NVM Express™ Management Interface

August 11, 2015

John Carroll, Storage Architect, Intel
Peter Onufryk, Storage Director, Product Development, PMC-Sierra
Austin Bolen, Storage Technologist, Dell

Agenda

- NVMe Management Interface Overview
  - Definition
  - Comparison to NVM Express Specification interface
  - Benefits over in-band management
  - To standardize or not to standardize
- NVMe-MI Usage
  - A real world example – Automated Remote Health Monitoring

- NVMe-MI Architecture
  - NVM Subsystem, Port, Management Endpoint, Command Slot
- Overview of Features/Functionality
  - NVMe Management Commands
  - NVMe Admin Commands
  - PCIe Commands
  - Control Primitives
  - VPD
- Standardization Status

NVMe Management Interface

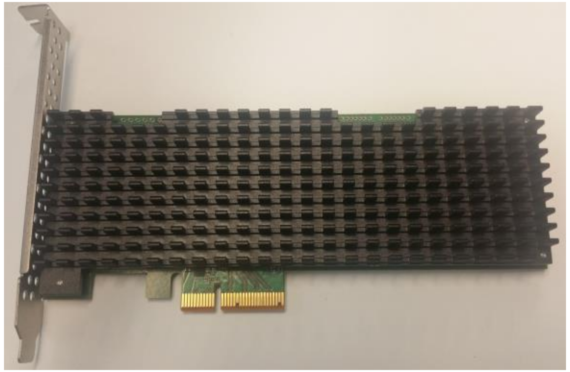
What is the NVMe Management Interface?

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

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.



Management Fundamentals

What is meant by "management"?

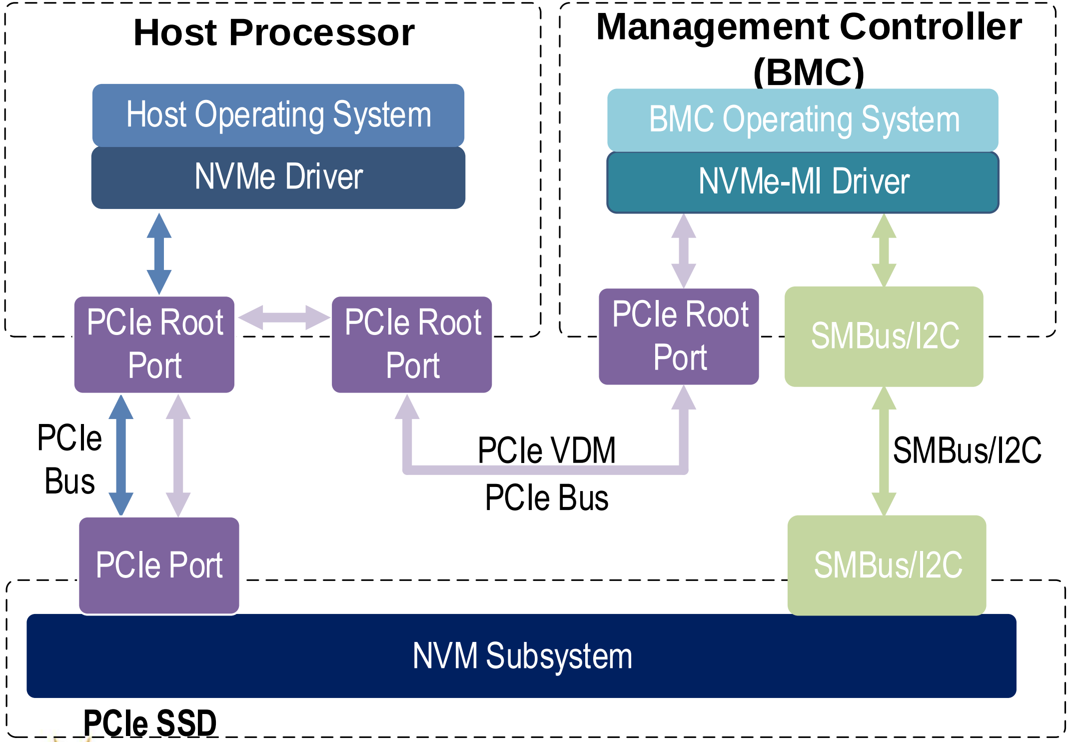
Four pillars of systems management:

- Inventory
- Configuration
- Monitoring
- Change Management

Management operational times:

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

In-Band vs Out-of-Band Management



- NVMe driver communicates to NVMe controllers over PCIe per NVMe Spec
- MC runs on its own OS on it own processor independent from host OS and driver
- Two OOB paths: PCIe VDM and SMBus
- 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.

In-Band Management Application

- Many host OSes to support (Windows, Linux, VMWare, etc.)
- Several different flavors/distros of each
- New revisions of OS and NVMe driver released over time
- Developing and maintaining a management application for every OS variant is resource/cost prohibitive
- Management features vary per OS

Out-of-Band Management Application

- Develop management application in one operating environment
- Works the same across any host OS the user installs
- Works across no OS cases (pre-boot, deployment)

Why Standardize NVMe Storage Device Management?

Reduces Cost and Broadens Adoption

- Allows OEMs to source storage devices from multiple suppliers
- Eliminates need for NVMe storage device suppliers to develop custom OEM specific management features

Consistent Feature Set

- All storage devices that implement management implement a common baseline feature set
- Optional features are implemented in a consistent manner

Industry Ecosystem

- Compliance tests / program
- Development tools

A Real World Example – Automated Remote Health Monitoring

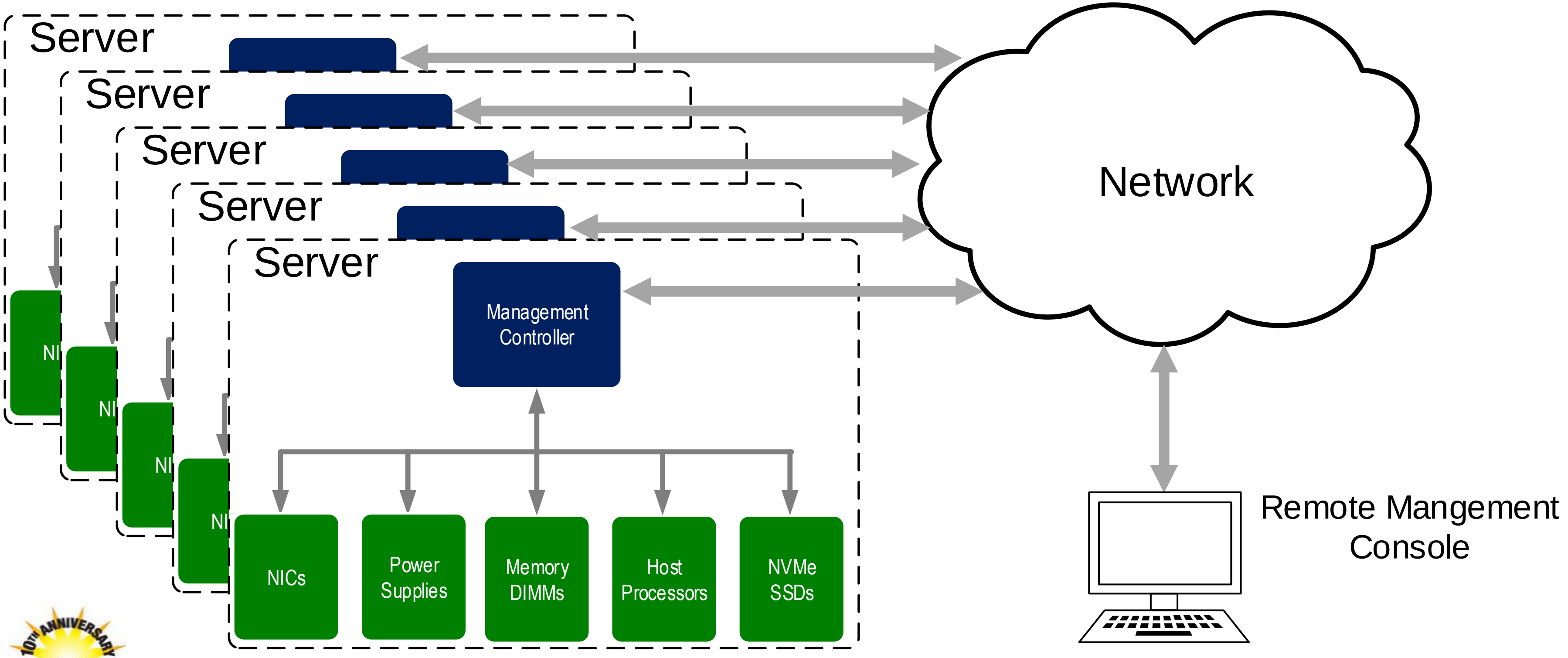
The Problem:

- Datacenter with hundreds of servers
- Each servers consists of dozens of Field Replaceable Units
- Some number of FRUs fail weekly (or even daily)
- Manually discovering and resolving issues due to failed FRUs is prohibitively time consuming and expensive

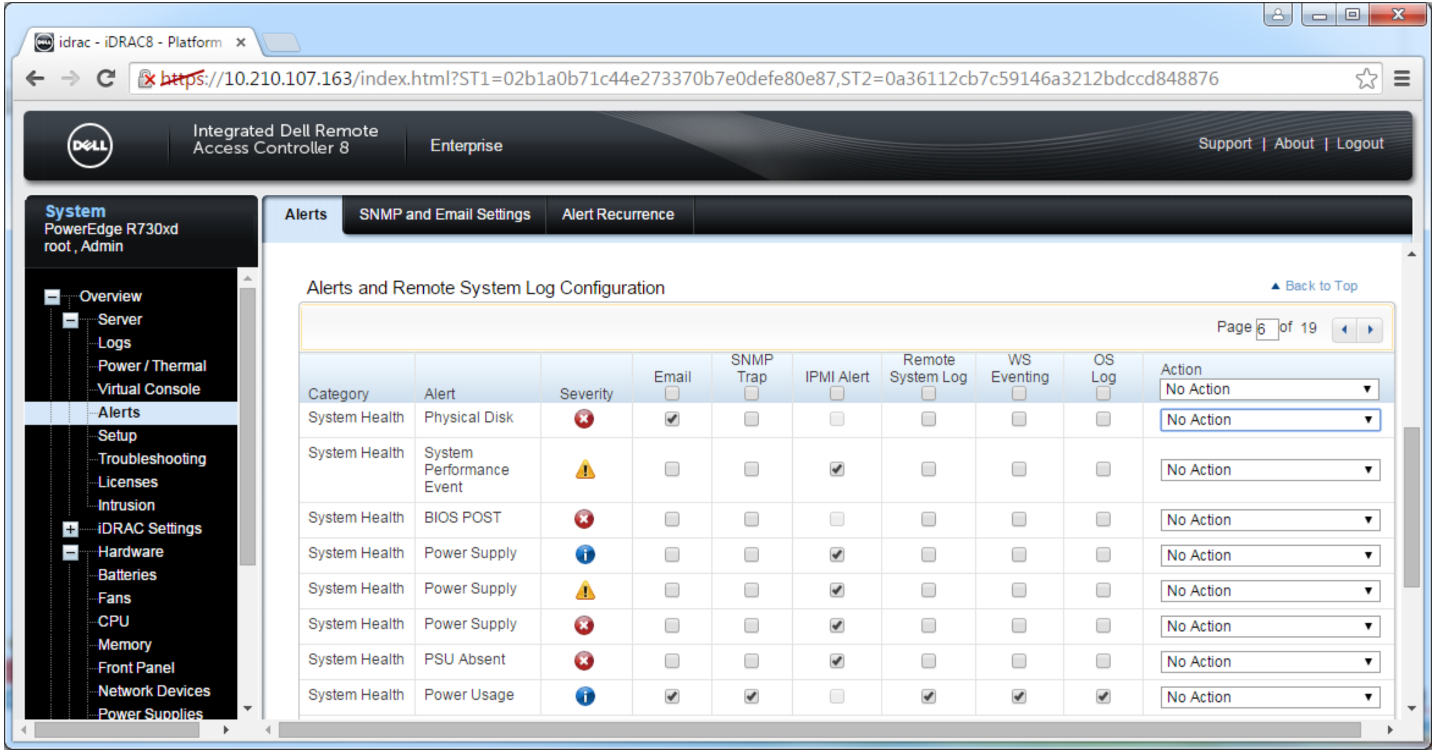
The Solution:

- Each server has a BMC to manage all FRUs
- Each BMC is connected to a network accessible via a remote management console
- BMC detects NVMe FRU failures using NVMe-MI and reports failures to a remote administrator

Remote Health Monitoring – Management Infrastructure

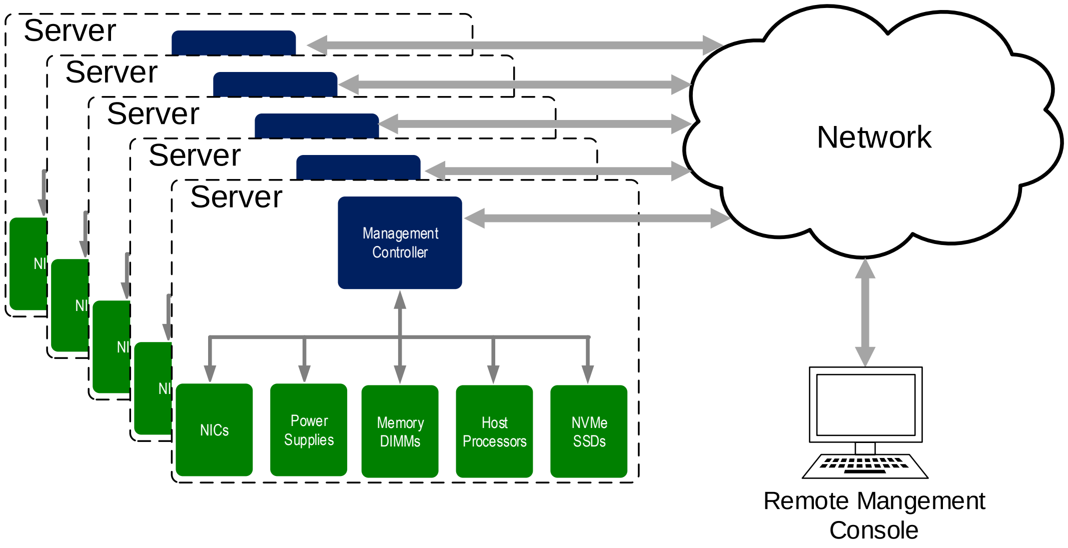


Remote Health Monitoring – Set up Alerts



Enable e-mail alerts for NVMe health events

Remote Health Monitoring – Detect Error Using NVMe-MI



- Management Controller issues NVM Subsystem Health Status Poll command to NVMe drive
- NVMe drive responds indicating a Critical Warning bit is set
- Management Controller then issues a Controller Health Status Poll command to the drive
- NVMe drive responds indicating a Reliability Degraded error occurred
- Management Controller sends email notification

Remote Health Monitoring – Receive E-Mail Alert

From: idrac-simpsons@smd.devops.dell.com
Sent: Monday, August 03, 2015 11:11 AM
To: Austin_Bolen@Dell.com
Subject: WIN-MMPK73JS9PO: Fault detected on drive in Bay ID 1 Slot ID 21.

System Host Name: WIN-MMPK73JS9PO
Event Message: The drive in Bay ID 1 Slot ID 21 reported a critical fault condition.
Date/Time: Mon Aug 03 2015 09:10:59
Severity: Critical

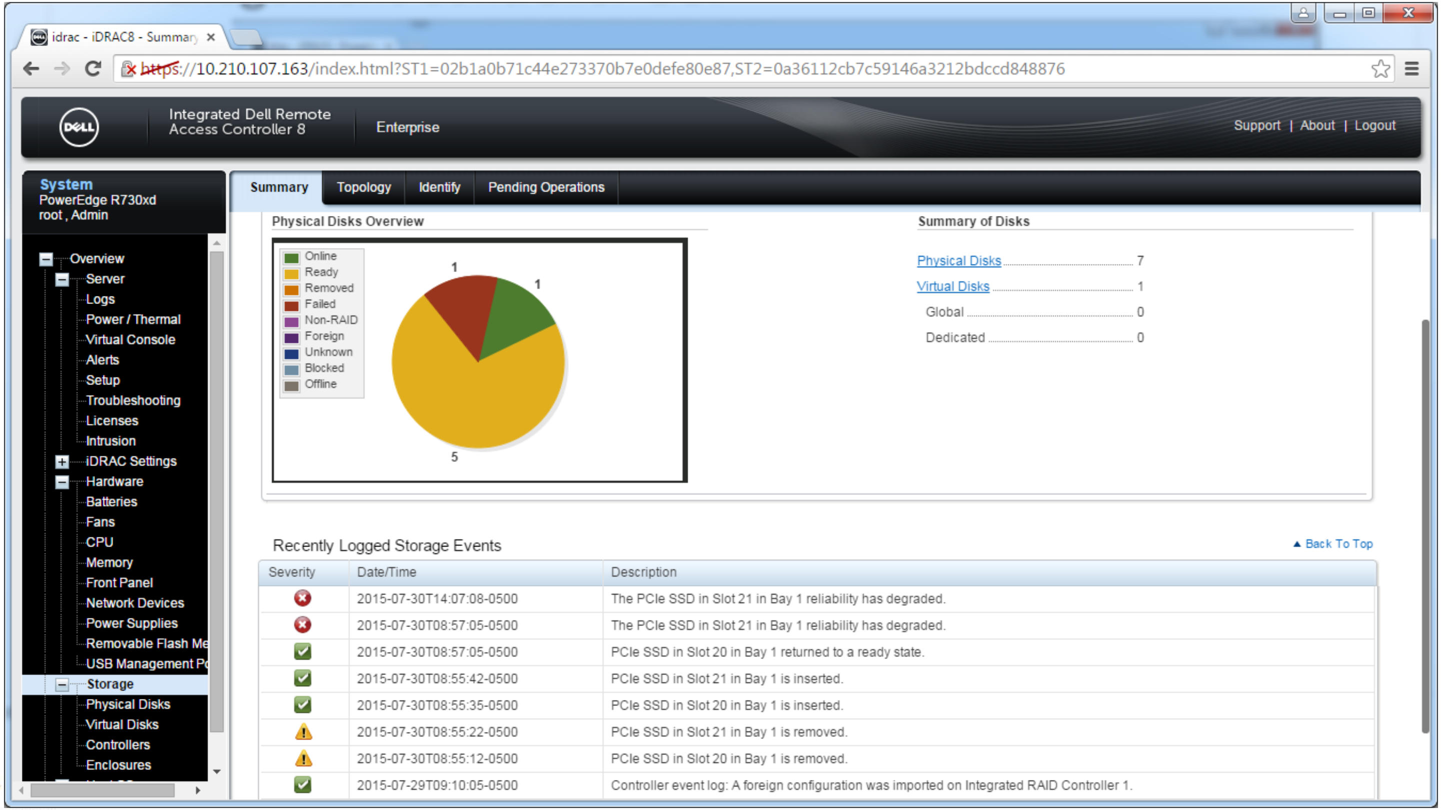
Detailed Description: Product documentation contains information on correct configuration. The failure could also be caused by a faulty component or related cabling. System performance may be degraded.
Recommended Action: Remove and re-seat the failed drive. If the issue persists, see Getting Help.
Message ID: PDR0001

System Model: PowerEdge R730xd
Service Tag: H9QVP22
Power State: ON
System Location: Rack 132 Slot 1 (2 U)

To launch the iDRAC Web Interface, click here: https://10.35.155.80. To launch the iDRAC Virtual Console, click here: https://10.35.155.80/console

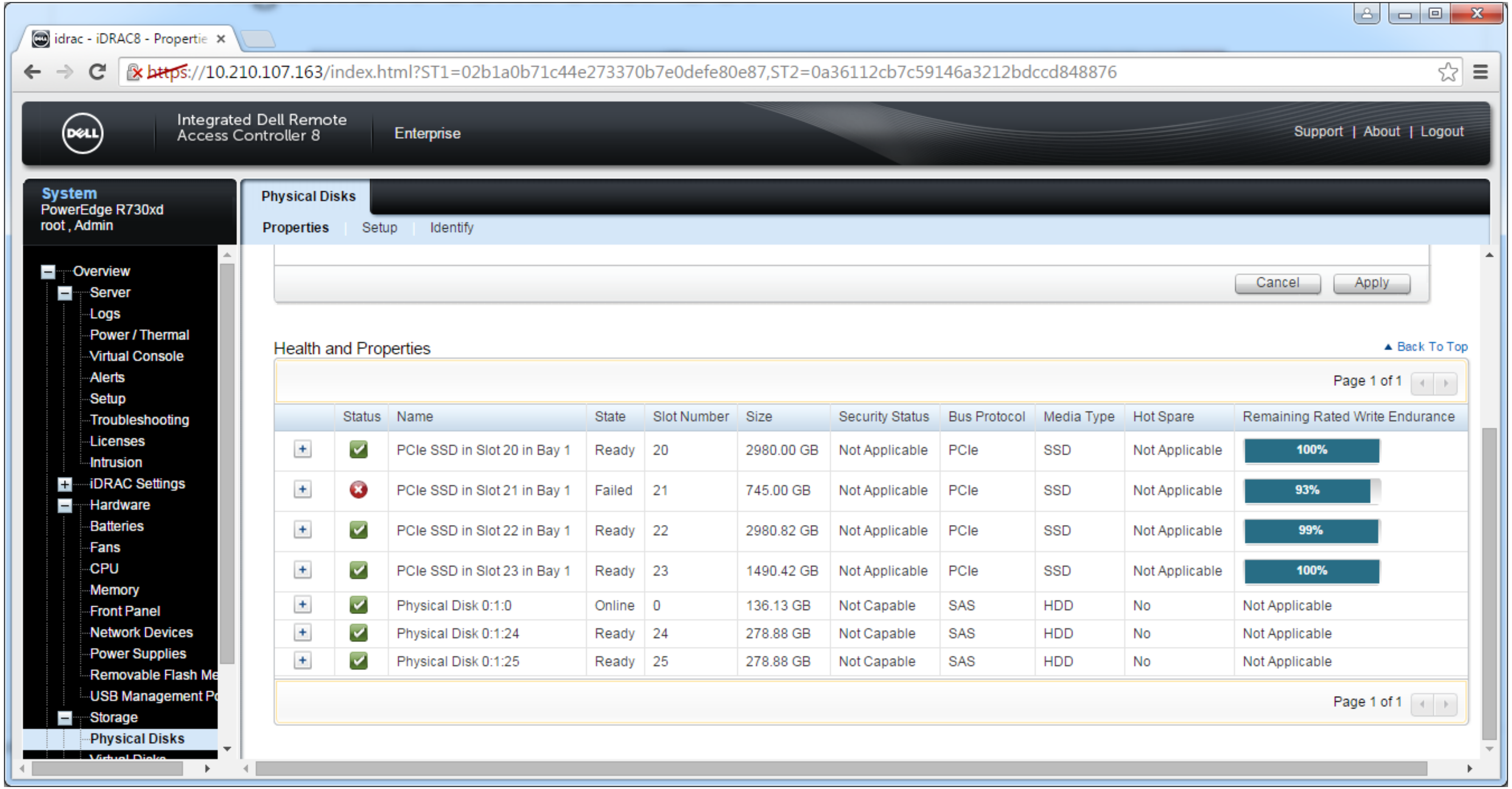
E-mail alert sent if a drive health event occurs

Remote Health Monitoring – Check Event Log



Check event log on remote system

Remote Health Monitoring – Drives Overview

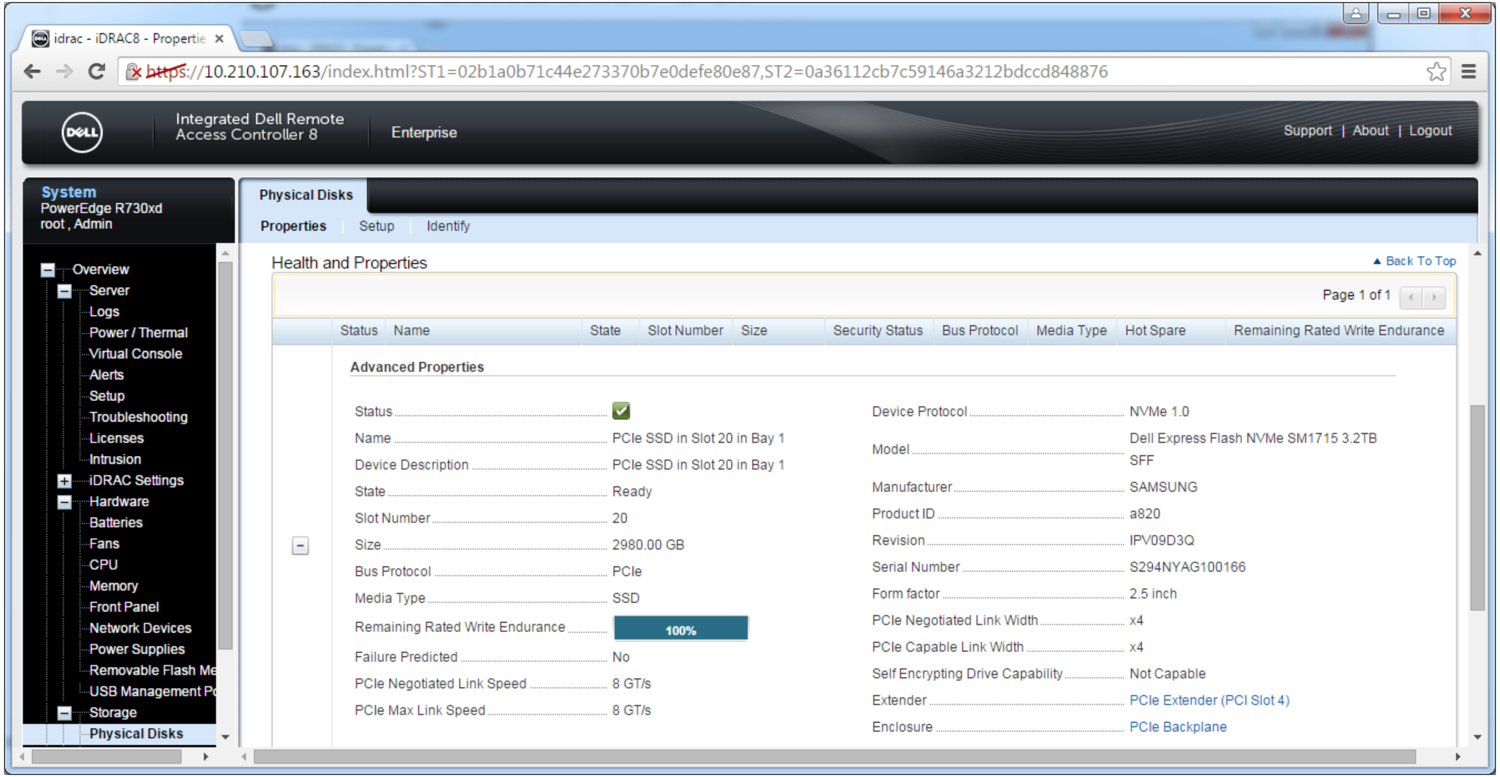


Status	Name	State	Slot Number	Size	Security Status	Bus Protocol	Media Type	Hot Spare	Remaining Rated Write Endurance
	PCIe SSD in Slot 20 in Bay 1	Ready	20	2980.00 GB	Not Applicable	PCIe	SSD	Not Applicable	100%
	PCIe SSD in Slot 21 in Bay 1	Failed	21	745.00 GB	Not Applicable	PCIe	SSD	Not Applicable	93%
	PCIe SSD in Slot 22 in Bay 1	Ready	22	2980.82 GB	Not Applicable	PCIe	SSD	Not Applicable	99%
	PCIe SSD in Slot 23 in Bay 1	Ready	23	1490.42 GB	Not Applicable	PCIe	SSD	Not Applicable	100%
	Physical Disk 0:1:0	Online	0	136.13 GB	Not Capable	SAS	HDD	No	Not Applicable
	Physical Disk 0:1:24	Ready	24	278.88 GB	Not Capable	SAS	HDD	No	Not Applicable
	Physical Disk 0:1:25	Ready	25	278.88 GB	Not Capable	SAS	HDD	No	Not Applicable

Page 1 of 1

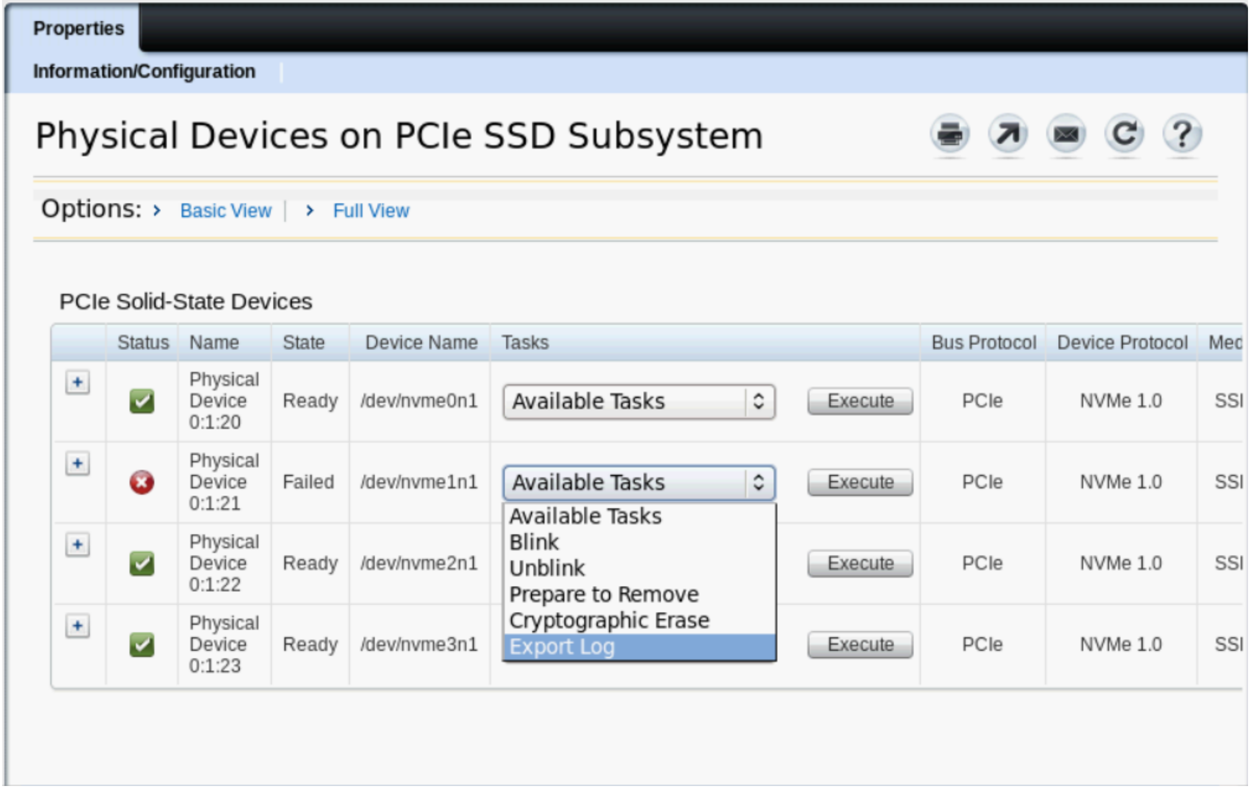
Review drive table to see status of all drives in the system

Remote Health Monitoring – Drive Detail



Expand a drive's view to see its details

Management Controller GUI – Export Log Files



Export logs from the drive

Remote Health Monitoring – Check Log File

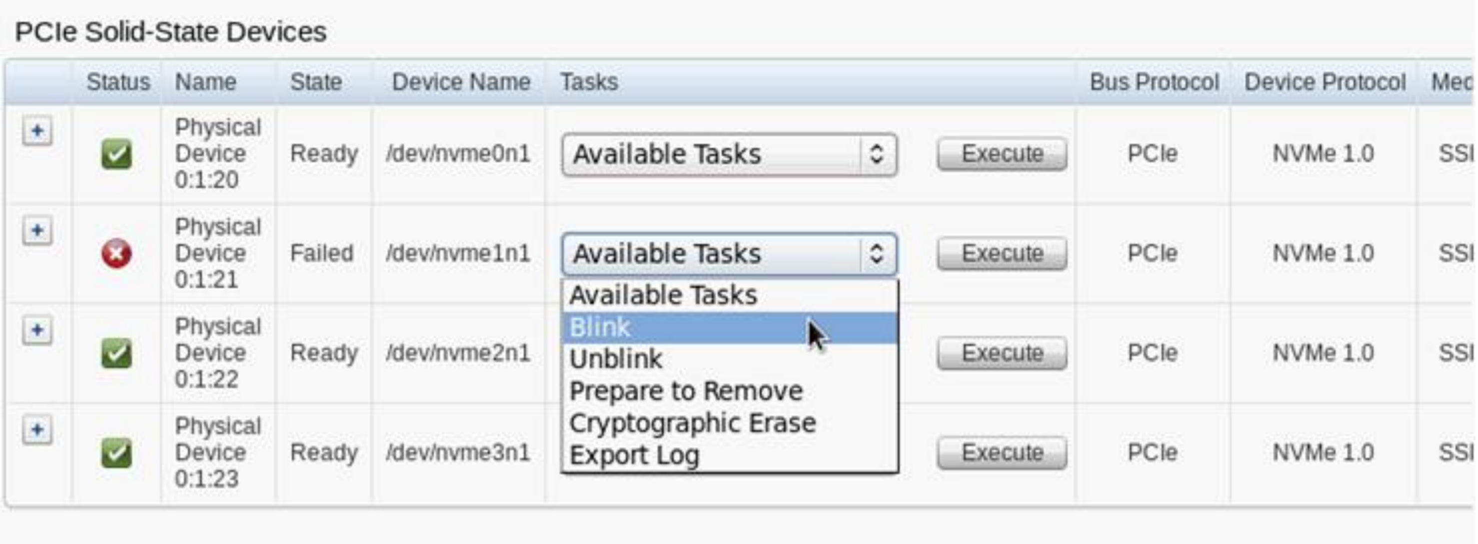
NVME_S1J1NYAF100038_0730095509.log

```
========== NVMe Device Identifier Information ================================
Model Number               = Dell Express Flash NVMe 400GB
Physical Device Location = Bay ID 1 Slot ID 21
Namespace Label            = /dev/nvme1n1
Firmware Revision          = 1.0.0
Serial Number              = S1J1NYAF100038
PCI Bus:Device.Function    = 133:00.0

========== NVMe SMART/Health Information Log ==========
Critical Warning:
  Available Space Fallen Below Threshold  = 0
  Temperature Exceeded Critical Threshold = 0
  NVM Subsystem Reliability Degraded      = 1
  Media Read Only Mode                    = 0
  Volatile Memory Backup Failed           = 0
Temperature                               = 34.85 Celsius (308 Kelvin)
Available Spare                           = 84%
Available Spare Threshold                 = 10%
Percentage Used                           = 7%
Data Read                                 = 1.20 PB (1,197,442,070,528,000 bytes)
Data Written                              = 1.21 PB (1,207,531,087,360,000 bytes)
```

Check the log for failure details

Remote Health Monitoring – Blink Drive LED



Admin blinks the indicator LED for the failed drive
Local Datacenter Technician finds and replaces faulty drive

Agenda

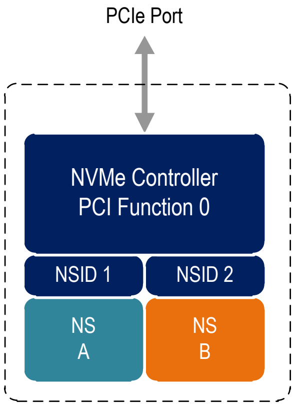
- NVMe Management Interface Overview
  - Definition
  - Comparison to NVM Express Specification interface
  - Benefits over in-band management
  - To standardize or not to standardize
- NVMe-MI Usage
  - A real world example – Automated Remote Health Monitoring

- NVMe-MI Architecture
  - NVM Subsystem, Port, Management Endpoint, Command Slot
- Overview of Features/Functionality
  - NVMe Management Commands
  - NVMe Admin Commands
  - PCIe Commands
  - Control Primitives
  - VPD
- Standardization Status

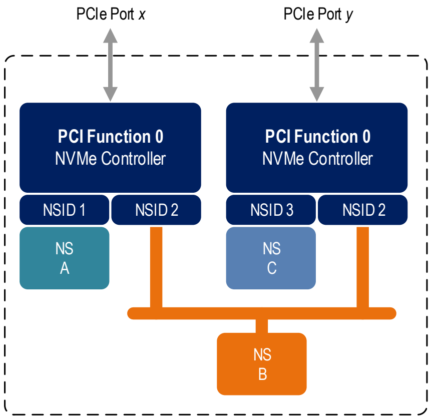
NVMe Architecture (review)

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

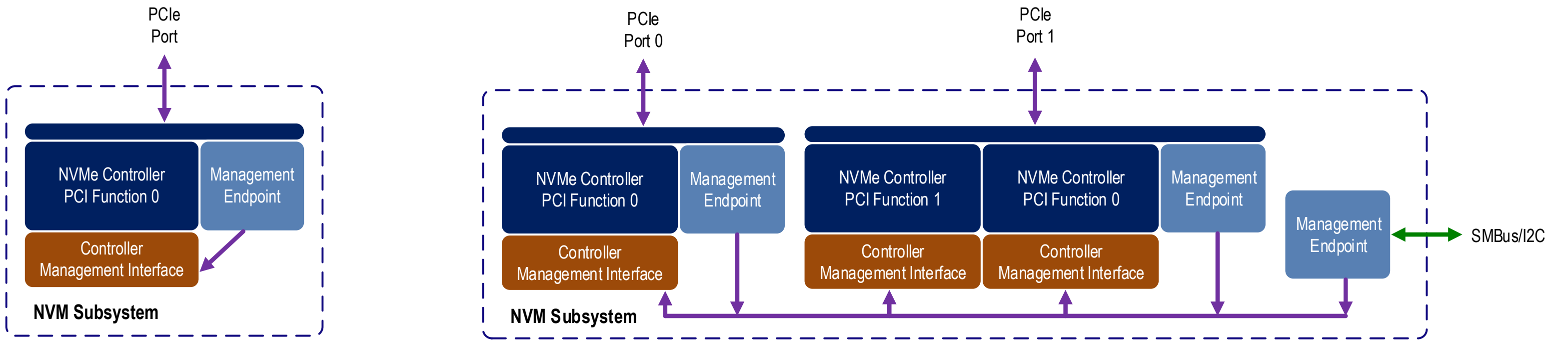
NVM Subsystem
One Controller/Port



NVM Subsystem
Two Controllers/Ports



NVMe Field Replaceable Units with NVMe-MI



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

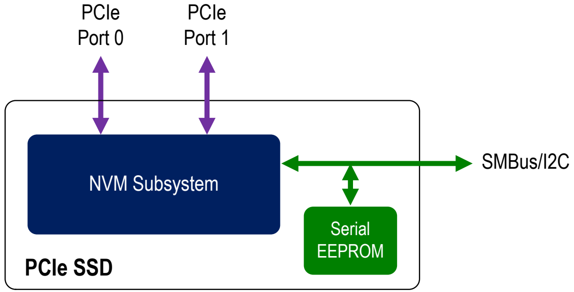
- One or more PCIe ports
- An optional SMBus/I2C interface
- One or more Management Endpoints

VPD – Vital Product Data

Vital Product Data typically available in a serial EEPROM

NVMe-MI defined standard VPD contents including:

- Device Form factor
- Initial and peak power usage by power rail
- RefClk/SRIS capability
- and more …



NVMe-MI makes VPD contents accessible out-of-band

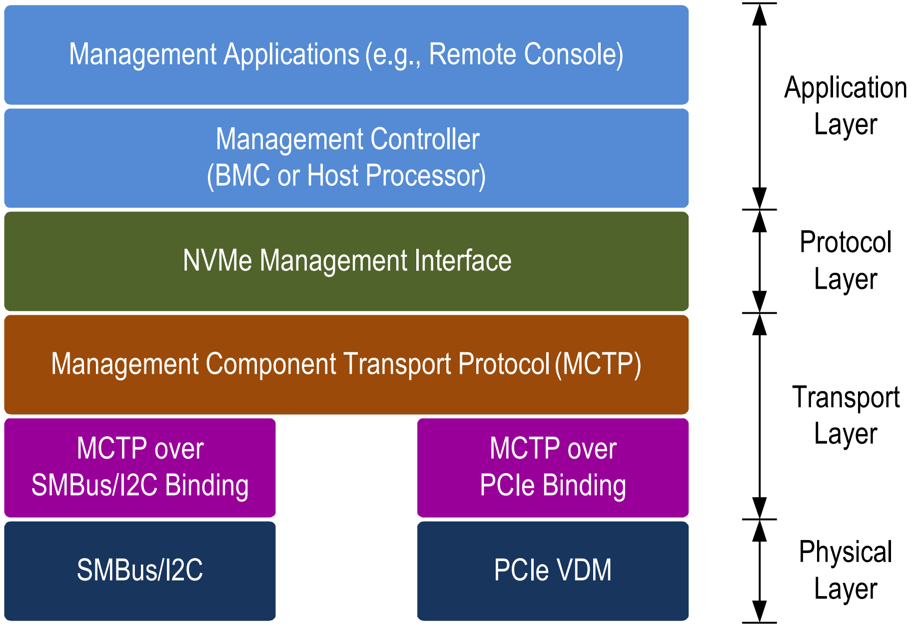
NVMe-MI Defines the Protocol for Managing NVMe

Leverage existing PCIe and SMBus

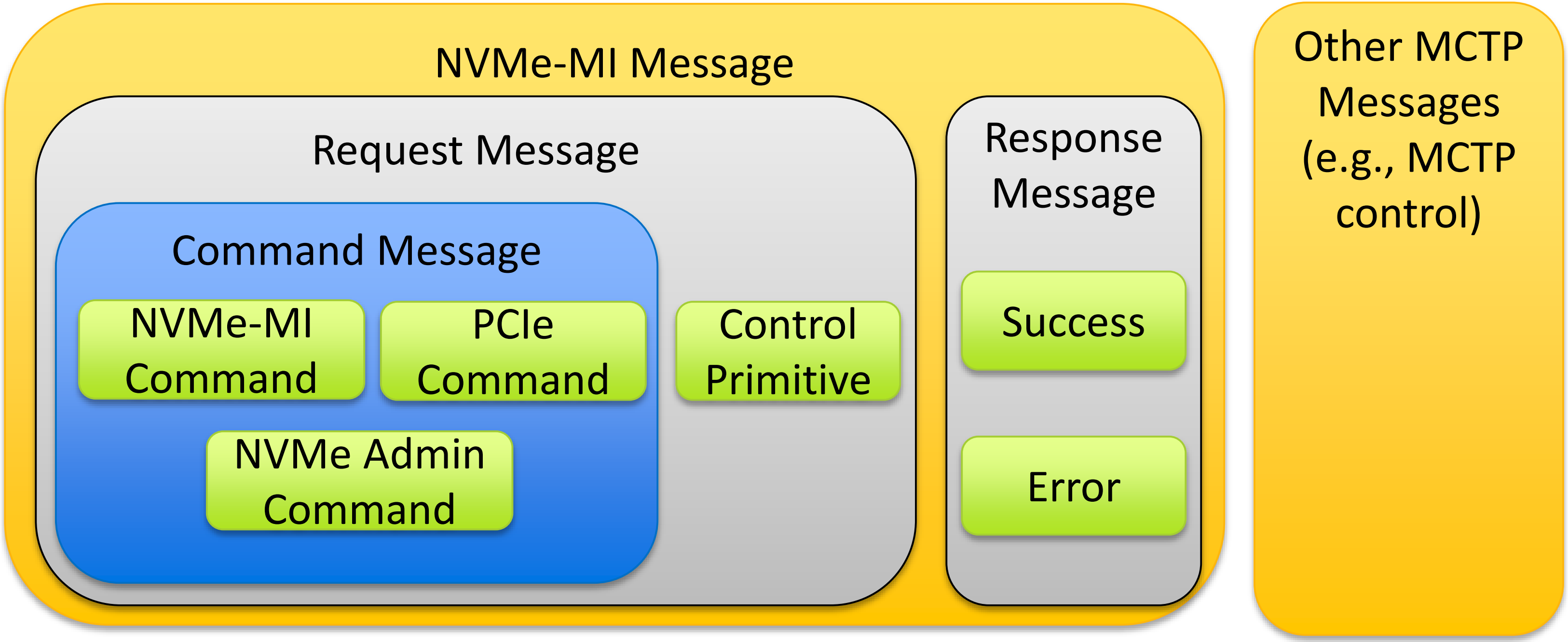
MCTP defines the transport layer

- Refer to http://dmtf.org/ for more info on MCTP

NVMe-MI is the protocol for applications to information

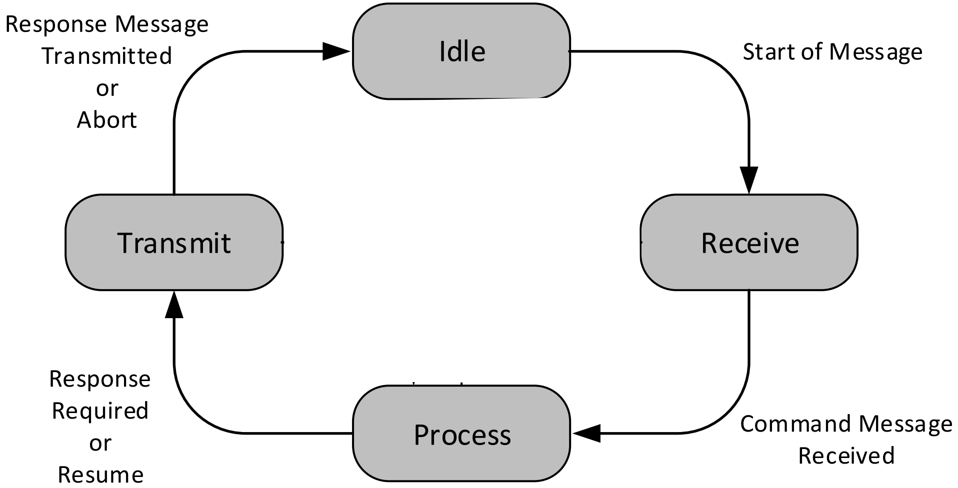


Types of MCTP Messages



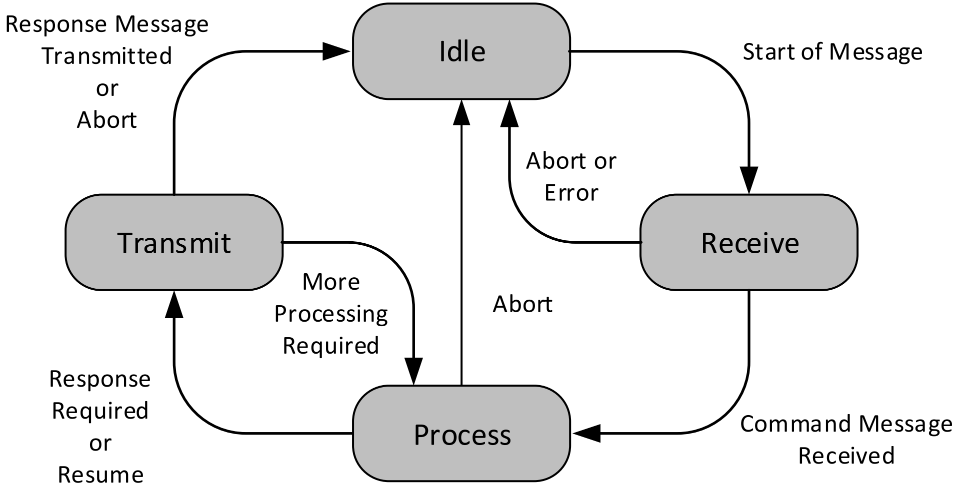
Command Slots

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



Command Slots

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



Management Interface Command Set

- Discover Device Capabilities
- Monitor Health Status
- Modify Configuration

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

Control Primitives

- Control Primitives enable a Management Controller 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

Summary

NVMe-MI standardizes out of band management to discover and configure NVMe devices

NVMe-MI 1.0 specification under member review – will be published on NVMe site after ratification

Join NVMe to shape the future of NVMe-MI nvm EXPRESS ™

Architected for Performance