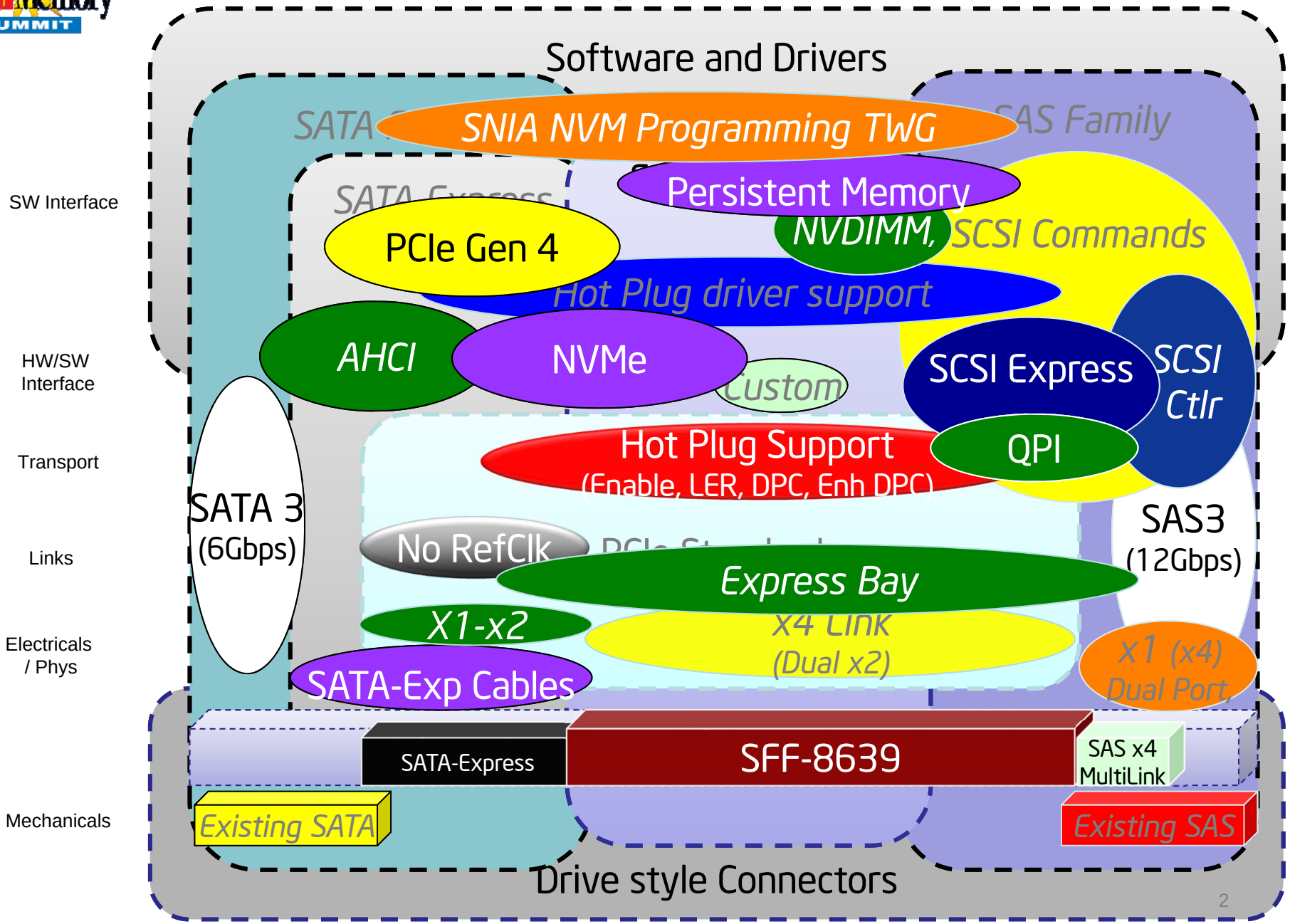
Annual Update on Interfaces

Session U-3

Jim Pappas

jim@intel.com

Agenda


Software and Drivers
SW Interface
HW/SW Interface
Transport
Links
Electricals / Phys
Mechanicals
SATA
SAS Family
SNIA NVM Programming TWG
Persistent Memory
NVDIMM, SCSI Commands
PCIe Gen 4
Hot Plug driver support
AHCI
NVMe
Custom
SCSI Express
SCSI Ctlr
QPI
Hot Plug Support
(Enable, LER, DPC, Enh DPC)
SATA 3
(6Gbps)
No RefClk
SAS3
(12Gbps)
Express Bay
X1-x2
X4 Link
(Dual x2)
x1 (x4)
Dual Port
SATA-Exp Cables
SATA-Express
SFF-8639
SAS x4
MultiLink
Existing SATA
Existing SAS
Drive style Connectors

Agenda

Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

Agenda

Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

PCIe Storage Strengths

Current PCIe SSD cards:
Well received by customers, have attained highest performance to date.

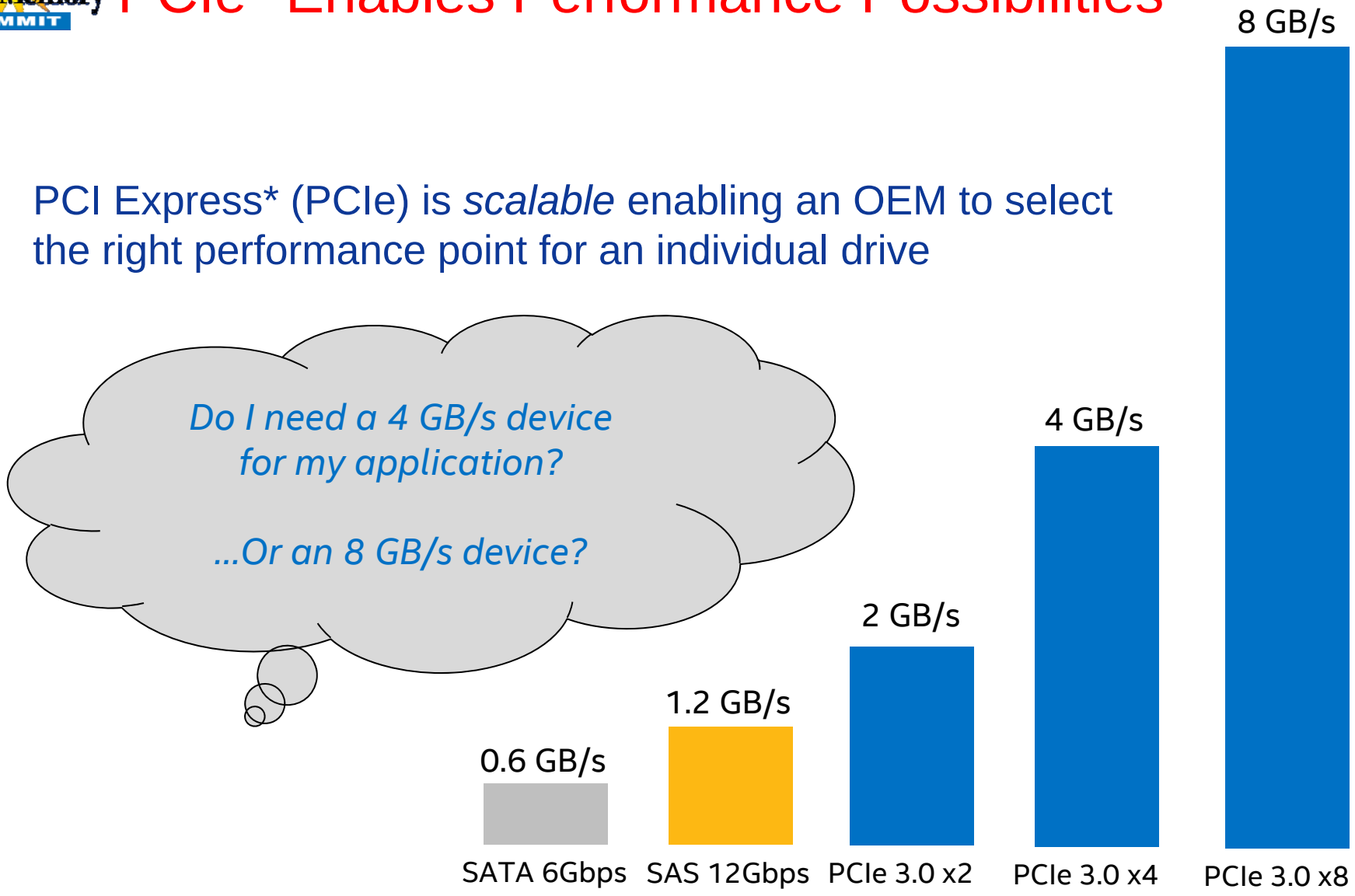
Complement existing storage protocols:
Providing highest IOPs and lowest latency for demanding applications

Obvious advantages: Reduced path components

- Lower costs
- Less real estate
- Less power
- Higher reliability
- Lower latency

PCIe* Enables Performance Possibilities

PCI Express* (PCIe) is *scalable* enabling an OEM to select the right performance point for an individual drive



Source: www.pcisig.com, www.t10.org, www.sata-io.org

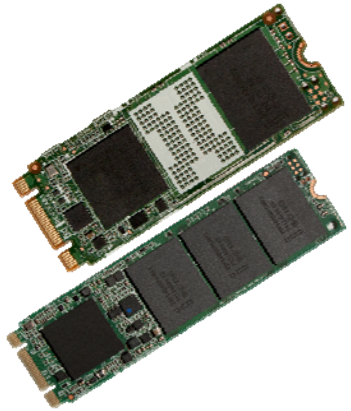
Agenda

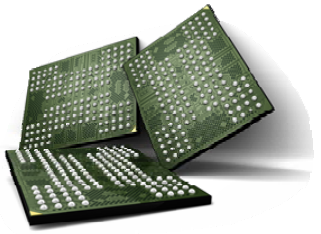
Transition to PCI Express Storage

- PCI Express Form Factors
- PCI Express Protocols
- PCI Express Electrical Signal Integrity
- PCI Express Scalability

PCI Express Form Factors
Client

- There are a variety of form factors depending on need
- M.2 is an optimized SSD only form factor for laptops
- The SATA Express connector supports 2.5” form factor drives when allowing for HDD/SSHD is needed
- Intel, SanDisk, Toshiba, and HP proposed a BGA solution for standardization in PCI SIG for behind-the-glass usages (e.g., 2-in-1 laptops) – join PCI SIG to participate!





VAIO*Pro 13 Ultrabook™

The world's lightest 13.3" touch Ultrabook[21].

Features:

- 4th gen Intel® Core™ i7 processor available
- Windows 8 Pro available
- Full HD TRILUMINOS IPS touchscreen (1920 x 1080)
- Super fast 512GB PCIe SSD available
- Ultra-light at just 2.34 lbs.

Form factors are there to support PCIe adoption in client in 2015.

SSD Form Factor Working Group

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

PCI Express Form Factors

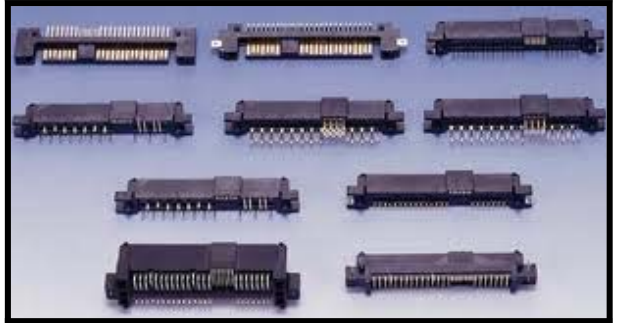
SSD Form Factor Working Group – SFF 8639

Form Factor

Benefit from current 2.5" HDD form factor

Expand power envelope

Connector



Multiple protocols:

PCIe 3.0, SAS 3.0, SATA 3.0

Management Bus

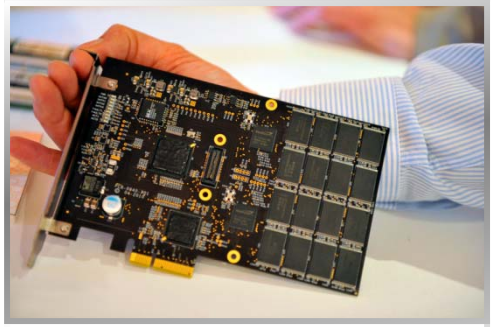
Dual port (PCIe)

Multi-lane capability (PCIe/SAS)

Power pins

SAS Drive Backward Compatibility

Hot-Plug



Hot-Plug Connector

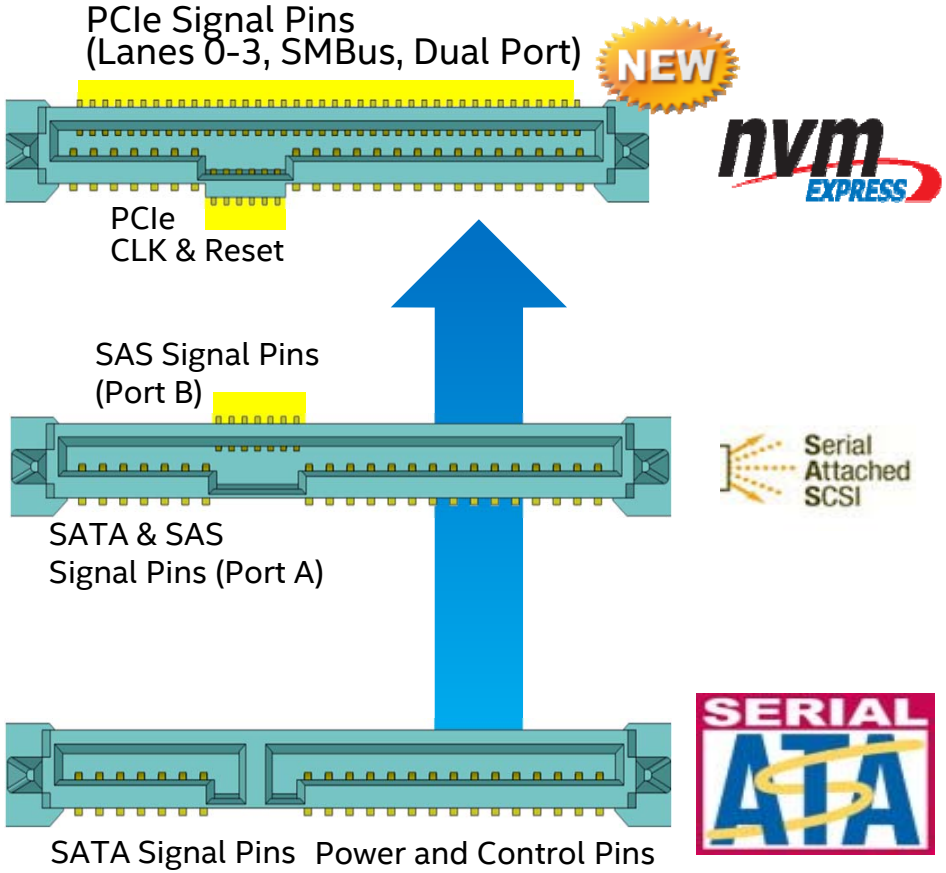
Identify desired drive behavior

Define required system behavior

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

PCI Express Form Factors

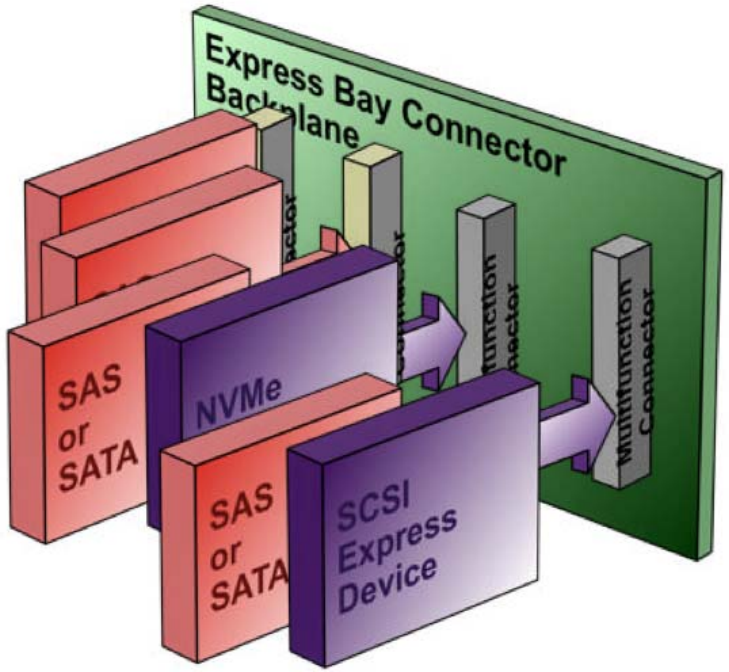
SSD Form Factor Working Group – SFF 8639



- A serviceable (hot pluggable) form factor is critical in Datacenter
- The SFF-8639 form factor / connector supports NVMe, SAS, and SATA

PCI Express Form Factors

SCSI Trade Association – Express Bay

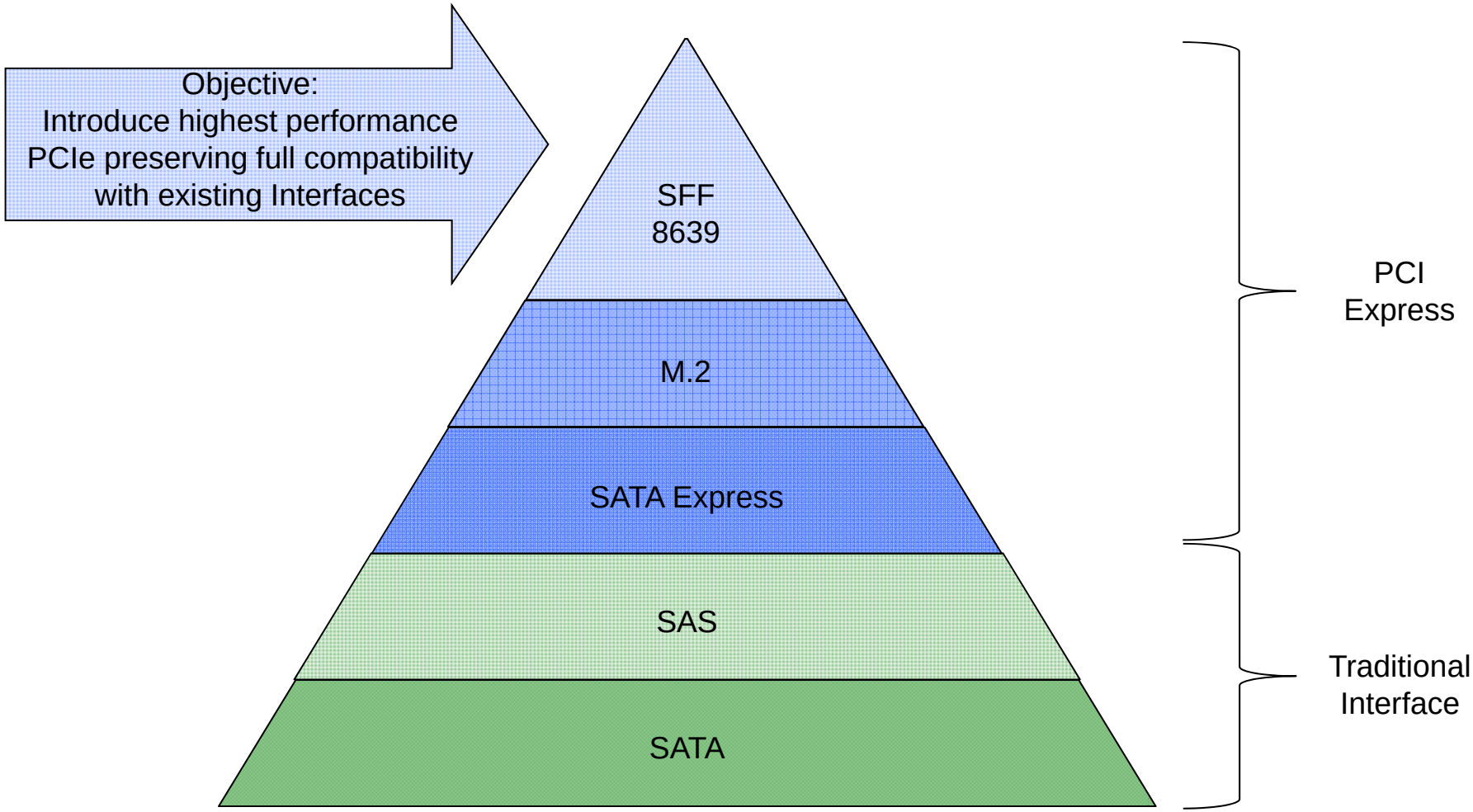


- Express Bay
  - Up to 25 Watts
    - For both SAS and PCIe
  - SFF-8639 connector
  - PCI-SIG electrical specification
- Objectives
  - Preserve the enterprise storage experience for PCI Express storage
  - Meet SSD performance demands
  - Open up new possibilities with serviceable, hot-pluggable Express Bay

Note: Specific configuration, protocol and power support will be OEM specific

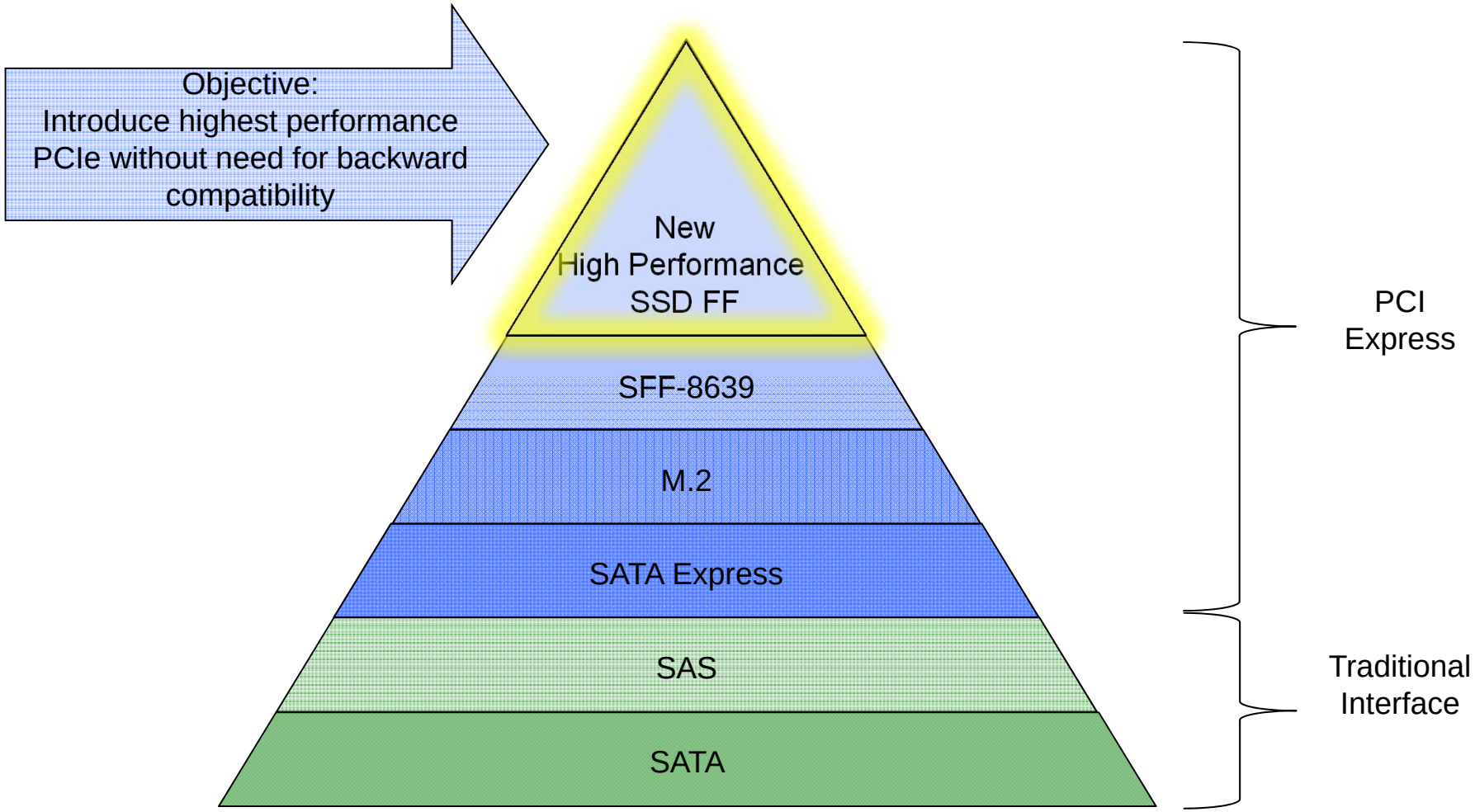
PCI Express Form Factors

SSD Form Factor Working Group – SFF 8639



PCI Express Form Factors

SSD Form Factor Working Group – Next Generation Form Factor



PCI Express Form Factors

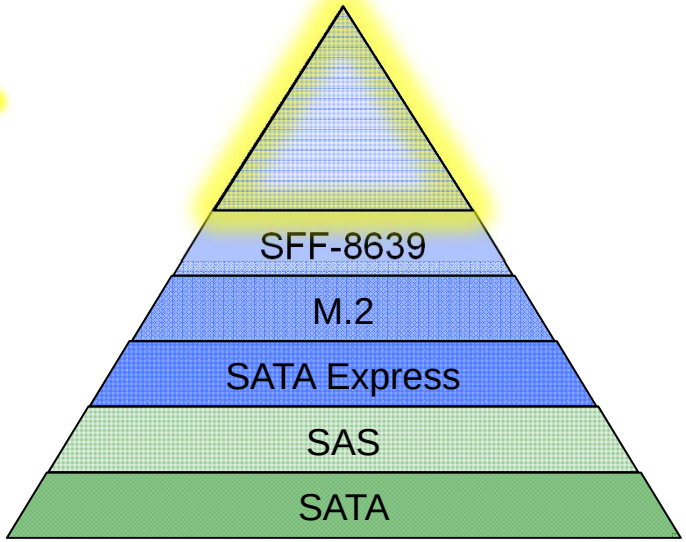
SSD Form Factor Working Group – Next Generation Form Factor

Announcing creation of a new project within the SSD FF WG!

Preliminary Design Targets:

- PCI Express Gen 4
- 8 Lanes (or dual-4)
- Up to 50W power

Each element is doubled from SFF-8639



Status:

- New working group approved by SSD FF WG Promoters
- Definition to begin Aug ’14
- Participation open to all 60 SSD FF WG members
- Join at http://www.ssdformfactor.org

Agenda

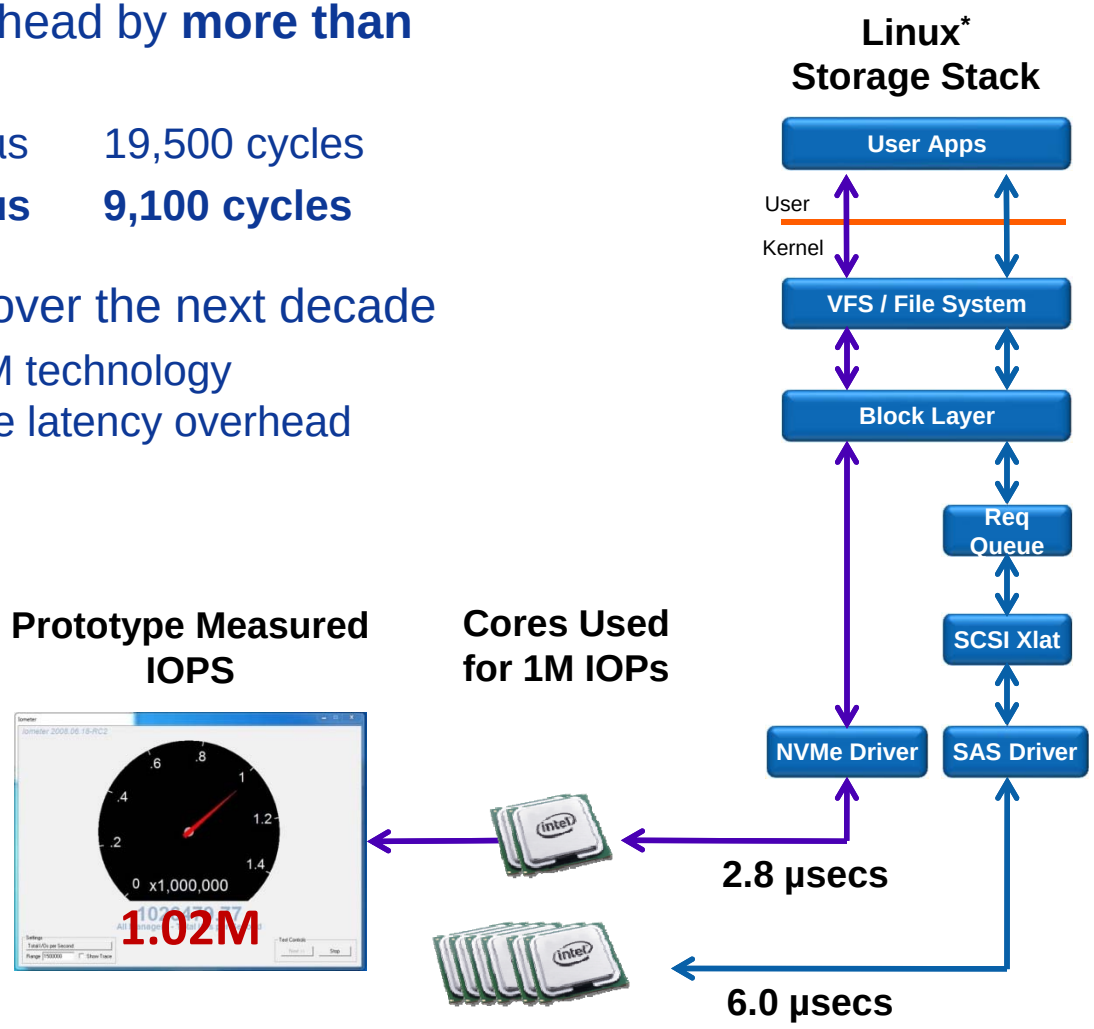
Transition to PCI Express Storage

- PCI Express Form Factors
- PCI Express Protocols
- PCI Express Electrical Signal Integrity
- 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**
- NVMe is designed to scale over the next decade
  - NVMe supports future NVM technology developments that will drive latency overhead below one microsecond

SCSI Express is a complementary effort to maintain SCSI compatibility over PCIe



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.

NVM Express Overview

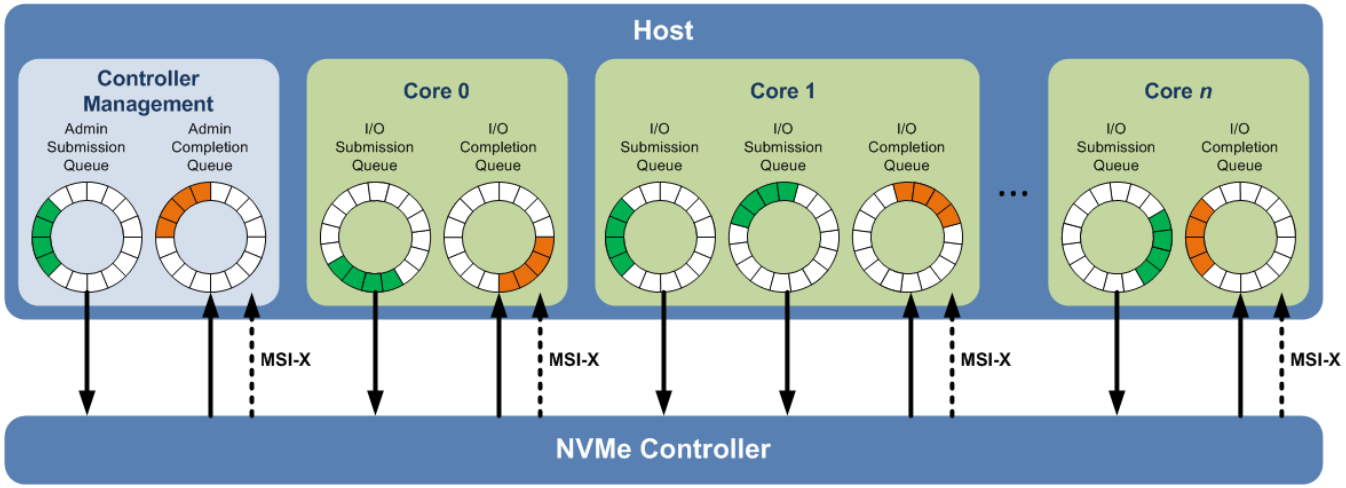
- NVM Express (NVMe) is a standardized high performance host controller interface for PCIe Storage, such as PCIe SSDs
  - Standardizes register set, feature set, and command set where there were only proprietary PCIe solutions before
  - Architected from the ground up for NAND and next generation NVM
  - Designed to scale from Enterprise to Client systems

- NVMe was developed by an open industry consortium and is directed by a 13 company Promoter Group

Technical Basics of NVMe

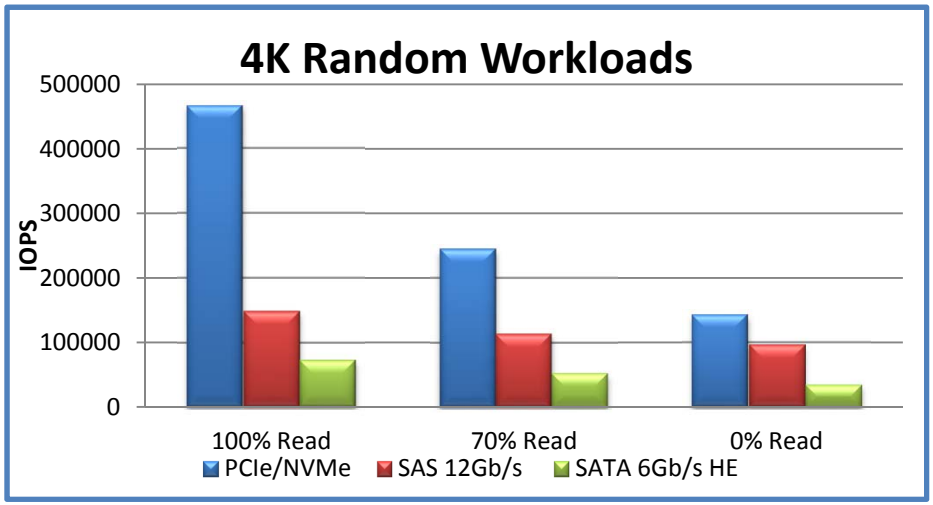
- All parameters for 4KB command in single 64B command
- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports MSI-X and interrupt steering
- Streamlined & simple command set (13 required commands)
- Optional features to address target segment (Client, Enterprise, etc.)

- Designed to scale for next generation NVM, agnostic to NVM type used

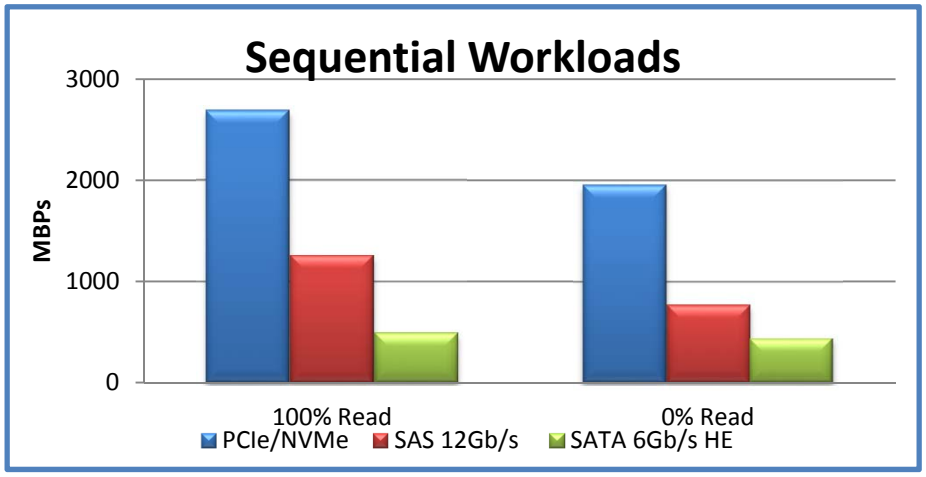


NVMe Delivers on Performance

- NVM Express delivers versus leadership SAS & SATA products
- Random Workloads
  - > 2X performance of SAS 12Gbps
  - 4-6X performance of SATA 6Gbps
- Sequential Workloads
  - Realize almost 3 GB/s reads
  - > 2X performance of SAS 12Gbps
  - > 4X performance of SATA 6Gbps



Source: Intel (PCIe/NVMe @ QD 128 ; SAS/SATA @ QD32



Source: Intel (PCIe/NVMe 128K @ QD 128 ; SATA 128K @ QD32; SAS 64K @QD32

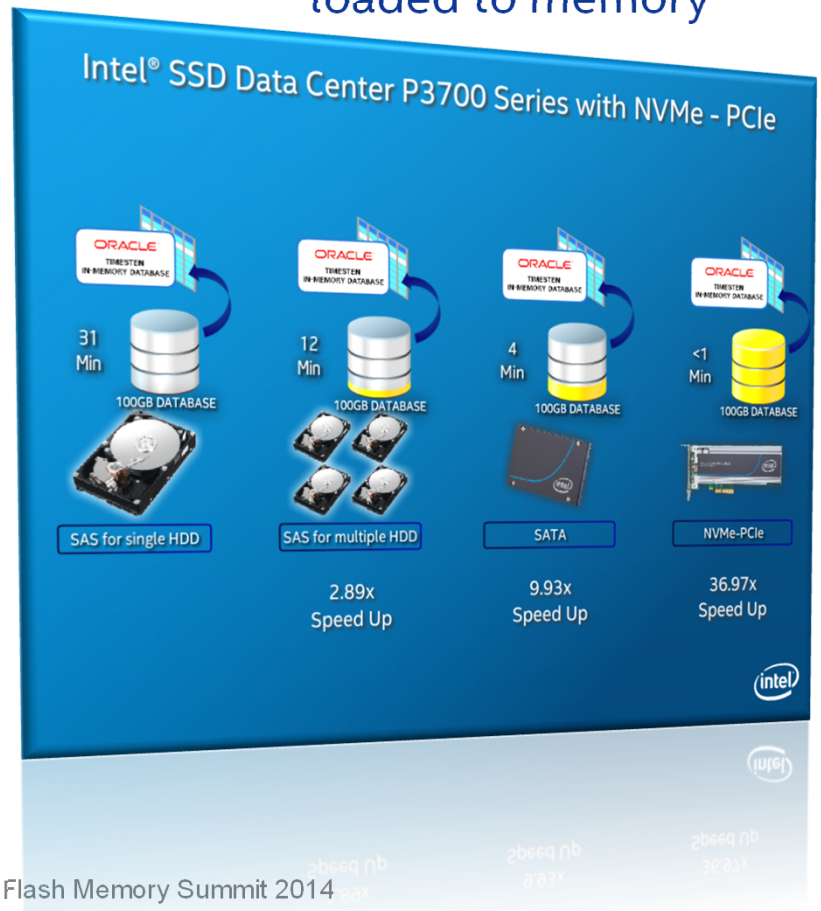
Note: PCIe/NVMe Measurements made on Intel(R)Core(TM) i7-3770S system @ 3.1GHz and 4GB Mem running Windows Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by IOmeter* tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar SSD800M/1000M (SAS) Solid State Drive Specification.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

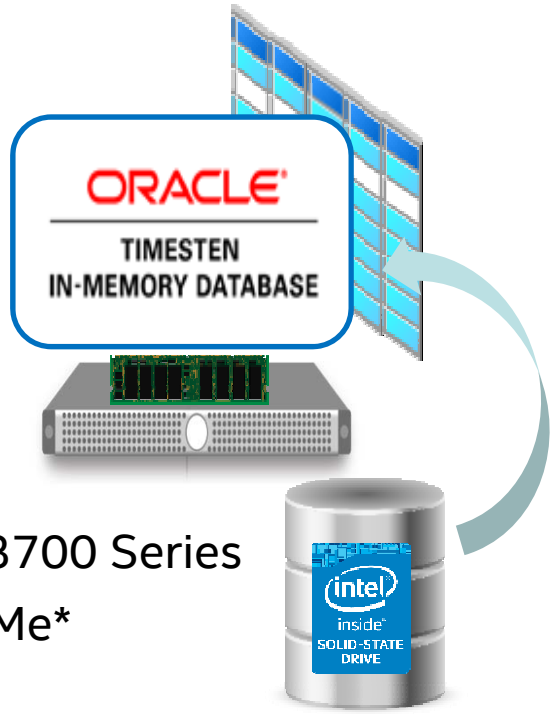
- Other names and brands are property of their respective owners

Real Workload Benefit of NVMe

Oracle TimesTen In-Memory Database Restart

Oracle TimesTen database checkpoint file loaded to memory





Intel® SSD DC P3700 Series with NVMe*

The products described in this document may vary based on design defects, numbers, or errors known as errata which may cause the product to deviate from published specifications.

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

Driver Ecosystem Flourishing

Windows*	• Windows* 8.1 and Windows* Server 2012 R2 include native driver • Open source driver in collaboration with OFA
Linux*	• Stable OS driver since Linux* kernel 3.12
Unix	• FreeBSD driver upstream
Solaris*	• Solaris driver will ship in S12
VMware*	• Open source vmklinux driver available on SourceForge
UEFI	• Open source driver available on SourceForge

Native OS drivers already available, with more coming.

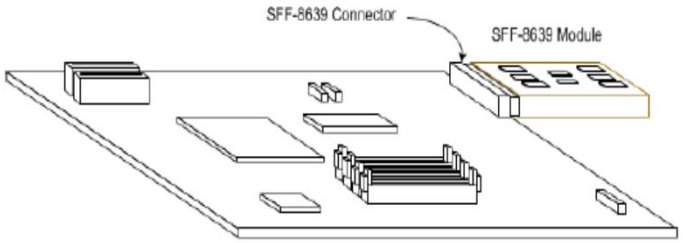
Agenda

Transition to PCI Express Storage

- PCI Express Form Factors
- PCI Express Protocols
- PCI Express Electrical Signal Integrity
- PCI Express Scalability

PCI Express Electrical Signal Integrity

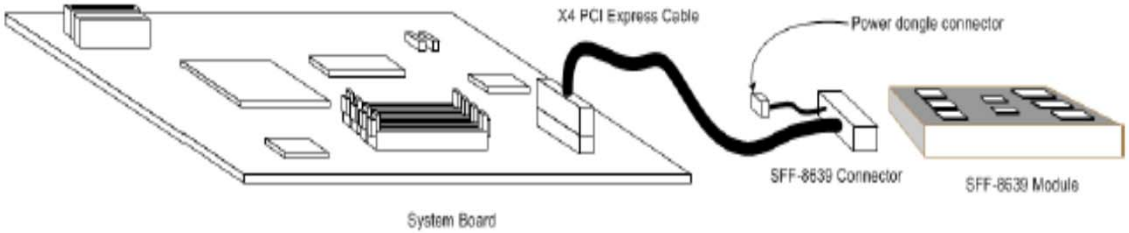
The richness of PCIe topologies introduce specific challenges to PCIe implementation. Some examples:

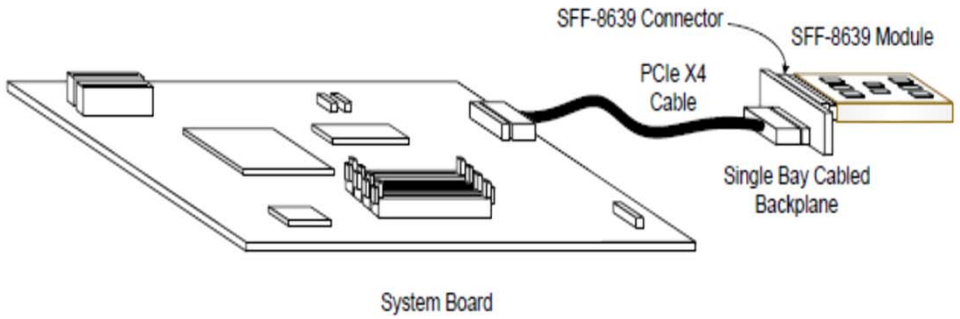


1: Single Connector – straight forward

2: Cable & Connector

- Now with 2 signal segments





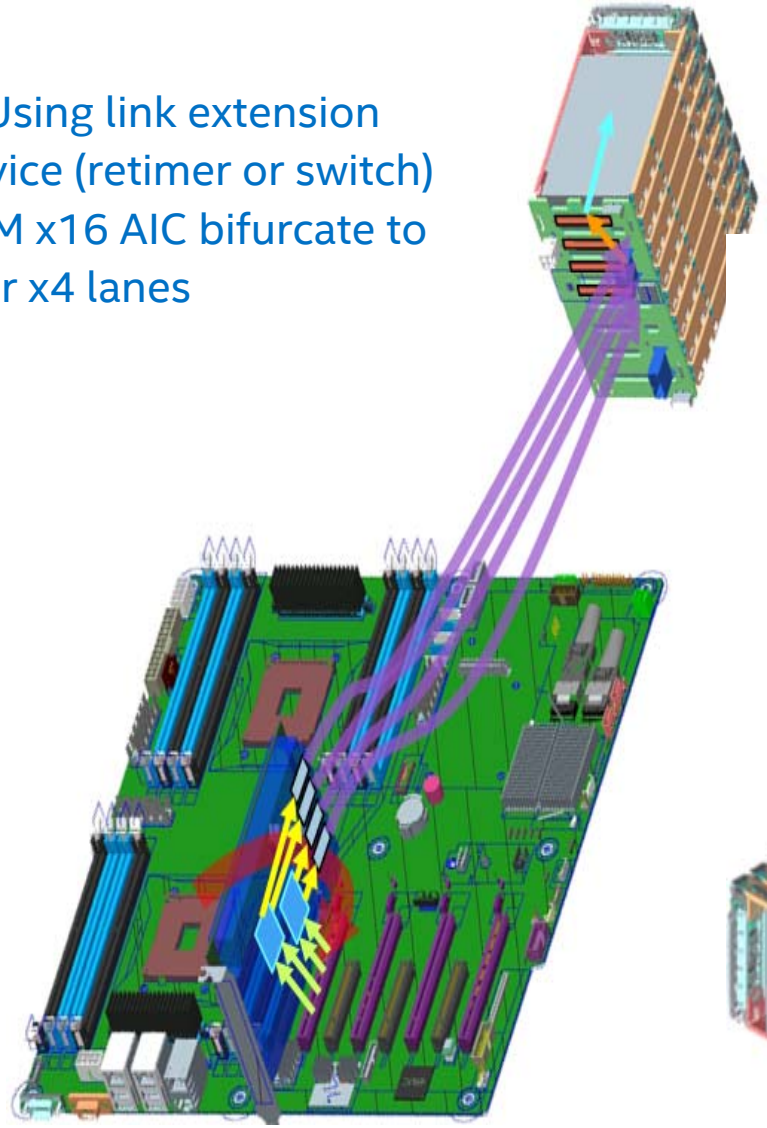
3: Cable, Connector, & backplane

- Now with 3 signal segments

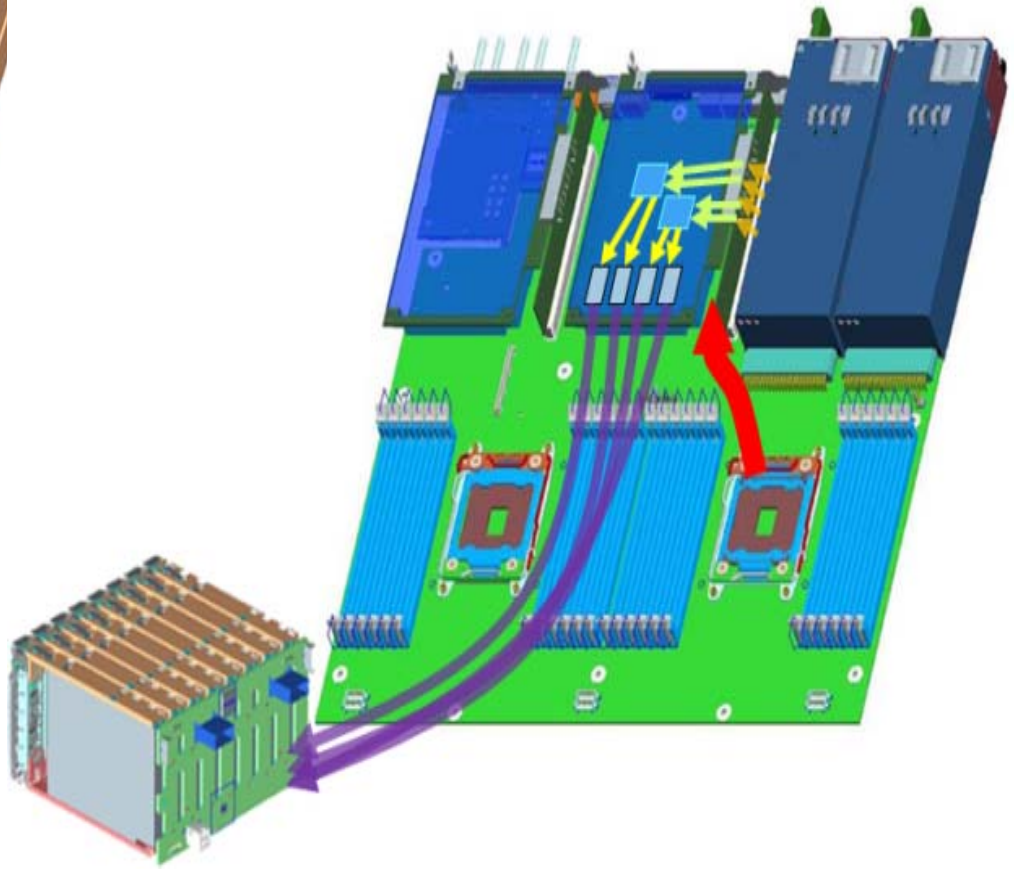
PCI Express Electrical Signal Integrity

More complex PCIe Topologies

4: Using link extension device (retimer or switch) CEM x16 AIC bifurcate to four x4 lanes



5: Using PCIe riser card plus a retimer or switch



PCI Express Electrical Signal Integrity

Summary

- Challenging Designs
  - High speed signals are hard to implement
  - SSDs are pushing the limits of PCIe implementations
  - Storage device companies are relatively inexperienced with PCIe
- Positive elements
  - We know how to do this!
  - The industry is taking responsibility to solve any issues
  - Plugfest, workshops, dedicated test fixtures, etc. are ongoing… and the issues are getting solved
- Call to action:
  - Attend Panel I-31: PCIe Storage - Thursday 8:30-10:50
    - Presentation: SFF 8639 PCIe SSD Ecosystem Readiness And Electrical Testing
    - Don Verner, JohnMichael Hands, Dan Froelich

Agenda

Transition to PCI Express Storage

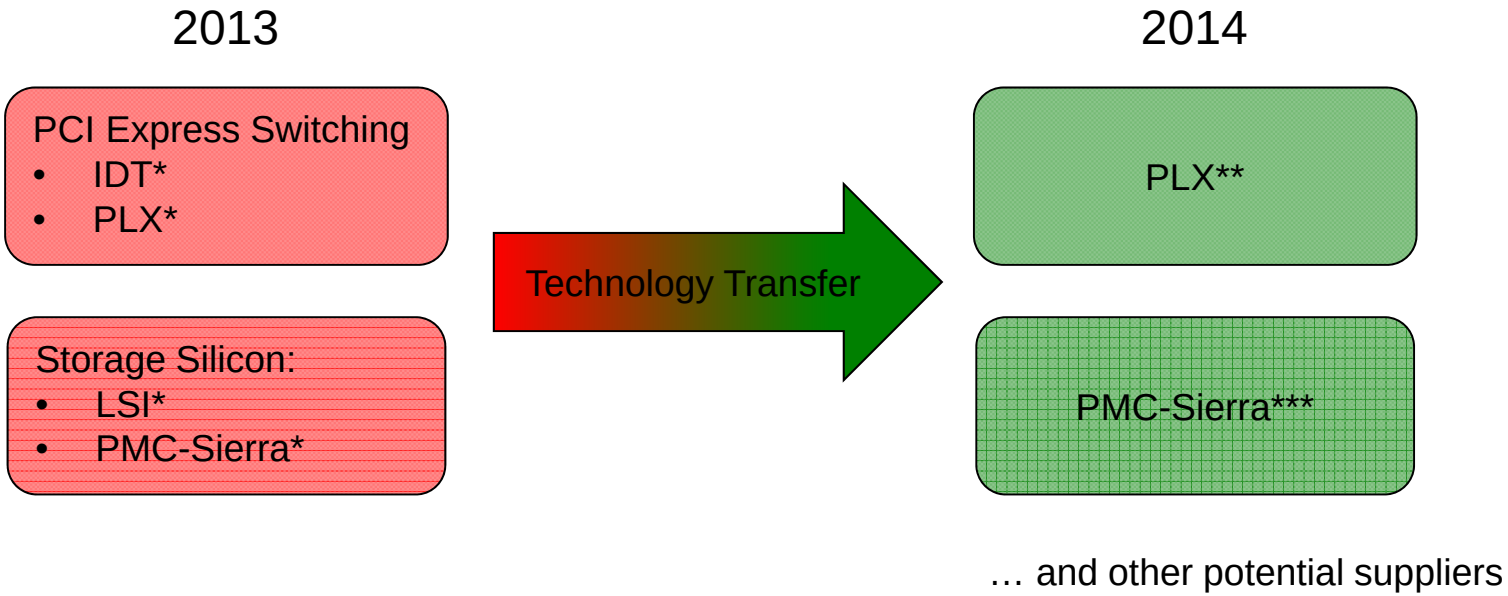
- PCI Express Form Factors
- PCI Express Protocols
- PCI Express Electrical Signal Integrity
- PCI Express Scalability

PCI Express Scalability

Scaling to dozens… or hundreds of storage nodes

Scaling PCIe requires expertise in 2 key technology areas:

- PCI Express Switching Silicon
- Storage Silicon



… and other potential suppliers

Industry is well poised for success in PCI Express Large Scale Deployments

* Other names and brands may be claimed as the property of others.
** PLX acquisition by Avago announced on June 23, 2014 is pending
*** PMC acquired certain PCI Express (PCIe) Switch assets from IDT on July 15, 2013

Agenda

Transition to PCI Express Storage

NVDIMMs

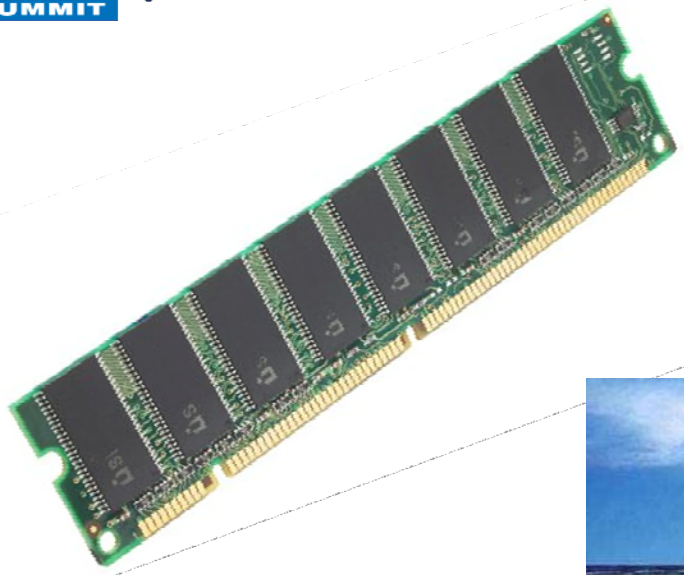
Software Interface Standards

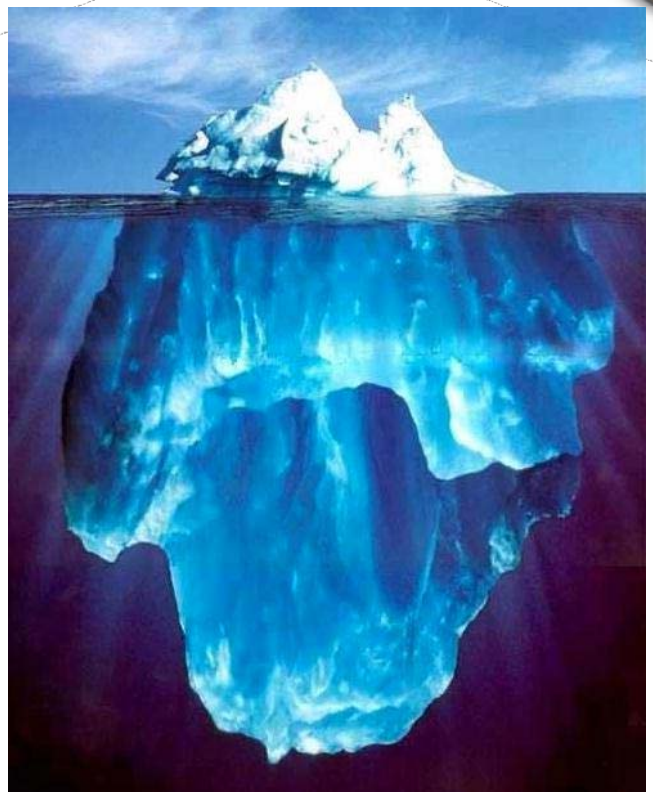
Some things in life do not always work well together

Neither Spoon nor Fork

My Personal Favorite

And at first blush…

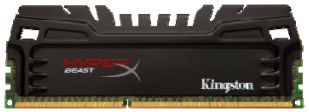




Convergence: "Memory Performance" and "Storage Durability"

The NVDIMM Family and Cousins

- DIMM
  - Dual In-line Memory Module



- NVDIMM
  - Non-Volatile Dual In-line Memory Module
- FLASH DIMM
  - Flash Storage on the Memory Bus

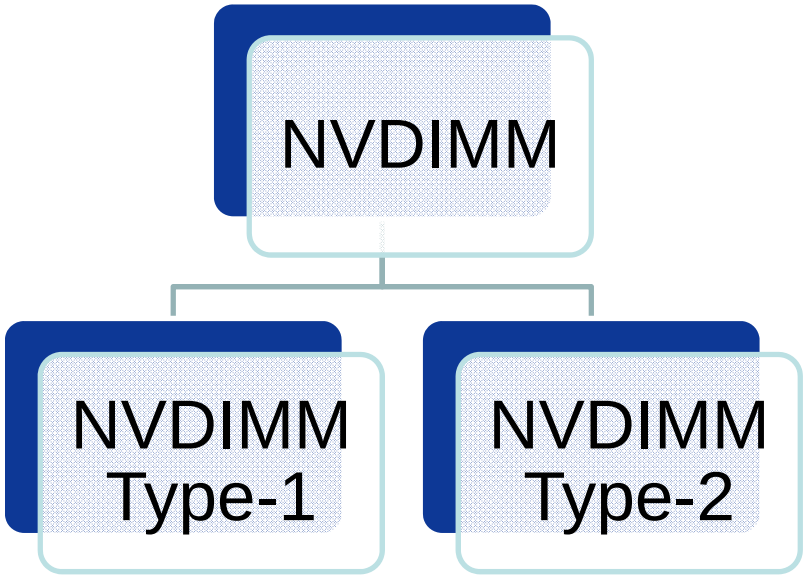
- SATA DIMM
  - DIMM form factor SSD, powered by Memory Bus

SNIA: NVDIMM SIG

Established: January 2014

- Providing education on how system vendors can design in NVDIMMs
- Communicating existing industry standards, and areas for vendor differentiation
- 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

NVDIMM Taxonomy



Add new "Type Definitions" as new NVDIMM solutions emerge

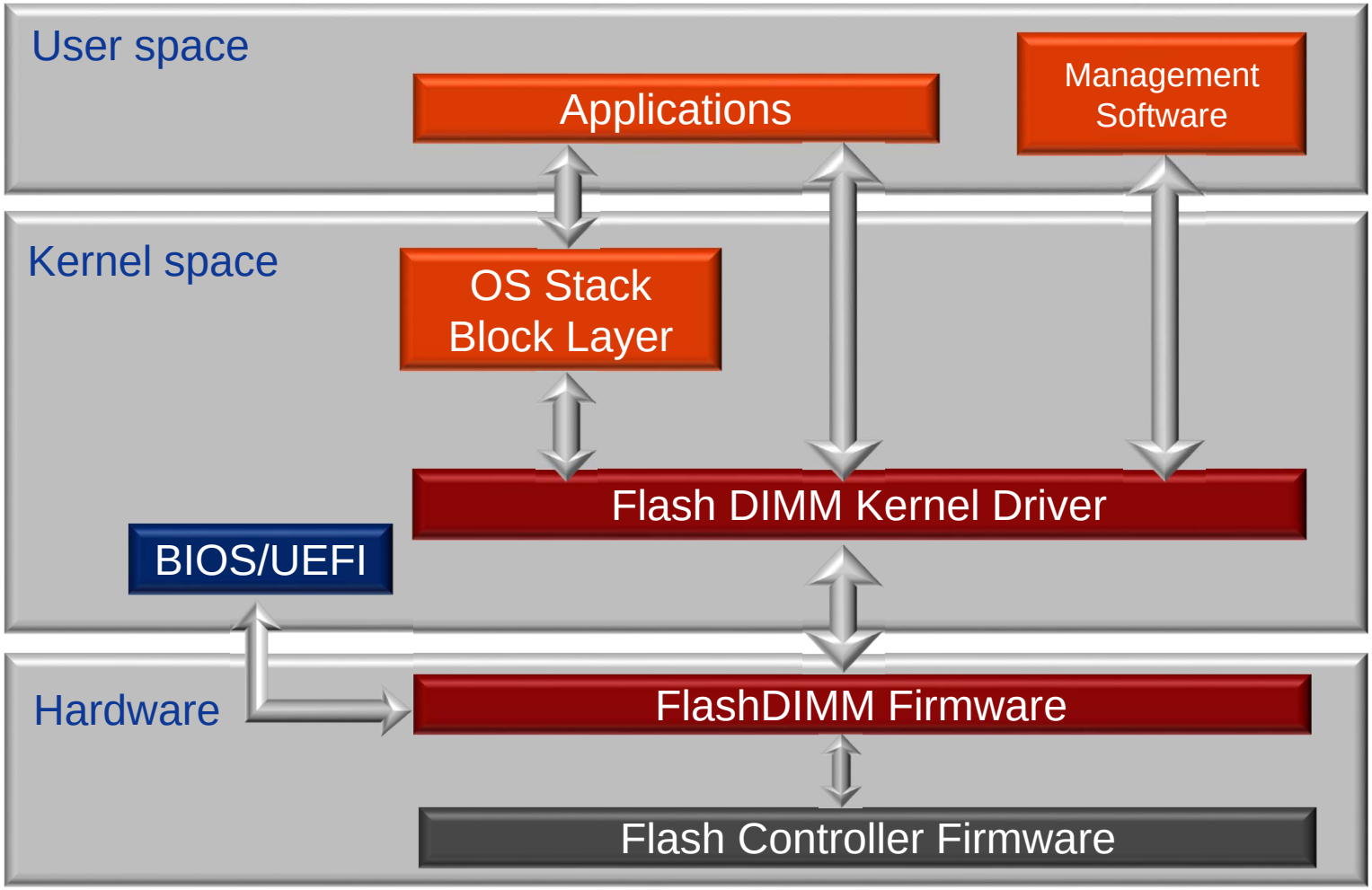
NVDIMM Type-1

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

NVDIMM Type-2

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

NVDIMM Software Architecture



NVDIMM SIG

- Join the NVDIMM SIG to help guide the direction of this new technology
- www.snia.org/forums/sssi/nvdimm

Agenda

Transition to PCI Express Storage

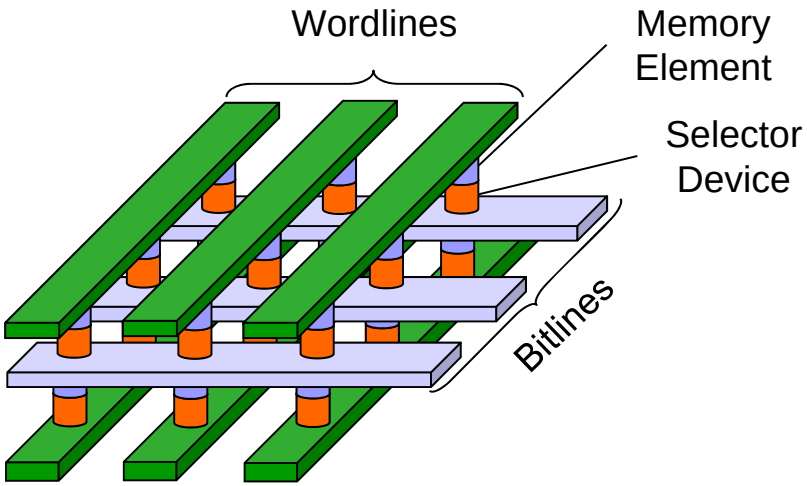
NVDIMMs

Software Interface Standards

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

Next Generation Scalable NVM

Scalable Resistive Memory Element



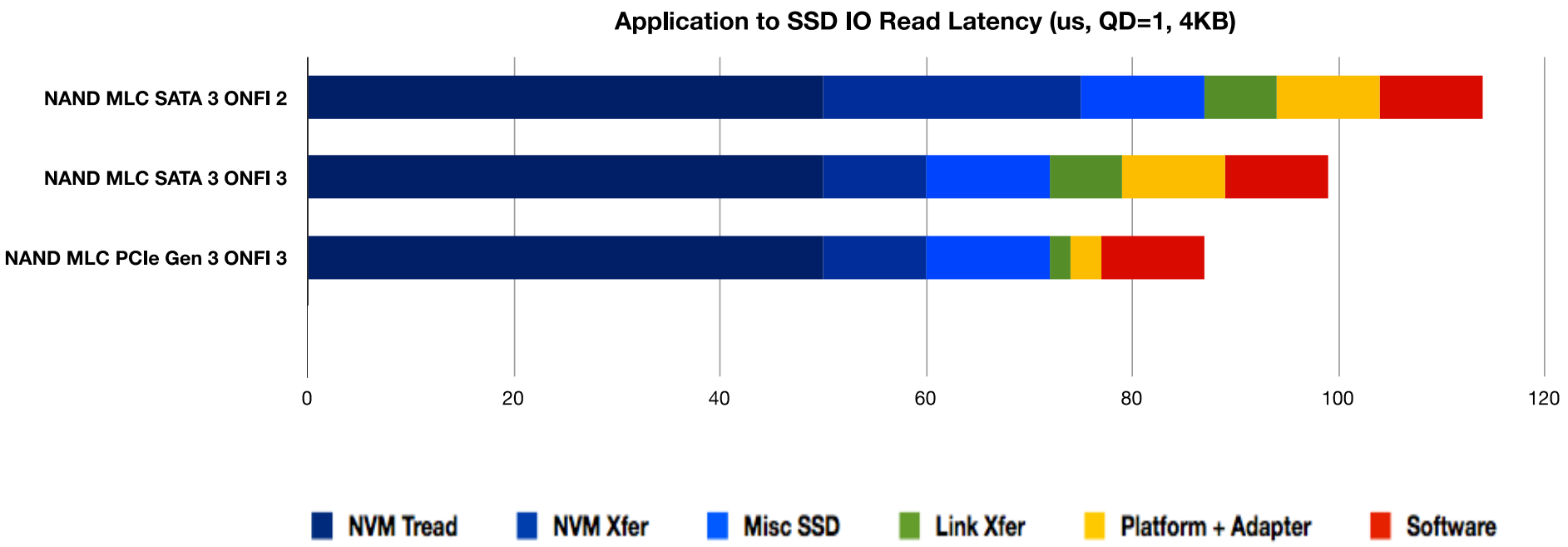
Cross Point Array in Backend Layers ~$4\lambda^2$ Cell

Resistive RAM NVM Options

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

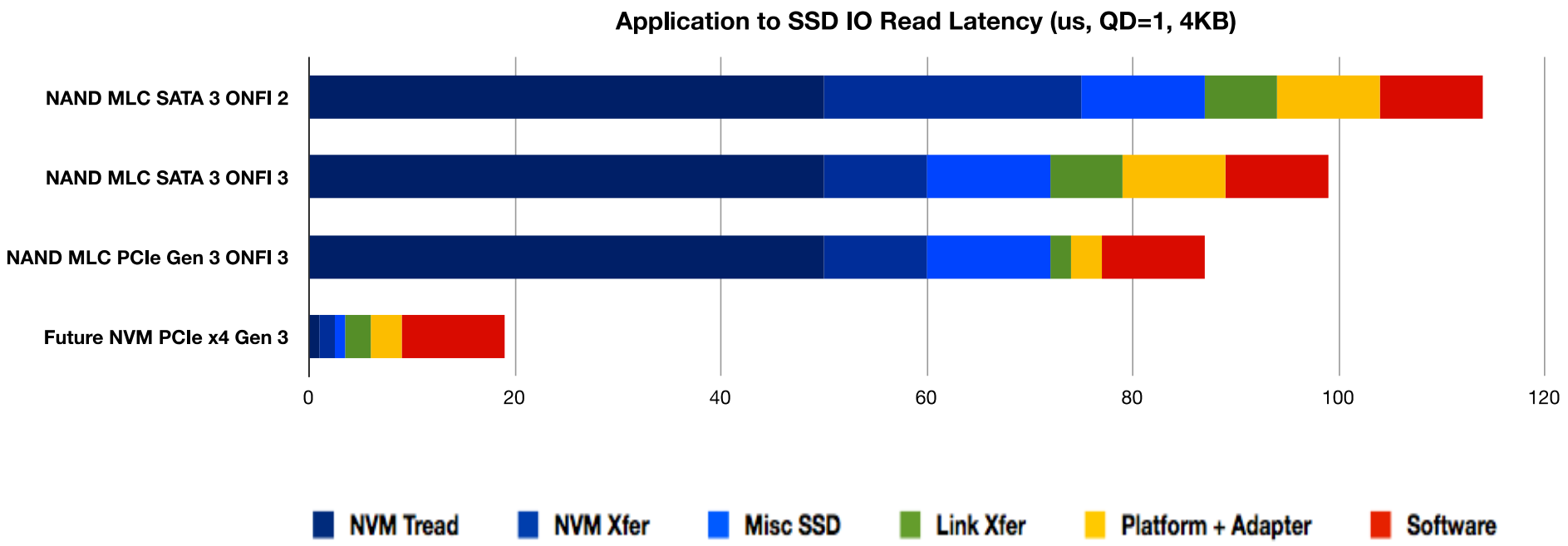
~ 1000x speed-up over NAND.

Opportunities with Next Generation NVM



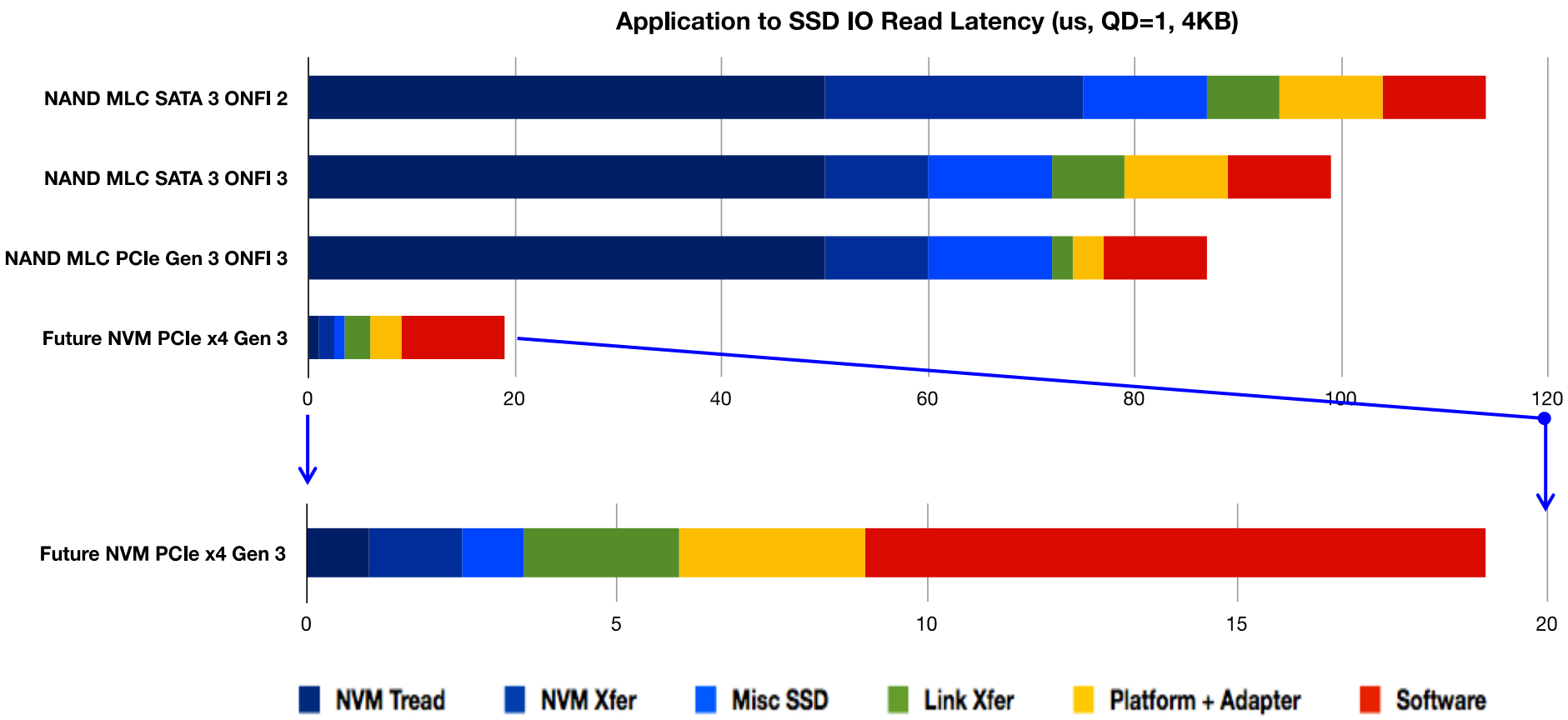
Normal Technology Advancements Drive Higher Performance

Opportunities with Next Generation NVM



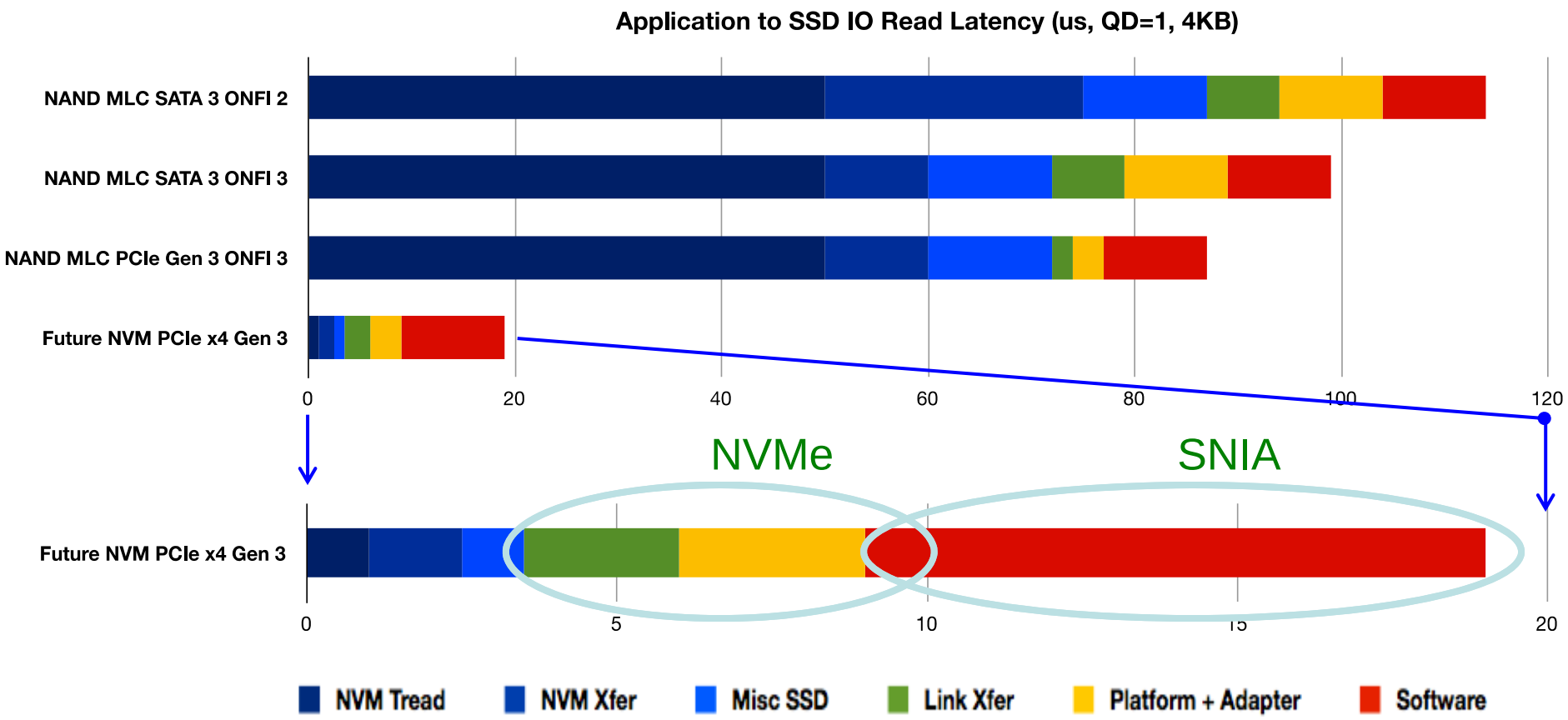
Next Generation NVM -- no longer the bottleneck

Opportunities with Next Generation NVM



Interconnect & Software Stack Dominate the Access Time

Opportunities with Next Generation NVM



NVM Express: Optimized platform storage interconnect & driver
SNIA NVM Programming TWG: Optimized system & application software

Agenda

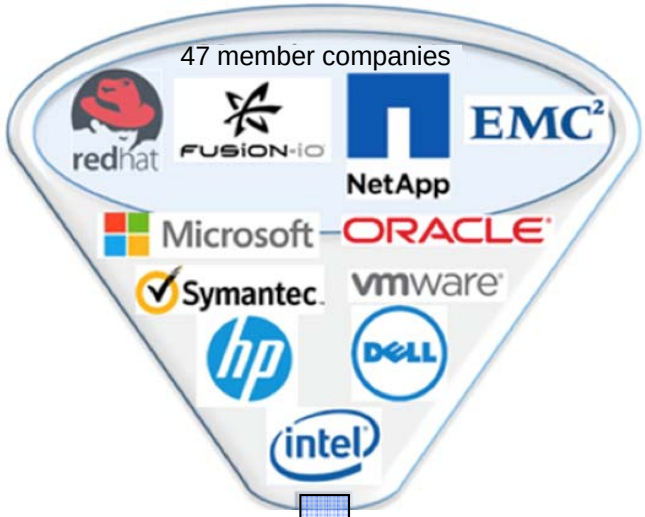
Transition to PCI Express Storage

NVDIMMs

Software Interface Standards

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

SNIA: NVM Programming TWG



SNIA

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 Dec '13 - (Publically available)

SNIA NVM Programing model

The Four Modes

Block Mode Innovation

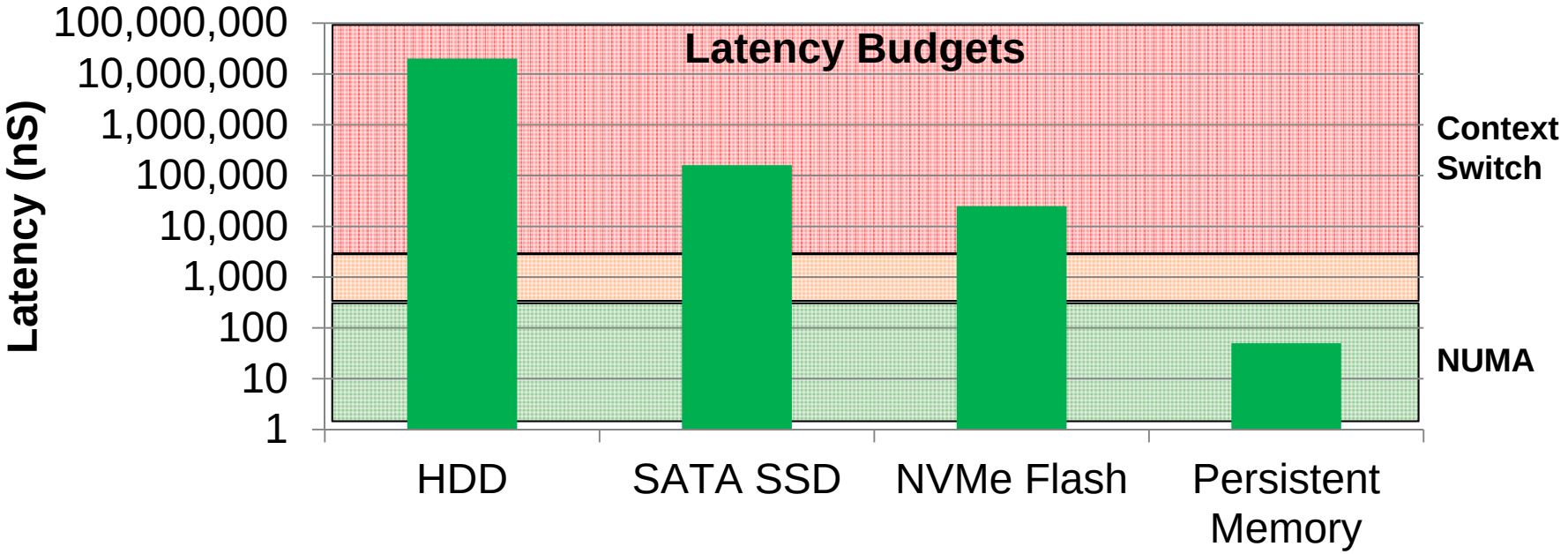
- Atomics
- Access hints
- NVM-oriented operations

Emerging NVM Technologies

- Performance
- Performance
- Perf… okay, cost

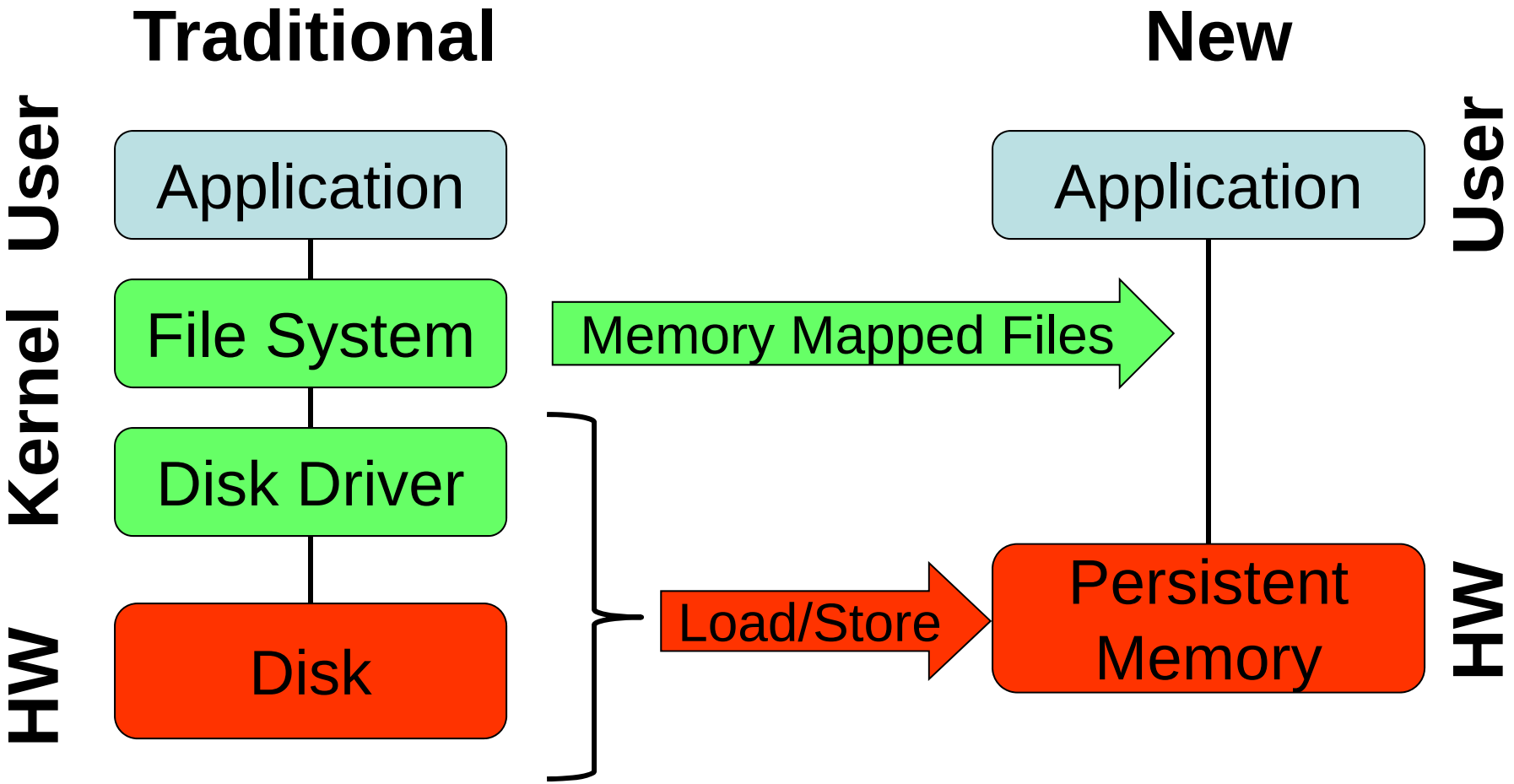
	Traditional	Persistent Memory
User View	NVM.FILE	NVM.PM.FILE
Kernel Protected	NVM.BLOCK	NVM.PM.VOLUME
Media Type	Disk Drive	Persistent Memory
NVDIMM	Disk-Like	Memory-Like

Application View of I/O Elimination

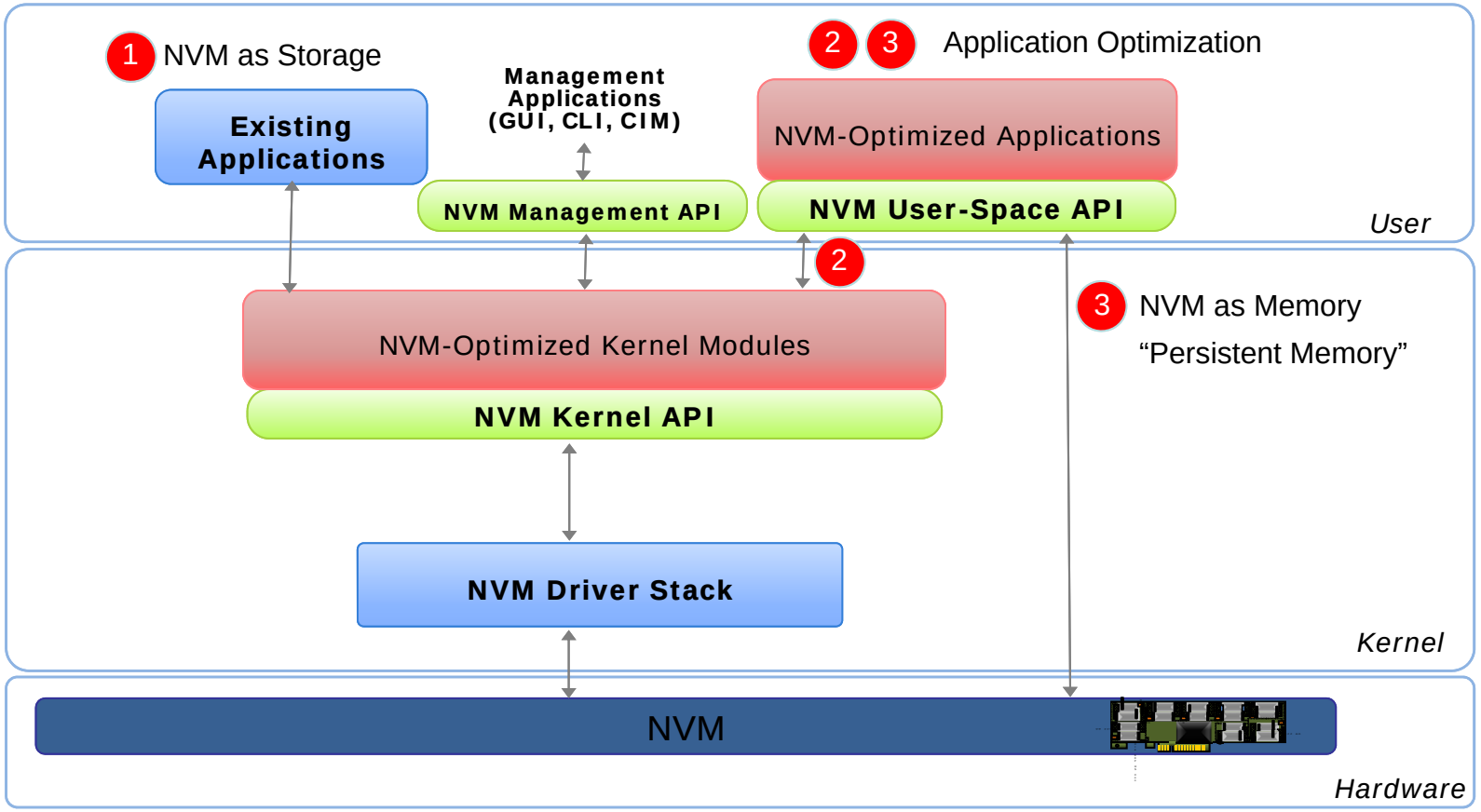


Typical NUMA range: 0 - 200 nS
Typical context switch range: above 2-3 uS

Eliminate File System Latency with Memory Mapped Files



Programming Model Context



SNIA NVM Programming TWG

Linux* Open Source Project, Microsoft *, other OSVs

Existing/Unchanged Infrastructure

Agenda

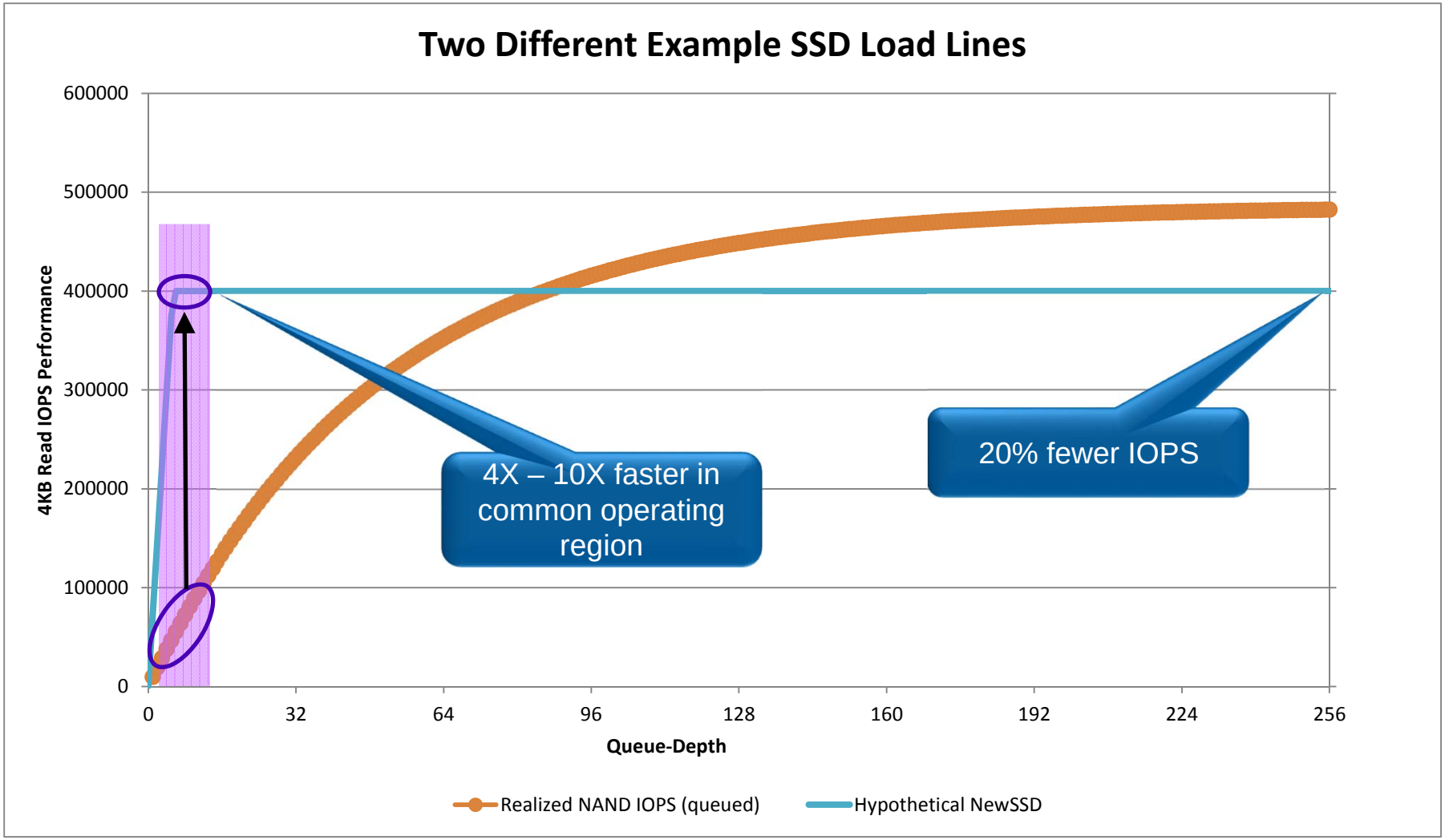
Transition to PCI Express Storage

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?



Conclusion: A New Metric for Measuring SSDs is Needed

Thank You!

Questions & Answers

Jim Pappas

jim@intel.com

Legal Disclaimer

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

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, 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.