NVMe™ over Fabrics: Updates for 2018
Sponsored by NVM Express™, Inc.

Brandon Hoff
Principle Software Architect, Broadcom
25 September 2018
Brandon Hoff

NVMe Marketing Workgroup Committee Member and Principle Software Architect at Broadcom
Agenda

• NVM Express™ Roadmap for NVMe over Fabrics (NVMe-oF)
• NVMe-oF™ Transports
• NVMe-oF Solutions
  • Enterprise AFAs
  • NVMe-oF Appliances
  • NVMe-oF JBOFs
• Interoperability Testing
Audience Poll

Are you considering deploying NVMe-oF?

a. Already deployed
b. Ready to deploy
c. Interested in deploying
d. Just learning about it
e. Not considering deploying
NVMe™ Feature Roadmap

- **NVMe™ Base**
  - NVMe™ 1.2 – Nov '14
    - Namespace Management
    - Controller Memory Buffer
    - Host Memory Buffer
    - Live Firmware Update
  - NVMe™ 1.2.1 May’16
    - Sanitize
    - Streams
    - Virtualization
  - NVMe™ 1.3
    - NVMe-oF™ 1.0 May’16
      - Transport and protocol
      - RDMA binding
    - NVMe-oF™-1.1*
      - Enhanced Discovery
      - TCP Transport Binding
  - NVMe™ 1.4*
    - IO Determinism
    - Persistent memory Region
    - Multipathing

- **NVMe™ oFabric**
  - NVMe-MI™ 1.0 Nov’15
    - Out-of-band management
    - Device discovery
    - Health & temp monitoring
    - Firmware Update
  - NVMe-MI™ 1.1
    - SES Based Enclosure Management
    - NVMe-MI™ In-band
    - Storage Device Enhancements
  - NVMe-MI™ In-band

- **NVMe™ MI**

* Subject to change

**Released NVMe™ specification**

**Planned release**
Scaling NVMe™ Requires a Network

- Many options, plenty of confusion
- Fibre Channel is the transport for the vast majority of today’s all flash arrays
  - FC-NVMe Standardized in Mid-2017
- RoCEv2, iWARP and InfiniBand are RDMA based but not compatible with each other
  - NVMe-oF™ RDMA Standardized in 2016
- FCoE as a fabric is an option, leverages the FC stack integrated into NVMe-oF™ 1.0
- NVMe/TCP - making its way through the standards
**NVMe-oF™/TCP**

- Defines a TCP Transport Binding layer for NVMe-oF
- Promoted by Facebook, Google, DELL EMC, Intel, Others. Sweet spots for JBOF/FBOFs
- Not RDMA-based
- Not yet part of the NVMe-oF standard, will likely be added in 2018/19
- Enables adoption of NVMe-oF into existing datacenter IP network environments that are not RDMA-enabled
- TCP offload required to leverage Flash potential
NVMe™/TCP Data Path Usage

Enables NVMe-oF™ I/O operations in existing IP Datacenter environments

- Software-only NVMe Host Driver with NVMe-TCP transport
- Provides an NVMe-oF alternative to iSCSI for Storage Systems with PCIe® NVMe SSDs
  - More efficient End-to-End NVMe Operations by eliminating SCSI to NVMe translations
- Co-exists with other NVMe-oF transports
  - Transport selection may be based on h/w support and/or policy
Storage Architectures
**NVMe™ over Fabrics – Storage Architectures**

**Enterprise Arrays - Traditional SAN**
- **APPs**
- **NVMe-oF**

**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
External Storage Market

• Current Status
  • Fibre Channel storage shows strong growth in capacity
  • The adoption of All Flash Arrays and NVMe™ storage will drive the need for faster networks
  • iSCSI is the dominant technology block over Ethernet
  • The only RDMA market for block storage is InfiniBand

• Top Vendor Announcements for NVMe-oF™
  • Tier 1 Vendors: Broadcom, Mellanox, IBM, Pure, NetApp, Toshiba, Marvell, EMC, Cisco, Intel, Microsemi, and a lot more
  • NVMe-oF is quickly becoming a leading block storage interface for external storage for applications that need the performance

Block Storage Capacity Shipped

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

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

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

**Back side of the Storage Array**
Performance improvement by moving from SAS/SATA drives to NVMe SSDs
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
Impact of NVMe™ For Media Access

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

- HDD: ~10 ms
- SAS TLC NAND SSD: ~80 µs
- NVMe TLC NAND SSD: ~80 µs
- NVMe SCM SSD (Local): ~2 µs
- NVMe-oF™ SCM SSD (Remote): ~6 µs

Drive Latency
IO Controller Latency
Software Latency

~ 10 ms
~ 25 µs
~ 10 µs
~ 80 µs
~ 20 µs
~ 10 µs
~ 80 µs
~ 5 µs
~ 2 µs
~ 5 µs
~ 6 µs
~ 5 µs
Enterprise Storage Solutions
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

Directions in Storage Networking

• 10GE ->100GE dominates the Cloud infrastructure
  • CSPs adopt new Ethernet technology faster than Enterprise
  • Less constrained by legacy install base
  • Some CSPs add additional networking functionality in their NICs

• FC continues link speed generations (now on Gen 6 at 32Gbps and Gen 7 at 64 Gps)
  • 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™ Enterprise Storage Architecture

High performance low latency storage solutions

- Persistent Memory (PMEM)
- Storage Class Memory (SCM) as Cache
- SSDs attached via NVMe
NVMe™ over Fibre Channel Performance on a A700s Single Node

Random Read 4KB Latency vs. IOPS

- 54% higher IOPS at 2300 µs
- At least 34% lower latency

Sequential Read 32KB Latency vs. Throughput

- 42% higher throughput at 1500 µs
- 43% higher throughput at 145 µs
- At least 11% lower latency

Random Read 4KB Latency vs. IOPS (zoom in)

- At least 34% lower latency

Sequential Read 64KB Latency vs. Throughput

- 23% higher throughput at 250 µs
- At least 15% lower latency

Note: all measurements taken on a single-node A700s. Standard implementations are dual-node.
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 I/O Latency: function of what the hardware can do

NVMe™ benefits from increased parallelization
NVMe-oF™ Enterprise Appliances and JBOFs
Hyperscale Infrastructure

Search

Front-End & Product

Hadoop

Object Store

Databases

Deep Learning AI

Deep Learning AI

Object Store

Hadoop

Front-End & Product

Search
Rack-As-A-Compute

**Right Sizing:**
- Clusters can use optimized ratio of compute and storage.
- Allows reducing wastage and improve performance

**Independent Scaling:**
Compute and storage capacities can be scaled per need.
The Composable Datacenter

- **Utilized SSDs**
  - App A: Needs 1 SSD
  - App B: Needs 2 SSDs
  - App C: Needs 3 SSDs

- **Spare SSDs**

**Spares / Expansion Pool**
- Minimize *Dark Flash!*
- Buy them only as needed
- Power them only as needed

**Other benefits**
- Dynamically allocate more or less storage
- Return SSDs to Pool as apps are retired
- Upgrade SSDs independently
Storage is Not Just About CPU I/O Anymore

- NVMe™ together with a PCIe fabric allow direct network to storage and accelerator to storage communications

Example:

1. Data transferred from network to NVMe™ CMB
2. NVMe block write operation imitated from CMB to NVM
   ... sometime later ...
3. NVMe block read operation initiated from NVM to CMB
4. GPU/Accelerator transfers data from NVMe CMB for processing
PCIe® NVMe™ JBOF

PCIe Switch

NVMe Host

NVMe SSD  NVMe SSD  NVMe SSD  NVMe SSD  NVMe SSD

Facebook Lightning PCIe NVMe JBOF
PCIe® JBOF Enclosure Management

• Native PCIe Enclosure Management (NPEM)
  • Submitted to the PCI-SIG® Protocol Workgroup (PWG) on behalf of the NVMe™ Management Interface (NVMe-MI™) Workgroup
  • Approved by PCI-SIG on August 10, 2017
  • Transport specific basic enclosure management

• SCSI Enclosure Services (SES) Based Enclosure Management
  • Technical proposal developed in the NVMe-MI workgroup
  • While the NVMe and SCSI architectures differ, the elements of an enclosure and capabilities to manage them are the same
    • Example enclosure elements: power supplies, fans, display or indicators, locks, temperature sensors, current sensors, voltage sensors, and ports
  • Comprehensive enclosure management for NVMe™ that leverages (SES), a standard developed by T10 for management of enclosures using the SCSI architecture
1U ruler based designs on PCIe attach being introduced into the market
- Designs provide high density NVMe™ but lack scalability
- Goal is to extend concept for cloud scale using NVMe-oF™
- Gain scalability of fabrics attached
- Simplify design by removing PCIe switch
NVMe™ Integrator’s List Conformance Testing
UNH-IOL

NVMe Conformance Test Cases: 220
NVMe Interop Test Cases: 9
NVMe-MI™ Conformance Test Cases: 53
NVMe-oF™ Conformance Test Cases: 132
NVMe-oF Interop Test Cases: 4
NVMe™ Integrator’s List Interoperability Testing

• NVMe interoperability requires running the technology against 5 unique configurations

• NVMe-MI™ interoperability is something that requires additional attention, no test plan today

• The NVMe-oF™ interoperability testing requires the following:
  • Target – run against two unique Initiator products
  • Switch – run against two unique Target products
  • Initiator – run against two unique Target products
NVMe.Next

Continual evolution of the NVMe™ Integrator’s List program in 2H18

• NVMe Plugfest #10 covering PCIe SSDs and NVMe-oF, October 2018
• TCP Conformance test offering
FCIA FC-NVMe™ Plugfest Events

Test Track 5  GEN6, GEN5 FC and FC-NVMe Dual Fabric HA Large Fabric Build

Implement pair wise zone for each I-T

Concurrent FC and FC-NVMe outlined in black

Analyzers inserted inline between switches and T328 and VIAVI cascaded; Initiator and Target separation enables ISL visibility to all I-T traffic.

https://fibrechannel.org
What Type of 3rd Party Testing is Available?

- Data Integrity
- Performance Analysis
- Interoperability
- Compliance and Pre-certification
  - PCI-SIG® PCIe Express®
  - NVMe™ Conformance Test
  - NVMe-MI™ Conformance Tests

http://teledynelecroy.com/protocolanalyzer/nvm-express/nvme-testing
Enterprise Support Ecosystem

- Enterprise Customers will want to get support from their vendors
  - Servers, storage, NIC/HBA, Network, and OSVs
- Solution is tested and supported by each vendor
- Solutions are documented by each vendor as supported
Audience Poll

What application(s) are you running on an NVMe-oF deployment?

a. Content/collaboration
b. Business applications (ERM/SCM/CRM)
c. Ecommerce
d. Dev Ops
e. Website operations
f. Data management (structured/unstructured)
Contributors

Brandon Hoff, Broadcom
Fazil Osman, Broadcom
Praveen Midha, Marvell
J Metz, Cisco
Clod Berrera, IBM
Mike Kieran, NetApp
Bryan Cowger, Kazan
Nishant Lodha, Marvell

Peter Onufryk, Microsemi
Sujoy Sen, Intel
Kamal Hyder, Toshiba
Manoj Wadekar, eBay
Yaniv Romem, Excelero
Tim Sheehan, UNH-IOL
Mark Jones, Broadcom
Nick Kriczky, Teledyne
For More Information

NVM Express™, Inc. partnered with FMS to organize a conference track devoted exclusively to NVM Express technology. View the slides from the NVMe™ sponsored track:

- NVME-101-1, Part 1: NVMe™: What you need to know for next year
- NVME-101-1, Part 2: NVMe™: Hardware Implementations and Key Benefits in Environments
- NVME-102-1, Part 1: NVMe™ Management Interface (NVMe-MI™) and Drivers Update
- NVMe-101-2, Part 1: “NVMe™ Management Interface (NVMe-MI™) Update
- NVME-102-1, Part 2: NVMe™ over Fabrics – Discussion on Transports
- NVME-201-1, Part 1: NVMe™ and NVMe-oF™ in Enterprise Arrays
- NVME-201-1, Part 2: NVMe-oF™ Enterprise Appliances
- NVME-202-1: NVMe-oF™ JBOFs

Video recordings of these presentations can be viewed on our YouTube Channel.

https://nvmexpress.org/about/flash-memory-summit-2018/