NVMe™ and NVMe-oF™ in Enterprise Arrays

Sponsored by NVM Express® organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

Speakers

Brandon Hoff

Clod Barrera

Mike Kieran

NVM Express Sponsored Track for Flash Memory Summit 2018

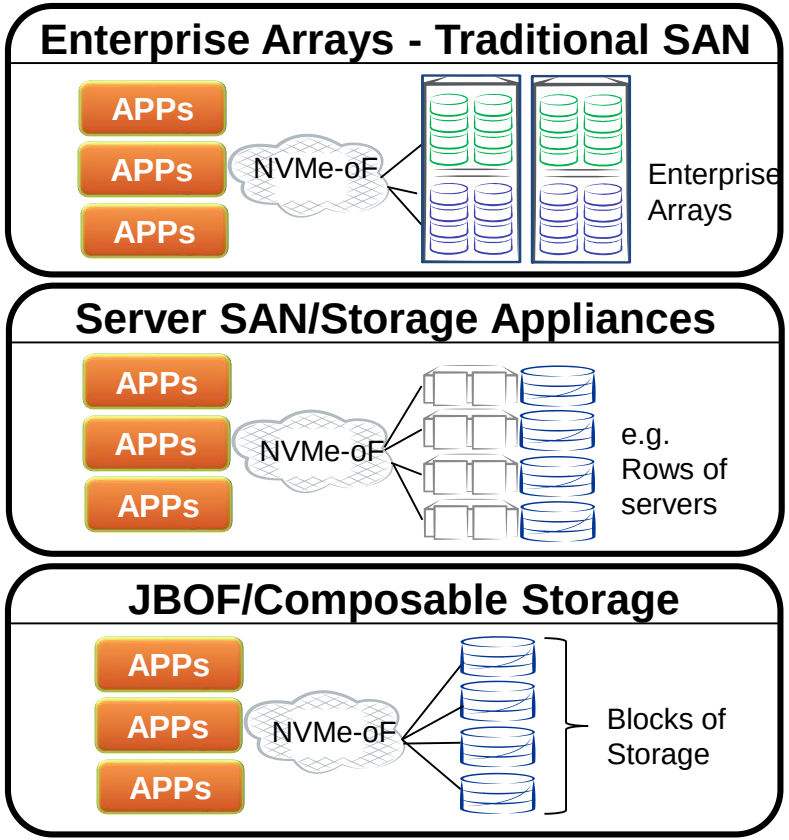
	Track	Title	Speakers	
NVMe-101-1	8/7/18 8:30-9:35	NVM Express: NVM Express roadmaps and market data for NVMe, NVMe-oF, and NVMe-MI - what you need to know the next year.	Janene Ellefson, Micron J Metz, Cisco	Amber Huffman, Intel David Allen, Segate
	8/7/18 9:45-10:50	NVMe architectures for in Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.	Janene Ellefson, Micron Chris Peterson, Facebook	Andy Yang, Toshiba Jonmichael Hands, Intel
NVMe-102-1	3:40-4:45 8/7/18	NVMe Drivers and Software: This session will cover the software and drivers required for NVMe-MI, NVMe, NVMe-oF and support from the top operating systems.	Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee Prewitt, Microsoft	Suds Jain, VMware David Minturn, Intel James Harris, Intel
	4:55-6:00 8/7/18	NVMe-oF Transports: We will cover for NVMe over Fibre Channel, NVMe over RDMA, and NVMe over TCP.	Brandon Hoff, Emulex Fazil Osman, Broadcom J Metz, Cisco	Curt Beckmann, Brocade Praveen Midha, Marvell
NVMe-201-1	8/8/18 8:30-9:35	NVMe-oF Enterprise Arrays: NVMe-oF and NVMe is improving the performance of classic storage arrays, a multi-billion dollar market.	Brandon Hoff, Emulex Clod Barrera, IBM	Mike Kieran, NetApp Brent Yardley, IBM
	8/8/18 9:45-10:50	NVMe-oF Appliances: We will discuss solutions that deliver high-performance and low-latency NVMe storage to automated orchestration-managed clouds.	Jeremy Warner, Toshiba Manoj Wadekar, eBay Kamal Hyder, Toshiba	Nishant Lodha, Marvell Lior Gal, Excelero
NVMe-202-1	8/8/18 3:20-4:25	NVMe-oF JBOFs: Replacing DAS storage with Composable Infrastructure (disaggregated storage), based on JBOFs as the storage target.	Bryan Cowger, Kazan Networks	Praveen Midha, Marvell Fazil Osman, Broadcom
	8/8/18 4:40-6:45	Testing and Interoperability: This session will cover testing for Conformance, Interoperability, Resilience/error injection testing to ensure interoperable solutions base on NVM Express solutions.	Brandon Hoff, Emulex Tim Sheehan, IOL Mark Jones, FCIA	Jason Rusch, Viavi Nick Kriczky, Teledyne

Abstract and Agenda

- Abstract:
  - Enterprise Arrays: NVMe-oF™ and NVMe™ is improving the performance of classic storage arrays, a multi-billion dollar market.

- NVMe-oF Panel
  - Storage Segmentation – Brandon Hoff, Emulex
  - NVMe over Fabrics Overview – Clod Barrera, IBM
  - NVMe over Fabrics on Enterprise Arrays, ANA, and more – Mike Kieran, NetApp
    - Performance Improvements at the Storage Array
  - Performance improvements in NVMe over Fabrics at the initiator and end-to-end – Brandon Hoff, Emulex
    - Performance Improvements in the Sever and End-to-End
  - Q&A

NVMe™ over Fabrics – Storage Architectures



Benefits:

- Storage services (dedup, compression, thin provisioning)
- High availability at the array
- Fully supported from the array vendor
- Example: NetApp/IBM

Benefits:

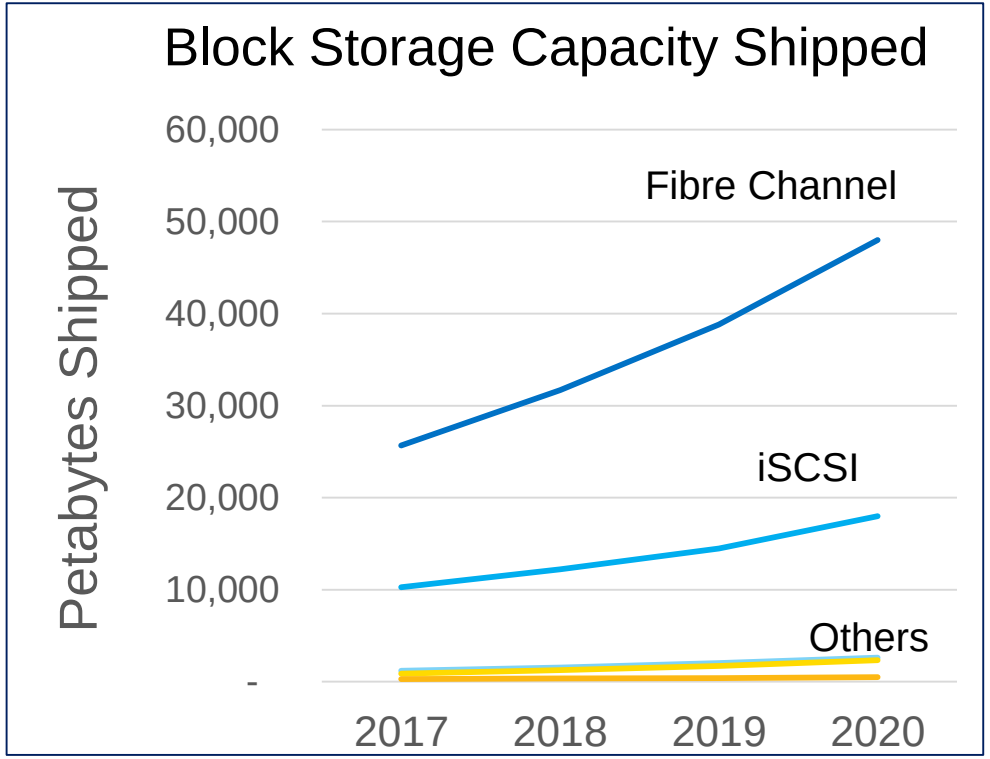
- High performance storage
- Lower cost that storage arrays, minimal storage services
- Roll-your-own support model
- Ex. SUSE on Servers configured to be storage targets

Benefits:

- Very low latency
- Low cost
- Great for a single rack/single switch
- Leverages NICs, smart NICs, and HBAs for NVMe-oF to PCIe®/NVMe™ translation

Enterprise Storage Market

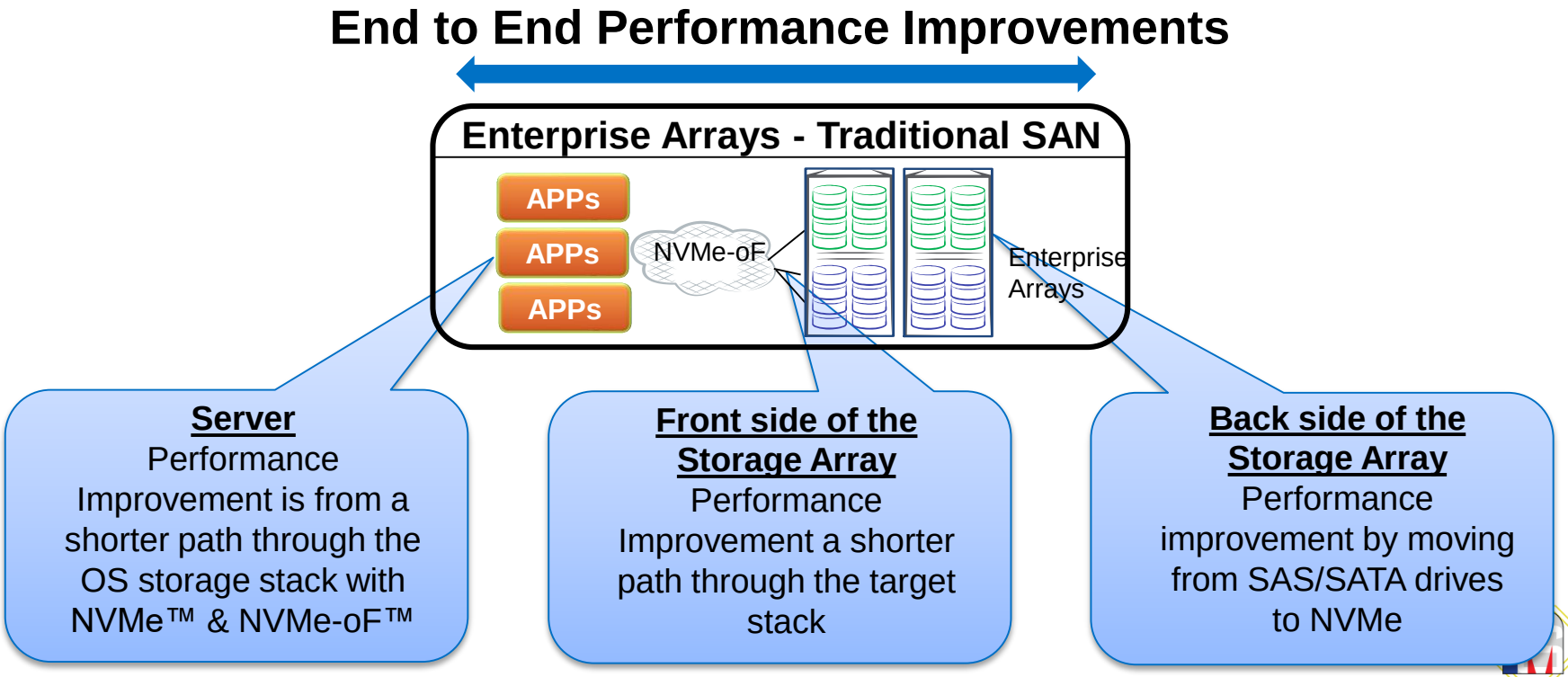
- Fibre Channel storage shows strong growth in capacity
  - Fibre Channel Storage capacity shipped is larger than all other types of external storage combined
- The adoption of All Flash Arrays and NVMe™ storage will drive the need for faster networks
- iSCSI is the dominate technology block over Ethernet
- The only RDMA market for block storage is Infiniband



Other Includes: FICON, FCoE, Infiniband, External SAS

IDC WW Capacity Shipped, 2016

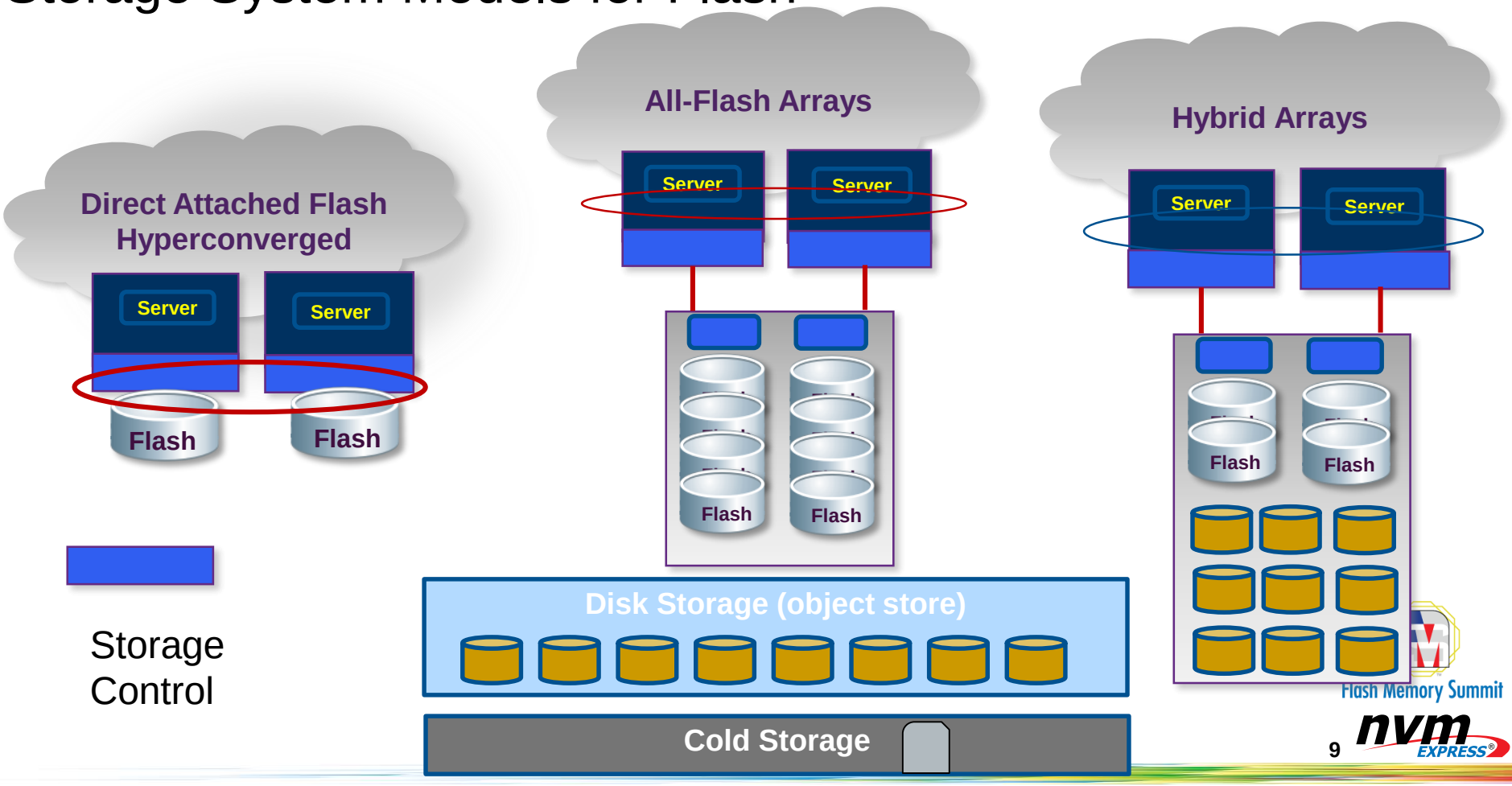
Three Areas of Performance Improvement



NVMe™ over Fabric for Enterprise Arrays

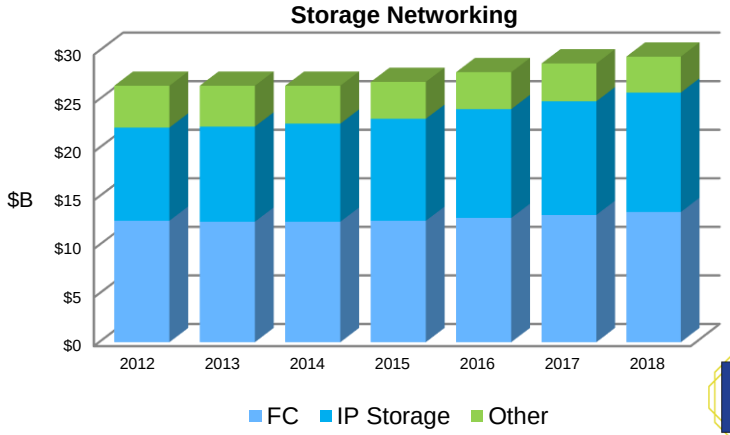
Clodoaldo Barrera and Brent Yardley, IBM

Storage System Models for Flash



Directions in Storage Networking

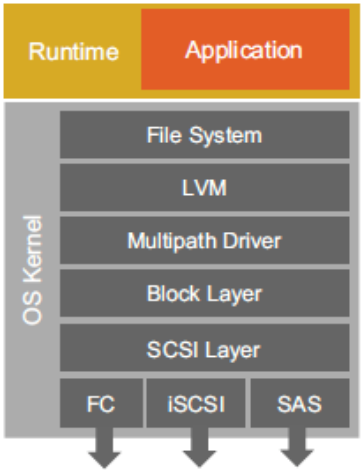
- **10GE ->100GE dominates the Cloud infrastructure**
  - CSPs adopt new Ethernet technology faster than Enterprise
  - Less constrained by legacy install base.
- **FC continues link speed generations (now on Gen 6 at 32Gbps**
  - Expect gradual decline in FC SAN share of storage attachment
  - Storage fabrics for new workloads, CSPs, Cold storage all favor IP storage attach – iSCSI, NAS, and REST Object Storage APIs.



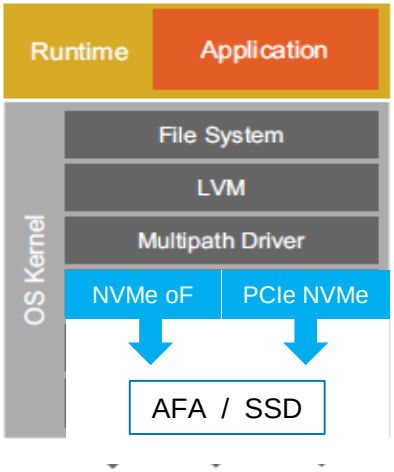
NVMe™ and NVMe-oF™

- NVMe protocol enables native parallelism within SSDs and All Flash Arrays (AFA)
- NVMe allows more efficient host software stacks for lower latency at application
- User-space drivers for selected software (e.g. In-memory DB) for maximum benefit

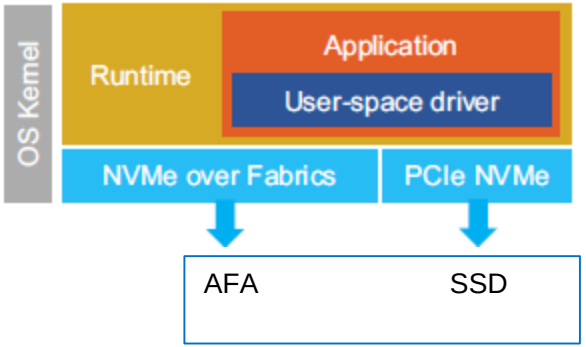
SCSI SAN/Local Storage



NVMe SAN/Local Storage

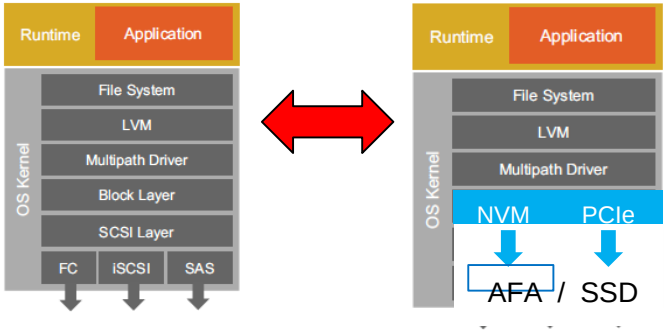


New Paradigm

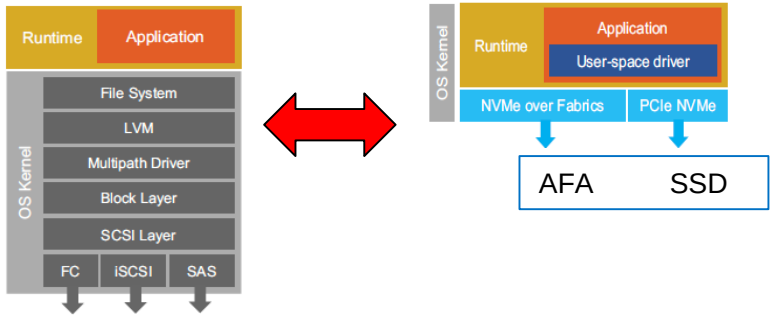


"IBM Storage and the NVM Express Revolution" Koltsidas & Hsu 2017 – IBM Redpaper

NVMe-oF™ Performance Benefits



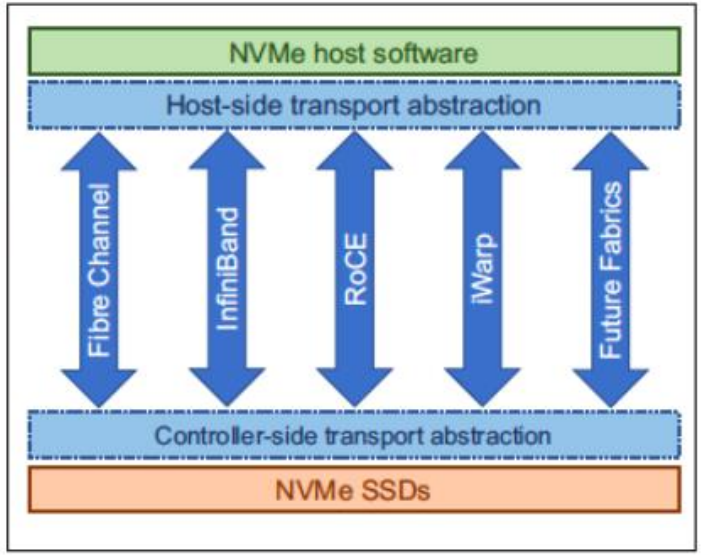
- NVMe™ and NVMe-oF have new kernel driver stacks in hosts to reduce lock contention and increase parallelism. Improved throughput and lower latency.
- For I/O-bound workloads, NVMe-oF lowers server I/O load and wait times.
- IBM benchmark on 16Gb FC and IBM FlashSystem AFA showed 30% lower CPU utilization from I/O



- From IBM Research – Spark application with RDMA connection to storage from user space showed up to 5X improvement in performance.
- Requires complete re-structure of I/O system and application awareness/modification

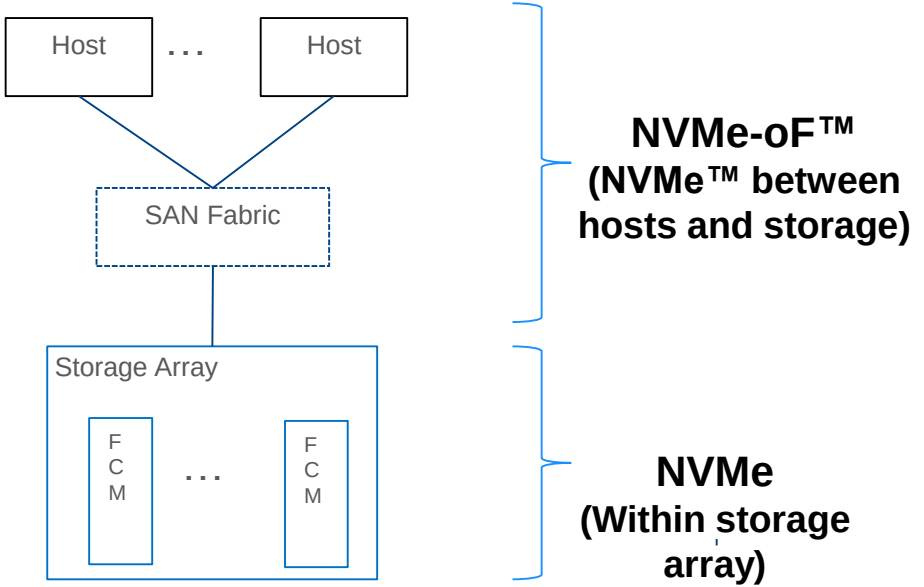
NVMe™ and NVMe™ over Fabric

- **Fast Media requires a new protocol with Memory/Storage semantics**
- **NVMe** is a new block memory/storage protocol that replaces SCSI. Flash storage is capable of higher IOP performance, throughput, and parallelism not possible on HDDs
- **NVMe over PCIe** – PCIe provides short distance connection for a processor to a small number of NVMe devices (SSDs)
- **NVMe-oF** - NVMe protocol is mapped to a fabric for distance and fanout. Supported fabrics include FC (Gen 5,6), Ethernet or IB SAN



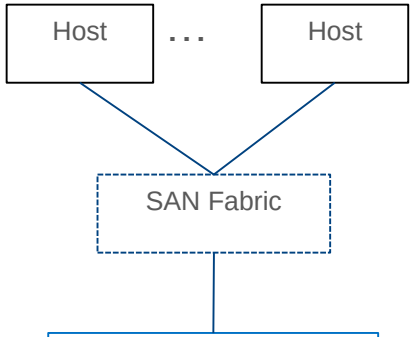
The Benefits of Continuity

- Storage Fabrics are a significant client investment
  - Management of full storage path
  - Performance and availability management
  - Audit controls
  - Upgrade migration process
  - Application and middleware compatibility testing
  - Security verification
  - Etc.…



Value of NVMe™ and NVMe-oF™

- Optimized for Flash
- Fast and Getting Faster
- Reduce Application License costs
- Future proof investment
- NVMe end-to-end strategy



NVMe-oF
(NVMe between hosts and storage)

NVMe
(Within storage array)

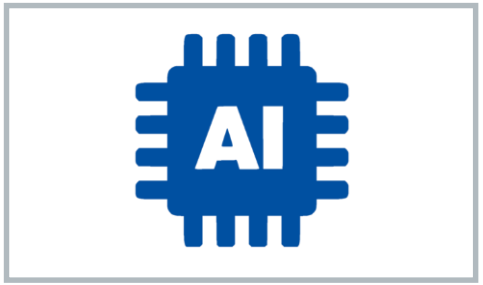
NVMe and NVMe-oF in Enterprise Arrays

Mike Kieran, Technical Marketing Engineer, NetApp

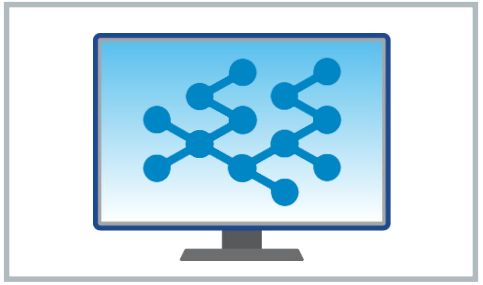
Real-Time Applications: The Next Phase of Digital Transformation

In-memory technologies will grow to ~$13B by 2020*

Artificial Intelligence



Machine Learning



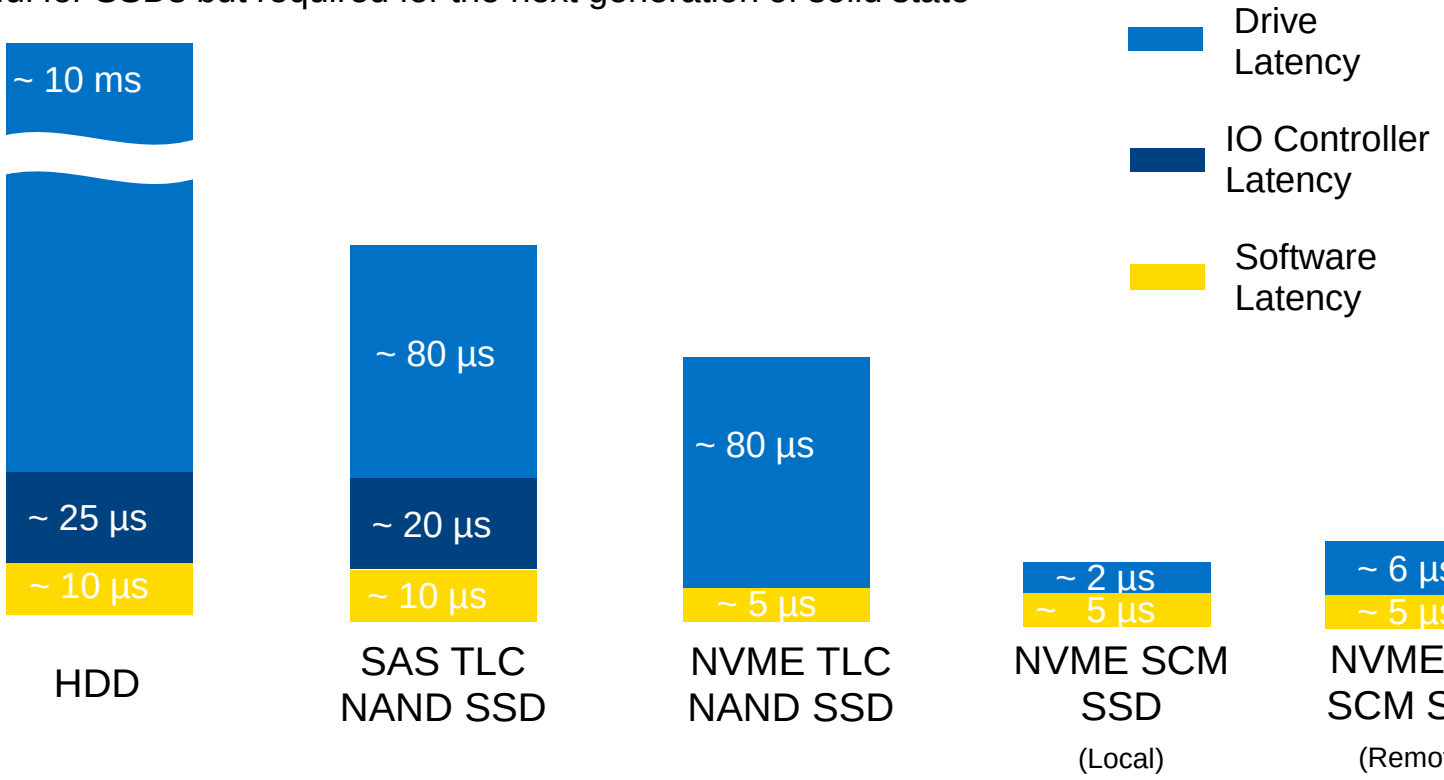
Real-Time Analytics

All demand lower latency and higher performance
from faster fabrics and faster media

* Gartner, Inc., Market Guide for In-Memory Computing Technologies, 16 January 2017

Impact of NVMe™ For Media Access

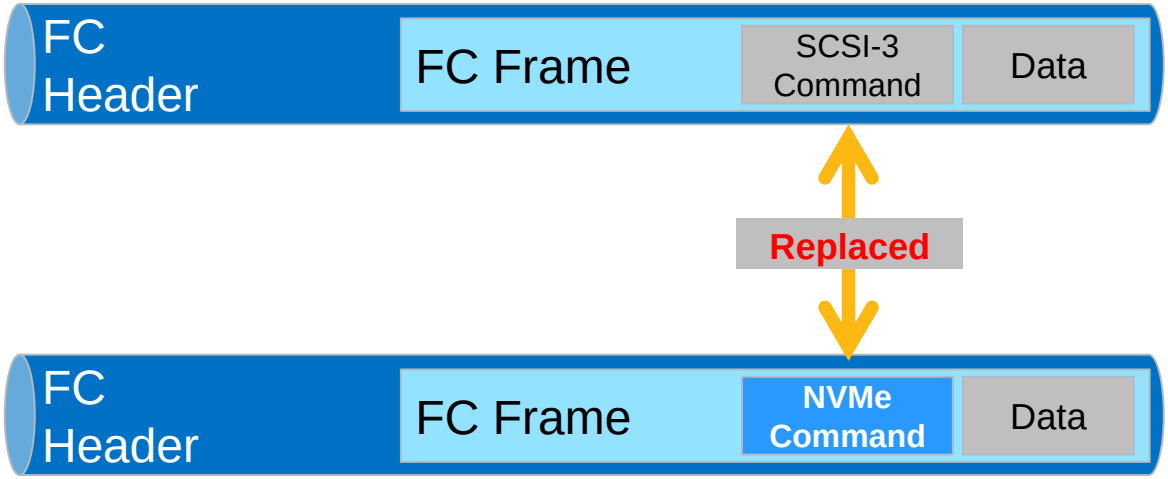
NVMe useful for SSDs but required for the next generation of solid state



NextGen Blocks - NVMe™

What are NVMe-oF™ and FC-NVMe?

- FCP - SCSI-3 command set encapsulated in an FC frame

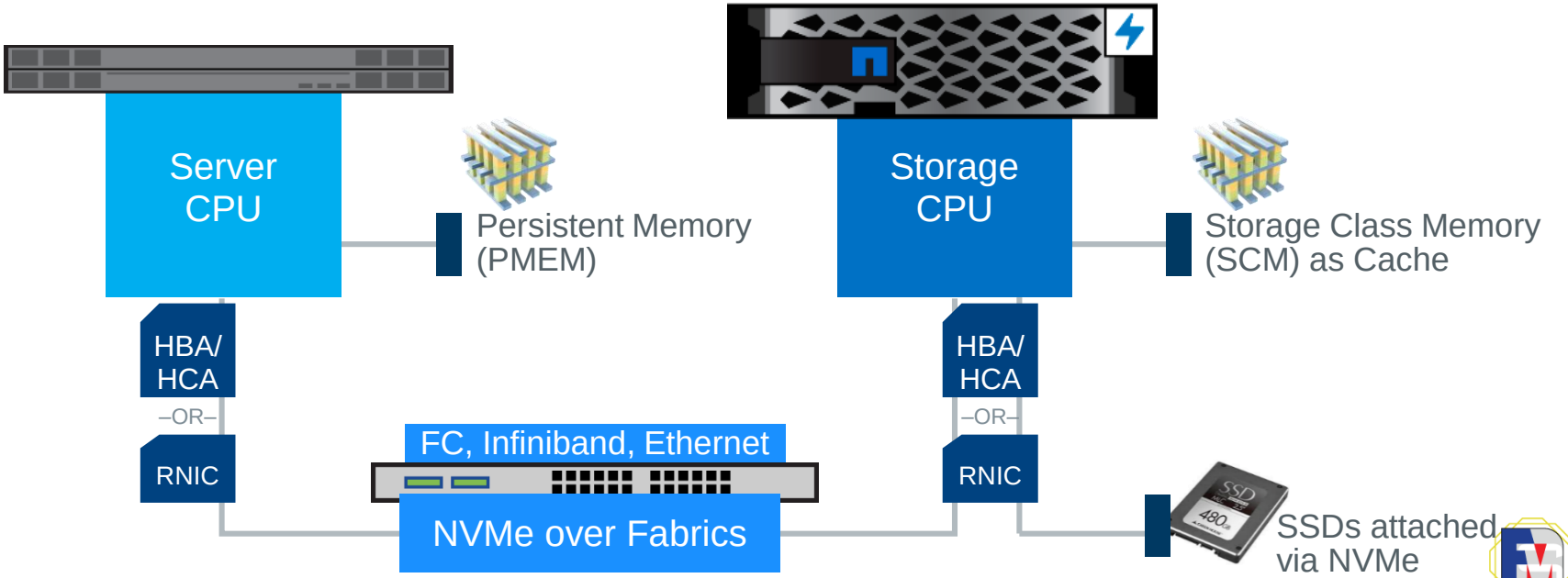


- FC-NVMe - NVMe command set encapsulated in an FC frame

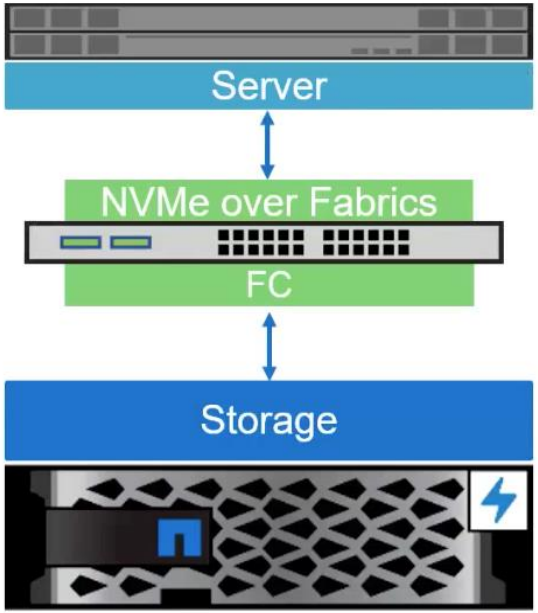
- Replaces SCSI-3 CDBs in a FC Frame
- Substantial performance boost because of:
  - Command streamlining
  - Reduced context switches
  - Increased multithreading - 64,000 queues with a maximum queue depth of 64,000

NetApp's NVMe™ Vision

Driving real value out of new technologies requires significant investment on multiple fronts from a market leader



FCP (SCSI) vs. NVMe™/FC Performance and Latency



NVMe™ Vocabulary Update

Getting used to new terminology as we migrate from SCSI to NVMe-oF™

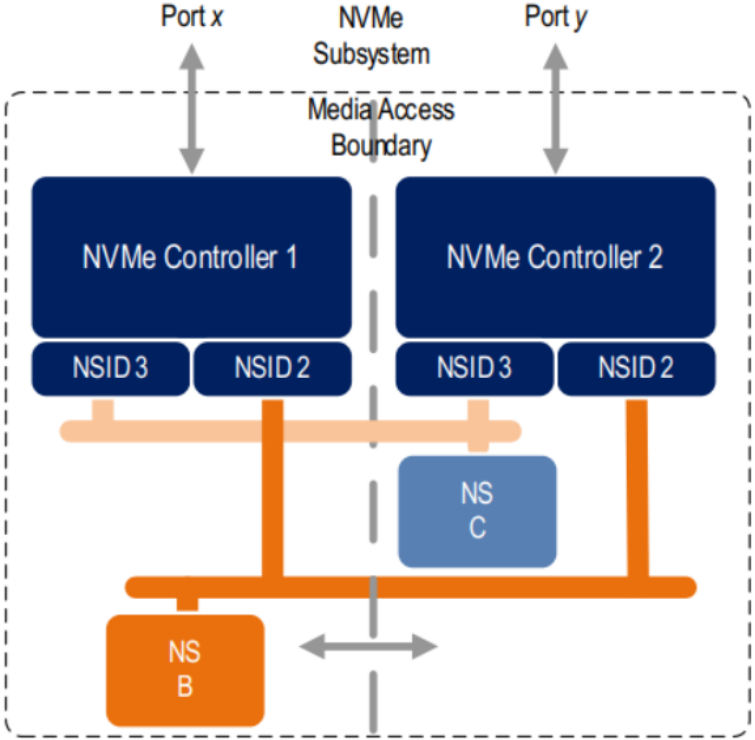
Protocol	Type	Example
NVMe	NQN	nqn.2014-08.com.vendor:nvme:nvm-subsystem-sn-d78432
iSCSI	IQN	iqn.1991-05.com.microsoft:dmrtk-srvr-m

FC	FC-NVMe
LUN	Namespace
WWPN	NQN
igroup	Subsystem
ALUA	ANA*

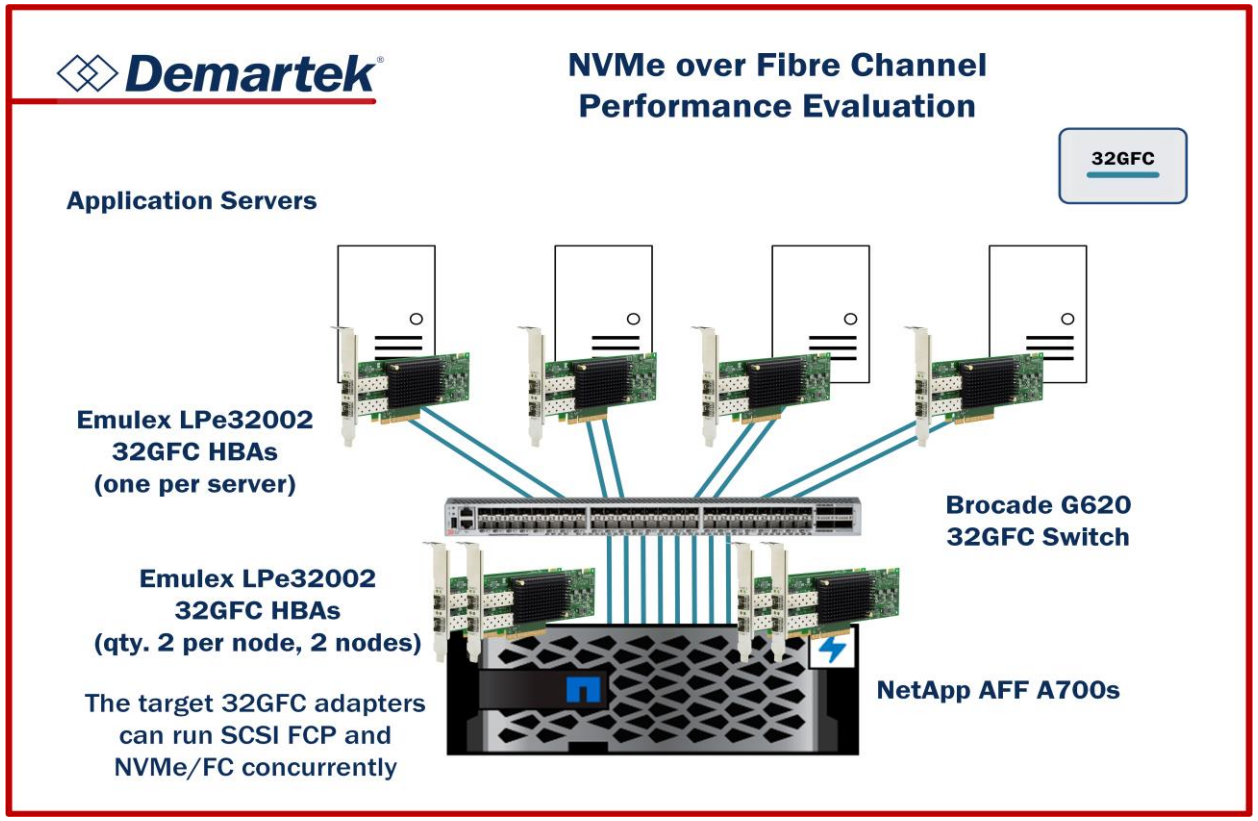
* Asymmetric Namespace Access (NetApp defined multipathing protocol for NVMe. Currently out for ratification by NVM Express® organization.

Ratified: Asymmetric Namespace Access

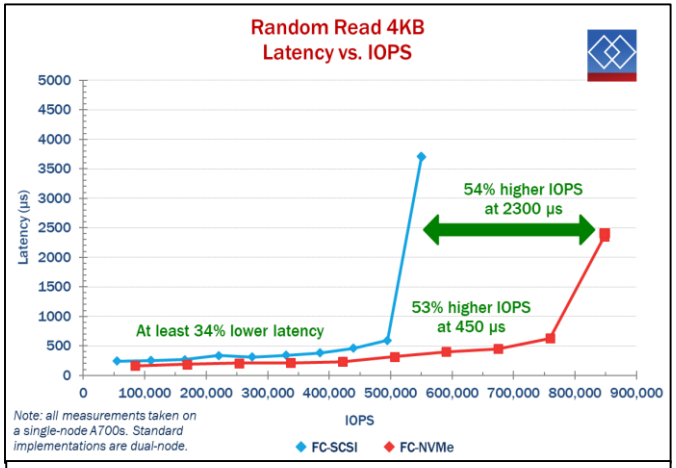
- Concept: Namespaces with multiple paths may have asymmetric properties
- Base protocol is ratified
- Domains and partitioning work is next

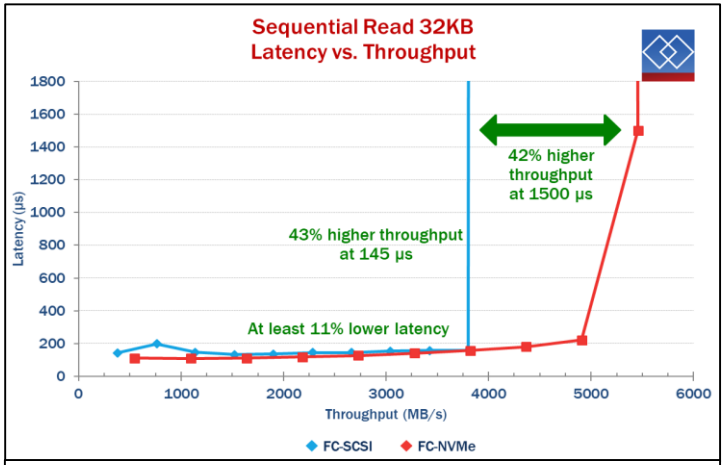


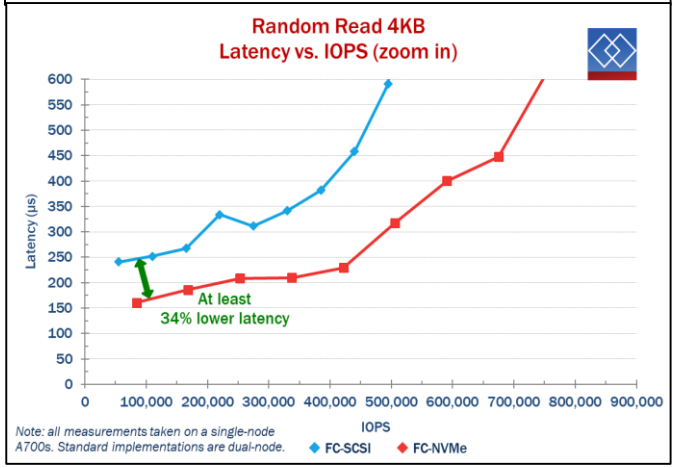
NVMe™ over Fibre Channel Performance Test

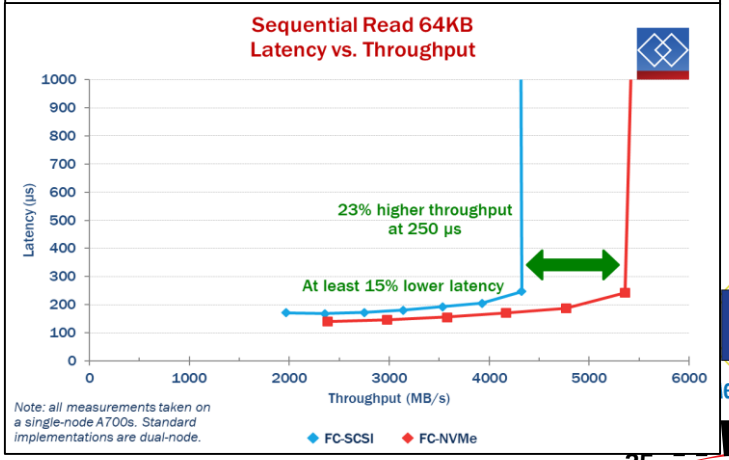


NVMe™ over Fibre Channel Performance on a A700s single node









Performance Improvements at the Initiator, and general storage performance improvements with NVMe over Fabrics

Server Test Configuration – Initiator performance

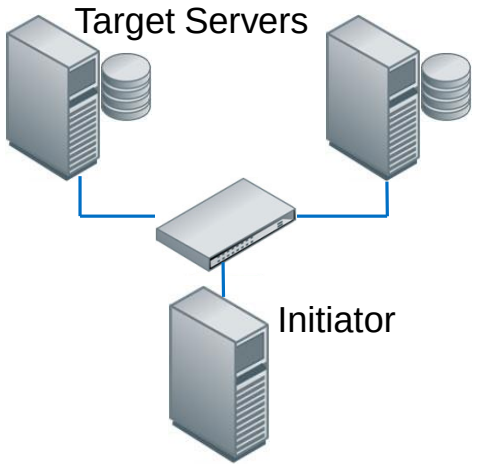
Target Servers – Qty 2

- Dual CPU - Purley
- 32G Dual-Port LPe32002 – 1 Port in use
- RHEL7.4 w/OCS-RAMd (SCSI Target)
- SLES12SP3 w/LPFC-T (NVMe Target)

Initiator

- Dual CPU - Purley
- 32G Dual-Port LPe32002 – 1 Port in use
- SLES12SP3 w/LPFC Driver (v.12.0.141.2)

Test Parameters: 32 threads and queue depth = 32



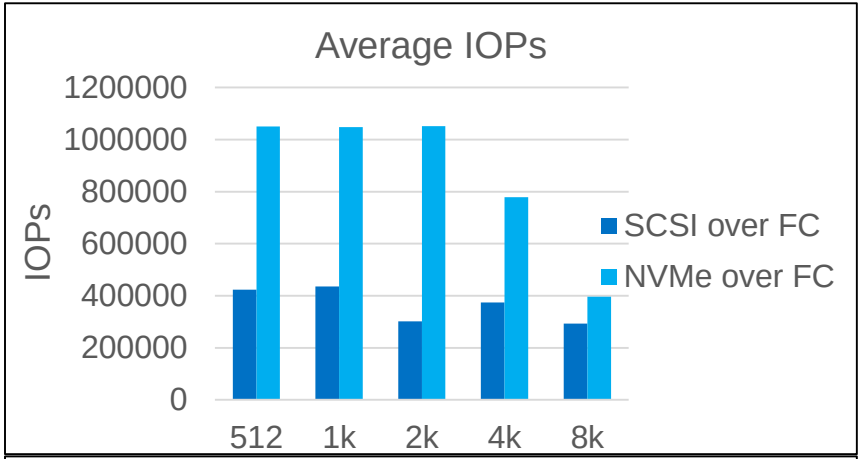
NVMe-oF™: Lean Stack Delivers more IOPs with less CPU

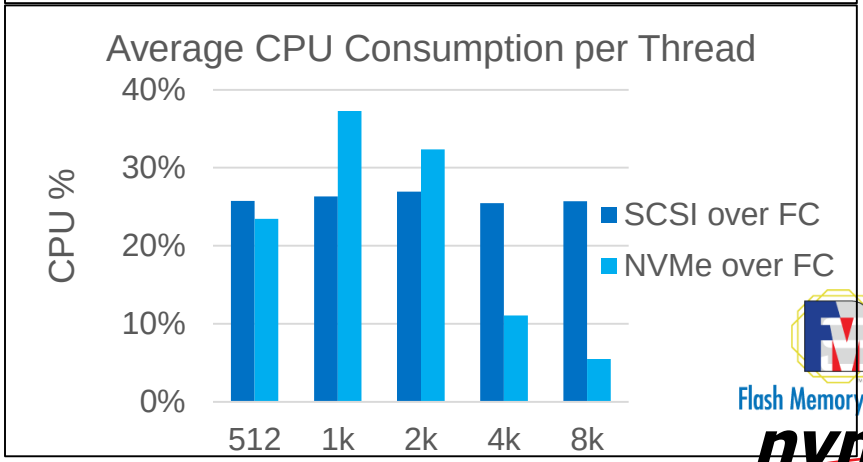
Customer Comments

- "NVMe™ over Fabrics delivers more transactions on the same storage footprint"
- "Our storage strategy going forward is based on NVMe over Fabrics," Large Health Care provider

Performance Benefits

- On average 2x-3x more IOPs at the same CPU consumption
- At 4k, we see 2x the IOPs at 50% of the CPU consumption





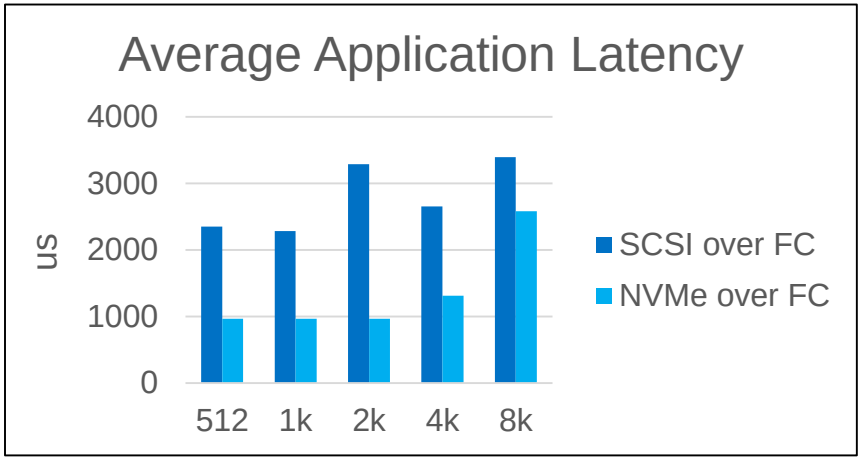
NVMe-oF™: Just runs faster

Application Latency: Response time as seen by the server application

- A function of the number of outstanding Ios
- For this example, 32 (QD) x 32 threads, which means 1024 outstanding IOs

Single IO Latency: Function of what the hardware can do

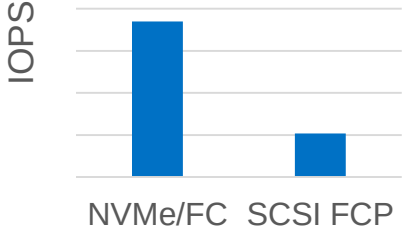
NVMe™ benefits from increased parallelization



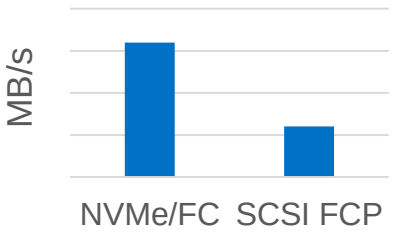
Performance Improvement of NVMe™ over Fabrics – End to End

NVMe/FC Vs. SCSI/FC Performance Improvement on the same hardware

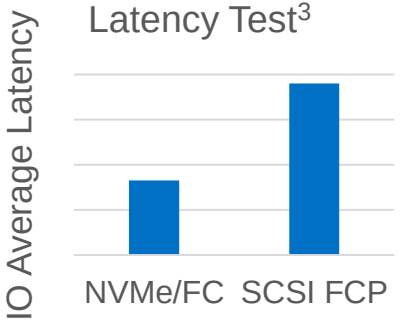
Simulated OLTP Workload IOPS[1]



Data Warehouse IO Throughput[2]



Batch Transaction Latency Test[3]



3.6x More Transactions

2.7x Higher Throughput

½ The Latency

[1]4K Random Read IOs, 16 Threads, Queue Depth of 16

[2]64K Random Read IOs, 16 Threads, Queue Depth of 16

[3]4K Random Read IOs, 8 Threads, Queue Depth of 1

Contact Information

For more information please contact the following:

Brandon Hoff	brandon.hoff@broadcom.com
Clod Berrera	barrerac@us.ibm.com
Mike Kieran	Michael.Kieran@netapp.com


nvm
EXPRESS®
Architected for Performance