NVMe™: What you need to know for next year

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

Janene Ellefson @jamminjanene

David Allen

J Metz @drjmetz
NVMe™ Agenda

Intro & 2 day Agenda
Market Outlook
NVMe™ Roadmap
NVMe-oF™
Q&A
<table>
<thead>
<tr>
<th>Track</th>
<th>Title</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe-101-1</td>
<td>NVM Express: NVM Express roadmaps and market data for NVMe, NVMe-oF,</td>
<td>Janene Ellefson, Micron J Metz, Cisco</td>
</tr>
<tr>
<td>8/7/18</td>
<td>and NVMe-MI - what you need to know for the next year.</td>
<td>Amber Huffman, Intel David Allen, Seagate</td>
</tr>
<tr>
<td>8/7/18</td>
<td>NVMe architectures for in Hyperscale Data Centers, Enterprise Data</td>
<td>Janene Ellefson, Micron Chris Peterson, Facebook</td>
</tr>
<tr>
<td>9:45-10:50</td>
<td>Centers, and in the Client and Laptop space.</td>
<td>Chander Chadha, Toshiba Jonmichael Hands, Intel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 3:40-4:45</td>
<td>Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee</td>
</tr>
<tr>
<td></td>
<td>NVMe-102-1</td>
<td>Prewitt, Microsoft</td>
</tr>
<tr>
<td></td>
<td>Drivers and Software: This session will cover the software and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>drivers required for NVMe-MI, NVMe, NVMe-oF and support from the top</td>
<td></td>
</tr>
<tr>
<td></td>
<td>operating systems.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 4:55-6:00</td>
<td>Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee</td>
</tr>
<tr>
<td></td>
<td>NVMe-oF Transports: We will cover for NVMe over Fibre Channel, NVMe</td>
<td>Suds Jain, VMware David Minturn, Intel James Harris, Intel</td>
</tr>
<tr>
<td></td>
<td>over RDMA, and NVMe over TCP.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 8:30-9:35</td>
<td>Brandon Hoff, Emulex Fazil Osman, Broadcom J Metz, Cisco</td>
</tr>
<tr>
<td></td>
<td>NVMe-oF Enterprise Arrays: NVMe-oF and NVMe is improving the</td>
<td>Brandon Hoff, Emulex Michael Peppers, NetApp Clod Barrera, IBM</td>
</tr>
<tr>
<td></td>
<td>performance of classic storage arrays, a multi-billion dollar</td>
<td>Fred Night, NetApp Brent Yardley, IBM</td>
</tr>
<tr>
<td></td>
<td>market.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 9:45-10:50</td>
<td>Jeremy Werner, Toshiba Manoj Wadekar, eBay Kamal Hyder, Toshiba</td>
</tr>
<tr>
<td></td>
<td>NVMe-oF Appliances: We will discuss solutions that deliver</td>
<td>Nishant Lodha, Marvell Yaniv Romem, CTO, Excelero</td>
</tr>
<tr>
<td></td>
<td>high-performance and low-latency NVMe storage to automated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>orchestration-managed clouds.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 3:20-4:25</td>
<td>Bryan Cowger, Kazan Networks</td>
</tr>
<tr>
<td></td>
<td>NVMe-oF JBOFs: Replacing DAS storage with Composable Infrastructure</td>
<td>Praveen Midha, Marvell Fazil Osman, Broadcom</td>
</tr>
<tr>
<td></td>
<td>(disaggregated storage), based on JBOFs as the storage target.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/7/18 4:40-6:45</td>
<td>Brandon Hoff, Emulex Tim Sheehan, IOL Mark Jones, FCIA</td>
</tr>
<tr>
<td></td>
<td>Testing and Interoperability: This session will cover testing for</td>
<td>Jason Rusch, Viavi Nick Kriczky, Teledyne</td>
</tr>
<tr>
<td></td>
<td>Conformance, Interoperability, Resilience/error injection testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>to ensure interoperable solutions base on NVM Express solutions.</td>
<td></td>
</tr>
</tbody>
</table>
About NVM Express™

- NVM Express (NVMe™) is an open collection of standards and information to fully expose the benefits of non-volatile memory in all types of computing environments from mobile to data center.
- NVMe™ is designed from the ground up to deliver high bandwidth and low latency storage access for current and future NVM technologies.

**NVM Express Base Specification**

The register interface and command set for PCI Express attached storage with industry standard software available for numerous operating systems. NVMe™ is widely considered the de facto industry standard for PCIe SSDs.

**NVM Express Management Interface (NVMe-MI™) Specification**

The command set and architecture for out of band management of NVM Express storage (i.e., discovering, monitoring, and updating NVMe™ devices using a BMC).

**NVM Express Over Fabrics (NVMe-oF™) Specification**

The extension to NVM Express that enables tunneling the NVM Express command set over additional transports beyond PCIe. NVMe over Fabrics™ extends the benefits of efficient storage architecture at scale in the world’s largest data centers by allowing the same protocol to extend over various networked interfaces.
NVMe™ Market Landscape
SSD Share of Revenue ($M)

SSD Share of Units Tam (M)

Source: Micron
NVMe™ Feature Roadmap

<table>
<thead>
<tr>
<th>Year</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NVMe™ Base**
- NVMe™ 1.2 – Nov ‘14
  - Namespace Management
  - Controller Memory Buffer
  - Host Memory Buffer
  - Live Firmware Update

**NVMe-oF™**
- NVMe-oF™ 1.0 May’16
  - Transport and protocol
  - RDMA binding

**NVMe-MI™**
- NVMe-MI™ 1.0 Nov’15
  - Out-of-band management
  - Device discovery
  - Health & temp monitoring
  - Firmware Update

**NVMe™ 1.2.1 May’16**
- Sanitize
- Streams
- Virtualization

**NVMe™ 1.3**
- IO Determinism
- Persistent Memory Region
- Multipathing

**NVMe™ 1.4**
- Subject to change

**NVMe-oF™-1.1**
- Enhanced Discovery
- TCP Transport Binding

**NVMe-MI™ 1.1**
- SES Based Enclosure Management
- NVMe-MI™ In-band
- Storage Device Enhancements

- Released NVMe specification
- Planned release

* Subject to change
Ever-Advancing Performance and Features

Data latency

- Improvement: I/O Determinism (IOD)

High Performance Non-Volatile data needs

- Improvement: Persistent Memory Region

Ease of Data sharing

- Improvements: Multi-Pathing access

NVMe™ 1.4+

- I/O Determinism
- Persistent memory Region
- Multipathing
Management Needs

Standardized Management for ease of adoption

• Industry standard tools and compliance

Improvements and updates to managing the subsystems and end devices

• Event logging
• Incorporating robust industry adopted enclosure management
• Diverse connections to end devices (SSDs)
  • Additional In-band mechanisms
Enterprise Networking Needs

- Robustness in networking topologies
  - Congestion Management
- New and interesting transport capabilities
  - TCP bindings for NVMe-oF™
- Improvements in automation
  - Discovery
- Security Enhancements
  - In-band authentication
NVMe™ 1.4

Projected completion: 2019
What is NVMe™ I/O Determinism?

- Service isolation region
- Increase Read I/OPs and reduce max latency
- Provides strict QoS profile
- Significantly improves P99 and P9999 for a well-behaved host
Persistent Memory Region (PMR)

Controller Memory Buffer (CMB)
- Introduced in NVMe™ 1.2
- PCI memory space exposed to host
- May be used to store commands and command data
- Contents do not persist across power cycles and resets

Persistent Memory Region (PMR)
- PCI memory space exposed to host
- May be used to store command data
- Content persist across power cycles and resets
NVMe™ Multipathing and Namespace Sharing

Technical Term: Asymmetric Namespace Access (ANA)

NVMe™ Multipathing I/O refers to two or more completely independent PCI Express paths between a single host and a namespace.

Namespace sharing enables two or more hosts to access a common shared namespace using different NVM Express controllers.

Both multi-path I/O and namespace sharing require that the NVM subsystem contain two or more controllers.
# NVMe™ 1.4 Well Underway

## NVMe 1.4 Development Status

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Ratified</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>10</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

### Released NVMe specification
- Device Self Test
- Set Timestamp
- Boot Partitions
- Sanitize
- Error Log Updates
- Globally Unique NGUID/EUI64
- SGL Dword Simplify
- Streams Directive
- Device Telemetry
- Virtualization
- Host Controlled Thermal Mgmt
- NVMe-MI Tunneling
- Grab Bag (incl. Strict Mode)
- Para-virtualized Dev Support

### Planned release
- Persistent Memory Region
- HMB Enhancements
- IO Determinism
- ANA Base Protocol
- Namespace Write Protect
- Transport SGL Descriptor
- NVM Sets & Read Recovery Levels
- Transport Error Codes

### 8 TPs Ratified Already

Ratified TPs available publically at:

http://nvmexpress.org/resources/specifications/
NVMe™ Management Interface (NVMe-MI™) 1.1

Projected completion: Early 2018
NVMe-MI™ 1.1 Key Work Items

- SCSI Enclosure Services (SES) Based Enclosure Management
  - Draft completed, NVMe-MI™ working through final technical items
  - Comprehensive enclosure management

- Support for In-Band NVMe-MI™
  - Draft complete and in workgroup review

- NVMe™ Storage Device Enhancement – In work

- Native PCIe Enclosure Management (NPEM)
  - Transport specific basic enclosure management
  - Approved by PCI-SIG® on August 10, 2017
NVMe-MI™ Out-of-Band Management

- **Out-of-Band Management** – Management that operates with hardware and components that are *independent of the operation system control*

- **NVMe™ Out-of-Band Management Interfaces**
  - SMBus/I2C
  - PCIe Vendor Defined Messages (VDM)
  - IPMI FRU Data (VPD) accessed over SMBus/I2C
In-Band Management and NVMe-MI™

- In-band mechanism allows application to tunnel NVMe-MI™ commands through NVMe™ driver
  - Two new NVMe™ Admin commands
    - NVMe-MI™ Send
    - NVMe-MI™ Receive

- Benefits
  - Provides management capabilities not available in-band via NVMe™ commands
    - Efficient NVM subsystem health status reporting
    - Ability to manage NVMe™ at a FRU level
    - Vital Product Data (VPD) access
    - Enclosure management
NVMe-oF™
NVMe™/TCP

Title: TP-8000 NVMe-oF™ TCP Transport Binding

Abstract:

  Provides extensions for defining a NVMe transport binding (“Fabrics”) for non-RDMA “vanilla” networks

Status: Phase 3
NVMe™/TCP

NVMe block storage protocol over standard TCP/IP transport

Enables disaggregation of NVMe SSDs without compromising latency and without requiring changes to networking infrastructure

Independently scale storage & compute to maximize resource utilization and optimize for specific workload requirements

Maintains NVMe model: sub-systems, controllers namespaces, admin queues, data queues
NVMe™/TCP in a Nutshell

NVMe-oF™ commands sent over standard TCP/IP sockets

Each NVMe queue pair mapped to a TCP connection

TCP provides a reliable transport layer for NVMe queueing model
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
NVMe™/TCP Control Path Usage

Enables use of NVMe-oF™ on Control-Path Networks (example: 1g Ethernet)

Discovery Service Usage

- Discovery controllers residing on a common control network that is separate from data-path networks

NVMe-MI™ Usage

- NVMe-MI endpoints on control processors (BMC, ..) with simple IP network stacks
  
  - NVMe-MI on separate control network

Source: Dave Minturn (Intel)
NVMe™/TCP Standardization

Expect NVMe over TCP standard to be ratified in 2H 2018
Discovery

A host connects to a DISCOVERY controller to find out what NVMe™ stuff is “out there”

• The discovery controller has a list of available devices (available NVMe subsystems, NVMe ports)
• The host can then connect to the things it has discovered and find namespaces to access
• One discovery service can point to other discovery services (nesting)

The “root” of discovery must be manually configured

A discovery service can’t tell a host if something changes
  ▪ Like if a new device shows up; or
  ▪ If a new port shows up; or
  ▪ If a completely new discovery service shows up

Special Thanks: Fred Knight, NetApp
Enhanced Discovery

- **How do I connect storage consumers to storage suppliers?**
  - Specification enhancement for efficient, dynamic resource management
  - Fabric-transport specific mechanisms to determine where to get provisioning information from
  - Allows the fabric to tell hosts when something changes
  - Allows hosts to perform *dynamic* discovery of new stuff;
    - Adapt to removal of stuff from the NVMe-oF™ environment
  - Dynamically find new paths; or know when old paths go away;
    - Now can be done over RDMA and TCP as well as FC

Special Thanks to Phil Cayton, Intel

Special Thanks: Fred Knight, NetApp
Issues with NVMe-oF™ Discovery and Management

The current NVMe-oF specification and Linux implementation lacks:

- Dynamic resource discovery and enumeration of remote resources
- Clear definition for methods of how to discover the proper discovery controller defining remote storage resource provisioning

To support large-scale deployment of NVMe-oF, more is needed:

- Specification enhancement for efficient, dynamic resource management
- Fabric-transport specific mechanisms to determine where to get provisioning information from
- Linux kernel driver stack changes as the specification evolves
- Management tools to enable NVMe-oF management and scale-out
- Finding the discovery root is still missing (manually configured)
- Discovery is still very weak on multiple fabric installations (no FABRIC ID in the discovery service, so while you have a name and a port, you don’t know which fabric to use to connect to it – IF you happen to be connected to multiple fabrics)
- Discovery is also still just discovery – NOT about management of the configuration or provisioning of anything

Special Thanks: Fred Knight, NetApp
Summary - The Future of NVMe™

**NVMe™ 1.4**
- IO Determinism
- Persistent Controller Mem Buffer and Event Log
- Multipathing (ANA)

**NVMe-MI™ 1.1**
- SCSI Enclosure Services (SES)
- NVMe-MI™ In-band
- Native Enclosure Management

**NVMe-oF™ 1.1**
- Enhanced Discovery
- TCP Transport Binding
<table>
<thead>
<tr>
<th>Track</th>
<th>Title</th>
<th>Chair</th>
<th>Speakers (Proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe-101-1</td>
<td>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</td>
<td>J Metz, Cisco, Amber Huffman, Intel, David Allen, Seagate</td>
</tr>
<tr>
<td></td>
<td>NVMe architectures for in Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.</td>
<td>Janene Ellefson, Micron</td>
<td>Chris Peterson, Facebook, Chander Chadah, Toshiba, Jonmichael Hands, Intel</td>
</tr>
<tr>
<td>NVMe-102-1</td>
<td>NVMe architecture for Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.</td>
<td>Uma Parepalli, Western Digital</td>
<td>Austin Bolen, Dell EMC, Myron Loewen, Intel, Lee Prewitt, Microsoft, Suds Jain, VMware, David Minturn, Intel, James Harris, Intel</td>
</tr>
<tr>
<td>3:40-4:45</td>
<td>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 such as NVMe-oF with Linux, RedHat, Suse, Oracle, Microsoft, Vmware as well as NVMe and NVMe-oF for SPDK.</td>
<td>Brandon Hoff, Broadcom</td>
<td>Fazil Osman, Broadcom, J Metz, Cisco, Curt Beckmann, Broadcom, Praveen Midha, Marvell</td>
</tr>
<tr>
<td>4:55-6:00</td>
<td>NVMe-oF Transports: NVMe over Fabrics is designed to be transport agnostic, with all transports being created equal from the perspective of NVM Express. We will cover for NVMe over Fibre Channel, NVMe over RDMA, and NVMe over TCP.</td>
<td>Brandon Hoff, Broadcom</td>
<td></td>
</tr>
<tr>
<td>8/8/18</td>
<td>NVMe-oF Enterprise Arrays: NVMe-oF and NVMe is improving the performance of classic storage arrays, a multi-billion dollar market. This session will cover NVMe and NVMe-oF for Enterprise All Flash Arrays (AFAs) including SPDK with NVMe-oF.</td>
<td>Brandon Hoff, Broadcom</td>
<td>Michael Peppers, NetApp, Clod Barrera, IBM</td>
</tr>
<tr>
<td>8/8/18</td>
<td>NVMe-oF Appliances: These solutions are different than Enterprise Arrays because the targets being more like JBOFs than Enterprise AFAs. We will discuss solutions that deliver high-performance and low-latency NVMe storage to automated orchestration-managed clouds.</td>
<td>Jeremy Warner, Toshiba</td>
<td>Manoj Wadekar, eBay, Kamal Hyder, Toshiba, Nishant Lodha, Marvell, Lior Gal, Excelero</td>
</tr>
<tr>
<td>8/8/18</td>
<td>NVMe-oF JBOFs: By replacing DAS storage with Composable Infrastructure (disaggregated storage), based on JBOFs as the storage target, end-users benefit in terms of business agility, ease of hardware upgrades, and lowering of both CAPEX and OPEX.</td>
<td>Bryan Cowger, Kazan Networkds</td>
<td>Praveen Midha, Marvell, Fazil Osman, Broadcom</td>
</tr>
<tr>
<td>8/8/18</td>
<td>Testing and Interoperability: There are at least 9 different standards that NVMe solutions leverage from PCIe to NVMe to Transports for NVMe-oF. This session will cover testing for Conformance, Interoperability, Resilience/error injection testing to ensure interoperable solutions.</td>
<td>Brandon Hoff, Broadcom</td>
<td>Tim Sheehan, IOL, Mark Jones, FCIA</td>
</tr>
</tbody>
</table>