NVM Express Introduction & Tutorial

David Akerson, Intel Corporation

#OFADevWorkshop
“If I had asked people what they wanted, they would have said faster horses.”
- Henry Ford
What is **nvm EXPRESS™**?

The industry standard for Enterprise and Client Non-volatile Memory based storage solutions

Architected from the ground up for Non-volatile memory to be more efficient, scalable, and manageable

Standardizes register set, feature set, and command set to deliver performance

Developed by an open industry consortium for Client, Enterprise and Data Center
NVM Express, Inc.
Consists of more than 80 companies from across the industry

Technical Workgroup
NVM Express Specifications - Queuing interface, NVMe I/O and Admin command set

Management Interface Workgroup
Out-of-band management over PCIe VDM and SMBus

Promoter Group
Led by a 13 company Board of Directors

Marketing Workgroup
NVM Express Awareness

NVM Express over Fabrics Workgroup
A flexible transport abstraction layer useful for many different fabrics
2014: An Amazing Year

“...our first look at an NVMe drive, and the results are impressive.”

AnandTech
NVMe™ Driver Ecosystem

Native / in-box

- Windows 8.1
- SLES 11 SP3
- SLES 12
- 13 | 14

Linux NVMe driver is open source

Install NVMe driver

- Windows Server 2012 R2 Certified
- Windows 7 Certified
- Windows Server 2008 R2 Certified
- VMware ESXi 5.5
## NVM Express Advantages over SATA

### PCIe for **scalable** performance, **flexible** form factors, and **industry stability**

### NVMe provides **lower latency** and increased **efficiency**: lower CPU utilization, lower power, lower TCO

### Increased **bandwidth**: 1 GB/s per lane – 1-16 lanes per drive

Directly attached to CPU, eliminate HBA cost and overhead

### **Low power** features from both PCIe and NVMe

### Security from Trusted Computing Group OPAL
NVM Express Technical Overview

- Supports deep queues of 64K commands per queue, up to 64K queues
- Supports MSI-X and interrupt steering, enables even performance scaling
- Streamlined & simple command set (13 required commands), optional features to address target segments
- Built for the future, ready for next gen NVM
NVM Express (NVMe) Delivers Best in Class IOPs … And Best in Class Sequential Performance

Compared to SAS 12 Gbps
• 100% random reads: >3X better IOPs
• 70% random reads: >2X better IOPs
• 100% random writes: ~1.5X better IOPs

Compared to SAS 12 Gbps
• 100% reads: >2X better performance
• 100% writes: >2.5X better performance

Note: PCI Express’ (PCIe')/NVM Express’ (NVMe) Measurements made on Intel® Core™ i7-3770S system @ 3.1GHz and 4GB Mem running Windows® Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by Iometer® tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar® SSD800M/1000M (SAS) Solid State Drive Specification. SATA Measurements from Intel Solid State Drive DC P3700 Series Product Specification.
Analyzing What Matters

- What matters in today’s Data Center is not just IOPs and bandwidth
- Let’s look at efficiency of the software stack, latency, and consistency

Server Setup

- Basic 4U Intel® Xeon® E5 processor based server
- Out of box software setup
- Moderate workload: 8 workers, QD=4, random reads

Storage Protocols Evaluated

<table>
<thead>
<tr>
<th>Interface</th>
<th>6Gb SATA</th>
<th>6Gb SATA</th>
<th>6Gb SAS</th>
<th>12Gb SAS</th>
<th>NVMe PCIe Gen 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attach Point</td>
<td>PCH chipset</td>
<td>6Gb SAS HBA</td>
<td>6Gb SAS HBA</td>
<td>12Gb SAS HBA</td>
<td>CPU</td>
</tr>
</tbody>
</table>

*Not strenuous on purpose – evaluate protocol and not the server.*

For detailed configuration information, refer to slide “Setup for Efficiency and Latency Analysis” in backup.
The Efficiency of NVM Express™

- CPU cycles in High Performance Computing are precious
  - Each CPU cycle required for an IO adds latency
- NVM Express takes less than half the CPU cycles per IO as SAS

With equivalent CPU cycles, NVM Express delivers over 2X the IOPs of SAS!
The Latency of NVM Express™

- The efficiency of NVM Express directly results in leadership latency
- When doubling from 6Gb to 12Gb, SAS only reduces latency by ~ 60 μS
- NVMe is more than 200 μs lower average latency than 12 Gb SAS

NVM Express delivers the lowest latency of any standard storage interface.
The Consistency of NVM Express

- NVM Express* (NVMe) leadership on latency and efficiency is **consistently** amazing.
- SAS is a mature software stack with over a decade of tuning, yet the first generation NVM Express software stack has 2 to 3X better consistency.

---

**Latency Consistency**

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6Gb SATA on PCH chipset</td>
<td>1.22%</td>
</tr>
<tr>
<td>6Gb SATA on 6Gb SAS HBA</td>
<td>1.01%</td>
</tr>
<tr>
<td>6Gb SAS on 6Gb SAS HBA</td>
<td>0.57%</td>
</tr>
<tr>
<td>12Gb SAS on 12Gb SAS HBA</td>
<td>0.97%</td>
</tr>
<tr>
<td>NVMe on CPU (PCIe Gen 3)</td>
<td>0.48%</td>
</tr>
</tbody>
</table>

*StDev / Avg Latency:

- 6Gb SATA on PCH chipset: 1.22%
- 6Gb SATA on 6Gb SAS HBA: 1.01%
- 6Gb SAS on 6Gb SAS HBA: 0.57%
- 12Gb SAS on 12Gb SAS HBA: 0.97%
- NVMe on CPU (PCIe Gen 3): 0.48%

**NVMe is already best in class, with more tuning yet to come.**
Fully Exploiting Next Gen NVM

- With Next Gen NVM, the NVM is no longer the bottleneck
- App to SSD read latency for 4KB transfer at Queue Depth of 1
The choice is yours…
For more information...

Visit nvmexpress.org
Thank You