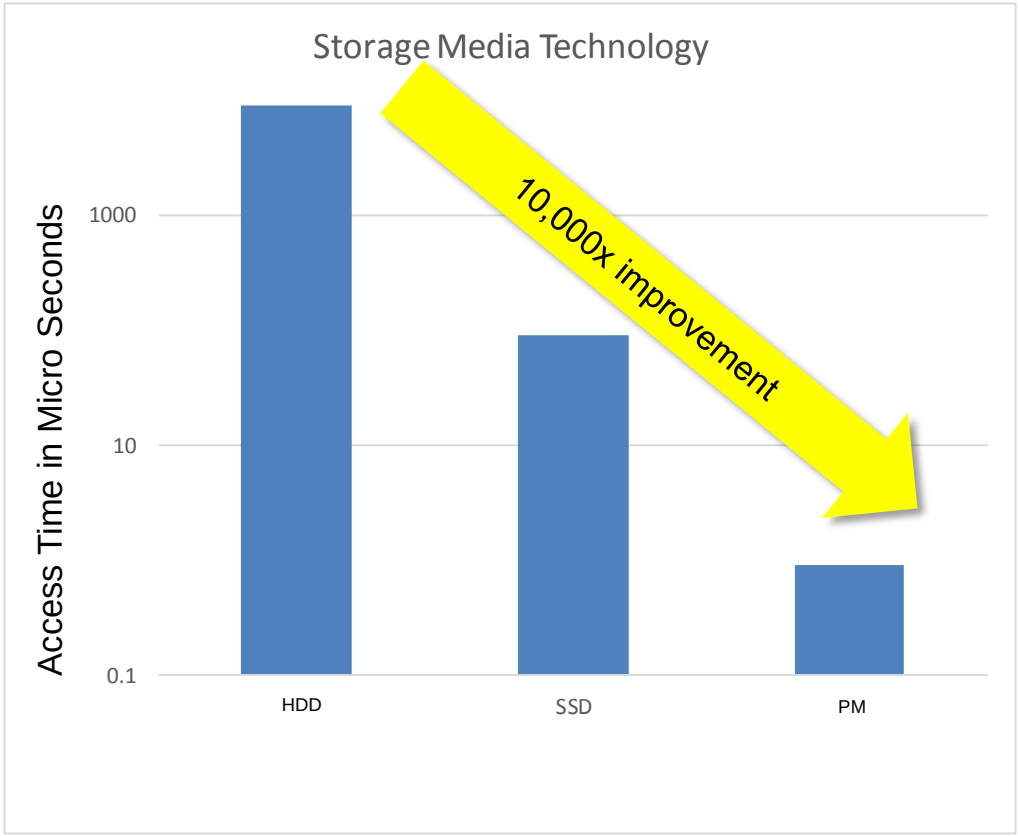
NVMe Over Fabrics (NVMe-oF)

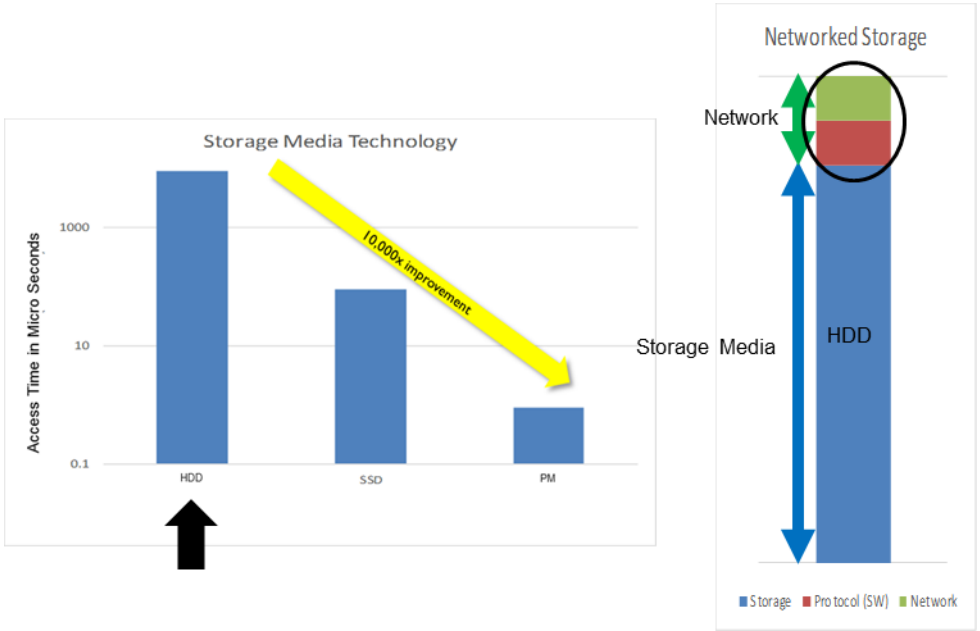
High Performance Flash Moves to Ethernet

Rob Davis
Vice President Storage Technology, Mellanox

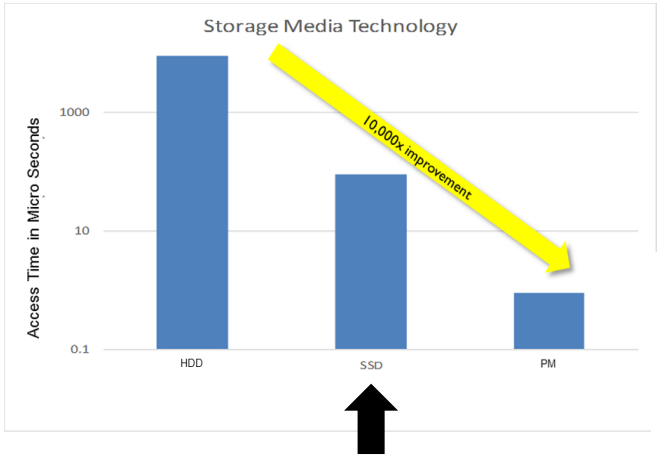
Why NVMe over Fabrics?

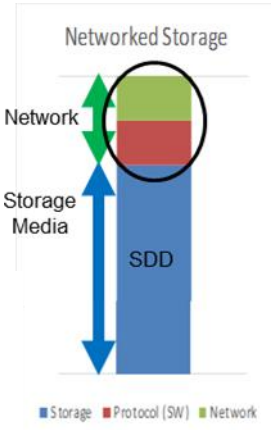


New Storage Performance Creates a Bottleneck

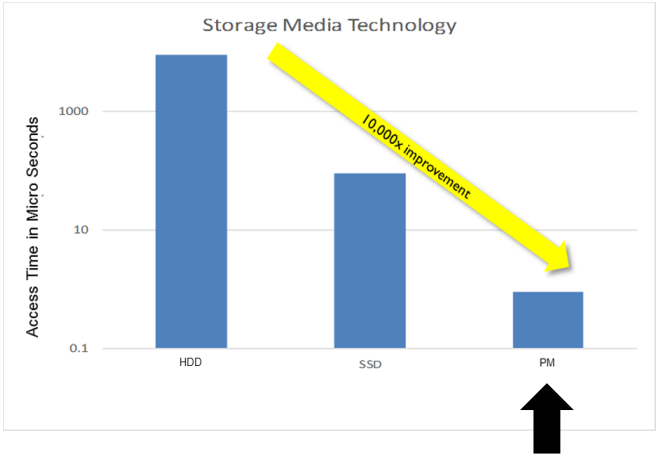


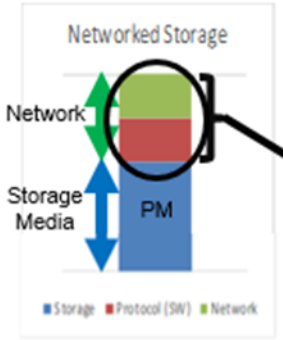
New Storage Performance Creates a Bottleneck

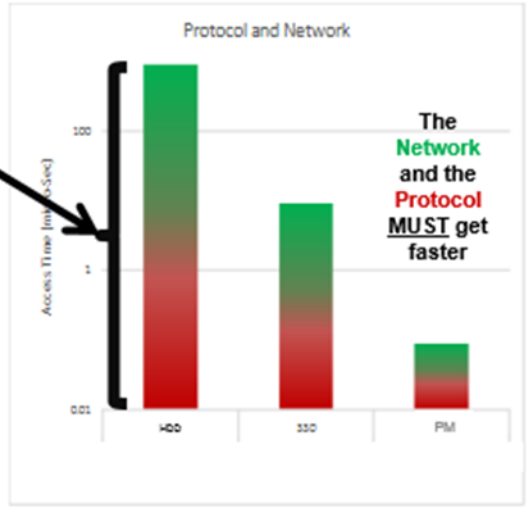




New Storage Performance Creates a Bottleneck

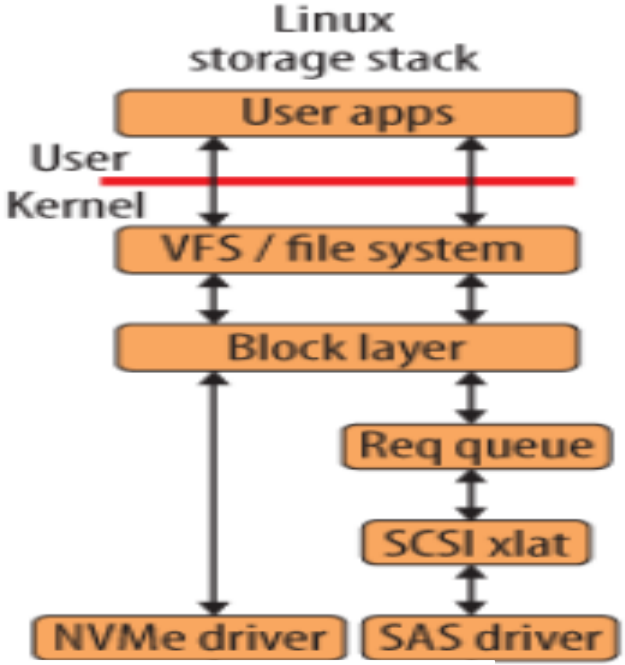




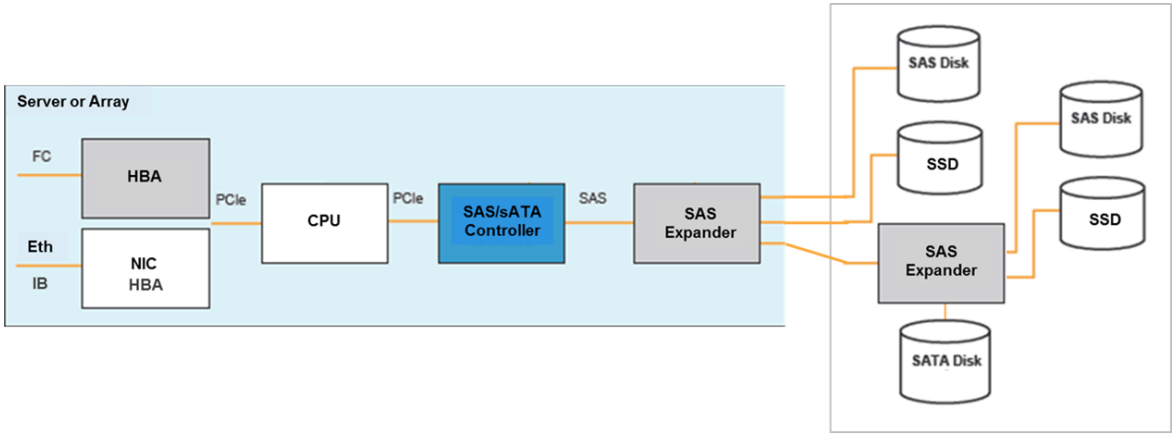


NVMe Technology Background

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



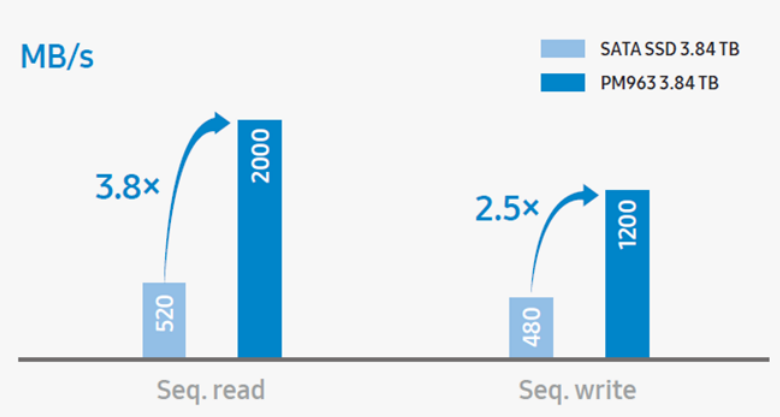
NVMe Technology Background

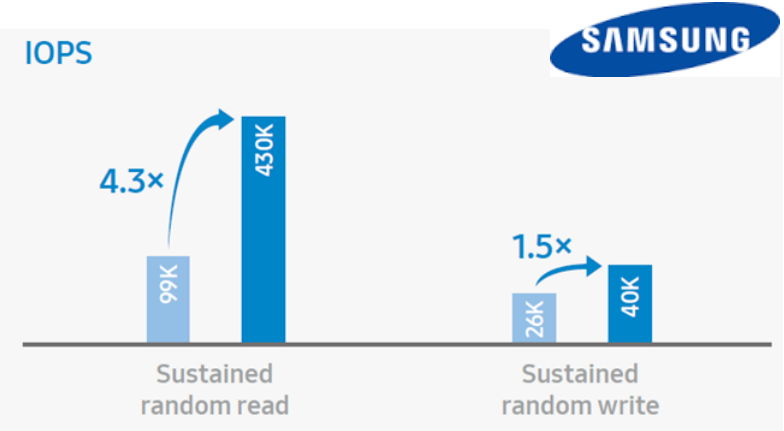


- Direct connection on PCIe from SSD to CPU
- No wire technology translation needed
  - Higher Performance
- Less components
  - Lower power
  - Lower cost(not yet)

NVMe Performance vs. sATA

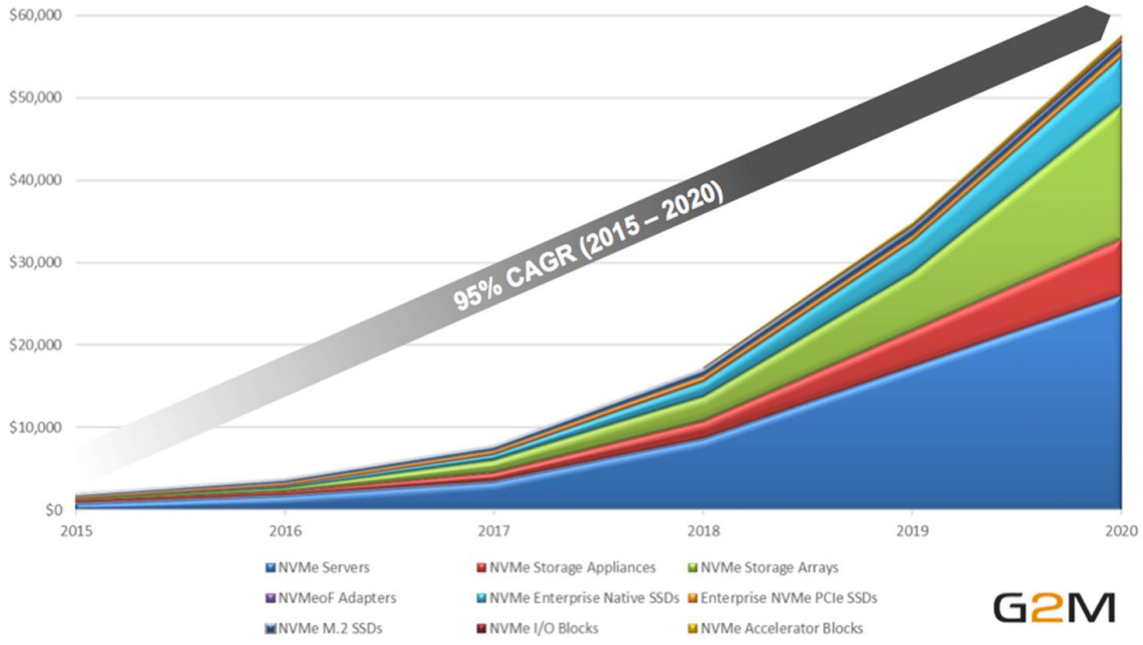
- 2.5x-4x more bandwidth
- 40-50% lower latency
- Up to 4x more IOPS





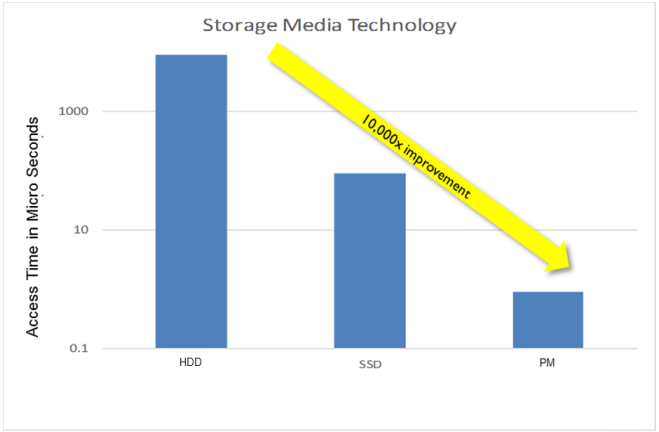
Analyst Predict Rapid Growth

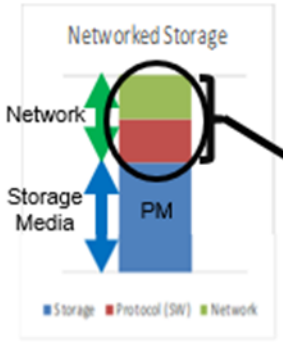
- ~50% of enterprise servers and storage appliances will support NVMe by 2020
- ~40% of all-flash arrays will be NVMe-based by 2020
- Shipments of NVMe SSDs will grow to 25+ million by 2020

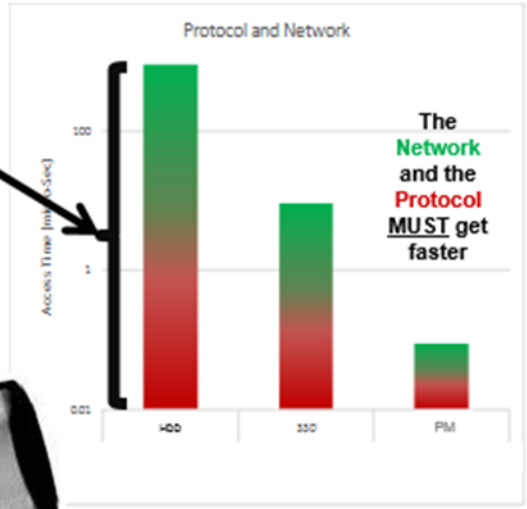


http://www.storagenewsletter.com/rubriques/market-reportsresearch/nvme-market-at-57-billion-by-2020-with-95-cagr-g2m-research/

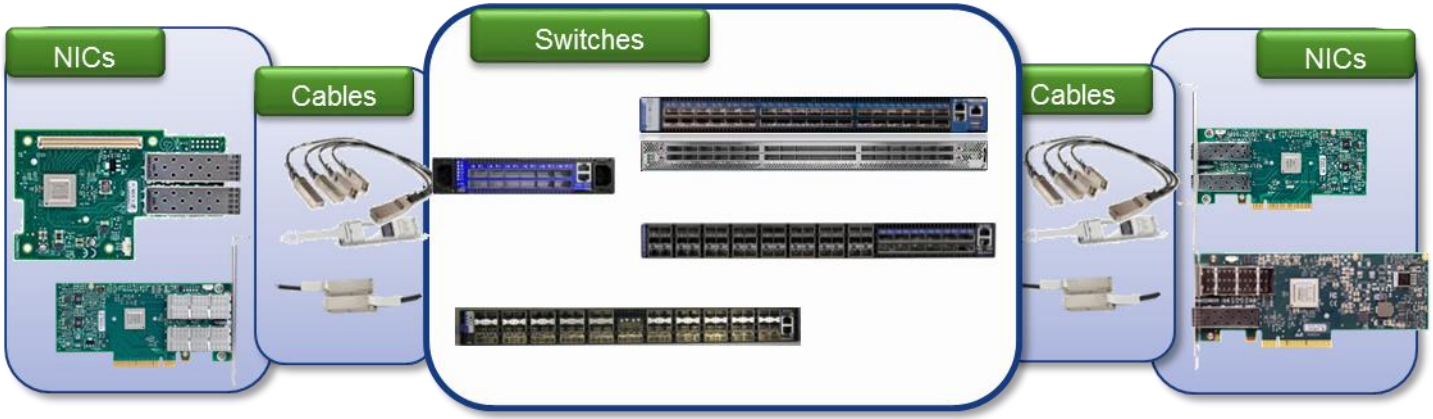
The Network and the Network Protocol Must get Faster





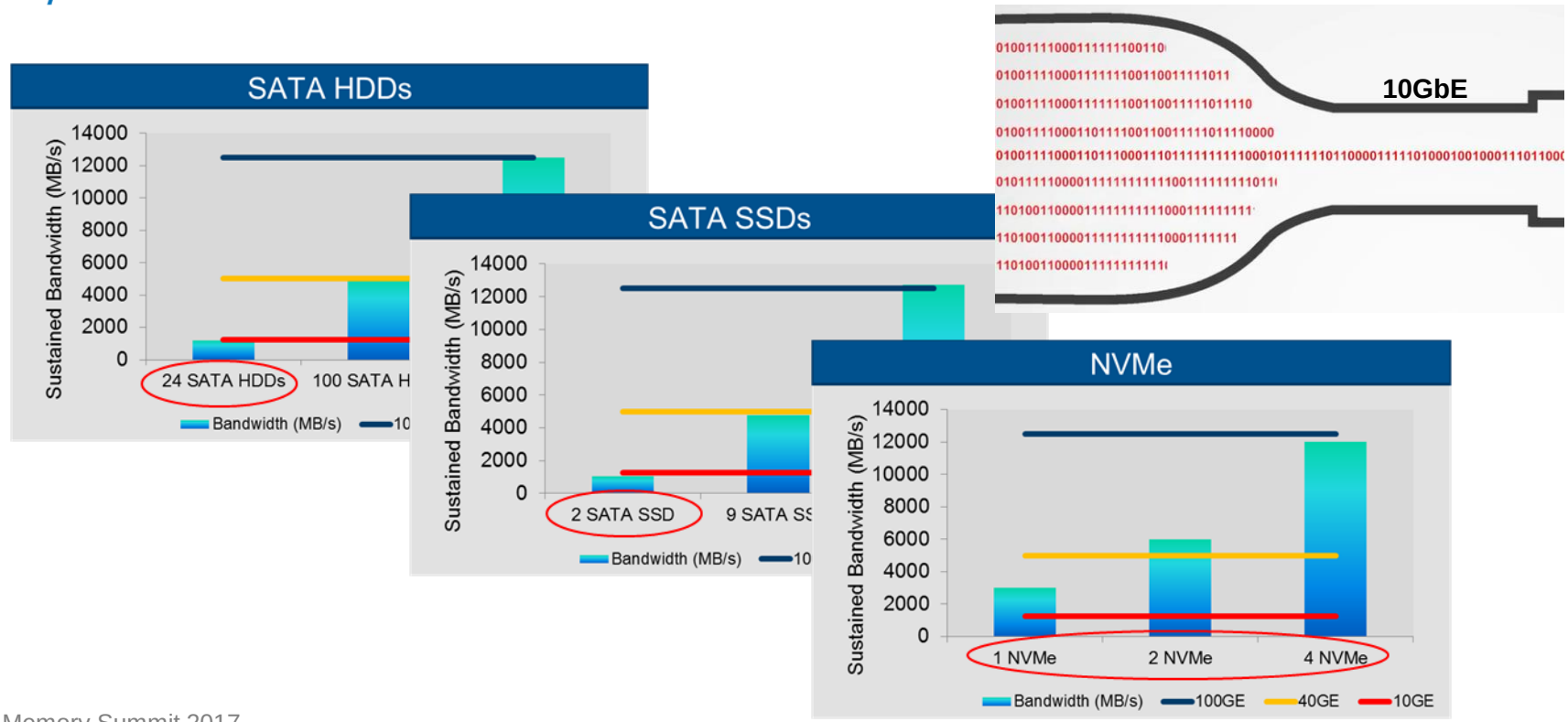


Faster Network Wires Solves ½ the Problem



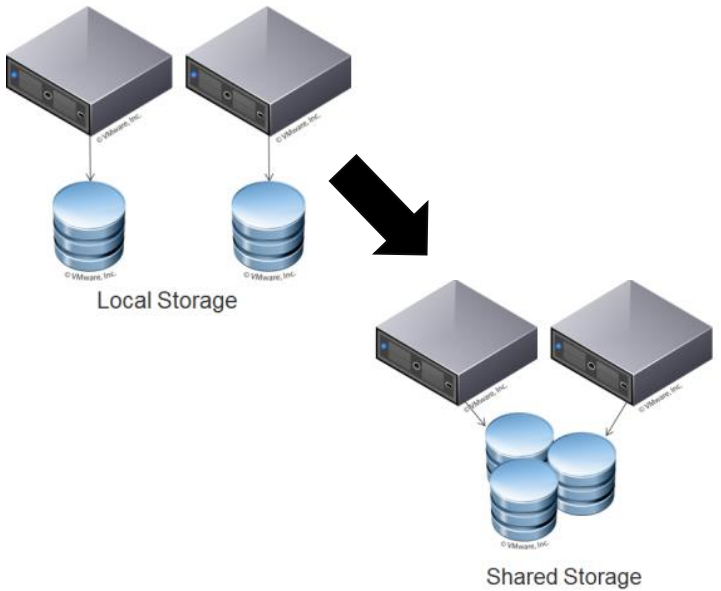
Ethernet & InfiniBand
End-to-End 25, 40, 50, 56, 100Gb
Going to 200 and 400Gb

Faster Storage Needs a Faster Network



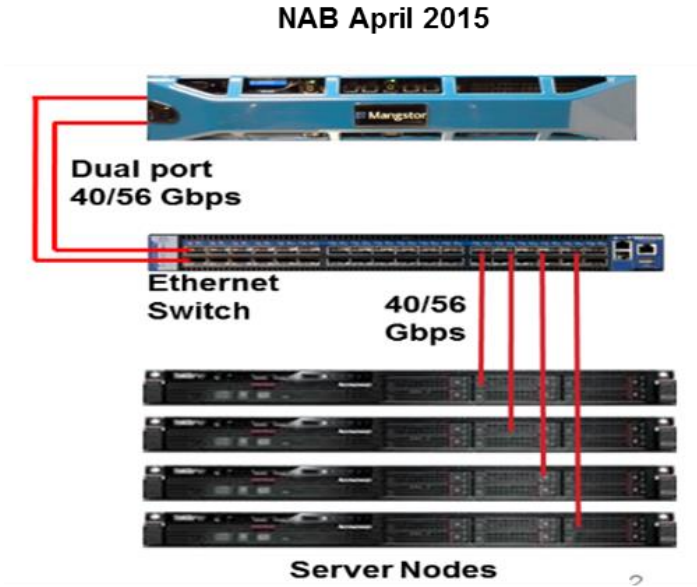
“NVMe over Fabrics” Enables Storage Networking of NVMe

- 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



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

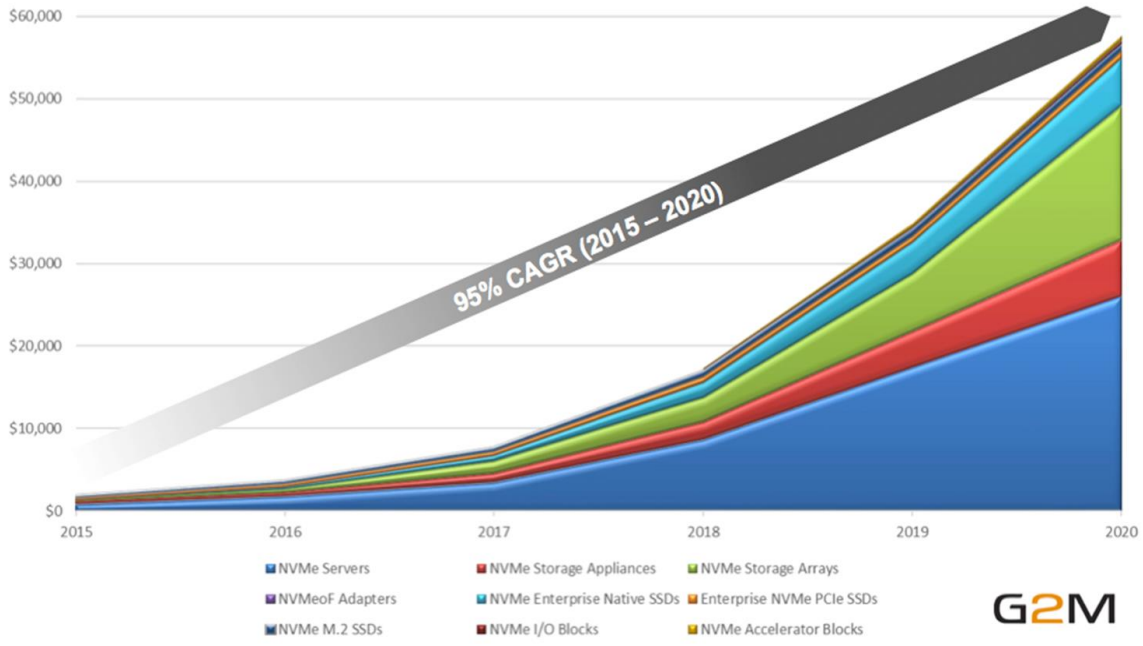
- Samsung
- HGST (WD)
- Intel
- Newisis (Sanmina)
- E8 Storage
- Seagate

NVMe-oF Demos FMS 2017

- E8, Micron, Celestica, Toshiba, Samsung, Mellanox, IBM, Kaminario, Excelero, MicroSemi, Newisys/Sanmina, Seagate/AIC, others

Analyst Predict Rapid Growth

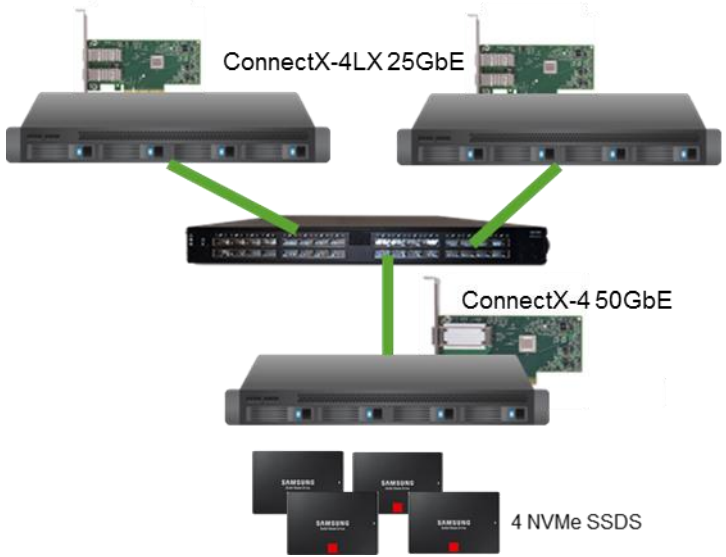
- 740,000 NVMe-oF adapter shipped by 2020
- RDMA NICs will claim >75% of the NVMe-oF market



http://www.storagenewsletter.com/rubriques/market-reportsresearch/nvme-market-at-57-billion-by-2020-with-95-cagr-g2m-research/

NVMe-oF Performance

- Open Source Linux NVMe-oF Software from NVMe.org
  - Accepted in upstream kernel
  - Will be in a future RHEL

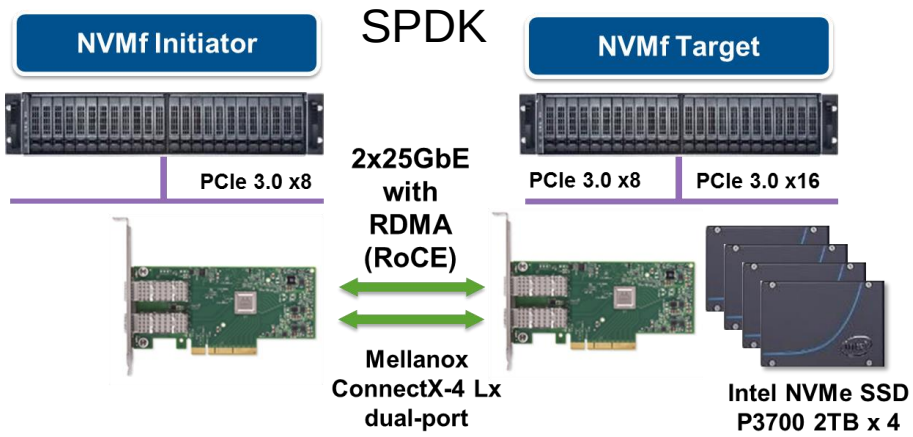


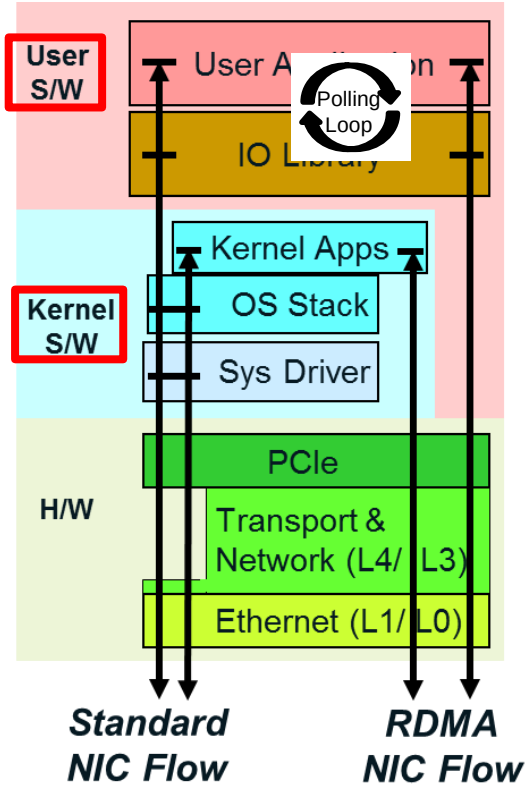
	Bandwidth (Target side)	IOPS (Target side)	Num. Online cores	Each core utilization
BS = 4KB, 16 jobs, IO depth = 64	**5.2GB/sec**	**1.3M**	**4**	**50%**

Added fabric latency
~12us, BS = 512b

https://community.mellanox.com/docs/DOC-2504

Kernel & User Space NVMe-oF

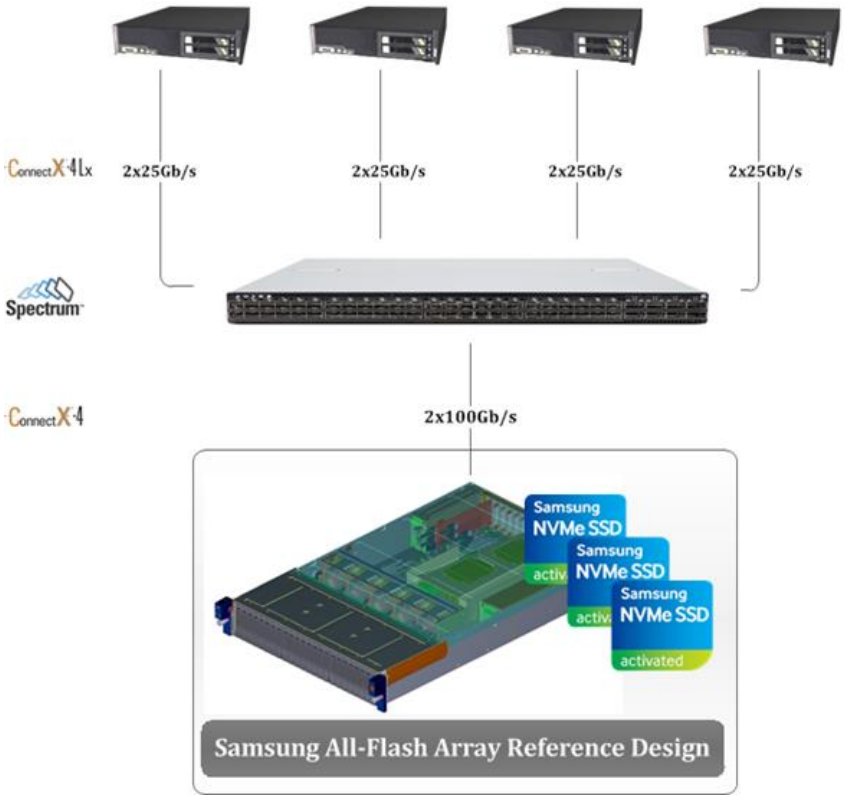




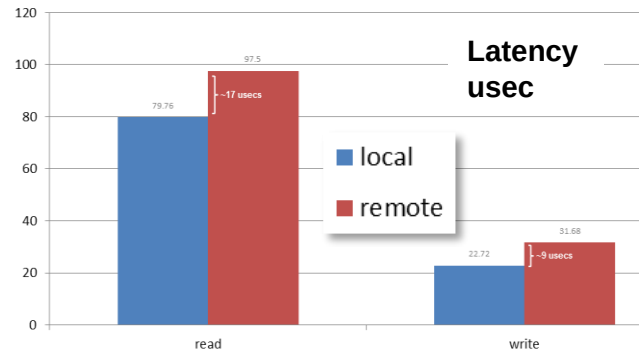
- 93 usec average round trip time measured from Initiator
- ~80 usec spent in reading I/O from NVMe SSD
- Total fabric latency ~12us
- Includes time spent in block layer of initiator submitting and receiving I/Os

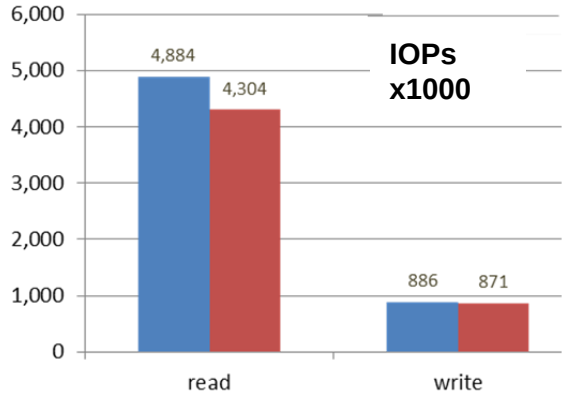
Full Array NVMe-oF Performance Configuration

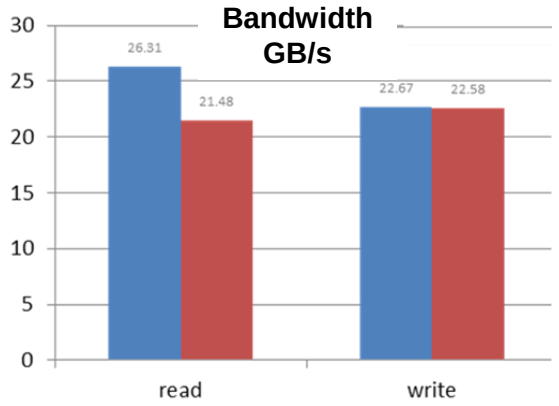
- Configuration
  - 1x NVMf target
  - 24x Samsung PM963 NVMe 2.5” 960GB SSDs
  - 2x 100Gb/s Mellanox ConnectX®-4 EN
  - 4x initiator hosts
  - 2x25Gb/s each
  - Open Source NVMe-oF kernel drivers



NVMe-oF RoCE Performance



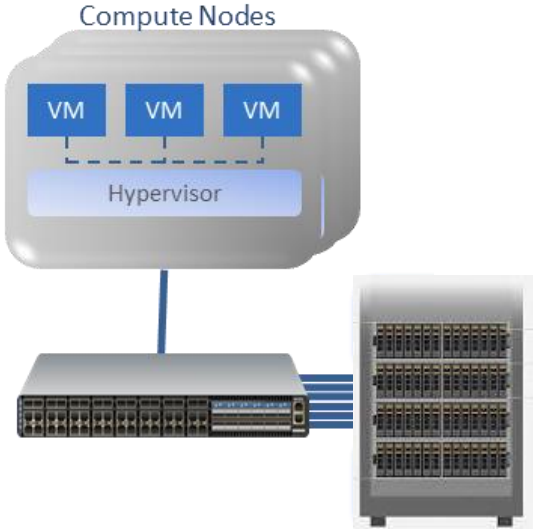




Performance Delta		1-drive	24-drive
Latency	Read	11%	15%
	Write	On par	On par
IOPS	Read	10%	12%
	Write	On par	2%
Throughput	Read	On par	18%
	Write	On par	On par

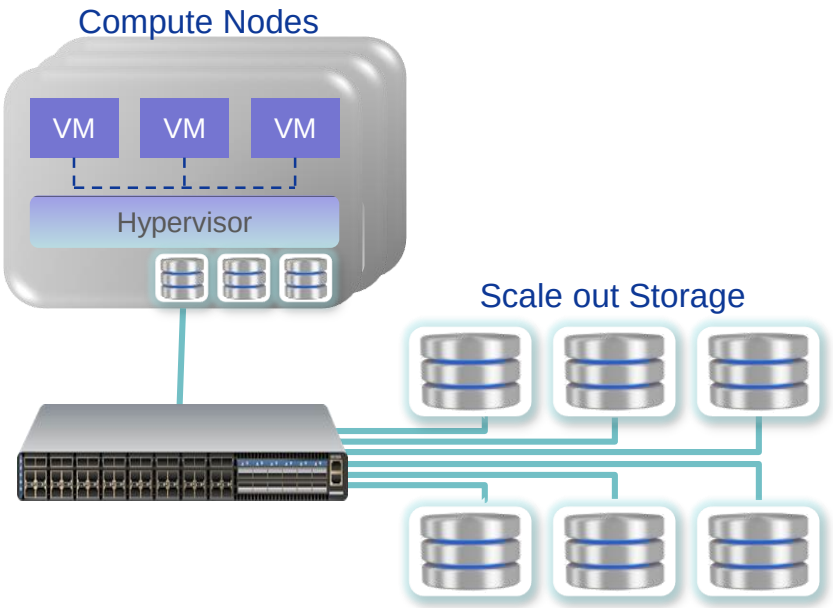
Applications for NVMe-oF

- Scale-Out Storage
  - Low latency
  - High bandwidth & IOPs

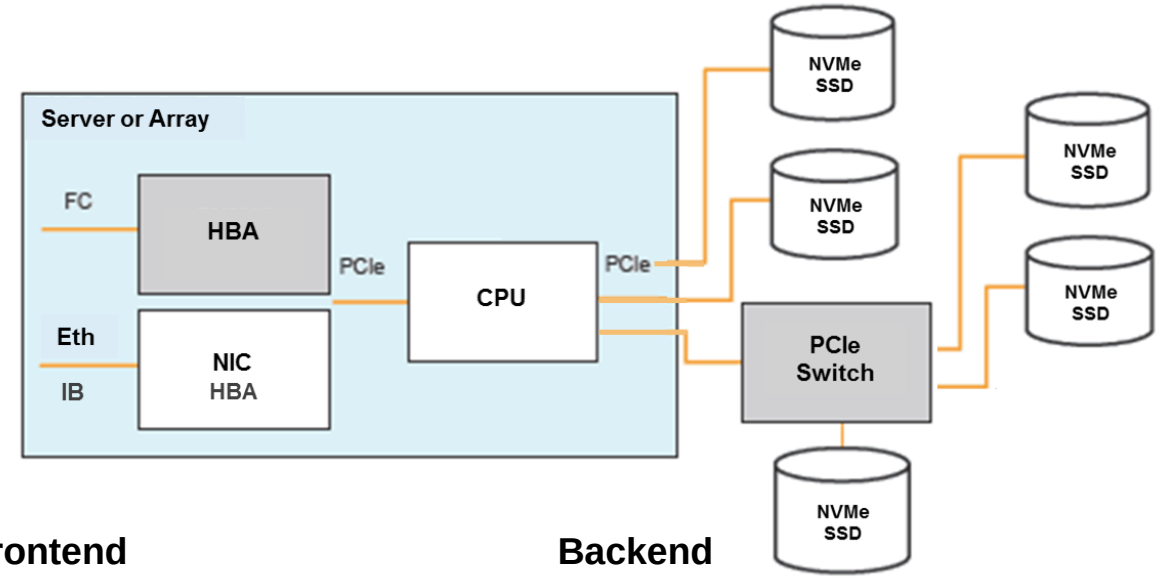


Applications for NVMe-oF

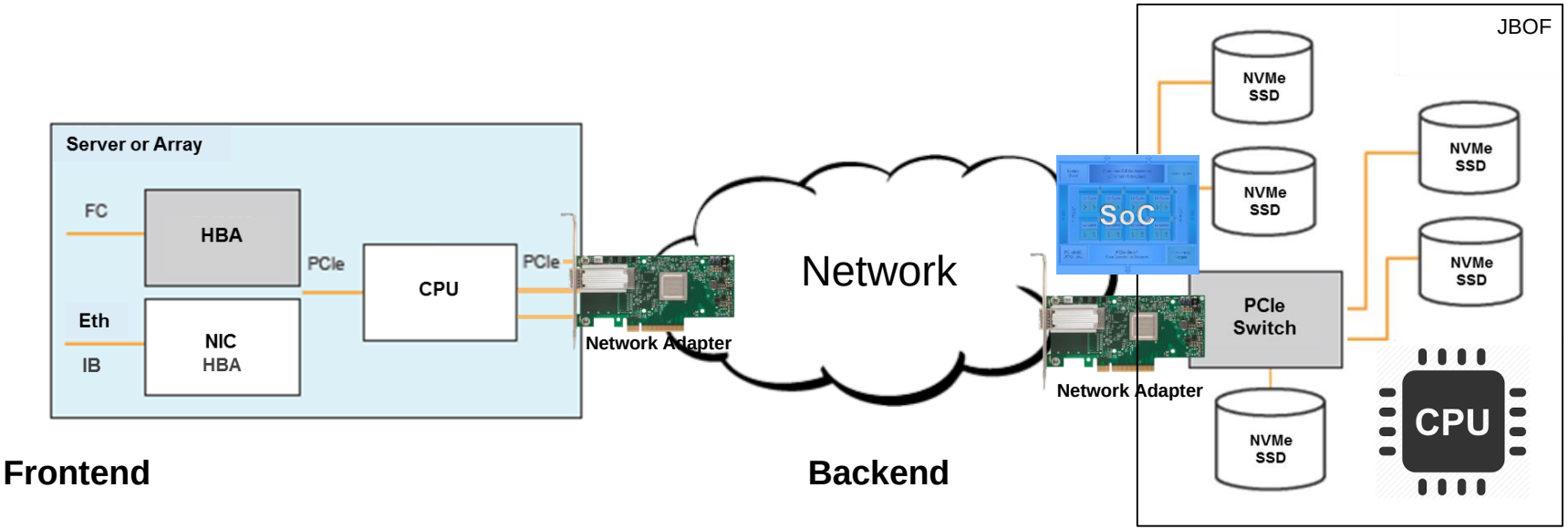
- Hyper-Converged
  - Collapse separate compute & storage
  - Integrated compute and storage nodes
  - Low latency and High bandwidth enable higher performance application support
  - Low CPU utilization



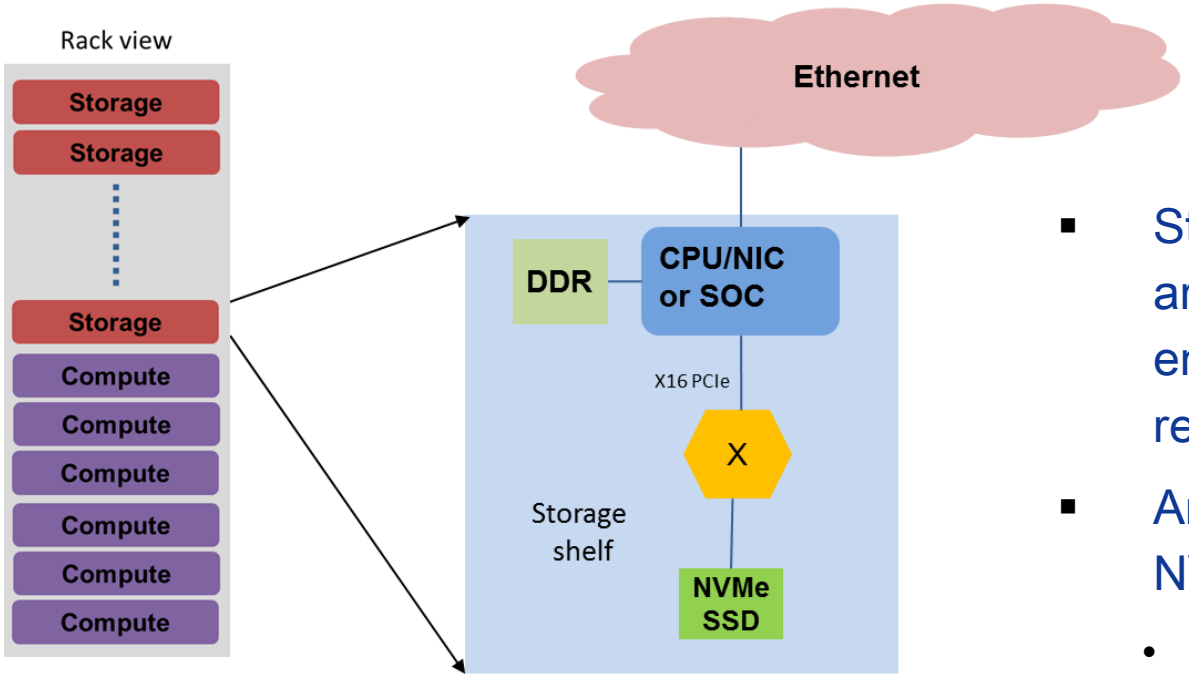
Storage Array or Appliance Backend Scale-out



Storage Array or Appliance Backend Scale-out



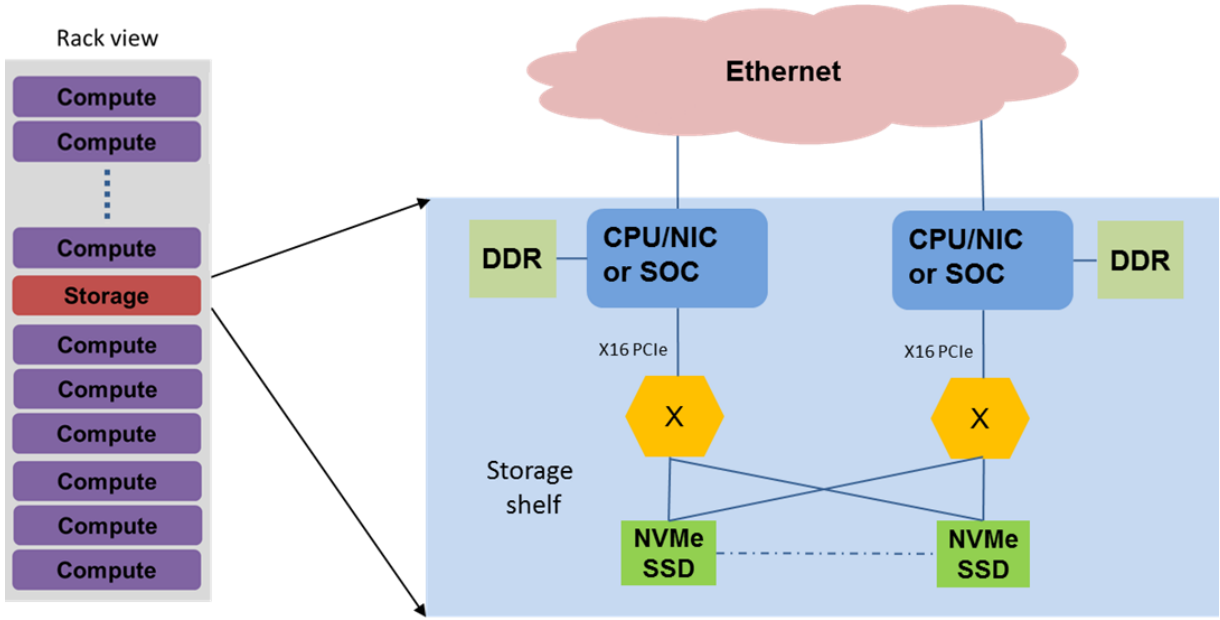
Compute/Storage Disaggregation



- Storage and Compute are not in the same enclosure – DAS replacement
- Architecture enabled by NVMe-oF
  - Low latency and High bandwidth a must

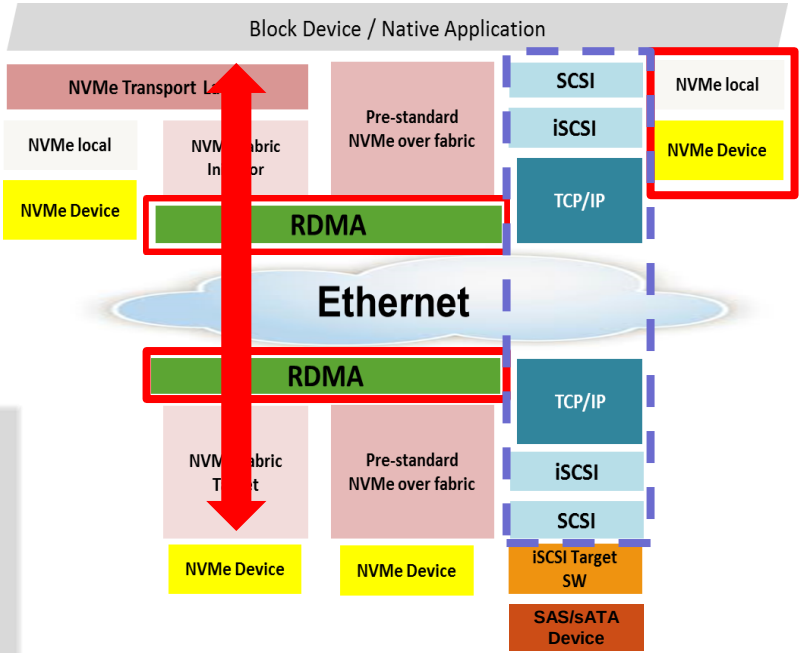
Classic SAN

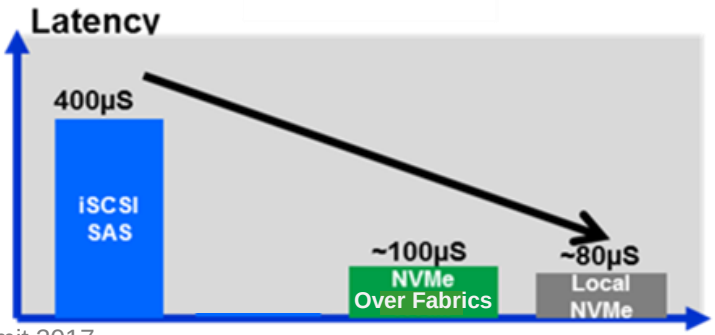
- Better utilization: capacity, rack space, and power
- Better scalability
- Management
- Fault isolation



How Does NVMe-oF Maintain NVMe Performance?

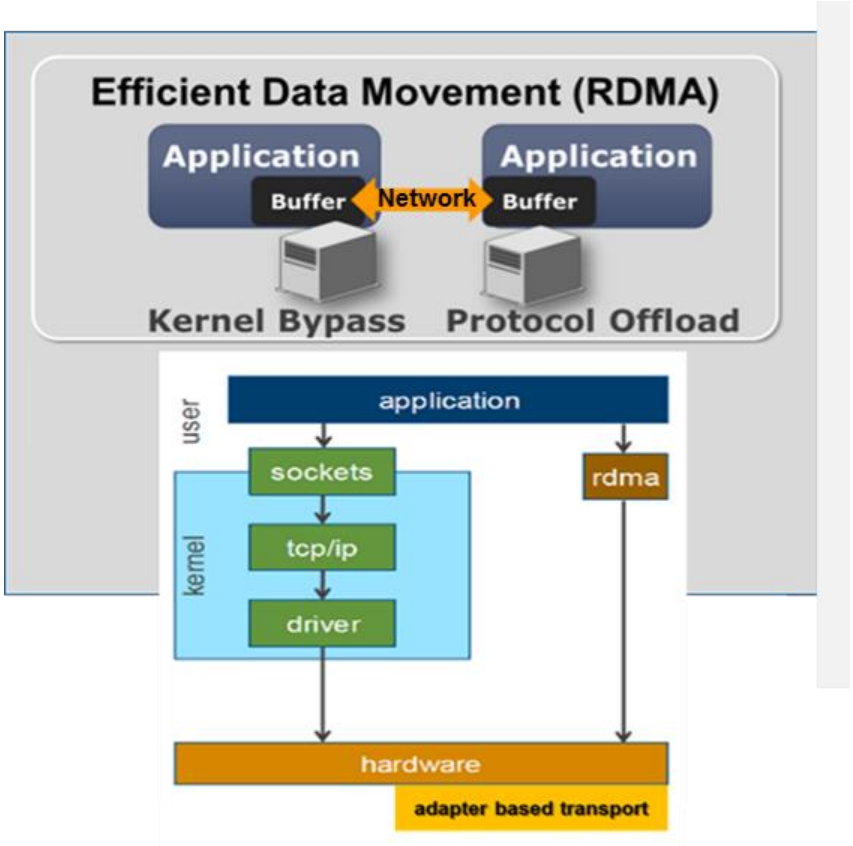
- Extends NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Relies on RDMA for performance
  - Bypassing TCP/IP



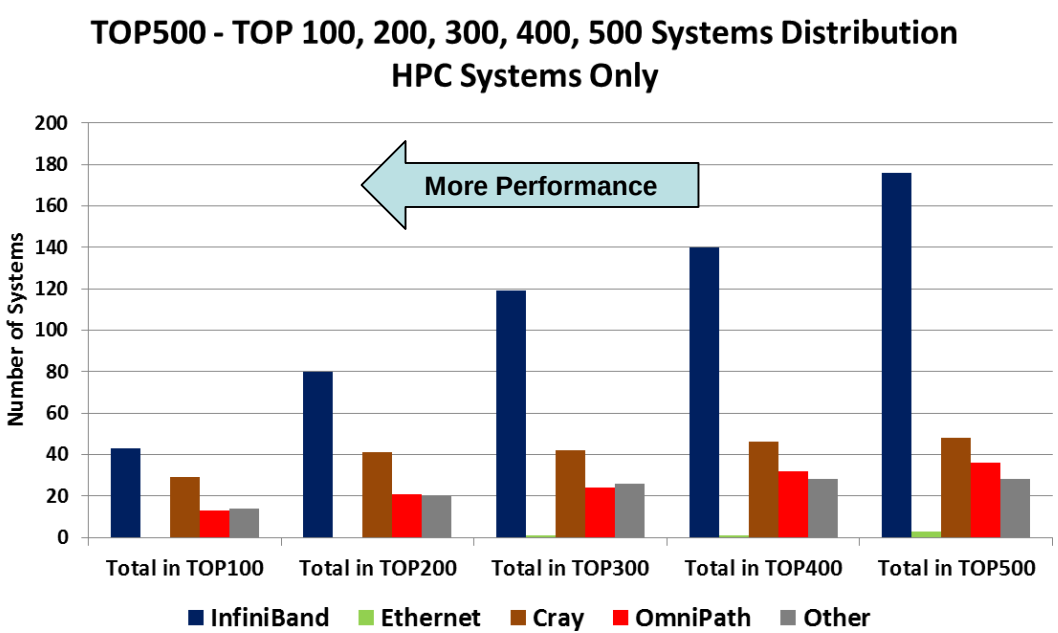


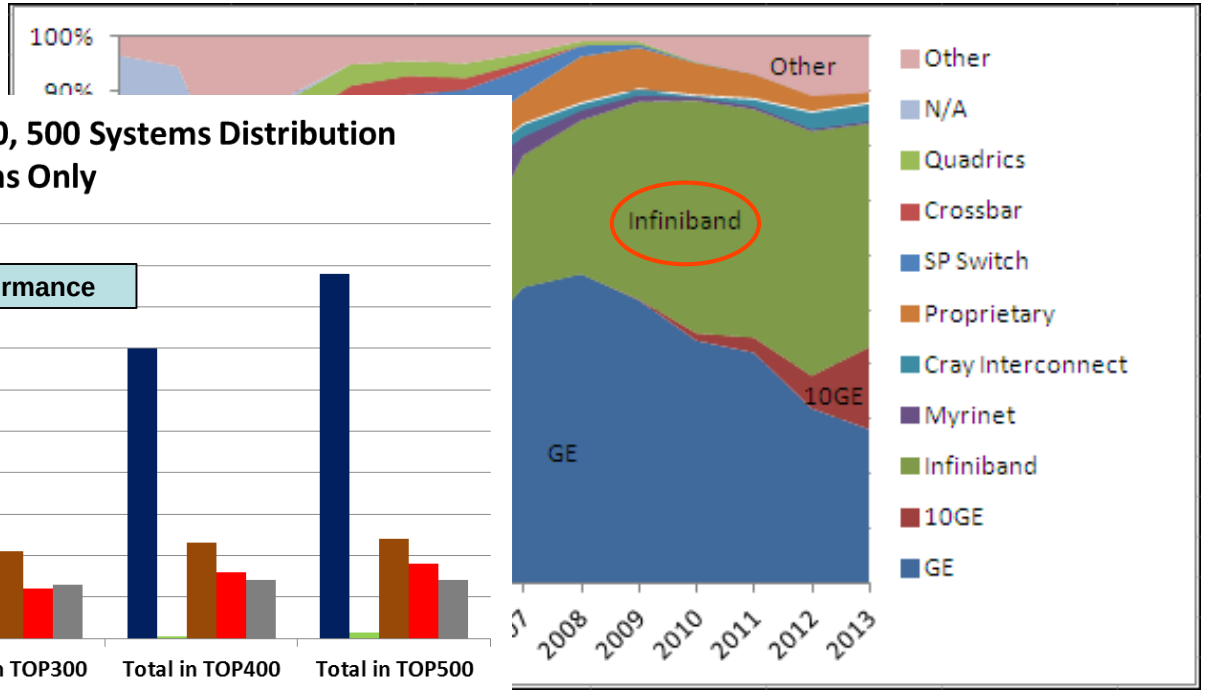
What is RDMA?

- Remote version of DMA(Direct Memory Access)
- Memory to memory move with out CPU
- Transport layer in RNIC
- RDMA protocol is part of the NVMe-oF standard
  - NVMe-oF version 1.0 includes a Transport binding specification for RDMA

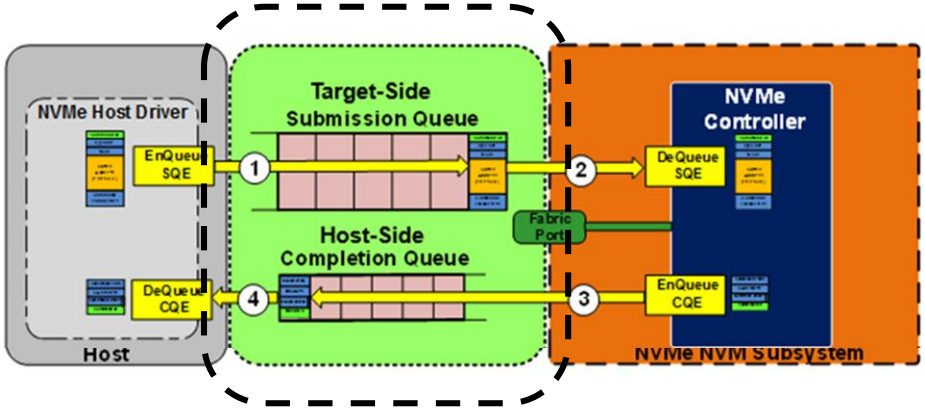


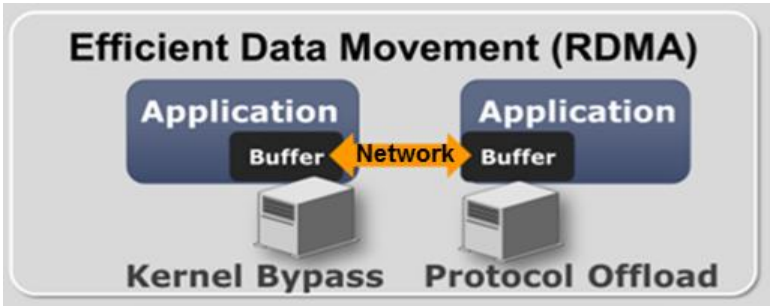
RDMA barrowed from HPC

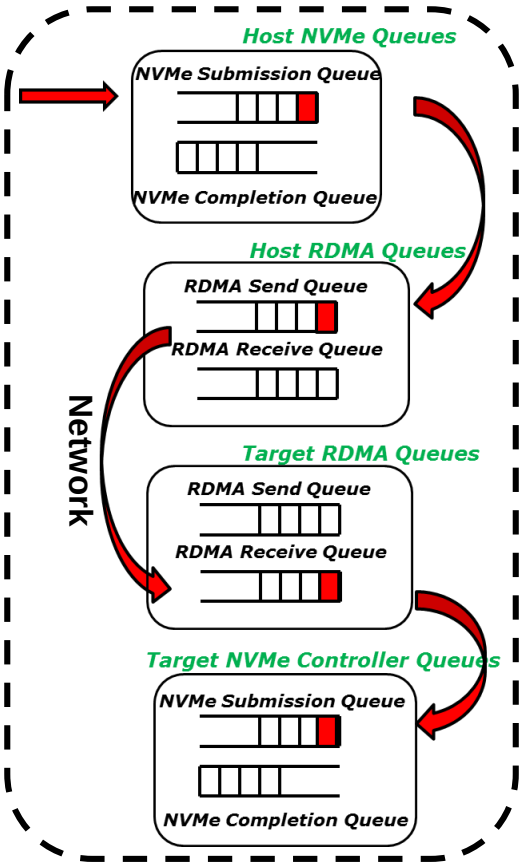




RDMA & NVMe-oF: A Perfect Fit

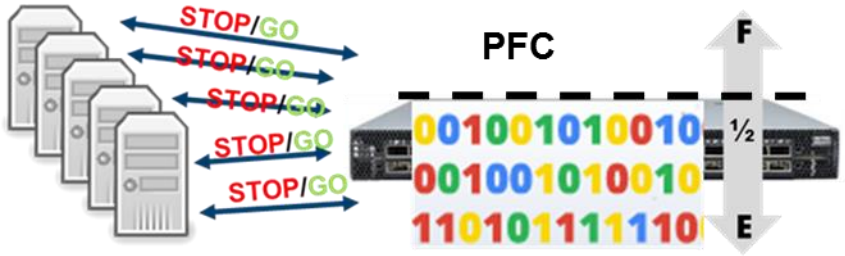


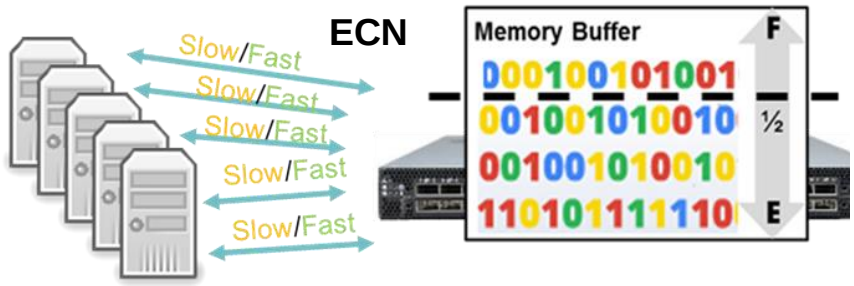




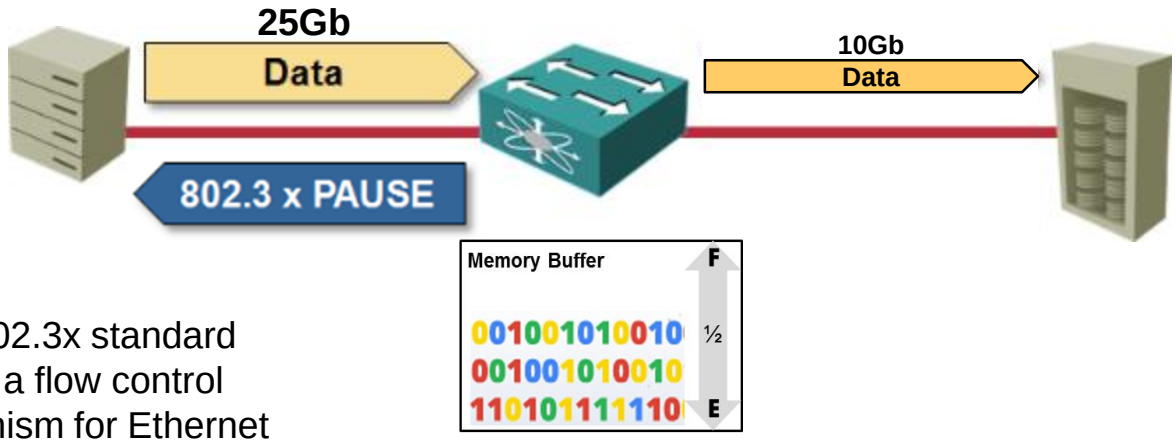
Network Congestion Management with RDMA

- Attention to congestion and data path quality are essential to maintain peak performance with RDMA on Ethernet
- Some of today's RoCE products require a lossless network implemented through PFC(IEEE Priority Flow Control)
- Some can also use ECN(IETF Explicit Congestion Notification) or both





Pause Frame

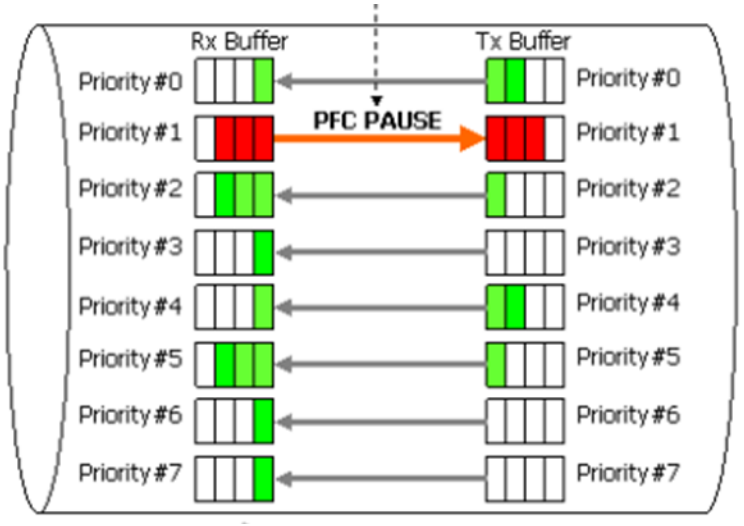


IEEE 802.3x standard defines a flow control mechanism for Ethernet called the pause frame

Priority Flow Control

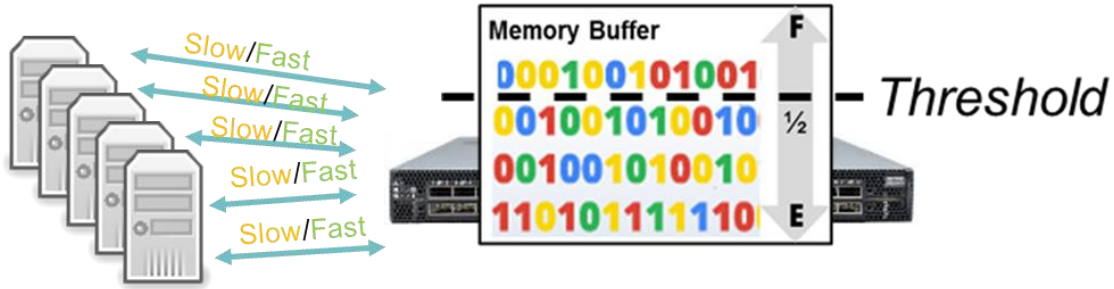
Priority Flow Control (PFC) is similar to 802.3x Pause, except seven priority levels are added. When the data in any of the eight buffers gets to a certain level a pause is sent causing the upstream device to stop sending data only for that priority level for a specified amount of time.

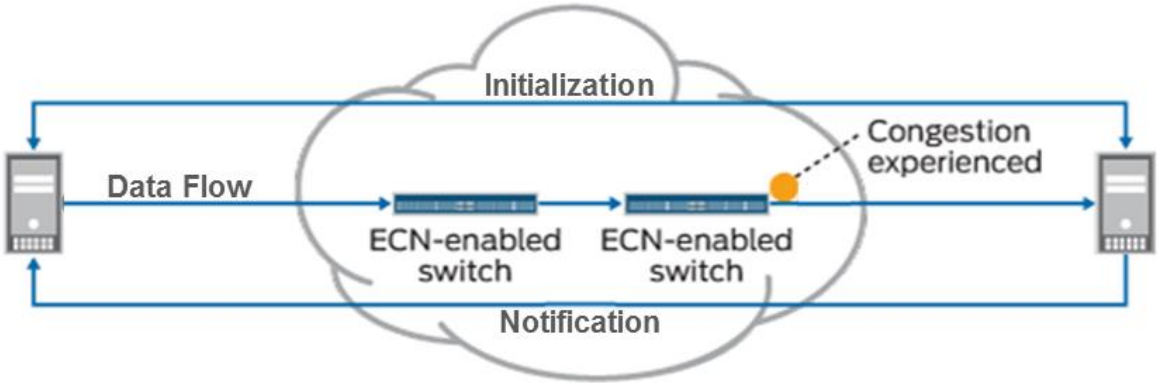
802.1Qbb - Priority-based Flow Control



Explicit Congestion Notification

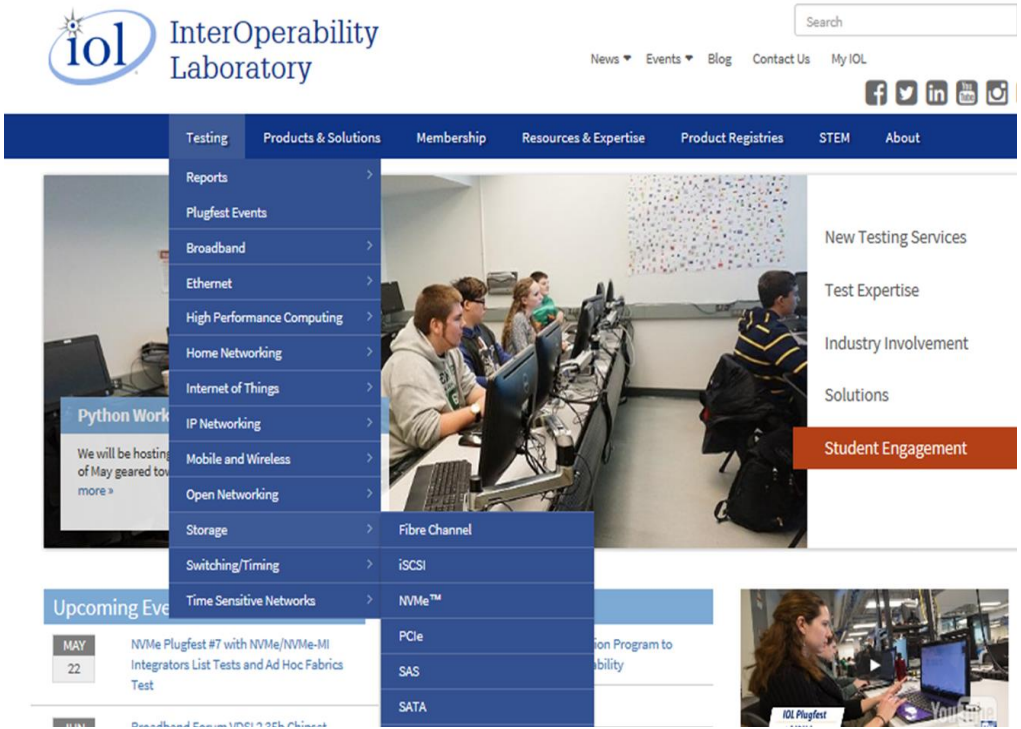
RFC 3168 Explicit Congestion Notification (ECN) slows down a explicit device's data rate that is believed to be overflowing another devices buffer.





UNH IOL Successful Multivendor Interoperability Test

- UNH-IOL neutral multi-vendor interoperability testing since 1988
- This May hosted the first test for NVMe-oF
- Test coincide with regularly scheduled bi-yearly NVMe testing
- Test plan called for participating vendors to mix and match their NICs in both Target and Initiator positions
- Testing was completely successful with near line rate performance at 25Gb/s also achieved



NVMe-oF Products Available Today

Just a sample of the market – not all inclusive list

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

Adapters/SOC

- Mellanox
- Cavium
- Broadcom
- Chelsio

SSD 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 a new protocol called NVMe over Fabrics(NVMe-oF)
- 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?

Thanks!

Rob Davis
Vice President Storage Technology, Mellanox