NVM Express™ Infrastructure - Exploring Data Center PCIe® Topologies

January 29, 2015

Jonmichael Hands – Product Marketing Manager, Intel Non-Volatile Memory Solutions Group

Peter Onufryk – Sr. Director Product Development – PMC-Sierra

Moderator: Ravi Chari – Storage Technologist – Oracle

View recorded webcast at https://www.brighttalk.com/webcast/12367/141221
Legal Disclaimer

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

No computer system can be absolutely secure. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

*Other names and brands may be claimed as the property of others.

Copyright © 2015 Intel Corporation. All rights reserved.
Agenda - NVM Express\textsuperscript{TM}(NVMe\textsuperscript{TM}) Infrastructure

- What is NVMe?
- NVMe advantages over SATA\textsuperscript{TM}
- NVMe driver ecosystem
- PCIe\textsuperscript{®} form factors, cables, and connectors
- Link extension and port expansion for PCIe
- PCIe Solid-State Drive Topologies
- NVMe Management
NVM Express™ is a standardized high performance software interface for PCI Express® Solid-State Drives.

Architected from the ground up for SSDs to be more efficient, scalable, and manageable.

NVMe is industry driven to be extensible for the needs of both the client and the data center.

“If I had asked people what they wanted, they would have said faster horses.”

- Henry Ford
NVM Express™ Community

NVM Express, Inc.
Consists of more than 75 companies from across the industry

Promoter Group
Led by 13 elected companies

Technical Workgroup
Queuing interface, NVMe I/O and Admin command set

Management Interface
Workgroup
Out-of-band management over PCIe® VDM and SMBus
What NVM Express™ brings to the DATA CENTER

Deployment at scale
Industry standard drivers, software, and management

Lower TCO
Efficiency of protocol, increased storage density, lower system power

Works out of the box
In standard operating systems
NVM Express™ (NVMe™) 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

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Configurations: Intel® S2600CP server, Intel® Xeon® E5-2690v2 x2, 64GB DDR3, Intel® SSD DC P3700 Series 400GB, LSI 9207-8i, Intel® SSD DC S3700, HGST 6GBps SAS
NVMe™ Driver Ecosystem

Linux NVMe driver is open source

Native / in-box

Install NVMe driver

*Other names and brands may be claimed as the property of others.
What do I need to start using an SSD?

- **Software:** NVMe™ driver
- **Hardware:** PCIe® infrastructure

NVMe sits on top of PCIe
Form Factors for PCI Express®

Data Center
- HD SSD FF
- SFF-8639
- Add in Card
- SATA Express
- M.2

Client
- BGA

AIC

2.5in SFF-8639

SATA Express™

M.2

BGA

M.2 BGA

HD SSD FF

SFF-8639 Add in Card SATA Express M.2

nvm EXPRESS™
Drive Connectors

**SFF-8639**
- Supports SATA, SAS, and PCIe® x4 or two x2
- PCIe data, reference clock, and side band

**SAS®**
- Backwards compatible with SATA
- Dual port

**SATA™**
- Keyed only for SATA drives
- Separate power and data
# SATA Express™ and SFF-8639 Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>SATAe</th>
<th>SFF-8639</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA / SAS®</td>
<td>SATA</td>
<td>SATA / SAS</td>
</tr>
<tr>
<td>PCI Express®</td>
<td>x2</td>
<td>x4 or dual x2</td>
</tr>
<tr>
<td>Host Mux</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ref Clock</td>
<td>Optional</td>
<td>Required</td>
</tr>
<tr>
<td>EMI</td>
<td>SRIS</td>
<td>Shielding</td>
</tr>
<tr>
<td>Height</td>
<td>7mm</td>
<td>15mm</td>
</tr>
<tr>
<td>Max Performance</td>
<td>2 GB/s</td>
<td>4 GB/s</td>
</tr>
<tr>
<td>Bottom Line</td>
<td>Flexibility &amp; Cost</td>
<td>Performance</td>
</tr>
</tbody>
</table>

SFF-8639 designed for data center, SATAe designed for Client

*Source: Seagate* (with permission)
# M.2 Form Factor Comparison

<table>
<thead>
<tr>
<th></th>
<th>M.2 Socket 2</th>
<th>M.2 Socket 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA</td>
<td>Yes, Shared</td>
<td>Yes, Shared</td>
</tr>
<tr>
<td>PCIe® x2</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PCIe x4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comms Support</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ref Clock</td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td>Max Performance</td>
<td>2 GB/s</td>
<td>4 GB/s</td>
</tr>
<tr>
<td>Bottom Line</td>
<td>Flexibility</td>
<td>Performance</td>
</tr>
</tbody>
</table>

**M.2 Socket 3 is the best option for Data Center PCIe SSDs**
miniSAS HD cables lightly modified for PCIe are being used due to the robust connector and high volume manufacturing.
Basic PCI Express® SSD Topology – 1 Connector

- SFF-8639 Connector directly attached to board
- Mostly used in small form factors such as compute node, blade, etc.
Basic PCI Express® SSD Topology – 2 Connector

- miniSAS HD Connector
- SFF-8639 Connector
- PCIe® Cable
- PCIe 3.0 x4 Enterprise SSD
- External Power
Basic PCI Express® SSD Topology – 3 Connector

1. miniSAS HD Connector
2. SSD Drive Carrier
3. SFF-8639 Connector

PCle® Cable

Backplane
Port Expansion Devices - Switches

Use Switches to expand number of PCIe® SSDs
Link Extension Devices – Switches and Retimers

Use Link Extension Devices for longer topologies
PCI Express® (PCIe®) Switches and Retimers

PCIe Switches
• Use for link extension and/or port expansion
• Hot-plug and error isolation
• High performance peer-to-peer transfers
• Extra software features

Retimers
• Mostly transparent to software
• Retimers should be more common in PCIe 4.0

Recommend using only switches or retimers for link extension of PCIe

Link Extension Devices
• Use when channel has > -20db loss: at 8GT/s PCIe 3.0

Retimer vs. Re-driver
• Repeater: A Retimer or a Re-driver
• Re-driver: Analog and not protocol aware
✓ Retimer: Physical Layer protocol aware, software transparent, Extension Device. Forms two separate electrical sub-links.
  ▪ Executes equalization procedure on each sub-link
High Function Switches

Diagram showing the connection between Hosts, PCIe Switches, and NVMe SSDs.
Complex PCI Express® Topology – 4 Connector

1. PCIe x16 slot
2. Cabled Add in card with Link Extension
3. miniSAS HD for PCIe
4. SFF-8639 Connector

Backplane

SSD Drive Carrier

PCIe® Cable
Complex PCI Express® Topology – 5 Connector

Cabled Add in card with Link Extension

PCle® x16 Riser

PCle x16 slot

Backplane

SSD Drive Carrier

miniSAS HD for PCle
PCI Express® cabling for future topologies - OCuLink*

<table>
<thead>
<tr>
<th>Category</th>
<th>OCuLink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Based</td>
<td>PCI-SIG®</td>
</tr>
<tr>
<td>PCIe® Lanes</td>
<td>X4</td>
</tr>
<tr>
<td>Layout</td>
<td>Smaller footprint</td>
</tr>
<tr>
<td>Signal Integrity</td>
<td>Similar on loss dominated channels</td>
</tr>
<tr>
<td>PCIe 4.0 ready</td>
<td>16GT/s target</td>
</tr>
<tr>
<td>Clock, power</td>
<td>Supports SRIS and 3.3/5V power</td>
</tr>
<tr>
<td>Production Availability</td>
<td>Mid 2015</td>
</tr>
</tbody>
</table>

OCuLink internal cables and connectors

Source: OCuLink internal cables and connectors
OCuLink* Provides Flexible Data Center Topologies

Cabled add in card

Backplane

PCIe® SSD

SFF-8639 Connector

Board to board connections
NVMe™ Storage Device Management

**Example Pre-boot Management**
- Inventory, Power Budgeting, Configuration, Firmware Update

**Example Out-of-Band Management During System Operation**
- Health Monitoring, Power/Thermal Management, Firmware Update, Configuration
Driver vs. Out-of-Band Management
Management Interface Protocol Layering

- Management Applications (e.g., Remote Console)
- Management Controller (BMC or Host Processor)
- NVMe Management Interface
- Management Component Transport Protocol (MCTP)
- MCTP over SMBus/I2C Binding
- MCTP over PCIe VDM Binding
- SMBus/I2C
- PCIe
- NVMe SSD
Questions?
NVMe™ 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
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
NVMe™ Development History

**NVMe 1.0 – Mar 2011**
- Queuing Interface
- Command Set
- End-to-End Protection
- Security
- PRPs

**NVMe 1.1 – Oct 2012**
- Multi-Path IO
- Namespace Sharing
- Reservations
- Autonomous Power Transition
- Scatter Gather Lists

**NVMe 1.2 – Q4 2014**
- Host Memory Buffer
- Replay Protected Area
- Active/Idle Power and RTD3
- Temperature Thresholds
- Namespace Management
- Controller Memory Buffer
- Live Firmware Update
- Atomicity Enhancements

**Timeline**
- 2011
- 2012
- 2013
- 2014
- 2015
nvm EXPRESS™
Architected for Performance