IDF15
INTEL DEVELOPER FORUM

NVM Express™: The Data Center and Client Storage Transformation

Amber Huffman – Senior Principal Engineer, Intel Corporation
Mike Shapiro – VP Software, EMC/DSSD

SSDS001

intel
experience
what's inside™

Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

NVM Express™ - Architected for NVM

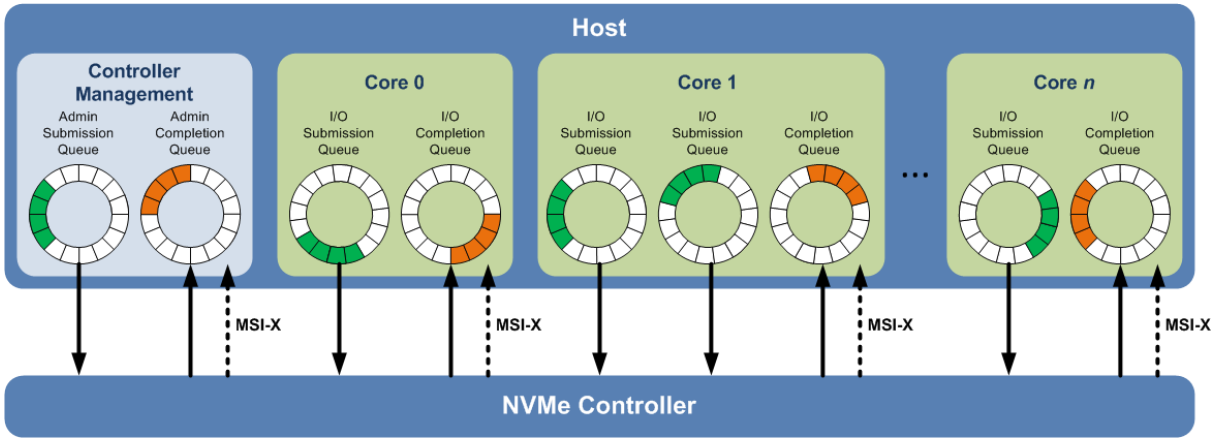
- NVM Express (NVMe) is the standard for PCI Express® (PCIe) SSDs
  - Standardizes register set, feature set, and command set where there were only proprietary PCIe solutions before
  - Architected from the ground up for NAND and next generation NVM
  - Designed to scale from Enterprise to Client systems

- NVMe is 80+ companies strong, with 13 company Board of Directors

Learn more at nvmexpress.org

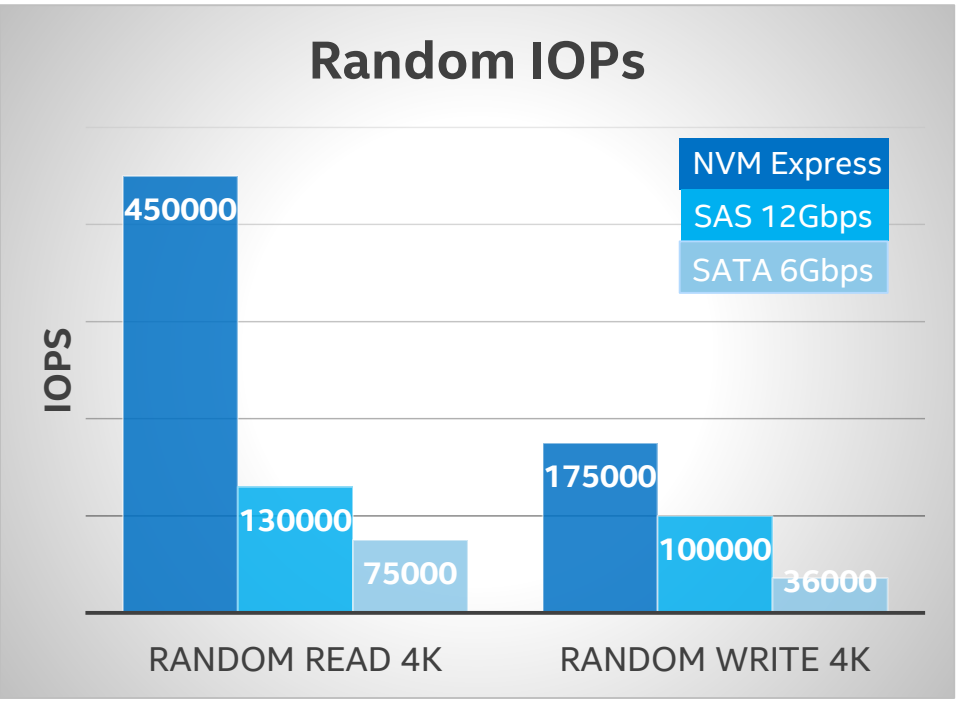
The Technical Basics

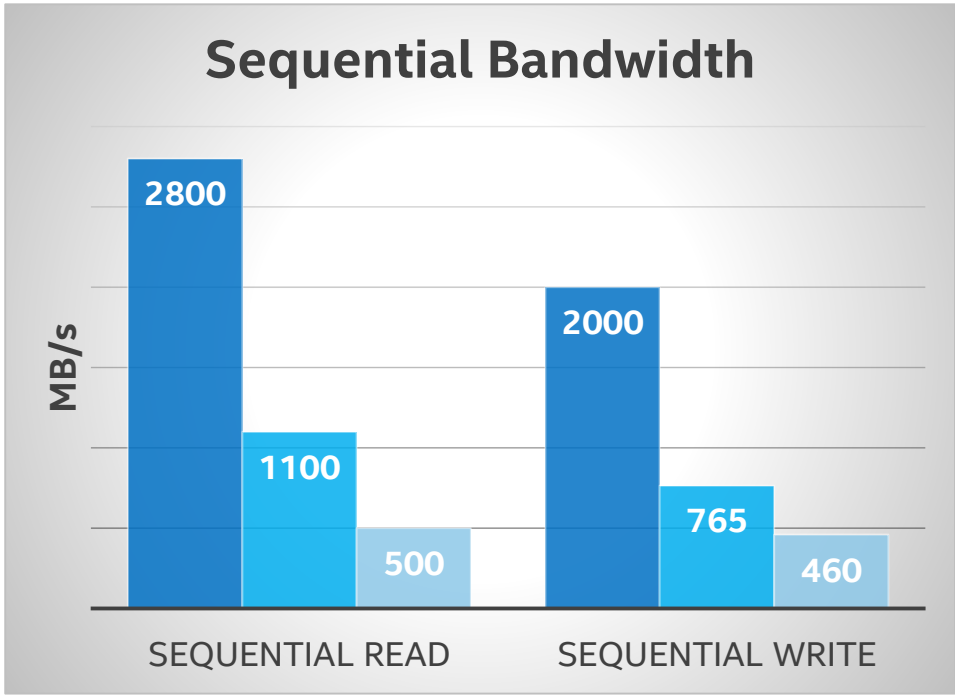
- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports MSI-X and interrupt steering
- Streamlined and simple command set (13 required commands)
- Optional features to address target segment
  - Data Center: Reservations, etc. Client: Power features, etc.
- Designed to scale for next generation NVM, agnostic to NVM type used



NVM Express™ Performance Leadership

Efficiency of NVM Express™ unlocks IOPs wall experienced in SATA* & SAS



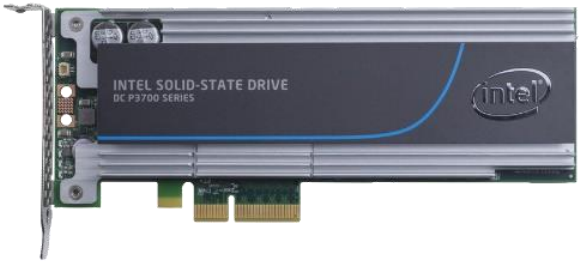


Results measured by Intel based on the following configurations. Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Configurations: Performance claims obtained from data sheet, sequential read/write at 128k block size for NVM Express and SATA*, 64k for SAS. Intel SSD C P3700 Series 2TB, SAS Ultrastar* SSD1600MM, Intel® SSD DC S3700 Series SATA 6Gbps. Intel® Core™ i7-3770K processor @ 3.50GHz, 8GB of system memory, Windows Server* 2012, IOMeter*. Random performance is collected with 4 workers each with 32 QD.

The Raw Capability of the Interface

If there is a single PCI Express® Gen3 x4 slot, what solution gives best IOPs and latency?

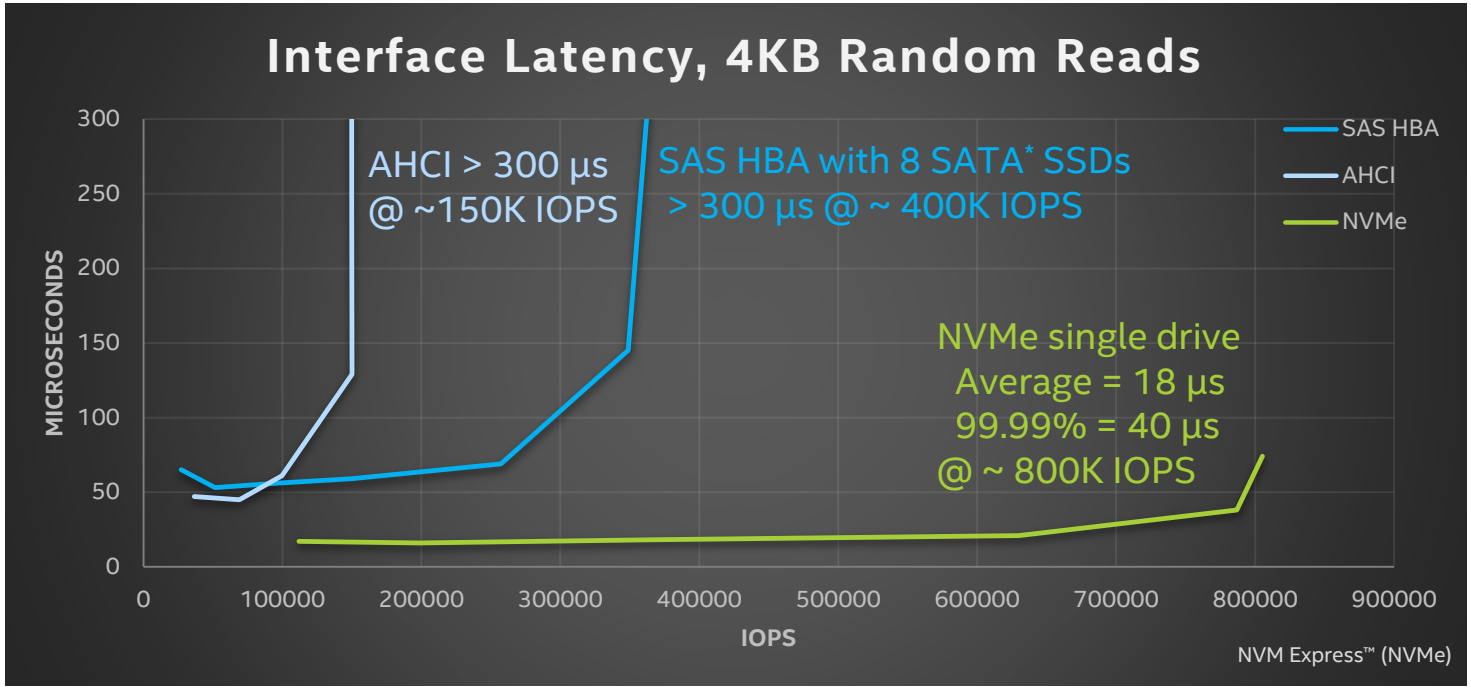
SAS HBA with 8 SATA* SSDs



1 NVM Express™ SSD

Analysis tested raw interface speed – no NVM access – with 4KB random reads

NVMe™ Delivers Higher IOPs and Better QoS



NVMe™ delivers 18 µs average and 40 µs 99.99% interface latency. Other interfaces have outliers in 100s of µs as interface reaches saturation.

Results measured by Intel based on the following configurations. Intel Server Board S2600WTT with 28 E5-2695 CPUs, 2 sockets, 2.3 GHz clock speed per CPU, Ubuntu* 14.04.1 LTS (GNU/Linux* 3.16.0-rc7tickles x86_64), idle=poll kernel settings, SAS HBA is LSI SAS9207-4i4e with controller LSI SAS 2308. SATA SSDs are Intel® SSD DC 3500 at 800 GB. NVMe SSD is Intel SSD P3700 at 1.6 TB. Workload details are Workload: 4K Random Reads using FIO – 4 + threads. Drives tested empty to test interface only (no NVM access.)

Proof Point – Database Analytics

A Lenovo* ThinkServer* RD650 with four NVM Express™ (NVMe) SSDs transformed the performance of the SQL database workload.

	NVMe SSDs	SATA SSDs	Benefit
Total Database Performance Queries/Hr Across All Instances	25,062.1	9,524.8	**2.63 X**
Single Database Performance Queries/Hr Across One Instance	6,265.5	4,762.4	**1.31 X**
Time to Answer Average Query Time	27.9 min	38.0 min	**10 MINUTES**

NVMe enables one server to replace four legacy servers

Detailed whitepaper at http://www.principledtechnologies.com/Lenovo/RD650_storage_performance_0415.pdf.

Intel is a sponsor and member of the BenchmarkXPRT Development Community, and was the major developer of the XPRT family of benchmarks. Principled Technologies is the publisher of the XPRT family of benchmarks. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases.

Proof Point – Developer Productivity

VirtualBox (virtualbox.org) is an open source virtualization product in common use in the app developer community.

	NVMe™ SSD	SATA* SSD	Time Saved	Productivity
Linux* Virtual Machine Clone Create VM within VirtualBox	144 sec	315 sec	171 sec	**2.2 X**
Dual Android* Compile Compile within VirtualBox VM	63.25 min	70.75 min	7.50 min	**1.13 X**

Upgrading to NVMe enhances developer productivity on common tasks

Configuration: Lenovo* P900 Workstation. Two Intel® Xeon® processor 2687W v3 with power saving and Intel® Hyper-Threading disabled, Memory 64GB Micron* DDR, Graphics NVidia* GTX 980, Windows® 8.1 Professional, Storage: Intel® SSD 730 Series 480GB or Intel SSD 750 Series 800GB (configured as 480GB). Linux Virtual Machine Clone uses Linux Ubuntu* VM within Virtualbox. Android Compile uses Android 5.0.1.r1 (Lollipop off of the AOSP branch) in VirtualBox VM.

NVM Express™ Driver Ecosystem is Robust

- Native support across Windows®, Linux*, VMware*, Solaris*, FreeBSD*, and UEFI

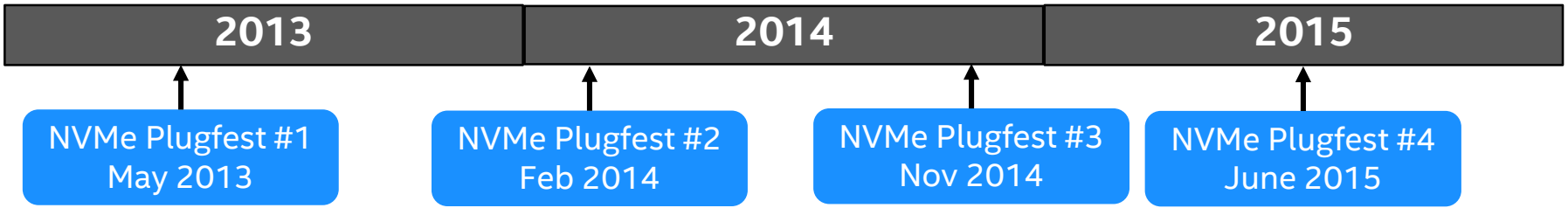
- Developments over past year include:
  - Microsoft added support to Windows 7 and Windows Server 2008 R2 (http://bit.ly/1edi7xl)
  - VMware released a driver for ESXi 5.5 and followed with inbox NVM Express™ (NVMe) driver for ESXi 6.0 (http://vmw.re/1TY37V4)
  - Linux converted to block-multiqueue that enabled use of kernel device mappers, merging contiguous I/Os, and finer grained stats
  - Chrome OS* added support for fast boot with NVMe by enhancing depthcharge (details in Google's repository at http://bit.ly/1JgI0ay)

Windows 8.1

NVM Express™ (NVMe) Interoperability Program

- The University of New Hampshire Interoperability Lab (UNH-IOL) has collaborated with NVM Express™ (NVMe) to deliver a robust interoperability program

- Four plugfests have been held, with the bar raised each time
  - E.g., The latest event successfully tested hot plug across a range of devices



- Check out the NVMe Integrator's List:

  https://www.iol.unh.edu/registry/nvme

Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

Client and Data Center Feature Guidance

NVM Express™ spans Client through Data Center – select the right features for your SSD implementation based on market segment targeted

Client Guidance

# of Queues	2 to 8
# of Namespaces	1 to 4
# of Power States	3 to 4

Low Power Features (e.g., APST)
Security Tunneling (e.g., Opal)
Host Memory Buffer

Data Center Guidance

# of Queues	16 to 128
# of Namespaces	1 to 16
# of Power States	1 to 4

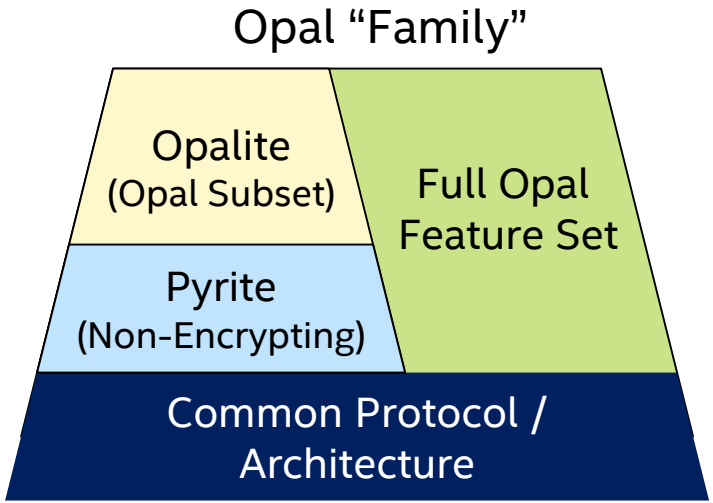
Security Tunneling (e.g., Opal)
End-to-end Data Protection

Storage Customer Features

Scatter/Gather List
Reservations (and dual port)
Controller Memory Buffer

Security for NVM Express™ Solutions

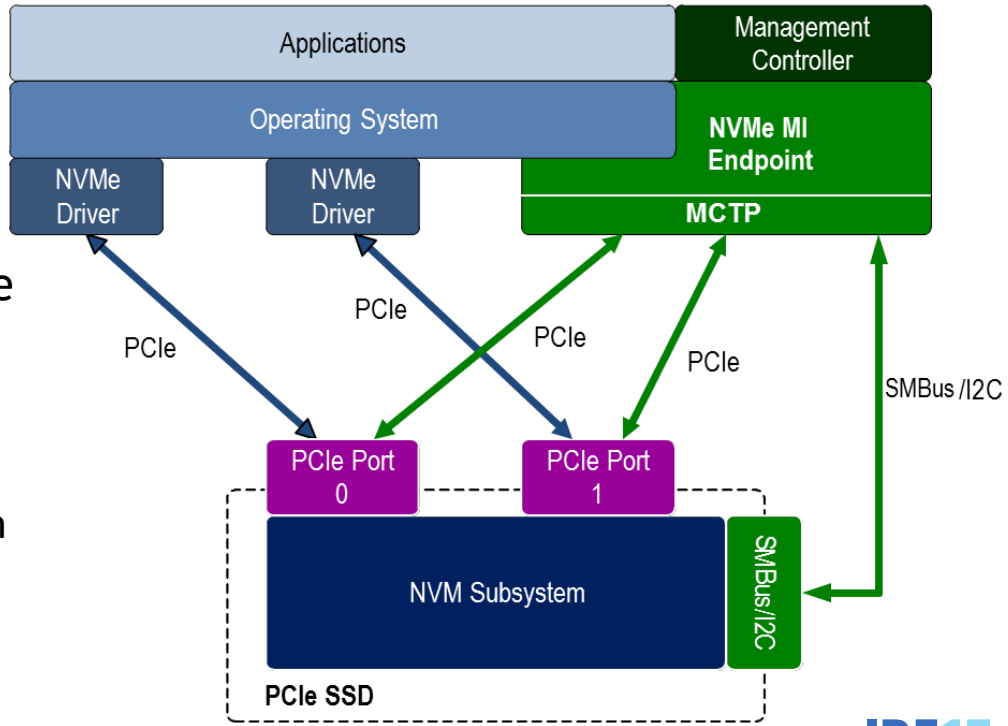
- NVMe™ is leveraging the security expertise of the Trusted Computing Group (TCG)
- TCG has developed a “family” of specifications to scale across the needs of NVMe in different Client and Enterprise solutions
- Data Center solutions should support Opal
- Client solutions may support Opal, Opalite, or Pyrite depending on security requirements



Management Interface for Data Center

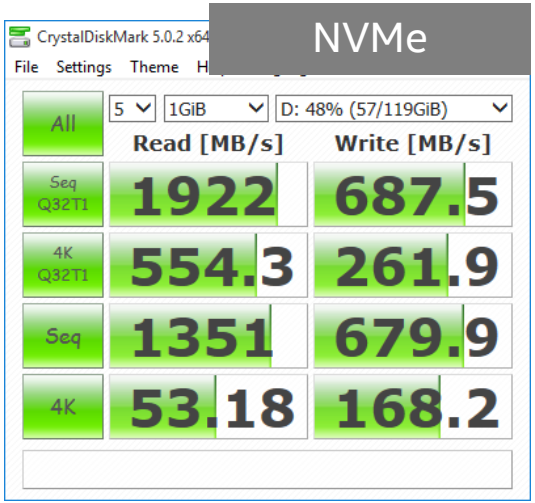
- Basic Management Command published in February
  - Defines simple SMBus command for gathering basic statistics
- Full Management spec now complete
  - Maps management interface to one or more physical interfaces
  - Works across deployment, pre-OS, runtime, aux power, and decommission

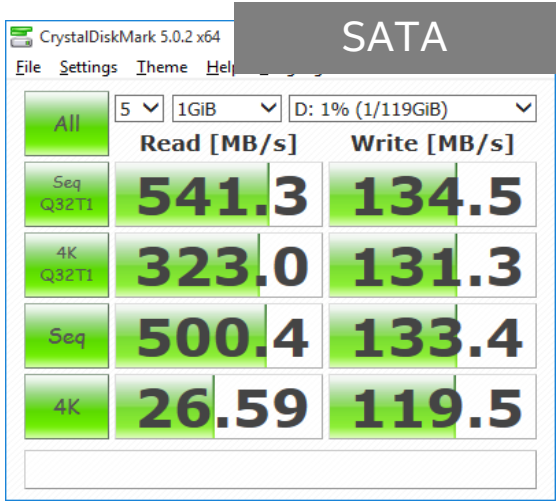
NVMe™ Management Interface

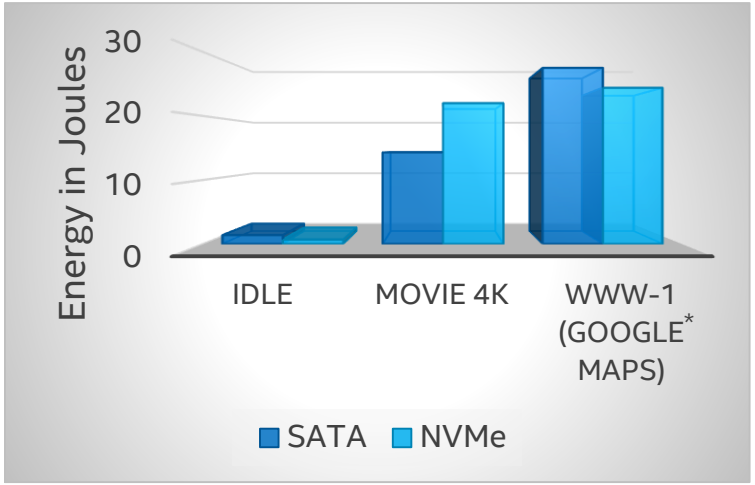


Performance with Low Power

- NVM Express™ delivers ~ 2 to 4X performance of SATA*
- And… with similar energy using the Intel® RST 14.6 Storage driver







Configuration: Next generation Intel® Core® platform running Windows® 10, Intel RST 14.6 Storage driver, Storage: Samsung M.2 SATA* SSD and Samsung* SM951 NVM Express SSD.

Race to Halt Benefits

File copy illustrates race to halt benefits of fast storage

File Copy (59740 files, 27.8 GB)

	SATA	NVMe™	Benefit
Time	281 sec	127 sec	**2.2 X**
Energy	606 J	250 J	**2.4 X**

Getting to idle faster delivers system level power savings

Configuration: Next generation Intel® Core® processor based platform running Windows® 10, Intel RST 14.6 Storage driver, Storage: Samsung M.2 SATA SSD and Samsung SM951 NVM Express™ SSD.

Architecting for Low Power in Client

NVMe has a flexible power management architecture with up to 32 power states

- Thermal throttling: Adjust performance to meet thermal system needs
- Non-operational states: Balance low power with fast resume for performance

Recommended Power States to Adapt to Client Workloads

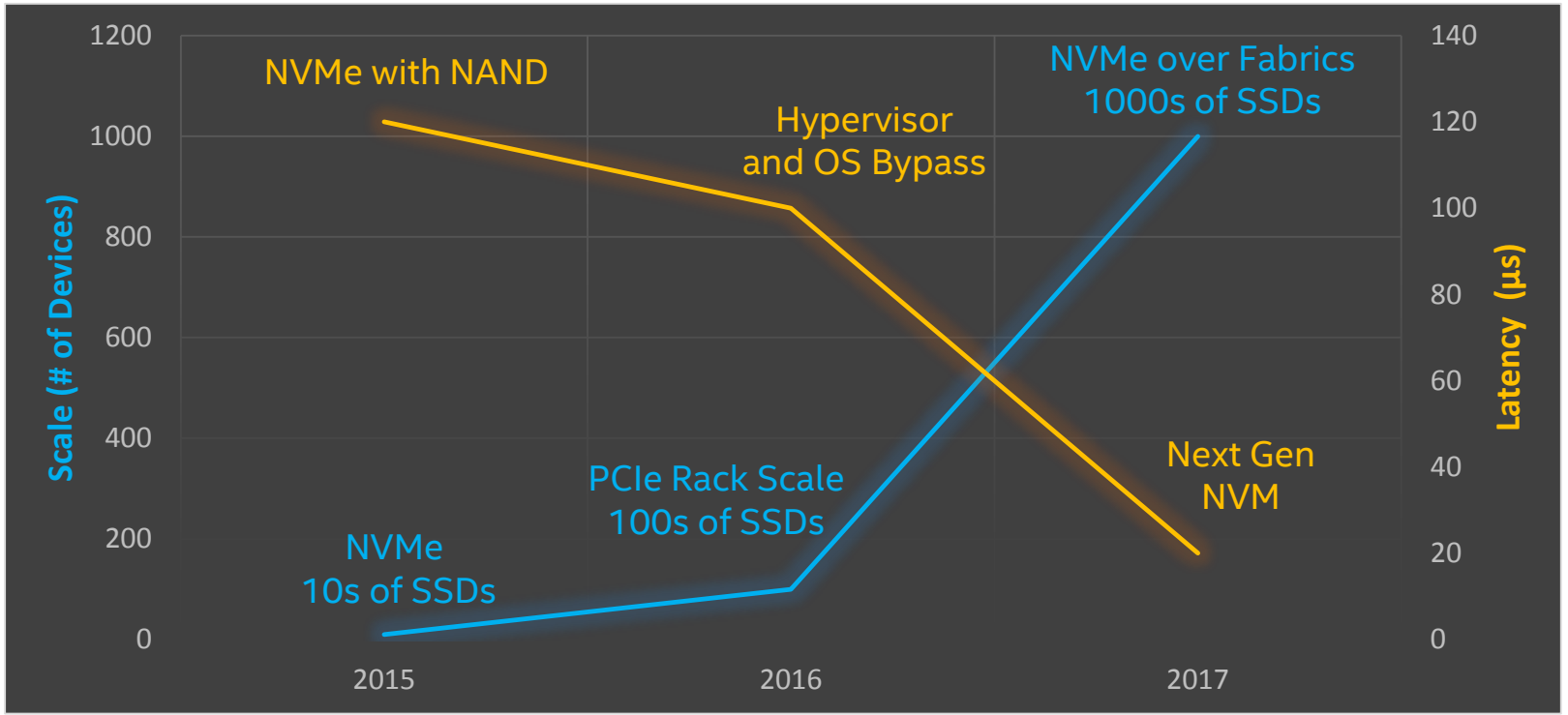
Power State	Description	Performance	Power	Resume Latency
PS0	Full Performance	100% Performance	No restriction	n/a
PS1	**Thermal Throttle**, Light	< 100% Performance	E.g., for M.2 do not exceed 2.4 W	n/a
PS2	**Thermal Throttle**, Heavy	< 100% Performance	E.g., for M.2 do not exceed 1.9 W	n/a
PS3	**Non-operational**, good power with fast recovery	n/a	Idle < 50 - 100 mW	< 1-10 ms
PS4	**Non-operational**, lowest power state	n/a	Idle < 5 mW	< 50-100 ms

Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

The EMC Perspective

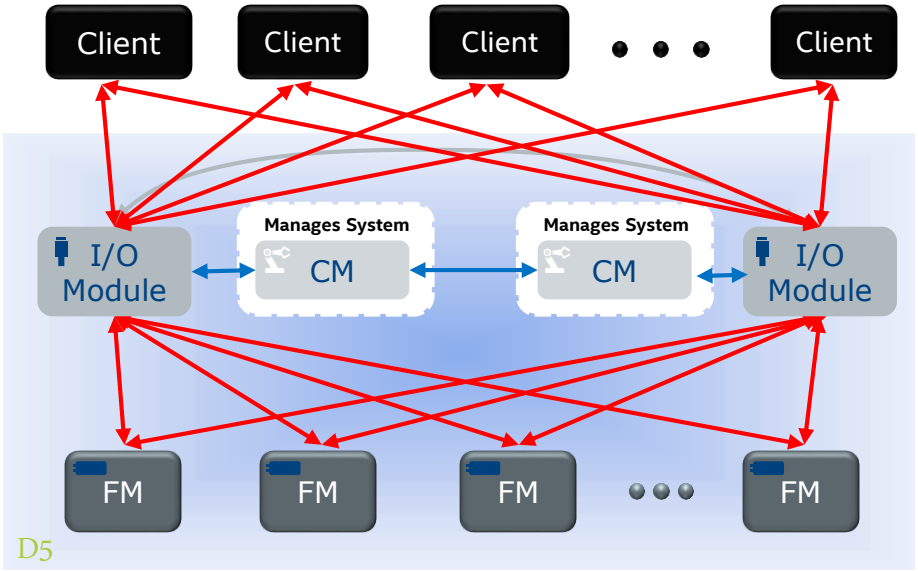
Next generation high-speed big-data apps require a new architecture



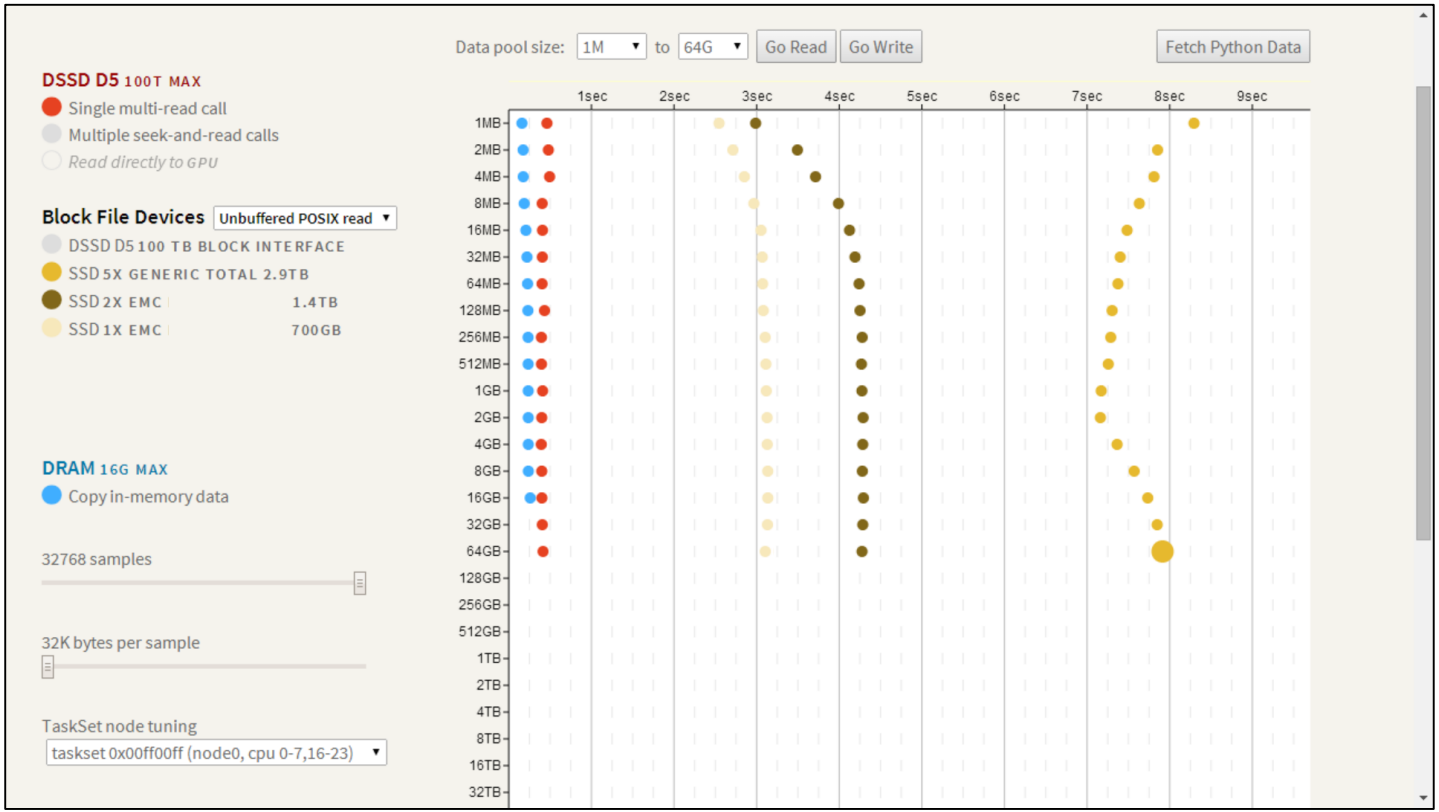
NVM Express™ (NVMe)
PCI Express® (PCIe)

What We've Been Working on …

- EMC / DSSD "D5" System, Previewed at EMC World
- First Shared, Virtualized NVMe™ storage device in the industry
- Dual-port NVMe Flash Modules, Multipath NVMe to Clients



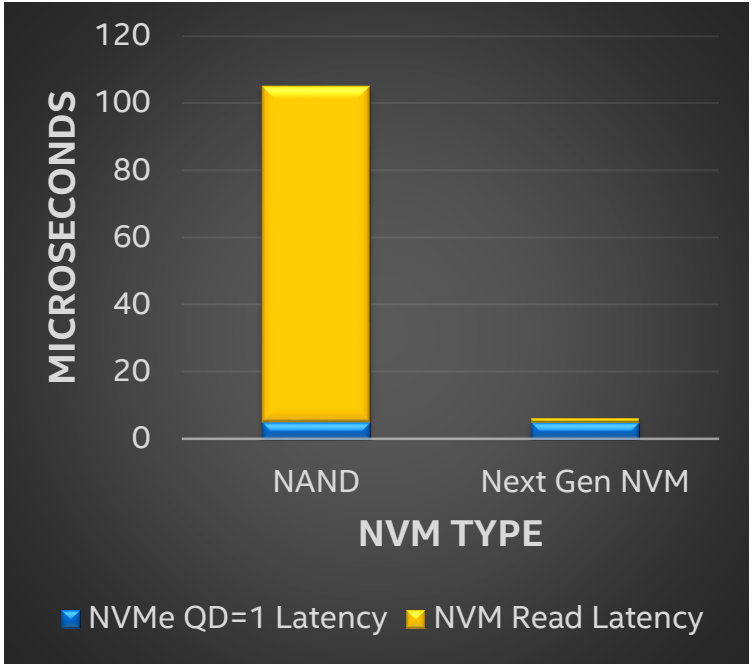
NVM Express™ Performance for Big Data



Visit EMC in the NVM Express™ Community at Booth #887.

The Need to Extend NVM Express™ Over Fabrics

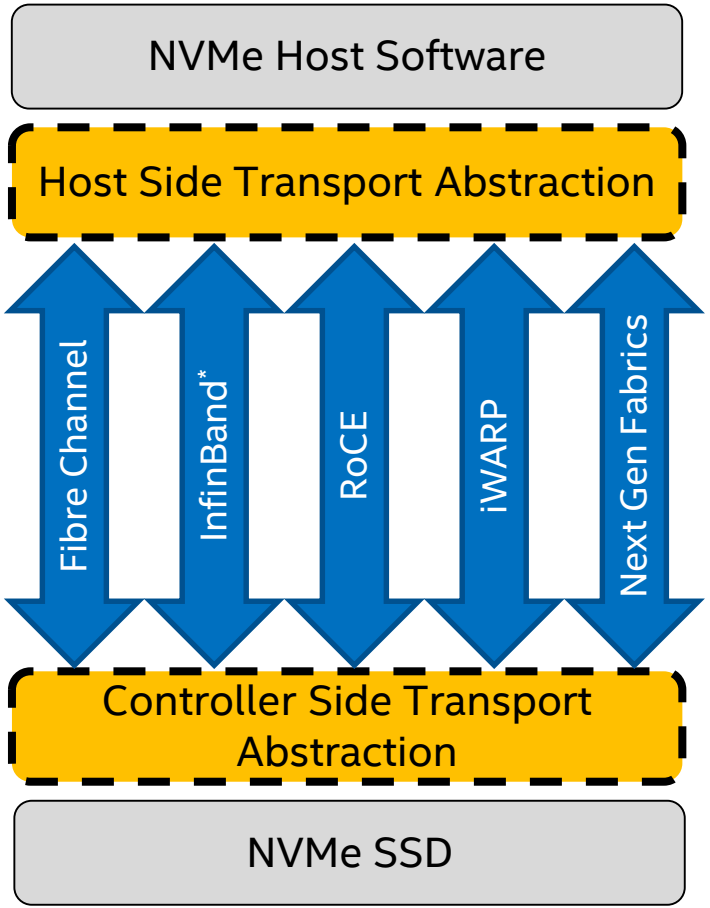
- PCI Express® ideal for in-server and in-rack, but difficult to scale beyond 100's of nodes:
  - Address routing rather than endpoint routing
  - Want to converge storage + networking at scale
  - Want to leverage standard switch infrastructure
- Existing Fabric interface (e.g., iSER / SRP) ecosystem is not well suited for this:
  - Inconsistent adoption across OS/VMs
  - Protocol is overly complex, adding latency
  - Issues even worse when we move to NG-NVM



Delivering < 20 µs across Fabric requires new, simple, efficient protocol

NVM Over Fabrics Overview

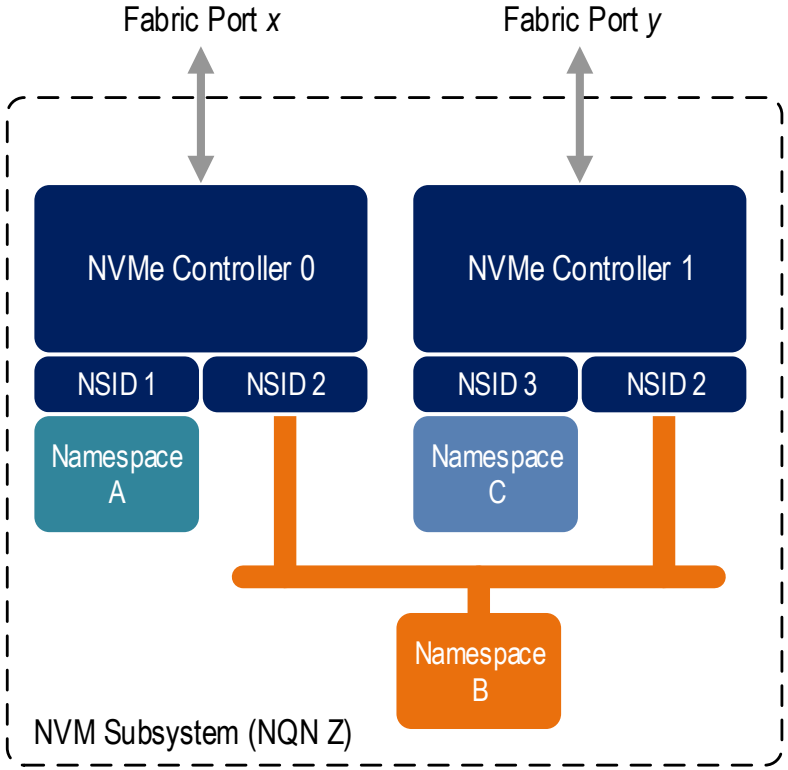
- The back-end of many deployments is PCIe Express® based NVM Express™ (NVMe) SSDs
- With 10-100Gb reliable RDMA fabric and NVMe SSDs, the remaining issue is the software necessary to execute the protocol
- Use NVMe end-to-end to get the simplicity, efficiency, and low latency
  - Simple protocol => Simple host and SSD software
  - No translation to/from another protocol like SCSI



Standard abstraction layer enables NVMe across range of Fabrics

Solid Architecture Foundation to Leverage

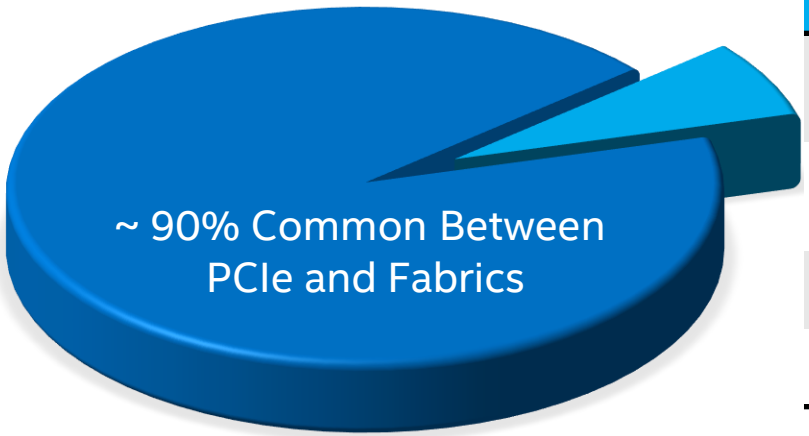
- NVM Express™ (NVMe) revision 1.2 defines solid architecture to leverage
- NVM Subsystem Architecture
  - Multiple NVMe Controllers and fabric ports
  - Multi-path I/O and multi-host support
- Namespace Architecture
  - Multiple (shareable) namespaces
  - Namespace management & reservations
- Multiple I/O Queue host interface
  - Simple command set, optimized for NVM
  - SGL based buffer descriptors



Commonality Between PCI Express® and Fabrics

- The vast majority of NVM Express™ (NVMe) is leveraged as-is for Fabrics
  - NVM Subsystem, Namespaces, Commands, Registers/Properties, Power States, Asynchronous Events, Reservations, etc.

- Primary differences reside in enumeration and queuing mechanism



Differences	PCI Express® (PCIe)	Fabrics
Identifier	Bus/Device/ Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queuing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

Capsules as Messaging Building Block

- Capsules are messages with **common** NVM Express™ content
  - Submission Queue (SQ) Command, Completion Queue (CQ) Status, Data, Metadata, SGLs

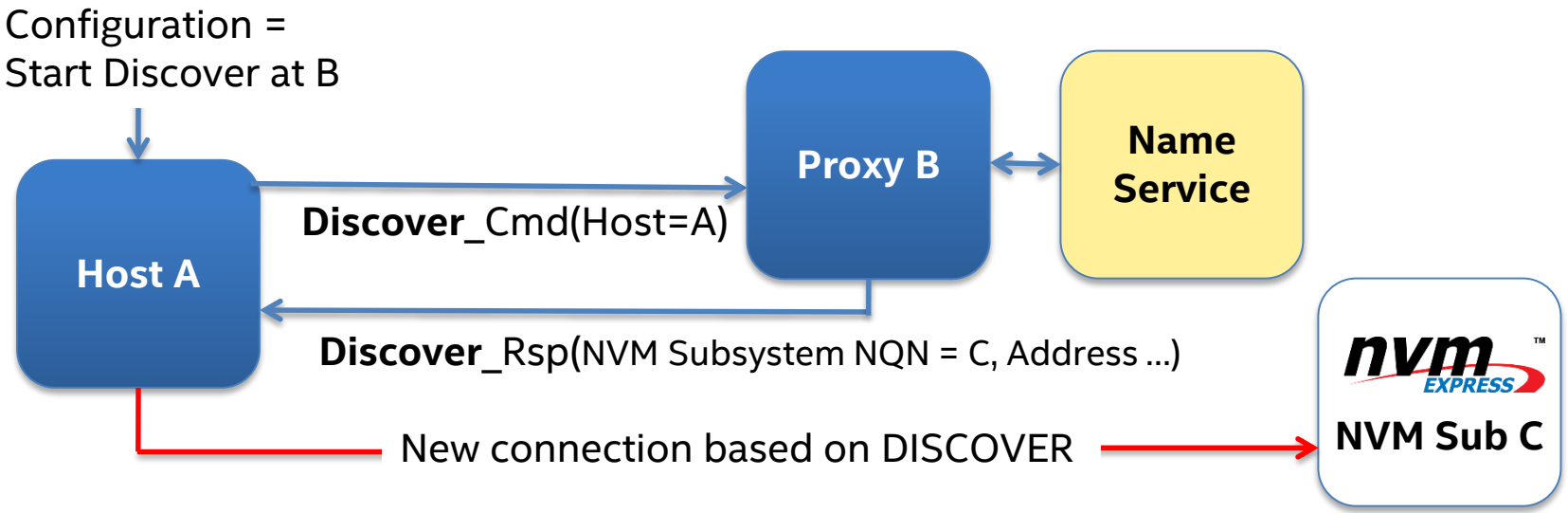
Host to SSD: RDMA_Send

SSD to Host: RDMA_Read (if additional data transfer required)

SSD to Host: RDMA_Send

Finding an NVM Subsystem

- The Discover command is used to find available NVM Subsystems
- A Name Service or another controller returns records one at a time
- The record includes the NVMe Qualified Name (NQN) of the NVM Subsystem, the address, and other transport specific details used to form a connection



NVM Express™ over Fabrics Development Status

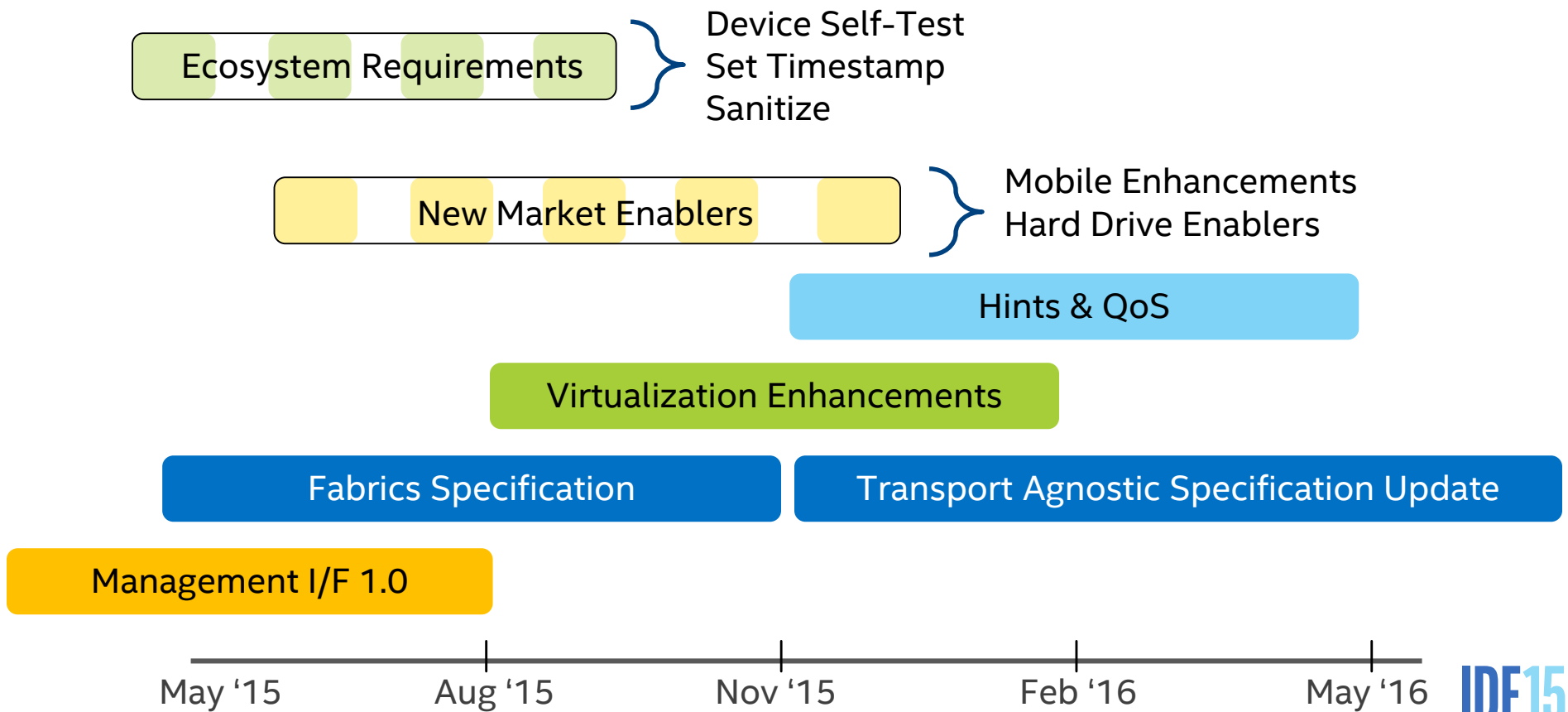
- NVM Express™ (NVMe) over Fabrics definition targeted to start ratification end of year
- Major concepts documented in ~ 0.5 draft level document
- Three independent prototypes have demonstrated the promise, all adding less than 10 µs of latency

Get ready for initial NVMe over Fabrics implementations in 2016

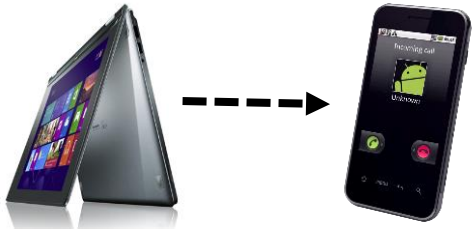
Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

NVM Express™ Technology Roadmap

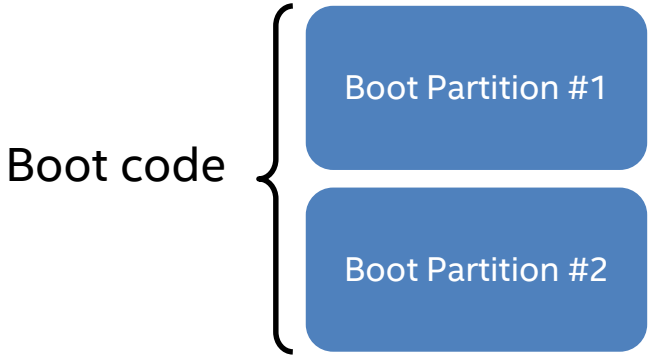


Enabling NVM Express™ in Mobile



- PCI Express® (PCIe) is a low power interface
  - In wireless mobile solutions today
  - PCIe on par with M-PHY as mobile interface solution
- Smaller BGA coming for mobile
- NVM Express™ (NVMe) defined "Boot Partitions" to enable non-BIOS boot
  - Read via MMIO registers
  - Write via enhanced firmware download commands
  - Protect/lock with Replay Block

Item	PCIe Gen3	PCIe Gen2	M-PHY Gear3
Line Speed [Gbps]	8	5	5.83
PHY Overhead	128/130,1[GB/s]	8/10, 500[MB/s]	8/10, 583[MB/s]
Active Power [mW]	60 (L0)	46 (L0)	58 (HS)
Standby Power [mW]	0.11 (L1.2)	0.11 (L1.2)	0.2 (Hibern8)
MB/mJ (higher better)	14-18	8-12	8-12

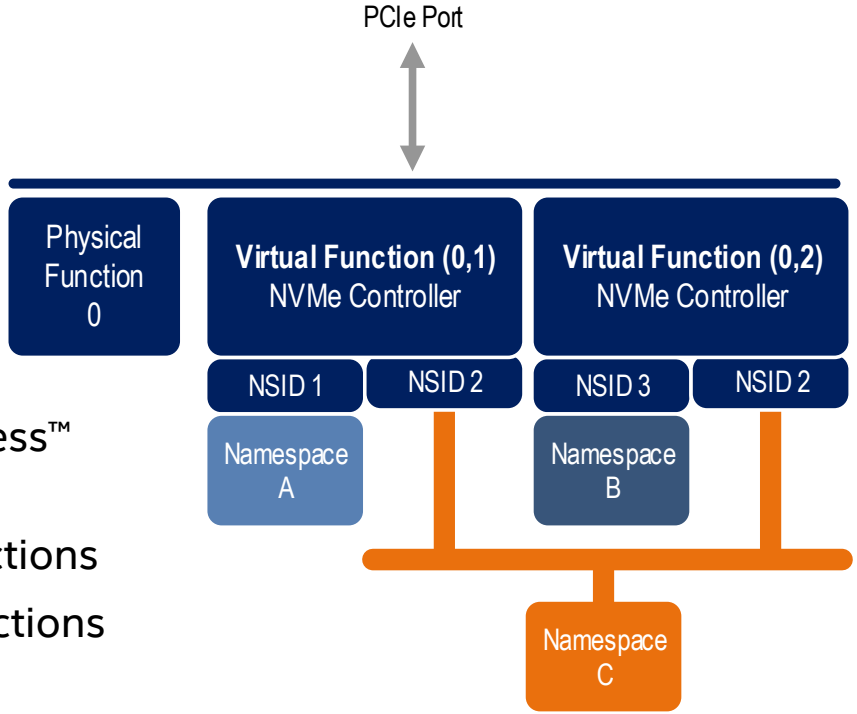


Source: SanDisk*

Virtualization Enhancements

- For lowest latency, SR-IOV is attractive in virtualized environments

- The Workgroup is targeting several enhancements for virtualization:
  - Standardize initialization flows for NVM Express™ (NVMe) physical and virtual functions
  - Assignment of physical queues to virtual functions
  - Streamlining features required for virtual functions
  - Cross controller Quality of Service



Agenda

- NVM Express™ (NVMe) Explained and Ecosystem Update
- Architecting Data Center and Client Solutions With NVMe
- NVMe Over Fabrics – NVM Express Across the Data Center
- Future NVMe Technology Development
- Summary and Learning More

Summary

- NVM Express™ (NVMe) products are delivering leadership performance and latency
- Architect your product with the right features for the targeted segment
- NVMe over Fabrics enables benefit of next generation NVMe SSDs to be realized across the Data Center
- The innovation continues – e.g., Mobile and Virtualization features

Get involved by joining NVM Express – nvmexpress.org

Intel and VMware*
500TB All-Flash Virtual SAN Array

- 64 Nodes achieving 6.7M IOPs!
- Configuration
  - 6,400 Virtual Machines with Windows Server*
  - 64 Hyper-Converged VMware ESXi* Hosts with 2 NVMe devices each used as write buffers
  - 2,304 Intel® Xeon® Processors
  - 500 TB Raw Flash

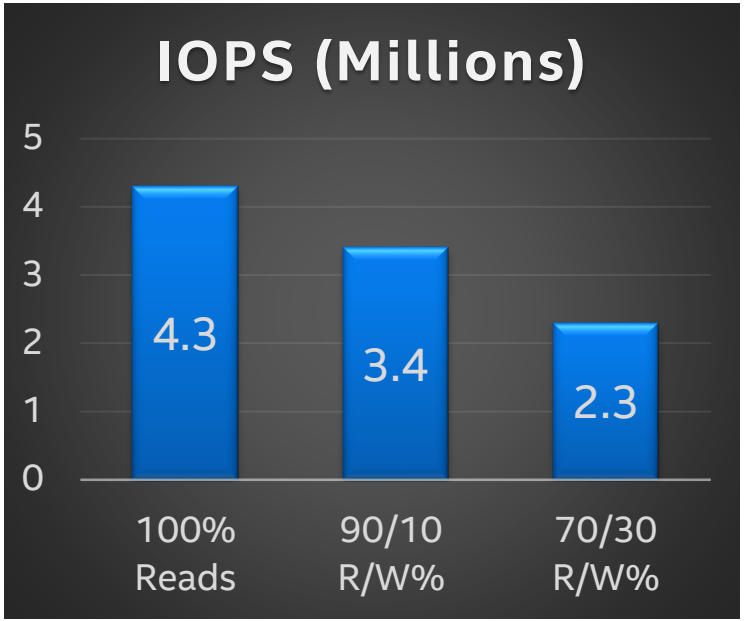
Intel collaborated with VMware* to deliver optimized solution. Check it out at the Intel® Solid-State Drive Pavilion.

Source: Intel*

NVM Express™ (NVMe)

Intel and Microsoft* Hyper-Converged All Flash Array Using Storage Spaces Direct

- 16 Nodes achieving 4.3M IOPs!
- Configuration
  - 16 Intel® Server Systems S2600WT
  - 128 VMs - 8 Virtual cores, 7.5GB memory per VM
  - Each server has 4 Intel® SSD DC P3700 Series devices
  - Data network is 10GbE RDMA



Intel collaborated with Microsoft* to deliver optimized solution. Check it out at the NVM Express™ Community (booth #879).

Source: Intel*

Learn More in the NVM Express™ Community

- Check out **today's** shipping NVM Express™ PCI Express® Solid-State Drive products

- Check out **tomorrow's** early prototypes from several IHVs of NVMe over Fabrics

Company	Booth #
Aperion Data Systems	873
EMC	887
HGST	886
Intel	871 & 881
JDSU	874
Kazan Networks	880
Keysight Technologies	876
Microsoft	879
PMC-Sierra	882
QLogic Corporation	883
Samsung Semiconductor	884
Seagate Technology	878
SK Hynix	885
Storage Networking Industry Association (SNIA)	888
Super Micro Computer	877
Teledyne LeCroy	872
Viking Technology	875

Additional Sources of Information

- A PDF of this presentation is available from our Technical Session Catalog: www.intel.com/idfsessionsSF. This URL is also printed on the top of Session Agenda Pages in the Pocket Guide.

- Demos in the showcase, outside NVM Express™ Community, include demonstrations in the Intel® Solid-State Drive Pavilion

- More web based info: www.nvmexpress.org and www.intel.com/ssd

Other Technical Sessions

Session ID	Title	Day	Time	Room
SSDS002	SSDs are Here – The Next Wave in Non-Volatile Memory-Driven Storage Modernizations	Tues	2:30	2008
SSDS003	What You Need to Know to Win the Storage Transition – Preparing for NVM Express™ in the Data Center	Tues	4:00	2008
SPCS006	Technology Insight: Intel Non-Volatile Memory Inside. The Speed of Possibility Outside	Tues	5:15	3016
SSDL001	Hands-on Lab: How to Unleash Your Storage Performance by Using NVM Express™ based PCI Express® Solid-State Drives	Wed	1:15; 4:00	2010
SSDC001	Tech Chat: Benchmarking Data Center Solid-State Drives – Insights Into Industry-Leading NVM Express™ SSD Performance Metrics	Wed and Thurs	10:30 Wed 9:30 Thurs	Tech Chat
SSDC002	Tech Chat: Insights into Intel® Solid-State Drives Data Retention and Endurance	Wed and Thurs	10:30 Wed 9:30 Thurs	Tech Chat
SSDC003	Tech Chat: NVM Express™ Features for High Availability and Storage Eco-System	Wed and Thurs	10:30 Wed 9:30 Thurs	Tech Chat
SSDS004	The Future of Storage Security	Thurs	1:00	2006
SFTS015	Next Generation Storage Architecture: Microsoft* Storage Spaces Direct and Intel® SSD Data Center Family for NVM Express™	Thurs	1:00	2008
SSDS005	New Software Capabilities and Experiences Through Innovation in Storage Architecture	Thurs	2:15	2006

Legal Notices and Disclaimers

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.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

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.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

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.

Intel, Core, Xeon, 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.

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the second 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 important factors that could cause actual results to differ materially from the company's expectations. Demand for Intel's products is highly variable and could differ from expectations due to factors including changes in business and economic conditions; consumer confidence or income levels; the introduction, availability and market acceptance of Intel's products, products used together with Intel products and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel's 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; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new products or incorporate new features into existing products, which may result in restructuring and asset impairment charges. 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. Results may also be affected by the formal or informal imposition by countries of new or revised export and/or import and doing-business regulations, which could be changed without prior notice. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. The amount, timing and execution of Intel's stock repurchase program could be affected by changes in Intel's priorities for the use of cash, such as operational spending, capital spending, acquisitions, and as a result of changes to Intel's cash flows or changes in tax laws. Product defects or errata (deviations from published specifications) may adversely impact our expenses, revenues and reputation. Intel's results could be affected by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues. 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. Intel's results may be affected by the timing of closing of acquisitions, divestitures and other significant transactions. 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.

Rev. 4/14/15