THE SOLID STATE TRANSFORMATION OF THE DATA CENTER

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

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](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](http://www.intel.com/benchmarks).

Intel's compilers 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)

- Endpoints (e.g. client, mobile, IoT)
- Edge
- Datacenters
- Real-time

25% CAGR

Source: IDC Data Age 2025 (2017)
TODAY’S TALK

1. OUR INDUSTRY TRANSFORMATION TO EMBRACE NVM

2. OUR CONTINUED OPPORTUNITIES TOGETHER
THE JOURNEY BEGINS ~ 2000

DRAM

HDD / TAPE
It was a question if SSDs would be a killer application for NAND.
NAND SSD TIER?
THE NEED FOR
OPEN NAND FLASH INTERFACE
There were many vendors, yet no standard interface, making it difficult to design SSDs.
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+ µs “hiccup” waiting for the interface bottleneck during every 4KB read.

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

ONFI 2.0 triples the legacy interface speed. ONFI 2.1 with more speed targeted for 2H’08.

ONFI defined 10x scaling of NAND interface in < 2 years.
ENABLING SCALE FOR SSD INDUSTRY

NAND Flash Revenue, ($M)

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

ONFI / JEDEC collaboration enabled unifying the industry to support scale of SSD ambitions.
ONFI Workgroup Continues To Produce Results!

<table>
<thead>
<tr>
<th>Major Revisions</th>
<th>ONFI 1.0: Standard electrical &amp; protocol interface, including basic command set</th>
<th>ONFI 2.0: Defined a high speed DDR 4 interface, including the traditional NAND bus speed in common use</th>
<th>ONFI 2.x: Additional features and support for bus speeds up to 200 MB/s</th>
<th>EZ NAND / ONFI 2.3: Enhanced ECC management feature</th>
<th>ONFI 3.0: Scaled high speed DDR 4 to 400 MT/s</th>
<th>ONFI 3.x: Scaled high speed DDR 4 to 533 MT/s</th>
<th>ONFI 4.0: Scaled high speed DDR 4 to 800 MT/s, Reduce power support to 1.2V (HV-DDR4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>50 MB/s</td>
<td>133 MB/s</td>
<td>200 MB/s</td>
<td>400 MB/s</td>
<td>533 MB/s</td>
<td>800 MB/s</td>
<td>ONFI – JEDEC Collaboration</td>
</tr>
</tbody>
</table>

Steady innovation...

ONFI has and continues to deliver innovation & interoperability enabling faster NAND adoption
KEEPING PACE WITH SSD NEEDS

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

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

- IP-PAES
- PHISON
- Microsoft
- SYNOPSYS
- INTEL
- MARVELL
- AMD
- GENESYS
- SANDISK
- AGIRA TECH
- IDT
- Hitachi
- Micron
- NVIDIA
- ULINK
- LENALI
- Mobile Semiconductor

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

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
NVMHCI FOR CLIENT WAS A MISFIRE
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
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”

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

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

0.70 revision achieved, available for Contributor review. Schedule enables product intercept in 2012.
Bucket 1: Eliminate performance bottlenecks seen in other interfaces.
- Do not require dedicated bus lanes
- Eight of 32 channels can be managed
- Support for many core systems

Bucket 2: Provides streamlining of command delivery
- End-to-end data protection
- (i.e., T10 DIF / DIX functionality)
- Firmware updates
- Encryption
- Comprehensive
- Health status reporting
- Robust error reporting

Bucket 3: Provides Enterprise features
- 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

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

**NVM Express** “BORN” IN MARCH 2011

NVMe*: Efficient SSD Performance

<table>
<thead>
<tr>
<th></th>
<th>AHCI1</th>
<th>NVMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncacheable Register Reads</td>
<td>4 per command</td>
<td>0 per command</td>
</tr>
<tr>
<td>Each consumes 2000 CPU cycles</td>
<td>8000 cycles, ~ 2.5 µs</td>
<td></td>
</tr>
<tr>
<td>MSI-X and Interrupt Steering</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Ensures one core not IOPs bottleneck</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallelism &amp; Multiple Threads</td>
<td>Requires synchronization lock to issue command</td>
<td>No locking, doorbell register per Queue</td>
</tr>
<tr>
<td>Ensures one core not IOPs bottleneck</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Queue Depth</td>
<td>32</td>
<td>64K Queues</td>
</tr>
<tr>
<td>Ensures one core not IOPs bottleneck</td>
<td></td>
<td>64K Commands per Queue</td>
</tr>
<tr>
<td>Efficiency for 4KB Commands</td>
<td>Command parameters require two serialized host DRAM fetches</td>
<td>Command parameters in one 64B fetch</td>
</tr>
<tr>
<td>4KB critical in Client and Enterprise</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*AHCI: Serial ATA programming interface. See http://www.intel.com/technology/serialata/ahci.htm*
Interoperability Program Underway

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

- 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.
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

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

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
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 Halts First Plugfest

WAKEFIELD, Mass., May 29, 2013 -- The NVM Express Workgroup, developer of the NVM Express specifications for accessing solid-state drives (SSDs) on a PCIe 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 request 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 Adesto Technologies, Dell Inc., F ouriert Systems, Inc., Hystor, a Western Digital company, Integrated Device Technology, Inc., Intel Corporation, Samsung Electronics Co., Ltd., SandDisk Corporation, Inc., Toshiba, Inc., and Western Digital Corporation.

Samsung Announces Industry’s First 2.5-Inch NVMe SSD

Samsung has announced the 93070H, a 2.5-inch Non-Volatile Memory Express (NVMe) PCIe SSD. According to Samsung, the 1.8TB 800GB NVMe SSD provides a sequential read speed at 3.8 GB/s, six times faster than the company’s current high-end enterprise SSD. The 1201050:1 sequential read performance is specified at up to 768 MB/s, more than 10 times as fast as existing SSD options.

NVMe products targeting Datacenter shipping this year

IDF13
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

And Best in Class Sequential Performance

- NVMe 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
The Efficiency of NVM Express® (NVMe)

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

![Efficiently and with low latency](image)

The Latency of NVM Express® (NVMe)

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

![Efficiently and with low latency](image)
WITH BETTER QUALITY OF SERVICE

NVMe™ Delivers Higher IOPs and Better QoS

Interface Latency, 4KB Random Reads

AHCI > 300 μs @ ~150K IOPS
SAS HBA with 8 SATA* SSDs > 300 μs @ ~ 400K IOPS
NVMe single drive
Average = 18 μs
99.99% = 40 μs
@ ~ 800K IOPS

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 S2600WT with 2x E5-2685 CPUs, 2 sockets, 2.3 GHz clock speed per CPU, Ubuntu 14.04.1 LTS (64-bit), Linux 3.15.0-rc7 (x86_64), iometer 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 x threads. Drives tested empty to test interface only (no NVMe access).
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
  - NVM 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

Data Center Interface Dynamics, One Level Deeper

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

Source: IDC

PCle SSDs lead the way by embracing industry standards.
NVM Express* (NVMe)
Development Timeline

NVMe 1.0 – Mar 2011
- Queuing Interface
- Command Set
- End-to-End Protection
- Security
- PRPs

NVMe 1.1 – Oct 2012
- Multi-Path IO
- Namespace Sharing
- Reservations
- Autonomous Power Transition
- Scatter Gather Lists

NVMe 1.2 – Q4 2014
- Host Memory Buffer
- Replay Protected Area
- Active/Idle Power and RTD3
- Temperature Thresholds
- Namespace Management
- Controller Memory Buffer
- Live Firmware Update
- Atomicity Enhancements

NVM Express* (NVMe) revision 1.2 undergoing ratification now.
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.

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.

Intel Developer Forum 2014
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

Driving Cloud Innovation

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

Intel Developer Forum 2016
EVEN MORE CAPABILITIES ... NVMe REVISION 1.3

NVMe* Development Timeline

- **NVMe* 1.0 – Mar ‘11**
  - Base spec for PCIe*
  - Queuing Interface
  - Command Set
  - E2E Data Protection
  - Security

- **NVMe 1.1 – Oct ‘12**
  - Multi-Path IO
  - Namespace Sharing
  - Reservations
  - Autonomous Power Transition
  - Scatter Gather Lists

- **NVMe 1.2 – Nov ‘14**
  - Namespace Management
  - Controller Memory Buffer
  - Temperature Thresholds
  - Active/Idle Power & RTD3
  - Host Memory Buffer
  - Live Firmware Update

- **NVMe 1.3 – Q4’16**
  - Sanitize
  - Virtualization
  - Directives
  - Self-Test & Telemetry
  - Boot Partitions
  - And more...

**NVMe Management Interface 1.0 – Nov ‘15**

- Out-of-band management
- Device discovery
- Health & temp monitoring
- Firmware Update
- And more....

**NVMe over Fabrics 1.0 – June ‘16**

- Enable NVMe SSDs to efficiently connect via fabrics like: Ethernet, Fibre Channel, InfiniBand™, etc

NVMe is the place for innovation in SSDs.

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

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.
**LATEST NVME ROADMAP**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe™ Base</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe™ 1.0/1.1 2011-2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Queuing I/O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• NVM Command Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Admin Command Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• PRP’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• SGLs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Multi-path IO &amp; Namespace Sharing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Reservations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Autonomous Power Transitions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe™ 1.2 – Nov ‘14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Namespace Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Controller Memory Buffer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Host Memory Buffer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Live Firmware Update</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe of Fabrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe-oF™ 1.0 May’16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Transport and protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• RDMA binding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe™ MI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe-MI™ 1.0 Nov’15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Out-of-band management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Device discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Health &amp; temp monitoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Firmware Update</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe of Fabrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe-oF™-1.1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Enhanced Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• TCP Transport Binding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Authentication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe™ 1.3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sanitize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Streams</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Virtualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe™ 1.4+</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• IO Determinism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Persistent Memory Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Multipathing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NVMe-MI™ 1.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• SES Based Enclosure Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• NVMe-MI™ In-band</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Storage Device Enhancements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TODAY’S CHALLENGE** QUALITY OF SERVICE AT SCALE

**Disaggregated Flash**
Applications change over time

**Disaggregated Flash**
Applications have different needs

**SSD Capacity = Shared Resource**

Flash Memory Summit 2017, courtesy Chris Petersen
TODAY’S CHALLENGE QUALITY OF SERVICE AT SCALE

Noisy Neighbors

Latency vs. Bandwidth

Image source: pixabay.com

Flash Memory Summit 2017, courtesy Chris Petersen
Today’s Challenge: Quality of Service at Scale

Read Latency Challenge

100% Random 4k Read Latency Distribution

Read Latency Challenge

90% Random 4k Read, 10% 4k Write Latency Distribution

35x!

Flash Memory Summit 2017, courtesy Chris Petersen
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
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 it.

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

<table>
<thead>
<tr>
<th>Differences</th>
<th>PCI Express® (PCIe)</th>
<th>Fabrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>Bus/Device/Function</td>
<td>NVMe Qualified Name (QN)</td>
</tr>
<tr>
<td>Discovery</td>
<td>Bus Enumeration</td>
<td>Discovery and Connect commands</td>
</tr>
<tr>
<td>Queuing</td>
<td>Memory-based</td>
<td>Message-based</td>
</tr>
<tr>
<td>Data Transfers</td>
<td>PRPs or SGLs</td>
<td>SGLs only, added Key</td>
</tr>
</tbody>
</table>
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 μs 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.
SPDK provided a boost with more IOPS/Core and less overhead.

http://mt.sohu.com/20170228/n48125433.shtml

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

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 PCIe Express* and SAS, used to create SFF-8639

Get involved – visit www ssdformfactor.org
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

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
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

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
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

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

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

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.
Invest for the Future with EDSFF

- 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

Scalable Family for Different Usages

**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

Santa Clara, CA
August 2018

Flash Memory Summit 2018
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 μs, 9100 cycles
  - Traditional: 6.0 μs, 19500 cycles
PREPARING THE WAY FOR AN NVM BREAKTHROUGH

Flash Memory Summit 2013

Next Generation Scalable NVM

Resistive RAM NVM Options

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

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

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

Fully Exploiting Next Gen NVM Requires Platform Improvements

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

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

- NAND MLC SATA 3.0 N0F12
- NAND MLC SATA 3.0 N0F3
- NAND MLC PCIe x4 Gen3 N0F3
- Future NVM PCIe x4 Gen3

Flash Memory Summit 2013
Santa Clara, CA
3D XPoint™ Technology

**Cross Point Structure**
Selectors allow dense packing and individual access to bits

**Breakthrough Material Advances**
Compatible switch and memory cell materials

**Scalable**
Memory layers can be stacked in a 3D manner

**High Performance**
Cell and array architecture that can switch states 1000x faster than NAND

*Technology 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

<table>
<thead>
<tr>
<th>Interfaces</th>
<th>Processor</th>
</tr>
</thead>
<tbody>
<tr>
<td>On Core CPU</td>
<td>SRAM Cache</td>
</tr>
<tr>
<td>On Die</td>
<td>DDR DRAM</td>
</tr>
<tr>
<td>Direct Attach</td>
<td>SSD</td>
</tr>
<tr>
<td>PCI Express*</td>
<td>HDD</td>
</tr>
<tr>
<td>NVM Express*</td>
<td></td>
</tr>
<tr>
<td>SAS, SATA</td>
<td></td>
</tr>
</tbody>
</table>

Relative Latency (reads)

- 1x
- 10x
- 100,000x
- 10,000,000x

*Technology 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

Intel® Optane™ SSDs change the game, operating at previously unreachable IOPs/Latency combinations.
# Proof Point: Cassandra 4.0* Database

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

<table>
<thead>
<tr>
<th>Read IOPS</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel® SSD DC P3700 (PCIe*)</td>
<td>1</td>
<td>2.5x</td>
<td>9x</td>
</tr>
<tr>
<td>+30% Intel® Optane™ SSD DC P4800X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intel® Optane™ SSD DC P4800X w/ Direct I/O* Java optimizations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intel® Optane™ DC Persistent Memory with app direct mode optimizations</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Deploy
available Intel® Optane™ DC SSDs

## Optimize
with available software tools

## Evolve
with next-generation memory technology

---

1. System configuration: Server model: 2x Intel® Xeon® E5 2699 v4 @ 4. Gzh, 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 name: 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

- DRAM
- OPTANE DC Persistent Memory
- OPTANE DC Solid State Drive
- NAND SSD
- HDD / TAPE
Global Digital Data Created (ZB)

- Endpoints (e.g. client, mobile, IoT)
- Edge
- Datacenters
- Real-time

25% CAGR

Source: IDC Data Age 2025 (2017)
WE HAVE INVENTED THE FUTURE TOGETHER

LET’S CONTINUE OUR WORK OVER THE NEXT DECADE