TCP Pacing in Data Center Networks

Monia Ghobadi, Yashar Ganjali

Department of Computer Science, University of Toronto
{monia, yganjali}@cs.toronto.edu

TCP, Oh TCP!

TCP, Oh TCP!

- TCP congestion control

TCP, Oh TCP!

- TCP congestion control
  - Focus on evolution of cwnd over RTT.

TCP, Oh TCP!

- TCP congestion control
  - Focus on evolution of cwnd over RTT.
- Damages

TCP, Oh TCP!

- TCP congestion control
  - Focus on evolution of cwnd over RTT.
- Damages
- TCP pacing

The Tortoise and the Hare: Why Bother Pacing?

The Tortoise and the Hare: Why Bother Pacing?

- Renewed interest in pacing for the data center environment
  - Small buffer switches
  - Small round-trip times
    - Disparity between the total capacity of the network and the capacity of individual queues

The Tortoise and the Hare: Why Bother Pacing?

- Renewed interest in pacing for the data center environment
  - Small buffer switches
  - Small round-trip times
    - Disparity between the total capacity of the network and the capacity of individual queues
  - Focus on tail latency cause by short-term unfairness in TCP

TCP Pacing's Potential

TCP Pacing's Potential

- Better link utilization on small switch buffers

TCP Pacing's Potential

- Better link utilization on small switch buffers
- Better short-term fairness among flows of similar RTTs:
  - Improves worst-flow latency

TCP Pacing's Potential

- Better link utilization on small switch buffers
- Better short-term fairness among flows of similar RTTs:
  - Improves worst-flow latency
- Allows slow-start to be circumvented
  - Saving many round-trip time
  - May allow much larger initial congestion window to be used safely

Contributions

Contributions

- Effectiveness of TCP pacing in data centers.

Contributions

- Effectiveness of TCP pacing in data centers.
- Benefits of using paced TCP diminish as we increase the number of concurrent connections beyond a certain threshold (Point of Inflection).

Contributions

- Effectiveness of TCP pacing in data centers.
- Benefits of using paced TCP diminish as we increase the number of concurrent connections beyond a certain threshold (Point of Inflection).
- Inconclusive results in previous works.

Contributions

- Effectiveness of TCP pacing in data centers.
- Benefits of using paced TCP diminish as we increase the number of concurrent connections beyond a certain threshold (Point of Inflection).
- Inconclusive results in previous works.
- Inter-flow bursts.

Contributions

- Effectiveness of TCP pacing in data centers.
- Benefits of using paced TCP diminish as we increase the number of concurrent connections beyond a certain threshold (Point of Inflection).
- Inconclusive results in previous works.
- Inter-flow bursts.
- Test-bed experiments.

Inter-flow Bursts

Inter-flow Bursts

- C: bottleneck link capacity

Inter-flow Bursts

- C: bottleneck link capacity
- B_{max} :buffer size

Inter-flow Bursts

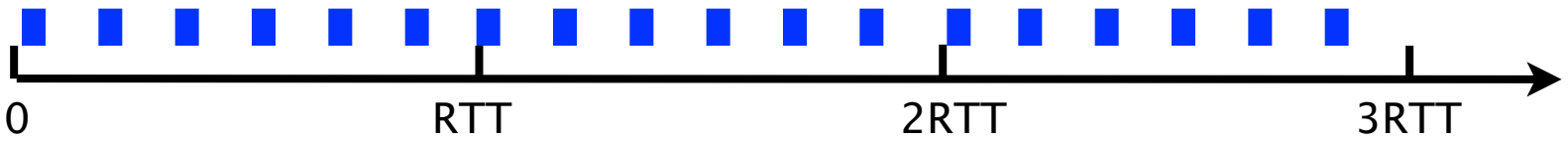
- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.

Inter-flow Bursts

- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.

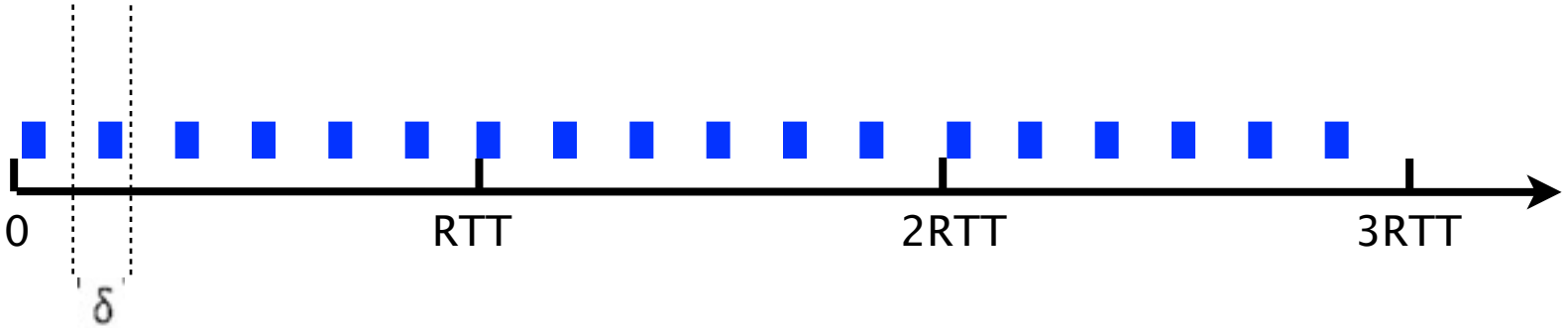
Inter-flow Bursts

- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.



Inter-flow Bursts

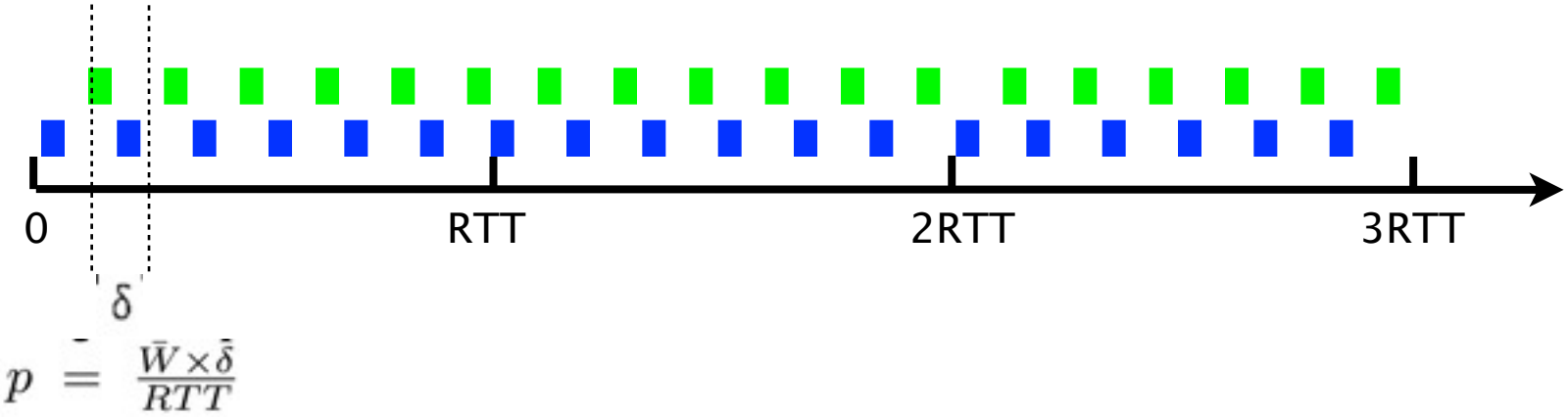
- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.



$$p = \frac{\bar{W} \times \bar{\delta}}{RTT}$$

Inter-flow Bursts

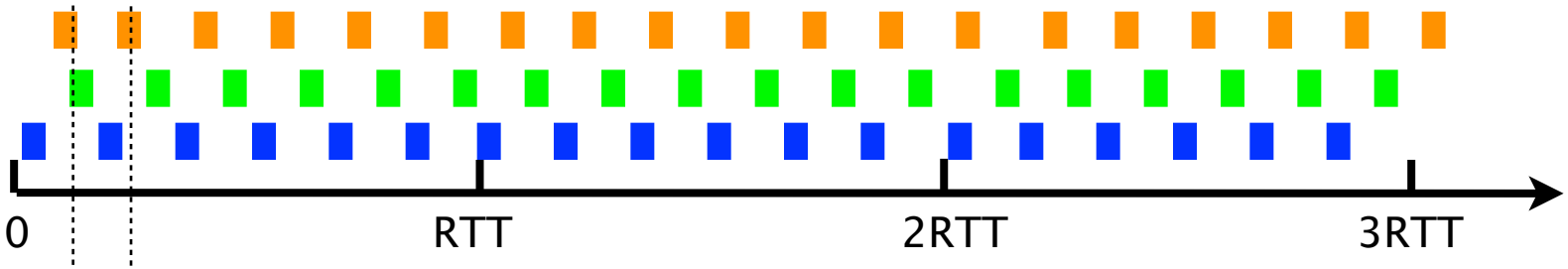
- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.



$$p = \frac{\bar{W} \times \bar{\delta}}{RTT}$$

Inter-flow Bursts

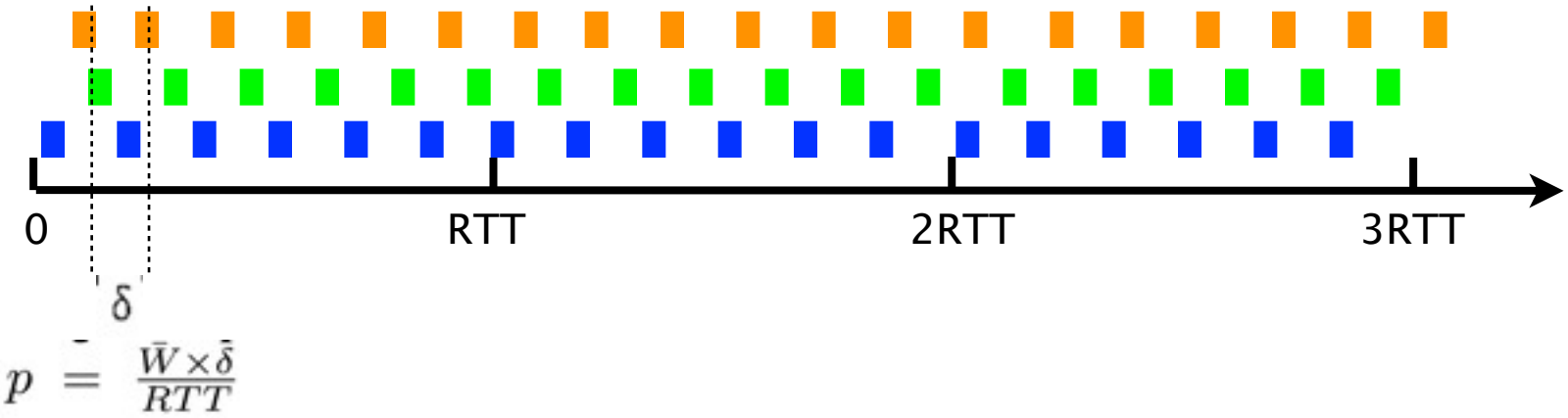
- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.



$$p = \frac{\bar{W} \times \bar{\delta}}{RTT}$$

Inter-flow Bursts

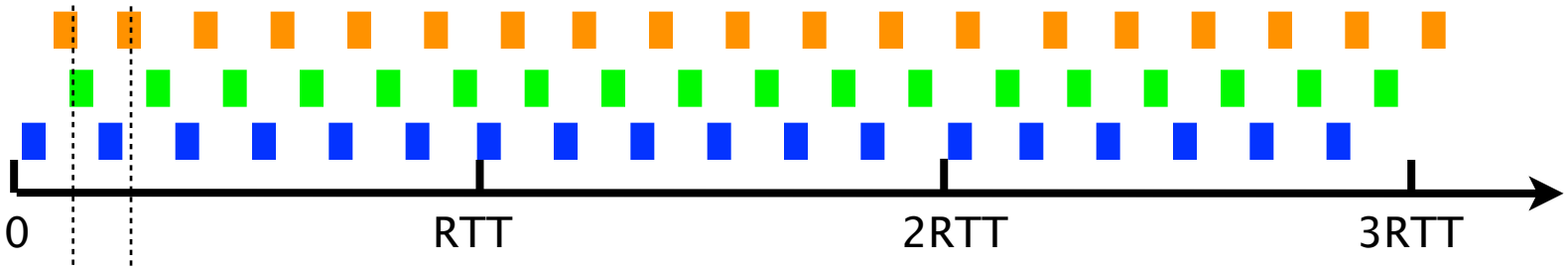
- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.
- X: Inter-flow burst



$$p = \frac{\bar{W} \times \bar{\delta}}{RTT}$$

Inter-flow Bursts

- C: bottleneck link capacity
- B_{max} :buffer size
- N: longed lived flows.
- W: packets in every RTT in paced or non-paced manner.
- X: Inter-flow burst $\sim B(N,p)$



$$p = \frac{\bar{W} \times \bar{\delta}}{RTT}$$

Modeling

Modeling

best case of non-paced

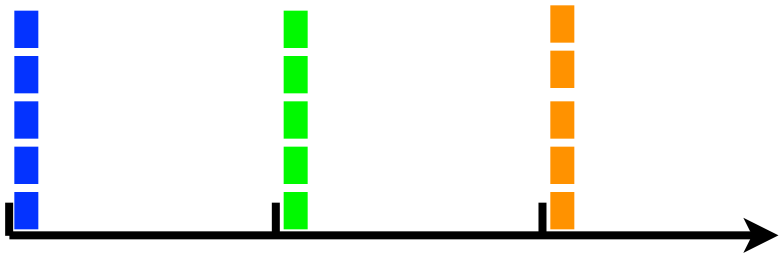
Modeling

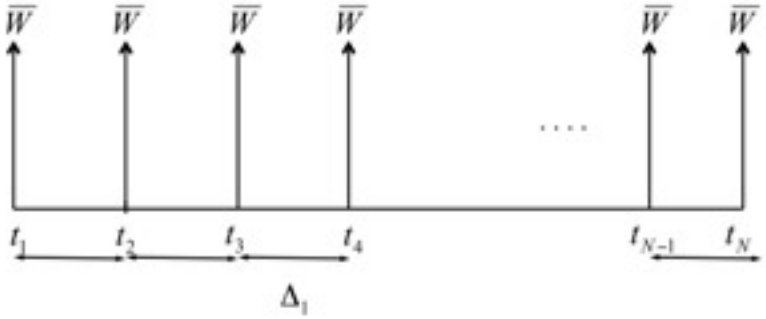
best case of non-paced

worst case of paced

Modeling

best case of non-paced

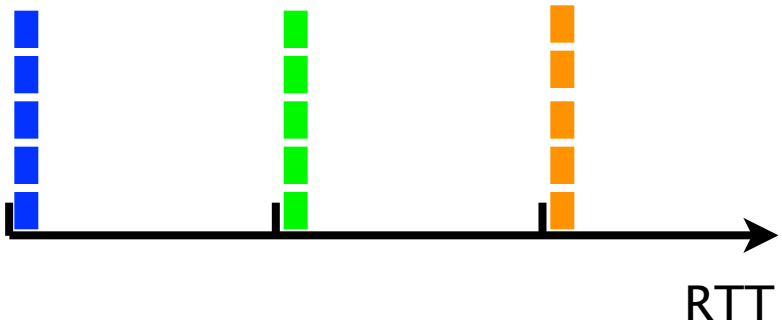




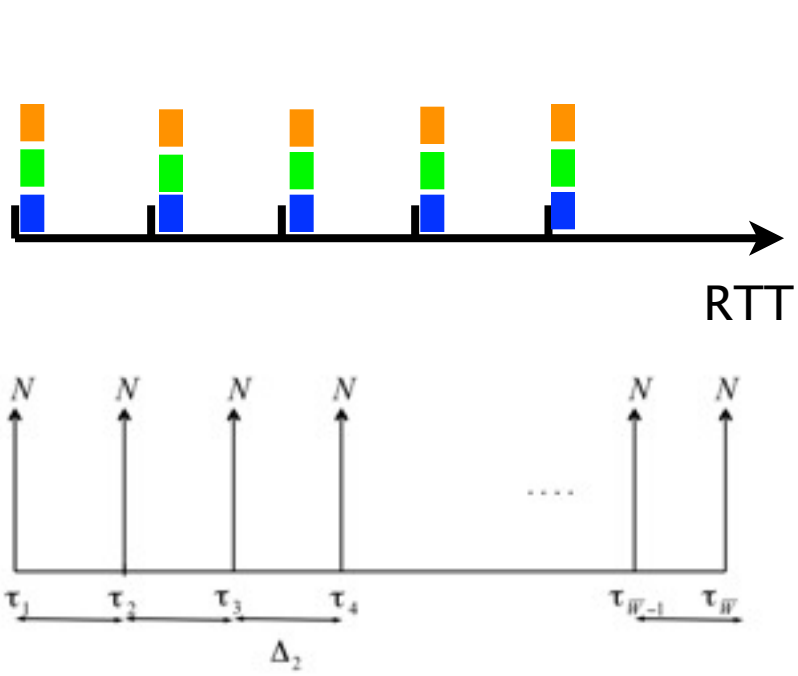
worst case of paced

Modeling

best case of non-paced

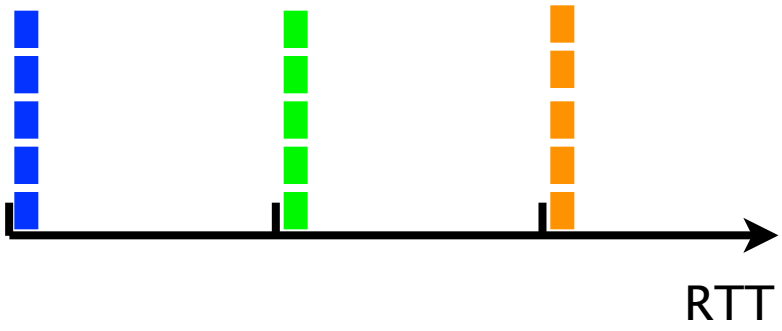


worst case of paced

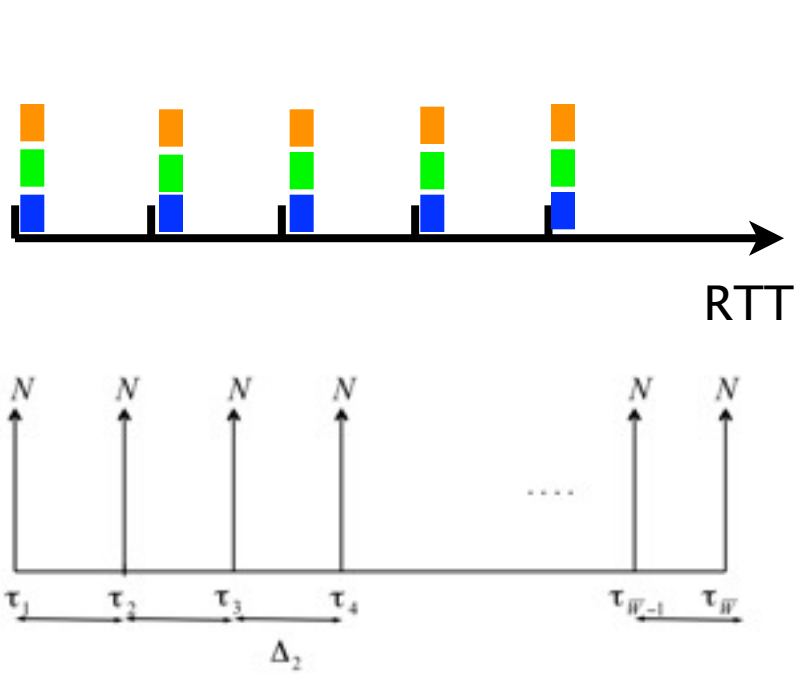


Modeling

best case of non-paced



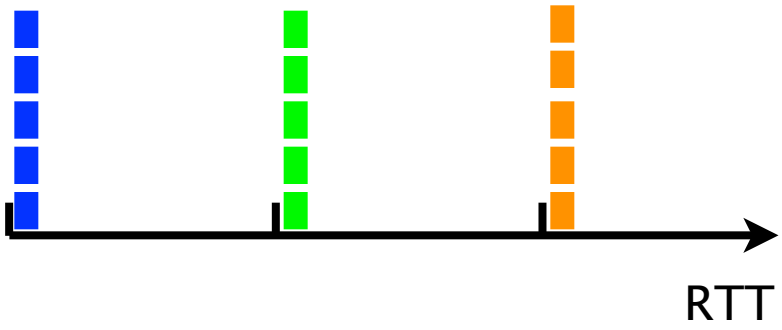
worst case of paced



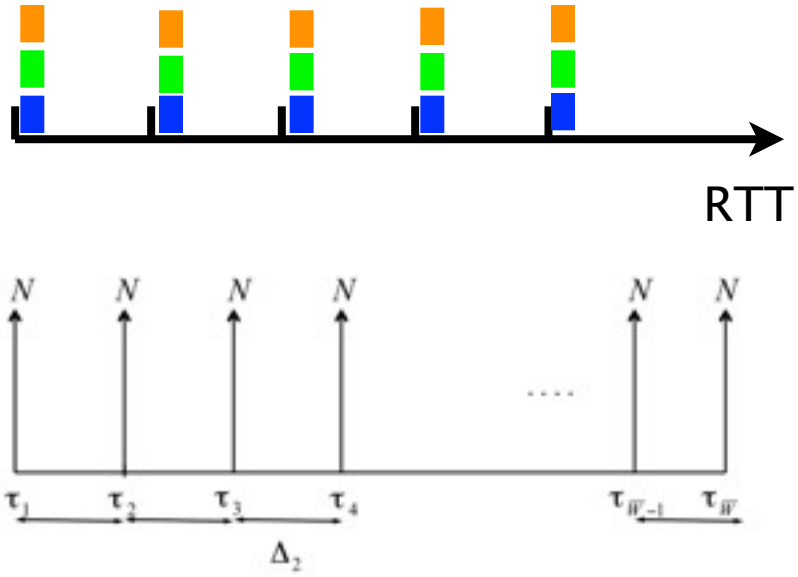
$$N < \bar{W} = \frac{C \times RTT}{B_{max}}$$

Modeling

best case of non-paced



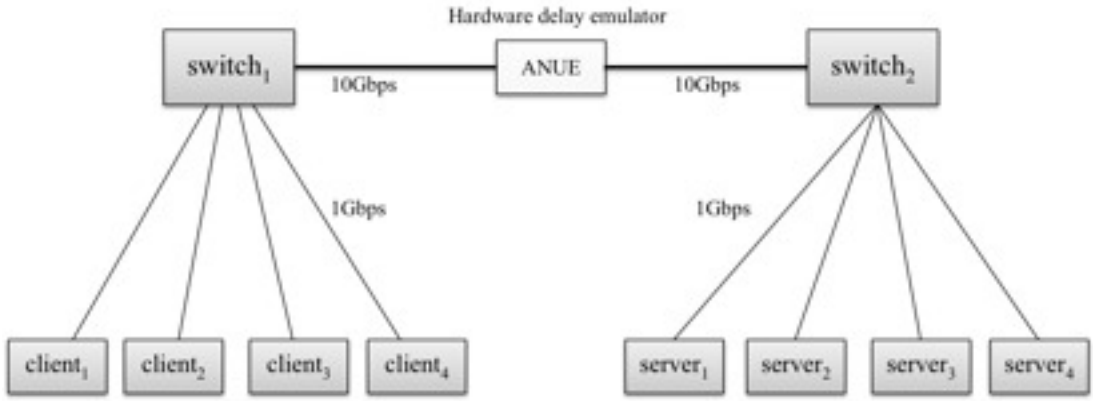
worst case of paced



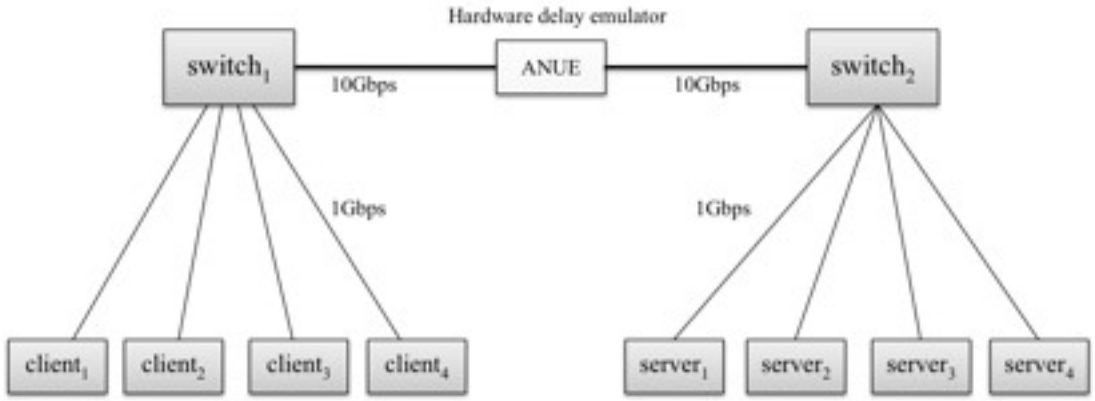
$$N < \bar{W} = \frac{C \times RTT}{B_{max}}$$

$$N^* = \Omega(\frac{C \times RTT}{B_{max}})$$

Experimental Studies

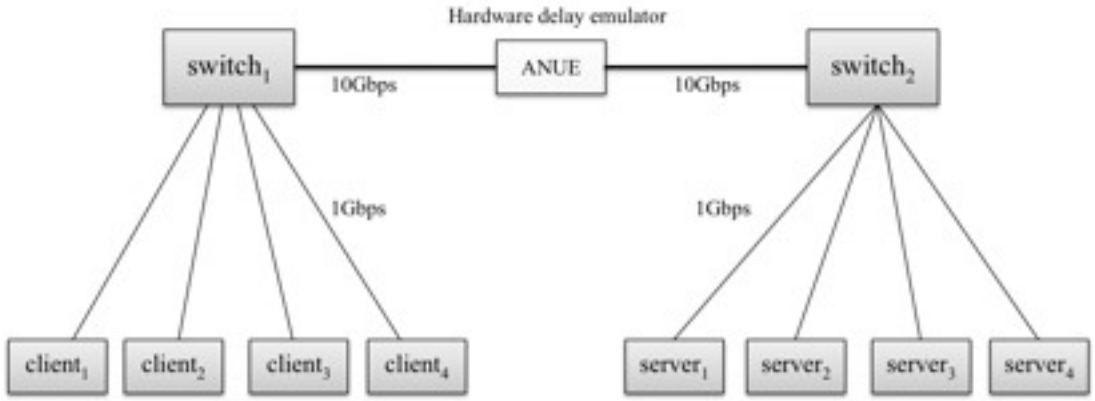


Experimental Studies



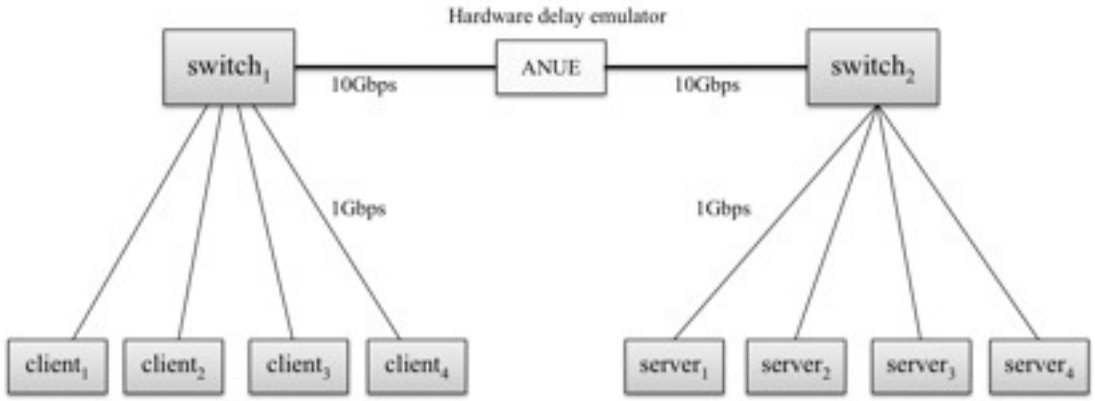
- Flow of sizes 1,2, 3 MB between servers and clients.

Experimental Studies



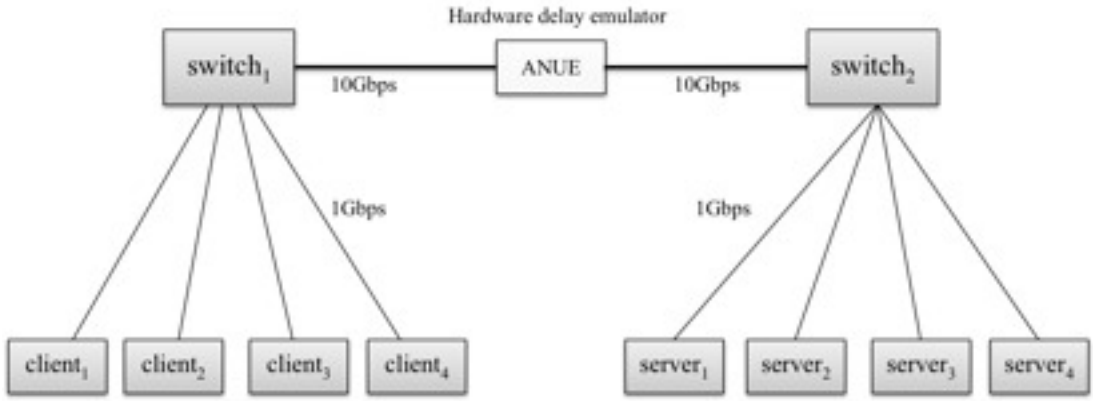
- Flow of sizes 1,2, 3 MB between servers and clients.
- Bottleneck BW: 1,2, 3 Gbps

Experimental Studies



- Flow of sizes 1,2, 3 MB between servers and clients.
- Bottleneck BW: 1,2, 3 Gbps
- RTT: 1 to 100ms

Experimental Studies



- Flow of sizes 1,2, 3 MB between servers and clients.
- Bottleneck BW: 1,2, 3 Gbps
- RTT: 1 to 100ms
- Bottleneck utilization, Drop rate, average and tail FCT

Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

No congestion

Base-Case Experiment:

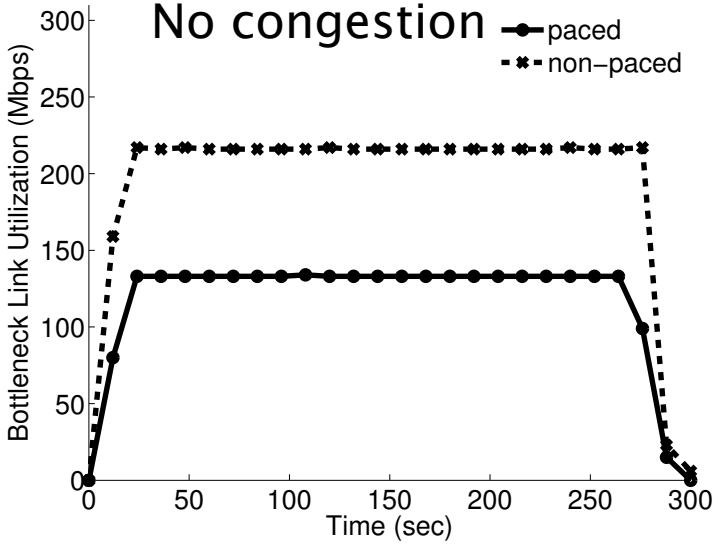
One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

No congestion

Congestion

Base-Case Experiment:

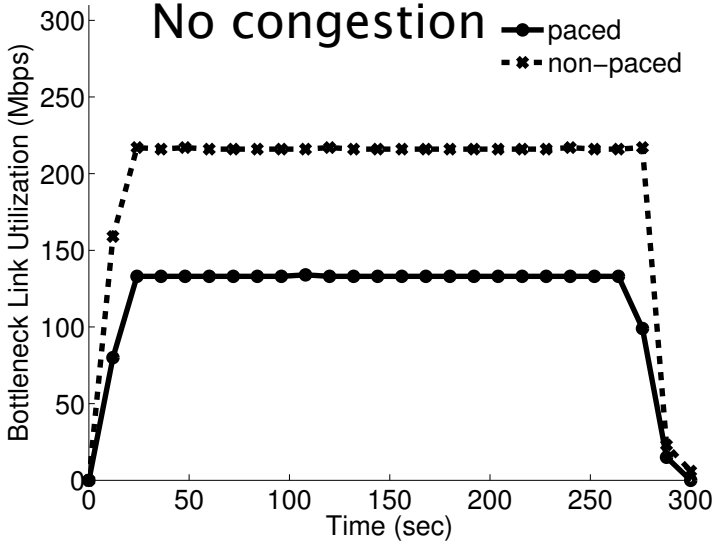
One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



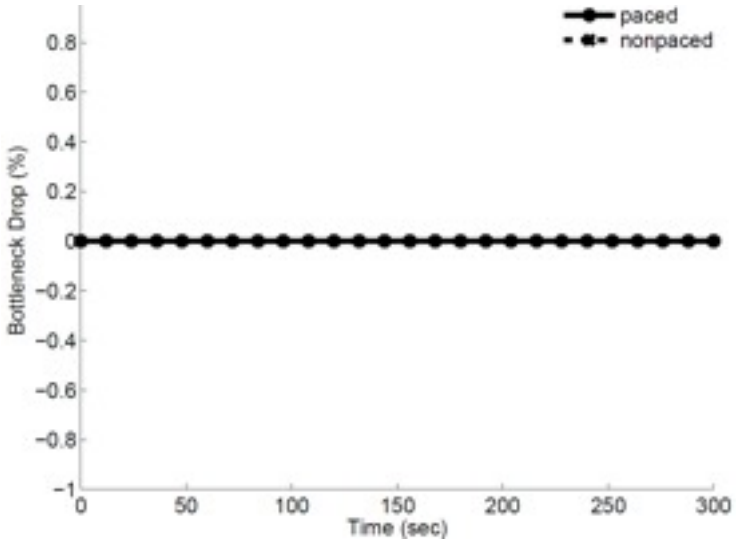
Congestion

Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

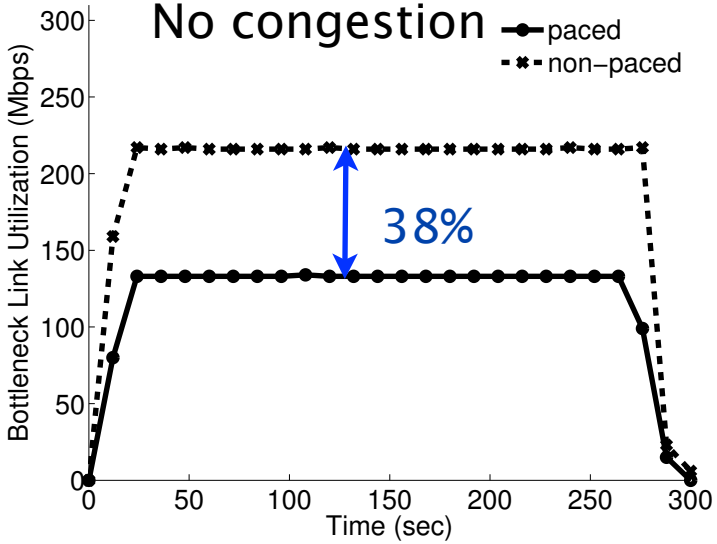


Congestion

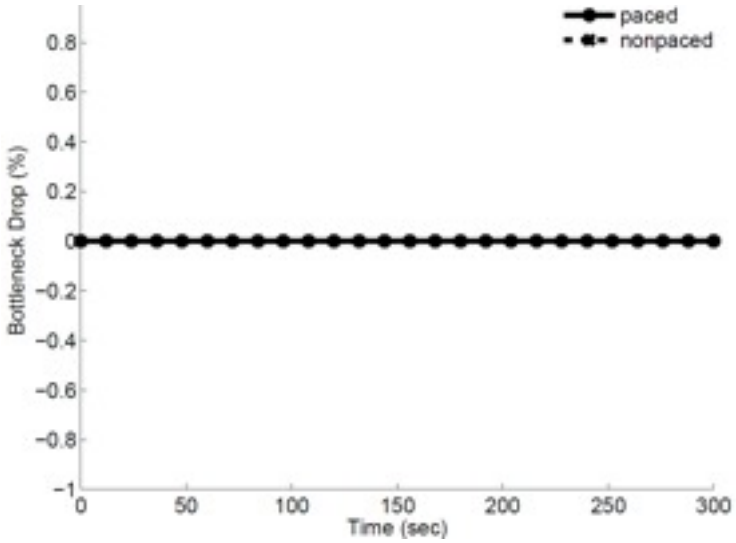


Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

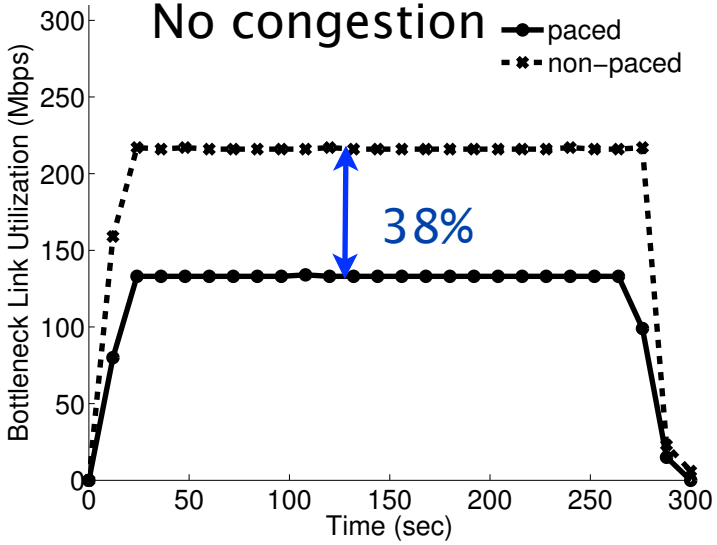


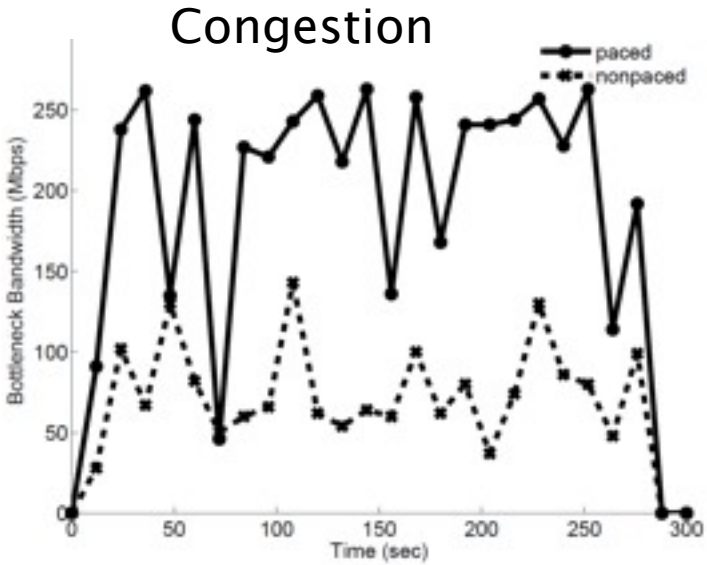
Congestion

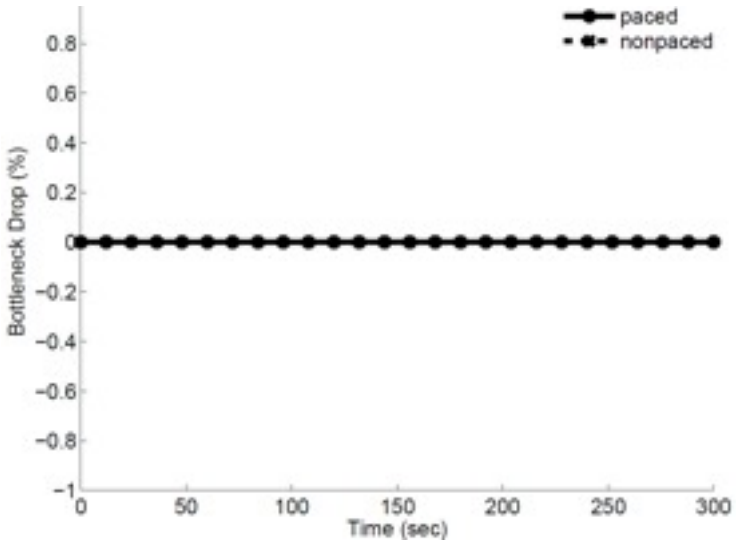


Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT

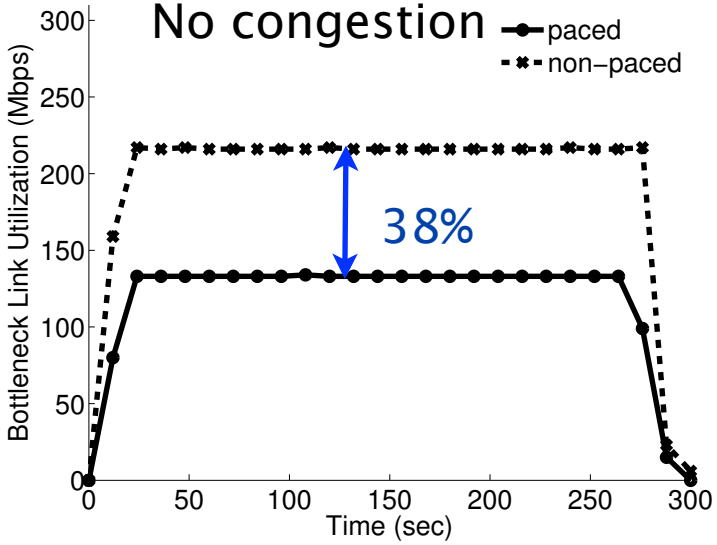


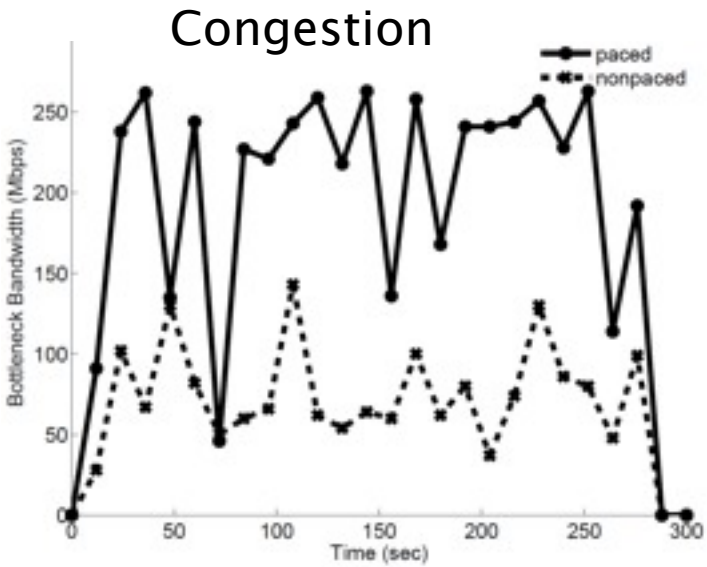


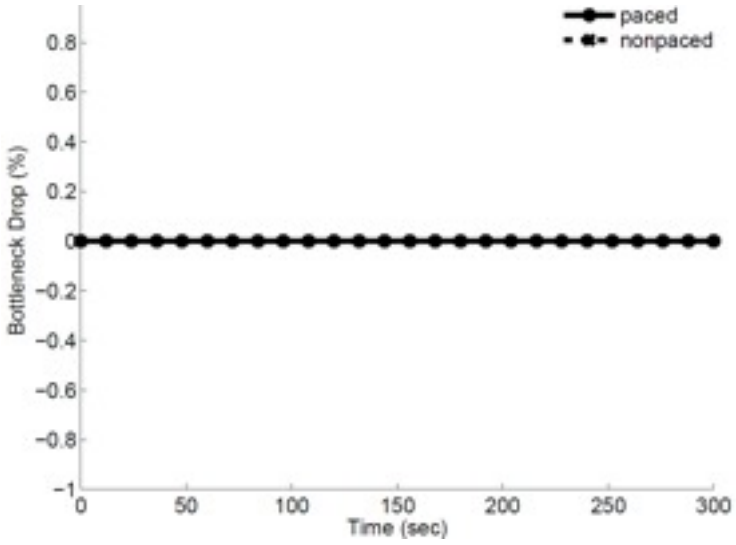


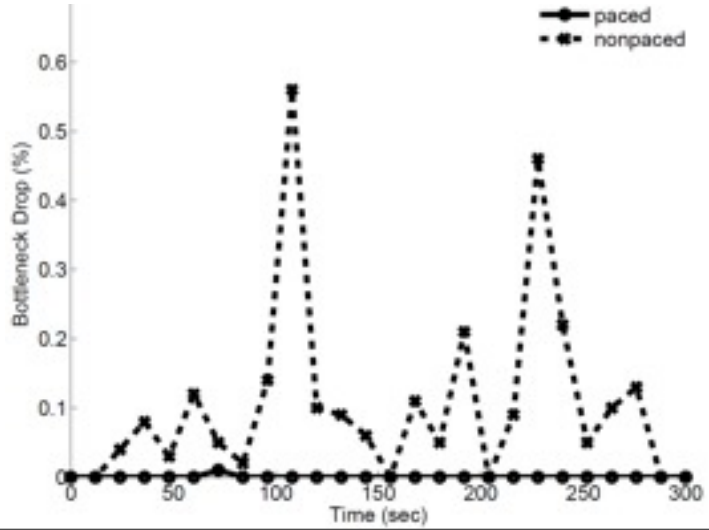
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



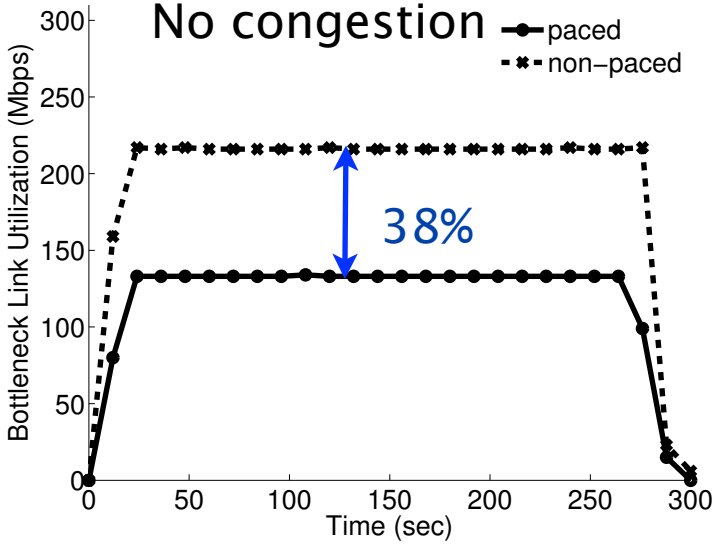


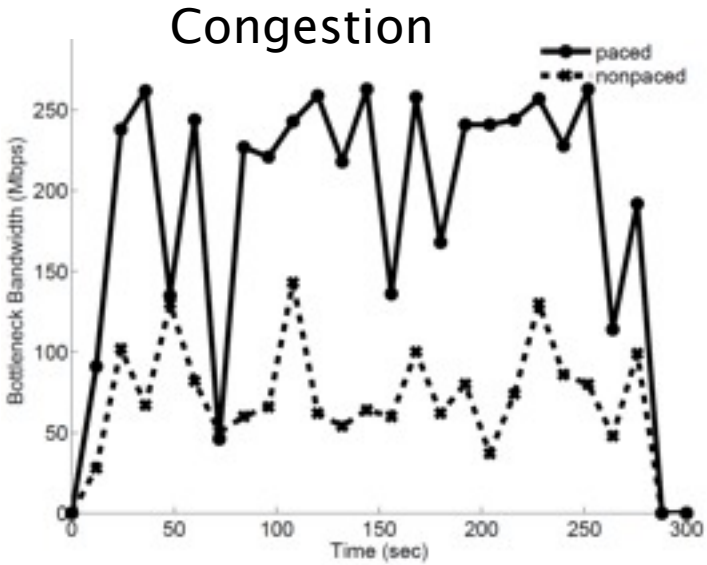


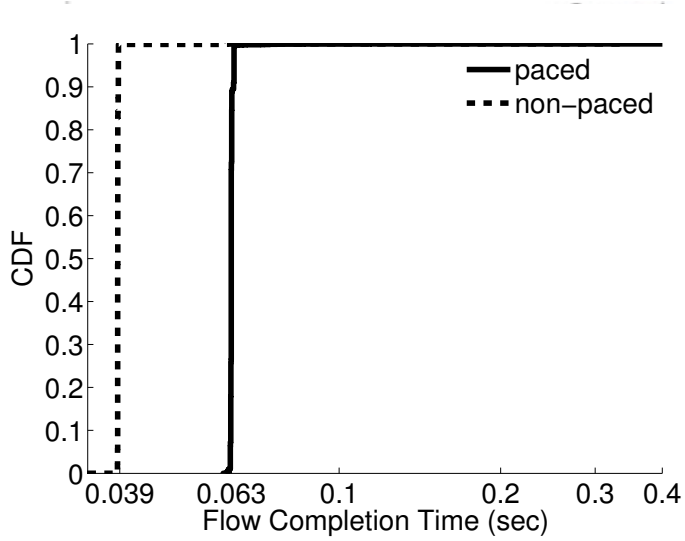


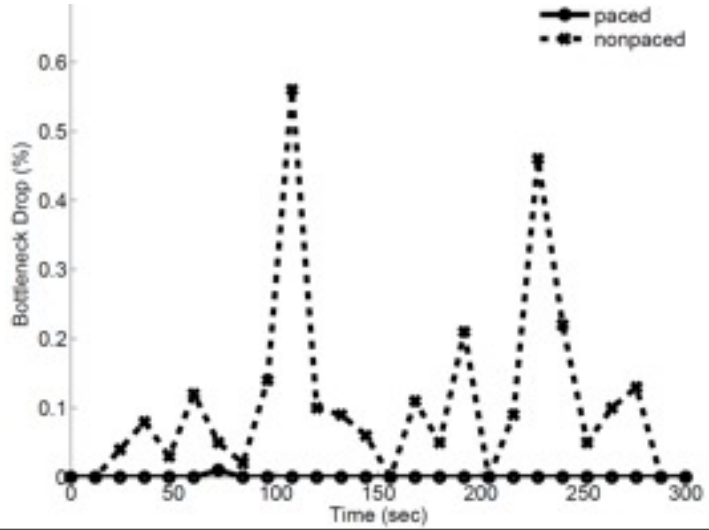
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



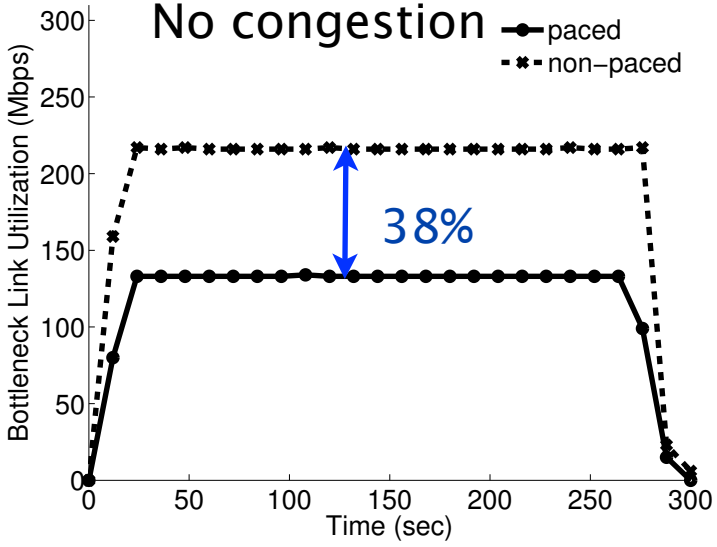


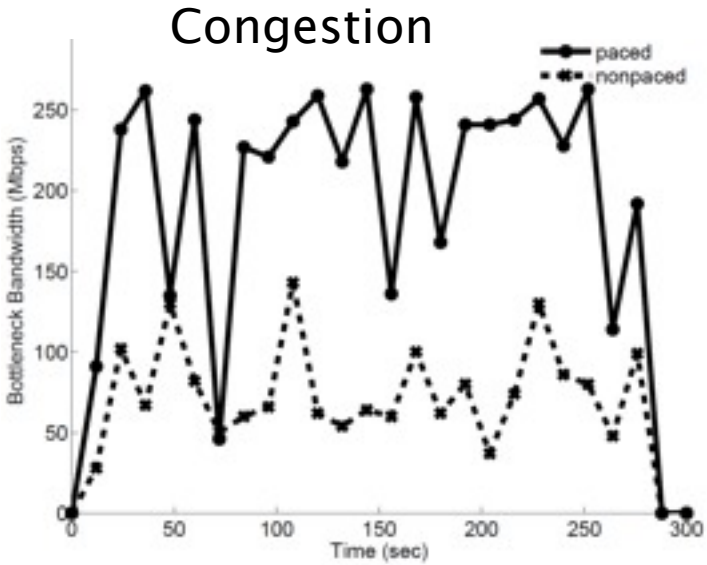


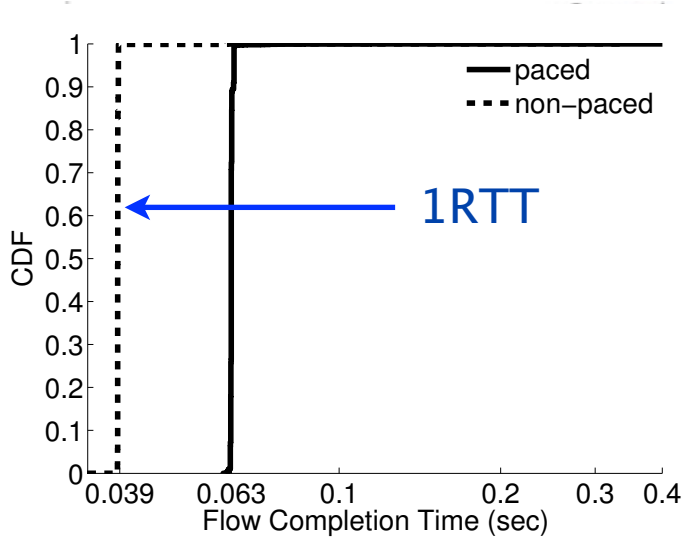


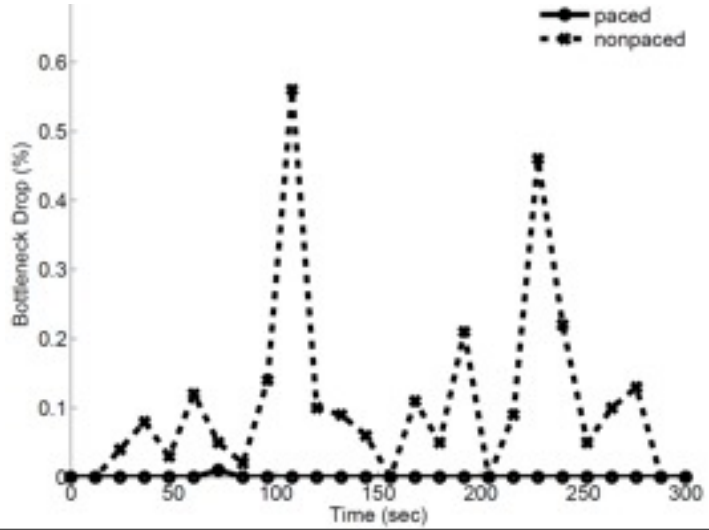
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



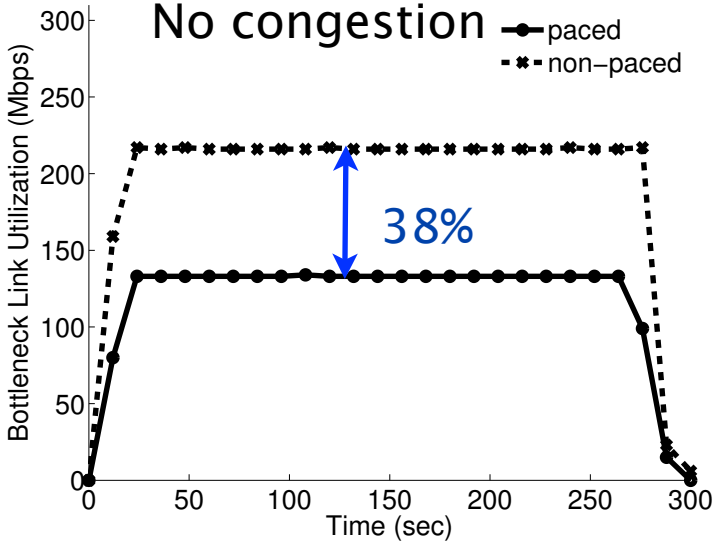


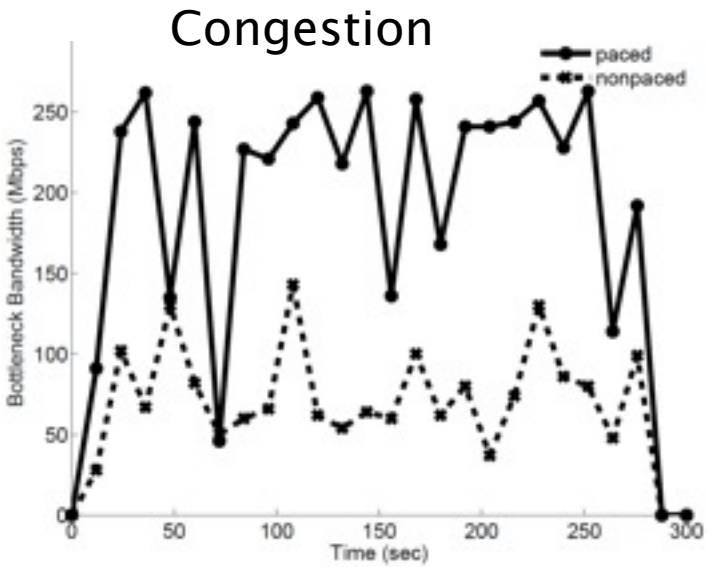


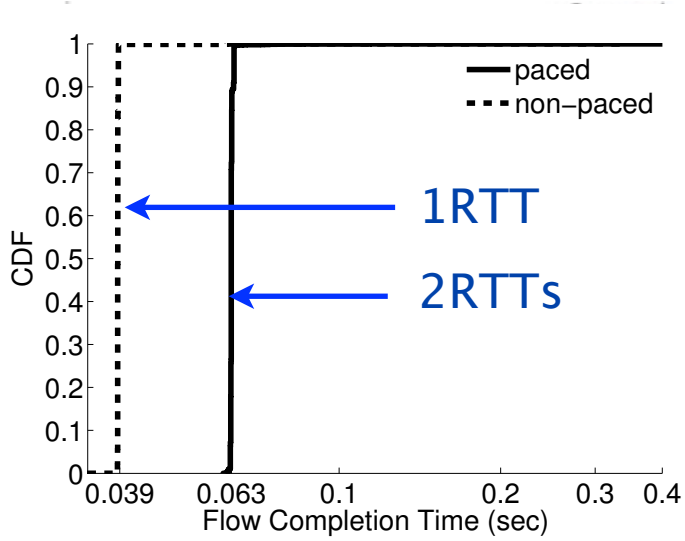


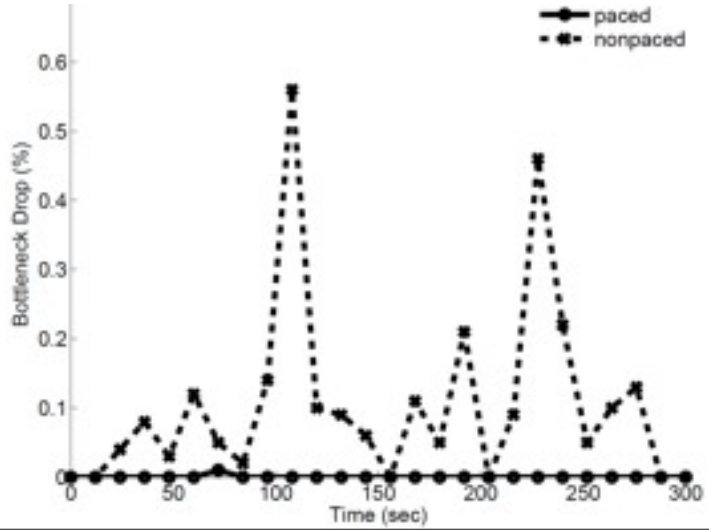
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



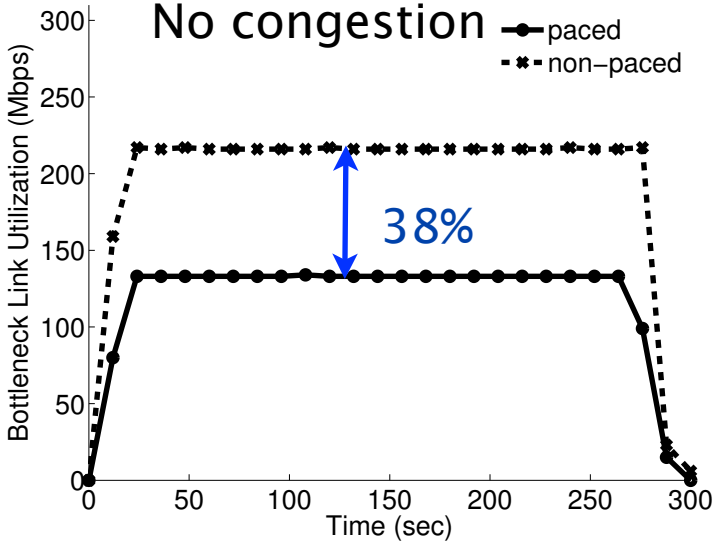


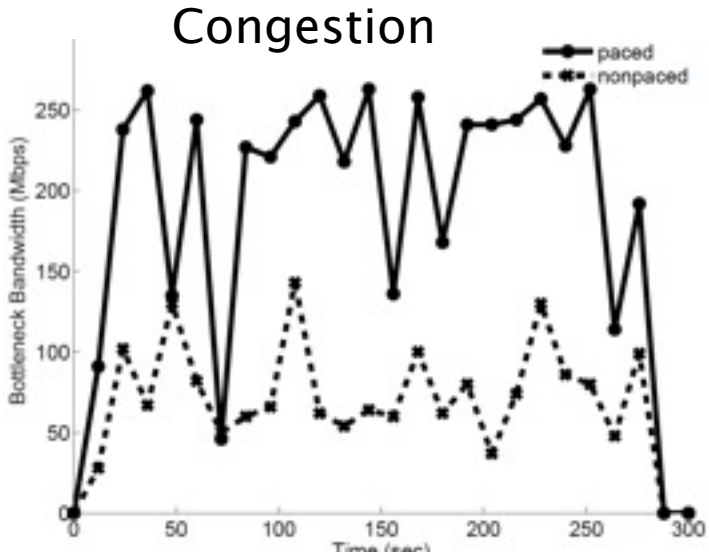


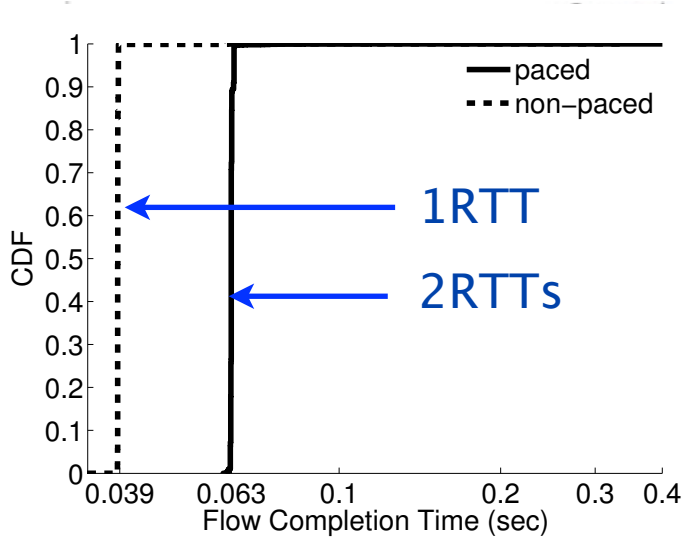


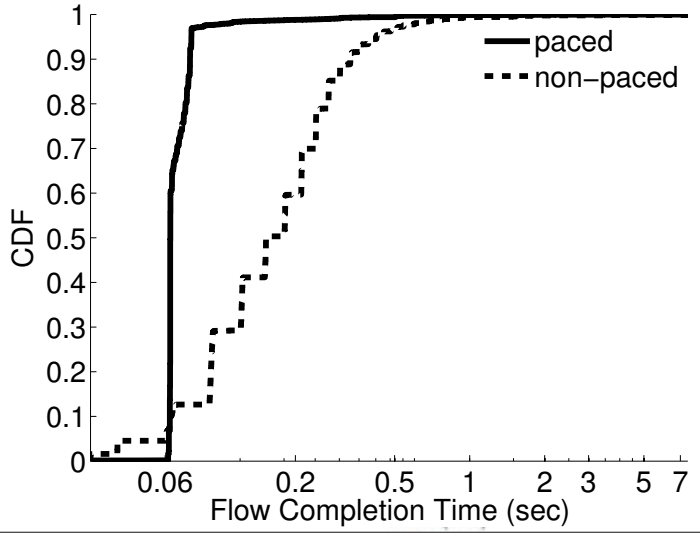
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT



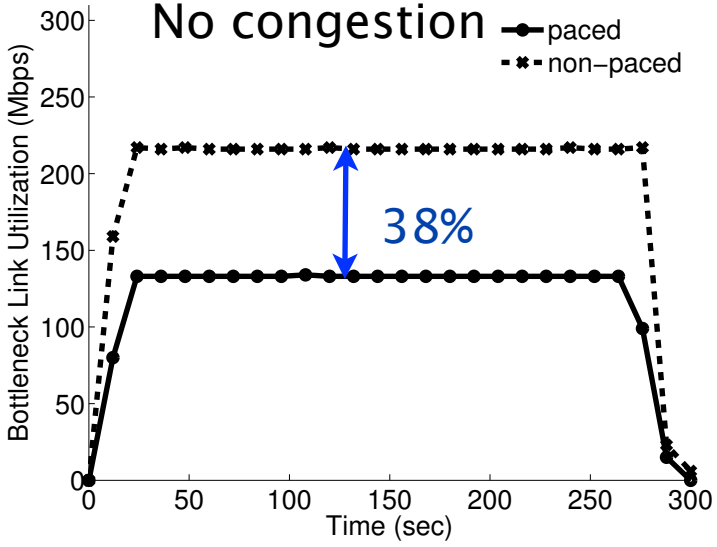


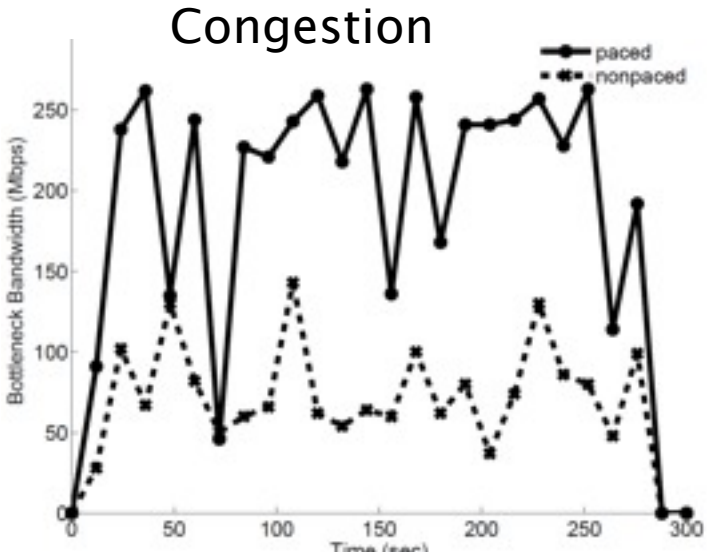


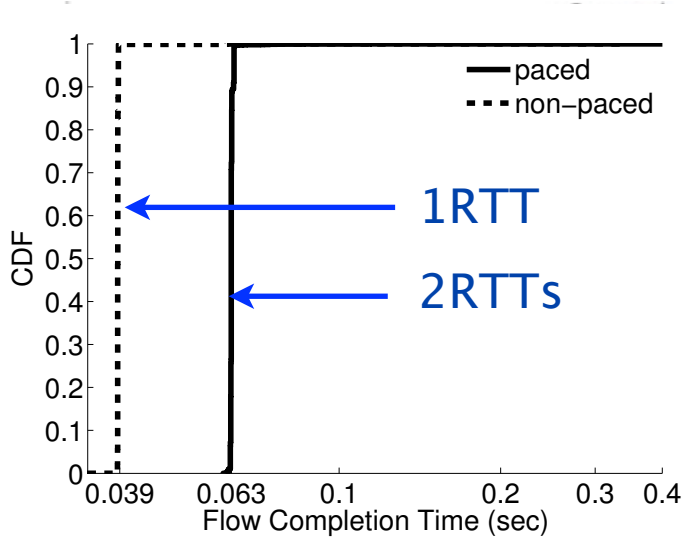


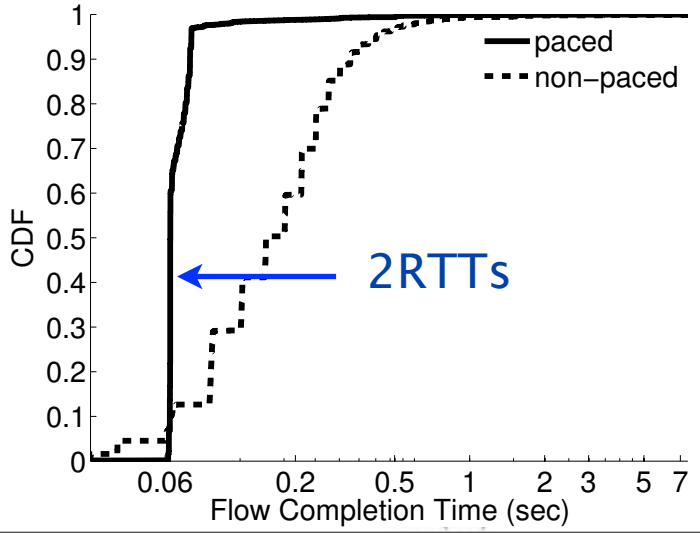
Base-Case Experiment:

One flow vs Two flows, 64KB of buffering, Utilization/Drop/FCT







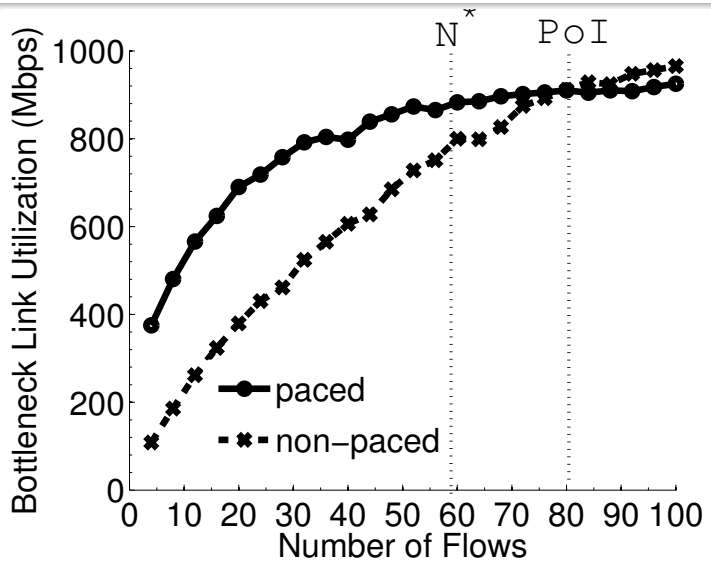


Multiple flows: Link Utilization/Drop/Latency

Buffer size 1.7% of BDP, varying number of flows

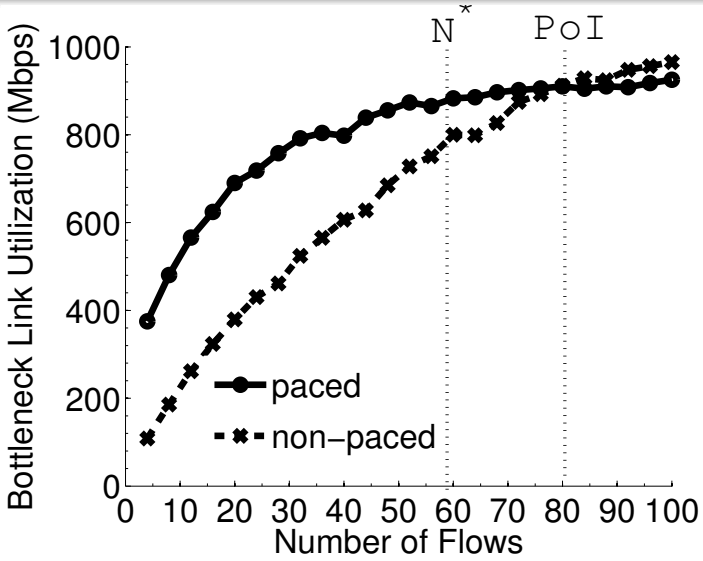
Multiple flows: Link Utilization/Drop/Latency

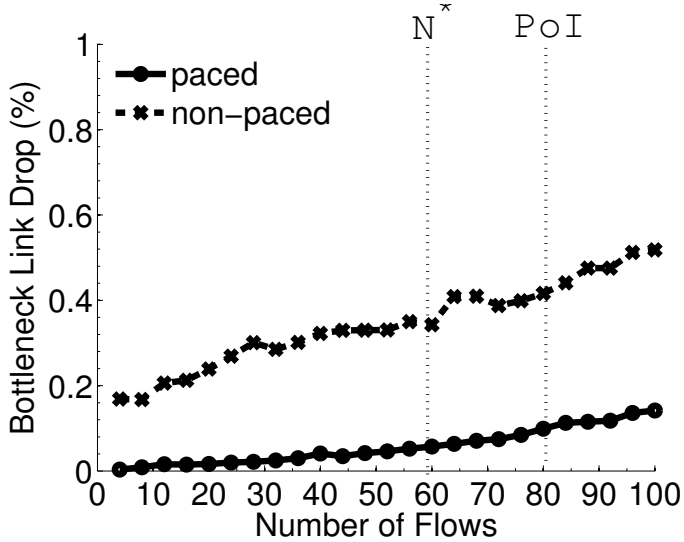
Buffer size 1.7% of BDP, varying number of flows



Multiple flows: Link Utilization/Drop/Latency

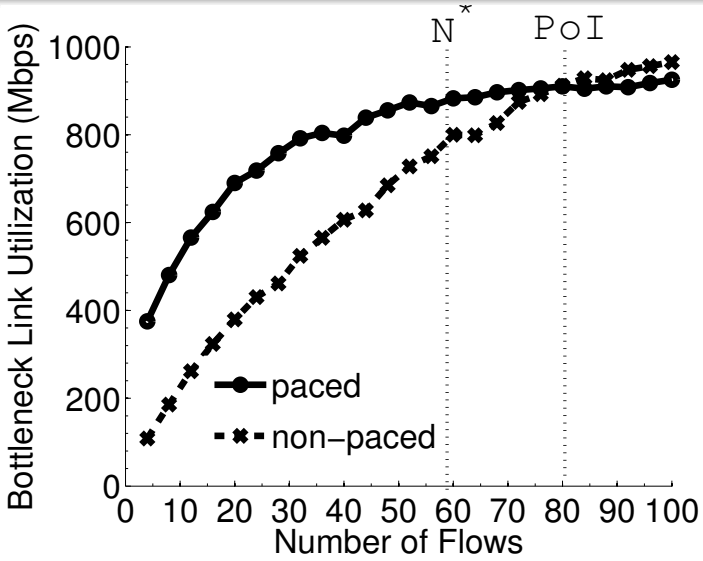
Buffer size 1.7% of BDP, varying number of flows

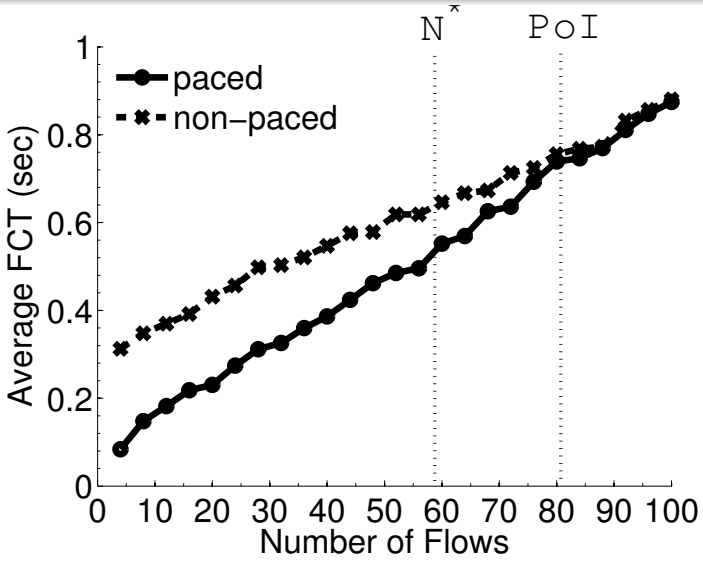


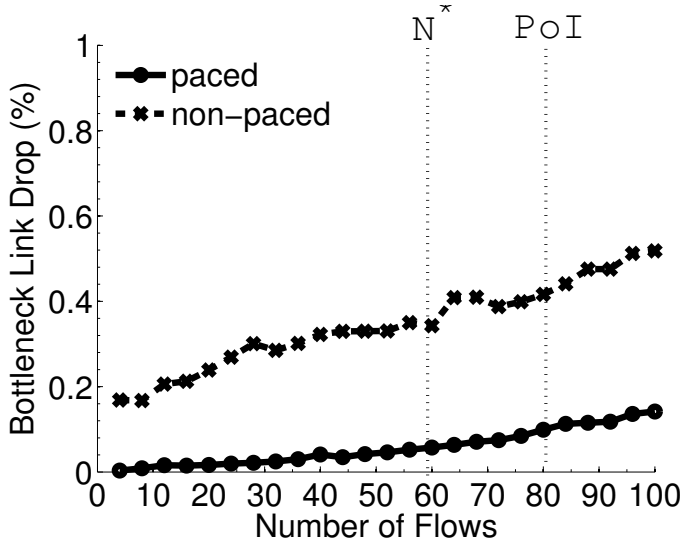


Multiple flows: Link Utilization/Drop/Latency

Buffer size 1.7% of BDP, varying number of flows

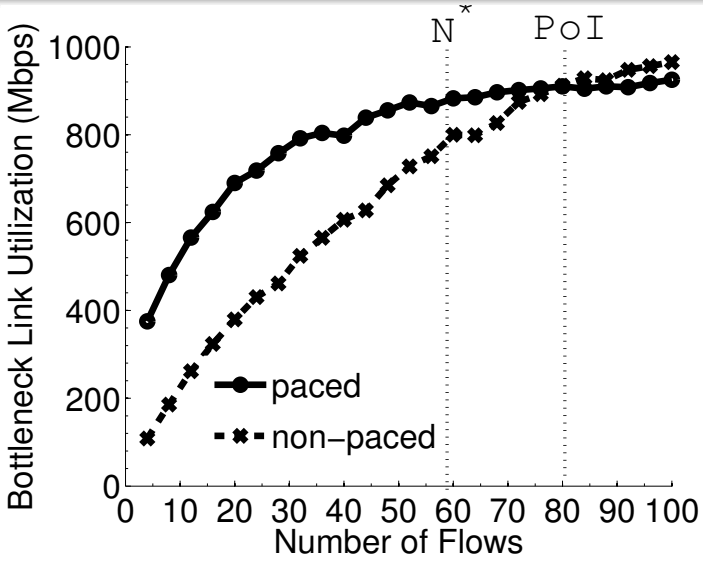


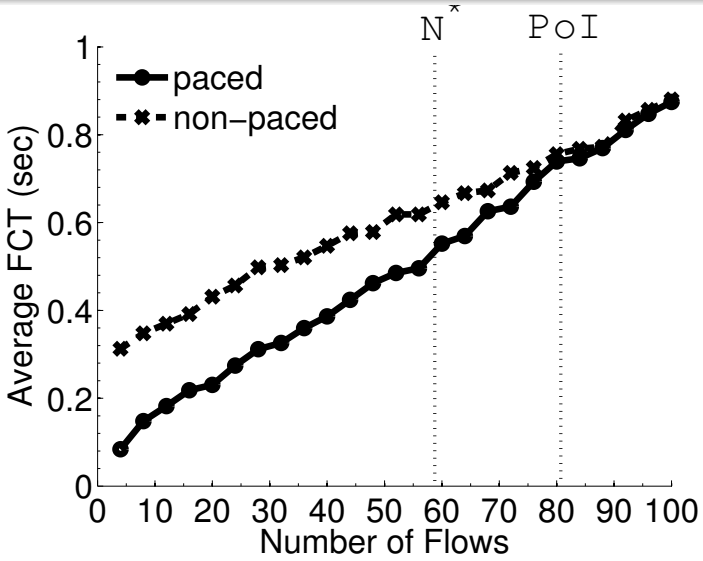


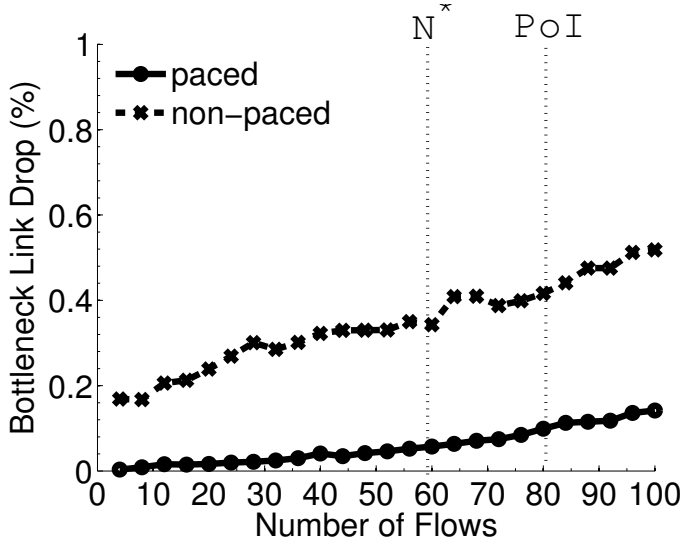


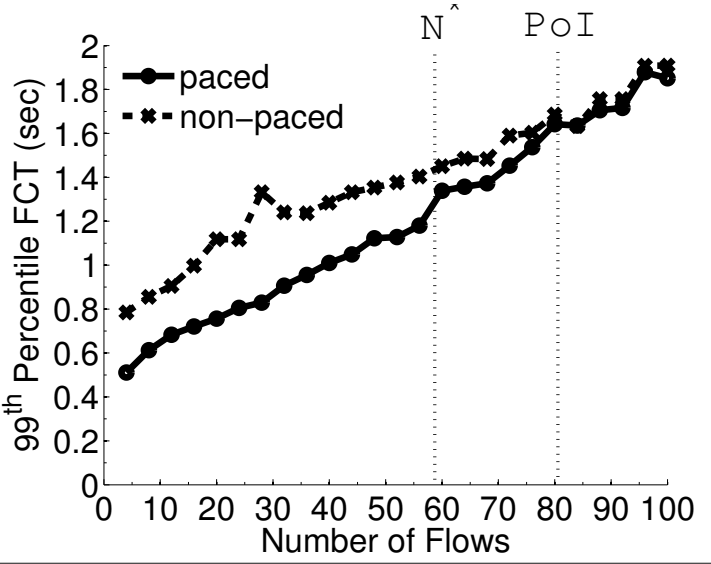
Multiple flows: Link Utilization/Drop/Latency

Buffer size 1.7% of BDP, varying number of flows



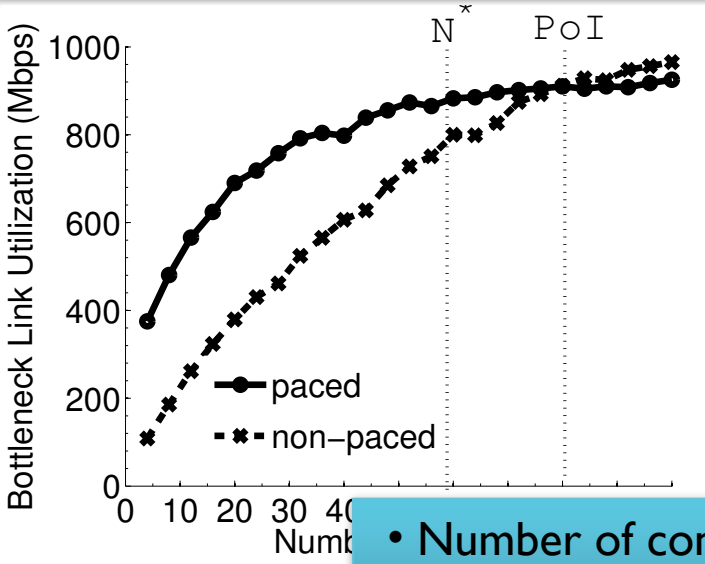


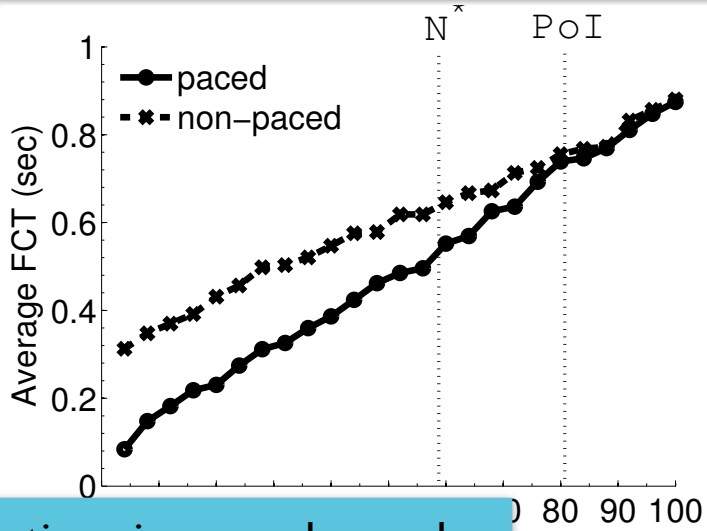




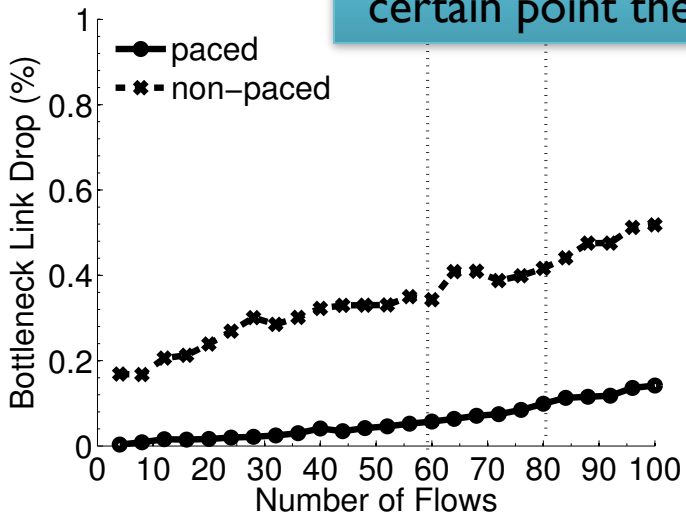
Multiple flows: Link Utilization/Drop/Latency

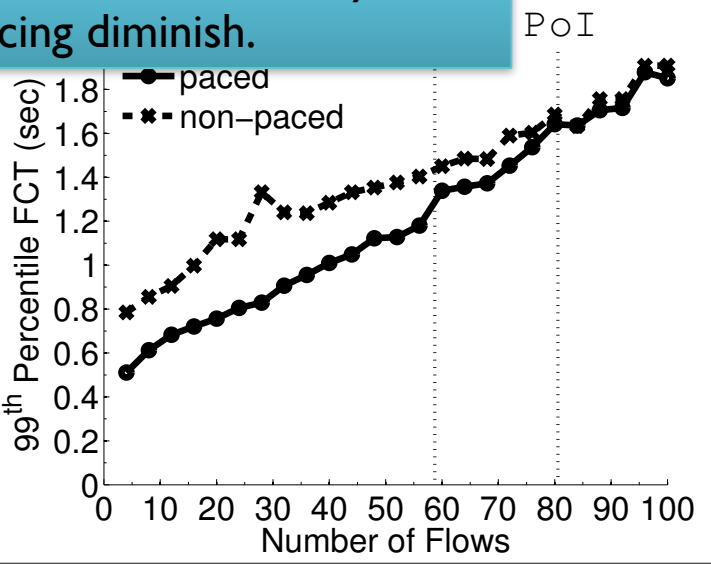
Buffer size 1.7% of BDP, varying number of flows





- Number of concurrent connections increase beyond a certain point the benefits of pacing diminish.



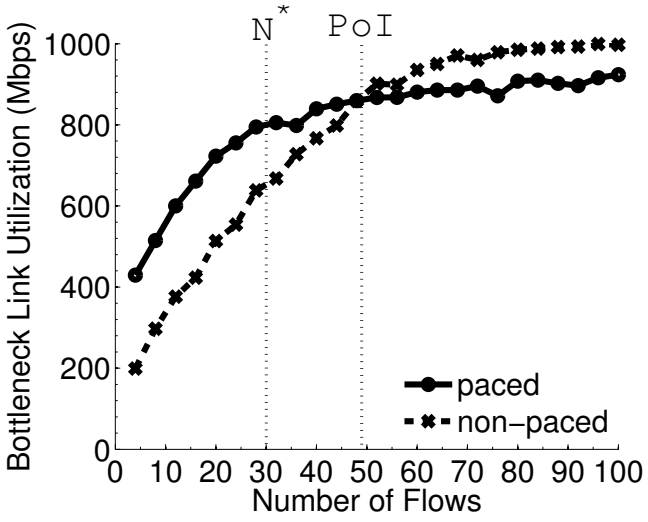


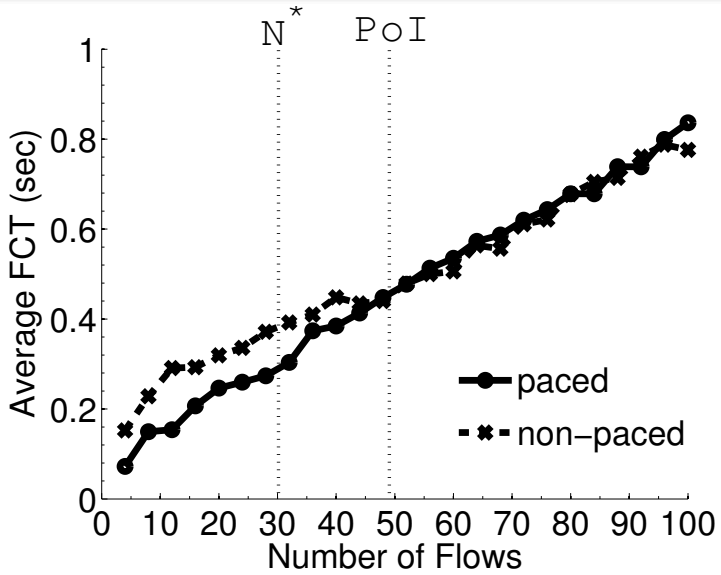
Multiple flows: Link Utilization/Drop/Latency

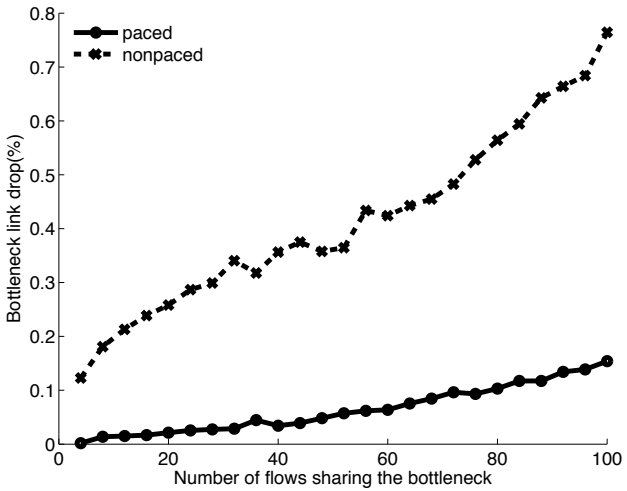
Buffer size 3.4% of BDP, varying number of flows

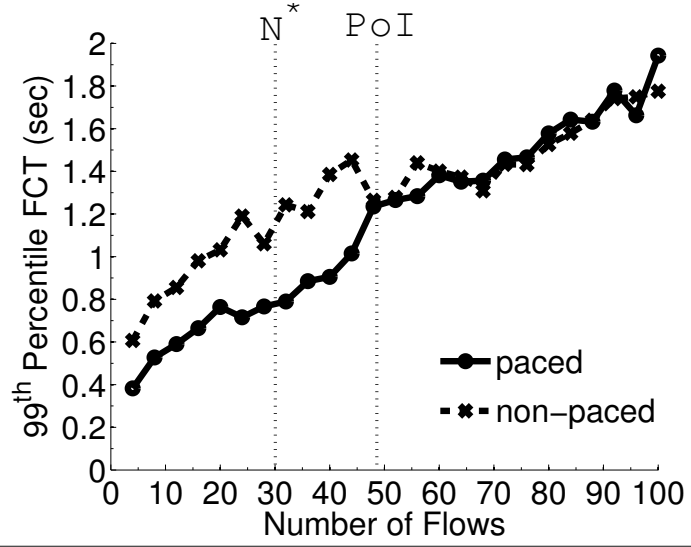
Multiple flows: Link Utilization/Drop/Latency

Buffer size 3.4% of BDP, varying number of flows



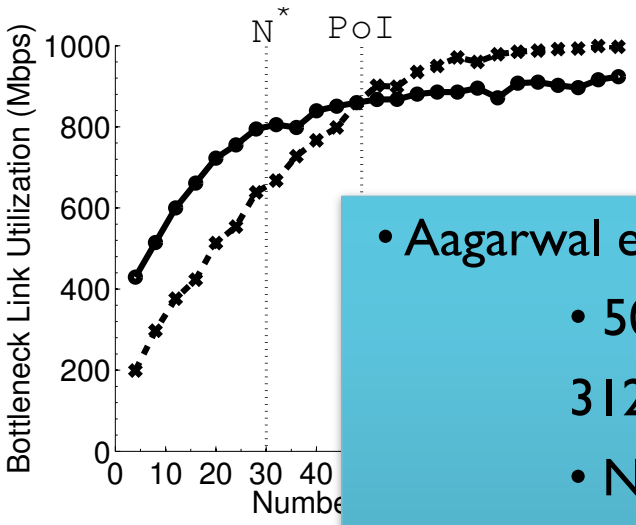


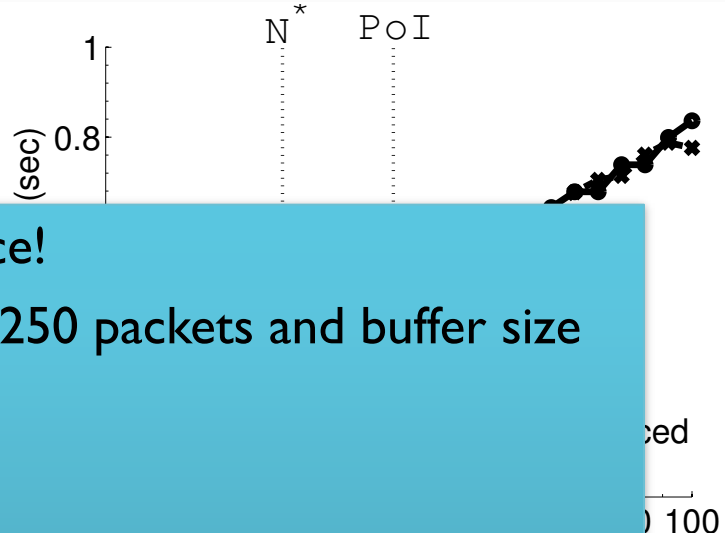




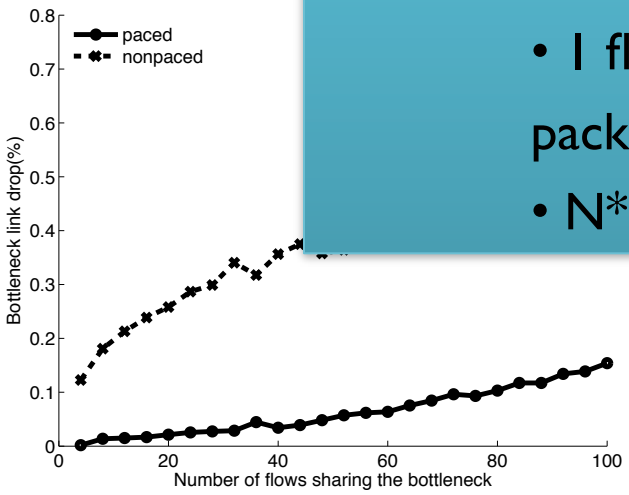
Multiple flows: Link Utilization/Drop/Latency

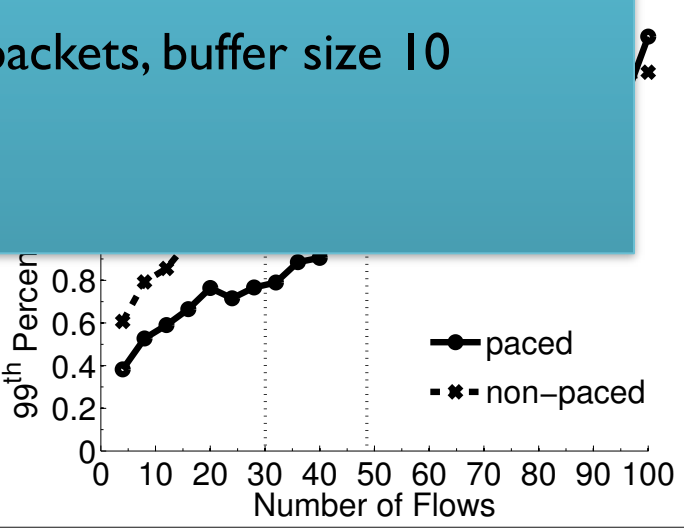
Buffer size 3.4% of BDP, varying number of flows





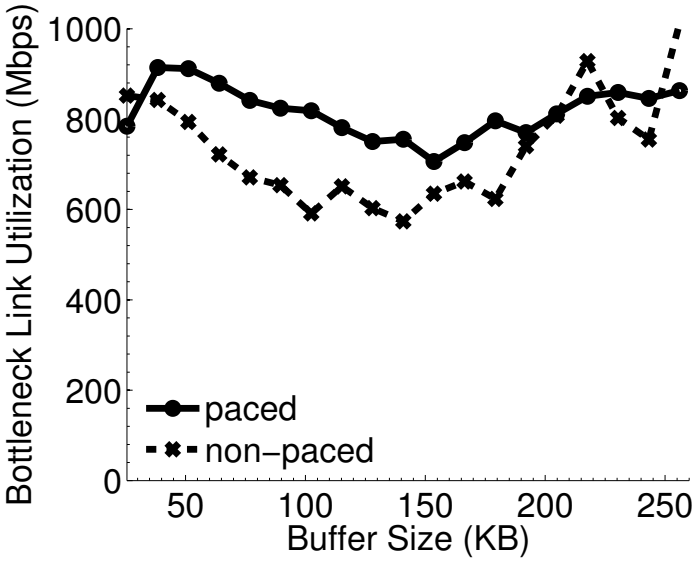
- Aagarwal et al.: Don't pace!
  - 50 flows, BDP 1250 packets and buffer size 312 packets
  - N* = 8 flows.
- Kulik et al.: Pace!
  - 1 flow, BDP 91 packets, buffer size 10 packets.
  - N* = 9 flows.



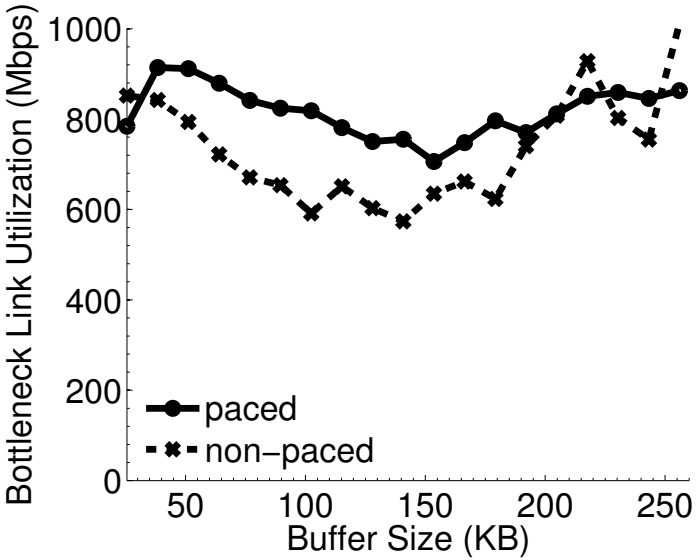


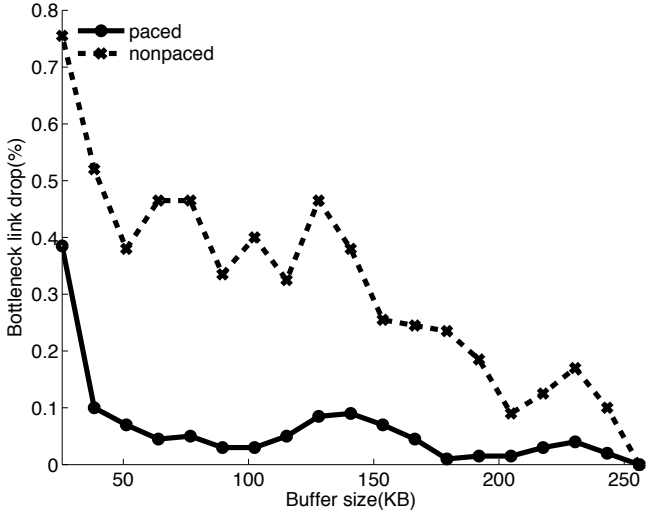
N* vs. Buffer

N* vs. Buffer

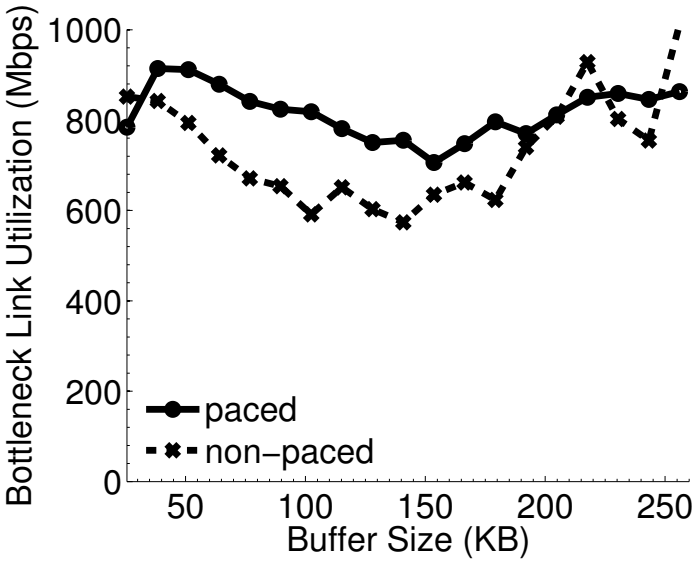

Bottleneck Link Utilization (Mbps)
1000
800
600
400
200
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)

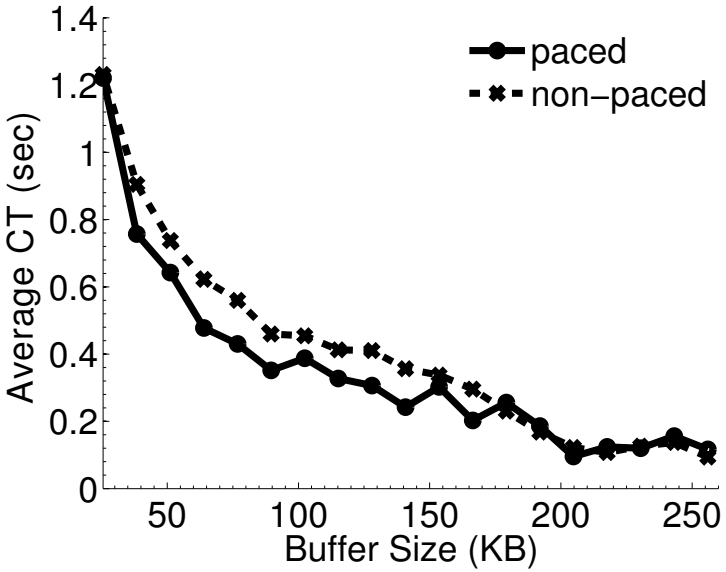
N* vs. Buffer

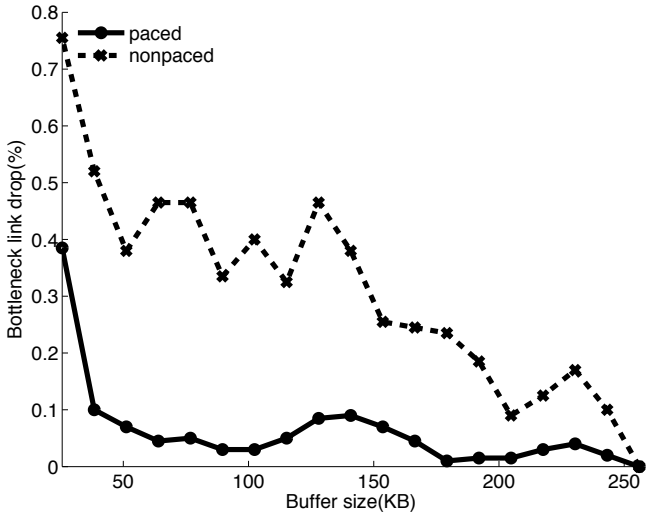

Bottleneck Link Utilization (Mbps)
1000
800
600
400
200
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
paced
nonpaced
Bottleneck link drop(%)
50
100
150
200
250
Buffer size(KB)

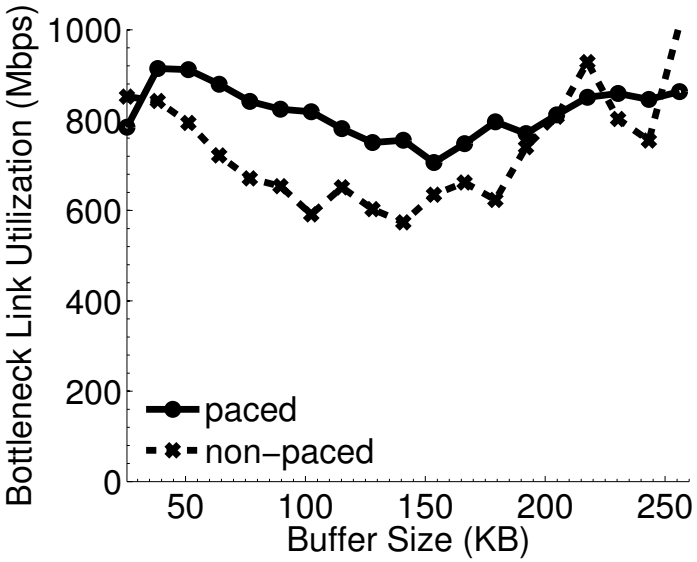
N* vs. Buffer

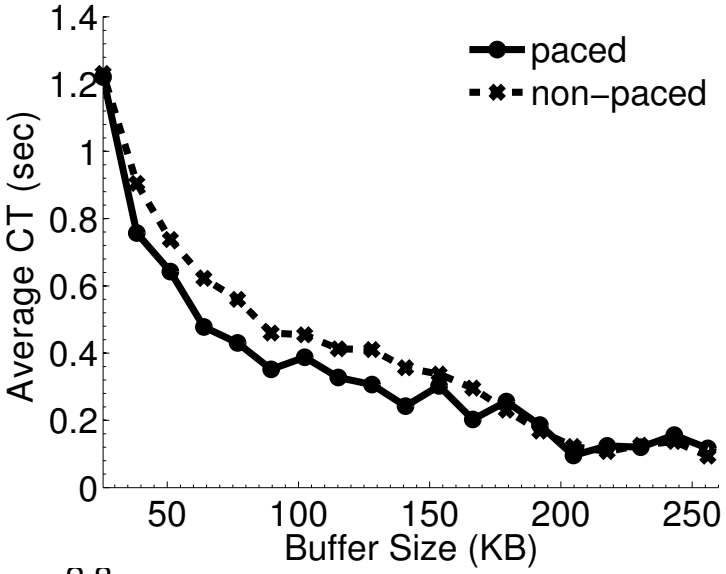

Bottleneck Link Utilization (Mbps)
1000
800
600
400
200
0
50
100
150
200
250
Buffer Size (KB)
paced
non−paced

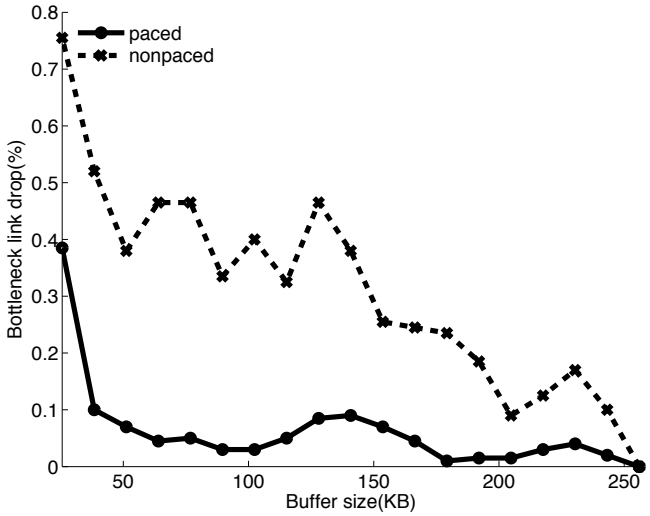

Average CT (sec)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
Buffer Size (KB)
paced
non−paced

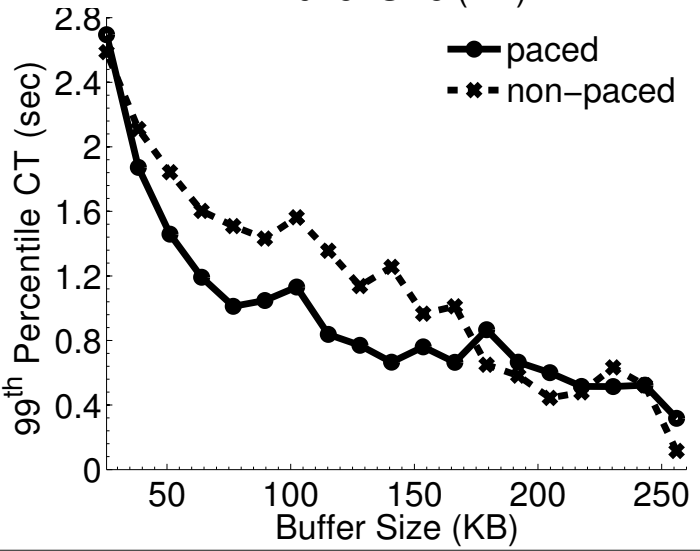

Bottleneck link drop(%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Buffer size(KB)
paced
nonpaced

N* vs. Buffer

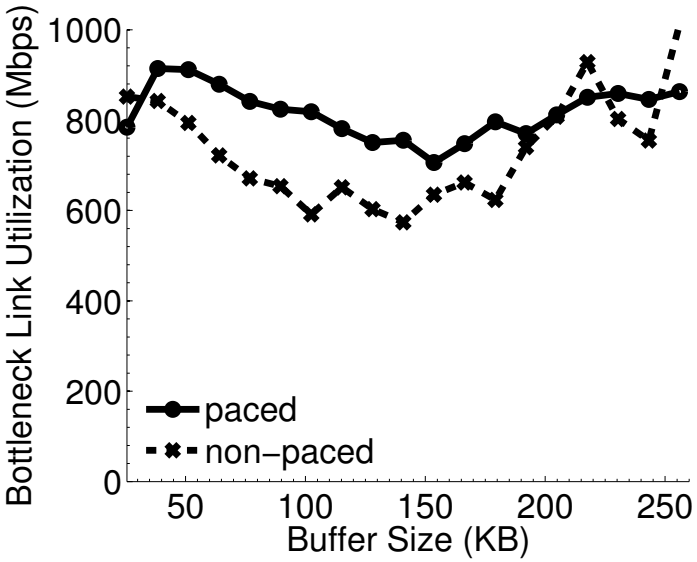

Bottleneck Link Utilization (Mbps)
1000
800
600
400
200
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)

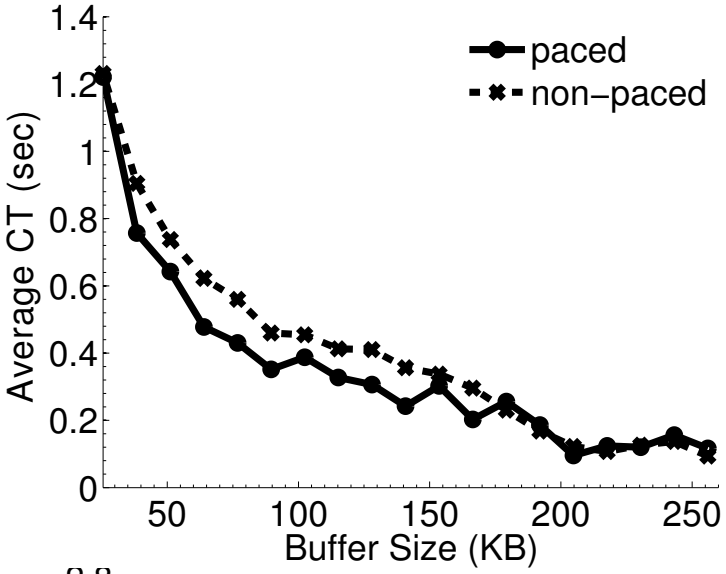

Average CT (sec)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)

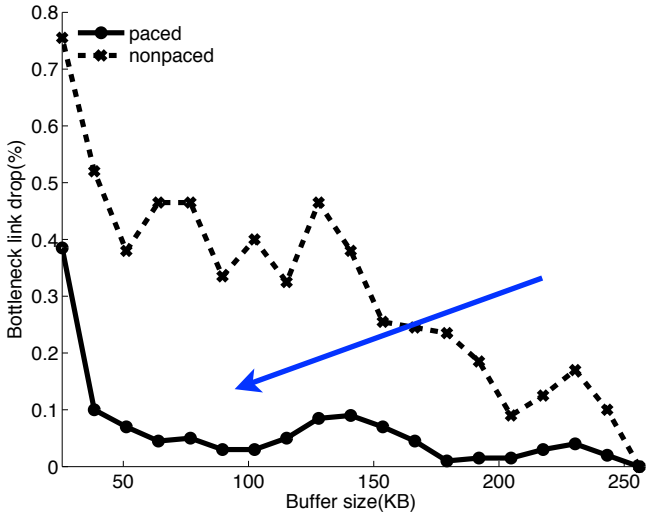

Bottleneck link drop(%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
paced
nonpaced
50
100
150
200
250
Buffer size(KB)

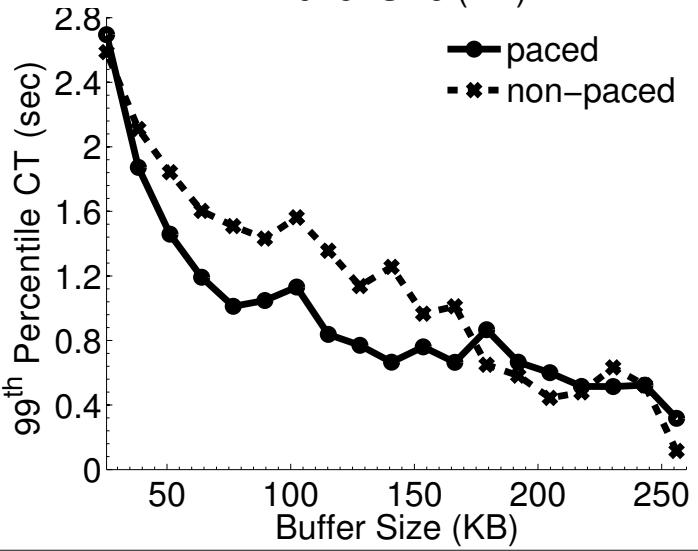

99th Percentile CT (sec)
2.8
2.4
2
1.6
1.2
0.8
0.4
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)

N* vs. Buffer


Bottleneck Link Utilization (Mbps)
1000
800
600
400
200
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)


Average CT (sec)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)


Bottleneck link drop(%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
paced
nonpaced
50
100
150
200
250
Buffer size(KB)

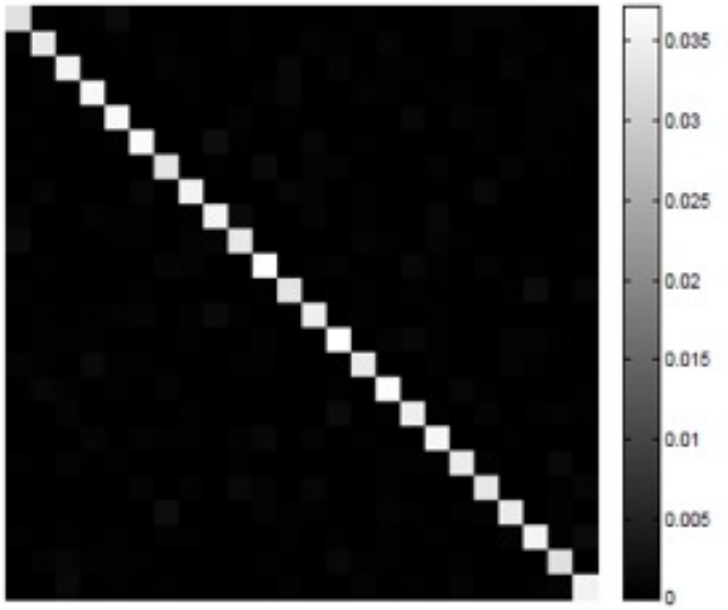

99th Percentile CT (sec)
2.8
2.4
2
1.6
1.2
0.8
0.4
0
paced
non−paced
50
100
150
200
250
Buffer Size (KB)

Clustering Effect:

The probability of packets from a flow being followed by packets from other flows

Clustering Effect:

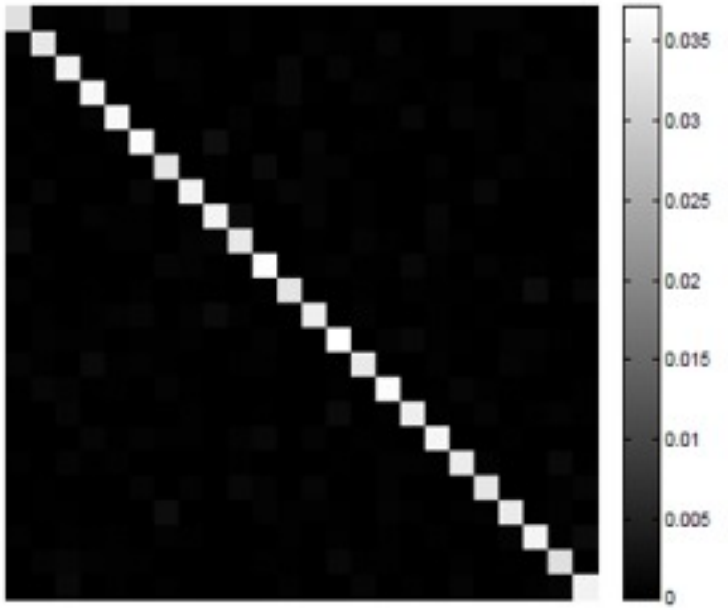
The probability of packets from a flow being followed by packets from other flows



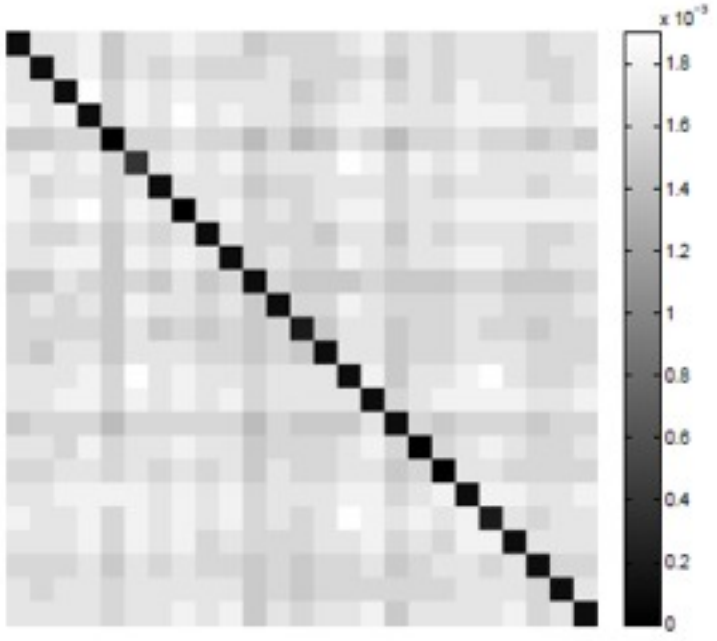
Non-paced: Packets of each flow are clustered together.

Clustering Effect:

The probability of packets from a flow being followed by packets from other flows



Non-paced: Packets of each flow are clustered together.



Paced: Packets of different flows are multiplexed.

Drop Synchronization:

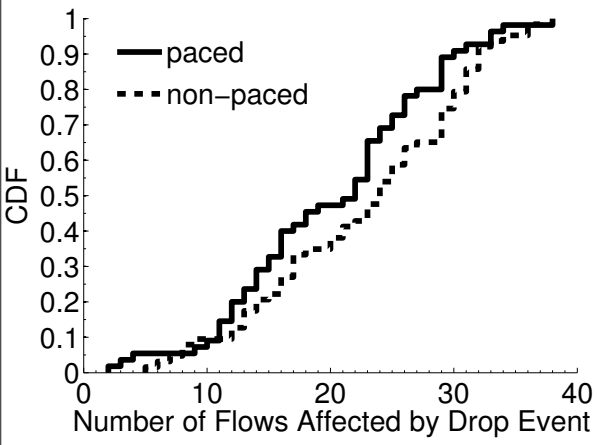
Number of Flows Affected by Drop Event

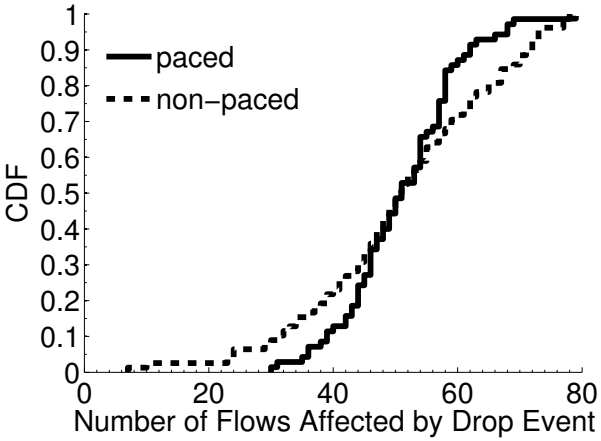
Drop Synchronization:
Number of Flows Affected by Drop Event

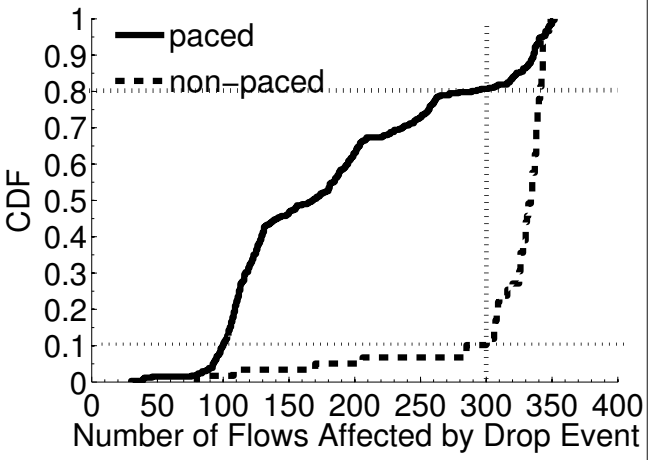
NetFPGA router to count the number of flows affected by drop events.

Drop Synchronization:
Number of Flows Affected by Drop Event

NetFPGA router to count the number of flows affected by drop events.







N: 48

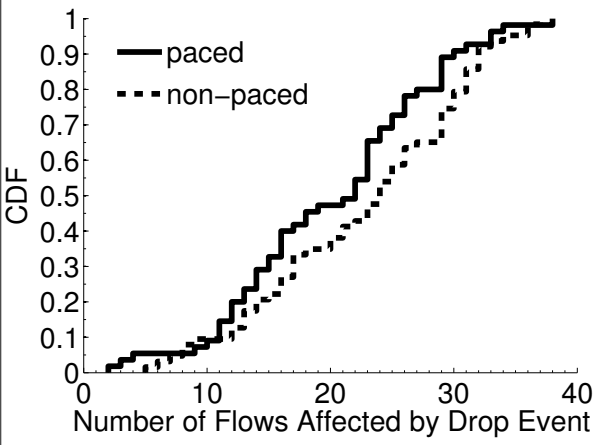
N: 96

N: 384

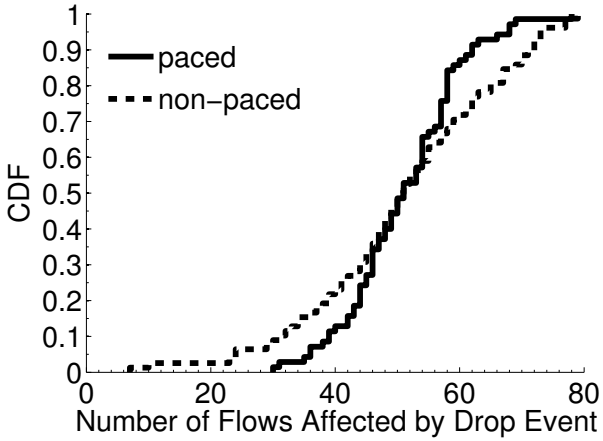
Drop Synchronization:

Number of Flows Affected by Drop Event

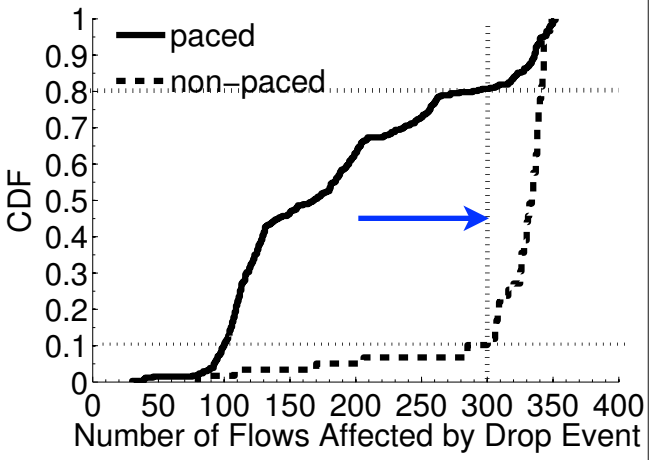
NetFPGA router to count the number of flows affected by drop events.



N: 48



N: 96



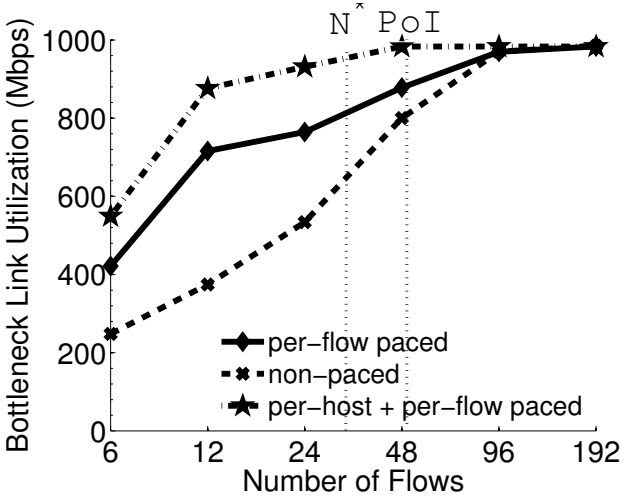
N: 384

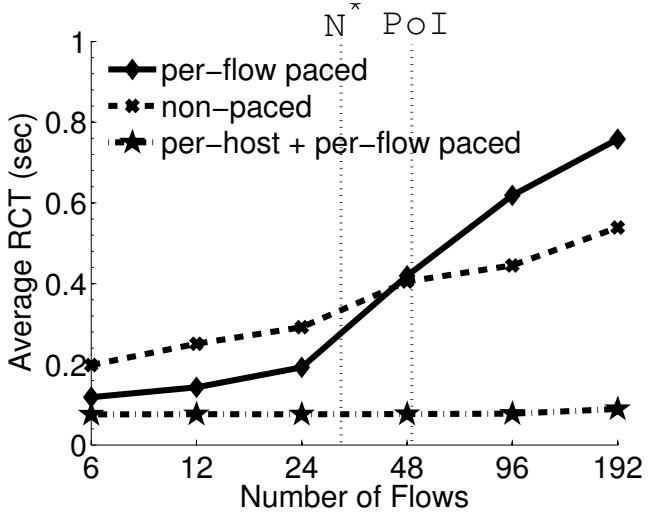
Future Trends for Pacing:

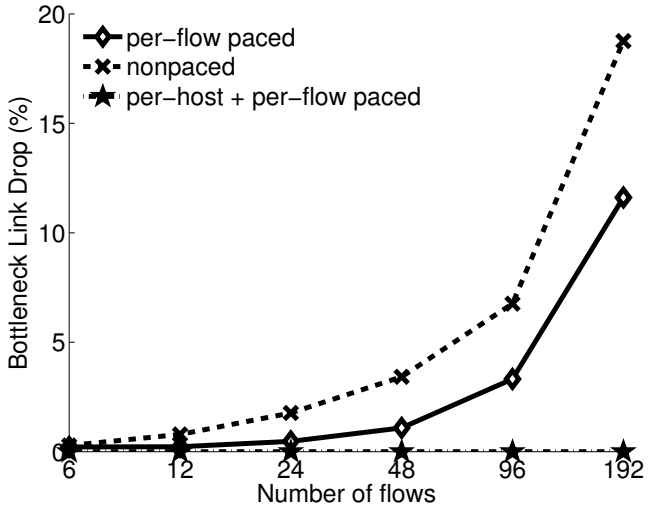
per-egress pacing.

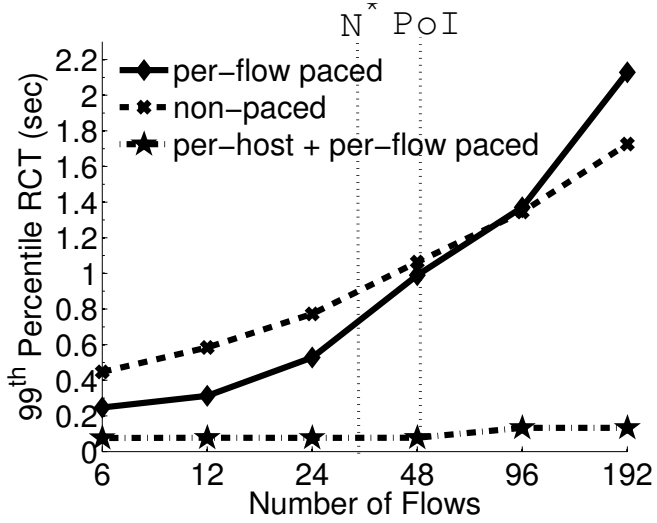
Future Trends for Pacing:

per-egress pacing.



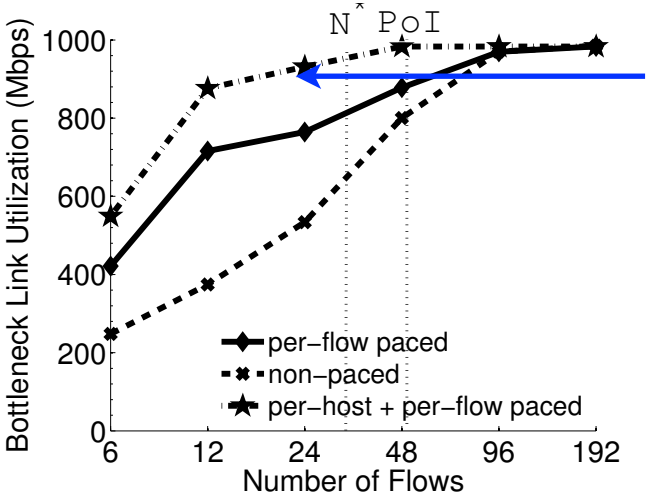


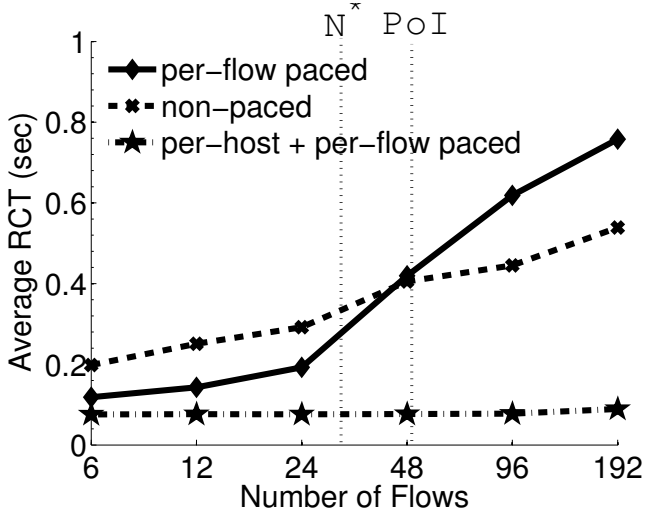


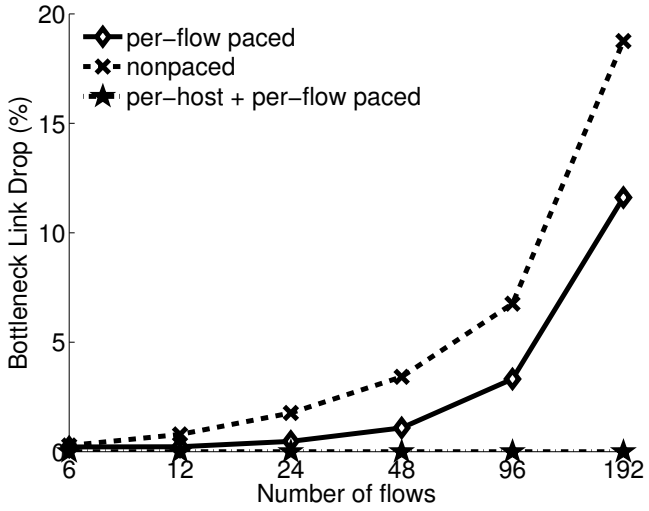


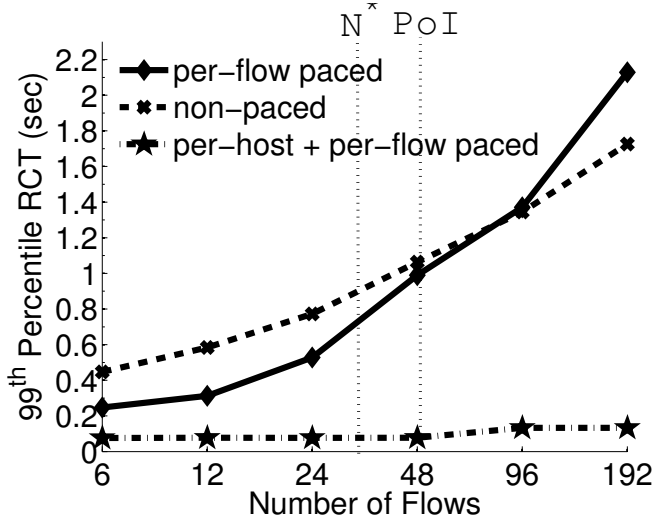
Future Trends for Pacing:

per-egress pacing.



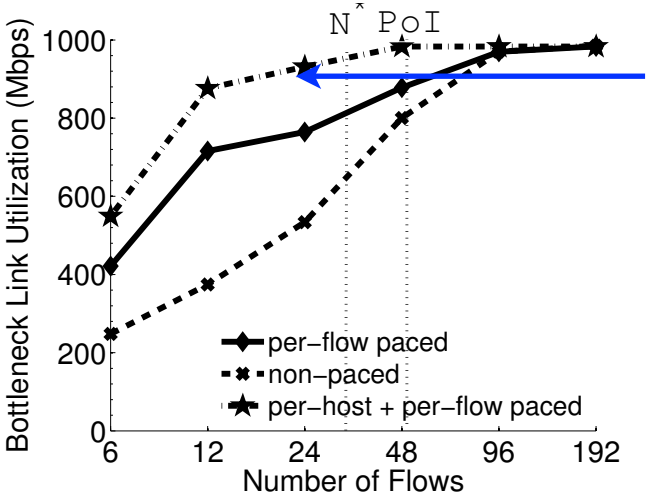


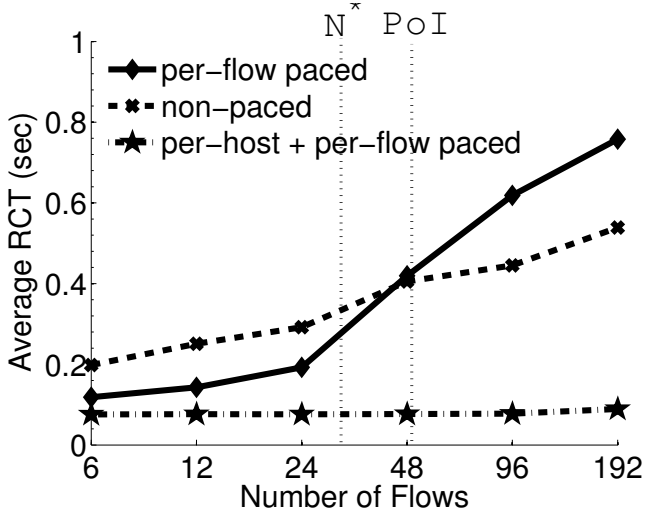


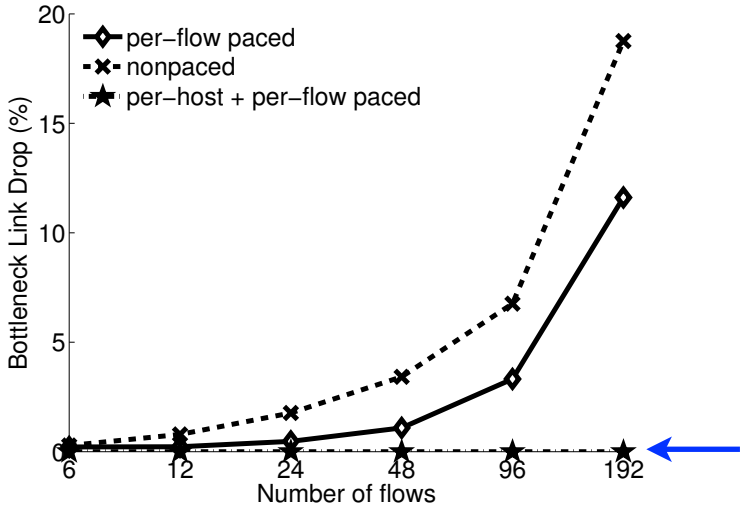


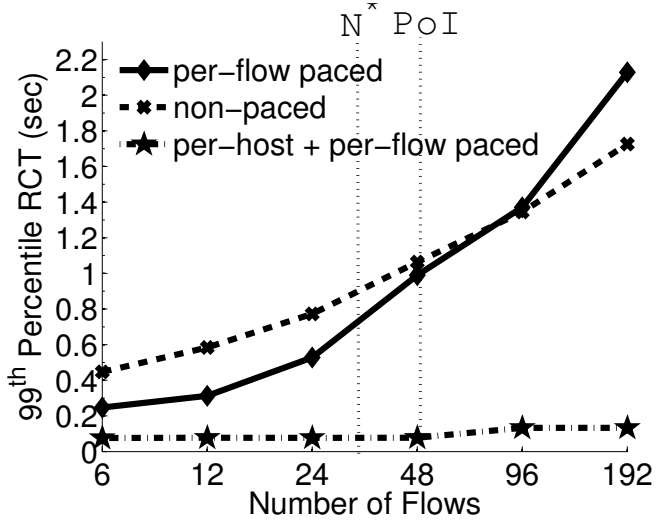
Future Trends for Pacing:

per-egress pacing.



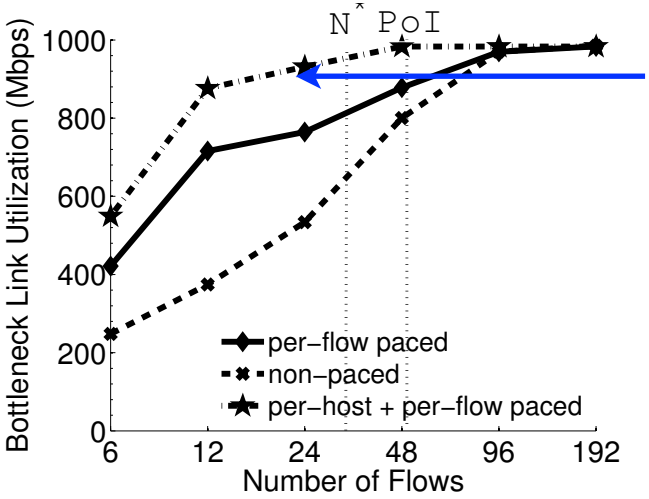


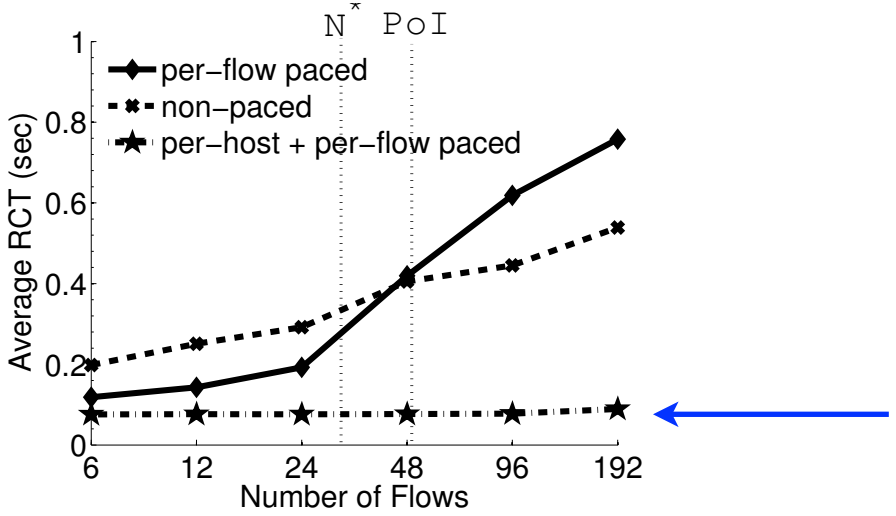


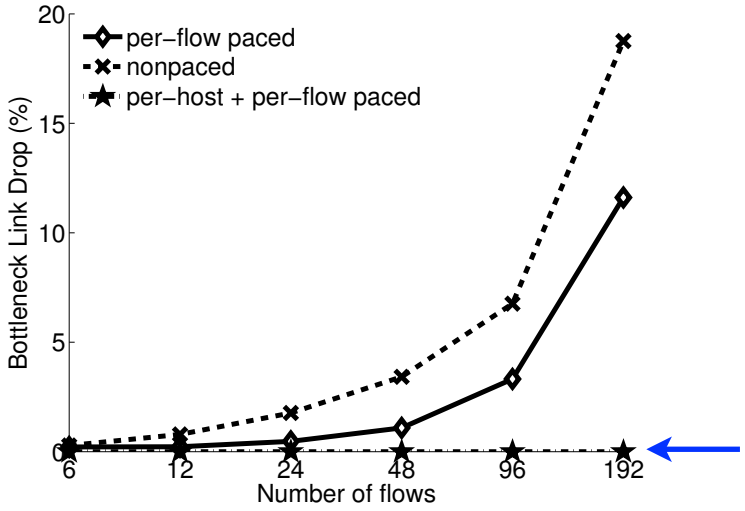


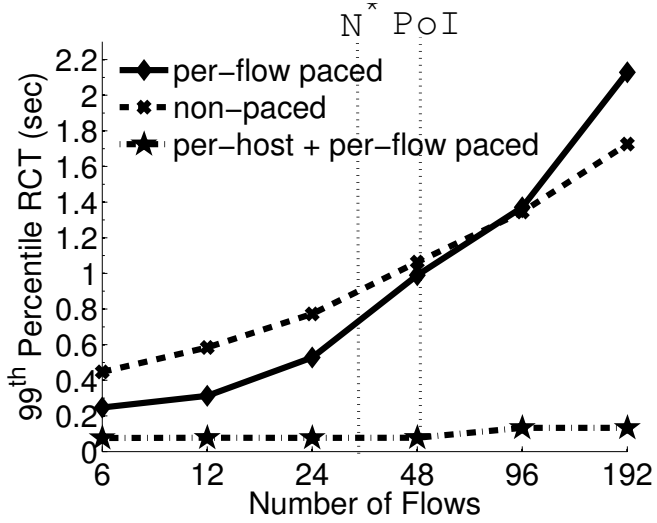
Future Trends for Pacing:

per-egress pacing.



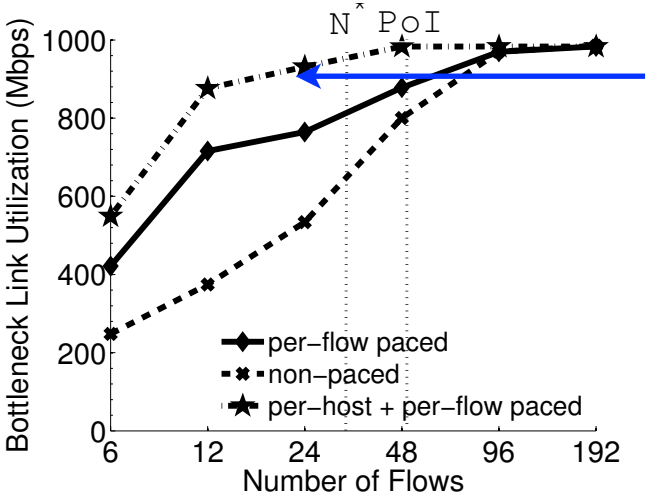


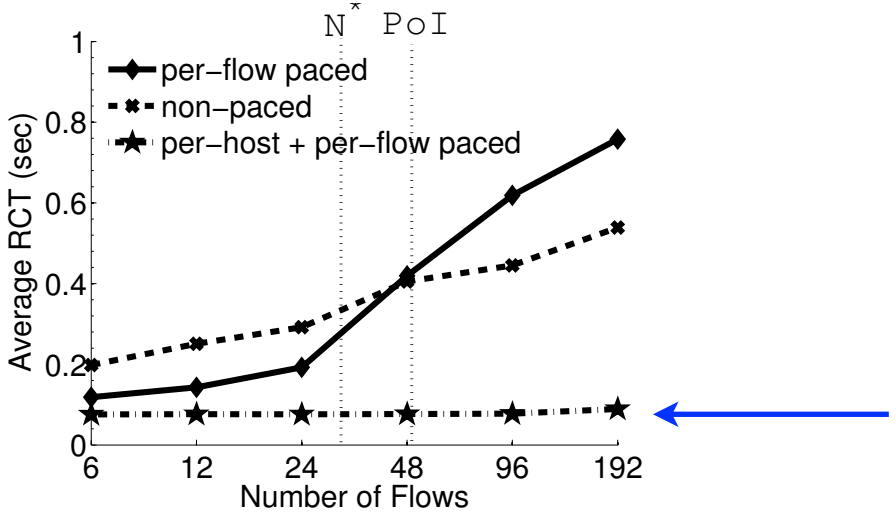


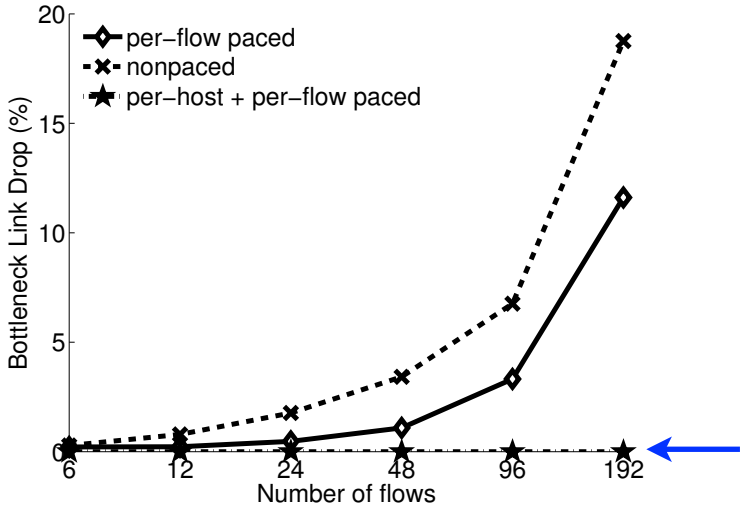


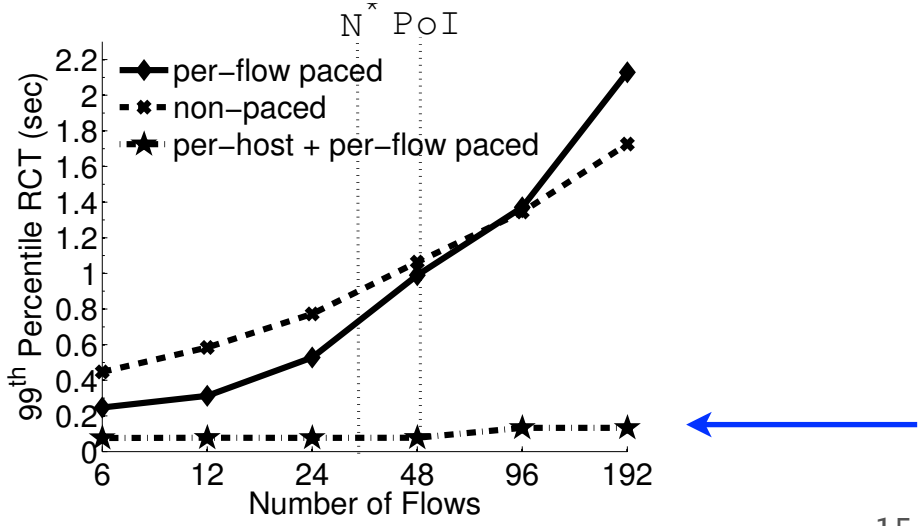
Future Trends for Pacing:

per-egress pacing.









Conclusions and Future work

- Re-examine TCP pacing's effectiveness:
  - Demonstrate when TCP pacing brings benefits in such environments.
- Inter-flow burstiness

- Burst-pacing vs. packet-pacing.
- Per-egress pacing.

Renewed Interest

[tcpm] (reducing) tcp bursts Inbox x

Yuchung Cheng via ietf.org

to tcpm

There are a lot of discussion on bursts across talks.
1. newcwv: idle-restart
2. tlp: how often is tail drops be caused by (higher) initial burst/send
3. burst (loss) after recovery due to snd.una + rwin jump.
4. I can throw in another one: video player application throttle sender by not reading the socket or clamp the receive buffer. But this causes TCP to burst when rwin opens up.

I think the working group should work on a general solution to reduce burst in the window-based, ack-clocked, TCP. I have heard solutions like
1. BSD/randy's max-burst solution
2. pace cwnd/rtt but in max-burst chunks
3. more ideas in http://www.isi.edu/touch/pubs/draft-hughes-restart-00.txt

We all know TCP is very smooth in bulk transfer. Unfortunately modern Apps are chatty even on video.

Thoughts?

---

tcpm mailing list
tcpm@ietf.org
https://www.ietf.org/mailman/listinfo/tcpm

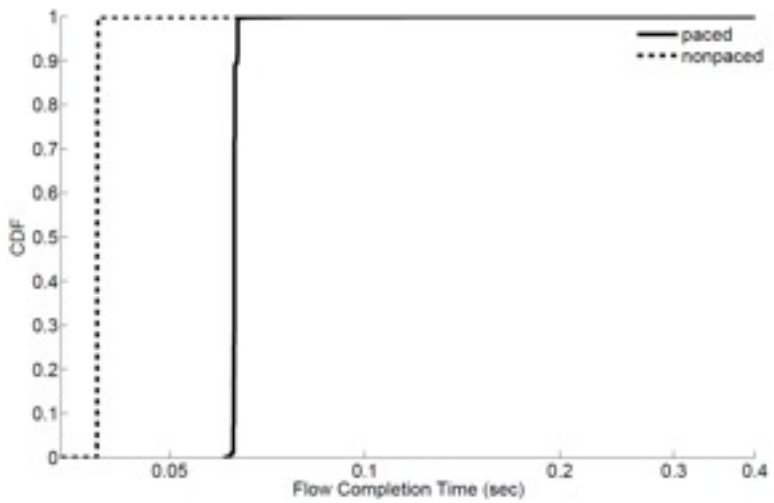
Traffic Burstiness Survey

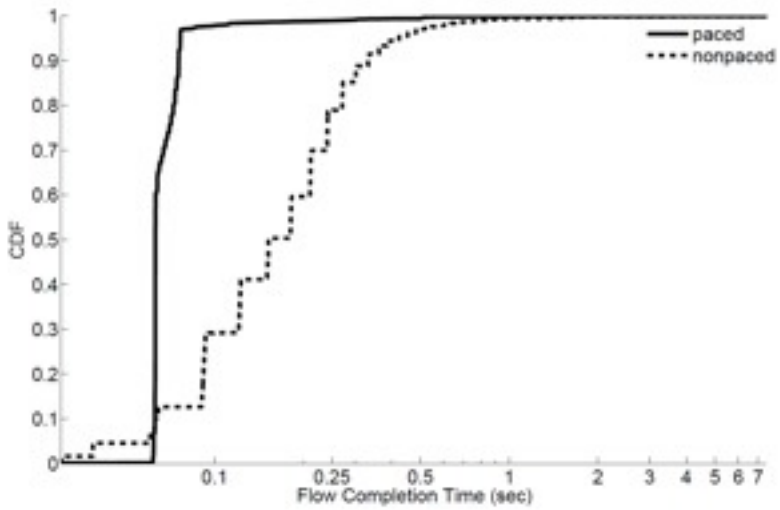
- 'Bursty' is a word with no agreed meaning. How do you define a bursty traffic?
- If you are involved with a data center, is your data center traffic bursty?
  - If yes, do you think that it will be useful to supress the burstiness in your traffic?
  - If no, are you already supressing the burstiness? How? Would you anticipate the traffic becoming burstier in the future?

monia@cs.toronto.edu

Base-Case Experiment:

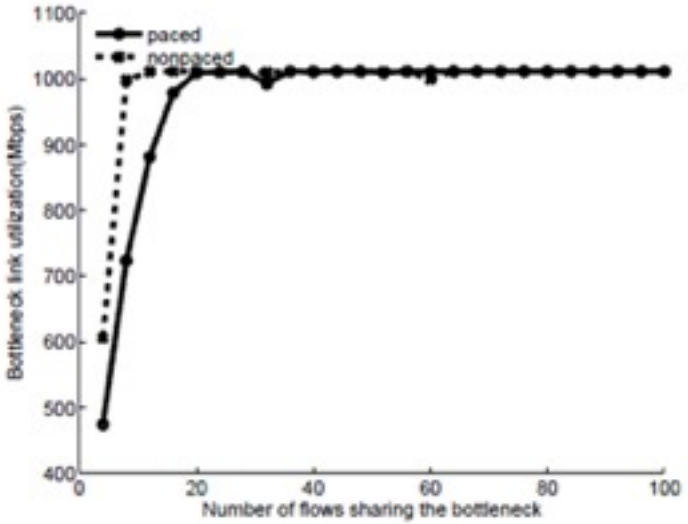
One RPC vs Two RPCs, 64KB of buffering, Latency

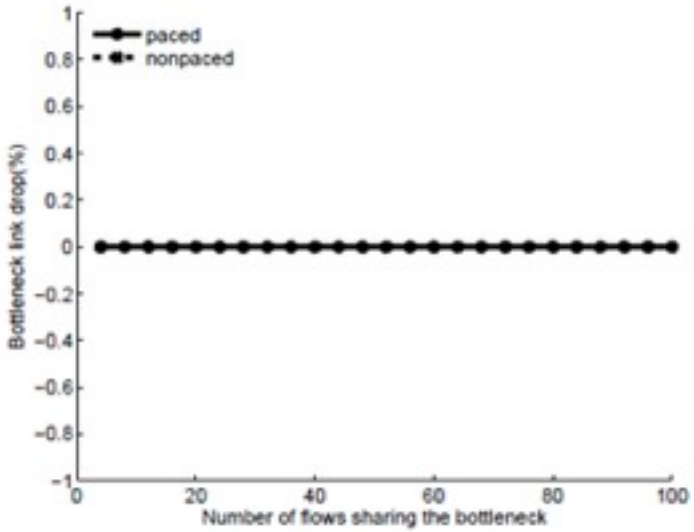


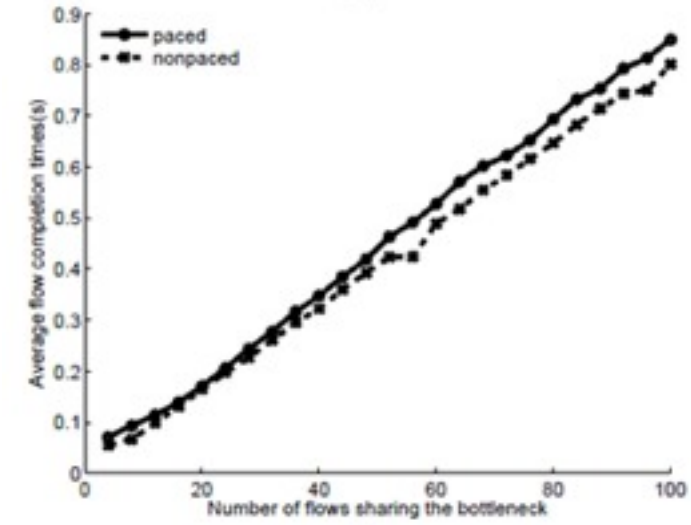


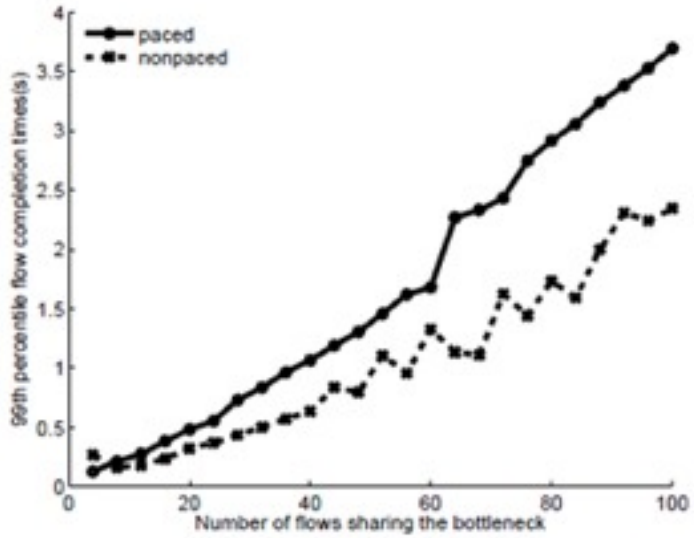
Multiple flows: Link Utilization/Drop/ Latency

Buffer size: 6% of BDP, varying number of



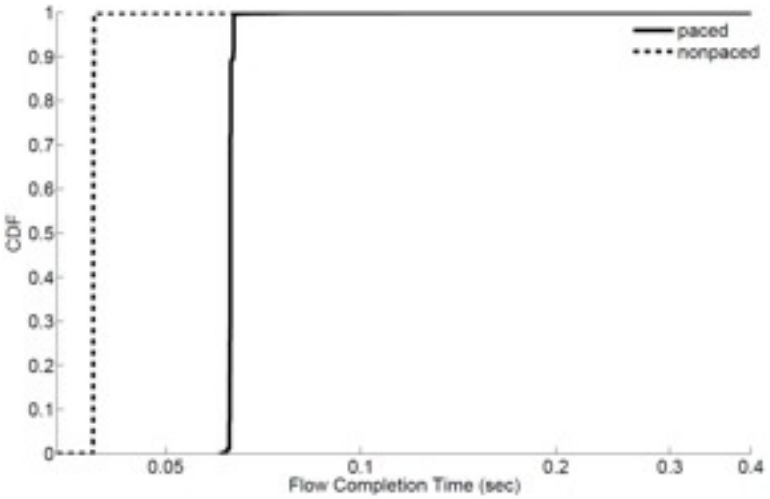






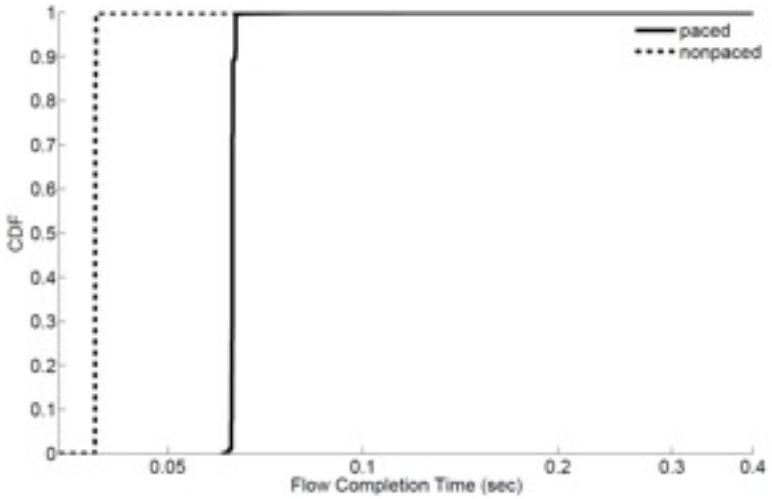
Base-Case Experiment:

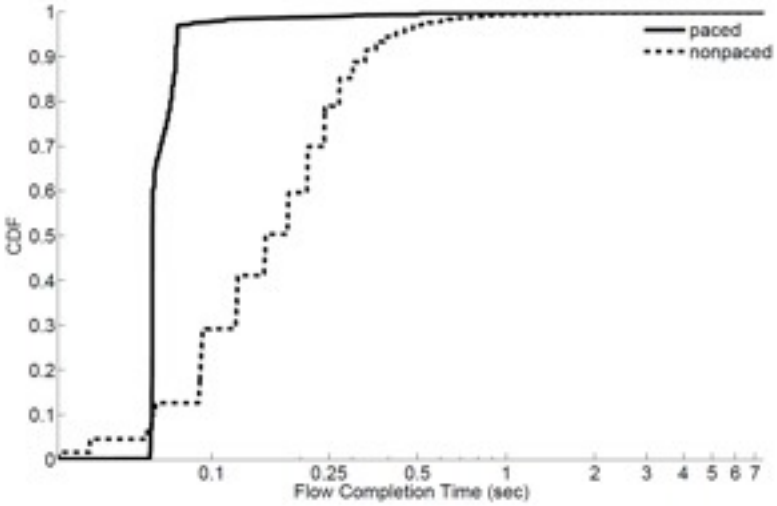
One RPC vs Two RPCs, 64KB of buffering, Latency / Queue Occupancy



Base-Case Experiment:

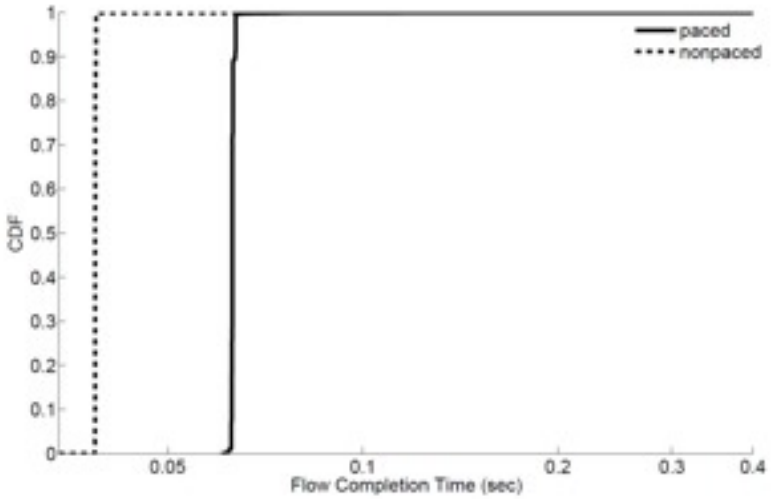
One RPC vs Two RPCs, 64KB of buffering, Latency / Queue Occupancy

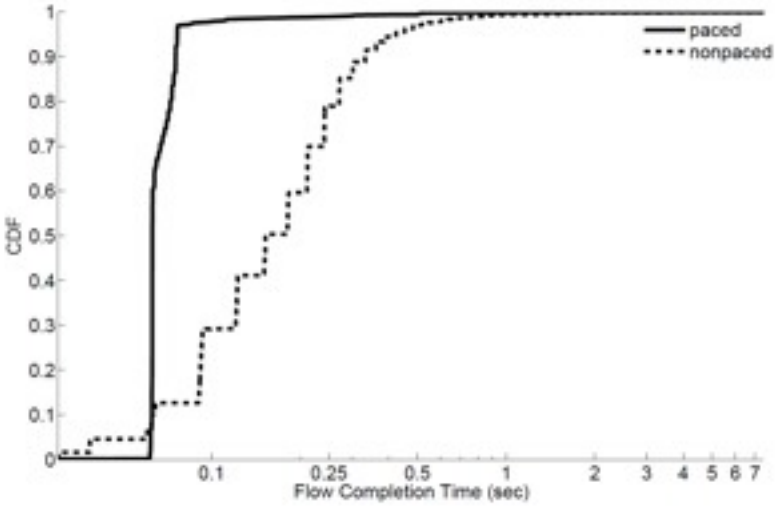


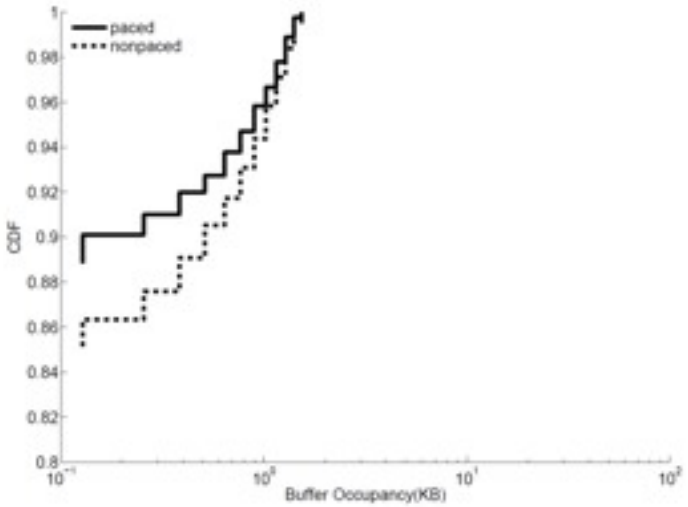


Base-Case Experiment:

One RPC vs Two RPCs, 64KB of buffering, Latency / Queue Occupancy

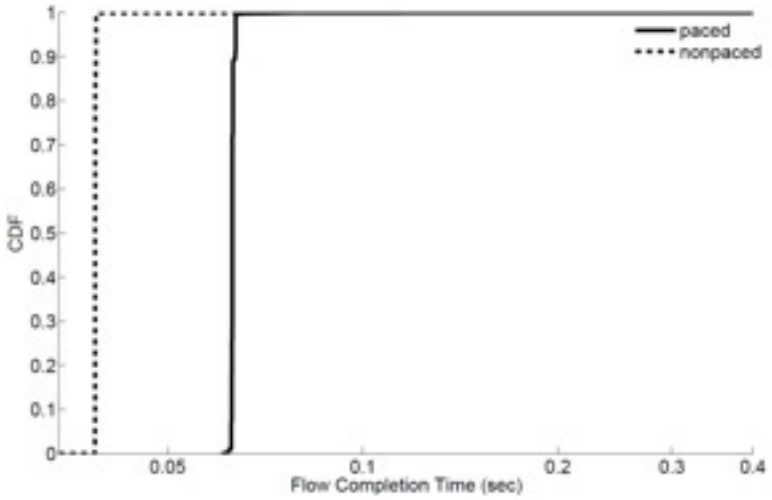


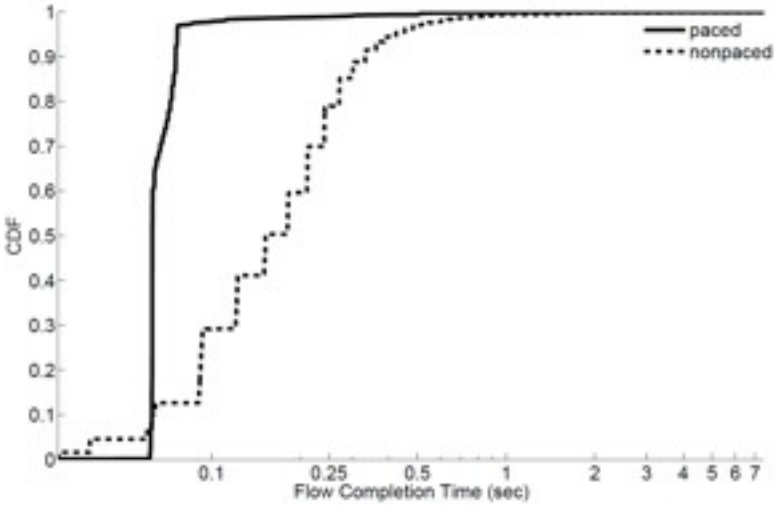


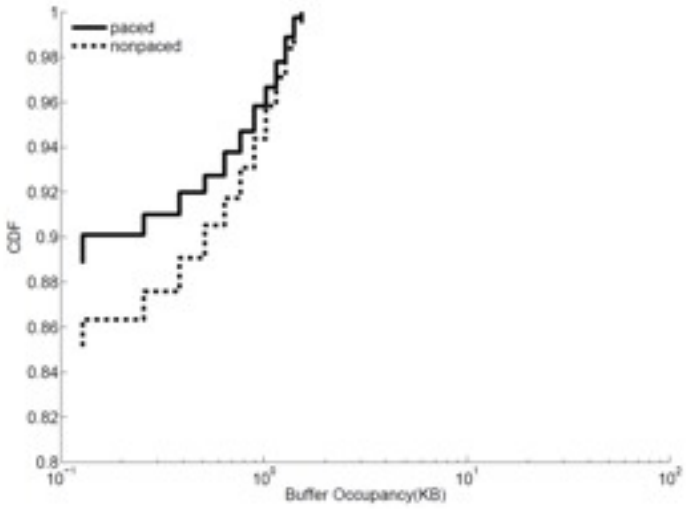


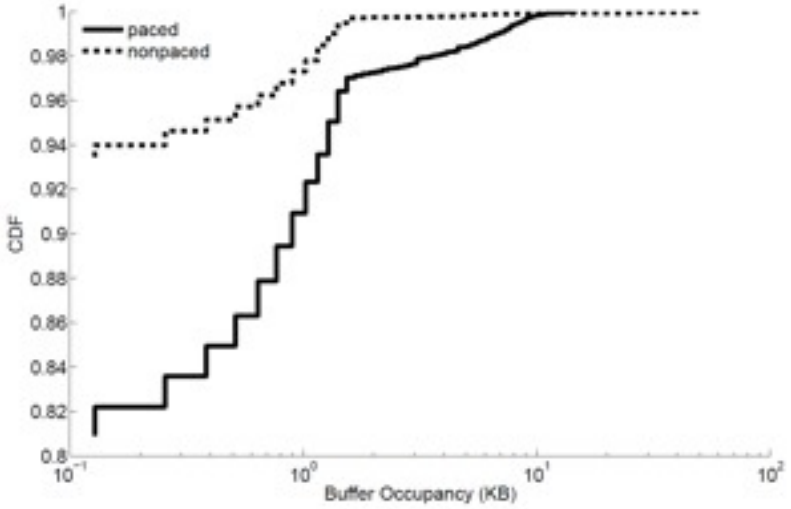
Base-Case Experiment:

One RPC vs Two RPCs, 64KB of buffering, Latency / Queue Occupancy

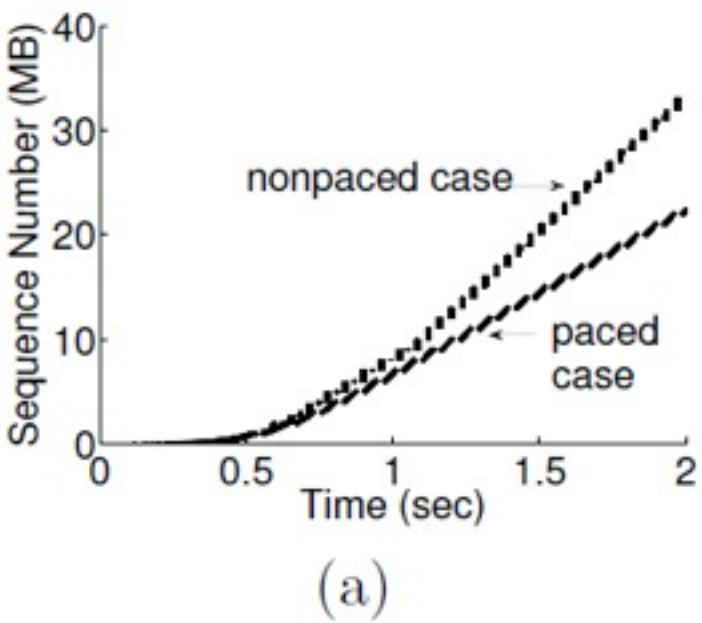






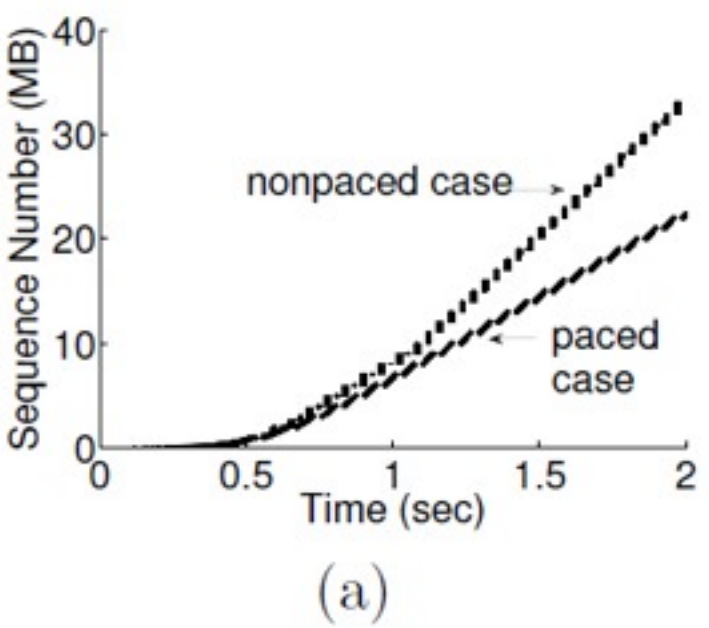


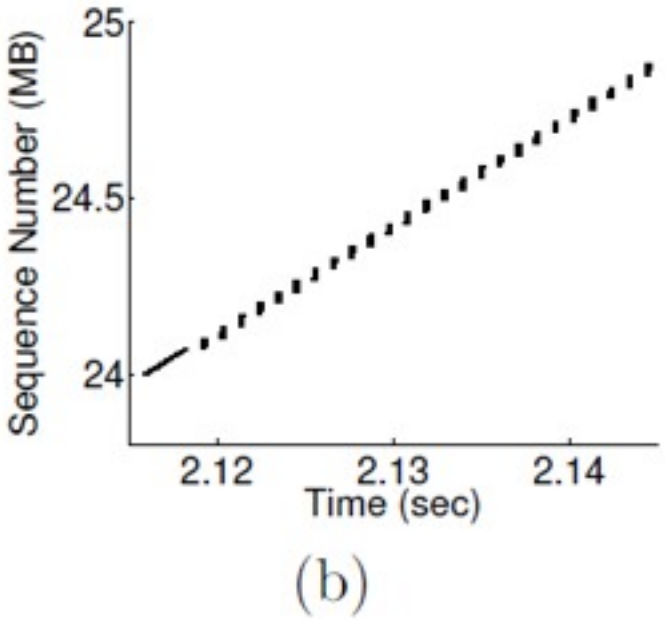
Functional test



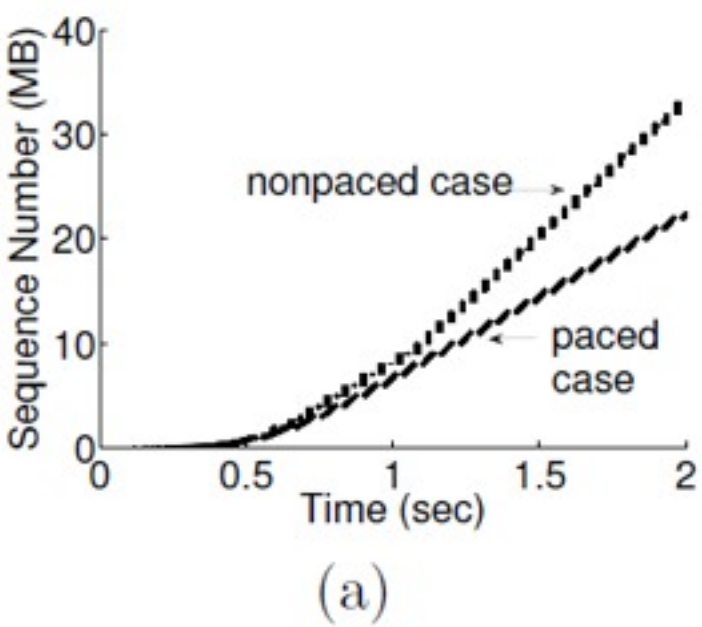
(a)

Functional test

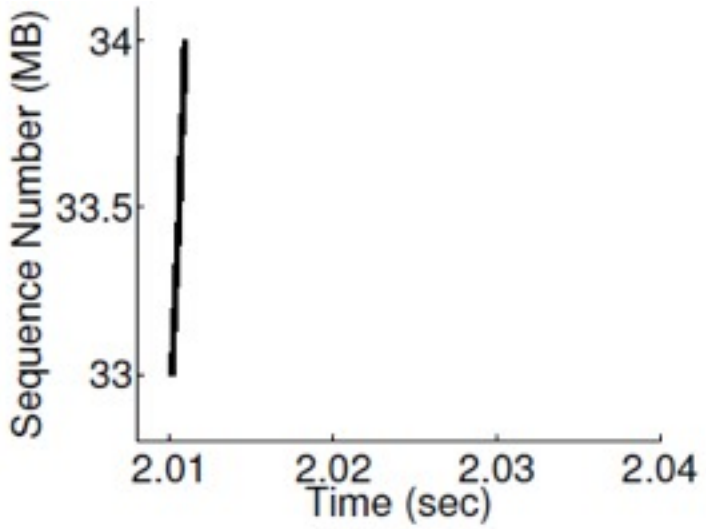




Functional test

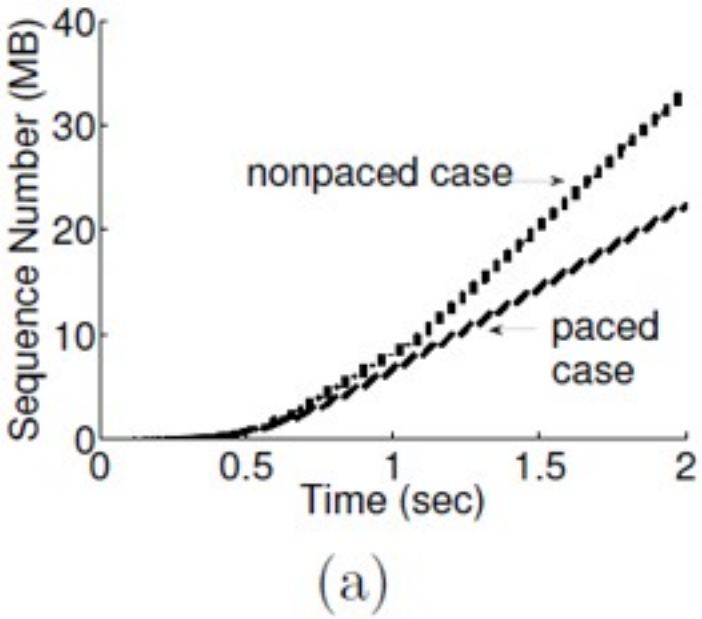


(a)

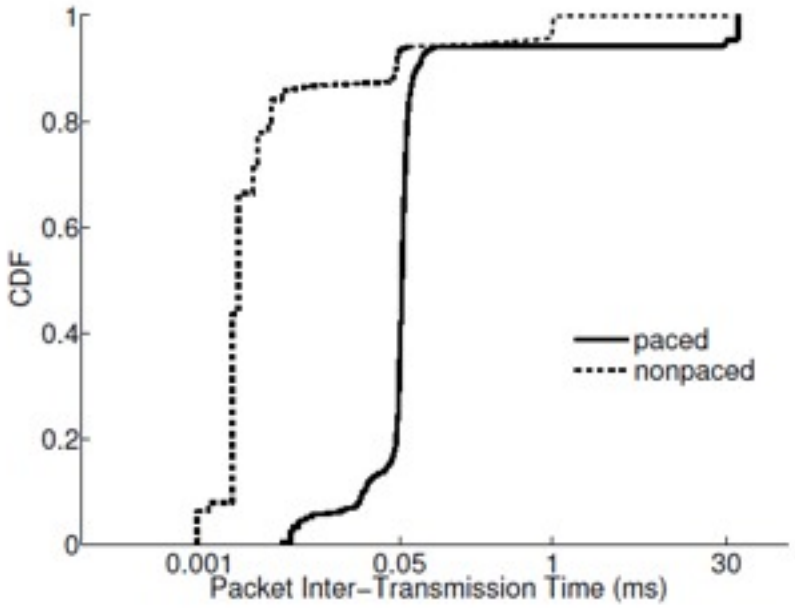


(c)

Functional test

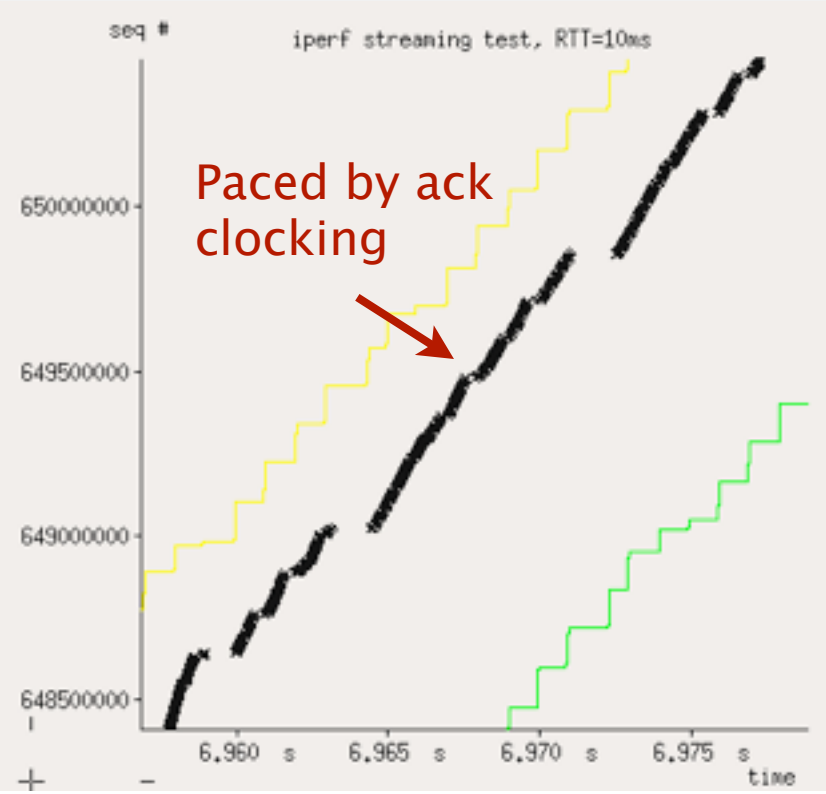


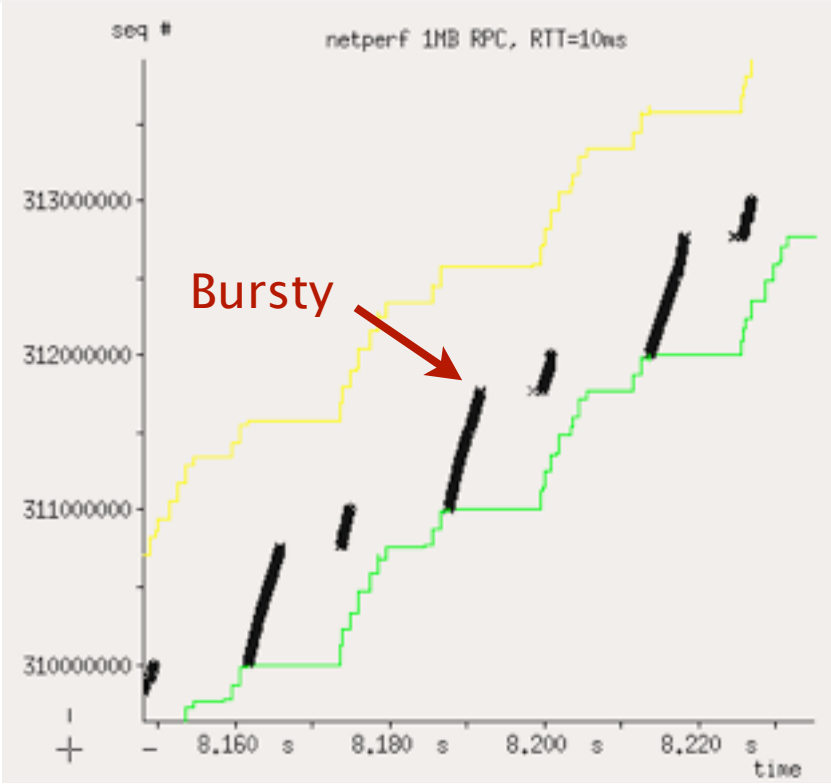
(a)

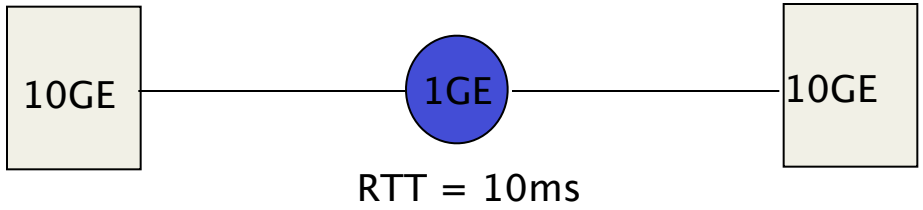


(d)

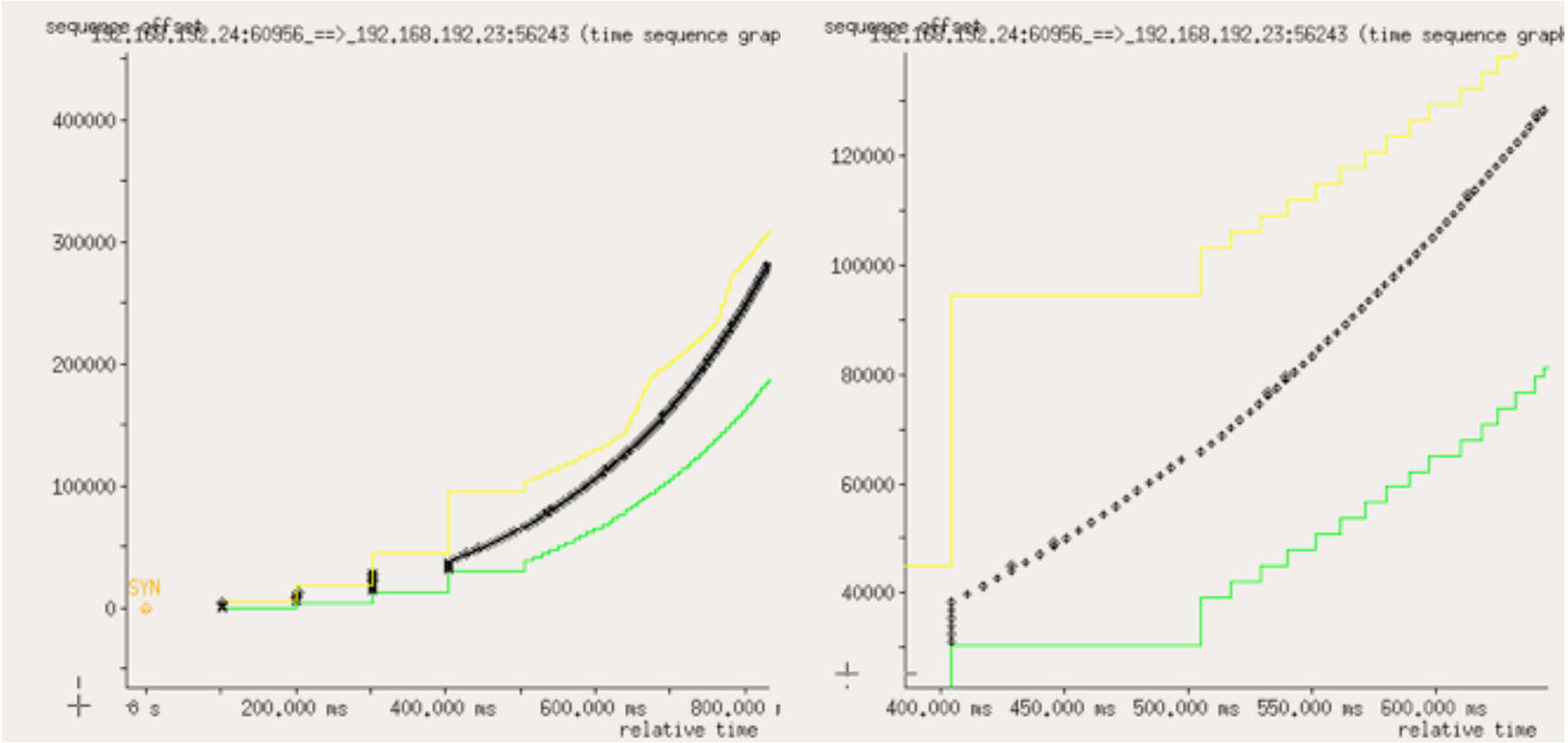
RPC vs. Streaming







Zooming in more on the paced flow

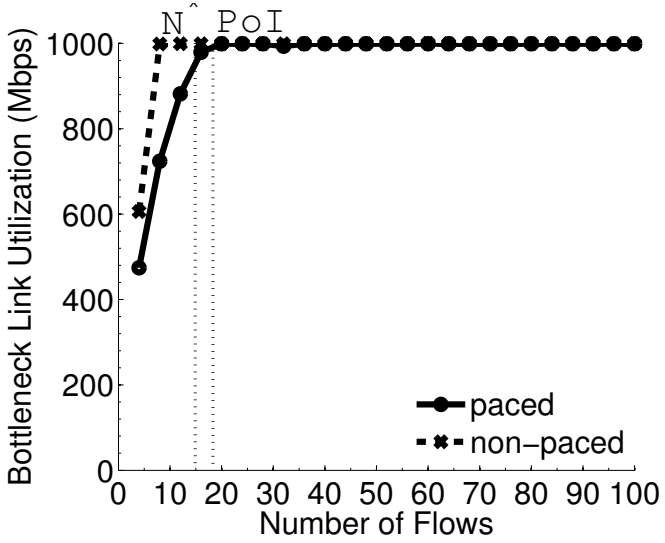


Multiple flows: Link Utilization/Drop/Latency

Buffer size 6.8% of BDP, varying number of flows

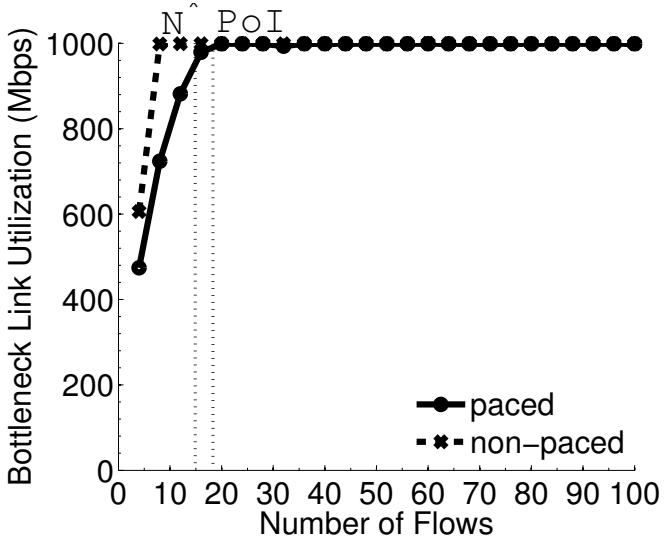
Multiple flows: Link Utilization/Drop/Latency

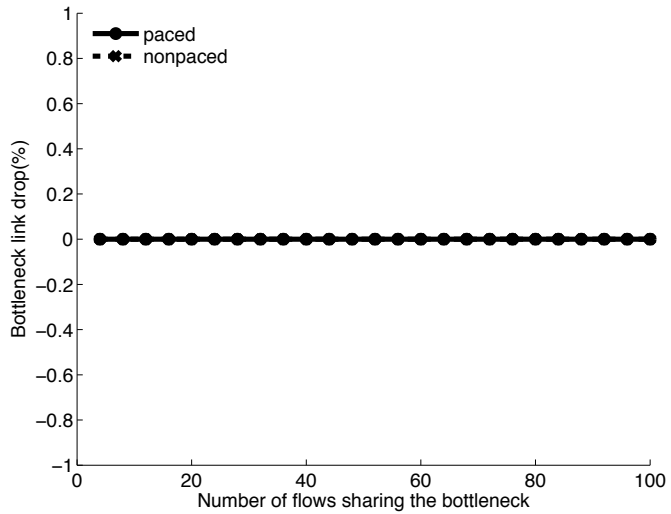
Buffer size 6.8% of BDP, varying number of flows



Multiple flows: Link Utilization/Drop/Latency

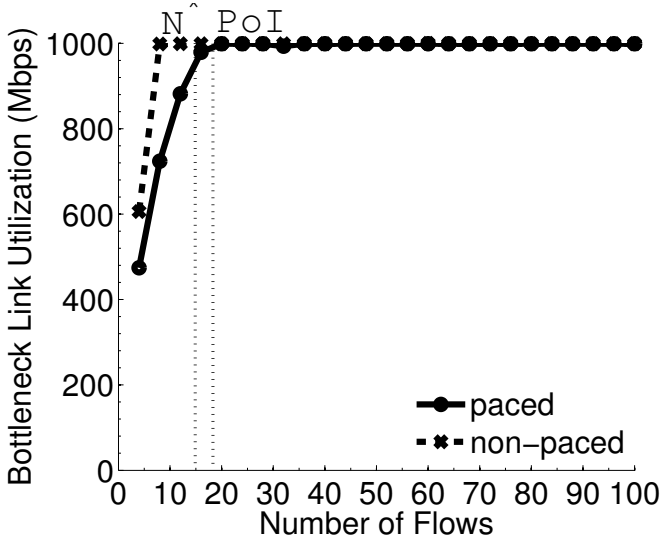
Buffer size 6.8% of BDP, varying number of flows

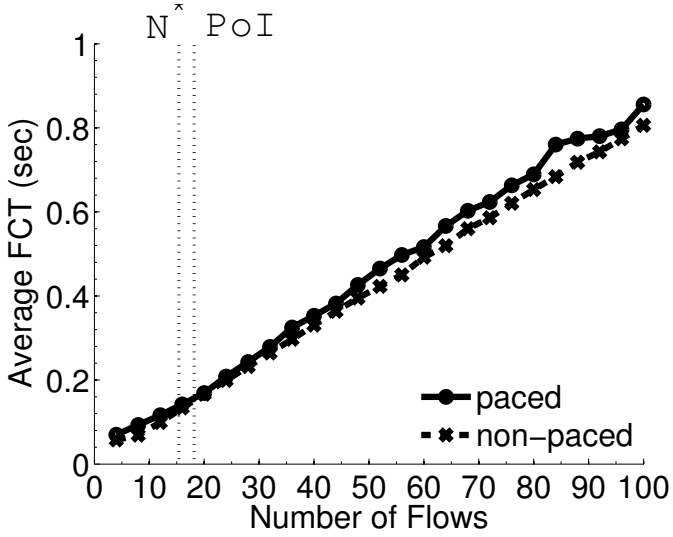


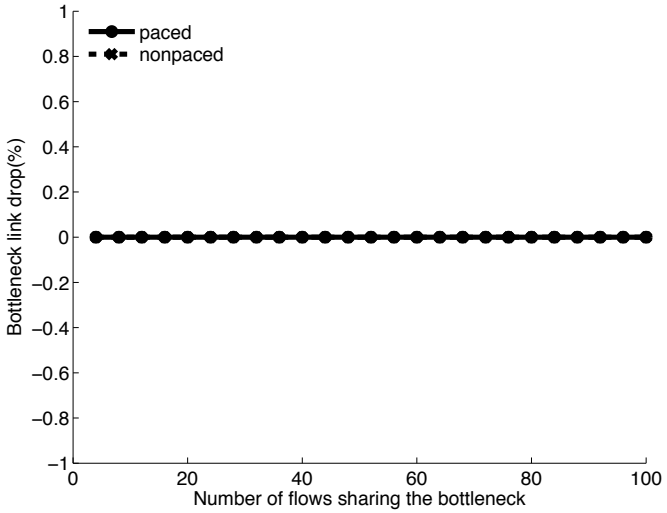


Multiple flows: Link Utilization/Drop/Latency

Buffer size 6.8% of BDP, varying number of flows







Multiple flows: Link Utilization/Drop/Latency

Buffer size 6.8% of BDP, varying number of flows