facebook

NVMe PCIe SSD Specification

Rev 0.82

Ta-Yu Wu, Ross Stenfort

Table of Contents

1 Overview

This Facebook specification provides a set of requirements targeting drives arriving in June 2019 for all PCIe-based SSDs that we will use in our infrastructure. This specification will be maintained continuously and periodically released to all SSD suppliers.

2 Scope

This document covers all PCIe-attached SSDs using the NVMe protocol including all form factors (add-in card, U.2, M.2, or any future designs) with the exception of boot/client SSDs. This document's requirements will apply to all Facebook server and storage systems. This document does NOT cover endurance or capacity requirements. Those will be addressed separately on an as needed basis.

3 Reliability

3.1 UBER

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

3.2 End to End Data Protection

Requirement ID	Description
REL-2A	All user data must be protected using overlapping ECC and CRC protection mechanisms throughout the entire read and write paths in the SSDs including all storage elements (registers, caches, SRAM, DRAM, NAND, etc.).
REL-2B	At least one bit of correction and 2 bits of detection is required.
REL-2C	The entire DRAM addressable space needs to be protected with at least one bit correction and 2 bits of detection scheme. This includes but not limited to the following: • Flash translation layer (FTL) • Mapping tables • Journal entries • Firmware scratch pad • System variables • Firmware code

REL-2D	Silent data corruption will not be tolerated and will result in immediate disqualification.
REL-2E	The SSD 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.
REL-2F	All SSD metadata, FW, FW variables, and other SSD system data should be protected by at least a single bit detection scheme.

3.3 Drive Recovery Behavior

Requirement ID	Description
REL-3A	SSD shall attempt to recover automatically due to any bit flips and continue normal operation. The SSD shall only crash (or assert) as a last resort.

3.4 Behavior on FW crash (assert)

Requirement ID	Description
REL-4A	FW shall not allow host read or write access to media
REL-4B	The SSD shall still support the ability to read any failure logs from the SSD to determine the nature of the failure. If not, a power cycle or reset shall recover the SSD enough to support the following functions: secure erase, error and SMART log collection, and internal drive log or crash dump collection.
REL-4C	All drive error logs shall be committed to non-volatile memory.

3.5 On time

Requirement ID	Description
REL-5A	The SSD will be powered up 100% of the time for the duration of its life.

3.6 Operational life

Requirement ID	Description
REL-6A	The warranty and design shall support a 5 year operational life.

3.7 AFR (Annual Failure Rate)

Requirement ID	Description
REL-7A	The SSD shall meet an MTBF of 2 million hours (AFR of <= 0.45%) under Facebook environmental conditions (70C drive composite temperature and up to 90% RH) throughout the life of the device.
REL-7B	Supplier must provide the temperature and humidity conditions used to determine the MTBF.

4 NVMe requirements

4.1 Overview

Requirement ID	Description
NVMe-1A	The SSD shall comply to the NVMe 1.3c spec unless otherwise specified.
NVMe-1B	A NVMe compliance report shall be provided.

4.2 Driver

Requirement ID	Description
NVMe-2A	The SSD shall support the upstream open-source Linux driver. Any OEM specific features, functionality, or fixes must be upstreamed.

4.3 NVMe Controller Configuration and Behavior

Requirement ID	Description
NVMe-3A	Weighted round-robin shall not be the default arbitration mechanism, Round-Robin shall be the Default Mechanism.
NVMe-3B	The SSD shall support a Maximum Data Transfer Size (MDTS) value of at least 256KB
NVMe-3C	Drive is expected to service I/O and ADMIN commands as soon as CSTS.RDY=1
NVMe-3D	The SSD Controller shall keep CSTS.RDY = 0 until the device comes ready internally and is able to service commands.
NVMe-3E	The Shutdown Notification completion (CSTS.SHST) shall be received within 5s of setting SHN bit, if RTD3 entry latency is not supported

NVMe-3F	The SSD Controller shall support the CC.SHN (Normal and Abrupt Shutdown Notifications) at a minimum.
NVMe-3G	Data Loss is not tolerated if SHN was completed by the Controller and it is expected to become ready without entering any recovery mode.
NVMe-3H	Shutdown Notification shall trigger flushing of all content within SSD's internal (SRAM/ DRAM) cache (*if one is present)*
NVMe-3I	The SSD Firmware shall support reporting of CSTS.CFS as indicated in the NVMe Spec.
NVMe-3J	The "Model Number" field in the Identify Controller Data Structure (CNS 01h, byte offset 24:63) must be identical to the Model Part Number (MPN) in the product datasheet provided to Facebook.

4.4 NVMe Reset Supported

Requirement ID	Description
NVMe-4A	NVMe Subsystem reset
NVMe-4B	NVMe controller reset

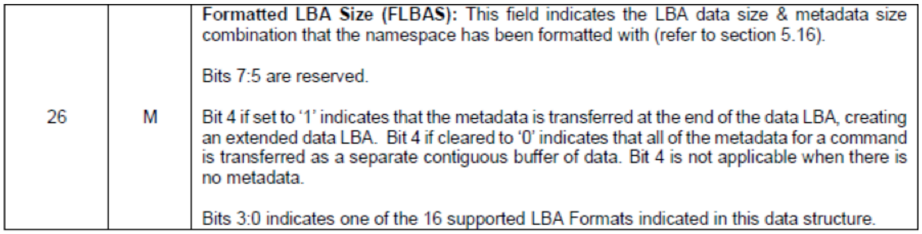
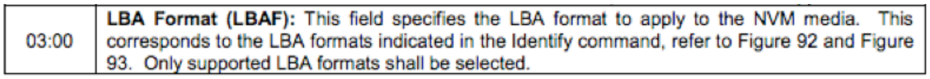
4.5 NVMe Admin Command Set

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

Requirement ID	Description
NVMe-5A	The SSD shall support all mandatory NVMe admin commands as specified by the NVMe version called out in Section 4.1 Overview.
NVMe-5B	Identify – In addition to supporting all the mandatory CNS values and the associated mandatory fields with the CNS, the following optional fields in the CNS must be supported: • Format progress indicator (FPI) • IO Performance and Endurance Hints (TP4025)
NVMe-5C	Firmware Commit – The following Commit Action (CA) shall be supported: • 000b – Download only • 001b – Download and activate upon reset • 010b – Activate upon reset • 011b – Activate immediately without reset
NVMe-5D	Firmware Image Download
NVMe-5E	Namespace Management
NVMe-5F	Namespace Attachment
NVMe-5G	Format NVM
NVMe-5H	Support for NVMe-MI Send and Receive is not required.

4.5.1 Namespace Management/Attachment commands

Facebook will use the namespace management command along with the attach/detach commands to increase SSD over-provisioning beyond the default minimum over-provisioning.

Requirement ID	Description
NVMe-6A	The namespace management commands shall be supported on all namespaces.
NVMe-6B	When creating a namespace, the default “Formatted LBA Size” parameter (FLBAS=0) shall correspond to a 4K sector size.  26 M **Formatted LBA Size (FLBAS):** This field indicates the LBA data size & metadata size combination that the namespace has been formatted with (refer to section 5.16). Bits 7:5 are reserved. Bit 4 if set to ‘1’ indicates that the metadata is transferred at the end of the data LBA, creating an extended data LBA. Bit 4 if cleared to ‘0’ indicates that all of the metadata for a command is transferred as a separate contiguous buffer of data. Bit 4 is not applicable when there is no metadata. Bits 3:0 indicates one of the 16 supported LBA Formats indicated in this data structure.
NVMe-6C	When formatting the drive with the Format command, the default “LBA Format” parameter (LBAF=0) shall correspond to a 4K sector size.  03:00 **LBA Format (LBAF):** This field specifies the LBA format to apply to the NVM media. This corresponds to the LBA formats indicated in the Identify command, refer to Figure 92 and Figure 93. Only supported LBA formats shall be selected.

4.5.2 Namespace Utilization (NUSE)

Requirement ID	Description
NVMe-7A	The NUSE must 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. 1. After a physical secure erase (SES=1), NUSE would be zero. And the usage data in “nvme list” would reflect that. 0.00 GB / 200.00 GB 2. After writing 1 GB worth of data, the usage data would show the following: 1.00 GB / 200.00 GB 3. After filling the drive, the usage data would show the following: 200.00 GB / 200.00 GB

	4. If the host issues a 10GB de-allocate command, the usage data would show the following: 190.00 GB / 200.00 GB

4.6 NVMe I/O Command Set

Requirement ID	Description
NVMe-8A	The SSD shall support all mandatory NVMe I/O commands as specified by the NVMe version called out in Section 4.1 Overview.
NVMe-8B	The SSD shall support the following optional NVMe I/O commands: 1. Dataset Management (De-allocate)

4.7 Optional NVMe feature support

The SSD shall also support the following NVMe features:

Requirement ID	Description
NVMe-9A	Telemetry
NVMe-9B	Timestamp
NVMe-9F	IO Performance and Endurance Hints (TP4025)

4.7.1 Telemetry Logging and Interface for Failure Analysis

The goal is to improve the ability to debug the SSD through event logging and utilize the NVMe telemetry host-initiated (07h) and controller-initiated (08h) log pages through NVMe CLI as the standard interface for log retrieval.

Requirement ID	Description
NVMe-12A	The FW must track the drive’s operational/event history and any critical parameters that can be used to debug issues.
NVMe-12B	All assert events and controller-initiated log captures will require an associated vendor-specific “Reason Identifier” that uniquely identifies the assert /controller condition.
NVMe-12C	The vendor must provide a table that categorizes the reason identifiers.
NVMe-12D	If any of the following list of conditions occur, the telemetry data must be committed to non-volatile storage so that the data is saved: 1. Ungraceful/graceful power cycle 2. Reboot 3. Any time the SMART critical warning changes to a non-zero value 4. Any type of FW assert

<table>
<tr><td></td><td>5. Retrieval of log via the host interface
6. The drive switches to a degraded mode during run-time
7. The SMART “SSD End to end correction” count is incremented</td></tr>
<tr><td>NVMe-12E</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 Ios</th></tr>
<tr><td><b>1</b></td><td>Periodic logging for monitoring trends/problems</td><td>Vendor-specific</td><td>&lt; 10ms max</td></tr>
</table>
</td></tr>
</table>

4.7.1.1 Telemetry CLI Plug-in specifications

Requirement ID	Description
NVMe-13A	The host needs to be able to retrieve the telemetry data via the ”vs-internal-log” NVMe plug-in sub-command which is mapped to the Telemetry host-initiated (0x07) and Telemetry controller-initiated (0x08) log pages.
NVMe-13B	The output can be packaged into a binary file and the reason identifier in “vs-error-reason-identifier” needs to be set.
NVMe-13C	The host also needs to be able to disable/enable the Telemetry controller-initiated (0x08) log page via the “vs-telemetry-controller-option” NVMe plug-in sub-command.
NVMe-13D	The default status is “DISABLED” for the controller-initiated log page.
NVMe-13E	Comply to Appendix C: NVMe-CLI Plugin Output for the required syntax on the following NVMe-CLI plugin subcommands: • vs-internal-log • vs-telemetry-controller-option

4.8 Additional features/commands

The SSD shall also support the following additional vendor unique commands (VUC) or data structures:

Requirement ID	Description
NVMe-14A	Clear PCIe correctable error counter 1. Needs to be part of the NVMe set/get feature admin command 2. Feature ID can be decided by the vendor

	3. Reset bit setting: 0 = NOP, 1 = Reset PCIe correctable counter in log page 0xCA 4. To execute the VUC, you just run the specified NVMe CLI plug in command "clear-pcie-correctable-errors"
NVMe-14B	Clear firmware activation history 1. Needs to be part of the NVMe set/get feature admin command 2. Feature ID can be decided by the vendor 3. Reset bit setting: 0 = NOP, 1 = Clear firmware activation history 4. To execute the VUC, you just run the specified NVMe CLI plug in command "clear-fw-activate-history".

4.9 Log page requirements

4.9.1 Mandatory Log page requirements

The SSD shall support the following mandatory log pages as defined in the NVMe specification version in Section 4.1:

Requirement ID	Description
NVMe-15A	Error Information (Log Identifier 01h)
NVMe-15B	SMART/Health Information (Log Identifier 02h)
NVMe-15C	Firmware Slot Information (Log Identifier 03h)
NVMe-15D	The host shall not be able to reset the "Percentage Used" field in the SMART/Health Information (Log Identifier 02h).

4.9.2 Optional NVMe Log pages

The SSD shall support the following additional log pages:

Requirement ID	Description
NVMe-17A	Telemetry Host-Initiated (Log Page 0x07)
NVMe-17B	Telemetry Controller-Initiated (Log Page 0x08)
NVMe-17D	Commands Supported and Effects (Log Page 0x05)

4.9.3 Additional Log pages

The following table details additional fields that need to be implemented in vendor specific log pages. The log page identifier is provided as a suggestion, but other identifiers may be used provided the NVMe CLI plugin requirements are met (see Section 4.10.1).

Requirement ID	Description
NVMe-18A	All values in the 0xCA log page shall be persistent across power cycles
NVMe-18B	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).

Req ID	**Field**	**# of Bytes**	**Field description**
NVMe -18D	Log page directory	Vendor defined	Provides a list of available log pages and corresponding log identifiers
NVMe -18E	Physical (NAND) bytes written	16	The number of bytes written to NAND. It must be possible to use this attribute in conjunction with another attribute to calculate the Write Amplification Factor (WAF). Any formulas required shall be provided.
NVMe -18F	Physical (NAND) bytes read	16	The number of bytes read from NAND.
NVMe -18G	Bad NAND block count	8 (2 bytes for normalized + 6 bytes for raw count)	Raw and normalized count of the number of NAND blocks that have been retired after the drive's manufacturing tests (i.e. grown bad blocks)
NVMe -18H	XOR Recovery count	8	Total number of times XOR was invoked to recover data. Data recovery may have succeeded or failed. See Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18I	Uncorrectable read error count	8	Total count of NAND reads that were not correctable by read retries, all levels of ECC, or XOR (as applicable). Data recovery fails, and an uncorrectable read error is returned to the host. See Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18J	Soft ECC error count	8	Total count of NAND reads that were not correctable by first-level ECC and requires invoking an intermediate recovery. Data recovery may have succeeded or failed. If the SSD have more than one intermediate recovery levels, then this counter only increments when intermediate recovery level 1 is

			invoked. Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18K	SSD End to end correction counts	8 (4 bytes for count of detected errors, 4 bytes for count of corrected errors)	A count of the detected and corrected errors by the SSD end to end error correction which includes DRAM, SRAM, or other storage element ECC/CRC protection mechanism (not NAND ECC). All correctable errors must result in a counter increase no matter what type of data the memory is protecting. All detected errors must result in a counter increase unless the error is uncorrectable and occurred in the system region. In the latter case, the incomplete shutdown flag must be flagged/incremented on the next power up.
NVMe -18L	System data % used	1	A normalized cumulative count of the number of erase cycles per block since leaving the factory for the system (FW and metadata) area. Starts at 0 and increments. 100 indicates that the estimated endurance has been consumed. Value is allowed to exceed 100 up to 255.
NVMe -18M	User data erase counts	8 (4 bytes for the maximum, 4 bytes for the minimum)	The maximum and minimum erase counts across all NAND blocks in the drive. The host shall not be able to reset this counter.
NVMe -18N	Refresh count	8	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.
NVMe -18O	Program fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Raw and normalized count of total program failures. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18P	User data erase fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Raw and normalized count of total erase failures in the user area. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18Q	System area erase fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Raw and normalized count of total erase failures in the system area. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18R	Thermal throttling status and count	2 (1 byte for the current status, 1 byte for the count)	The current status of thermal throttling (enabled or disabled) and a count of the number of thermal throttling events.
NVMe -18S	PCIe Correctable Error count	8	Summation counter of all PCIe correctable errors (Bad TLP, Bad DLLP, Receiver error, Replay timeouts, Replay rollovers)

NVMe -18T	Incomplete shutdowns	4	A count of the number of shutdowns that have occurred that did not complete properly
NVMe -18U	% 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 mean both blocks that have been erased and blocks that have all invalid data. Invalid blocks are blocks that are either marked invalid by drive FW OR by the host (via TRIM or overwrite). For example, if the total number of spare blocks is 100 and garbage collection has been able to reclaim 20 blocks, then this field reports 20%.

4.9.4 SMART Log Persistence

Requirement ID	Description
NVMe-19A	The SSD shall not lose any of the SMART data log which is more than 1 hour old including across power cycles/resets.
NVMe-19B	The SSD shall not lose any super cap failures and SMART critical warnings including across power cycles/resets.

4.9.5 PCIe Error Logging

The following table includes the PCIe physical layer error counters that need to be implemented. This is in addition to aggregated PCIe error counters defined above in Section 4.9.3.

NVMe-20A:

Event	*Counted in PCIe correctable error counter?*	*Description*
Unsupported Request Error Status (URES)	No	PCIe configuration space error types. This should be reported in the PCIe standard configuration space registers (PCIe Base Specification 3.1 Section 7.10.2 and 7.10.5)
ECRC Error Status (ECRCES)	No	
Malformed TLP Status (MTS)	No	
Receiver Overflow Status (ROS)	No	
Unexpected Completion Status (UCS)	No	
Completer Abort Status (CAS)	No	

Completion Timeout Status (CTS)	No	
Flow Control Protocol Error Status (FCPES)	No	
Poisoned TLP Status (PTS)	No	
Data Link Protocol Error Status (DLPES)	No	
Advisory Non-Fatal Error Status (ANFES)	No	
Replay Timer Timeout Status (RTS)	Yes	
REPLAY_NUM Rollover Status (RRS)	Yes	
Bad DLLP Status (BDS)	Yes	
Bad TLP Status (BTS)	Yes	
Receiver Error Status (RES)	Yes	

4.10 Utility

4.10.1 Management Utility

Facebook will be using the NVMeCLI utility (https://github.com/linux-nvme/nvme-cli) as the management utility for NVMe SSDs.

Requirement ID	Description
UTIL-1A	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

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

Requirement ID	Description
UTIL-1B	A single, common plugin for all of the supplier's NVMe-based products
UTIL-1C	Vendor and additional log page decoding including into a human readable format and JSON output
UTIL-1D	Access to OEM commands
UTIL-1E	The ability to pull crash dumps or FW logs (binary output is acceptable)
UTIL-1F	The plugin's subcommand nomenclature must adhere to the table below and cannot change across versions unless approved by Facebook.

Requirement ID	NVMe CLI Nomenclature	Purpose
UTIL-1G	vs-smart-add-log	Retrieve extended Facebook SMART Information from section 4.9.3.
UTIL-1H	vs-internal-log	Retrieve drive telemetry logging. See 4.7.1.1 for the definition and refer to Appendix C: NVMe-CLI Plugin Output for the required syntax.
UTIL-1I	vs-telemetry-controller-option	Controls the controller-initiated telemetry. The default state is to DISABLE controller-initiated telemetry. See 4.7.1.1 for the definition and refer to Appendix C: NVMe-CLI Plugin Output for the required syntax.
UTIL-1J	vs-error-reason-identifier	Retrieves the error reason identifier from the telemetry log page. See section 4.7.1.1 for the definition.
UTIL-1K	vs-fw-activate-history	Lists the last twenty firmware that were activated (not downloaded) on the drive. Each entry must have a POH timestamp that correlates to the current POH in the standard SMART. See Appendix C: NVMe-CLI Plugin Output for the output rules on this command.
UTIL-1L	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 Facebook 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-1M	clear-pcie-correctable-errors	VUC that clears the correctable PCIe error counter
UTIL-1N	clear-fw-activate-history	VUC that clears the output of the “vs-fw-activate-history”
UTIL-1O	log-page-directory	VUC that lists all the log pages and a description of their contents
UTIL-1P	plugin-version	Shows the plugin’s version information
UTIL-1Q	Help	Display this help

4.10.2 PCIe eye capture

Requirement ID	Description
UTIL-2A	A utility shall be provided that will allow Facebook to grab the internal eye of the device in order to tune the signal integrity of the device to the target platform.

5 Functional

5.1 PCIe requirements

5.1.1 Lane width

Requirement ID	Description
FUNC-1A	The device shall support a x4 lane width.

5.1.2 Maximum Payload Size

Requirement ID	Description
FUNC-2A	The SSD shall support a PCIe Maximum Payload Size of 256 bytes.

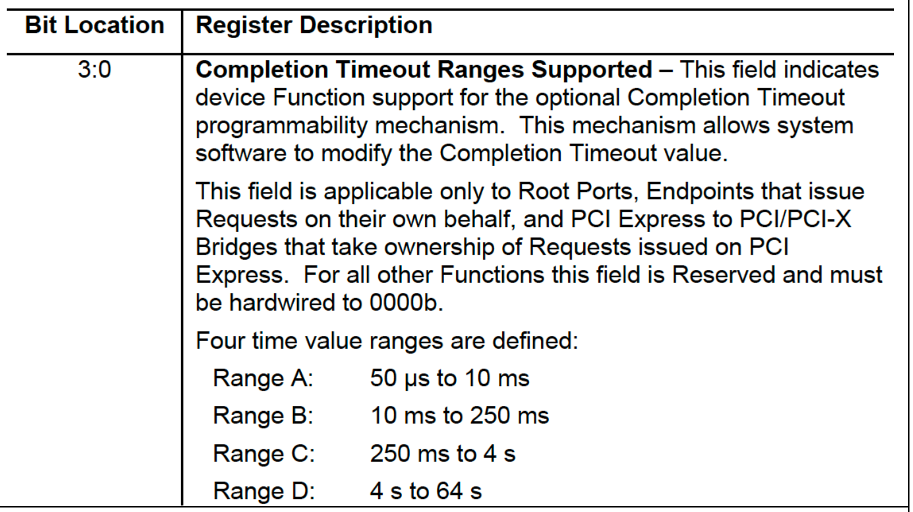
5.1.3 Lane reversal

Requirement ID	Description
FUNC-3A	The SSD must support lane reversal with all lanes connected or partially connected lanes. (e.g. a x4 device must support it for x4, x2, and x1 connections).

5.1.4 PCIe Compliance

Requirement ID	Description
FUNC-4A	Must be compliant to PCIe base specification 3.1a.
FUNC-4B	Comply to PLI_1.8V_USB_Higher_Power ECN, which also includes pin out changes.
FUNC-4C	Provide PCIe compliance report.

5.1.5 PCIe Timeout Support

<table>
<tr><th>Requirement ID</th><th>Description</th></tr>
<tr><td>FUNC-5A</td><td>The SSD Controller shall support modification of PCIe TLP completion timeout range as defined by the PCIe Base Spec.</td></tr>
<tr><td>FUNC-5B</td><td>The vendor must disclose the vendor-specific timeout range definition if the controller deviates from the base specification below:

<table>
<tr><th>Bit Location</th><th>Register Description</th></tr>
<tr><td>3:0</td><td><b>Completion Timeout Ranges Supported –</b> This field indicates device Function support for the optional Completion Timeout programmability mechanism. This mechanism allows system software to modify the Completion Timeout value.
This field is applicable only to Root Ports, Endpoints that issue Requests on their own behalf, and PCI Express to PCI/PCI-X Bridges that take ownership of Requests issued on PCI Express. For all other Functions this field is Reserved and must be hardwired to 0000b.
Four time value ranges are defined:
Range A: 50 µs to 10 ms
Range B: 10 ms to 250 ms
Range C: 250 ms to 4 s
Range D: 4 s to 64 s</td></tr>
</table></td></tr>
<tr><td>FUNC-5C</td><td>Disabling of PCIe Completion Timeout shall also be supported by the SSD Controller</td></tr>
</table>

5.1.6 PCIe Reset Supported

Requirement ID	Description

FUNC-6A	PCIe Conventional Reset: • PCIe Cold or Warm Reset (*achieved by toggling of PERST#*)
FUNC-6B	PCIe Function Level Reset
FUNC-6C	PCIe Hot Reset

5.2 Boot requirements

Requirement ID	Description
FUNC-7A	The SSD does not need to be boot-able, but does need to be visible in UEFI/BIOS. An option ROM shall not be included.

5.3 TRIM

Requirement ID	Description
FUNC-8A	The SSD shall support TRIMs.
FUNC-8B	For data that has been De-Allocated (TRIM) the NVMe specification requires it to be 0, 1, or unchanged when read. FB does not require this to be followed. FB requires once the data has been de-allocated and then written to that the data read matches what was written. When the LBA is in the de-allocated state there are no requirements involving TRIM for what the data is or that it does not change.
FUNC-8C	If data has been de-allocated and not written to and then an unsafe power down event happens FB has no requirements around TRIM for these LBAs with respect to what data is returned when power is re-applied to the SSD.

5.4 Sector size support

Requirement ID	Description
FUNC-9A	The SSD shall support 4096 byte sectors and shall be formatted to this sector size from the factory.

5.5 Data protection

Requirement ID	Description
FUNC-10A	The SSD shall support a protection scheme that protects against NAND block level failures.
FUNC-10B	The protection scheme must also support NAND plane level failures without data or metadata loss.

5.6 Power-loss protection support

5.6.1 Support requirements

Requirement ID	Description
FUNC-11A	The SSD shall support full power-loss protection for all acknowledged data and metadata.
FUNC-11B	The Power-loss protection health check shall not impact IO latency.
FUNC-11C	Metadata rebuild due to an unexpected power loss shall not exceed 120 seconds and the SSD must be fully operational after this.
FUNC-11D	Power-loss protection health check shall be performed by the firmware at least once every 24 hours.
FUNC-11E	While performing the power-loss protection health check, the SSD must still have enough charge be able to handle an ungraceful power loss properly.
FUNC-11F	In case of a graceful shutdown operation (*CC.SHN=1 set by the NVMe driver*), no data loss is tolerated.
FUNC-11G	An ungraceful shutdown event shall not make the drive non-functional under any conditions.
FUNC-11H	The firmware algorithm must 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. The safeguard algorithm must be reviewed with Facebook.

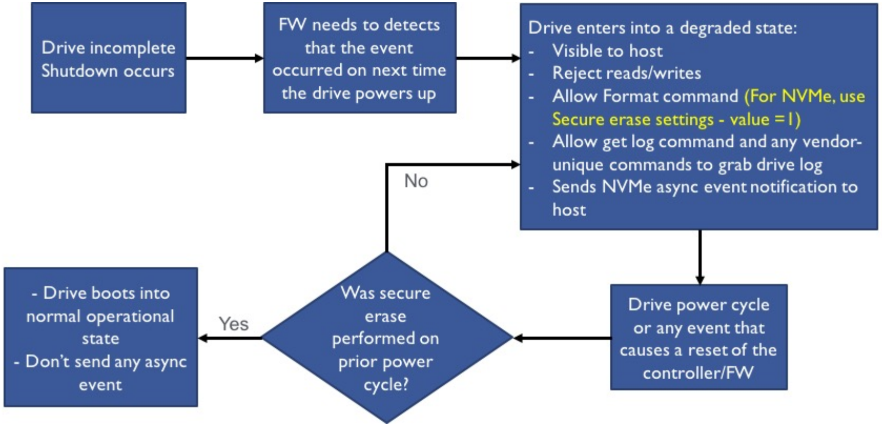
5.6.2 Power-loss protection failure

Requirement ID	Description
FUNC-12A	When the power-loss protection mechanism fails for any reason while power is applied, the SSD shall switch to “Read-only” mode and shall not enter into write-through mode.
FUNC-12B	The SSD shall still support data eradication as defined in 5.7 even if it is operating in “Read-only” mode, and it shall support admin commands to enable reading the sensor or SMART data.

5.6.3 Incomplete shutdown

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

Requirement ID	Description

FUNC-13A	The SSD shall incorporate a shutdown checksum or flag as the very last piece of data written to flash to detect incomplete shutdown.
FUNC-13B	This checksum in FUNC-10A must be used on power-up to confirm that the previous shutdown was 100% successful.
FUNC-13C	The incomplete shutdown will result in an increase in the Facebook SMART “incomplete shutdown” counter and the NVMe standard SMART log “critical warning” field shall have bit 2 set (NVM subsystem reliability).
FUNC-13D	The following diagram explains the desired SSD behavior if a shutdown was incomplete.  Drive incomplete Shutdown occurs FW needs to detects that the event occurred on next time the drive powers up Drive enters into a degraded state: - Visible to host - Reject reads/writes - Allow Format command (For NVMe, use Secure erase settings - value = 1) - Allow get log command and any vendor-unique commands to grab drive log - Sends NVMe async event notification to host No Drive power cycle or any event that causes a reset of the controller/FW Was secure erase performed on prior power cycle? Yes - Drive boots into normal operational state - Don't send any async event *Figure 1: Incomplete shutdown sequence and expected behavior*

5.7 Data Encryption and Eradication

Requirement ID	Description
FUNC-14A	The SSD shall support AES-256 encryption (or better), or NAND-level data eradication using the NVMe Format feature defined in the NVMe specification version in Section 4.1.
FUNC-14B	If encryption is implemented and appropriate test documentation can be provided, then NAND-level data eradication is not required.
FUNC-14C	The following is a list of NAND-level functionality that is required if encryption and crypto erase is not supported: a. Performs a physical NAND-level erase on every NAND block including any grown bad blocks (factory bad blocks which can not contain any application data can be excluded). b. The operation shall PASS only if ALL NAND blocks are erased successfully. This include grown bad blocks. c. The operation shall FAIL if the above operations fails or it is not possible to physically erase any NAND block for any reason even if it is on the grown bad block list.

	d. The SSD shall report a failure to the host. The supplier shall determine the error code to return, but it must be defined and described in the documentation. e. This functionality will be assigned to the Secure Erase Setting (SES) = 011h in the Format command.

5.8 FW updates

Requirement ID	Description
FUNC-15A	FW updates shall not require a power cycle, any types of PCIe convention reset, or reboot. A NVMe controller reset must be sufficient to activate the new FW.
FUNC-15B	A FW activation history log must be recorded and retrievable via NVMe CLI plug-in. See "vs-fw-activate-history" command in section 4.10.1 and Appendix C: NVMe-CLI Plugin Output for the output rules.
FUNC-15C	Any drives used for qualification by Facebook or its ODMs cannot have any restrictions on the # of firmware downloads.
FUNC-15D	For firmware commit action 010b (firmware activation without reset), the SSD shall complete the firmware activation process and be ready to accept host IO and admin commands within 5 seconds from the receipt of the firmware commit command.

5.8.1 FW Downgrade protection

Requirement ID	Description
FUNC-16A	The FW needs to prevent any FW update operations from completing if the FW downgrade is incompatible with the current version of FW. These FW limitations must be clearly communicated to Facebook.

5.9 Time to Ready

Requirement ID	Description
FUNC-17A	The SSD Controller shall come on-line within 20 seconds, i.e. the time taken after CC.EN = 1, before CSTS.RDY = 1 shall not be > 20s. This is referred to as "Time to Ready". In no event shall the drive take longer than CAP.TO (including worst case scenarios e.g. Abrupt Shutdown).

5.10 Background data refresh

Requirement ID	Description
FUNC-18A	The SSD shall support background data refresh will the SSD is powered on to ensure there is no data-loss due to power-on retention issues.
FUNC-18B	This must be designed and tested to support the normal NAND operating temperature as described in section 0. In other words, if the SSD is cooled to a composite temperature of 70C which in turn implies a NAND temperature of 80C this must be taken into account.
FUNC-18C	Background data refresh should cover the entire drive including the over-provisioned area and be designed to continuously run in the background and not just during idle periods.

5.11 Low power Modes

Requirement ID	Description
FUNC-20A	Facebook does not require the drive to support ASPM (Active State Power Management).
FUNC-20B	The default firmware should disable these PCIe power management features.

5.12 Command Timeout

The Facebook kernel adheres to the following command timeouts as defined by the *Linux NVMe Inbox Driver*. The SSD supplier must disclose any I/O scenario that could violate these command timeouts.

Requirement ID	Description
FUNC-21A	ADMIN Commands: 60 seconds I/O Commands: 30 seconds

6 Endurance

6.1 Endurance data

Requirement ID	Description
ENDU-1A	The SSD documentation shall include the number of physical bytes that can be written to the SSD assuming a write amplification of 1. This will be used in the formula shown below. The units should be GB (10^9 bytes). $\text{Physical Drive Writes per Day (pDWPD)} = \frac{\text{Physical Bytes Written @ WAF = 1}}{\text{(5 years x 365 days x usable capacity)}}$

6.2 Endurance conditions

Since there are a number of factors that impact the SSD endurance, the table below provides the requirements for Facebook’s environment, which may be different from “standard” requirements.

Requirement ID	Description
ENDU-2A	Powered-off data retention (end of life) to be at least 1 week at 25C.
ENDU-2B	Powered-on data retention to be at least 5 years.
ENDU-2C	The SSD shall not throttle its performance based on the endurance metric (AKA endurance throttling).

6.3 Shelf life

Requirement ID	Description
ENDU-3A	A new SSD may be kept as a datacenter spare and therefore must be fully functional even if it sits on the shelf for up to 1 year @ 40C before getting installed in the server. When installed the drive will be formatted.

6.4 End-of-Life (EOL) Testing

Requirement ID	Description
ENDU-4A	Facebook requires various types of samples for EOL testing in accordance to the quantities outlined in the “Facebook NVMe EOL Test Process” document. These are separate samples used for qualification.
ENDU-4B	The EOL workloads are the ones defined in Appendix A: Performance targets.

7 Electrical

7.1 PCIe Add-in Card

7.1.1 Hot-swap

Requirement ID	Description
ELEC-1A	The SSD does not need to support hot-swap.

7.1.2 Power/Activity LED

Requirement ID	Description
ELEC-2A	The SSD shall support a power or activity LED.

7.1.3 Power consumption

Requirement ID	Description
ELEC-3A	The SSD shall not consume more than 40W under any conditions.

7.2 2.5” U.2 (SFF-8639)

7.2.1 Hot-swap

Requirement ID	Description
ELEC-4A	2.5” SSDs shall support hot-swap.

7.2.2 Power/Activity LED

Requirement ID	Description
ELEC-5A	The SSD shall support driving an activity LED through the connector.
ELEC-5B	The LED should be lit solidly when power is applied and flashing when there is traffic going to the SSD.

7.2.3 Power consumption

Requirement ID	Description
ELEC-6A	The SSD shall not consume more than 14W under any conditions.

7.2.4 In-rush current

Requirement ID	Description
ELEC-7A	The SSD shall not exceed 2.3A during a power-on or hot-swap event.

7.3 M.2

7.3.1 Hot-swap

Requirement ID	Description
ELEC-8A	M.2 SSDs shall support hot-swap with the use of a Facebook designed carrier.

7.3.2 Power/Activity LED

Requirement ID	Description
ELEC-9A	The SSD shall support driving an activity LED through the connector.
ELEC-9B	The LED should be lit solidly when power is applied and flashing when there is traffic going to the SSD.

7.3.3 Power consumption

Requirement ID	Description
ELEC-10A	The SSD maximum average power consumption over 500ms for any workload shall not exceed 8.5W with a sampling rate of 2ms or better. The measurement duration must be at least 15 minutes on a pre-conditioned drive.
ELEC-10B	The SSD peak power shall not exceed 13W in a 100 us window with a sampling rate of 4uS or better. The measurement duration must be at least 15 minutes on a pre-conditioned drive.

7.3.4 SMBUS support

Requirement ID	Description
ELEC-11A	The SSD shall support 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)
ELEC-11B	The SSD's SMBUS spec shall comply to version 3.1. (http://smbus.org/specs/SMBus_3_1_20180319.pdf)

8 Thermal

8.1 Operating Conditions

8.1.1 Data Center Altitude

Requirement ID	Description
THRM-1A	Support for data centers being located at an altitude of up to 2000 meters above sea level is required

8.1.2 Cold-Aisle temperature

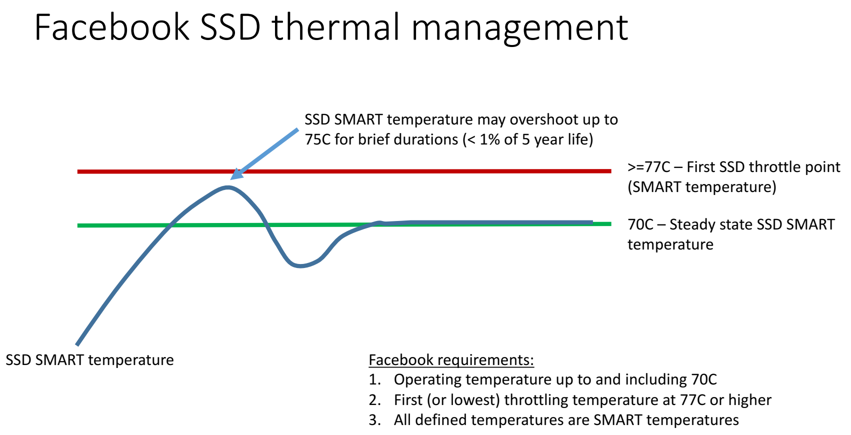
Requirement ID	Description
THRM-2A	Thermal study with each chassis being qualified is required. Some background information related to this is the following: The data centers will maintain the cold aisle temperatures between 18°C and 30°C (65°F to 85°F). The mean temperature in the cold aisle is 24°C with 3°C standard deviation. The cold aisle temperature in a data center may fluctuate minutely depending to the outside air temperature of data center.

8.1.3 Relative Humidity

Requirement ID	Description
THRM-3A	The SSD shall operate normally with relative humidity to be between 10% and 90%.

8.2 Thermal throttling

Requirement ID	Description
THRM-4A	The SSD shall implement a thermal throttling mechanism to protect the SSD in case of a failure or excursion that causes the SSD temperatures to increase above its maximum specified temperature.
THRM-4B	The SSD shall begin thermal throttling at a composite temperature of 77C or higher.
THRM-4C	Facebook requires a single throttling point at the highest possible temperature. Multiple throttling steps are not acceptable.
THRM-4D	However, under normal operating conditions the SSD shall not engage in thermal throttling when installed in Facebook systems as Facebook will maintain the SSD temperature at or below 70C (as reported by the composite temperature).

THRM-4E	Thermal throttling shall only engage under certain failure conditions such as excessive server ambient temperature or multiple fan failures. The desired behavior is illustrated below in Figure 1.  *Figure 1: Facebook's thermal management scheme*
THRM-4F	The firmware algorithm must 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. The safeguard algorithm must be reviewed with Facebook.

8.3 Temperature reporting

Requirement ID	Description
THRM-5A	The SSD shall expose the raw sensor readings from all of the sensors on the SSD. The raw sensor readings shall by placed in the "Temperature Sensor X" fields in the NVMe CLI smart-log output.
THRM-5B	The SSD's drive-to-drive composite temperature variation shall be +/- 1 degrees C. Two different drives shall not report a composite temperature greater than 2 degrees apart under the same environmental conditions, slot location, and workload.
THRM-5C	The SSD's within drive composite temperature variation shall be +/- 1 degrees C. A single drive's composite temperature shall not vary by more than 2 degrees once it is in a steady state under the same environmental conditions, slot location, and workload.
THRM-5D	The supplier shall provide the equation, settings, and thresholds used to calculate the composite temperature. Any changes in the thermal equation, settings, or thresholds must be clearly communicated to Facebook.

8.4 Thermal Shutdown

Requirement ID	Description
THRM-6A	If the SSD implemented a mechanism to shut down or halt the drive at a given temperature, that temperature value must be at 85°C composite temperature or higher.

9 Mechanical

9.1 PCIe Add-in Card

Requirement ID	Description
MECH-1A	The SSD shall adhere to the PCIe CEM 3.0 specification. Either half-height or full-height, half-length cards are acceptable. The airflow direction in our servers has the upstream or inlet air coming from through the PCIe bracket.

9.2 2.5" U.2

Requirement ID	Description
MECH-2A	The SSD shall adhere to the SFF-8639 (U.2) specification.
MECH-2B	The SSD shall have a thickness of 7mm.

9.3 M.2

Requirement ID	Description
MECH-3A	The SSD shall adhere to the M.2 specification with a size of 22mm x 110mm.
MECH-3B	The bottom-side height shall not exceed 1.5mm.
MECH-3C	The top-side height shall not exceed 2mm.
MECH-3D	The SSD shall use an M key.

10 SMBUS support

10.1 Temperature

Requirement ID	Description
SBUS-1A	The SSD shall support the NVMe Simple Management Interface specification 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.
SBUS-1B	Both SMBUS block read and byte read commands must be supported.
SBUS-1C	Facebook requires additional vendor-specific information that needs to be outputted in the Subsystem Management Data Structure. See Appendix B: Vendor-specific NVMe-MI output.

10.2 VPD

Requirement ID	Description
SBUS-2A	VPD support is optional, but desirable if available.

11 Security

Requirement ID	Description
SEC-1	Shall support signed firmware binary update which is checked before firmware is activated.
SEC-2	Shall have XTS-AES-256 or AES-256 hardware-based data encryption or better is required.
SEC-3	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	Shall support crypto erase.
SEC-5	Shall support Secure Boot.
SEC-6	Must have a method of identifying a secure boot failure which does not require physical access to the SSD

SEC-7	Shall be FIPS 140-2 capable (not required to get FIPS certificate).
SEC-8	Shall support Key revocation allowing a new key to be used for firmware validation on update. Preferred implementation is for up to 4 keys
SEC-9	Shall support Ruby Version 1.00, Revision 1.00 and Configurable Namespace Locking (CNL) feature set Version 1.00, revision 1.00). (mandatory support for Namespace Global Range Locking object, optional support for Namespace Non-Global Range Locking object)
SEC-10	Secure development processes must be followed. Annual 3rd party audits must be conducted, and results shared with FB.
SEC-11	All signing keys must be stored in a hardware security module (HSM).
SEC-12	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 must be provided to FB.
SEC-13	Uncontrolled remote access to intrusive debug capabilities (JTAG/UART, etc) must not be possible.
SEC-14	Adversarial testing using red-teams shall be conducted before qualification start. A report of items attempted and results must be provided to FB.

12 Labeling

Requirement ID	Description
LABL-1A	2 separate labels: 1 for manufacturer part number, and 1 for serial number
LABL-1B	Each label contains a single data element with no spaces. Dashes are acceptable
LABL-1C	Minimum feature size that can be captured: 10 mils (0.254 mm)
LABL-1D	Data contained in the 2D bar code should also be printed in the label margin area in human-readable format. Minimum font size: 6
LABL-1E	Labels must be visible on the product with or without TIP (gap pad) material applied to the key components (controller, NAND, DRAM).
LABL-1F	Facebook also has the following key requirements with regards to the serial number: • The S/N needs to be a 2D QR Data-Matrix • It shall be at least 5mm x 5mm by XY dimension • S/N cannot contain vendor ID, it needs to be a S/N only • S/N must match the electrical read out S/N from drive • Human readable S/N is also required by the side of 2D SN QR code.

13 Compliance

13.1 ROHS Compliance

Requirement ID	Description
ENV-1A	The Supplier shall adhere to the latest version Facebook Materials of Concern Standard 4.0 (Agile PN: 18-000142), which may be updated over time. Facebook will notify the Supplier with updates to the Standard.
ENV-1B	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
ENV-2A	SSD manufacturer needs to provide ESD immunity level (HBM-Human Body Model) measured in accordance with ANSI/ESDA/JEDEC JS-001-2010 spec.

14 Shock and Vibration

Below are the shock and vibration specifications for the Facebook M.2 SSDs at the M.2 module level.

Requirement ID	Description
SV-1A	The operational shock requirement is 700G, half-sine, 0.5ms, total 6 shocks, along all three axes (+/-)
SV-1B	The non-operational shock requirement is 700G, half-sine, 0.5ms, total 6 shocks, along all three axes (+/-)
SV-1C	The vibration requirement during operation is: 1.8G_{rms}, 5-500 Hz, Random Vibe, 20 min along all three axes
SV-1D	The vibration requirement during non-operation is: 15G_{rms}, 5-1500 Hz, total 6 sweeps along all three axes

15 Future Considerations

The items listed in this section are not requirements today but have a high chance that they will turn into requirements in the future. The goal is to provide a preview of the requirements and features to help with technology alignment. The items are not listed in order of priority.

1. Read Recovery Level (TP4018a)
2. The Appendix A: Performance targets will include additional workloads
3. PWRDIS and/or PLN# pin support for M.2 (power disable)
4. Increased maximum average power limit on M.2 form factor
5. PCIe maximum payload size support of 512 bytes
6. MDTS value support of 512KB
7. Data Security. Requirements for data security should be expected to continue to increase in the future.
8. About 3 to 5 FB-FIO Synth Flash additional workloads is considered for the next round of spec update.
9. NVM Sets is under investigation for the future. Please see Appendix E: NVM Set Design Guidance for further information.
10. Persistent log page support

16 Required documentation

Please see the “Document Checklist” sheet in the latest “Pre-Qual checklist” document.

17 Rules of Engagement During Qualification

Requirement ID	Description
QUAL-1A	Full release notes with all the changes must be provided with each FW release and must comply with the format outlined in the "Product Release Note Details" document.
QUAL-1B	All changes related to the following areas must be clearly communicated to Facebook prior to implementation: • Performance changes in terms of bandwidth or latency for reads/writes/trims • Thermal equation, settings, or thresholds • Hardware changes • Endurance or TBW @ WAF=1 • HW/FW Changes that results in a change in Write amplification factor
QUAL-1C	The qualification drive samples that Facebook receives must not have any hard restrictions in terms of the following: • # of power cycles/resets performed • # of firmware downloads performed • # of secure erases performed
QUAL-1D	Every firmware binary release that is not backwards compatible with the previous firmware should include an a "firmware download companion" binary that is exactly the same but with a different version number to facilitate the testing of FW upgrade and downgrade.
QUAL-1E	Silent data corruption shall result in immediate disqualification.

18 Revision History

Author	**Revision**	**Description**
Chris Petersen	0.1	Initial release
Chris Petersen	0.2	Added additional details to the utility, endurance, and ERAD sections.
Chris Petersen	0.3	Added reliability section, some additional tweaks
Chris Petersen	0.4	Outlined additional logging needs, changed the utility to NVMeCLI, and lots of wording improvements
Chris Petersen	0.5	Updated altitude, thermal behavior, NVMe version, namespace commands, shutdown behavior

Chris Petersen	0.6	Added performance targets, ability to clear PCIe counters, ROHS requirement, log page clarifications
Chris Petersen, Ta-Yu Wu	0.7	Thermal requirements improved, added IOD and Set requirements, added telemetry requirements, improved log page definitions, updated the performance targets, updated compliance/ESD requirements
Ta-Yu Wu, Ross Stenfort	0.8	1. Updated document format to include requirement labels 2. Included a target timeframe for when the spec will take effect 3. Modified the E2E datapath protection section 4. Added a NVMe controller configuration and behavior section 5. Added requirements for format progress indicator and IO performance and Endurance Hints 6. Modified IOD and Set requirements including associated log page modifications 7. Added requirements for Command Supports and Effects log page. 8. Added requirements for modifying PCIe completion timeouts 9. Modified the TRIM requirements 10. Modified XOR protection requirement to data protection requirement 11. Added SSD security requirements 12. Added EOL testing requirements 13. Modified power requirements 14. Added Shock and Vibration requirements 15. Added a future consideration section 16. Clarified in the performance targets to use “kyber” as the default scheduler 17. Added a sequential write bandwidth requirement 18. Added new appendix with details for NVMe-MI bytes, NVMe-CLI plugin outputs, and SSD read-recovery level definitions 19. Misc. changes and clarifications to existing specifications
Ross Stenfort, Ta-Yu Wu	0.81	1. Updated Section THRM-3A, NVME-10K. 2. Moved PLM section to future requirements section. 3. Removed old security section (5.11) and added section 11 for security along with new security requirements for section 11. 4. Modified the performance targets to capacity based and updated the latency targets.

		5. Updated the target date to June 2019. 6. Updated ELEC-10A and ELEC-10B to include sampling rate. 7. Added safeguard requirements to the temperature sensor and the PLP health check.
Ta-Yu Wu, Ross Stenfort	0.82	1. Updated the shock and vibration requirements 2. Moved NVM Set requirements to the future requirements section. 3. Added requirements for firmware activation without reset 4. Update to the latest compliance specification 5. Clarified the requirements for a few of the SMART counters 6. Removed the Sanitize command requirements 7. Removed Predictable Latency Mode requirements from the future section

Appendix A: Performance targets

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 and not a pass/fail criteria as that will be determined by Facebook on a case by case basis.

The targets are broken down into five segments:

1. FB-FIO Synth Flash Targets by drive capacity
    a) 4TB Targets
    b) 2TB Targets
    c) 1TB Targets
2. fb-FIOSynthFlash TRIM Rate target
3. IO.go benchmark target
4. Fileappend benchmark target
5. Sequential write bandwidth

All targets shall be achieved by using “kyber” as the I/O scheduler.

1a. 4TB SSD Capacity Performance Targets for FB-FIO Synth Flash - HE_Flash_Short_TRIM

Workload	Read IOPS per TB	Write MiB/s per TB	P99 Read Latency	P99.99 Read Latency	P99.9999 Read Latency	P99.99 Write Latency	P99.9999 Write Latency
4K_L2R6DWPD_ wTRIM	8,750 (34MiB/s)	72 MiB/s	3,000 us	6,500 us	11,000 us	20,000 us	35,000 us
4K_L2R9DWPD_ wTRIM	8,750 (34MiB/s)	93 MiB/s	3,200 us	7,000 us	12,000 us	20,000 us	35,000 us
MyRocks_Heavy_ wTRIM	4,375 (34MiB/s)	82 MiB/s	3,000 us	6,500 us	10,000 us	15,000 us	20,000 us
Fleaf	35,000 (410MiB/s)	87 MiB/s	3,300 us	7,000 us	12,000 us	25,000 us	35,000 us

1b. 2TB SSD Capacity Performance Targets for FB-FIO Synth Flash - HE_Flash_Short_TRIM

Workload	**Read IOPS per TB**	**Write MiB/s per TB**	**P99 Read Latency**	**P99.99 Read Latency**	**P99.9999 Read Latency**	**P99.99 Write Latency**	**P99.9999 Write Latency**
4K_L2R6DWPD_wTRIM	8,750 (34MiB/s)	72 MiB/s	2,000 us	5,000 us	8,500 us	15,000 us	25,000 us
4K_L2R9DWPD_wTRIM	8,750 (34MiB/s)	93 MiB/s	2,200 us	5,500 us	9,500 us	15,000 us	25,000 us
MyRocks_Heavy_wTRIM	4,375 (34MiB/s)	82 MiB/s	2,000 us	5,000 us	8,500 us	10,000 us	15,000 us
Fleaf	35,000 (410MiB/s)	87 MiB/s	3,000 us	6,000 us	10,000 us	20,000 us	25,000 us

1c. 1TB SSD Capacity Performance Targets for FB-FIO Synth Flash – Search Index and Search LM

Workload	**Read IOPS per TB**	**Write MiB/s per TB**	**P99 Read Latency**	**P99.99 Read Latency**	**P99.9999 Read Latency**	**P99.99 Write Latency**	**P99.9999 Write Latency**
SearchIndex_wTRIM	76,000 (560MiB/s)	20 MiB/s	3,500 us	10,000 us	15,000 us	20,000 us	25,000 us
SearchLM_wTRIM	72,500 (2,000MiB/s)	12 MiB/s	1,500 us	10,000 us	15,000 us	20,000 us	25,000 us

2. 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

3. 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

4. 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

5. 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

Appendix B: Vendor-specific NVMe-MI output

The following table outlines the additional vendor-specific NVMe-MI output required by Facebook for SSDs. Byte offset starts at 32 (decimal).

Command Code (Decimal)	Byte Offset (Decimal)	Definition	Value	Description
32 (Device Identifier)	32	LOI	0x35 (53 decimal)	Length of Identification: Indicates the number of bytes before PEC is encountered.
	33	VERSION	0x1	Version: This is the version of the FB Standard Device Identifier.
	73:34	PP/MN		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.
	74	MEFF		Management End Point Form Factor. This field should be populated based on the updated "NVMe MultiRecord Area" form factor list in TP6007 for section 9.2.3 of the NVMe Management Interface 1.0a.
	75	FFI_0	FFI_0[3:0] 0x0 FFI_0 [7:4] Reserved	Form Factor Information 0 Register
	85:76	Reserved		Reserved for future use
	86	PEC		The 8-bit code used to verify the address/data.
87 (Storage)	87	LOI	0x7 (7 decimal)	Length of Identification: Indicates the number of bytes before PEC is encountered.
	88	SVERSION	0x1	Storage Version. This is the version of the storage data structure.
	90:89	Capacity		This is the raw capacity in GB in Hex (2048 GB in raw capacity = 0x800). Does not include any extra spare blocks within the NAND.
	91	PWR		POWER: This is RMS power 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
	92	SINFO_0	SINFO_0 [1:0] 0x0 - Power Loss Protection not	Storage Information 0

			supported 0x1 - Power Loss Protection supported 0x2-0x3 - Reserved SINFO_0 [7:2] - Reserved	
	93-94	Reserved		Reserved
	95	PEC		The 8-bit code used to verify the address/data.

Appendix C: NVMe-CLI Plugin Output

The following section outlines the output format and rules for certain NVMe-CLI plugin sub-commands.

CLI plug in subcommand:
“vs-smart-add-log”

Drive-level data for drives without NVM Sets:
Drive - physical bytes written
Drive - physical bytes read

Example Output:

```
Drive - physical bytes written: 100
Drive - physical bytes read: 100
```

Drives with NVM Sets needs to have additional NVM Set-level data outputted. Below shows only a snippet of the attribute in the 0xCA log page. The entire 0xCA log page output is required in the actual implementation.

Drive-level data for drives with NVM Sets:
Drive - physical bytes written
Drive - physical bytes read

NVM Set-level data (needs to be repeated for each set) for drives with NVM Sets:
[Set #ID][NS #ID] - physical bytes written
[Set #ID][NS #ID] - physical bytes read

Example Output:

```
Drive - physical bytes written: 100
Drive - physical bytes read: 100
[Set 1][NS 1] - physical bytes written: 20
[Set 1][NS 1] - physical bytes read: 20
[Set 2][NS 2] - physical bytes written: 25
[Set 2][NS 2] - physical bytes read: 25
[Set 3][NS 3] - physical bytes written: 35
[Set 3][NS 3] - physical bytes read: 35
[Set 4][NS 4] - physical bytes written: 20
[Set 4][NS 4] - physical bytes read: 20
```

CLI plug in subcommand:
“vs-internal log”

Parameters		Output expectations
telemetry_type	**telemetry_data_area**	
none	None	- Default vendor-specific data capture
host	[1][2][3]	- Data area 1, 2, or 3 binary capture
controller	[1][2][3]	- Data area 1, 2, or 3 binary capture
Any other combination		- Returns “unsupported parameters entered” and a message on how to enter a valid combination
Note 1: If controller-initiated log page is selected, the Telemetry Controller-Initiated Data Generation Number (byte 383) must be included in the output		
Note 2: Reason identifier must be returned and set in the "vs-error-reason-identifier" nvme-CLI plug-in sub-command for all FW assert and controller-initiated captures		

CLI plug in subcommand:
“**vs-telemetry-controller-option**” **(default status is DISABLED)**

Parameter value	Output expectations
enable	- Enables controller-initiated log page and indicates if the operation is successful or not
disable	- Disables controller-initiated log page and indicates if the operation is successful or not
status	- Returns controller-initiated log page state as “disabled” or “enabled”

CLI plug in subcommand:
“vs-fw-activation-history”

Output:
20 entries of information. Oldest entry on top. When the drive is first shipped from the factory, there are no entries recorded.

Output columns:

Entry Number	Power on Hour	Power cycle count	Current firmware	New FW activated	Slot number	Commit Action Type	Result

Entry number – Increments every time a firmware activation is attempted no matter if the result is good or bad. Acts as a firmware activation counter.
Power on hour (POH) – Displays the POH of the SSD when the firmware activation happened. Accuracy needs to be down to the second at least.
Power cycle (PC) count – Display the power cycle count that the firmware activation occurred.
Current firmware – Displays the firmware currently running on the SSD before the firmware activation took place
New FW activated – Displays the activated firmware version that is running on the SSD after the firmware activation took place
Slot number – Displays the slot that the firmware is being activated from
Commit Action Type – Displays the Commit action type associated with the firmware activation event
Results – Records the results of the firmware activation event. The output shall follow the table below:

Results	Output
Pass	Pass
Fail	Failed + error code

Entry recording rules:

1. An entry must be recorded whenever a FW activation is taking place (does not matter if there's a reset or not). FW downloads do not generate an entry.
2. 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:
    a. POH is within 1 minute from the last RECORDED entry
    b. Power cycle count is the same
    c. Current firmware is the same
    d. New FW activated is the same
    e. Slot number is the same
    f. Commit Action Type is the same
    g. Results are the same

Examples:

FW Activation Examples:

Host-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
Get activation-history

NVMe-CLI Plugin Output:

Entry Number	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-events and initial states:

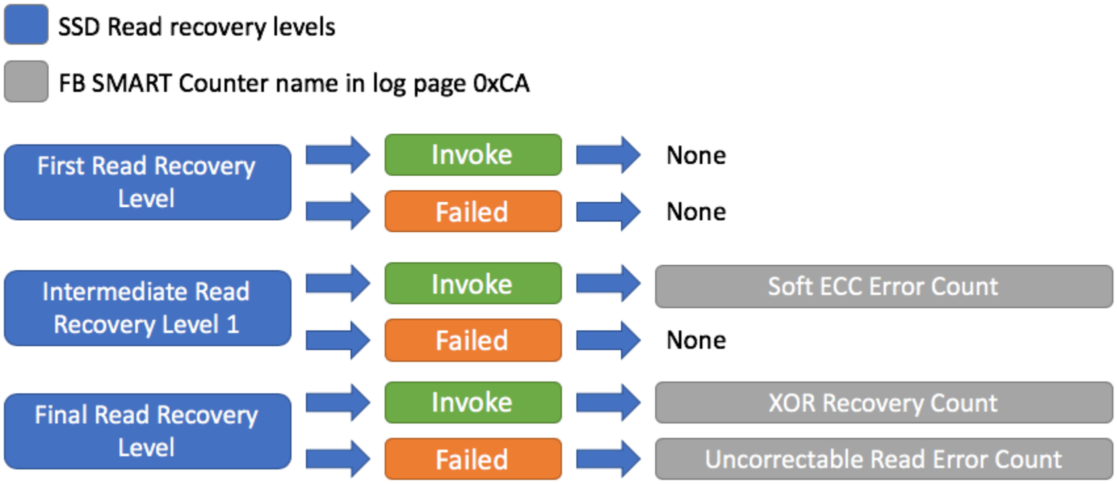
Initial State: Slot1=101
POH 1:00:01, PC 1, FW Commit CA=011b Slot=1 FW=102, pass
POH 1:00:10, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1
POH 1:00:30, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1 (not recorded)
POH 1:01:15, PC 1, FW Commit CA=0011b Slot=1 FW=102, fail reason #1 (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:

Entry Number	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 D: SSD Read Recovery Level Definition

The following section outlines the definition of various read-recovery levels within a SSD and how they align to the FB SMART counters in Section 4.9.3.


SSD Read recovery levels
FB SMART Counter name in log page 0xCA
First Read Recovery Level
Invoke
None
Failed
None
Intermediate Read Recovery Level 1
Invoke
Soft ECC Error Count
Failed
None
Final Read Recovery Level
Invoke
XOR Recovery Count
Failed
Uncorrectable Read Error Count

Appendix E: NVM Set Design Guidance

Functional requirements for Drives with NVM Sets Only:

Requirement ID	Description
NVMe-10A	Facebook requires that the SSD support NVM Sets.
NVMe-10B	These NVM Sets characteristics are to be configured at the SSD manufacturer.
NVMe-10C	Each Set shall have a capacity of 1TB (before over-provisioning)
NVMe-10D	Each Set shall have a single namespace Multiple namespace per set support is NOT required. Each Set and the associated NVMe namespace shall have the same ID number. For example, Set 1 is associated with Namespace 1. Commands that do not follow this association shall be aborted.
NVMe-10E	Each Set shall be an independent group of NAND die which are not shared across Sets. One endurance group per Set and wear leveling is performed only within the endurance group and not across groups.
NVMe-10F	Each Set shall implement the Endurance Group log page.
NVMe-10G	Each Set shall have one or more dedicated NAND channels which are not shared across Sets
NVMe-10H	Each Set shall have independent buffers to minimize contention or interference
NVMe-10I	Each Set shall support the UBER defined in section 3.1.
NVMe-10J	Each Set shall support the error protection requirements as defined in section 5.5.
NVMe-10K	Each Set shall use the smallest possible over-provisioning needed to maintain requirement of NVMe-10I above
NVMe-10L	The number of Sets shall scale as the capacity scales without increasing the Set capacity (i.e. a 4TB SSD has 4x 1TB Sets and an 8TB SSD has 8x 1TB Sets)
NVMe-10M	Each Set shall exhibit the following isolation characteristics: 1. I/O and Data Set Management De-allocate (TRIM) to 1 NVM Set shall not affect any other NVM Set on the same SSD by more than 15us for P99.9999 latencies. 2. The following admin commands shall not affect any other NVM Set on the same SSD by more than 15us for P99.9999 latencies a. Format (SES=0, 1, or 2) b. Namespace management c. Namespace Attachment d. Get Log Page (i.e. SMART log 0x02 or 0xCA) e. Set Feature (identifier 0x14)

	3. All other Admin commands such as reset, or FW activate are expected to impact the entire SSD and therefore all Sets.
NVMe-10N	Read Recovery Levels do not need to be supported at this time (potentially in the future).
NVMe-10O	All Sets shall have the following minimum performance characteristics assuming all Sets are active and 1TB in capacity (performance targets are per Set): 1. 75k 4k random read IOPs at queue depth 32 or higher 2. 250MB/s of sequential write bandwidth
NVMe-10P	Wear leveling across NVM Sets is Facebook's responsibility, but if the endurance of one NVM Set is exceeded for any reason it shall not affect the operation of other NVM Sets within that SSD (e.g. the entire drive shall not enter read-only mode). Wear leveling within a set is the drive providers responsibility
NVMe-10Q	TPAR 4050 Endurance Log Enhancements. This is not approved yet by NVMe, but it is expected to be approved and will need to be supported. This essentially adds logs that are per endurance group and also AER support for the endurance logs.
NVMe-10R	TPAR 4045 SQ/Sets granularity. This is not approved yet by NVMe, but it is expected to be approved and will need to be supported. This allows the drive to associated SQs on a per set basis.
NVMe-10S	The order of the Sets and the Set's associated NVMe namespace shall not change under any condition. For example, if namespace 2 (associated with Set 2) is deleted, the FW needs to preserve the association of namespace 3 with Set 3 even after a reset or a power cycle. The FW is allowed to recreate namespace 2 before the reset or power cycle in this scenario to preserve the proper ordering.
NVMe-10T	The NVMe CLI plugin subcommand "vs-smart-add-log" output needs to be modified to support the additional Facebook log page 0xCA data generated from an NVM Sets enabled SSD. Please refer to Appendix C: NVMe-CLI Plugin Output for the syntax.

SMART Log Page (02h) requirements for Drives with NVM Sets Only

The following table defines the logic for the attributes in the standard smart log page for SSDs with NVM Sets. Drives without NVM Sets shall follow the wording in the NVMe specification version in Section 4.1.

Requirement ID	Standard SMART Log page (controller-level)	Definition of the Attributes
NVMe-16A	critical warning – available spares	Trigger only if the value falls below threshold for any of the endurance group/namespace

NVMe-16B	critical warning – temperature threshold	Trigger if the composite temperature goes above this threshold
NVMe-16C	critical warning – subsystem reliability	Trigger only if any of the namespace/endurance group has lost subsystem reliability or if a critical threshold is reached where the reliability is lost at the controller level
NVMe-16D	critical warning – read-only	Trigger only if any of the namespace/endurance group is in read-only
NVMe-16E	critical warning – volatile memory backup	Trigger if the PLP solution failed
NVMe-16F	available spares value	Report as out of total remaining spares for the controller
NVMe-16G	available spares threshold	Vendor Specific
NVMe-16H	percentage used	Report as out of total percentage used for the controller
NVMe-16I	data units read	Report the value for the controller
NVMe-16J	data units written	Report the value for the controller

Facebook Vendor-specific SMART Log Page Requirements for Drives with NVM Sets Only

The following table details additional fields that need to be implemented in vendor specific log pages for drives with NVM Sets.

Requirement ID	Description
NVMe-18C	The attributes in the 0xCA log page shall report a value at the drive level and a value for each of the NVM Sets as applicable, see table below. When reporting the value per set it must be clear which endurance group the data is for in the NVMe-CLI plugin output. See Appendix C: NVMe-CLI Plugin Output.

Req ID	**Field**	**# of Bytes**	**Drive Value**	**NVM Set Value (Drives with Sets only)**	**Field description**
NVMe-18D	Log page directory	Vendor defined	Sum of all Sets	Yes, Per Set	Provides a list of available log pages and corresponding log identifiers
NVMe-18E	Physical (NAND) bytes written	16	Sum of all Sets	Yes, Per Set	The number of bytes written to NAND. It must be possible to use this attribute in conjunction with another attribute to calculate the Write Amplification Factor (WAF). Any formulas required shall be provided.
NVMe-18F	Physical (NAND) bytes read	16	Sum of all Sets	Yes, Per Set	The number of bytes read from NAND.

NVMe -18G	Bad NAND block count	8 (2 bytes for normalized + 6 bytes for raw count)	Sum of all Sets	Yes, Per Set	Raw and normalized count of the number of NAND blocks that have been retired after the drive's manufacturing tests (i.e. grown bad blocks)
NVMe -18H	XOR Recovery count	8	Sum of all Sets	Yes, Per Set	Total number of times XOR was invoked to recover data. Data recovery may have succeeded or failed. See Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18I	Uncorrectable read error count	8	Sum of all Sets	Yes, Per Set	Total count of NAND reads that were not correctable by read retries, all levels of ECC, or XOR (as applicable). Data recovery fails, and an uncorrectable read error is returned to the host. See Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18J	Soft ECC error count	8	Sum of all Sets	Yes, Per Set	Total count of NAND reads that were not correctable by first-level ECC. Data is recovered by an intermediate recovery mechanism and returned correctly to the host. See Appendix D: SSD Read Recovery Level Definition for more details.
NVMe -18K	SSD End to end correction counts	8 (4 bytes for count of detected errors, 4 bytes for count of corrected errors)	Drive level count	No	A count of the detected and corrected errors by the SSD end to end error correction which includes DRAM, SRAM, or other storage element ECC/CRC protection mechanism (not NAND ECC). All correctable errors must result in a counter increase no matter what type of data the memory is protecting. All detected errors must result in a counter increase unless the error is uncorrectable and occurred in the system region. In the latter case, the incomplete shutdown flag must be flagged/incremented on the next power up.
NVMe -18L	System data % used	1	Drive level count	No	A normalized cumulative count of the number of erase cycles per block since leaving the factory for the system (FW and metadata) area. Starts at 0 and increments. 100 indicates that the estimated endurance has been consumed. Value is allowed to exceed 100 up to 255.

NVMe -18M	User data erase counts	8 (4 bytes for the maximum, 4 bytes for the minimum)	Sum of all Sets	Yes, Per Set	The maximum and minimum erase counts across all NAND blocks in the drive. The host shall not be able to reset this counter.
NVMe -18N	Refresh count	8	Sum of all Sets	Yes, Per Set	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.
NVMe -18O	Program fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Sum of all Sets	Yes, Per Set	Raw and normalized count of total program failures. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18P	User data erase fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Sum of all Sets	Yes, Per Set	Raw and normalized count of total erase failures in the user area. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18Q	System area erase fail count	8 (2 bytes for normalized + 6 bytes for raw count)	Drive level count	No	Raw and normalized count of total erase failures in the system area. Normalized count starts at 100 and shows the percent of remaining allowable failures.
NVMe -18R	Thermal throttling status and count	2 (1 byte for the current status, 1 byte for the count)	Drive level count	No	The current status of thermal throttling (enabled or disabled) and a count of the number of thermal throttling events.
NVMe -18S	PCIe Correctable Error count	8	Drive level count	No	Summation counter of all PCIe correctable errors (Bad TLP, Bad DLLP, Receiver error, Replay timeouts, Replay rollovers)
NVMe -18T	Incomplete shutdowns	4	Drive level count	No	A count of the number of shutdowns that have occurred that did not complete properly
NVMe -18U	% Free Blocks	1	Sum of all Sets	Yes, Per Set	A normalized count of the number of blocks that are currently free (available) out of the total pool of spare (invalid) blocks. Free blocks mean both blocks that have been erased and blocks that have all invalid data. Invalid blocks are blocks that are either marked invalid by drive FW OR by the host (via TRIM or overwrite). For example, if the total number of spare blocks is 100 and garbage collection has

					been able to reclaim 20 blocks, then this field reports 20%.

NVMe CLI Utility:
Identify NVM Set and endurance log page command must be supported.

Telemetry:

NVMe-12D	If any of the following list of conditions occur, the telemetry data must be committed to non-volatile storage so that the data is saved: 8. The following critical warnings in the endurance log page for any of the NVM Sets changes to a non-zero value. • Available spares • Sub-system reliability • Read-only mode

Power:
The power limit in the spec must be obeyed when all Sets are active