The Transition to PCI Express* for Client SSDs

Amber Huffman

Senior Principal Engineer

Intel

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

Legal Notices and Disclaimers

- 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. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards 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.
- Any code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.
- Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance
- Intel, Intel Inside, the Intel logo, Centrino, Intel Core, Intel Atom, Pentium and UltraBook are trademarks of Intel Corporation in the United States and other countries.
- **Material in this presentation is intended as product positioning and *not* approved end user messaging.**
- **This document contains information on products in the design phase of development.**
- *Other names and brands may be claimed as the property of others.
-

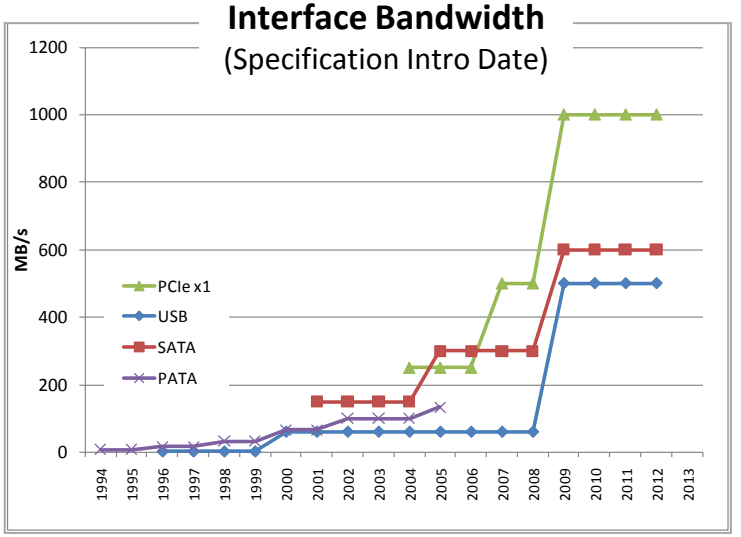
Agenda

- Why PCI Express* (PCIe*) for Client Storage ?
- SATA Express Demystified
- Form Factors & Connectors
  - SATA Express card (i.e., Next Generation Form Factor – NGFF)
  - SFF-8639 for Enterprise
  - 2.5” SATA Express connector
  - Enabling inexpensive PCIe cabling
- Software Interface Options
  - The benefits of NVM Express

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

Why PCIe for Client Storage ?

- SSDs can be built that exceed SATA Gen3 (600 MB/s) today
- Enabling SATA beyond 600 MB/s is a long term development effort
  - Single lane scaling beyond ~ 8Gbps is challenging & requires trade-offs
  - Multi-lane SATA requires a new connector and modified chipset SATA controllers to make multi-lane software transparent
- To enable higher speed client SSDs in near term (‘13 / ’14), PCIe is the only choice
  - PCIe has bandwidth lead (1 GB/s with Gen3)
  - PCIe has multi-lane for scalability (x2, x4, …)
  - Software compatible PCIe SSDs can be built as a single port AHCI device



PCIe can deliver the performance client SSDs need today.

SATA Express Demystified

What is SATA Express?
SATA Express is pure PCIe. There is no SATA link or transport layer, so there's no translation overhead – users will see the full performance of PCIe. Perhaps a good way to think about SATA Express is as the standardization of PCIe as an interface for a client storage device in an HDD-type form factor.

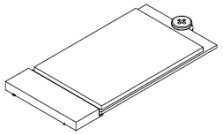
SATA-IO Whitepaper: http://sata-io.org/documents/sata_express_article_2012.pdf

- SATA Express **IS** a marketing name
- SATA Express **DOES** define form factors / connectors that support either SATA or PCIe based SSDs/HDDs/hybrids
- SATA Express **DOES NOT** define the software interface
  - AHCI or NVM Express software interfaces may be used

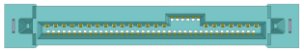
Form Factor & Connector Landscape

Ultrabook™ Mobile All-in-one Desktop WS Server

CEM add-in card



SATA Express card
(i.e., NGFF)



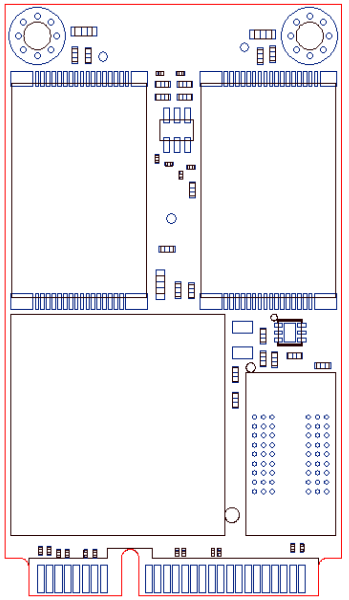
SFF-8639

2.5” SATA Express Connector

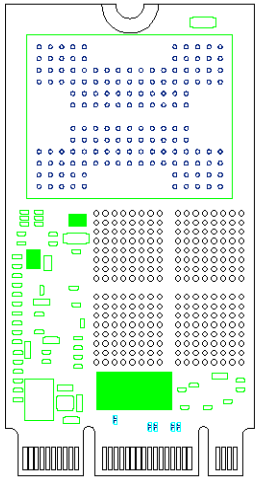
- CEM add-in card supports high speed SSDs with up to 4 lanes of PCIe
- SATA Express card (i.e., NGFF) is designed for the unique needs of Ultrabook™
- SFF-8639 designed for Enterprise use – supports 2.5” PCIe, SAS, SATA
- 2.5” SATA Express connector designed for client SSDs & HDDs/hybrids

Optimizing for Ultrabook™

- mSATA is the small FF for SATA SSDs today
- mSATA has significant shortcomings looking ahead to the PCIe transition, specifically:
  - Too thick: z-height ~ 5mm
  - Challenging capacity: difficult to efficiently add NAND packages
  - Limited performance: only one lane
- Ultrabook™ needs an optimized form factor that addresses these issues
  - Path to < 2.5mm z-height
  - Efficient capacity scaling to enable small 32GB caches to 512GB SSDs
  - Scalable speed for future products (up to 4GB/s)



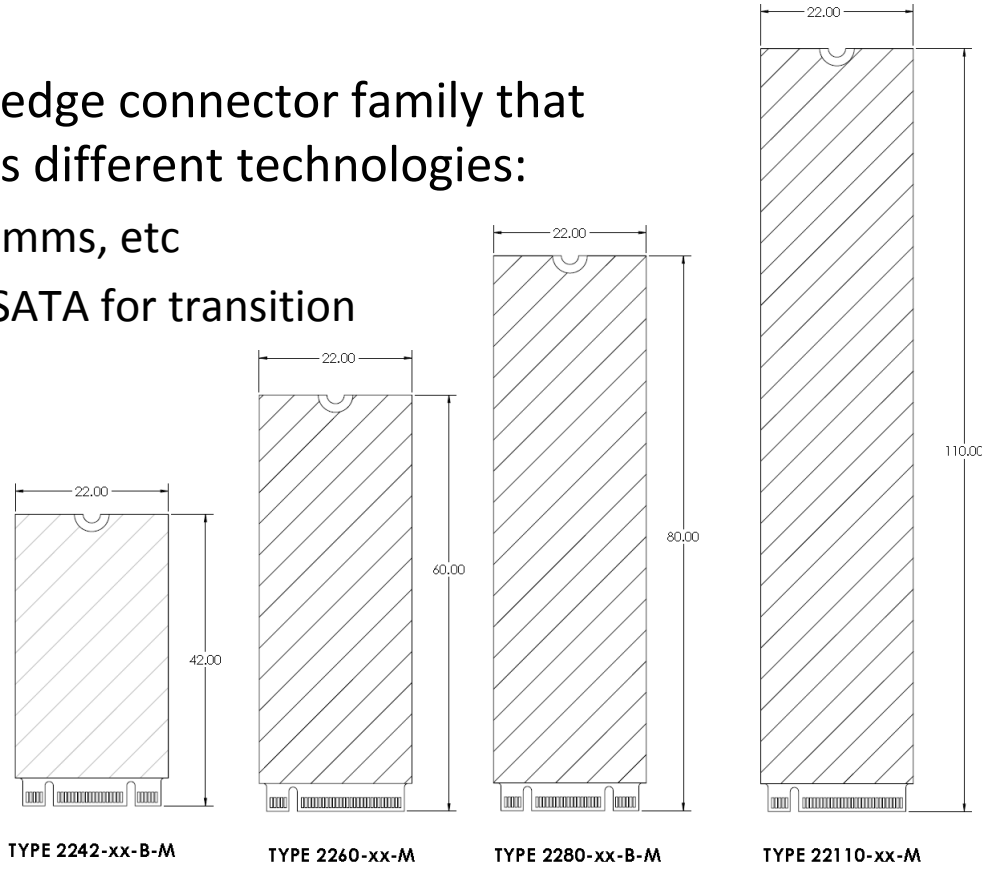
mSATA



Optimized cache FF

Introducing SATA Express card (also known as NGFF)

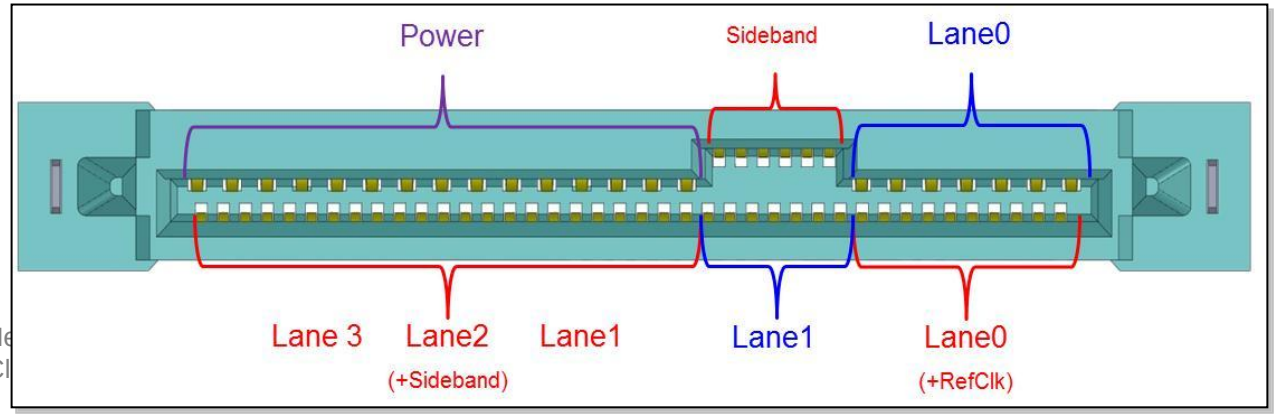
- The challenges for mSATA also apply to mobile add-in cards in general (i.e., PCI Express* Mini Card)
- SATA Express card is a common card edge connector family that supports multiple module sizes across different technologies:
  - SSDs/caches, Wi-Fi, WWAN, multi-comms, etc
  - SSD: Lane 0 muxed between PCIe & SATA for transition
- Three families of modules:
  - Socket 1: Wi-Fi only
  - Socket 2: SSD, cache, WWAN, other
  - Socket 3: storage only (SSD, cache)
- Which socket to use?
  - Socket 2: Flexible usage, 2 lanes only
  - Socket 3: 4 lanes for future scaling



Type 2242 SSD & Cache	Type 2260 SSD & Cache	Type 2280 SSD	Type 22110 SSD

SFF-8639 Connector for Enterprise

- SFF-8639 is an Enterprise backplane connector for 2.5” storage (SSD or HDD) covering PCIe, SATA, and SAS
  - 2.5” is critical Enterprise form factor due to hot swap backplanes
  - Dell’s 12th generation servers launched in March included SFF-8639
- SFF-8639 includes 6 lanes, only 4 lanes are used at one time (not muxed)
  - 4 lanes (red below) are PCIe, envisioned to connect to the CPU PCIe lanes
  - 2 lanes (blue below) are envisioned to connect to an HBA/RAID controller or chipset for SAS & SATA support
- Desire: Enable client PCIe SSDs to be used in Enterprise backplanes

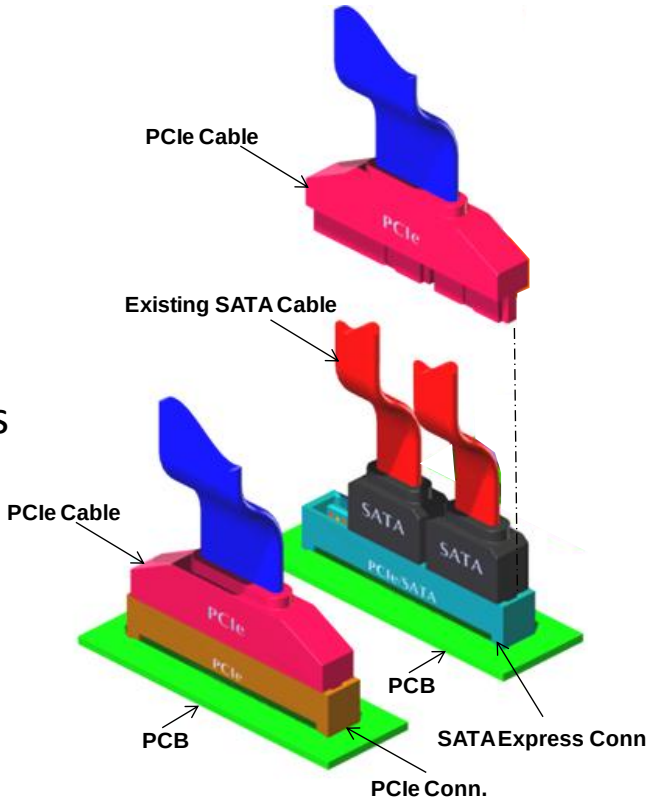


SFF-8639

Blue = SAS/SATA
Red = Enterprise PCIe

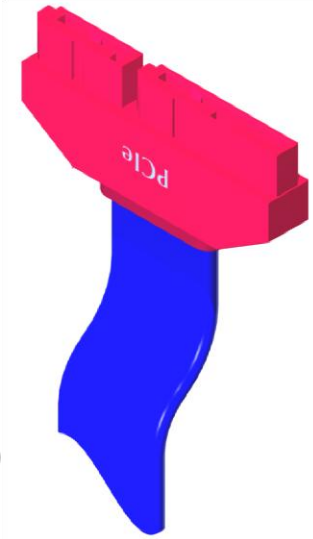
2.5” SATA Express Connector

- SATA Express is a 2.5” connector enabling transition from SATA to PCIe
- SATA Express includes 2 lanes muxed between SATA & PCIe on the host
  - The host chipset can dynamically or statically select SATA or PCIe
  - If SATA selected, enables two cabled SATA devices to be attached
  - If PCIe selected, one x2 PCIe device may be attached
- The SATA Express device connector is mechanically compatible to SFF-8639, enabling a client PCIe device to be used in an Enterprise backplane



Enabling Inexpensive Cabling

- PCI Express requires a reference clock when using spread spectrum
- A reference clock in a cable causes EMI issues, requiring a more expensive connector/cable solution
  - SATA and USB do not include a clock, SATA cables ~ $0.30 in volume
  - An equivalent PCIe cable with reference clock would add > $1 in BOM
- A PCI SIG ECN is under discussion to address issue, changes:
  - Requires use of large elasticity buffer
  - More frequent insertions of SKIP ordered sets (similar to SATA ALIGN)
  - Requires receiver changes (clock data recovery)
- This feature enables inexpensive PCIe cabling for 2.5” SATA Express, as well as other lower cost external cabled PCIe opportunities



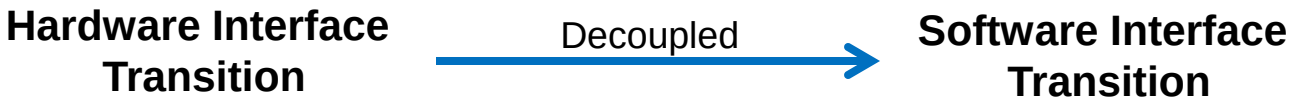
Form Factor & Connector Wrap-up

- SATA Express card (i.e., NGFF) and 2.5” SATA Express connector support SATA & PCIe muxing to ease transition from SATA to PCIe in ‘13 – ’15
- SATA Express card standardization ongoing in SATA-IO and PCI SIG, products in ‘13
  - SATA-IO focused on storage usages, PCI SIG focused on wireless usages
  - SATA-IO and PCI SIG collaborating to realize flexible usage model for OEMs
  - Expect SSD/cache SATA-based products in ’13, transitioning to PCIe in ‘14
- 2.5” SATA Express connector completing definition in SATA-IO in Q3
  - ECN for independent clock + spread spectrum under development in PCI SIG to enable inexpensive PCIe SSD cabling solution

Get involved in SATA-IO and PCI SIG to drive next generation form factors & connectors for the PCIe storage transition.

Software Interface Options

- IDE was the legacy Parallel ATA programming interface
- AHCI was introduced as the Serial ATA programming interface in 2004
  - Designed for hard drives
  - Key features: Native Command Queuing support, power management features (Slumber, Partial, etc)
- NVM Express is the PCIe SSD programming interface, architected from the ground up for performance
  - Designed for SSDs, with scalability for future NVM technologies
  - Key features: Optimized interrupt architecture for scalable IOPs, large scale parallelism supported, deep queues, etc
- Client PCIe storage transition will be similar to SATA transition:



NVM Express Overview

- NVM Express is a scalable host controller interface designed for Enterprise and client systems that use PCI Express* SSDs
- NVMe was developed by industry consortium of 80+ members and is directed by a 13 company Promoter Group

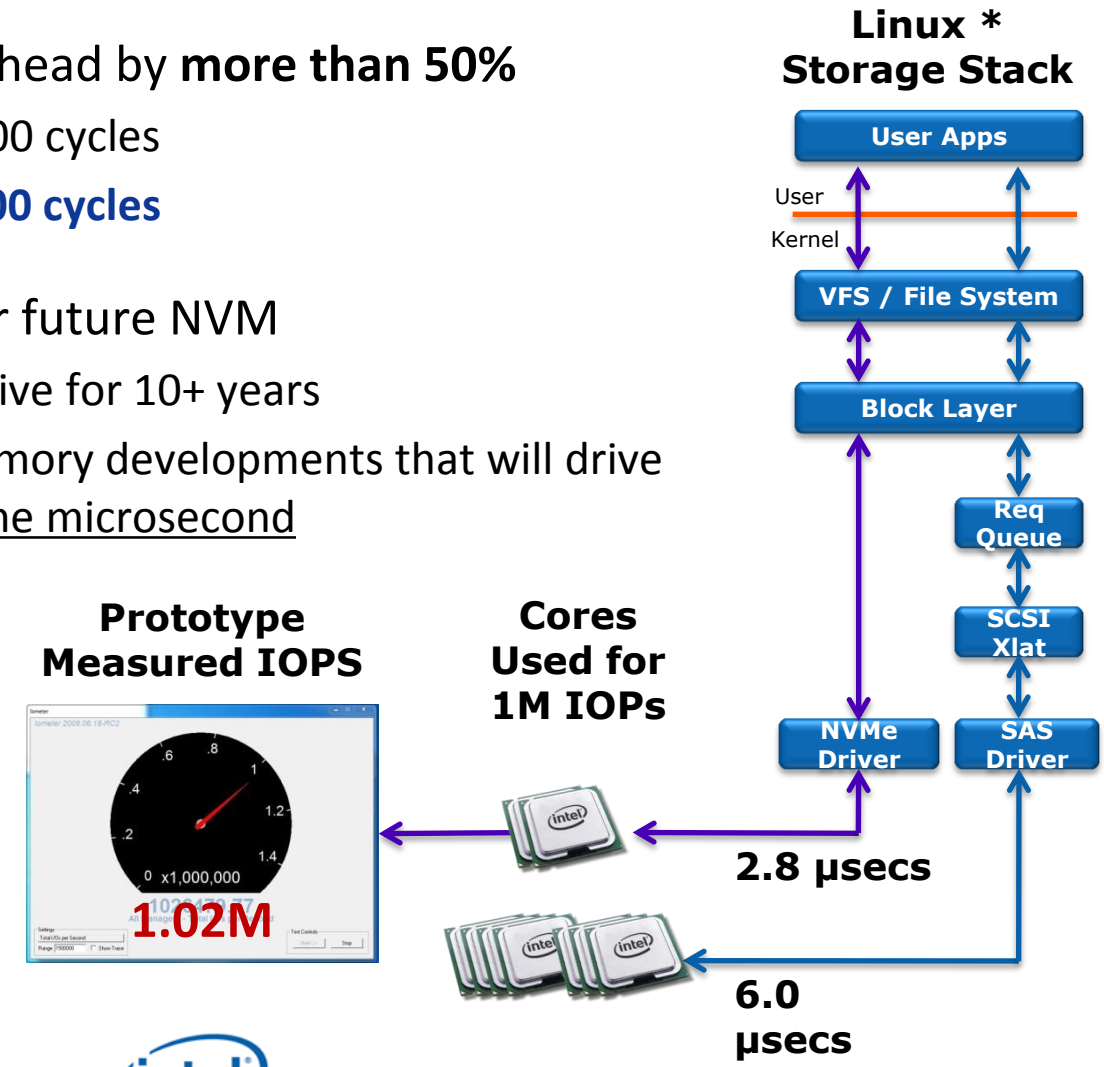
- NVMe 1.0 was published March 1, 2011
- Open source reference drivers for Linux and Windows are available
- Product introductions later this year, first in Enterprise

Architectural Differences Between AHCI and NVM Express

	AHCI	NVM EXPRESS
Uncacheable Register Reads Each consumes 2000 CPU cycles	4 per command 8000 cycles, ~ 2.5 µs	**0** per command
MSI-X and Interrupt Steering Ensures one core not IOPs bottleneck	No	Yes
Parallelism & Multiple Threads Ensures one core not IOPs bottleneck	Requires synchronization lock to issue command	No locking, doorbell register per Queue
Maximum Queue Depth Ensures one core not IOPs bottleneck	1 Queue 32 Commands per Q	64K Queues 64K Commands per Q
Efficiency for 4KB Commands 4KB critical in Client and Enterprise	Command parameters require two serialized host DRAM fetches	Command parameters in one 64B fetch

NVMe Delivers Cutting Edge Performance

- 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 defined to scale for future NVM
  - Host controller standards live for 10+ years
  - NVMe supports future memory developments that will drive latency overhead below one microsecond



Chatham NVMe Prototype

Prototype Measured IOPS

Cores Used for 1M IOPs

*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.

Summary

- PCIe can deliver the performance client SSDs need today
- There is a plethora of form factor / connector options to satisfy unique needs of each unique platform
  - SATA Express card (i.e., NGFF), CEM add-in card, 2.5” SATA Express, SFF-8639
- NVM Express is the best long term software interface for PCIe SSDs
- Get involved in the standards organizations driving the transition
  - SATA-IO: www.sata-io.org
  - PCI SIG: www.pcisig.com
  - NVM Express: www.nvmexpress.org