NVMe over Fabrics: Learning from early developments

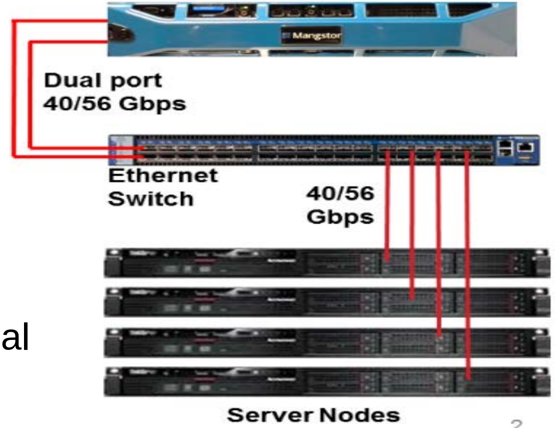
Rob Davis

robd@mellanox.com

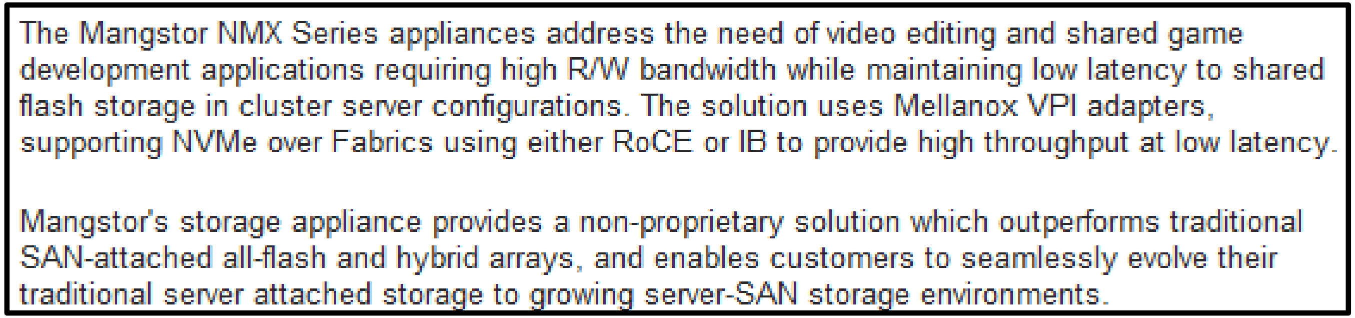
NAB Demo of NVMe over Fabric

NAB Show April 11 - 16, 2015 Las Vegas

10GB/s reads
8GB/s writes
2.5M random read IOPS
(4KB block size)
Max latency 6-8us over local



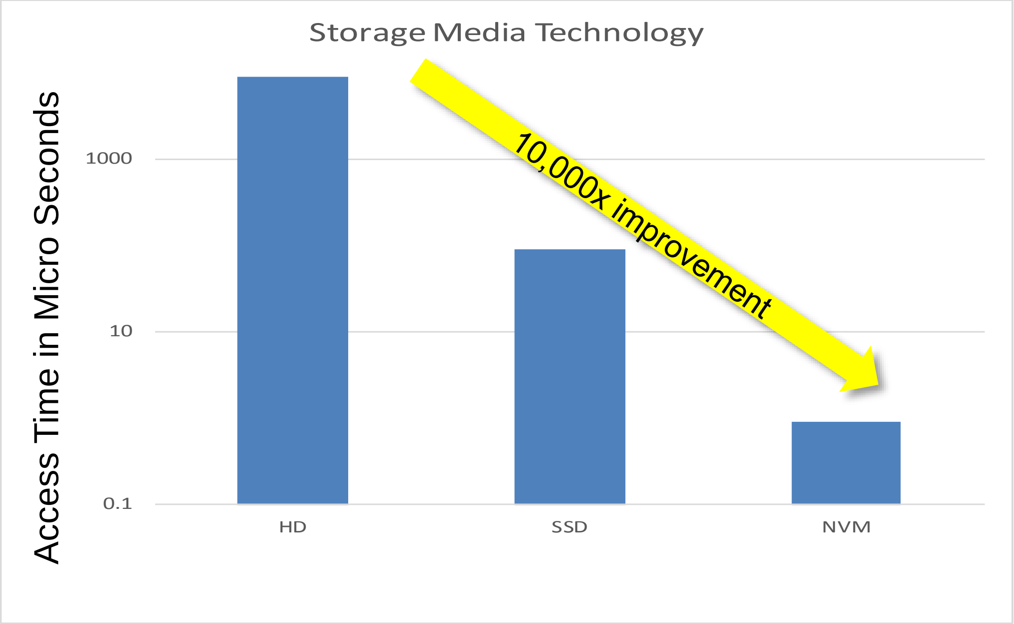
From PR



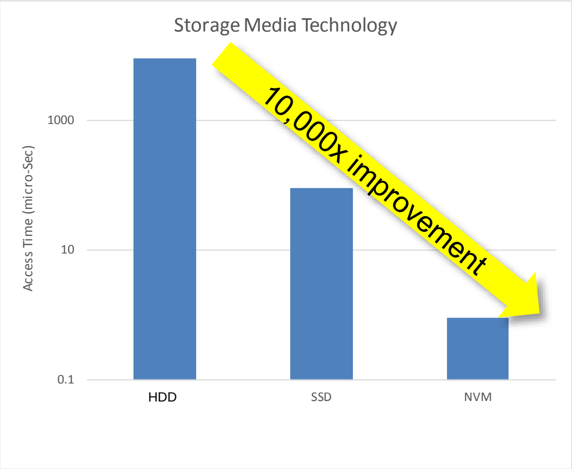
The Mangstor NMX Series appliances address the need of video editing and shared game development applications requiring high R/W bandwidth while maintaining low latency to shared flash storage in cluster server configurations. The solution uses Mellanox VPI adapters, supporting NVMe over Fabrics using either RoCE or IB to provide high throughput at low latency.

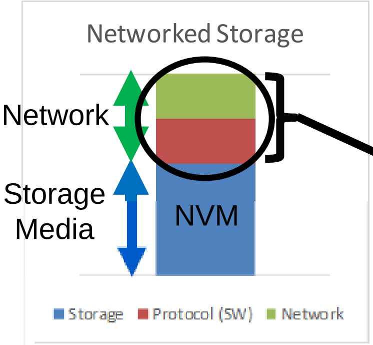
Mangstor's storage appliance provides a non-proprietary solution which outperforms traditional SAN-attached all-flash and hybrid arrays, and enables customers to seamlessly evolve their traditional server attached storage to growing server-SAN storage environments.

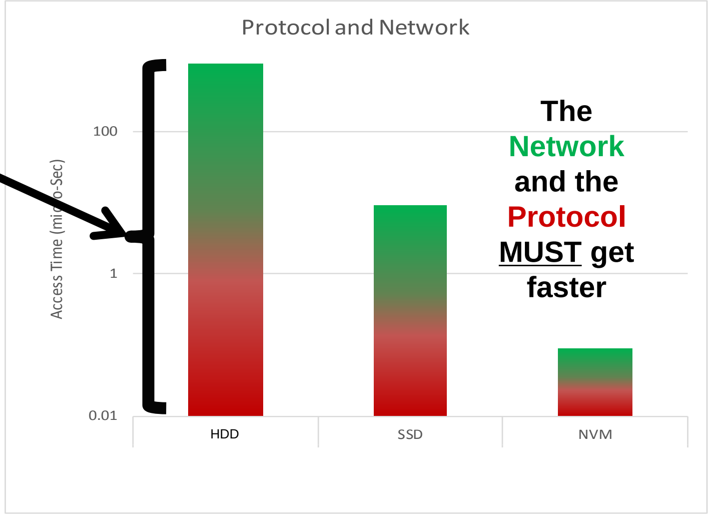
Storage Market Evolution to Non-Volatile Memory(NVM)



Flash Performance Creates Bottleneck at Network Layer

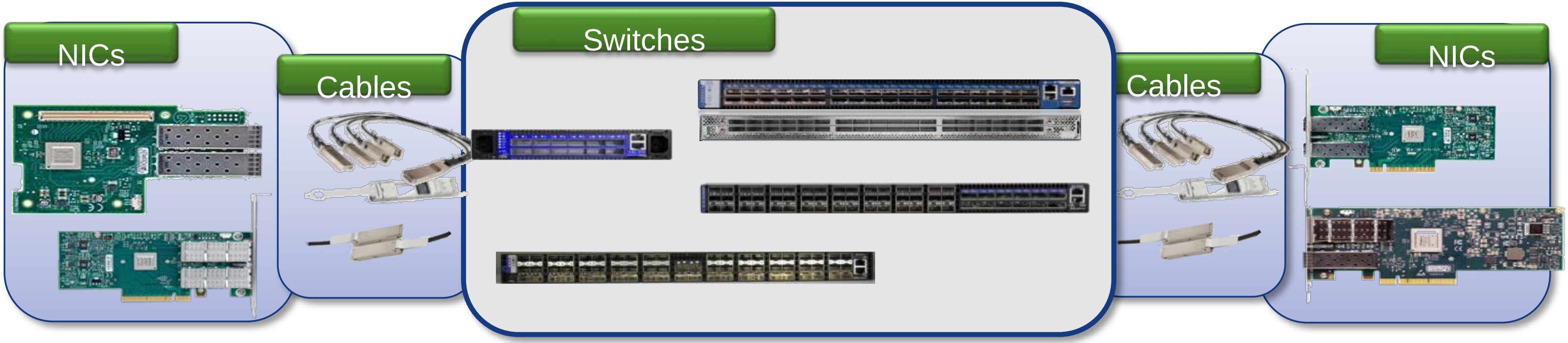






The Network and the Protocol <u>MUST</u> get faster

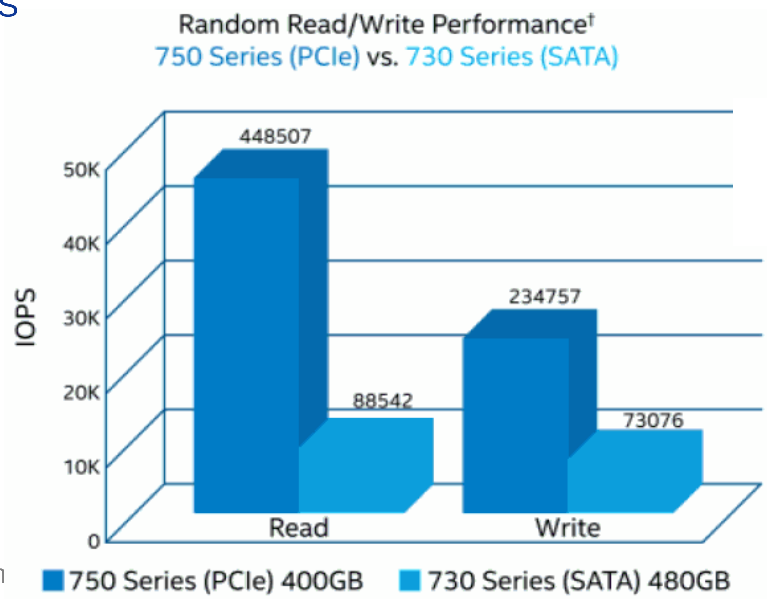
Faster Networking is Here Today:

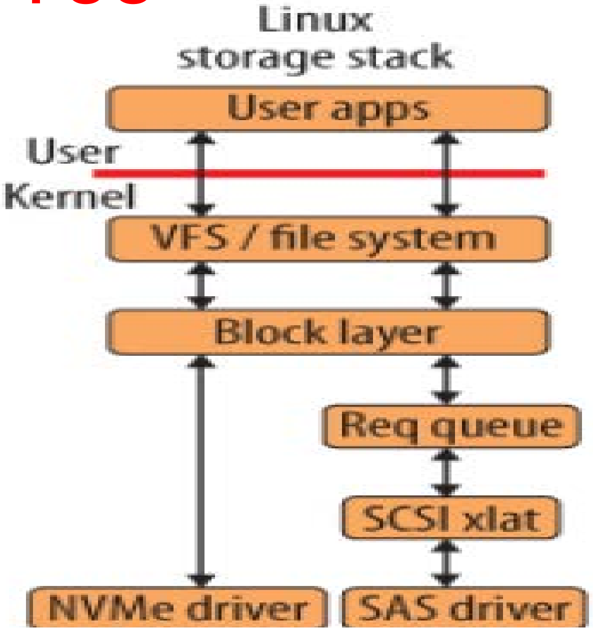


End-to-End 25, 32, 40, 50, 56, 100Gb Ethernet, Fibre Channel & InfiniBand

Faster Protocol is Here Too

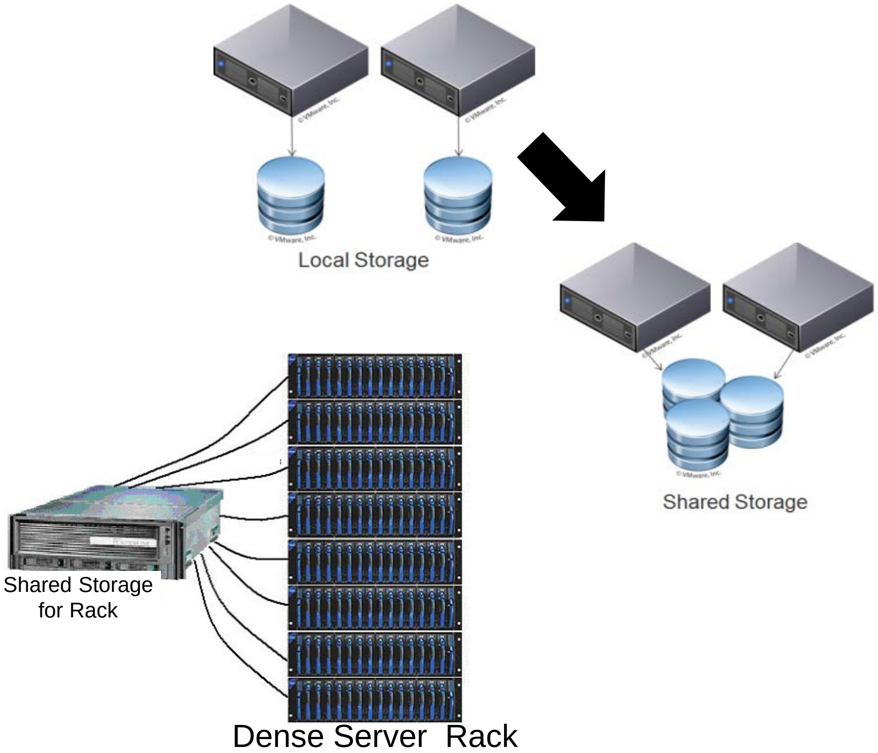
- NVMe: Optimized for flash and next-gen NV-memory
  - Traditional SCSI interfaces designed for spinning disk
  - NVMe bypasses unneeded layers
- NVMe Flash Outperforms SAS/SATA Flash
  - 2x-2.5x more bandwidth, 40-50% lower latency, Up to 3x more IOPS





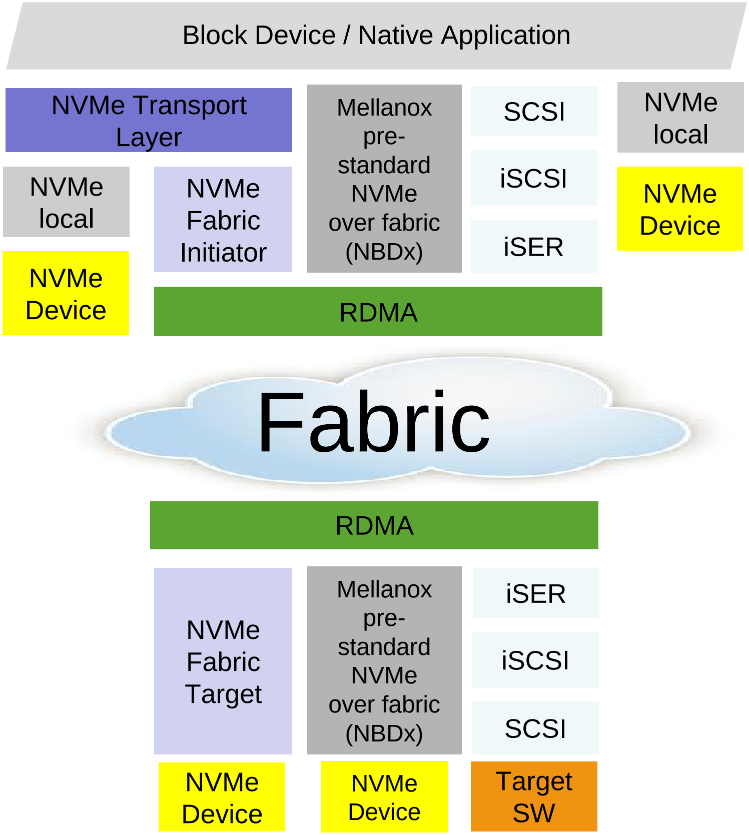
“NVMe Over Fabrics” is the Logical and Historical Next Step

- Sharing NVMe based storage across multiple Servers
- Driven by the need for the best possible compute efficiency and shared advantages
  - Utilization of capacity, rack space, power, cost
  - Ease of scalability, management, fault isolation
- Shared storage requires a Network/Fabric
- NVMe over Fabrics standard in development
  - Version 1.0 in 4Q15



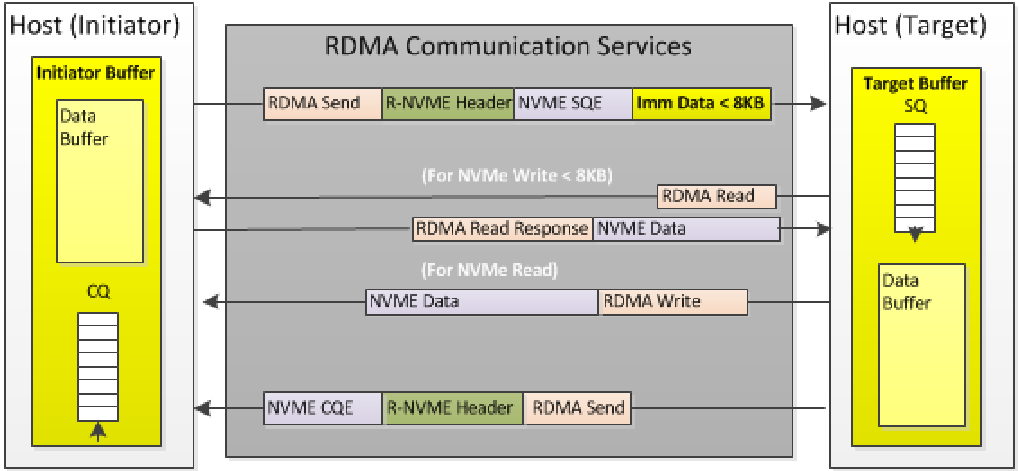
How does “NVMe over Fabrics” work?

- The idea is to extend the efficiency of local NVMe interface over the fabric
  - NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
  - RoCE, iWARP, IB
  - iSER uses SCSI for transport
- A few pre-standard versions are available today



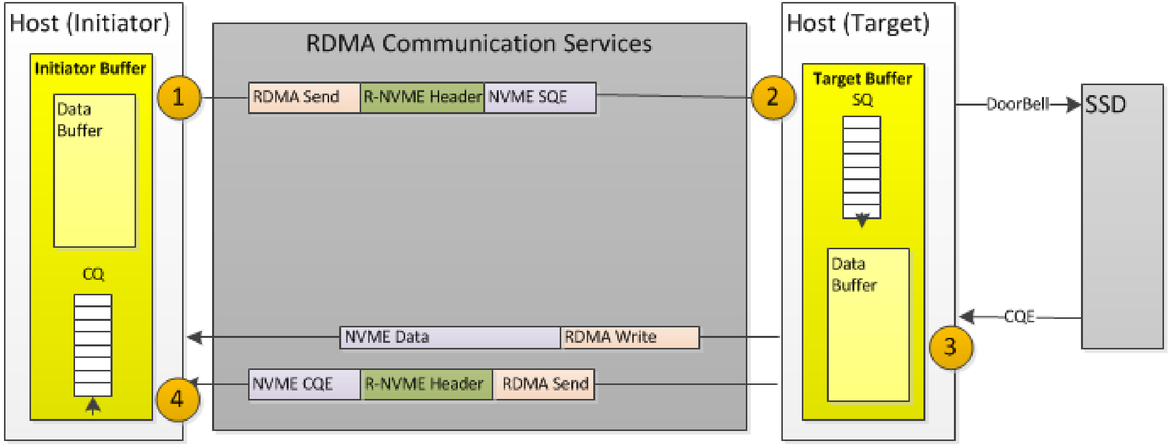
NVMe over Fabrics Transport Model

- One-to-one mapping between NVMe queues and RDMA queues
- RDMA-Send used to push encapsulated NVMe SQE and CQE
- NVMe SQs are target resident, NVMe CQs are initiator resident
- NVMe data exchange is done by RDMA operations
  - NVMe write data can be transferred within the command as immediate data
    - Limited to write < 8KB



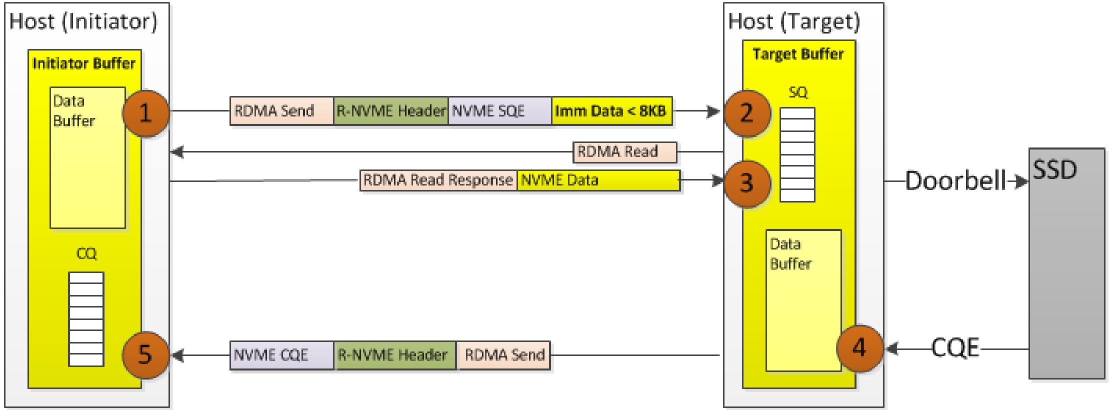
Remote Read Operation Algorithm

- Initiator
  - Post SQE send to NIC
- Target
  - Upon RDMA Send Completion (2)
    - Allocate Memory for Data
    - Post SQE to SQ
    - Post DB to SSD
    - Wait for SSD completion (3)
    - Send data with RDMA
    - Send NVMe CQE
    - Free memory



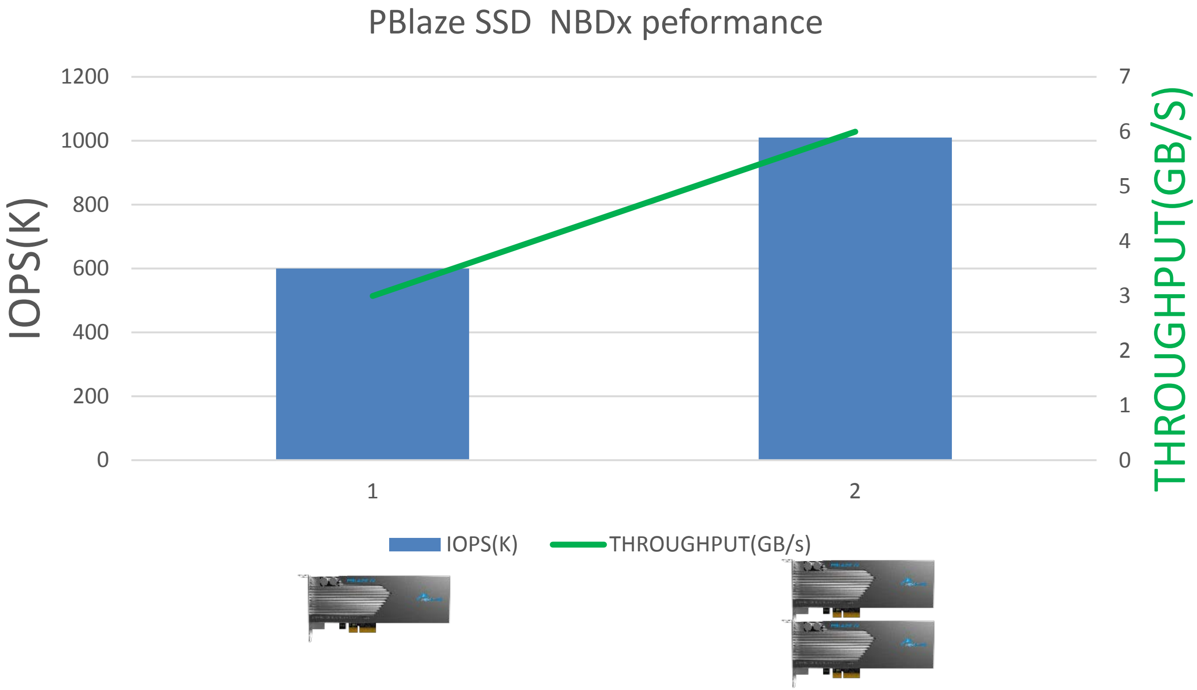
Remote Write Operation Algorithm

- Initiator
  - Post SQE send to NIC
- Target
  - Upon RDMA Send Completion (2)
    - Allocate Memory for Data
    - Post SQE to SQ
    - If (immediate data)
      - Write data to memory
    - Else
      - Post RDMA to fetch data
      - Wait for RDMA completion (3)
    - Post DB to SSD
    - Wait for SSD completion (4)
    - Send NVMe CQE
    - Free memory



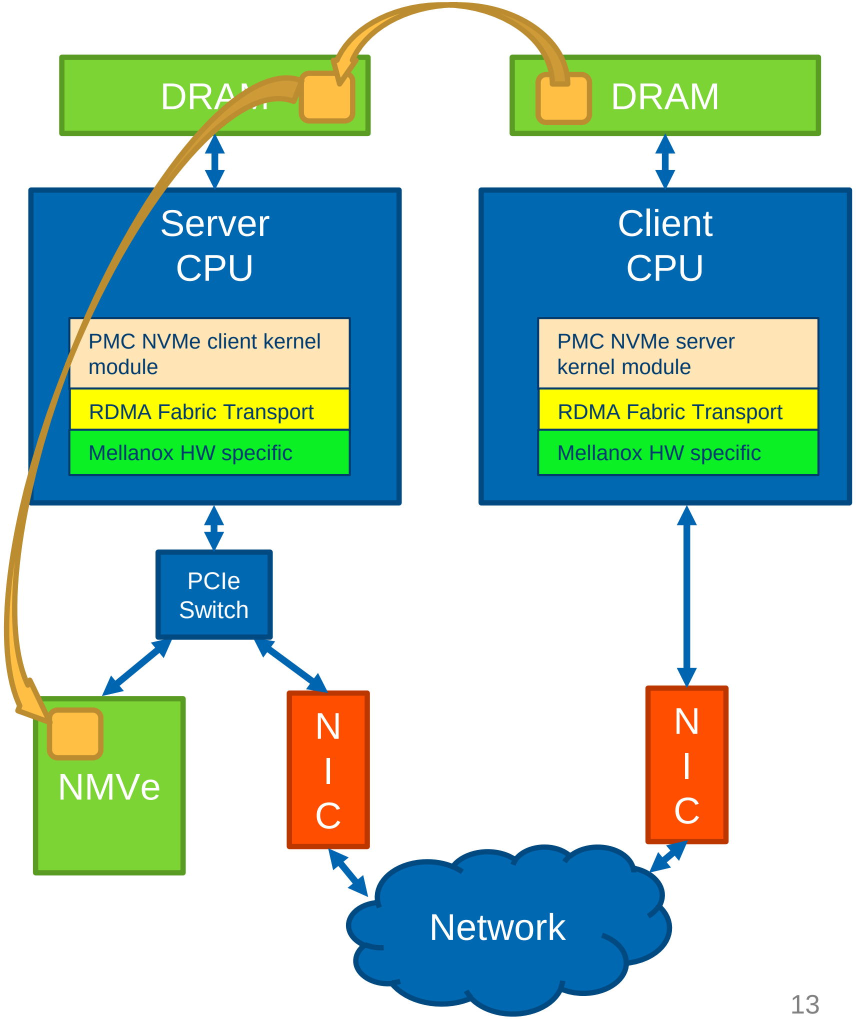
Memblaze

PBLAZE IV

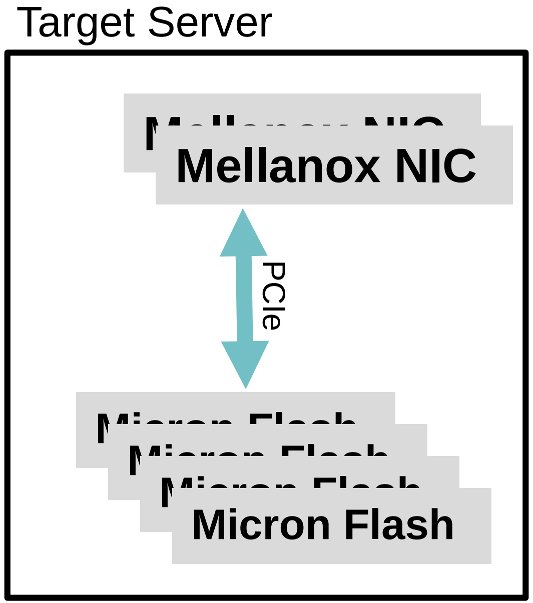


PMC

- Development platform to enable testing of NVMe with RDMA
  - Mellanox RDMA NIC
  - PMCS high performance NVMe device
- IO performance for remote NVMe transactions similar to local device
  - No impact to IO throughput
    - Fully utilizing RDMA bandwidth with 4K IO
  - Latency impact is currently <6us on 4KB random Read/Write



Micron

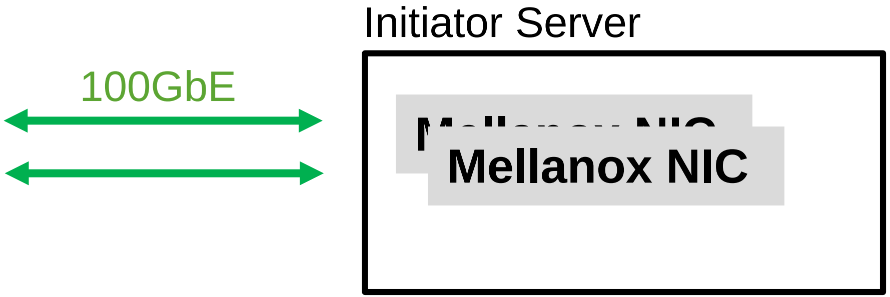


Target Server

Mellanox NIC

PCIe

Micron Flash



100GbE

Initiator Server

Mellanox NIC

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

Conclusions

- New Non-Volatile storage technology is moving the performance bottle neck for SANs from the storage devices to the IO
- The IO industry is responding with new products to address this
- RDMA technology is essential to these new products
    - RoCE is the only Ethernet RDMA protocol shipping today that makes sense for NVMe over Fabrics
    - iWARP adds back the TCP/IP layers, negating the performance gains from removing SCSI
- Some IO vendors are already sampling Pre-standard products
    - Public demos of NVMe over Fabrics(NVMeOF) are happening
    - Expect to see products next year

Questions?

Rob Davis

robd@mellanox.com