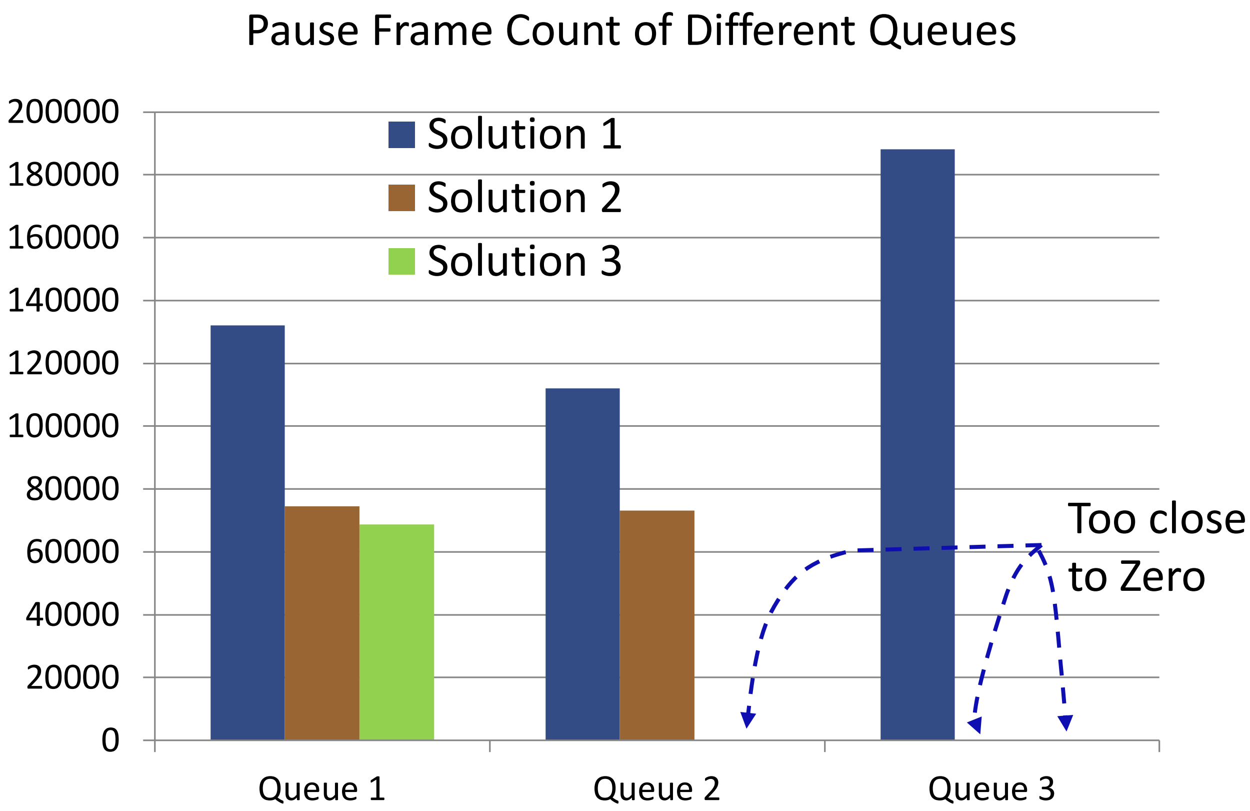

Pause Frame Count Received by Switch

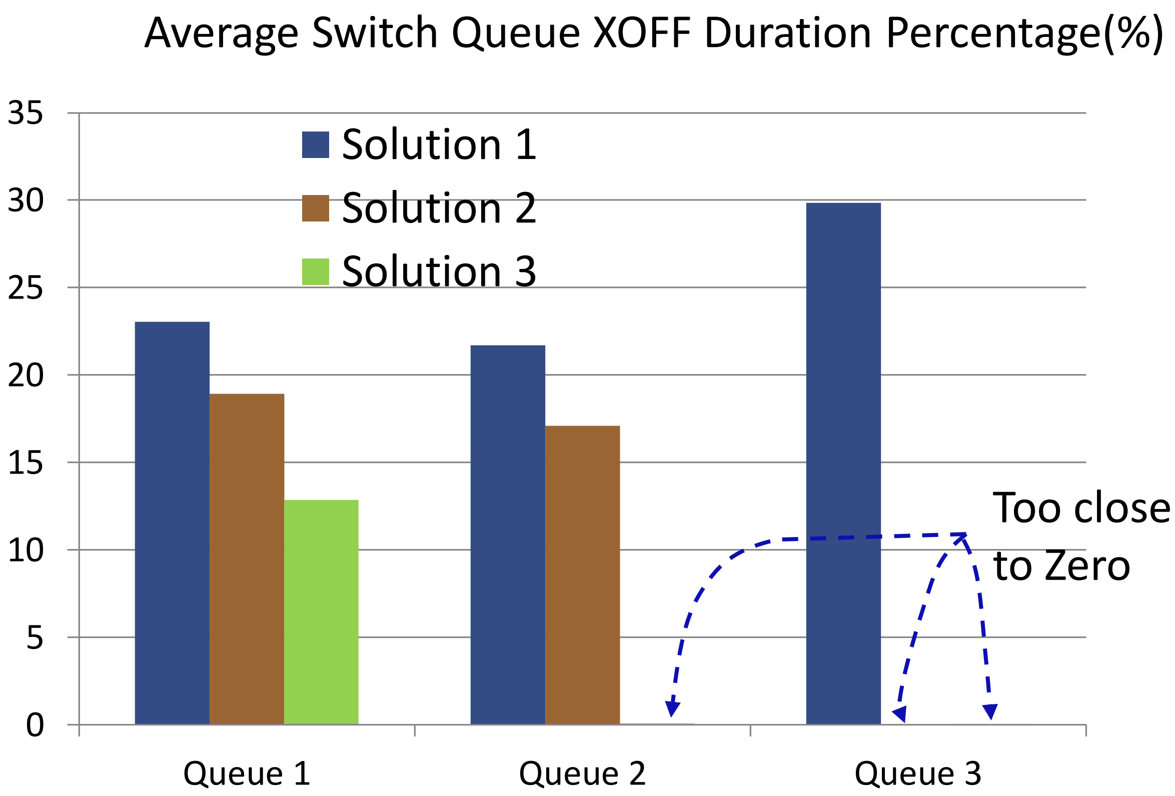

Average Switch Queue XOFF Duration Percentage(%)

- CI can reduce Pause frame count and XOFF duration significantly.
- 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.

Solution Comparison

- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI



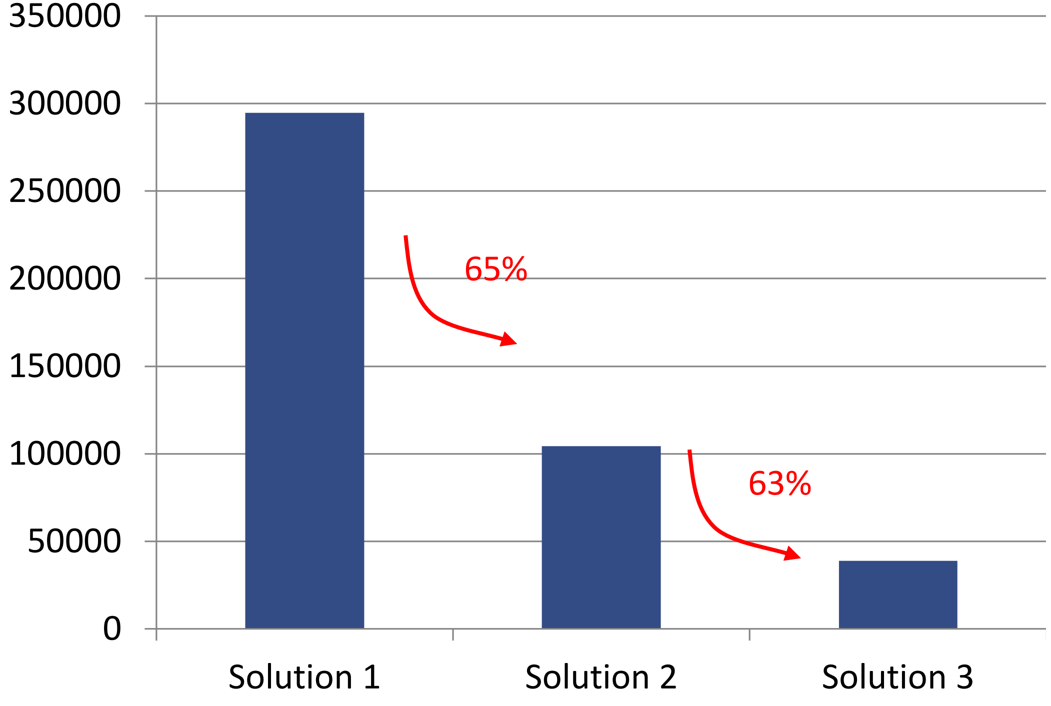


- CI can reduce Pause frame count and XOFF duration for all queues.
- Almost 100% decrease for queue 2 and 3, namely mice flow queue and elephant flow queue compared with solution 1, in which queue 2 and queue 3 are normal flow queue.

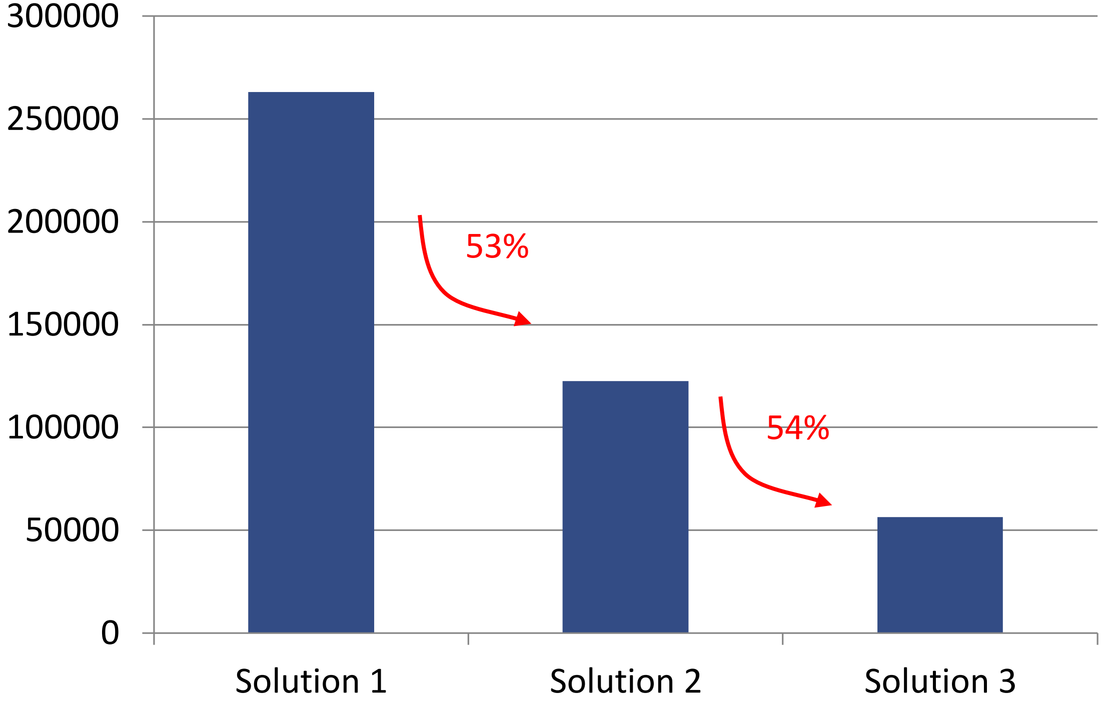
Solution Comparison

- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI

Pause Frame Count Receive by Servers



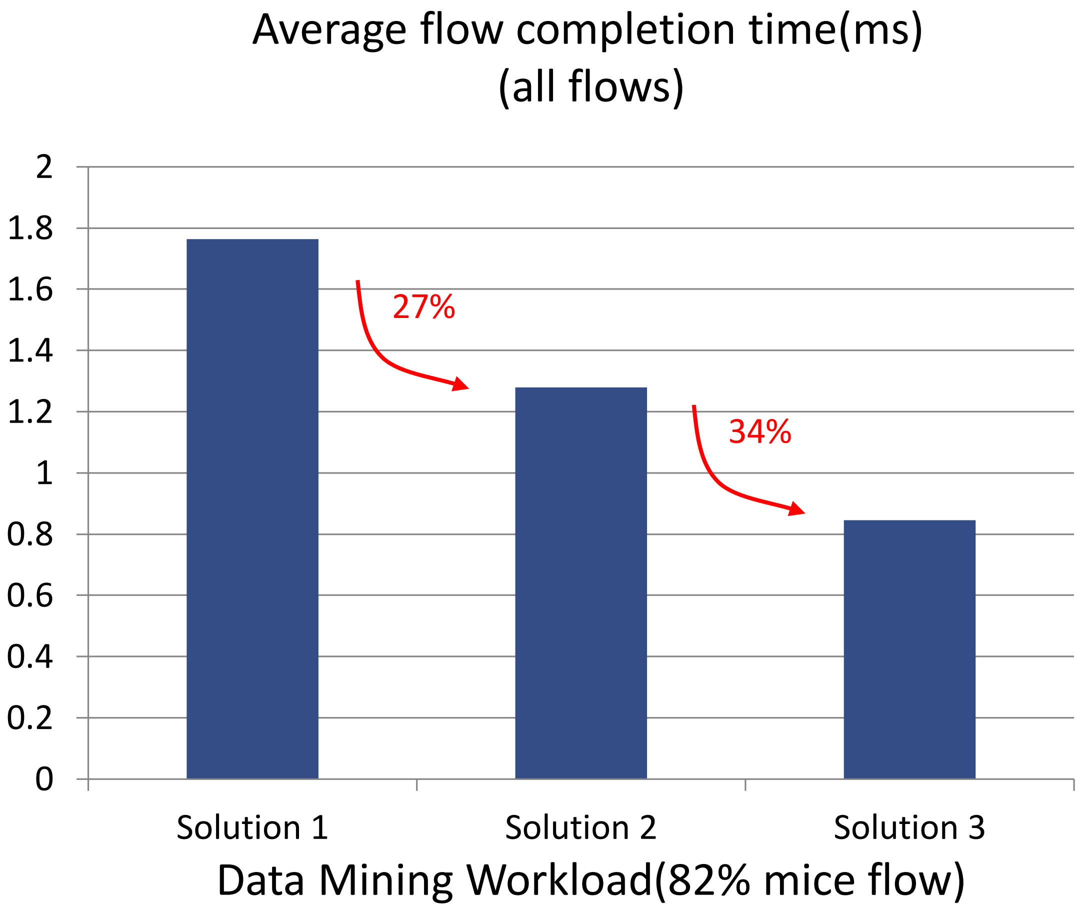
CNP Count Received by Servers



- CI can reduce Pause frame count and CIP count significantly on the server.

Solution Comparison

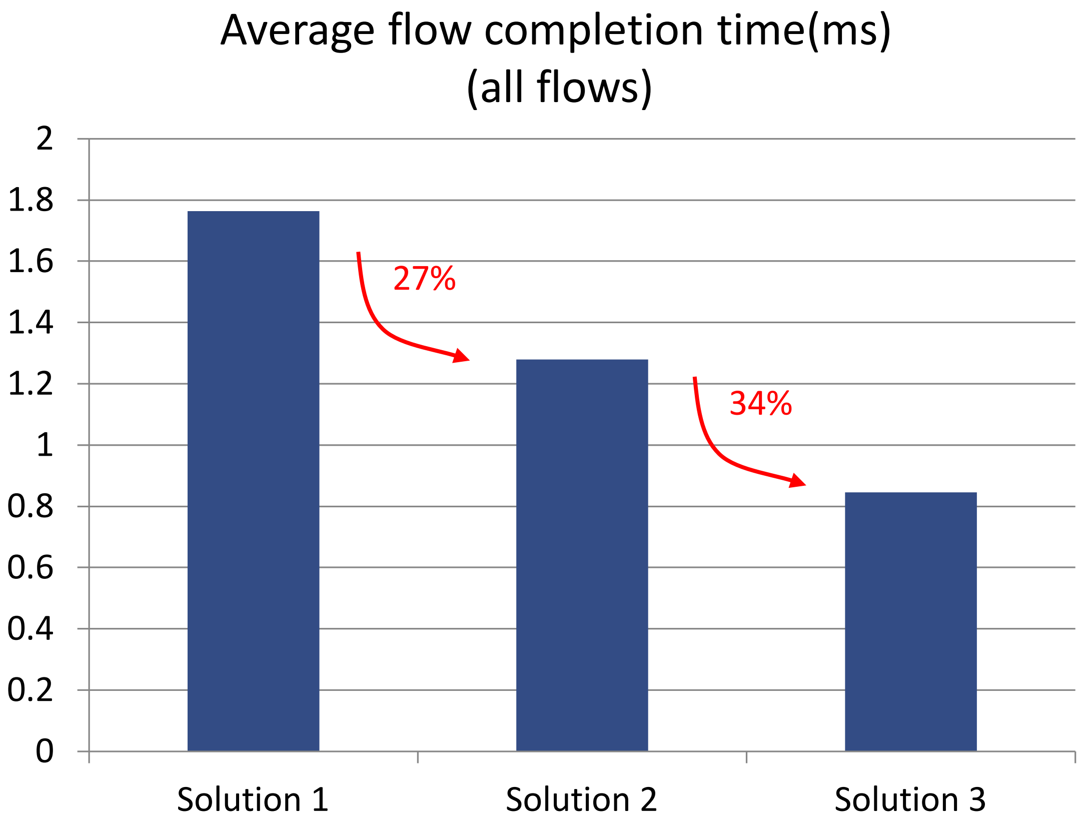
- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI



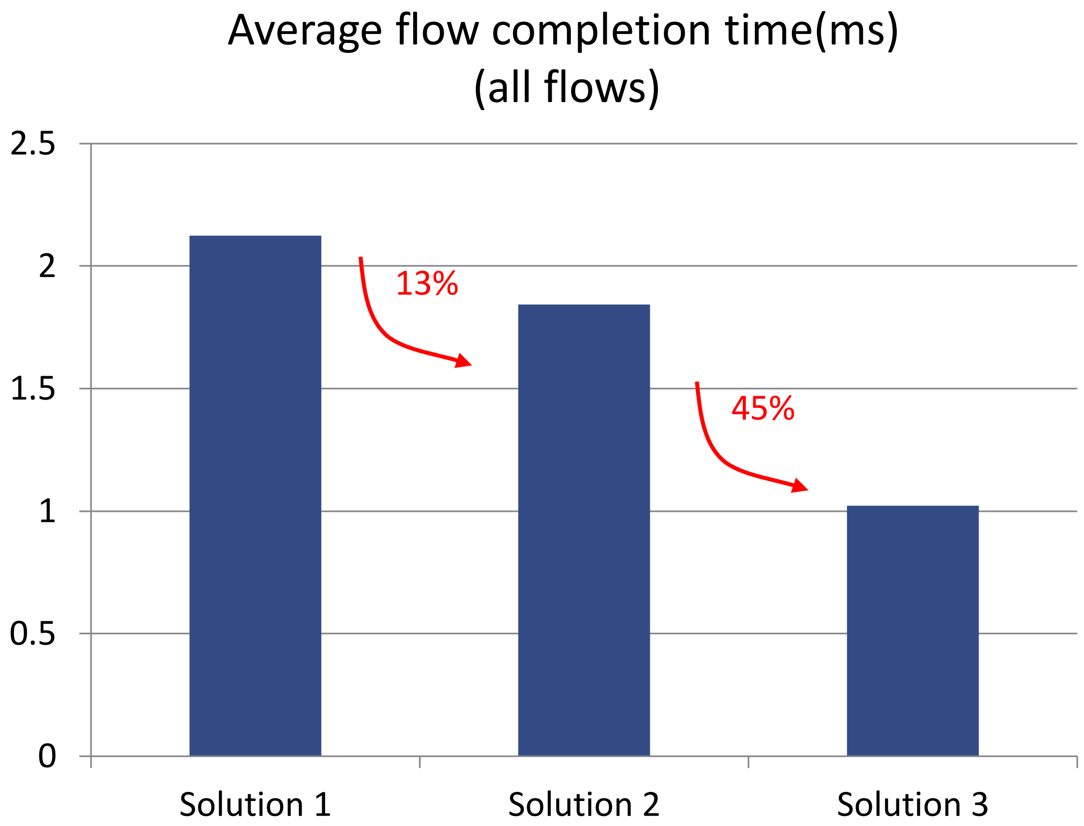
- All these bring a big upgrade of performance.

Solution Comparison

- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI



Data Mining Workload(82% mice flow)

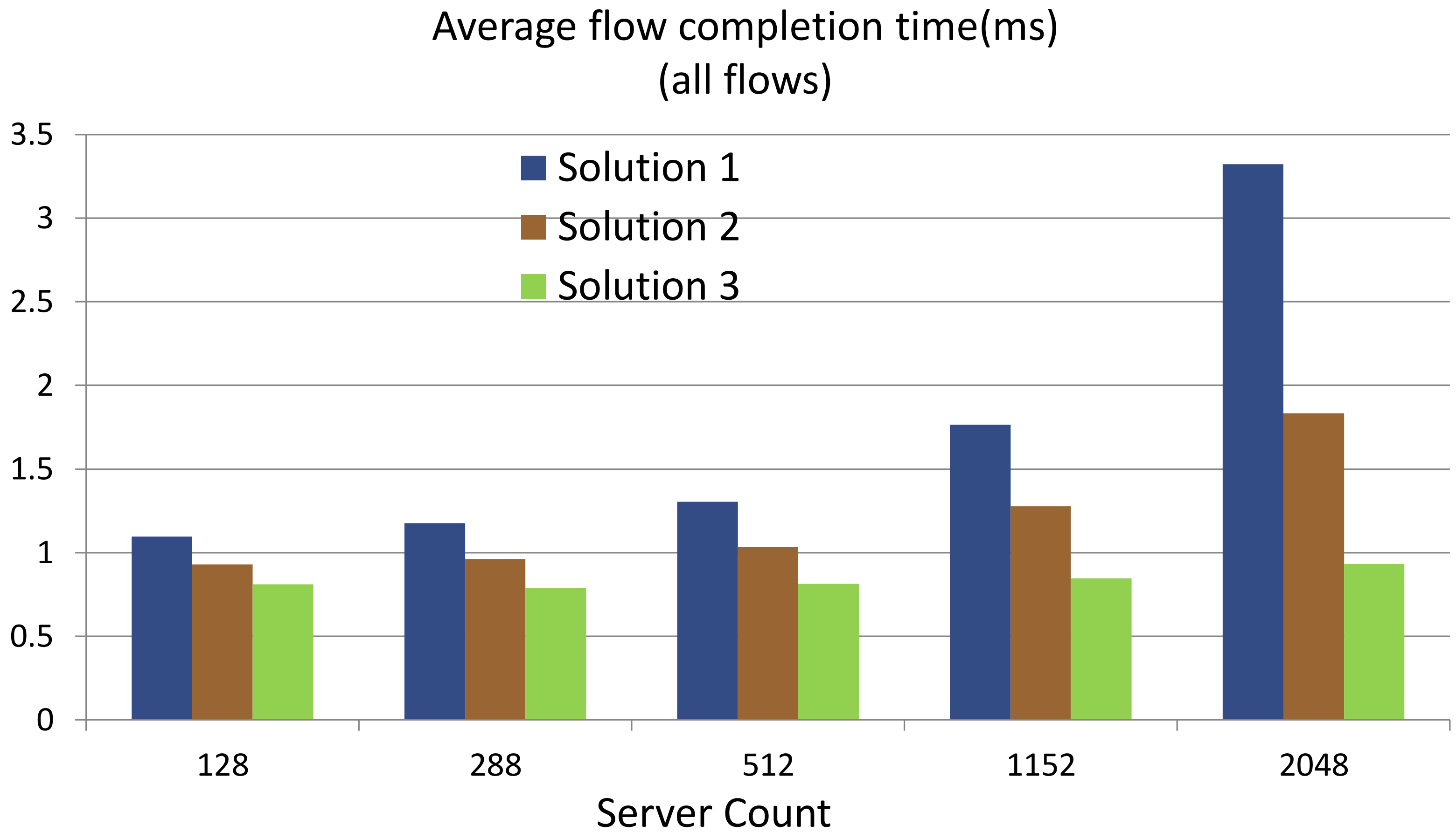


Cache Follower Workload(60% mice flow)

- Solution 2(mice prioritization) can not bring big improvement in less mice flow scenario. CI can.
- Seems like CI is a traffic pattern independent solution.

Solution Comparison

- Solution 1: PFC + ECN
- Solution 2: PFC + ECN with mice prioritization
- Solution 3: PFC + ECN with mice prioritization and CI



- The performance of Solution 1 and Solution 2 degrades when scales out. CI does not.
- Seems like CI is a scale independent solution.

Questions?