NVM Express Overview & Ecosystem Update

Amber Huffman

Sr Principal Engineer, Intel

August 13, 2013

Agenda

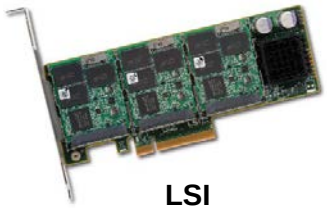
- Importance of Architecting for NVM
- Overview of NVM Express
- Storage Drivers
- Form Factors and Connectors

PCIe* SSDs for the Datacenter

- PCI Express* is a great interface for SSDs
  - Stunning performance — 1 GB/s per lane (PCIe* Gen3 x1)
  - With PCIe* scalability — 8 GB/s per device (PCIe* Gen3 x8) or more
  - Lower latency — Platform+Adapter: 10 µsec down to 3 µsec
  - Lower power — No external SAS IOC saves 7-10 W
  - Lower cost — No external SAS IOC saves ~ $15
  - PCIe* lanes off the CPU — 40 Gen3 (**80** in dual socket)

Virident

Fusion-io

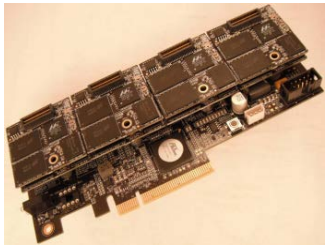


LSI

OCZ

Micron

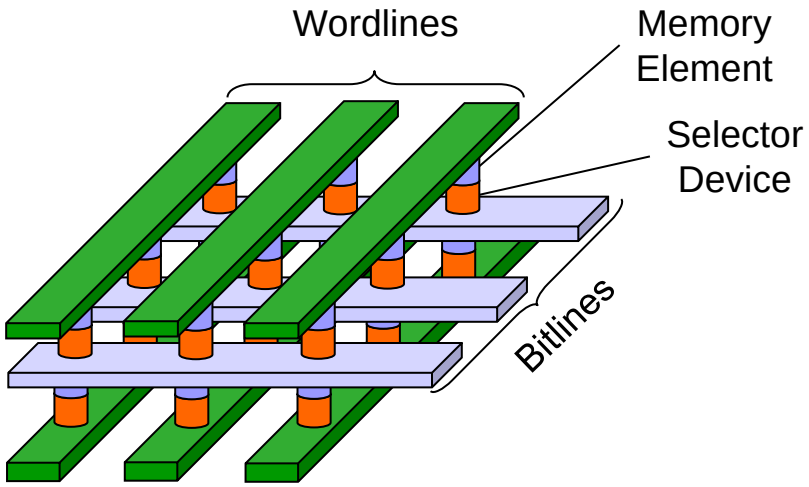
Intel



Marvell

Next Generation Scalable NVM

Scalable Resistive Memory Element



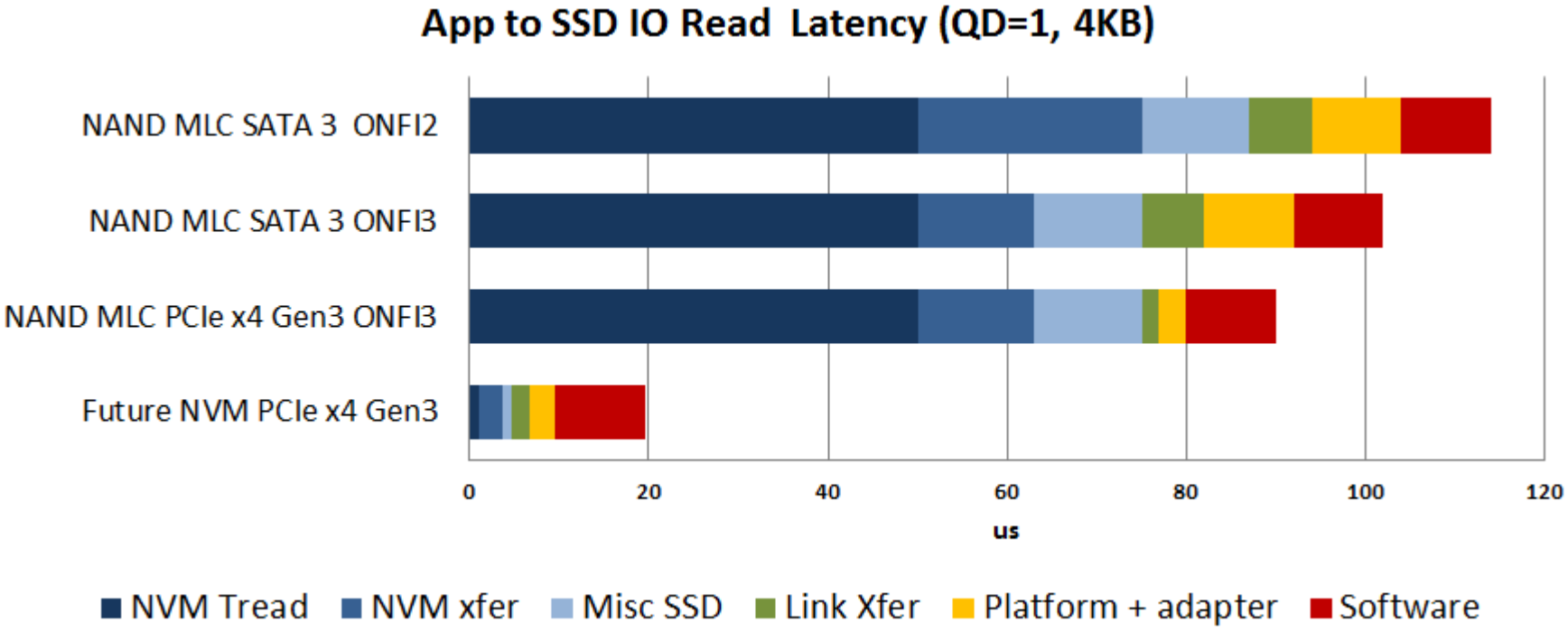
Cross Point Array in Backend Layers ~$4\lambda^2$ Cell

Resistive RAM NVM Options

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

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

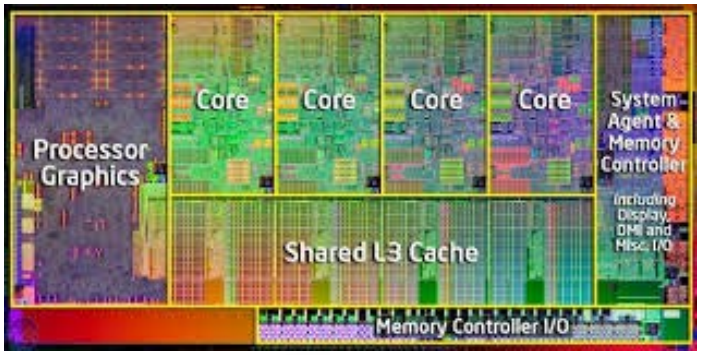
Fully Exploiting Next Gen NVM
Requires Platform Improvements



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

Transformation Required

- Transformation was needed for full benefits of multi-core CPU
  - App and OS level changes required
- To date, SSDs have used the legacy interfaces of hard drives
  - Based on a single, slow rotating platter
- SSDs are inherently parallel and next gen NVM approaches DRAM-like latencies



For full SSD benefits, must architect for NVM from the ground up.
NVMe is architected for NAND today and next gen NVM of tomorrow.

Agenda

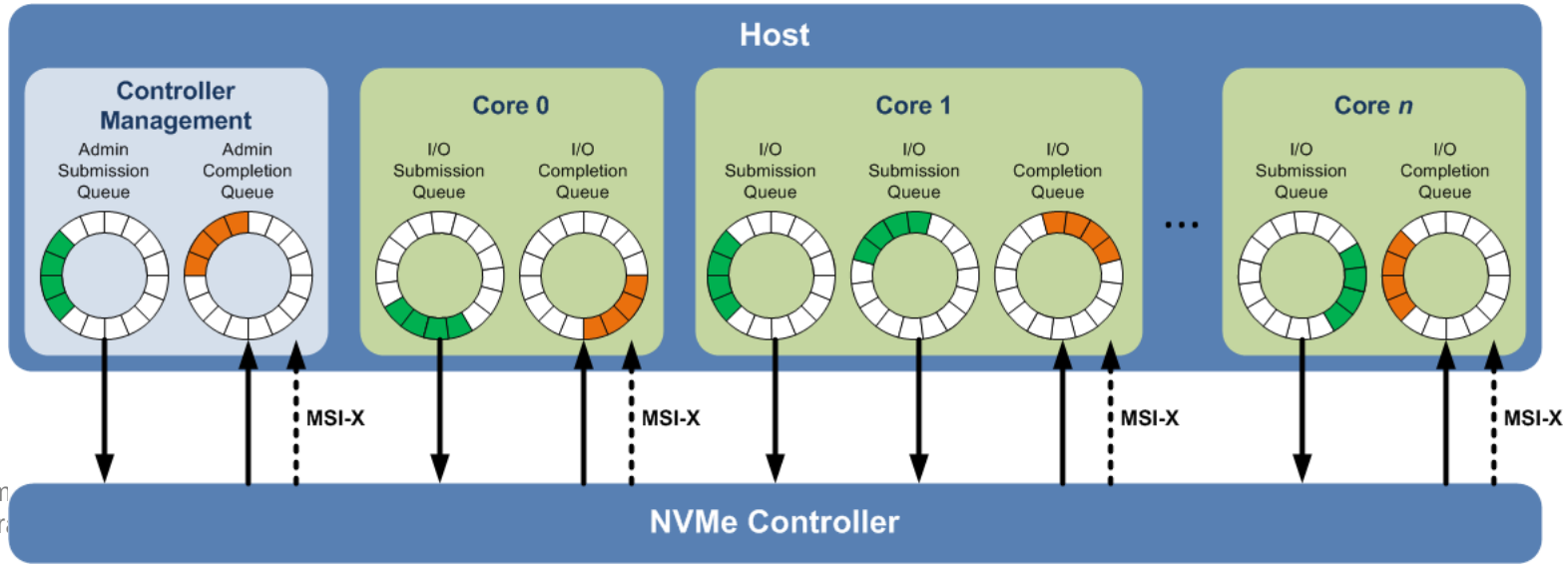
- Importance of Architecting for NVM
- Overview of NVM Express
- Storage Drivers
- Form Factors and Connectors

NVM Express

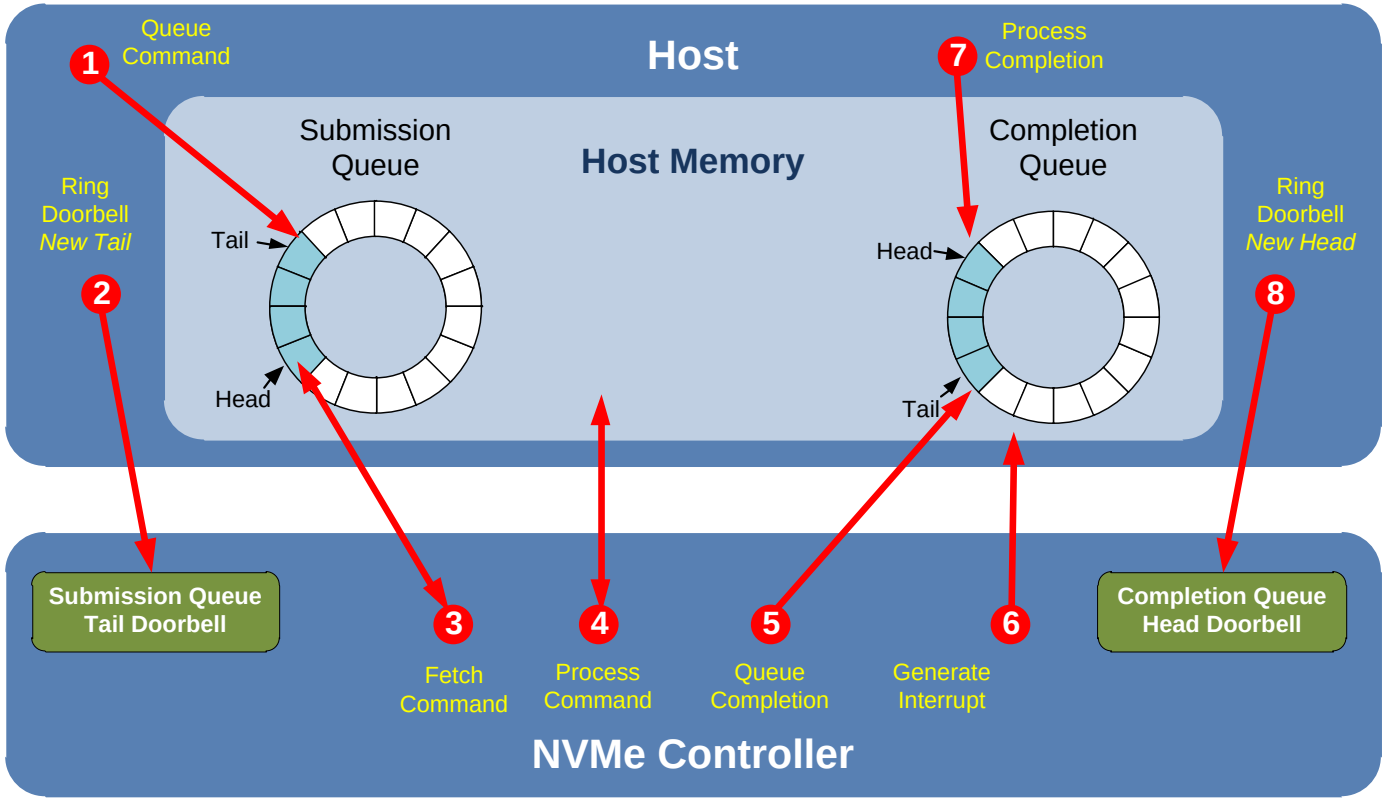
- NVM Express (NVMe) is the standardized high performance host controller interface for PCIe* SSDs
- NVMe was architected from the ground up for non-volatile memory, scaling from Enterprise to Client
  - The architecture focuses on latency, parallelism/performance, and low power
  - The interface is explicitly designed with next generation NVM in mind
- NVMe was developed by an open industry consortium of 90+ members and is directed by a 13 company Promoter Group

Technical Basics

- All parameters for 4KB command in single 64B command
- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports MSI-X and interrupt steering
- Streamlined & simple command set (13 required commands)
- Optional features to address target segment (Client, Enterprise, etc)
  - Enterprise: End-to-end data protection, reservations, etc
  - Client: Autonomous power state transitions, etc
- Designed to scale for next generation NVM, agnostic to NVM type used



Queuing Interface
Command Submission & Processing



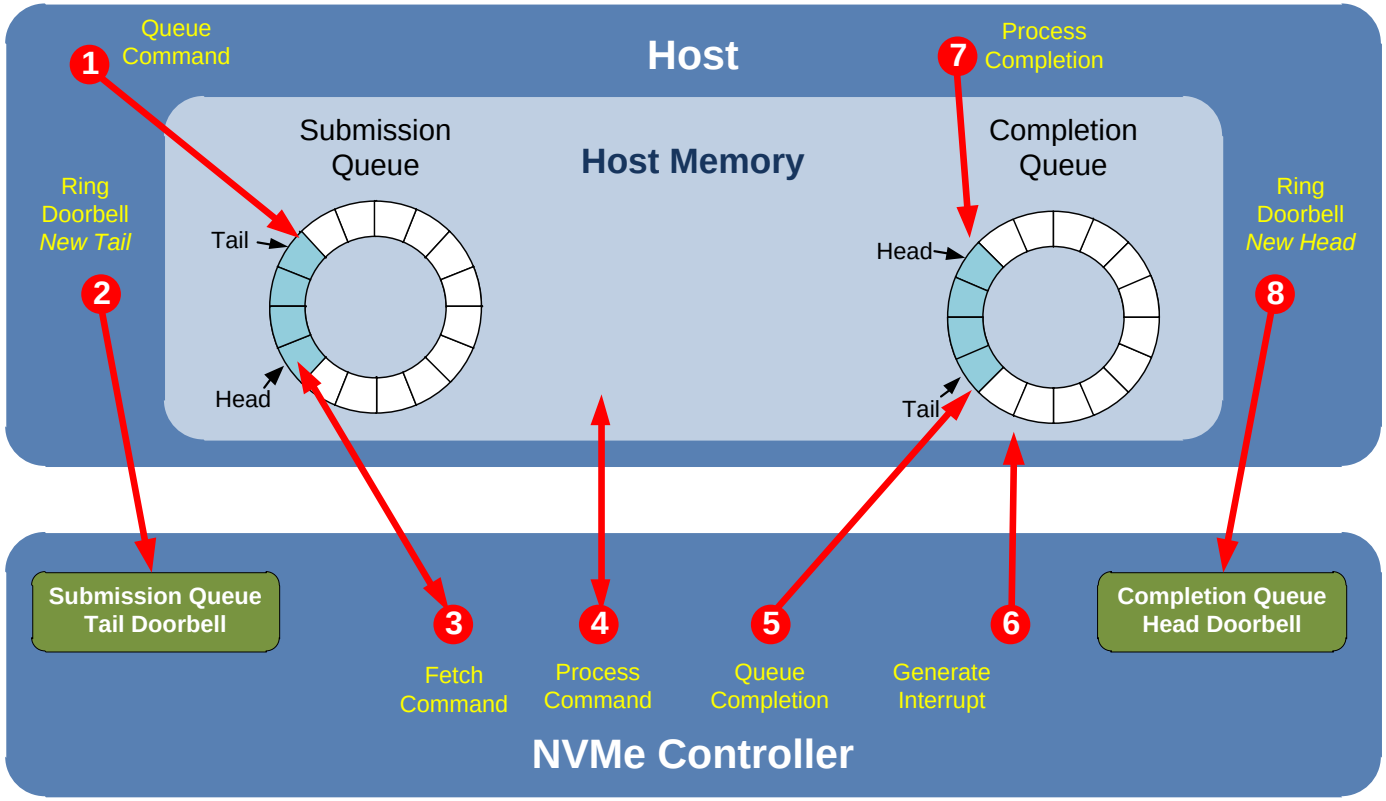
Command Submission

1. Host writes command to Submission Queue
2. Host writes updated Submission Queue tail pointer to doorbell

Command Processing

3. Controller fetches command
4. Controller processes command

Queuing Interface
Command Completion



Command Completion

5. Controller writes completion to Completion Queue
6. Controller generates MSI-X interrupt
7. Host processes completion
8. Host writes updated Completion Queue head pointer to doorbell

Simple Optimized Command Set

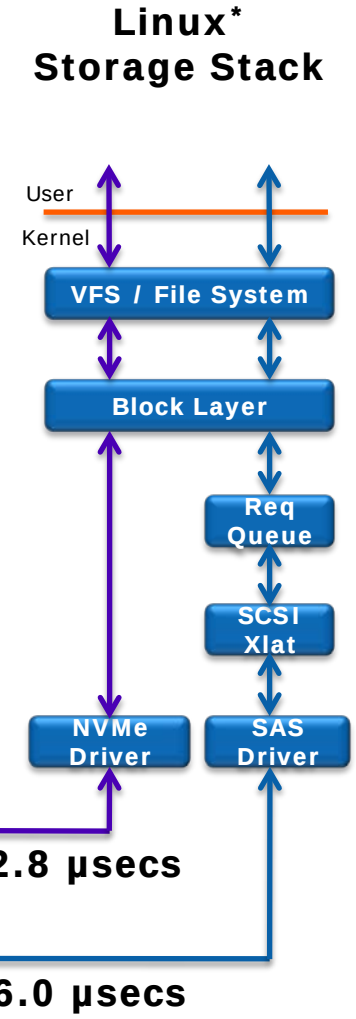
Admin Commands
Create I/O Submission Queue
Delete I/O Submission Queue
Create I/O Completion Queue
Delete I/O Completion Queue
Get Log Page
Identify
Abort
Set Features
Get Features
Asynchronous Event Request
Firmware Activate (optional)
Firmware Image Download (optional)
Format NVM (optional)
Security Send (optional)
Security Receive (optional)

NVM I/O Commands
Read
Write
Flush
Write Uncorrectable (optional)
Compare (optional)
Dataset Management (optional)
Write Zeros (optional)
Reservation Register (optional)
Reservation Report (optional)
Reservation Acquire (optional)
Reservation Release (optional)

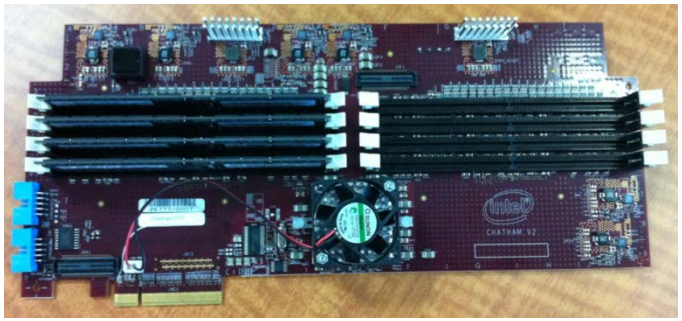
Only 10 Admin and 3 I/O commands required.

Proof Point: NVMe Latency

- NVMe reduces latency overhead by more than 50%
  - SCSI/SAS: 6.0 µs 19,500 cycles
  - **NVMe: 2.8 µs 9,100 cycles**
- Increased focus on storage stack / OS needed to reduce latency even further



Chatham NVMe Prototype



Prototype Measured IOPS

Cores Used for 1M IOPs

Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat* EL6.0 2.6.32-71 Kernel.

NVMe Deployment Beginning

- NVM Express 1.0 specification published in March 2011
  - Additional Enterprise and Client capabilities included in NVMe 1.1 (Oct 2012)
- First plugfest held May 2013 with 11 companies participating
  - Interoperability program run by University of New Hampshire Interoperability Lab, a leader in PCIe*, SAS, and SATA compliance programs

"NVMe was designed from the ground up to enable non-volatile memory to better address the needs of enterprise and client systems. This Plugfest highlights the rapid development and maturity of the NVMe specification and the surrounding infrastructure as well as supporting PCIe SSD devices."

JH Lee, Vice President,
Flash Memory Product Planning and Enabling,
Samsung Electronics

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). NVM 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.

NVM Express products targeting Datacenter expected to ship 2H'13.

Agenda

- Importance of Architecting for NVM
- Overview of NVM Express
- Storage Drivers
- Form Factors and Connectors

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!

Windows* Open Source Driver Update

Release 1	• Q2 2012 (released) • 64-bit support on Windows* 7, Windows* Server 2008 R2 • Mandatory features
Release 1.1	• Q4 2012 (released) • Added 64-bit support Windows* 8 • Public IOCTLs and Windows* 8 Storport updates
Release 1.2	• Q2 2013 (released) • Added 64-bit support on Windows* Server 2012 • Signed executable drivers
Release 1.3	• Target: Q4 2013 • Added 32-bit support on all supported OS versions • End-to-end Data Protection

Three major releases of the Windows* community driver since 2012.
Code contributions from Huawei, IDT, Intel, LSI, and SanDisk.

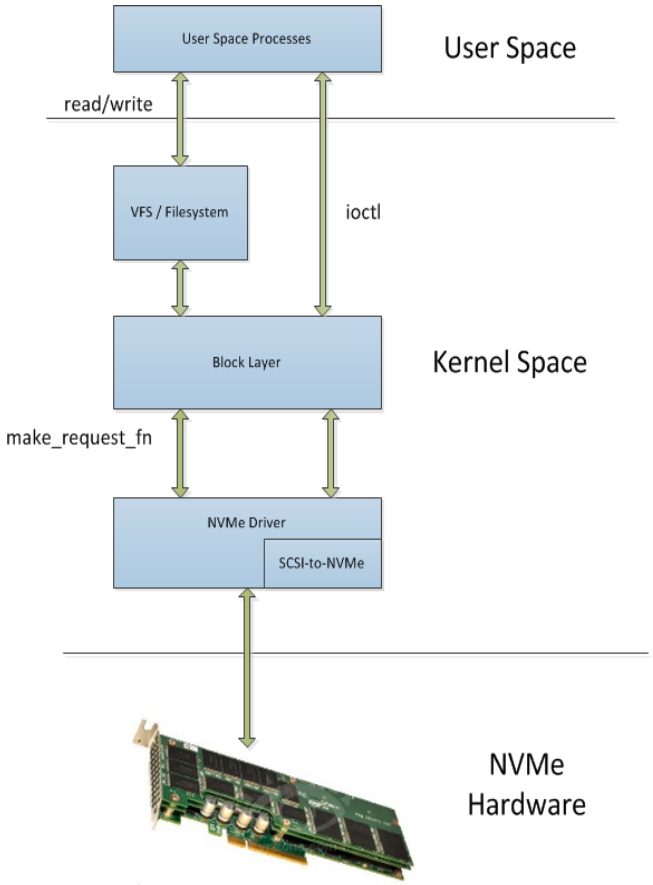
Linux* Driver Update

Recent Feature Additions

- Deallocate (i.e., Trim support)
- 4KB sector support (in addition to 512B)
- SCSI IOCTL support
- MSI support (in addition to MSI-X)
- Disk I/O statistics
- Many bug fixes

Community Effort

- Contributions from Fastor, IDT, Intel, Linaro, Oracle, SanDisk, and Trend Micro
- 59 changes since integrated into kernel



Current work includes power management, support for end-to-end data protection, sysfs manageability, and NUMA optimizations.

FreeBSD Driver Update

- NVMe support is upstream in the head and stable/9 branches
- FreeBSD 9.2 will be the first official release with NVMe support, slated for end of August

FreeBSD NVMe Modules

nvmecontrol

User space utility, including firmware update

nvd

NVMe/block layer shim driver

nvme

Core NVMe driver

Solaris* Driver Update

- Current Status from Oracle team:
  - Stable and efficient working prototype conforming to 1.0c
  - Direct block interfaces bypassing complex SCSI code path
  - NUMA optimized queue/interrupt allocation
  - Support 8K memory page size on SPARC system
  - Plan to validate driver against Oracle SSD partners
  - Plan to integrate into S12 and a future S11 Update Release

- Future Development Plans:
  - Boot & install on SPARC and X86
  - Surprise removal support
  - Multipath, SR-IOV, SGL, etc

VMware Driver Update

- Initial “vmklinux” based driver in final stages of development
  - First release in mid-Oct, 2013
  - Certified release in Dec, 2013
- Native NVMe driver with pluggable Vendor Extensions planned for the future
- VMware‘s IOVP program includes workflow for bugs/issues

UEFI Driver

- The UEFI 2.4 specification available at www.UEFI.org contains updates for NVMe

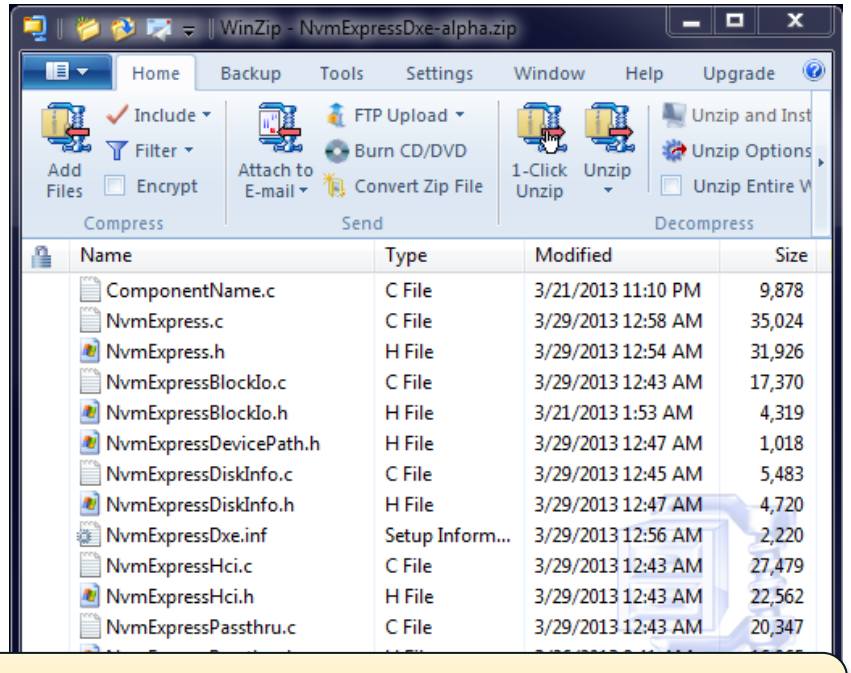
- An open source version of an NVMe implementation for UEFI is at:

  https://sourceforge.net/projects/edk2/files/EDK%20II%20Releases/other/NvmExpressDxe-alpha.zip/download

- The UEFI driver has been validated using the qemu virtual platform
  - The team is ready to test against real hardware as it rolls in

"AMI is working with vendors of NVMe devices and plans for full BIOS support of NVMe in 2014."

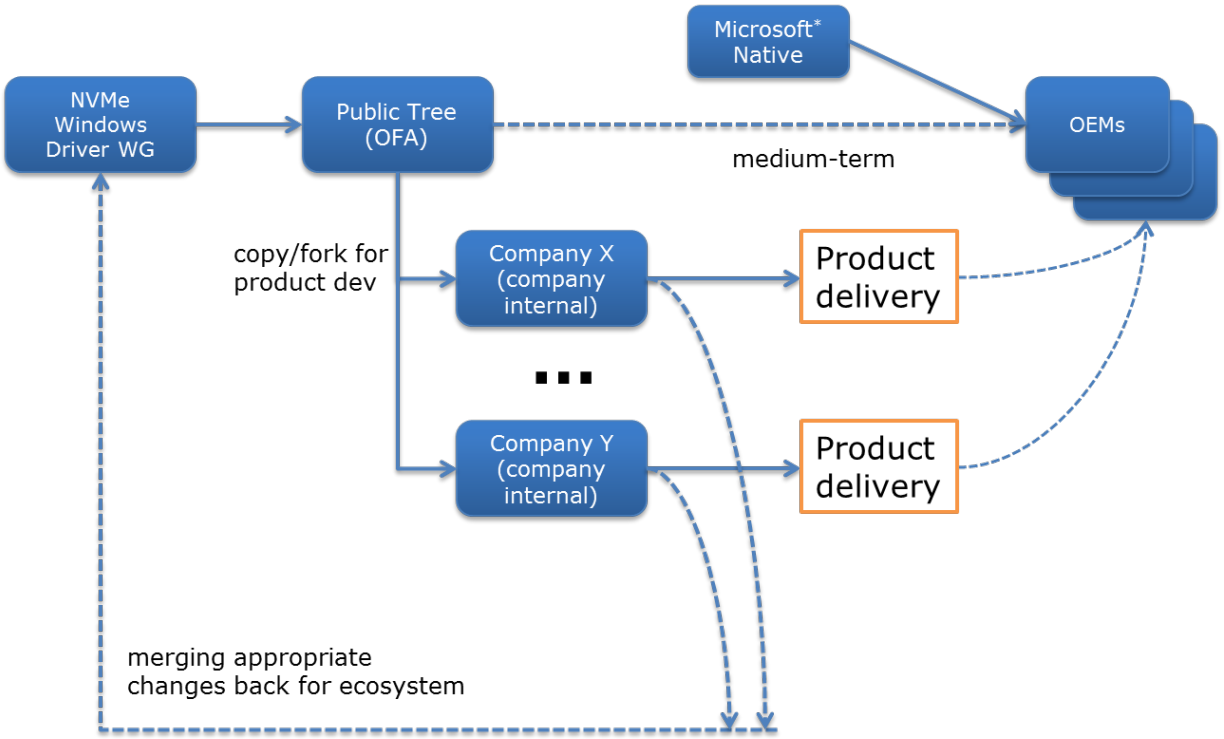
Sandip Datta Roy
VP BIOS R&D, AMI



NVMe boot support with UEFI will start percolating releases from Independent BIOS Vendors in 2014.

Fork and Merge for IHV Value-Add Drivers

- The NVMe Workgroup continues to recommend a “Fork and Merge” approach when IHVs provide their own value add driver
- Benefits of this strategy include: 1) maximum reference code re-use, 2) continuous improvement of reference code, 3) enables product team to focus on delivery (not basic driver)



Agenda

- Importance of Architecting for NVM
- Overview of NVM Express
- Storage Drivers
- Form Factors and Connectors

M.2 Emerging as Primary Client Form Factor

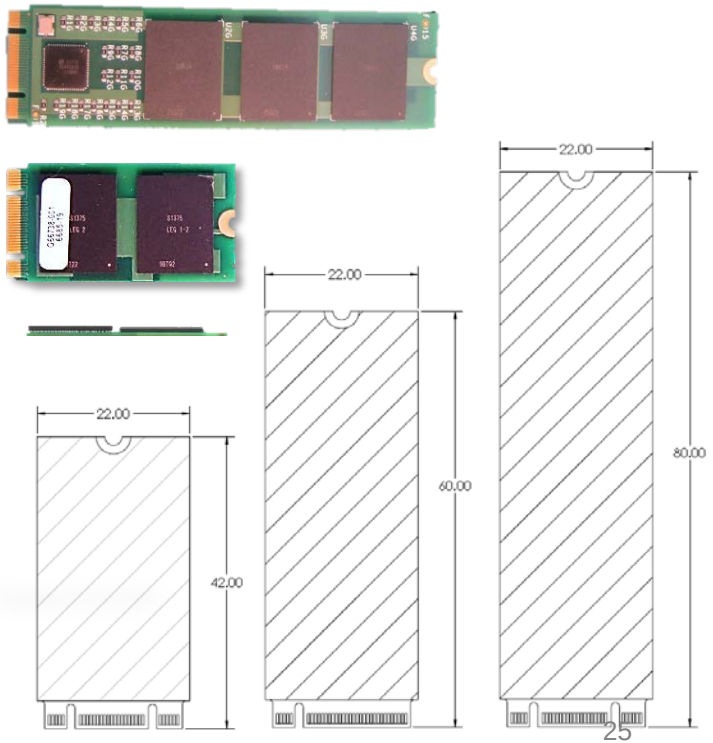
- In client, as SSDs move first to PCIe, OEMs are using the optimized M.2 form factor
- As native OS support of NVMe becomes pervasive, OEMs will move from AHCI to NVMe to take full advantage of PCIe

VAIO* Pro 13 Ultrabook™

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

Features:

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

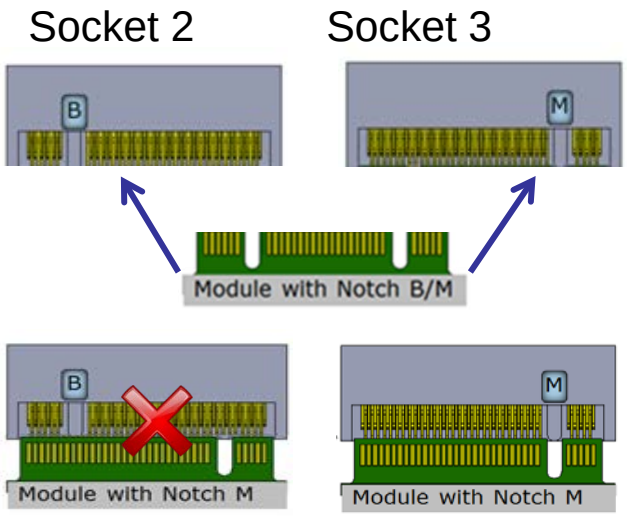


*Other names and brands may be claimed as the property of others.

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)

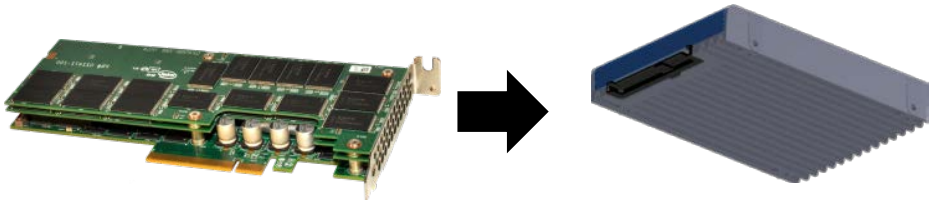
- OEMs choose the socket to include
  - Socket 2: Most flexibility
  - Socket 3: Highest performance

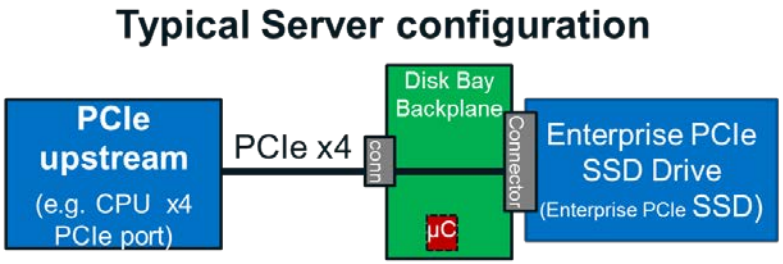


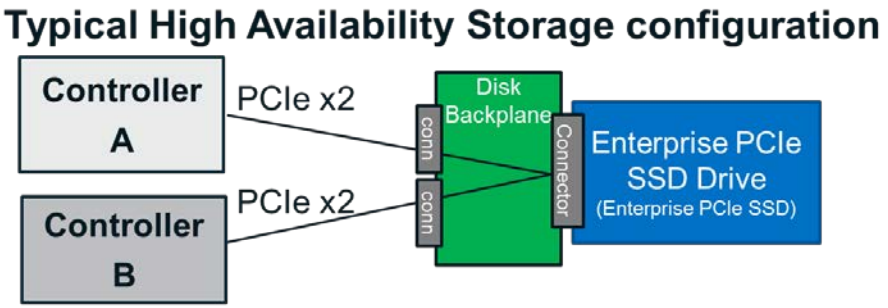
With M.2, client OEMs can choose maximum performance with 4 lanes, or they can choose flexibility with SATA and WWAN options.

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

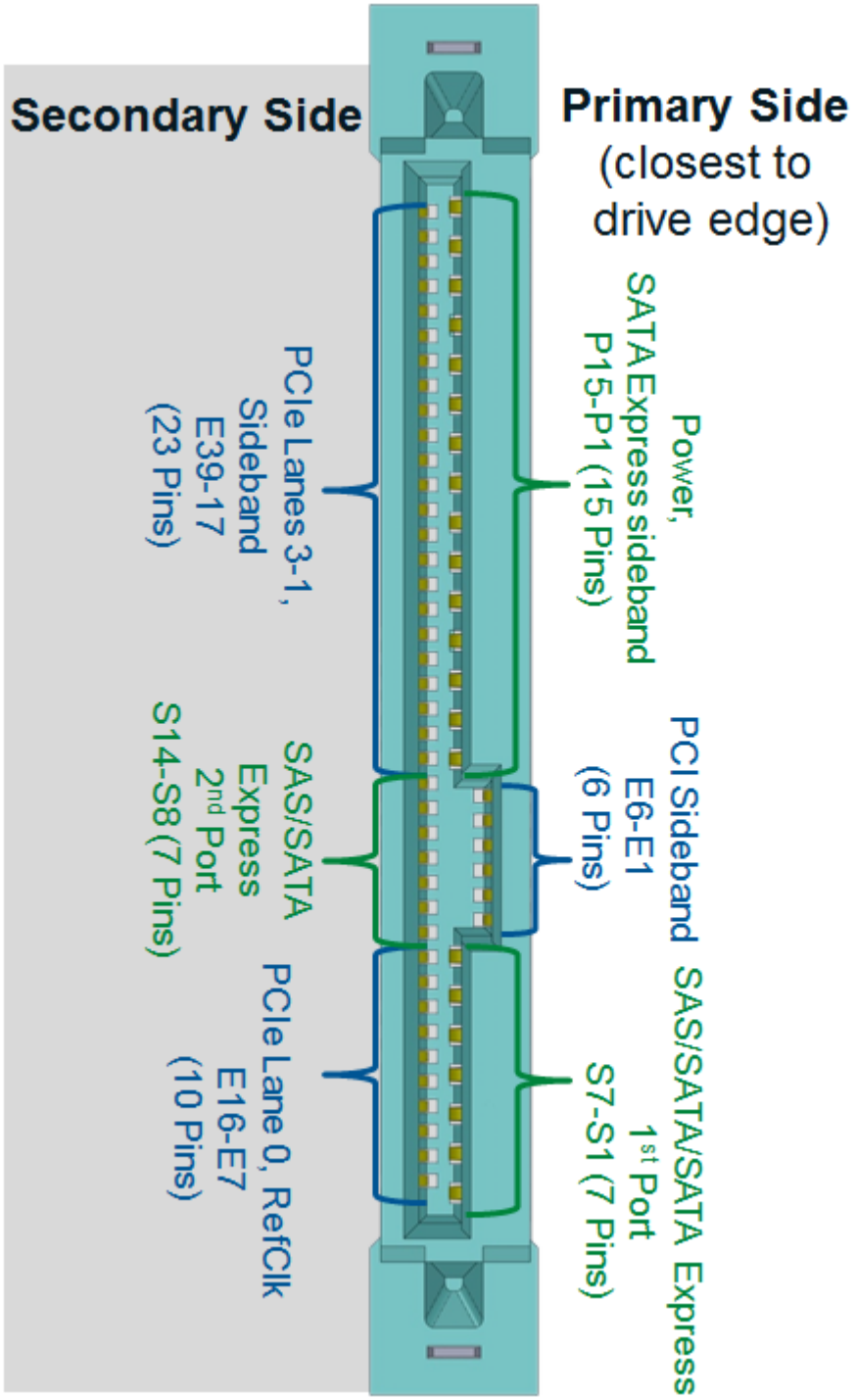






SFF-8639 Flexibility

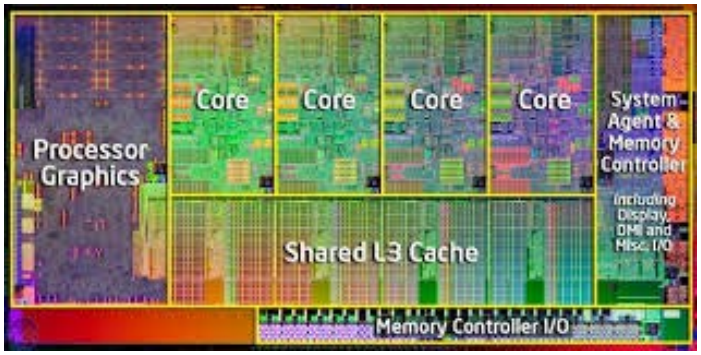
- SFF-8639 supports:
  - Enterprise PCIe x4 SSDs
  - Existing SAS drive (dual port)
  - Existing SATA drives
- As ecosystem develops:
  - Client 2.5” PCIe (often referred to as SATA Express)
  - x4 SAS
- Supports flexible backplanes
  - Enterprise x4 PCIe SSDs
  - SAS/SATA HDDs



Summary: Transformation Required

Recall:

- Transformation was needed for full benefits of multi-core CPU
  - Application and OS level changes required
- To date, SSDs have used the legacy interfaces of hard drives
  - Based on a single, slow rotating platter..
- SSDs are inherently parallel and next gen NVM approaches DRAM-like latencies



For full SSD benefits, must architect for NVM from the ground up.
Future proof your PCIe storage investment by adopting NVMe.