NVMe over Fabrics

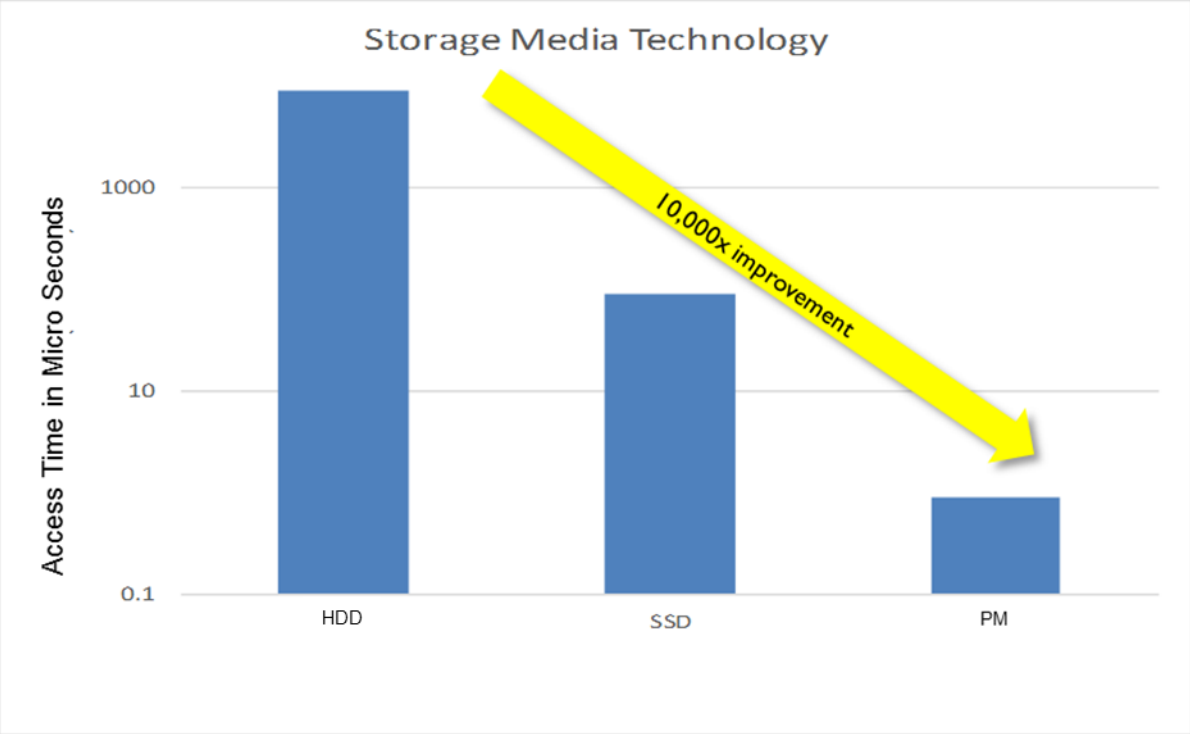
High Performance SSDs networked over Ethernet

Rob Davis
Vice President Storage Technology, Mellanox

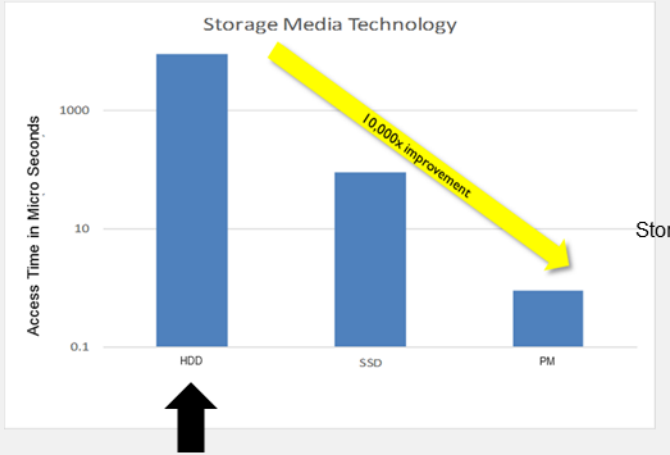
Ilker Cebeli
Senior Director of Product Planning, Samsung

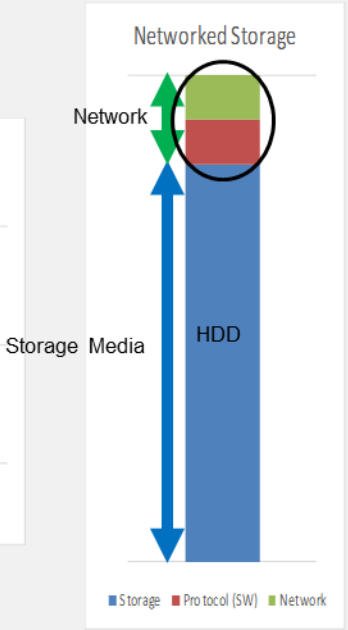
May 3, 2017

Storage Performance Dramaticly Increasses

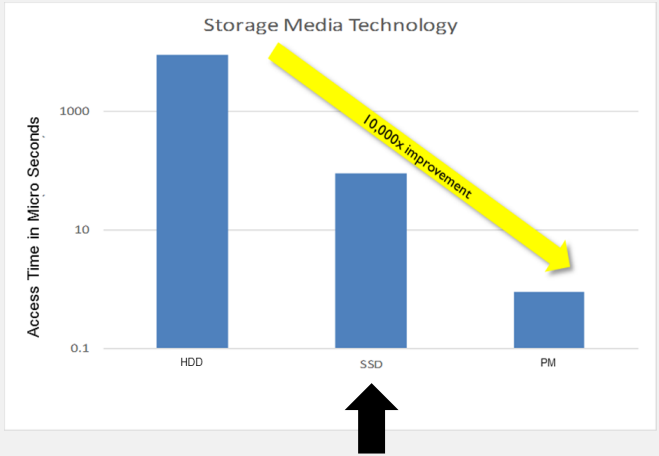


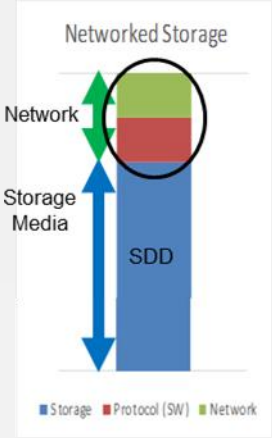
New Storage Performance Creates Bottleneck



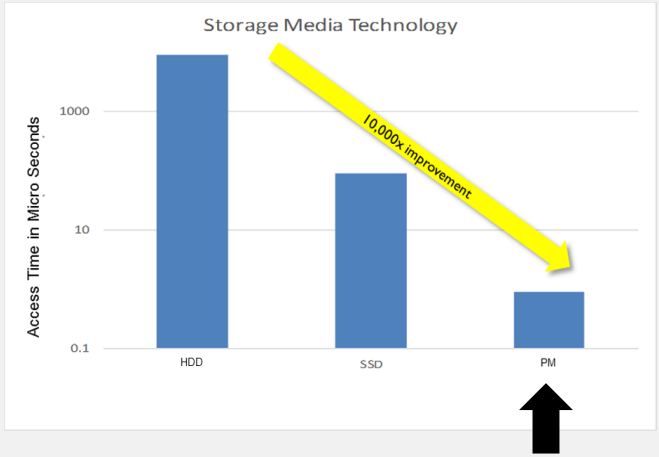


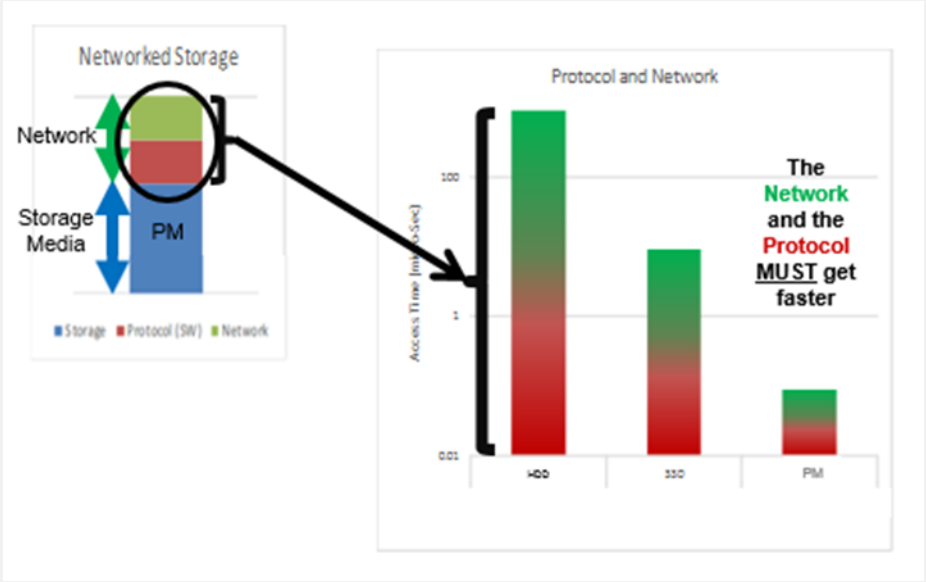
New Storage Performance Creates Bottleneck





New Storage Performance Creates Bottleneck

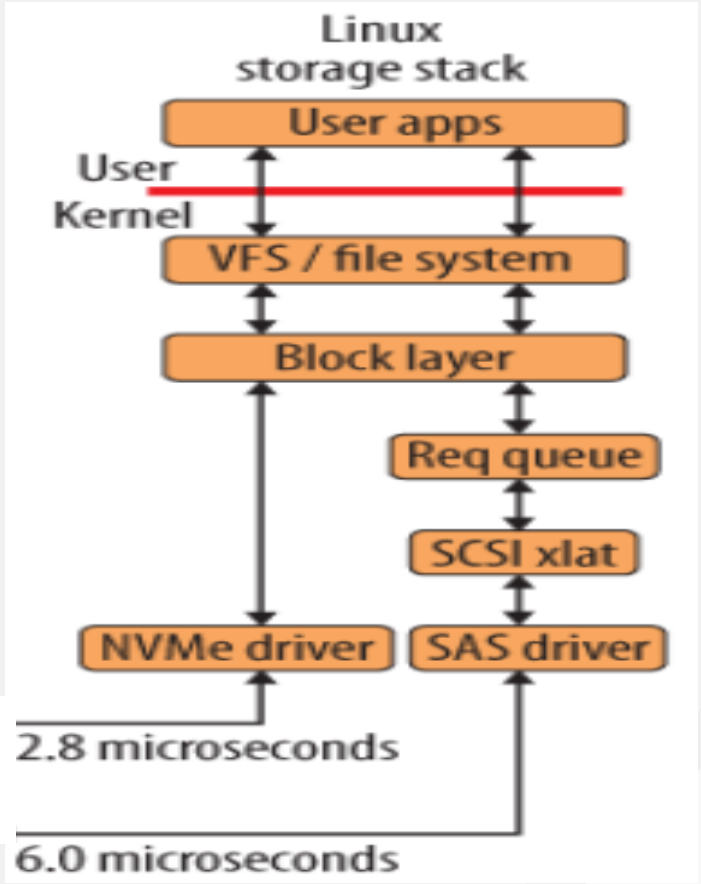




NVMe Technology – Background

Optimized for flash

- Traditional SCSI designed for disk
- NVMe bypasses unneeded layers
- Dramatically reducing latency



NVMe Design Advantages

- Lower latency
- Direct connection to CPU's PCIe lanes
- Higher bandwidth
- Scales with number of PCIe lanes
- Best in class latency consistency
- Lower cycles/IO, fewer cmds, better queueing
- Lower system power
- No HBA required

NVMe SSD Product Example

Samsung PM963 NVMe SSD

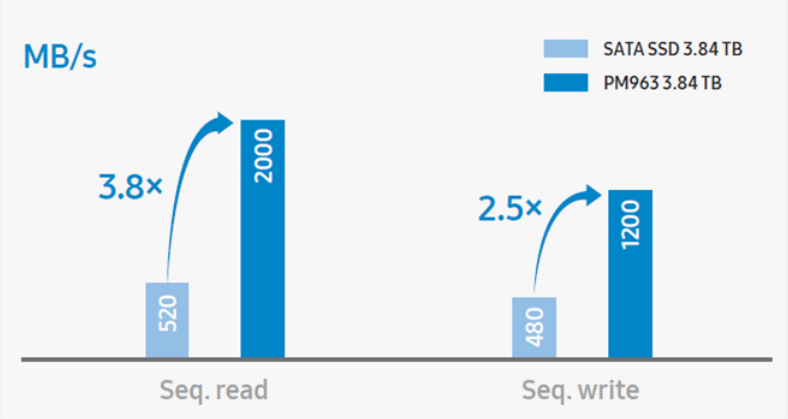
Samsung PM963 Specification	
Form Factor	2.5”
Host Interface	PCIe Gen3 x4
Capacities	800GB, 1.6TB, 3.2TB
Sequential Read	2000 MB/s
Sequential Write	1200 MB/s
Random Read	Up to 430KIOPS
Random Write	Up to 40KIOPS

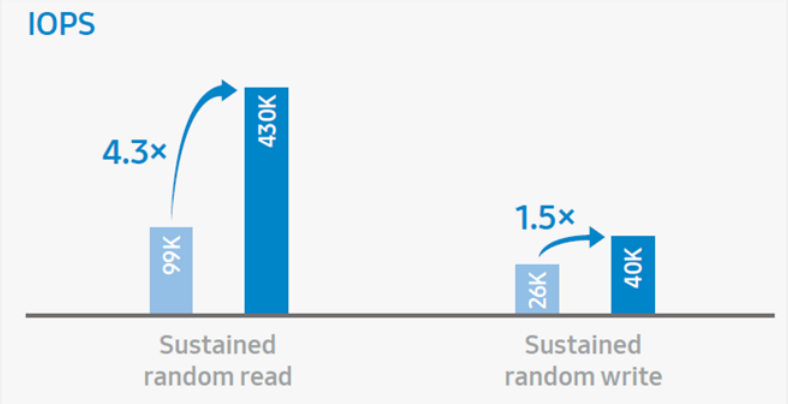
- Leverages latest VNAND technology
- Delivers consistent low latency

NVMe Performance

NVMe outperforms SATA SSDs

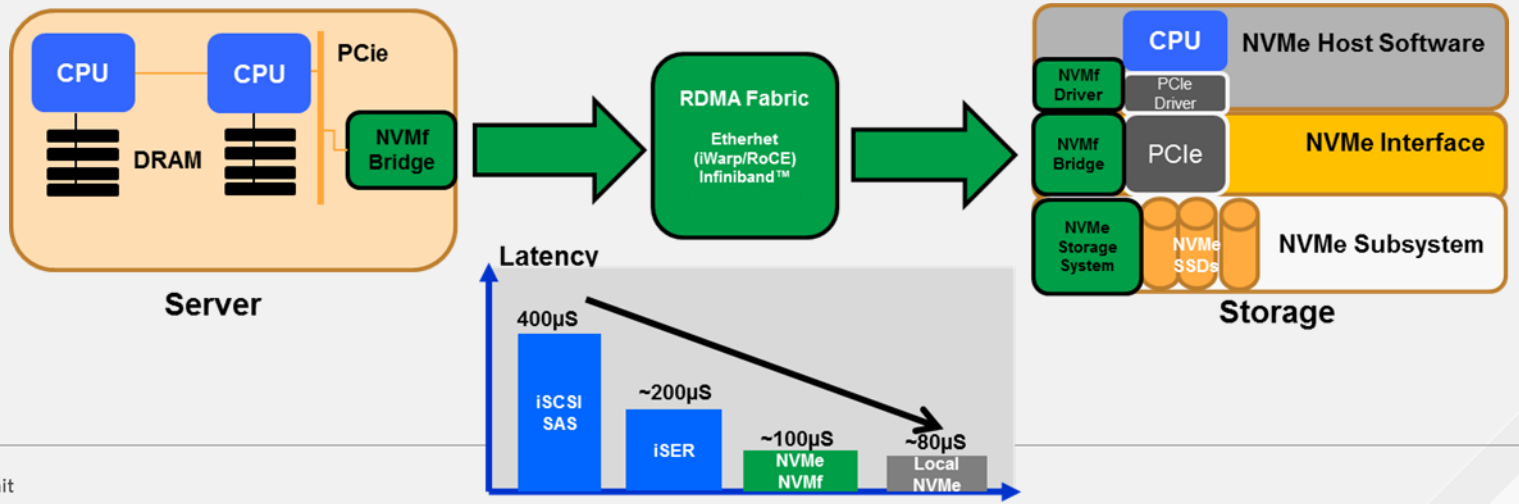
- 2.5x-4x more bandwidth,
- 40-50% lower latency
- Up to 4x more IOPS





What is NVM Express Over Fabrics (NVMe-oF)?

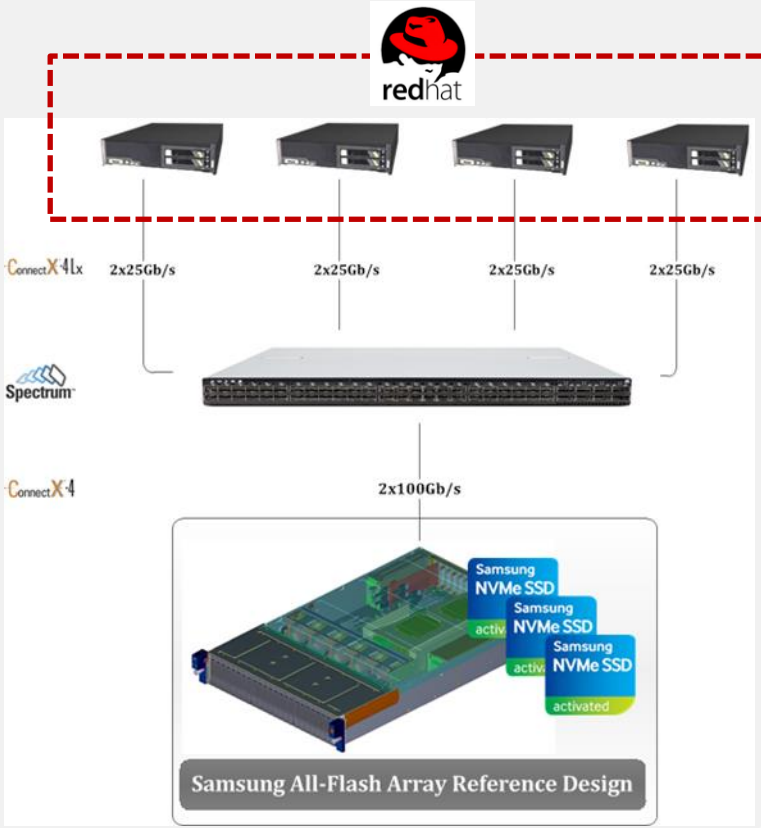
- A protocol interface to NVMe that enable operation over other interconnects (e.g., Ethernet, InfiniBand™, Fibre Channel).
- Shares the same base architecture and NVMe Host Software as PCIe
- Enables NVMe Scale-Out and low latency (<10µS latency) operations on Data Center Fabrics
- Avoids protocol translation (avoid SCSI)



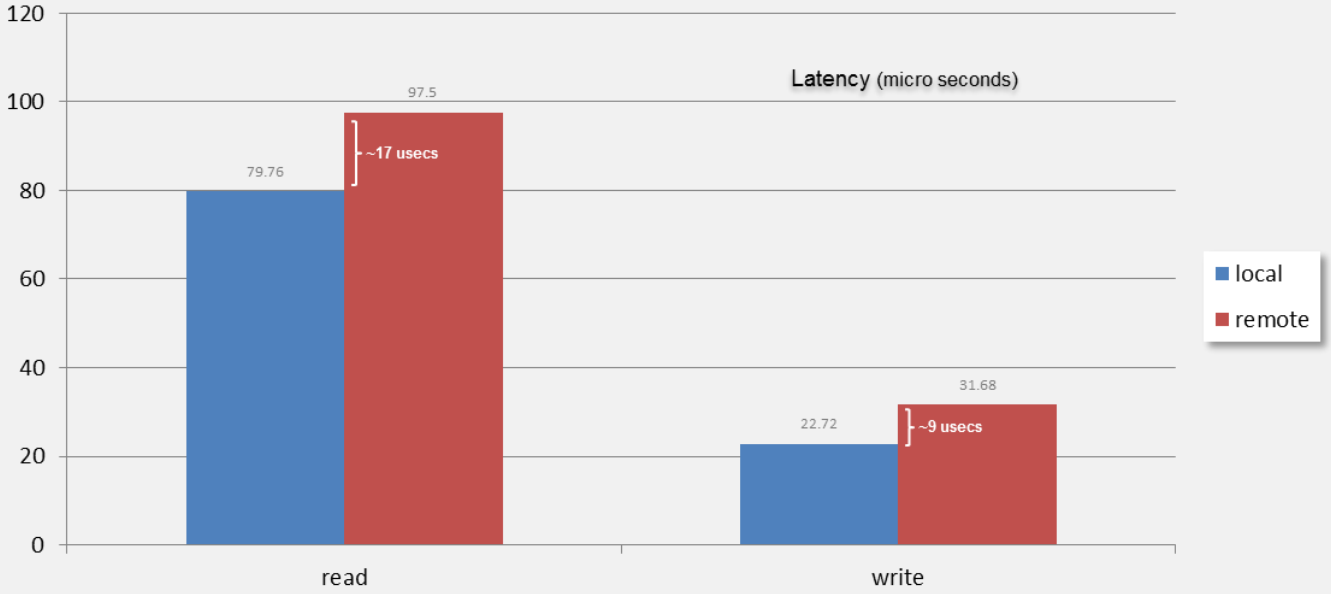
NVMe-oF Performance Test

Configuration

- 1x NVMf target
- 24x Samsung PM963 NVMe 2.5” 960GB SSDs
- 2x 100Gb/s Mellanox ConnectX®-4 EN
- 4x initiator hosts
- 2x25Gb/s each
- Open Source NVMe-oF kernel drivers

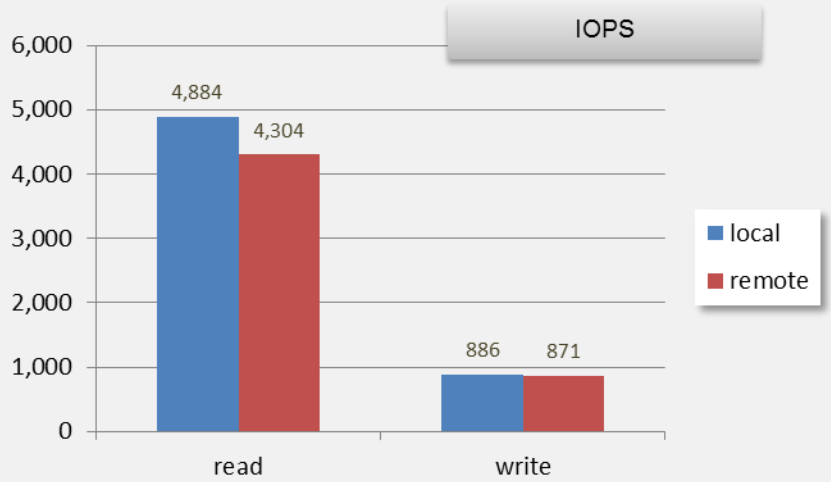


Latency Comparison



- Random IO at QD1, 1 job
- Round-trip delta: Reads ~17usecs; Writes ~9usecs

Performance (24 SSDs)



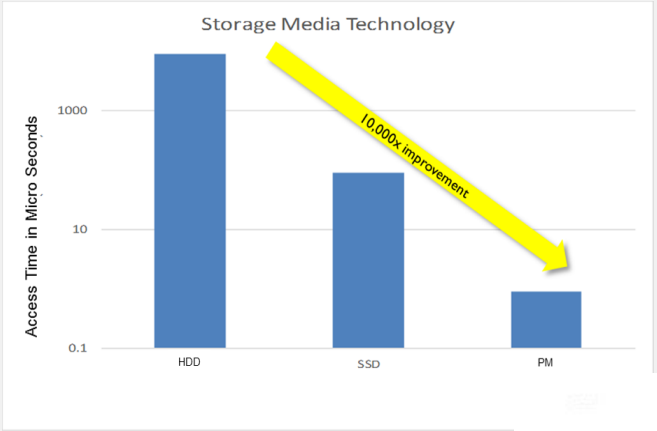
- High aggregate NVMe-oF performance: 4.3M IOPS & 21.5GB/s throughput

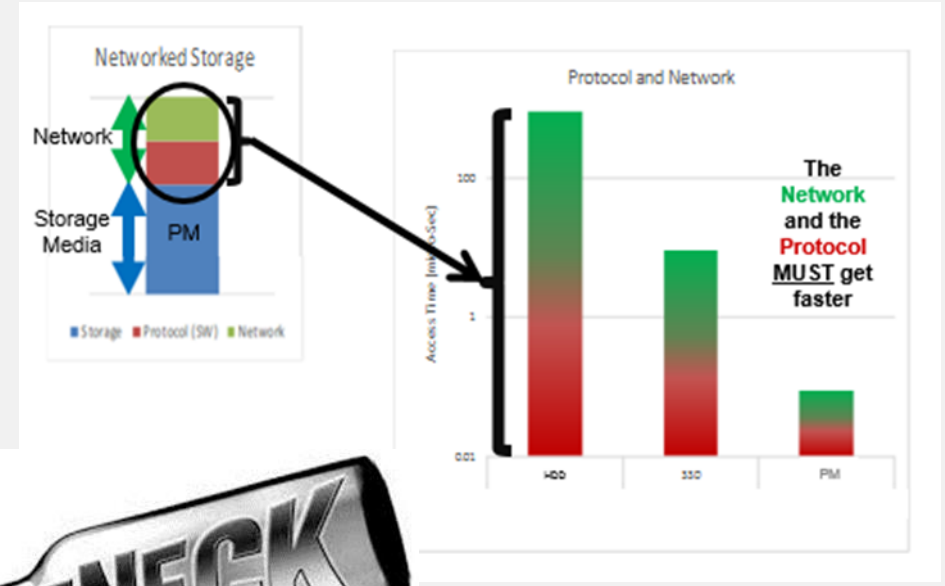
Summary: NVMe Local vs. Remote

Performance Delta		1-drive	24-drive
Latency	Read	11%	15%
	Write	On par	On par
IOPS	Read	10%	12%
	Write	On par	2%
Throughput	Read	On par	18%
	Write	On par	On par

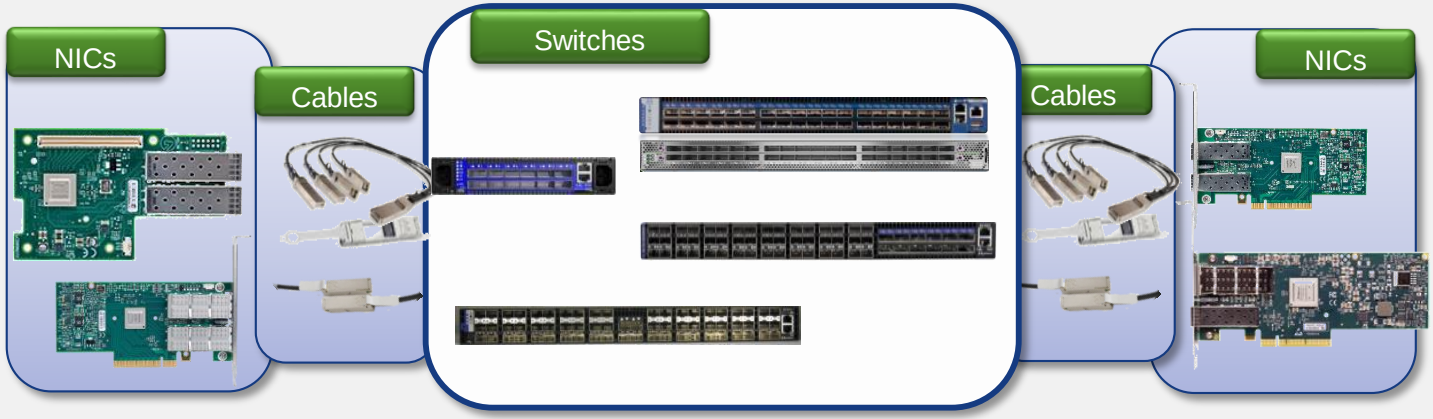
LEARN. NETWORK.
EXPERIENCE
OPEN SOURCE.

New Storage Performance Creates Bottleneck





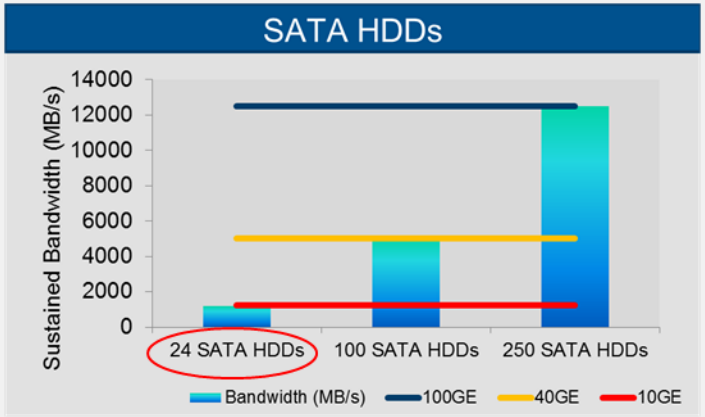
Faster Networking is Here Today



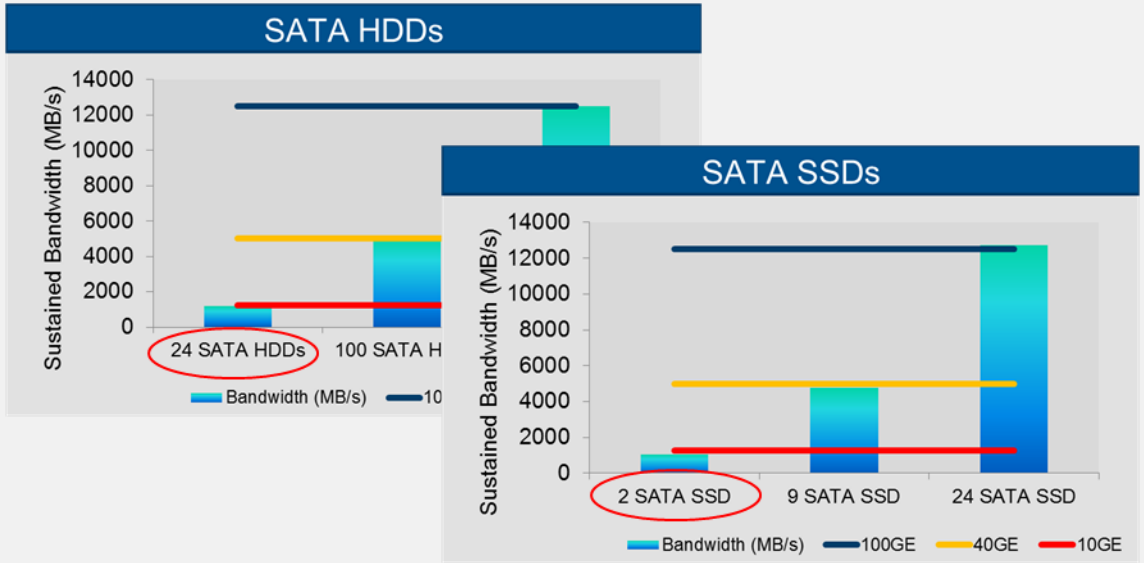
Ethernet & InfiniBand

End-to-End 25, 40, 50, 56, 100Gb

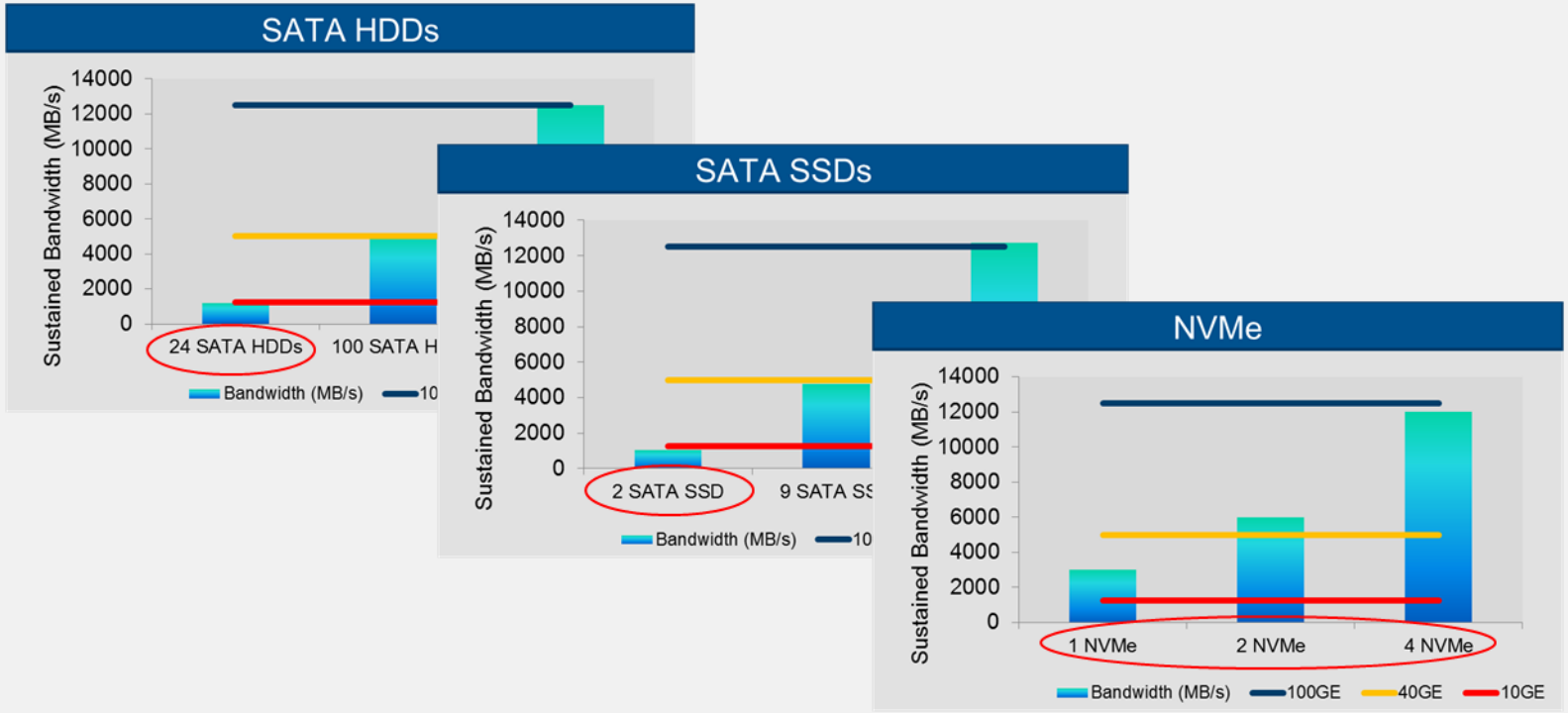
Faster Storage Needs a Faster Network



Faster Storage Needs a Faster Network



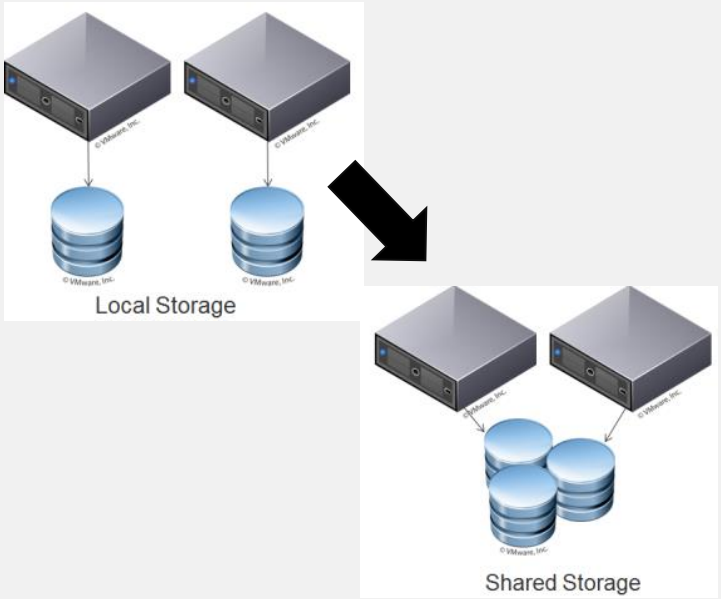
Faster Storage Needs a Faster Network



“NVMe over Fabrics” Enables Storage Networking of NVMe

Sharing NVMe-based storage with multiple servers

- Better utilization: capacity, rack space, and power
- Better scalability
- Management
- Fault isolation



NVMe over Fabrics (NVMe-oF) industry standard

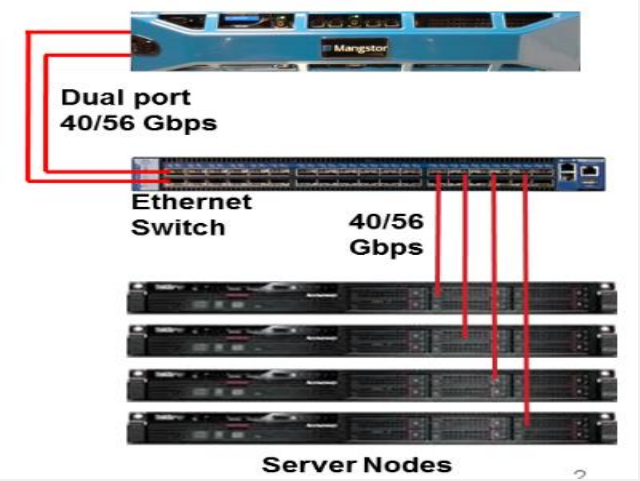
NVMe.org developed the specification

- Many contributing companies
- Version 1.0 completed in June 2016

Early pre-standard demos:

- Mellanox, Samsung, Intel, Micron, PMC, Mangstor, WD, others
- Version 1.0 at Flash Memory Summit August of 2016

NAB April 2015



Showed high IOPs and bandwidth and extremely low latency

Some NVMe-oF Demos at FMS and IDF 2016

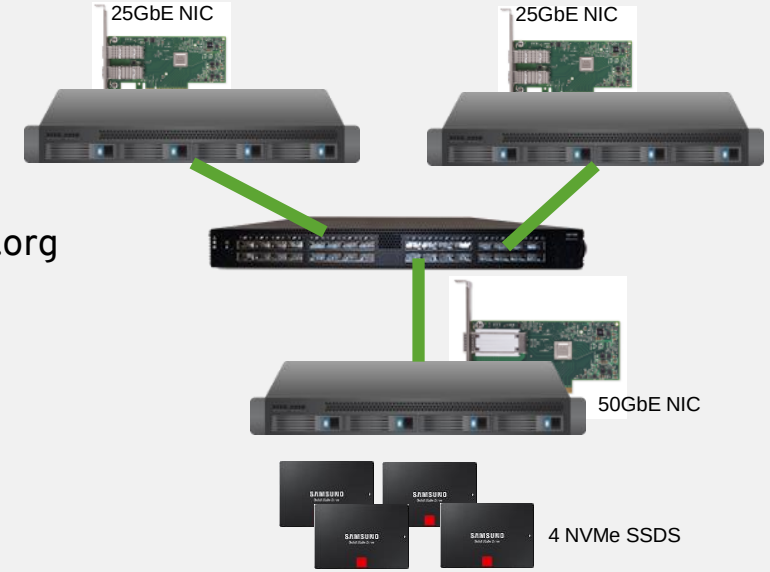
Flash Memory Summit

- Samsung
- E8 Storage
- Micron
- Newisis (Sanmina)
- Pavilion Data - in Seagate booth
- Mangstor

Intel Developer Forum

- Samsung
- HGST (WD)
- Intel
- Newisis (Sanmina)
- E8 Storage
- Seagate

NVMe-oF Performance



Open Source Linux NVMe-oF Software from NVMe.org

- Accepted in upstream kernel
- Will be in a future RHEL

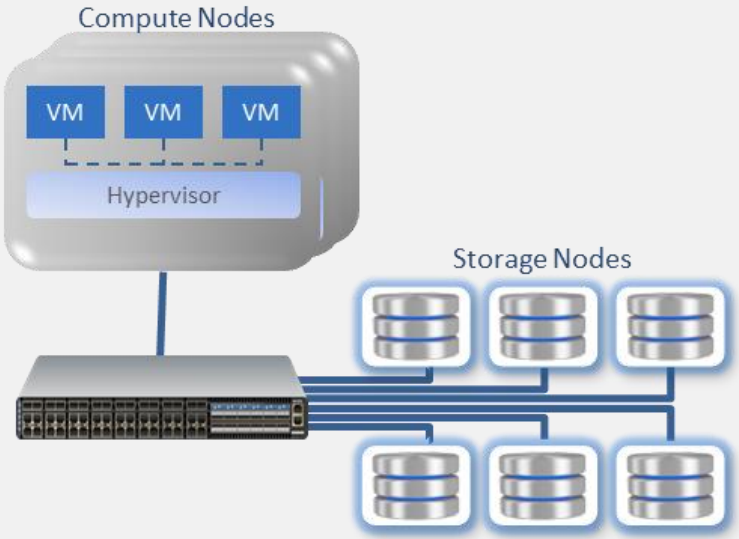
Added fabric latency
~12us, BS = 512b

	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%

Applications for NVMe-oF

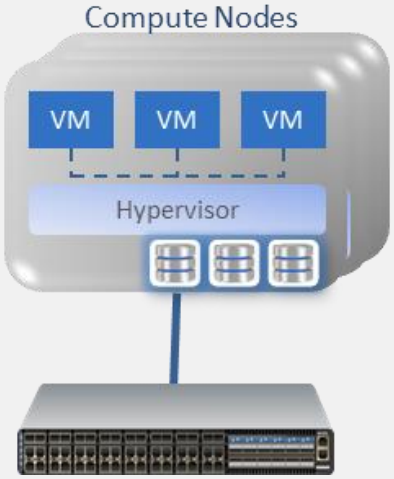
Scale-Out Storage

- Low latency
- High bandwidth
- Enables low TCO with high performance

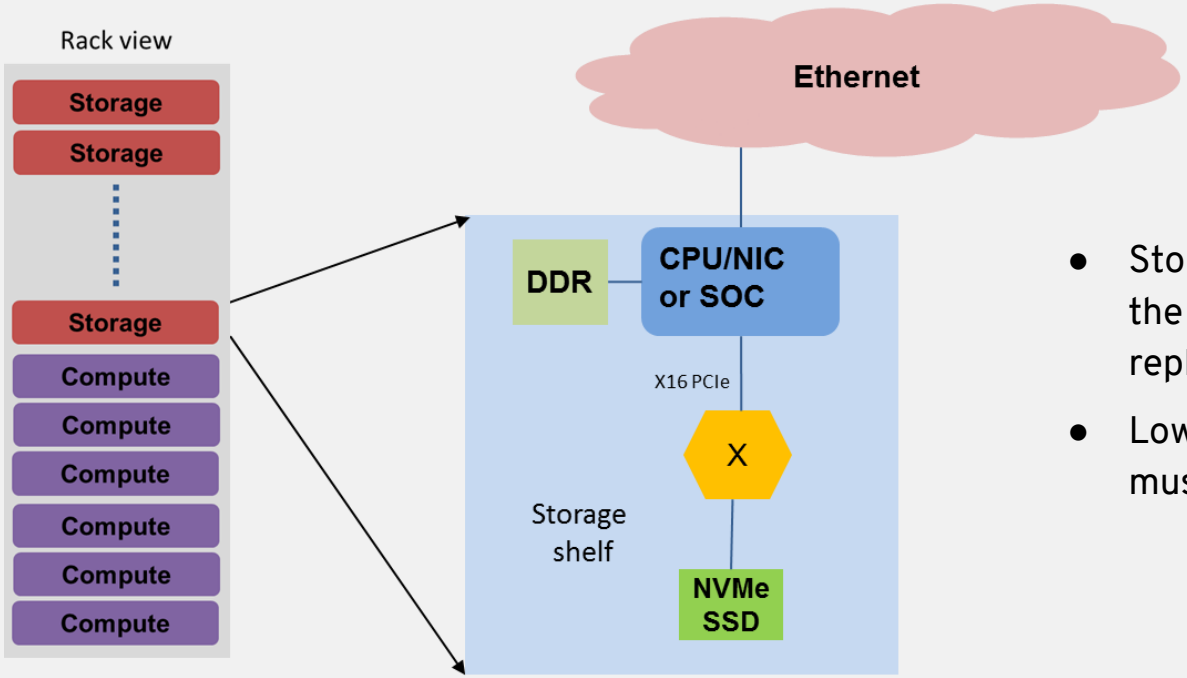


Hyper-Converged

- Collapse separate compute & storage
- Integrated compute and storage nodes
- Low latency and High bandwidth enable higher performance application support



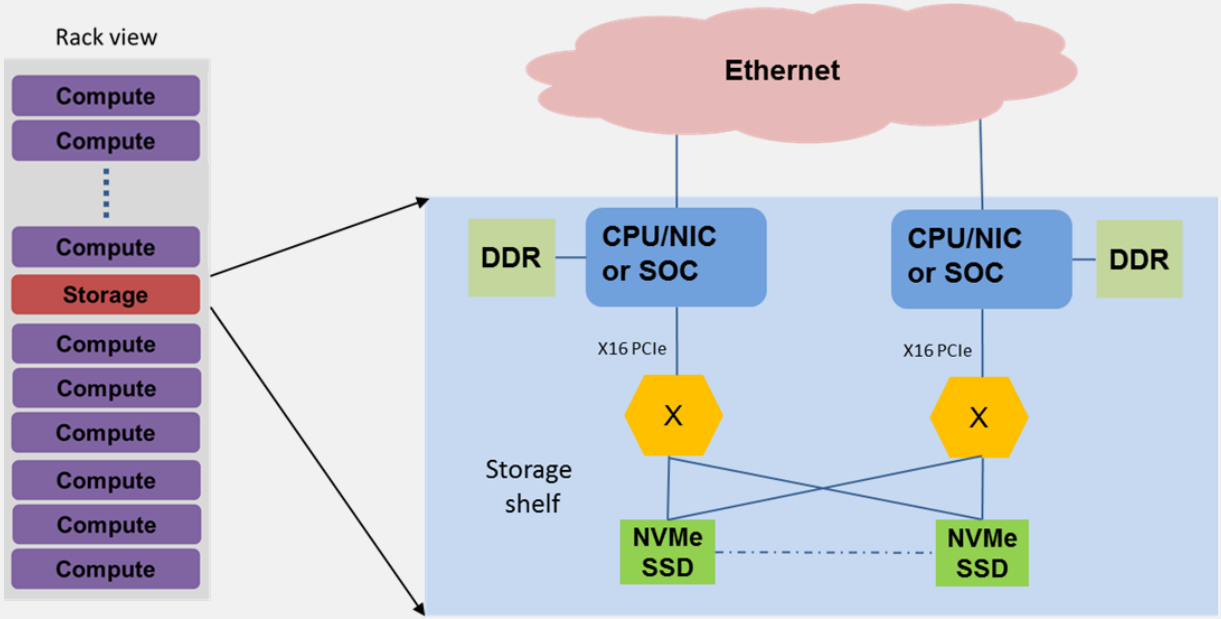
Compute/Storage Disaggregation



- Storage and Compute are not in the same enclosure – DAS replacement
- Low latency and High bandwidth a must

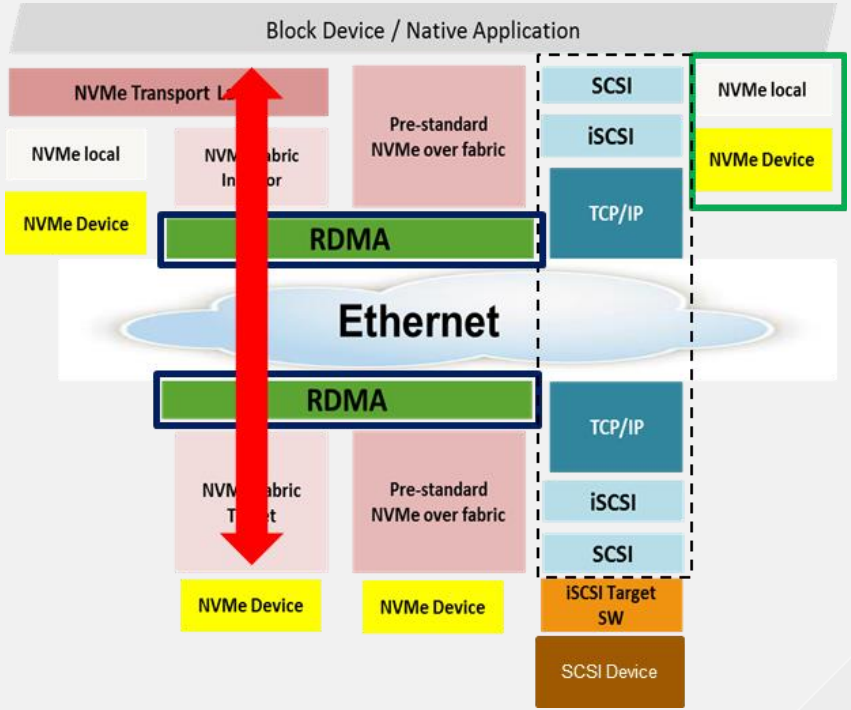
Classic SAN

- Better utilization: capacity, rack space, and power
- Better scalability
- Management
- Fault isolation



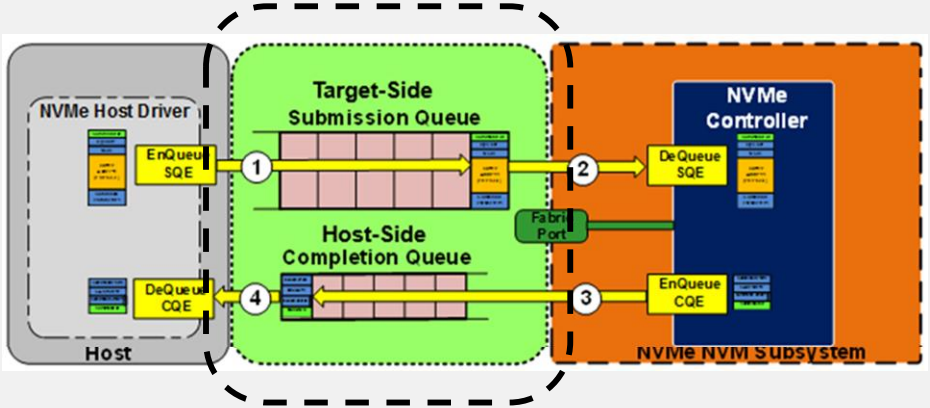
Why is NVMe-oF so Fast

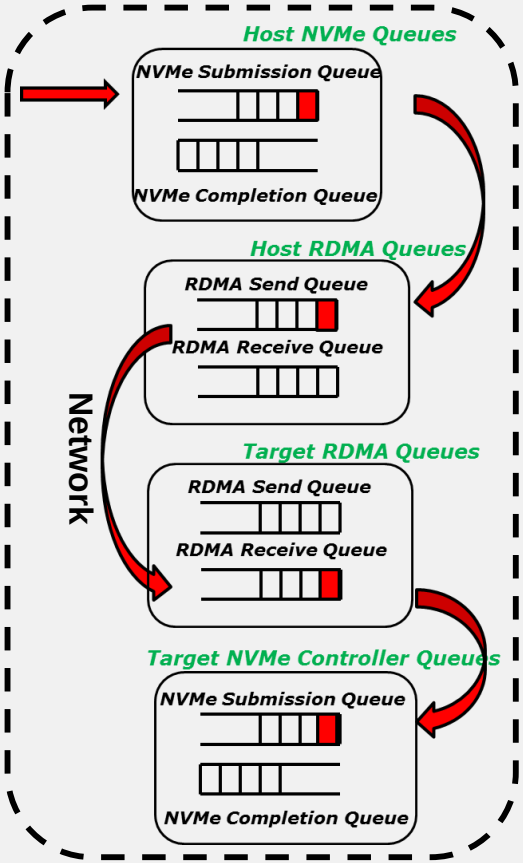
- Extends NVMe efficiency over a fabric
- NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
- Bypassing TCP/IP



https://community.mellanox.com/docs/DOC-2186

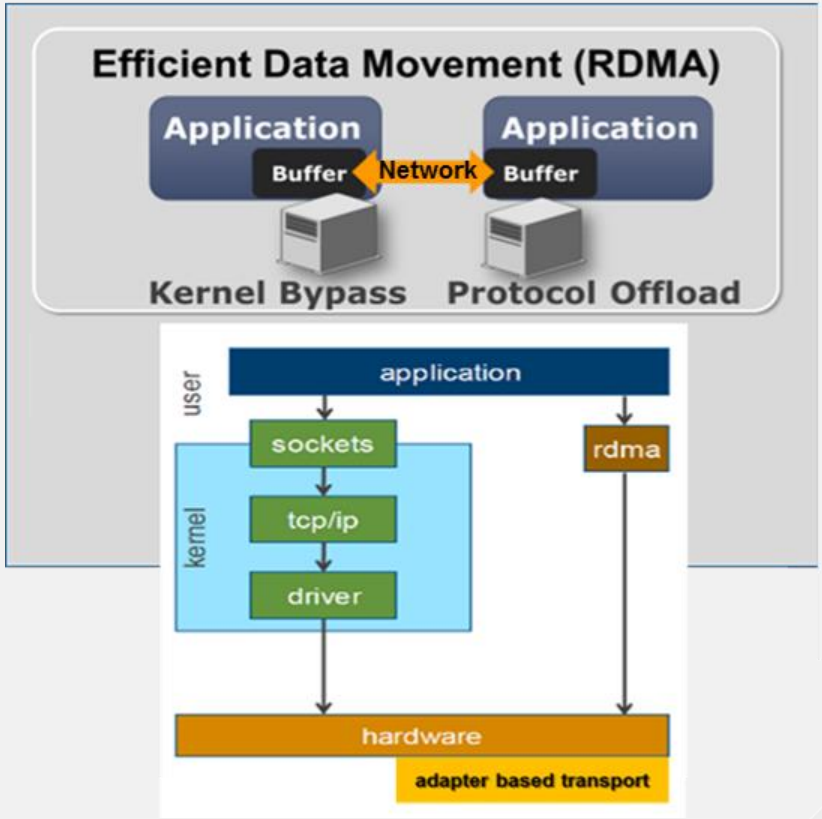
RDMA & NVMe: A Perfect Fit





What is RDMA

- Remote version of DMA(Direct Memory Access)
- Memory to memory move with out CPU
- TCP/IP stack bypass
- Transport layer in RNIC



NVMe-oF Products Available Today

Just a sample of the market – not all inclusive list

- SuperMicro
- Pavillion
- Mangstor
- E8
- Liqid
- Excelero
- Pavilion
- AIC
- Sanmina

Reference Designs

- Samsung
- Micron
- Toshiba
- Kingston
- WD
- Seagate

Conclusions

- New storage technology is moving the performance bottle neck for networked storage from the storage devices to the network – **“Faster Storage needs Faster Networks”**
- The Industry is responding with faster speeds and NVMe-oF protocol
- RDMA technology is essential to high NVMe-oF performance
- This performance will enable many new networked storage solutions
- Early products and SSD vendor reference designs are already available

Questions?


plus.google.com/+RedHat


facebook.com/redhatinc


linkedin.com/company/red-hat


twitter.com/RedHatNews


youtube.com/user/RedHatVideos

THANK YOU


plus.google.com/+RedHat


facebook.com/redhatinc


linkedin.com/company/red-hat


twitter.com/RedHatNews

youtube.com/user/RedHatVideos