Congestion Control for High-speed Extremely Shallow-buffered Data Center Networks

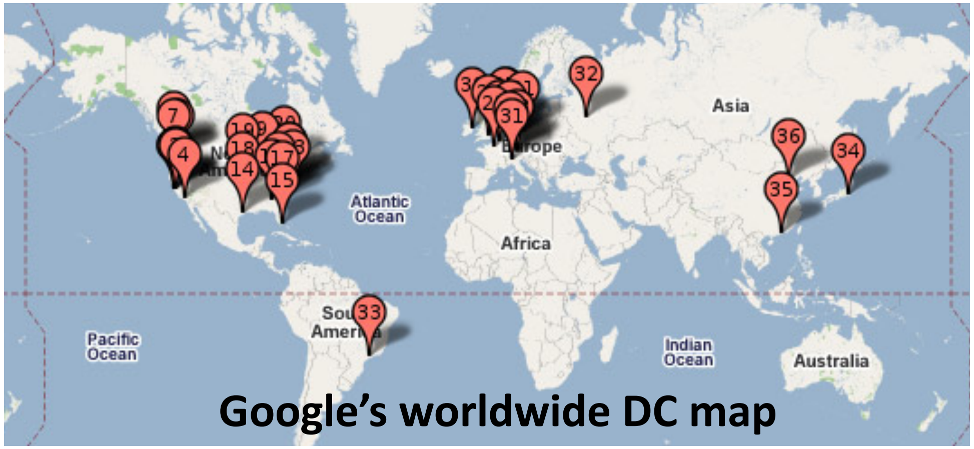
Kai Chen

July 4, 2017 @ SJTU

SingLab HKUST

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

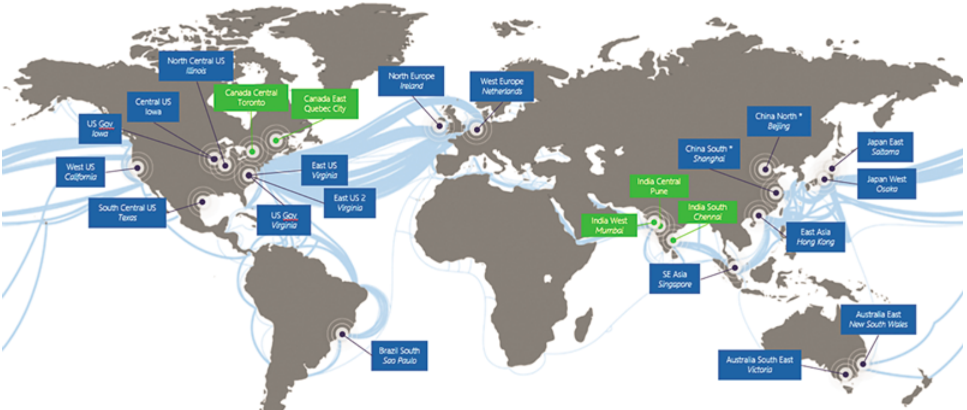
Data centers around the world



Google's worldwide DC map

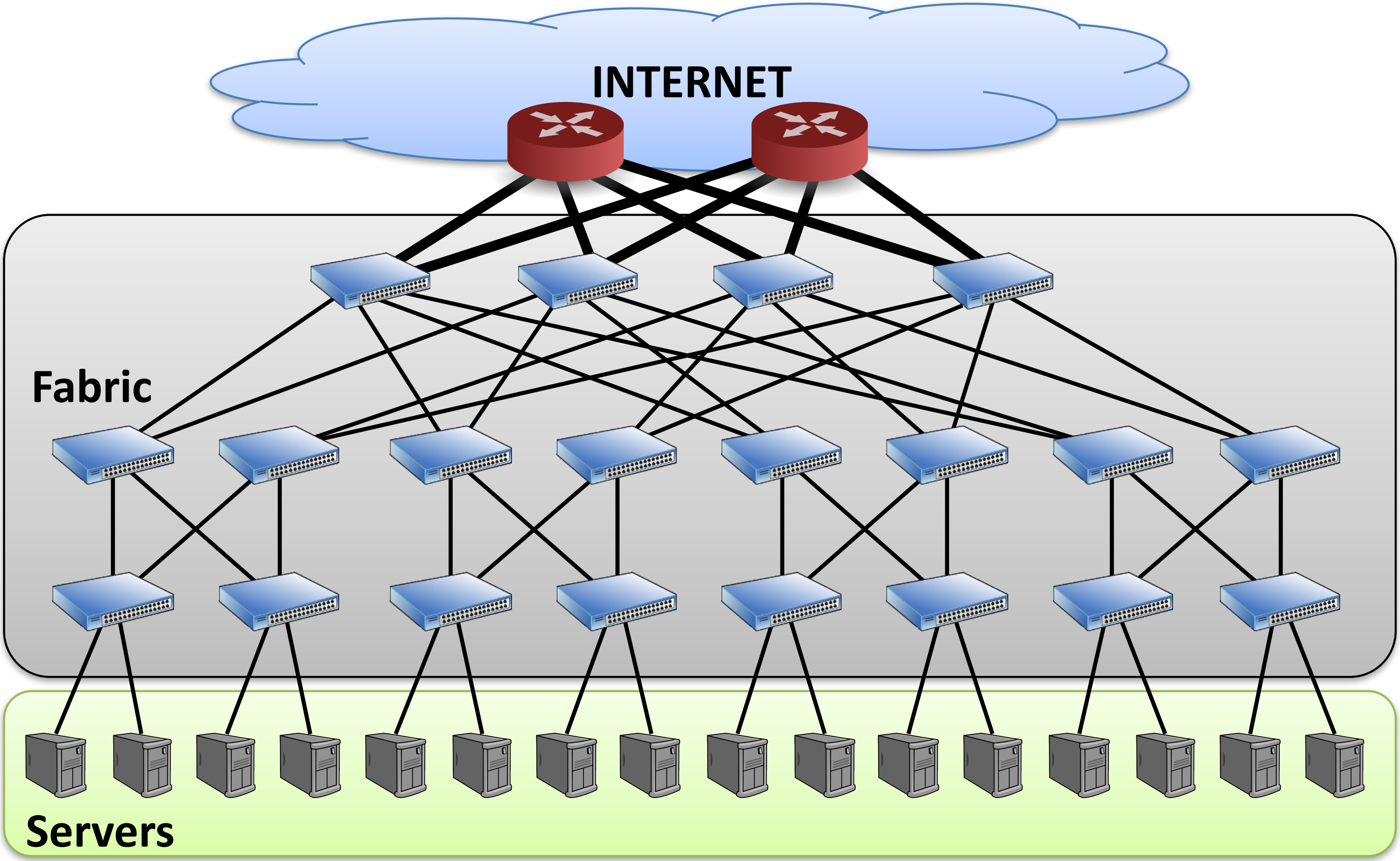
Facebook DC interior

Microsoft's DC in Dublin, Ireland



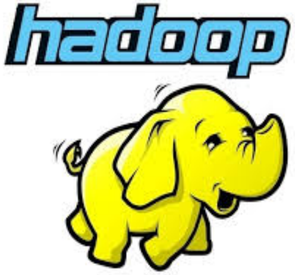
Global Microsoft Azure DC Footprint

Data Center Network (DCN)



Data Center Applications

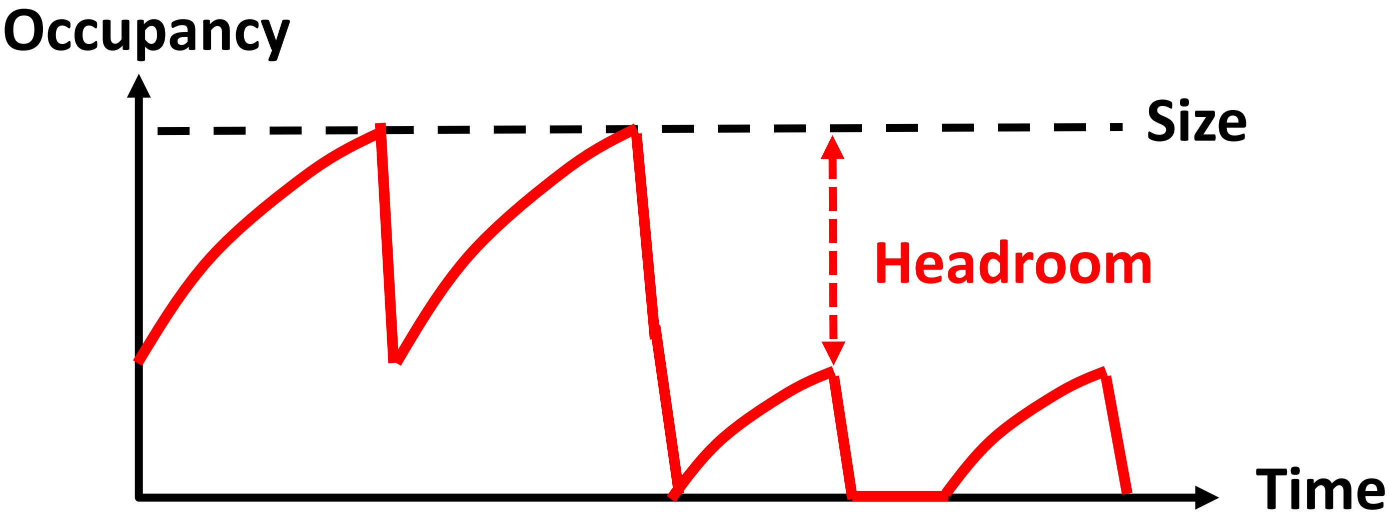
- Network Requirements
  - – Desire low latency for short messages
  - – Desire high throughput for large flows



The challenge is to achieve both goals simultaneously

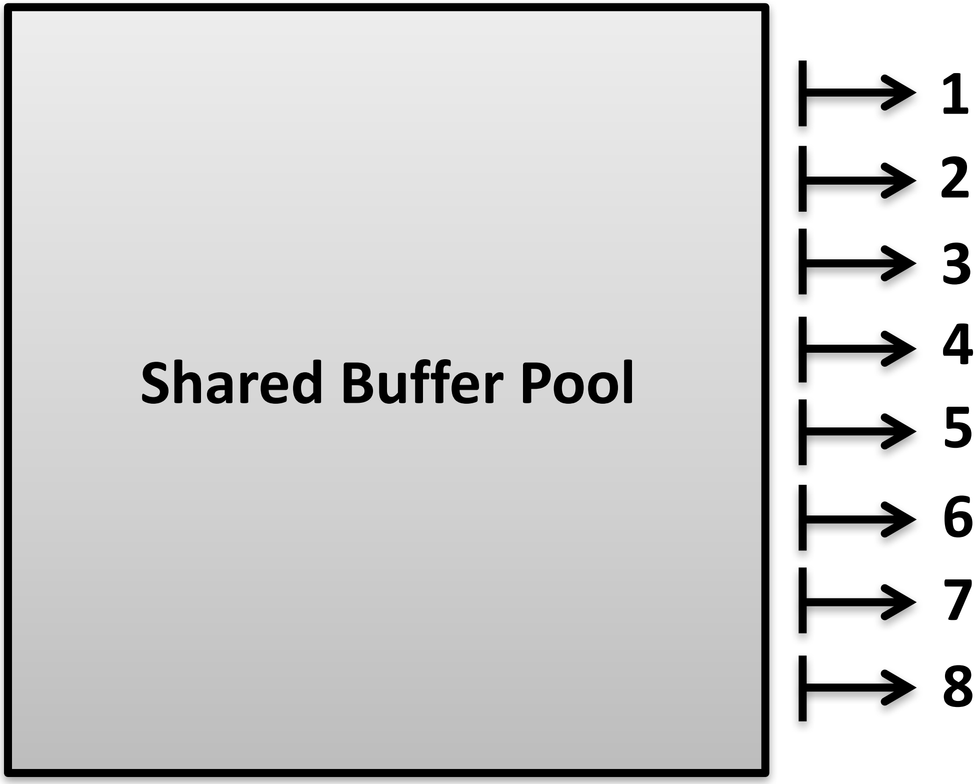
Tension Between Requirements (From buffer's perspective)

- High throughput: **large** switch buffer occupancies
- Low latency: **small** switch buffer occupancies
  - Reduce **queueing delay**
  - Reduce **packet losses** with large headroom



What is current practice?

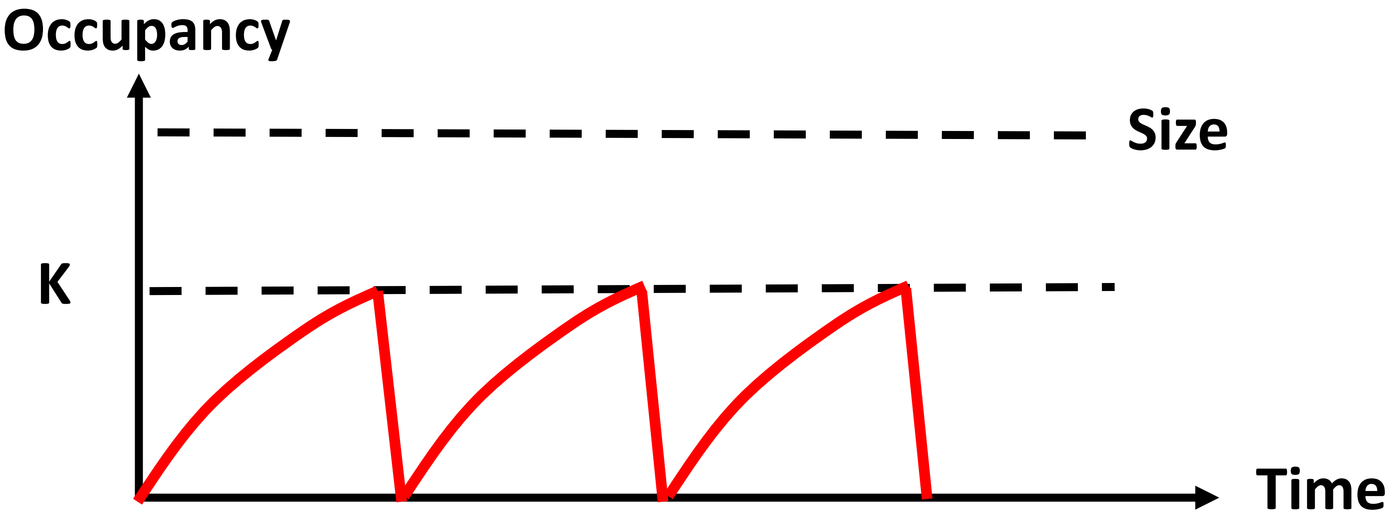
- Dynamic buffer allocation at switch
  - Reduce packet losses



What is current practice?

- Dynamic buffer allocation at switch
  - Reduce packet losses

- ECN-based transports (e.g., DCTCP Sigcomm'10)



What is current practice?

- Dynamic buffer allocation at switch
  - Reduce packet losses

- ECN-based transports (e.g., DCTCP Sigcomm’10)
  - Low buffer occupancies → Low queueing delay
  - Leave headroom → Reduce packet losses
  - $K = C \times RTT \times \lambda$ threshold → 100% throughput

What is current practice?

- Dynamic buffer allocation at switch
  - Reduce packet losses

- ECN-based transports (e.g., DCTCP Sigcomm’10)
  - Low buffer occupancies → Low queueing delay
  - Leave headroom → Reduce packet losses
  - $K = C \times RTT \times \lambda$ threshold → 100% throughput

Basic Buffer Requirement

Current Practice

- Dynamic buffer allocation at switch
  - Reduce packet losses
- ECN-based transports
  - Low buffer occupancies → Low queueing delay
  - Leave headroom → Reduce packet losses
  - $C \times RTT \times \lambda$ threshold → 100% throughput

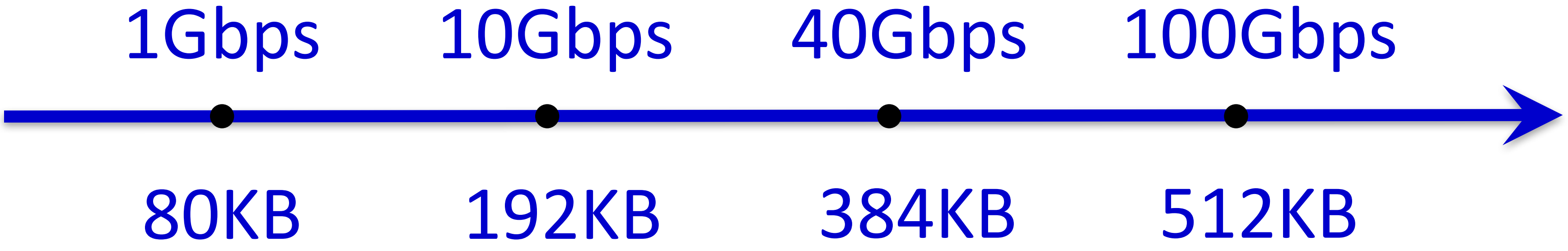
Basic Buffer Requirement

Is current practice good enough?

No with recent trends!

Recent Trends in DCNs

- The link speed scales up quickly
  - 100Gbps and beyond
- The switch buffer does not increase as expected
  - Reasons: cost, price, etc.



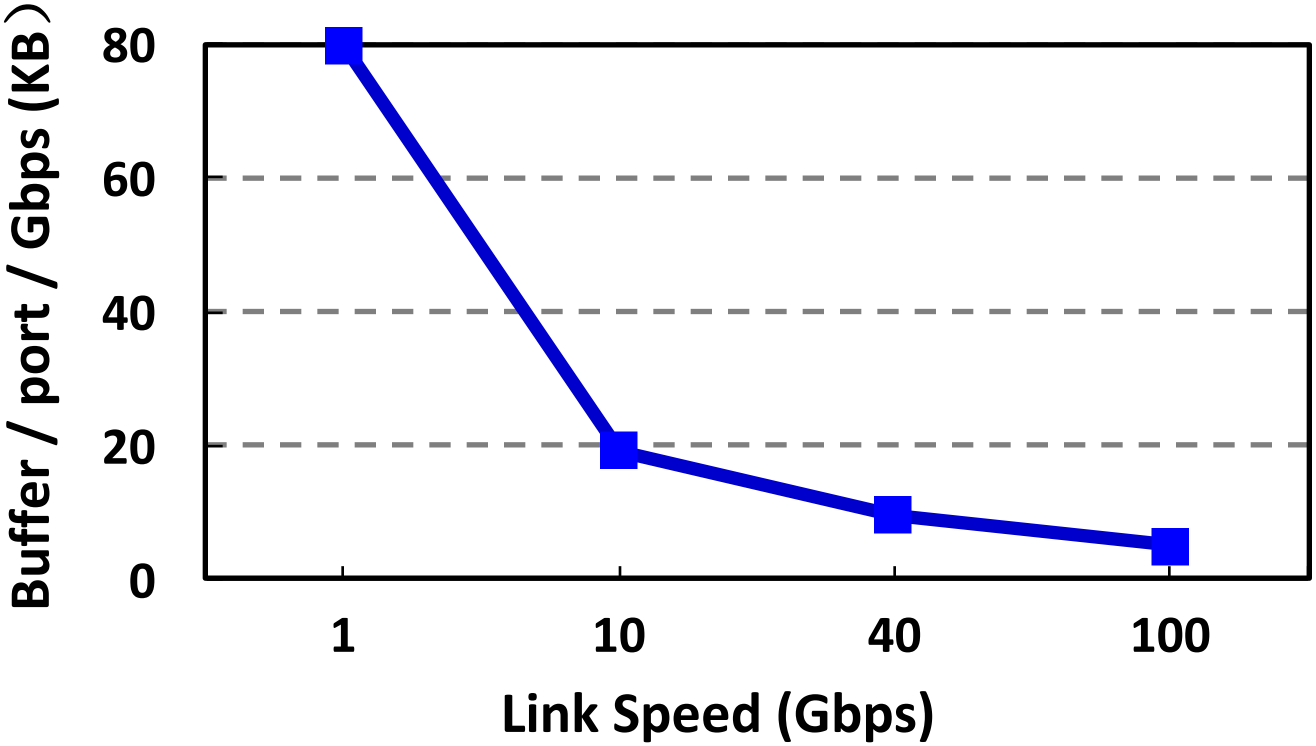
1Gbps 10Gbps 40Gbps 100Gbps

80KB 192KB 384KB 512KB

Buffer / port of Broadcom chips

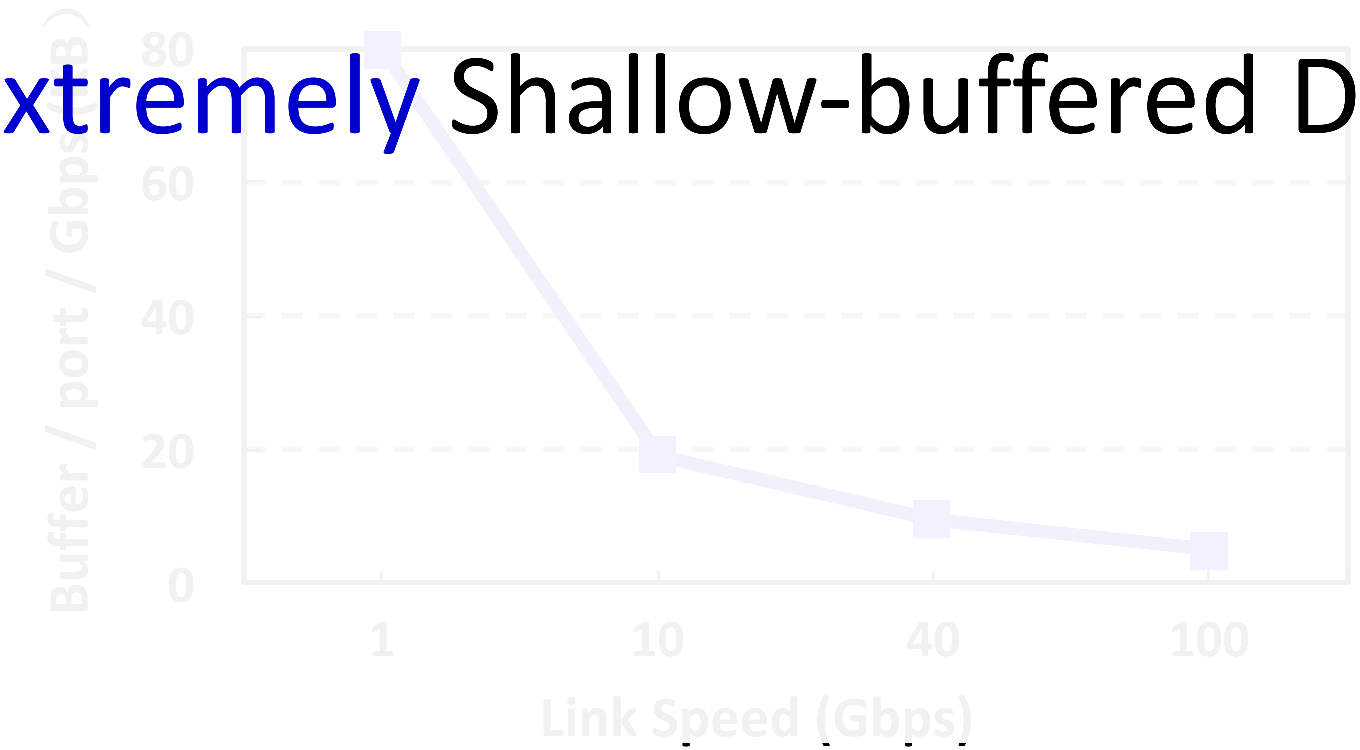
Making it worse …

- Switch buffer becomes increasingly shallow
  - Buffer per port per Gbps keeps decreasing



Observation

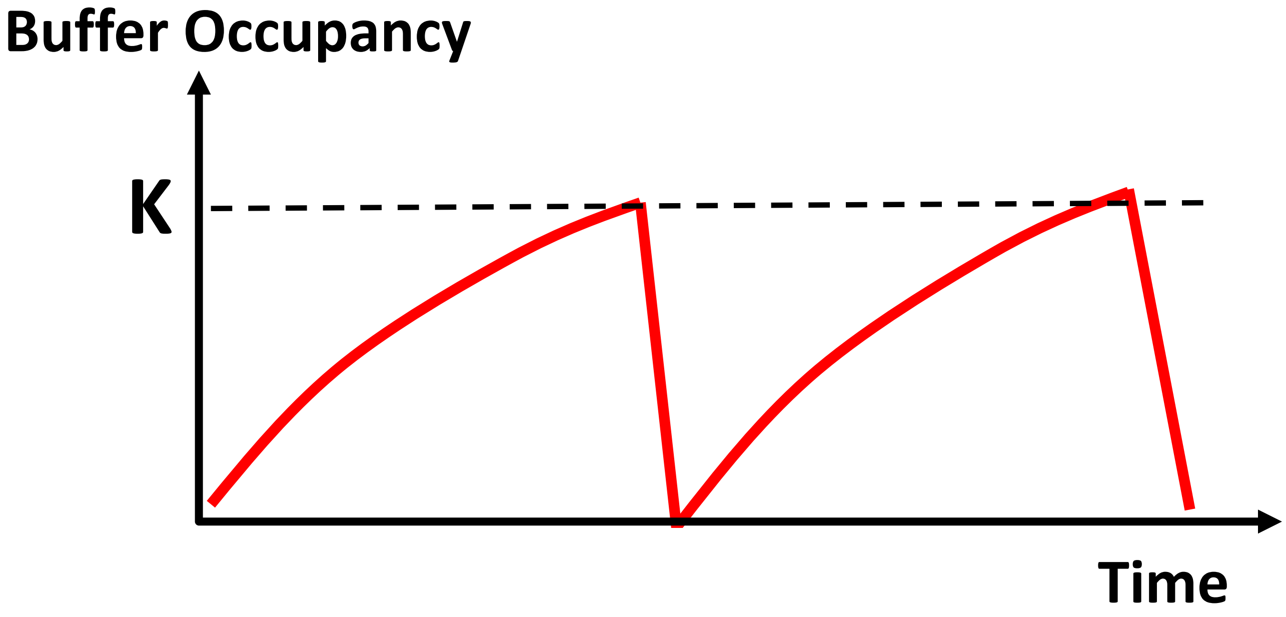
- More and more shallow switch buffer
  - Buffer per port per Gbps keeps decreasing



Extremely Shallow-buffered DCNs

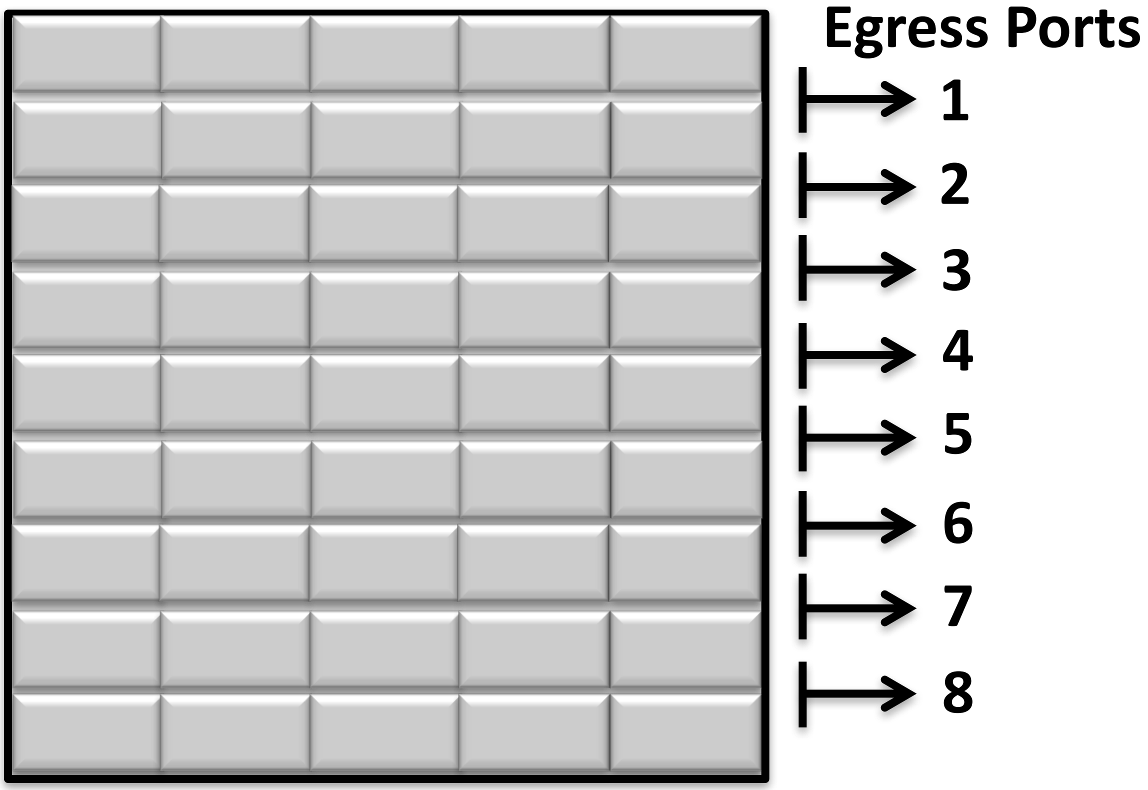
Problems of Existing Solutions (1)

- Standard ECN configuration (current practice)
  - $C \times RTT \times \lambda$ per port for high throughput



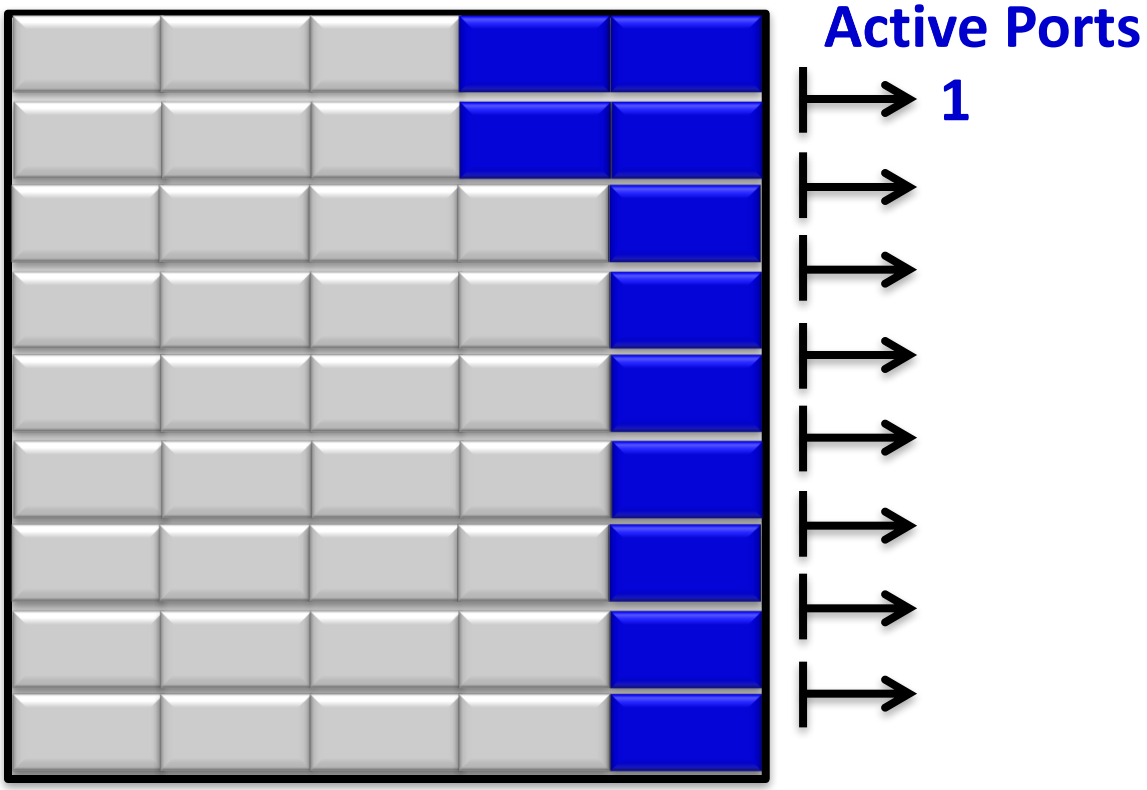
Problems of Existing Solutions (1)

- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput



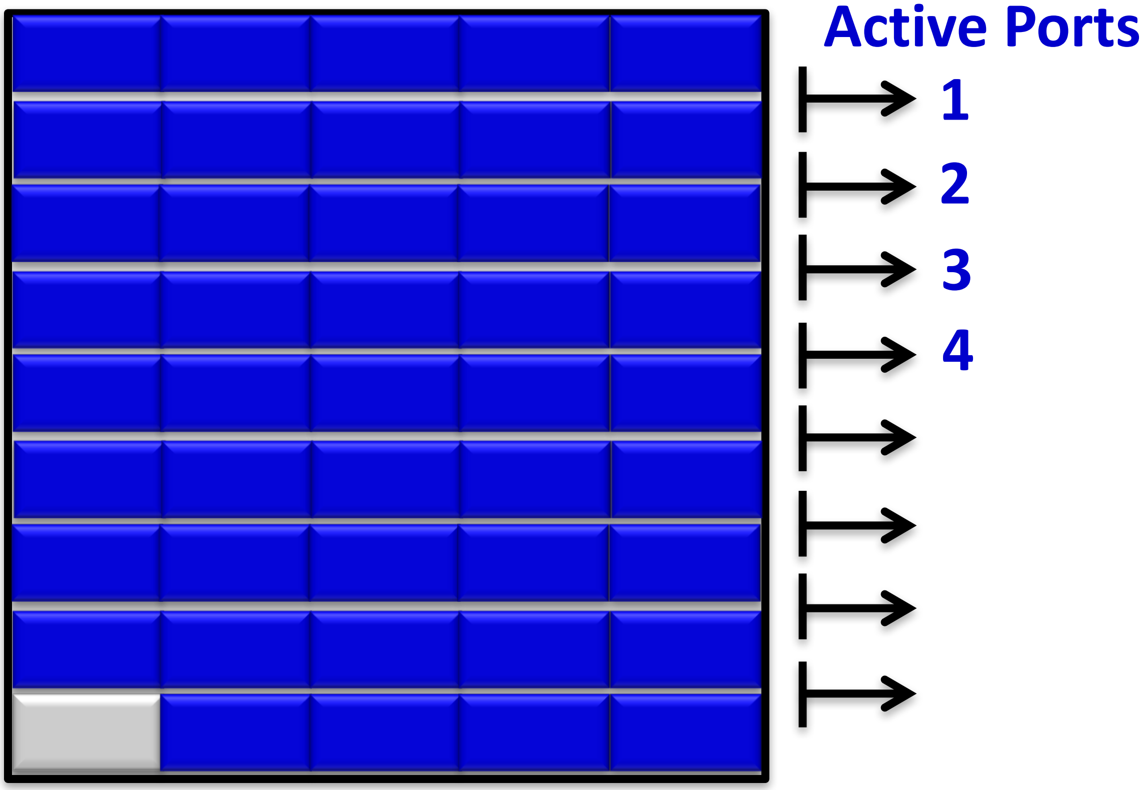
Problems of Existing Solutions (1)

- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput



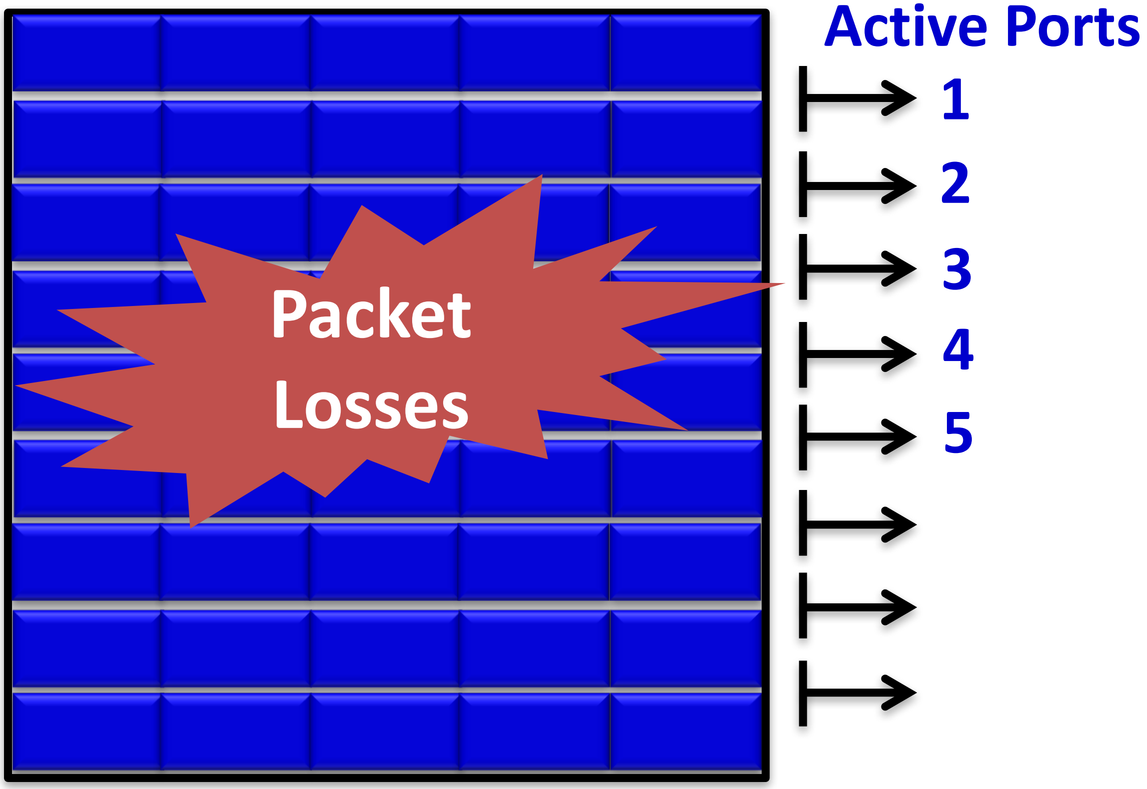
Problems of Existing Solutions (1)

- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput



Problems of Existing Solutions (1)

- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput



Problems of Existing Solutions (1)

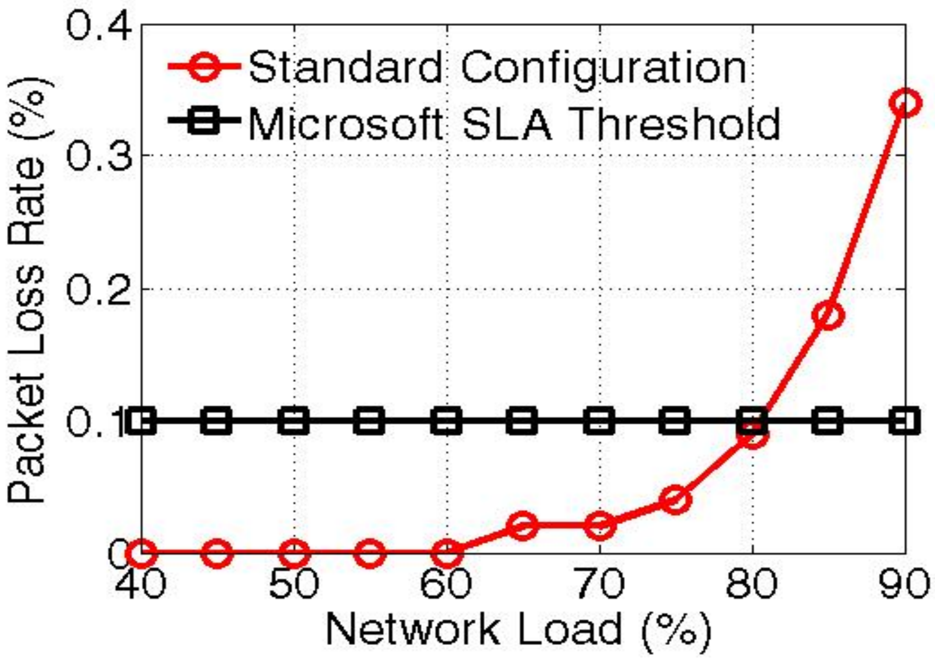
- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput
  - Excessive packet losses with many active ports

Example: Broadcom Tomhawk

- 16MB shared buffer for 32 x 100Gbps ports
- 1MB ($100Gbps \times 80\mu s$) per port buffering
- $\geq$ 50% of ports are active → buffer overflow

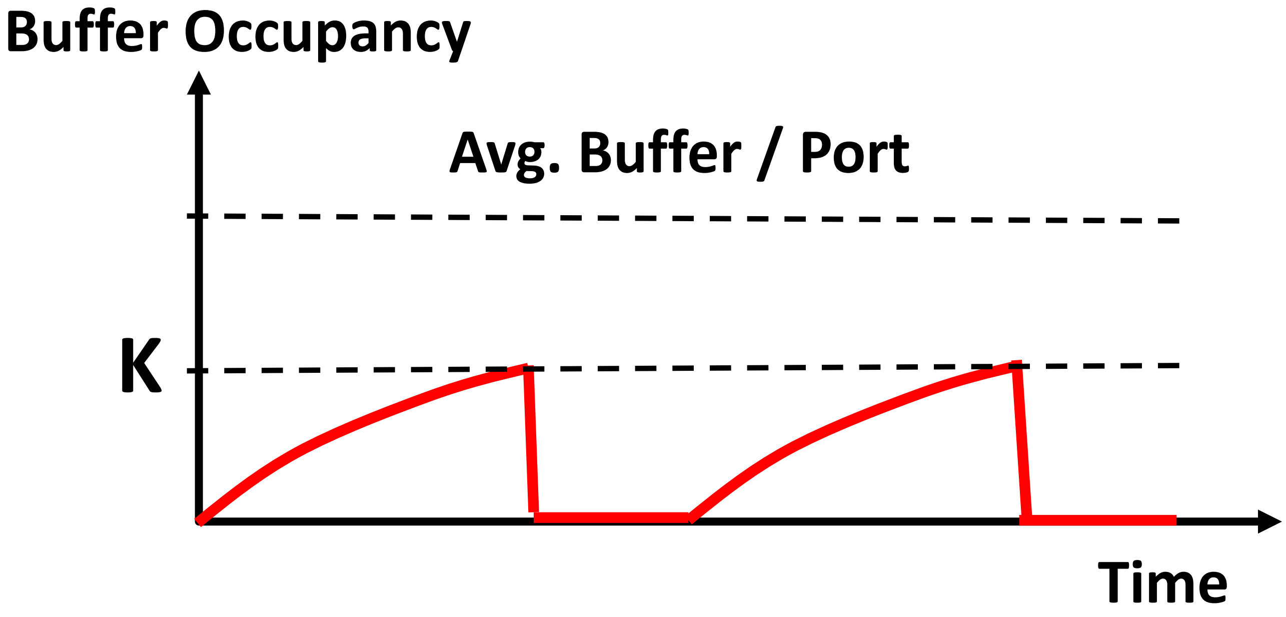
Problems of Existing Solutions (1)

- Standard ECN configuration
  - – $C \times RTT \times \lambda$ per port for high throughput
  - – Excessive packet losses with many active ports



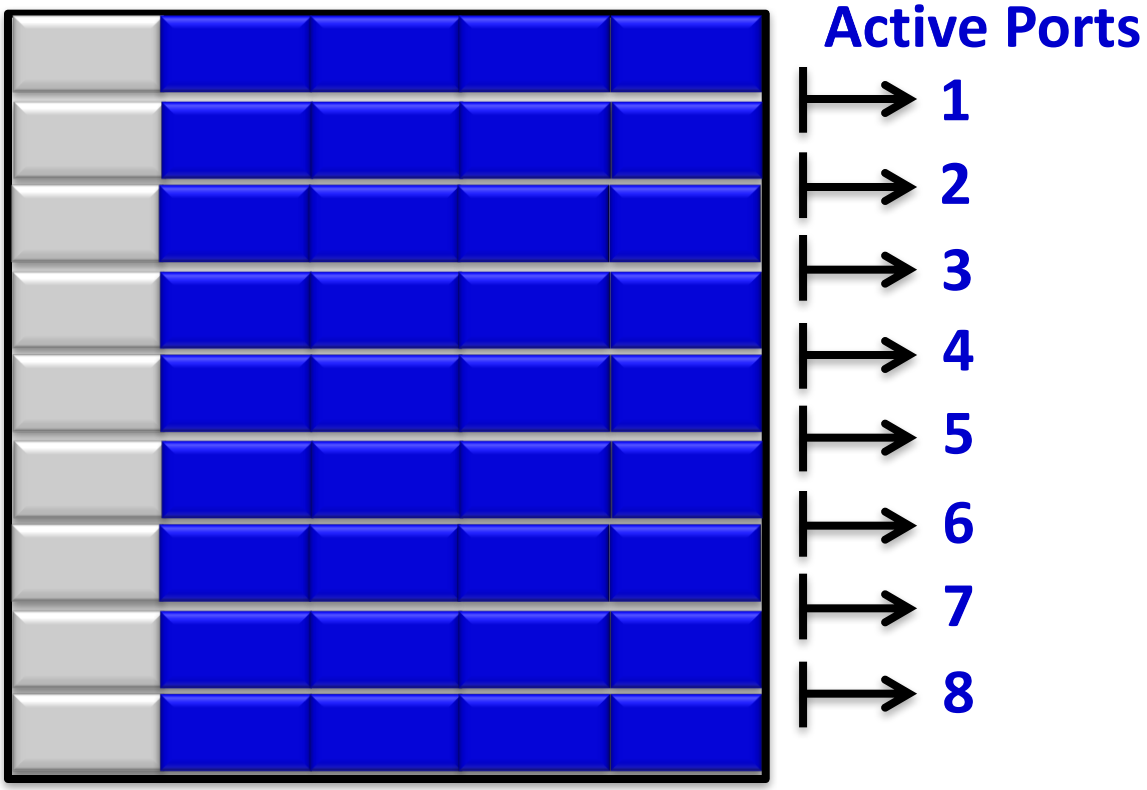
Problems of Existing Solutions (2)

- Conservative ECN configuration
  - Leave headroom for low packet loss rate



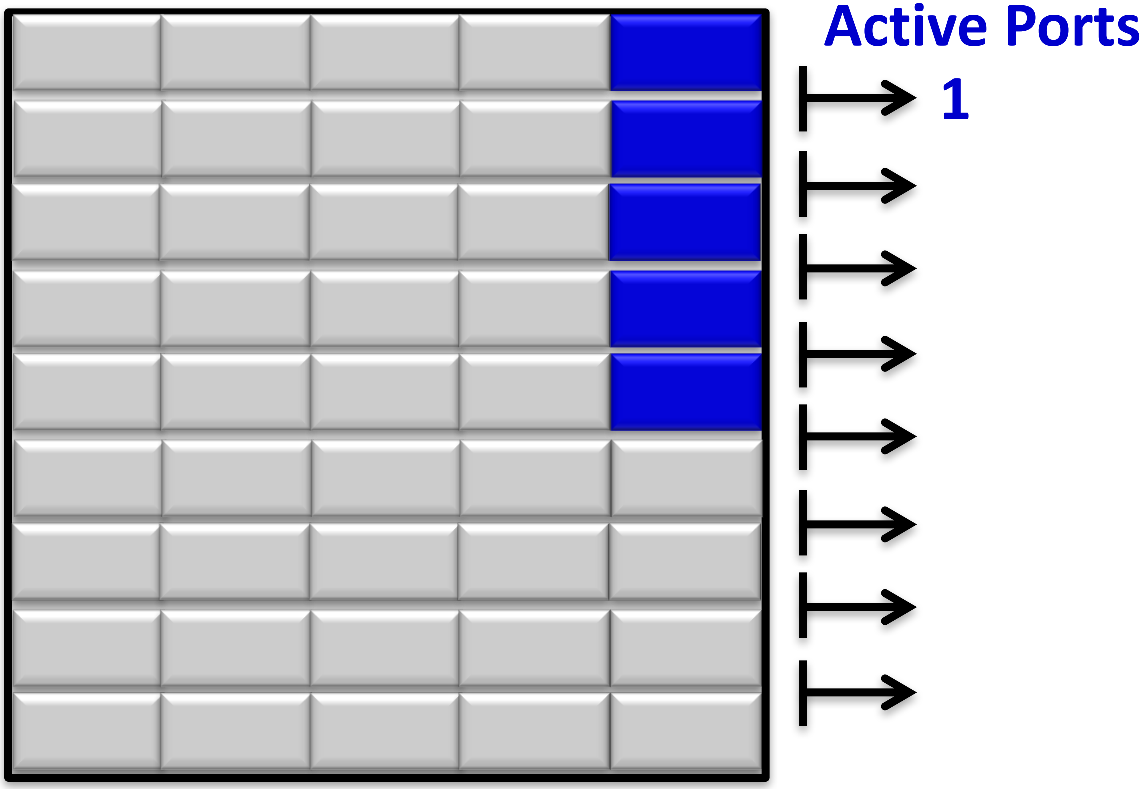
Problems of Existing Solutions (2)

- Conservative ECN configuration
  - Leave headroom for low packet loss rate



Problems of Existing Solutions (2)

- Conservative ECN configuration
  - Leave headroom for low packet loss rate

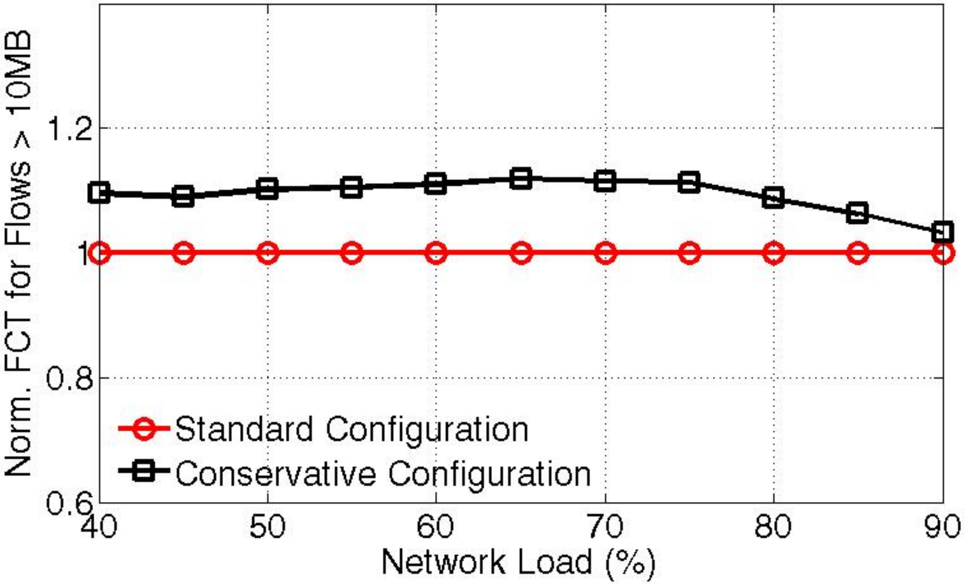


Problems of Existing Solutions (2)

- Conservative ECN configuration
  - Leave headroom for low packet loss rate
  - Significant throughput degradation with few active ports

Problems of Existing Solutions (2)

- Conservative ECN configuration
  - Leave headroom for low packet loss rate
  - Significant throughput degradation with few active ports



Summary of Problems

- Standard ECN configuration
  - $C \times RTT \times \lambda$ per port for high throughput
  - Excessive packet losses with many active ports

- Conservative ECN configuration
  - Leave headroom for low packet loss rate
  - Significant throughput degradation with few active ports

Design Goals

- High Throughput
- Low Packet Loss Rate
- When many ports are active?
  - Packet loss rate prioritized over throughput
- Readily-deployable
  - Legacy Network Stacks & Commodity Switch ASIC

Our Solution

- High Throughput
- Low Packet Loss Rate
- When many ports are active?
  - Packet loss rate prioritized over throughput
- Readily-deployable
  - Legacy Network Stacks & Commodity Switch ASIC

Buffer-aware Congestion Control

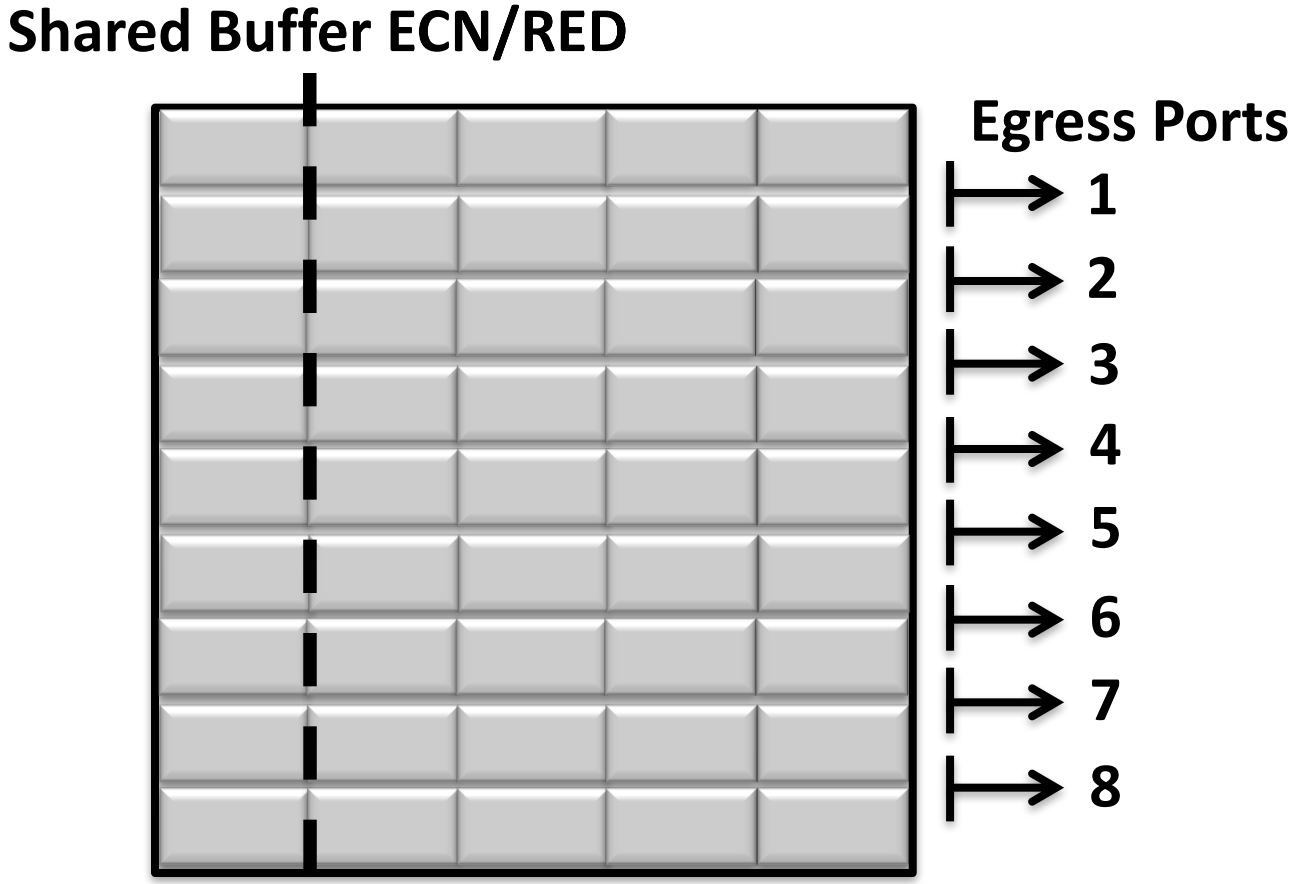
BCC Mechanisms

- End-host
  - Legacy ECN-based transports

- Switch
  - Per port standard ECN configuration
  - Shared buffer ECN/RED

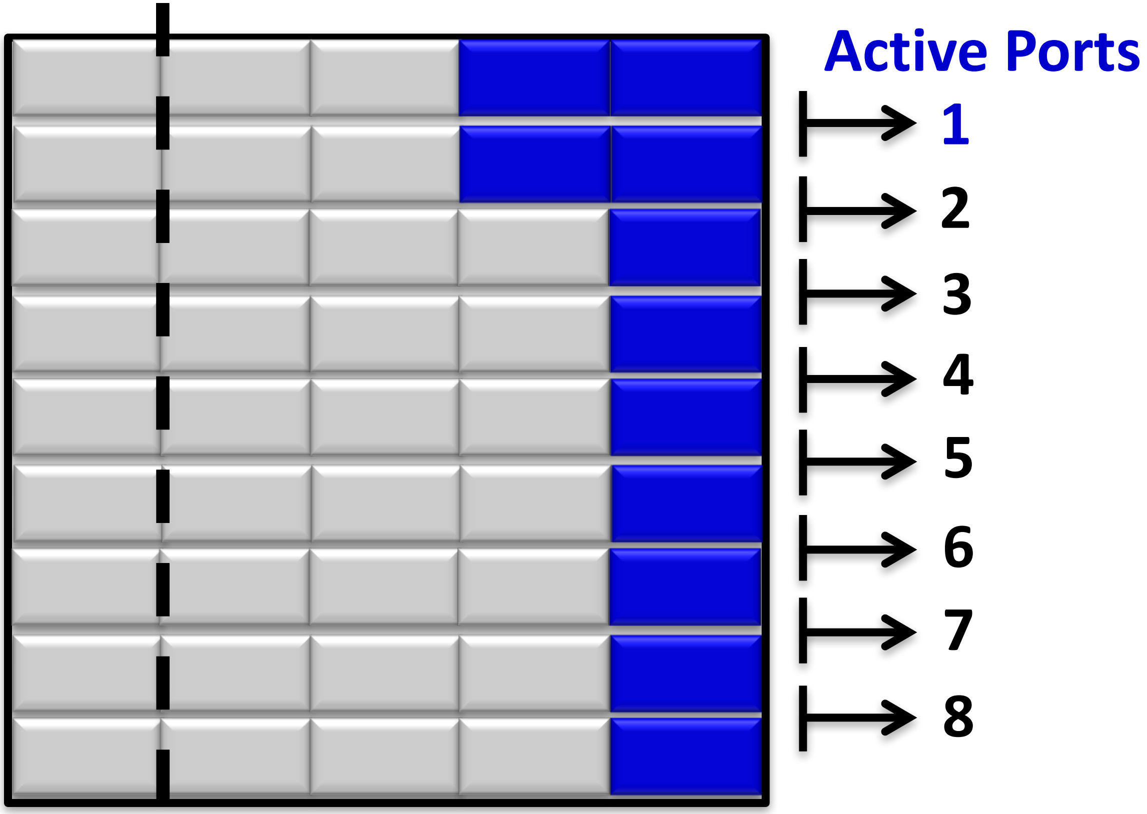
OR

How BCC works?



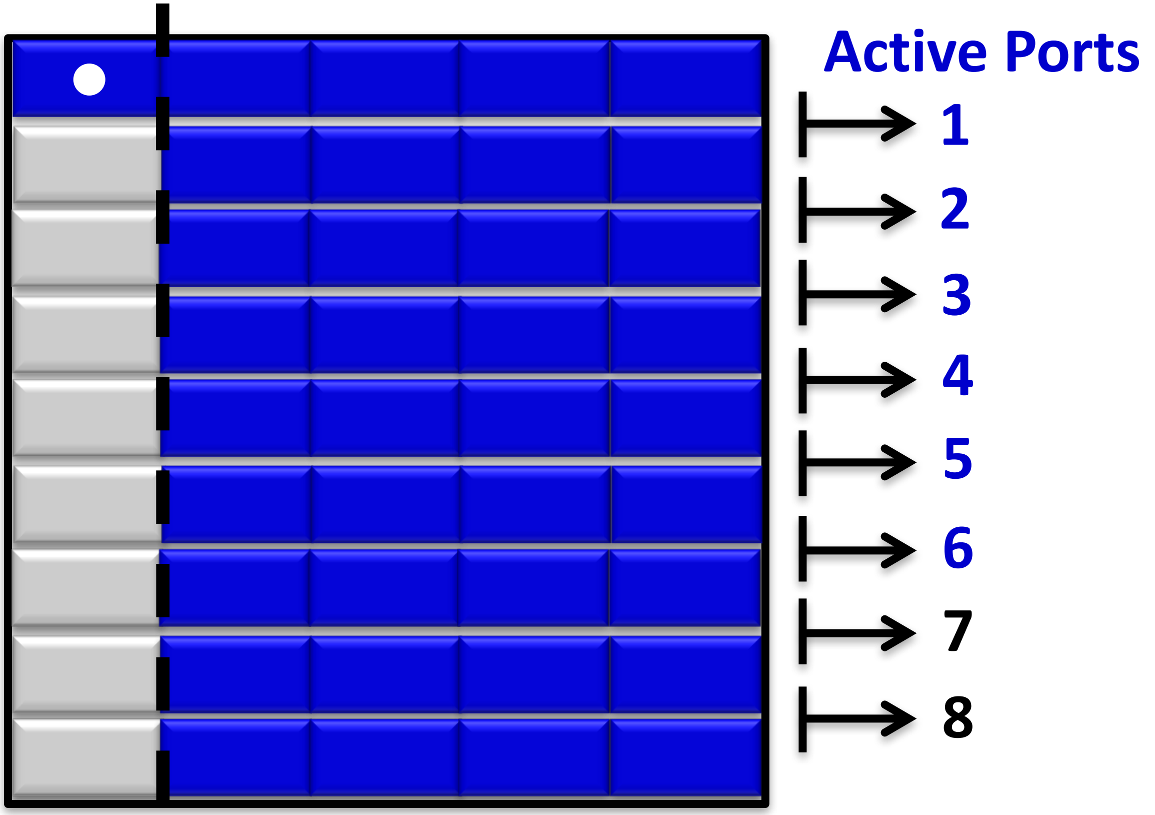
When few ports are active

- Per port standard ECN configuration ensures high throughput & low packet loss rate



When many ports are active

- Shared buffer ECN/RED achieves low packet loss rate at the cost of a small throughput loss



BCC in 1 Slide

- Few Active Ports → Abundant Buffer
  - Per port standard ECN configuration
  - Achieve high throughput & low packet loss rate
- Many Active Ports → Scarce Buffer
  - Shared buffer ECN/RED
  - Trade a little throughput for low packet loss rate

Buffer Aware

BCC in 1 Slide

- Few Active Ports → Abundant Buffer
  - Per port standard ECN configuration
  - Achieve high throughput & low packet loss rate
- Many Active Ports → Scarce Buffer
  - Shared buffer ECN/RED
  - Trade a little throughput for low packet loss rate
- One More ECN Configuration at the Switch

Testbed Validation

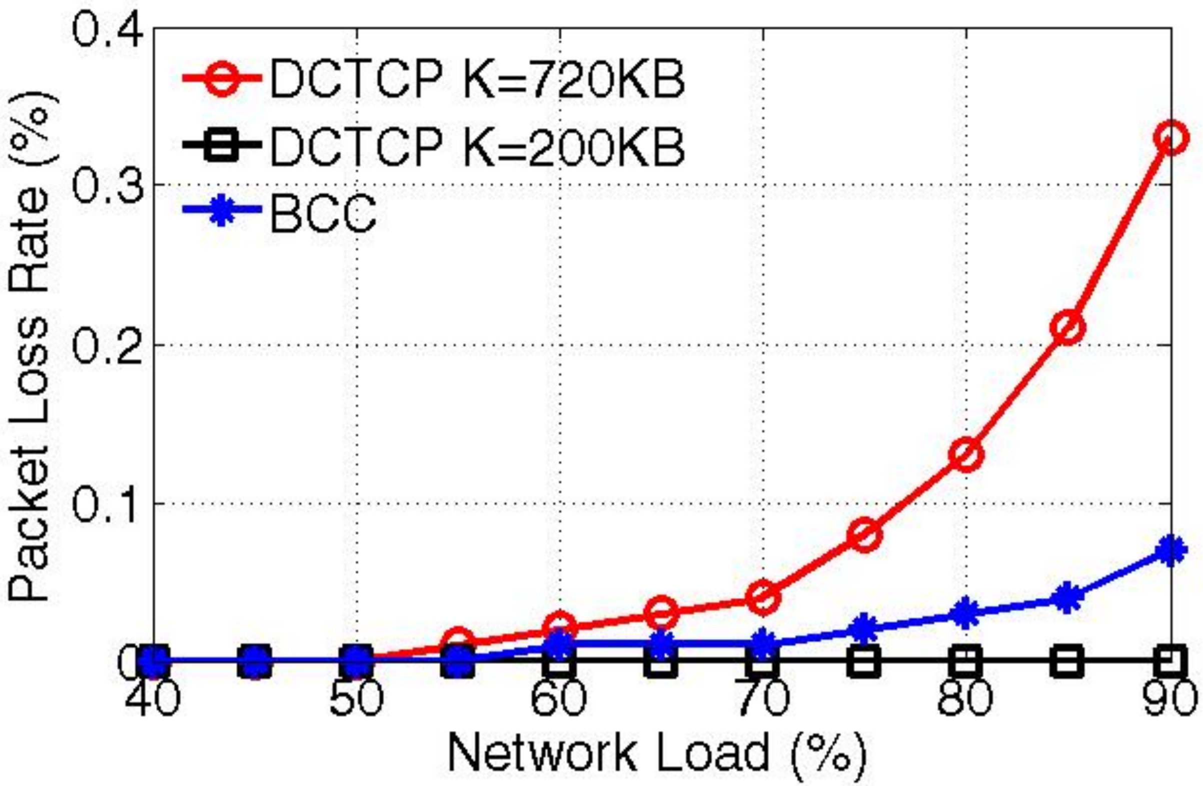
- Functionality Validation at Arista 7060CX-32S 100G switch

```
switch(config)# qos random-detect ecn global-buffer minimum-threshold 500 kbytes maximum-threshold 500 kbytes
```

Large Scale Simulations

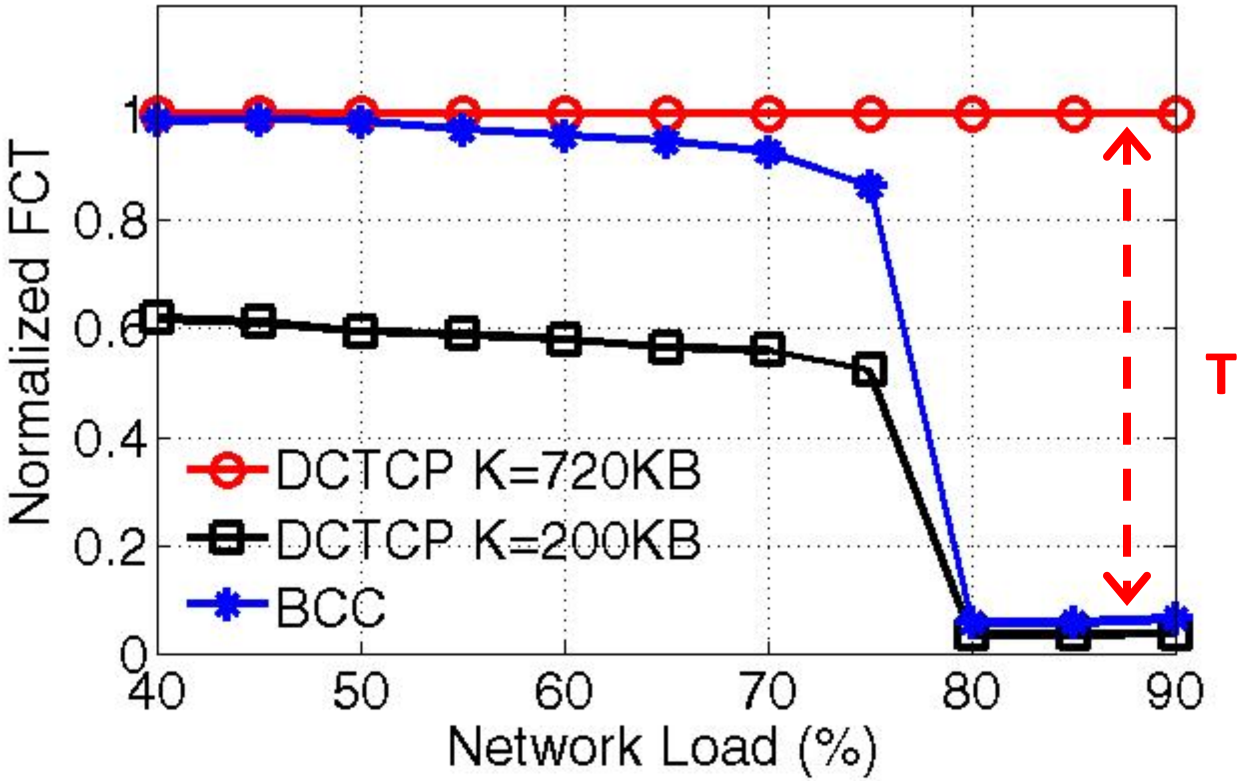
- Settings:
  - 128-host 100Gbps spine-leaf fabric
  - Realistic web search traffic
- Schemes compared
  - Standard per port ECN/RED (K = 720KB)
  - Conservative per port ECN/RED (K = 200KB)
- Metrics:
  - Flow Completion Time (FCT) & Packet Loss Rate

Packet Loss Rate



BCC keeps low packet loss rate

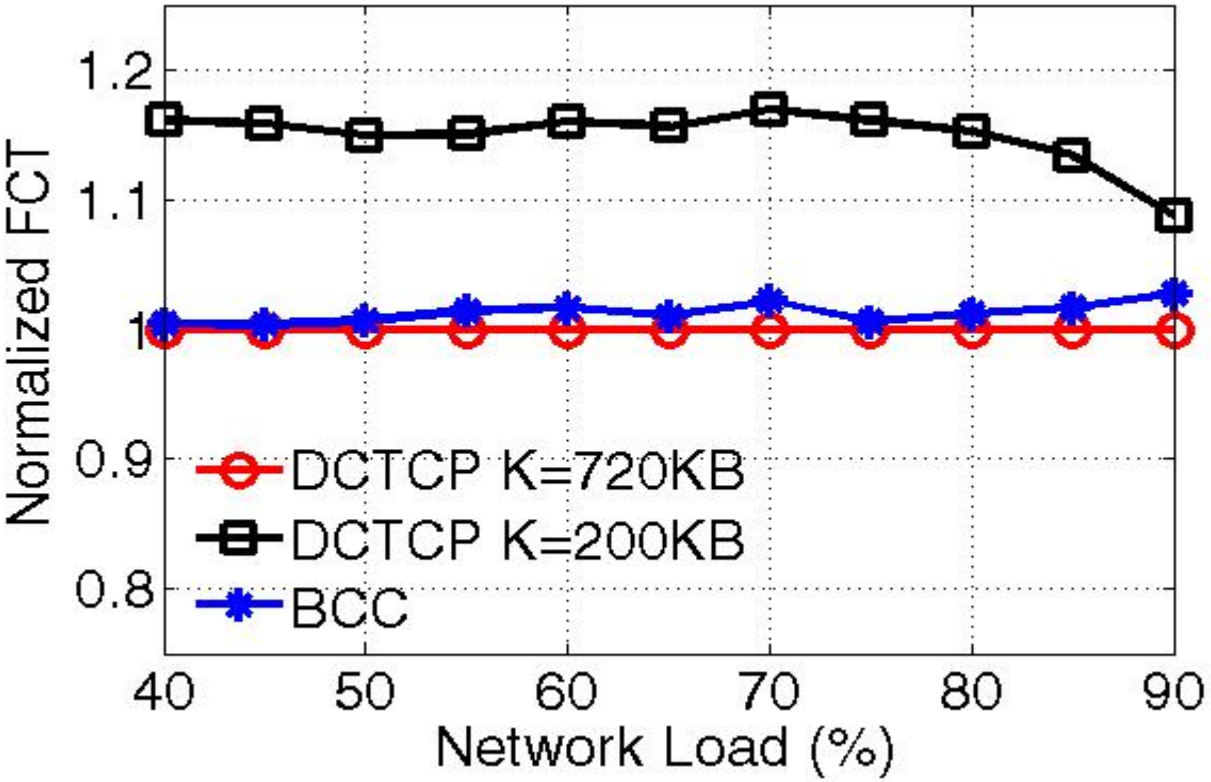
99th percentile FCT for Flows <100KB



TCP RTO

BCC keeps low packet loss rate

Average FCT for Flows > 10MB



BCC only trades a little throughput

BCC Recap

- Abundant Buffer
  - Deliver high throughput & low packet loss rate
- Scarce Buffer
  - Trade a little throughput for low packet loss rate
- Readily-deployable
  - One more ECN configuration is enough

Thanks!