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
• NVMe driver communicates to NVMe controllers over PCIe per NVMe Spec

• MC runs on its own OS on its 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 server 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

Network

Server

Management Controller

NICs
Power Supplies
Memory DIMMs
Host Processors
NVMe SSDs

Remote Management Console
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: H9QVF22
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

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

![Management Controller GUI](image-url)

**Physical Devices on PCIe SSD Subsystem**

Options:  
- Basic View  
- Full View

### PCIe Solid-State Devices

<table>
<thead>
<tr>
<th>Status</th>
<th>Name</th>
<th>State</th>
<th>Device Name</th>
<th>Tasks</th>
<th>Bus Protocol</th>
<th>Device Protocol</th>
<th>Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Physical Device 0:1:20</td>
<td>Ready</td>
<td>/dev/nvme0n1</td>
<td>Available Tasks</td>
<td>PCIe</td>
<td>NVMe 1.0</td>
<td>SSI</td>
</tr>
<tr>
<td></td>
<td>Physical Device 0:1:21</td>
<td>Failed</td>
<td>/dev/nvme1n1</td>
<td>Available Tasks</td>
<td>PCIe</td>
<td>NVMe 1.0</td>
<td>SSI</td>
</tr>
<tr>
<td></td>
<td>Physical Device 0:1:22</td>
<td>Ready</td>
<td>/dev/nvme2n1</td>
<td>Available Tasks</td>
<td>PCIe</td>
<td>NVMe 1.0</td>
<td>SSI</td>
</tr>
</tbody>
</table>
|          | Physical Device 0:1:23 | Ready  | /dev/nvme3n1     | Available Tasks  
  - Blink  
  - Unblink  
  - Prepare to Remove  
  - Cryptographic Erase  
  - Export Log | PCIe         | NVMe 1.0        | SSI       |

**Export logs from the drive**
Remote Health Monitoring – Check Log File

<table>
<thead>
<tr>
<th>No.</th>
<th>NVMe Device Identifier Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Model Number = Dell Express Flash NVMe 400GB</td>
</tr>
<tr>
<td>2</td>
<td>Physical Device Location = Bay ID 1 Slot ID 21</td>
</tr>
<tr>
<td>3</td>
<td>Namespace Label = /dev/nvme1n1</td>
</tr>
<tr>
<td>4</td>
<td>Firmware Revision = 1.0.0</td>
</tr>
<tr>
<td>5</td>
<td>Serial Number = S1J1NYAF100038</td>
</tr>
<tr>
<td>6</td>
<td>PCI Bus:Device.Function = 133:00.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>NVMe SMART/Health Information Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Critical Warning:</td>
</tr>
<tr>
<td>10</td>
<td>Available Space Fallen Below Threshold = 0</td>
</tr>
<tr>
<td>11</td>
<td>Temperature Exceeded Critical Threshold = 0</td>
</tr>
<tr>
<td>12</td>
<td>NVM Subsystem Reliability Degraded = 1</td>
</tr>
<tr>
<td>13</td>
<td>Media Read Only Mode = 0</td>
</tr>
<tr>
<td>14</td>
<td>Volatile Memory Backup Failed = 0</td>
</tr>
<tr>
<td>15</td>
<td>Temperature = 34.85 Celsius (308 Kelvin)</td>
</tr>
<tr>
<td>16</td>
<td>Available Spare = 84%</td>
</tr>
<tr>
<td>17</td>
<td>Available Spare Threshold = 10%</td>
</tr>
<tr>
<td>18</td>
<td>Percentage Used = 7%</td>
</tr>
<tr>
<td>19</td>
<td>Data Read = 1.20 PB (1,197,442,070,528,000 bytes)</td>
</tr>
<tr>
<td>20</td>
<td>Data Written = 1.21 PB (1,207,531,087,360,000 bytes)</td>
</tr>
</tbody>
</table>

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.
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

NVMe-MI Message

Request Message

Command Message

- NVMe-MI Command
- PCIe Command
- NVMe Admin Command

Control Primitive

Response Message

- Success
- Error

Other MCTP Messages (e.g., MCTP control)
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

<table>
<thead>
<tr>
<th>Command</th>
<th>O/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration Set</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Configuration Get</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Controller Health Status Poll</td>
<td>Mandatory</td>
</tr>
<tr>
<td>NVM Subsystem Health Status Poll</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Read NVMe-MI Data Structure</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Reset</td>
<td>Mandatory</td>
</tr>
<tr>
<td>VPD Read</td>
<td>Mandatory</td>
</tr>
<tr>
<td>VPD Write</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Vendor Specific</td>
<td>Optional</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Command</th>
<th>O/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get Features</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Get Log Page</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Identify</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Firmware Activate/Commit</td>
<td>Optional</td>
</tr>
<tr>
<td>Firmware Image Download</td>
<td>Optional</td>
</tr>
<tr>
<td>Format NVM</td>
<td>Optional</td>
</tr>
<tr>
<td>Namespace Management</td>
<td>Optional</td>
</tr>
<tr>
<td>Security Send</td>
<td>Optional</td>
</tr>
<tr>
<td>Security Receive</td>
<td>Optional</td>
</tr>
<tr>
<td>Set Features</td>
<td>Optional</td>
</tr>
<tr>
<td>Vendor Specific</td>
<td>Optional</td>
</tr>
</tbody>
</table>
PCIe Commands

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

<table>
<thead>
<tr>
<th>Command</th>
<th>O/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe Configuration Read</td>
<td>Optional</td>
</tr>
<tr>
<td>PCIe Configuration Write</td>
<td>Optional</td>
</tr>
<tr>
<td>PCIe Memory Read</td>
<td>Optional</td>
</tr>
<tr>
<td>PCIe Memory Write</td>
<td>Optional</td>
</tr>
<tr>
<td>PCIe I/O Read</td>
<td>Optional</td>
</tr>
<tr>
<td>PCIe I/O Write</td>
<td>Optional</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Control Primitive</th>
<th>O/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pause</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Resume</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Abort</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Get State</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Replay</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>
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
Architected for Performance