Backpressure Flow Control

Prateesh Goyal, Preey Shah, Naveen Sharma,
Kevin Zhao, Mohammad Alizadeh,
Tom Anderson

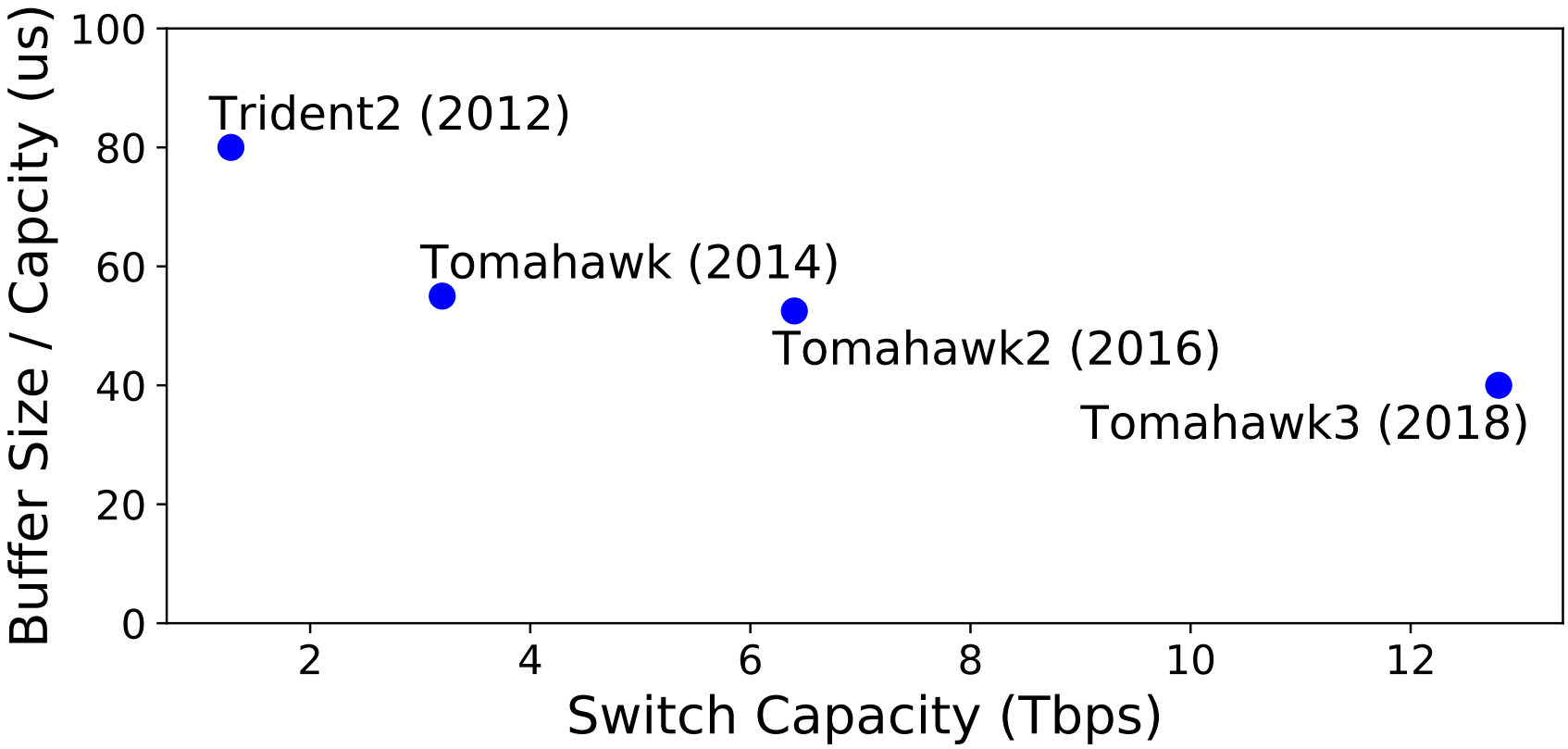
Two Types of Congestion Control

- End to End: action delayed by at least one RT
  - Sources send initial window
  - Adjust rate based on feedback
  - Complex control loop: topology and signalling
- Hop by Hop: short control loops
  - Sources send at line rate, pushback at switch
  - Per-flow state and head-of-line blocking
  - Widely used at server and rack level

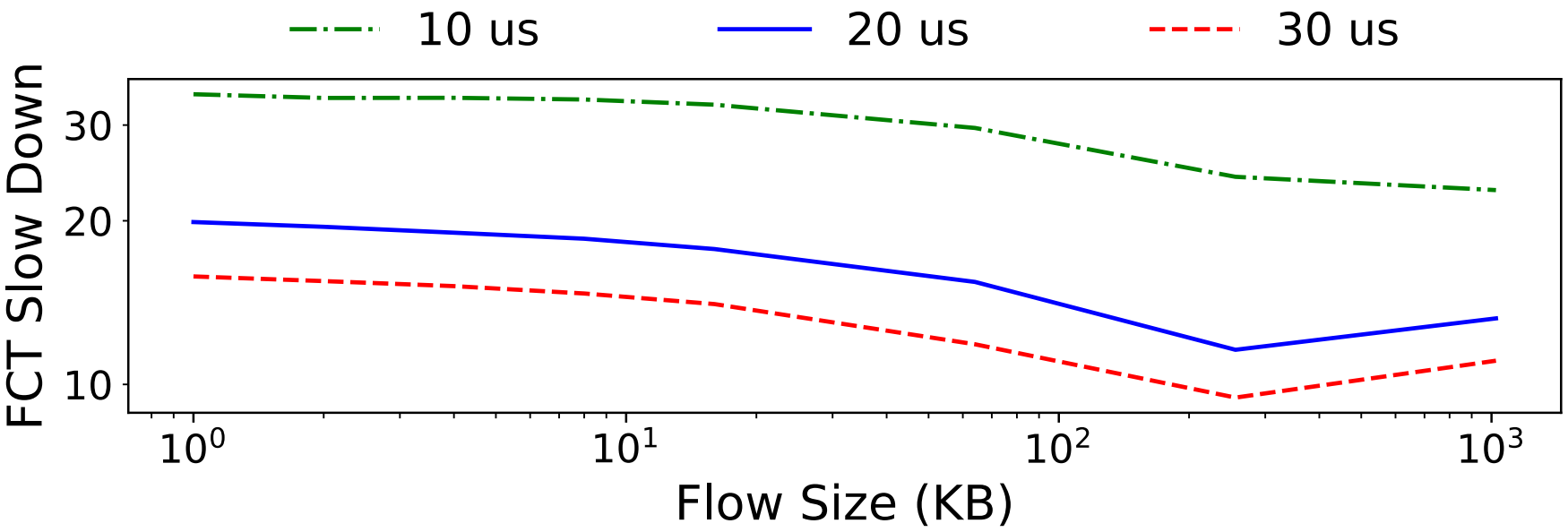
Why Now?

- Data center bandwidth increasing rapidly
  - Soon, most traffic will fit in a single round trip
- Latency and tail latency a dominant concern
  - Increasing percentage of RDMA
- Traffic patterns are highly bursty
  - Hard to control what isn't stable
- Network operational costs important
  - E2E: lower utilization for same tail latency

Switch Capacity Increasing

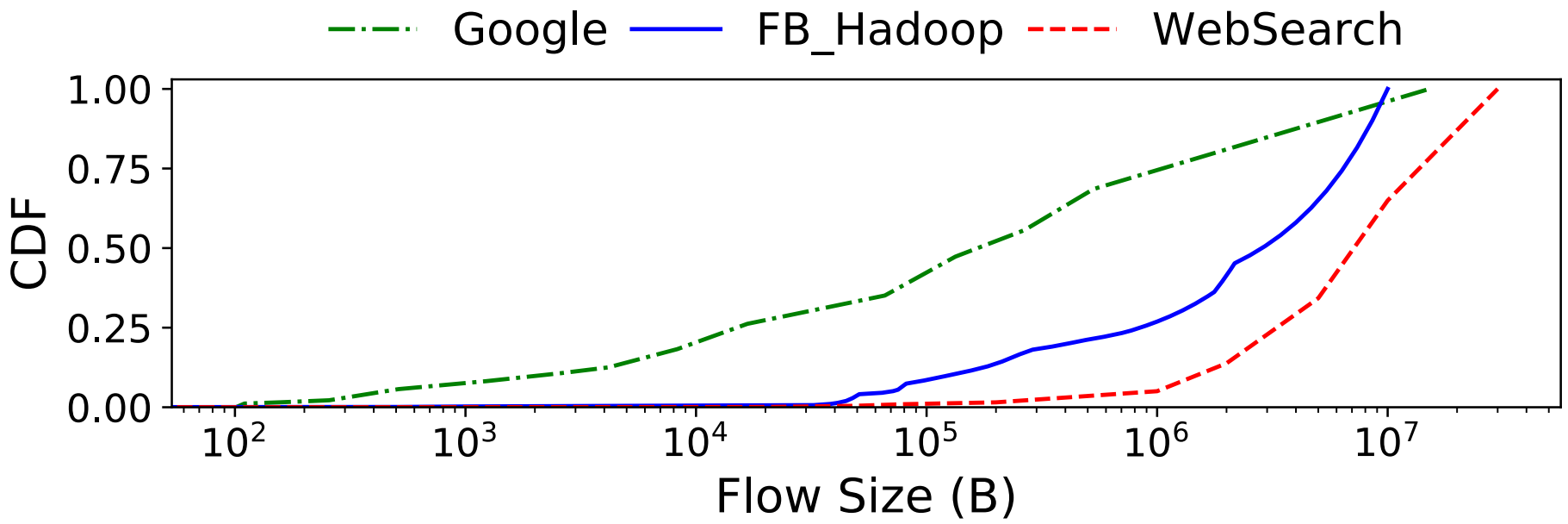


Buffering Matters to Tail Latency



DCQCN, 99% tail latency, Google workload, 75% util+incast

Elephants Are Mice



Weighted by flow size

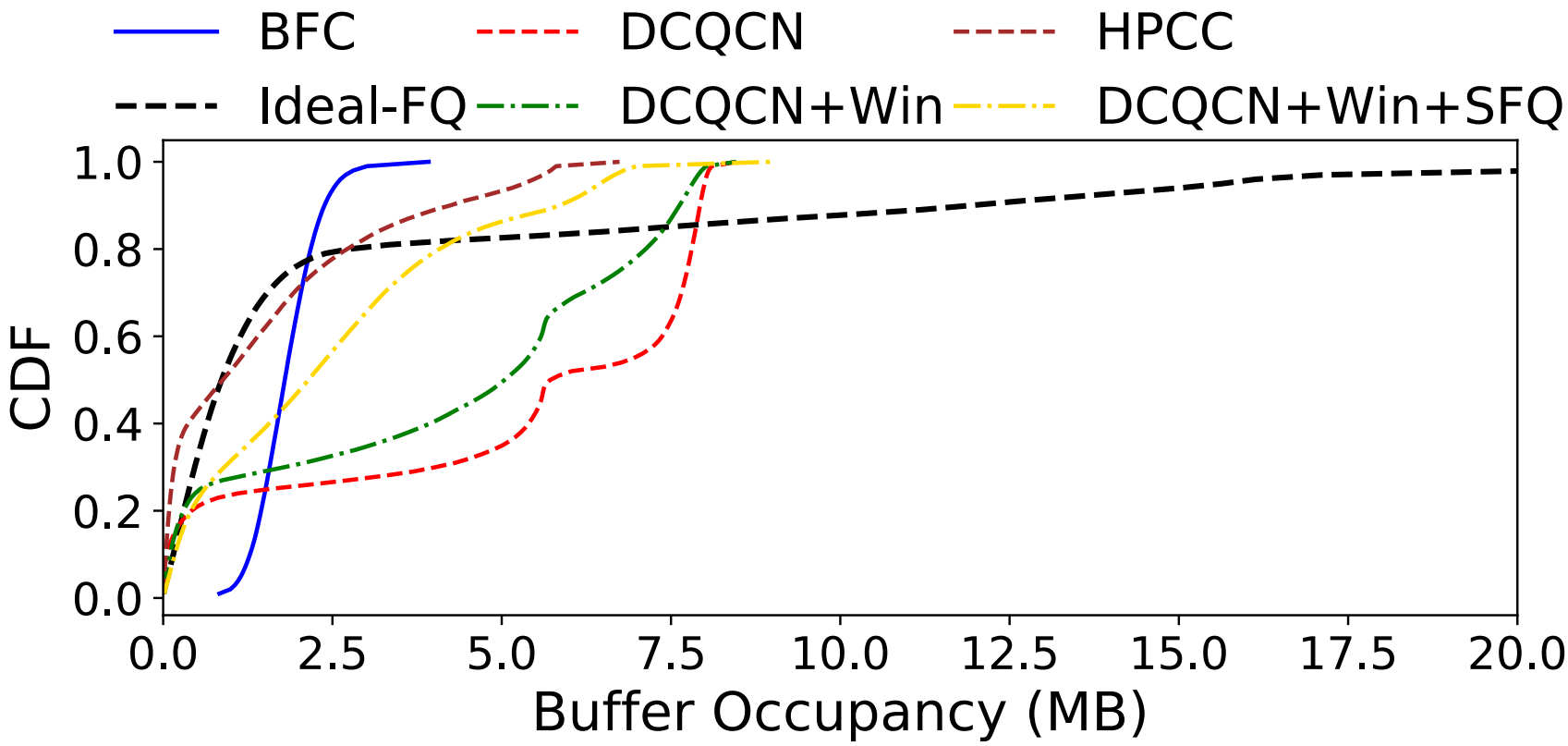
100 Gb

1 Tb

Backpressure Flow Control

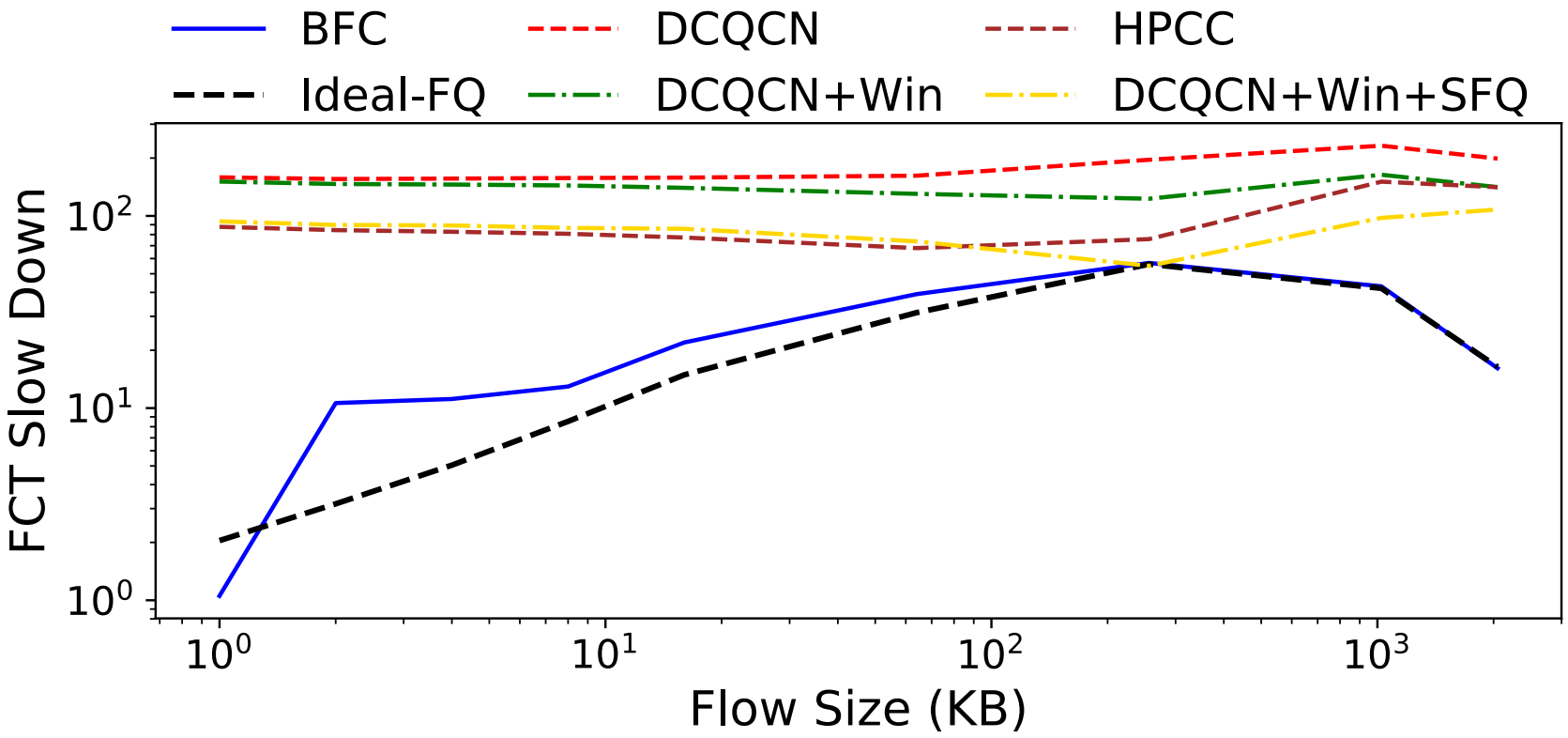
- Assumptions (Tofino like)
  - Limited number of egress queues (e.g., 32)
  - Queues can be paused/unpaused
  - Deficit RR among unpaused queues
- Dynamic assignment of flows to queues
- Per-hop pause frames, bloom filter for flows
  - Aggressive: push queueing upstream unless needed to keep egress busy
- Switch state α number of queued *flows*

Buffer Occupancy



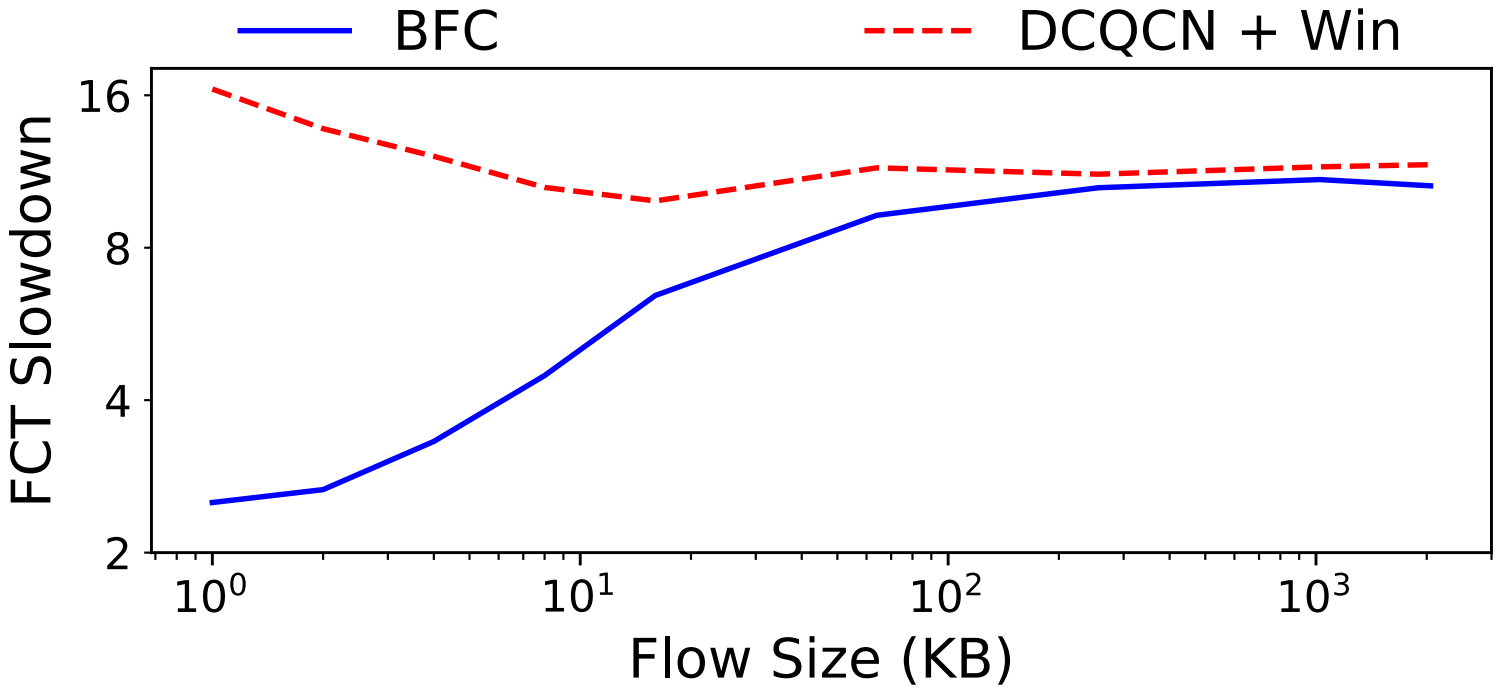
BFC buffers α number of queued *flows*

Tail Latency



99%, Google workload, 60% util + incast

Cross-Data Center Traffic



99% tail latency, intra-DC traffic in presence of cross-DC traffic

Network Cut Theorem

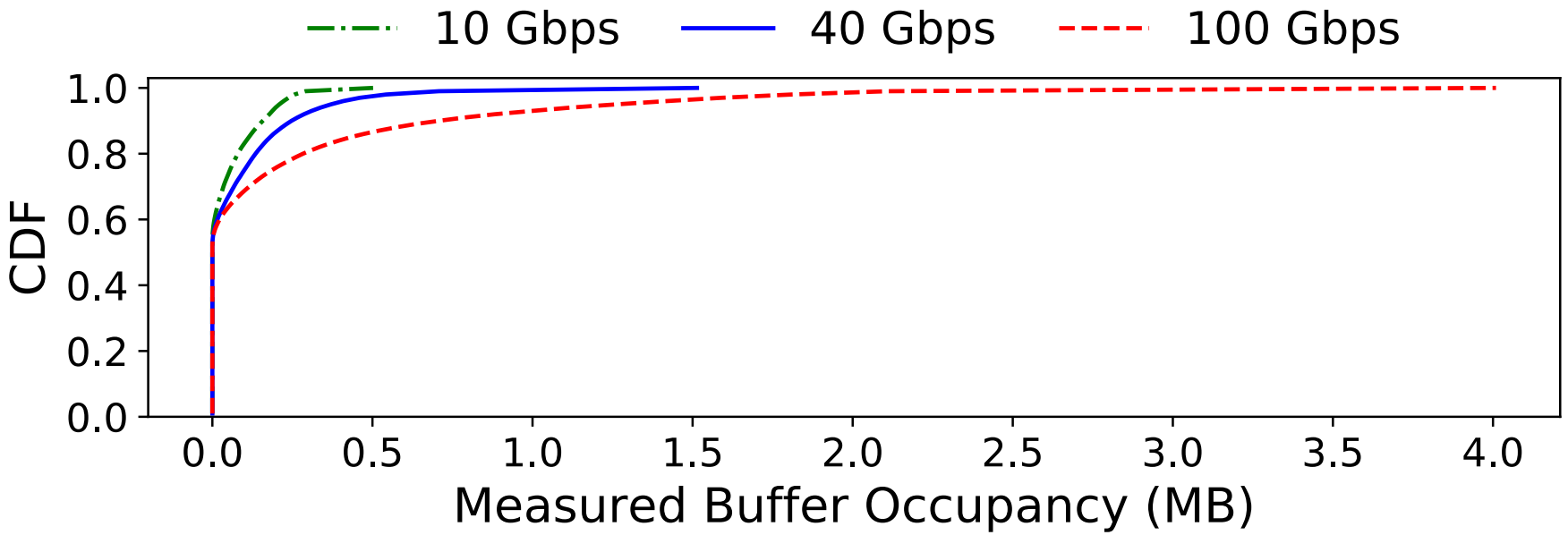
For today's bursty traffic patterns and flow sizes, e2e cc *cannot* provide all of (choose at most 2):

- C: High link capacity
- U: Efficient link utilization
- T: Low tail latency

Hop by hop flow control can provide all three

Backup

Faster Links Harder to Control



DCQCN, Google workload, 75% util+incast