



THE SOLID STATE TRANSFORMATION OF THE DATA CENTER

Amber Huffman
Intel Fellow, Data Center Group, Intel Corporation
President, NVM Express, Inc.
September 26, 2018

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. **No product 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. For more complete information about performance and benchmark results, visit <http://www.intel.com/benchmarks>.

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 complete information visit <http://www.intel.com/benchmarks>.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.

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.

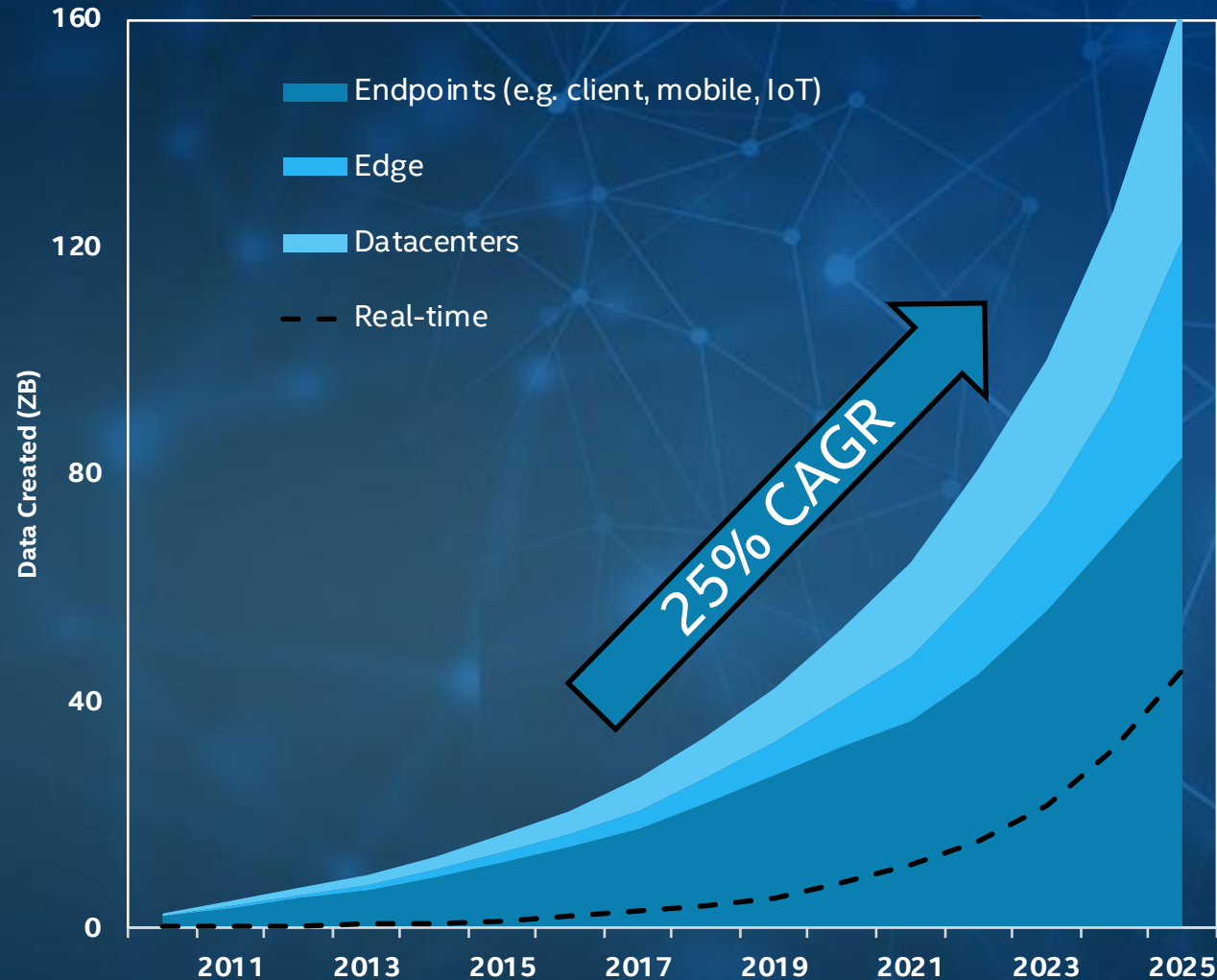
© 2018 Intel Corporation.

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

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

ZETABYTES AND MORE ZETABYTES ...

Global Digital Data Created (ZB)

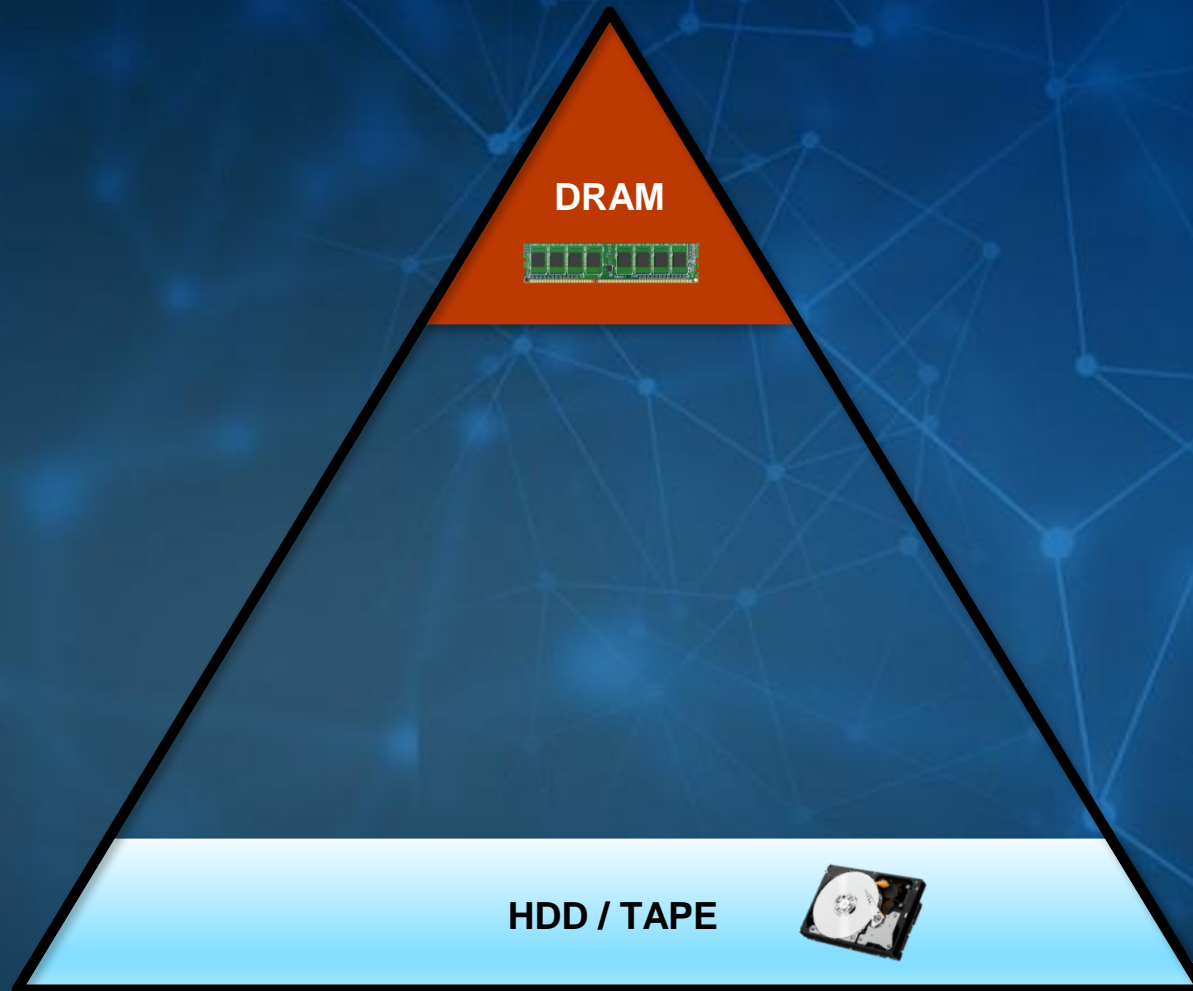


TODAY'S TALK

1. OUR INDUSTRY **TRANSFORMATION** TO EMBRACE NVM

2. OUR CONTINUED **OPPORTUNITIES** TOGETHER

THE JOURNEY BEGINS ~ 2000



ROLL BACK A **DECADE+**

Intel Developer Forum 2007

DRAMeXchange Projects SSD to Become the New Killer Application of NAND Flash in 2008

Tuesday, January 09, 2007 | DRAMeXchange

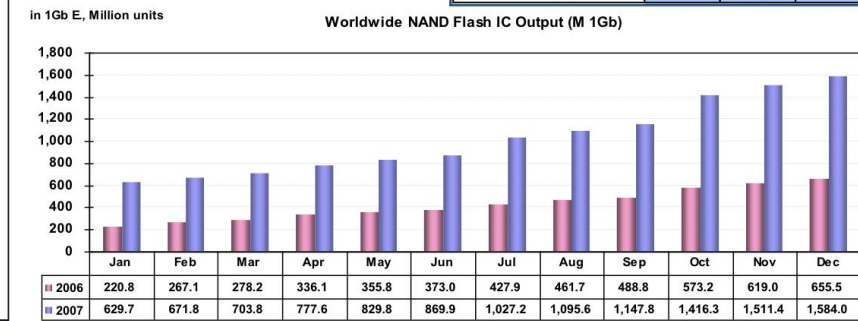
SanDisk, who is a prominent Flash memory card maker, unveiled a 1.8" 32Gb Solid State Drive (SSD) storage device on January 4th, 2007. The product is initially expected to target enterprises. As the unit production costs for NAND Flash are expected to drop further, SanDisk expects the general public (especially business executives who travel frequently) to begin showing interest in purchasing NBs that are equipped with an SSD in the near future.

PC is a Driver of NAND Growth

- NAND growth is projected to be 140% YoY for 06/07
- Largest growth area is "Others"
 - > 50% projected QoQ growth for Q3/Q4
- A key component of the "Others" category is PC uses

Unit: in 1Gb E., Million units

	Q307F		Q407F	
	Shipment	Demand	Shipment	Demand
Digital Still Cameras	26.8	697.9	31.5	884.6
	13.7%	32.0%	17.4%	27.0%
Cell phones	271.7	880.2	309.7	1,271.5
	7.0%	38.0%	14.0%	44.0%
USB Drives	35.1	615.2	38.9	734.5
	14.8%	29.0%	10.7%	19.0%
Flash-based MP3/PMP	32.6	779.7	46.7	1,223.4
	9.5%	32.0%	43.3%	57.0%
Other (DVs - Game)			54.0%	61.0%
Total NAND Flash Demand			33.0%	42.0%



Intel Developer FORUM

Source: DRAMeXchange, 4/2007



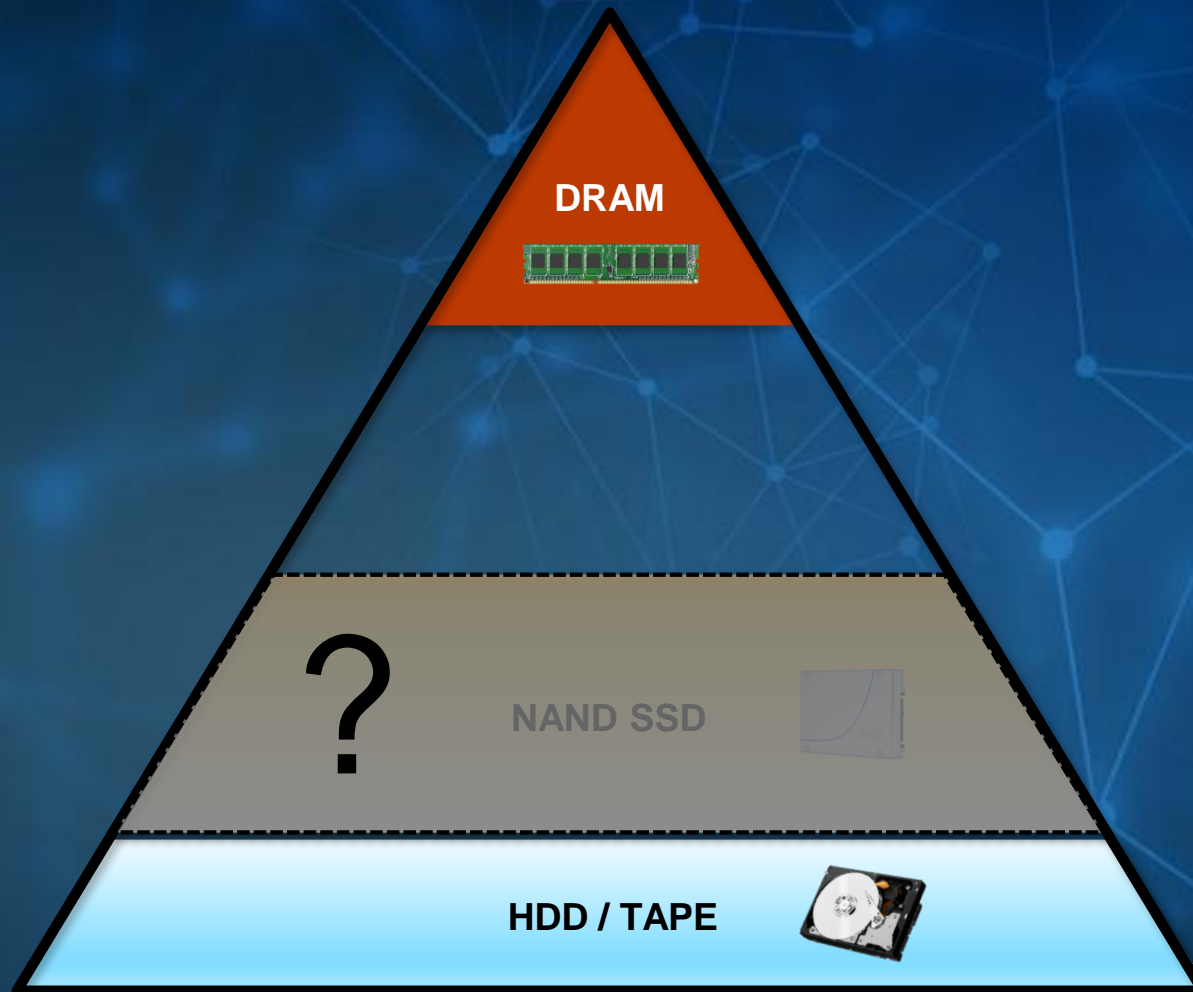
4

It was a question if SSDs would be a killer application for NAND.



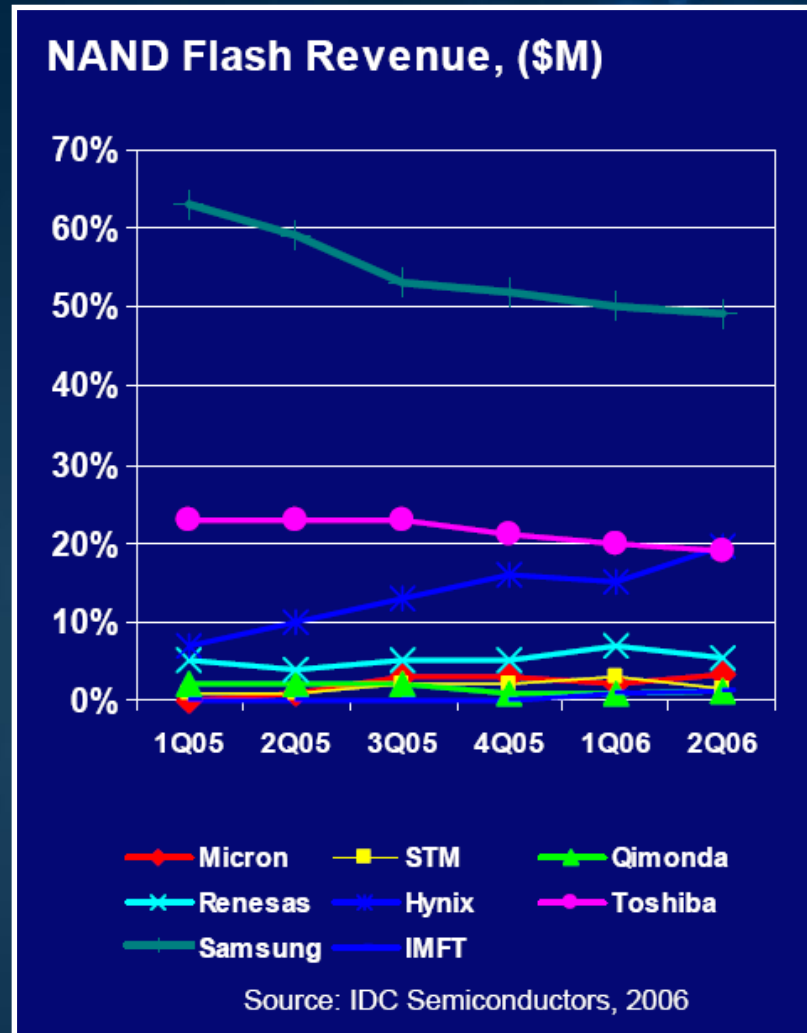
6

NAND SSD TIER ?



THE NEED FOR OPEN NAND FLASH INTERFACE

NO STANDARD NAND INTERFACE



There were many vendors, yet no standard interface, making it difficult to design SSDs.

ENTER OPEN NAND FLASH INTERFACE (ONFI)

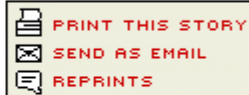
[EE Times: Semi News](#)

Intel, Hynix, Micron, Sony form NAND group

[Mark LaPedus](#)

[EE Times](#)

(05/09/2006 4:24 PM EST)



SAN JOSE, Calif. — Seeking to accelerate the time-to-market for NAND-based flash memories in the marketplace, Hynix, Intel, Micron, Phison and Sony are among the founding companies that on Tuesday (May 9) announced the formation of a new and long-awaited working group in the arena.

ONLINE
EETIMES



Open NAND Flash Interface Specification

Revision 1.0
28-December-2006

Hynix Semiconductor
Intel Corporation
Micron Technology, Inc.
Phison Electronics Corp.
Sony Corporation
STMicroelectronics

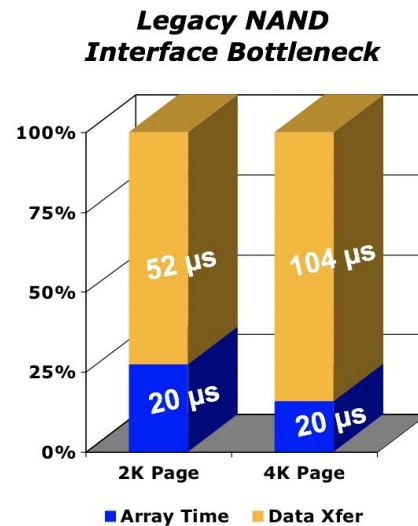
ONFI codified commonalities, and then started to scale for SSDs.

INCREASING PERFORMANCE FOR SSDS

Intel Developer Forum 2008

Legacy Interface Bottleneck

- NAND performance is determined by two elements
 - NAND array access time
 - Data transfer time across the bus
- For legacy NAND reads, the dominant factor is the bus!
 - Performance is limited to 40 MB/s
 - With interface improvements data could be read at over 150 MB/s
- The issue gets **significantly worse** as page size increases



80+ μ s "hiccup" waiting for the interface bottleneck during every 4KB read.

Intel Developer
FORUM

10

ONFi 2.0 Eliminates the Bottleneck

- ONFi 2.0 was published in February
- Adds a synchronous DDR interface option for high speed
 - 133 MT/s in first generation
 - Scalability to 400 MT/s
 - 3.3V and 1.8V VccQ options
 - Optimized BGA package
- Additional headroom found, ONFi 2.1 underway now and will add 166 MT/s and 200 MT/s speeds

Interface Roadmap	
Legacy	40 MB/s
Gen1	~ 133 MB/s
Gen2	~ 266 MB/s
Gen3	400 MB/s +

ONFi 2.0 triples the legacy interface speed. ONFi 2.1 with more speed targeted for 2H'08.



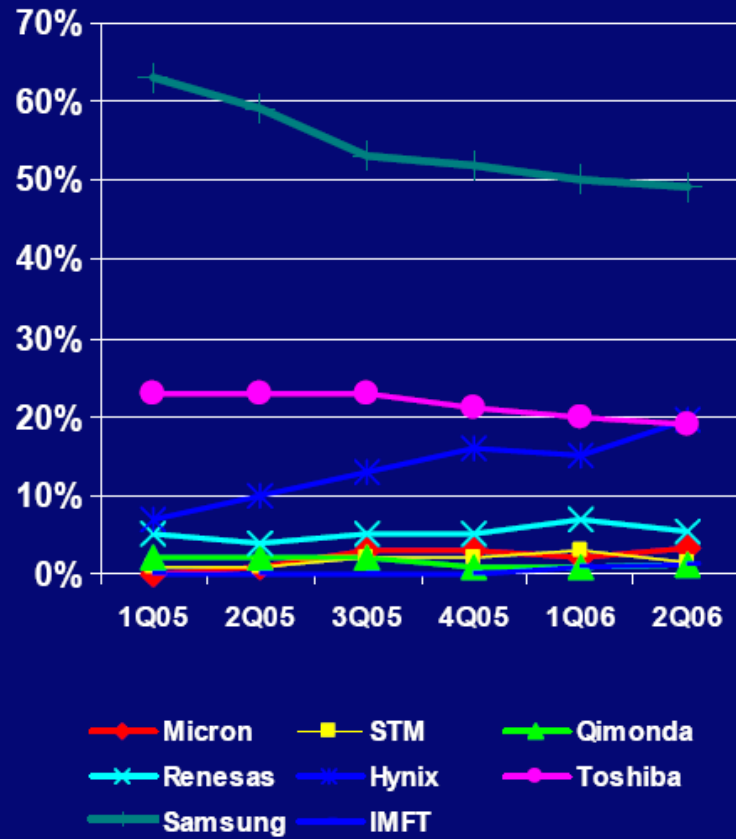
Intel Developer
FORUM

11

ONFI defined 10x scaling of NAND interface in < 2 years.

ENABLING **SCALE** FOR SSD INDUSTRY

NAND Flash Revenue, (\$M)



Source: IDC Semiconductors, 2006

JEDEC and ONFi Collaboration

- The ONFi Workgroup is pleased to team up with JEDEC on NAND standardization moving forward
- ONFi is submitting the ONFi 2.0 specification as part of the joint effort



Intel Developer
FORUM

13

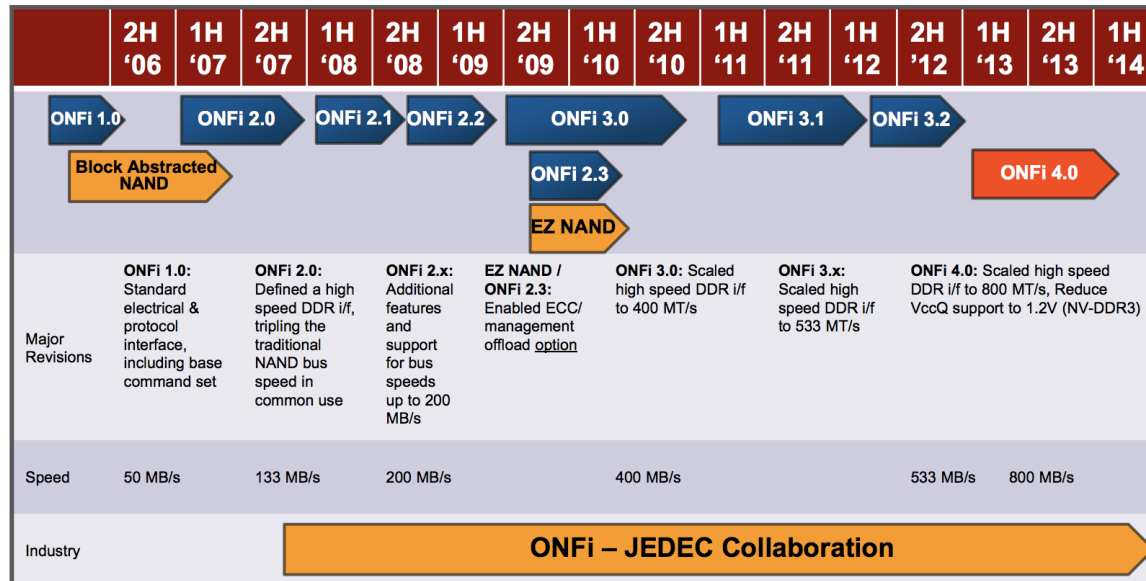
ONFI / JEDEC collaboration enabled unifying the industry to support scale of SSD ambitions.

ONFI CONTINUED INNOVATION

Flash Memory Summit 2014



ONFI Workgroup Continues To Produce Results!



ONFI has and continues to deliver innovation & interoperability enabling faster NAND adoption

Steady innovation...

KEEPING **PACE** WITH SSD NEEDS



Open NAND Flash Interface Specification

Revision 4.1
12 12 2017

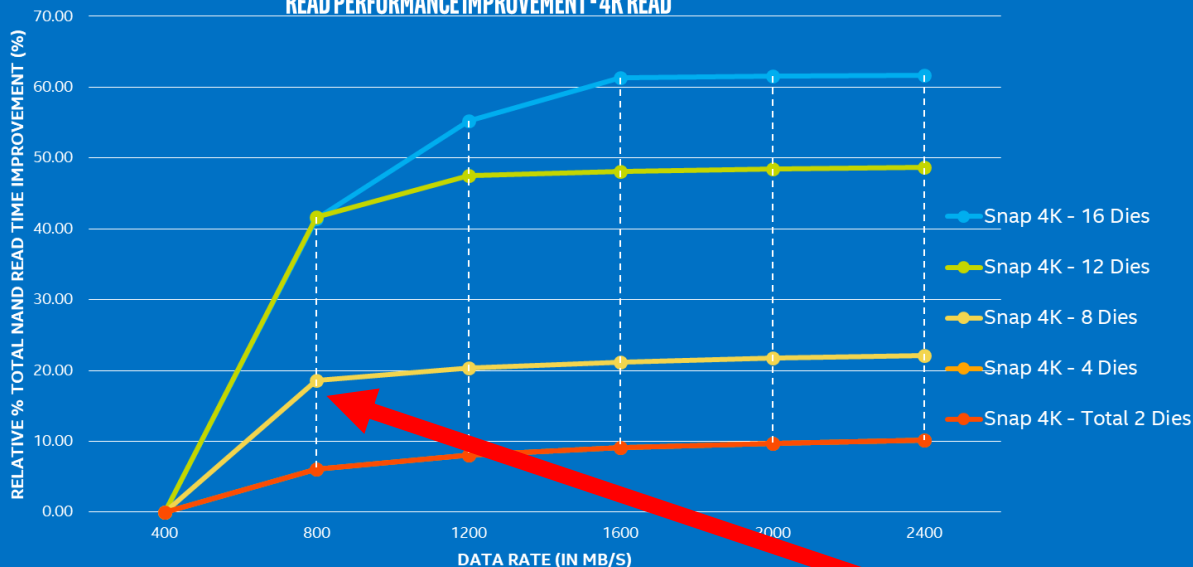
Intel Corporation
Micron Technology, Inc.
Phison Electronics Corp.
Western Digital Corporation
SK Hynix, Inc.
Sony Corporation

**ONFI 4.0 in 2014 scaled to 800
MT/s.**

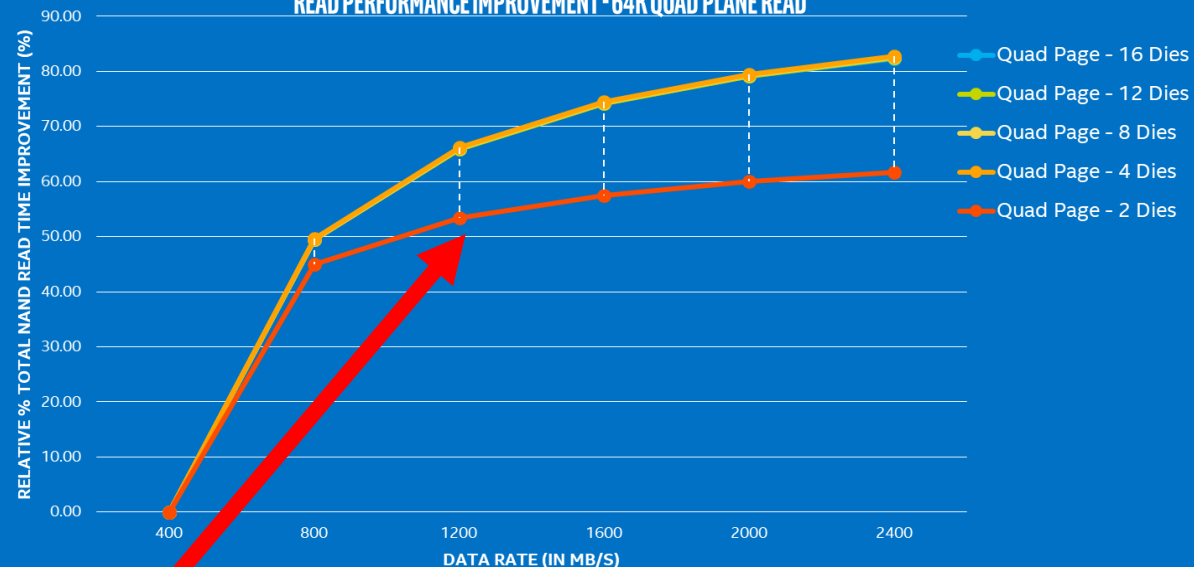
**ONFI 4.1 in 2017 scaled to 1200
MT/s.**

NAND PERFORMANCE IMPROVEMENTS ARE SLOWING

READ PERFORMANCE IMPROVEMENT - 4K READ



READ PERFORMANCE IMPROVEMENT - 64K QUAD PLANE READ



ROI reduction at higher transfer rates due to NAND performance. Keep ONFI steady, unless breakthrough in NAND media.

THE PATH TO **NVM EXPRESS**

THE ORIGINAL "NVMHCI"

Flash Memory Summit 2009



Remember NVMHCI: An Optimized Interface for NVM

- NVMHCI: Non-Volatile Memory Host Controller Interface
- NVMHCI is a clean and optimized interface for SSDs and caches
- NVM equivalent of the SATA AHCI controller interface



Companies Driving NVMHCI

The NVMHCI Workgroup includes 40+ members, focused on delivering streamlined NVM solutions.

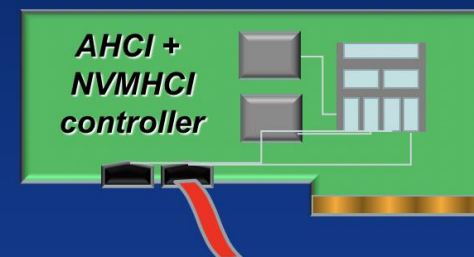
Santa Clara, CA USA
August 2009

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



Technical Essence of NVMHCI

- NVMHCI defines a standard programming interface for non-volatile memory subsystems
- Leverage AHCI to provide best infrastructure for caching
 - One driver for HDDs and NAND
- Allows NVMHCI registers to appear as:
 - A separate PCI device
 - A port within an existing AHCI controller
- NVMHCI is a logical interface
 - All NAND management abstracted out: NAND technology changes too quickly
 - All caching algorithms are outside the spec: NVMHCI only defines how caching software gets access to the NAND
- Optimized interface for both cache and SSD usage models



Santa Clara, CA USA
August 2009

4



NVMHCI FOR CLIENT WAS A MISFIRE

NVMHCI AS A **SPRING BOARD** FOR ENTERPRISE

Flash Memory Summit 2009

SMART Modular Technologies Announces Enterprise Class PCI Express Storage Solution



Delivering 140K random performance and 200x power reduction, the new 400GB XceedOPS PCIe raises the bar in next-generation solid-state storage.

NEWARK, CA, June 1, 2009 - SMART Modular Technologies (WVH), Inc. ("SMART" or the manufacturer) has announced its reaction to Fusion-io's patented RAID controller. SMART's new XceedOPS PCIe is the first in a line of applications.

The Register
Biting the hand that feeds IT

Micron barges into PCIe SSD business
Partners up with IDT
By [Chris Mellor](#) - [Get more from this author](#)
Posted in Storage | 28th July 2009 12:51 GMT

Fusion-io unveils 80GB ioXtreme PCI Express SSD

By Matthew DeCarlo, TechSpot.com
Published: June 8, 2009, 9:15 AM EST

TECHSPOT
PC TECHNOLOGY NEWS AND ANALYSIS

Fusion-io is launching a new "FatalIty" branded product as they deliver an enthusiast-oriented PCI Express solid state drive. The ioXtreme SSD will make use of the PCI-E x4 interface and bear a non-volatile 80GB capacity based on MLC NAND technology.



RamSan-20 highlights:

- 450GB Flash storage
- 120,000 IOPS
- 700 MB/s random sustained external throughput
- ECC and RAID
- Embedded CPU controller.

To order, please contact [Texas Memory Systems Sales](#)

Super Talent Debuts PCI Express SSD RAIDrive

Wednesday, April 01, 2009 - by [Shawn Oliver](#)

Please [Super Talent](#), let this not be a joke. Here on the 1st of April, Super Talent has announced its reaction to Fusion-io's patented RAID controller. Super Talent's new RAIDrive is the first in a line of applications.

HOT HARDWARE
THE HOTTEST TECH, TESTED AND BURNED IN

OCZ's New Blazing Fast 1TB Z SSD Drive

7:00 PM - March 4, 2009 by Steve Seguin

OCZ was demonstrating its new Z Drive with 1 TB of storage at CeBIT 2009 this week.

The new Z Drive from OCZ is a storage device that connects to an x8 PCIe slot and offers 1 terabyte of [storage capacity](#). The Z Drive is about the same size a dual-slot graphics card, so its not exactly small, but the device is stated to offer maximum read and write speeds of up to 600 MB/sec. and 500 MB/sec., respectively.



Enterprise class PCIe SSDs are coming to market. However, they have proprietary interfaces impacting adoption.

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



Extend NVMHCI for Enterprise class PCIe SSDs

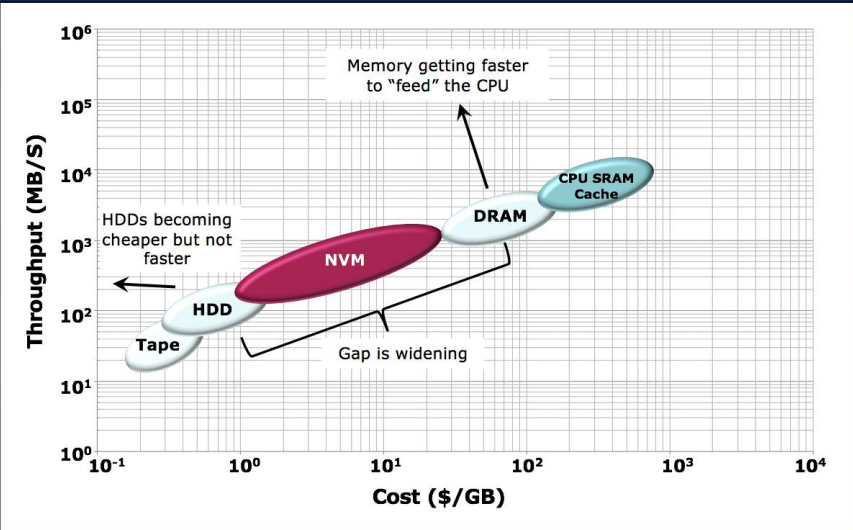
- Extend NVMHCI to meet the needs of Enterprise PCIe SSDs
 - Address Enterprise server scenarios
 - Enables SSD vendors to focus on building a great SSD
 - Enables OS vendors to deliver a great driver for all PCIe SSDs
 - Enables OEMs to qualify a single driver on each OS, with features implemented in a consistent fashion, reducing time to market
- Leverage NVMHCI interface, software infrastructure, and Workgroup to fill this gap quickly with a streamlined solution
 - Make NVMHCI an ideal interface for Enterprise PCIe SSDs
 - Take advantage of drivers already written or underway
 - Take advantage of existing Workgroup, an efficient team that can execute quickly

Santa Clara, CA USA
August 2009

ADDRESSING THE GAP ...



Gap in the Storage/Memory Hierarchy is Growing



NVM is filling the price/performance gap between DRAM and HDD, thereby creating the "I/O Memory Tier"

Santa Clara
August 2010



3



Enterprise NVMHCI Goals & Timeline

- Goals for standard:
 - Address Enterprise usage scenarios
 - Enable an efficient & scalable interface, from very high-end to client
 - Ensure no interface impediments to exceeding > 1M IOPs
 - Enable OS vendors to deliver standard high performance drivers
 - Provide a consistent feature set to enable SSD interoperability
 - Reduce TTM for PCIe SSDs by enabling OEMs to validate/qual one PCIe SSD driver for each OS and one consistent feature set
- To get involved, join the NVMHCI Workgroup
 - Details at <http://www.intel.com/standards/nvmhci>

	Apr '10	May '10	Jun '10	Jul '10	Aug '10	Sep '10	Oct '10	Nov '10	Dec '10
Revision	0.5		0.7			0.9		RC	1.0
Definition	0.5: Basic capabilities and approach defined.			0.7: Basic definition complete for all features. Feature freeze.			0.9: Erratum only. RC: Member review. 1.0: Published.		

0.70 revision achieved, available for Contributor review. Schedule enables product intercept in 2012.

Santa Clara
August 2010

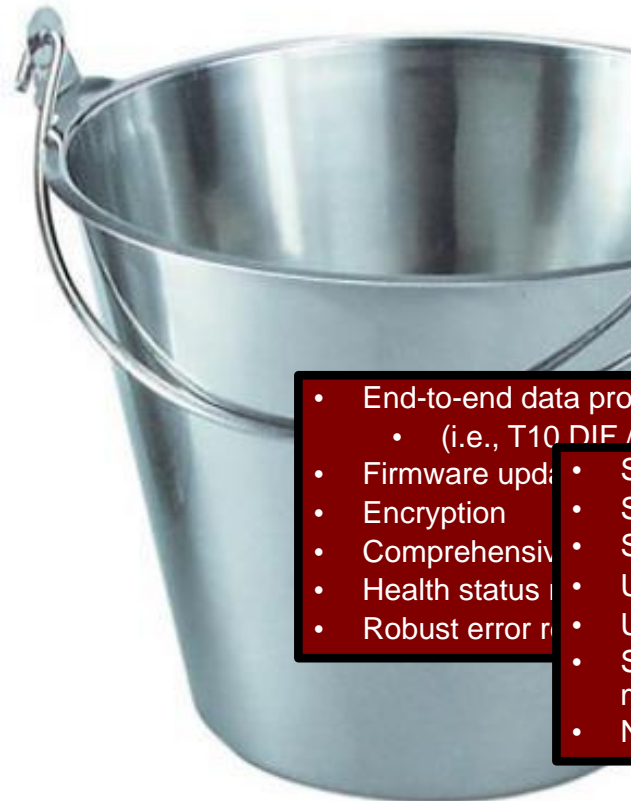
PRINCIPLES OF ENTERPRISE NVMHCI



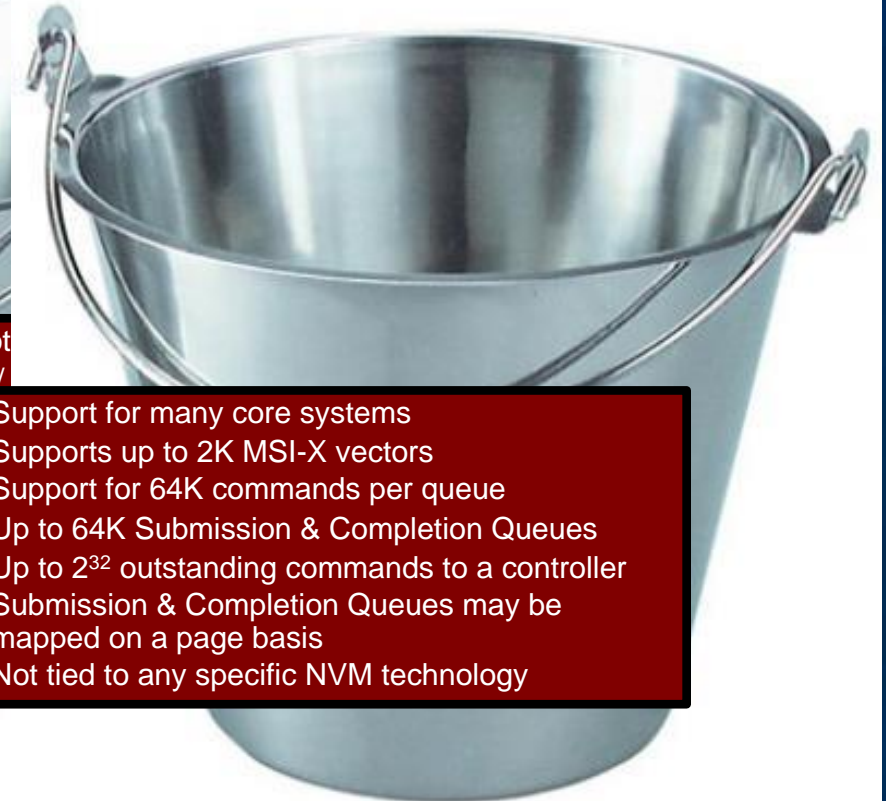
Bucket 1: Eliminated
seen in



Bucket 2: Provides
streamlined co



Bucket 3:
Provides Enterprise feature



Bucket 4: Provides scalable architecture
for now & the future.

-
-
-
-
-
-

- Do not
- Eight d
- Efficient
- Efficient
- Support
- Support

- End-to-end data prot
- (i.e., T10 DIF /
- Firmware upd
- Encryption
- Comprehensive
- Health status
- Robust error r

- Support for many core systems
- Supports up to 2K MSI-X vectors
- Support for 64K commands per queue
- Up to 64K Submission & Completion Queues
- Up to 2^{32} outstanding commands to a controller
- Submission & Completion Queues may be mapped on a page basis
- Not tied to any specific NVM technology

NVM EXPRESS “BORN” IN MARCH 2011

NVM Express* Overview

- NVM Express is a scalable host controller interface designed for Enterprise and Client systems that use PCI Express* SSDs
 - Includes optimized register interface and command set
- NVMe was developed by industry consortium of 80+ members and is directed by a 10 company Promoter Group




- NVMe 1.0 published on March 1st, available at nvmexpress.org

NVMe standardization effort is complete & stable
Product introductions coming in 2012

IDF2011
INTEL DEVELOPER FORUM

9

NVMe*: Efficient SSD Performance

	AHCI ¹	
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	32	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 designed for high parallelism and low latency

¹AHCI: Serial ATA programming interface. See <http://www.intel.com/technology/serialata/ahci.htm>

IDF2011
INTEL DEVELOPER FORUM

DEVELOPING THE ECOSYSTEM

Intel Developer Forum 2012

Interoperability Program Underway

- The NVM Express Workgroup is collaborating with an industry leader, UNH-IOL, to develop the NVMe Interoperability program

March 12, 2012 08:00 AM Eastern Daylight Time

Industry Leaders Develop New, High-Performance PCIe SSD Solutions at UNH-IOL



University of New Hampshire
InterOperability
Laboratory

Lab Now Accepting Founding Member Companies for NVMe Consortium

DURHAM, N.H. --(BUSINESS WIRE)--The University of New Hampshire InterOperability Laboratory (UNH-IOL), an independent provider of broad-based testing and standards conformance services for the networking and storage industries, is accepting founding members for the laboratory's new Non-Volatile Memory Express (NVMe) Consortium. The NVMe Consortium will focus on developing an interoperability test suite for NVMe compliant software and devices. Founding members will join industry leaders Dell, EMC, IDT, Intel, LSI Corporation, NetApp, Oracle and SanDisk to create new, innovative, high-performance storage solutions based on the NVMe standard for PCIe SSDs.

- UNH-IOL has extensive experience in conformance and interop test services for leading industry standards in storage & networking (SATA, SAS, Fibre Channel, etc.)
- Since late 2011 UNH-IOL has been working with the NVMe Promoter Group to develop NVMe test documentation and tools

NVMe is working with UNH-IOL to ensure an interoperable ecosystem that OEMs can count on

33

NVMe = NVM Express

Flash Memory Summit 2013



Driver Developments on Major OSes

Windows*	• Windows* 8.1 includes inbox driver • Open source driver in collaboration with OFA
Linux*	• Native OS driver since Linux* 3.3 (Jan 2012)
Unix	• FreeBSD driver upstream; ready for release
Solaris*	• Solaris driver will ship in S12
VMware	• vmklinux driver certified release in Dec 2013
UEFI	• Open source driver available on SourceForge

Native OS drivers already available in Windows and Linux!

Flash Memory Summit 2013
Santa Clara, CA



16

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

BUILDING OUT ENTERPRISE FEATURES

NVM Express* 1.1 Overview

- The NVM Express 1.1 specification, published in October of 2012, adds additional optional client and Enterprise features

Multi-path Support

- Reservations
- Unique Identifier per Namespace
- Subsystem Reset

Power Optimizations

- Autonomous Power State Transitions

Command Enhancements

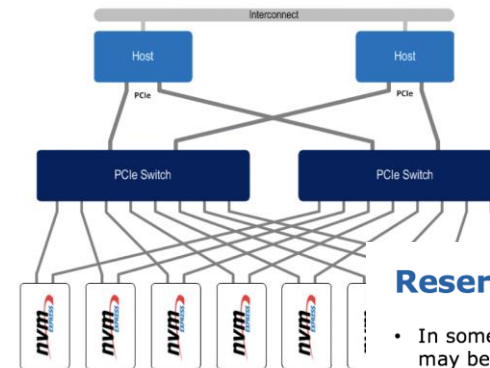
- Scatter Gather List support
- Active Namespace Reporting
- Persistent Features Across Power States
- Write Zeros Command

IDF13

21

Multi-path Support

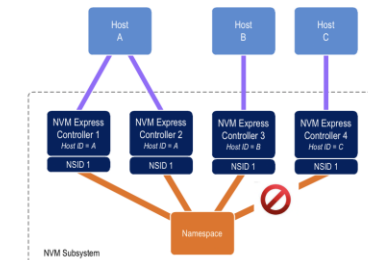
- Multi-path includes the traditional dual port model
- With PCI Express*, it extends further with switches



22

Reservations

- In some multi-host environments, like Windows* clusters, reservations may be used to coordinate host access
- NVMe 1.1 includes a simplified reservations mechanism that is compatible with implementations that use SCSI reservations
- What is a reservation? Enables two or more hosts to coordinate access to a shared namespace.
 - A reservation may allow Host A and Host B access, but disallow Host C



23

IDF13

NVMe SHIPS IN 2H 2013

NVM Express* Deployment is Starting

- First plugfest held May 2013 with 11 companies participating
 - Three devices on Integrator's List
 - Next plugfest planned for Q4
- Samsung announced first NVM Express* (NVMe) product in July

FOR IMMEDIATE RELEASE

NVM Express Workgroup Holds First Plugfest

Milestone in Process to Deliver Standards-based Interoperability for PCI Express Solid-State Drives

WAKEFIELD, Mass., May 29, 2013 – The [NVM Express Workgroup](#), developer of the NVM Express specification for accessing solid-state drives (SSDs) on a PCI Express (PCIe) bus, held its first Plugfest at the University of New Hampshire InterOperability Lab in Durham, N.H., May 13-16, 2013. This event provided an opportunity for participants to measure their products' compliance with the NVM Express (NVMe) specification and to test interoperability with other NVMe products.

The NVMe specification defines an optimized register interface, command set and feature set for PCIe-based Solid-State Drives (SSDs). NVMe refers to non-volatile memory, as used in SSDs. The goal of NVMe is to unlock the potential of PCIe SSDs now and in the future, and to standardize the PCIe SSD interface. Participating in the Plugfest were Agilent Technologies, Dell Inc., Fastor Systems, Inc., HGST, a Western Digital company, Integrated Device Technology, Inc., Intel Corporation, Samsung Electronics Co., Ltd., SanDisk Corporation, sTec, Inc., Teledyne LeCroy, and Western Digital Corporation.

JULY 18TH, 2013 by Josh Linden

Samsung Announces Industry's First 2.5-Inch NVMe SSD

Samsung has announced the XS1715, a 2.5-inch Non-Volatile Memory Express (NVM Express) PCIe SSD. According to Samsung, the 1.6TB SFF-8639 NVMe SSD provides a sequential read speed at 3,000MB/s, six times faster than the company's current high-end enterprise SSD. The XS1715's random read performance is specified at up to 740,000 IOPS, more than 10 times as fast as existing SSD options.



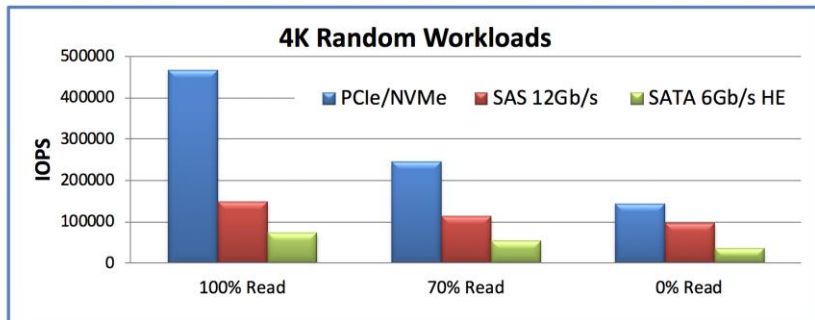
NVMe products targeting Datacenter shipping this year

IDF13

NVMe DELIVERS IN PERFORMANCE

NVM Express* (NVMe) Delivers Best in Class IOPs

- 100% random reads: NVMe has >3X better IOPs than SAS 12Gbps
- 70% random reads: NVMe has >2X better IOPs than SAS 12Gbps
- 100% random writes: NVMe has ~ 1.5X better IOPs than SAS 12Gbps



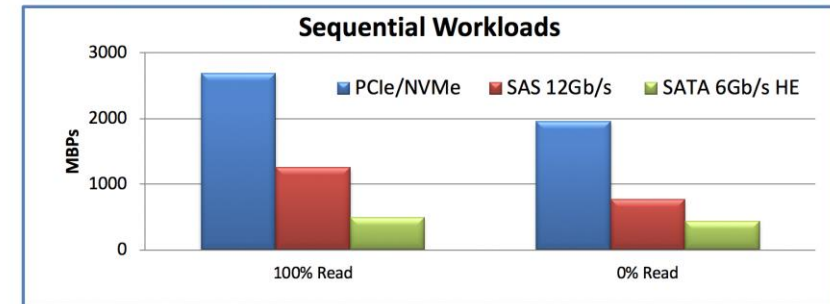
Note: PCI Express (PCIe)/NVM Express (NVMe) Measurements made on Intel® Core™ i7-3770S system @ 3.1GHz and 4GB Mem running Windows® Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by iometer® tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar® SSD800M/1000M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

IDF14

14

And Best in Class Sequential Performance

- NVM Express* (NVMe) delivers > 2.5GB/s of read and ~ 2 GB/s of write performance
 - 100% reads: NVMe has >2X better performance than SAS 12Gbps
 - 100% writes: NVMe has >2.5X better performance than SAS 12Gbps



Note: PCI Express (PCIe)/NVMe Measurements made on Intel® Core™ 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. Source: Intel Internal Testing. 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.

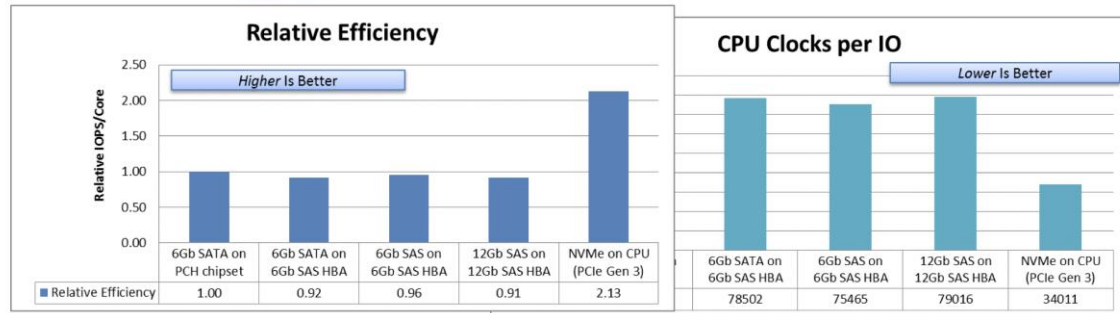
IDF14

15

EFFICIENTLY AND WITH LOW LATENCY

The Efficiency of NVM Express* (NVMe)

- CPU cycles in a Data Center are precious
- And, each CPU cycle required for an IO adds latency
- NVM Express* (NVMe) takes less than half the CPU cycles per IO as SAS

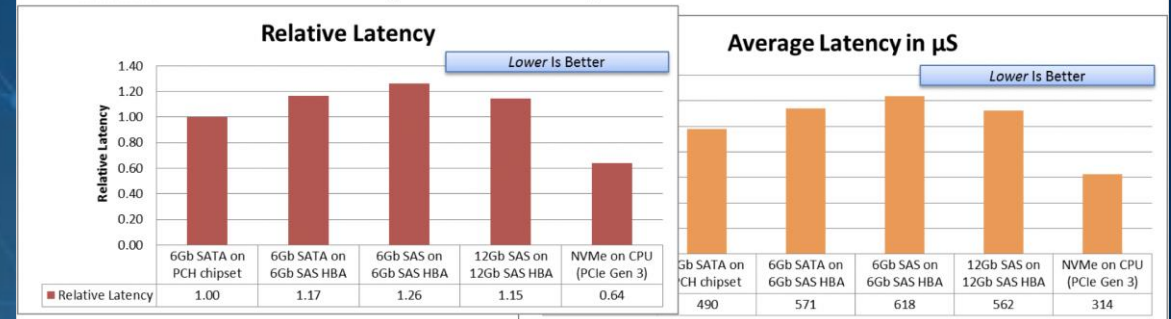


With equivalent CPU cycles, NVMe delivers over 2X the IOPs of SAS!

IDF14

The Latency of NVM Express* (NVMe)

- The efficiency of NVM Express* (NVMe) directly results in leadership latency
- When doubling from 6Gb to 12Gb, SAS only reduces latency by ~ 60 μ S
- NVMe is more than 200 μ s lower latency than 12 Gb SAS

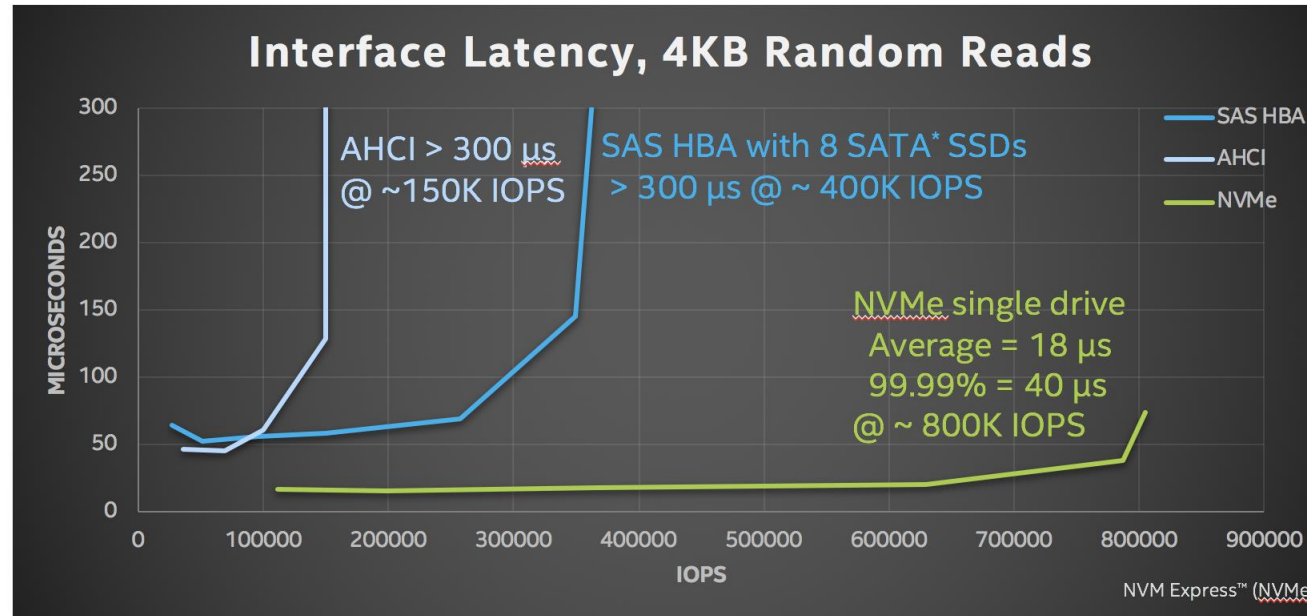


NVMe delivers the lowest latency of any standard storage interface.

IDF14

WITH BETTER QUALITY OF SERVICE

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).

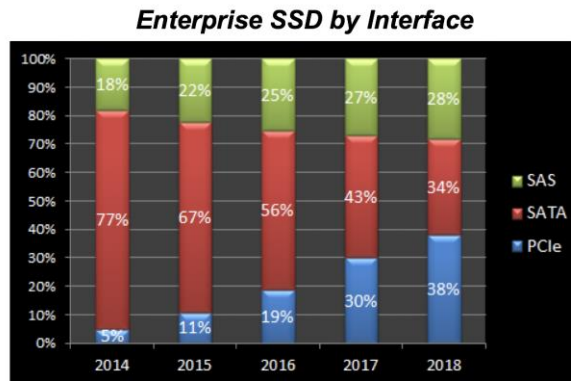
9

IDF15
INTEL DEVELOPER FORUM

AND ANALYSTS NOTICE

PCI Express* (PCIe*) SSDs Projected to Lead in Data Center

- PCI Express* (PCIe*) projected as leading SSD interface in DC by 2018
- PCIe leads in performance
 - PCIe bandwidth is significantly higher than SAS or SATA
 - NVMe Express* (NVMe) has lower latency than SAS or SATA
- Industry standards for PCIe in place
 - NVMe is the software interface
 - SFF-8639 defines a 2.5" form factor



Source: IDC

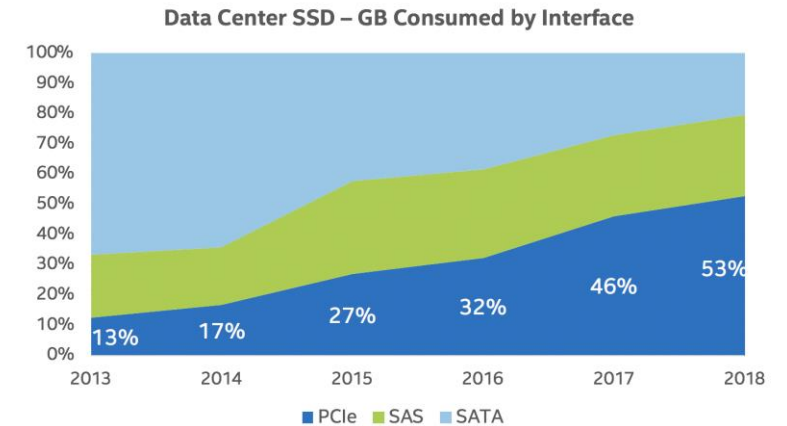
PCIe SSDs lead the way by embracing industry standards.

IDF14

6

Data Center Interface Dynamics, One Level Deeper

- PCI Express* (PCIe*) is projected to lead even sooner by capacity
- More NVMe is shipped in each PCIe SSD than with other interfaces



PCIe projected to lead in NVM shipped to Data Center in 2016.

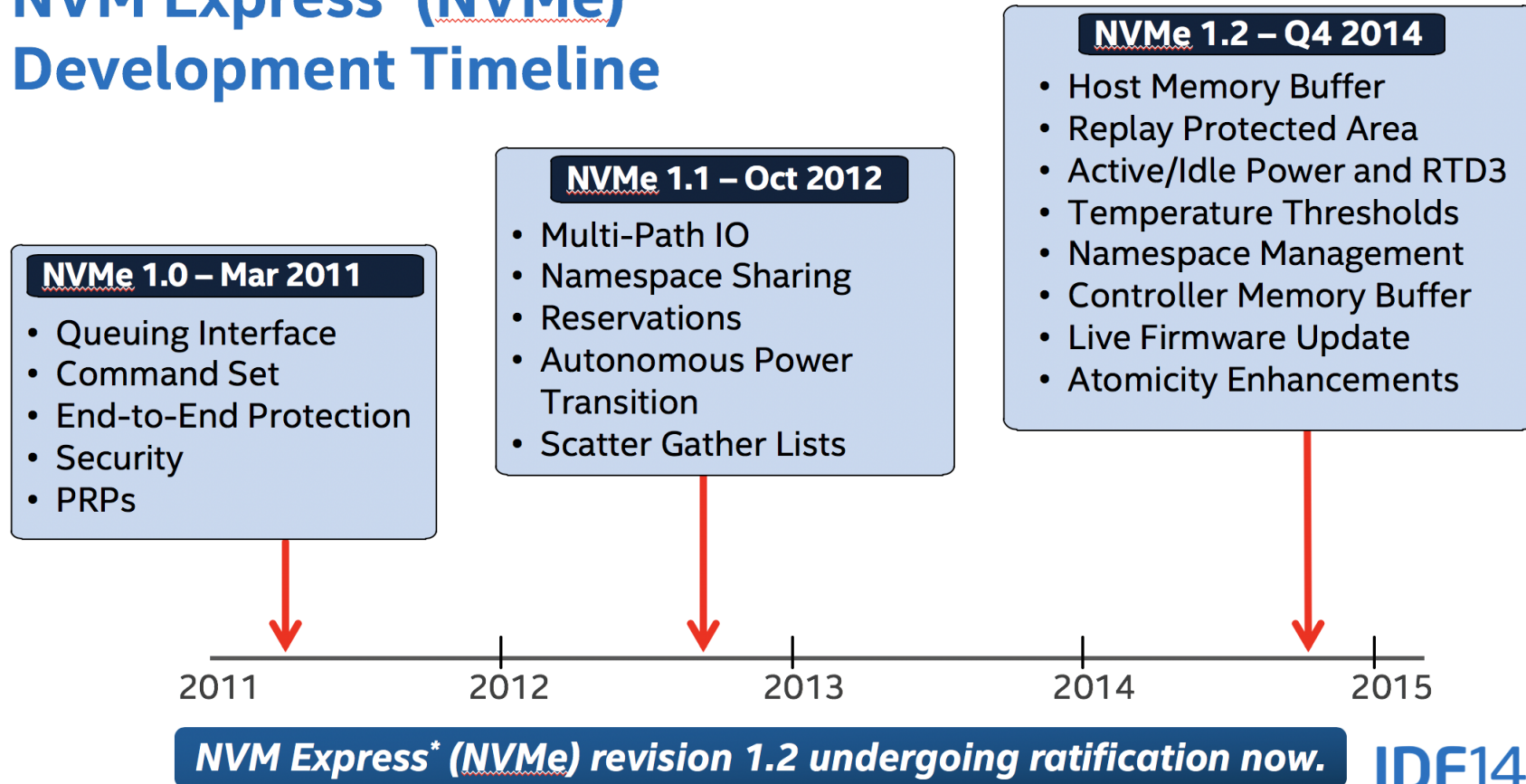
Source: Intel Market Model and multiple industry analysts

IDF14

7

NVMe CONTINUES TO ADD CAPABILITIES

NVM Express* (NVMe) Development Timeline

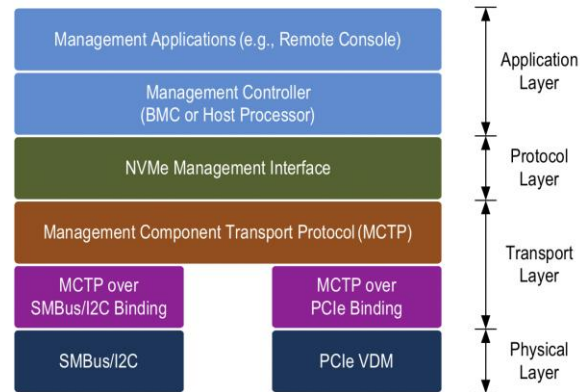


26

AND HEADS TO NEW FRONTIERS ...

NVM Express* (NVMe) Management Interface

- Defines out-of-band management that is independent of the physical transport and protocol
- Maps the management interface to one or more out-of-band physical interfaces (e.g., I2C, PCI Express*)
- Specifies a management command set for NVM Express* (NVMe) devices



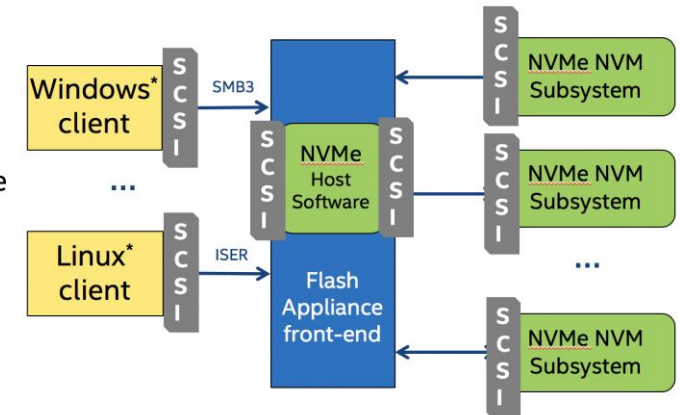
The management interface is targeted for completion end of year.

IDF14

27

NVM Express* (NVMe) in Fabric Environments

- A primary use case for NVM Express* (NVMe) is in a Flash appliance
- Hundreds or more SSDs may be attached – too many for PCI Express* based attach
- Concern: Remote SSD attach over a fabric uses SCSI based protocols today – requiring protocol translation

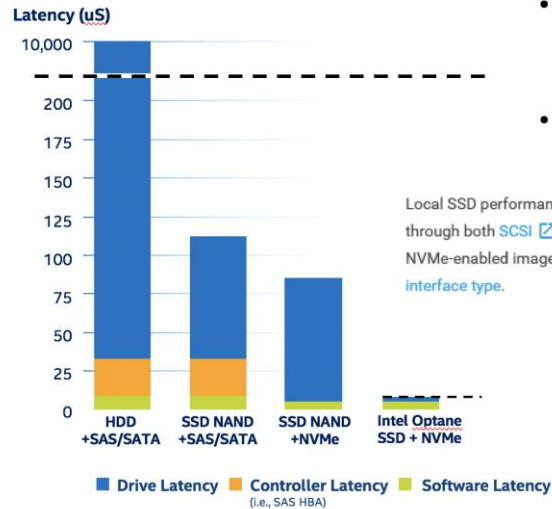


Desire best performance and latency from SSD investment over fabrics like Ethernet, InfiniBand™, Fibre Channel, and Intel® Omni Scale Fabric.

28

DRIVING INNOVATION IN CLOUD

Broad Adoption of NVM Express*



- NVM Express (NVMe*) delivers speed required by Cloud Service Providers
- NVMe is ready for Intel® Optane™ SSDs

Local SSD performance depends heavily on which interface you select. Local SSDs are available through both [SCSI](#) and [NVMe](#) interfaces. If you choose to use NVMe, you must use a special NVMe-enabled image to achieve the best performance. For more information, see [Choosing a disk interface type](#).



Source: Storage Technologies Group, Intel. Comparisons between memory technologies based on in-market product specifications and internal Intel specifications.



Driving Cloud Innovation

- Cloud storage innovation is centering on NVMe*
- A great example is Facebook's Lightning design

DDP U.S. SUMMIT 2016

Lightning: A flexible NVMe JBOD

Chris Petersen
HARDWARE SYSTEMS ENGINEER, FACEBOOK

Mike Yan
HARDWARE ENGINEER, FACEBOOK

Clark Shao
HARDWARE ENGINEER, FACEBOOK

Why NVMe?



It scales!



It's open source!



Multiple form factors!



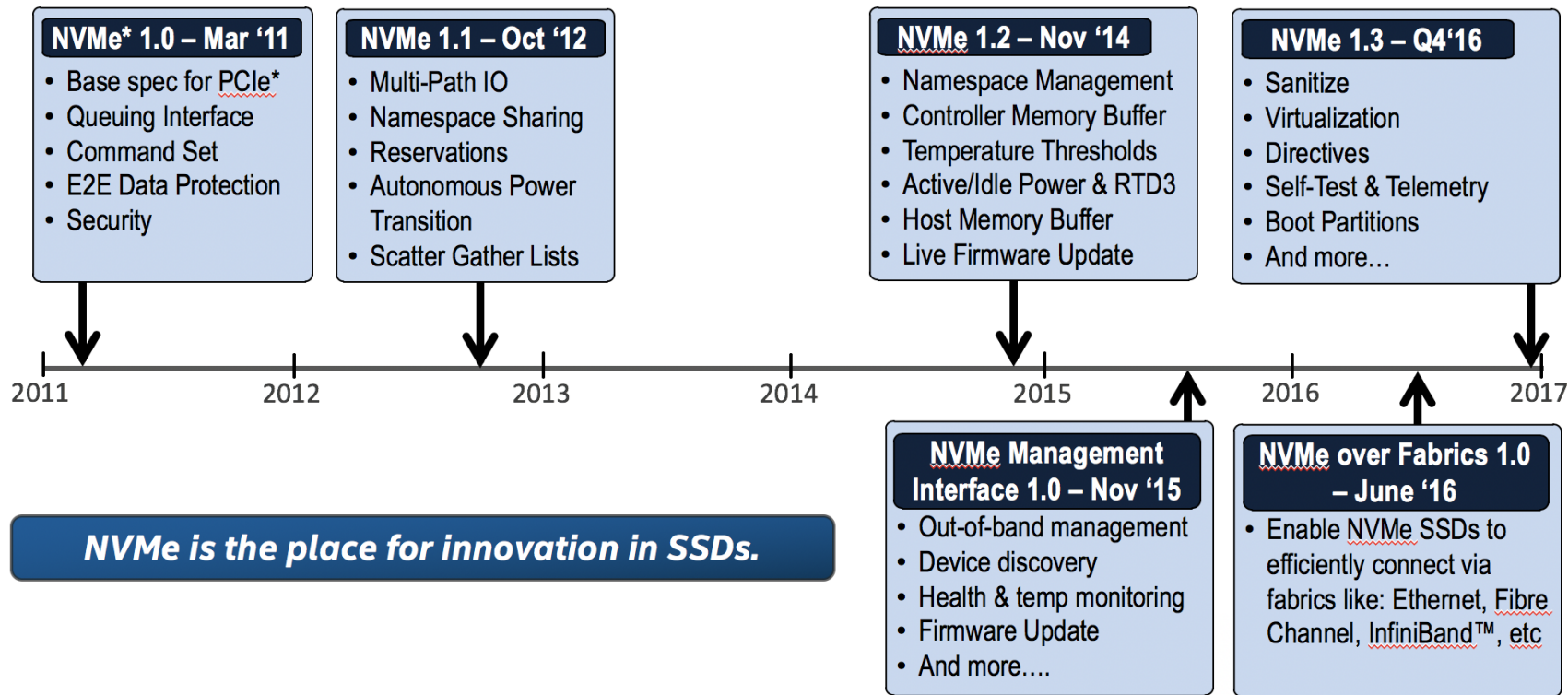
Design objectives

- Modular
- Flexible configurations
- Compatible with upcoming standards



EVEN MORE CAPABILITIES ... NVMe REVISION 1.3

NVMe* Development Timeline



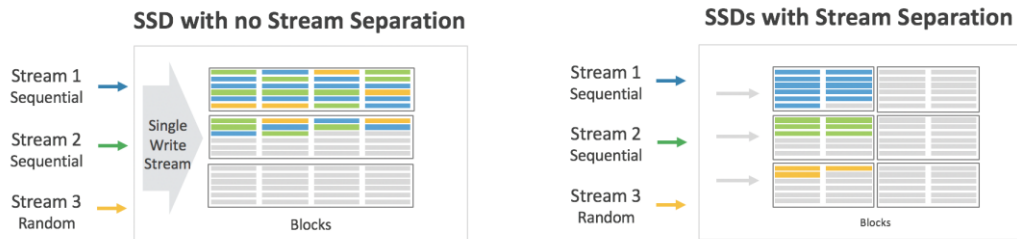
*Other names and brands may be claimed as the property of others.

30

STREAMS, VIRTUALIZATION, AND MORE

Directives : Streams

- Allows a host to physically segregate ~ 10 – 20 streams of data
- If host manages data well, reduces write amplification
 - E.g., stream 3 no longer interferes with stream 1 and stream 2

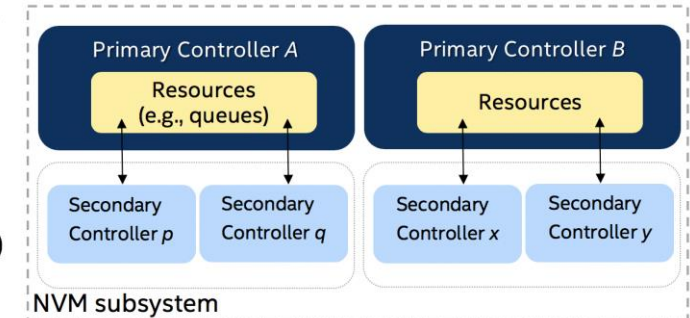


IDF16
INTEL DEVELOPER FORUM

32

Direct Assignment Support in NVMe*

- There is a hierarchy of *primary* and *secondary* controllers
- The near term approach maps onto PCIe* SR-IOV
 - *primary* = physical function (PF)
 - *secondary* = virtual function (VF)
- Abstraction allows future mechanisms beyond SR-IOV

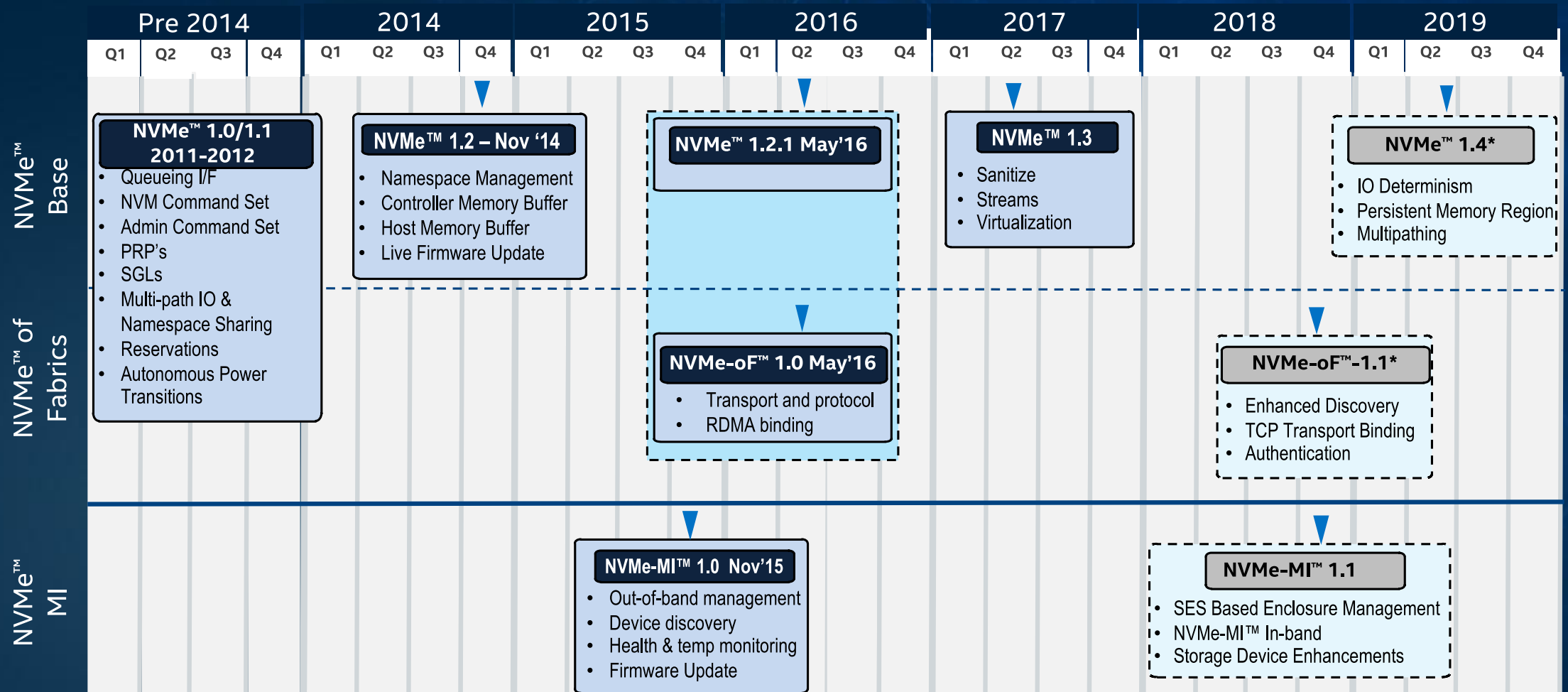


Industry's first definition of standard SR-IOV driver across vendors.

IDF16
INTEL DEVELOPER FORUM

25

LATEST NVMe ROADMAP



TODAY'S CHALLENGE QUALITY OF SERVICE AT SCALE

Facebook @ Scale

facebook Community Update 7.26.2017
Bringing the world closer together

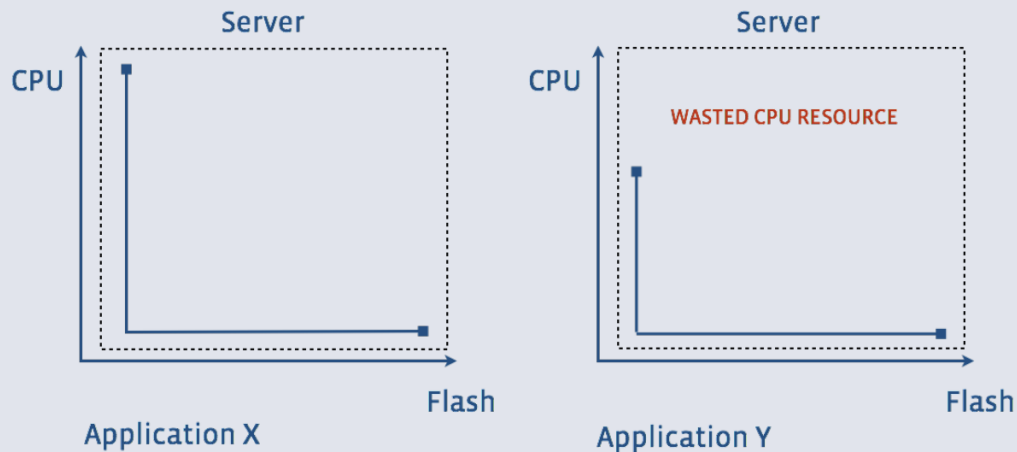
Disaggregated Flash

Applications change over time

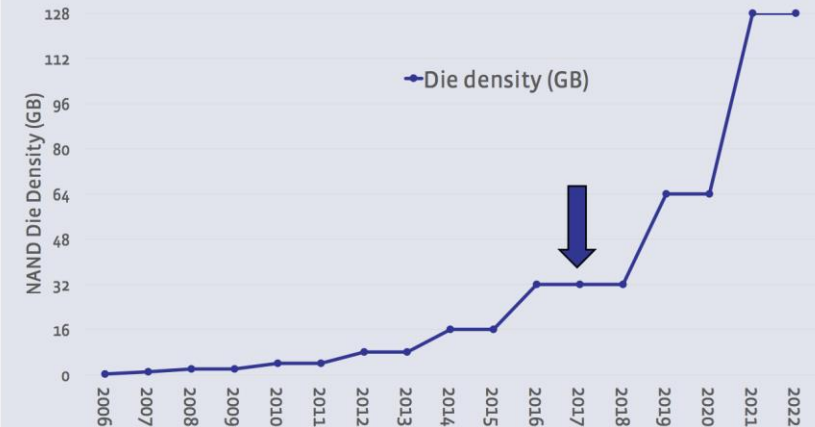


Disaggregated Flash

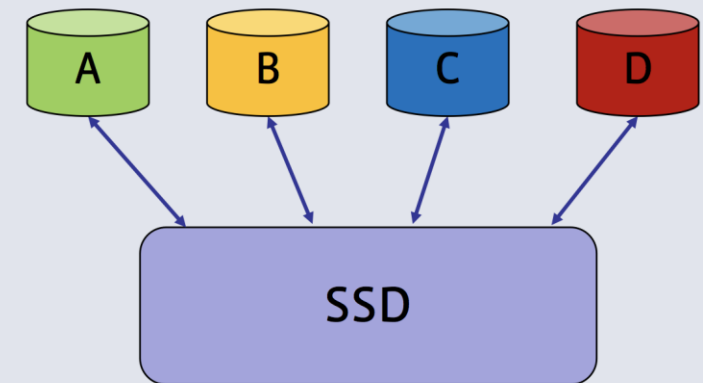
Applications have different needs



NAND Flash Trend

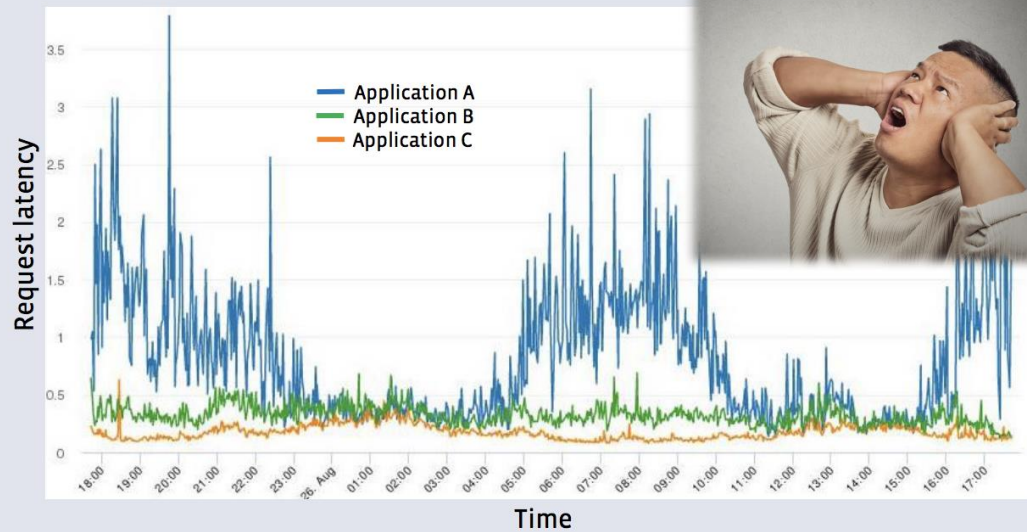


SSD Capacity = Shared Resource



TODAY'S CHALLENGE QUALITY OF SERVICE AT SCALE

Noisy Neighbors



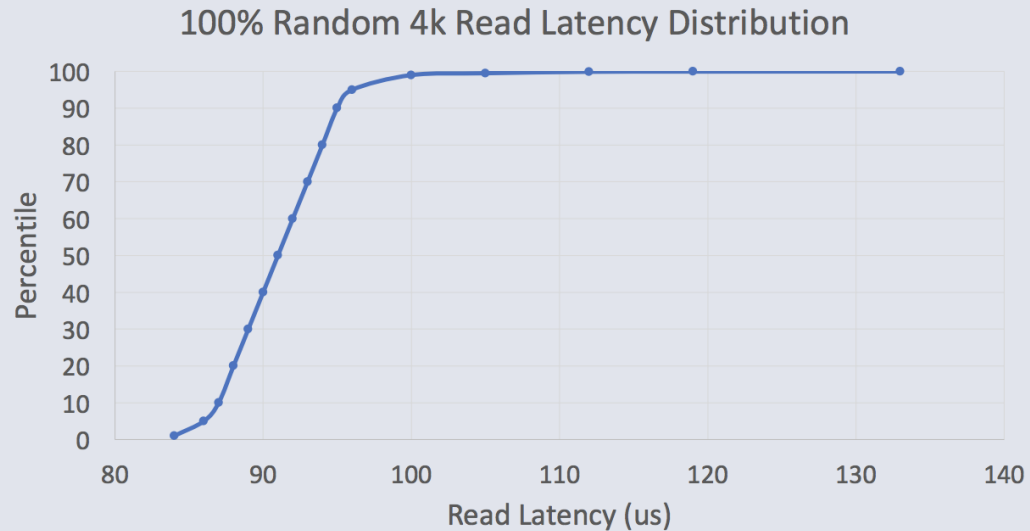
Latency vs. Bandwidth



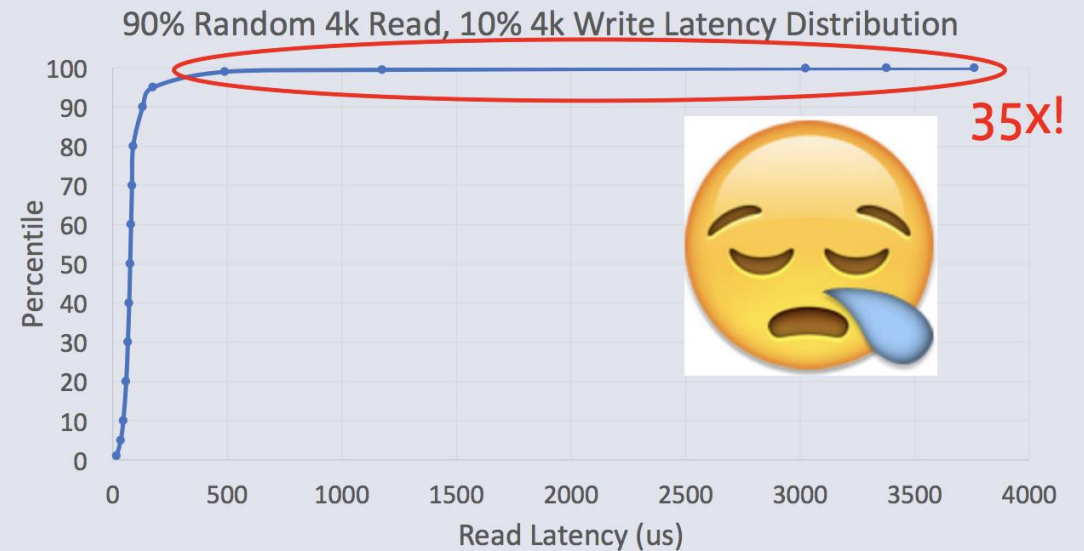
Image source: pixabay.com

TODAY'S CHALLENGE QUALITY OF SERVICE AT SCALE

Read Latency Challenge



Read Latency Challenge



CALL TO ACTION

SOLVE QOS AT SCALE

There are many approaches to solve quality of service at scale.
Collectively, embrace **ONE** path forward and **SCALE** the solution.

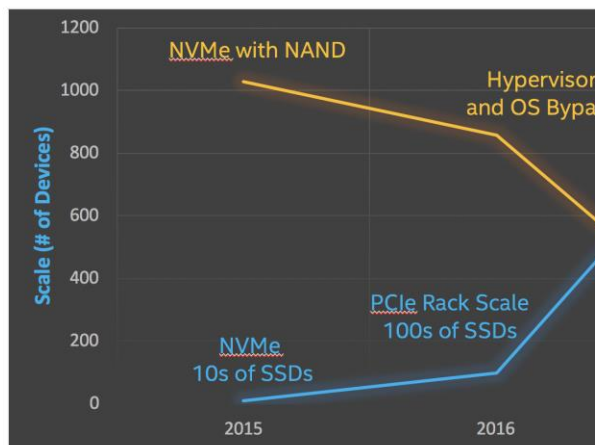
SCALING THE NUMBER OF SSDS

ENTER FABRICS

WHY NVMe OVER FABRICS

The EMC Perspective

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



The Need to Extend NVM Express™ Over Fabrics

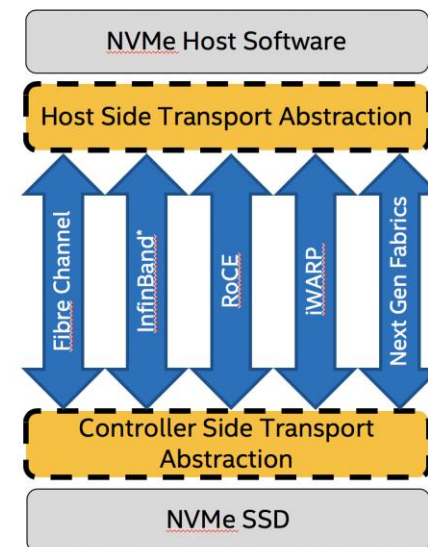
- PCI Express® ideal for in-server and in-rack, but difficult to scale beyond 100's of nodes:
 - Address routing rather than endpoint routing
 - Want to converge storage + networking at scale
 - Want to leverage standard switch infrastructure
- Existing Fabric interface (e.g., iSER / SRP) ecosystem is 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 n



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

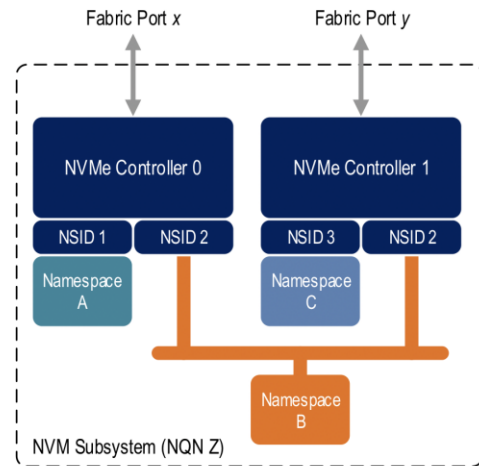


Standard abstraction layer enables NVMe across range of Fabrics

ARCHITECTURE 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

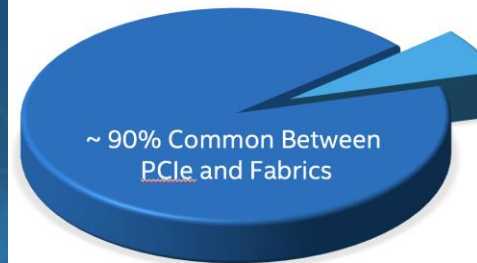


IDF15
INTEL DEVELOPER FORUM

27

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

IDF15
INTEL DEVELOPER FORUM

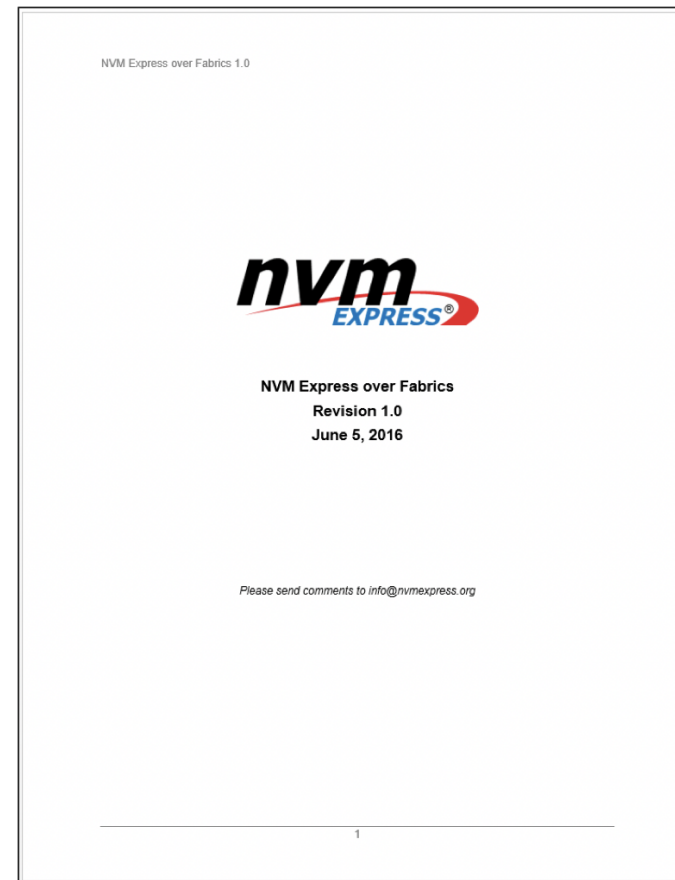
28

NVME OVER FABRICS 1.0 DELIVERED IN 2016

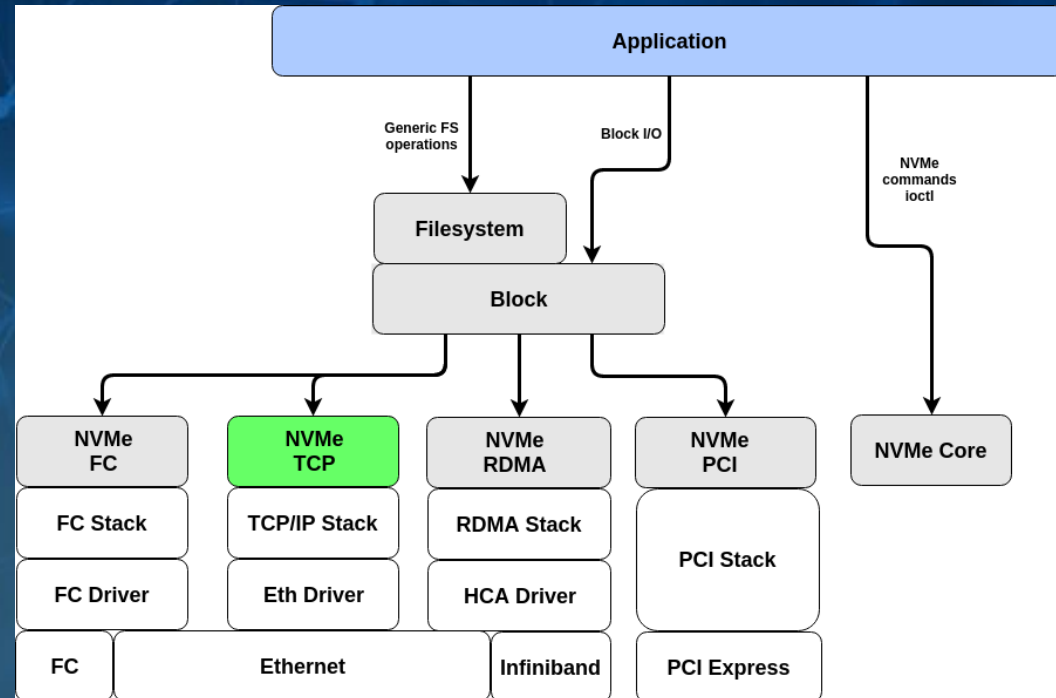
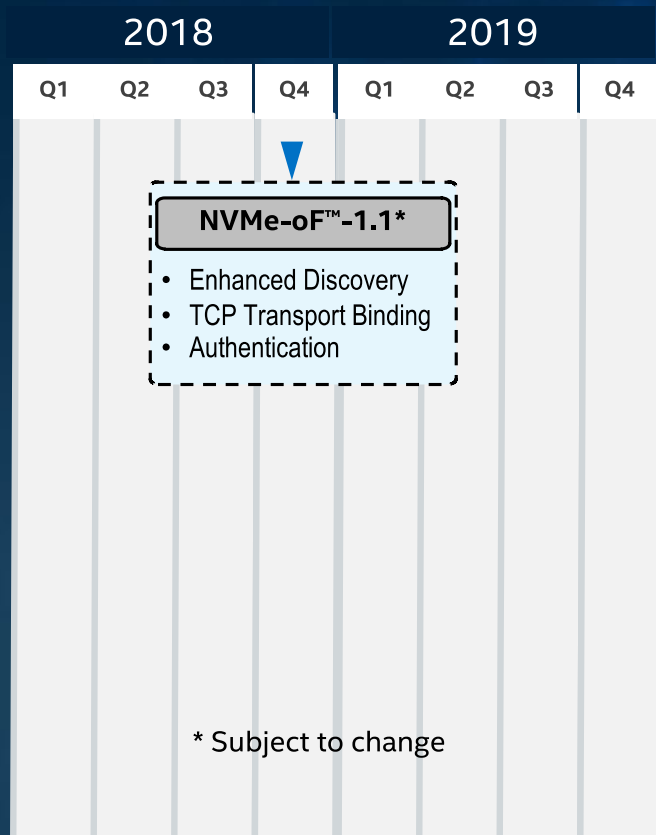
Solutions Coming Soon

- Revision 1.0 published on June 5
- Linux* host and target driver published simultaneously
 - More than 20 companies participated
- Proof of concepts shown by ~10 companies with products imminent
 - Showing << 10 μ s latency added

Get ready for NVM Express* over Fabrics products appearing later this year!

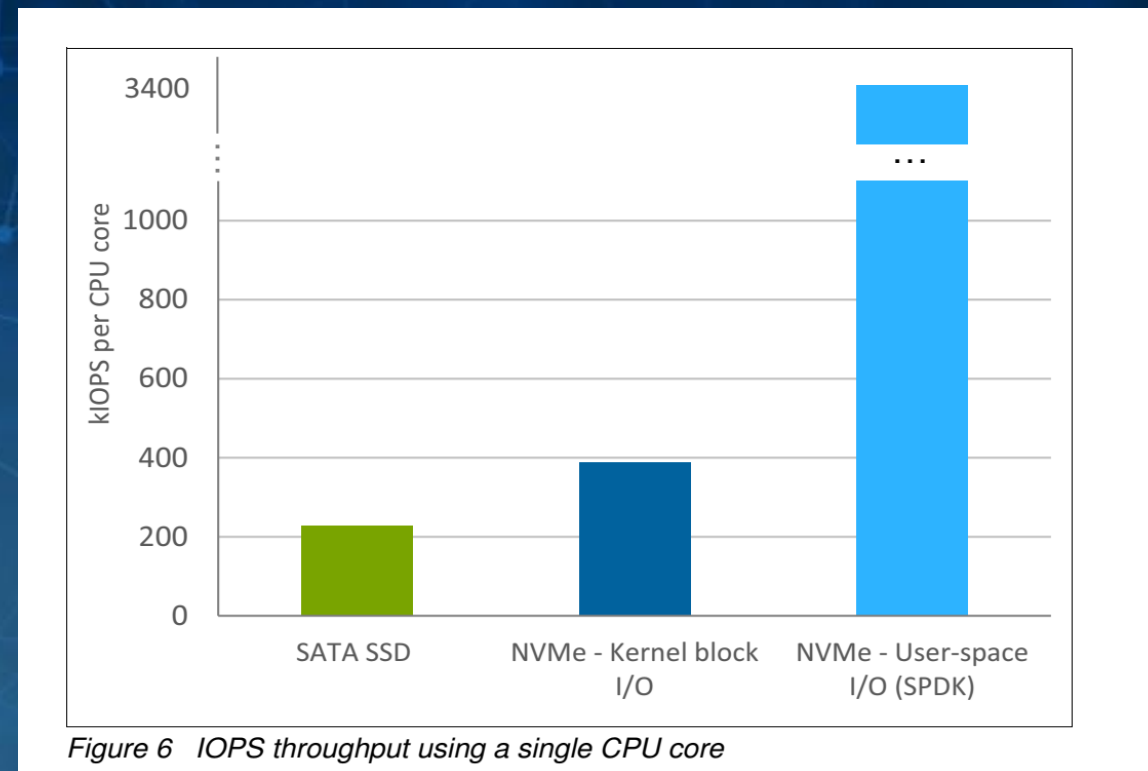
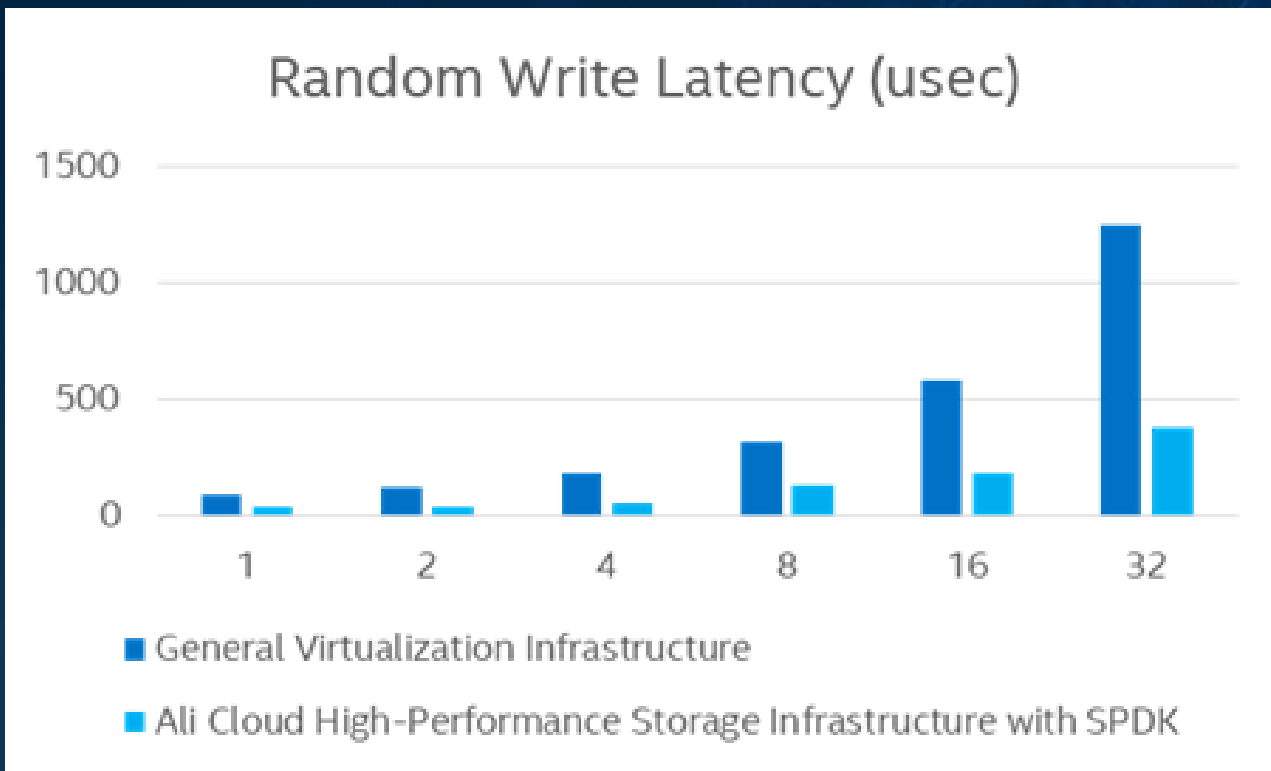


INTRODUCING **TCP** AS A TRANSPORT



Alternative to iSCSI to realize full benefits of NVMe end to end.

SOFTWARE PLAYS KEY ROLE



SPDK provided a boost with more IOPS/Core and less overhead.

CONNECTORS, FORM FACTORS, OH MY...

ENABLING 2.5" PCIe* SSD, NOW KNOWN AS U.2

SSD Form Factor Working Group *Driving Enterprise SSD Infrastructure*

Form Factor



Benefit from current 2.5" HDD form factor

Expand power envelope

Connector



Support Multiple Protocols

- PCI Express* 3.0, SAS 3.0, SATA 3.0

Management Bus

Dual port (PCIe)

Multi-lane capability (PCIe/SAS)

SAS & SATA Backward Compatibility

Hot-Plug



Hot-Plug Connector

Identify desired drive behavior

Define required system behavior

IDF2011
INTEL DEVELOPER FORUM

23

Enterprise Connector Status

- The SSD Form Factor Workgroup is focused solely on Enterprise SSD infrastructure
 - Five Promoters direct effort: Dell*, EMC*, Fujitsu*, Intel, and IBM*
 - 45 members contribute to the definition
- 2.5" Form Factor specification released
 - Rev 0.85 released to Workgroup members on 9/6
 - Previous release (0.72) released to Workgroup members on 2/11
- Drive Connector Pin-out specification released
 - Revision 1.0 in March '11 and revision 1.1 in May '11
 - Submitted to SFF to actively drive industry alignment for "one connector" for PCI Express* and SAS, used to create SFF-8639

Get involved – visit www.ssdformfactor.org

IDF2011
INTEL DEVELOPER FORUM

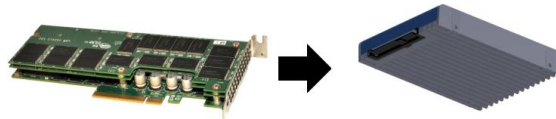
24

U.2 TAKES SHAPE

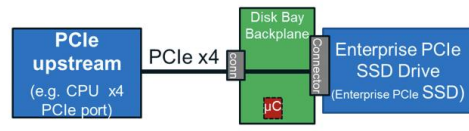


SFF-8639 Brings Full Storage Capabilities to Enterprise

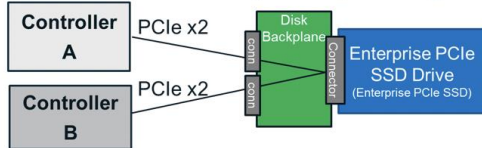
- SFF-8639 brings a 2.5" pluggable form factor to the Enterprise
- For Enterprise PCIe SSDs, this includes support for a typical server and storage configuration
- Server: Single x4 PCIe SSD
- Storage: High availability dual ported solution



Typical Server configuration



Typical High Availability Storage configuration



Flash Memory Summit 2013
Santa Clara, CA

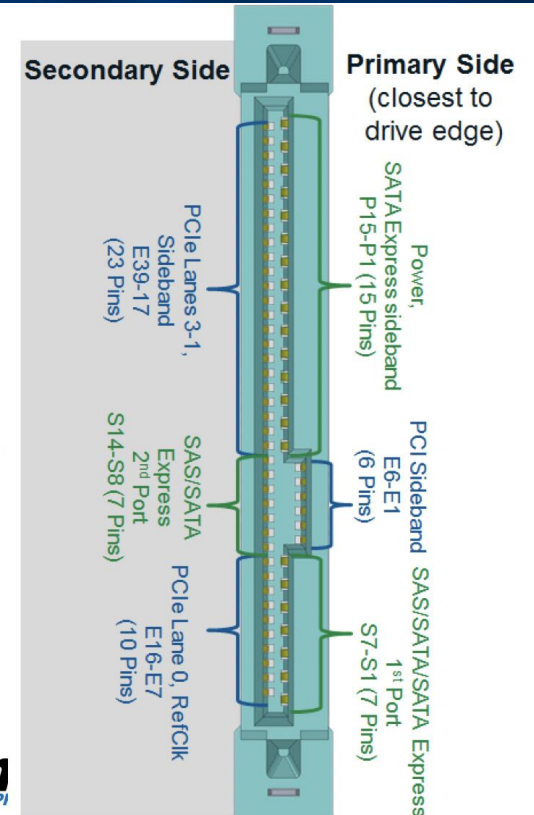


27



SFF-8639 Flexibility

- SFF-8639 supports:
 - Enterprise PCIe x4 SSDs
 - Existing SAS drive (dual port)
 - Existing SATA drives
- As ecosystem develops:
 - Client 2.5" PCIe (often referred to as SATA Express)
 - x4 SAS
- Supports flexible backplanes
 - Enterprise x4 PCIe SSDs
 - SAS/SATA HDDs



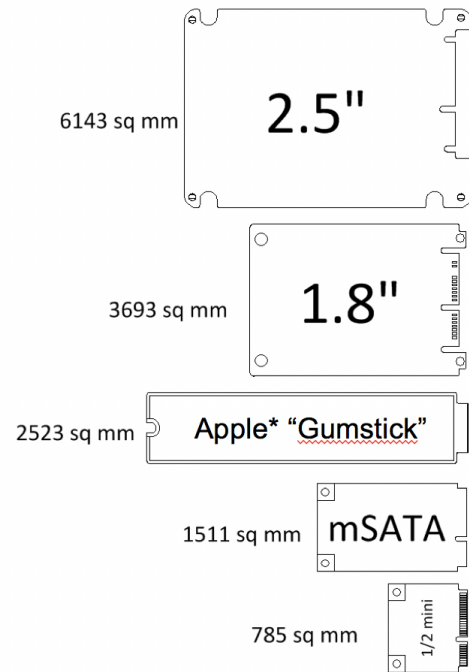
Flash Memory Summit 2013
Santa Clara, CA



INVENTING M.2

Client Form Factors

- The outline of each client SSD used today is shown to the right
- 2.5" and 1.8" are cased form factors, drop-in HDD replacements
- SSD vendors are being pushed for "one off" custom form factors leading to increased cost
 - e.g. gumstick design
- A standard optimized caseless SSD form factor is needed



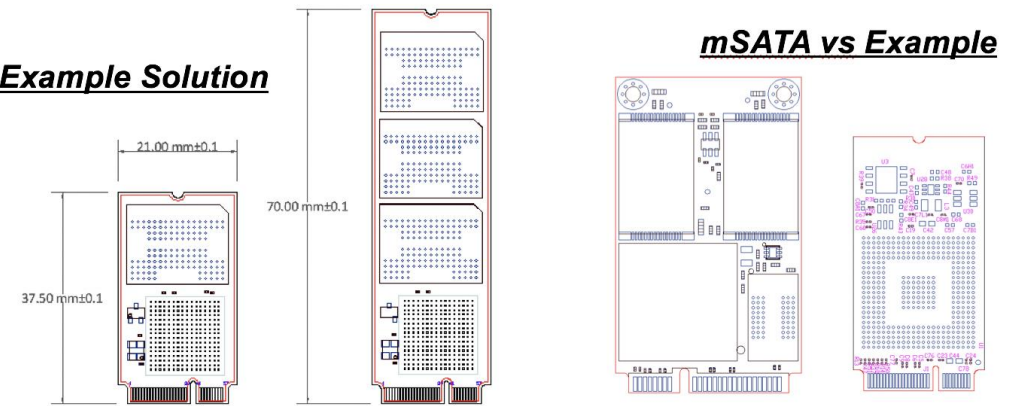
IDF2011
INTEL DEVELOPER FORUM

28

An Optimized Caseless Form Factor

- Attributes to strive for in a new standardized caseless SSD FF:
 - Scalable from small to large capacity points
 - Support SATA Gen3 and two lanes of PCI Express* Gen3
 - Optimize for Z height (e.g. board edge connector, reduce PCB thickness)
 - Mounting strategy will limit board area and reduce fasteners
 - Optimize board size based on BGA NAND package & ensure efficient tiling

Example Solution



Watch for new standard caseless form factor effort

IDF2011
INTEL DEVELOPER FORUM

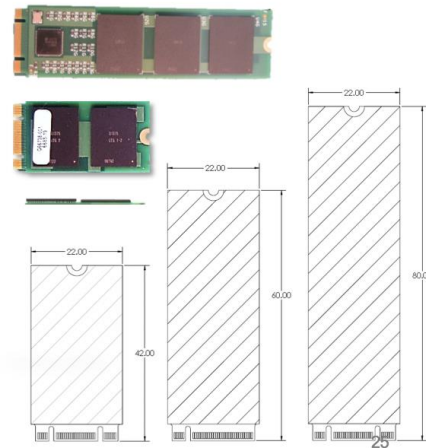
29

M.2 EMERGES IN CLIENT



M.2 Emerging as Primary Client Form Factor

- In client, as SSDs move first to PCIe, OEMs are using the optimized M.2 form factor
- As native OS support of NVMe becomes pervasive, OEMs will move from AHCI to NVMe to take full advantage of PCIe



VAIO* Pro 13 Ultrabook™
The world's lightest 13.3" touch Ultrabook²¹.

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.

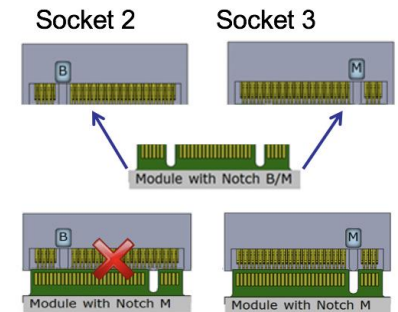


*Other names and brands may be claimed as the property of others.



M.2 Provides OEM Choice: Max Performance or Flexibility

- Three families of modules:
 - Socket 1: Wi-Fi/Connectivity only
 - Socket 2: WWAN, Storage (SATA*, PCIe* x1, PCIe* x2), other
 - Socket 3: Storage only (SATA*, PCIe* x1, PCIe* x2, PCIe* x4)
- OEMs choose the socket to include
 - Socket 2: Most flexibility
 - Socket 3: Highest performance




With M.2, client OEMs can choose maximum performance with 4 lanes, or they can choose flexibility with SATA and WWAN options.





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

BUT ... M.2 AND U.2 KEEP US IN THE “LEGACY BOX”

 Flash Memory Summit

The Dilemma of Defining The System of Tomorrow – Today

- Typical design point is 2 socket, 1U server
- Configurability is Critical
 - Needed today does not mean needed tomorrow
 - More stranded IOs = Opportunity lost, wasted \$\$\$
 - Ideal scenario: All **precious** IOs are utilized
- New technologies (e.g., FPGAs) increase the challenge



Santa Clara, CA
August 2018

2

 Flash Memory Summit

Challenges to Address



- **Need More NVM Sites**
less packages/SSD = more dies/package = lower yield/package
- **Support SSDs and MORE**
Legacy connectors have been SSD only.
- **Optimize for NVM**
Legacy form factors in Enterprise and Datacenter based on HDDs or client SSDs.
- **Thermals and TCO Matter**
Legacy SSDs not thermally optimized. Airflow to CPU restricted.

Santa Clara, CA
August 2018

3

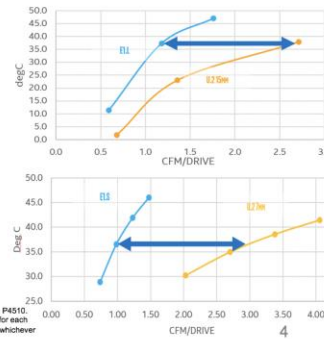
INVEST FOR THE FUTURE WITH EDSFF



Enter EDSFF 😎

Flash Memory Summit

- General purpose scalable connector
 - Flexible: Multiple orientations, widths, PCIe* Gen 5+ support
 - Supports interoperable specs (EDSFF, OCP Mezz, Gen Z, etc.)
- Break free of legacy to optimize for NVM
 - 50-100% increase in media package sites
- Improved thermal efficiency
 - 2-3x less airflow needed
 - Or support higher power devices



Santa Clara, CA
August 2018

Source - Intel, Comparing airflow required to maintain equivalent temperature of a 4TB U.2 15mm Intel® SSD DC P4500 to a 4TB "1U-L" form factor for Intel® SSD DC P4500. Results have been estimated or simulated using internal analysis or architectural simulation or modeling, and provided for informational purposes. Simulation involves three drives for each form factor in a sheet metal representation of a server, 12.5mm pitch for "Ruler" form factor, 1500m elevation, limiting SSD on case temp of TCC or thermal throttling performance, whichever comes first. SC guard band. Results used as a proxy for airflow anticipated on EDSFF spec compliant "Ruler" form factor Intel® SSD P4510.



Scalable Family for Different Usages

Flash Memory Summit



- E1.L (SFF-TA-1007)**
- 318.75 x 38.4 mm
- Supports > 40W
- Up to 48 Standard NAND sites



- E1.S (SFF-TA-1006)**
- 111.5 x 31.5 mm
- Supports >12W
- Up to 12 Standard NAND sites



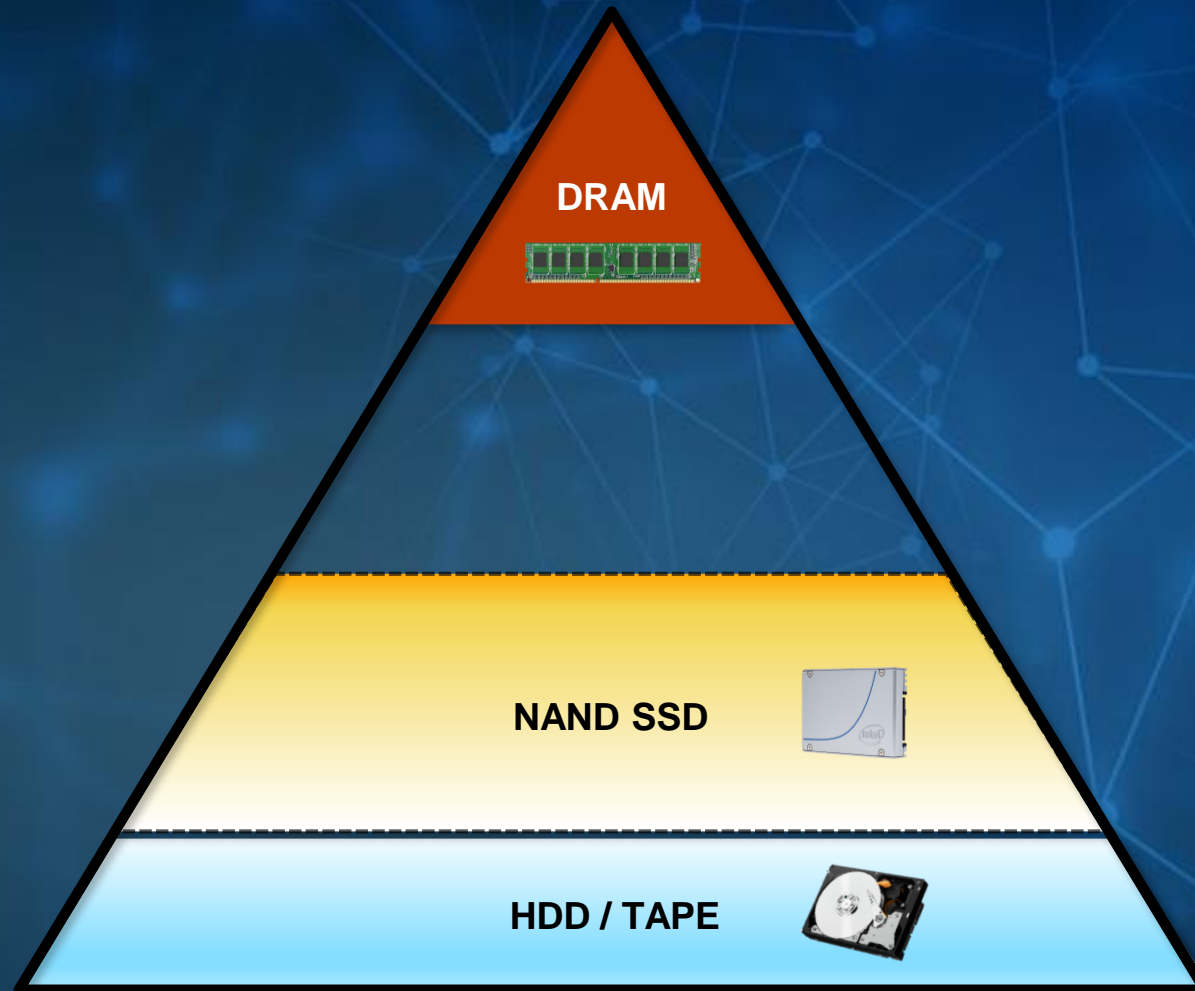
- E3 (SFF-TA-1008)**
- (104.9/142.2) x 76mm
- Supports up to 70W
- Up to 48 Standard NAND sites

- Same Protocol: NVMe
- Same Interface: PCIe
- Same Connector: SFF-TA-1002
- Same Pinout and Functions (hot plug, serviceable)
- Different Usages, Same Expectations!

Santa Clara, CA
August 2018

5

WE DELIVERED NAND SSD TIER AT SCALE, **TOGETHER**



FUTURE NVM IS NOW

FOCUS OF NVMe – ENABLE FUTURE NVM

Scalability for Future NVM

- NVMe* is defined to scale for future NVM
 - Host controller standards live for 10+ years
 - Future NVM may have sub microsecond latencies
- 1M IOPS needs highly efficient driver approach
 - Benefits from removing OS queues, IO scheduler, and SCSI layer while optimizing for NVMe
- Block layer attach reduces overhead > 50%
 - Block layer: 2.8 μ s, 9100 cycles
 - Traditional: 6.0 μ s, 19500 cycles

Chatham NVMe Prototype



Prototype Measured IOPS



Cores Used for 1M IOPS



Linux * Storage Stack

User Apps

User
Kernel

VFS / File System

Block Layer

Req Queue

SCSI Xlat

NVMe Driver

SAS Driver

2.8 μ secs

6.0 μ secs

IDF2011
INTEL DEVELOPER FORUM

15

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

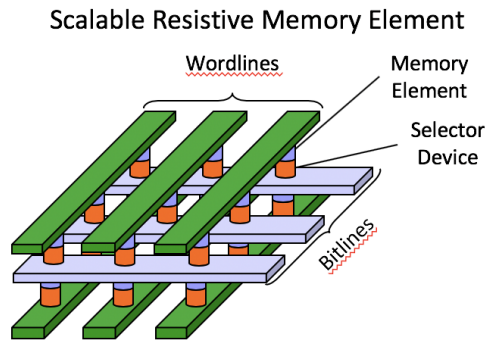
PREPARING THE WAY FOR AN NVM BREAKTHROUGH

Flash Memory Summit 2013



Next Generation Scalable NVM

Resistive RAM NVM Options



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

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

Many candidate next generation NVM technologies.
Offer $\sim 1000x$ speed-up over NAND, closer to DRAM speeds.

Flash Memory Summit 2013
Santa Clara, CA

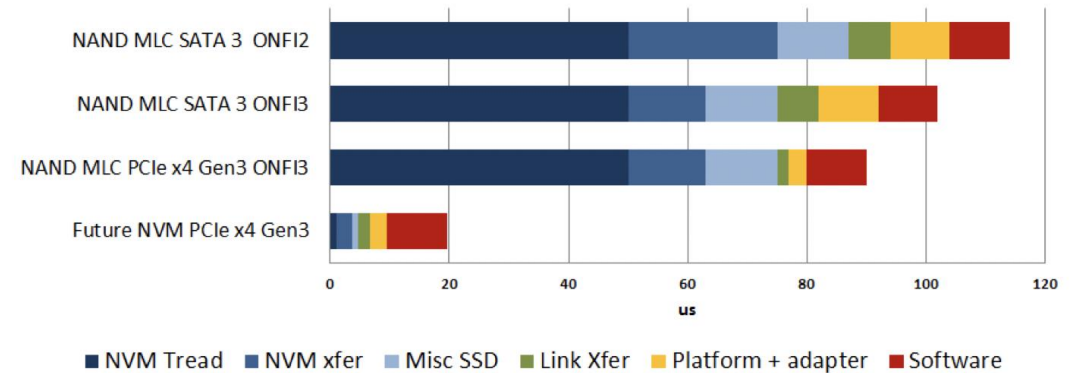


4



Fully Exploiting Next Gen NVM Requires Platform Improvements

App to SSD IO Read Latency (QD=1, 4KB)



- With Next Gen NVM, the NVM is no longer the bottleneck
 - Need optimized platform storage interconnect
 - Need optimized software storage access methods

Flash Memory Summit 2013
Santa Clara, CA



5

For full SSD benefits, must architect for NVM from ground up.

THE BREAKTHROUGH

3D XPoint™ Technology

IDF16
INTEL DEVELOPER FORUM

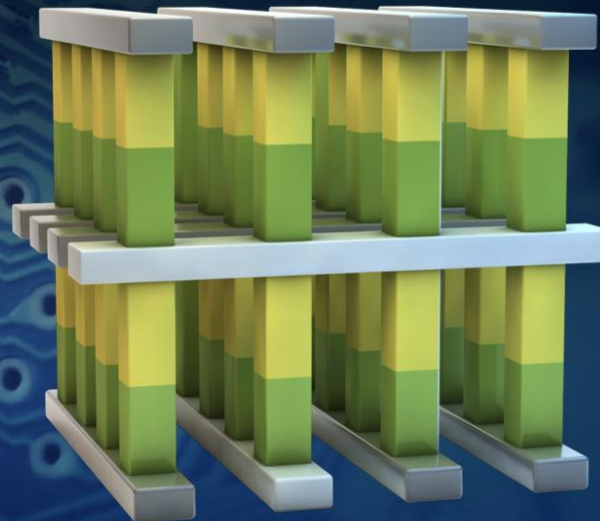
Cross Point Structure

Selectors allow dense packing and individual access to bits



Scalable

Memory layers can be stacked in a 3D manner



Breakthrough Material Advances

Compatible switch and memory cell materials



High Performance

Cell and array architecture that can switch states 1000x¹ faster than NAND

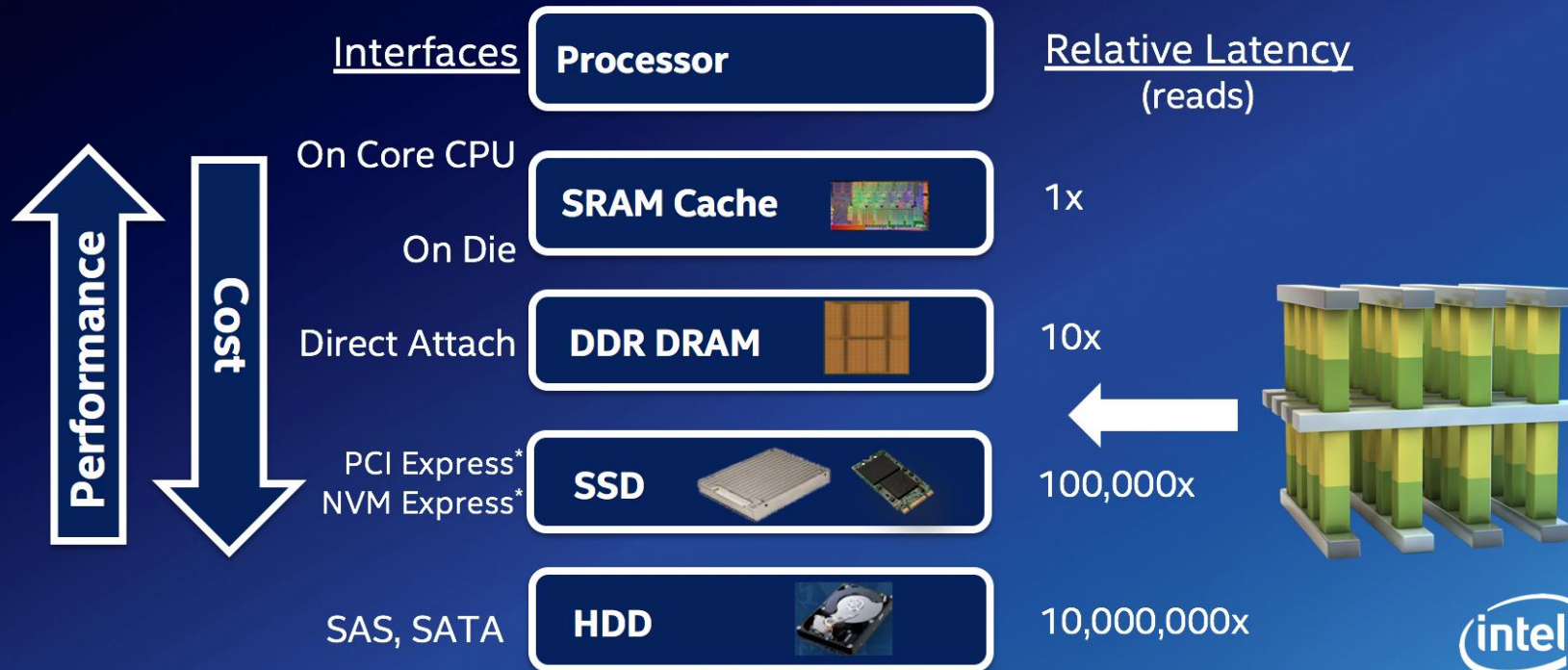


¹Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

REIMAGINING THE HIERARCHY

Memory and Storage Platform Connection

IDF16
INTEL DEVELOPER FORUM

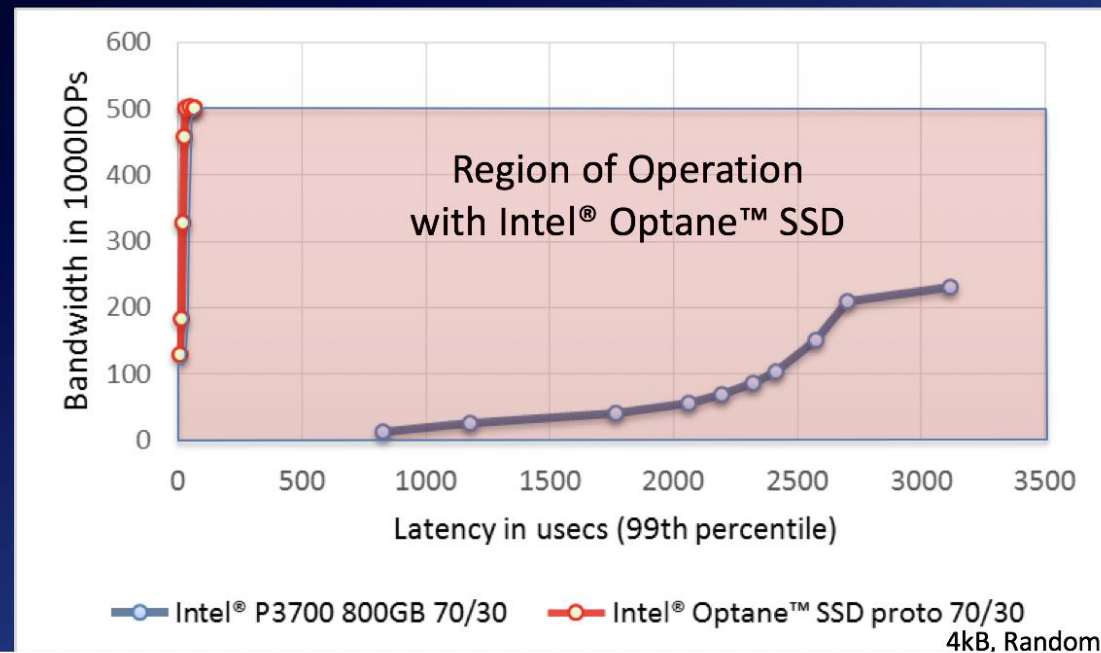


¹Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

WITH PREVIOUSLY UNREACHABLE PERFORMANCE

SSD Performance

IDF16
INTEL DEVELOPER FORUM



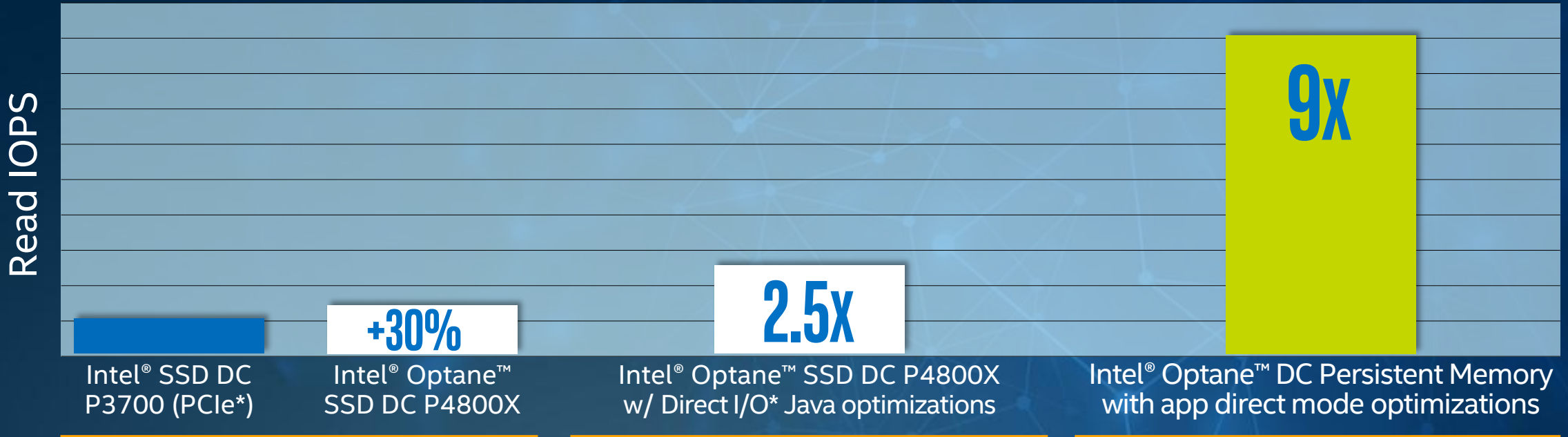
Intel® Optane™ SSDs change the game, operating at previously unreachable IOPs/Latency combinations



¹⁷ Config: I7-6700K Turbo to 4.3GHz, ASUS Z170m-plus 4x4GB DDR-2133, Hyperthreading disabled CPU C-state disabled, Ubuntu 14.04 LTS 64 bit server, kernel 4.4 (polling enabled), FIO 2.1.11

PROOF POINT: CASSANDRA 4.0* DATABASE

IOPS performance vs. Comparable Server System with DRAM and NAND SSD



1 DEPLOY

available Intel® Optane™ DC SSDs

2 OPTIMIZE

with available software tools

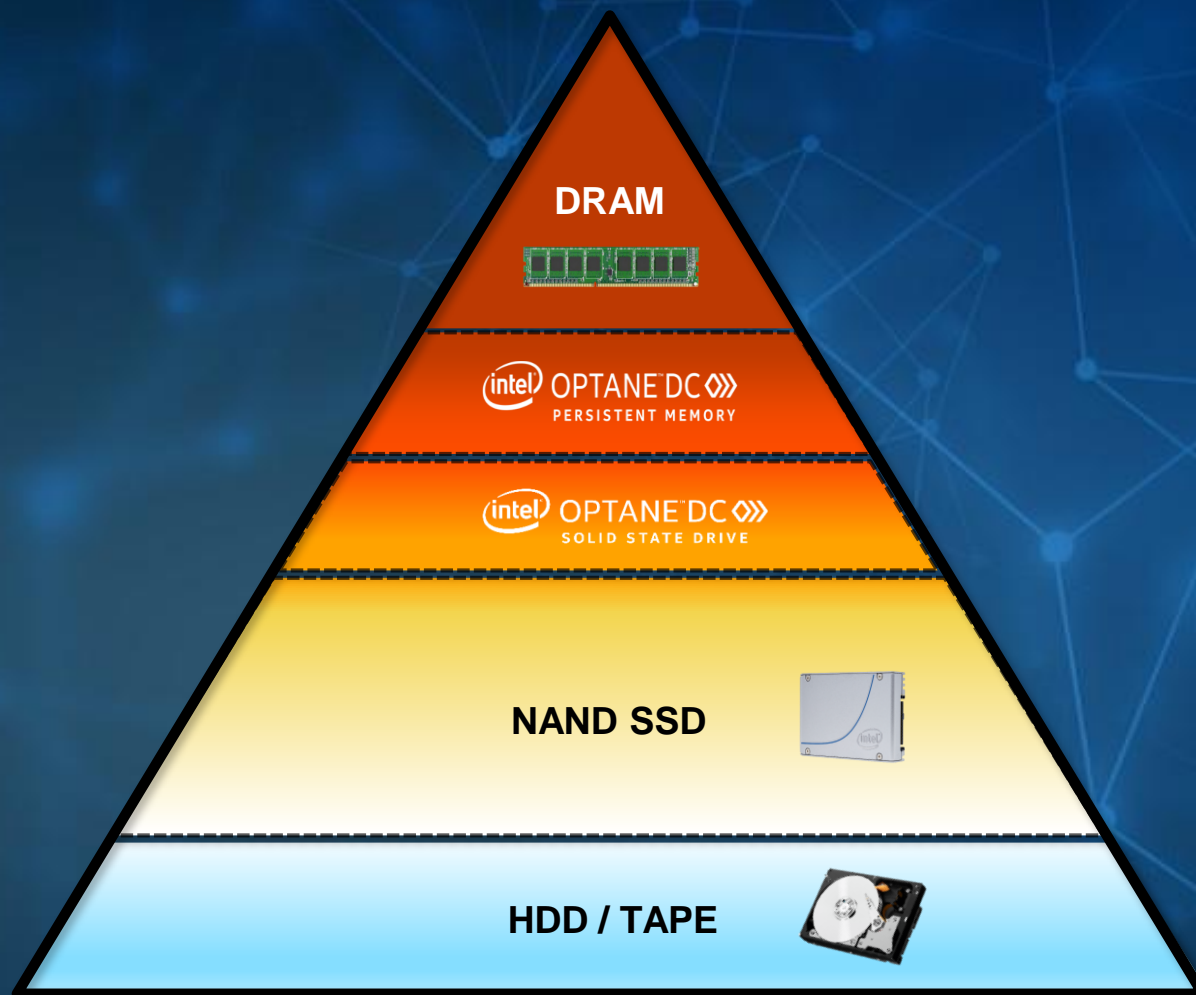
3 EVOLVE

with next-generation memory technology

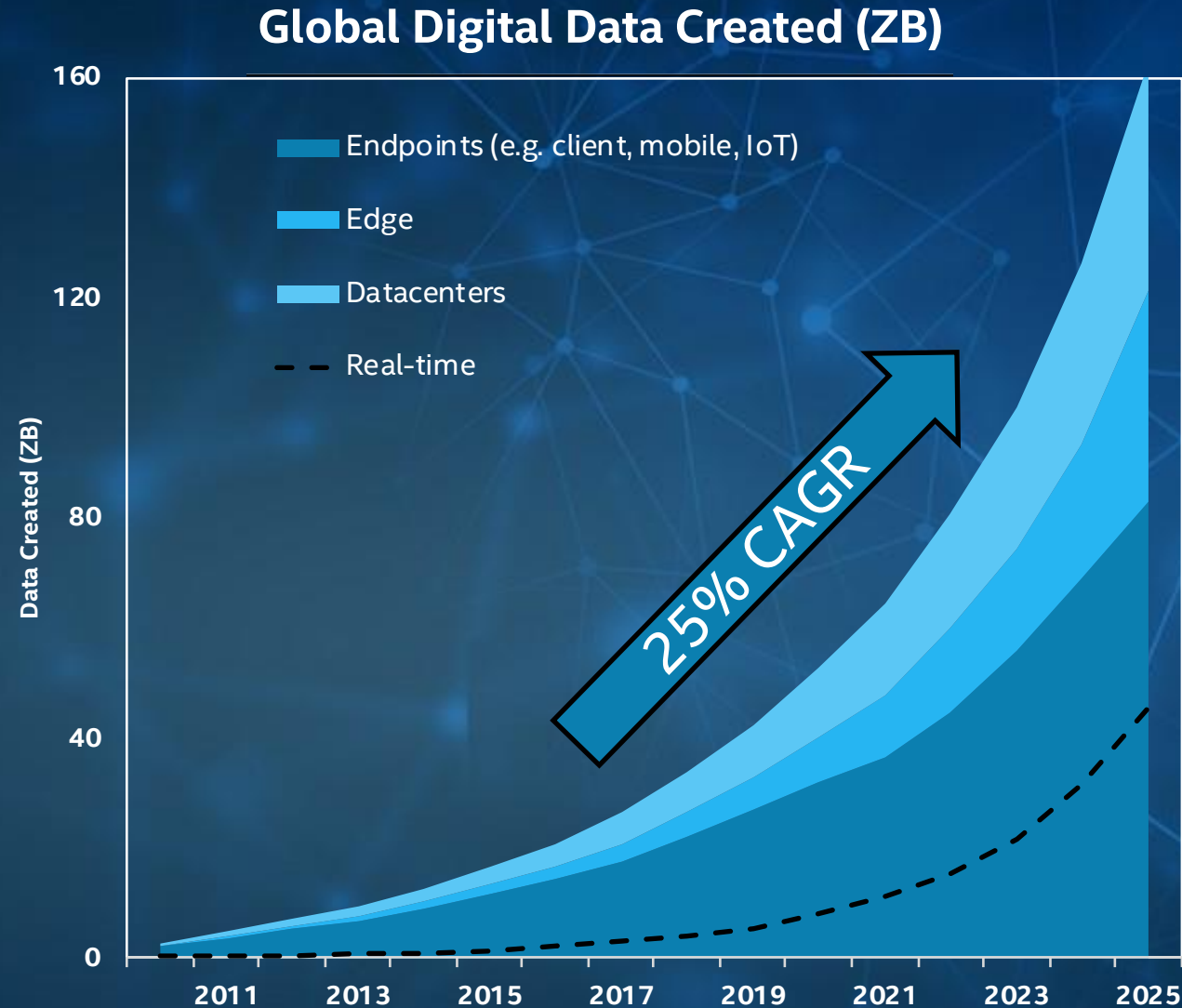
*System configuration: Server model: 2x Intel® Xeon® E5 2699 v4 @ 4. Ghz, Intel system board S2600WFWF, 384GB DDR4 @ 2667Mhz, 4x Intel® Optane DC SSD 375GB; CentOS 7.3.1611 (kernel 4.17.6), Network is 10GbE. Apache Cassandra version 4.0-SNAPSHOT (DirectIO from Intel-based Java DirectIO Development team). Cassandra-stress tool used for benchmarking embedded into the Cassandra version build 4.0. Java heap size 64GB, Java Garbage collector G1GC, Java Version Oracle JDK 10.01 that embeds with Cassandra. Experimental release used for Optane Persistent Memory based system. Baseline nvme NAND Intel Drives – Intel SSD DC P4510. Baseline consists of Operating System OS page cache (not DirectIO) and best methods per Datastax and lead Companies of the Apache Cassandra Open Source version Project Management Committee. Performance results are based on testing as of July 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

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

THE EMERGING HIERARCHY



REMEMBER THE EVER INCREASING ZETABYTES



WE HAVE INVENTED THE FUTURE **TOGETHER**

LET'S CONTINUE **OUR WORK** OVER THE NEXT DECADE

