NVMe™ Management Interface (NVMe-MI™) and Drivers Update

Sponsored by NVM Express® organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

Speakers

Uma Parepalli

Austin Bolen

Myron Loewen

Lee Prewitt

Dave Minturn

Suds Jain

Jim Harris

Agenda

- Session Introduction – Uma Parepalli, Marvell (Session Chair)
- NVMe Management Interface - Austin Bolen, Dell EMC and Myron Loewen, Intel
- NVMe Driver Updates
  - NVMe Driver Ecosystem and UEFI Drivers - Uma Parepalli, Marvell
  - Microsoft Inbox Drivers - Lee Prewitt, Microsoft
  - Linux Drivers - Dave Minturn, Intel
  - VMware Drivers -Suds Jain, VMWare
- SPDK Updates - Jim Harris, Intel

NVMe™ Management Interface (NVMe-MI™) Workgroup Update

Austin Bolen, Dell EMC

Myron Loewen, Intel

Agenda

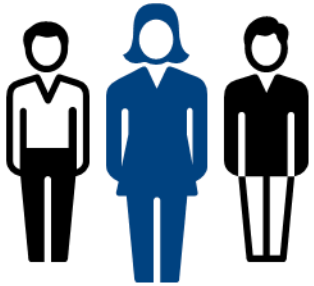
- NVMe-MI™ Workgroup Update
- NVMe-MI 1.0a Overview
- What's new in NVMe-MI 1.1
  - In-band NVMe-MI
  - Enclosure Management
  - Managing Multi NVM Subsystem Devices
- Summary

NVM Express®, Inc. 120+ Companies defining NVMe™ together

Board of Directors

13 elected companies, stewards of the technology & driving processes

Chair: Amber Huffman



Marketing Workgroup

NVMexpress.org, webcasts, tradeshows, social media, and press

Co-Chairs: Janene Ellefson and Jonmichael Hands

Technical Workgroup

NVMe Base and NVMe Over Fabrics

Chair: Amber Huffman

Management Intf. Workgroup

Out-of-band management over SMBus and PCIe® VDM

Chair: Peter Onufryk

Vice Chair: Austin Bolen

Interop (ICC) Workgroup

Interop & Conformance Testing in collaboration with UNH-IOL

Chair: Ryan Holmqvist

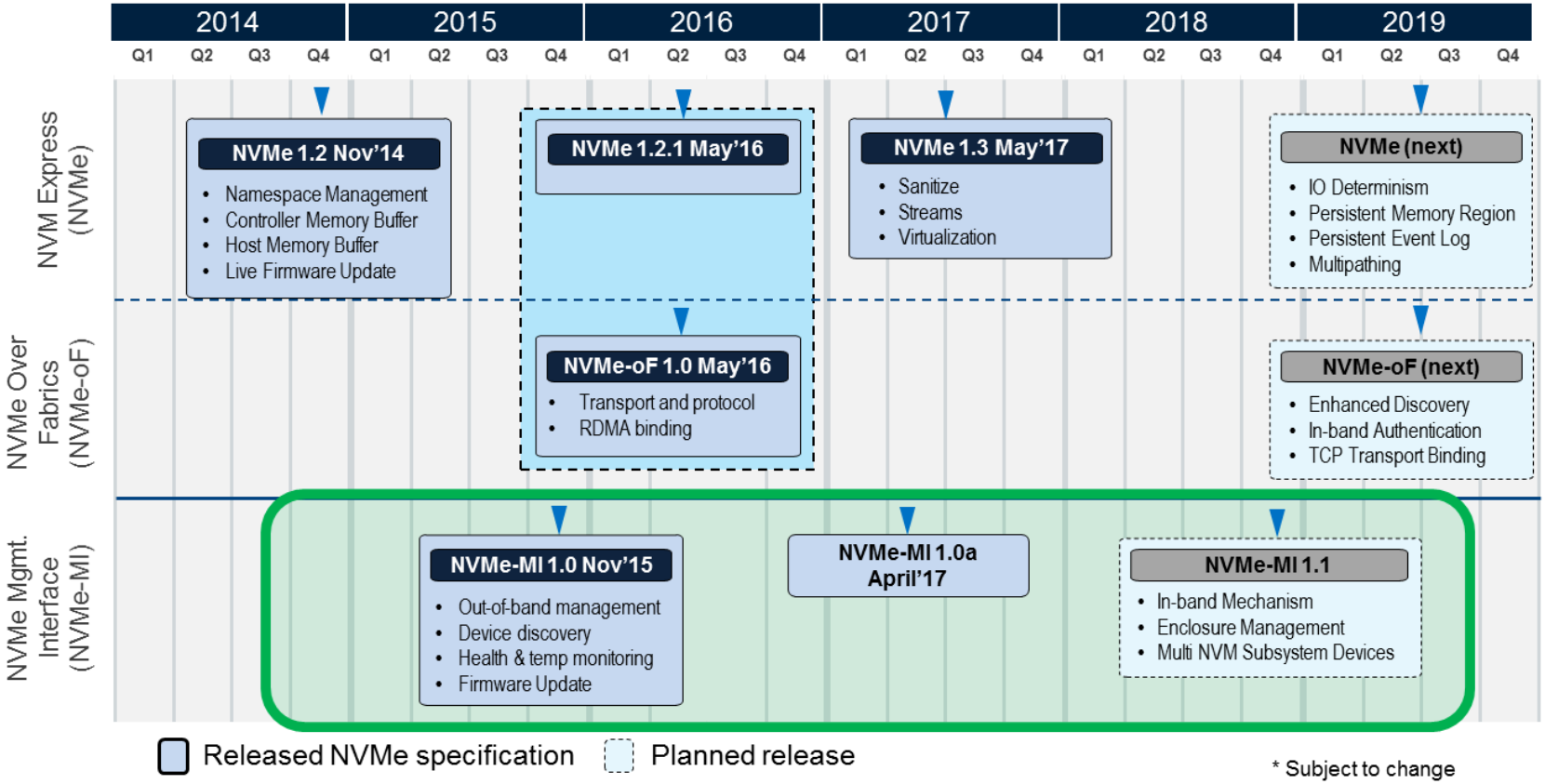
facebook Microsoft intel cisco DELL EMC SEAGATE TOSHIBA

Micron ORACLE SAMSUNG Microsemi NetApp WD Western Digital

Flash Memory Summit

nvm EXPRESS®

NVM Express™ Roadmap



NVMe-MI™ Ecosystem

- Commercial test equipment for NVMe-MI
- NVMe-MI 1.0a compliance testing program has been developed
  - Compliance testing started in the May 2017 NVMe™ Plugfest conducted by the University of New Hampshire Interoperability Laboratory (UNH-IOL)
  - 7 devices from multiple vendors have passed compliance testing and are on the NVMe-MI Integrators List
- Servers are shipping that support NVMe-MI

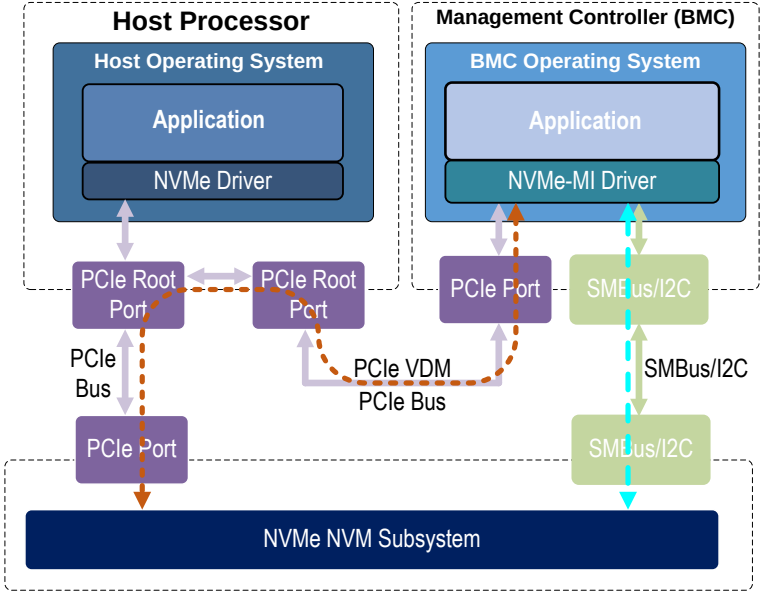
What is the NVMe™ Management Interface 1.0a?

A programming interface that allows out-of-band management of an NVMe Storage Device Field Replaceable Unit (FRU)

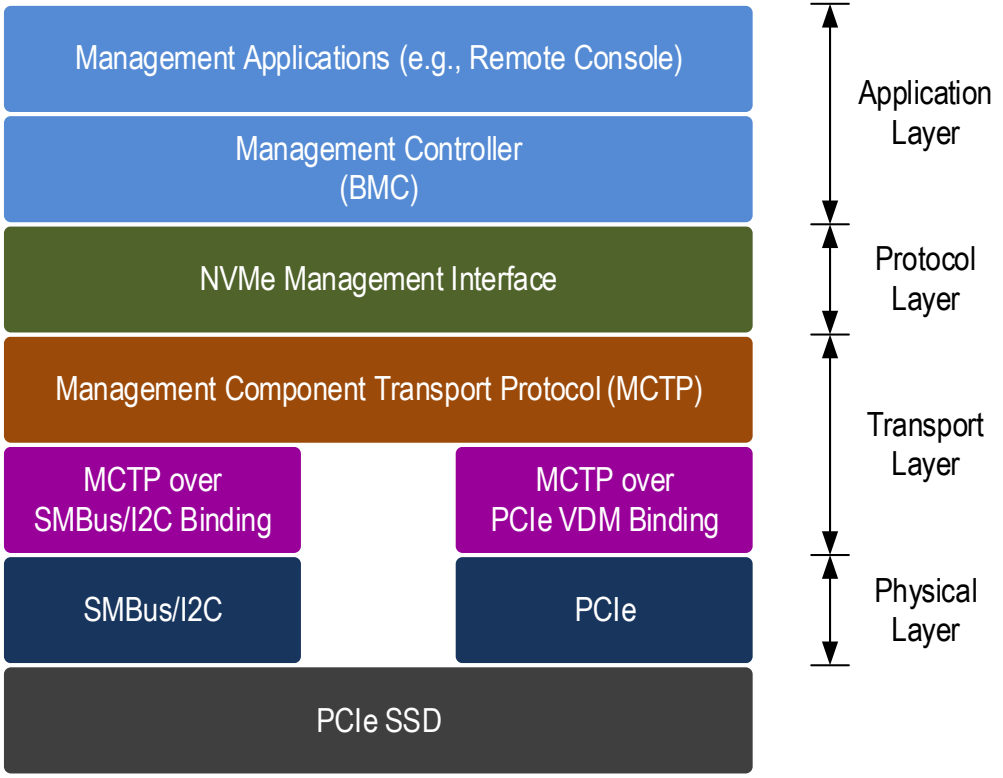
Out-of-Band Management and NVMe-MI™

- **Out-of-Band Management** – Management that operates with hardware resources and components that are ***independent of the host operating system control***

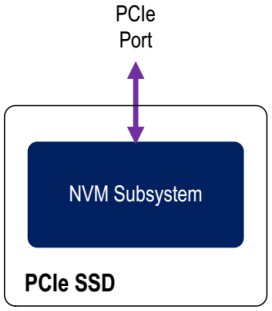
- **NVMe™ Out-of-Band Management Interfaces**
  - SMBus/I2C
  - PCIe Vendor Defined Messages (VDM)

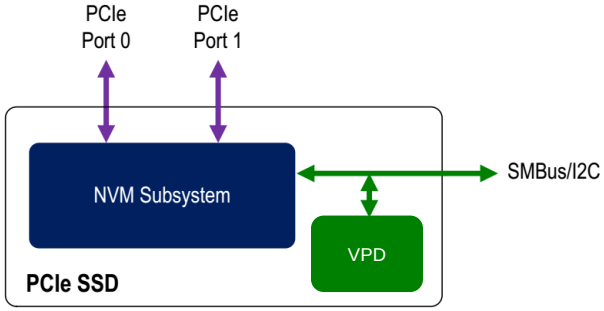


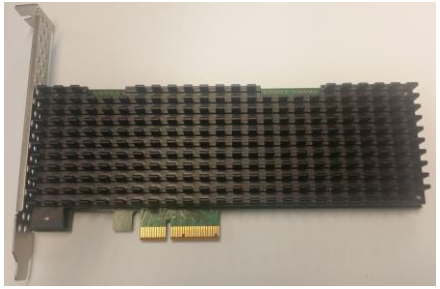
NVMe-MI™ Out-of-Band Protocol Layering



NVMe™ Storage Device in 1.0a

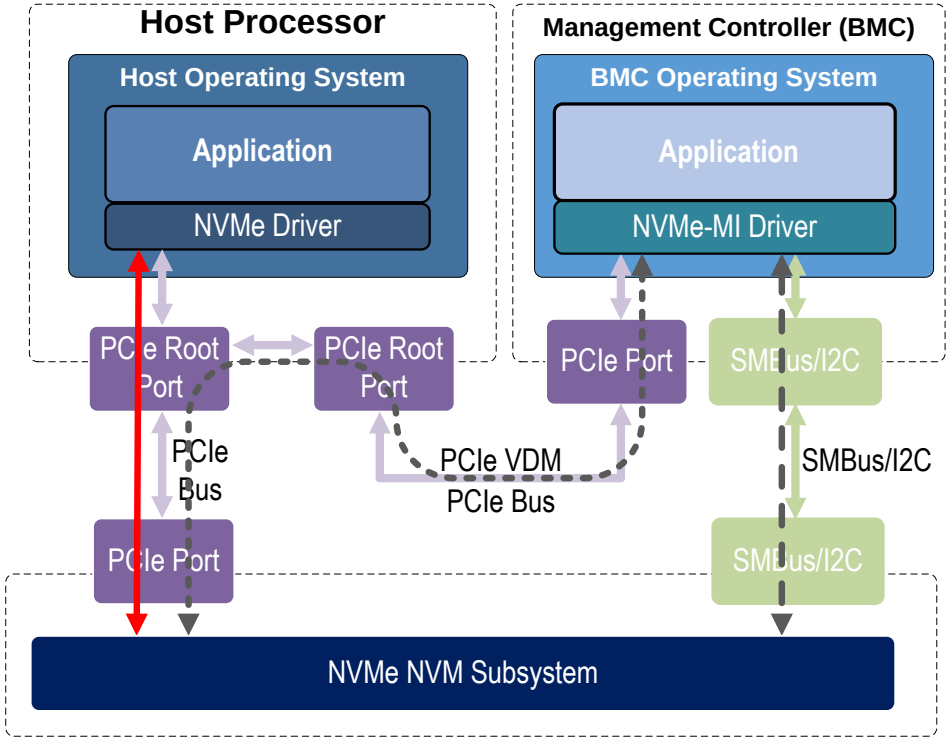






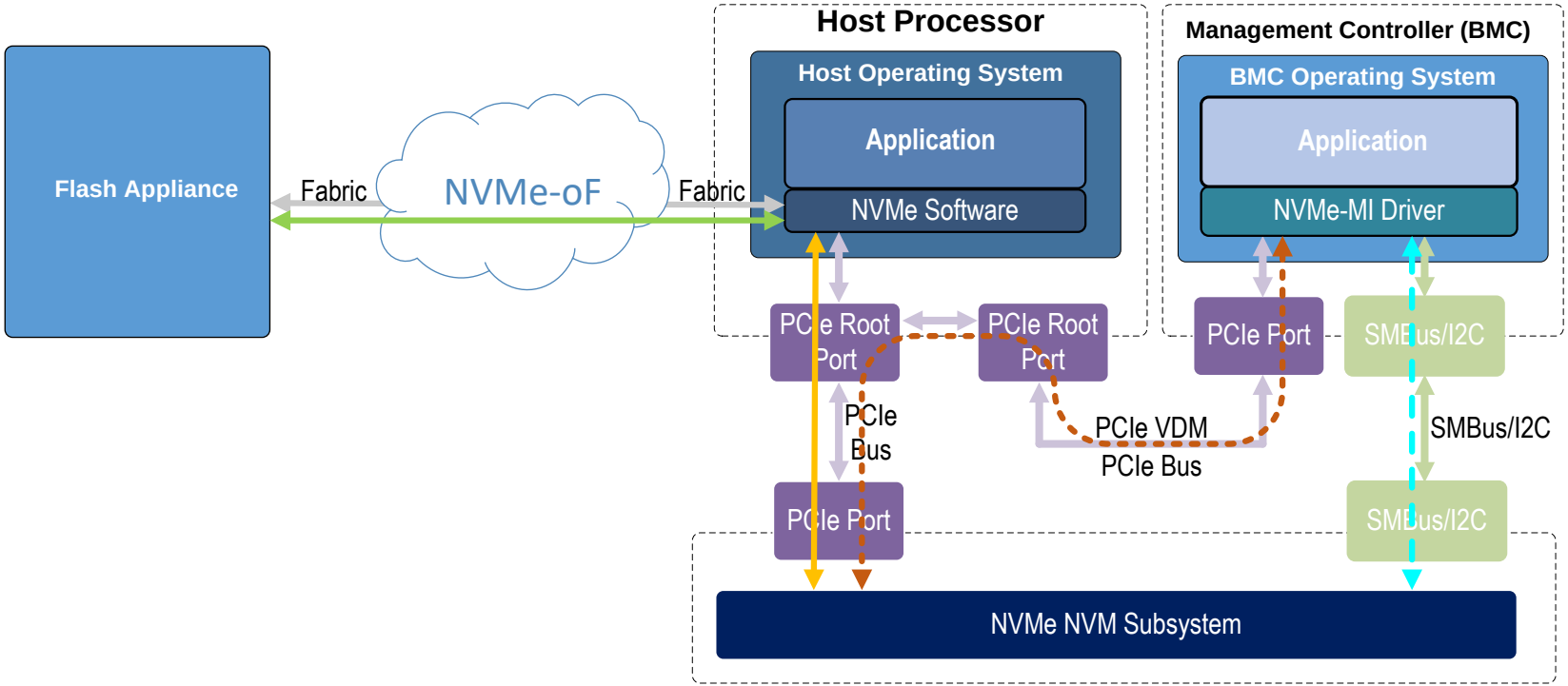
- **NVMe Storage Device** – One NVM Subsystem with one or more ports, vital product data (VPD), and an optional SMBus/I2C interface

In-Band Management and NVMe-MI™



- In-band mechanism allows application to tunnel NVMe-MI commands through NVMe™ driver
  - Two new NVMe Admin commands
    - NVMe-MI Send
    - NVMe-MI Receive
- Benefits
  - Provides management capabilities not available in-band via NVMe commands
    - Efficient NVM Subsystem health status reporting
    - Ability to manage NVMe at a FRU level
    - Vital Product Data (VPD) access
    - Enclosure management

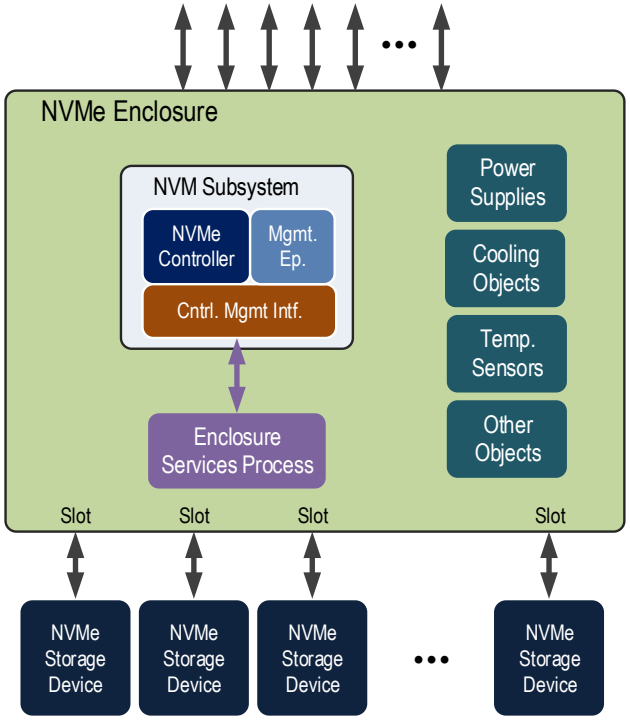
NVMe-MI™ over NVMe-oF™



Plumbing in place for NVMe-MI over NVMe-oF

Enclosure Management

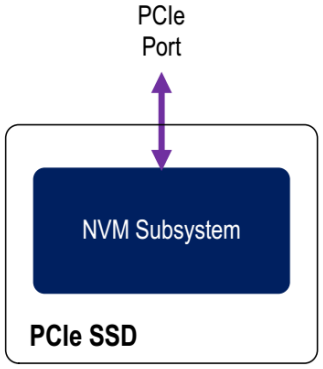
- SES Based Enclosure Management
  - Technical proposal developed in NVMe-MI™ workgroup
  - While the NVMe™ and SCSI architectures differ, the elements of an enclosure and the capabilities required to manage these elements are the same
    - Example enclosure elements: power supplies, fans, display or indicators, locks, temperature sensors, current sensors, voltage sensors, and ports
  - Comprehensive enclosure management that leverages SCSI Enclosure Services (SES), a standard developed by T10 for management of enclosures using the SCSI architecture



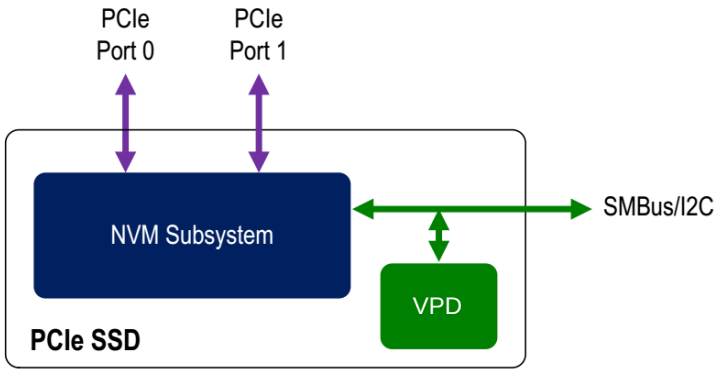
Multi NVM Subsystem Management

NVMe-MI™ 1.0a NVMe™ Storage Device

- **NVM Storage Device** – One NVM Subsystem with one or more ports and an optional SMBus/I2C interface

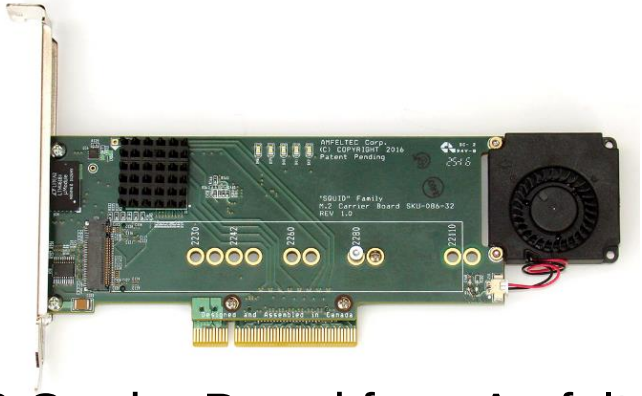


Single Ported PCIe SSD



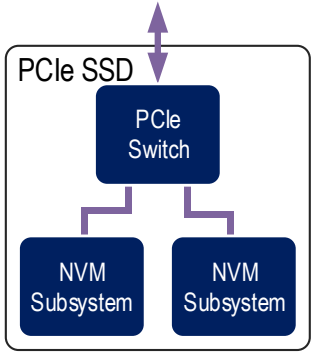
Dual Ported PCIe SSD with SMBus/I2C

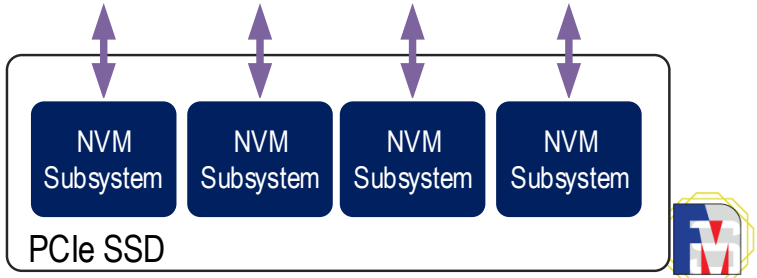
NVMe™ Storage Device with Multiple NVM Subsystems



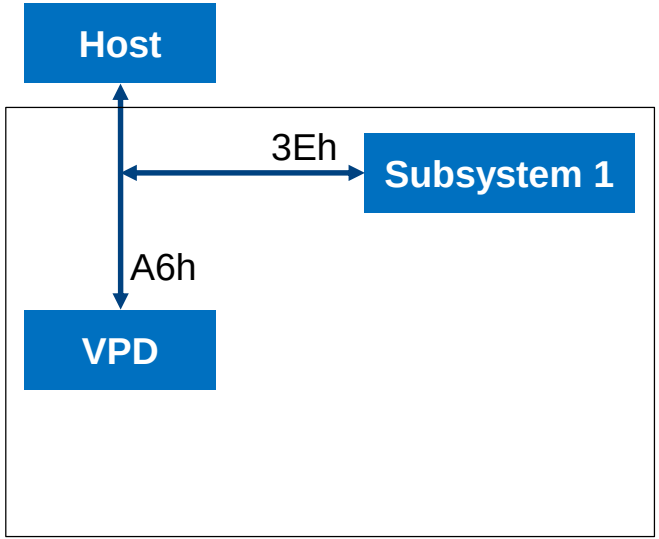
M.2 Carrier Board from Amfeltec

ANA Carrier Board from Facebook



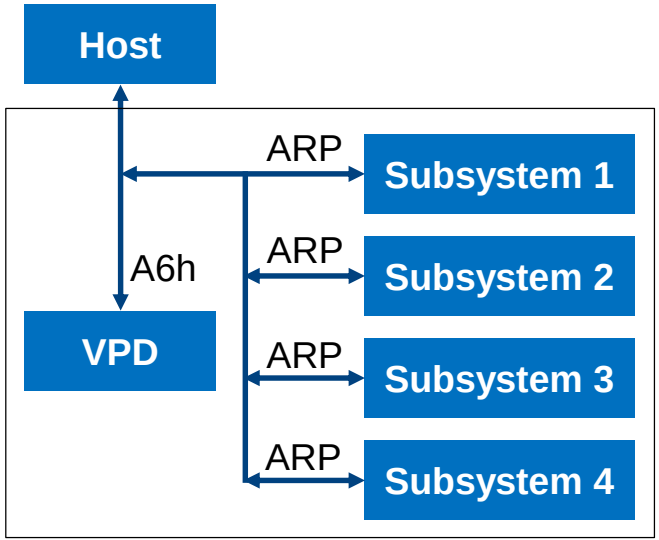


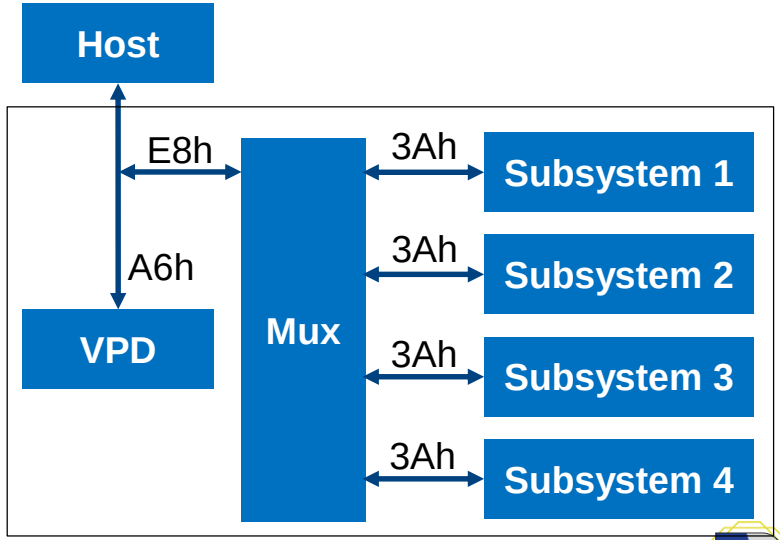
SMBus Topology for NVMe-MI™ 1.0



Multiple NVM Subsystems on a single SMBus Port

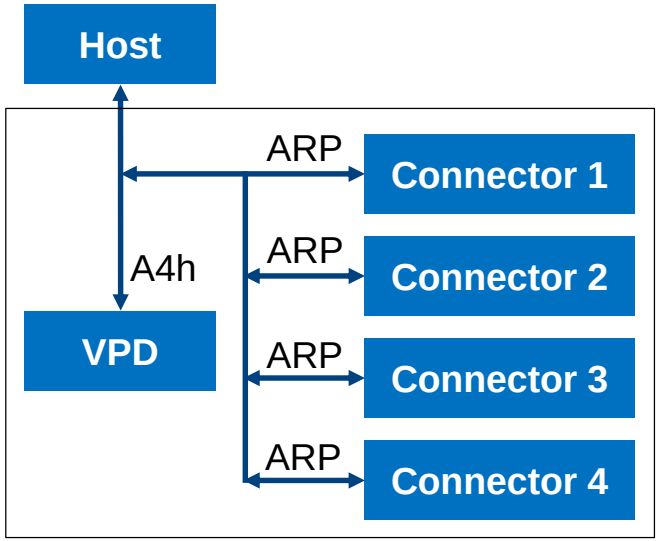
- Describe topology in new VPD MultiRecord
- Add UDID types for additional devices like Mux

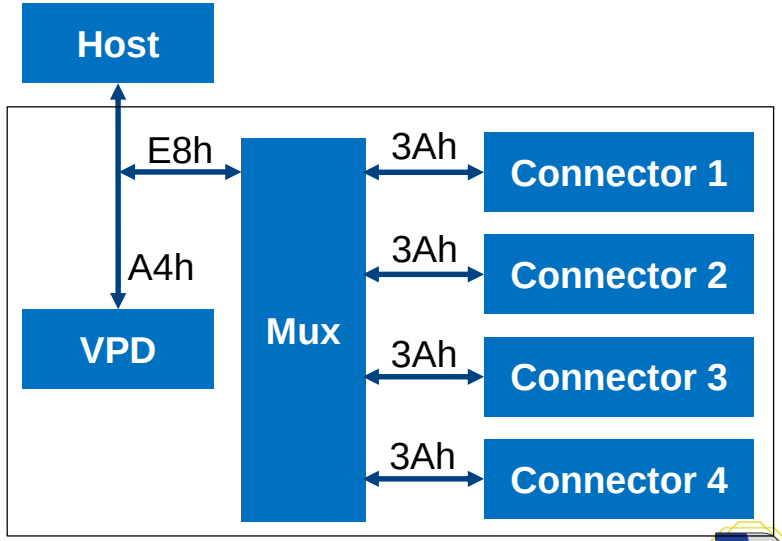




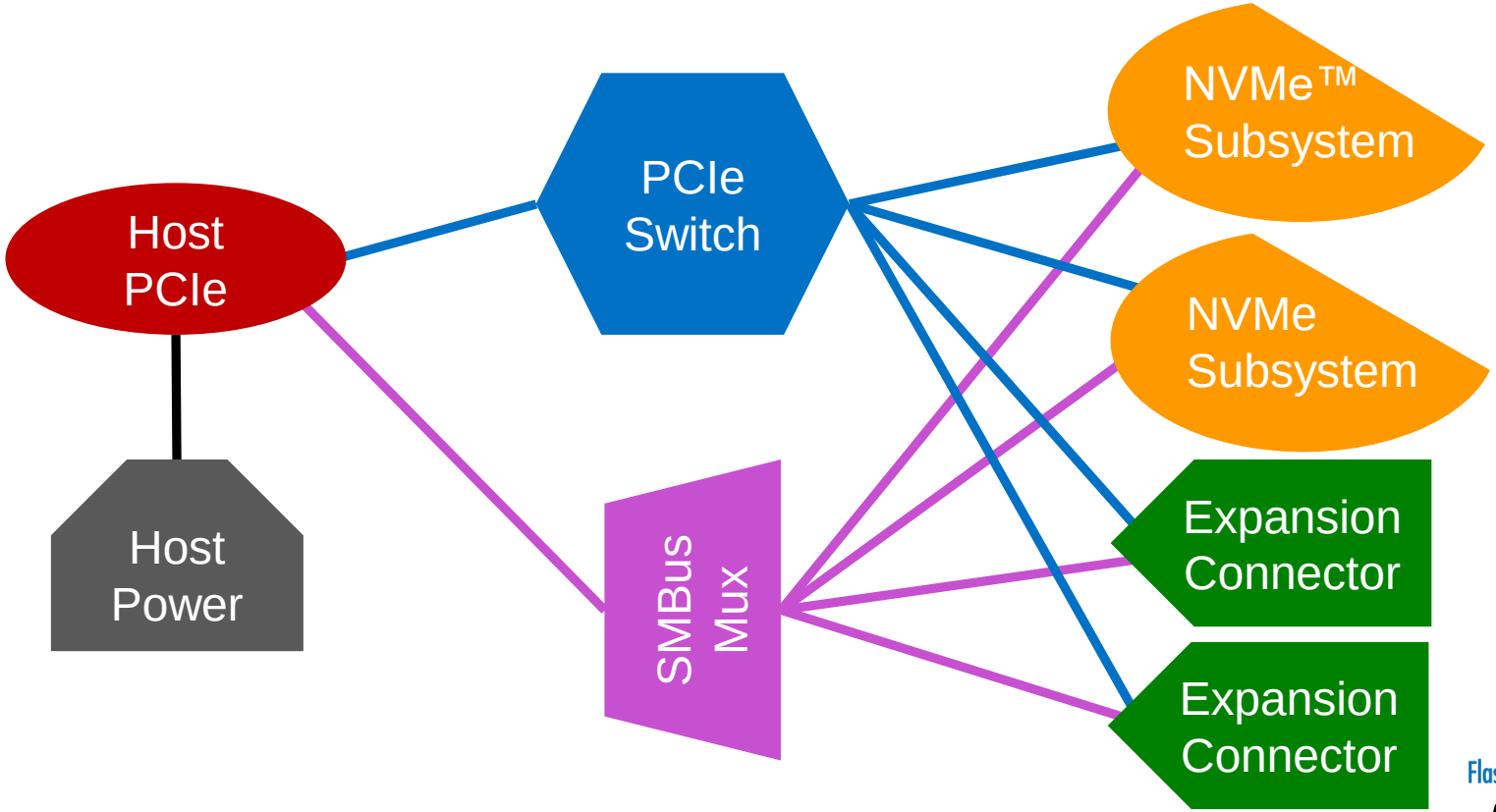
Support Expansion Connectors

- New VPD address to avoid conflicts with plugged in devices
- Optional Labels for each connector to assist technicians

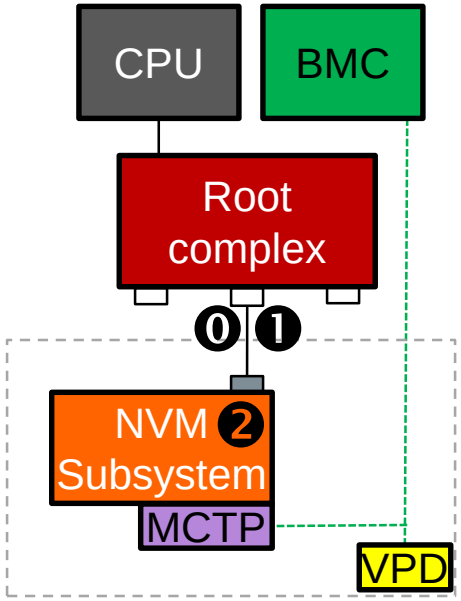




A Connection Graph Between Element Types

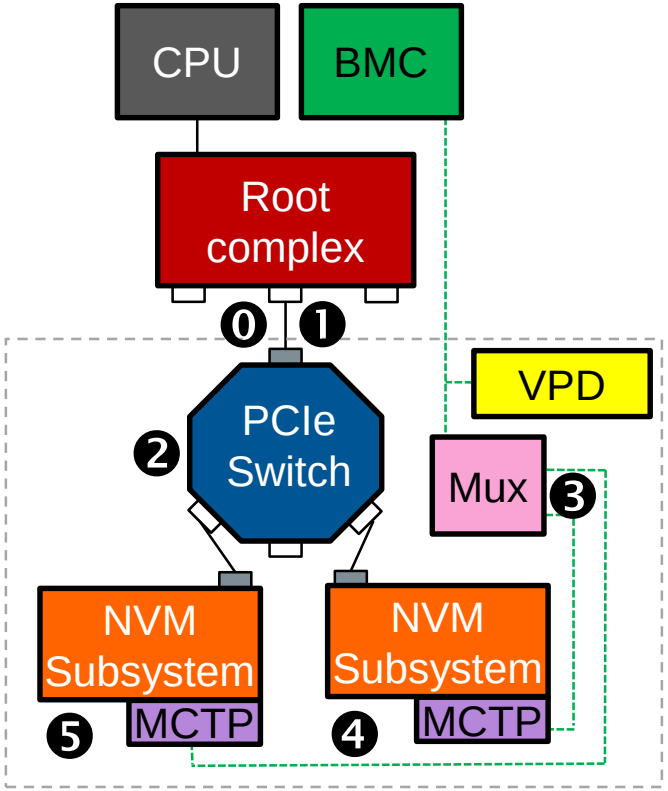


Single Port Example (35 bytes of 256B EEPROM)

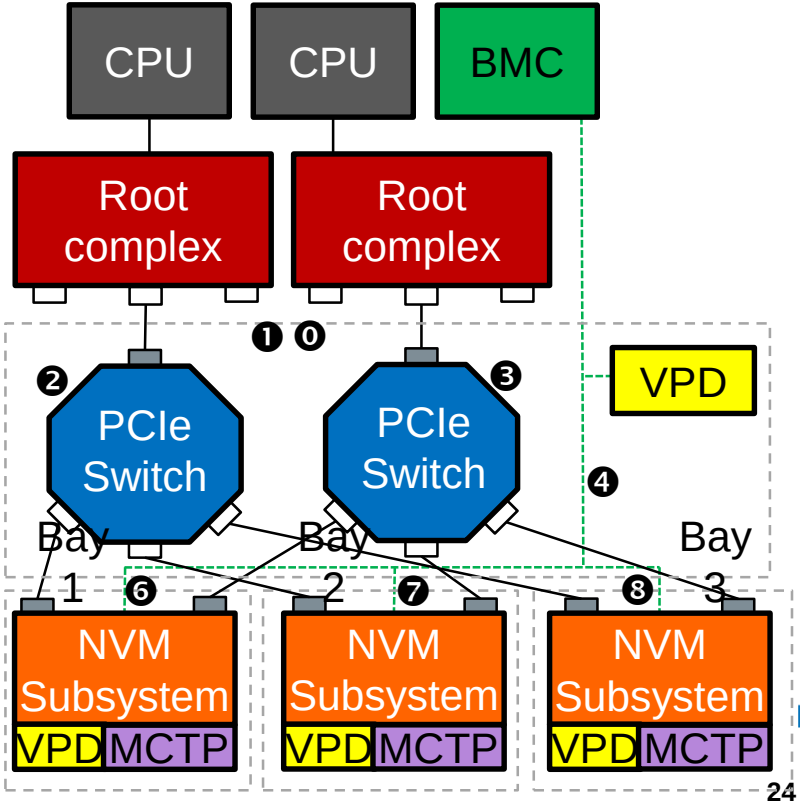


Header	Record: Element 0Dh	Record Format 82h	Record Length 23h	Record Chcksm 34h	Header Chcksm 75h	Version Number 00h	Rsvd 00h	Element Count 03h	
Element 0	Type: Host 01h	Element Length 08h	Form Factor 12h	SMBus Dest 02h	Link Options 00h	Link 0 Width 84h	Link 0 Start 00h	Link 0 Dest. 02h	
Element 1	Type: Power 02h	Element Length 08h	Thermal Load 0Fh	Vaux Load 32h	Rail Options 00h	Rail Voltage 78h	12V initial 08h	12V max 0Fh	
Element 2	Type: NVMe 09h	Element Length 13h	MCTP Address 3Ah	SMBus speed 01h	PCIe Ports 12h	Port 0 Speed 0Fh	Port 0 Flags 01h	Total NVM Capacity (MSB first) 000000000000000000000000h	

2 NVM Subsystems with Mux (82)



Dual Port with Expansion Connectors (78)



Summary

- NVMe-MI™ 1.0a is gaining market acceptance and is available in shipping products
- NVMe-MI 1.1 is nearing completion
  - Significant new features
    - In-band mechanism
    - Enclosure management
    - Support for multi NVM subsystem management
- It is time to start thinking about anchor features for NVMe-MI 1.2

Additional Material on NVMe-MI™

- BrightTALK Webinar
  - https://www.brighttalk.com/webcast/12367/282765/the-nvme-management-interface-nvme-mi-learn-whats-new
- Flash Memory Summit 2017
  - Slides: https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FA12_PartA.pdf
  - Video:
    - https://www.youtube.com/watch?v=daKL7tIvNII
    - https://www.youtube.com/watch?v=Daqj-XqlCo8
- Flash Memory Summit 2015
  - Slides: https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2015/20150811_FA11_Carroll.pdf
- Flash Memory Summit 2014
  - Slides: https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2014/20140804_SeminarF_Onufryk_Bolen.pdf
- NVMe-MI Specification
  - https://nvmexpress.org/resources/specifications/

References

MCTP Overview: http://dmtf.org/sites/default/files/standards/documents/DSP2016.pdf

MCTP Base Spec: https://www.dmtf.org/sites/default/files/standards/documents/DSP0236_1.3.0.pdf

MCTP SMBus/I2C Binding:
https://www.dmtf.org/sites/default/files/standards/documents/DSP0237_1.1.0.pdf

MCTP PCIe VDM Binding:
https://www.dmtf.org/sites/default/files/standards/documents/DSP0238_1.0.2.pdf

IPMI Platform Management FRU Information Storage Definition:
https://www.intel.la/content/www/xl/es/servers/ipmi/ipmi-platform-mgt-fru-infostorage-def-v1-0-rev-1-3-spec-update.html


nvm
EXPRESS®
Architected for Performance

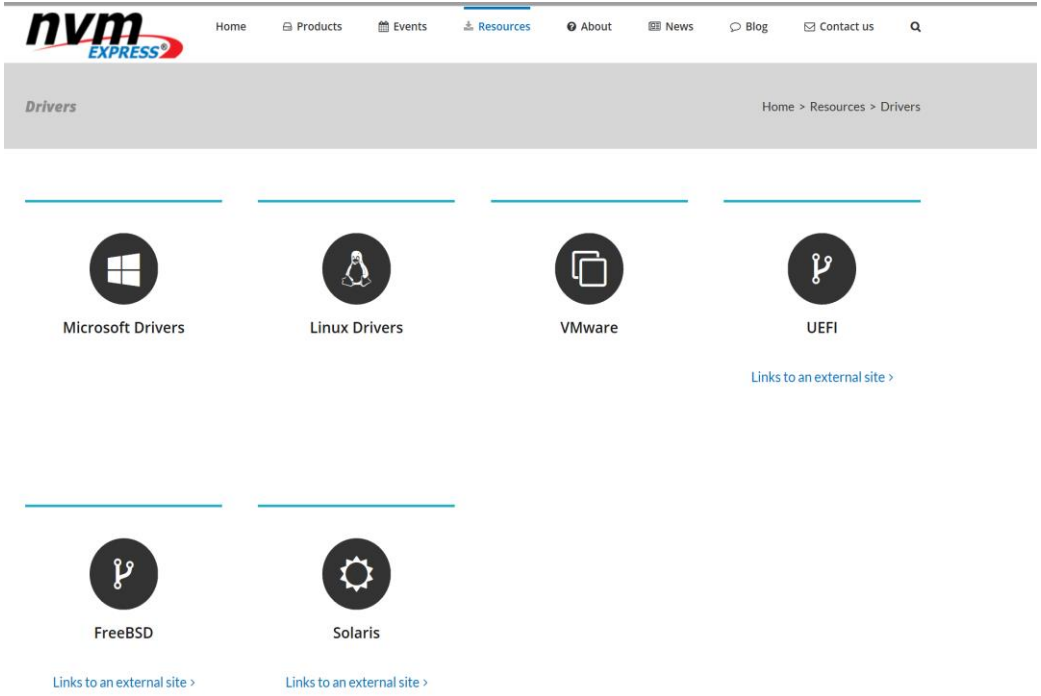
UEFI NVMe™ Drivers Update

Uma Parepalli, Marvell

NVMe™ Driver Ecosystem

Robust drivers available on all major platforms

NVM Express® Website – Drivers Home Page

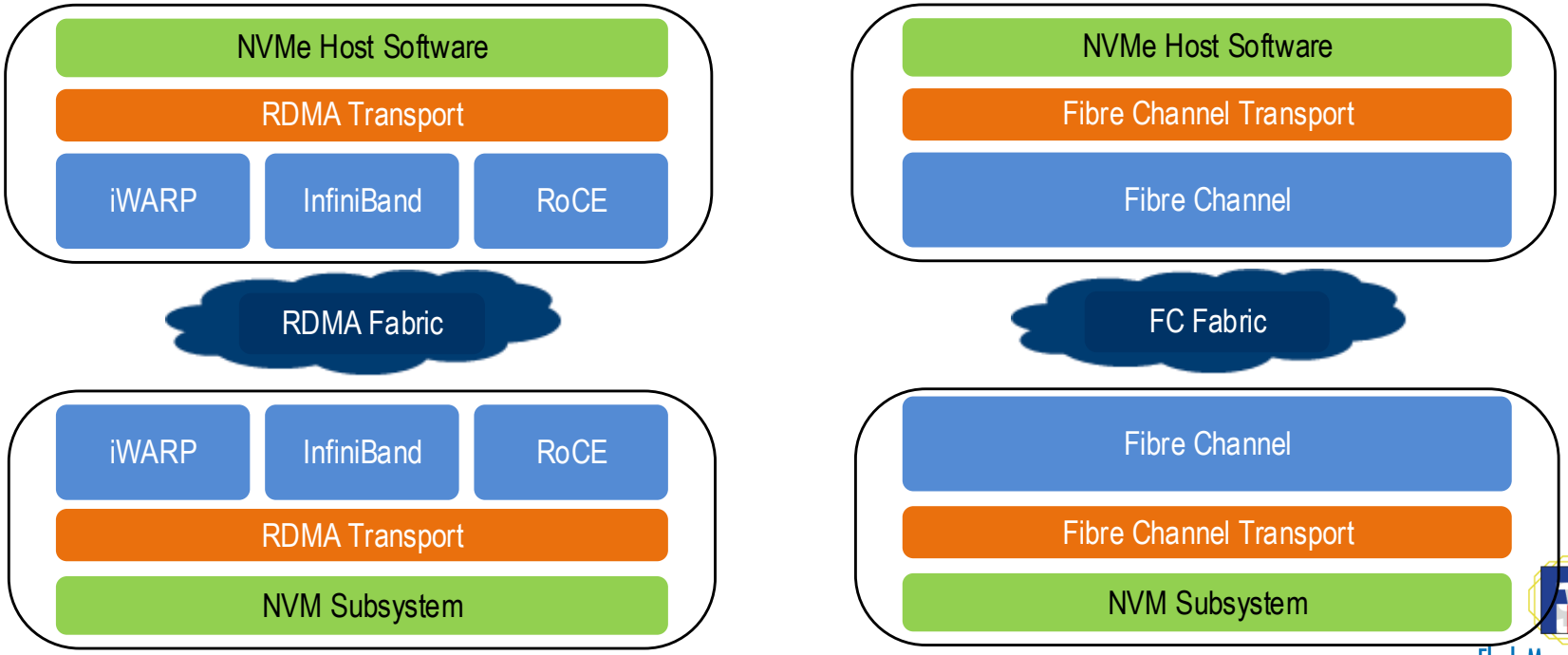


UEFI NVMe™ Drivers – What is new

- UEFI drivers available for a while on Intel platforms.
- ARM processor based systems now have built-in NVMe specification compliant UEFI driver and boot to Windows and Linux Operating Systems.

Linux NVMe™ over Fabrics Drivers

Supporting NVMe over RDMA, Fibre Channel, TCP and iWARP



Marvell (former Cavium) contributed the NVMe-oF™ Drivers to the Linux Upstream

33


nvm
EXPRESS®
Architected for Performance

Windows Inbox NVMe™ Driver

Lee Prewitt, Microsoft

Agenda

- New Additions for Spring Update (RS4)
- New Additions for Fall Update (RS5)
- Futures

NVMe™ Additions for Spring Update (RS4)

- DMA remapping support for StorNVMe
- F-State stair stepping when not in Modern Standby

New Additions for Fall Update (RS5)

- Asynchronous event request support for Namespace Change Notification
- Device Telemetry
- Support for extended log page

Futures*

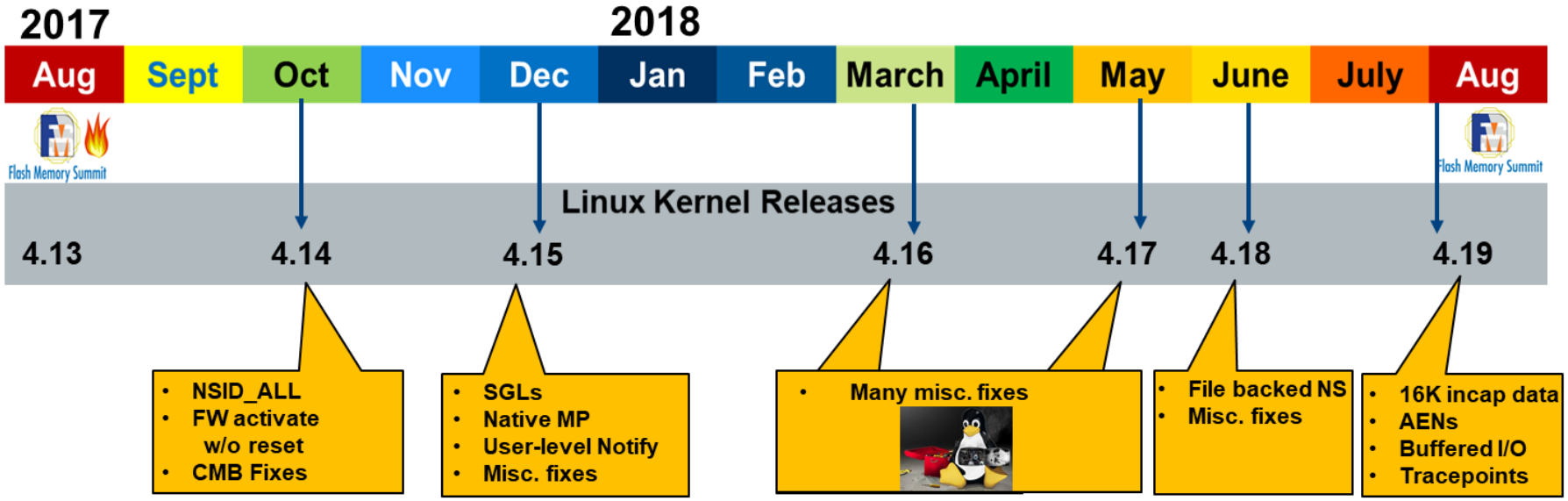
- D3 enabled by default on lowest power state
- Support for interface to Host Controlled Thermal Management
- Support for NVM Sets
- Support for Endurance Group Information
- Support for Namespace Management

*Not plan of record

Linux NVMe™ Driver Update

Dave Minturn, Intel Corp.

A Year in the Life of Linux NVMe™ Drivers



Projected NVMe™ Driver Features For Next 12 Months

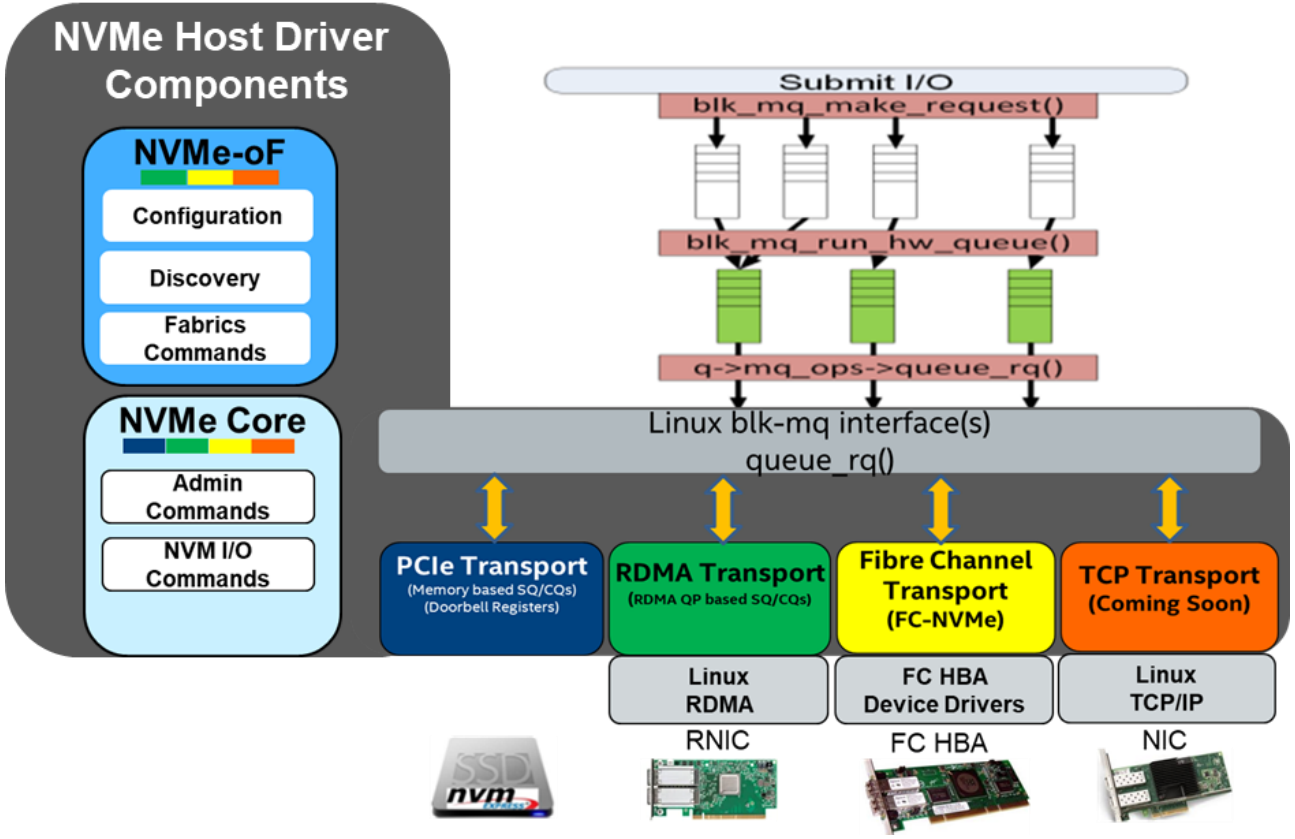
NVMe-oF™ Host/Target Driver functionality based on NVMe-oF 1.1 features

- NVMe/TCP Transport (available today to NVMe.org Driver WG members)
- Discovery Log AEN
- Flow Control Negotiation
- Authentication
- Transport SGLs and Error Codes

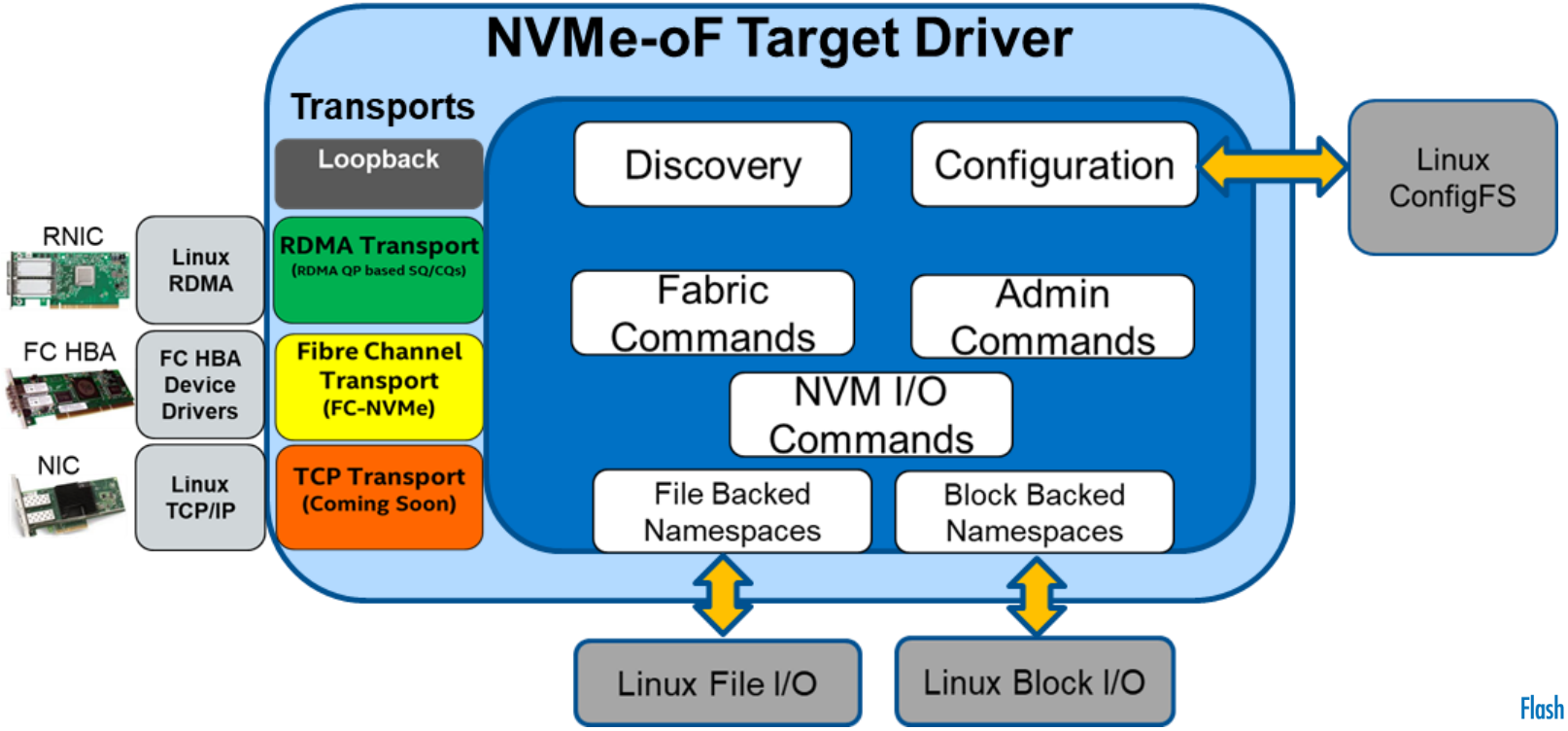
NVMe Host Core and PCIe transport functionality based on NVMe 1.4 features

- Asymmetric Namespace Access (ANA)
- Persistent Memory Region (PMR)
- Determinism and NVM Sets
- Host Memory Buffers

NVMe™ Host Driver Components



NVMe-oF™ Target Driver Components



Linux NVMe™ Driver References

- NVMe Specifications and Ratified TPs available publically at:
  http://nvmexpress.org/resources/specifications/
- NVMe Linux Drivers Sources
  www.kernel.org (mainline and stable)
- NVMe Linux Driver Reflector (for the latest patches and RFCs)
  https://lists.infradead.org/mailman/listinfo/linux-nvme
- NVMExpress.org Linux Fabrics Driver Working Group (members only)
  - Access to NVMe-oF Drivers based on non-public specifications


nvm
EXPRESS®
Architected for Performance

NVM Express® in vSphere Environment

Sudhanshu (Suds) Jain, VMware

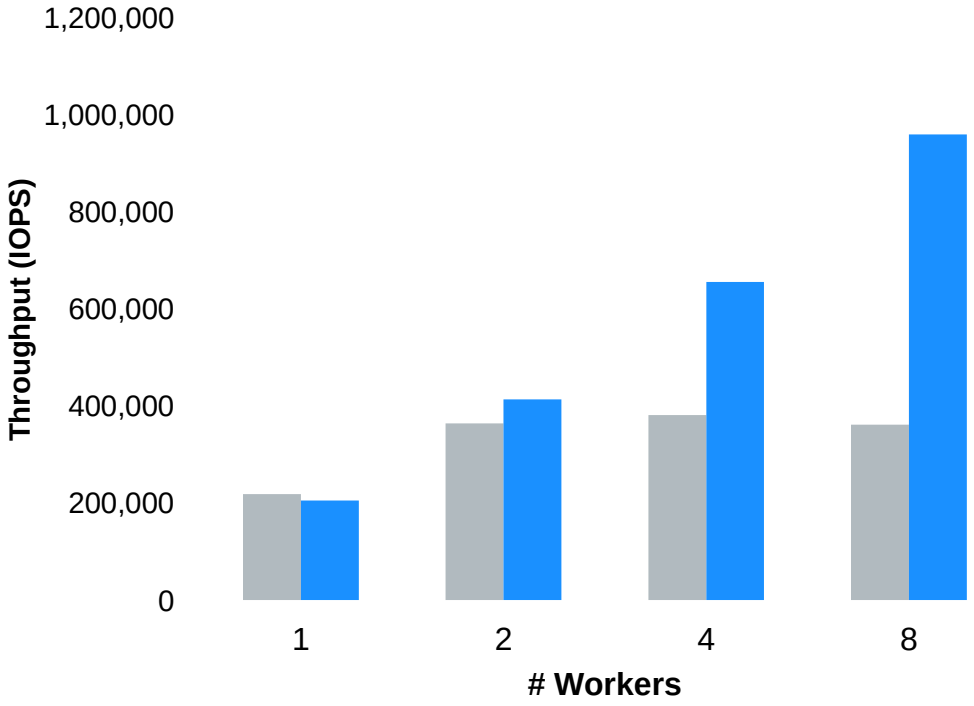
Agenda

- NVMe™ Driver EcoSystem in vSphere 6.7
- Future Direction

NVMe™ Focus @VMWare

	vSphere 6.5	vSphere 6.7	Future Direction
Driver	• Boot (UEFI) • Firmware Update • End-to-end protection • Deallocate/TRIM/Unmap • 4K • SMART, Planned hot-remove	• Performance enhancements • Extended CLI • Name space management • Async event error handling • Enhance diagnostic logs	• NVMe Over Fabric • Multiple fabric option • SR-IOV
Core Stack	• Reduced serialization • Locality improvements • vNVMe Adaption layer • Multiple completion worlds support in NVMe	• Optimized stack - Highly parallel execution for single path local NVMe devices • Reach target of 90%+ performance of device spec • NVMe Multi-pathing	• Next Generation Storage Stack with ultra-high IOPS • End-to-end NVMe Stack
Virtual Devices	• NVMe 1.0e spec • Hot-plug support • VM orchestration	• Performance improvements • Async mode support • unmap support	• Rev the specification • Parallel execution @backend • 4K Support • Scatter-gather support • Interrupt coalescing

NVMe™ Performance Boost



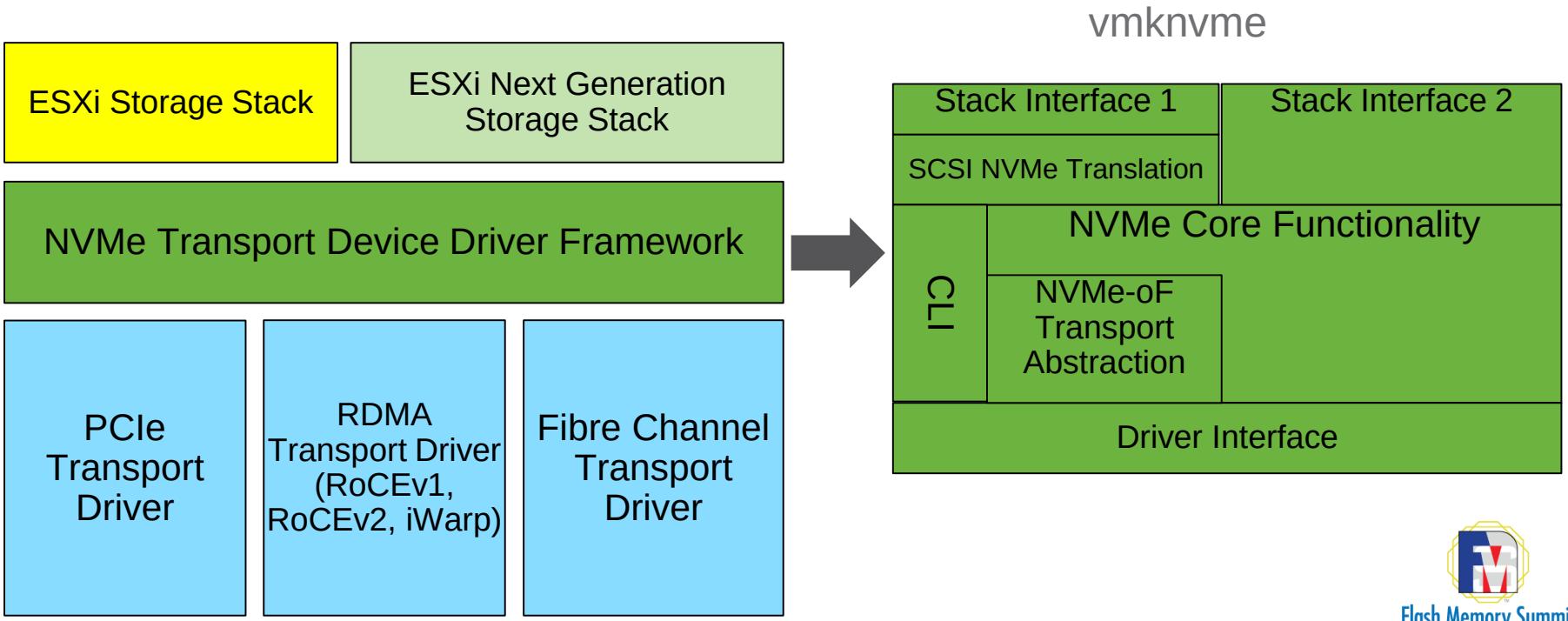
Hardware:

- Intel® Xeon® E5-2687W v3 @3.10GHz (10 cores + HT)
- 64 GB RAM
- NVM Express* 1M IOPS @ 4K Reads

Software:

- vSphere* 6.0U2 vs. Future prototype
- 1 VM, 8 VCPU, Windows* 2012, 4 VMDK eager-zeroed
- IOMeter:
  - 4K seq reads, 64 OIOs per worker, even distribution of workers to VMDK

(Future) NVMe™ Driver Architecture



NVMe™ Driver Ecosystem

- Available as part of base ESXi image from vSphere 6.0 onwards
  - Faster innovation with async release of VMware NVMe driver
- VMware led vSphere NVMe Open Source Driver project to encourage ecosystem to innovate
  - https://github.com/vmware/nvme
- Broad NVMe Ecosystem on VMware NVMe Driver

  https://www.vmware.com/resources/compatibility/search.php?deviceCategory=io
  - Close to 300 third party NVMe devices certified on VMware NVMe driver


nvm
EXPRESS®
Architected for Performance

Storage Performance Development Kit and NVM Express®

Jim Harris, Intel Data Center Group

Notices and disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- Some results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance..
- Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.
- 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.
- Optimization Notice: 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. Notice Revision #20110804.
- 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.
- 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 www.intel.com/benchmarks.
- Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.
- 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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
- The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.
- No computer system can be absolutely secure.
-
- *Other names and brands may be claimed as the property of others.

NVMe™ Software Overhead

- NVMe Specification enables highly optimized drivers
  - No register reads in I/O path
  - Multiple I/O queues allows lockless submission from multiple CPU cores in parallel
- But even best of class kernel mode drivers have non-trivial software overhead
  - 3-5us of software overhead per I/O
  - 500K+ IO/s per SSD, 4-24 SSDs per server
  - <10us latency with latest media (i.e. Intel Optane™ SSD)
- Enter the Storage Performance Development Kit
  - Includes polled-mode and user-space drivers for NVMe

Storage Performance Development Kit (SPDK)

- Open Source Software Project
  - BSD licensed
  - Source code: http://github.com/spdk
  - Project website: http://spdk.io
- Set of software building blocks for scalable efficient storage applications
  - Polled-mode and user-space drivers and protocol libraries (including NVMe™)
- Designed for NAND and latest generation NVM media latencies

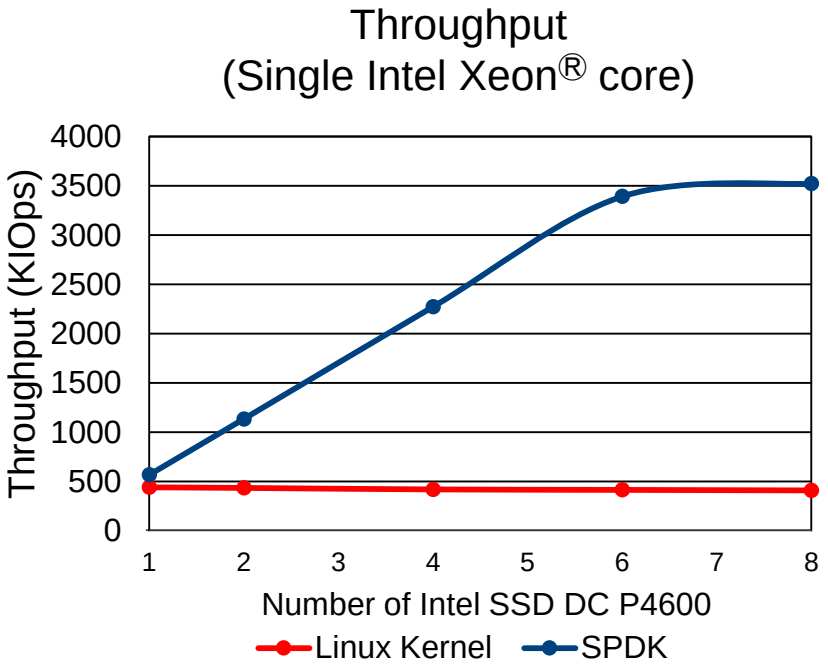
NVMe™ Driver Key Characteristics

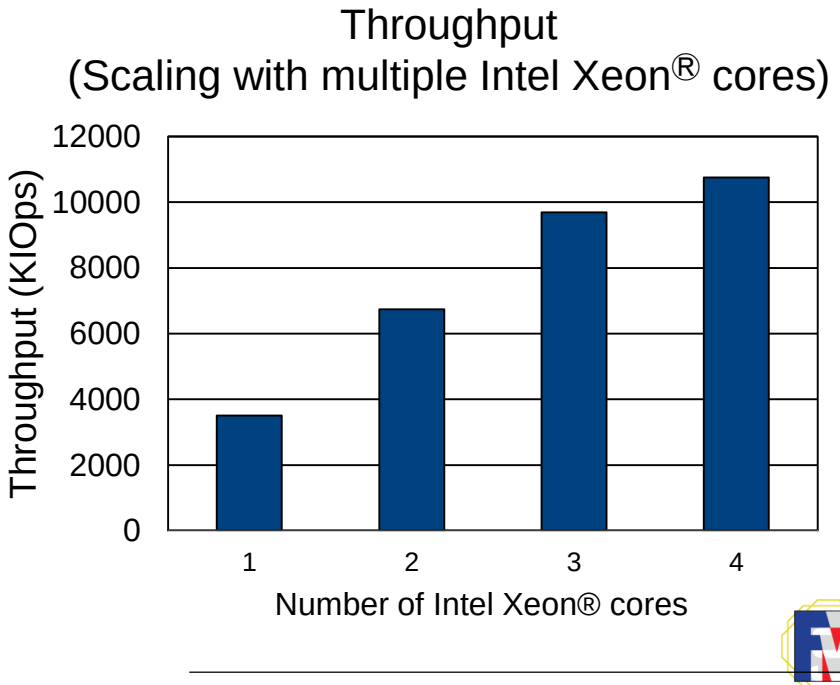
- Supports NVMe 1.3 spec-compliant devices
- Userspace Asynchronous Polled Mode operation
- Application owns I/O queue allocation and synchronization
- NVMe Features supported include:
  - End-to-end Data Protection
  - SGL
  - Reservations
  - Namespace Management
  - Weighted Round-Robin
  - Controller Memory Buffer
  - Firmware Update
  - Asynchronous Event Requests

NVMe™ Driver Key Characteristics

- Driver Features and Capabilities include:
  - Hotplug
  - Error Injection
  - Open Channel
  - Device Quirks
  - Configurable Timeouts
  - Configurable I/O Queue Sizes
  - Raw Command APIs
  - NVMe™ over Fabrics
  - fio plugin

NVMe™ Driver Performance Comparison





System Configuration: 2S Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz, 192GB DDR4 Memory, 6x Memory Channels per socket, 1 16GB 2667 DIMM per channel, Fedora 27, Linux kernel 4.15.15-300.fc27.x86_64, BIOS: HT enabled, p-states enabled, turbo enabled, SPDK 18.04, numjobs=1, direct=1**,** block size 4k 22 Intel® SSD DC P4600 (2 TB, 2.5in PCIe 3.1 x4, 3D1 TLC) 8 on socket 0 and 14 on socket 1.

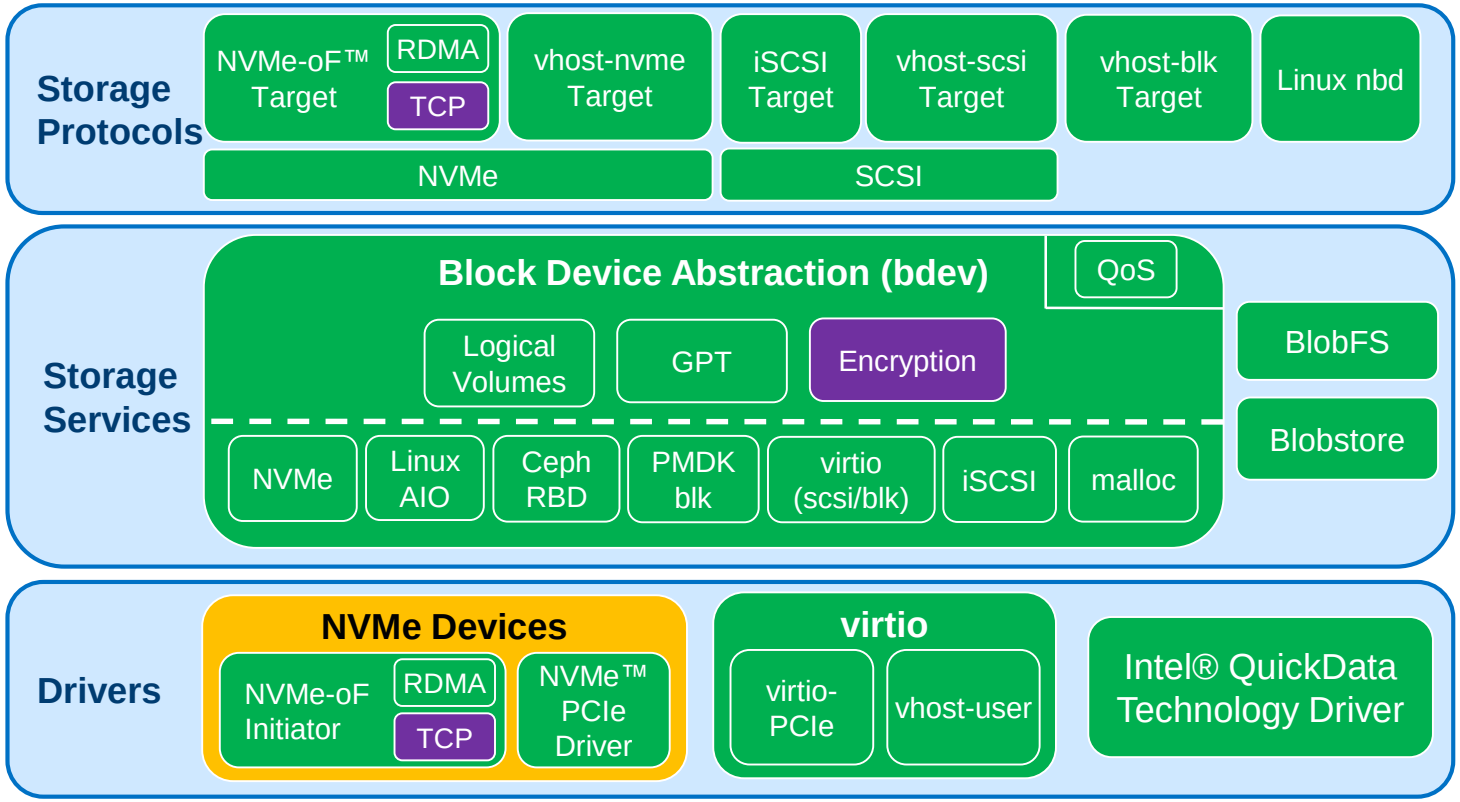
NVMe-oF™ Initiator

- Common API for local and remote access
  - Differentiated by probe parameters
- Pluggable fabric transport
  - RDMA supported currently (using libibverbs)
  - Allows for future transports (i.e. TCP, FC)

SPDK Architecture

SPDK 18.07

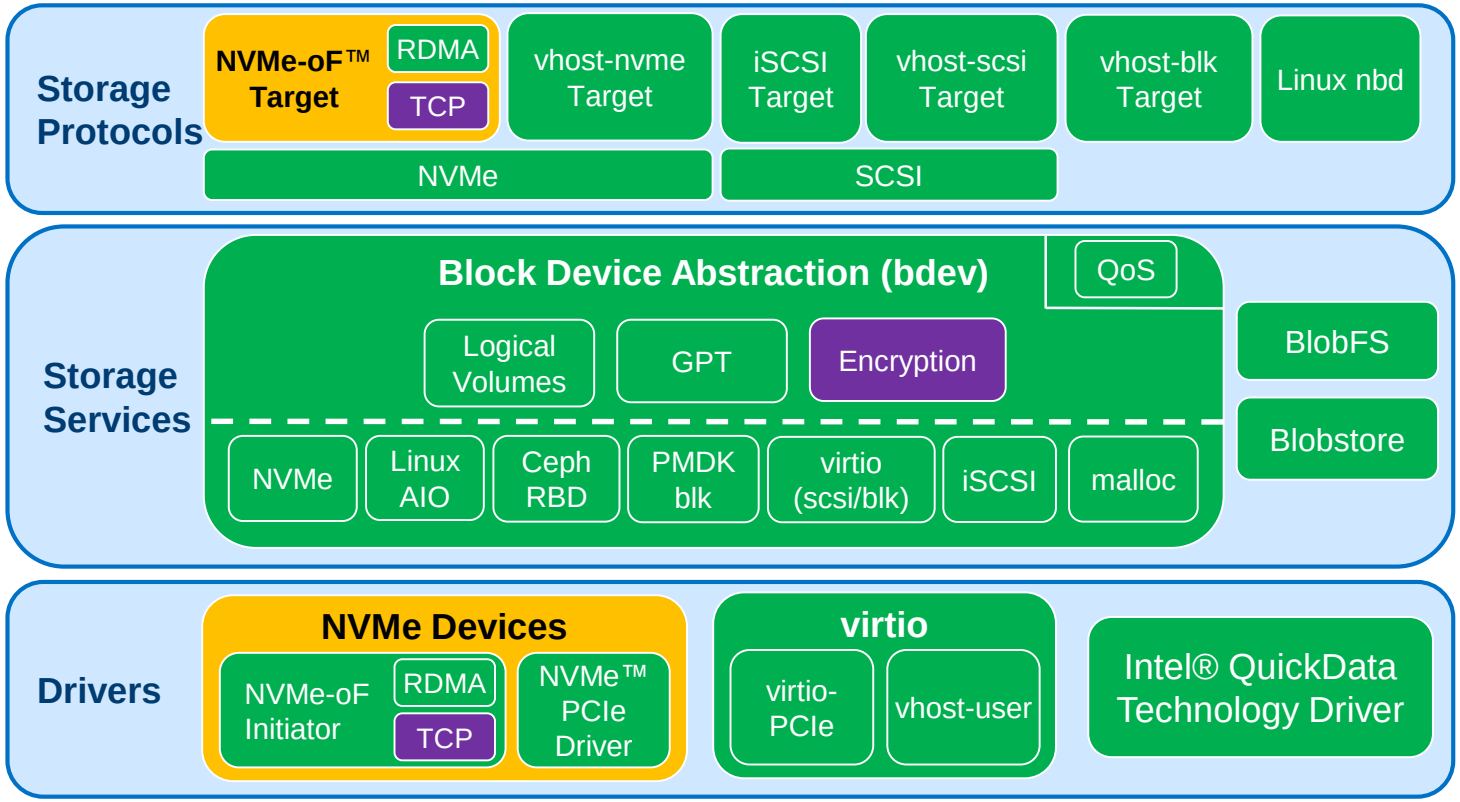
In Progress



SPDK Architecture

SPDK 18.07

In Progress



NVMe-oF™ Target

- Polled-mode userspace NVMe-oF target implementation
  - Pluggable fabric transport (similar to NVMe-oF initiator)
  - Presents SPDK block devices as namespaces
    - Locally-attached namespaces
    - Logical volumes
    - etc.
- SOFT-202-1 – Wednesday 3:20-5:45pm
  - Ben Walker – NVMe-oF: Scaling up with SPDK

Call to Action

- Check out SPDK!
  - Source code: http://github.com/spdk
  - Project website: http://spdk.io
    - Getting Started Guide (including Vagrant environment)
    - Mailing List
    - IRC
    - GerritHub