LEGAL NOTICE:

This Technical Proposal is proprietary to the NVM Express, Inc. (also referred to as “Company”) and/or its successors and assigns.

NOTICE TO USERS WHO ARE NVM EXPRESS, INC. MEMBERS: Members of NVM Express, Inc. have the right to use and implement this Technical Proposal subject, however, to the Member’s continued compliance with the Company’s Intellectual Property Policy and Bylaws and the Member’s Participation Agreement.

NOTICE TO NON-MEMBERS OF NVM EXPRESS, INC.: If you are not a Member of NVM Express, Inc. and you have obtained a copy of this document, you only have a right to review this document or make reference to or cite this document. Any such references or citations to this document must acknowledge NVM Express, Inc. copyright ownership of this document. The proper copyright citation or reference is as follows: **“© 2008 to 2024 NVM Express, Inc. ALL RIGHTS RESERVED.”** When making any such citations or references to this document you are not permitted to revise, alter, modify, make any derivatives of, or otherwise amend the referenced portion of this document in any way without the prior express written permission of NVM Express, Inc. Nothing contained in this document shall be deemed as granting you any kind of license to implement or use this document or the specification described therein, or any of its contents, either expressly or impliedly, or to any intellectual property owned or controlled by NVM Express, Inc., including, without limitation, any trademarks of NVM Express, Inc.

LEGAL DISCLAIMER:

THIS DOCUMENT AND THE INFORMATION CONTAINED HEREIN IS PROVIDED ON AN “**AS IS**” BASIS. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, NVM EXPRESS, INC. (ALONG WITH THE CONTRIBUTORS TO THIS DOCUMENT) HEREBY DISCLAIM ALL REPRESENTATIONS, WARRANTIES AND/OR COVENANTS, EITHER EXPRESS OR IMPLIED, STATUTORY OR AT COMMON LAW, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE, VALIDITY, AND/OR NONINFRINGEMENT.

All product names, trademarks, registered trademarks, and/or servicemarks may be claimed as the property of their respective owners.

The NVM Express® design mark is a registered trademark of NVM Express, Inc.

NVM Express Workgroup
c/o VTM, Inc.
3855 SW 153rd Drive
Beaverton, OR 97003
USA
info@nvmexpress.org

NVM Express® Technical Proposal (TP)

Technical Proposal ID	6034a
Revision Date	2024.05.22
Builds on Specification(s)	NVM Express Management Interface Specification, Revision 1.2c NVM Express Base Specification, Revision 2.0c
References	TP6021 Status Reporting Enhancements TP6027b Reset Behavior Clarifications

Technical Proposal Author(s)

Name	**Company**
Austin Bolen	Dell Technologies
Manjunath AM	Dell Technologies

Technical Proposal Overview

This TP adds support for multiple Management Endpoints per port which allows multiple Management Controllers to manage an NVMe Subsystem via NVMe-MI for high-availability use cases.

Revision History

Revision Date	**Author**	**Change Description**
2023.09.11	Austin Bolen	• Initial draft.
2023.09.12	Austin Bolen	• Added that several fields in the SMBus/I2C Element UDID apply to the NVM Subsystem instead of the Management Endpoint. • Cleanup.
2023.10.30	Austin Bolen	• Moved some fixes out this TP and to errata. • Updated the definition of Management Endpoint Capabilities in Identify Controller to account for multiple Management Endpoints per port. • More updates to indicate that status bits are per Management Endpoint in the out-of-band mechanism and per Controller in the in-band tunneling mechanism.
2023.11.06	Austin Bolen	• Editorial updates based on feedback from Mike Allison.
2023.12.28	Devin Allison	• Integrated
2024.01.03	Devin Allison	• Editorial updates based on feedback from Austin Bolen and Mike Allison.
2024.03.02	Austin Bolen	• Initial draft of TP6034a. • Made several clarifications to which status bits where being referenced in the Controller Health Status Poll command.
2024.04.29	Austin Bolen	• Minor wordsmithing based on feedback from Mike Allison.
2024.05.22	Devin Allison	• Integrated

Description for Changes Document for the NVM Express Management Interface Specification

New Features:

- Multiple Management Endpoints per Port (Optional)
    - Adds support for multiple Management Endpoints per port using MCTP bridging which allows multiple Management Controllers to manage an NVMe Subsystem via NVMe-MI over SMBus/I2C for high-availability use cases.
    - **Incompatible change**
        - Status bits in the out-of-band mechanism now have an instance per Management Endpoint instead of a single instance shared among all Management Endpoints in the NVM Subsystem.
    - References
        - TP6021 Status Reporting Enhancements
        - TP6027b Reset Behavior Clarifications

Description for Changes Document for the NVM Express Base Specification

New Features:

- Multiple Management Endpoints per Port (Optional)
    - Adds support for multiple Management Endpoints per port using MCTP bridging which allows multiple Management Controllers to manage an NVMe Subsystem via NVMe-MI over SMBus/I2C for high-availability use cases.

Markup Conventions:

Black:	Unchanged (however, hot links are removed)
~~Red Strikethrough~~:	Deleted
Blue:	New
Blue Highlighted:	TBD values, anchors, and links to be inserted in new text.
<Green Bracketed>:	Notes to editor or reader
Orange:	Text is pulled in from a referenced Technical Proposal
~~Orange~~:	Deleted text from a referenced Technical Proposal

Description of Specification Changes for NVM Express Management Interface Specification

1 Introduction

…

1.4 NVM Subsystem Architectural Model

…

The PCIe ports and SMBus/I2C port of an NVM Subsystem may ~~optionally~~ each contain ~~a single~~ zero or more ~~NVMe~~ Management Endpoints~~ (hereafter referred to as simply Management Endpoint)~~. A Management Endpoint is an MCTP endpoint that is the terminus and origin of MCTP packets/messages and is responsible for implementing the MCTP Base Protocol, processing MCTP Control Messages, and internal routing of Command Messages. Each Management Endpoint in an NVM Subsystem has a Port Identifier that is less than or equal to the Number of Ports (NUMP) field value in the NVM Subsystem Information Data Structure. If multiple Management Endpoints are supported on a port, then the NVMe Subsystem shall support MCTP bridging for MCTP endpoint ID discovery and assignment on that port (refer to the MCTP Base Specification).

…

Figure 3 illustrates an example NVM Subsystem. The NVM Subsystem contains a single Controller and there is a Management Endpoint associated with the PCIe port.

Figure 3: NVM Subsystem Associated with Single PCIe Port

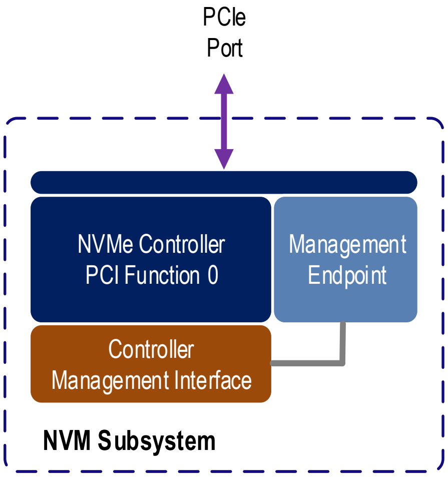


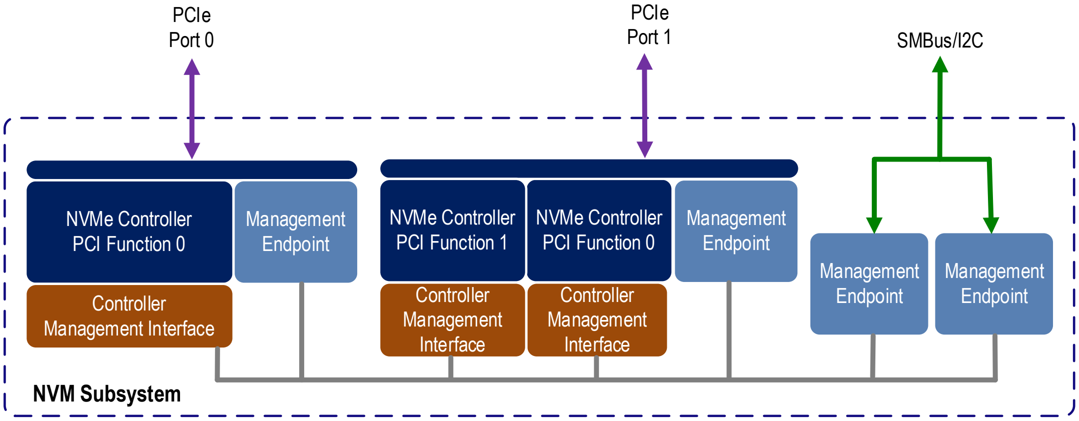
Figure 4 illustrates an example NVM Subsystem that is associated with a dual ported PCIe SSD. The NVM Subsystem contains one Controller associated with PCIe Port 0 and two Controllers associated with PCIe Port 1. There is a Management Endpoint associated with ~~the~~ each PCIe port and the SMBus/I2C port. Since the NVM Subsystem contains a Management Endpoint, all Controllers have an associated Controller Management Interface.

Dual-port PCIe SSDs are typically used in systems that provide two in-band hosts and two Management Controllers for redundancy in high-availability use cases. One Management Controller is connected to the PCIe VDM Management Endpoint on PCIe port 0 and the other Management Controller is connected to the PCIe VDM Management Endpoint on PCIe port 1.

PCIe SSD connectors typically only have a single SMBus/I2C port. To accommodate two Management Controllers, the PCIe SSD in this example implements two Management Endpoints on the SMBus/I2C port using MCTP bridging for discovery and assignment of multiple MCTP endpoint IDs (refer to the MCTP Base Specification). The method for determining which Management Controller communicates with which SMBus/I2C Management Endpoint and the method for the Management Controllers to arbitrate for control of the SMBus/I2C port are outside the scope of this specification.

<Note to editor: Update “Figure 4: NVM Subsystem with Dual-Ported PCIe Ports with SMBus/I2C Port” to add a second Management Endpoint on SMBus/I2C:>

Figure 4: NVM Subsystem with Dual ~~Ported~~ PCIe Ports and an SMBus/I2C Port



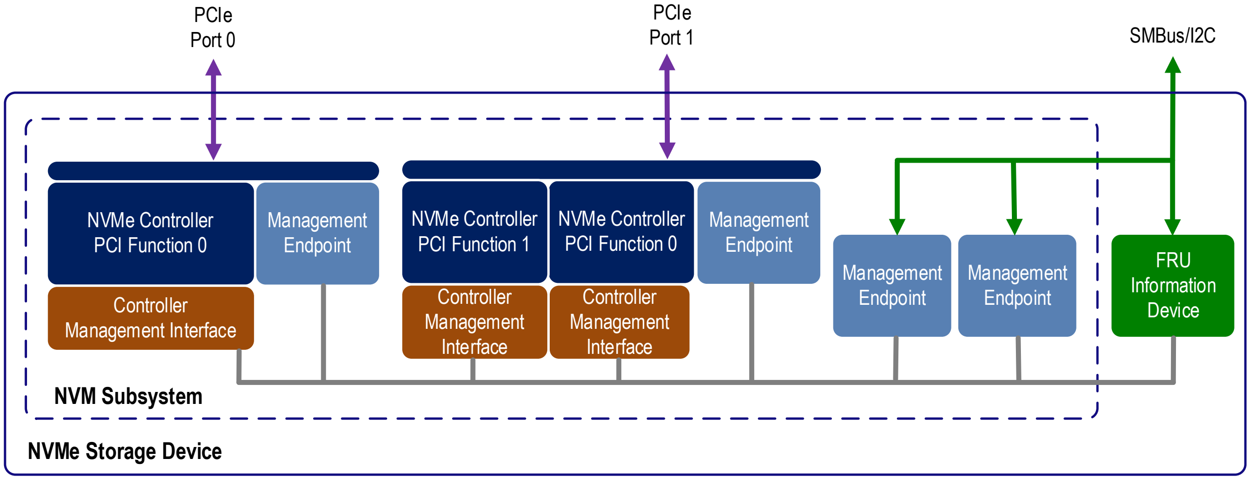
1.5 NVMe Storage Device Architectural Model

…

Figure 6 illustrates an NVMe Storage Device that is a dual-port PCIe SSD with an SMBus/I2C port and a FRU Information Device implemented using a Serial EEPROM.

<Note to editor: Update “Figure 6: Dual-Port PCIe SSD with SMBus/I2C” to add a second Management Endpoint on SMBus/I2C as follows:>

Figure 6: Dual-Port PCIe SSD with SMBus/I2C



…

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant’s agreement. Copyright © 2008 to 2024 NVM Express, Inc.

1.7 Conventions

Modify a portion of Section 1.7 (Conventions) as follows:

…

~~Reset – For the out-of-band mechanism, this column indicates the value the field is initialized to by an NVM Subsystem Reset. For the in-band tunneling mechanism, this column indicates the value the field is initialized to by a Controller Level Reset (refer to the NVM Express Base Specification).~~

1.8 Definitions

…

1.8.14 Management Endpoint ~~or NVMe Management Endpoint~~

An MCTP endpoint associated with an NVM Subsystem (e.g., an NVMe SSD or NVMe Enclosure) that is the terminus and origin of MCTP packets/messages and which processes Request Messages and transmits Response Messages.

1.8.26 PCIe Reset

A mechanism used to reset ~~a~~ one or more PCIe VDM Management Endpoints in an NVMe Storage Device or NVMe Enclosure. For more information, see section 8.3.5.

2 Physical Layer

This section describes the physical layers supported by this specification for NVMe Storage Devices or NVMe Enclosures.

2.1 PCI Express

PCI Express is used as a physical layer in both the out-of-band mechanism and the in-band tunneling mechanism in this specification.

For the out-of-band mechanism, a PCIe port in an NVMe Storage Device or NVMe Enclosure may implement ~~a~~ one or more Management Endpoints. If the PCIe port implements ~~a~~ one or more Management Endpoints, then:

a) the PCIe port shall support MCTP over PCIe Vendor Defined Messages (VDMs) as specified by the MCTP PCIe VDM Transport Binding Specification;

b) the Management Endpoint should be associated with PCIe Function 0 on the upstream PCIe bus on the NVMe Storage Device or NVMe Enclosure; and

c) the Management Endpoint should not be associated with a PCIe SR-IOV Virtual Function.

For the in-band tunneling mechanism, host software issues NVMe Admin Commands (NVMe-MI Send and NVMe-MI Receive) to the NVMe Admin Queue over PCI Express. Refer to the NVM Express Base Specification and section 4.3 of this specification for additional details on the NVMe-MI Send and NVMe-MI Receive commands.

2.2 SMBus/I2C

This section defines the requirements for an NVMe Storage Device or NVMe Enclosure that implements an SMBus/I2C port. The SMBus/I2C physical layer is only applicable for the out-of-band mechanism.

If an NVMe Storage Device or NVMe Enclosure implements an NVM Subsystem with ~~a~~ one or more Management Endpoints associated with an SMBus/I2C port, then that port shall comply to the MCTP SMBus/I2C Transport Binding Specification.

…

Figure 16: SMBus/I2C Elements and Requirements

SMBus/I2C Element	Default SMBus/I2C Address		SMBus ARP Support	Required Element Presence
	Hex Format	Binary Format [1]		
FRU Information Device	A6h	1010_011xb	Optional	Required on an NVMe Storage Device with *no* Expansion Connectors. Undefined on NVMe Enclosures.
FRU Information Device	A4h	1010_010xb	Optional	Required on Carriers (i.e., an NVMe Storage Device with *one or more* Expansion Connectors). Undefined on NVMe Enclosures.
SMBus/I2C Management Endpoint	3Ah	0011_101xb	Optional	Required if an NVMe Storage Device or NVMe Enclosure contains ~~an~~ one or more SMBus/I2C Management Endpoints.
SMBus/I2C Mux	E8h	1110_100xb	Optional	For NVMe Storage Devices, required if there is more than one SMBus/I2C element on any SMBus/I2C channel with the same SMBus/I2C address that does not support ARP. Undefined on NVMe Enclosures.
Basic Management Command [2]	D4h	1101_010xb	Optional	For NVMe Storage Devices, not recommended for new designs. Undefined on NVMe Enclosures.
NOTES: 1. The x represents the SMBus/I2C read/write bit. 2. The NVMe Basic Management Command is defined in Appendix A as an informative technical note.				

…

Figure 17: SMBus/I2C Element UDID

<table>
<tr><th>Bits</th><th>Field</th><th>Description</th></tr>
<tr><td colspan="3">…</td></tr>
<tr><td>111:96</td><td>Vendor ID</td><td>This field contains the PCI-SIG vendor ID for the NVM Subsystem ~~Management Endpoint~~.</td></tr>
<tr><td>95:80</td><td>Device ID</td><td>This field contains a vendor assigned device ID for the NVM Subsystem ~~Management Endpoint~~.</td></tr>
<tr><td>79:64</td><td>Interface</td><td>This field defines the SMBus version and the Interface Protocols supported.
<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>15:08</td><td>Reserved</td></tr>
<tr><td>07</td><td><b>ZONE:</b> This bit shall be cleared to ‘0’.</td></tr>
<tr><td>06</td><td><b>IPMI:</b> This bit shall be cleared to ‘0’.</td></tr>
<tr><td>05</td><td><b>ASF:</b> This bit shall be set to ‘1’. Refer to the MCTP SMBus/I2C Transport Binding Specification.</td></tr>
<tr><td>04</td><td><b>OEM:</b> This bit shall be set to ‘1'.</td></tr>
<tr><td>03:00</td><td><b>SMBus Version:</b> This field shall be set to 4h for SMBus Version 2.0, or to 5h for SMBus Version 3.0 and 3.1.</td></tr>
</table></td></tr>
<tr><td>63:48</td><td>Subsystem Vendor ID</td><td>This field contains the PCI-SIG vendor ID for the NVM Subsystem ~~Management Endpoint~~.</td></tr>
<tr><td>47:32</td><td>Subsystem Device ID</td><td>This field contains a vendor assigned device ID for the NVM Subsystem ~~Management Endpoint~~.</td></tr>
<tr><td colspan="3">…</td></tr>
</table>

5 Management Interface Command Set

…

5.1 Configuration Get

…

5.1.1 SMBus/I2C Frequency (Configuration Identifier 01h)

The SMBus/I2C Frequency configuration indicates the current frequency of each Management Endpoint on the SMBus port, if applicable.

The configuration specific fields in the NVMe Management Dword 0 field are shown in Figure 67. The configuration specific fields in the NVMe Management Dword 1 field are reserved. The current SMBus/I2C Frequency configuration is returned in the NVMe Management Response field as shown in Figure 68.

Figure 67: SMBus/I2C Frequency – NVMe Management Dword 0

Bits	Description
31:24	**Port Identifier:** This field specifies the port whose SMBus/I2C Frequency is indicated.
23:08	Reserved
07:00	**Configuration Identifier:** This field specifies the identifier of the Configuration that is being read. Refer to Figure 66.

Figure 68: SMBus/I2C Frequency – NVMe Management Response

<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>23:04</td><td>Reserved</td></tr>
<tr><td>03:00</td><td><b>SMBus/I2C Frequency:</b> The current frequency of each Management Endpoint on the SMBus/I2C~~. The default value for~~ port.
A Management Endpoint Reset (refer to section 8.3.3) shall set this field ~~following a reset or power cycle is~~ to 1h~~, if SMBus is supported~~.
<table>
<tr><th>Value</th><th>Description</th></tr>
<tr><td>0h</td><td>~~SMBus is not supported or disabled~~ This value is obsolete for implementations compliant with versions of this specification later than 1.2.</td></tr>
<tr><td>1h</td><td>100 kHz</td></tr>
<tr><td>2h</td><td>400 kHz</td></tr>
<tr><td>3h</td><td>1 MHz</td></tr>
<tr><td>4h to Fh</td><td>Reserved</td></tr>
</table>
</td></tr>
</table>

…

5.1.3 MCTP Transmission Unit Size (Configuration Identifier 03h)

The MCTP Transmission Unit Size configuration indicates the current MCTP Transmission Unit Size of each Management Endpoint on the port corresponding to the Port Identifier specified in the NVMe Management Dword 0 field.

The configuration specific fields in the NVMe Management Dword 0 field are shown in Figure 69. The configuration specific fields in the NVMe Management Dword 1 field are reserved. The current Transmission unit size of the specified port is returned in the NVMe Management Response field as shown in Figure 70.

Figure 69: MCTP Transmission Unit Size – NVMe Management Dword 0

Bits	Description
31:24	**Port Identifier:** This field specifies the port whose MCTP Transmission Unit Size is indicated.
23:08	Reserved
07:00	**Configuration Identifier:** This field specifies the identifier of the Configuration that is being read. Refer to Figure 66.

Figure 70: MCTP Transmission Unit Size – NVMe Management Response

Bits	Description
23:16	Reserved
15:00	**MCTP Transmission Unit Size:** This field contains the MCTP Transmission Unit Size in bytes to be used by each Management Endpoint on the port.~~ The default value for~~ A Management Endpoint Reset (refer to section 8.3.3) shall cause this field ~~following a reset or power cycle is ~~to be set to 40h (64).

5.2 Configuration Set

…

5.2.1 SMBus/I2C Frequency (Configuration Identifier 01h)

…

Figure 73: SMBus/I2C Frequency – NVMe Management Dword 0

<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>31:24</td><td><b>Port Identifier:</b> This field specifies the port whose SMBus/I2C Frequency is specified.</td></tr>
<tr><td>23:12</td><td>Reserved</td></tr>
<tr><td>11:08</td><td><b>SMBus/I2C Frequency:</b> This field specifies the new frequency for each Management Endpoint on the specified SMBus/I2C port.
<table>
<tr><th>Value</th><th>Description</th></tr>
<tr><td>0h</td><td>Reserved</td></tr>
<tr><td>1h</td><td>100 kHz</td></tr>
<tr><td>2h</td><td>400 kHz</td></tr>
<tr><td>3h</td><td>1 MHz</td></tr>
<tr><td>4h to Fh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>07:00</td><td><b>Configuration Identifier:</b> This field specifies the identifier of the Configuration that is being written. Refer to Figure 66.</td></tr>
</table>

5.2.2 Health Status Change (Configuration Identifier 02h)

…

<Note to editor: Globally add “the” in front of NVMe Management Dword 1 and “field” after.>

A Configuration Set command that selects Health Status Change ~~may be used to clear~~ clears corresponding bits selected in the NVMe Management Dword 1 field of the Composite Controller Status Flags field to ‘0’.

A Configuration Set command that selects Health Status Change operates independently for each Management Endpoint in the out-of-band mechanism and each Controller in the in-band tunneling mechanism.

<Note to editor: Globally add “the” in front of Composite Controller Status Flags and “field” after.>

An NVMe Storage Device or NVMe Enclosure supporting the Health Status Change Configuration Identifier in the out-of-band mechanism shall have an independent ~~copy~~ instance of the Composite Controller Status Flags field dedicated to each Management Endpoint~~ the out-of-band mechanism~~. In the out-of-band mechanism, a Configuration Set command that selects Health Status Change only applies to the ~~copy~~ instance of the Composite Controller Status Flags field dedicated to the Management Endpoint to which the Configuration Set command was issued~~ out-of-band mechanism~~. Refer to Figure 92~~ section 5.4~~ for more details on the Composite Controller Status Flags field.

An NVMe Storage Device or NVMe Enclosure supporting the Health Status Change Configuration Identifier in the in-band tunneling mechanism shall have an independent ~~copy~~ instance of the Composite Controller Status Flags field dedicated to each Controller ~~the in-band tunneling mechanism~~. In the in-band tunneling mechanism, a Configuration Set command that selects Health Status Change only applies to the ~~copy~~ instance of the Composite Controller Status Flags field dedicated to the Controller to which the Configuration Set command was issued ~~in-band tunneling mechanism~~.

…

5.2.3 MCTP Transmission Unit Size (Configuration Identifier 03h)

The MCTP Transmission Unit Size configuration specifies a new MCTP Transmission Unit Size for each Management Endpoint on the port corresponding to the specified Port Identifier. A Management Controller should check the maximum MCTP Transmission Unit Size for the port reported by the Management Endpoint using the Read NVMe-MI Data Structure command (refer to Figure 96).

…

Figure 77: MCTP Transmission Unit Size – NVMe Management Dword 1

Bits	Description
31:16	Reserved
15:00	**MCTP Transmission Unit Size:** This field contains the MCTP Transmission Unit Size in bytes to be used by each Management Endpoint on the port.

5.3 Controller Health Status Poll

…

The Controller Health Status Poll command operates independently for each Management Endpoint in the out-of-band mechanism and each Controller in the in-band tunneling mechanism.

<Note to editor: Globally add “the” in front of Controller Health Status Changed Flags and “field” after.>

An NVMe Storage Device or NVMe Enclosure supporting the Controller Health Status Poll command in the out-of-band mechanism shall have an independent ~~copy~~ instance of both the Controller Health Data Structure (refer to Figure 81) and the Controller Health Status Changed Flags field (refer to Figure 82) for each Controller in the NVM Subsystem dedicated to each Management Endpoint ~~the out-of-band mechanism~~. In the out-of-band mechanism, a Controller Health Status Poll command only applies to the ~~copy~~ instance of the Controller Health Data Structure and the Controller Health Status Changed Flags field dedicated to the Management Endpoint to which the Controller Health Status Poll command was issued ~~out-of-band mechanism~~.

An NVMe Storage Device or NVMe Enclosure supporting the Controller Health Status Poll command in the in-band tunneling mechanism shall have an independent ~~copy~~ instance of both the Controller Health Data Structure and the Controller Health Status Changed Flags field for each Controller in the NVM Subsystem dedicated to each Controller ~~the in-band tunneling mechanism~~. In the in-band tunneling mechanism, a Controller Health Status Poll command only applies to the ~~copy~~ instance of the Controller Health Data Structure and the Controller Health Status Changed Flags field dedicated to the Controller to which the Controller Health Status Poll command was issued ~~in-band tunneling mechanism~~.

…

Figure 79: Controller Health Status Poll – NVMe Management Dword 1

Bits	Description
31	**Clear Changed Flags (CCF):** If this bit is set to ‘1’, then the Management Endpoint shall perform the following steps atomically in the order listed: 1. perform the selection criteria based on the Controller Health Status Changed Flags field as described in section 5.3.1.2; 2. for Controllers whose Controller Health Data Structure is returned, copy the instance of the Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted to the corresponding Controller Health Status Changed field in the Controller Health Data Structure; and 3. for Controllers whose Controller Health Data Structure is returned, clear each bit in the instance of the Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted to ‘0’~~ in Controllers whose Controller Health Data Structure is contained in the Response Data~~. If this bit is set to ‘1’, then the following bits in the Controller Status field in Controllers whose~~ the~~ Controller Health Data Structure is returned (refer to Figure 80) shall be cleared to ‘0’: - Namespace Attribute Changed (NAC); - Firmware Activated (FA); and - Telemetry Controller-Initiated Data Available (TCIDA). The Controller Health Status Changed Flags field and the following bits in the Controller Status field in the Controller Health Data Structure shall not be modified in Controllers whose Controller Health Data Structure is not returned~~ contained in the Response Data~~: - Namespace Attribute Changed (NAC); - Firmware Activated (FA); and - Telemetry Controller-Initiated Data Available (TCIDA). If this bit is cleared to ‘0’, then the Controller Health Status Changed Flags field and the following bits in the Controller Status field in the Controller Health Data Structure shall not be modified in any Controller: - Namespace Attribute Changed (NAC); - Firmware Activated (FA); and - Telemetry Controller-Initiated Data Available (TCIDA).
30:05	Reserved
04	**Critical Warning (CWARN):** If this bit is set to ‘1’, then a Controller Health Data Structure shall be returned for Controllers with the Critical Warning bit set to ‘1’ in the~~ir~~ instance of that Controller’s Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted unless excluded by other selection criteria as described in section 5.3.1. If this bit is set to ‘1’, then a Controller Health Data Structure shall not be returned for Controllers with the Critical Warning bit cleared to ‘0’ in their Controller Health Status Changed Flags field unless included by other selection criteria as described in section 5.3.1. If this bit is cleared to ‘0’, then the Critical Warning bit in the Controller Health Status Changed Flags field shall not be included in the selection criteria described in section 5.3.1.
03	**Available Spare (SPARE):** If this bit is set to ‘1’, then a Controller Health Data Structure shall be returned for Controllers with the Available Spare bit set to ‘1’ in the~~ir~~ instance of that Controller’s Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted unless excluded by other selection criteria as described in section 5.3.1. If this bit is set to ‘1’, then a Controller Health Data Structure shall not be returned for Controllers with the Available Spare bit cleared to ‘0’ in their Controller Health Status Changed Flags field unless included by other selection criteria as described in section 5.3.1. If this bit is cleared to ‘0’, then the Available Spare bit in the Controller Health Status Changed Flags field shall not be included in the selection criteria described in section 5.3.1.

Figure 79: Controller Health Status Poll – NVMe Management Dword 1

Bits	Description
02	**Percentage Used (PDLU):** If this bit is set to ‘1’, then a Controller Health Data Structure shall be returned for Controllers with the Percent Used bit set to ‘1’ in the~~ir~~ instance of that Controller’s Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted unless excluded by other selection criteria as described in section 5.3.1. If this bit is set to ‘1’, then a Controller Health Data Structure shall not be returned for Controllers with the Percent Used bit cleared to ‘0’ in their Controller Health Status Changed Flags field unless included by other selection criteria as described in section 5.3.1. If this bit is cleared to ‘0’, then the Percent Used bit in the Controller Health Status Changed Flags field shall not be included in the selection criteria described in section 5.3.1.
01	**Composite Temperature Changes (CTEMP):** If this bit is set to ‘1’, then a Controller Health Data Structure shall be returned for Controllers with the Composite Temperature bit set to ‘1’ in the~~ir~~ instance of that Controller’s Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted unless excluded by other selection criteria as described in section 5.3.1. If this bit is set to ‘1’, then a Controller Health Data Structure shall not be returned for Controllers with the Composite Temperature bit cleared to ‘0’ in their Controller Health Status Changed Flags field unless included by other selection criteria as described in section 5.3.1. If this bit is cleared to ‘0’, then the Composite Temperature bit in the Controller Health Status Changed Flags field shall not be included in the selection criteria described in section 5.3.1.
00	**Controller Status Changes (CSTS):** If this bit is set to ‘1’, then a Controller Health Data Structure shall be returned for Controllers with the Controller Status Change bit set to ‘1’ in the~~ir~~ instance of that Controller’s Controller Health Status Changed Flags field dedicated to the Responder to which the Controller Health Status Poll command was submitted unless excluded by other selection criteria as described in section 5.3.1. If this bit is set to ‘1’, then a Controller Health Data Structure shall not be returned for Controllers with the Controller Status Change bit cleared to ‘0’ in their Controller Health Status Changed Flags field unless included by other selection criteria as described in section 5.3.1. If this bit is cleared to ‘0’, then the Controller Status Change bit in the Controller Health Status Changed Flags field shall not be included in the selection criteria described in section 5.3.1.

…

Figure 81: Controller Health Data Structure (CHDS)

Description
…
NOTES: 1. An NVM Subsystem Reset shall reset the instance of the Controller Status field dedicated to each Management Endpoint in the NVM Subsystem ~~the out-of-band mechanism~~ and the instance of the Controller Status field dedicated to each Controller in the NVM Subsystem. The instance of the Controller Status field dedicated to a Controller shall be reset by a Controller Level Reset (refer to the NVM Express Base Specification) of that Controller. Note that a Controller Level Reset may affect the Controller Status field in the out-of-band mechanism (e.g., a Controller Level Reset causes the CECO bit in the instance of the Controller Status bits dedicated to the out-of-band mechanism to be set to ‘1’). The instance of the Controller Status field dedicated to a Management Endpoint shall be reset by a Management Endpoint Reset of that Management Endpoint. No instance of the Controller Status field shall be reset by any other resets other than the resets documented by this note.

…

5.3.2 Filtering by Controller Health Status Changed Flags

…

Figure 82: Controller Health Status Changed Flags (CHSCF)

Bits	Reset [1]	Description
NOTES: 1. An NVM Subsystem Reset shall reset the instance of the Controller Health Status Changed Flags field dedicated to each Management Endpoint in the NVM Subsystem ~~the out-of-band mechanism~~ and the instance of the Controller Health Status Changed Flags field dedicated to each Controller in the NVM Subsystem. The instance of the Controller Health Status Changed Flags field dedicated to a Controller shall be reset by a Controller Level Reset (refer to the NVM Express Base Specification) of that Controller. Note that a Controller Level Reset may affect the Controller Health Status Changed Flags field in the out-of-band mechanism (e.g., a Controller Level Reset causes the CECO bit in the instance of the Controller Health Status Changed Flags field dedicated to the out-of-band mechanism to be set to ‘1’). The instance of the Controller Health Status Changed Flags field dedicated to a Management Endpoint shall be reset by a Management Endpoint Reset of that Management Endpoint. No instance of the Controller Health Status Changed Flags field shall be reset by any other resets other than the resets documented by this note.		

5.6 NVM Subsystem Health Status Poll

…

The NVM Subsystem Health Status Poll command operates independently ~~using~~ for each Management Endpoint in the out-of-band mechanism and each Controller in the in-band tunneling mechanism.

An NVMe Storage Device or NVMe Enclosure supporting the NVM Subsystem Health Status Poll command using the out-of-band mechanism shall have an independent ~~copy~~ instance of the NVM Subsystem Health Data Structure (refer to Figure 91) dedicated to each Management Endpoint ~~the out-of-band mechanism~~. In the out-of-band mechanism, an NVM Subsystem Health Status Poll command only applies to the ~~copy~~ instance of the NVM Subsystem Health Data Structure dedicated to the Management Endpoint to which the NVM Subsystem Health Status Poll command was issued ~~out-of-band mechanism~~.

An NVMe Storage Device or NVMe Enclosure supporting the NVM Subsystem Health Status Poll command using the in-band tunneling mechanism shall have an independent ~~copy~~ instance of the NVM Subsystem Health Data Structure dedicated to each Controller ~~the in-band tunneling mechanism~~. In the in-band tunneling mechanism, an NVM Subsystem Health Status Poll command only applies to the ~~copy~~ instance of the NVM Subsystem Health Data Structure dedicated to the Controller to which the NVM Subsystem Health Status Poll command was issued ~~in-band tunneling mechanism~~.

The NVM Subsystem Health Status Poll command uses the NVMe Management Dword 1 field as shown in Figure 90.

…

Figure TBD2: Composite Controller Status Data Structure (CCSDS)

Bytes	Description
…	

NOTES:

1. An NVM Subsystem Reset shall reset the instance of the Composite Controller Status Flags field dedicated to each Management Endpoint in the NVM Subsystem ~~the out-of-band mechanism~~ and the instance of the Composite Controller Status Flags field dedicated to each Controller in the NVM Subsystem.

   The instance of the Composite Controller Status Flags field dedicated to a Controller shall be reset by a Controller Level Reset (refer to the NVM Express Base Specification) of that Controller. Note that a Controller Level Reset may affect the Composite Controller Status Flags field in the out-of-band mechanism (e.g., a Controller Level Reset causes the CECO bit in the instance of the Composite Controller Status Flags field dedicated to the out-of-band mechanism to be set to ‘1’).

   The instance of the Composite Controller Status Flags field dedicated to a Management Endpoint shall be reset by a Management Endpoint Reset of that Management Endpoint.

   No instance of the Composite Controller Status Flags field shall be reset by any other resets other than the resets documented by this note.

5.7 Read NVMe-MI Data Structure

…

The Port Information data structure contains information about a port within the NVM Subsystem. The Port Identifier specifies the port. The Controller Identifier fields are reserved. The format is shown in Figure 96.

Figure 96: Port Information Data Structure

<table>
<tr><th>Bytes</th><th>Description</th></tr>
<tr><td>00</td><td><b>Port Type:</b> Specifies the port type.
<table>
<tr><th>Value</th><th>Definition</th></tr>
<tr><td>0h</td><td>Inactive</td></tr>
<tr><td>1h</td><td>PCIe</td></tr>
<tr><td>2h</td><td>SMBus</td></tr>
<tr><td>3h to FFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>01</td><td><b>Port Capabilities:</b> This field contains information about the capabilities of the port.
<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>7:1</td><td>Reserved</td></tr>
<tr><td>0</td><td><b>Command Initiated Auto Pause Supported (CIAPS):</b> If this bit is set to ‘1’, then the Command Initiated Auto Pause (CIAP) bit is supported in Command Messages on this port. If this bit is cleared to ‘0’, then the CIAP bit is not supported in Command Messages on this port.</td></tr>
</table></td></tr>
<tr><td>03:02</td><td><b>Maximum MCTP Transmission Unit Size:</b> The maximum MCTP Transmission Unit size that all Management Endpoints on the port are ~~is~~ capable of sending and receiving.
If the port does not support MCTP, then this field shall be cleared to 0h.
If the Port Type is PCIe and the port supports MCTP, then this field shall be set to a value between 64 bytes and the PCIe Max Payload Size Supported (refer to the PCI Express Base Specification), inclusive. All PCIe ports within an NVM Subsystem should report the same value in this field.
If the Port Type is SMBus and the port supports MCTP, then this field shall be set to a value between 64 bytes and 250 bytes, inclusive.</td></tr>
<tr><td>07:04</td><td><b>Management Endpoint Buffer Size:</b> This field specifies the size of the Management Endpoint Buffer in bytes when a Management Endpoint Buffer is supported.
A value of 0h in this field indicates that the Management Endpoint does not support a Management Endpoint Buffer.</td></tr>
<tr><td>31:08</td><td>Port Type Specific (refer to Figure 97 and Figure 98)</td></tr>
</table>

…

Figure 98: SMBus Port Specific Data

<table>
<tr><th>Bytes</th><th>Description</th></tr>
<tr><td colspan="2">…</td></tr>
<tr><td>11</td><td><b>Maximum Management Endpoint SMBus/I2C Frequency:</b> This field indicates the maximum SMBus/I2C frequency supported by ~~the~~ all Management Endpoints on the port.
<table>
<tr><th>Value</th><th>Definition</th></tr>
<tr><td>0h</td><td>Not supported</td></tr>
<tr><td>1h</td><td>100 kHz</td></tr>
<tr><td>2h</td><td>400 kHz</td></tr>
<tr><td>3h</td><td>1 MHz</td></tr>
<tr><td>4h to FFh</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td colspan="2">…</td></tr>
</table>

8 Management Architecture

…

8.2 Vital Product Data

…

8.2.4 NVMe PCIe Port MultiRecord Area

…

Figure 154: NVMe PCIe Port MultiRecord Area

<table>
<tr><th>Bytes</th><th>Factory Default</th><th>Description</th></tr>
<tr><td colspan="3">…</td></tr>
<tr><td>10</td><td>Impl Spec</td><td><b>MCTP Support:</b> This field contains a bit vector that specifies the level of support for the NVMe Management Interface.
<table>
<tr><th>Bits</th><th>Definition</th></tr>
<tr><td>7:1</td><td>Reserved</td></tr>
<tr><td>0</td><td>If this bit is set to ‘1’, then MCTP-based management commands are supported on all PCIe VDM Management Endpoints on the PCIe port. If this bit is cleared to ‘0’, then MCTP-based management commands are not supported on the PCIe port.</td></tr>
</table></td></tr>
<tr><td colspan="3">…</td></tr>
</table>

…

8.2.5 Topology MultiRecord Area

…

8.2.5.7 NVM Subsystem Element Descriptor

The NVM Subsystem Element Descriptor is shown in Figure 172 and is used to describe an NVM Subsystem contained in the NVMe Storage Device.

Figure 172: NVM Subsystem Element Descriptor

<table>
<tr><th>Bytes</th><th>Factory Default</th><th>Description</th></tr>
<tr><td colspan="3">…</td></tr>
<tr><td>03</td><td>3Ah
or
3Bh</td><td>**SMBus/I2C Address Info:** If the NVM Subsystem supports ~~an~~ MCTP ~~over~~ on all SMBus/I2C Management Endpoints on the SMBus/I2C port, then this field indicates the SMBus/I2C address for the MCTP over SMBus/I2C port and whether or not SMBus ARP is supported; otherwise, this field shall be cleared to 0h.
<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>7:1</td><td>**SMBus/I2C Address:** This field contains the 7-bit SMBus/I2C address. Refer to Figure 16 for requirements.</td></tr>
<tr><td>0</td><td>**ARP Capable:** This bit is set to ‘1’ if SMBus ARP is supported, else it is cleared to ‘0’. Refer to Figure 16 for requirements.</td></tr>
</table></td></tr>
<tr><td colspan="3">…</td></tr>
</table>

…

8.3 Reset Architecture

…

8.3.3 Management Endpoint Reset

…

Additional requirements and recommendations for Management Endpoint Resets are specified elsewhere in this specification. For example, a Management Endpoint Reset:

- resets bits and fields that are dedicated to each Management Endpoint in the out-of-band mechanism as defined in Figure 80, Figure 81, and Figure 90;

<Note to reader: The Composite Controller Status Flags field is added by TP6027b.>

- resets the value of the Composite Controller Status Flags field as defined by Figure TBD2;
- resets the value of the SMBus/I2C Frequency field as defined by Figure 68;
- resets the value of the MCTP Transmission Unit Size field as defined by Figure 69; and
- clears the Control Primitive Specific Response field to 0h as defined in Figure 42.

8.3.4 SMBus Reset

SMBus clock-low recovery is the ability to reset communication on all SMBus/I2C Management Endpoints on an SMBus/I2C port when the SMBus/I2C clock on that SMBus/I2C port is low for longer than $t_{TIMEOUT,MIN}$ (refer to the SMBus Specification). SMBus/I2C Management Endpoints shall support SMBus clock-low recovery. It is strongly recommended that any SMBus/I2C element other than the SMBus/I2C Management Endpoint (refer to Figure 16 for a list of SMBus/I2C elements) should support SMBus clock-low recovery. SMBus clock-low recovery shall cause an SMBus Reset. An SMBus Reset caused by SMBus clock-low recovery shall not cause ARP-assigned addresses to be reset to their default values. ~~All SMBus/I2C elements should support the recommendation for SMBus Reset when the SMBus/I2C clock is low for longer than $t_{TIMEOUT,MIN}$ (refer to the SMBus Specification).~~

Some form factors may also specify one or more form factor-specific ~~separate SMBus Reset~~ mechanisms to reset the SMBus (e.g., SMBRST# as defined in SFF-TA-1009 or the rising edge of the +3.3 Vaux rail as defined in the PCI Express SFF-8639 Module Specification). ~~If such mechanisms are~~ Any form factor-specific mechanisms to reset the SMBus supported by an NVMe Storage Device or NVMe Enclosure ~~NVM Subsystem, then the NVM Subsystem~~ shall cause an SMBus Reset. ~~propagate the reset to all SMBus/I2C elements on the NVM Subsystem and translate the reset, if needed, to Expansion Connector form factors.~~

An SMBus Reset shall cause a Management Endpoint Reset of ~~the~~ all SMBus/I2C Management Endpoints on the SMBus/I2C port. For any SMBus/I2C element other than the SMBus/I2C Management Endpoint (refer to Figure 16 for a list of SMBus/I2C elements), it is strongly recommended that an SMBus Reset should reset that SMBus/I2C element.

An SMBus Reset shall cause an SMBus reset mechanism defined for the Expansion Connector to be applied to each Expansion Connector in the NVMe Storage Device.

If ~~the~~ an SMBus/I2C Management Endpoint ~~port element on an NVM Subsystem~~ is transmitting a Response Message, then an SMBus Reset shall cause the SMBus/I2C port ~~it~~ to attempt to generate a STOP condition ~~as defined in~~ (refer to the SMBus Specification) within 5 ms from the assertion of SMBus Reset ~~or after the current data byte in the transfer process~~. The ~~NVM Subsystem shall remain idle on~~ SMBus/I2C port shall remain in the bus idle condition (refer to the SMBus Specification) for the remainder of the SMBus Reset assertion even if ~~other SMBus/I2C elements~~ a Management Controller attempts to ~~address it~~access the SMBus/I2C port. An ~~NVM Subsystem~~ SMBus/I2C port shall support SMBus/I2C accesses starting from the de-assertion of SMBus Reset within the same timing constraints as are applicable to transitioning from an unsupported to a supported power state as defined in section 8.1 ~~be ready to receive a START condition as defined in the SMBus Specification within 10 ms after SMBus Reset de-assertion~~.

…

8.3.TBD PCIe Reset

A PCIe Reset is generated by:

a) a Conventional Reset (refer to the PCI Express Base Specification); or

b) a Function Level Reset (refer to the PCI Express Base Specification).

PCIe Resets have additional impacts on in-band traffic and NVMe Controller operations which are outside the scope of this specification.

A Conventional Reset shall cause a Management Endpoint Reset of ~~the~~ all PCIe VDM Management Endpoints associated with the PCI Express port being reset. A Function Level Reset shall cause a Management Endpoint Reset of the PCIe VDM Management Endpoint associated with the Function being reset.

A PCIe VDM Management Endpoint shall support PCIe MCTP accesses after a PCIe Reset is de-asserted within the same timing constraints as are applicable to transitioning from an unsupported to a supported NVM Subsystem power state as defined by section 8.1.

Description of Specification Changes for NVM Express Base Specification

5 Admin Command Set

…

5.17 Identify command

…

5.17.2 Identify Data Structures

…

5.17.2.1 Identify Controller Data Structure (CNS 01h)

…

Figure 275: Identify – Identify Controller Data Structure, I/O Command Set Independent

<table>
<tr><th>Bytes</th><th>I/O[1]</th><th>Admin[1]</th><th>Disc[1]</th><th>Description</th></tr>
<tr><td colspan="5">…</td></tr>
<tr><td>255</td><td>M</td><td>M</td><td>M</td><td>**Management Endpoint Capabilities (MEC):** This field indicates the ~~capabilities of the~~ support for Management Endpoints in the NVM subsystem. Refer to the NVM Express Management Interface Specification for details.
<table>
<tr><th>Bits</th><th>Description</th></tr>
<tr><td>7:2</td><td>Reserved</td></tr>
<tr><td>1</td><td>**PCIe Port Management Endpoint (PCIEME):** If ~~set to '1', then~~ the NVM subsystem contains ~~a~~ one or more Management Endpoints on ~~a~~ one or more PCIe ports, then this bit shall be set to '1'; otherwise, this bit shall be cleared to '0'.</td></tr>
<tr><td>0</td><td>**SMBus/I2C Port Management Endpoint (SMBUSME):** If ~~set to '1', then~~ the NVM subsystem contains ~~a~~ one or more Management Endpoints on ~~an~~ the SMBus/I2C port, then this bit shall be set to '1'; otherwise, this bit shall be cleared to '0'.</td></tr>
</table></td></tr>
<tr><td colspan="5">…</td></tr>
</table>