Annual Update on Interfaces
Session U-3

Jim Pappas
Intel Corporation: Director of Technology Initiatives
SNIA: Board of Directors & Executive Committee
jim@intel.com
Agenda

- Transition to PCI Express Storage
- NVDIMMs
- Software Interface Standards
Agenda

Transition to PCI Express Storage

• PCI Express SSD Market
• PCI Express Form Factors
• PCI Express Protocols (NVMe)
• PCI Express Scalability
PCI Express* (PCIe) is scalable enabling an OEM to select the right performance point for an individual drive.

Do I need a 4 GB/s device for my application?

...Or an 8 GB/s device?


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

Massive data growth is driving SSDs into the data center with NVMe as the interface of choice.

- Data Center SSD Market: Will be approaching $10B in 2018, was $4.6B in 2014.
- 2018 DC Storage TAM: More than 40% of revenue projected to be SSDs, the rest on HDDs.
- NVMe by 2017: Half the data center SSD market is NVMe by 2017.

Source: Forward Insights Q1'15, Intel Q2'15
NVMe™ Driving PCIe SSDs in the Data Center

Data Center SSD Units by Interface

Data Center SSD total GB by Interface

Source: Forward Insights Q1'15, Intel Q2'15
Agenda

Transition to PCI Express Storage

- PCI Express SSD Market
- PCI Express Form Factors
- PCI Express Protocols (NVMe)
- PCI Express Scalability
Data Center Form Factors for

42, 80, and 110mm lengths, Smallest footprint of PCIe, use for boot or for max storage density

2.5in makes up the majority of SSDs sold today because of ease of deployment, hotplug, serviceability, and small form factor.

Add-in-card (AIC) has maximum system compatibility with existing servers and most reliable compliance program. Higher power envelope, and options for height and length

Source: Forward Insights Q2'15
SSD Form Factor Working Group

Driving industry standardization, innovation, and performance for the benefit of our customers.

http://www.ssdformfactor.org
# Storage Drives Interface Profiles

<table>
<thead>
<tr>
<th>Interface Profile (indicates connector signals &amp; location)</th>
<th>Plug Connector (Drive)</th>
<th>Receptacle Connector (backplane/Motherboard/cable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M.2 PCIe</td>
<td>Card-edge</td>
<td>M.2 Connector Hx.x-x-Optx</td>
</tr>
<tr>
<td>M.2 SATA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SATA</td>
<td>SATA</td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td>SFF-8680</td>
<td>SFF-8680</td>
</tr>
<tr>
<td>SAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U.2 PCIe</td>
<td>SFF-8639</td>
<td></td>
</tr>
<tr>
<td>U.2 SAS</td>
<td>SFF-8680</td>
<td>SFF-8639</td>
</tr>
<tr>
<td>U.2 SATA</td>
<td>SATA</td>
<td></td>
</tr>
<tr>
<td>Express Bay PCIe</td>
<td>SFF-8639</td>
<td></td>
</tr>
<tr>
<td>Express Bay SAS</td>
<td>SFF-8680</td>
<td></td>
</tr>
<tr>
<td>Express Bay SATA</td>
<td>SATA</td>
<td></td>
</tr>
<tr>
<td>Express Bay Multi Link SAS (x4 SAS)</td>
<td>SFF-8639</td>
<td></td>
</tr>
<tr>
<td>Express Bay SATA-Express (x2 PCIe)</td>
<td>SATAe</td>
<td></td>
</tr>
</tbody>
</table>

Source: Gary Kotzur, Dell, Q3'15

- Not Included
  - USB
  - Ethernet-1
  - Ethernet-2
One 2.5” Connector for All

U.2
(Formerly SFF-8639)

Note: Always pronounced “U dot 2”
End the Confusion: M.2 and U.2

- Easy to remember
- Easy to understand
U.2 drive / U.2 SSD
SAS/SATA/PCIe® 2.0 or 3.0, x1, x2, x4

U.2 connector
Connector can be on cable or backplane
Supports PCIe, SAS, and SATA

U.2 backplane
On a server or workstation

U.2 cable
A cable that connects a U.2 drive

U.2 host connector
U.2 SSDs are available in a variety of power configurations:

- 8W
- 25W Max

U.2 Power Envelope
Comparing Clocking Mechanisms for PCIe®

Recommendation: SSDs supporting SRIS also support RefClk
Current Generation Form Factor

- SATA
- SAS
- M.2
- U.2

PCI Express

Traditional Interface
Next Generation Form Factor
(under development)

Preliminary Design Targets:
• PCI Express Gen 4
• 8 Lanes (or dual-4)
• Up to 50W power

http://www.ssdformfactor.org
Agenda

Transition to PCI Express Storage

• PCI Express SSD Market
• PCI Express Form Factors
• PCI Express Protocols (NVMe)
• PCI Express Scalability
NVMe Latency Measurements

- NVMe reduces latency overhead by more than 50%
  - SCSI/SAS: 6.0 µs 19,500 cycles
  - NVMe: 2.8 µs 9,100 cycles

- Designed for future NVM technology with sub-microsecond latency

Prototype Measured IOPS: 1.02M

Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat® EL6.0 2.6.32-71 Kernel.
NVMe™ Driver Ecosystem

Windows 8.1
6.5, 6.6, 6.7
7.0, 7.1
SLES 11
SP3 SLES 12
13, 14

Native / in-box
Install NVMe driver
Transition to PCI Express Storage

- PCI SSD Market
- PCI Form Factors
- PCI Express Protocols (NVMe)
- PCI Express Scalability
## U.2/NVMe Scalability

<table>
<thead>
<tr>
<th>Topology</th>
<th>Key reason to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU attach</td>
<td>Highest performance, lowest cost</td>
</tr>
<tr>
<td>Switch</td>
<td>Flexibility (x2, x4), high drive count (capacity)</td>
</tr>
<tr>
<td>PCH</td>
<td>Cheap attach, x1-x4 PCIe, don't use CPU lanes</td>
</tr>
<tr>
<td>Retimer</td>
<td>Inexpensive link extension of PCIe, routing distance</td>
</tr>
<tr>
<td>HBA/RAID card</td>
<td>Hardware RAID of NVMe devices, support for multiple protocols (SAS, SATA, PCIe)</td>
</tr>
</tbody>
</table>

![Diagram showing the connection between different components: PCIe 3.0 x4, Switch, Retimer, PCH, and HBA/RAID card.](image-url)
Goal: Establish viability of remote NVMe vs. local NVMe

- Within ~ 10 µs latency
- Minimal IOPS decrease

Result:

- 8 µs latency
- Zero IOPS decrease
NVMe over Fabrics

- Simplicity, Efficiency and End-to-End NVM Express (NVMe) Model
- Supports all known fabrics
- Specification targeted EOY '15, ratification Q1 ‘16
- Get involved – visit nvmexpress.org
Agenda

Transition to PCI Express Storage

NVDIMMs

Software Interface Standards
The NVDIMM Family and Cousins

- **DIMM**
  - Dual In-line Memory Module

- **NVDIMM-N**
  - Non-Volatile Dual In-line Memory Module

- **NVDIMM-F**
  - Flash Storage on the Memory Bus

NVDIMM Technology is Jointly Driven by JEDEC & SNIA
NVDIMM Taxonomy

NVDIMM

NVDIMM-N
- MCU interaction w/ DRAM only
- SW Access = Load/Store (Block &/or Byte)
- No data copy during access
- Capacity = DRAM

NVDIMM-F
- MCU interaction w/ RAM buffer only
- SW Access = Block
- Minimum one data copy required during access
- Capacity = Non Volatile Memory (NVM)

Engineering is developed by JEDEC
NVDIMM Software Architecture

- **User space**
  - Applications
  - Management Software

- **Kernel space**
  - OS Stack
    - Block Layer
  - Flash DIMM Kernel Driver

- **Hardware**
  - BIOS/UEFI
  - FlashDIMM Firmware
  - Flash Controller Firmware

Diagram showing the software architecture of NVDIMM with layers for user space, kernel space, and hardware, including components like Applications, Management Software, OS Stack, Block Layer, Flash DIMM Kernel Driver, FlashDIMM Firmware, and Flash Controller Firmware.
SNIA: NVDIMM SIG

- **Education**
  - Helping technology and solution vendors whose products integrate NVDIMMs to communicate their benefits and value to the greater market
  - Developing vendor-agnostic user perspective case studies, best practices, and vertical industry requirements

- **Software Engineering**
  - NVM Programming TWG
Persistent memory programming model
(pmем-aware file system)

- With Pmem, no paging between persistence and volatile memory
- Memory map command causes pmem file to be mapped to app's virtual memory
- Sync command flushes CPU cache
- Load/Store commands directly access pmem
## NVDIMM Enabling on Intel Platforms

<table>
<thead>
<tr>
<th>Document</th>
<th>Description</th>
<th>Ownership/Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACPI 6.0 Spec</td>
<td>ACPI 6.0 is a BIOS specification that allows NVDIMMs to be described by platform firmware to OS/VMM via Nonvolatile Memory Firmware Interface Table (NFIT).</td>
<td>Developed by ACPI Working Group Published by UEFI forum Owned by ACPI</td>
</tr>
<tr>
<td>NFIT Device Driver source code</td>
<td>Open Source code that will serve as the foundation for developing a Linux NVDIMM NFIT-compatible driver for integration into Linux kernels in 2015</td>
<td>Intel Copyrighted GPL licensed published to Linux Kernel developers repository</td>
</tr>
<tr>
<td>NVDIMM Namespace Specification</td>
<td>Describes a method for sub-dividing persistent memory into <em>namespaces</em>, which are analogous to NVM Express Namespaces. The document also describes the <em>Block Translation Table</em> (BTT) layout which provides single sector write atomicity for block devices built on pmem</td>
<td>Developed by Intel. <a href="http://pmem.io/">http://pmem.io/</a></td>
</tr>
<tr>
<td>NVDIMM DSM Interface Example</td>
<td>Targeted to writers of BIOS and OS drivers for NVDIMMs whose design adheres to the NFIT Tables in the ACPI V6.0 specification. The document specifically discusses the NVDIMM Device Specific Method (_DSM) example.</td>
<td>Developed by Intel. <a href="http://pmem.io/">http://pmem.io/</a></td>
</tr>
<tr>
<td>NVDIMM Driver Writer's Guide</td>
<td>Targeted to driver writers for NVDIMMs that adhere to the NFIT tables in the Advanced Configuration and Power Interface (ACPI) V6.0 specification, the Device Specific Method (DSM) specification and the NVDIMM Namespace Specification. This document specifically discusses the block window HW interface and persistent memory interface that Intel is proposing for NVDIMMs.</td>
<td>Developed by Intel. <a href="http://pmem.io/">http://pmem.io/</a></td>
</tr>
</tbody>
</table>
Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

- New media that enables broad memory/storage convergence
- NVM Programming Technical Working Group
- Benchmarking after the transition to new media
Next Generation Scalable NVM

Resistive RAM NVM Options

<table>
<thead>
<tr>
<th>Family</th>
<th>Defining Switching Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase Change Memory</td>
<td>Energy (heat) converts material between crystalline (conductive) and amorphous (resistive) phases</td>
</tr>
<tr>
<td>Magnetic Tunnel Junction (MTJ)</td>
<td>Switching of magnetic resistive layer by spin-polarized electrons</td>
</tr>
<tr>
<td>Electrochemical Cells (ECM)</td>
<td>Formation / dissolution of &quot;nano-bridge&quot; by electrochemistry</td>
</tr>
<tr>
<td>Binary Oxide Filament Cells</td>
<td>Reversible filament formation by Oxidation-Reduction</td>
</tr>
<tr>
<td>Interfacial Switching</td>
<td>Oxygen vacancy drift diffusion induced barrier modulation</td>
</tr>
</tbody>
</table>

Scalable Resistive Memory Element

Cross Point Array in Backend Layers \(\sim 4\lambda^2\) Cell

~ 1000x speed-up over NAND.
Opportunities with Next Generation NVM

Normal Technology Advancements Drive Higher Performance
Opportunities with Next Generation NVM

Next Generation NVM -- no longer the bottleneck
Opportunities with Next Generation NVM

**Interconnect & Software Stack Dominate the Access Time**

- NAND MLC SATA 3 ONFI 2
- NAND MLC SATA 3 ONFI 3
- NAND MLC PCIe Gen 3 ONFI 3
- Future NVM PCIe x4 Gen 3
- Future NVM PCIe x4 Gen 3

Diagram showing the contribution of different components to SSD IO Read Latency (us, QD=1, 4KB): NVM Tread, NVM Xfer, Misc SSD, Link Xfer, Platform + Adapter, Software.
Opportunities with Next Generation NVM

NVM Express: Optimized platform storage interconnect & driver
SNIA NVM Programming TWG: Optimized system & application software
Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

- New media that enables broad memory/storage convergence
- NVM Programming Technical Working Group
- Benchmarking after the transition to new media
SNIA: NVM Programming TWG

SNIA TWG Focus
- 4 Modes of NVM Programming
  - NVM File, NVM Block, NVM PM Volumes, NVM PM Files
  - Ensure Programming model covers the needs of ISVs

Unified NVM Programming Model
Version 1.0 approved by SNIA - (Publically available)
## SNIA NVM Programming Model

### The Four Modes

<table>
<thead>
<tr>
<th>Block Mode Innovation</th>
<th>Emerging NVM Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Atomics</td>
<td>• Performance</td>
</tr>
<tr>
<td>• Access hints</td>
<td>• Performance</td>
</tr>
<tr>
<td>• NVM-oriented operations</td>
<td>• Perf... okay, cost</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Traditional</th>
<th>Persistent Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>User View</td>
<td>NVM.FILE</td>
<td>NVM.PM.FILE</td>
</tr>
<tr>
<td>Kernel Protected</td>
<td>NVM.BLOCK</td>
<td>NVM.PM.VOLUME</td>
</tr>
<tr>
<td>Media Type</td>
<td>Disk Drive</td>
<td>Persistent Memory</td>
</tr>
<tr>
<td>NVDIMM</td>
<td>Disk-Like</td>
<td>Memory-Like</td>
</tr>
</tbody>
</table>

Flash Memory Summit 2014
Santa Clara, CA
Application View of I/O Elimination

Typical NUMA range: 0 - 200 nS
Typical context switch range: above 2-3 uS
Memory Mapped Files
Eliminate File System Latency

Traditional

User
Application

Kernel
File System
Disk Driver

HW
Disk

New

User
Application

HW
Persistent Memory
Load/Store
Memory Mapped Files
Programming Model Context

1. NVM as Storage
   - Existing Applications
   - Management Applications (GUI, CLI, CIM)

2. NVM-Optimized Applications
   - NVM User-Space API
   - NVM-Optimized Kernel Modules
     - NVM Kernel API
     - NVM Driver Stack

3. NVM as Memory
   - "Persistent Memory"

- SNIA NVM Programming TWG
- Linux* Open Source Project, Microsoft *, other OSVs
- Existing/Unchanged Infrastructure
Agenda

Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

• New media that enables broad memory/storage convergence
• NVM Programming Technical Working Group
• Benchmarking after the transition to new media
Comparison – which is the “better” SSD?

Two Different Example SSD Load Lines

4X – 100X faster in common operating region

20% fewer IOPS

Conclusion: A New Metric for Measuring SSDs is Needed
Thank You!
Questions & Answers

Jim Pappas
jim@intel.com
INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY
ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN
INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY
EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES
RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER
INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or
death. SHOULD YOU PURCHASE OR USE INTEL’S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND
HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH,
HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY
OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL
APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE
INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics
of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for
conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with
this information.

The products described in this document 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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-
4725, or go to: http://www.intel.com/design/literature.htm

Intel, Look Inside and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

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

Copyright ©2013 Intel Corporation.
Risk Factors

The above statements and any others in this document that refer to plans and expectations for the third quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel’s and competitors’ products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel's response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, and political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel's results could be affected by the timing of closing of acquisitions and divestitures. Intel’s results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel's SEC filings, including the company’s most recent reports on Form 10-Q, Form 10-K and earnings release.