NVMe Over Fabrics—High Performance Flash Moves to Ethernet

PCIe Storage Track

John F. Kim, Director of Storage Marketing at Mellanox Technologies

Traditional Ways to Connect Flash

- Inside the Server
  - SCSI (SATA, SAS)
  - PCIe (proprietary standards)
  - USB

- Outside the Server
  - DAS: SAS, SATA, USB, InfiniBand
  - Fibre Channel SAN
  - FCoE
  - Ethernet TCP (iSCSI, NAS, Object)

New Ways to Connect Flash

- Inside the Server
  - PCIe with standardized NVMe
  - NVDIMM
- Outside the Server
  - DAS: Thunderbolt, USB 3.0
  - Ethernet RDMA (iSER, SMB Direct)
  - Clustered file system (InfiniBand or Ethernet)
  - Hyper-converged (Ethernet)
  - NVMe over Fabrics

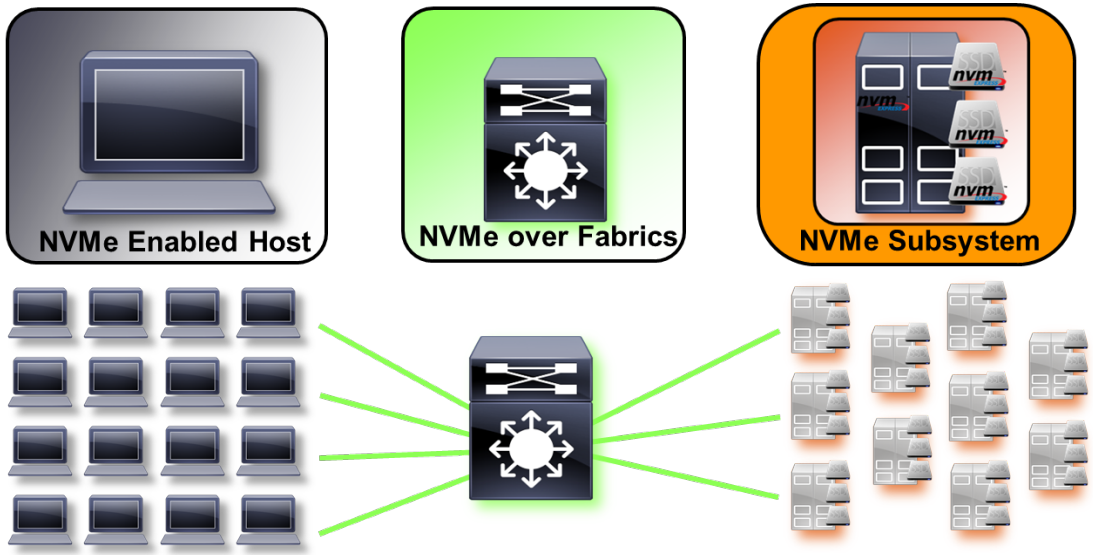
Options for NVMf Network

- InfiniBand
  - Best performance
- Ethernet (RoCE or iWARP)
  - Most widely used
- Fibre Channel (FC-NVMe standard in progress)
  - Most popular network for flash arrays in the enterprise

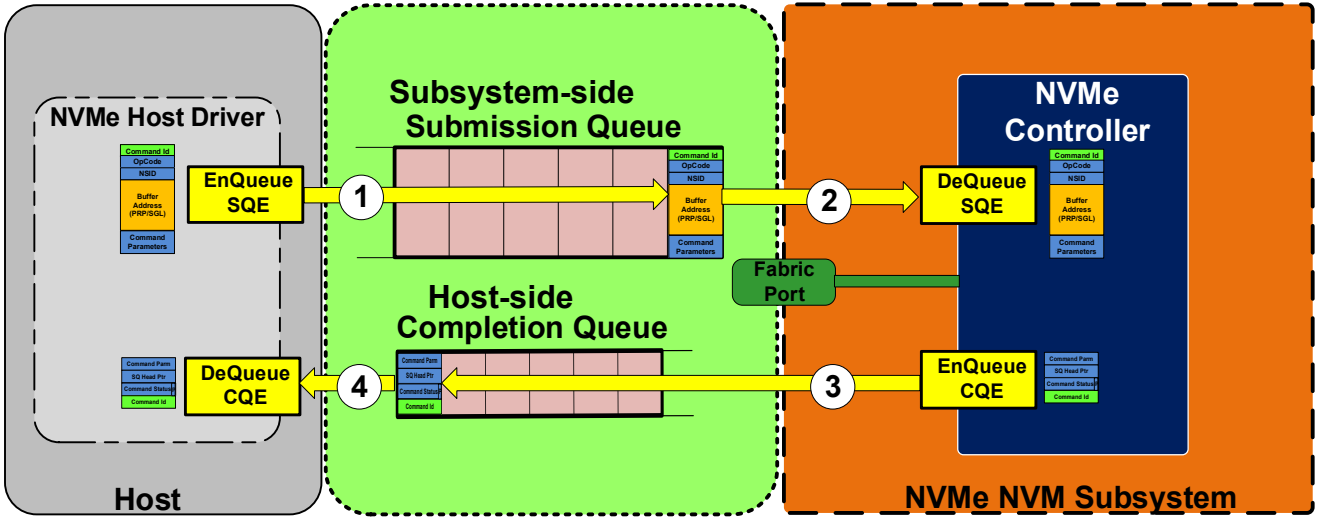
Why NVMe Over Fabrics

End-to-End NVMe semantics across a range of topologies

[diagram from SNIA ESF webcast]



NVMe Queuing Operational Model



- 1. Host Driver enqueues the Submission Queue Entries into the SQ
- 2. NVMe Controller dequeues Submission Queue Entries
- 3. NVMe Controller enqueues Completion Queue Entries into the CQ
- 4. Host Driver dequeues Completion Queue Entries

Why Ethernet for NVMf?

- Ubiquity
- 25/50/100GbE today, 200GbE soon
- RDMA (RoCE and iWARP)
- Scalability to 100’s of thousands nodes
- Standards-based, multi (many)-vendor
- Lowest Price/Port
- Converged Fabric capability
- Standard cables extend to long distance

RDMA Options on Ethernet

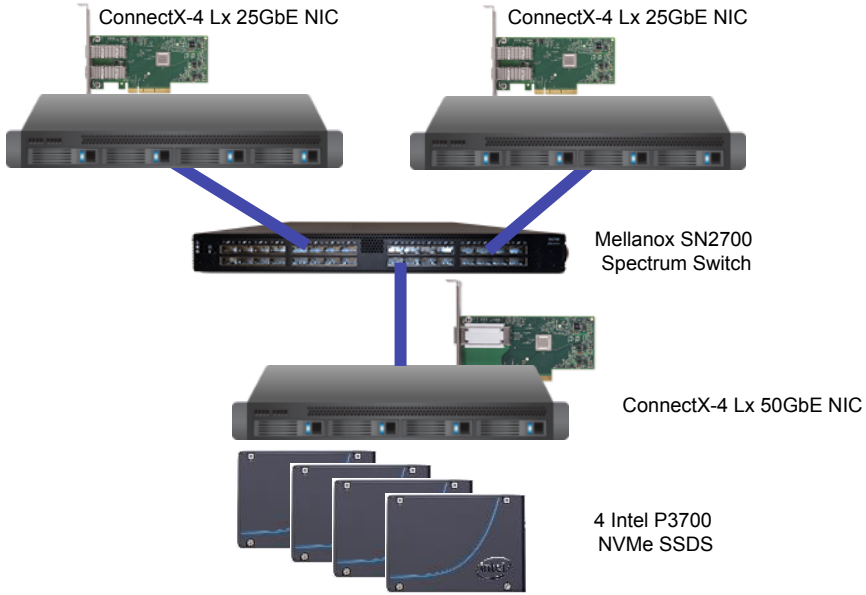
- RoCE uses UDP/IP on Ethernet, no TCP
  - InfiniBand transport on top of Ethernet
- iWARP layers on top of TCP/IP
  - Offloaded TCP/IP flow control and management
  - Might include TCP Offload Engine (TOE)
- Both RoCE and iWARP support RDMA verbs
  - NVMf using verbs can be run on top of either transport
- Other options without RDMA

Who Supports RDMA on Ethernet?

Vendor	RoCE	iWARP	25/50/100
Mellanox	Yes		Yes (GA)
Broadcom (Emulex)	Yes (in beta)		?
Broadcom (NetXtreme)	?		
Chelsio		Yes (GA)	Announced
QLogic	Yes (announced)	Yes (announced)	Yes (in beta)

Mellanox NVMf Demo on RoCE

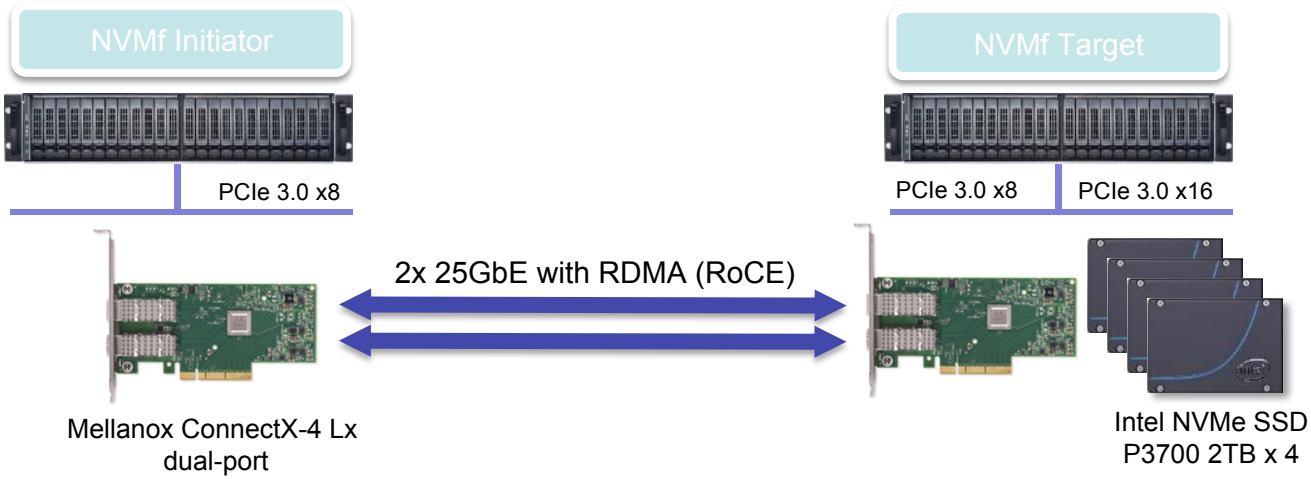
- Topology –
  - Two compute nodes
    - ConnectX4-Lx 25Gbps port
  - One storage node
    - ConnectX4-Lx 50Gbps port
    - 4 X Intel NVMe devices (P3700/750 series)
  - Nodes connected through switch



Added fabric latency
~12us

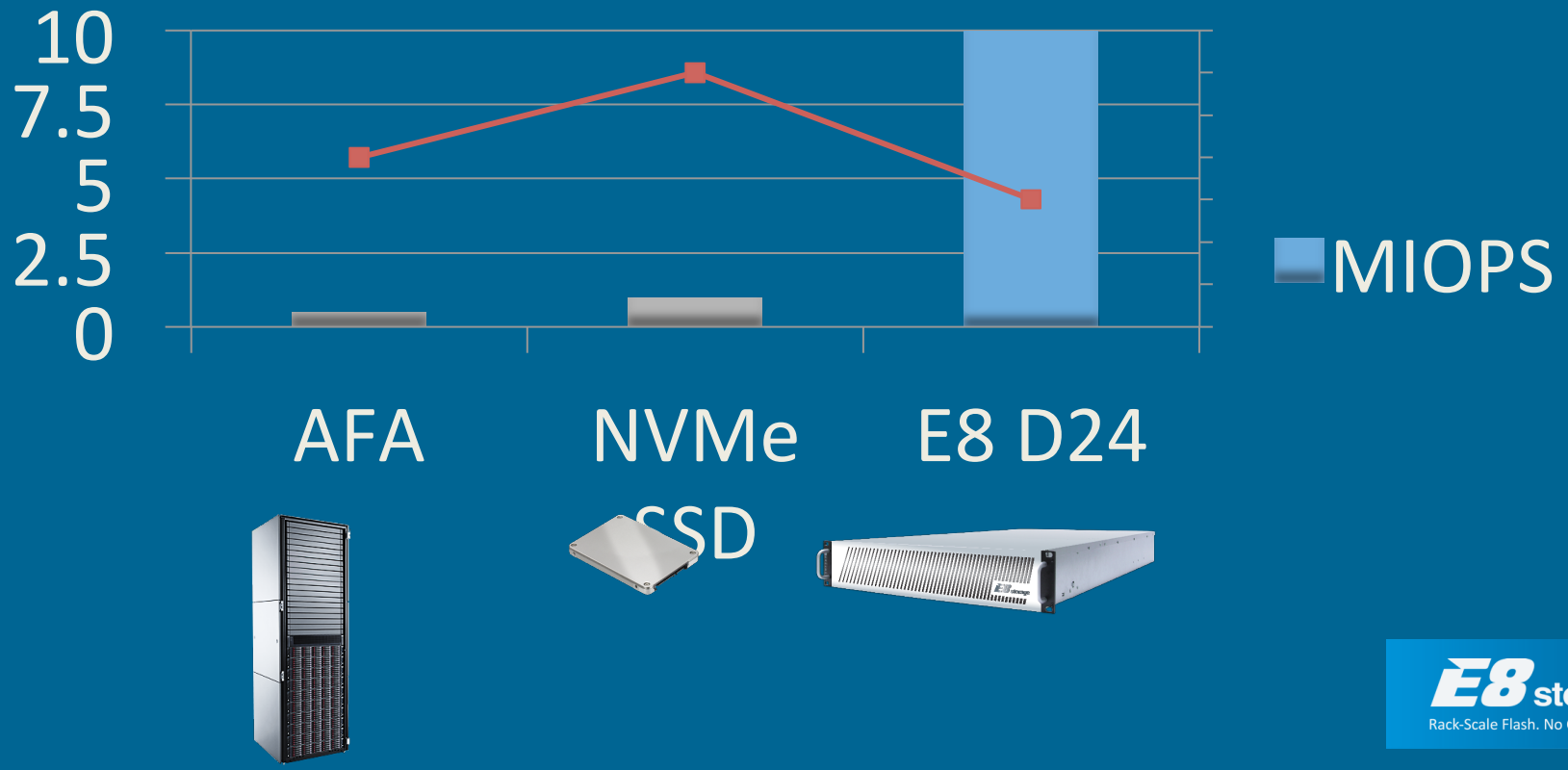
	Bandwidth (Target side)	IOPS (Target side)	Num. Online cores	Each core utilization
BS = 4KB, 16 jobs, IO depth = 64	5.2GB/sec	1.3M	4	50%

Intel SPDK NVMf Demo



- Intel SPDK software with user-space target and polling mode
- Mellanox ConnnectX-4 Lx NICs, running RoCE
- 4 Intel NVMe P3700 SSDs connected to shared PCIe Gen3x16 bus
- Intel Xeon-D 1567 CPU on target side; Intel Xeon E5 2600 V3 on initiator side

E8 Storage: x10 Performance, Half the TCO

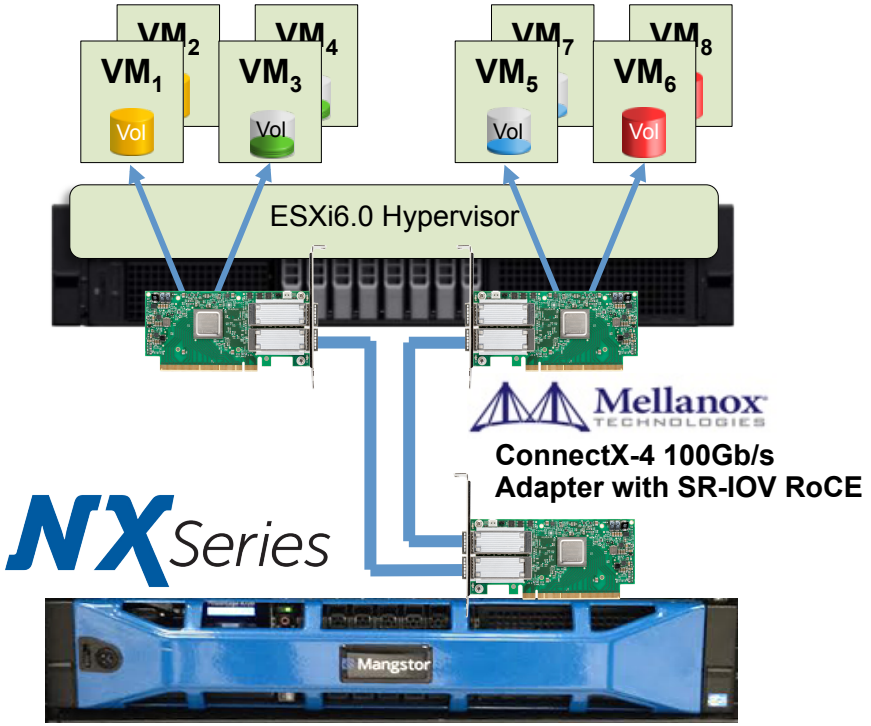


E8 storage
Rack-Scale Flash. No Compromise.

NVMe over 100GbE RoCE Fabric Demo

NVMf in ACTION

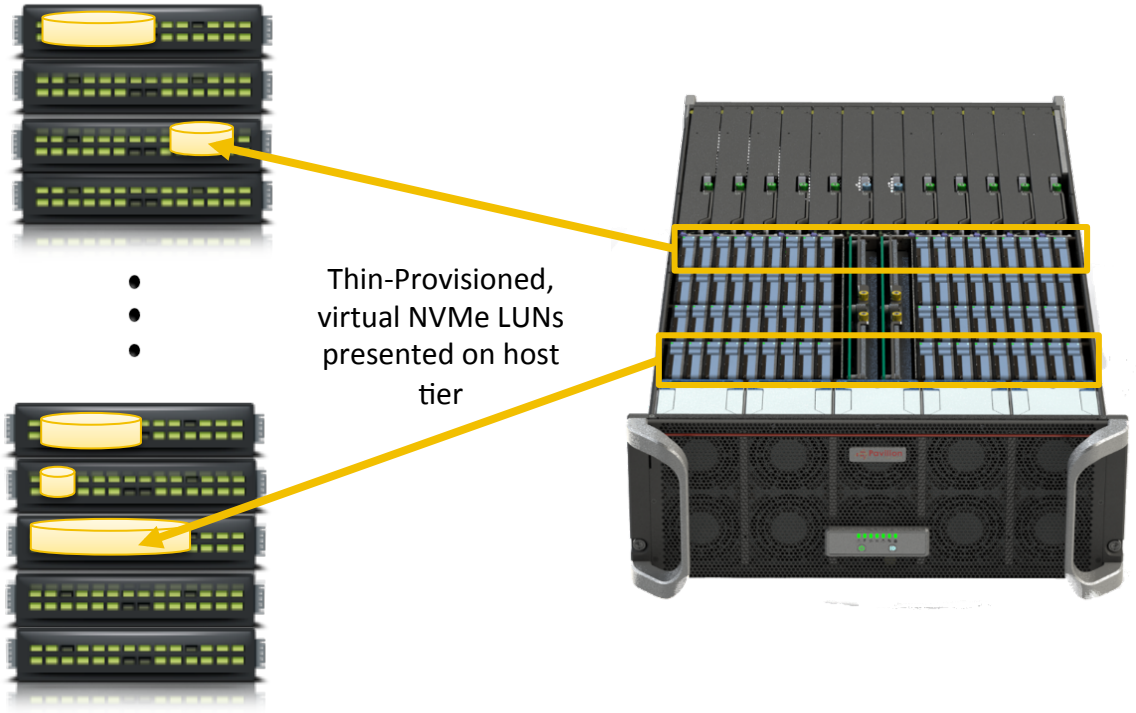
- **12-, 16-, 32-TB Flash Arrays**
- **2.25M, 1.75M Rd/Wr IOPS**
- **10.0GB/s, 6.0GB/s Rr/Wr BW**
- **< 200us Latency**



@FMS16 Hall D

FIO Test Tools Suite illustrates Mangstor's NX6320 High Performance, Low Latency NVMe over Fabric Devices.

Low Cost Of Ownership + Enterprise Reliability



- ✓ No Custom Software or Hardware required in hosts/application tier
- ✓ Standard Ethernet Connectivity
- ✓ Erasure Coded Volumes
- ✓ Thin Provisioned Volumes
- ✓ QOS Policy per Volume

Power AND Density, in an Enterprise-Class Array

120 GB/S Bandwidth

20 Million 4K IOPS

Half PB Capacity (4U)

*1 PB in Q1-2017

Pay-As-You Grow Scalability in 4U Chassis

Zero Host Footprint

Up To 40 x 40 GB/S Ports

High Speed Data Copies/ Clones

Thin Provisioning

Erasure Coding

Hot Swappable Components

NVMf: High Performance Flash Moves to Ethernet

Thank You!