Western Digital.

A Journey into NVMe-oF™: Options, Trade-offs and Challenges

Ihab Hamadi

Fellow, Western Digital

August 8, 2019

Flash Memory Summit

Agenda

1 Background

2 Landscape of NVMe™ Fabrics

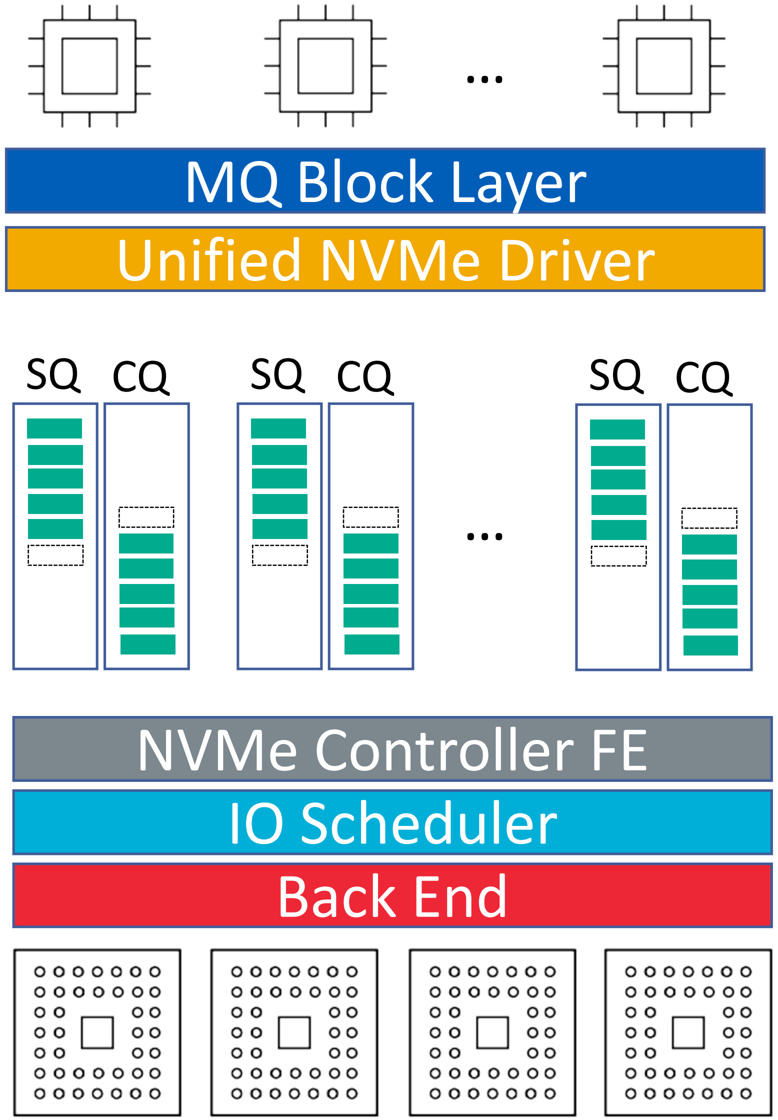
3 Lossless vs. Lossy

4 Fabric Selection Criteria

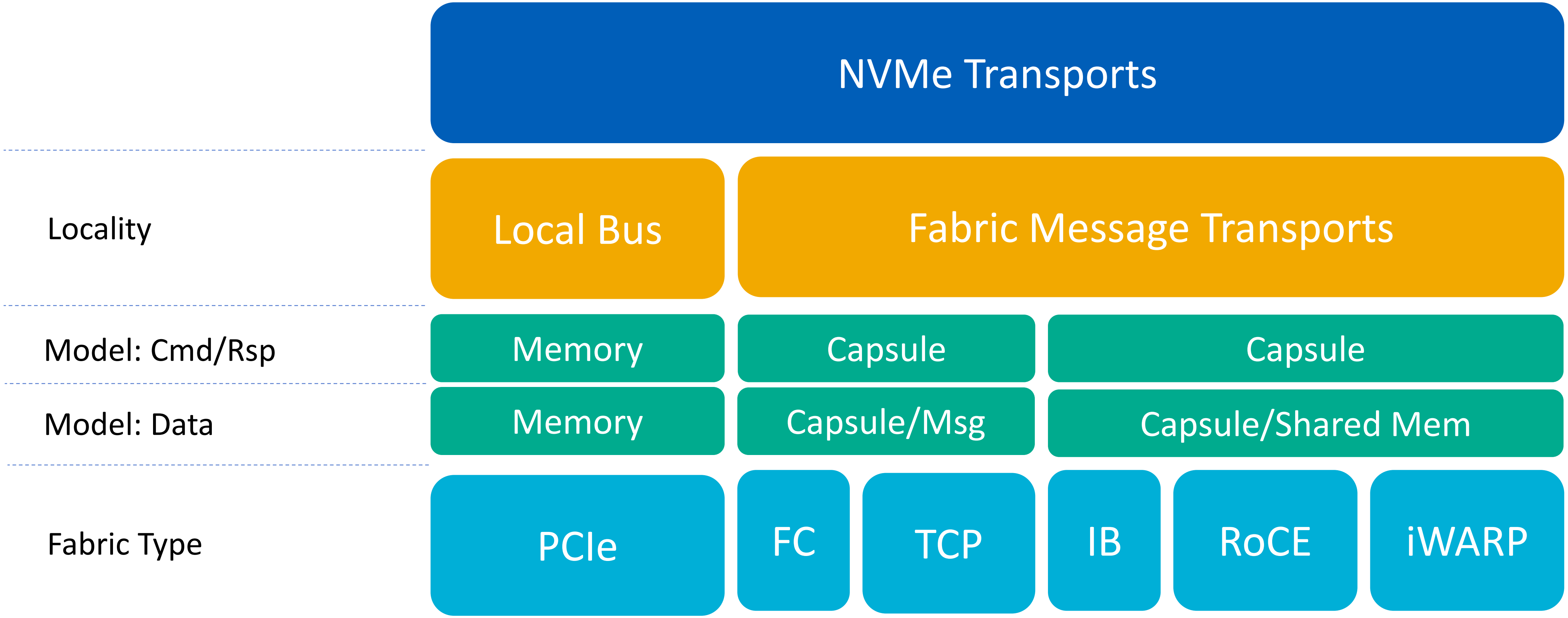
5 Case Study

Background: Why NVMe? Why NVMe-oF?

- Parallelism fits multi-core CPUs
  - Also reduces/spreads host CPU load
- Removes some cost components
  - Some common HW blocks
  - One driver
- Storage System Benefits
  - Lower latency (average & tail)
  - Higher BW
- NVMe-oF Motivation:
  - Extend benefits end-to-end



NVMe Transport Model

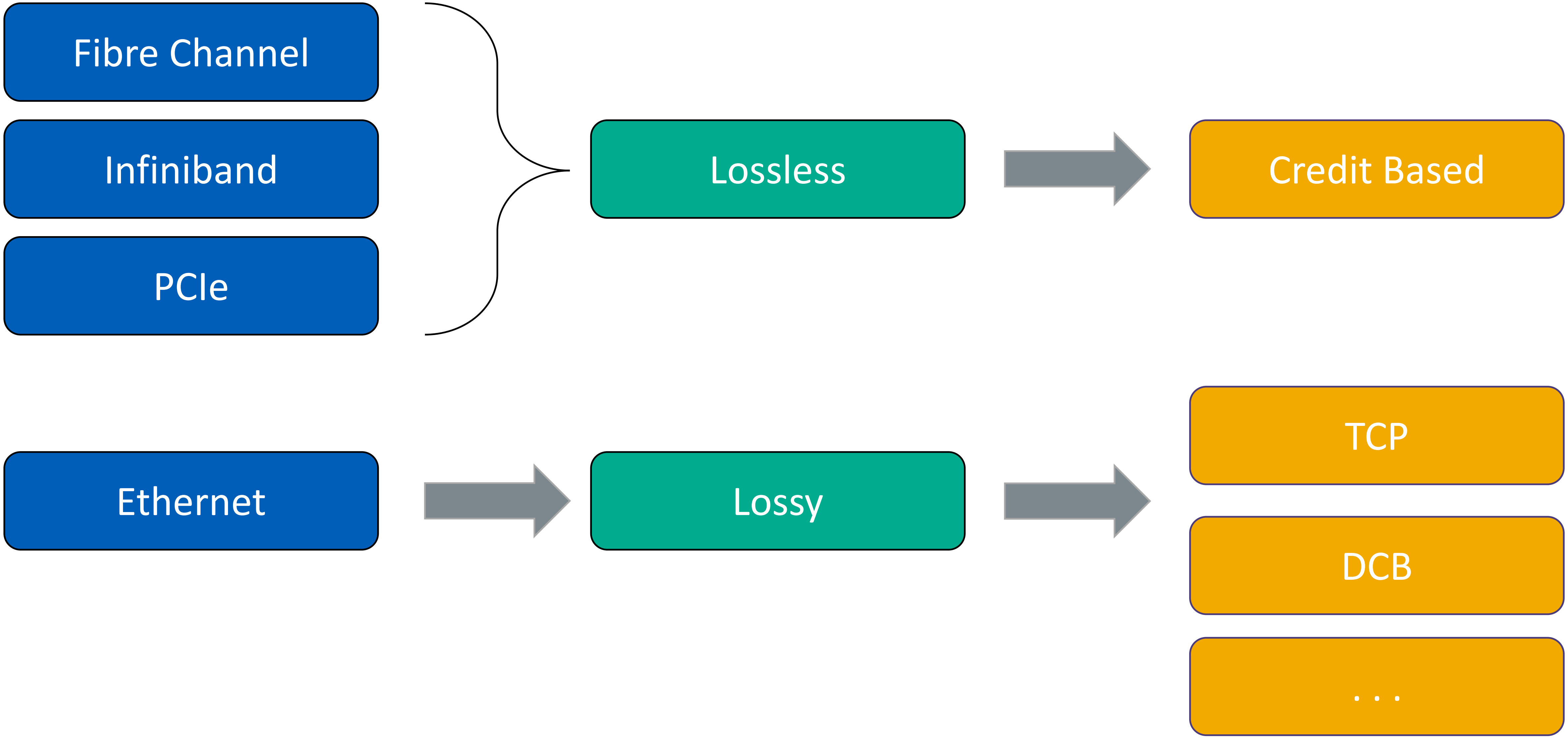


	NVMe Transports		
Locality	Local Bus	Fabric Message Transports	
Model: Cmd/Rsp	Memory	Capsule	Capsule
Model: Data	Memory	Capsule/Msg	Capsule/Shared Mem
Fabric Type	PCIe	FC TCP	IB RoCE iWARP

Western Digital.

Flash Memory Summit 2019, Santa Clara, CA

Fabric 101: Lossy vs. Lossless Fabrics



Data Center Bridging (DCB)

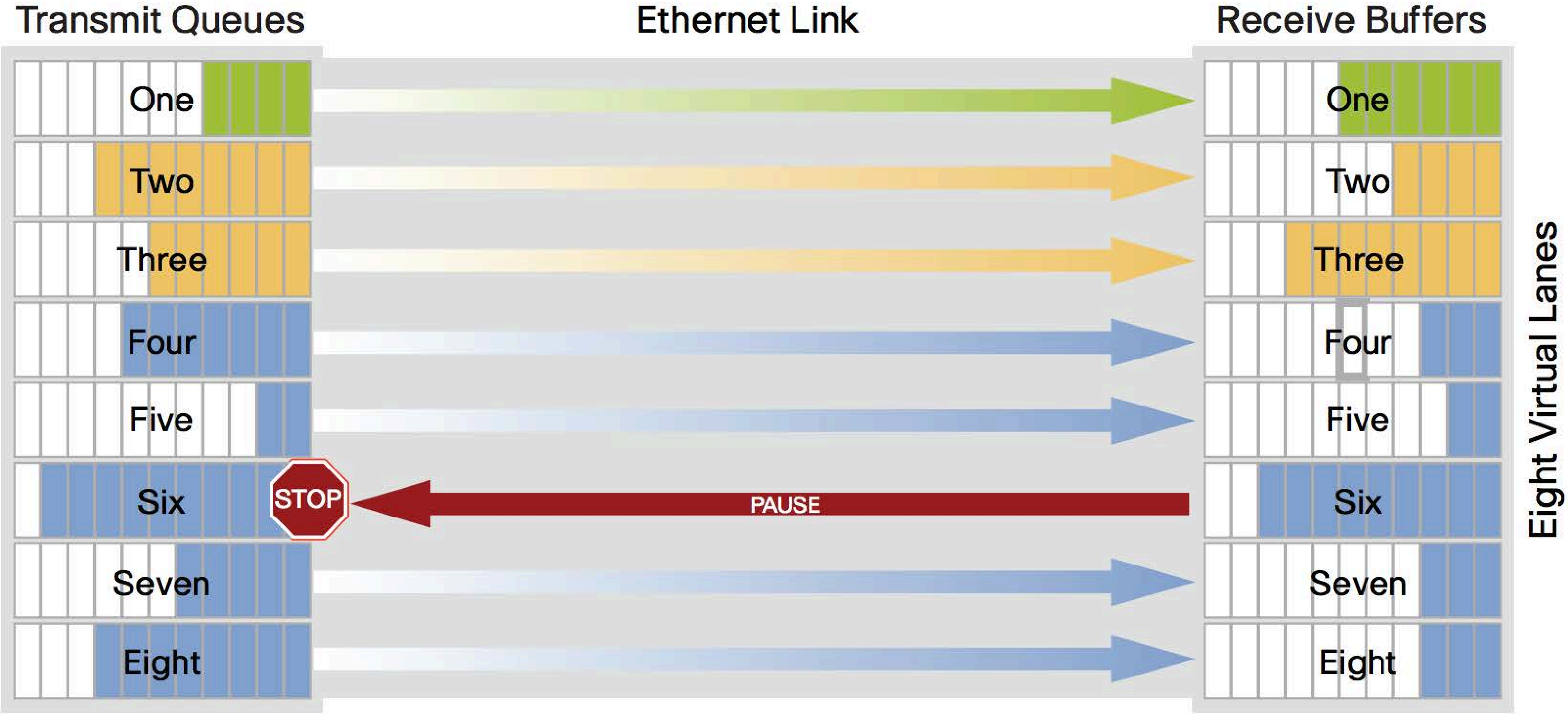
PFC

ECN

DCBx

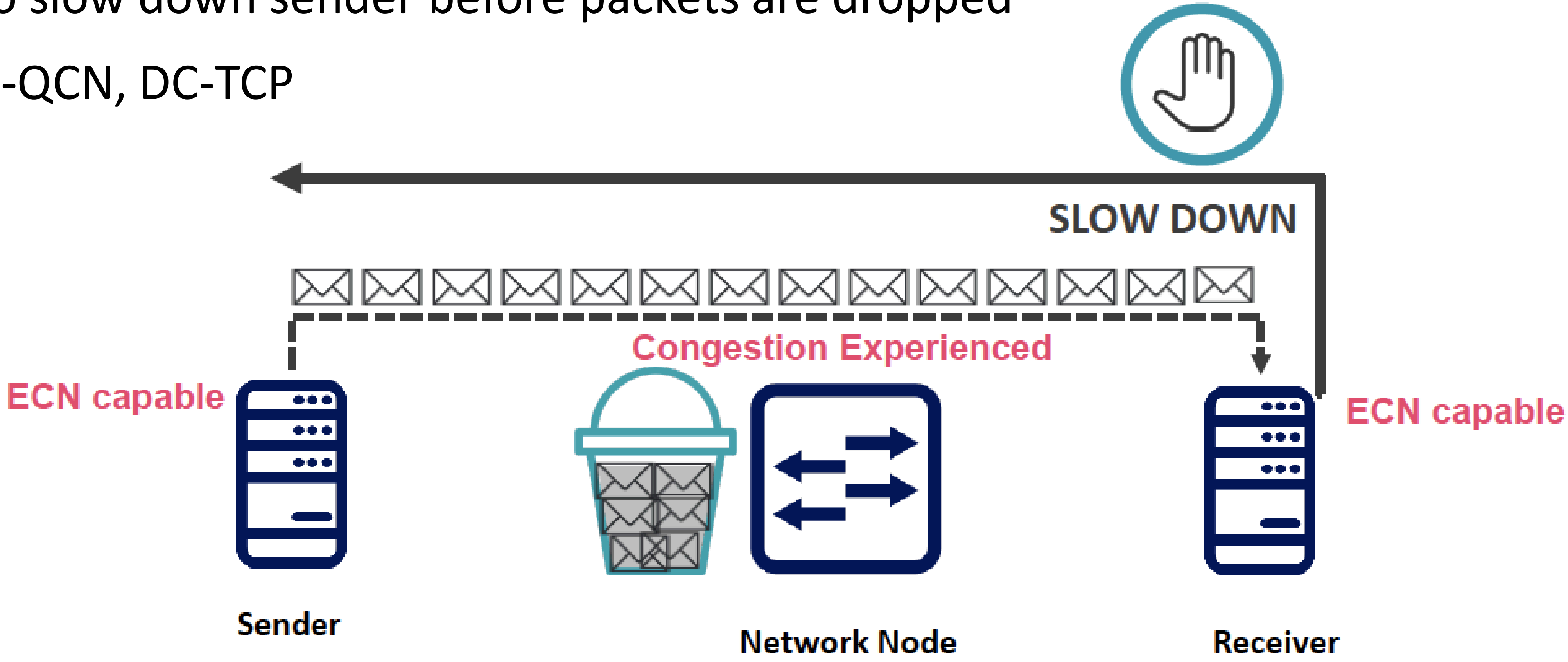
ETS

DCB: PFC


Transmit Queues
Ethernet Link
Receive Buffers
One
Two
Three
Four
Five
Six
STOP
PAUSE
Seven
Eight
One
Two
Three
Four
Five
Six
Seven
Eight
Eight Virtual Lanes

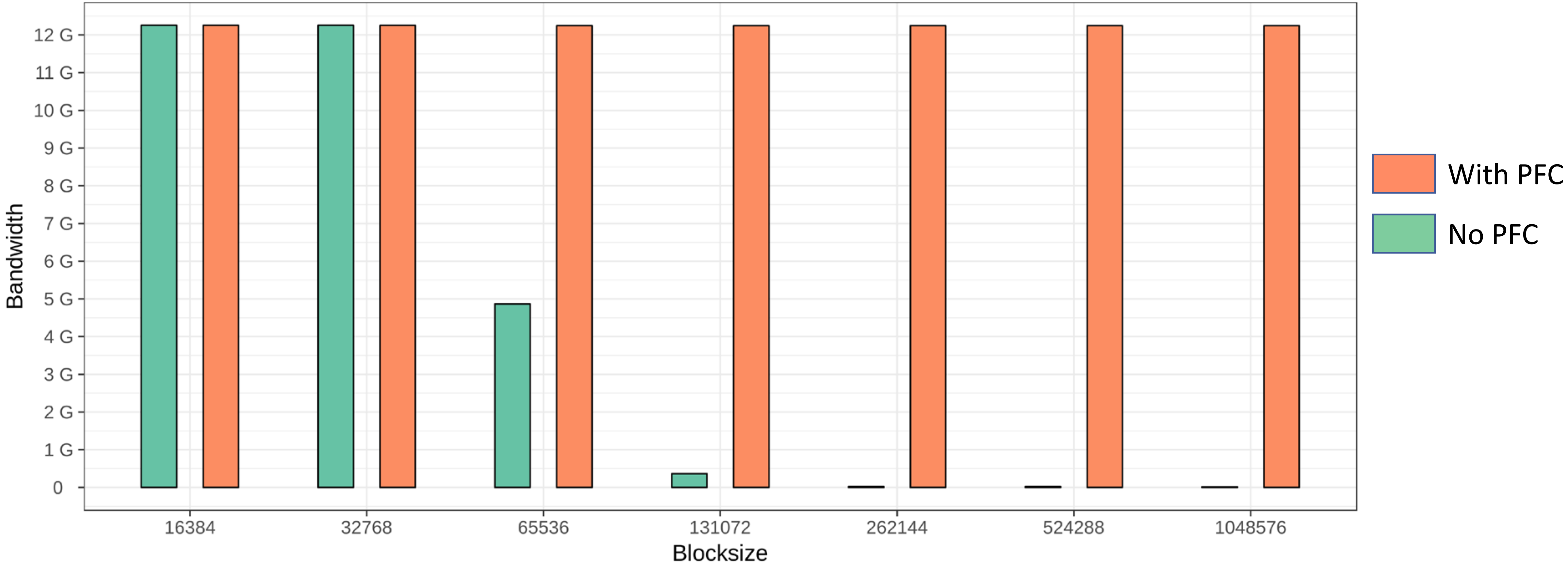
DCB: ECN

- ECN is end-to-end congestion management mechanism
- Three roles: Sender (RP), Switch (CP), Receiver (NP)
- Goal is to slow down sender before packets are dropped
- QCN, DC-QCN, DC-TCP



Do I Really Need DCB (Lossless Net) with RoCE?

BW vs. IO Size



Source: Western Digital Performance Tests

Flash Memory Summit 2019, Santa Clara, CA

8/9/2019 9

Fabric Selection Criteria

Environment

Target Loc

Accessibility

Distance

Existing Fabrics

Consumer Loc

Regulatory

Multi-tenancy

Metrics

Perf: Latency
Perf: Predictability
Perf: Consistency
Perf: Bandwidth

Cost: $/Port
Cost: CPU/BW
Cost: CPU/IOPS

Scale

Single Rack
Multi-rack
Clos architecture
Oversubscription
Link aggregation
Redundancy

Operations

Onboarding
Configuration
Automation
Adv Telemtry
Intent Based
SW Defined <x>

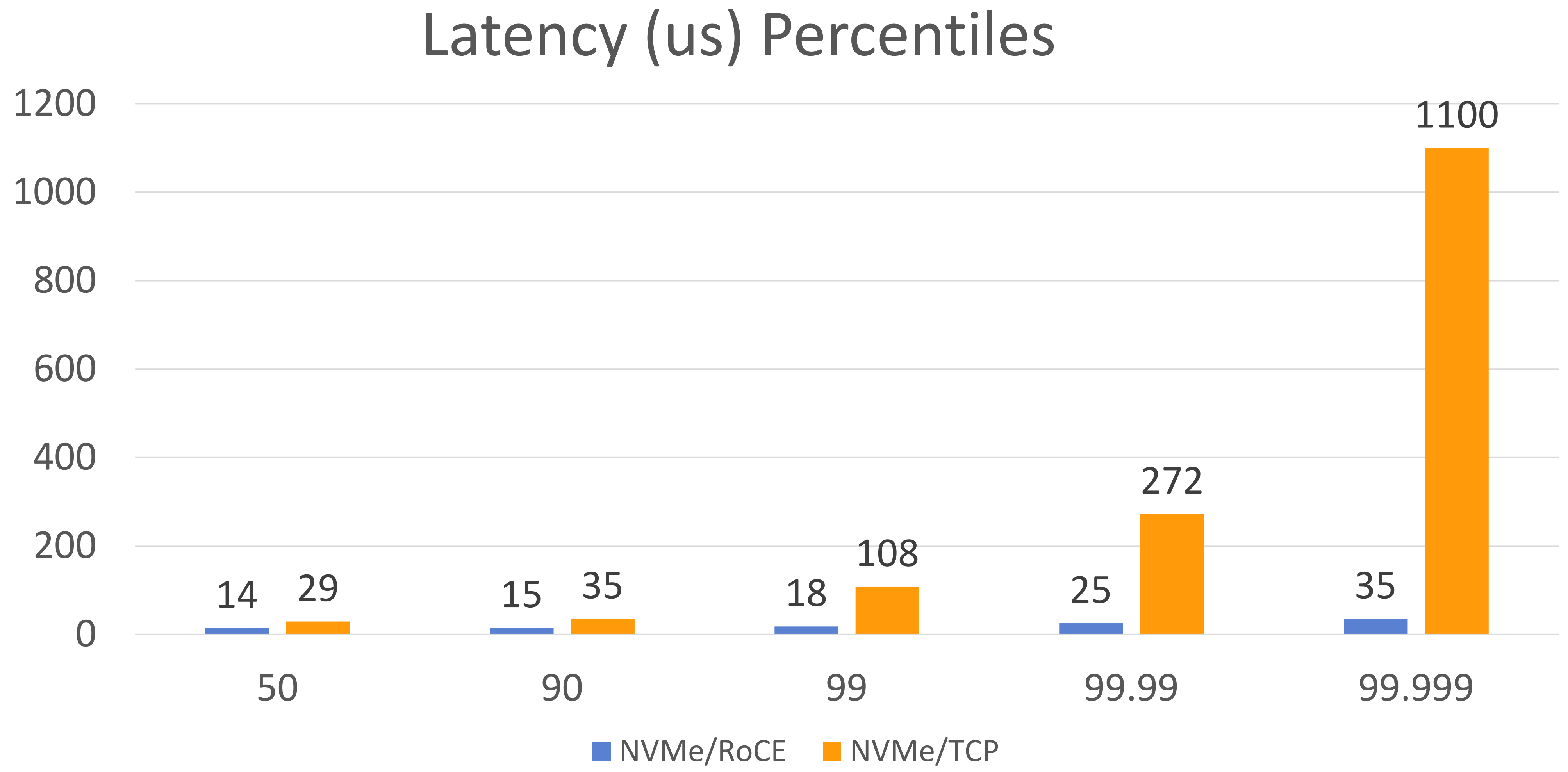
Future

Future Roadmap
Scale-up
Upgrade

Case Study: Fabrics Comparison (partial sample)

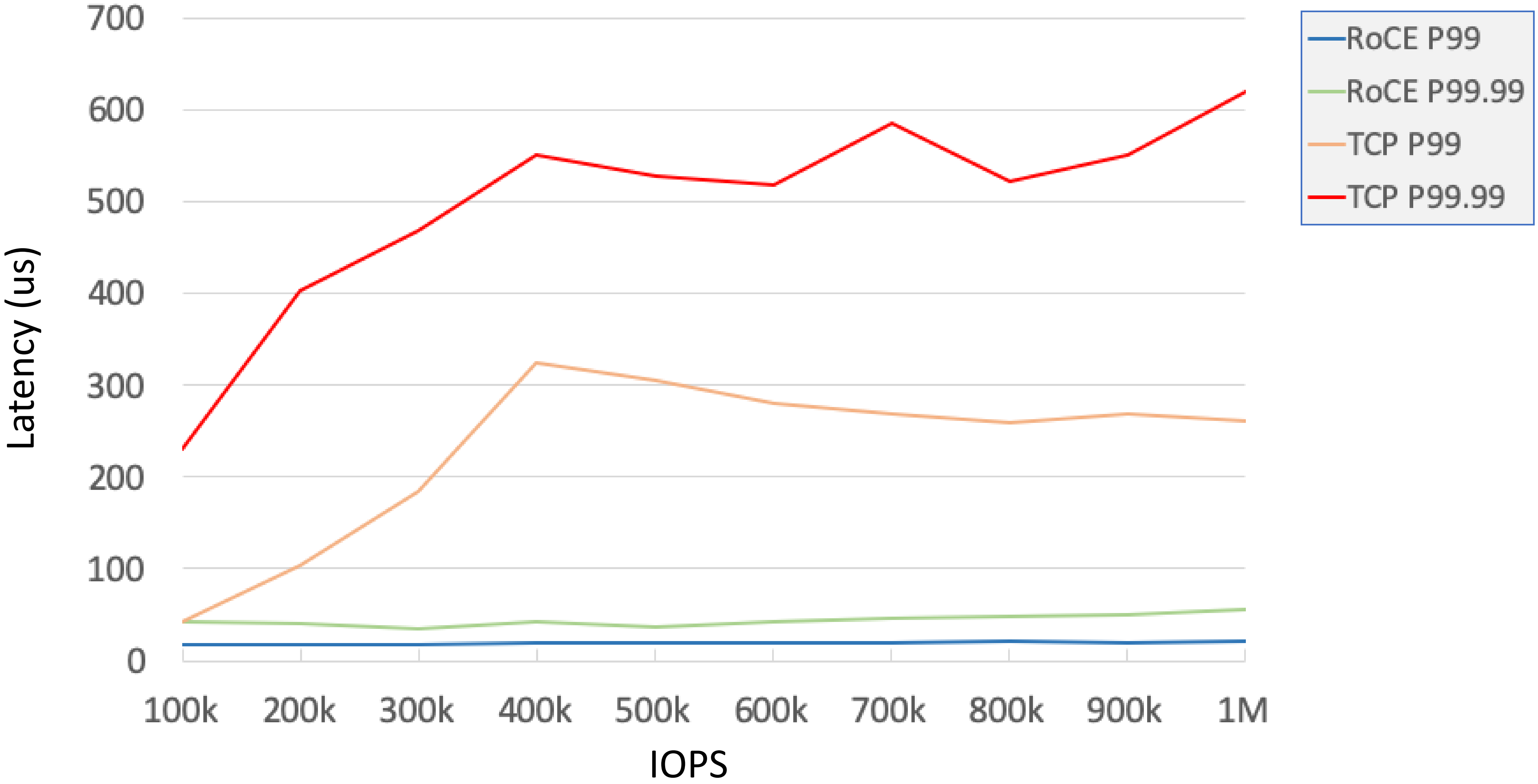
	NVMe/RoCE	NVMe/TCP
Max Speed (current->next gen)	200G → 400G	200G → 400G
Link Aggregation	Yes. HW based	Yes. HW based
½ Round Trip Transport Latency	1.4us	8-30us
4k Write Latency (50th Percentile)	14us	31us
4k Write Latency – Tail/QoS (99.99th percentile)	25us	272us
Encapsulation	UDP	TCP
Routability	Routable UDP based	Routable TCP based
Scale	Multi Rack	Multi Rack
Convergence with other traffic	Yes	Yes
Switch ASIC (Merchant Silicon)	Yes	Yes
Disaggregated Switches	Yes	Yes
SDN	Yes	Yes

Latency Comparison



Source: Western Digital Performance Tests

Latency vs. IOPS



Source: Western Digital Performance Tests

Test Setup

- Linux kernel 5.0
- Mellanox ConnectX-5 NIC
- Mellanox 2700 32x100G switch
- Intel® Xeon® Gold 6150 CPU @ 2.70GHz
- 100G RAM disk

Summary

- NVMe/RoCE and NVMe/TCP are complimentary technologies
- RoCE has lower and more consistent latency
- RoCE needs DCB
- RoCE uses less CPU cycles
- TCP does not need DCB
- TCP appears less optimized for performance and efficiency
- No “One Size Fits All”

Western Digital®

Western Digital and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. Intel and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. The NVMe and NVMe-oF word marks are trademarks of NVM Express, Inc. All other marks are the property of their respective owners.