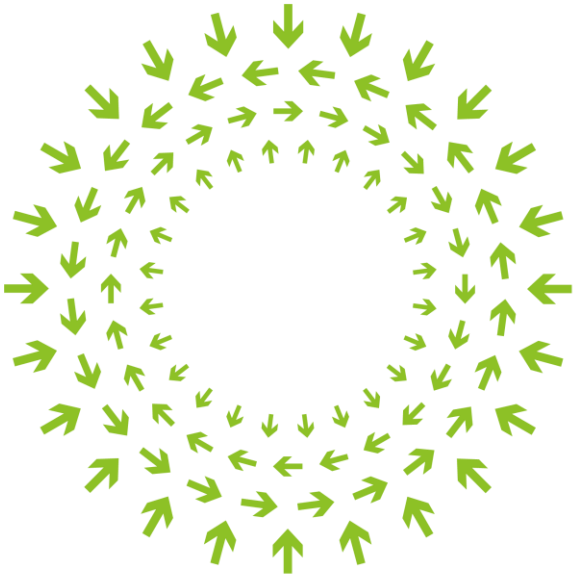


OPEN
Compute Project

NVMe Cloud SSD Specification

Version 1.0 (03182020)

Author: Ross Stenfort, Ta-Yu Wu, Facebook
Author: Lee Prewitt, Microsoft

Table of Contents

1 License OWF Option

Contributions to this Specification are made under the terms and conditions set forth in Open Web Foundation Contributor License Agreement ("OWF CLA 1.0") ("Contribution License") by:

Facebook
Microsoft

Usage of this Specification is governed by the terms and conditions set forth in the Open Web Foundation Final Specification Agreement ("OWFa 1.0").
Note: The following clarifications, which distinguish technology licensed in the Contribution License and/or Specification License from those technologies merely referenced (but not licensed), were accepted by the Incubation Committee of the OCP:

NONE

2 Overview

This document is to define the requirements for a cloud based NVMe™ SSD for use in data centers.

3 Scope

This document covers PCIe-attached SSDs using NVM Express.

4 NVM Express Requirements

4.1 Overview

Requirement ID	Description
NVMe-1	The device shall comply with all required features of the NVMe 1.4 specification. Optional features shall be implemented per the requirements of this specifications.
NVMe-2	Any optional features supported by the device not described in this document shall be clearly documented and disclosed to the customer.

NVMe-3	Any vendor unique features supported by the device not described in this document shall be clearly documented and disclosed to the customer.

4.2 NVMe Reset Supported

Requirement ID	Description
NVMeR-1	NVMe Subsystem reset shall be supported.
NVMeR-2	NVMe controller reset shall be supported.

4.3 NVMe Controller Configuration and Behavior

Requirement ID	Description
NVMe-CFG-1	The default arbitration shall be Round-Robin. Weighted Round Robin with urgent Class Priority shall be supported.
NVMe-CFG-2	The device shall support a Maximum Data Transfer Size (MDTS) value of at least 256KB.
NVMe-CFG-3	The device firmware shall support reporting of CSTS.CFS as indicated in the NVMe Specification.
NVMe-CFG-4	The "Model Number" field in the Identify Controller Data Structure (CNS 01h, byte offset 24:63) shall be identical to the Model Part Number (MPN) on the label and in the product datasheet provided to the customer.
NVMe-CFG-5	The minimum supported queue depth shall be 1024 per submission queue.
NVMe-CFG-6	The minimum number of IO Queue Pairs shall be 64.
NVMe-CFG-7	Device shall support EIU64 to differentiate namespaces.
NVMe-CFG-8	Device shall support an NGUID per Namespace.

4.4 NVMe Admin Command Set

The device shall support the following mandatory and optional NVMe admin commands:

Requirement ID	Description
NVMe-AD-1	The device shall support all mandatory NVMe admin commands.
NVMe-AD-2	Identify – In addition to supporting all the mandatory CNS values and the associated mandatory fields within the CNS, the following optional fields in the CNS shall be supported: • Format progress indicator (FPI) • IO Performance and Endurance Hints ○ NSFEAT bit 4 = 1b
NVMe-AD-3	Namespace Management command shall be supported.

NVMe-AD-4	Namespace Attachment command shall be supported.
NVMe-AD-5	Format NVM command shall be supported. Secure Erase Settings (SES) values 000b, 001b and 010b shall be supported.
NVMe-AD-6	Support for NVMe-MI Send and Receive is not required.

4.4.1 Namespace Management/Attachment Commands

The namespace management command along with the attach/detach commands is used to increase device over-provisioning beyond the default minimum over-provisioning.

Requirement ID	Description
NSM-1	The namespace management commands shall be supported on all namespaces.
NSM-2	When creating a namespace, the default "Formatted LBA Size" parameter (FLBAS=0) in the Identify Namespace Data Structure (Byte 26) shall correspond to the default sector size set at the factory.
NSM-3	When formatting the device with the Format command, the default "LBA Format" parameter (LBAF=0) in Command Dword 10 bits 3:0 shall correspond to the default sector size set at the factory.

4.4.2 Namespace Utilization (NUSE)

Requirement ID	Description
NUSE-1	The NUSE shall be equal to the number of logical blocks currently allocated in the namespace. NUSE cannot be hardcoded to be equal to NCAP. See below for an example on a 200GB device: 1. After a physical secure erase (SES = 001b), NUSE would be zero. And the usage data would reflect that: 0.00 GB 2. After writing 1 GB worth of data, the usage data would show the following: 1.00 GB 3. After filling the device, the usage data would show the following: 200.00 GB 4. If the host issues a 10GB de-allocate command, the usage data would show the following: 190.00 GB

4.5 NVMe I/O Command Set

Requirement ID	Description
NVMe-IO-1	The device shall support all mandatory NVMe I/O commands.
NVMe-IO-2	The device shall support Dataset Management and at a minimum De-Allocate.

4.6 Optional NVMe Feature Support

The device shall also support the following NVMe features:

Requirement ID	Description
NVMe-OPT-1	Telemetry shall be supported. Both Host Initiated Telemetry and Controller Initiated Telemetry shall be supported.
NVMe-OPT-2	Timestamp shall be supported to align the devices internal logs.

4.7 Command Timeout

Device supplier shall disclose any I/O scenario that could violate these command timeouts.

Requirement ID	Description
CTO-1	ADMIN Commands shall take no more than 10 seconds from submission to completion.
CTO-2	The only exceptions to CTO-1 shall be Format, Self-Test and Sanitize and the TCG commands Revert, Revert SP and Change Key.
CTO-3	When CSTS.RDY is set to 1b, I/O Commands shall take no more than 8 seconds from submission to completion. The device shall not have more than 7 IOs take more than 2 seconds in one hour.
CTO-4	I/O command processing time shall not be a function of device capacity.

4.8 Log Page Requirements

4.8.1 Standard Log Page Requirements

Requirement ID	Description
STD-LOG-1	Error Information (Log Identifier 01h)
STD-LOG-2	SMART/Health Information (Log Identifier 02h)
STD-LOG-3	Under no conditions shall the Percentage Used field in the SMART/Health Information (Log Identifier 02h) be reset.
STD-LOG-4	The Percentage Used field in the SMART/Health Information (Log Identifier 02h) shall be based on the average P/E cycle of the device. In addition, this field shall be based on the actual P/E cycle count of the media and not on the Power On Hours (POH) of the device.
STD-LOG-5	Firmware Slot Information (Log Identifier 03h)
STD-LOG-6	Commands Supported and Effects (Log Page 0x05)
STD-LOG-7	Telemetry Host-Initiated (Log Page 0x07)
STD-LOG-8	Telemetry Controller-Initiated (Log Page 0x08)

4.8.2 Telemetry Logging and Interface for Failure Analysis

The following applies to telemetry logging as the ability to quickly debug failures is required:

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>TEL-1</td><td>The device shall track the operational/event history and any critical parameters that can be used to debug issues.</td></tr>
<tr><td>TEL-2</td><td>The supplier shall provide a table that categorizes the reason identifiers.</td></tr>
<tr><td>TEL-3</td><td>If any of the following list of conditions occur, the telemetry data shall be committed to non-volatile storage so that the data is saved:
1. Ungraceful/graceful power cycle
2. Reboot
3. Any time a SMART critical warning changes to a non-zero value
4. Any type of firmware asserts
5. Retrieval of log via the host interface
6. The device switches to a degraded mode during run-time
7. The SMART “End to End Correction Counts” count is incremented</td></tr>
<tr><td>TEL-4</td><td>The Reason Identifier field shall be the most recent failure identifier and shall not be cleared by a power cycle or reset.</td></tr>
<tr><td>TEL-5</td><td>The table below provides the specifications for the controller-initiated and the host-initiated log page “data areas”. Implementation of Data areas 2 and 3 are optional.
<table>
<tr><th>Data Area</th><th>Purpose</th><th>Data Area Size</th><th>Latency Impact to IO</th></tr>
<tr><td>1</td><td>Periodic logging for monitoring trends/problems</td><td>Vendor-specific</td><td>< 10ms max</td></tr>
</table>
</td></tr>
<tr><td>TEL-6</td><td>The default status is “DISABLED” for the controller-initiated log page.</td></tr>
<tr><td>TEL-7</td><td>All device error logs shall be committed to non-volatile memory.</td></tr>
</table>

4.8.3 SMART Cloud Heath Log (0xC0) - Vendor Unique Log page

Below are the requirements for the Cloud Health Log Page located at 0xC0:

Requirement ID	Description
SLOG-1	All values in the Vendor Log pages shall be persistent across power cycles unless otherwise specified.
SLOG-2	All counters shall be saturating counters (i.e. if the counter reaches the maximum allowable size it stops incrementing and does NOT roll back to 0 unless otherwise specified).

SLOG-3	All values in logs shall be little endian format.
SLOG-4	A normalized counter, unless otherwise specified, shall be reported as the following: 100% shall represent the number at factory exit. 1% shall represent the minimum amount to be reliable. A value of 0% means the device shall no longer be considered reliable. 100% shall be represented as 0x64.
SLOG-5	Devices shall support the attributes listed in section 5.14.1.2 of the NVMe specification version 1.4.
SLOG-6	A Read of any of the SMART logs (0x02 or 0xC0) shall not require an update of the SMART values. It shall be a simple read of the current data and shall not block IO.
SLOG-7	Unless otherwise specified, the device shall update these values in the background at least once every ten minutes.
SLOG-8	The composite and raw temperature sensor values shall be updated when the log page is accessed.
SLOG-9	All assert events and controller-initiated log captures will require an associated vendor-specific “Reason Identifier” that uniquely identifies the assert /controller condition.
SLOG-10	The device shall not lose any of the SMART (Health 0x02 or Cloud Health 0xC0) data logs which are more than 10 minutes old including across power cycles/resets.
SLOG-11	The device shall not lose any back up energy source failure information or SMART (Health 0x02 or Cloud Health 0xC0) critical warnings including across power cycles/resets.

4.8.4 SMART Cloud Attributes Log Page

The vendor-specific log page, 0xC0 (Cloud Attribute Log Page) shall be 512-bytes and defines the following attributes:

Req ID	Byte Address	Field	# of Bytes	Field description
SMART-1	15:0	Physical Media Units Written	16	Contains the number of bytes written to the media; this value includes both user and metadata written to the user and system areas. It shall be possible to use this attribute to calculate the Write Amplification Factor (WAF).
SMART-2	31:16	Physical Media Units Read	16	Contains the number of bytes read from the media from both the user and system areas.

<table>
<tr><td>SMART-3</td><td>39:32</td><td>Bad User NAND Blocks</td><td>8</td><td>The Raw count specifies the number of user NAND blocks that have been retired. The Normalized value is the percent of user blocks still available. The Normalized value shall be set to 0x64 and the Raw count shall be set to zero on factory exit. It should be noted there are 2 bytes for normalized and 6 bytes for raw count. See normalized definition above.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>39:38</td><td>Normalized value</td></tr><tr><td>37:32</td><td>Raw count</td></tr></table></td></tr>
<tr><td>SMART-4</td><td>47:40</td><td>Bad System NAND Blocks</td><td>8</td><td>The Raw count specifies the number of system NAND blocks that have been retired. The Normalized value is the percent of system spare blocks still available. The normalized value shall be set to 0x64 and the Raw count shall be set to zero on factory exit. It should be noted there are 2 bytes for normalized and 6 bytes for raw count. See normalized definition above.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>47:46</td><td>Normalized value</td></tr><tr><td>45:40</td><td>Raw count</td></tr></table>
A value of 0xFFFF_FFFF_FFFF_FFFF indicates that the Bad User NAND block count field above represents all blocks on the device and Bad System NAND block count field is invalid.</td></tr>
<tr><td>SMART-5</td><td>55:48</td><td>XOR Recovery count</td><td>8</td><td>Total number of times XOR was invoked to recover data in NAND. This shall cover all reads from NAND. Data recovery may have succeeded or failed. This shall be set to zero on factory exit.</td></tr>
<tr><td>SMART-6</td><td>63:56</td><td>Uncorrectable read error count</td><td>8</td><td>Total count of NAND reads that were not correctable by read retries, all levels of ECC, or XOR. This shall be a count of the number of times data recovery fails and an uncorrectable read error is returned to the host.</td></tr>
</table>

<table>
<tr><td>SMART-7</td><td>71:64</td><td>Soft ECC error count</td><td>8</td><td>Total count of NAND reads that were not correctable by first level ECC and requires invoking an intermediate recovery. This shall cover all NAND read accesses. Data recovery may have succeeded or failed. If the device has more than one intermediate recovery level, then this counter only increments when intermediate recovery level 1 is invoked.</td></tr>
<tr><td>SMART-8</td><td>79:72</td><td>End to End Correction Counts</td><td>8</td><td>A count of the detected and corrected errors by the end to end error correction which includes DRAM, SRAM, or other storage element ECC/CRC protection mechanism (not NAND ECC). All correctable errors shall result in a counter increase no matter what type of data the memory is protecting. All detected errors shall result in a counter increase unless the error is uncorrectable and occurred in the system region. In the latter case, the incomplete shutdown flag shall be flagged/incremented on the next power up. It should be noted there are 4 bytes for count of detected errors and 4 bytes for count of corrected errors.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>79:76</td><td>Corrected Errors</td></tr><tr><td>75:72</td><td>Detected Errors</td></tr></table></td></tr>
<tr><td>SMART-9</td><td>80</td><td>System data % used</td><td>1</td><td>A normalized cumulative count of the number of erase cycles per block since leaving the factory for the system (firmware and metadata) area. Starts at 0 and increments. 100 indicates that the estimated endurance has been consumed. Value may exceed 100 up to 255. This count shall increment regardless of what the backing media of the blocks are (e.g. SLC and TLC). If system data is split between media types, then this shall report the worst-case count so that the device wear out is clearly understood. This counter has a different behavior than the normalized counter definition in LOG-4. 100% (0x64) represents the device many</td></tr>
</table>

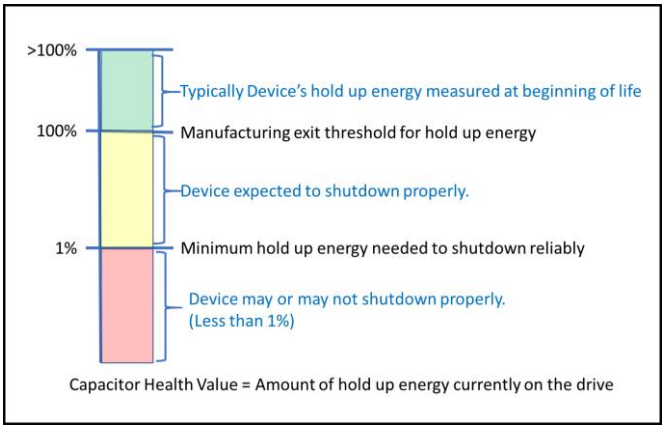
<table>
<tr><td></td><td></td><td></td><td></td><td>no longer function reliably as the max erase cycles has been hit.</td></tr>
<tr><td>SMART-10</td><td>87:81</td><td>Refresh Counts</td><td>7</td><td>This is a count of the number of blocks that have been re-allocated to maintain data integrity. This counter does not include creating free space due to garbage collection.</td></tr>
<tr><td>SMART-11</td><td>95:88</td><td>User data erase counts</td><td>8</td><td>The maximum and minimum erase counts across the user NAND blocks in the device. The host shall not be able to reset this counter. It should be noted there are 4 bytes for the maximum and 4 bytes for the minimum. Bad blocks shall not be included in this count.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>95:92</td><td>Minimum User Data Erase Count</td></tr><tr><td>91:88</td><td>Maximum User Data Erase Count</td></tr></table></td></tr>
<tr><td>SMART-12</td><td>97:96</td><td>Thermal throttling status and count</td><td>2</td><td>The current status of thermal throttling (enabled or disabled) and a count of the number of thermal throttling events. Note that there is 1 byte for the current status and 1 byte for the count. For devices that only have 1 throttle point only the first level throttle bit shall be set. This shall be set to zero on factory exit.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>97</td><td>Current Throttling Status

Current Status Byte definition:
0x00= unthrottled
0x01 = first level throttle
0x02 = 2nd level throttle
0x03 = 3rd level throttle
0x04-0xFF = Reserved</td></tr><tr><td>96</td><td>Number of thermal throttling events</td></tr></table></td></tr>
</table>

SMART-13	103:98	Reserved	6	Shall be set to 0x0.
SMART-14	111:104	PCIe Correctable Error count	8	Summation counter of all PCIe correctable errors (bad TLP, bad DLLP, receiver error, replay timeouts, replay rollovers). These counts shall only increment during run time. They shall not increment during training or power fail. This shall be set to zero on factory exit.
SMART-15	115:112	Incomplete Shutdowns	4	A count of the number of shutdowns that have occurred that did not completely flush all required user data and metadata to non-volatile memory for any reason. This shall be set to zero on factory exit.
SMART-16	119:116	Reserved	4	Shall be set to 0x0.
SMART-17	120	% Free Blocks	1	A normalized count of the number of blocks that are currently free (available) out of the total pool of spare (invalid) blocks. Free blocks means both blocks that have been erased and blocks that have all invalid data. Invalid blocks are blocks that are either marked invalid by device firmware or by the host (via de-allocate or overwrite). For example, if the total number of spare blocks is 100 and garbage collection has been able to reclaim (garbage collection and erase) 20 blocks, then this field reports 20%.
SMART-18	127:121	Reserved	7	Shall be set to 0x0.
SMART-19	129:128	Capacitor Health	2	This field is an indicator of the capacitor health and represents the capacitor holdup energy margin during operation. If no capacitor is present a value of 0XFFFF shall be reported. 100% represents the passing hold up energy threshold when a device leaves manufacturing. Thus, a device will typically report greater than 100% in this field after leaving manufacturing at beginning of life. 1% is the minimum hold up energy required to conduct a proper shutdown reliably. A value of 0% may or

				may not result in a device failing to shutdown properly. This value shall never go negative. Zero is the minimum. 
SMART-20	135:130	Reserved	6	Shall be set to 0x0.
SMART-21	143:136	Unaligned I/O	8	This is a count of the number of write IOs performed by the device that are not aligned to the indirection unit size (IU) of the device. Alignment indicates only the start of each IO. The length does not affect this count. This counter shall reset on power cycle. This counter shall not wrap. This shall be set to zero on factory exit.
SMART-22	151:144	Security Version Number	8	This is the security version number of the firmware image. The supplier increments this number any time it includes a fix of a security issue in any firmware image for this device.
SMART-23	159:152	NUSE	8	Namespace Utilization. This is a copy of the Namespace Utilization field defined in the Identify Namespace Data Structure Bytes 23:16.
SMART-24	175:160	PLP Start Count	16	This is a count of the number of times the device has initiated its power loss protection process due to supply voltage drop. This counter shall be incremented on the initial detection of the power loss condition. This does not include PLP health check operations.

SMART-25	191:176	Endurance Estimate	16	This field is an estimate of the total number of data bytes that may be written to the device over its lifetime assuming a write amplification of 1. (i.e., no increase in the number of write operations performed by the device beyond the number of write operations requested by a host). This value shall be equivalent to the Endurance Estimate field in the Endurance Group Log (Log Identifier 09h).
SMART-26	493:192	Reserved	302	Shall be set to 0x0.
SMART-27	495:494	Log Page Version	2	This indicates the version of the mapping this log page uses. Shall be set to 0x0002.
SMART-28	511:496	Log Page GUID	16	Shall be set to 0xAFD514C97C6F4F9CA4f2BFEA2810AFC5.

4.8.5 Error Recovery Log Page

The vendor-specific log page, 0xC1 shall be 512-bytes and define the following attributes:

Req ID	Byte Address	Field	# of Bytes	Field description
EREC-1	1:0	Panic Reset Wait Time	2	The amount of time the host should wait for the device panic workflow to complete in msec.
EREC-2	2	Panic Reset Action	1	Bit field indicating potential reset actions that can be taken. If no reset action needed, do not set any of the bits. More than 1 bit can be set and it's up to the host to decide the sequence of action(s) to take. Use Bit 0 if possible. If Bit 0 is not possible use Bit 1, etc.

Byte Address	Field Description
2	Panic Reset Action Byte definition: • Bit 0 = NVMe Controller Reset • Bit 1 = NVMe Subsystem Reset • Bit 2 = PCIe Function Level Reset

<table>
<tr><td></td><td></td><td></td><td></td><td><table><tr><td></td><td>• Bit 3 = PERST
• Bit 4 = Power Cycle Reset
• Bit 5 = PCIe Conventional Hot Reset
• Bit 7:6 = Reserved</td></tr></table></td></tr>
<tr><td>EREC-3</td><td>3</td><td>Device Recovery Action</td><td>1</td><td>The recovery action to take for handling a device panic condition. Value is dependent on the panic condition. Use 0x00 if possible.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>3</td><td>Device Recovery Action Byte definition:
• 0x00 = No Action Required
• 0x01 = Format NVM Required
• 0x02 = Vendor Specific Command Required
• 0x03 = Vendor Analysis Required
• 0x04 = Device Replacement Required
• 0x05 = Sanitize Required
• 0x06-0xFF = Reserved</td></tr></table></td></tr>
<tr><td>EREC-4</td><td>11:4</td><td>Panic ID</td><td>8</td><td>ID to identify the panic condition encountered. A Zero value indicates no panic. Value is dependent on the panic condition.
The following Panic ID values are reserved for Host defined fault codes for known panic conditions:
○ 0x00000000 00000000h – 0x00000000 0000FFFFh

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>11:4</td><td>Panic ID definition:
• 0x00000000 00000001h – Panic caused by flush failures or data loss during power loss handling.</td></tr></table></td></tr>
</table>

<table>
<tr><td>EREC-5</td><td>15:12</td><td>Device Capabilities</td><td>4</td><td>Field to indicate device capabilities.

<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>15:12</td><td>Device Capabilities definition:
Bit 0 = Panic AEN Supported: If set, indicates device supports using AEN to notify host of a panic condition*.
Bit 1 = Panic CFS Supported: If set, indicates device supports using CFS to notify host of a panic condition*.
Bit 31:2 = Reserved</td></tr></table>
*Note: It is valid for a device to indicate support for both Panic AEN Supported and Panic Controller Fatal Status Supported. If the device supports both, the device shall only use one of the panic notification mechanisms when reporting a given panic event.</td></tr>
<tr><td>EREC-6</td><td>16</td><td>Vendor Specific Recovery Opcode</td><td>1</td><td>Vendor specific command opcode to recover device from panic condition. Only valid when Device Recovery Action field value is 0x2. When Device Recovery Action field value is not 0x2, this field shall be set to 0x0.</td></tr>
<tr><td>EREC-7</td><td>19:17</td><td>Reserved</td><td>3</td><td>Shall be set to 0x0</td></tr>
<tr><td>EREC-8</td><td>23:20</td><td>Vendor Specific Command CDW12</td><td>4</td><td>CDW12 value for the Vendor Specific command to recover device from panic condition. Only valid when Device Recovery Action field value is 0x2. When Device Recovery Action field value is not 0x2, this field shall be set to 0x0.</td></tr>
<tr><td>EREC-9</td><td>27:24</td><td>Vendor Specific Command CDW13</td><td>4</td><td>CDW13 value for the Vendor Specific command to recover device from panic condition. Only valid when Device Recovery Action field value is 0x2. When Device Recovery Action field value is not 0x2, this field shall be set to 0x0.</td></tr>
<tr><td>EREC-10</td><td>493:28</td><td>Reserved</td><td>466</td><td>Shall be set to 0x0.</td></tr>
</table>

EREC-11	495:494	Log Page Version	2	This indicates the version of the mapping this log page uses. Shall be set to 0x0001.
EREC-12	511:496	Log Page GUID	16	Shall be set to 0x5A1983BA3DFD4DABAE3430FE2131D944.

4.8.6 Firmware Activation History

The vendor-specific log page, 0xC2 shall be 4096-bytes with the following functional requirements and field format.

Requirement ID	Description
FWHST-LOG-1	Lists the last twenty firmware images that were activated (not downloaded) on the drive. This is a circular buffer where the 21st entry is placed in entry 0 (byte offset 36 decimal).
FWHST-LOG-2	When the drive is first shipped from the factory, there are no entries recorded.
FWHST-LOG-3	An entry shall be recorded whenever a Firmware Commit command is received regardless of the commit action. Firmware downloads shall not generate an entry.
FWHST-LOG-4	Redundant activation events shall not generate a new entry to prevent the scrolling out of useful information. An entry shall be considered redundant if it meets ALL the criteria below: 1. Power on Hours is within 1 minute from the last RECORDED entry 2. Power cycle count is the same 3. Current firmware is the same 4. New firmware activated is the same 5. Slot number is the same 6. Commit Action Type is the same 7. The Result field has not changed
FWHST-LOG-5	Firmware Activation History's log page format shall follow the requirements below.

4.8.6.1 Firmware Activation History Log Page Format (Log Identifier 0xC2)

This log page defines the format for recording the Firmware Activation History.

Req ID	Byte Address	Field	# of Bytes	Field description
FAHL-1	0	Log Identifier	1	This field shall be set to 0xC2.

FAHL-2	3:1	Reserved	3	Shall be set to 0x0.
FAHL-3	7:4	Valid Firmware Activation History Entries	4	Contains the number of event entries in the log that are valid. Starts at 0 from the factory or after a Clear Firmware Update Activation History Set Features (See Section 3.11.2). Increments on each new log entry (see FWHST-LOG-4).
FAHL-4	71:8	Firmware Activation History Entry 0	64	This field contains the first firmware activation entry.
	…	…	…	…
	1287:1224	Firmware Activation History Entry 20	64	This field contains the last firmware activation entry.
FAHL-5	4077:1288	Reserved	2788	Shall be set to 0x0.
FAHL-6	4079:4078	Log Page Version	2	This indicates the version of the mapping this log page uses. Shall be set to 0x0001.
FAHL-7	4095:4080	Log Page GUID	16	Shall be set to 0xD11Cf3AC8AB24DE2A3F6DAB4769A796D.

4.8.6.2 Firmware Activation History Entry Format

This defines the History Entry format for recording Firmware Activation History events.

Req ID	Byte Address	Field	# of Bytes	Field description
FAHE-1	0	Entry Version Number	1	Indicates the version of this entry format used in the device. Shall be set to '1' (0x01).
FAHE-2	1	Entry Length (EL)	1	This field indicates the length in bytes of the entry log event data. Shall be set to '64' (0x40).
FAHE-3	3:2	Reserved	2	Shall be set to 0x0.

FAHE-4	5:4	Valid Firmware Activation History Entries	2	This field shall increment every time a firmware activation is attempted regardless of the result. This value shall be set to '0' (0x0) when the drive is shipped from manufacturing. This field shall be a saturating counter.
FAHE-5	13:6	Timestamp	8	This field shall indicate the Timestamp of when the firmware activation occurred. The format of this field shall be as defined in section 5.21.1.14 Timestamp (Feature Identifier 0Eh) of the NVMe 1.4 specification.
FAHE-6	21:14	Reserved	8	Shall be set to 0x0.
FAHE-7	29:22	Power Cycle Count	8	This field shall indicate the power cycle count in which the firmware activation occurred.
FAHE-8	37:30	Previous Firmware	8	This field shall indicate the previous firmware version running on the device before this firmware activation took place. The format of this field shall be as defined in field Firmware Revision (FR) section 5.15.2.2 Identify Controller Data Structure of the NVMe 1.4 specification.
FAHE-9	45:38	Current Firmware	8	This field shall indicate the activated firmware version that is running on the device after the firmware activation took place. The format of this field shall be as defined in field Firmware Revision (FR) section 5.15.2.2 Identify Controller Data Structure of the NVMe 1.4 specification.
FAHE-10	46	Slot Number	1	This field shall indicate the slot that the activated firmware is in.
FAHE-11	47	Commit Action Type	1	This field shall indicate the Commit action type associated with the firmware activation event.
FAHE-12	49:48	Result	2	This field shall indicate the results of the firmware activation event. A value of 0x0 shall represent the firmware commit was successful. A non-zero value shall represent the firmware commit was unsuccessful and the value represents the status code associated with the failure.

FAHE-13	63:50	Reserved	14	Shall be set to 0x0.

4.8.7 Firmware Update Requirements

This defines the requirements for firmware update in the device.

Requirement ID	Description
FWUP-1	A firmware activation history log shall be recorded. See section 4.8.6.
FWUP-2	Devices shall not have any restrictions on the number of firmware downloads supported.
FWUP-3	The firmware Commit command with the following Commit Action (CA) codes shall be supported: • 000b – Download only • 001b – Download and activate upon reset • 010b – Activate upon reset • 011b – Activate immediately without reset
FWUP-4	Firmware Image Download Command shall be supported.
FWUP-5	The Firmware Update Granularity (FWUG) field shall be set to 0x01h indicating that the granularity and alignment requirement of the firmware image being updated is 4096 Bytes.
FWUP-6	The device shall support a minimum of 2 slots for firmware update and may support up to 7.
FWUP-7	For firmware commit action 011b (firmware activation without reset), the device shall complete the firmware activation process and be ready to accept host IO and admin commands within 1 second from the receipt of the firmware commit command. The Maximum Time for Firmware Activation (MTFA) field shall not exceed Ah.
FWUP-8	The firmware shall prevent any firmware update operations from completing if the firmware downgrade is incompatible with the current version of firmware. The firmware rollback protection shall cover all cases including security.
FWUP-9	A single corrupted firmware image shall not result in the device no longer functioning. Multiple copies of the same firmware image shall be maintained to ensure the device can reliably boot.

4.9 De-Allocation Requirements

Requirement ID	Description
TRIM-1	The device shall support De-Allocate/TRIM.

TRIM-2	For data that has been De-Allocated (TRIM) the NVMe specification requires it to be 0, 1, or unchanged when read. Data returned shall only be 0, 1 or unchanged on a sector by sector basis.
TRIM-3	If data has been de-allocated and not written to when an unsafe power down event happens, the data shall be 0, 1 or unchanged when read.
TRIM-4	De-allocated addresses shall provide the performance and reliability benefits of overprovisioned space.
TRIM-5	The device shall support Garbage Collection during periods of no IO (Idle GC).
TRIM-6	Read latency shall not change more than 5% from baseline when the host is issuing De-Allocate/TRIM commands.
TRIM-7	Read latency shall not change more than 5% from baseline when the device is performing Idle GC.

4.10 Sector Size and Namespace Support

Requirement ID	Description
SECTOR-1	Devices 8 TB or less shall support both 4096-byte and 512-byte sectors and shall be formatted to one of these sector sizes from the factory.
SECTOR-2	Devices greater than 8 TB shall support 4096-byte sectors.
SECTOR-3	The device shall have one Namespace as shipped from the factory.

4.11 Set/Get Features Requirements

The device shall support the following additional vendor unique Set/Get Features Log Pages.

Requirement ID	Description
GETF-1	For any Get Feature Identifier defined in this section (3.11) Selection (SEL) values 00b to 11b in Dword 10 shall be supported.

4.11.1 Error Injection Set Feature Identifier (0xC0)

Feature to inject one or more error conditions to be reported by the device. If multiple Set Features commands for this feature are processed, then only information from the most recent successful command is retained (i.e., subsequent commands replace information provided by previous commands).

Req ID	Dword	Field	Bits	Field description
SERRI-1	0	Command Identifier (CID)	31:16	Shall be set as defined in NVMe Specification version 1.4.

SERRI-2	0	PRP or SGL for Data Transfer (PSDT)	15:14	Shall be set to 00b
SERRI-3	0	Reserved	13:10	Shall be set to zero
SERRI-4	0	Fused Operation (FUSE)	9:8	Shall be set to 00b
SERRI-5	0	Opcode (OPC)	7:0	Shall be set to 09h
SERRI-6	1	Namespace Identifier (NSID)	31:0	Shall be set to zero
SERRI-7	2:3	Reserved	31:0	Shall be set to zero
SERRI-8	4:5	Metadata Pointer (MPTR)	31:0	Shall be set to zero
SERRI-9	6:9	Data Pointer (DPTR)	31:0	Shall point to a physically contiguous 4096-byte address range containing 0 to 127 Error Injections Data Structure Entries
SERRI-10	10	Save (SV)	31	Shall be set to 0b
SERRI-11	10	Reserved	30:8	Shall be set to zero
SERRI-12	10	Feature Identifier (FID)	7:0	Shall be set to C0h
SERRI-13	11	Reserved	31:7	Shall be set to zero
SERRI-14	11	Number of Error Injections	6:0	This field shall specify the number of valid Error Injection Data Entries described in the address range pointed to by the Data Pointer (DPTR) field.
SERRI-15	12:15	Reserved	31:0	Shall be set to zero

Requirement ID	Description
ERRI-1	The maximum number of entries in the Number of Error Injections field shall be 127.
ERRI-2	A value of 0000000b in the Number of Error Injections field shall clear any outstanding error injection events.
ERRI-3	The error injections shall not overlap and may be listed in any order (e.g., ordering by error injection type is not required).
ERRI-4	Any unused entries in the Error Injection data structure shall have all fields set to 0 and shall be ignored by the device.
ERRI-5	The device shall abort the Error Injection Set Feature command if the request contains an error injection type that is not supported or the Single Instance value for the given Error Injection Type is not valid.
ERRI-6	Once the trigger conditions specified in an Error Injection Entry are met, the device shall inject the defined error event such that the host can detect the error through either an AEN being sent, the CFS bit being set or command being aborted.

4.11.1.1 Error Injection – Data Structure Entry

<table>
<tr><th>Req ID</th><th>Byte Address</th><th>Field</th><th># of Bytes</th><th>Field description</th></tr>
<tr><td>ERRIE-1</td><td>0</td><td>Error Entry Flags</td><td>1</td><td>Error Entry Flags definition:
Bit 0 = Error Injection Enable: If cleared to 0, indicates error injection is disabled. If set to 1, indicates error injection is enabled.
Bit 1 = Single Instance: If cleared to 0, indicates error injection is enabled until disabled. If set to 1, indicates a single instance error injection where a single error shall be injected. After a single instance error has been created, the error injection is considered disabled.
Bit 7:2 = Reserved</td></tr>
<tr><td>ERRIE-2</td><td>1</td><td>Reserved</td><td>1</td><td>Shall be set to zero</td></tr>
<tr><td>ERRIE-3</td><td>3:2</td><td>Error Injection Type</td><td>2</td><td>Error Injection type definition:
<table><tr><th>Value</th><th>Field Description</th></tr><tr><td>0h</td><td>Reserved</td></tr></table></td></tr>
</table>

				1h – Device Panic – CPU/Controller Hang 2h – Device Panic – NAND Hang 3h – Device Panic – PLP Defect 4h – Device Panic – Logical Firmware Error 5h – Device Panic – DRAM Corruption Critical Path 6h – Device Panic – DRAM Corruption Non-Critical Path 7h – Device Panic – NAND Corruption 8h – Device Panic – SRAM Corruption 9h – Device Panic – HW Malfunction Ah – Device Panic – No More NAND Spares Available Bh to FFFFh – Reserved
ERRIE-4	31:4	Error Injection Type Specific Definition	28	Error Injection Type specific definition.

4.11.1.2 Device Panic Error Injection Type

The device shall inject a device panic that the host can detect through either an AEN or the CFS bit being set. For the Device Panic type, a Single Instance value of 0 is not valid. Host shall perform the Panic Reset and Device Recovery actions specified in Log Page 0xC1.

Req ID	Byte Address	Field	# of Bytes	Field description
ERRIEDP-1	0	Error Entry Flags	1	Device Panic Error Entry Flags: • Bit 0 = Shall be set to 1 • Bit 1 = Shall be set to 1 • Bit 7:2 = Reserved
ERRIEDP-2	1	Reserved	1	Shall be set to zero

<table>
<tr><td>ERRIEDP-3</td><td>3:2</td><td>Error Injection Type</td><td>2</td><td>Shall be set to the range of 1h to 9h</td></tr>
<tr><td>ERRIEDP-4</td><td>31:4</td><td>Error Injection Type Specific Definition</td><td>28</td><td>Device Panic Error Injection information:
<table><tr><th>Byte Address</th><th>Field Description</th></tr><tr><td>5:4</td><td>Number of Reads to Trigger Device Panic (NRTDP): Indicates the number of Read commands the device shall process and complete before triggering a device panic.</td></tr><tr><td>31:6</td><td>Reserved: Shall be set to zero</td></tr></table></td></tr>
</table>

4.11.2 Error Injection Get Feature Identifier (0xC0)

This Get Feature returns the set of error injections that are enabled on the device. The attributes specified in section 3.11.2.1 are returned in Dword 0 of the completion queue entry and the Error Inject data structure specified section 3.11.1.1 is returned for each error injection in the data buffer for that command. If there are no currently enabled error injections, the data buffer returned shall contain all zeros. The device shall clear to zero all unused entries in the Error Injection data structure.

4.11.2.1 Error Injection – Get Features Completion Queue Entry Dword 0

Req ID	Field	Bits	Field description
GERRI-1	Reserved	31:7	Shall be set to zero
GERRI-2	Number of Error Injections (NUM)	6:0	This field indicates the number of enabled error injections returned in the command data buffer. See section 3.11.1.1 for the format of the entries.

4.11.3 Clear Firmware Update History Set Feature Identifier (0xC1)

Req ID	Dword	Field	Bits	Field description
CFUH-1	0	Command Identifier (CID)	31:16	Shall be set as defined in NVMe Specification version 1.4.

CFUH-2	0	PRP or SGL for Data Transfer (PSDT)	15:14	Shall be set to 00b
CFUH-3	0	Reserved	13:10	Shall be set to zero
CFUH-4	0	Fused Operation (FUSE)	9:8	Shall be set to 00b
CFUH-5	0	Opcode (OPC)	7:0	Shall be set to 09h
CFUH-6	1	Namespace Identifier (NSID)	31:0	Shall be set to zero
CFUH-7	2:3	Reserved	31:0	Shall be set to zero
CFUH-8	4:5	Metadata Pointer (MPTR)	31:0	Shall be set to zero
CFUH-9	6:9	Data Pointer (DPTR)	31:0	Shall be set to zero
CFUH-10	10	Save (SV)	31	Shall be set to 0b
CFUH-11	10	Reserved	30:8	Shall be set to zero
CFUH-12	10	Feature Identifier (FID)	7:0	Shall be set to C1h
CFUH-13	11	Clear Firmware Update History Log	31	Set to 1b to clear the Firmware Activation History Log Page (0xC2). The NVMe CLI plug in command “clear-fw-activate-history” can also perform this operation.
CFUH-14	11	Reserved	30:0	Shall be set to zero
CFUH-15	12:15	Reserved	31:0	Shall be set to zero

4.11.4 Read Only/Write Through Mode Set Feature Identifier (0xC2)

This Set Feature defines the mode to which the device shall transition at End of Life (EOL) or on failure of the Power Loss Protection (PLP) circuitry.

Requirement ID	Description
ROWTM-1	The device shall default from the factory to Read Only Mode (ROM) (01b).

Req ID	Dword	Field	Bits	Field description
SROWTM-1	0	Command Identifier (CID)	31:16	Shall be set as defined in NVMe Specification version 1.4.
SROWTM-2	0	PRP or SGL for Data Transfer (PSDT)	15:14	Shall be set to 00b
SROWTM-3	0	Reserved	13:10	Shall be set to zero
SROWTM-4	0	Fused Operation (FUSE)	9:8	Shall be set to 00b
SROWTM-5	0	Opcode (OPC)	7:0	Shall be set to 09h
SROWTM-6	1	Namespace Identifier (NSID)	31:0	Shall be set to zero
SROWTM-7	2:3	Reserved	31:0	Shall be set to zero
SROWTM-8	4:5	Metadata Pointer (MPTR)	31:0	Shall be set to zero
SROWTM-9	6:9	Data Pointer (DPTR)	31:0	Shall be set to zero

<table>
<tr><td>SROWTM-10</td><td>10</td><td>Save (SV)</td><td>31</td><td>Shall be set to 1b</td></tr>
<tr><td>SROWTM-11</td><td>10</td><td>Reserved</td><td>30:8</td><td>Shall be set to zero</td></tr>
<tr><td>SROWTM-12</td><td>10</td><td>Feature Identifier (FID)</td><td>7:0</td><td>Shall be set to C2h</td></tr>
<tr><td>SROWTM-13</td><td>11</td><td>End of Life Behavior</td><td>31:30</td><td>Field to indicate device write behavior at End of Life (EOL) or in the event of loss of PLP functionality.

<table><tr><th>Value</th><th>Field Description</th></tr><tr><td>00b</td><td>Reserved</td></tr><tr><td>01b</td><td>The device shall transition to Read Only Mode (ROM)</td></tr><tr><td>10b</td><td>The device shall transition to Write Through Mode (WTM)</td></tr><tr><td>11b</td><td>Reserved</td></tr></table></td></tr>
<tr><td>SROWTM-14</td><td>11</td><td>Reserved</td><td>29:0</td><td>Shall be set to zero</td></tr>
<tr><td>SROWTM-15</td><td>12:15</td><td>Reserved</td><td>31:0</td><td>Shall be set to zero</td></tr>
</table>

4.11.5 Read Only/Write Through Mode Get Feature Identifier (0xC2)

Dword 0 of command completion queue entry.

Req ID	Field	Bits	Field description
GROWTM-1	Reserved	31:3	Shall be set to zero
GROWTM-2	End of Life Behavior	2:0	Field to indicate what the device write behavior is configured for at End of Life (EOL) or in the event of loss of PLP functionality. The tables below define the required return values for each Selection (SEL) state. All other bit values are reserved.

<table>
<tr><td></td><td></td><td></td><td>
Current state (Selection (SEL) set to 00b)
<table>
<tr><th>Value</th><th>Field Description</th></tr>
<tr><td>001b</td><td>The device will transition to Read Only Mode (ROM) at End of Life (EOL) or on PLP failure</td></tr>
<tr><td>010b</td><td>The device will transition to Write Through Mode (WTM) at End of Life (EOL) or on PLP failure</td></tr>
</table>
Default state (Selection (SEL) set to 01b)
<table>
<tr><th>Value</th><th>Field Description</th></tr>
<tr><td>001b</td><td>Read Only Mode (ROM) is the factory default</td></tr>
<tr><td>010b</td><td>The Write Through Mode (WTM) is the factory default</td></tr>
</table>
Saved state (Selection (SEL) set to 10b)
<table>
<tr><th>Value</th><th>Field Description</th></tr>
<tr><td>001b</td><td>The saved state is set to Read Only Mode (ROM)</td></tr>
<tr><td>010b</td><td>The saved state is set to Write Through Mode (WTM)</td></tr>
</table>
Capabilities (Selection (SEL) set to 11b)
<table>
<tr><th>Value</th><th>Field Description</th></tr>
<tr><td>101b</td><td>This feature is saveable, changeable and not namespace specific</td></tr>
</table>
</td></tr>
</table>

4.11.6 Clear PCIe Correctable Error Counters Set Feature Identifier (0xC3)

Req ID	Dword	Field	Bits	Field description
CPCIE-1	0	Command Identifier (CID)	31:16	Shall be set as defined in NVMe Specification version 1.4.
CPCIE -2	0	PRP or SGL for Data	15:14	Shall be set to 00b

		Transfer (PSDT)		
CPCIE -3	0	Reserved	13:10	Shall be set to zero
CPCIE -4	0	Fused Operation (FUSE)	9:8	Shall be set to 00b
CPCIE-5	0	Opcode (OPC)	7:0	Shall be set to 09h
CPCIE-6	1	Namespace Identifier (NSID)	31:0	Shall be set to zero
CPCIE-7	2:3	Reserved	31:0	Shall be set to zero
CPCIE-8	4:5	Metadata Pointer (MPTR)	31:0	Shall be set to zero
CPCIE-9	6:9	Data Pointer (DPTR)	31:0	Shall be set to zero
CPCIE-10	10	Save (SV)	31	Shall be set to 0b
CPCIE-11	10	Reserved	30:8	Shall be set to zero
CPCIE-12	10	Feature Identifier (FID)	7:0	Shall be set to C3h
CPCIE-13	11	Clear PCIe Error Counters	31	Set to 1b to clear all PCIe correctable error counters in log page 0xCA. The NVMe CLI plug in command “clear-pcie-correctable-errors” can also perform this operation.
CPCIE-14	11	Reserved	30:0	Shall be set to zero
CPCIE-15	12:15	Reserved	31:0	Shall be set to zero

4.11.7 Enable IEEE1667 Silo Set Feature Identifier (0xC4)

This Set Feature shall return an error if the OPAL Security state is not manufactured_inactive.

Req ID	Dword	Field	Bits	Field description
S1667-1	0	Command Identifier (CID)	31:16	Shall be set as defined in NVMe Specification version 1.4.
S1667-2	0	PRP or SGL for Data Transfer (PSDT)	15:14	Shall be set to 00b
S1667-3	0	Reserved	13:10	Shall be set to zero
S1667-4	0	Fused Operation (FUSE)	9:8	Shall be set to 00b
S1667-5	0	Opcode (OPC)	7:0	Shall be set to 09h
S1667-6	1	Namespace Identifier (NSID)	31:0	Shall be set to zero
S1667-7	2:3	Reserved	31:0	Shall be set to zero
S1667-8	4:5	Metadata Pointer (MPTR)	31:0	Shall be set to zero
S1667-9	6:9	Data Pointer (DPTR)	31:0	Shall be set to zero
S1667-10	10	Save (SV)	31	Shall be set to 1b
S1667-11	10	Reserved	30:8	Shall be set to zero

S1667-12	10	Feature Identifier (FID)	7:0	Shall be set to C4h
S1667-13	11	Enable IEEE1667 Silo	31	If set to 0b, the IEEE 1667 silo shall be disabled no later than the next power cycle. If set to 1b, the IEEE 1667 silo shall be enabled no later than the next power cycle.
S1667-14	11	Reserved	30:0	Shall be set to zero
S1667-15	12:15	Reserved	31:0	Shall be set to zero

4.11.8 Enable IEEE1667 Silo Get Feature Identifier (0xC4)

Dword 0 of command completion queue entry.

Req ID	Field	Bits	Field description
G1667-1	Reserved	31:3	Shall be set to zero
G1667-2	IEEE1667 Silo Enabled	2:0	The tables below define the required return values for each Selection (SEL) state. All other values are illegal. Current state (Selection (SEL) set to 00b) **Value** / **Field Description** 000b: The IEEE1667 silo is currently disabled 001b: The IEEE1667 silo is currently enabled Default state (Selection (SEL) set to 01b) **Value** / **Field Description** 000b: The IEEE1667 silo factory default is disabled 001b: The IEEE1667 silo factory default is enabled Saved state (Selection (SEL) set to 10b) **Value** / **Field Description**

			000b: The IEEE1667 silo saved state is disabled 001b: The IEEE1667 silo saved state is enabled Capabilities (Selection (SEL) set to 11b) **Value** / **Field Description** 101b: This feature is saveable, changeable and not namespace specific

5 PCIe Requirements

The following are PCIe requirements.

Requirement ID	Description
PCI-1	The device shall support a PCIe Maximum Payload Size (MPS) of 256 bytes or larger.
PCI-2	The device Controller shall support modification of PCIe TLP completion timeout range as defined by the PCIe Base Spec.
PCI-3	The vendor shall disclose the vendor-specific timeout range definition if the controller deviates from the PCI Express Base Spec 3.1a Table 7-25 which defines Ranges A, B, C and D.
PCI-4	Disabling of PCIe Completion Timeout shall also be supported by the device Controller.
PCI-5	PCIe Conventional Reset Shall be supported: PCIe Cold or Warm Reset (*achieved by toggling of PERST#*)
PCI-6	PCIe Function Level Reset shall be supported.
PCI-7	PCIe Hot Reset shall be supported.

5.1 Boot Requirements

Requirement ID	Description
BOOT-1	The device shall support UEFI.
BOOT-2	An option ROM shall not be included.

5.2 PCIe Error Logging

The following table indicates where the PCIe physical layer error counters shall be logged. This is in addition to the aggregated PCIe error counters defined in section 3.8.4 SMART Cloud Attributes Log Page (0xC0).

Requirement ID	Event	Logging mechanism
PCIERR-1	Unsupported Request Error Status (URES)	Uncorrectable Error Status Register, Offset 0x4 in PCIe Base Specification 3.1a Section 7.10.2
PCIERR-2	ECRC Error Status (ECRCES)	
PCIERR-3	Malformed TLP Status (MTS)	
PCIERR-4	Receiver Overflow Status (ROS)	
PCIERR-5	Unexpected Completion Status (UCS)	
PCIERR-6	Completer Abort Status (CAS)	
PCIERR-7	Completion Timeout Status (CTS)	
PCIERR-8	Flow Control Protocol Error Status (FCPES)	
PCIERR-9	Poisoned TLP Status (PTS)	
PCIERR-10	Data Link Protocol Error Status (DLPES)	
PCIERR-11	Advisory Non-Fatal Error Status (ANFES)	Uncorrectable Error Status Register, Offset 0x10 in PCIe Base Specification 3.1a Section 7.10.5
PCIERR-12	Replay Timer Timeout Status (RTS)	Correctable PCIe Error Count in the SMART Cloud Attributes Log Page (0xC0).
PCIERR-13	REPLAY_NUM Rollover Status (RRS)	
PCIERR-14	Bad DLLP Status (BDS)	
PCIERR-15	Bad TLP Status (BTS)	
PCIERR-16	Receiver Error Status (RES)	

5.3 Low Power Modes

Requirement ID	Description
LPWR-1	If Active State Power Management (ASPM) is supported, the default firmware state shall be disabled.

5.4 PCIe Eye Capture

Requirement ID	Description
EYE-1	A utility shall be provided that will allow the user to capture the internal eye of the device in order to tune the signal integrity of the device to the target platform.

6 Reliability

6.1 UBER

Requirement ID	Description
UBER-1	The device shall support an Uncorrectable Bit Error Rate (UBER) of < 1 sector per 10^{17} bits read.

6.2 Power On/Off Requirements

6.2.1 Time to Ready and Shutdown Requirements

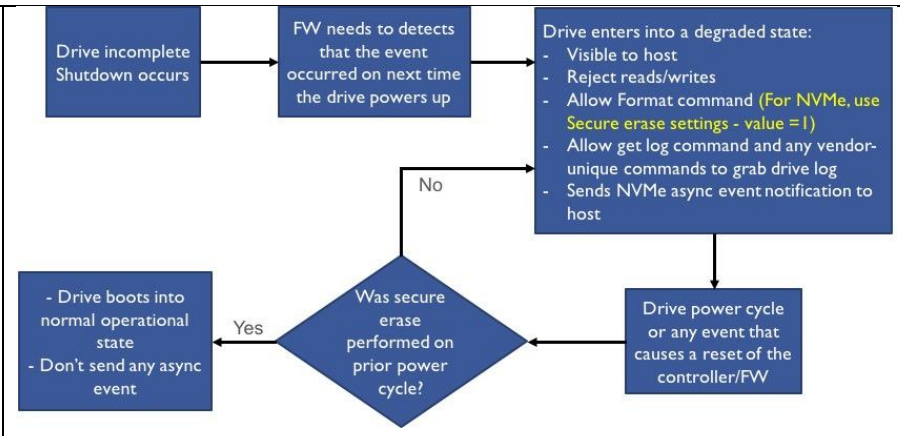
Requirement ID	Description
TTR-1	The device shall respond to the identify command 2 seconds from PRST#. PRST# may be sent as part of the power on sequence.
TTR-2	The device shall respond to I/O with 20 seconds of PRST#. PRST# may be sent as part of the power on sequence.
TTR-3	The device is expected to service ADMIN commands as soon as CSTS.RDY = 1.
TTR-4	The device shall keep CSTS.RDY = 0 until the device comes ready internally and is able to service commands.
TTR-5	The Shutdown Notification completion (CSTS.SHST) shall be received within 10 seconds of setting SHN bit.
TTR-6	The device Controller shall support the CC.SHN (Normal and Abrupt Shutdown Notifications) at a minimum.
TTR-7	When safe shutdown is completed successfully, the device shall not enter a rebuild/recovery mode on the next power on.
TTR-8	Shutdown Notification shall trigger flushing of all content within the device's internal (SRAM/ DRAM) cache(s) (*if one is present)*
TTR-9	The device firmware shall support reporting of CSTS.CFS as indicated in the NVMe Spec.
TTR-10	The device shall support full power-loss protection for all acknowledged data and metadata.
TTR-11	The Power-loss protection health check shall not impact IO latency.
TTR-12	Metadata rebuild due to an unexpected power loss shall not exceed 120 seconds and the device shall be fully operational after this.
TTR-13	Power-loss protection health check shall be performed by the firmware at least once every 24 hours.
TTR-14	While performing the power-loss protection health check, the device shall still have enough charge to be able to handle an ungraceful power loss properly.
TTR-15	In case of a graceful shutdown operation (*CC.SHN = 1 set by the NVMe device driver*), no data loss is tolerated.

TTR-16	An ungraceful shutdown event shall not make the device non-functional under any conditions.
TTR-17	The firmware/hardware algorithm shall deploy safeguards to prevent a false detection of power loss protection failure. Example of a false detection would be a glitch in any of the power loss circuitry readings which would cause a transient event to trigger a false power loss protection failure when the power loss protection hardware is healthy.
TTR-18	The device shall implement a power-loss protection (PLP) health check which can detect the capacitor holdup energy margin for the capacity health SMART attribute SMART-19. The PLP health check shall not just check for open/short capacitor conditions but shall measure the true available margin energy.

6.2.2 Incomplete/ Unsuccessful Shutdown

An incomplete/ unsuccessful shutdown is a graceful or ungraceful power down that did not complete 100% of the shutdown sequence for any reason (firmware hang/crash, capacitor failure, PLP circuit failure, etc.).

Requirement ID	Description
INCS-1	When the power-loss protection mechanism fails for any reason while power is applied, the device shall generate an AEN to the host and transition to the write behavior as defined in the Power Loss Behavior Set Feature Identifier (0xC2).
INCS-2	The device shall incorporate a shutdown checksum or flag as the very last piece of data written to flash to detect incomplete shutdown.
INCS-3	This checksum in INCS-2 shall be used on power-up to confirm that the previous shutdown was 100% successful.
INCS-4	The incomplete shutdown shall result in an increase in the Cloud SMART “Incomplete Shutdowns” counter and the NVMe standard SMART log (Log page 0x02) “critical warning” field shall have bit 2 set (NVM subsystem reliability).
INCS-5	The device shall still support data eradication as defined in section 11.3 Data Encryption and Eradication even if it is operating in “Read-only” mode, and it shall support admin commands to enable reading the sensor or SMART data.
INCS-6	When the device increments the “Incomplete Shutdowns” counter (SMART-15), it shall use the following recovery procedure at the next power up:



6.3 End to End Data Protection

Requirement ID	Description
E2E-1	All user data shall be protected using overlapping ECC and CRC protection mechanisms throughout the entire read and write path in the device including all storage elements (registers, caches, SRAM, DRAM, NAND, etc.).
E2E-2	At least one bit of correction and 2 bits of detection is required for all memories. This shall be for all memories regardless of function.
E2E-3	The entire DRAM addressable space shall to be protected with at least one-bit correction and 2 bits of detection scheme (SECDED). This includes but not limited to the following: • Flash translation layer (FTL) • Mapping tables (including metadata related to deallocated LBAs) • Journal entries • Firmware scratch pad • System variables • Firmware code
E2E-4	Silent data corruption shall not be tolerated under any circumstances.
E2E-5	The device shall include a mechanism to protect against returning the data from the wrong logical block address (LBA) to the host. It is acceptable that device stores additional/modified information to provides protection against returning wrong data to host. Device shall perform host LBA integrity checking on all transfers to and from the media.
E2E-6	All device metadata, firmware, firmware variables, and other device system data shall be protected by at least a single bit detection scheme.

6.4 Behavior on Firmware Crash, Panic or Assert

Requirement ID	Description
CRASH-1	After a firmware crash, panic or assert the device shall not allow write access to the media. The device shall allow read access only if it can guarantee data integrity.
CRASH-2	After a firmware crash, panic or assert, that is not a controller hardware failure, the device shall still support ADMIN commands including the ability to read any failure logs from the device to determine the nature of the failure.
CRASH-3	After the host performs the action specified in Device Recovery Action (EREC-3), the device shall allow full read and write access at full performance.
CRASH-4	If after a firmware crash, panic or assert there is the possibility of user data corruption, the Device Recovery Action shall require a Format.

6.5 Annual Failure Rate (AFR)

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>REL-1</td><td>The device shall meet an MTBF of 2.5 million hours (AFR of <= 0.35%) under the following environmental conditions:
<table>
<tr><th>Specification</th><th>Environment</th><th>Requirement</th></tr>
<tr><td>Temperature</td><td>Operational</td><td>• 0˚C to 50˚C (32˚F to 112˚F)</td></tr>
<tr><td rowspan="2">Humidity</td><td>Operational</td><td>• 10% to 90% non-condensing
• Yearly weighted average: < 80% RH
○ 90% of year: < 80%
○ 10% of year: 80% to 90%
• Maximum dewpoint: 29.4°C (85°F)</td></tr>
<tr><td>Non-Operational</td><td>• 5% to 95% non-condensing
• 38°C (100.4°F) maximum wet bulb temperature</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td>REL-2</td><td>The device shall meet an MTBF of 2.0 million hours (AFR of <= 0.44%) under the following environmental conditions:

<table>
<tr><th>Specification</th><th>Environment</th><th>Requirement</th></tr>
<tr><td>Temperature</td><td>Operational</td><td>• 0˚C to 55˚C (32˚F to 158˚F)</td></tr>
<tr><td rowspan="2">Humidity</td><td>Operational</td><td>• 10% to 90% non-condensing
• Maximum dewpoint: 29.4°C (85°F)</td></tr>
<tr><td>Non-operational</td><td>• 5% to 95% non-condensing
• 38°C (100.4°F) maximum wet bulb temperature</td></tr>
</table>
</td></tr>
<tr><td>REL-3</td><td>Supplier shall provide the temperature and humidity conditions used to determine the MTBF.</td></tr>
<tr><td>REL-4</td><td>Supplier shall provide UBER and AFR de-rating curves for the combined Temperature and Relative Humidity range shown in requirement REL-2 for up to 70˚C (158˚F).</td></tr>
</table>

6.6 Background Data Refresh

Requirement ID	Description
BKGND-1	The device shall support background data refresh while the device is powered on to ensure there is no data-loss due to power-on retention issues.
BKGND-2	The device shall be designed and tested to support the normal NAND operating temperature. For example, if the device is cooled to a composite temperature of 70°C (158°F) which in turn implies a NAND temperature of 80°C (176°F) this shall be accounted for.
BKGND-3	Background data refresh shall cover the entire device and be designed to continuously run in the background and not just during idle periods.

6.7 Wear-leveling

Requirement ID	Description
WRL-1	The device shall utilize the entire Endurance Group media capacity range whenever the device needs to wear-level a block. The device shall not restrict the wear-leveling range to a subset of the Endurance Group media capacity unless otherwise specified. If the device does not support Endurance Groups, it shall wear-level across the entire physical media of the Namespace.

7 Endurance

7.1 Endurance Data

Requirement ID	Description
ENDUD-1	The device documentation shall include the number of physical bytes able to be written to the device assuming a write amplification of 1. The units should be gigabytes (10^9 bytes).
ENDUD-2	The conditions to test device Endurance are: • 50/50 Read/Write workload (by number of I/Os) • 4kiB Read accesses aligned to 4kiB boundaries • 128kiB Write accesses aligned to 128kiB boundaries • Random pattern of Read addresses • Sequential pattern of Write addresses • 100% active range • 80% full device (80% data, 20% free space) • 0% compressible data • Ambient temperature 35°C (95°F) • Short stroked device if capacity is 2TB or greater (See EOL-2).
ENDUD-3	The Percentage Used in the SMART Heath Log shall track linearly with bytes written and at 100% it shall match the EOL value specified in ENDUD-1.

7.2 Retention Conditions

Since there are several factors that impact the device endurance, the table below provides the requirements for the datacenter environment.

Requirement ID	Description
RETC-1	Non-Operational (Powered-off) data retention (end of life) shall be at least 1 month at 25°C (77°F).
RETC-2	Operating (Powered-on) data retention shall be at least 7 years. For purposes of this requirement, the assumption is that the Terabytes Written (TBW) capability of the devices is used linearly over the lifetime. This requirement does not imply any specific warranty period.
RETC-3	The device shall not throttle its performance based on the endurance metric (endurance throttling).

7.3 Shelf Life

Requirement ID	Description
SLIFE-1	A new device may be kept as a datacenter spare and therefore shall be fully functional even if it sits on the shelf for up to 1 year at 40°C (104°F) before getting installed in the server. Device can be considered to be in new in box factory state.

7.4 End-of-Life (EOL)

Requirement ID	Description
EOL-1	Various types of samples are required for EOL testing: 1. Beginning of Life (Short stroked if required by EOL-2) 2. End of Life (Short stroked if required by EOL-2) 3. End of Life (Not short stroked if different than #2)
EOL-2	On 2 TB or larger devices, there shall be a method to “short stroke” the device. Media reserved for background operations shall be proportionally adjusted. If a “short stroked” firmware or tool is required, the “short stroked” capacity shall be 10% of the native device capacity.
EOL-3	Upon reaching 100% of specified device endurance, the device shall notify the host with an AEN.
EOL-4	The device shall continue to operate in a read/write mode as long as data is not at risk.
EOL-5	The device shall switch to read-only mode when the available spares field in the SMART / Health Information (Log Identifier 02h) reaches 0%. A value of 0% represents the device state where there is an insufficient number of spare blocks to support Host writes. After the drive switches to read-only mode, bit 2 of the Critical Warning field of section 5.14.1.2 SMART Attributes in the NVMe specification shall be set. The device shall generate a Critical Warning async notification (AEN) when the available spares value falls below the available spare threshold.

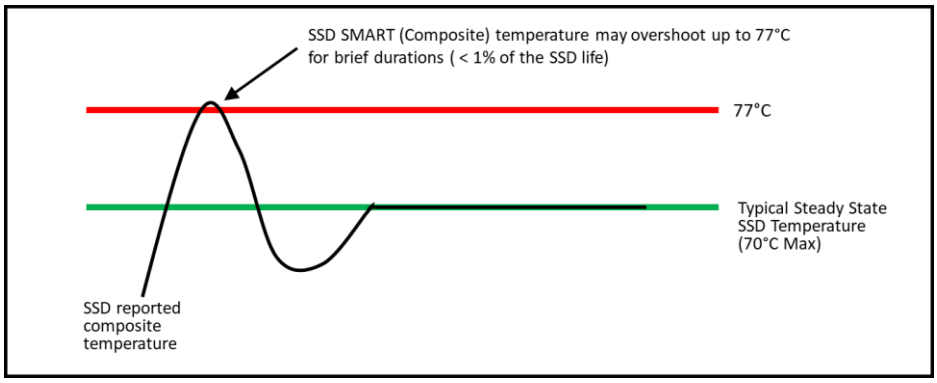
8 Thermal

8.1 Data Center Altitude

Requirement ID	Description

THERM-1	Support for data centers being located at an altitude of up to 10,000 feet above sea level is required. There shall be no de-rating up to 6,000 feet above sea level. Above 6,000 feet the derating shall be 0.9°C (1.6°F)/1000ft.
THRM-2	A thermal study with each platform is required. The thermal design shall be validated up to 35°C (95°F) ambient temperature for the platform with a worst-case airflow of 1.5 meters per second.
THRM-3	The device shall operate normally with relative humidity to be between 10% and 90%.

8.2 Thermal Throttling

Requirement ID	Description
TTROTTLE-1	The device shall implement a thermal throttling mechanism to protect the device in case of a failure or excursion that causes the device temperature to increase above its maximum specified temperature.
TTROTTLE-2	When a temperature throttle occurs, a Temperature Async Event Notification shall be issued to the host.
TTROTTLE-3	Thermal throttling shall only engage under certain failure conditions such as excessive server ambient temperature or multiple fan failures. The required behavior is illustrated below: 
TTROTTLE-4	The firmware algorithm shall deploy safeguards to prevent a false activation of either thermal throttling or thermal shutdown. Example of a false activation would be a glitch in any of the sensor readings which would cause the composite temperature to reach the thermal throttling or thermal shutdown limit.
TTROTTLE-5	A composite temperature of 77°C (170.6°F) shall be used for throttling.
TTROTTLE-6	Thermal throttling shall not start based on the rate of temperature increase or slew rate.

8.3 Temperature Reporting

Requirement ID	Description
TRPT-1	The device shall expose the current raw sensor readings from all the sensors on the device.
TRPT-2	The device's device-to-device composite temperature variation shall be +/- 1 degrees C. Two different devices shall not report a composite temperature greater than 2 degrees apart under the same environmental conditions, slot location, and workload.
TRPT-3	A single device's composite temperature shall not vary by more than +/-1°C (1.8°F) degrees once it is in a steady state under the same environmental conditions, slot location, and workload.
TRPT-4	The supplier shall provide the equation, settings, and thresholds used to calculate the composite temperature.

8.4 Thermal Shutdown

Requirement ID	Description
THRMS-1	If the device implemented a mechanism to shut down or halt the device at a given temperature, that temperature value shall be at 85°C (185°F) composite temperature or higher.

9 Form Factor Requirements

9.1 Generic Form Factor Requirements

Requirement ID	Description
GFF-1	The device shall be compliant to PCIe base specification 3.1a.
GFF-2	Vendor shall provide a PCIe compliance report.
GFF-3	The device shall support lane reversal with all lanes connected or partially connected lanes. (e.g. a x4 device shall support it for x4, x2, and x1 connections).
GFF-4	The device shall train to x1 when only one upstream port is available, to x2 when the upstream device provides only 2 lanes per device and to x4 when 4 lanes are present.
GFF-5	The device shall support hot swap on form factors that support hot swap.

9.2 Power Consumption Measurement Methodology

Requirement ID	Description
PCM-1	The device max average power consumption for any workload shall not exceed the maximum average power as configured in PWR-2 in a 500ms window with a sampling rate of 2ms or better. The measurement duration shall be at least 15 minutes on a pre-conditioned device.
PCM-2	The device peak power for any workload shall not exceed the max form factor power in a 100us window with a sampling rate of 4uS or better. The measurement duration shall be at least 15 minutes on a pre-conditioned device.
PCM-3	The peak power shall be no more than 1.5X higher than the max average power as defined in PWR-3.

9.3 Power Levels

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>PWR-1</td><td>The Power Management Set Feature (0x02) and the Power State descriptor table shall be supported.</td></tr>
<tr><td>PWR-2</td><td>The following power state descriptor table shall be supported to describe the devices operational power states. Additional power states greater than 35W are allowed.
<table>
<tr><th>Power State</th><th>Maximum Average Power (MAP)</th><th>Entry Latency (ENTLAT)</th><th>Exit Latency (EXLAT)</th></tr>
<tr><td>0</td><td>35W</td><td>IHV*</td><td>IHV</td></tr>
<tr><td>1</td><td>25W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>2</td><td>20W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>3</td><td>18W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>4</td><td>16W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>5</td><td>14W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>6</td><td>12W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>7</td><td>10W</td><td>IHV</td><td>IHV</td></tr>
<tr><td>8</td><td>8.25W</td><td>IHV</td><td>IHV</td></tr>
</table>
*Table entries containing IHV (Independent Hardware Vendor) are to be filled out by the manufacturer. If a given power state is not supported, then the device shall report FFFFFFFFh for the Entry Latency (ENTLAT) and Exit Latency (EXLAT).
All other entries in the power state descriptor table are “Don’t Care”.</td></tr>
</table>

PWR-3	The method of measurement for Maximum Average Power (MAP) is defined in PCM-1.
PWR-4	Power state entries above the maximum rated power envelope of the device shall not be in the table.
PWR-5	The Set Features for Power Management with the SV bit 31 in Command Dword 10 shall be supported so that the power level can be set and will be saved across power cycles.
PWR-6	The device, regardless of form factor or capacity, shall have an idle power of 5 Watts or less per European regulation.

9.4 M.2 Form Factor Requirements

Requirement ID	Description
FFM2-1	The device shall adhere to the M.2 specification with a size of 22mm x 110mm.
FFM2-2	The bottom-side height shall not exceed 1.5mm.
FFM2-3	The top-side height shall not exceed 2mm.
FFM2-4	The device shall use an M key.
FFM2-5	The device shall support PCIe Gen3 x4.
FFM2-6	The device shall support driving an activity LED through the connector.
FFM2-7	The LED should be lit solidly when power is applied and flashing when there is traffic going to the SSD.
FFM2-8	The device shall not use any pins that are defined in the m.2 specification for vendor unique functionality.
FFM2-9	The device shall support a protection scheme that protects against NAND block level failures.
FFM2-10	The protection scheme must also support NAND plane level failures without data or metadata loss.
FFM2-11	The Label shall be placed on the top side of the device.
FFM2-12	The device electricals shall follow the SMBUS connection as described below and in the PCI SIG ECN. (https://pcisig.com/sites/default/files/specification_documents/4_SMBus_interface_for_SSD_Socket_2_and_Socket_3.pdf)
FFM2-13	The device's SMBUS protocol shall comply to version 3.1 (http://smbus.org/specs/SMBus_3_1_20180319.pdf)

9.5 E1.S Form Factor Requirements

Requirement ID	Description
FFE1S-1	The device shall adhere to the latest revision of SFF-TA-1006.
FFE1S-2	At a minimum the device shall support PCIe Gen3 x4.

FFE1S-3	The device shall support activity and error LEDs.
FFE1S-4	The activity LED shall be lit solidly when power is applied and flashing when there is traffic going to the device.
FFE1S-5	The device shall support a protection scheme that protects against NAND block level failures.
FFE1S-6	The protection scheme must also support NAND plane level failures without data or metadata loss.
FFE1S-7	The amber LED shall meet the requirements specified in SFF TA-1009. The functioning of the Amber LED shall be independent of the 12V, 3.3Vaux and the state of the PWRDIS pin.
FFE1S-8	The thermal performance of the 9.5mm and 25mm cases and their associated pressure drops needs to be provided.
FFE1S-9	The PWRDIS pin shall be supported.
FFE1S-10	The SMBUS electrical connections shall follow the “DC Specification For 3.3V Logic Signaling” as defined in SFF-TA-1009 revision 2.0. Including Vih1 with a max of 3.465V.
FFE1S-11	The device’s SMBUS protocol shall comply to version 2.0 (http://smbus.org/specs/smbus20.pdf)

9.6 E1.L Form Factor Requirements

Requirement ID	Description
FFE1L-1	The device shall adhere to the latest revision of SFF-TA-1007.
FFE1L-2	At a minimum the device shall support PCIe Gen3 x4.
FFE1L-5	The device shall support activity and error LEDs.
FFE1L-6	The activity LED shall be lit solidly when power is applied and flashing when there is traffic going to the device.
FFE1L-7	The amber LED shall meet the requirements specified in SFF TA-1009. The functioning of the Amber LED shall be independent of the 12V, 3.3Vaux and the state of the PWRDIS pin.
FFE1L-8	The thermal performance of the 9.5mm and 18mm cases and their associated pressure drops shall be provided.
FFE1L-9	The PWRDIS pin shall be supported.
FFE1L-10	The SMBUS electrical connections shall follow the “DC Specification For 3.3V Logic Signaling” as defined in SFF-TA-1009 revision 2.0. Including Vih1 with a max of 3.465V.
FFE1L-11	The device’s SMBUS protocol shall comply to version 2.0 (http://smbus.org/specs/smbus20.pdf)

10 SMBUS support

10.1 SMBUS Requirements

Requirement ID	Description
SMBUS-1	The device shall support the NVMe Basic Management Command as defined in Appendix A of the NVMe Management Interface 1.0a specification. (http://www.nvmexpress.org/wp-content/uploads/NVM_Express_Management_Interface_1_0a_2017.04.08_-_gold.pdf). The primary purpose is for sideband access to temperature information for fan control. There's no requirement to implement anything else outside of Appendix A in the NVMe Management Interface specification.
SMBUS-2	Both SMBUS block read/write and byte read/write commands shall be supported.
SMBUS-3	The device shall implement the SMBus format as show in section 10.2.
SMBUS-4	Unless otherwise noted, the default value for the Firmware Update Flags field (Byte 91) in the SMBUS Data structure shall be set 0xFF.
SMBUS-5	The Secure Boot Failure Feature Reporting Supported bit at offset 243 shall be supported and set to 0x1.
SMBUS-6	When there is a secure boot failure it shall be reported with the following behavior: Command Code 0: Bit 6: Drive not ready = 1 Bit 5 Drive Functional = 0 Command Code 242: Bit 7 Secure Boot Failure Feature Reporting Supported = 1 Bit 6 Secure Boot Failure Status = 1
SMBUS-7	The device shall take no longer than the CAP.TO timeout value to produce stable SMBUS output through the NVMe Simple Management Interface protocol.

10.2 SMBUS Data Format

Command Code (Decimal)	Offset (Decimal)	Description
0	0	Defined in NVM Express Management Interface 1.0a.
8	8	Defined in NVM Express Management Interface 1.0a.

32	32	Length of GUID. This is the number of bytes until the PEC code. This shall be 16 decimal (0x10).
	48:33	GUID: This is a 16-byte Global Unique Identifier. The GUID shall be 738920e5-6bee-4258-9a7a-cebdb35f0085
	49	PEC: An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm is defined in SMBus Specifications.
50	50	Length of Telemetry: Indicated the number of additional bytes to read before encountering PEC. This value should always be 3**8** (0x2**6**) in implementations of this version of the spec.
	51	Temperature Flags: This field reports the effect of temperature on the device's performance. Temperature Throttling – Bit 7 is set to 1b when the drive is throttling performance to prevent overheating. **Clear to 0b when the device is not throttling.** Bits 6-0 shall be set to 1b.
	52	Max Power Supported: This shall denote the max average power supported by this device rounded to the nearest watt. Some examples of how to use this is a 50W device is 0x32, a 25W device is 0x19, a 15W device is 0xF, an 8.25W device is 8W which is 0x8.
	84:53	Configured Power state. This is the power level entry that is currently set based on Set Features.
	88:85	This is the device raw capacity in GB in Hex (2048 GB in raw capacity = 0x800). Does not include any extra spare blocks within the NAND.
	89	PEC: An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm in SMBus Specifications.
90	90	Length of Status of firmware Update Field: Indicates number of additional bytes to read before encountering PEC. This value should always be 4 (0x04) in implementations of this version of the spec.
	91	Firmware Update Flags: This field allows the host to control whether the current firmware allows new firmware images to be activated. Please see the "Security" section of this specification for more information.

<table>
<tr><td></td><td></td><td>Enable Firmware Update -- Bit 7
Written by host, read by drive<table><tr><td>1</td><td>Unlock Firmware Update
Drive shall enable Firmware update</td></tr><tr><td>0</td><td>Lock Firmware Update
Drive shall block and error on Firmware download and activate commands</td></tr></table>
Firmware Update Enabled -- Bit 6
Written by drive, read by host<table><tr><td>1</td><td>Firmware Update Unlocked
Drive shall allow Firmware download and activate commands</td></tr><tr><td>0</td><td>Firmware Update Locked
Drive shall block and error on Firmware download and activate commands</td></tr></table>
The default shall persist across power cycles. Bits 5-0 shall be set to ‘1’.</td></tr>
<tr><td></td><td>94:92</td><td>Reserved. Shall be set to 0x0.</td></tr>
<tr><td></td><td>95</td><td>PEC:
An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm in SMBus Specifications.</td></tr>
<tr><td>96</td><td>96</td><td>Length of Version: Indicates number of additional bytes to read before encountering PEC. This value should always be 56 (38h) in implementations of this version of the spec.</td></tr>
<tr><td></td><td>104:97</td><td>Firmware Version Number:
This field shall indicate the activated firmware version that is running on the device after the firmware activation took place. The format of this field shall be as defined in field Firmware Revision (FR) section 5.15.2.2 Identify Controller Data Structure of the NVMe specification.</td></tr>
<tr><td></td><td>112:105</td><td>Reserved:
Shall be set to 0x0.</td></tr>
<tr><td></td><td>152:113</td><td>Product Part/Model Number. The reason for 40 bytes is to keep this consistent with NVMe that already has this field in the identify command.</td></tr>
<tr><td></td><td>153</td><td>PEC:
An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm in SMBus Specifications.</td></tr>
<tr><td>154</td><td>241:154</td><td>Reserved:</td></tr>
</table>

		Shall be set to 0x0.
242	242	Length of Version: Indicates number of additional bytes to read before encountering PEC. This value should always be 4 (0x4).
	243	Bit 7: Secure Boot Failure Feature Reporting Supported When set to 0x1 the secure boot feature reporting is supported. When set to 0x0 the secure boot failure feature reporting is not supported. Bit 6: Secure Boot Failure Status: When set to 0b there is no secure boot failure. When set to 1b there is a secure boot failure. This bit shall only be set if the Secure Boot Feature Supported bit is set to 1b and there is a secure boot failure. Bit 5:0 Reserved. Shall be set to 0x0.
	246:244	Reserved: Shall be set to 0x0.
	247	PEC: An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm in SMBus Specifications.
248	248	Length of Specification Version: Indicates number of additional bytes to read before encountering PEC. This value should always be 6 (0x06) in implementations of this version of the spec.
	250:249	Log Page Version Number: Indicates the version of this mapping used in the device. Shall be set to '3' (0x03) after an SMBus block read is completed.
	254:251	Reserved: Shall be set to 0000h.
	255	PEC: An 8-bit CRC calculated over the slice address, command code, second slave address and returned data. Algorithm in SMBus Specifications.

11 Security

11.1 Basic Security Requirements

Requirement ID	Description
SEC-1	The device shall support signed firmware binary update which is checked before firmware is activated. The device firmware shall be

	authenticated using cryptographic keys on every reboot and during firmware update.
SEC-2	The device shall have XTS-AES-256 or AES-256 hardware-based data encryption or better is required.
SEC-3	The device shall have anti-rollback protection for firmware. The anti-rollback protection shall be implemented with a security version which is different than the firmware version. If the security version of the firmware being activated is greater or equal to the current security version the firmware may be activated. If the security version of the firmware being activated is not equal or greater than the firmware being activated the firmware update shall fail.
SEC-4	The device shall support Crypto Erase.
SEC-5	The device shall support Secure Boot.
SEC-6	The device shall have a method of identifying a secure boot failure which does not require physical access to the device.
SEC-7	The device shall be FIPS 140-3 capable (not required to get FIPS certificate) and shall follow the NIST 800-90 specification.
SEC-8	The device shall implement only FIPS and NIST approved implementations and algorithms.
SEC-9	The device shall support Key revocation allowing a new key to be used for firmware validation on update. Preferred implementation is to allow for up to 3 key revocations.
SEC-10	The device shall support Opal v2.01 with mandatory support for the Locking feature, the Opal SSC feature and the Datastore Table feature.
SEC-11	The device shall support Single User Mode feature set Version 1.00, revision 1.00.
SEC-12	The device shall support Configurable Namespace Locking (CNL) feature set Version 1.00, revision 1.00 with mandatory support for the Namespace Global Range Locking object. The Namespace Non-Global Range Locking object may be supported.
SEC-13	For some models, the IEEE 1667 silo will be required for eDrive support.
SEC-14	Supplier shall follow the Security Development Lifecycle (SDL), and provide a report with the following for each qualification-ready or production-ready firmware version: • The Threat Model • Fuzz & Pen Tests • Static Analysis • Build Logs and Compiler Security Settings

	Additional information about the SDL is available here: https://www.microsoft.com/en-us/sdl/default.aspx
SEC-15	Security audits, including firmware source code review, shall be required.
SEC-16	All signing keys shall be stored in a hardware security module (HSM).
SEC-17	Access/use of signing keys should be restricted to a small set of developers, following the principle of least privilege. Number of people with access and their corresponding roles shall be provided.
SEC-18	All debug ports shall be disabled before the device leaves the factory. Alternatively, the port may be accessible in the field only after a strong crypto authentication process.
SEC-19	All vendor unique commands, log pages or set features that are not explicitly defined in this specification shall be disabled before the device leaves the factory. Alternatively, the commands may be accessible in the field only after a strong crypto authentication process.
SEC-20	Adversarial testing using red teams shall be conducted before qualification start. A report of items attempted, and results shall be provided.
SEC-21	Vendor shall provide timely notification of security issues and delivery of fixes: • Vendor shall document all security fixes with each firmware update • Vendor shall notify end customer within 7 days of discovering security issues in the device hardware or firmware. • Notification of issues shall include the process and timeline of the vendor's commitment to fix the issue: ○ For privately disclosed vulnerabilities, the duration shall be no longer than 90 days ○ For publicly disclosed vulnerabilities, the duration shall be no longer than 7 days ○ Vendors shall notify the customers about the known CVEs and security issues and provide security-related updates before public announcement
SEC-22	All Telemetry and debugging logs shall be human readable.
SEC-23	The device shall not include user data, passwords, keys and any secret information in any Telemetry or debug logs.

11.2 Secure Boot

The device shall support Secure Boot. There are two fundamental things to address for secure boot:

1. Secure boot rooted in hardware
2. Core Root of Trust Measurement

The vendor should follow the recommendations in the *TCG Publication for Hardware Requirements for a Device Identifier Composition Engine*. DICE coupled with RIOT Core and Source for RIOT can help implement Cryptographic Identity with implicit attestation.

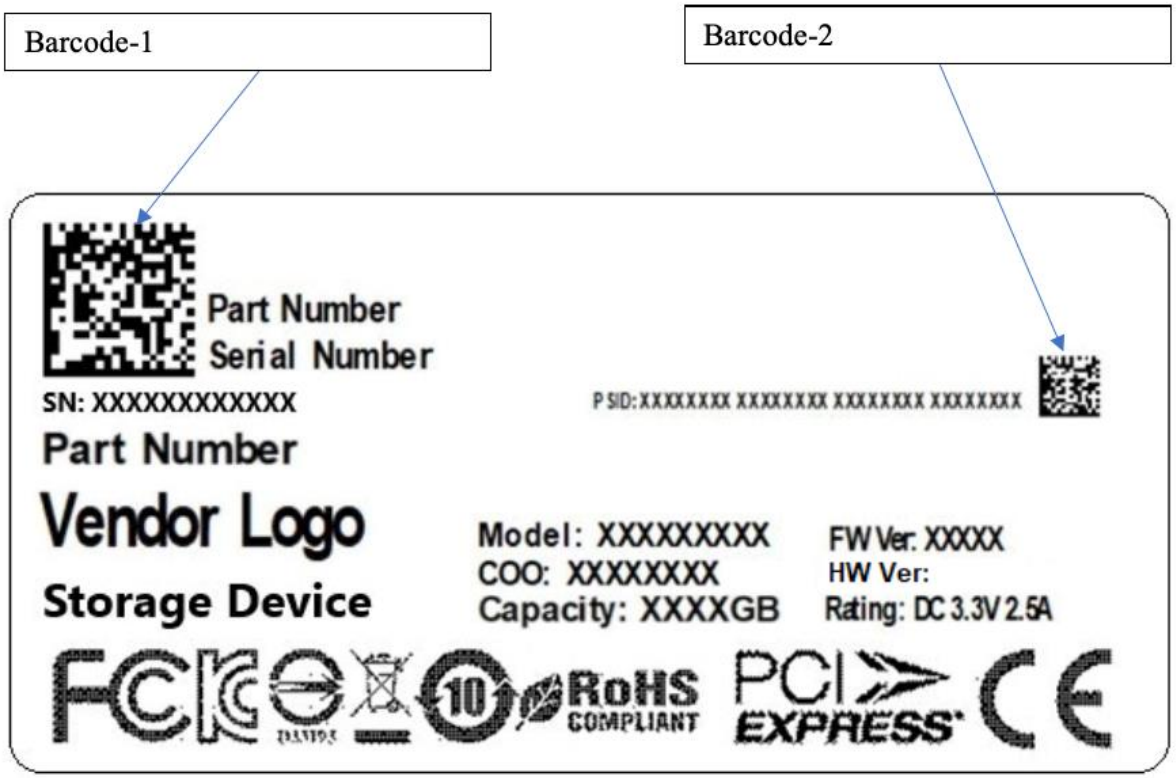
Requirement ID	Description
SBT-1	The device shall comply with the *FIPS 186-4* Digital Signature Standard (DSS) and the *Open Compute Security Project Publication for Secure Boot Requirements*.
SBT-2	For Core Root of Trust measurement, each device shall have a Cryptographic Device Identity.
SBT-3	The TCG DICE standard, or hardware based cryptographic identity shall be implemented.
SBT-4	The device should following the guidance in the Commercial National Security Algorithm Suite regarding quantum resistant algorithms and key sizes.
SBT-5	Secure boot flow shall be immutable for exploitation and use immutable public keys.

11.3 Data Encryption and Eradication

Requirement ID	Description
DATAE-1	The device shall support AES-256 encryption (or better), or NAND-level data eradication using the NVMe Format Feature.

12 Labeling

The following sample label is meant to be used to refer to the label requirements in section 11.1:



12.1 Label Requirements

Requirement ID	Description

<table>
<tr><td>LABL-1</td><td>The following fields are required information that shall be placed on the label:

<table>
<tr><th>Item</th><th>Format</th><th>Text Required</th><th>Barcode Required</th><th>Barcode Type</th></tr>
<tr><td>Barcode-1</td><td>‘Part Number’ ‘Underscore’ ‘Serial Number’ /n</td><td>No</td><td>Yes</td><td>2d</td></tr>
<tr><td>Part Number</td><td>Same as the MPN used for Ordering.</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Serial Number</td><td>Alpha-Numeric. 12-20 digits with first 4 digits indicating:
Date of Manufacturing in Work Week and Year
WWYY1234567890123456</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Model Name</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Capacity</td><td>Number of GB</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>STORAGE DEVICE</td><td>Text shall be “STORAGE DEVICE”</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>PSID</td><td>TCG-OPAL Spec</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Barcode-2</td><td>‘PSID’ /n</td><td>No</td><td>Yes</td><td>2d</td></tr>
<tr><td>HW Revision</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Firmware Name & Revision</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Regulatory Mark</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Country Certificatio n Numbers</td><td>If device has certain country certifications, they shall be displayed</td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>Certificatio n Logos</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
<tr><td>RoHS/ Green</td><td></td><td>Yes</td><td>No</td><td>N/A</td></tr>
</table>
</td></tr>
<tr><td>LABL-2</td><td>The Model Name on the shipping label shall match the model name used during qualification.</td></tr>
<tr><td>LABL-3</td><td>The minimum font size shall be 3 points and the typical size should be 6 points.</td></tr>
<tr><td>LABL-4</td><td>For the Capacity field, if there are space constraints, the manufacturer may remove “Capacity:” and just show “XXXXGB”.</td></tr>
<tr><td>LABL-5</td><td>To distinguish Part Number and Serial Number, the label shall have an underscore “_” between the Part Number and the Serial Number. Examples:</td></tr>
</table>

<table>
<tr><td></td><td>Part Number: SSD0001
Serial Number: abcdefghi

SSD0001_abcdefghi</td></tr>
<tr><td>LABL-6</td><td>There shall be a line with “STORAGE DEVICE”.</td></tr>
<tr><td>LABL-7</td><td>The following fields are optional information that can be placed on the label at the discretion of the device maker. Placement is also at the device makers discretion as long as such information does not interfere with the mandatory information above. No additional barcode shall be present.
<table>
<tr><th>Item</th><th>Format</th><th>Label Required</th><th>Barcode Required</th><th>Barcode Type</th></tr>
<tr><td>Processor Code (BA)</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Maker Logo</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Rated Voltage & Current</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Production Date</td><td>DDMMYYYY: DD(Date), MM(Month), YYYY(Year)</td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Weekly Code</td><td>YYWW: YY(Year), WW(Week)</td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Warranty VOID IF REMOVED</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Country of Origin</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Makers Own Label Material Number</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Website, Company Address</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>SSD</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>Product Series Name</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
<tr><td>SA: Value used within</td><td></td><td>Optional</td><td>No</td><td>N/A</td></tr>
</table></td></tr>
</table>

	manufacturing				
	PBA: Physical Board Address (identifies the physical configuration of the device)		Optional	No	N/A
	WWN: World Wide Number (unique for each device)		Optional	No	N/A
	US FCC Regulatory		Optional	No	N/A
LABL-8	To ensure that datacenter operations personnel can quickly and easily identify devices that have been ticketed for field replacement, it is mandatory to have the proper identifying fields on the label(s), in the format specified below.				
LABL-9	The label shall not degrade over the standard SSD lifetime under standard operating conditions.				
LABL-10	For each formfactor, the label shall be placed as specified below: • M.2: the label shall be placed on the top side of the device as defined in the PCI-Sig M.2 formfactor specification • E1.S: the label shall be placed on the Primary side of the device as defined in the SFF TA-1006 formfactor specification • E1.L: the label shall be placed on the either the Primary or Secondary side of the device as defined in the SFF TA-1007 formfactor specification				

13 Compliance

13.1 ROHS Compliance

Requirement ID	Description

ROHS-1	The Supplier shall provide component-level reporting on the use of listed materials by concentration (ppm) for all homogenous materials.

13.2 ESD Compliance

Requirement ID	Description
ESD-1	Device manufacturer needs to provide ESD immunity level (HBM-Human Body Model) measured in accordance withIEC-61000-4-2.

14 Shock and Vibration

Below are the shock and vibration specifications for storage devices:

Requirement ID	Description
SV-1	The non-operational shock requirement is 700G, half-sine, 0.5ms, total 6 shocks, along all three axes (+/-).
SV-2	The vibration requirement during operation is: 1.8G_{rms}, 5-500-5 Hz, Random Vibe, 20 min along all three axes.
SV-3	The vibration requirement during non-operation is: 3.13G_{rms}, 5-800-5 Hz, total 6 sweeps along all three axes, 20 minutes per sweep.
SV-4	Validation flow for Shock and Vibration: 1. UUT (Unit Under Test), test fixture should be visually inspected and ensured that everything is torqued or secured as needed. Pictures of test fixture with and w/o UUT should be provided. 2. Baseline performance of device should be gathered and used as a reference against post S&V data to ensure no performance impact incurred. 3. Once S&V testing is completed, repeat visual inspection to the UUT and test fixture to ensure no physical damage or performance impact has occurred to the UUT or test fixture. 4. Re-run stress test on the UUT in case of non-op test and provide data indicating no performance impact incurred to the unit

15 NVMe Linux CLI Plug-In Requirements

15.1 NVMe CLI Management Utility

The NVMeCLI utility (https://github.com/linux-nvme/nvme-cli) shall be used as the management utility for NVMe devices.

Requirement ID	Description
UTIL-1	The SSD supplier must test their SSDs with this utility and ensure compatibility. The following is the minimum list of commands that need to be tested with NVMeCLI: • Format • Secure erase • FW update • Controller reset to load FW • Health status • Log page reads including vendor log pages • SMART status • List devices • Get/set features • Namespace management • Identify controller and namespace • Effects log page

15.2 NVMe CLI Plugin Requirements

The device supplier shall develop and provide a Linux NVMe CLI plugin that meets the following requirements:

Requirement ID	Description
UTIL-PI-1	A single, common plugin for all of the supplier's NVMe-based products
UTIL-PI-2	Vendor and additional log page decoding including into a human readable format and JSON output
UTIL-PI-3	Access to OEM commands
UTIL-PI-4	The ability to pull crash dumps or FW logs (binary output is acceptable)
UTIL-PI-5	The plugin's subcommand nomenclature must adhere to section 15.2.1 and cannot change across versions.
UTIL-PI-6	The plugin shall use the existing NVMe CLI interface to access any vendor unique commands that are supported by the device.
UTIL-PI-7	If the NVMe CLI interface needs to transfer greater than 16 MB of data, the NVMe vendor unique Command shall have the ability to do multiple scatter/gather elements on the data buffer.

15.2.1 NVMe CLI Plug-In Nomenclature/Functional Requirements

The NVMe CLI plugin must meets the following naming and functional requirements:

Requirement ID	NVMe CLI Nomenclature	Purpose
UTIL-NM-1	vs-smart-add-log	Retrieve extended SMART cloud Information from section 4.8.4. The SMART cloud attributes must use the exact same attribute name as indicated in that section.
UTIL-NM-2	vs-internal-log	Retrieves internal drive telemetry/debug logging.
UTIL-NM-3	vs-fw-activate-history	Outputs the firmware activation history log page (0xC2) in table format. See section 14.2.2.1for the table output format.
UTIL-NM-4	vs-drive-info	Outputs the following information: 1. **Drive_HW_revision** – Displays the current HW rev of the drive. Any BOM or HW change must increment this version number. The value starts at 0 for pre-MP units and starts at 1.0 for MP units. The value increments by 0.1 for any HW changes in the pre-MP or MP stage. Qualification samples sent to Customer ODMs at the beginning of qualification is considered MP stage and needs to start at 1.0. 2. **FTL_unit_size** – Display FTL unit size. Units are in KB, so "4" means the FTL unit size is 4KB.
UTIL-NM-5	clear-pcie-correctable-errors	Vendor Unique Set feature that clears the correctable PCIe error counter. See section 3.11.6 for more details.
UTIL-NM-6	clear-fw-activate-history	Vendor Unique Set feature that clears the output of the "vs-fw-activate-history" and also the Firmware activation History Log page (0xC2). See section 3.11.3 for more details.
UTIL-NM-7	log-page-directory	The NVMe command that lists all the log pages and a description of their contents
UTIL-NM-8	cloud-SSD-plugin-version	Prints version "1.0"
UTIL-NM-9	Help	Display this help

15.2.2 NVMe CLI Plug-In FW Activation History Requirements

Requirement ID	Description
UTIL-FWHST-1	A table with entries that indicate the history of Firmware activation on the device
UTIL-FWHST-2	Using the plugin command in UTIL-NM-4 will retrieve the table
UTIL-FWHST-3	Lists the last twenty firmware that were activated (not downloaded) on the drive. Oldest entries are on top.
UTIL-FWHST-4	When the drive is first shipped from the factory, there are no entries recorded.
UTIL-FWHST-5	An entry must be recorded whenever a FW activation is taking place and does not matter if there's a reset or not. FW downloads do not generate an entry.
UTIL-FWHST-6	Redundant activation events shall not generate a new entry to prevent the scrolling out of useful information. An entry is considered to be redundant if they meet ALL the criteria below: 8. POH is within 1 minute from the last RECORDED entry 9. Power cycle count is the same 10. Current firmware is the same 11. New FW activated is the same 12. Slot number is the same 13. Commit Action Type is the same 14. Results are the same
UTIL-FWHST-7	Firmware Activation History's output column headers shall follow the requirements below.

Requirement ID	Firmware Activation History Column Header	Purpose
UTIL- **FWHST** -8	Firmware Activation Counter	Increments every time a firmware activation is attempted no matter if the result is good or bad. When the drive is shipped from manufacturing, this value is '0x0'.
UTIL- FWHST -9	Power on Hour	Displays the POH of the SSD when the firmware activation happened. Accuracy needs to be down to the second.
UTIL- FWHST -10	Power Cycle Count	Display the power cycle count that the firmware activation occurred.

UTIL- FWHST -11	Current Firmware	Displays the firmware currently running on the SSD before the firmware activation took place
UTIL- FWHST -12	New Firmware Activated	Displays the activated firmware version that is running on the SSD after the firmware activation took place.
UTIL- FWHST -13	Slot Number	Displays the slot that the firmware is being activated from.
UTIL- FWHST -14	Commit Action Type	Displays the Commit action type associated with the firmware activation event.
UTIL- FWHST -15	Result	Records the results of the firmware activation event. A passing event shall state a "Pass" for the result. A failing event shall state a "Failed" + the error code associated with the failure.

15.2.2.1 NVMe CLI Plug-In FW Activation History Example Outputs

FW Activation Examples:

Host FW download and activation events and initial states:
Initial State: Slot1=101
POH 1:00:00, PC 1, FW Commit CA=011b Slot=1 FW=102
POH 2:00:00, PC 1, FW Commit CA=001b Slot=1 FW=103
POH 3:00:00, PC 1, FW Commit CA=001b Slot=1 FW=104
POH 4:00:00, PC 1, FW Commit CA=001b Slot=1 FW=105
Reset
POH 5:00:00, PC 1, FW Commit CA=011b Slot=1 FW=106
POH 6:00:00, PC 1, FW Commit CA=001b Slot=1 FW=107
Power Cycle
POH 7:00:00, PC 2, FW Commit CA=001b Slot=1 FW=108

NVMe-CLI Plugin Output:

Firmware Activation Counter	Power on Hour	Power cycle count	Current firmware	New FW activated	Slot number	Commit Action Type	Result
1	1:00:00	1	101	102	1	011b	pass

2	4:00:00	1	102	105	1	001b	pass
3	5:00:00	1	105	106	1	011b	pass
4	7:00:00	2	106	107	1	001b	pass

Repeated Activation Events examples:

Host FW download and activation events and initial states:

Initial State: Slot1=101
POH 1:00:01, PC 1, FW Commit CA=011b Slot=1 FW=102, pass
POH 1:00:10, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1
POH 1:00:30, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1 (not recorded)
POH 1:01:15, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1 (recorded as the time difference is greater than 1 minute from the last recorded event)
POH 1:01:25, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #2 (recorded as the failure reason changed)

NVMe-CLI Plugin Output:

Firmware Activation Counter	Power on Hour	Power cycle count	Current firmware	New FW activated	Slot number	Commit Action Type	Result
1	1:00:01	1	101	102	1	011b	pass
2	1:00:10	1	102	102	1	011b	Fail #1
3	1:01:15	1	102	102	1	011b	Fail #1
4	1:01:25	1	102	102	1	011b	Fail #2

Appendix A – Facebook Specific Items

The following items apply specifically to devices delivered to Facebook.

1 Configuration Specifics

Requirement ID	Description
FB-CONF-1	Devices shall be formatted to 4096-byte sectors from the factory.
FB-CONF-2	IEEE 1667 shall not be supported. Devices shall not support Set Features for IEEE1667 Silo Set Feature Identifier (0xC4).
FB-CONF-3	SMBUS byte 91 bit 6, “Firmware Update Unlocked”, bit shall be set to 0x1 by default from the factory.
FB-CONF-4	Devices shall not support Set Features (0xC0) Error Injection.

2 Performance Requirements

The following numbers are the Facebook performance targets for data storage SSD across all form factors. They are provided to serve as a guidance for SSD Vendors. Items related to the items below may be available on GitHub at https://github.com/facebook

The targets are broken down into the following segments:

Requirement ID	Description
FB_PERF-1	FB-FIO Synth Flash Targets (for all capacities)
FB-PERF-2	fb-FIOSynthFlash TRIM Rate targets
FB-PERF-3	IO.go benchmark target
FB-PERF-4	Fileappend benchmark target
FB-PERF-5	Sequential write bandwidth
FB-PERF-6	Cache bench target
FB-PERF-7	All targets shall be achieved by using “kyber” as the I/O scheduler.
FB-PERF-8	All targets shall be achieved with the SSD max average power consumption not exceeding 10W based on the power methodology described in PCM-1, PCM-2 and PCM-3.

1a. Performance Targets for FB-FIO Synth Flash - HE_Flash_Short_TRIM_2H19 (for all capacities)_

Workload	Read MiB/s per TB	Write MiB/s per TB	TRIM BW per TB	P99 Read Latency	P99.99 Read Latency	P99.9999 Read Latency	P99.99 Write Latency	P99.9999 Write Latency
4K_L2R6DWPD_wTRIM	68 MiB/s	72 MiB/s	117 MiB/s	2,000 us	5,000 us	8,500 us	15,000 us	25,000 us
4K_L2R9DWPD_wTRIM	68 MiB/s	93 MiB/s	156 MiB/s	2,200 us	5,500 us	9,500 us	15,000 us	25,000 us
MyRocks_Heavy_wTRIM	120 MiB/s	101 MiB/s	22 MiB/s	2,000 us	5,000 us	8,500 us	10,000 us	15,000 us
Fleaf	320MiB/s	87 MiB/s	89 MiB/s	3,000 us	6,000 us	10,000 us	20,000 us	25,000 us

Performance Targets for FB-FIO Synth Flash – Cache

Workload	Read MiB/s per TB	Write MiB/s per TB	TRIM BW per TB	P99 Read Latency	P99.99 Read Latency	P99.9999 Read Latency	P99.99 Write Latency	P99.9999 Write Latency
B_Cache	164 MiB/s	96 MiB/s	0 MiB/s	2,000us	5,500 us	15,000 us	20,000 us	25,000 us

Performance Targets for FB-FIO Synth Flash – Search_2H19

Workload	Read MiB per Node	Write MiB/s per Node	TRIM MiB/s per Node	P99 Read Latency	P99.99 Read Latency	P99.9999 Read Latency	P99.99 Write Latency	P99.9999 Write Latency
SearchLM_wTRIM	2,550 MiB/s	12 MiB/s	130 MiB/s	1,500 us	10,000 us	15,000 us	20,000 us	25,000 us

1b. Trim Rate Targets

- This test measures raw trim performance which no background I/O
- 64M trim >= 50GiB/s & <= 10ms P99 trim latency
- 3GB trim >= 500GiB/s & <= 10ms P99 trim latency

1c. IO.go Benchmark Targets

- This test measures how long the file system is blocked from writing/overwriting a file while a different file is deleted
- Less than 4 file sizes total with latency outliers > 10ms
- No more than 2 latency outliers per file size
- No single latency outlier above 15ms

1d. Fileappend Benchmark Targets

- This test measures how long the file system is blocked from appending to a file while a different file is deleted.
- No measurable stalls reported by this tool
- Max acceptable latency outlier is 10ms when deleting 1GiB or 2GiB file

1e. Sequential Write Bandwidth

- Full drive (all available user capacity, all namespaces) must be written/filled in 180 minutes or less
- Simple single-threaded sequential write FIO script to fill drive

1f. Cache bench target

- A benchmarking tool that's a supplement for FB FIO Synth Flash tool on measuring performance for cache applications. This is different than the "B Cache" workload in FB FIO Synth Flash.
- Two workloads need to be tested:
  - Tao Leader
  - Memcache
- The final allocator and throughput stats from the benchmark will be used to see if the targets are met.
- Vendor NVMe CLI plug-in with "physical NAND bytes written" metric in the SMART Cloud Health Log (0xC0) needs to be working to get the write amplification.

Workload	Get Rate	Set Rate	Read Latency P99 (us)	Read Latency P99.99 (us)	Read Latency P100 (us)	Write Latency P99.99 (us)	Write Latency P100 (us)	Write Amp
Tao Leader	87,000	16,000	400	3,000	12,000	700	8,000	1.3
Memcache WC	3,200	1,500	1,700	14,000	15,000	7,000	8,000	1.4

Appendix B – Microsoft Specific Items

The following items apply specifically to Microsoft.

1 Configuration Specifics

Requirement ID	Description
MS-CONF-1	E1.S and M.2 devices shall be formatted to 512-byte sectors from the factory
MS-CONF-2	E1.L devices shall be formatted to 4096-byte sectors from the factory.
MS-CONF-3	IEEE 1667 shall be supported.

MS-CONF-4	For E1.S and M.2 SMBUS byte 91 bit 6, “Firmware Update Unlocked”, bit shall be set to 0x1 (Firmware Update is Enabled) by default from the factory.
MS-CONF-5	For E1.L SMBUS byte 91 bit 6, “Firmware Update Unlocked”, bit shall be set to 0x0 (Firmware Update is Disabled) by default from the factory.

2 Performance Requirements

Requirement ID	Description
MS-PERF-1	The device shall meet the performance targets with these assumptions: • Entropy of all workloads is 100% (uncompressible) • Active range is 100% • Maximum power draw as specified • Operations are “naturally aligned,” meaning they are aligned to IO-sized boundaries
MS-PERF-2	The vendor shall provide a performance test report using the preconditioning methodology to achieve steady state performance for the given device, e.g. SNIA Solid State Storage Performance Test Specification (PTSE).
MS-PERF-3	As the devices will support Power Loss Protection (PLP), the performance shall not be degraded by the following: • PLP - backed cache(s) shall enable fast performance. • FUA – forced unit access shall not incur a performance penalty. • Flush Cache – flush cache shall be ignored but acknowledged. • SET FEATURE write-cache disable - command shall be ignored but acknowledged.
MS-PERF-4	Random read latency shall match or beat the distribution listed in each form factor performance requirements below under the following test conditions: • Queue Depth = 1 • Device is near End-of-Life, with or without utilizing “short stroked” firmware • Starting state: Device is trimmed then written sequentially with 1MiB accesses • 256KiB sequential writes with the rate adjusted so that 10% of the volume is writes, or

	256KiB random writes with the rate adjusted so that 5% of the volume is writes (whichever is worse) 4KiB, 8KiB, or 64KiB random reads with the rate adjusted so that 90% of the volume is reads

2.1 M.2 Performance Requirements

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>MS-PFM2-1</td><td>When the device is set to an RMS average of 8.25W it shall meet or exceed the minimum throughput numbers for the following workloads (up to the limit of the PCIe bus):

<table>
<tr><th>Metric</th><th>Workload</th><th>Minimum Throughput</th></tr>
<tr><td>Sequential Read (MiB/s)</td><td>128kiB, QD = 128, 0% writes</td><td>400/TiB</td></tr>
<tr><td>Sequential Write (MiB/s)</td><td>128kiB, QD = 128, 100% writes</td><td>200/TiB</td></tr>
<tr><td>Random Read (IOPS)</td><td>4kiB, QD = 2/TiB, 0% writes</td><td>10k/TiB</td></tr>
<tr><td>Random Write (IOPS)</td><td>4kiB, QD = 128, 100% write commands</td><td>8k/TiB</td></tr>
<tr><td>Random Mix (IOPS)</td><td>4kiB, QD = 128, 30% write commands</td><td>50k/TiB</td></tr>
</table>
</td></tr>
<tr><td>MS-PFM2-2</td><td>Device shall not exceed the following latency requirements:

<table>
<tr><th></th><th></th><th>4KiB (µs)</th><th>8KiB (µs)</th><th>64KiB (µs)</th><th>Operations Needed in Test*</th></tr>
<tr><td>Average</td><td></td><td>240</td><td>250</td><td>450</td><td>--</td></tr>
<tr><td>99 %</td><td>(2 nines)</td><td>300</td><td>360</td><td>770</td><td>>100</td></tr>
<tr><td>99.9 %</td><td>(3 nines)</td><td>500</td><td>650</td><td>1,200</td><td>>1,000</td></tr>
<tr><td>99.99 %</td><td>(4 nines)</td><td>900</td><td>1,000</td><td>2,000</td><td>>10,000</td></tr>
<tr><td>99.999 %</td><td>(5 nines)</td><td>1,200</td><td>2,000</td><td>3,500</td><td>>1e5</td></tr>
<tr><td>Maximum Timeout</td><td></td><td>8 seconds</td><td>8 seconds</td><td>8 seconds</td><td>--</td></tr>
</table>

*The test shall apply the minimum number of operations listed in the right-most column.</td></tr>
</table>

2.2 E1.S Performance Requirements

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>MS-PFE1S-1</td><td>When the device is set to an RMS average of 20W it shall meet or exceed the minimum throughput numbers for the following workloads (up to the limit of the PCIe bus):
<table>
<tr><th>Metric</th><th>Workload</th><th>Minimum Throughput</th></tr>
<tr><td>Sequential Read (MiB/s)</td><td>128kiB, QD = 128, 0% writes</td><td>200/TiB</td></tr>
<tr><td>Sequential Write (MiB/s)</td><td>128kiB, QD = 128, 100% writes</td><td>100/TiB</td></tr>
<tr><td>Random Read (IOPS)</td><td>4kiB, QD = 2/TiB, 0% writes</td><td>10k/TiB</td></tr>
<tr><td>Random Write (IOPS)</td><td>4kiB, QD = 128, 100% write volume</td><td>5k/TiB</td></tr>
<tr><td>Random Mix (IOPS)</td><td>4kiB, QD = 128, 10% write volume</td><td>7.5k/TiB</td></tr>
</table></td></tr>
<tr><td>PFE1S-2</td><td>Device shall not exceed the following latency requirements:
<table>
<tr><th></th><th></th><th>4KiB (µs)</th><th>8KiB (µs)</th><th>64KiB (µs)</th><th>Operations Needed in Test*</th></tr>
<tr><td>Average</td><td></td><td>240</td><td>250</td><td>450</td><td>--</td></tr>
<tr><td>99 %</td><td>(2 nines)</td><td>300</td><td>360</td><td>770</td><td>>100</td></tr>
<tr><td>99.9 %</td><td>(3 nines)</td><td>500</td><td>650</td><td>1,200</td><td>>1,000</td></tr>
<tr><td>99.99 %</td><td>(4 nines)</td><td>900</td><td>1,000</td><td>2,000</td><td>>10,000</td></tr>
<tr><td>99.999 %</td><td>(5 nines)</td><td>1,200</td><td>2,000</td><td>3,500</td><td>>1e5</td></tr>
<tr><td>Maximum Timeout</td><td></td><td>8 seconds</td><td>8 seconds</td><td>8 seconds</td><td>--</td></tr>
</table>
*The test shall apply the minimum number of operations listed in the right-most column.</td></tr>
</table>

2.3 E1.L Performance Requirements

Requirement ID	Description

<table>
<tr><td>MS-PFE1L-1</td><td>Bandwidth and Throughput for use in xDirect: When the device is set to an RMS average of 20W it shall meet or exceed the minimum throughput numbers for the following workloads (up to the limit of the PCIe bus):

<table>
<tr><th>Metric</th><th>Workload</th><th>Minimum Throughput</th></tr>
<tr><td>Sequential Read (MiB/s)</td><td>128kiB, QD = 128, 0% writes</td><td>200/TiB</td></tr>
<tr><td>Sequential Write (MiB/s)</td><td>128kiB, QD = 128, 100% writes</td><td>100/TiB</td></tr>
<tr><td>Random Read (IOPS)</td><td>4kiB, QD = 2/TiB, 0% writes</td><td>10k/TiB</td></tr>
<tr><td>Random Write (IOPS)</td><td>4kiB, QD = 128, 100% write volume</td><td>5k/TiB</td></tr>
<tr><td>Random Mix (IOPS)</td><td>4kiB, QD = 128, 10% write volume</td><td>7.5k/TiB</td></tr>
</table></td></tr>
<tr><td>MS-PFE1L-2</td><td>Bandwidth and Throughput for use in XIO: When the device is set to a max RMS of 20W it shall meet or exceed the minimum throughput numbers for the following workloads (up to the limit of the PCIe bus):

<table>
<tr><th>Metric</th><th>Workload</th><th>Minimum Throughput</th></tr>
<tr><td>Sequential Read (MiB/s)</td><td>128kiB, QD = 128, 0% writes</td><td>200/TiB</td></tr>
<tr><td>Sequential Write (MiB/s)</td><td>128kiB, QD = 128, 100% writes</td><td>50/TiB</td></tr>
<tr><td>Random Read (IOPS)</td><td>4kiB, QD = 2/TiB, 0% writes</td><td>6k/TiB</td></tr>
<tr><td>Random Write (IOPS)</td><td>16kiB, QD = 128, 100% write volume</td><td>5k/TiB</td></tr>
<tr><td>Random Mix (IOPS)</td><td>4kiB random read, 16kiB sequential write, QD = 128, 10% write volume</td><td>5k/TiB</td></tr>
</table></td></tr>
<tr><td>PFE1L-3</td><td>Device shall not exceed the following latency requirements:</td></tr>
</table>

<table>
<tr><td></td><td>
<table>
<tr><th></th><th></th><th>4KiB (µs)</th><th>8KiB (µs)</th><th>64KiB (µs)</th><th>Operations Needed in Test*</th></tr>
<tr><td>Average</td><td></td><td>240</td><td>250</td><td>450</td><td>--</td></tr>
<tr><td>99 %</td><td>(2 nines)</td><td>300</td><td>360</td><td>770</td><td>>100</td></tr>
<tr><td>99.9 %</td><td>(3 nines)</td><td>500</td><td>650</td><td>1,200</td><td>>1,000</td></tr>
<tr><td>99.99 %</td><td>(4 nines)</td><td>900</td><td>1,000</td><td>2,000</td><td>>10,000</td></tr>
<tr><td>99.999 %</td><td>(5 nines)</td><td>1,200</td><td>2,000</td><td>3,500</td><td>>1e5</td></tr>
<tr><td>Maximum Timeout</td><td></td><td>8 seconds</td><td>8 seconds</td><td>8 seconds</td><td>--</td></tr>
</table>
*The test shall apply the minimum number of operations listed in the right-most column.
</td></tr>
<tr><td>PFE1L-4</td><td>No more than 5 IOs in one hour shall take more than 2 seconds to complete.</td></tr>
</table>