SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE
Innovation in Storage Products, Services, and Solutions
June 13-15, 2016 | Marriott San Mateo | San Mateo, CA

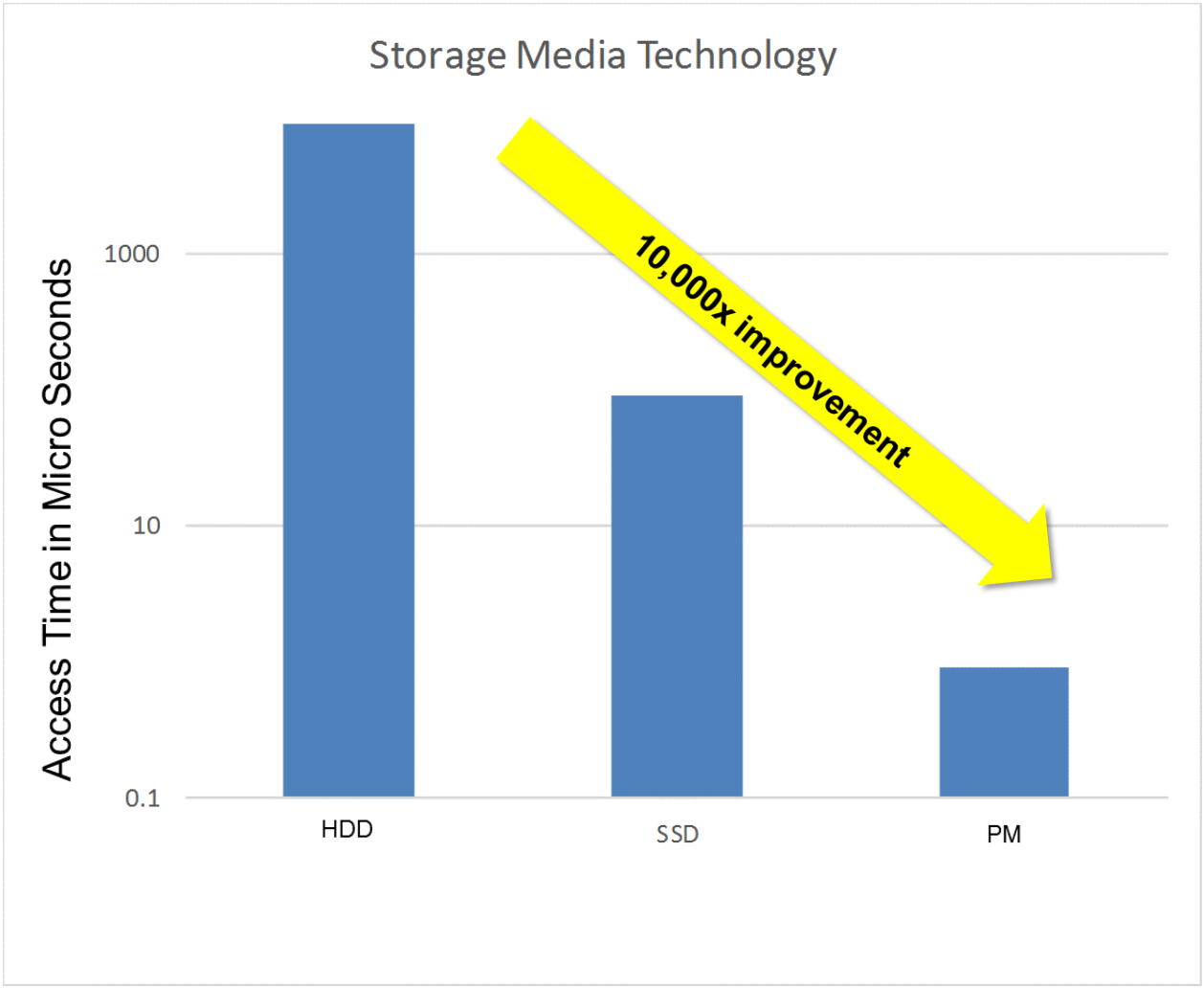
Demonstrating NVMe Over Fabrics Performance on Ethernet Fabrics

Rob Davis

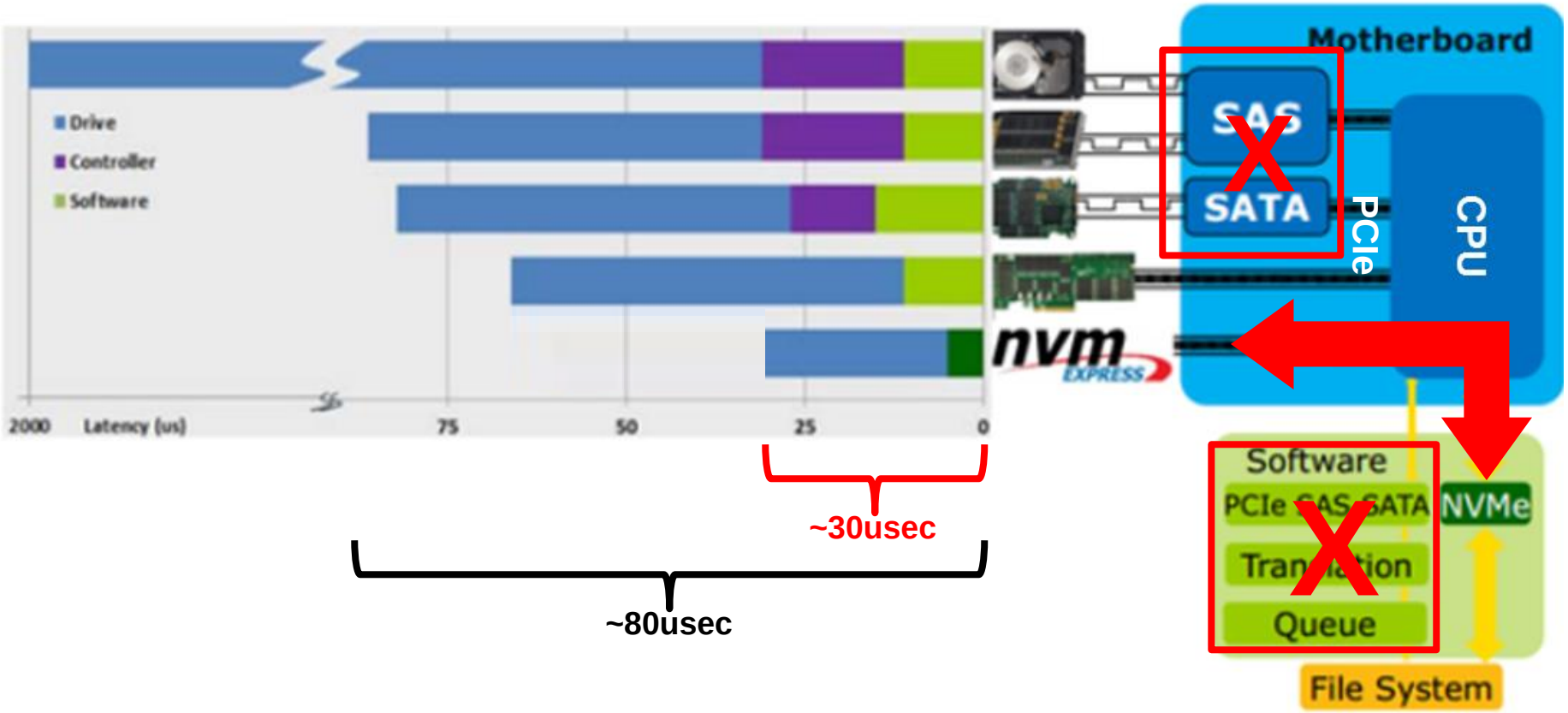
VP of Storage Technology

Mellanox

Why NVMe and NVMe over Fabrics(NVMf)

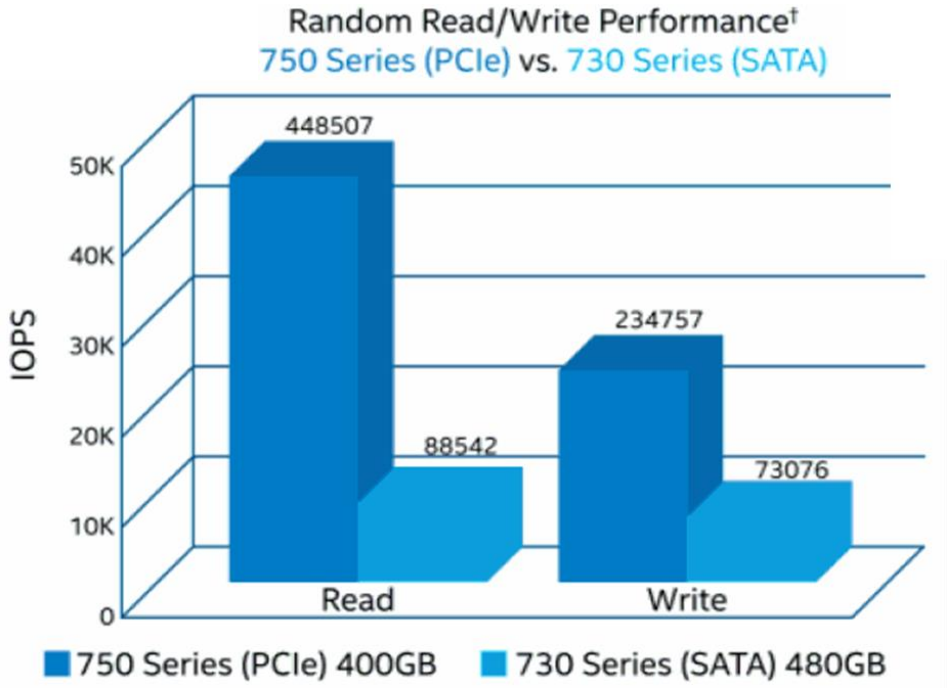


What Makes NVMe Faster



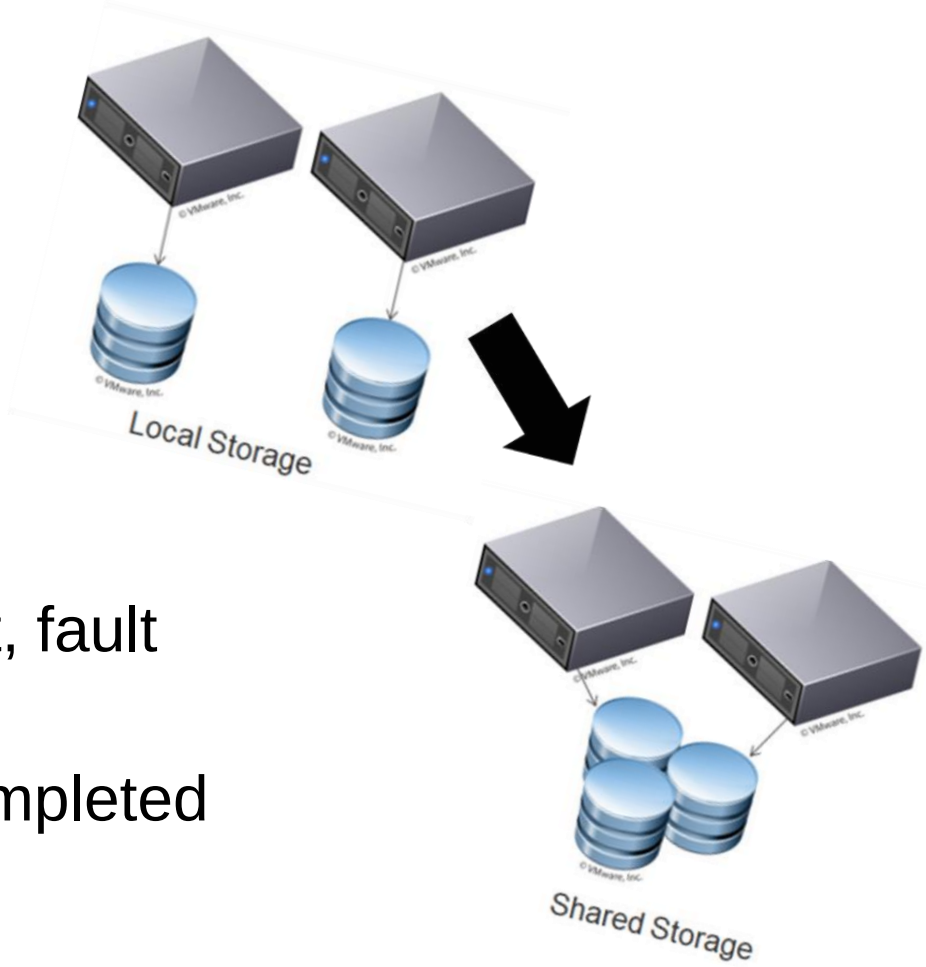
NVMe Performance

- NVMe flash outperforms SAS/SATA flash
  - 2x-4x more bandwidth, 50-60% lower latency, Up to 5x more IOPS
- NVMe is optimized for flash and next-gen persistent memory(PM)
  - Traditional SCSI interfaces designed for spinning disk
  - NVMe bypasses unneeded layers



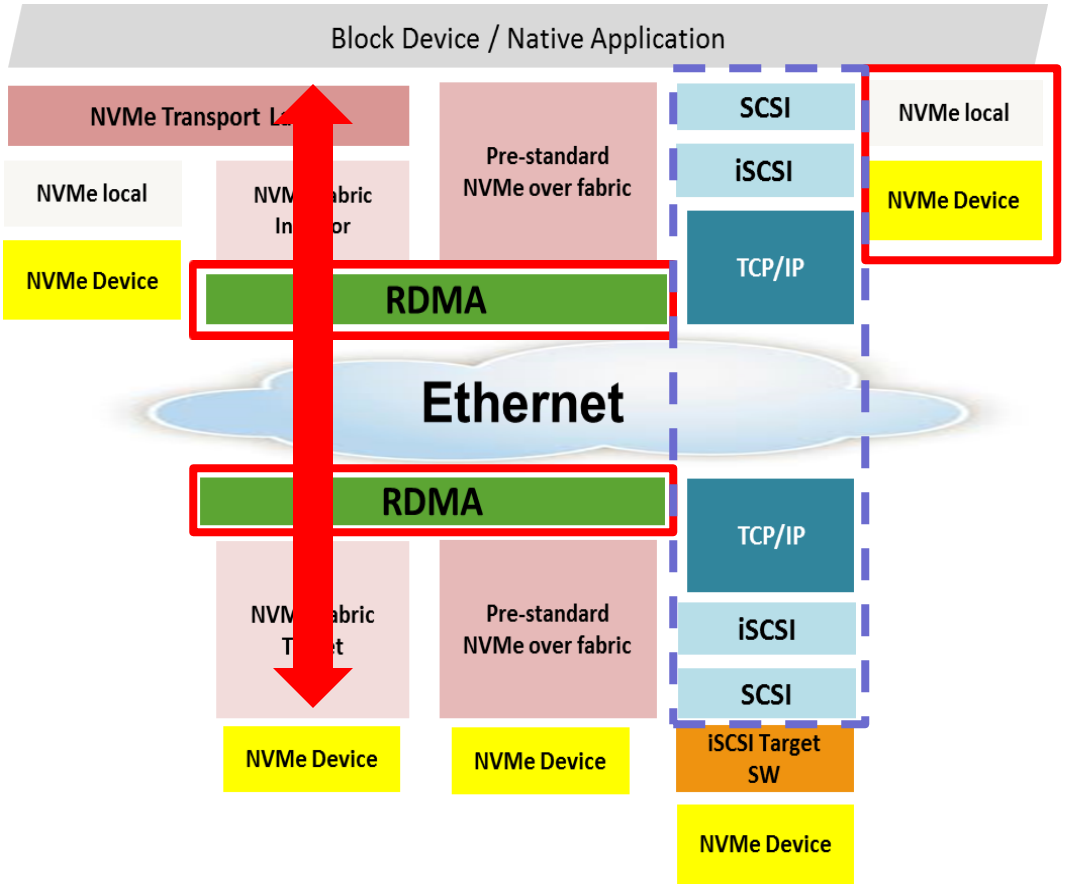
NVMf is the Logical and Historical next step

- Sharing NVMe based storage across multiple servers/CPUs
  - Better utilization: capacity, rack space, power
  - Scalability, management, fault isolation
- NVMf Standard 1.0 was completed in early June

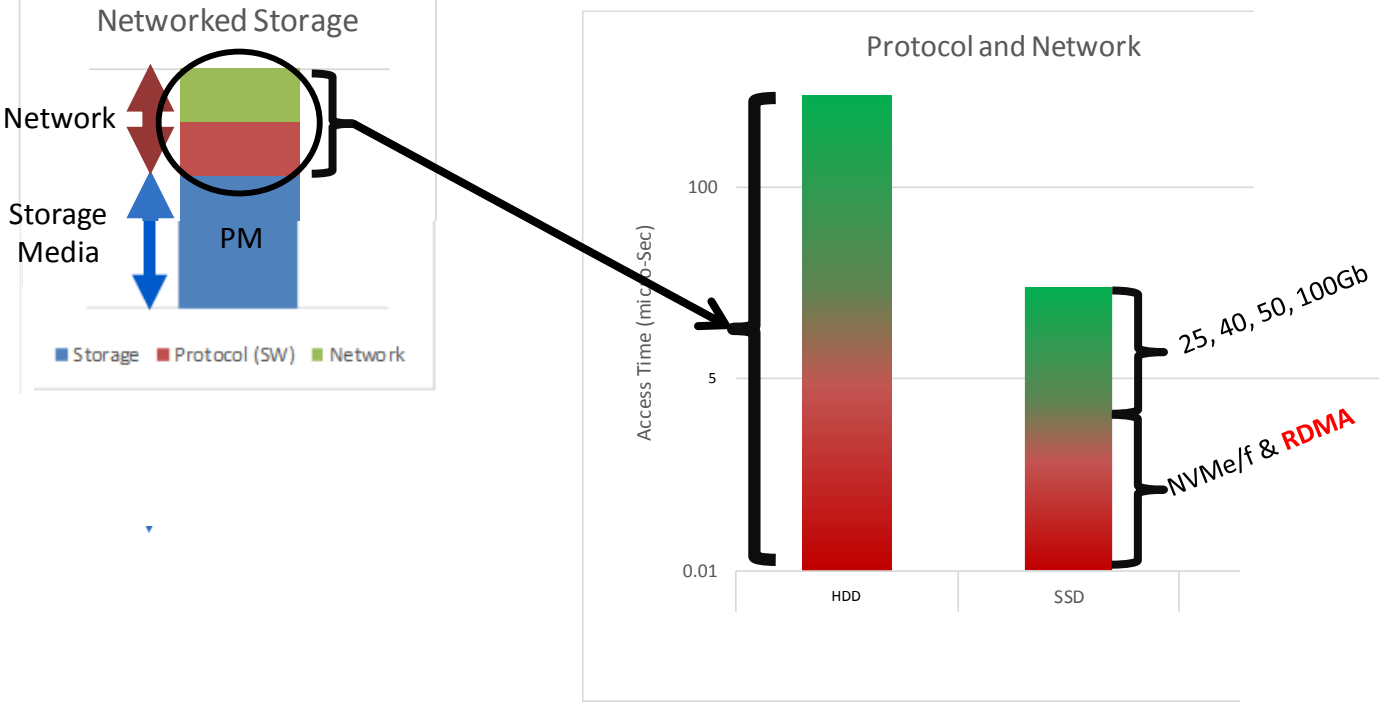


How Does NVMf Maintain Performance

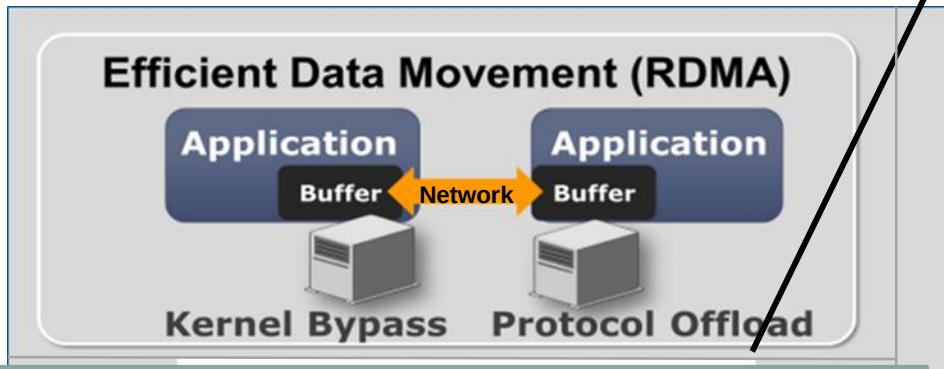
- The idea is to extend the efficiency of the local NVMe interface over a fabric
  - Ethernet or IB
  - NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
  - Bypassing TCP/IP

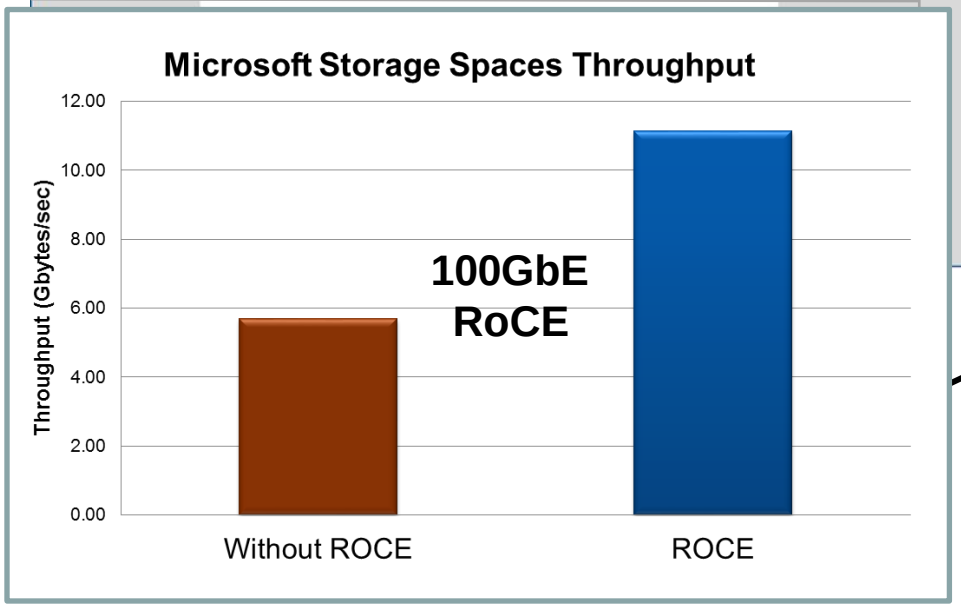


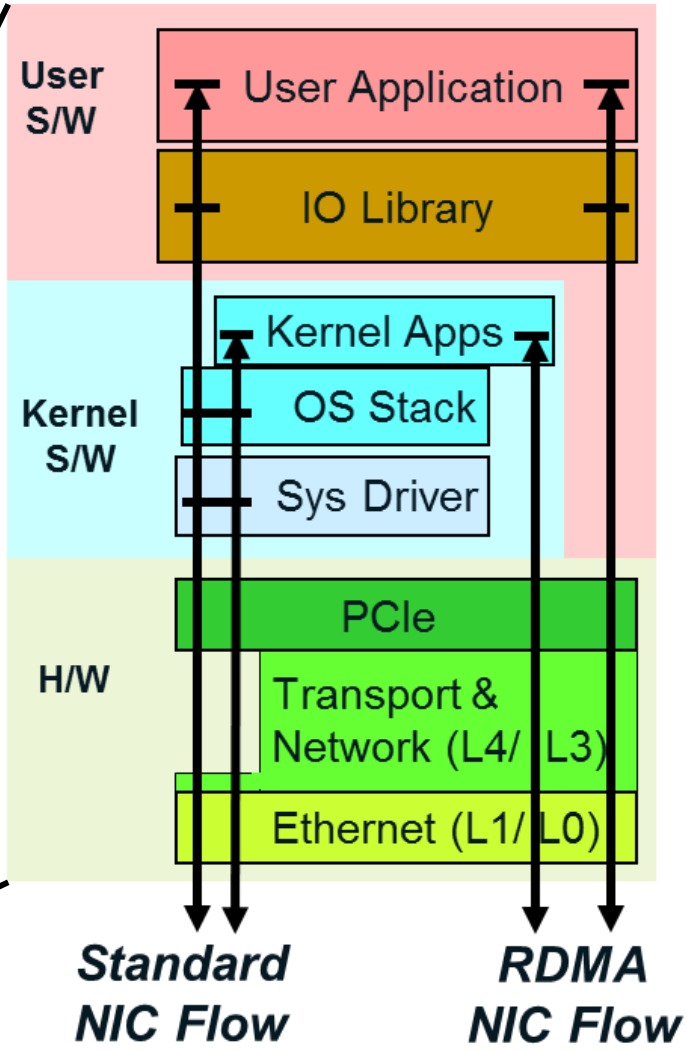
Why not Traditional TCP/IP Network Stack



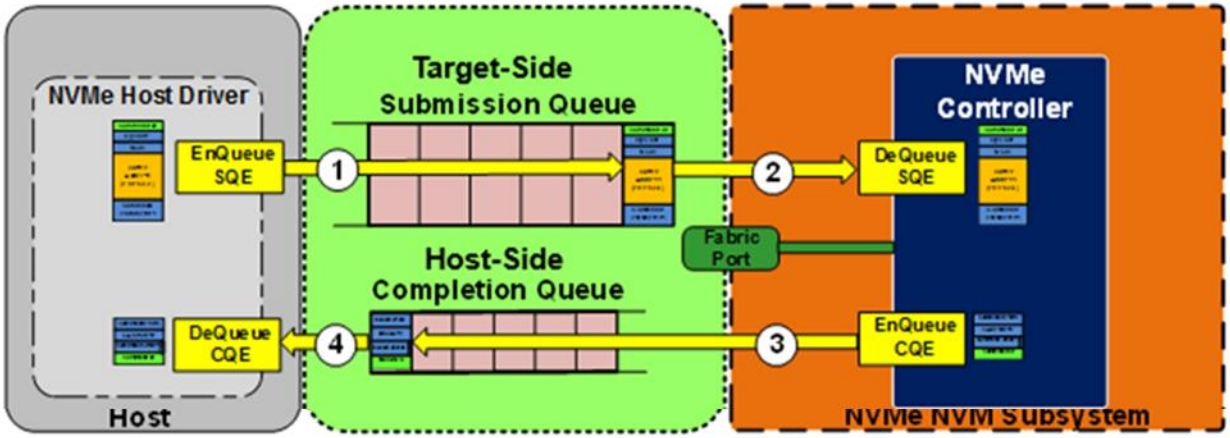
What is RDMA

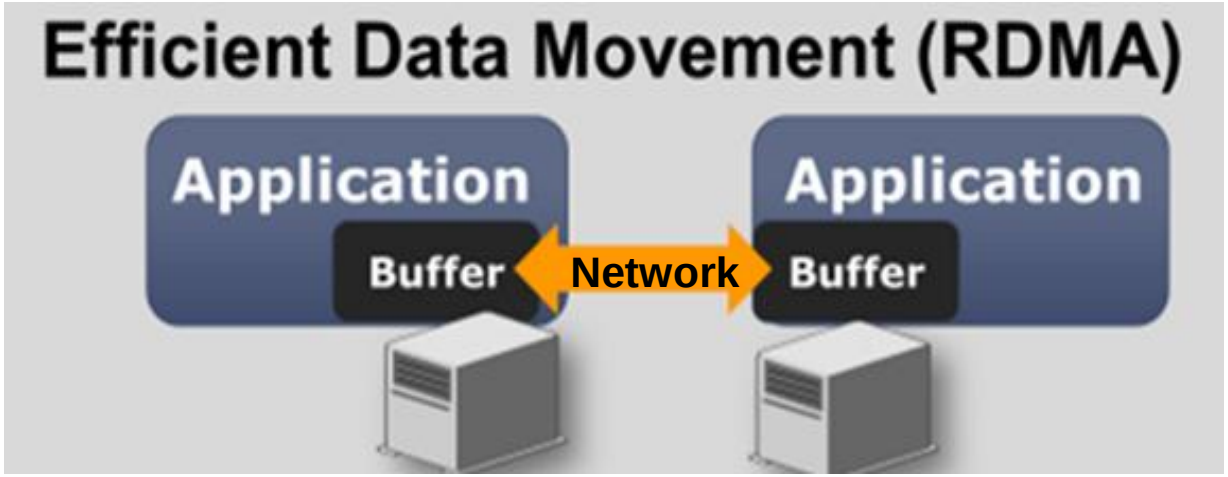


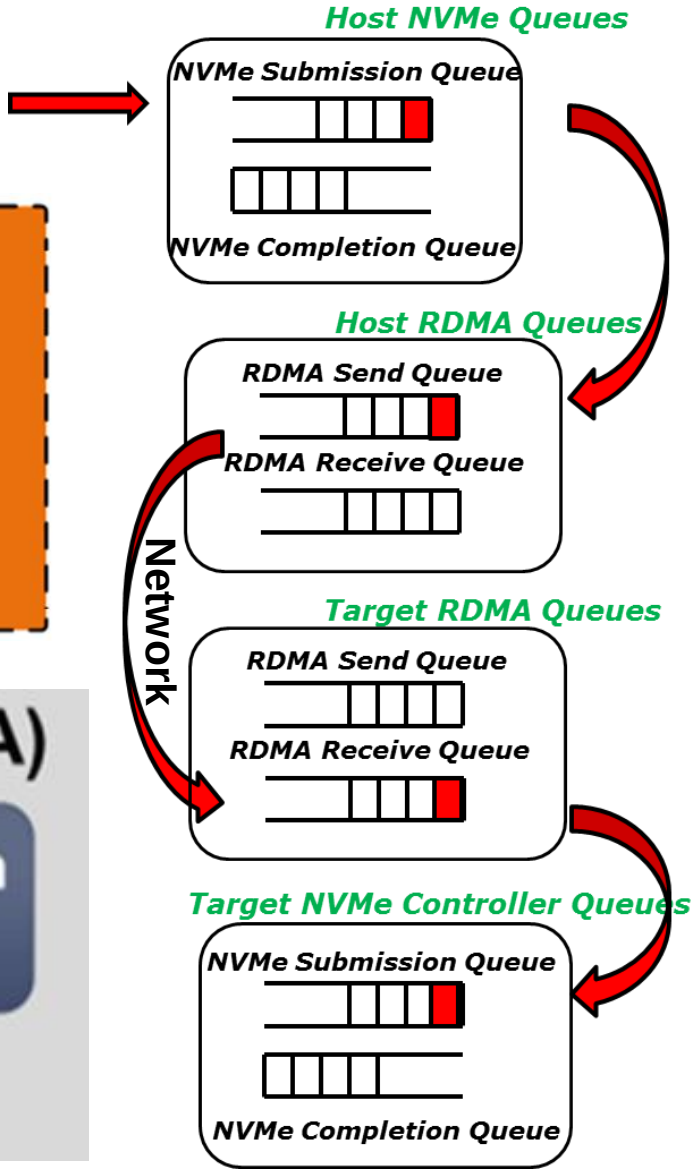




RDMA & NVMe: A Perfect Fit

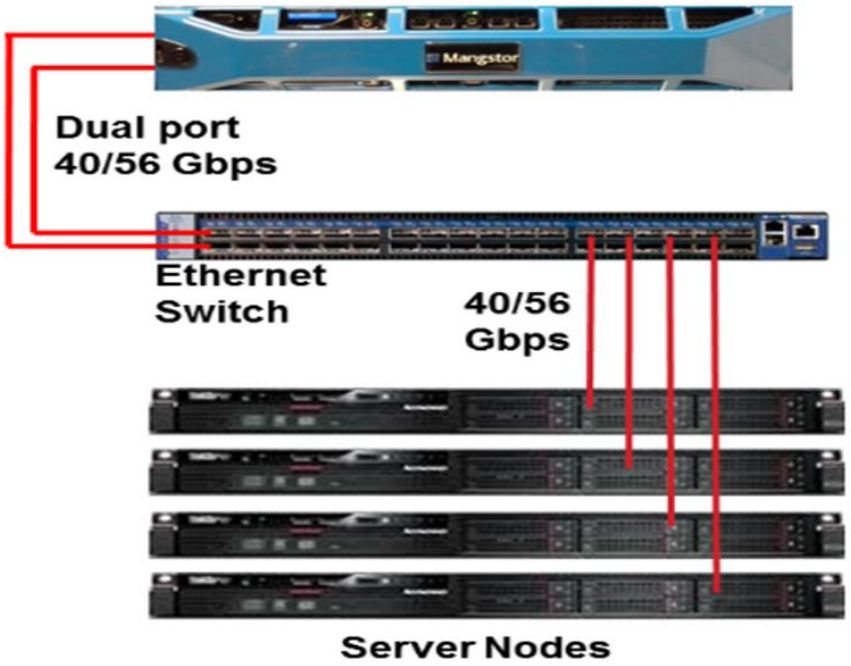


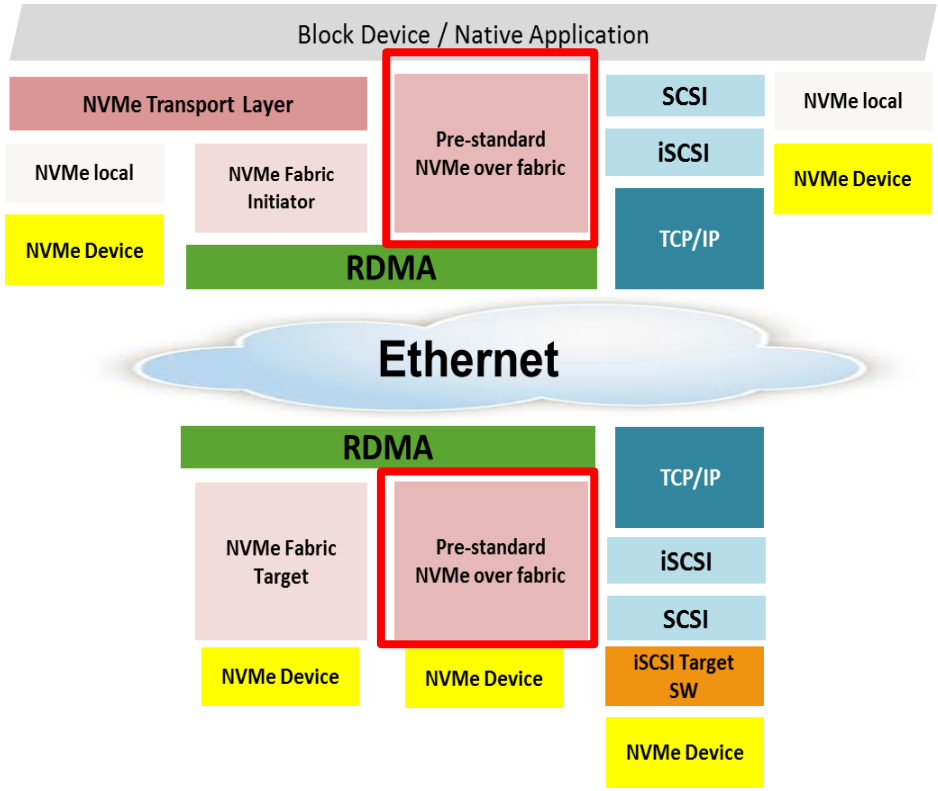




Early Pre-standard Demonstrations

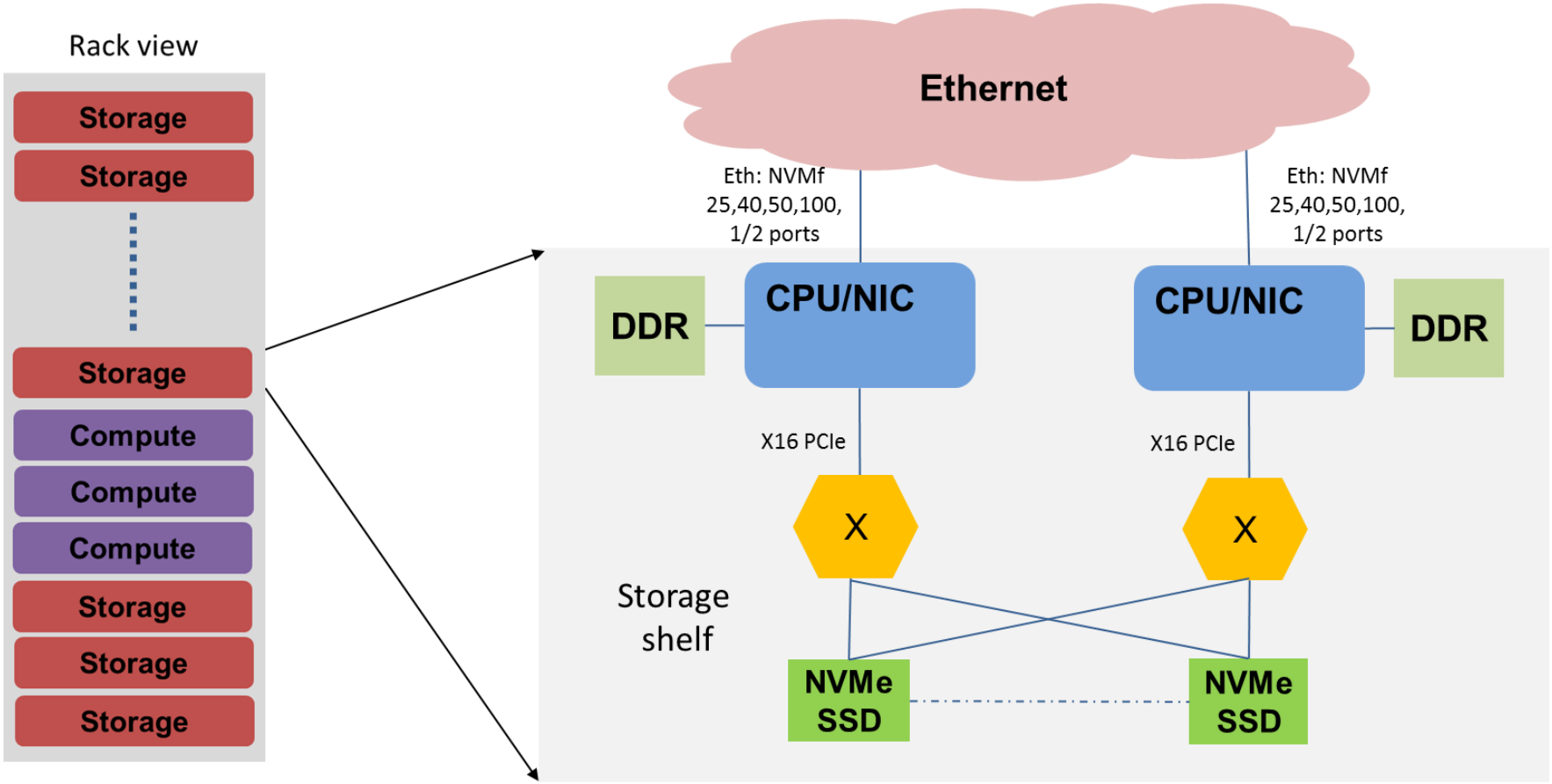
- April 2015
  - NAB Las Vegas



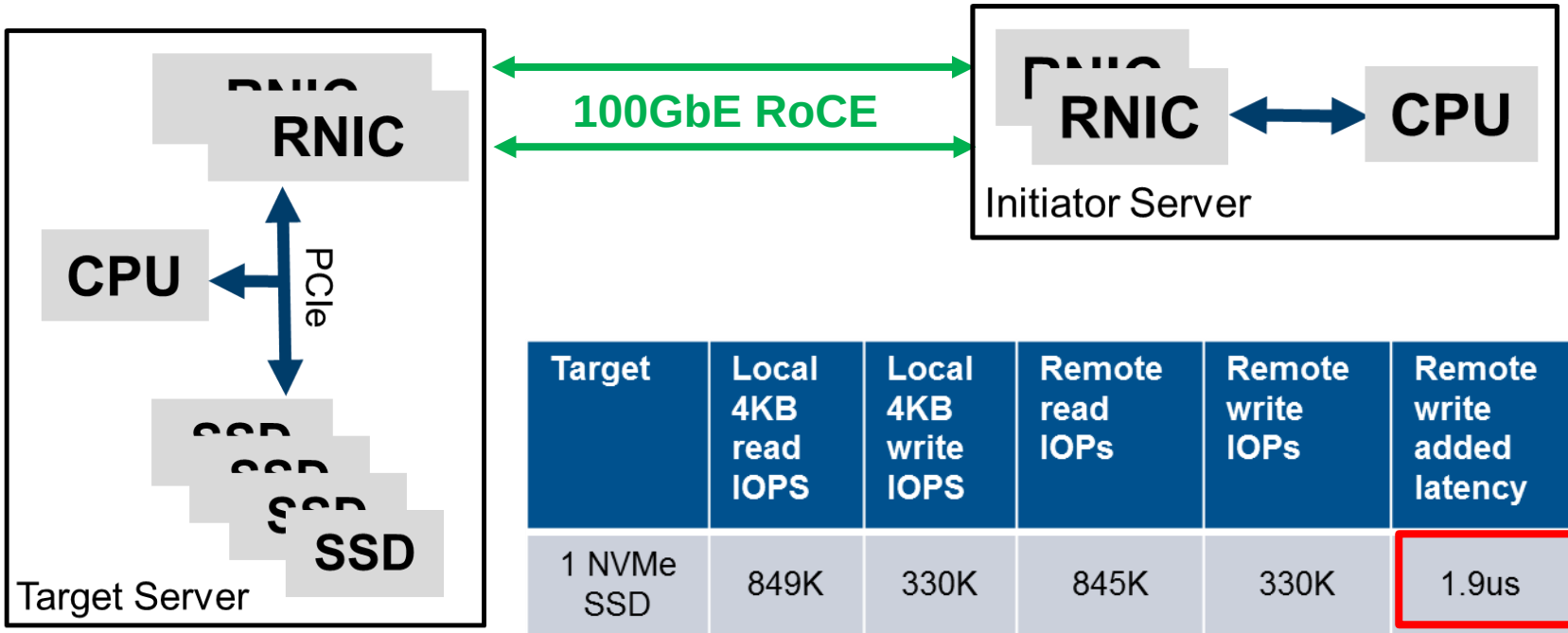


- 10Gb/s Reads, 8Gb/s Writes
- 2.5M Random Read 4 KB IOPs
- Latency ~8usec over local

Compute/Storage Disaggregation

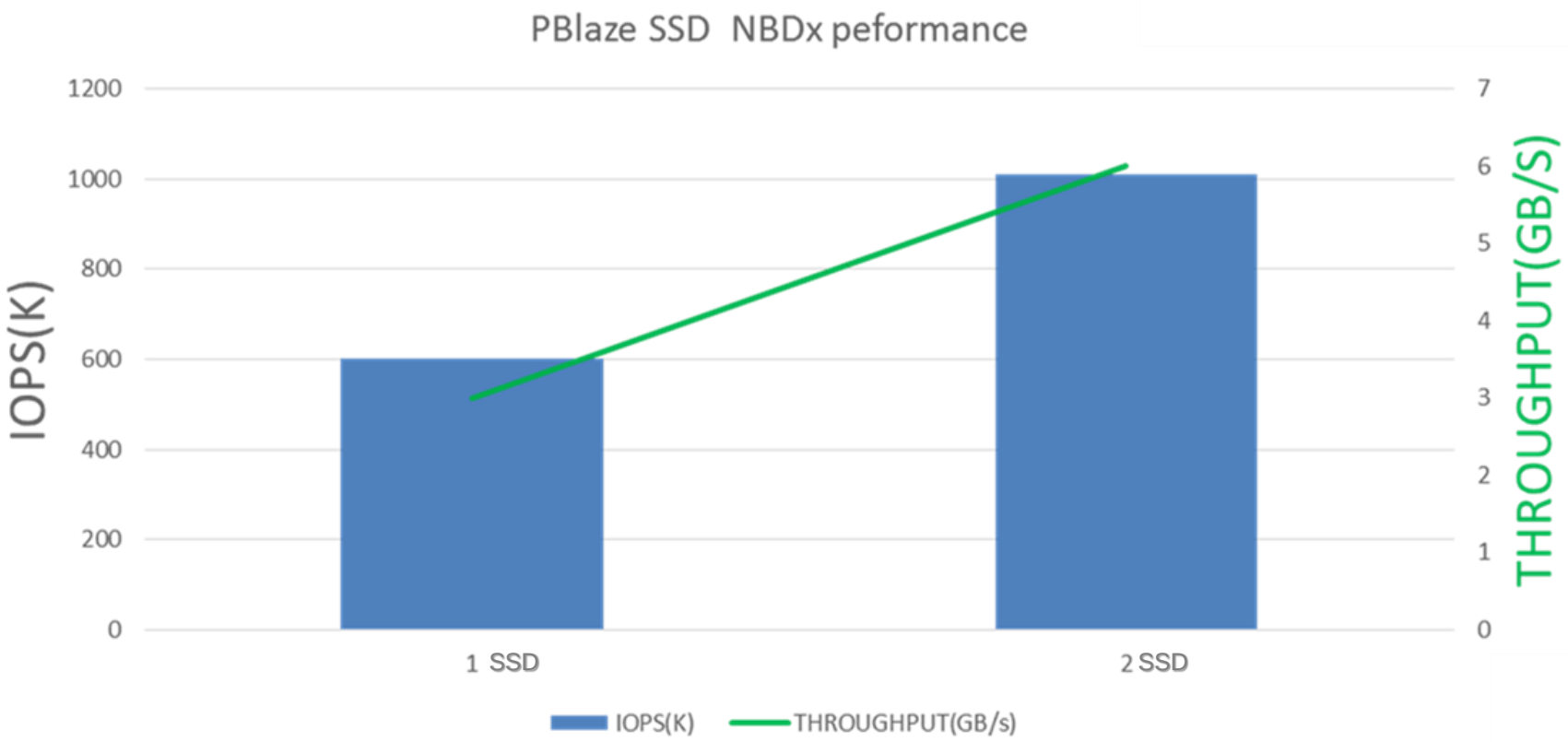


Micron FMS 2015 Demonstration



Target	Local 4KB read IOPS	Local 4KB write IOPS	Remote read IOPs	Remote write IOPs	Remote write added latency	Remote read added latency
1 NVMe SSD	849K	330K	845K	330K	1.9us	4.76us
4 NVMe SSDs	3406K	1333K	3388K	1332K	N/A	N/A

MemBlaze Demonstrations – 40GbE/RoCE



Pre-standard Drivers Converge to V1.0

Demo	NVMe Hardware	Software / Drivers	Network
Mangstor	Mangstor	Mangstor NVMeoF	RoCE or IB
PMC Sierra	PMC	PMC NVMeoF	40Gb RoCE
HGST	HGST	HGST NVMeoF	56Gb InfiniBand
Micron	Micron	Mellanox NBDx	100Gb RoCE
Memblaze	Memblaze	Mellanox NBDx	40Gb RoCE
Samsung at FMS15	Samsung	iSER / Ceph / SMB Direct	40Gb RoCE
Intel at IDF14	Intel	Intel/Chelsio NVMeoF	40Gb iWARP
Stealth startups	Any / Intel NVMe	Startup's NVMeoF	40Gb RoCE

NVMf Standard 1.0 Community Open Source Driver Development

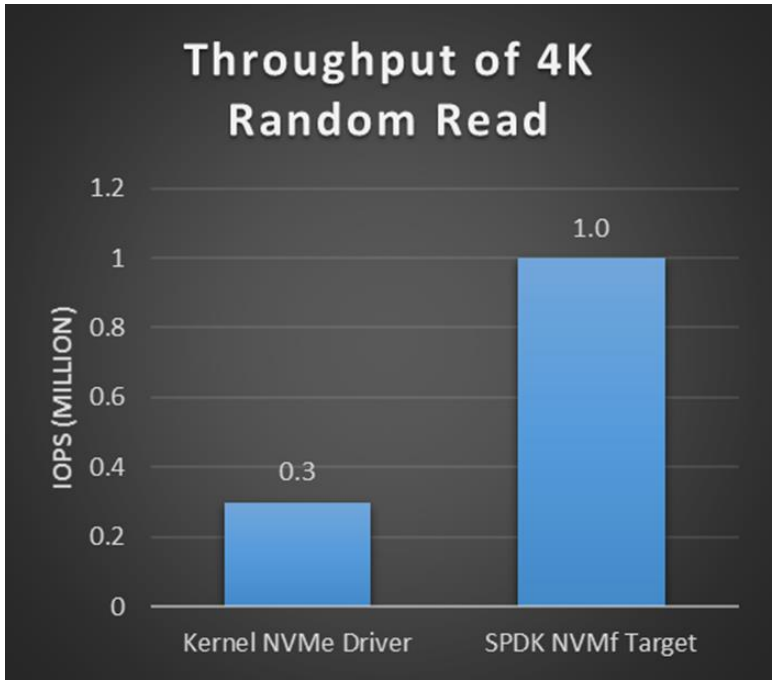
- Mellanox
- Intel
- HGST
- EMC
- Apeiron Data Systems
- Broadcom Corporation
- Chelsio Communications, Inc
- Excelero
- Hewlett Packard Enterprise
- Kazan Networks
- Kenneth Okin Consulting
- Mangstor
- NetApp
- Oracle America Inc.
- PMC
- Qlogic Corporation
- Samsung
- SK hynix Inc.

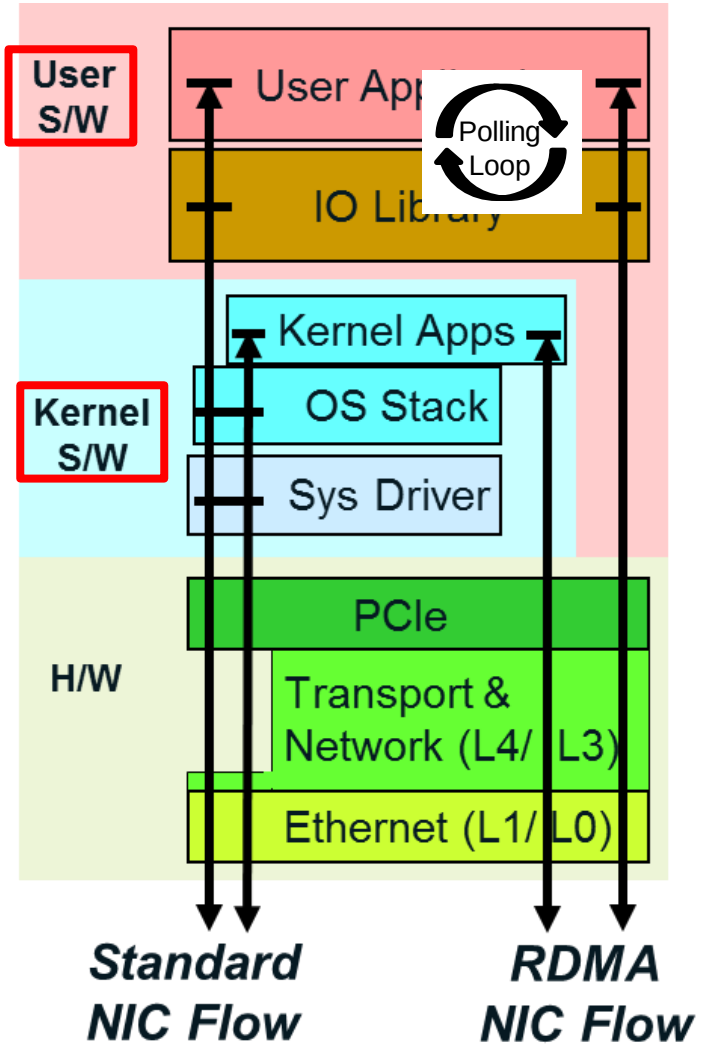
Early Community Driver Performance

- Topology –
  - Two compute nodes
    - ConnectX4 25GbE RoCE
  - One storage node
    - ConnectX4-LX 50GbE RoCE
    - 4 X Intel NVMe device (P3700/750 series)
  - Nodes connected through switch
- BS = 4k, 16 jobs, IO depth = 64
  - 4 cores @ 50% utilization

Bandwidth	IOPS	Added latency
5.2GB/sec	1.3M	~12us

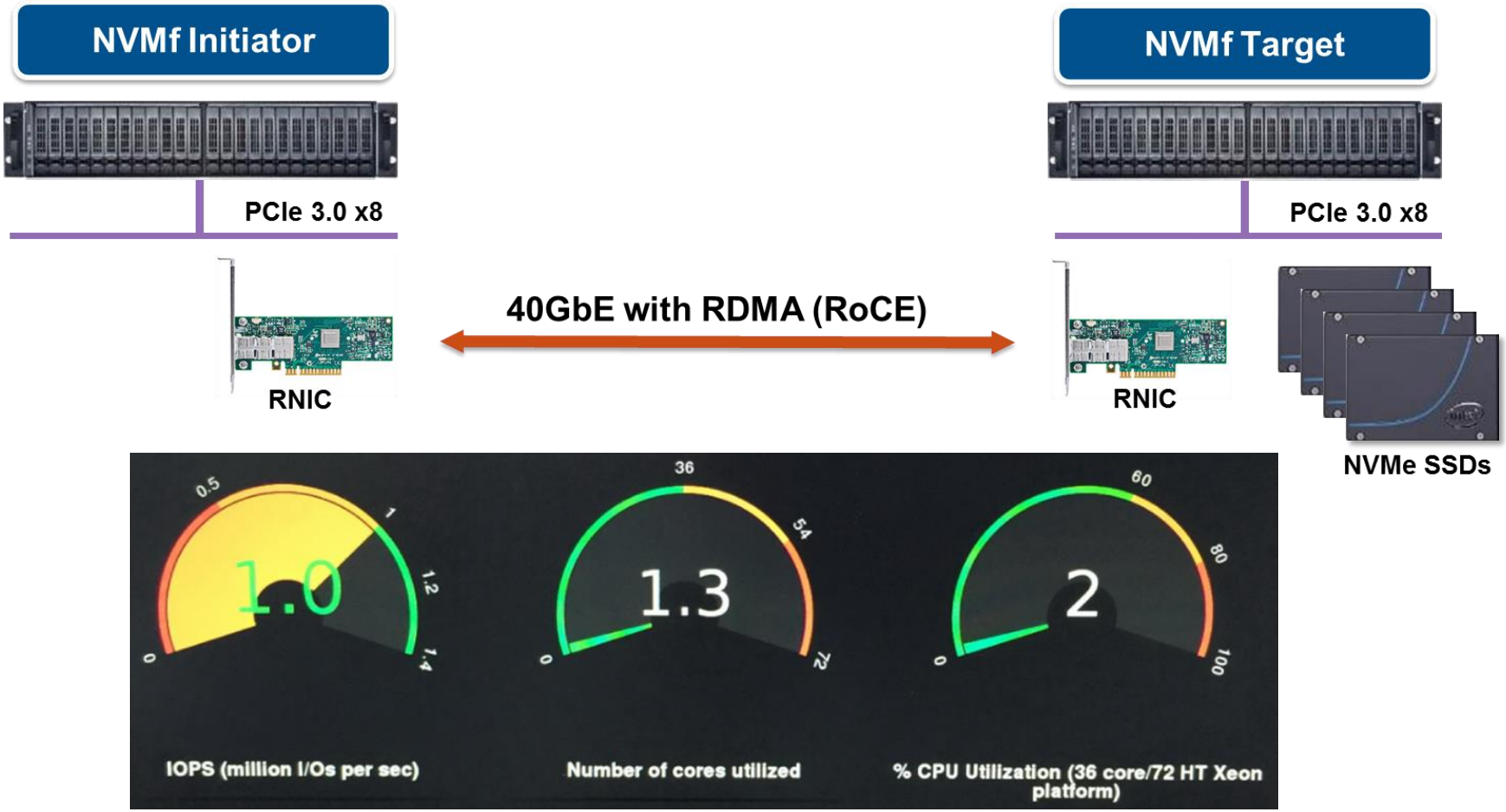
Kernel & User Based NVMf





- Throughput of NVMf with polling user driver can reach ~1.0M IOPS, with only 1 CPU cores utilized

Intel NVMf IDF Shenzhen Demo



Conclusions

- Future Storage solutions will be able to deliver DAS storage performance over a network if:
  - NVMe SSDs – new NVMe protocol eliminates HDD legacy bottlenecks
  - Fast network – “Faster storage needs faster networks!”
  - NVMf with RDMA – new NVMf protocol running over RDMA is within microseconds of DAS

Thanks!

Rob Davis

robd@mellanox.com