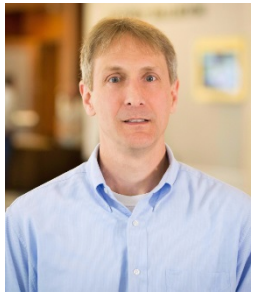
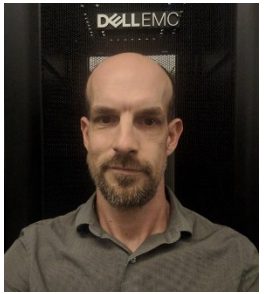
NVMe Management Interface (NVMe-MI)



Peter Onufryk
Microsemi Corp.
NVMe-MI Workgroup Chair



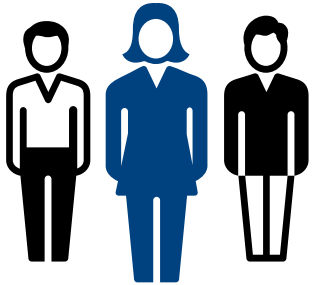
Austin Bolen
Dell EMC
NVMe-MI Workgroup Vice Chair

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

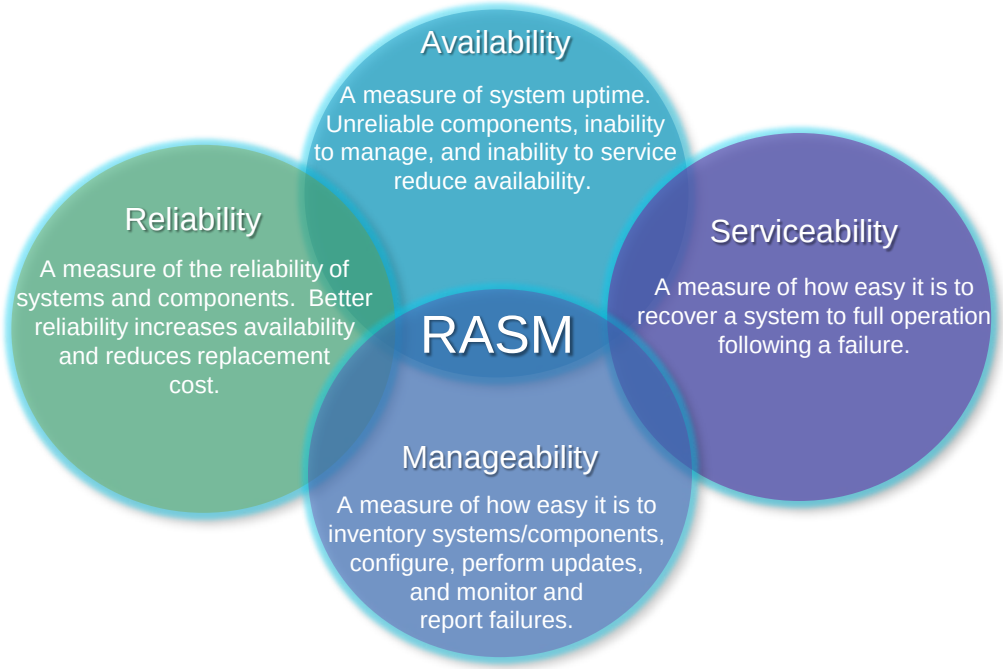
Chair: Ryan Holmqvist

facebook Microsoft intel cisco DELL EMC SEAGATE TOSHIBA

Micron ORACLE SAMSUNG Microsemi NetApp WD Western Digital

nvm EXPRESS®

RASM (Reliability, Availability, Serviceability, Manageability)



"Customers choose suppliers who provide the features that are important to them. Customers care about TCO (Total Cost of Ownership). Consequently, in the server space, MHz is not the only thing that's important: TCO is greatly affected by the RASM features of the servers. When server OEMs and users talk, their focus is RASM: Reliability, Availability, Serviceability, and Manageability. To a customer, RASM means dollars. Adding or improving on RASM reduces TCO.

The cost of downtime is extremely high. According to IMEX Research*, the average cost of an unplanned outage runs into the hundreds of thousand of dollars."(Reference 2)

Better RASM = Reduced TCO

Management Fundamentals

Pillars of Systems Management

- Inventory
- Configuration
- Monitoring
- Change Management

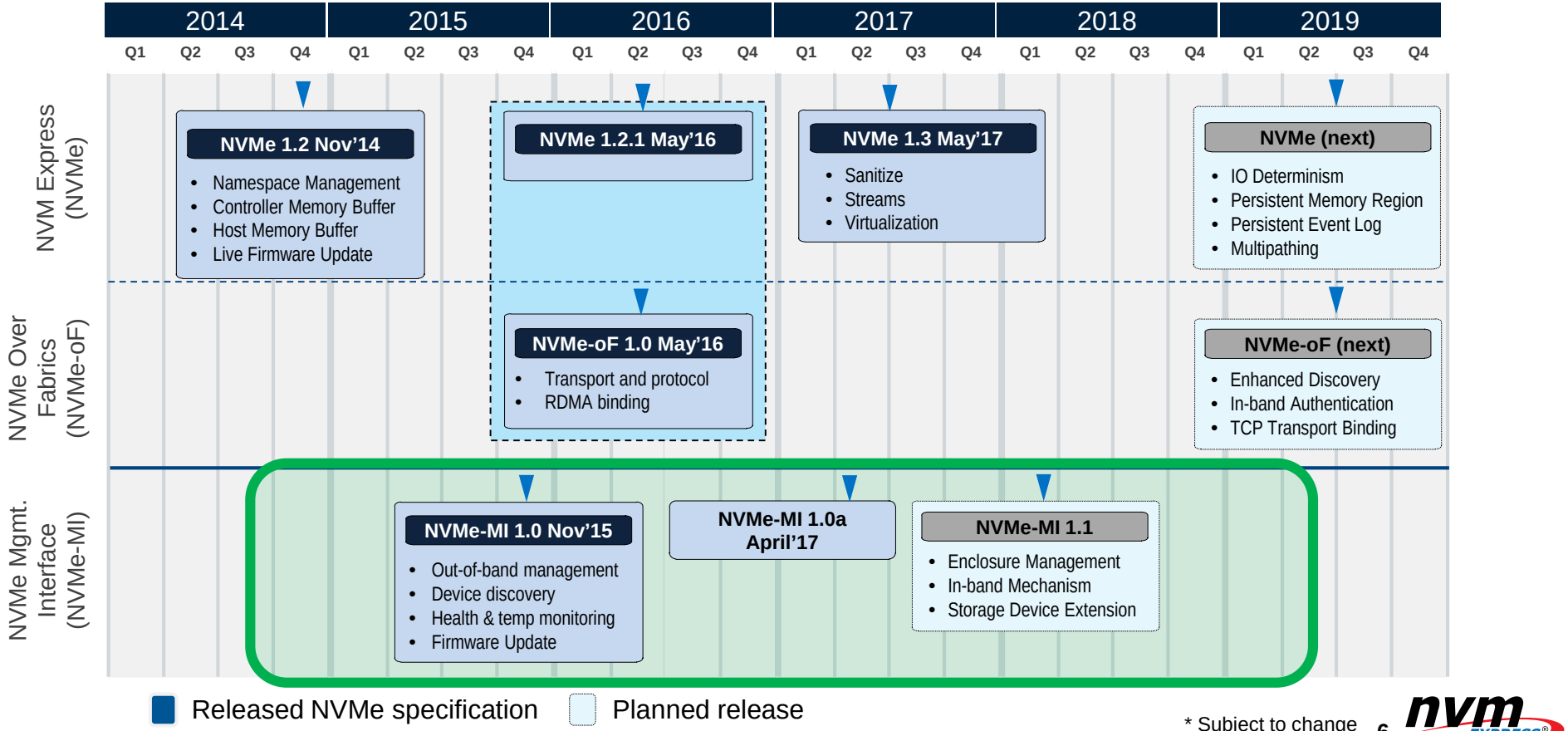
Management Operational Times

- Deployment (No OS)
- Pre-OS (e.g. UEFI/BIOS)
- Runtime
- Auxiliary Power
- Decommissioning

What is the NVMe Management Interface 1.0a?

A programmable interface that allows out-of-band management of an NVMe Storage Device Field Replaceable Unit

NVM Express Roadmap



* Subject to change

Benefits of NVMe-MI and Standardization

Benefit	OEM	Drive Vendor	End User
Clear requirements and specification	✓	✓	
Industry standard compliance program	✓	✓	
Industry standard tools	✓	✓	
Ability to source NVMe-MI drives from multiple vendors	✓		✓
Reduces need for drive vendors to develop proprietary management features		✓	
Lower TCO over life of NVMe Storage Device			✓
Allows product differentiation	✓	✓	

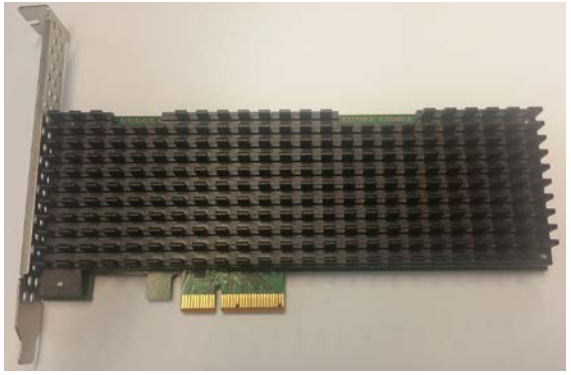
Sample End User Use Cases

Use Case	Benefit
Inventorying	Asset management. Re-provisioning systems. Track quality of components.
Health monitoring	Identify bad drives for quick replacement.
Wear monitoring	Replace drives nearing wear-out to avoid failure.
Temp. monitoring	Fan throttling reduces power, noise, and fan wear.
Power monitoring and configuration	Power throttling to save energy and cool system.
Perf. monitoring	Look for performance bottlenecks.
Configuring	Format drives for initial use. Crypto erase drives for re-provisioning or decommissioning.
Change Mgmt.	Update drive firmware for bug fixes and security patches.

Field Replaceable Unit (FRU)

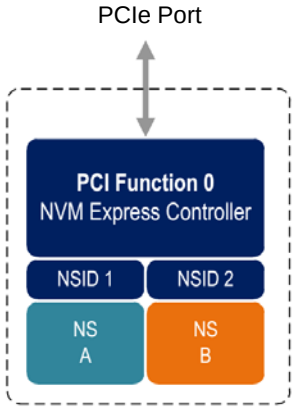
FRU Definition (Wikipedia)

A circuit board, part or assembly that can be quickly and easily removed from a computer or other piece of electronic equipment, and replaced by the user or a technician without having to send the entire product or system to a repair facility.

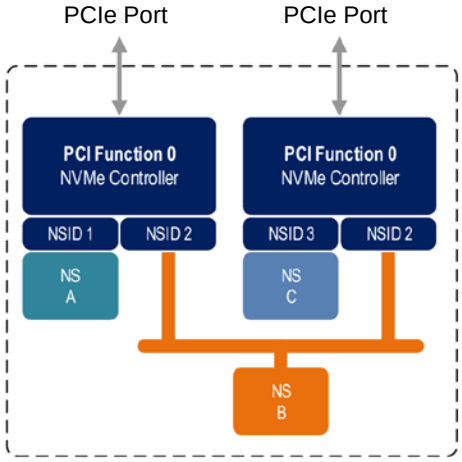


NVMe Architecture (review)

- **NVM Subsystem** - one or more controllers, one or more namespaces, one or more PCI Express ports, a non-volatile memory storage medium, and an interface between the controller(s) and non-volatile memory storage medium



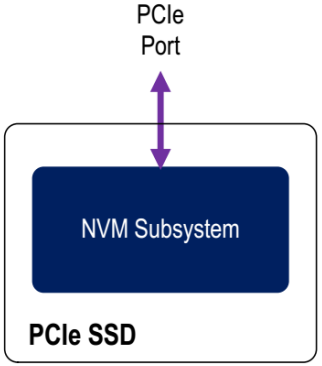
NVM Subsystem with One Controller and One Port



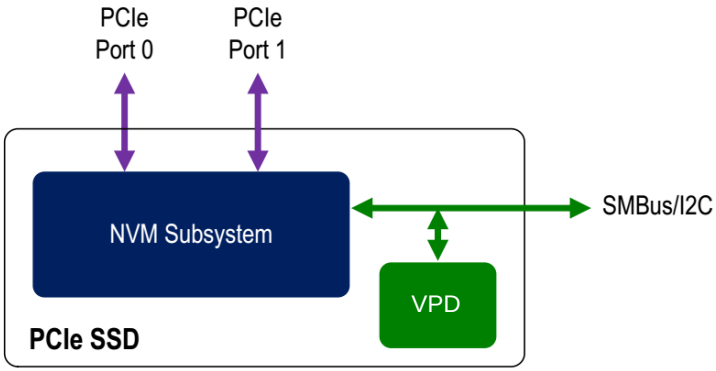
NVM Subsystem with Two Controllers and Two Ports

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

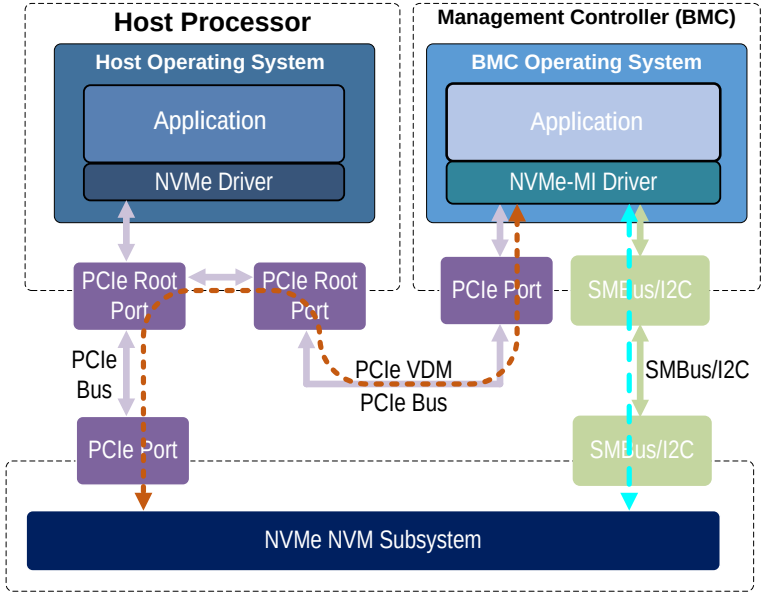
Vital Product Data (VPD)

- Utilizes IPMI Platform Management FRU Information Storage Definition with NVMe-MI extensions
- The VPD may be accessed using two methods
  - NVMe-MI commands over MCTP
  - SMBus/I2C interface using I2C operations as defined by IMPI Platform Management FRU Information Storage Definition

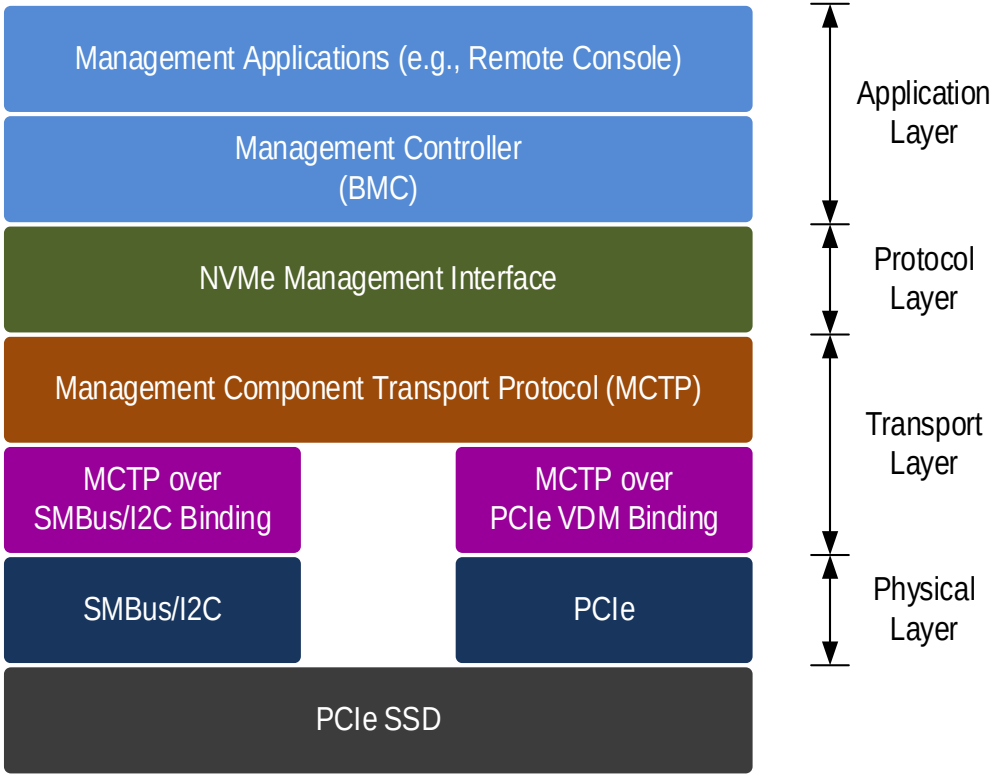
VPD Elements
Common Header
Product Info Area (optional)
NVMe MultiRecord Area
NVMe PCIe Port MultiRecord Area
Internal Use Area (optional)
Chassis Info Area (optional)
Board Info Area (optional)

Out-of-Band Management and NVMe-MI

- **Out-of-Band Management** – Management that operates with hardware resources and components that are ***independent of the operation system control***
- **NVMe Out-of-Band Management Interfaces**
  - SMBus/I2C
  - PCIe Vendor Defined Messages (VDM)
  - IPMI FRU Data (VPD) accessed over SMBus/I2C

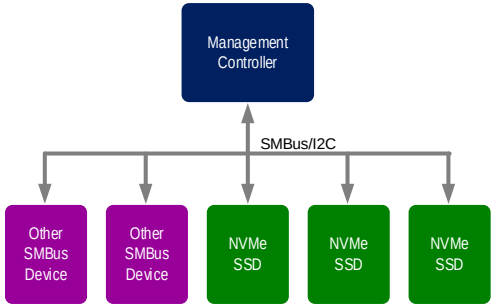


NVMe-MI Protocol Layering



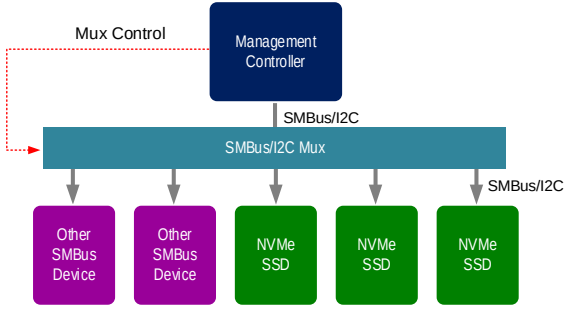
SMBus/I2C Topologies and Addressing

Shared SMBus/I2C



Requires Unique SMBus/I2C addresses

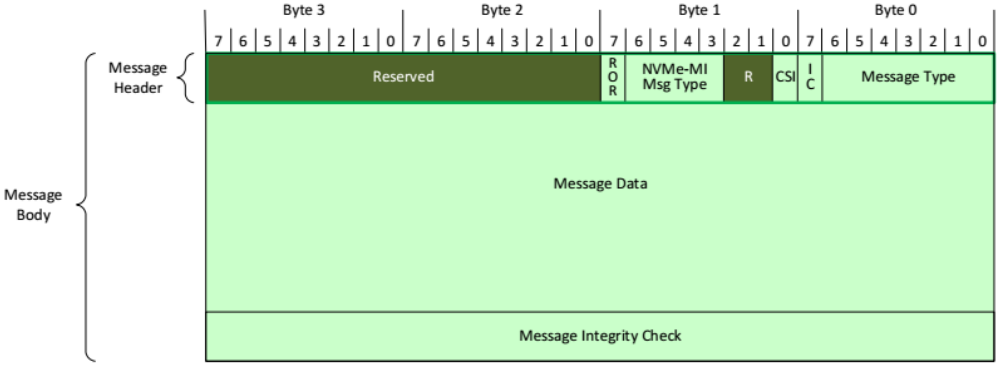
Segmented SMBus/I2C



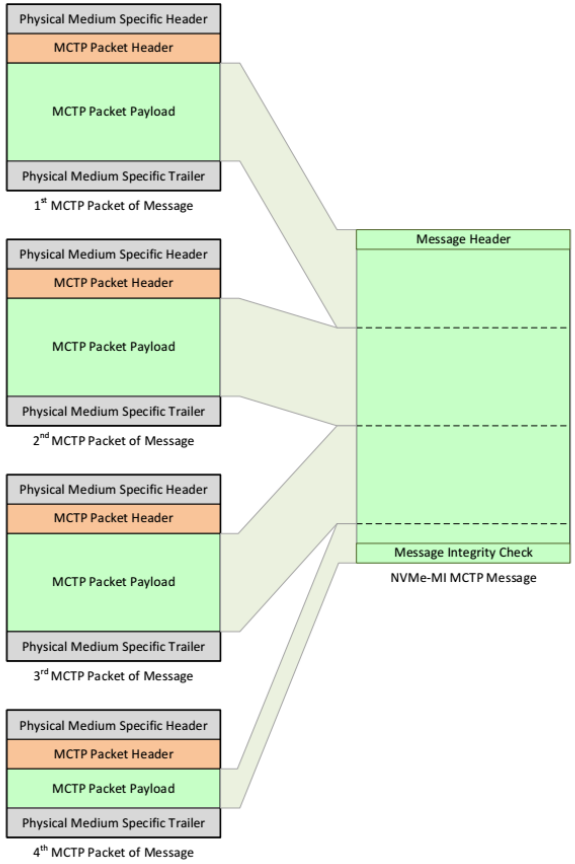
Repeated SMBus/I2C Addresses Supported

- During Auxiliary Power (if supported)
  - I2C serial EEPROM read/write access at default SMBus/I2C address 0xA6, but may be modified using ARP
- During Main Power
  - MCTP Endpoint at default SMBus/I2C address 0x3A, but may be modified using ARP
  - I2C serial EEPROM read/write access
    - If auxiliary power was provided, then SMBus/I2C address shall be maintained if modified using ARP; otherwise, the default address is 0xA6
    - SMBus/I2C address may be modified using ARP

NVMe-MI Message

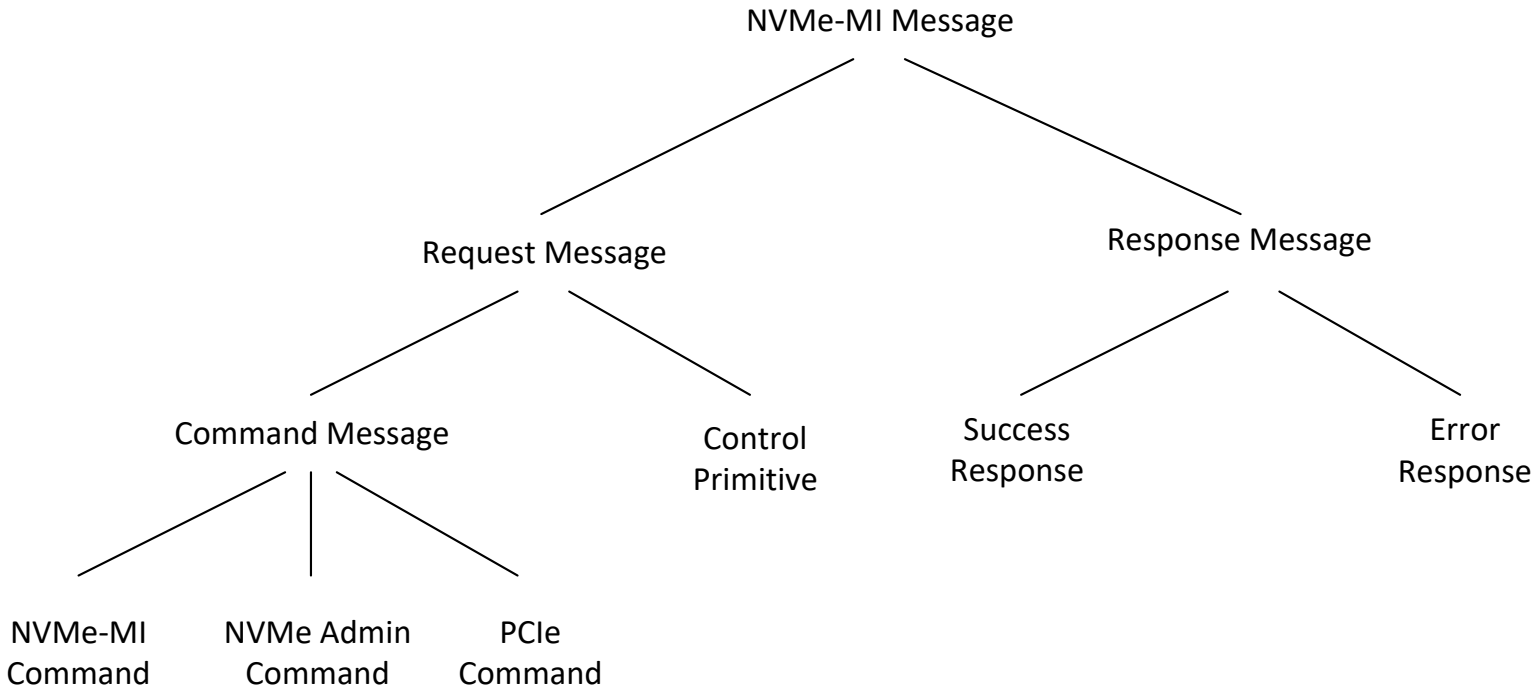


NVMe-MI Message

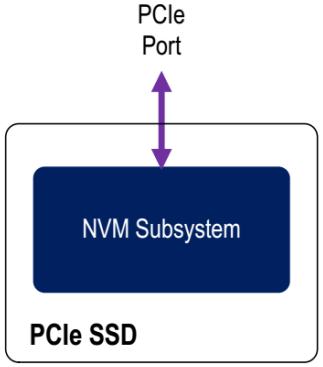


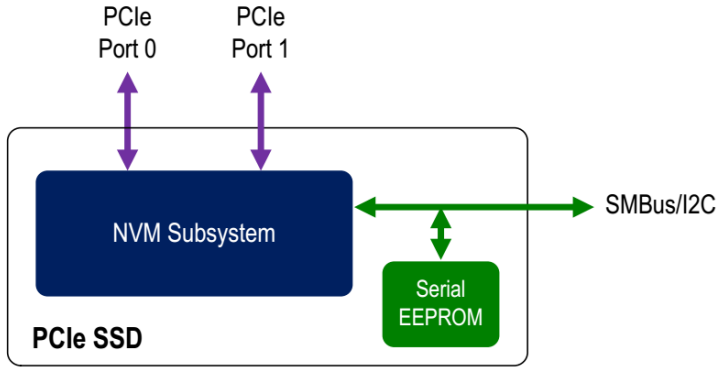
NVMe-MI MCTP Message Assembly

NVMe-MI Message Taxonomy

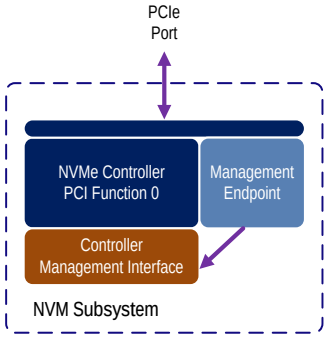


NVMe Storage Device

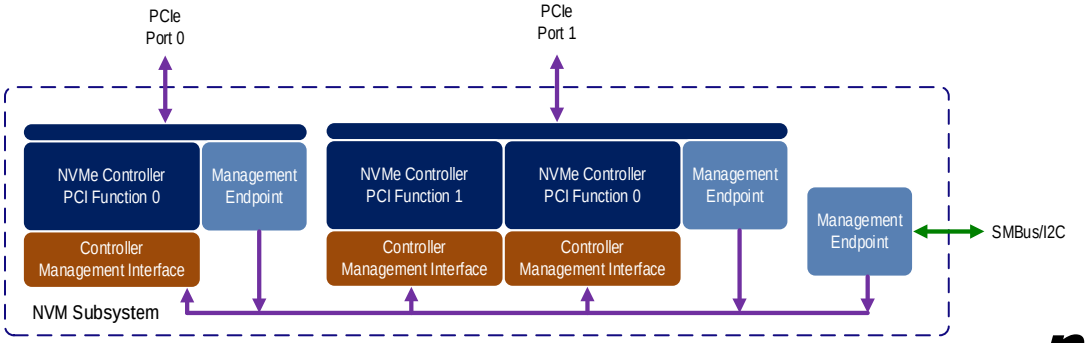




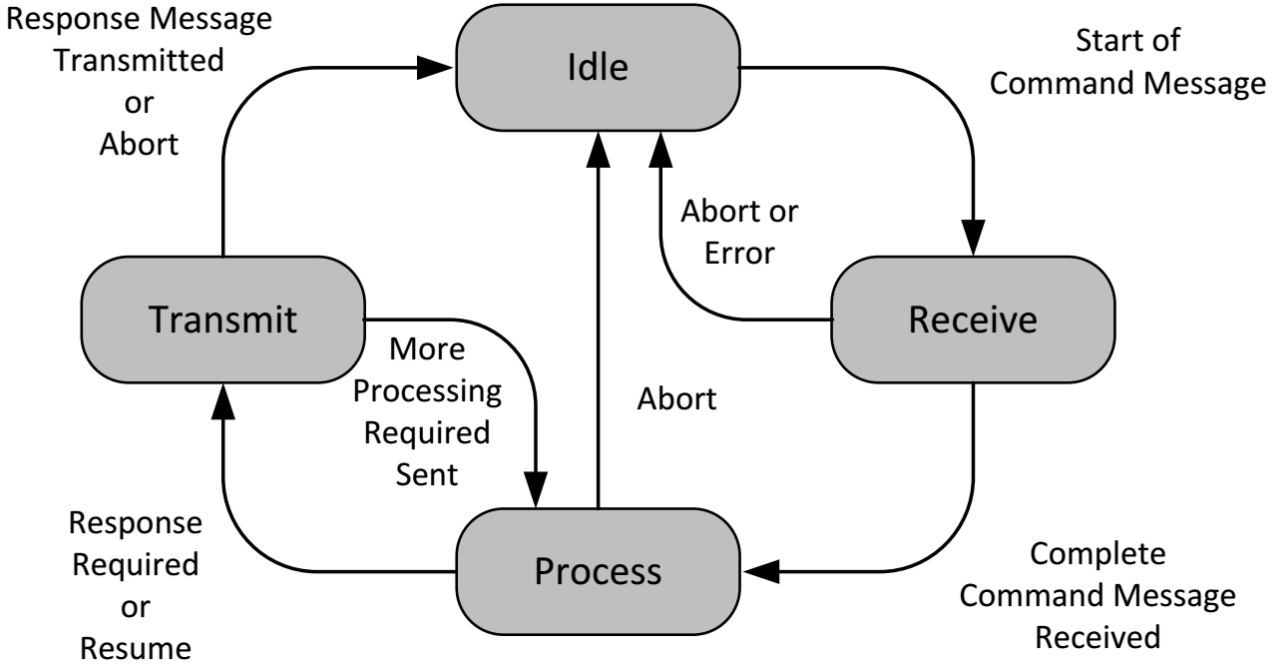
Single Ported PCIe SSD



Dual Ported PCIe SSD with SMBus/I2C



Command Servicing State Diagram for Command Slots



Control Primitives

Control Primitive	Description
Pause	Suspend transmission
Resume	Resume paused transmission
Abort	Reinitialize command slot
Get State	Retrieve state (e.g., errors) associated with a command slot
Replay	Retransmit response message for last command message processed in a command slot

NVMe-MI 1.0a Command Set Overview

Command Type	Command
NVMe Management Interface Specific Commands	Read NVMe-MI Data Structure
	NVM Subsystem Health Status Poll
	Controller Health Status Poll
	Configuration Get
	Configuration Set
	VPD Read
	VPD Write
	Reset
	Vendor Specific
PCIe Command	PCIe Configuration Read
	PCIe Configuration write
	PCIe I/O Read
	PCIe I/O Write
	PCIe Memory Read
	PCIe Memory Write
	Vendor Specific

Command Type	Command
NVMe Admin Commands	Firmware Activate/Commit
	Firmware Image Download
	Format NVM
	Get Features
	Get Log Page
	Identify
	Namespace Management
	Namespace Attachment
	Security Send
	Security Receive
	Set Features
	Vendor Specific

NVMe Management Interface Specific Commands

Command	O/M*	Description
Read NVMe-MI Data Structure	M	Retrieve information about the NVM Subsystem, Management Endpoint, or NVMe Controllers • NVM Subsystem Information • Port Information • Controller Information • Optional Commands Supported
NVM Subsystem Health Status Poll	M	Used to efficiently determine changes in health status attributes associated with the NVM Subsystem (e.g., Unrecoverable error, reset required, PCIe status, Controller SMART / Health Information, composite temperature, composite, and controller status)
Controller Health Status Poll	M	Efficiently determines changes in health status attributes associated with one or more Controllers in the NVM Subsystem
Configuration Get	M	Get NVMe-MI configuration parameter (e.g., SMBus/I2C frequency and MCTP transmission unit size)
Configuration Set	M	Set NVMe-MI configuration parameter
VPD Read	M	Read Vital Product Data (VPD)
VPD Write	M	Write Vital Product Data (VPD)
Reset	O	Reset NVM Subsystem

*O = Optional, M=Mandatory

NVMe Admin Commands

Command	O/M*	Description
Firmware Activate/Commit	O	Verifies that a valid firmware image has been downloaded and commits that revision to a specific firmware slot
Firmware Image Download	O	Download all of a portion of a firmware image for a future update to the controller
Format NVM	O	Low level format of the NVM media associated with one or more Namespaces
Get Features	M	Get NVMe configuration parameter
Set Features	O	Set NVMe configuration parameter
Get Log Page	M	Retrieve NVMe log page
Identify	M	Retrieve information about the Controllers, Namespaces, or NVM Subsystem
Namespace Management	O	Create or delete a Namespace
Namespace Attachment	O	Attach or detach a Namespace from a Controller
Security Send	O	Transfer command/data associated with security protocol
Security Receive	O	Transfer command/data associated with security protocol

*O = Optional, M=Mandatory

PCIe Commands

Command	O/M*	Description
PCIe Configuration Read	O	Read PCI Express configuration space
PCIe Configuration Write	O	Write PCI Express configuration space
PCIe I/O Read	O	Read PCI Express I/O space
PCIe I/O Write	O	Write PCI Express I/O space
PCIe Memory Read	O	Read PCI Express memory space (BAR memory & MMIO)
PCIe Memory Write	O	Write PCI Express memory space (BAR memory & MMIO)

*O = Optional, M=Mandatory

NVMe-MI Operational Times

Power State	Main Power	Auxiliary Power
Powered Off	Off	Off
Auxiliary Power	Off	On
Main Power	On	On
Main Power with No Auxiliary Power	On	Off

Power States

Operation	Powered Off	Auxiliary Power	Main Power (with Auxillary Power)	Main Power with No Auxiliary Power
VPD I2C Access	Not Supported	Supported	Supported	Inplementation Specific
SMBus/I2C MCTP Access	Not Supported	Optional[1]	Supported	Supported
PCIe MCTP Access	Not Supported	Not Supported	Supported	Supported

NOTES:
1. An implementation that supports SMBus/I2C MCTP Access during Auxiliary Power may support a subset of commands during this power state. The commands that are supported are implementation specific.

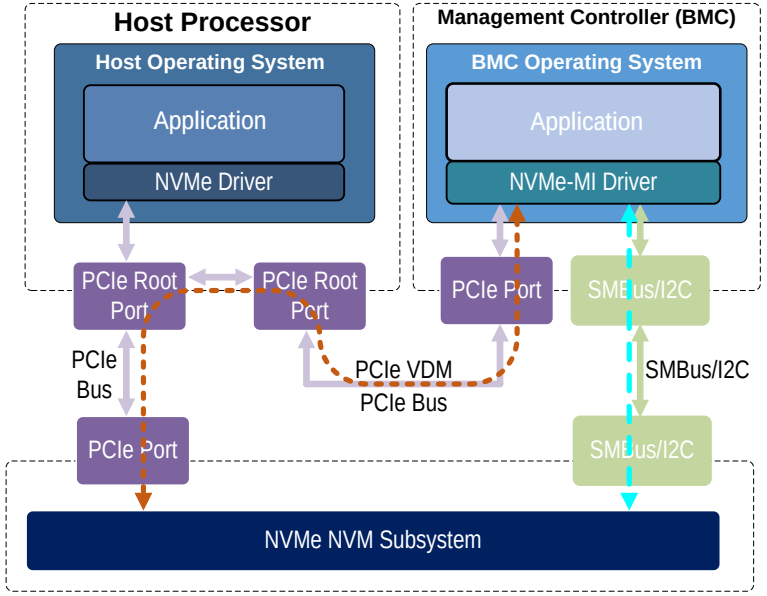
Operations Supported During Power States

New Features Targeted for NVMe-MI 1.1

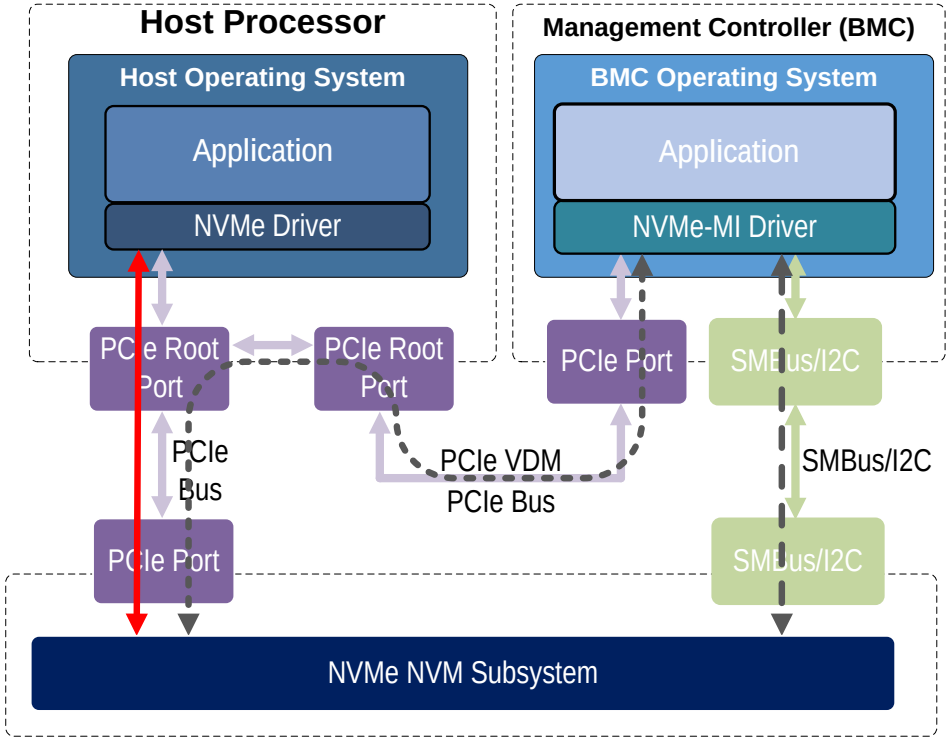
- In-Band NVMe-MI
- Enclosure Management
- NVMe Storage Device Enhancement

Out-of-Band Management and NVMe-MI

- **Out-of-Band Management** – Management that operates with hardware resources and components that are ***independent of the operation system control***
- **NVMe Out-of-Band Management Interfaces**
  - SMBus/I2C
  - PCIe Vendor Defined Messages (VDM)
  - IPMI FRU Data (VPD) accessed over SMBus/I2C

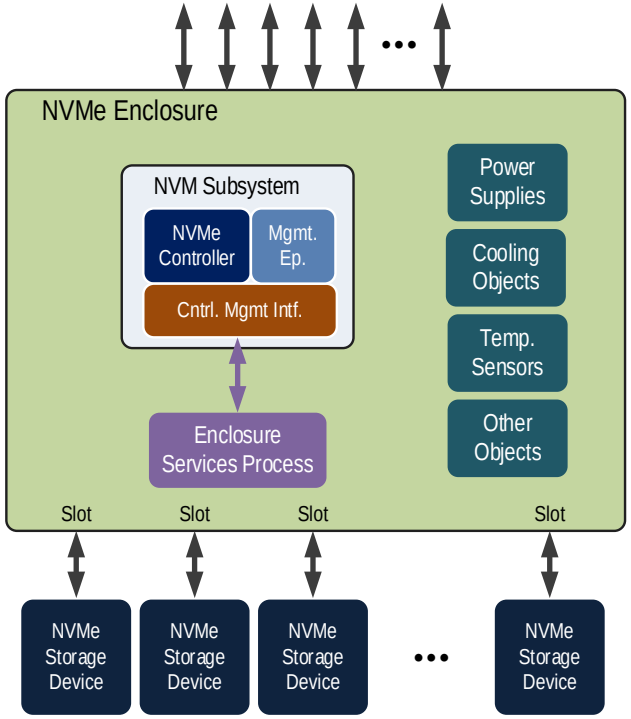


In-Band Management and NVMe-MI



- 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

Example Enclosure



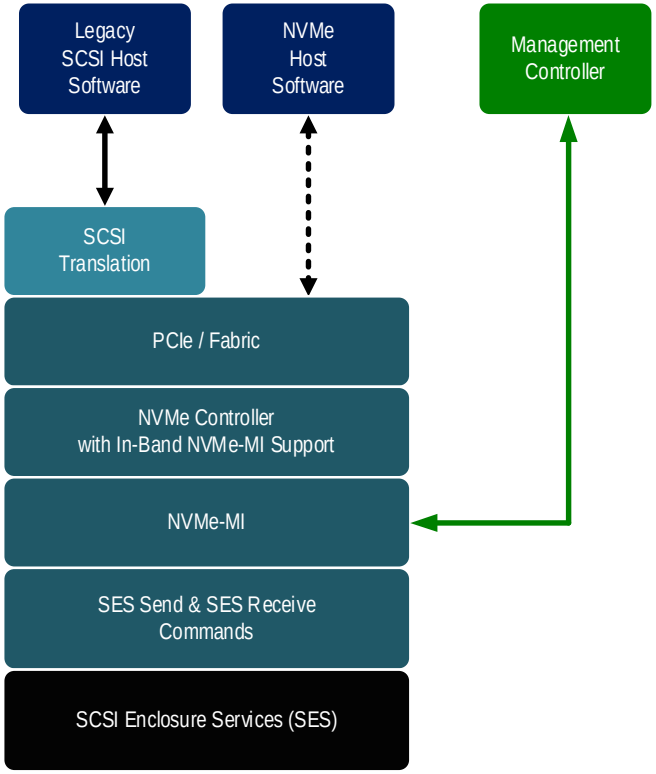
Enclosure Management

- Native PCIe Enclosure Management (NPEM)
  - Transport specific basic enclosure management
  - Submitted to the PCI-SIG Protocol Workgroup (PWG) on behalf of the NVMe Management Interface Workgroup
  - Approved by PCI-SIG on August 10, 2017
- SES Based Enclosure Management
  - Technical proposal being developed in NVMe-MI workgroup
  - Comprehensive enclosure management

SES Based Enclosure Management

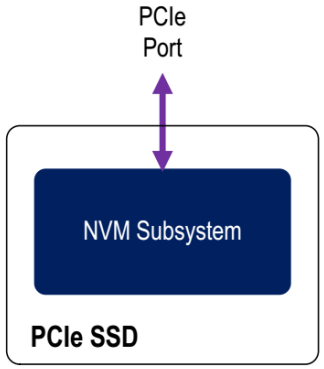
- SCSI Enclosure Services (SES) is a standard developed by T10 for management of enclosures using the SCSI architecture
- While the NVMe and SCSI architectures differ, the elements of an enclosure and the capabilities required to manage these elements are the same
  - Example enclosure elements: power supplies, fans, display or indicators, locks, temperature sensors, current sensors, voltage sensors, and ports
- NVMe-MI leverages SES for enclosure management
  - SES manages the elements of an enclosure using control and status diagnostic pages transferred using SCSI commands (SCSI SEND DIAGNOSTIC & SCSI RECEIVE DIAGNOSTIC RESULTS)
  - NVMe-MI uses these same control and status diagnostic pages, but transfers them using the SES Send and SES Receive commands

Enclosure Management Protocol Layering

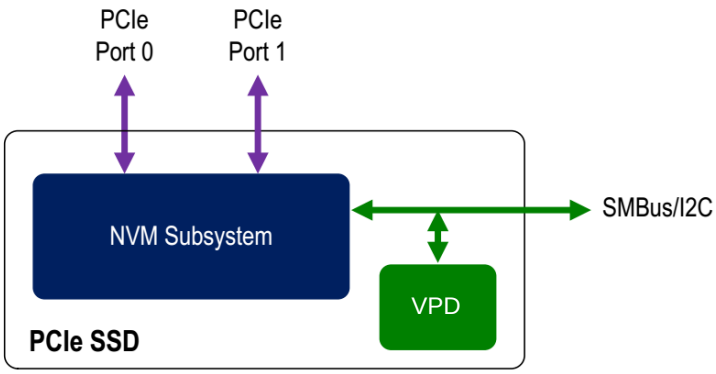


NVMe-MI 1.0a NVMe Storage Device

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

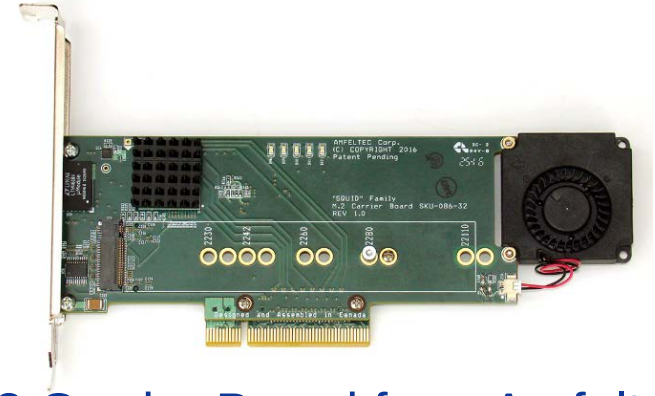


Single Ported PCIe SSD



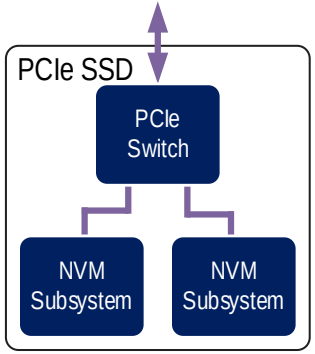
Dual Ported PCIe SSD with SMBus/I2C

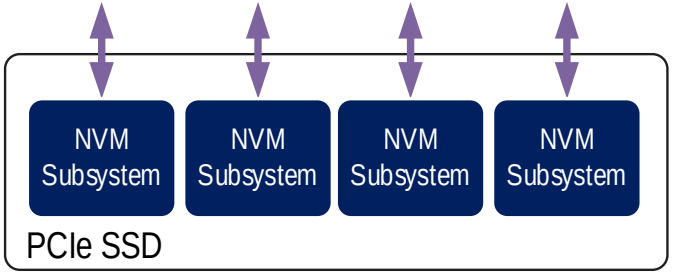
NVMe Storage Device with Multiple NVM Subsystems



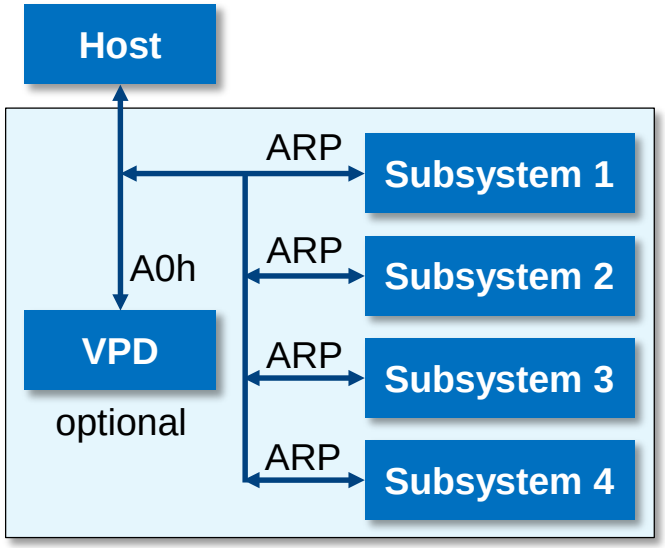
M.2 Carrier Board from Amfeltec

ANA Carrier Board from Facebook

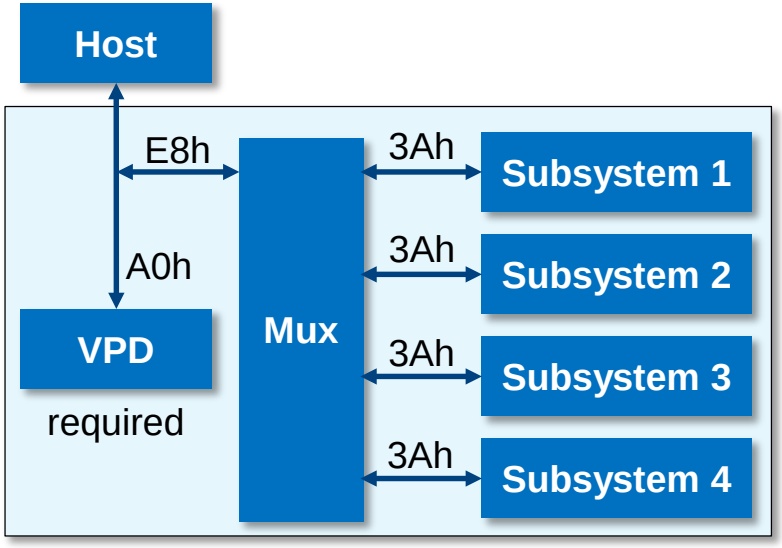




Multiple NVM Subsystems on an NVMe Storage Device and I2C/SMBus Topologies



Shared SMBus / I2C



Segmented SMBus / I2C

NVMe-MI Ecosystem

- Commercial test equipment and conformance tests exist 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)
  - 6 devices have passed compliance testing and are on the NVMe-MI Integrators List
- Servers are shipping that support NVMe-MI

Summary

- NVMe-MI 1.0a has been released
  - Focused on managing NVMe Storage Devices (e.g., SSDs)
  - SSDs and systems are shipping that support NVMe-MI 1.0a
- NVMe-MI 1.1 is nearing completion
  - Technical work is scheduled for completion this year and a ratified specification is expected in Q1’18
  - Key new features in NVMe-MI 1.1
    - In-band NVMe-MI
    - Enclosure Management
    - NVMe Storage Device Enhancements

References

1. NVMe/NVMe-MI - http://nvmexpress.org/
2. RASM - https://software.intel.com/en-us/articles/rasm-a-primer-for-isv-applications-engineers
3. RASM - http://www.ni.com/white-paper/14410/en/
3. Manageability - http://www.ni.com/white-paper/14415/en/
4. Reliability - http://www.ni.com/white-paper/14412/en/
5. Serviceability - http://www.ni.com/white-paper/14414/en/
6. Availability - http://www.ni.com/white-paper/14413/en/

Don’t Miss the Next Webcast!

Join us to learn about the evolution of the NVMe storage protocol and what’s in store for its future, in 2018 and beyond in our next webcast titled:

The Evolution and Future of NVMe

Tuesday, December 19th at 9:00am PT / 12:00pm ET.

https://www.brighttalk.com/webcast/12367/290529

David Allen, NVMe Board Member and Seagate's Senior Director of Marketing

Dr. J Metz, Board Member, and R&D Engineer for the Office of the CTO for Cisco

Questions?

Visit www.nvmexpress.org for more information on NVM Express technology

Follow us:

Twitter: https://twitter.com/NVMexpress
LinkedIn: https://www.linkedin.com/company/11106843/
Facebook: https://www.facebook.com/NVM-Express-137395516813022/


nvm
EXPRESS®
Architected for Performance