Architected for Performance

Overview of the NVMe Management Interface Specification

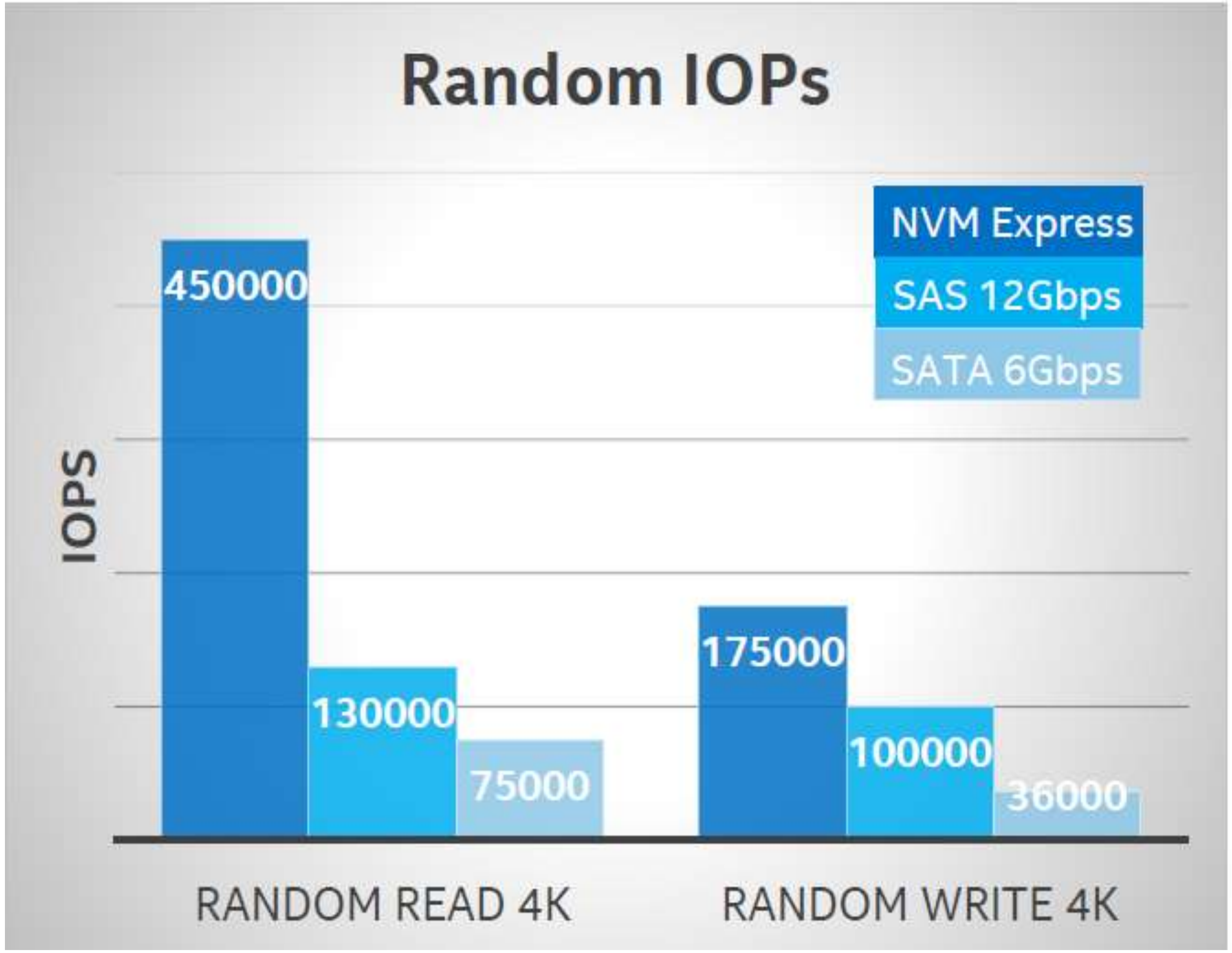
SNIA SDC INDIA,
May 26th 2016

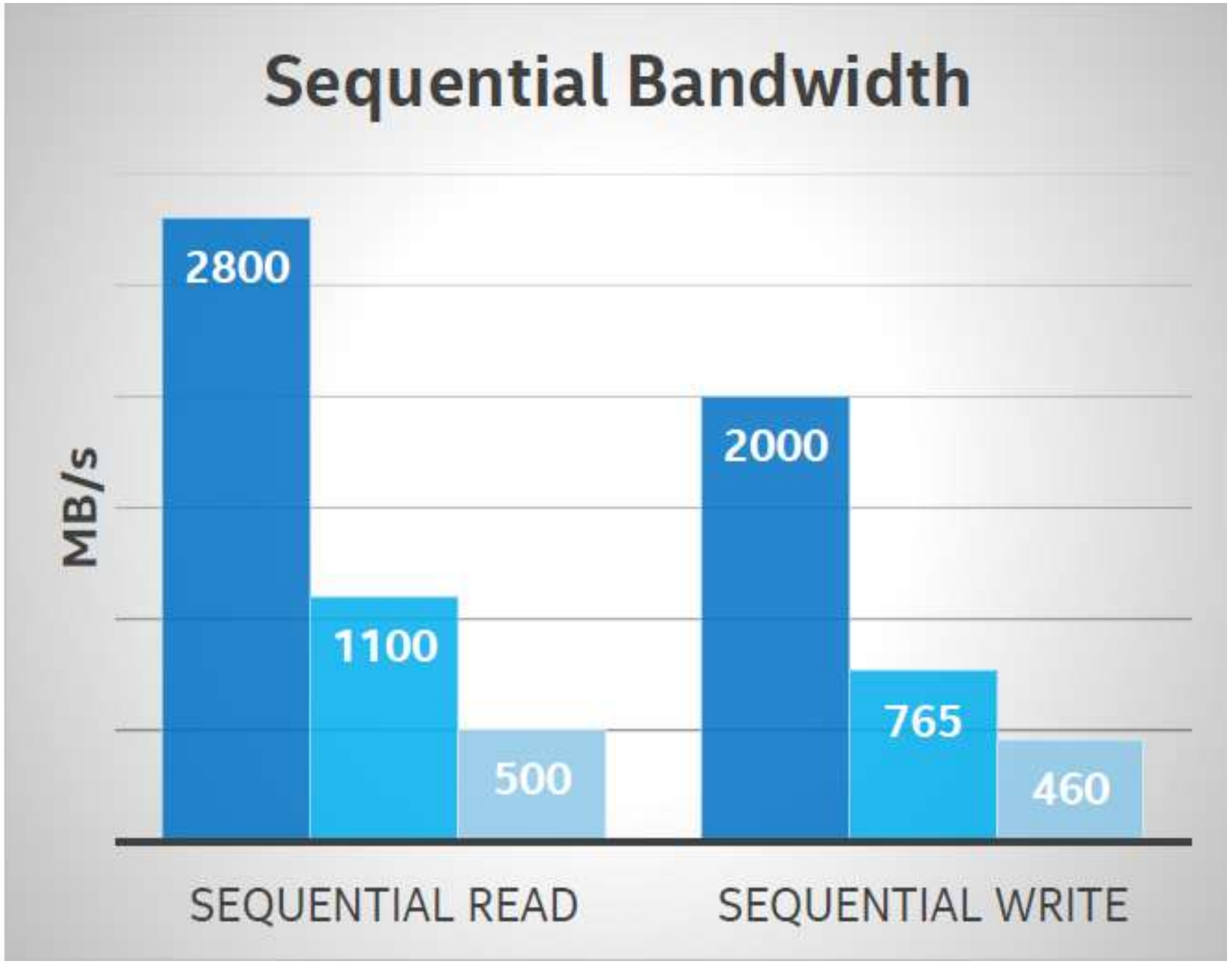
Karthik Venkatasubba
Manjunath AM
Dell R&D, Bangalore

Agenda

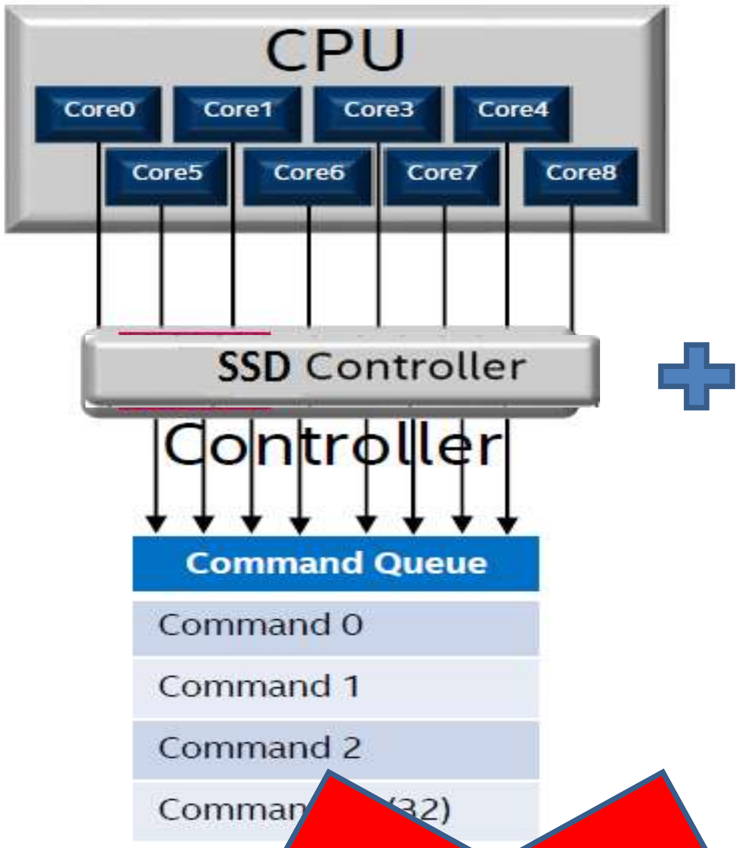
- NVMe
  - Legacy storage stacks on modern SSS – no performance improvement
  - NVMe's improvisation over legacy protocols exploiting flash characteristics
- NVMe Management
  - In-band v/s Out-of-Band management paradigms
  - Out-of-band mgmt. protocol framework
  - OSI Model
  - Architectural Model
- Overview of Features in NVMe Management Interface
  - Control Primitives
  - NVMe Management Commands
  - NVMe Admin Commands
  - PCIe Commands
- Q & A

NVMe’s Comparative Performance

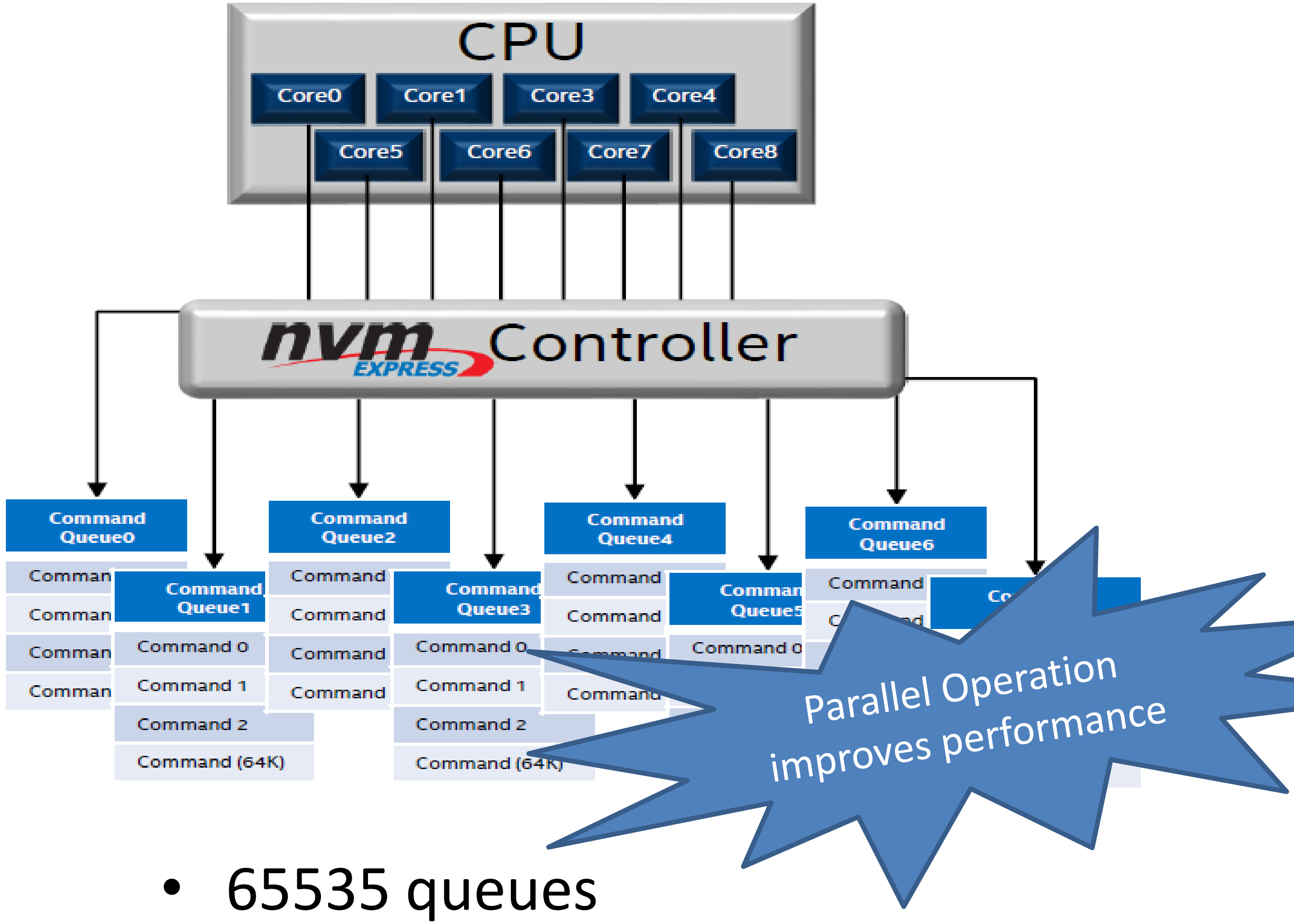




Current Performance Bottleneck (Resolved)



\+ NVMe =



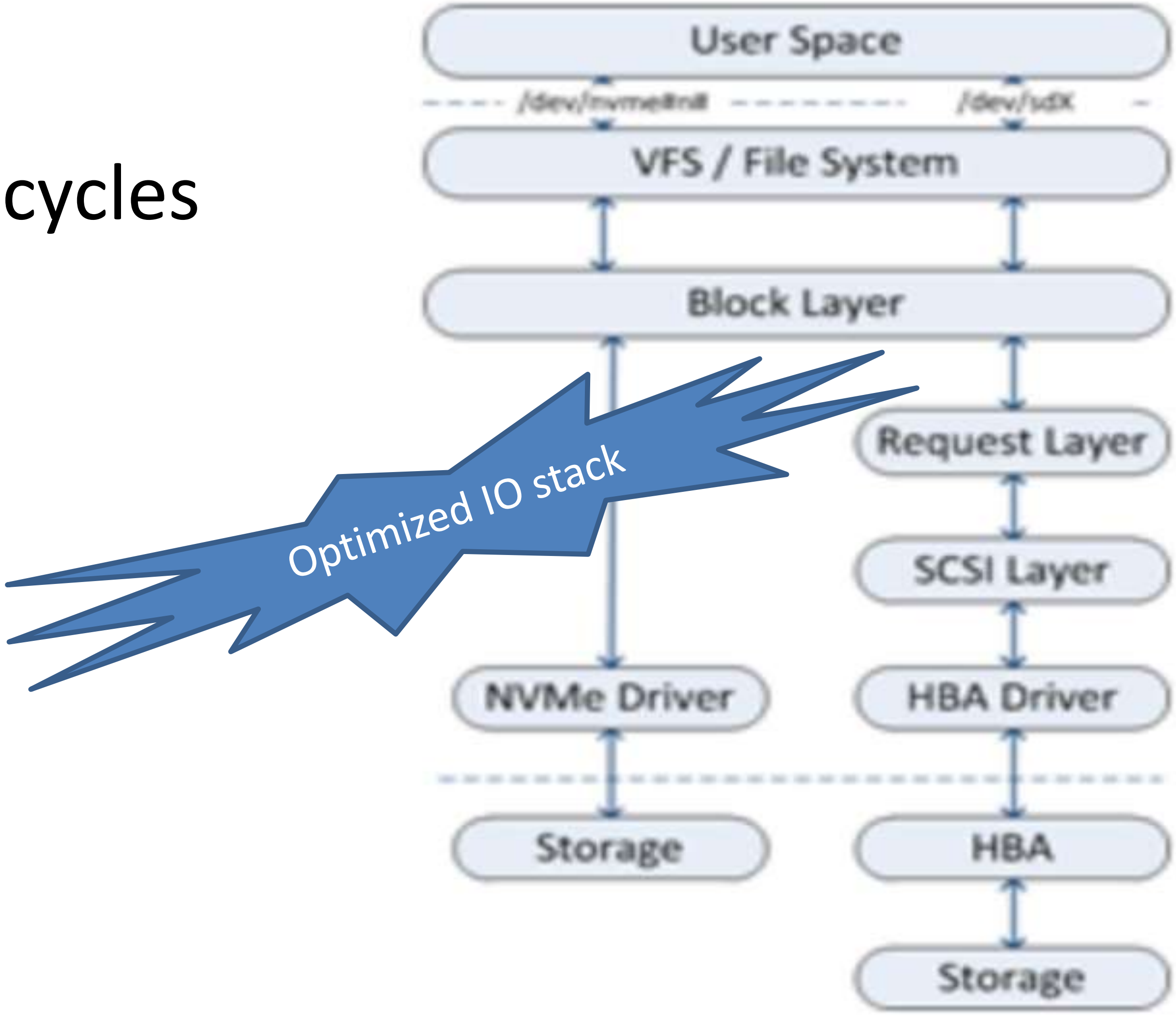
- 1 queue
- 32 commands/queue

- 65535 queues
- 64,000 commands/queues

Software Stack Improvements

Architected for Performance

- Submission latency and CPU cycles reduced > 50 %
  - – SAS: 6.0 us, 19,500 cycles
  - – NVMe: 2.8 us, 9,100 cylces

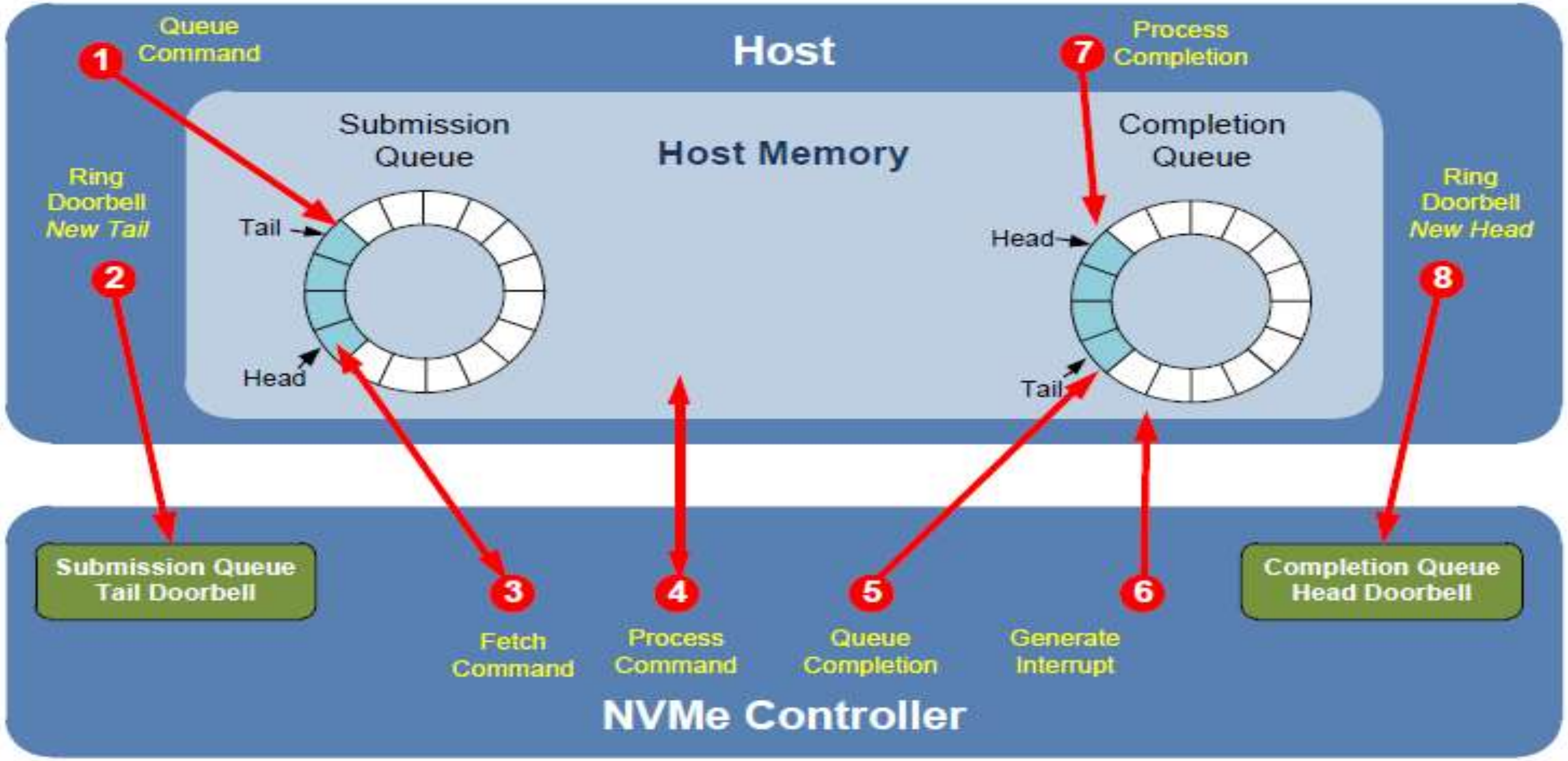


Storage Protocols Compared

nvm EXPRESS
Architected for Performance

	SATA	SAS		PCIe	
	SATA	SAS	Multilink	SOP/PQI	**NVMExpress**
DriveForm Factors	1.8",2.5", 3.5"	2.5", 3.5"	2.5"	2.5"	2.5", Card
No of Ports/ Lanes	1	1,2	1, 2, 4	1, 2, 4	1,2, 4(8 on card)
Command Set/Que Interface	ATA / SATA-IO	SCSI / SAS	SCSI / SAS	SCSI /SOP/PQI	NVM Express
Transfer Rate	6Gb/s	12Gb/s	12Gb/s	8 Gb/s	8 Gb/s
DriveConnector	SFF-xxxx	SFF-8680	SFF-8639	SFF-8639	SFF-8639 (2.5"),CEM (Edge-Card)
Express Bay Compatible?	Yes, 2.5"	Yes, 2.5"	Yes, 2.5"	Yes, 2.5"	Yes, 2.5"
Drive Power (Typical)	9W Typical	9W Typical	Upto 25W	Upto 25W	Upto 25W
Max Bandwidth	0.6GB/s	4. 8 GB/s (x2)	9.6GB/s (x4)	8 GB/s (x4)	8 GB/s (x4)
Host DriverStack (Stg Cntlr/Direct Drives)	AHCI	IHV	IHV	Common Driver (SOP/PQI)	Common Driver (NVMExpress)

Command Submission



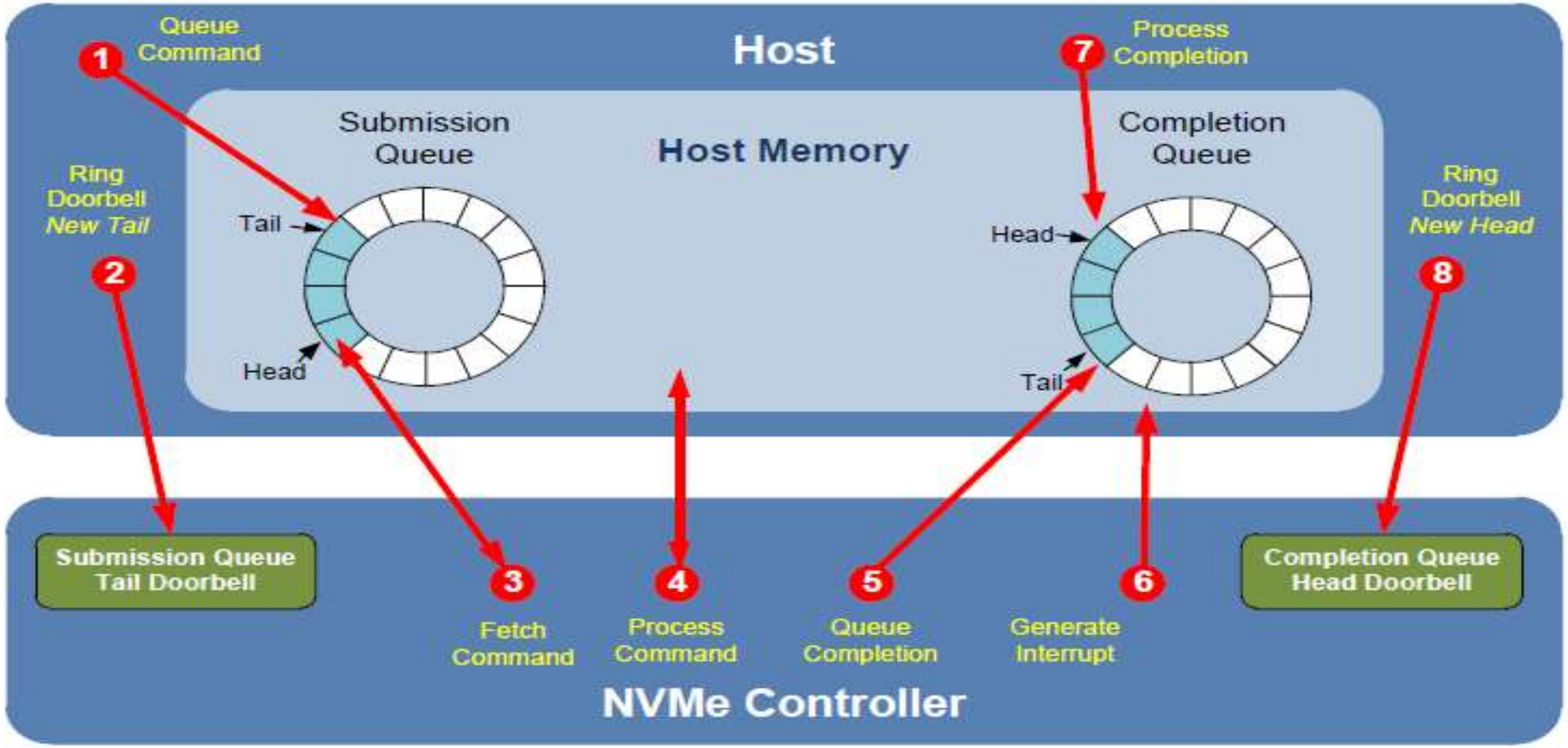
Command Submission

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

Command Completion

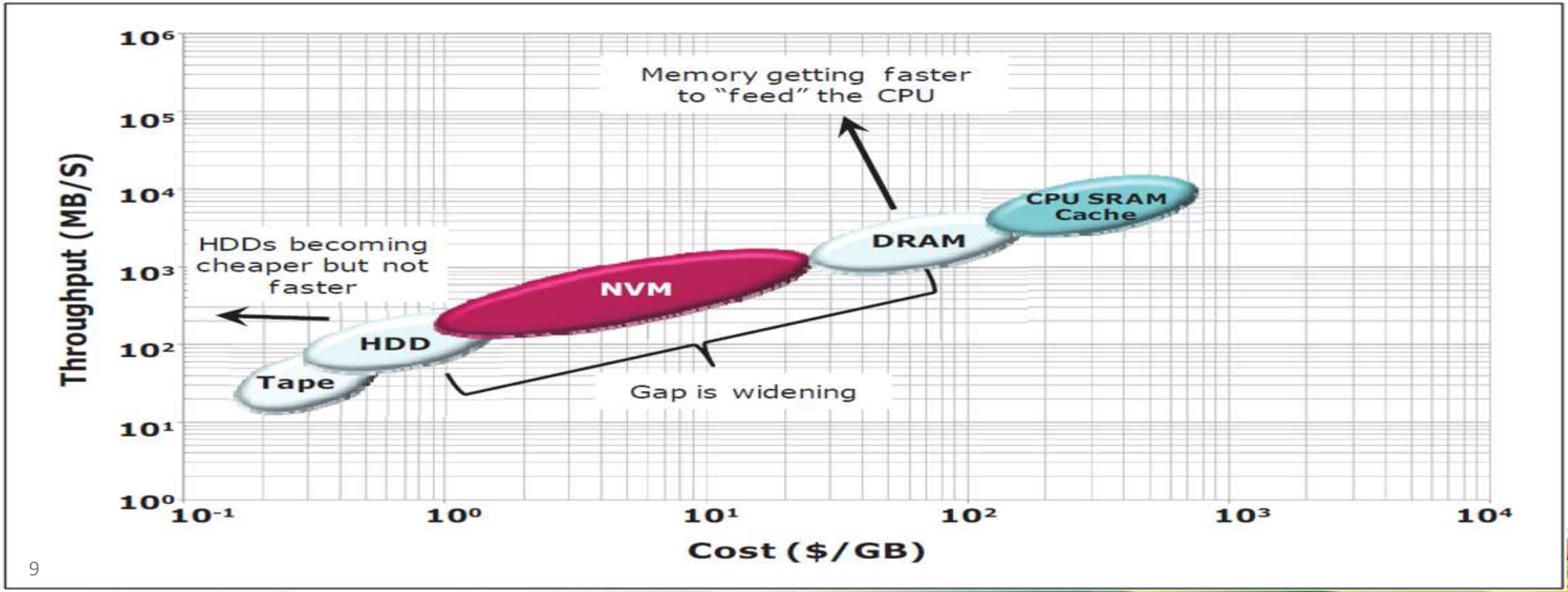


Command Completion

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

Why NVMe is becoming popular?

- There is an increasing gap in the performance of DRAM and hard drives. NVMe in the form of Solid State Drives is filling this gap.



PCIe SSD Form Factors

- Add-in Card (AIC)

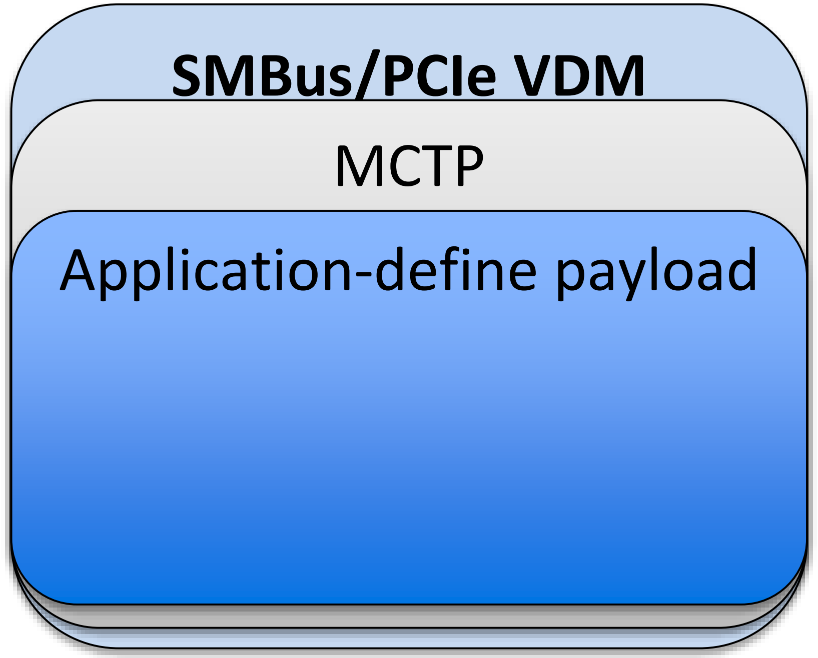
- 2.5” SSD FF (hot plug)

NVME MANAGEMENT

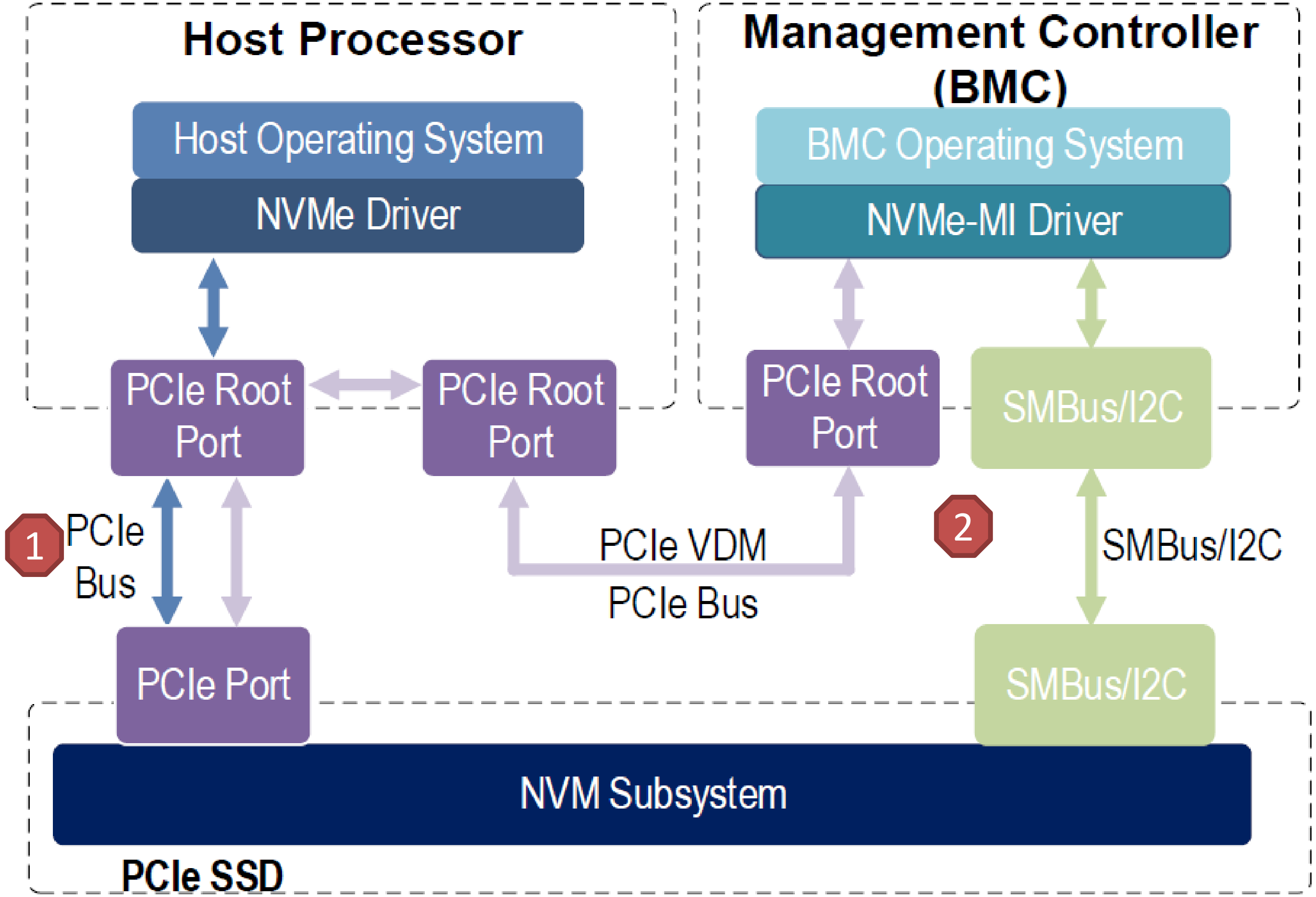
Management Protocol Stack

Protocol	Transmission unit	Endpoints	Operations	Direction
SMBus	Bytes	Master, Slave	Commands - READ/WRITE	Half duplex
PCIe VDM	TLP	Requester/Completer	Transactions - Memory, IO, configuration	Full duplex
MCTP binding spec	Physical medium-specific			
MCTP	Messages	Source/Destination	Control commands	Medium-specific
Application-defined	Application-defined			

Packet layering



In-Band vs Out-of-Band Management



1. NVMe driver communicates to NVMe controllers over PCIe per NVMe Spec

2. Two OOB paths: PCIe VDM and SMBus

- Note: PCIe VDMs are completely separate from in-band PCIe traffic though they share the same physical connection

In-band vs Out-of-Band Management (cont’d)

- In-Band Management (OS agents)
  - Many host OSes to support (Windows, Linux, VMWare, etc.)
  - Several different flavors/distros of each OS
  - Developing/maintaining/validating a management application for every OS variant is resource/cost-prohibitive
  - New revisions of OS and NVMe driver released over time
  - If given a choice, customers would want to do away with installing management agents in the OS which continuously consume CPU cycles
  - Security implications
  - Management features vary per OS

- Out-of-Band Management (Agent-free)
  - Develop management application in one operating environment (i.e. BMC)
  - Works the same across any host OS
  - Works across no OS cases (pre-boot, deployment)
  - Doesn’t consume host CPU cycles

NVMe-MI

- A programming interface that allows *out-of-band* *management* of an NVMe *Field Replaceable Unit* (FRU) or an embedded NVMe NVM Subsystem

Four pillars of systems management:

- Inventorying
- Configuration
- Monitoring
- Change Management

Management operational times:

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

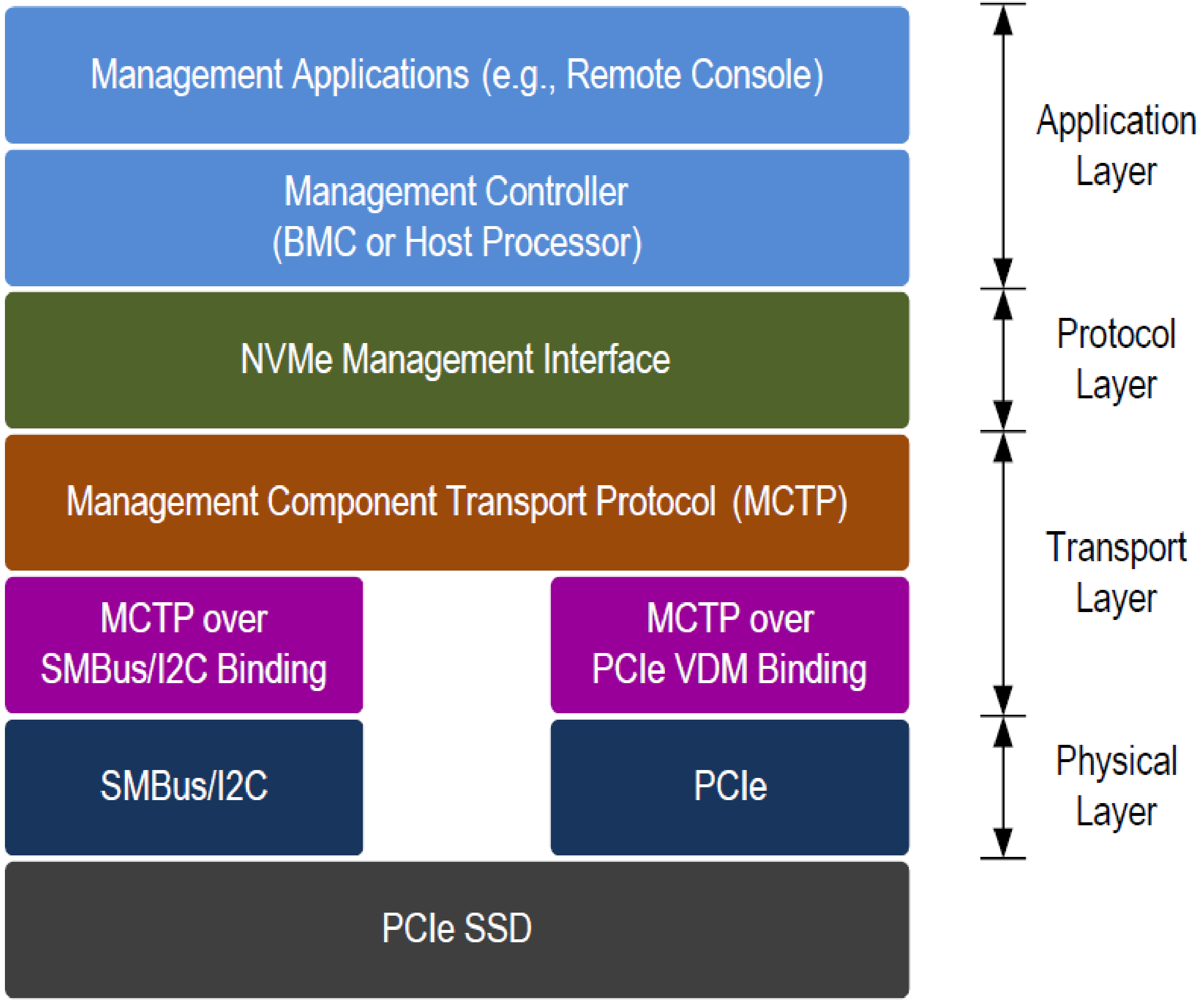
NVMe-MI OSI Model

MCTP defines the transport layer

- Refer to DMTF Specs: DSP0236, DSP0237, DSP0238, DSP0235

NVMe-MI defines:

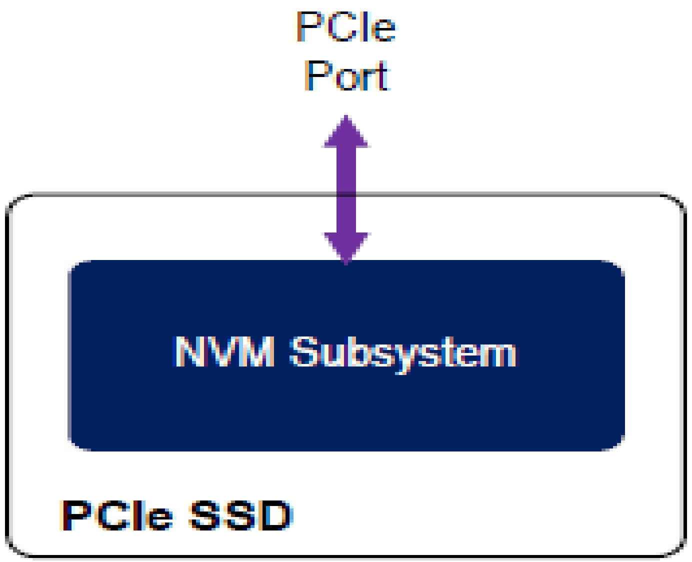
- Messages for BMC (aka SP or MC) to NVMe (aka device or PCIe SSD) out-of-band communication
- Additional flow control and exception handling on top of MCTP
- VPD access



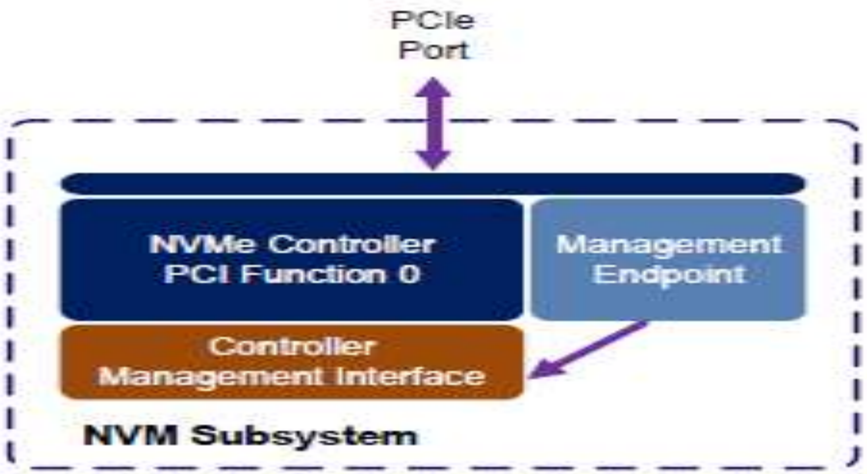
NVMe-MI Architectural Model

- **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

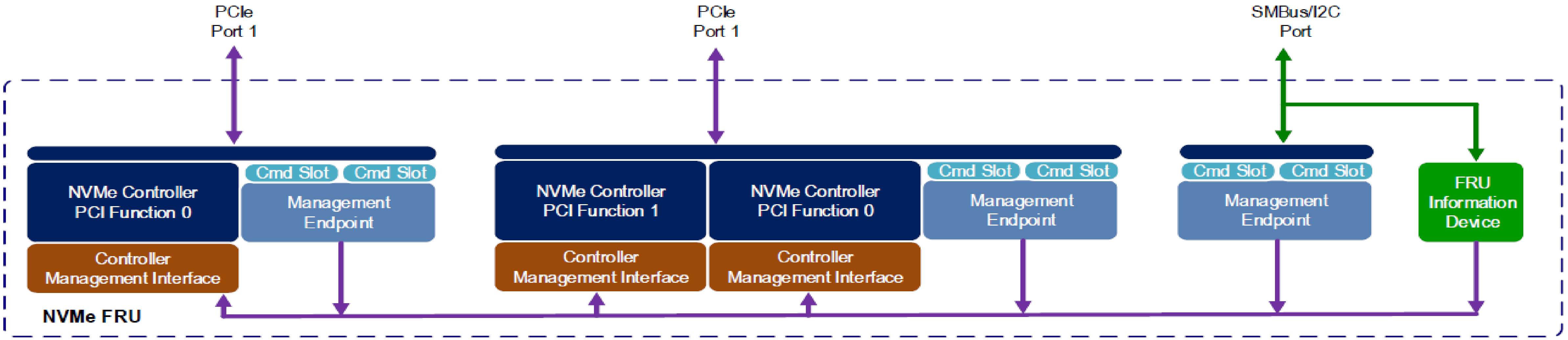
NVM Subsystem :
One Controller/Port



NVM Subsystem's anatomy



NVMe-MI Architectural Model (cont'd)



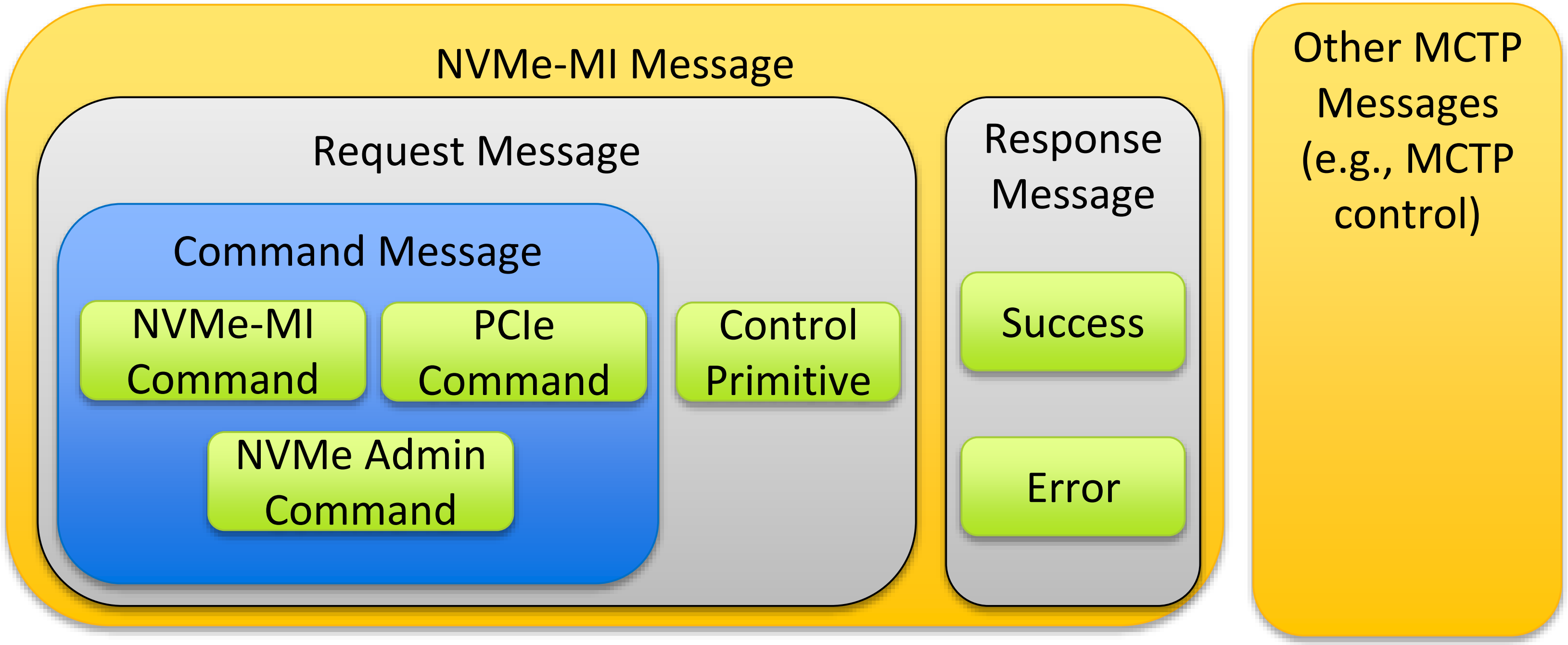
An NVMe FRU consists of one and only one NVM Subsystem with

- One or more PCIe ports (PCIe VDM)
- Optional SMBus/I2C port
- Management Endpoint per port
- Two Command Slots per Management Endpoint
- Controller Management Interface per NVMe Controller
- FRU Information Device

NVME MANAGEMENT COMMANDS

NVMe-MI Message Types

Types of MCTP Messages



Control Primitives

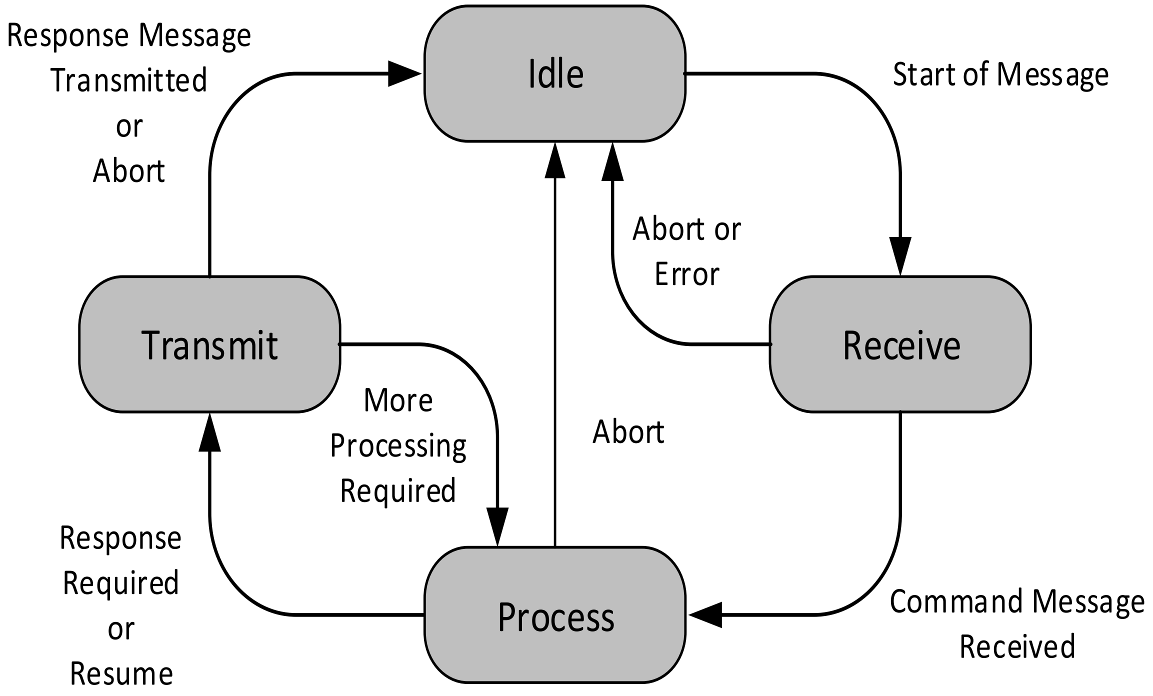
- Control Primitives enable a Management Controller to utilize flow control and to detect and recover from errors

- Control Primitives fit into a single packet and do not require message assembly

Control Primitive	O/M
Pause	Mandatory
Resume	Mandatory
Abort	Mandatory
Get State	Mandatory
Replay	Mandatory

Command Slots

- Each NVMe-MI Management Endpoint has two Command Slots to service Command Messages
- Each Command Slot follows this state machine



Management Interface Command Set

- Discover Capabilities
- Optimized Health Monitoring/polling
- Initialize & troubleshoot NVMe-MI
- Efficiently manage NVMe at the FRU level
- Sub-system level

Command	O/M
Configuration Set	Mandatory
Configuration Get	Mandatory
Controller Health Status Poll	Mandatory
NVM Subsystem Health Status Poll	Mandatory
Read NVMe-MI Data Structure	Mandatory
Reset	Mandatory
VPD Read	Mandatory
VPD Write	Mandatory
Vendor Specific	Optional

NVMe Admin Commands

- NVMe-MI defines mechanism to send existing NVMe Admin Commands out-of-band
- Admin Commands target a controller in the NVM subsystem

Command	O/M
Get Features	Mandatory
Get Log Page	Mandatory
Identify	Mandatory
Firmware Activate/Commit	Optional
Firmware Image Download	Optional
Format NVM	Optional
Namespace Management	Optional
Security Send	Optional
Security Receive	Optional
Set Features	Optional
Vendor Specific	Optional

PCIe Commands

- PCIe Commands provide optional functionality to read and modify PCIe memory

Command	O/M
PCIe Configuration Read	Optional
PCIe Configuration Write	Optional
PCIe Memory Read	Optional
PCIe Memory Write	Optional
PCIe I/O Read	Optional
PCIe I/O Write	Optional

Basic Management Command

- Simple and optional command
- Intended for Vendors/System integrators looking for a light-weight NVMe out-of-band device monitoring
- Does not use MCTP
- Limited set of attributes could be monitored by the host via SMBus like : Temperature, Critical Warnings and Life Remaining
- Mode of operation is very much like a typical VPD access to FRU information device

Example SMBus block read of the drive's status (status flags, SMART warnings, temperature):

Start	Addr	W	Ack	Cmd Code	Ack	Restart	Addr	R	Ack	Length	Ack	Status Flags	Ack	SMART Warnings	Ack	Temp	Ack	Drive Life Used	Ack	Reserved	Ack	Reserved	Ack	PEC	Ack	Stop
	D4h			00h			D5h			06h		BFh		FFh		1Eh		01h		00h		00h		10h		

Summary

NVMe-MI standardizes an **out-of-band** management interface to discover, monitor and configure NVMe devices

NVMe-MI adds the ability to manage NVMe **at the FRU level**

NVMe Management Interface Specification Revision 1.0 ratified and available at http://www.nvmexpress.org/.

Thank You!

References

- SSD Form Factor Working Group, http://www.ssdformfactor.org/
- SMBus, http://smbus.org/
- PCI SIG, https://**pcisig**.com/
- DMTF, http://dmtf.org/
- NVMe, www.**nvme**xpress.org