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

EMULEX

IBM

NetApp
<table>
<thead>
<tr>
<th>Track</th>
<th>Title</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe-101-1</td>
<td>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.</td>
<td>Janene Ellefson, Micron J Metz, Cisco</td>
</tr>
<tr>
<td></td>
<td>8/7/18 9:45-10:50 NVMe architectures for in Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.</td>
<td>Janene Ellefson, Micron Chris Peterson, Facebook</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee Prewitt, Microsoft</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Brandon Hoff, Emulex Fazil Osman, Broadcom J Metz, Cisco</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Brandon Hoff, Emulex Clod Barrera, IBM</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Jeremy Warner, Toshiba Manoj Wadekar, eBay Kamal Hyder, Toshiba</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Bryan Cowger, Kazan Networks</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td>Brandon Hoff, Emulex Tim Sheehan, IOL Mark Jones, FCIA</td>
</tr>
</tbody>
</table>
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

**Enterprise Arrays - Traditional SAN**

- APPs
- NVMe-oF
- Enterprise Arrays

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

**Server SAN/Storage Appliances**

- APPs
- NVMe-oF
- e.g. Rows of servers

**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

**JBOF/Composable Storage**

- APPs
- NVMe-oF
- Blocks of Storage

**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

![Graph showing block storage capacity shipped from 2017 to 2020 for Fibre Channel, iSCSI, and Others. Other includes: FICON, FCoE, Infiniband, External SAS. IDC WW Capacity Shipped, 2016.]
Three Areas of Performance Improvement

End to End Performance Improvements

Enterprise Arrays - Traditional SAN

- APPs
- APPs
- APPs

Server
Performance Improvement is from a shorter path through the OS storage stack with NVMe™ & NVMe-oF™

Front side of the Storage Array
Performance Improvement a shorter path through the target stack

Back side of the Storage Array
Performance improvement by moving from SAS/SATA drives to NVMe
NVMe™ over Fabric for Enterprise Arrays

Clodoaldo Barrera and Brent Yardley, IBM
Storage System Models for Flash

- Direct Attached Flash
- Hyperconverged
- Server
- Server
- Flash
- Flash

All-Flash Arrays

- Server
- Server
- Flash
- Flash

Hybrid Arrays

- Server
- Server
- Flash
- Flash

Disk Storage (object store)

Cold Storage
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

“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....

SAN Fabric

NVMe-oF™ (NVMe™ between hosts and storage)

NVMe (Within storage array)
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 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

Impact of NVMe™ For Media Access

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

- HDD:
  - Drive Latency: ~10 ms
  - IO Controller Latency: ~25 µs
  - Software Latency: ~10 µs

- SAS TLC NAND SSD:
  - Drive Latency: ~80 µs
  - IO Controller Latency: ~20 µs
  - Software Latency: ~10 µs

- NVME TLC NAND SSD:
  - Drive Latency: ~80 µs
  - IO Controller Latency: ~5 µs
  - Software Latency: ~5 µs

- NVME SCM SSD (Local):
  - Drive Latency: ~2 µs
  - IO Controller Latency: ~5 µs
  - Software Latency: ~5 µs

- NVME-of SCM SSD (Remote):
  - Drive Latency: ~6 µs
  - IO Controller Latency: ~5 µs
  - Software Latency: ~5 µs
NextGen Blocks - NVMe™

What are NVMe-oF™ and FC-NVMe?

- FCP - SCSI-3 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

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

- Single Port Performance
  - 3x improvement
  - 50% increase in peak IOPS
  - ~80 μS reduction in average latency
NVMe™ Vocabulary Update

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

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSCSI</td>
<td>IQN</td>
<td>ign.1991-05.com.microsoft:dmrtk-srvr-m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FC</th>
<th>FC-NVMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUN</td>
<td>Namespace</td>
</tr>
<tr>
<td>WWPN</td>
<td>NQN</td>
</tr>
<tr>
<td>igroup</td>
<td>Subsystem</td>
</tr>
<tr>
<td>ALUA</td>
<td>ANA*</td>
</tr>
</tbody>
</table>

* 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

Application Servers

Emulex LPe32002 32GFC HBAs (one per server)

Emulex LPe32002 32GFC HBAs (qty. 2 per node, 2 nodes)

The target 32GFC adapters can run SCSI FCP and NVMe/FC concurrently

Brocade G620 32GFC Switch

NetApp AFF A700s
NVMe™ over Fibre Channel Performance on a A700s single node

### Random Read 4KB

**Latency vs. IOPS**

- 54% higher IOPS at 2300 µs
- 53% higher IOPS at 450 µs

**Note:** all measurements taken on a single-node A700s. Standard implementations are dual-node.

### Sequential Read 32KB

**Latency vs. Throughput**

- 43% higher throughput at 145 µs

**Note:** all measurements taken on a single-node A700s. Standard implementations are dual-node.

### Random Read 4KB (zoom in)

**Latency vs. IOPS**

- At least 34% lower latency

**Note:** all measurements taken on a single-node A700s. Standard implementations are dual-node.

### Sequential Read 64KB

**Latency vs. Throughput**

- 23% higher throughput at 250 µs

**Note:** all measurements taken on a single-node A700s. Standard implementations are dual-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
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™: Lean Stack Delivers more IOPs with less CPU"
**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

![Average Application Latency](chart_image)
Performance Improvement of NVMe™ over Fabrics – End to End

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

Simulated OLTP Workload IOPS\(^1\)

![Graph showing IOPS comparison between NVMe/FC and SCSI FCP](image1.png)

Data Warehouse IO Throughput\(^2\)

![Graph showing MB/s comparison between NVMe/FC and SCSI FCP](image2.png)

Batch Transaction Latency Test\(^3\)

![Graph showing IO Average Latency comparison between NVMe/FC and SCSI FCP](image3.png)

**3.6x More Transactions**

**2.7x Higher Throughput**

**\(\frac{1}{2}\) 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
Architected for Performance

nvm EXPRESS®

Architected for Performance