Improving ECN Marking Scheme with Micro-burst Traffic in Data Center Networks

Danfeng Shan, Fengyuan Ren

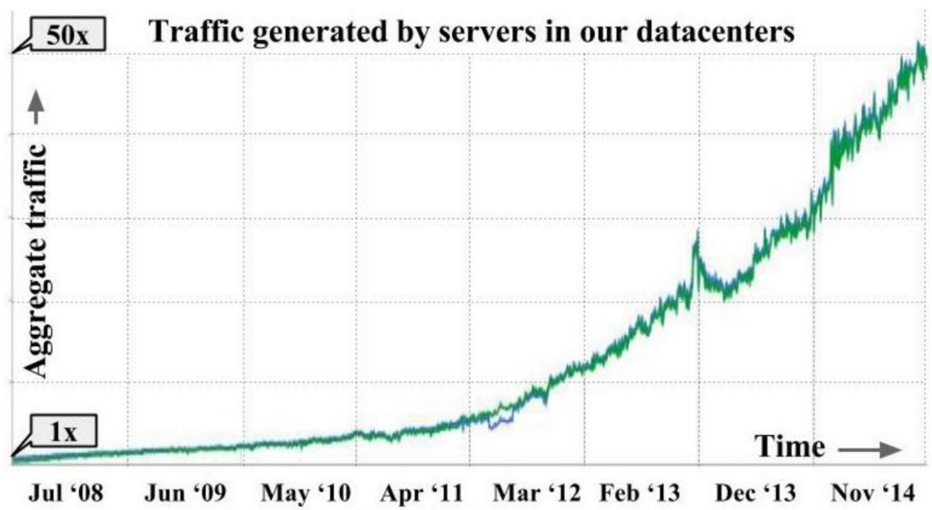
IEEE INFOCOM 2017

Outline

- Background & Motivation
- Analysis
- Solution
- Evaluation
- Conclusion

Data Center Networks

- **Intra DC**
  - Distributed applications
    - High throughput & Low latency
  - Growing traffic

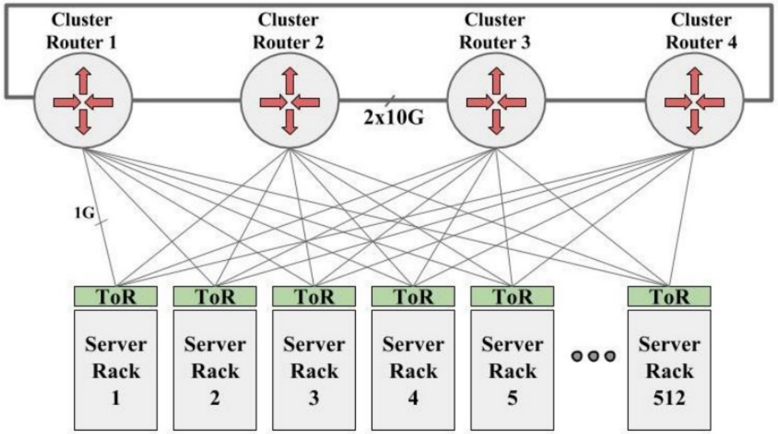


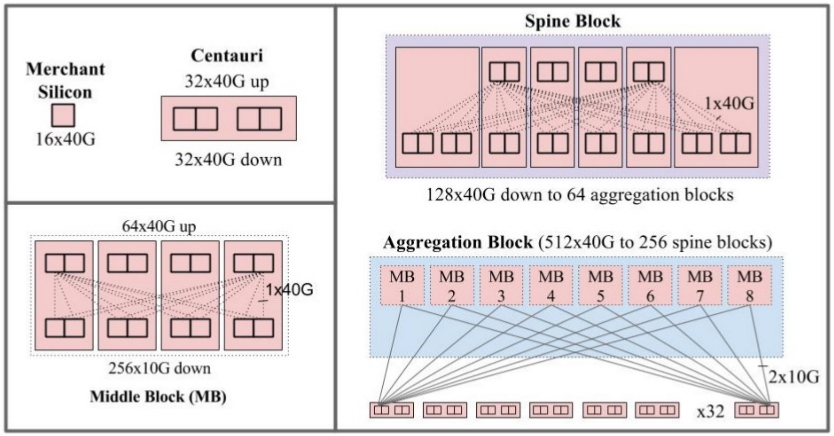
Google, SIGCOMM15

Data Center Networks

- DCN architecture

Google, SIGCOMM15





1/10Gbps Network (2004)

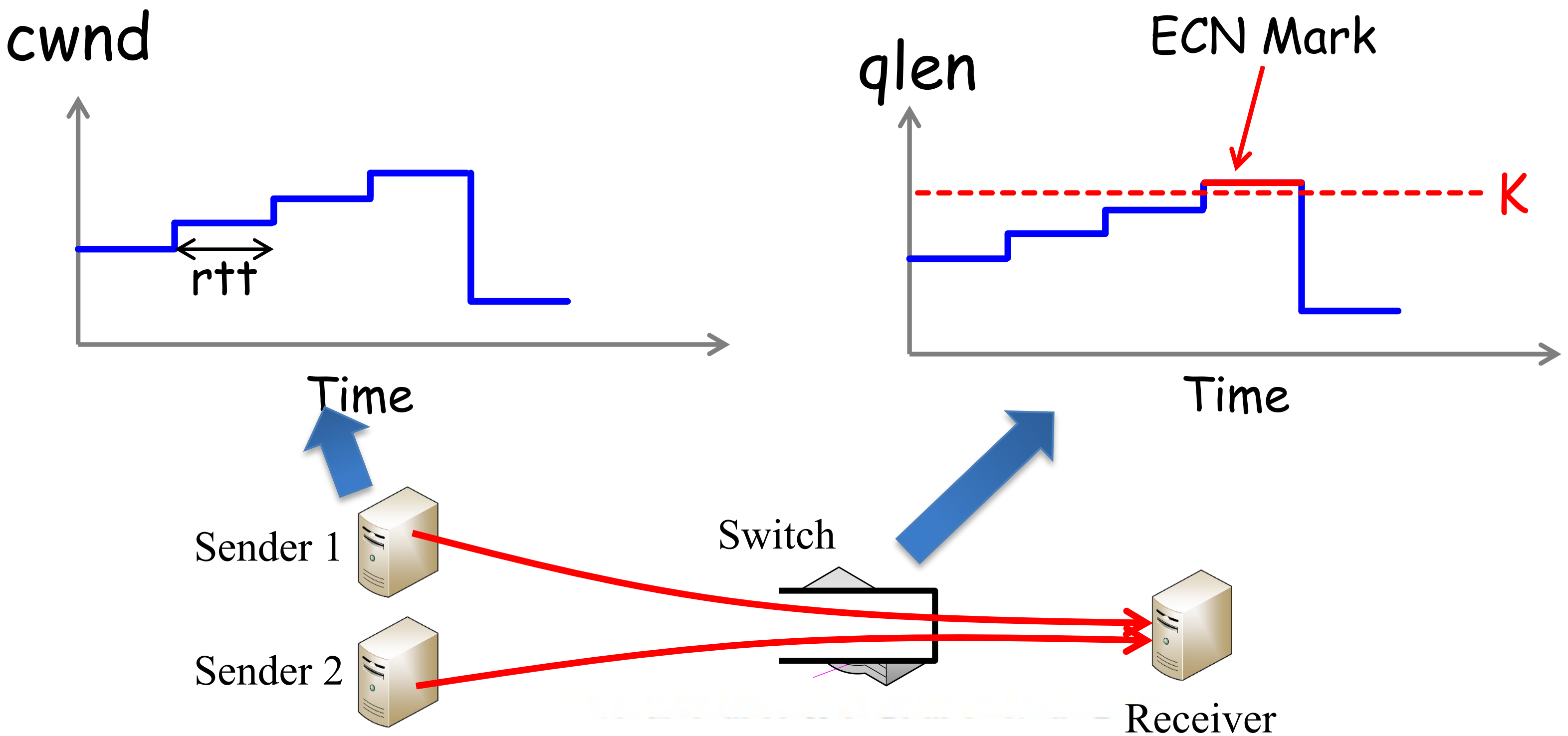
10/40Gbps Network (2015)

Data Center Networks

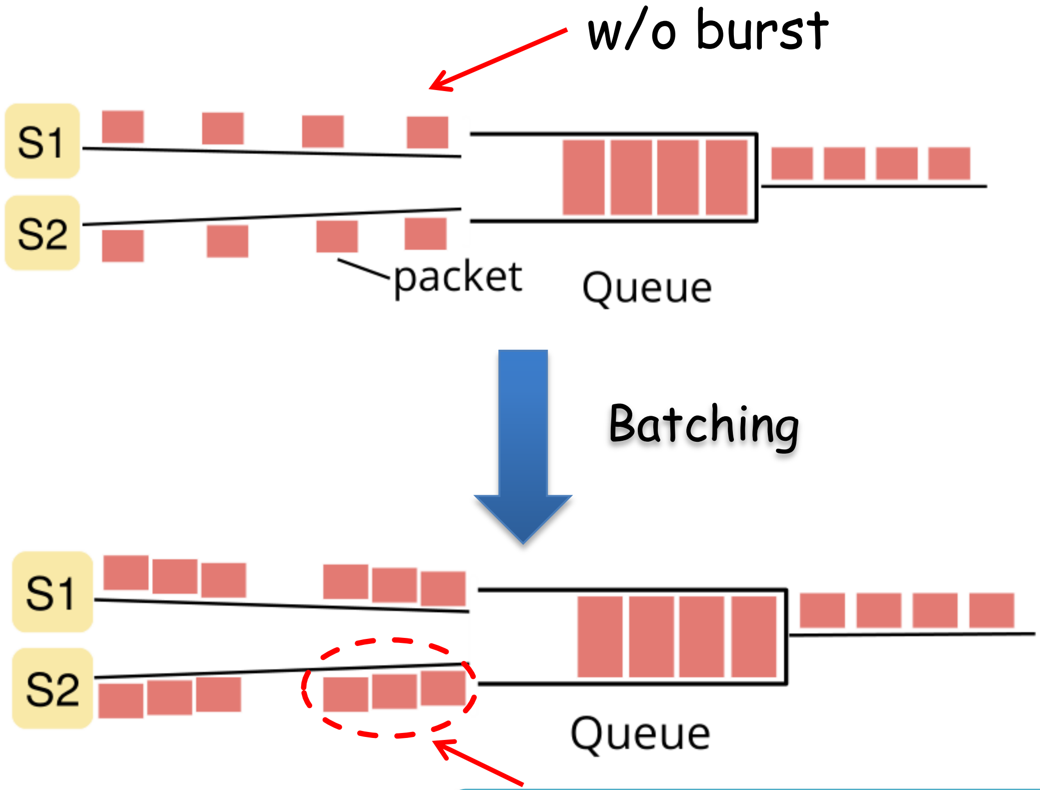
- **Reducing CPU overhead: batching**
  - Large Segment Offload: TSO, GSO
  - Receive Side Offload: RSC, LRO, GRO
  - Interrupt Coalescing (IC)
  - Jumbo Frame
  - …

ECN Marking in DCN

- ECN marking
  - DCTCP, ECN*, DCQCN, ……
  - Single ECN threshold, Instant queue length
    - If Qlen > K, mark packets with ECN
    - Senders slow down according to ECN feedbacks

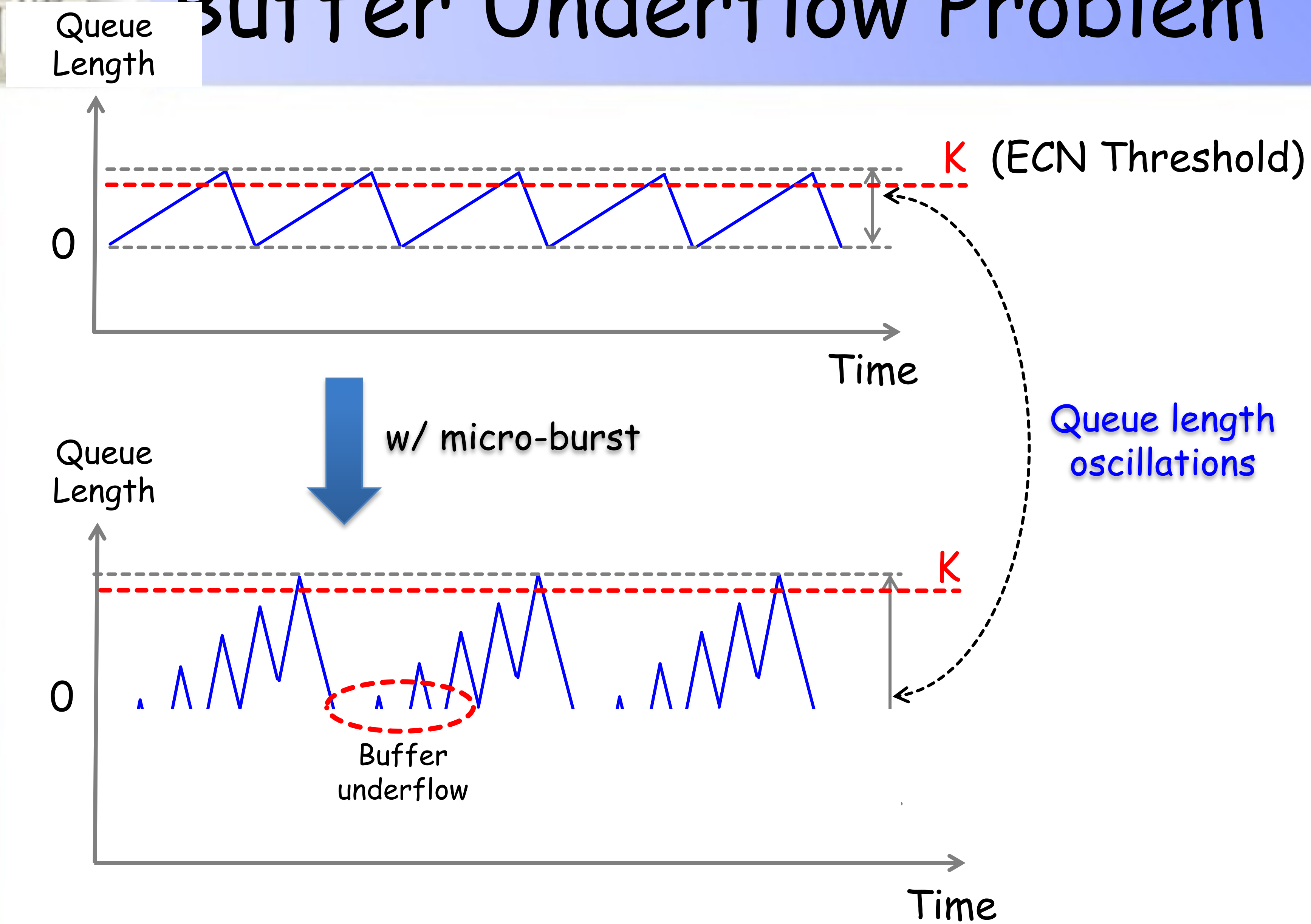


Buffer Underflow Problem

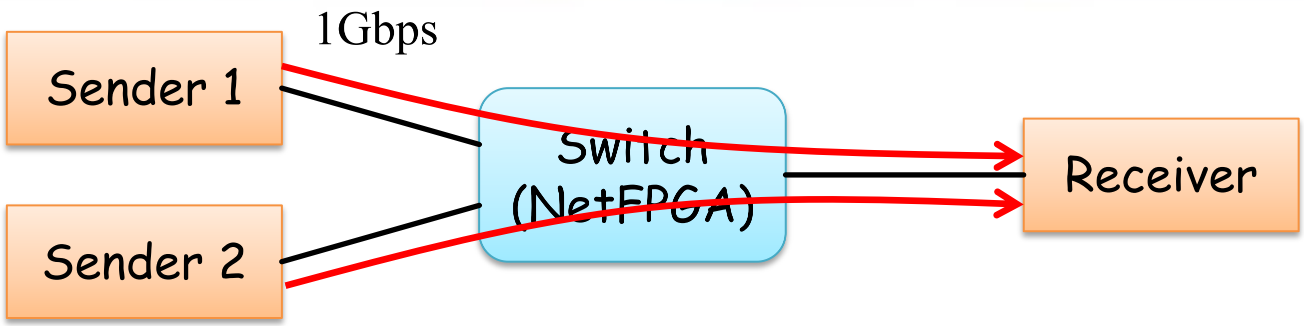


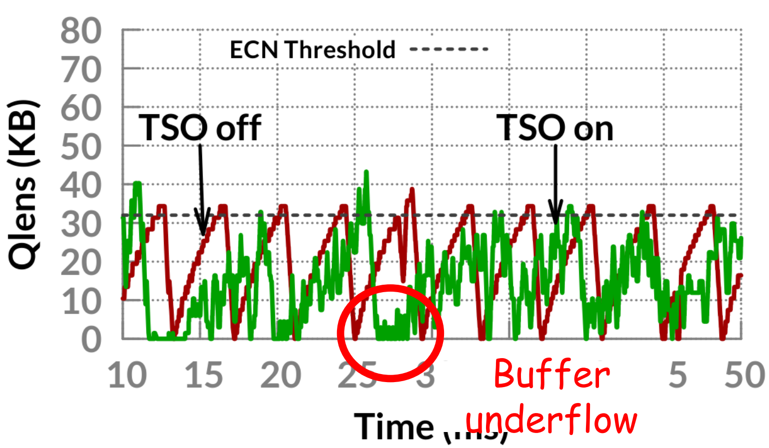
Micro-burst:
traffic burst in sub-RTT level

Buffer Underflow Problem

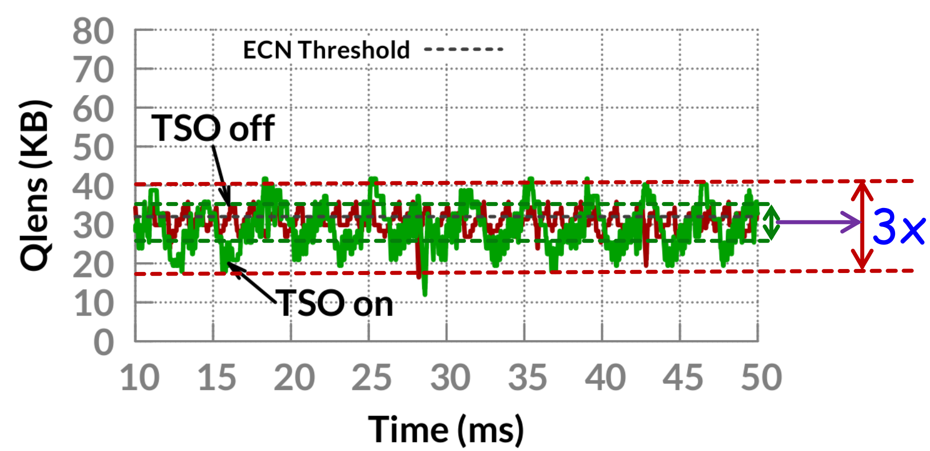


Buffer Underflow Problem



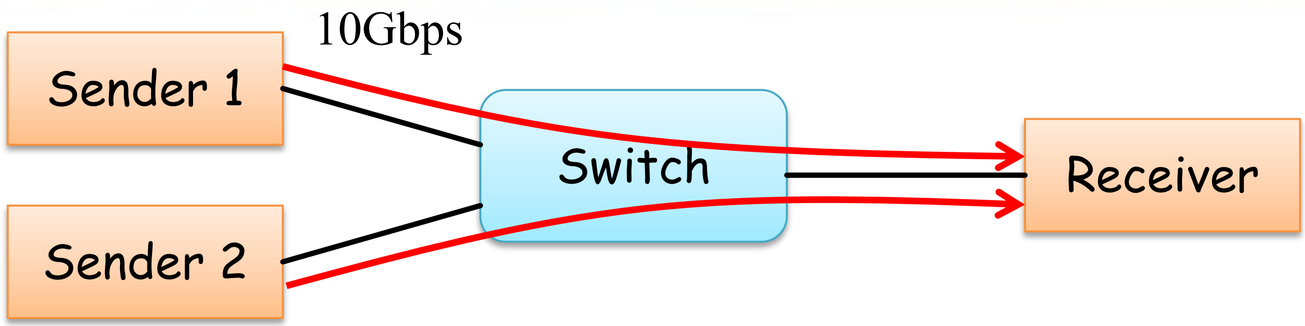


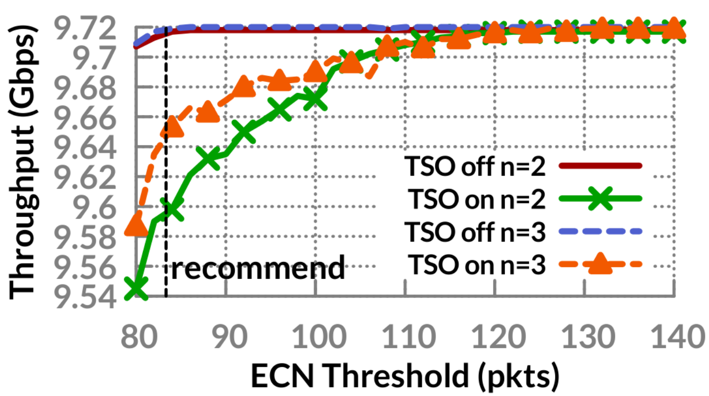
ECN*



DCTCP

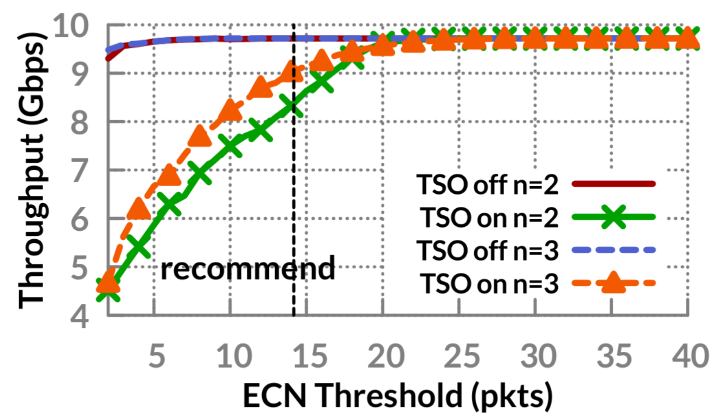
Buffer Underflow Problem





ECN*

127Mbps throughput loss



DCTCP

1.391Gbps throughput loss

Brief Summary

Batching → Micro-burst → Severe queue length oscillations → Buffer underflow → Throughput loss

The amplitude of queue length oscillations?

How to set ECN threshold with micro-burst traffic?

Outline

- Background & Motivation
- Analysis
- Solution
- Evaluation
- Conclusion

Analysis

- **Amplitude of queue length oscillations (DCTCP)**

N: # of flows
C: link capacity
d: Base RTT
K: ECN threshold
A: Amplitude of oscillation

w/o micro-burst

$$A = \frac{1}{2}\sqrt{2N(Cd + K)}$$

BDP: Bandwidth Delay Product

w/ TSO-induced micro-burst

$$\frac{\beta(N-1)}{N-\beta}(Cd+K) \le A \le \frac{\beta(N-1)}{N(1-\beta)}(Cd+K) + N\sqrt{\frac{(1-\beta)(Cd+K)}{2}}$$

w/o micro-burst traffic
$O(\sqrt{BDP})$

→

w/ micro-burst traffic
$O(BDP)$

Analysis

N: # of flows
C: link capacity
d: Base RTT
K: ECN threshold

- **ECN threshold settings to achieve 100% throughput (DCTCP)**

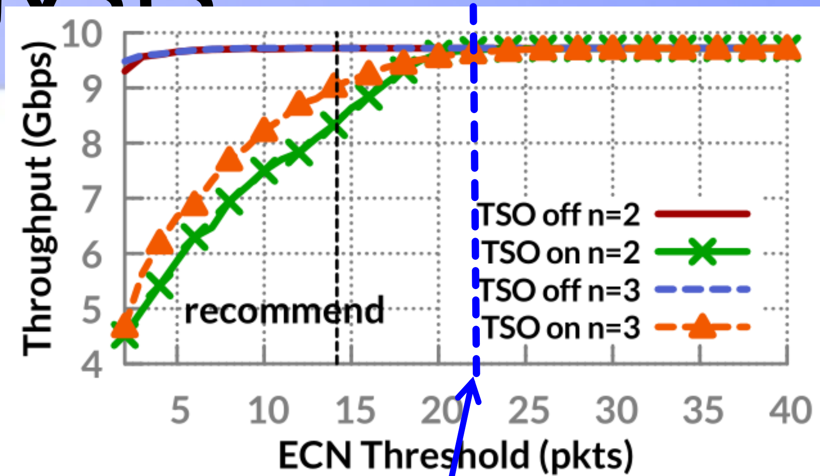
w/o micro-burst

$$K \geq 0.17Cd$$

w/ TSO-induced micro-burst

$$\begin{cases} K \geqslant \frac{\beta(N-1)}{N(1-\beta)}Cd + \frac{N(N-\beta)^2}{8(N-1)\beta} \\ \quad if \left[\frac{N(N-\beta)}{\beta(N-1)}\right]^2 \frac{1-\beta}{8} \leqslant Cd + K \\ K \geqslant \frac{N^2(1-\beta)+N\sqrt{N^2(1-\beta)^2+8Cd(1-\beta)}}{4} \\ \quad if \left[\frac{N(N-\beta)}{\beta(N-1)}\right]^2 \frac{1-\beta}{8} > Cd + K \end{cases}$$

Analysis



- **An example**
  - $N = 2, \beta = \frac{1}{3}$
  - $BDP = 83.3$ packets
    - RTT=100us, C=10Gbps

w/o micro-burst	$K \geq 16.7$ packets
w/ TSO-induced micro-burst	$K \geq 22.9$ packets

With TSO, the ECN threshold
should be 61.6% larger than that without batching.

Outline

- Background & Motivation
- Analysis
- Solution
- Evaluation
- Conclusion

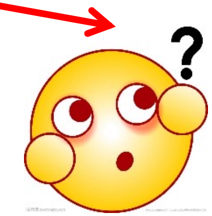
What Can We Do?

Why the problem occurs?

- Bursty traffic
  - Eliminate bursty traffic
    - TCP pacing
    - Pacing at NIC

  Not enough:
  TSO

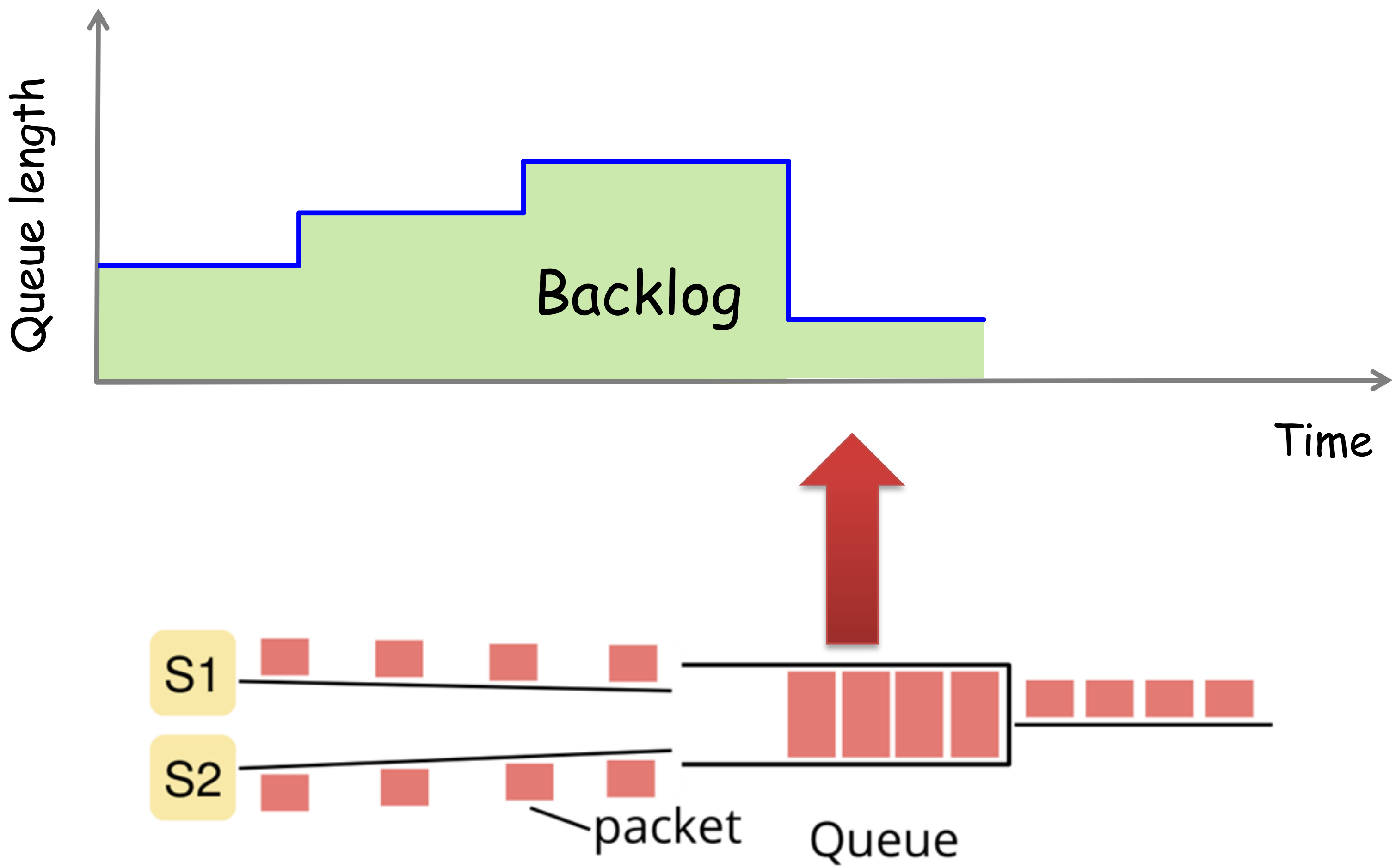
  Hard

- ECN mis-marking
  - Improve ECN marking scheme
    - Increasing ECN threshold
    - Average queue length
    - 

- Increase queueing latency
- Hard to determine ECN threshold

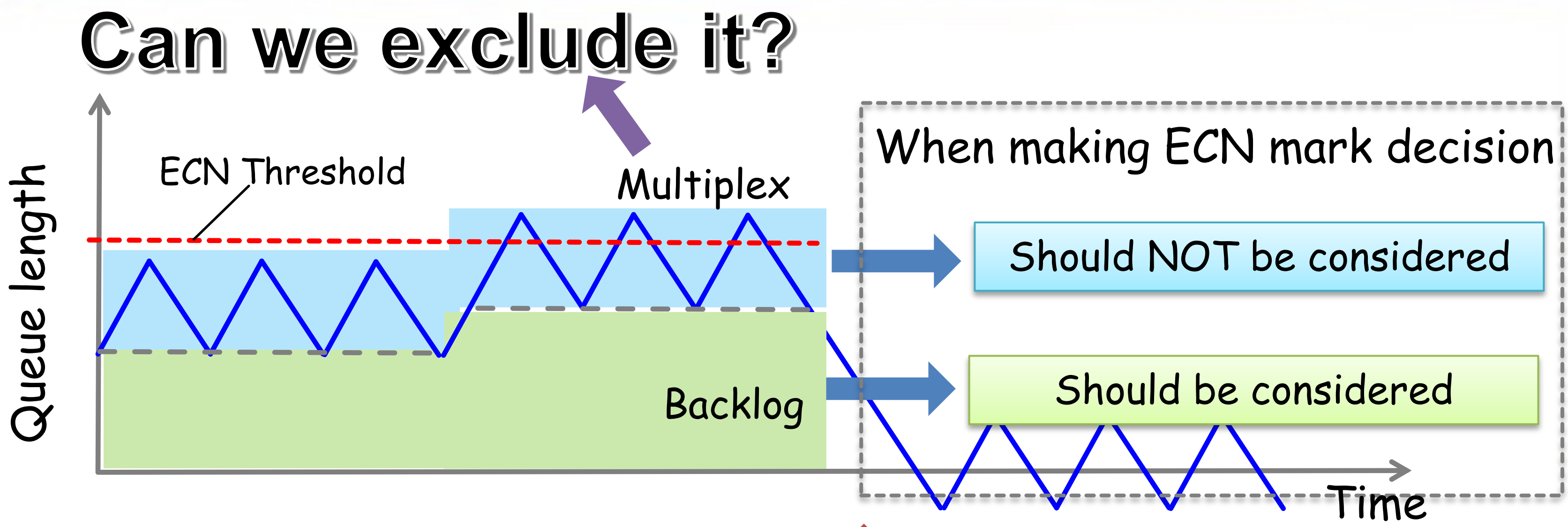
Reduce responsiveness

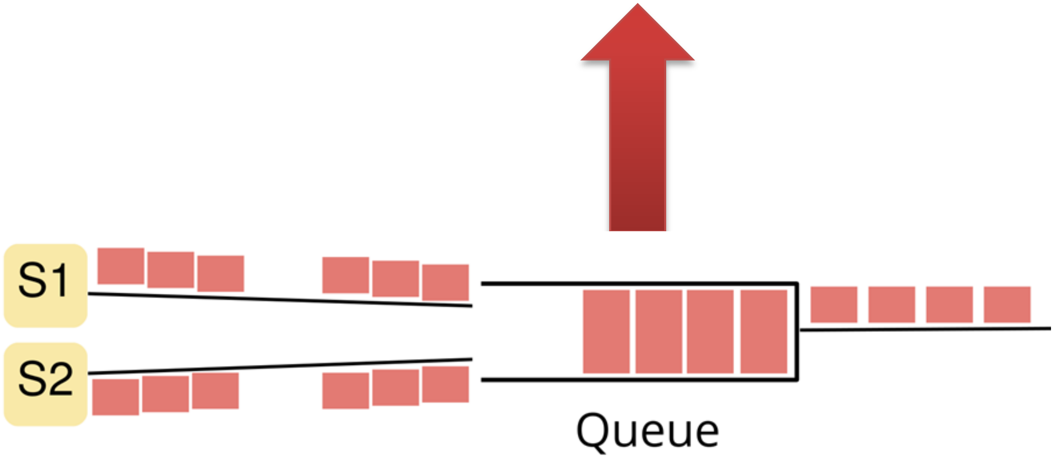
Queue Composition



Uniform traffic

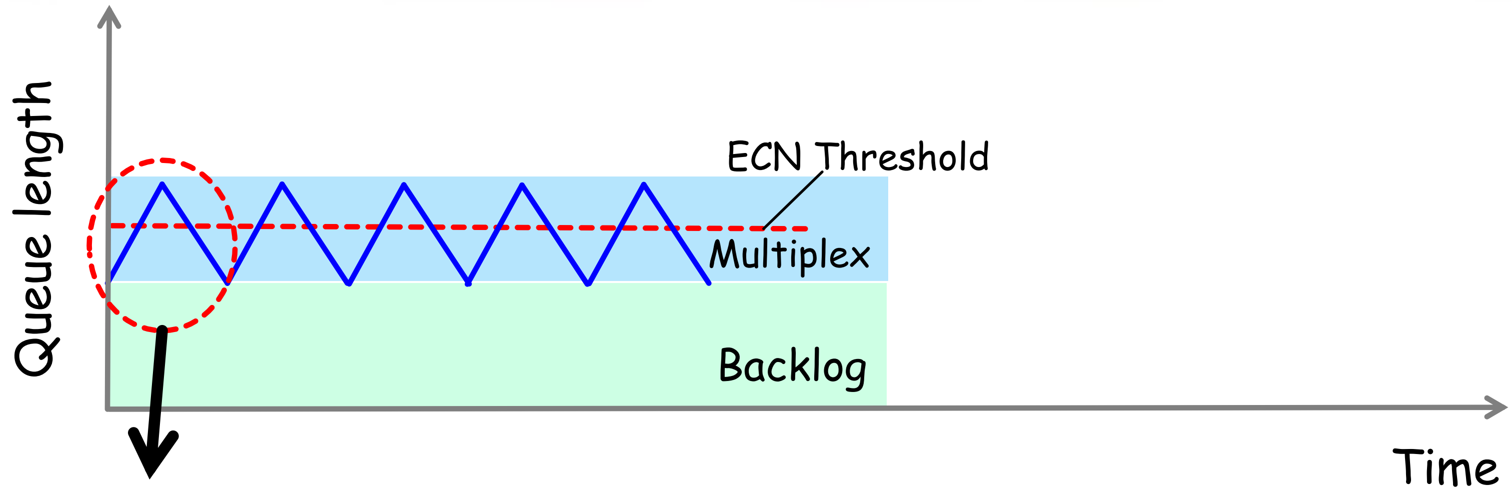
Queue Composition



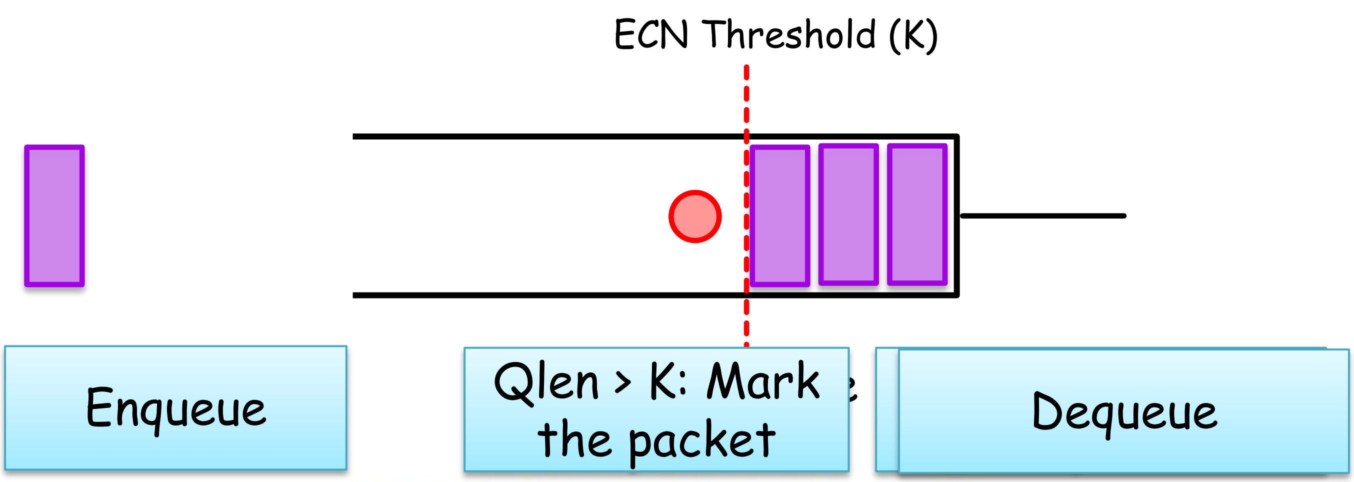


Bursty traffic

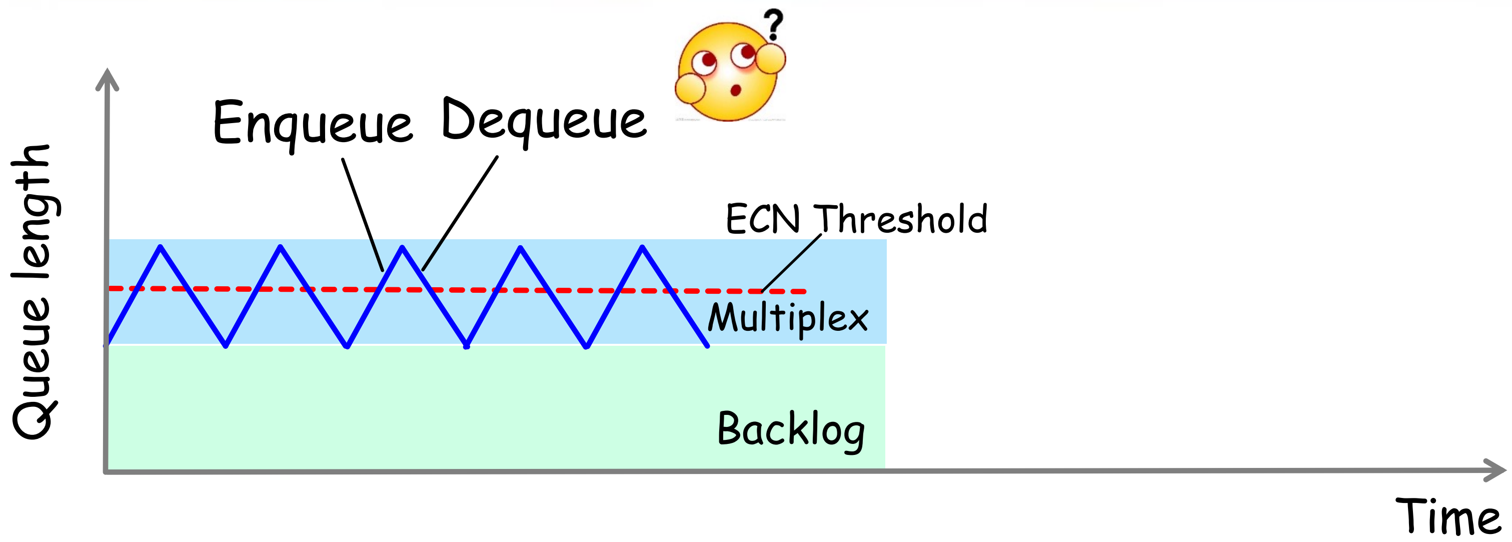
Basic Idea



Queue length increasing is transient



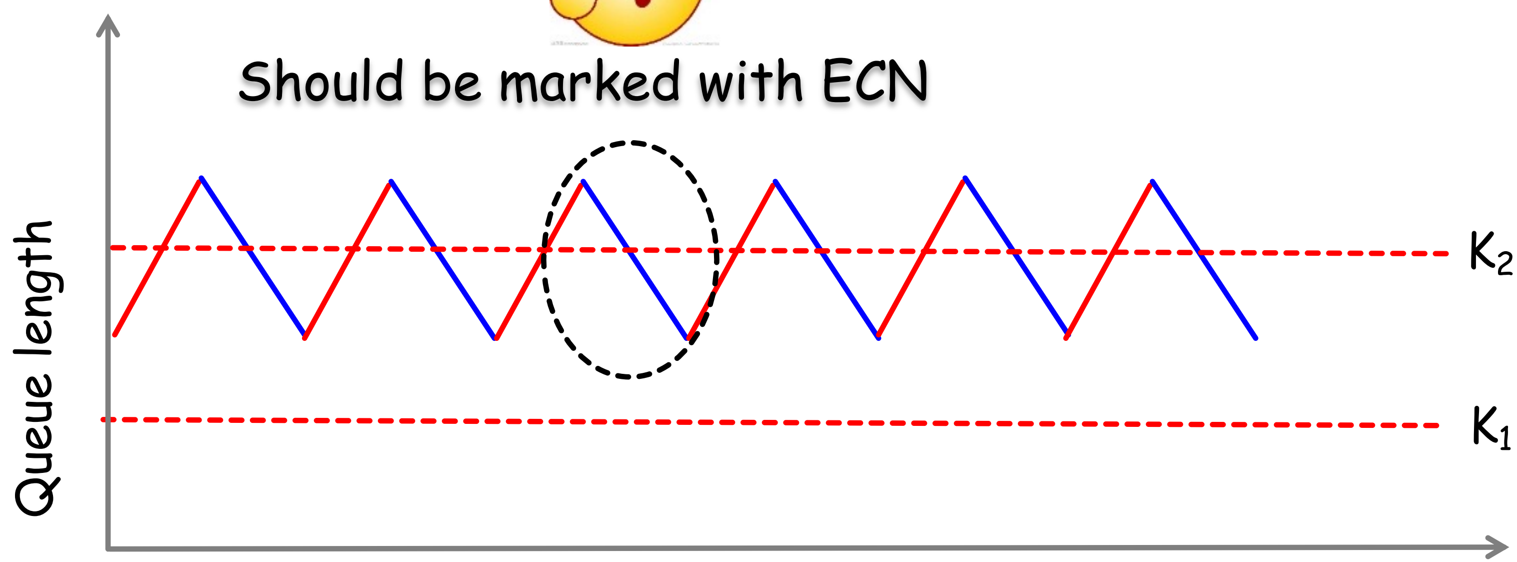
Basic Idea



The queue length is decreasing: slope < 0

Our Solution: CEDM

Should be marked with ECN



If qlen > K_2, Mark packets anyway

Our Solution: CEDM

- Combined Enqueue and Dequeue ECN Marking

Packet enqueue

Mark packet if
qlen > K_2

Or else:
Mark packet if
1. qlen >= K_1 *and*
2. slope >= 0

Packet dequeue

Unmark packet if
1. qlen < K_1 *or*
2. qlen > K_2 and slope < 0

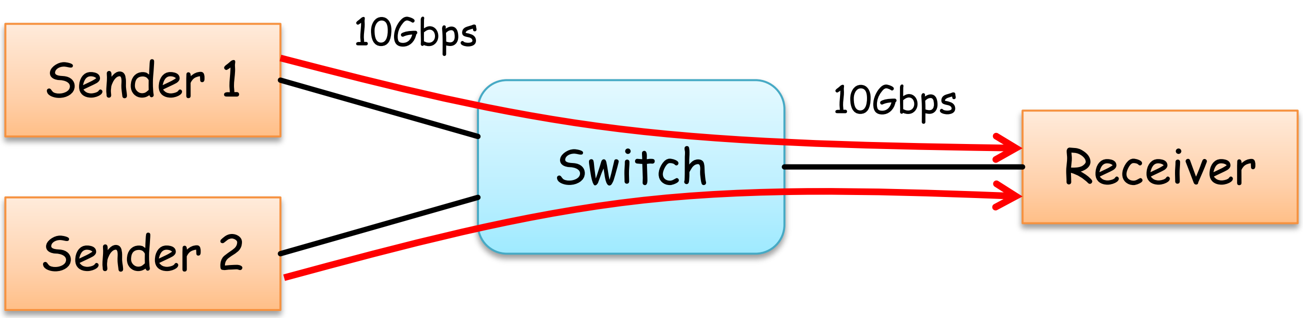
qlen: queue length
slope: derivative of queue length
K: ECN threshold

Outline

- Background & Motivation
- Analysis
- Solution
- Evaluation
- Conclusion

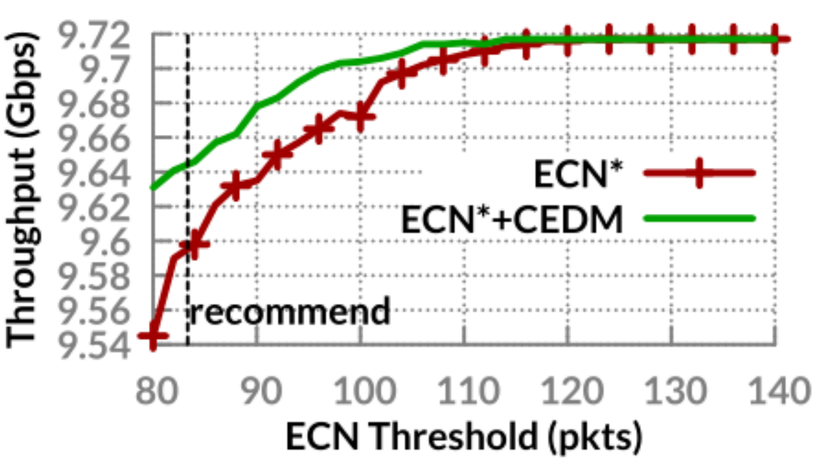
Evaluation

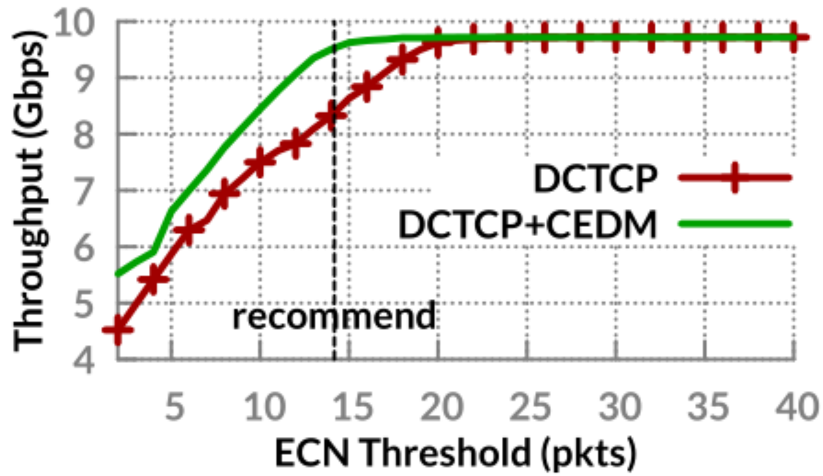
- Throughput



Evaluation

- Throughput





Throughput loss is reduced by 1.6X

Throughput loss is reduced by 6X

Evaluation

- **Incast performance**

DCTCP: K=19 packets

DCTCP+CEDM:
K=14 packets ($0.17 \times C \times RTT$)

Buffer size: 150KB
Link rate: 10Gbps

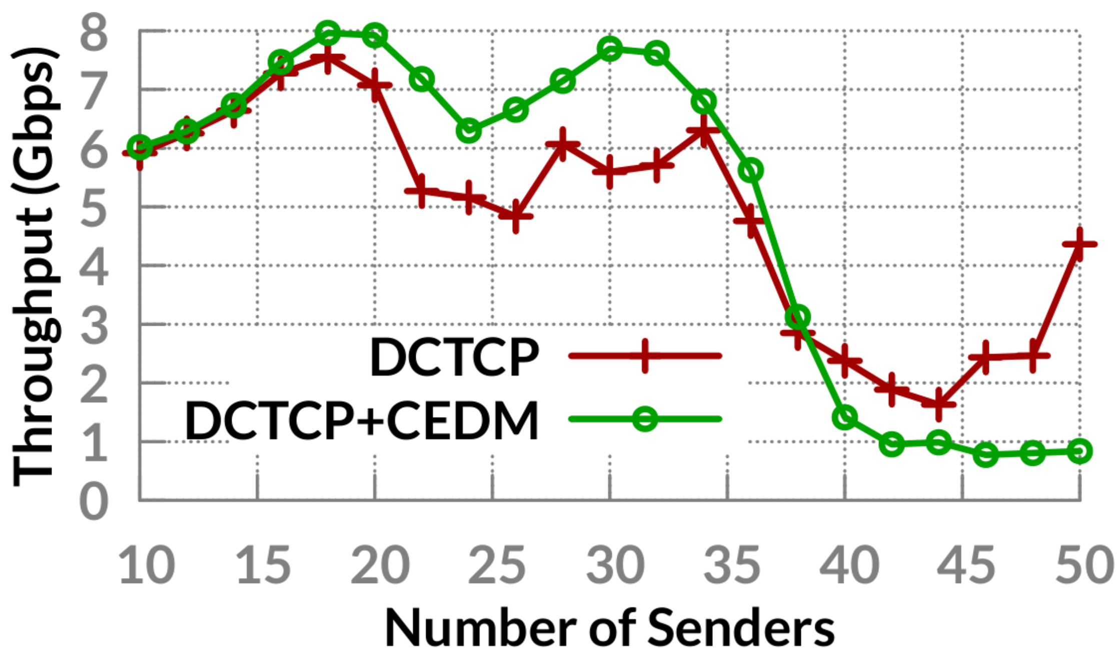


Fig. 11. Incast performance

Evaluation

- **Large-scale simulations**
  - Leaf-spine topology with 144 servers
  - 10/40Gbps network

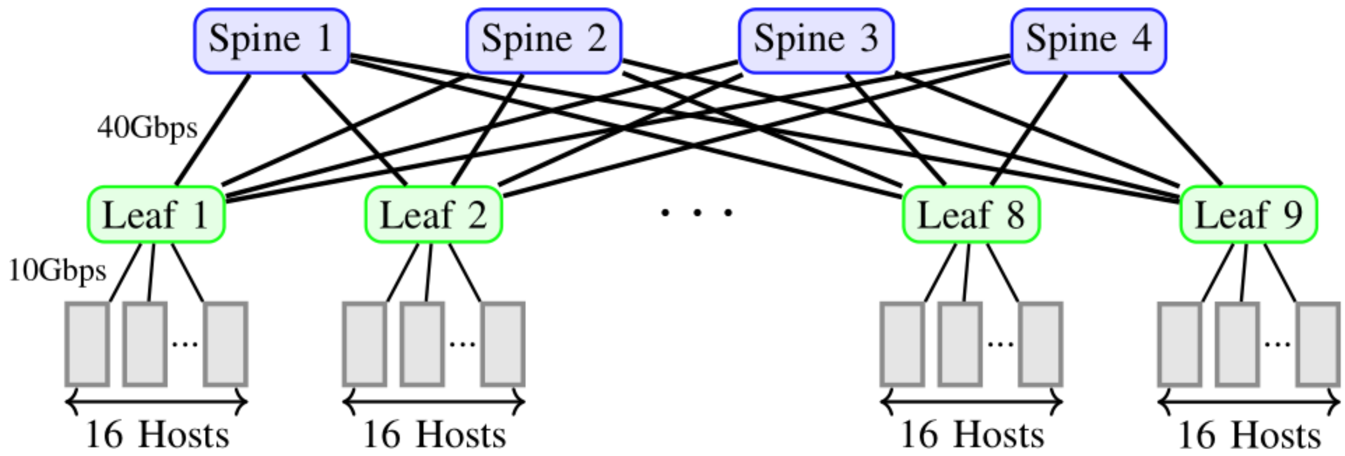
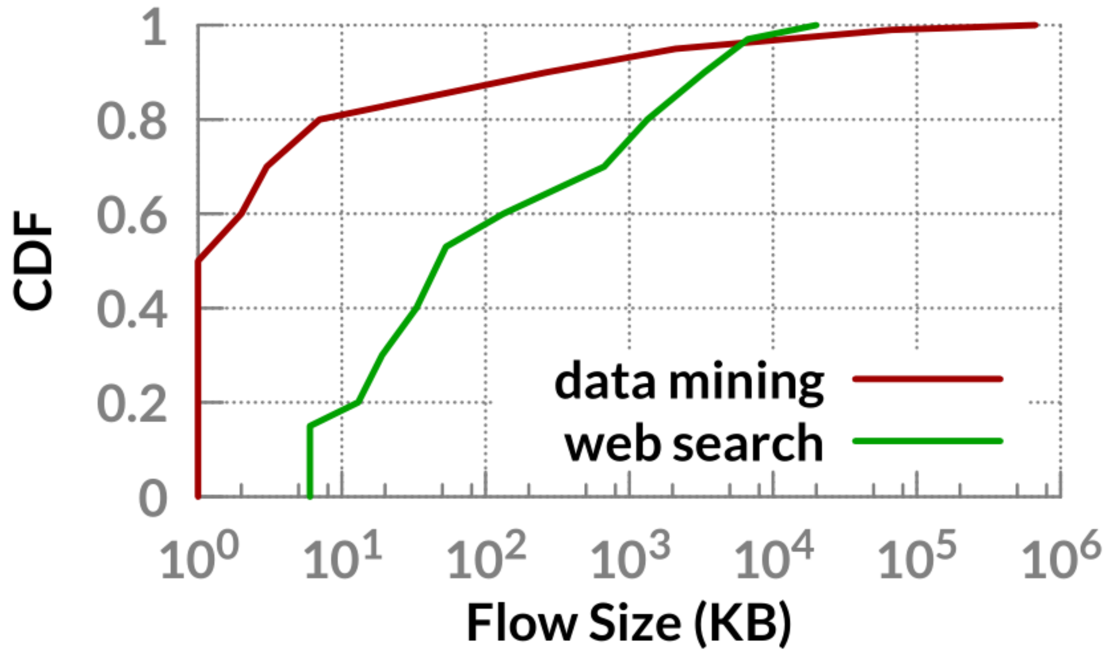


Fig. 14. Leaf-spine topology in large-scale simulations

Evaluation

- **Large-scale simulations**
  - Two widely used flow size distributions
    - Data mining workload
    - Web search workload



Evaluation

- **Large-scale simulations**

Recommended settings in 10Gbps network [1]

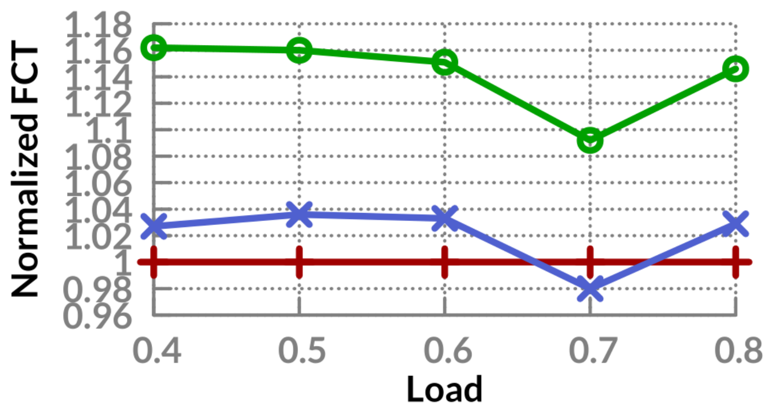
Protocols	ECN threshold (packet)
DCTCP	K=65
DCTCP+CEDM	K=12
DCTCP	K=12

Theoretical setting: $0.17 \times C \times RTT$ [2]

[1] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data Center TCP (DCTCP)," in SIGCOMM, 2010.
[2] M. Alizadeh, A. Javanmard, and B. Prabhakar, "Analysis of DCTCP: Stability, Convergence, and Fairness," in SIGMETRICS, 2011.

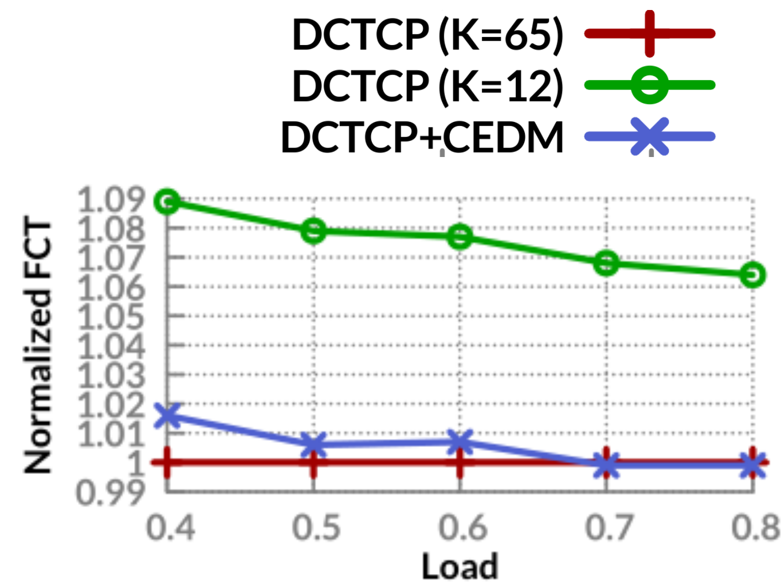
Evaluation

- **Large-scale simulations**
  - Large flows



(d) (10MB, ∞): Average

Data mining workload

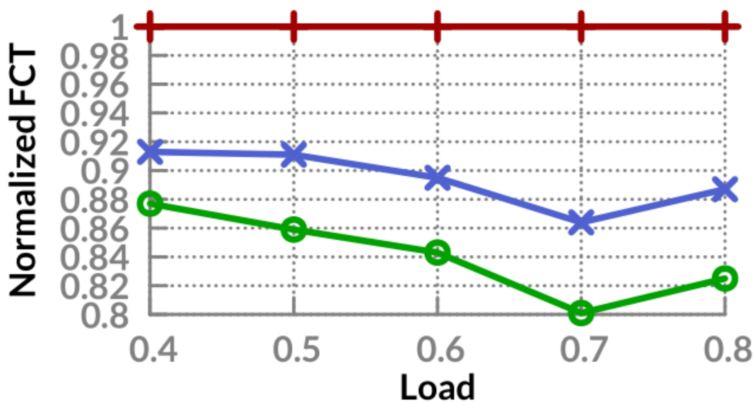


(d) (10MB, ∞): Average

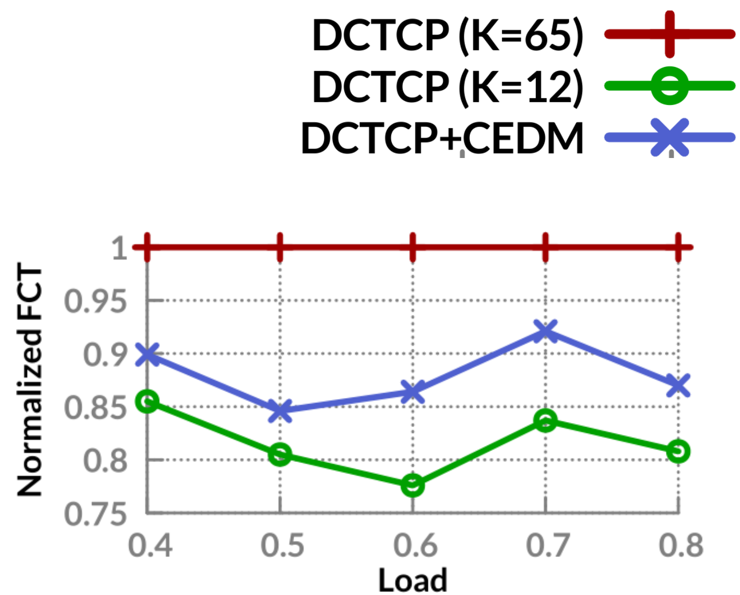
Web search workload

Evaluation

- **Large-scale simulations**
  - Small flows



(b) (0, 100KB]: Average



(b) (0, 100KB]: Average

Data mining workload

Web search workload

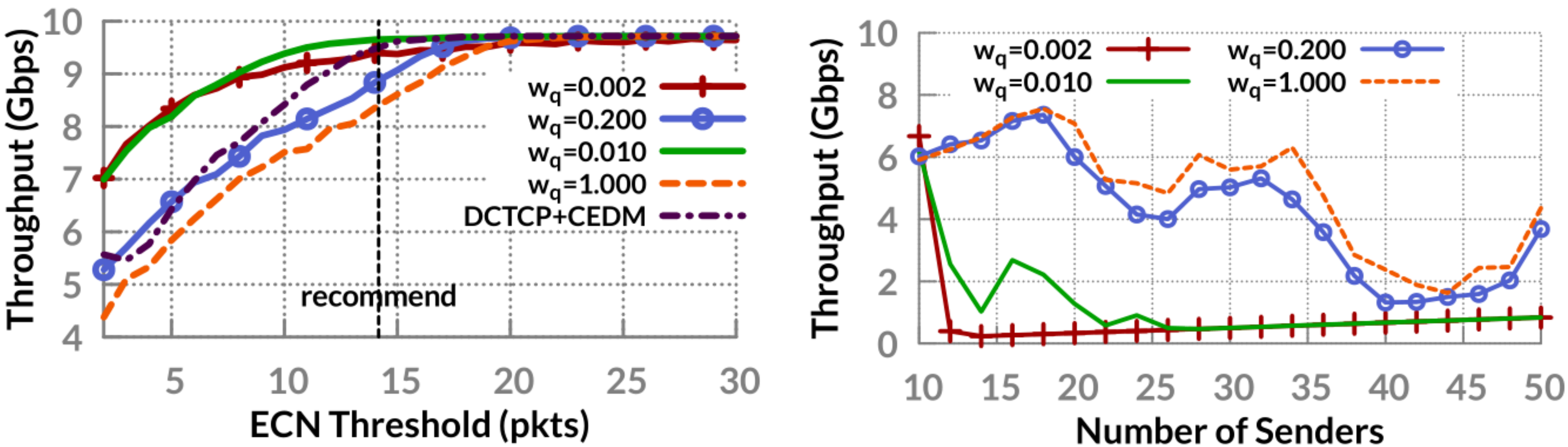
Conclusion

- **Reveal the buffer underflow problem caused by**
  - Instantaneous-queue-length-based ECN marking scheme
  - Batching-scheme-induced micro-burst traffic
- **Theoretically deduce the amplitude of queue length oscillations**
- **CEDM: a simple ECN marking scheme**
  - Exclude transient queue occupancy caused by multiplexing of micro-burst traffic
  - High throughput under low ECN threshold


Q
&
A

Backup Slide

- *Average queue length*

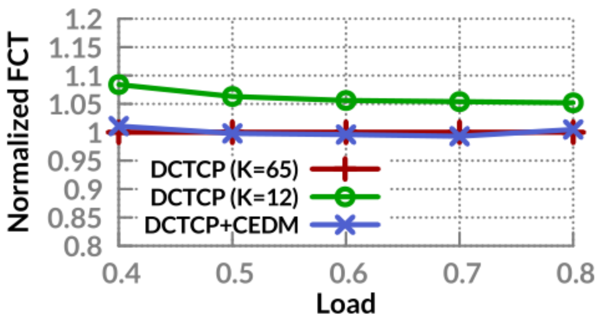


(a) overall throughput (b) incast performance

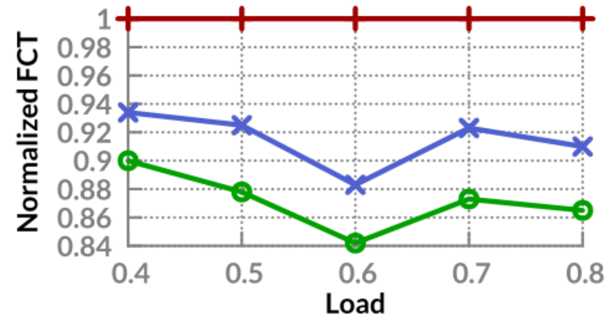
Fig. 12. Performance when ECN is marked according to average queue length

Backup Slide

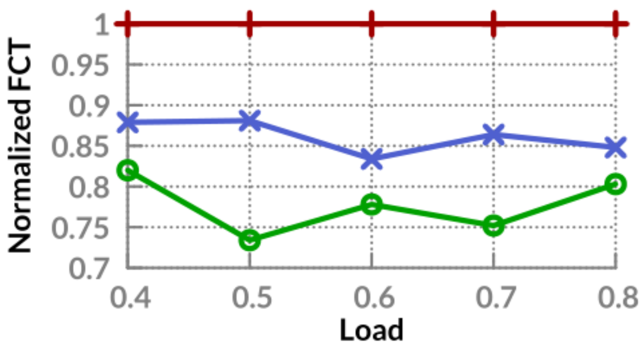
- 2:1 oversubscribed network



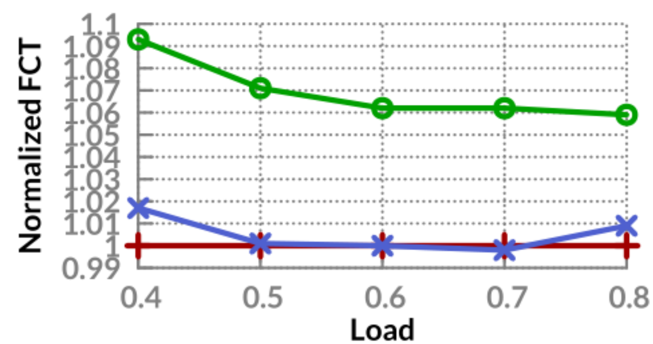
(a) Overall: Average



(b) (0, 100KB]: Average



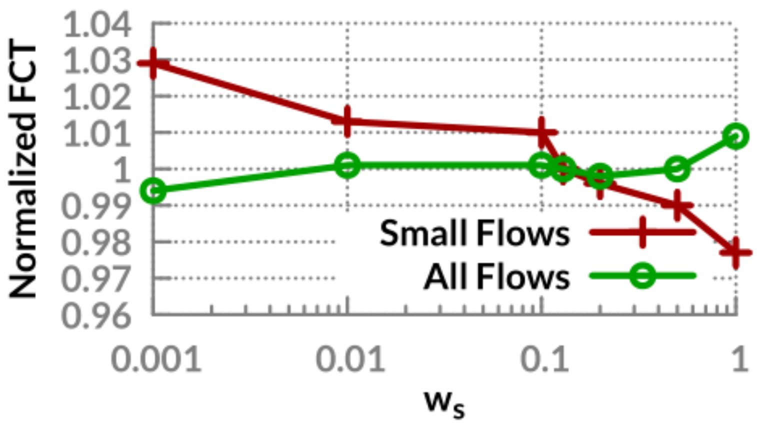
(c) (0, 100KB]: 99th percentile

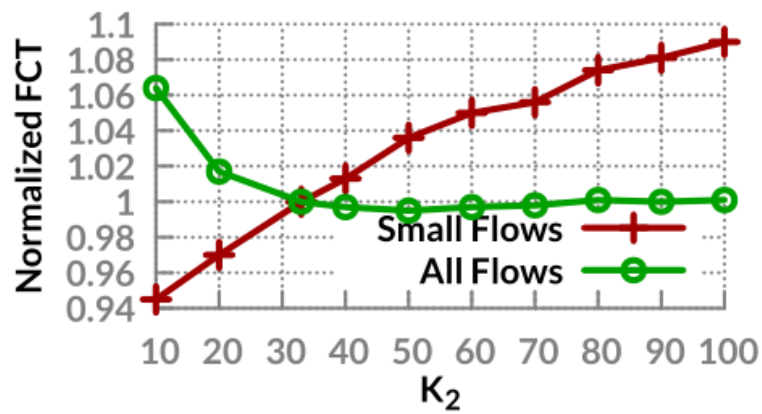


(d) (10MB, ∞): Average

Backup Slide

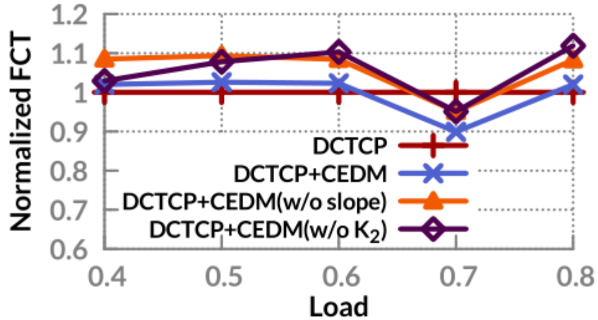
- Parameter sensitivity



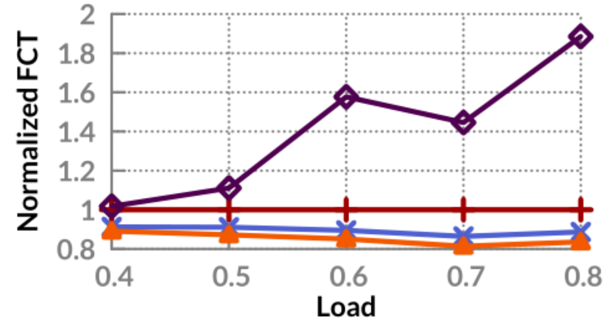


Backup Slide

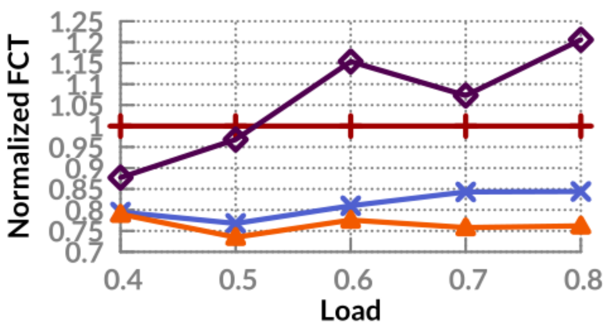
- Effective of slope and double threshold



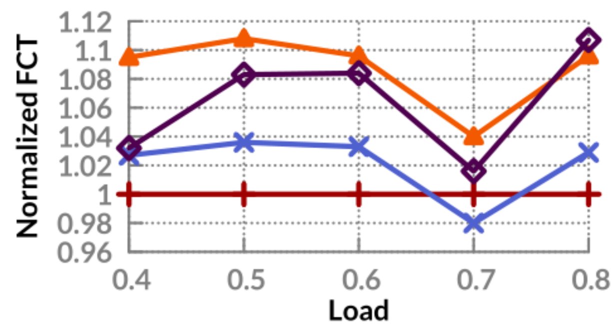
(a) Overall: Average



(b) (0, 100KB]: Average



(c) (0, 100KB]: 99th percentile



(d) $(10\text{MB}, \infty)$: Average