Architected for Performance

NVMe™ SSD Management, Error Reporting and Logging Capabilities

Sponsored by NVM Express

June 30, 2020

Speakers

Jonmichael Hands

Sr. Strategic Planner & Product Manager

intel

Rohit Gupta

Segment Marketing

Western Digital.

Bill Martin

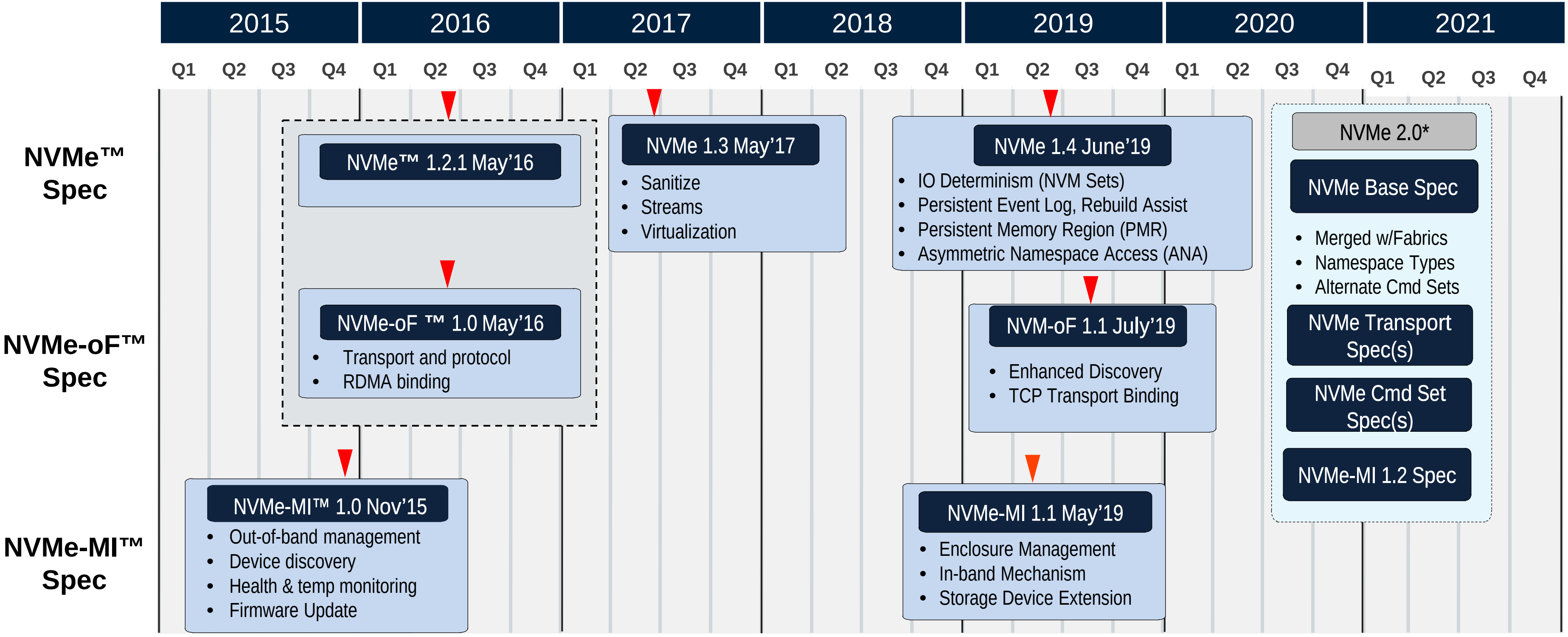
SSD IO Standards

SAMSUNG

NVMe™ Technology Features for Errors, Logging and Health Monitoring

Jonmichael Hands, Sr. Strategic Planner & Product Manager, Intel SSDs,
Co-Chair NVMe Marketing WG

NVM Express Technology Specification Roadmap



2015 2016 2017 2018 2019 2020 2021

Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4

NVMe™ Spec

NVMe™ 1.2.1 May'16

NVMe 1.3 May'17

- Sanitize
- Streams
- Virtualization

NVMe 1.4 June'19

- IO Determinism (NVM Sets)
- Persistent Event Log, Rebuild Assist
- Persistent Memory Region (PMR)
- Asymmetric Namespace Access (ANA)

NVMe 2.0*

NVMe Base Spec

- Merged w/Fabrics
- Namespace Types
- Alternate Cmd Sets

NVMe Transport Spec(s)

NVMe Cmd Set Spec(s)

NVMe-MI 1.2 Spec

NVMe-oF™ Spec

NVMe-oF ™ 1.0 May'16

- Transport and protocol
- RDMA binding

NVM-oF 1.1 July'19

- Enhanced Discovery
- TCP Transport Binding

NVMe-MI™ Spec

NVMe-MI™ 1.0 Nov'15

- Out-of-band management
- Device discovery
- Health & temp monitoring
- Firmware Update

NVMe-MI 1.1 May'19

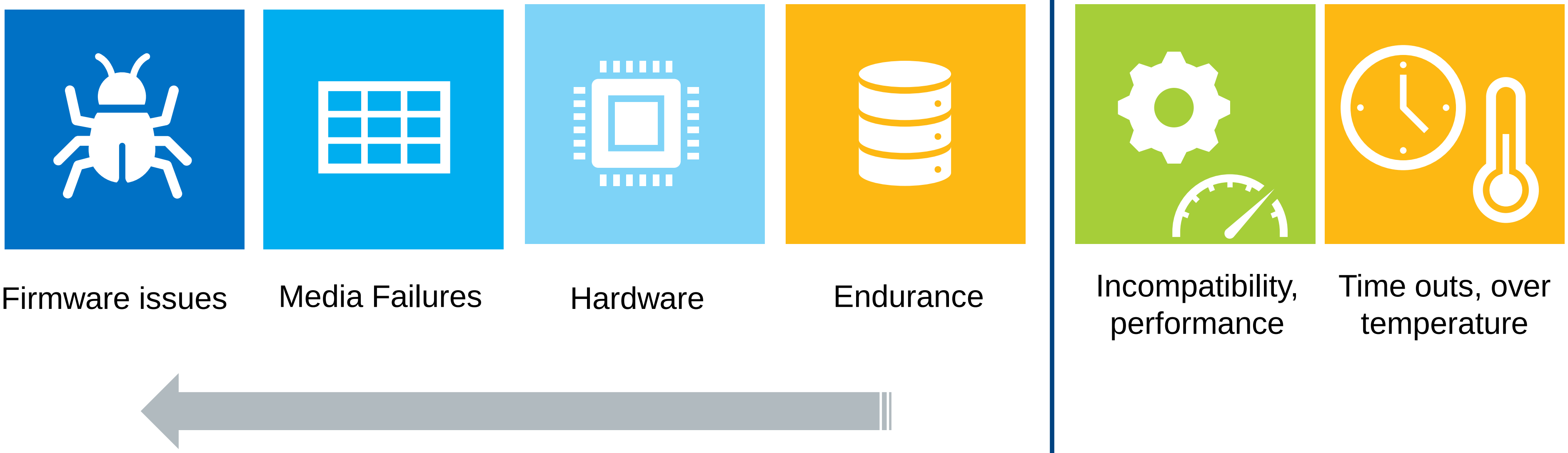
- Enclosure Management
- In-band Mechanism
- Storage Device Extension

▭ Released NVMe specification ⬚ Planned release

How Do SSDs Fail?

Failures

Returns



Firmware issues

Media Failures

Hardware

Endurance

Incompatibility, performance

Time outs, over temperature

Increasing prevalence

Case Studies

Minimizing Customer Interruptions Due to SSD Failures, Brennan Watt, Microsoft

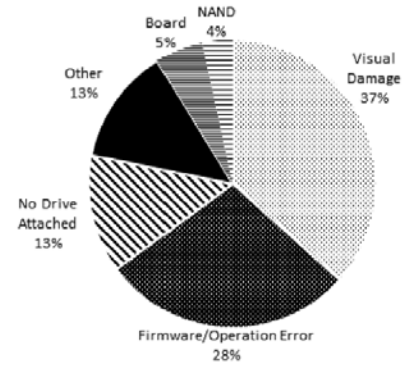
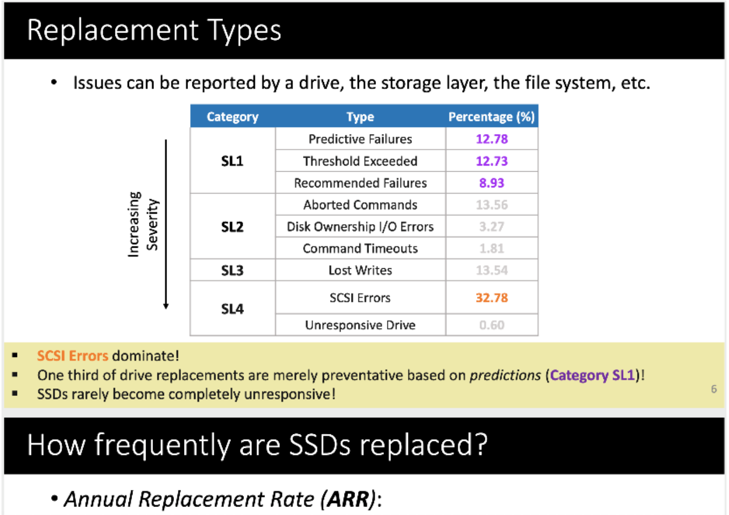


Fig. 20. ***Breakdown of field failures for the S3500. Visual damage***

Reliability of Solid-State Drives Based on NAND Flash Memory, 2017

A Study of SSD Reliability in Large Scale Enterprise Storage Deployments
https://www.usenix.org/conference/fast20/presentation/maneas



Replacement Types

- Issues can be reported by a drive, the storage layer, the file system, etc.

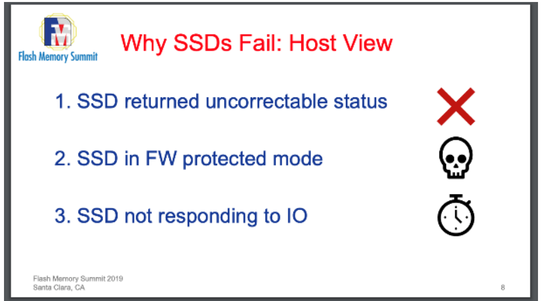
Increasing Severity

Category	Type	Percentage (%)
SL1	Predictive Failures	12.78
	Threshold Exceeded	12.73
	Recommended Failures	8.93
SL2	Aborted Commands	13.56
	Disk Ownership I/O Errors	3.27
	Command Timeouts	1.81
SL3	Lost Writes	13.54
SL4	SCSI Errors	32.78
	Unresponsive Drive	0.60

- SCSI Errors dominate!
- One third of drive replacements are merely preventative based on *predictions* (**Category SL1**)!
- SSDs rarely become completely unresponsive!

How frequently are SSDs replaced?

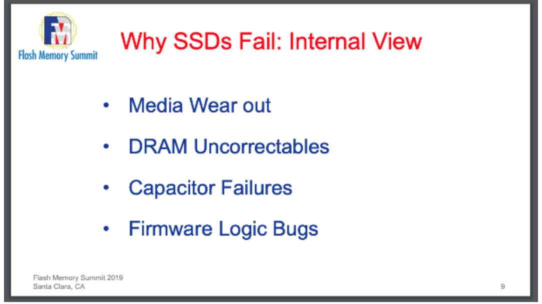
- *Annual Replacement Rate (**ARR**)*:



Flash Memory Summit

Why SSDs Fail: Host View

1. SSD returned uncorrectable status ✗
2. SSD in FW protected mode ☠
3. SSD not responding to IO ⏱



Flash Memory Summit

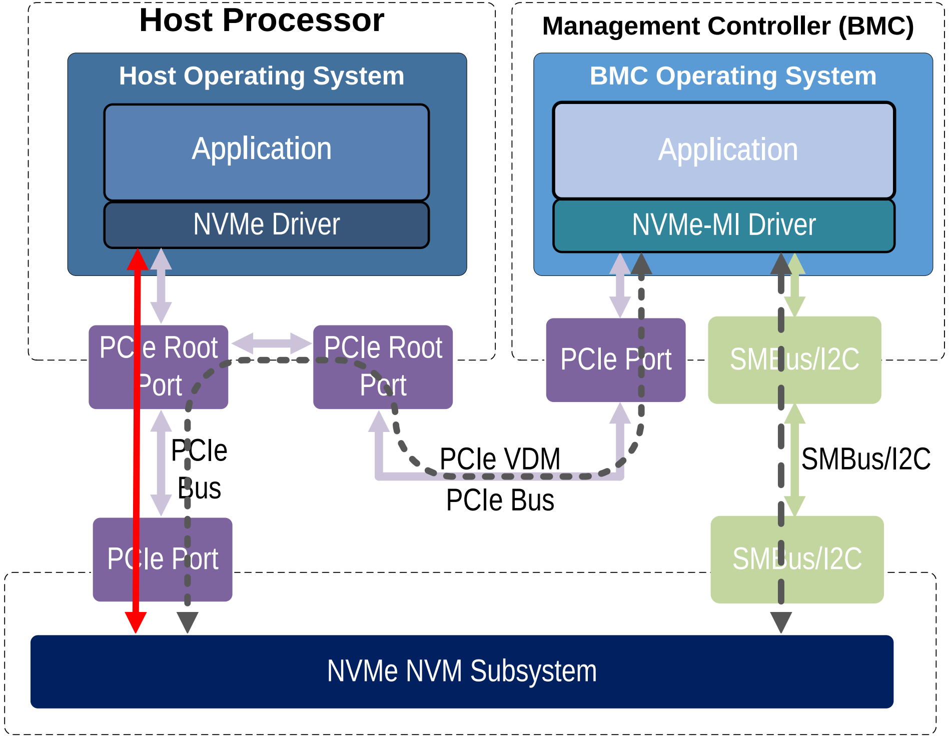
Why SSDs Fail: Internal View

- Media Wear out
- DRAM Uncorrectables
- Capacitor Failures
- Firmware Logic Bugs

NVMe™ Features for Errors, Logging and Health Monitoring

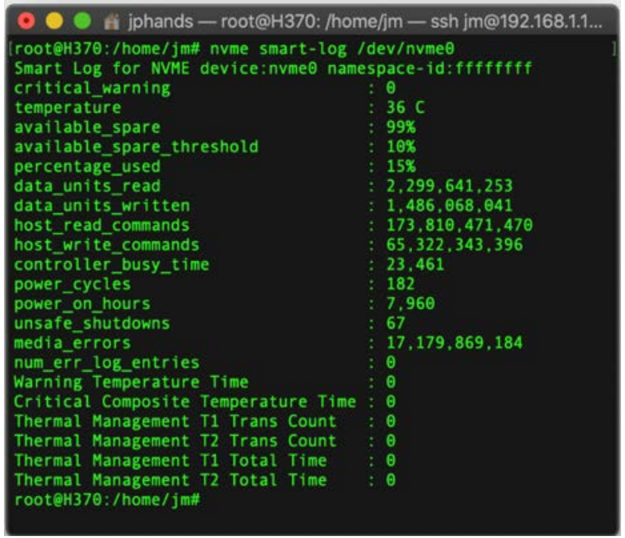
Feature	Description	Use case
SMART Log Page / Critical Warning	The SMART log page is used to report on general health information about the drive. Its main health indicator is called the **critical warning**	Main health monitoring dashboard
Error Log Page	This log page maintains important information regarding the number of errors, which queue they came from, and which data and namespaces were affected	Main error dashboard
Persistent Event Log	human readable & timestamped log of events occurring on the SSD such as errors, updating firmware, format, etc.	Human readable log, SSD “black box” recorder
Telemetry	Telemetry enables manufacturers to collect internal data logs to improve the functionality and reliability of products	Triage of field failures, periodic health monitoring, root cause firmware bugs
Asynchronous event support	Asynchronous events are used to notify host software of status, error, and health information as these events occur.	Operating system to get notified of events
Device Self-Test	diagnostic testing sequence that tests the integrity and functionality of the controller and may include testing of the media associated with namespaces	Factory integration, testing
End-to-end data protection (PI)	To provide robust data protection from the application to the NVM media and back to the application itself	Protect against data corruption from host to device

NVMe™ Management Interface (NVMe-MI™) 1.1 Specification



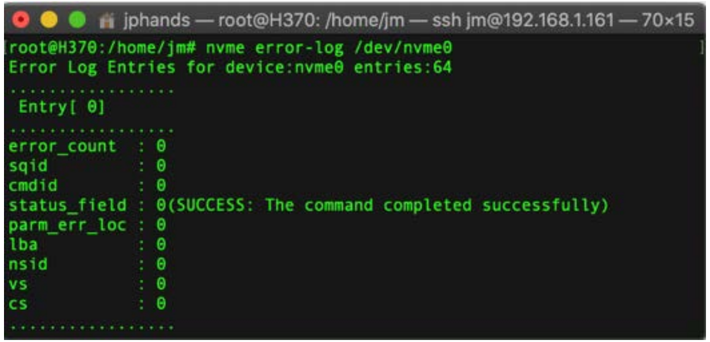
- **Out-of-Band Management** – Management that operates with hardware resources and components that are ***independent of the host operating system control***
- **NVMe™ Out-of-Band Management Interfaces:** SMBus/I2C, PCIe Vendor Defined Messages (VDM)
- In-band mechanism allows application to tunnel NVMe-MI™ commands through NVMe driver
- Benefits: Provides management capabilities not available in-band via NVMe commands
  - Efficient NVM Subsystem health status reporting
  - Ability to manage NVMe at a FRU level
  - Vital Product Data (VPD) access
  - Enclosure management

NVMe™ SMART Log, Error Log



```
root@H370:/home/jm# nvme smart-log /dev/nvme0
Smart Log for NVME device:nvme0 namespace-id:ffffffff
critical_warning                    : 0
temperature                         : 36 C
available_spare                     : 99%
available_spare_threshold           : 10%
percentage_used                     : 15%
data_units_read                     : 2,299,641,253
data_units_written                  : 1,486,068,041
host_read_commands                  : 173,810,471,470
host_write_commands                 : 65,322,343,396
controller_busy_time                : 23,461
power_cycles                        : 182
power_on_hours                      : 7,960
unsafe_shutdowns                    : 67
media_errors                        : 17,179,869,184
num_err_log_entries                 : 0
Warning Temperature Time            : 0
Critical Composite Temperature Time : 0
Thermal Management T1 Trans Count   : 0
Thermal Management T2 Trans Count   : 0
Thermal Management T1 Total Time    : 0
Thermal Management T2 Total Time    : 0
root@H370:/home/jm#
```

SMART log critical warning is main indicator


jphands — root@H370: /home/jm — ssh jm@192.168.1.161 — 70×15

```
root@H370:/home/jm# nvme error-log /dev/nvme0
Error Log Entries for device:nvme0 entries:64
.................
 Entry[ 0]
.................
error_count  : 0
sqid         : 0
cmdid        : 0
status_field : 0(SUCCESS: The command completed successfully)
parm_err_loc : 0
lba          : 0
nsid         : 0
vs           : 0
cs           : 0
.................
```

Errors are logged here

Telemetry

- NVMe™ 1.3 specification defines Telemetry with two new log pages:
  - Host Initiated Telemetry Log (log page identifier 0x07)
  - Controller Initiated Telemetry Log (log page identifier 0x08)
- The NVMe 1.3 Telemetry specification defines that the Log Page return data contains:
  - Standard header as specified
  - Data requested must be multiple of 512 Bytes
  - Up to three consecutive data areas

Host Initiated Telemetry Log

Controller Initiated Telemetry Log

The Telemetry log can consist of 3 data areas:

Data Area 1:
small size, designed for operational periodic data pulls (health monitoring, performance) during operation, contains critical drive data

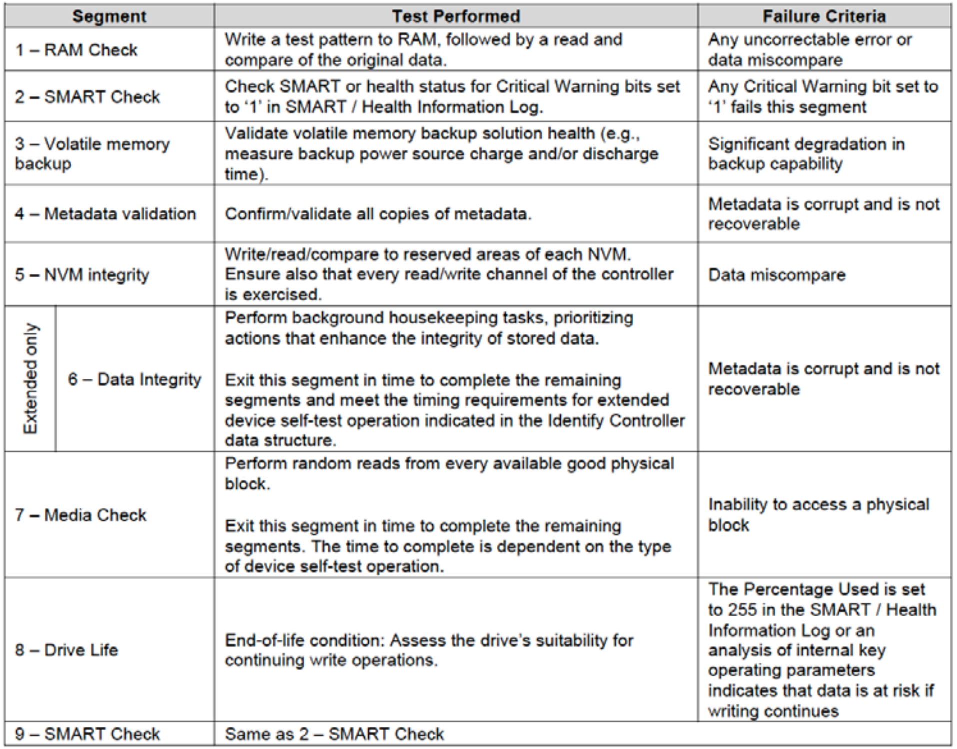
Data Area 2:
medium, scale up for additional content

Data Area 3:
large, designed to be comprehensive for failure triage and root cause analysis

Device Self-Test Operation

- Offline diagnostic test, often done at factory or system integrator to ensure SSD working properly
- Short test – 2 min or less
- An extended device self-test operation persist across reset
- Both can be interrupted by format, sanitize, or another self-test command

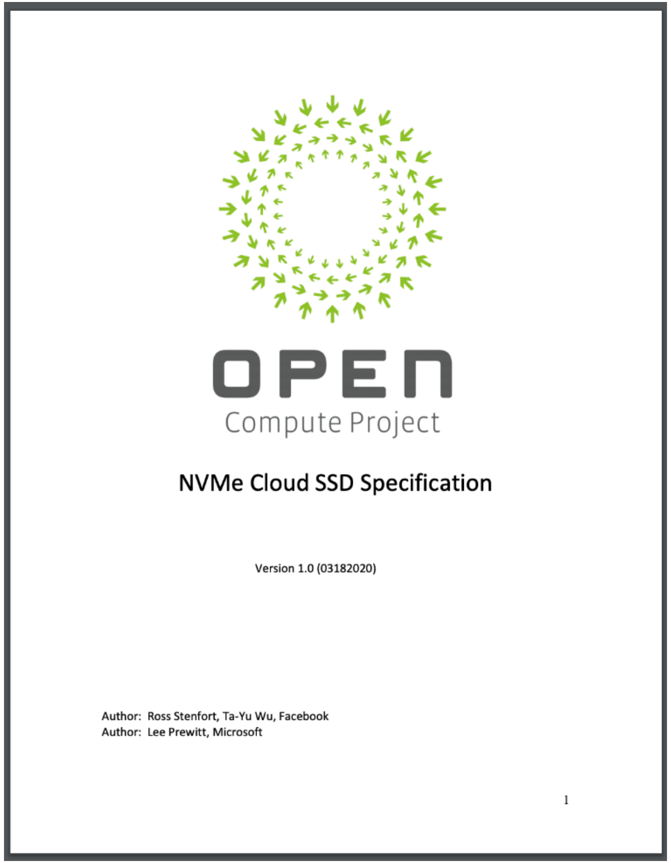
Figure 476: Example Device Self-test Operation (Informative)



Segment	Test Performed	Failure Criteria
1 – RAM Check	Write a test pattern to RAM, followed by a read and compare of the original data.	Any uncorrectable error or data miscompare
2 – SMART Check	Check SMART or health status for Critical Warning bits set to '1' in SMART / Health Information Log.	Any Critical Warning bit set to '1' fails this segment
3 – Volatile memory backup	Validate volatile memory backup solution health (e.g., measure backup power source charge and/or discharge time).	Significant degradation in backup capability
4 – Metadata validation	Confirm/validate all copies of metadata.	Metadata is corrupt and is not recoverable
5 – NVM integrity	Write/read/compare to reserved areas of each NVM. Ensure also that every read/write channel of the controller is exercised.	Data miscompare
Extended only 6 – Data Integrity	Perform background housekeeping tasks, prioritizing actions that enhance the integrity of stored data. Exit this segment in time to complete the remaining segments and meet the timing requirements for extended device self-test operation indicated in the Identify Controller data structure.	Metadata is corrupt and is not recoverable
7 – Media Check	Perform random reads from every available good physical block. Exit this segment in time to complete the remaining segments. The time to complete is dependent on the type of device self-test operation.	Inability to access a physical block
8 – Drive Life	End-of-life condition: Assess the drive's suitability for continuing write operations.	The Percentage Used is set to 255 in the SMART / Health Information Log or an analysis of internal key operating parameters indicates that data is at risk if writing continues
9 – SMART Check	Same as 2 – SMART Check	

OCP Cloud NVMe™ SSD Spec

- NVM Express™ Specification Features
  - Vendor unique log pages for cloud SSDs
- PCI Express® Specification Features
- SMART Log Requirements
- Thermal Requirements
- Quality Requirements
- Power Requirements
- SMBUS data layout
- Security Requirements
- Form Factor Requirements
- Open source tool access requirements



OPEN
Compute Project

NVMe Cloud SSD Specification

Version 1.0 (03182020)

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

1

OCP Cloud NVMe™ SSD Specification

SMART Cloud Attributes Log Page, C0

- Physical media units read/written (to calculate WAF)
- Bad user and system NAND blocks
- XOR recoveries
- Uncorrectable error count
- Soft ECC errors
- End-to-end correction counts
- System data % used
- Refresh counts
- User data erase counts
- Thermal throttling status and counts
- PCIe correctable errors
- Incomplete shutdowns
- % free blocks
- Capacitor health
- Unaligned IO
- Security version
- PLP status
- Endurance estimate

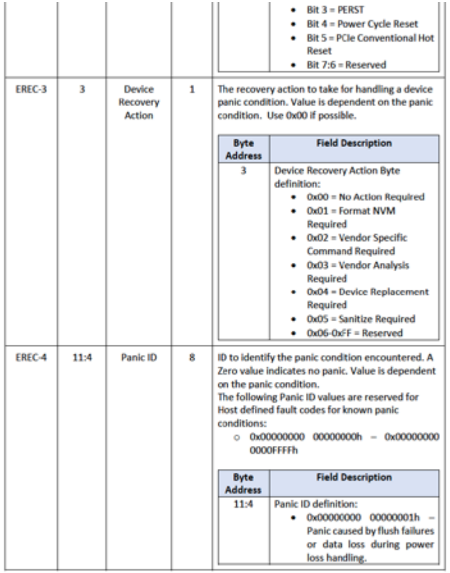
C0 log page allows for deeper predictive analytics and health monitoring

OCP Cloud NVMe™ SSD Specification

Error Recovery Log Page, C1

- Panic Reset Wait Time
- Panic Reset Action
- Device Recovery Action
- Panic ID
- Device Capabilities
- Vendor Specific Recover opcode

https://www.opencompute.org/documents/nvme-cloud-ssd-specification-v1-0-3-pdf



				• Bit 3 = PERST • Bit 4 = Power Cycle Reset • Bit 5 = PCIe Conventional Hot Reset • Bit 7:6 = Reserved
EREC-3	3	Device Recovery Action	1	The recovery action to take for handling a device panic condition. Value is dependent on the panic condition. Use 0x00 if possible. Byte Address: 3 — Field Description: 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
EREC-4	11:4	Panic ID	8	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 Byte Address: 11:4 — Field Description: Panic ID definition: • 0x00000000 00000001h – Panic caused by flush failures or data loss during power loss handling.

NVMe™ 1.4 Specification Features

Category	Feature	Benefit
Hyperscale performance	NVM Sets	Improved multi tenant quality of service through physical isolation / separation
	Read Recovery Levels	Improved read latency with host to drive tradeoff on UBER
	IO Determinism	Read only like latencies for mixed read/write workloads
	Multi-Host Shared Write Streams	Improve SSD endurance by tagging data into streams, new use cases on dealing with data from multiple hosts
New Use Cases	Persistent Memory Region	Multi purpose persistent memory for innovative use cases
Manageability / Triage	Administrative Controller	Splits NVMe™ controller up into administrative, I/O, and discovery controllers. Admin controller used for enclosure management.
	Persistent Event Log	SSD keeps log of events that host (e.g. OS) can read
NVMe-oF™ Spec	Multipathing and Namespace Sharing (ANA)	Discover optimal path to namespace
Data integrity, configurations	Rebuild Assist	Drive can discover unrecoverable data and ask host to rebuild from other copies
	Enhanced Command Retry	Host configurable retry status for commands with time delay
	Namespace Granularity	Create namespace size that is optimal for the SSD media layout
	Verify	Verify data integrity on drive without sending data to host
	Namespace write protect	Lockdown namespace for read only and boot use cases

Persistent Event Log

First version (TP 4007)	Second version	Future work
SMART / Health Log Snapshot	Subsystem hardware error	Power Excursion
Firmware Commit Event	Set Feature	Voltage Excursion
Timestamp Change	Format	Rebuild assist notification
Power-On or Reset	Sanitize	NVMe-MI™ failures
Vendor Specific	Namespace Create/Delete	IO Determinism
	TCG	Performance stats
	Temperature Excursion	

The log is intended to persistently capture significant events for use by software/system vendors that are not the NVMe™ subsystem manufacturer such as operating systems, management software, storage system vendors, etc.

Admin Command Set and Persistent Event Log

Rohit Gupta, Segment Marketing, Western Digital

NVMe™ 1.0 Specification Admin Command Set

Specifications	Transports	Commands	
NVM Express Specification nvm EXPRESS®	PCI EXPRESS®	**Admin Commands** Create IO Submission Queue Create IO Completion Queue Delete IO Submission Queue Delete IO Completion Queue Abort Command Asynchronous Event Requests Get Log Page Identify Get Feature Set Feature Firmware Download Firmware Activate Format NVM Security Send Security Receive	**NVM Commands** Flush Read Write Compare Write Uncorrectable Dataset Management

NVMe™ 1.4 Specification Admin Command Set

Specifications	Transports	Commands
NVM Express Specification NVMe™ over Fabrics NVMe- MI™ Specification	PCI EXPRESS RDMA Remote Direct Memory Access FIBRE CHANNEL	Admin Commands IO Commands

Admin Commands	
Identify	Directive Send
Firmware Download	Directive Receive
Firmware Commit	Get LBA Status
Security Send	Namespace Management
Security Receive	Namespace Attach
Log Page	Sanitize
Format NVM	Virtualization Management
Get Feature	Device Self Test
Set Feature	Fabrics
MI Send	Keep Alive
MI Receive	

IO Commands
Flush
Read
Write
Compare
Write Uncorrectable
Dataset Management
Write Zeroes
Verify
Reservation Register
Reservation Acquire
Reservation Release
Reservation Report

NVMe™ 1.4 Specification Admin Command Set

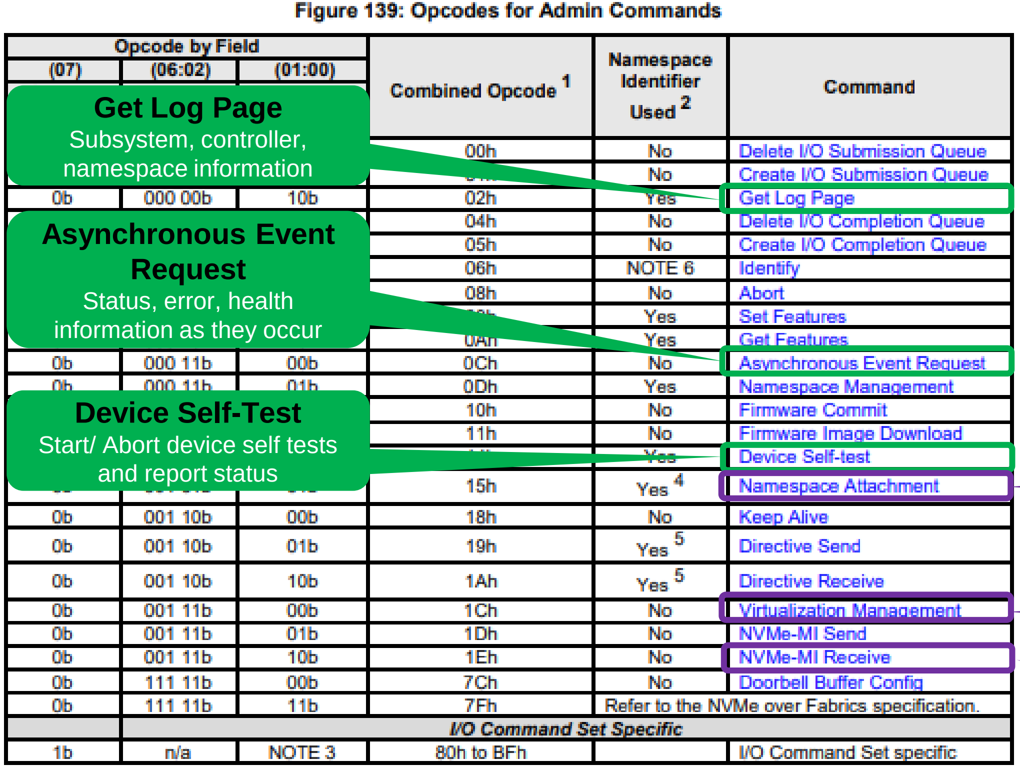


Figure 139: Opcodes for Admin Commands

Opcode by Field (07)	Opcode by Field (06:02)	Opcode by Field (01:00)	Combined Opcode [1]	Namespace Identifier Used [2]	Command
[illegible]	[illegible]	[illegible]	00h	No	Delete I/O Submission Queue
[illegible]	[illegible]	[illegible]	[illegible]	No	Create I/O Submission Queue
0b	000 00b	10b	02h	Yes	Get Log Page
[illegible]	[illegible]	[illegible]	04h	No	Delete I/O Completion Queue
[illegible]	[illegible]	[illegible]	05h	No	Create I/O Completion Queue
[illegible]	[illegible]	[illegible]	06h	NOTE 6	Identify
[illegible]	[illegible]	[illegible]	08h	No	Abort
[illegible]	[illegible]	[illegible]	[illegible]	Yes	Set Features
[illegible]	[illegible]	[illegible]	0Ah	Yes	Get Features
0b	000 11b	00b	0Ch	No	Asynchronous Event Request
0b	000 11b	01b	0Dh	Yes	Namespace Management
[illegible]	[illegible]	[illegible]	10h	No	Firmware Commit
[illegible]	[illegible]	[illegible]	11h	No	Firmware Image Download
[illegible]	[illegible]	[illegible]	[illegible]	Yes	Device Self-test
[illegible]	[illegible]	[illegible]	15h	Yes [4]	Namespace Attachment
0b	001 10b	00b	18h	No	Keep Alive
0b	001 10b	01b	19h	Yes [5]	Directive Send
0b	001 10b	10b	1Ah	Yes [5]	Directive Receive
0b	001 11b	00b	1Ch	No	Virtualization Management
0b	001 11b	01b	1Dh	No	NVMe-MI Send
0b	001 11b	10b	1Eh	No	NVMe-MI Receive
0b	111 11b	00b	7Ch	No	Doorbell Buffer Config
0b	111 11b	11b	7Fh	Refer to the NVMe over Fabrics specification.	
	I/O Command Set Specific				
1b	n/a	NOTE 3	80h to BFh		I/O Command Set specific

Get Log Page
Subsystem, controller, namespace information

Asynchronous Event Request
Status, error, health information as they occur

Device Self-Test
Start/ Abort device self tests and report status

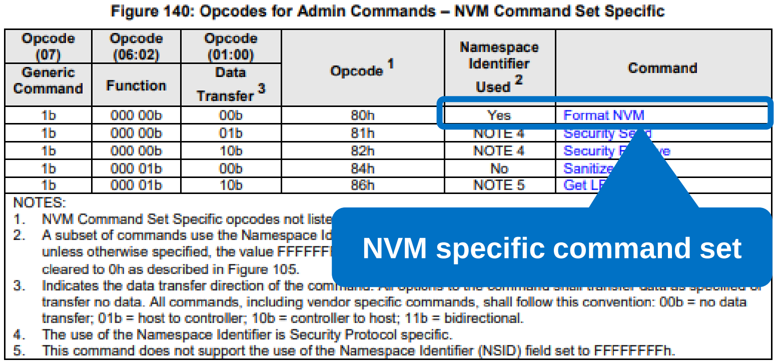


Figure 140: Opcodes for Admin Commands – NVM Command Set Specific

Opcode (07) Generic Command	Opcode (06:02) Function	Opcode (01:00) Data Transfer [3]	Opcode [1]	Namespace Identifier Used [2]	Command
1b	000 00b	00b	80h	Yes	Format NVM
1b	000 00b	01b	81h	NOTE 4	Security Se[illegible]d
1b	000 00b	10b	82h	NOTE 4	Security R[illegible]ve
1b	000 01b	00b	84h	No	Sanitize
1b	000 01b	10b	86h	NOTE 5	Get L[illegible]

NOTES:
1. NVM Command Set Specific opcodes not liste[illegible]
2. A subset of commands use the Namespace Id[illegible] unless otherwise specified, the value FFFFFF[illegible] cleared to 0h as described in Figure 105.
3. Indicates the data transfer direction of the com[illegible] transfer no data. All commands, including vendor specific commands, shall follow this convention: 00b = no data transfer; 01b = host to controller; 10b = controller to host; 11b = bidirectional.
4. The use of the Namespace Identifier is Security Protocol specific.
5. This command does not support the use of the Namespace Identifier (NSID) field set to FFFFFFFFh.

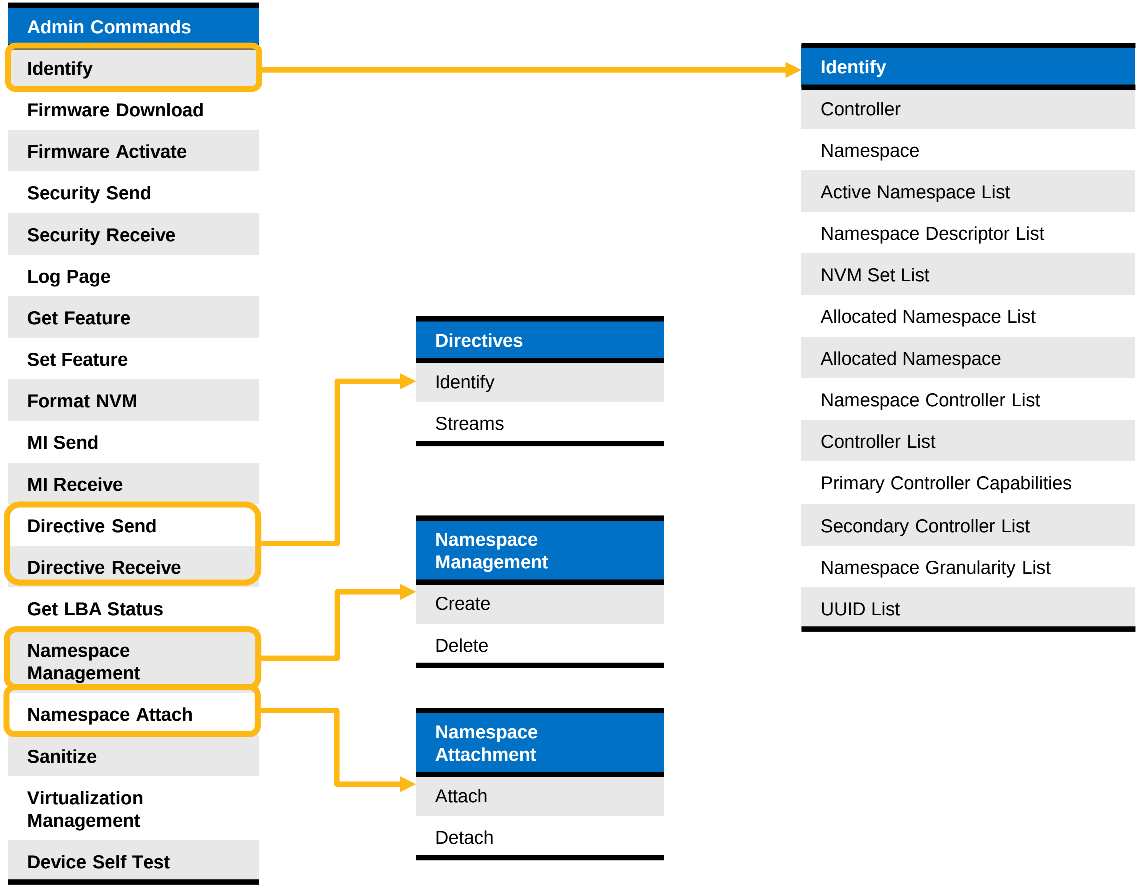
NVM specific command set

Namespace Attachment
Attach/ detach, manage controllers w/ namespace

Virtualization Management
To support virtualization enhancement capabilities

NVMe-MI Receive
In-Band tunneling message service model

NVMe™ 1.4 Specification Admin Sub-Commands



Admin Commands
Identify
Firmware Download
Firmware Activate
Security Send
Security Receive
Log Page
Get Feature
Set Feature
Format NVM
MI Send
MI Receive
Directive Send
Directive Receive
Get LBA Status
Namespace Management
Namespace Attach
Sanitize
Virtualization Management
Device Self Test

Directives
Identify
Streams

Namespace Management
Create
Delete

Namespace Attachment
Attach
Detach

Identify
Controller
Namespace
Active Namespace List
Namespace Descriptor List
NVM Set List
Allocated Namespace List
Allocated Namespace
Namespace Controller List
Controller List
Primary Controller Capabilities
Secondary Controller List
Namespace Granularity List
UUID List

NVMe™ 1.4 Specification Admin Sub-Commands: Get/Set Feature

Admin Commands
Identify
Firmware Download
Firmware Activate
Security Send
Security Receive
Log Page
Get Feature
Set Feature
Format NVM
MI Send
MI Receive
Directive Send
Directive Receive
Get LBA Status
Namespace Management
Namespace Attach
Sanitize
Virtualization Management
Device Self Test

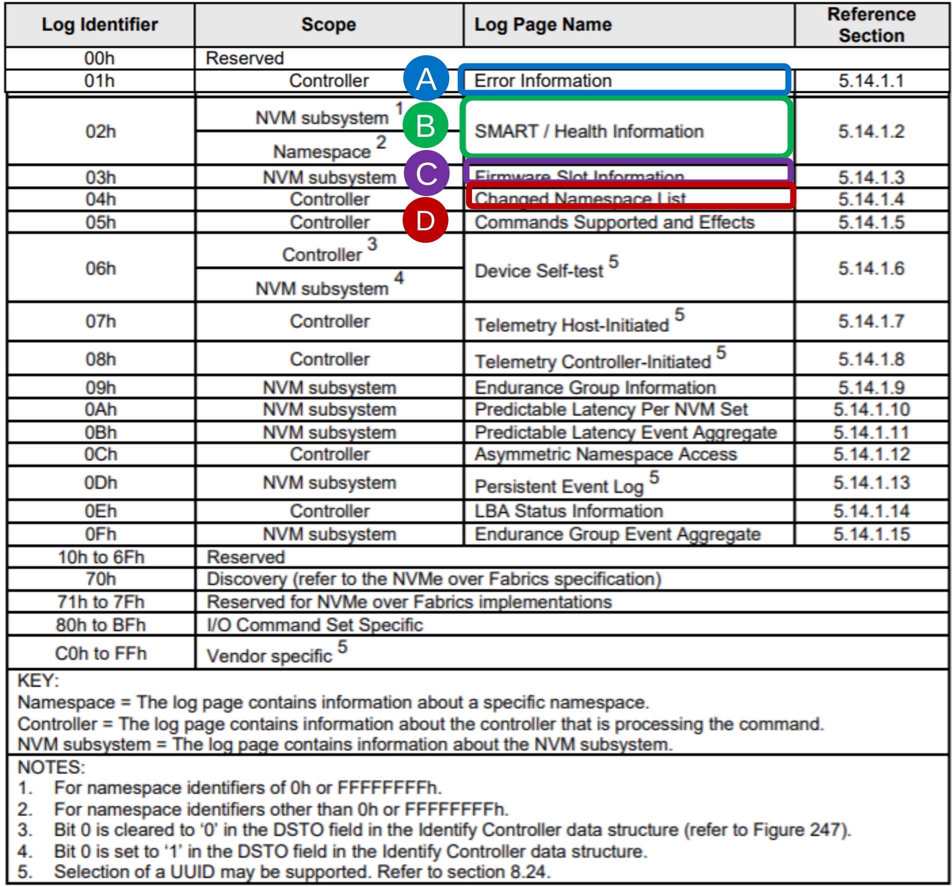
Features	
Arbitration	Host Controlled Thermal Management
Power Management	Non-operational Power State Config
LBA Range Type	Read Recovery Levels Config
Temperature Threshold	Predictable Latency Mode Config
Error Recovery	Predictable Latency Window
Volatile Write Cache	LBA Status Attributes
Number of Queues	Host Behavior
Interrpt Coalescing	Sanitize Config
Interrupt Vector Config	Endurance Group Event Config
Write Atomicity	Software Progress Marker
Asynchronous Event Config	Host Identifier
Auto Power State Management	Reservation Notification Mask
Host Memory Buffer	Reservation Persistence
Timestamp	Namespace Write Protect
Keep Alive Timeout	

NVMe™ 1.4 Specification Admin Sub-commands: Log Pages

Admin Commands
Identify
Firmware Download
Firmware Activate
Security Send
Security Receive
Log Page
Get Feature
Set Feature
Format NVM
MI Send
MI Receive
Directive Send
Directive Receive
Get LBA Status
Namespace Management
Namespace Attach
Sanitize
Virtualization Management
Device Self Test

Log Pages
Error
SMART
Firmware Info
Changed Namespace List
Command Effects
Device Self Test
Host Telemetry
Controller Telemetry
Endurance Group Information
NVM Set Predictable Latency
Predictable Latency Event Aggregate
LBA Status Information
Endurance Group Event Aggregate
Discover
Reservation Notification
Sanitize Status
Asymmetric Namespace Access
Persistent Event Log

Log Page Details



Log Identifier	Scope	Log Page Name	Reference Section
00h	Reserved		
01h	Controller	Error Information	5.14.1.1
02h	NVM subsystem [1] Namespace [2]	SMART / Health Information	5.14.1.2
03h	NVM subsystem	Firmware Slot Information	5.14.1.3
04h	Controller	Changed Namespace List	5.14.1.4
05h	Controller	Commands Supported and Effects	5.14.1.5
06h	Controller [3] NVM subsystem [4]	Device Self-test [5]	5.14.1.6
07h	Controller	Telemetry Host-Initiated [5]	5.14.1.7
08h	Controller	Telemetry Controller-Initiated [5]	5.14.1.8
09h	NVM subsystem	Endurance Group Information	5.14.1.9
0Ah	NVM subsystem	Predictable Latency Per NVM Set	5.14.1.10
0Bh	NVM subsystem	Predictable Latency Event Aggregate	5.14.1.11
0Ch	Controller	Asymmetric Namespace Access	5.14.1.12
0Dh	NVM subsystem	Persistent Event Log [5]	5.14.1.13
0Eh	Controller	LBA Status Information	5.14.1.14
0Fh	NVM subsystem	Endurance Group Event Aggregate	5.14.1.15
10h to 6Fh	Reserved		
70h	Discovery (refer to the NVMe over Fabrics specification)		
71h to 7Fh	Reserved for NVMe over Fabrics implementations		
80h to BFh	I/O Command Set Specific		
C0h to FFh	Vendor specific [5]		

KEY:
Namespace = The log page contains information about a specific namespace.
Controller = The log page contains information about the controller that is processing the command.
NVM subsystem = The log page contains information about the NVM subsystem.

NOTES:
1. For namespace identifiers of 0h or FFFFFFFFh.
2. For namespace identifiers other than 0h or FFFFFFFFh.
3. Bit 0 is cleared to '0' in the DSTO field in the Identify Controller data structure (refer to Figure 247).
4. Bit 0 is set to '1' in the DSTO field in the Identify Controller data structure.
5. Selection of a UUID may be supported. Refer to section 8.24.

A

- Reports error information for a command that completed with error or errors agnostic to particular command
- Host software asks for "n" error logs, then the error logs for the most recent "n" errors reported
- Controller clears the log page entries on power cycle and controller level reset

B

- Provides SMART and general health information over the life of the controller, retained across power cycles.
- Critical health warnings may be indicated via async. event notification, configured using the set features command

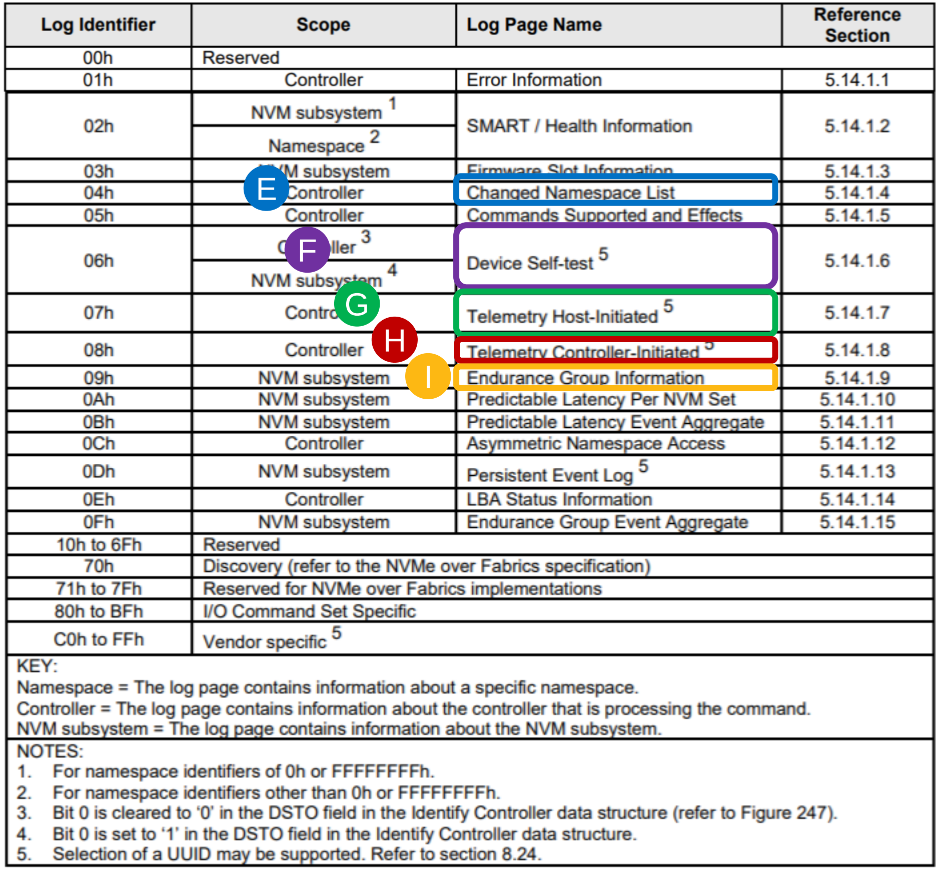
C

- Describes the firmware rev. in each firmware slot supported, indicates the active slot number and the slot that is going to be activated at the next controller level reset

D

- Reports attached namespaces changes such as identify namespace data structure, been added or deleted
- Log page contains a namespace list with up to 1,024 entries

Log Page Details



Log Identifier	Scope	Log Page Name	Reference Section
00h	Reserved		
01h	Controller	Error Information	5.14.1.1
02h	NVM subsystem [1] Namespace [2]	SMART / Health Information	5.14.1.2
03h	NVM subsystem	Firmware Slot Information	5.14.1.3
04h	Controller	Changed Namespace List	5.14.1.4
05h	Controller	Commands Supported and Effects	5.14.1.5
06h	Controller [3] NVM subsystem [4]	Device Self-test [5]	5.14.1.6
07h	Controller	Telemetry Host-Initiated [5]	5.14.1.7
08h	Controller	Telemetry Controller-Initiated [5]	5.14.1.8
09h	NVM subsystem	Endurance Group Information	5.14.1.9
0Ah	NVM subsystem	Predictable Latency Per NVM Set	5.14.1.10
0Bh	NVM subsystem	Predictable Latency Event Aggregate	5.14.1.11
0Ch	Controller	Asymmetric Namespace Access	5.14.1.12
0Dh	NVM subsystem	Persistent Event Log [5]	5.14.1.13
0Eh	Controller	LBA Status Information	5.14.1.14
0Fh	NVM subsystem	Endurance Group Event Aggregate	5.14.1.15
10h to 6Fh	Reserved		
70h	Discovery (refer to the NVMe over Fabrics specification)		
71h to 7Fh	Reserved for NVMe over Fabrics implementations		
80h to BFh	I/O Command Set Specific		
C0h to FFh	Vendor specific [5]		

KEY:
Namespace = The log page contains information about a specific namespace.
Controller = The log page contains information about the controller that is processing the command.
NVM subsystem = The log page contains information about the NVM subsystem.

NOTES:
1. For namespace identifiers of 0h or FFFFFFFFh.
2. For namespace identifiers other than 0h or FFFFFFFFh.
3. Bit 0 is cleared to '0' in the DSTO field in the Identify Controller data structure (refer to Figure 247).
4. Bit 0 is set to '1' in the DSTO field in the Identify Controller data structure.
5. Selection of a UUID may be supported. Refer to section 8.24.

E

List the commands that the controller supports and the effects of those commands on the state of the NVM subsystem

F

- Reports the status of any device self-test operation in progress and the percentage complete of that operation and results of the last 20 device self-test operations

G

- Telemetry Host-Initiated Data bit set to '1', controller captures states in this log, all Telemetry Data Blocks are 512 bytes
- The Telemetry Host-Initiated Data consists of three areas: Data Area 1, Data Area 2, and Data Area 3

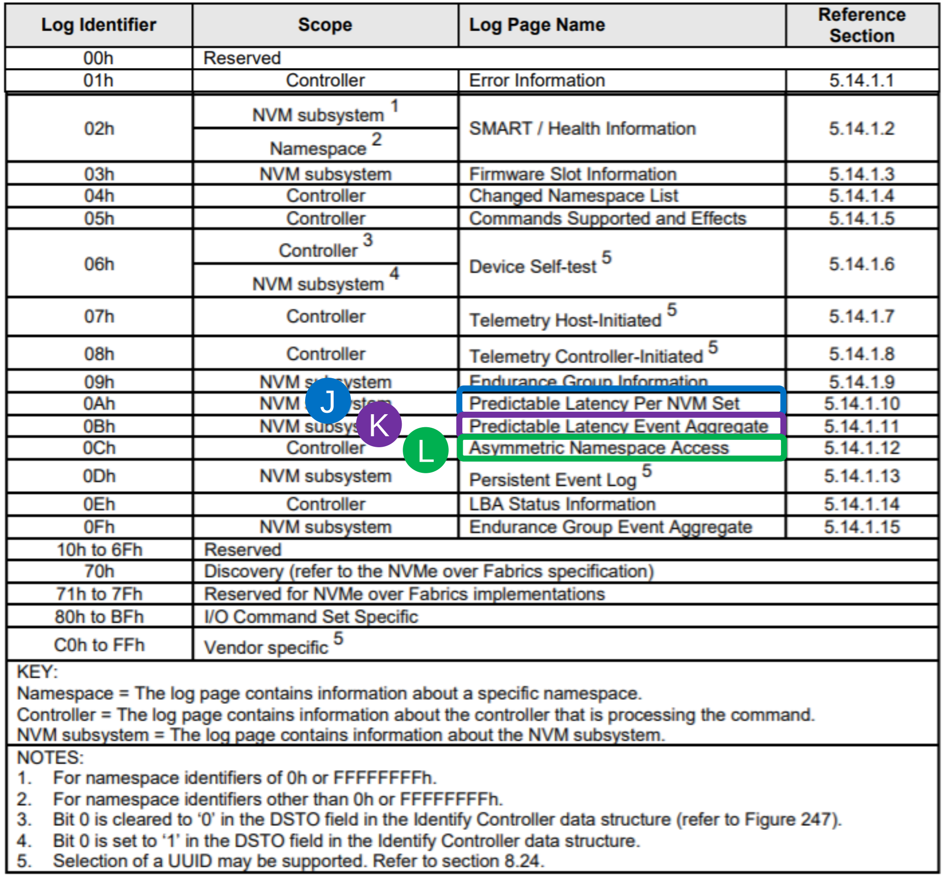
H

- Controller initiated and captures internal states. The Telemetry Controller-Initiated Data persist across all resets
- Telemetry Controller-Initiated Data consists of three areas: Data Area 1, Data Area 2, and Data Area 3

I

- Provides endurance information based on the Endurance Group (EG), the information provided over the life of the EG

Log Page Details



Log Identifier	Scope	Log Page Name	Reference Section
00h	Reserved		
01h	Controller	Error Information	5.14.1.1
02h	NVM subsystem [1] Namespace [2]	SMART / Health Information	5.14.1.2
03h	NVM subsystem	Firmware Slot Information	5.14.1.3
04h	Controller	Changed Namespace List	5.14.1.4
05h	Controller	Commands Supported and Effects	5.14.1.5
06h	Controller [3] NVM subsystem [4]	Device Self-test [5]	5.14.1.6
07h	Controller	Telemetry Host-Initiated [5]	5.14.1.7
08h	Controller	Telemetry Controller-Initiated [5]	5.14.1.8
09h	NVM subsystem	Endurance Group Information	5.14.1.9
0Ah	NVM subsystem	Predictable Latency Per NVM Set	5.14.1.10
0Bh	NVM subsystem	Predictable Latency Event Aggregate	5.14.1.11
0Ch	Controller	Asymmetric Namespace Access	5.14.1.12
0Dh	NVM subsystem	Persistent Event Log [5]	5.14.1.13
0Eh	Controller	LBA Status Information	5.14.1.14
0Fh	NVM subsystem	Endurance Group Event Aggregate	5.14.1.15
10h to 6Fh	Reserved		
70h	Discovery (refer to the NVMe over Fabrics specification)		
71h to 7Fh	Reserved for NVMe over Fabrics implementations		
80h to BFh	I/O Command Set Specific		
C0h to FFh	Vendor specific [5]		

KEY:
Namespace = The log page contains information about a specific namespace.
Controller = The log page contains information about the controller that is processing the command.
NVM subsystem = The log page contains information about the NVM subsystem.

NOTES:
1. For namespace identifiers of 0h or FFFFFFFFh.
2. For namespace identifiers other than 0h or FFFFFFFFh.
3. Bit 0 is cleared to '0' in the DSTO field in the Identify Controller data structure (refer to Figure 247).
4. Bit 0 is set to '1' in the DSTO field in the Identify Controller data structure.
5. Selection of a UUID may be supported. Refer to section 8.24.

J

- Determine the current window for the specified NVM Set when Predictable Latency Mode is enabled and any events occurred

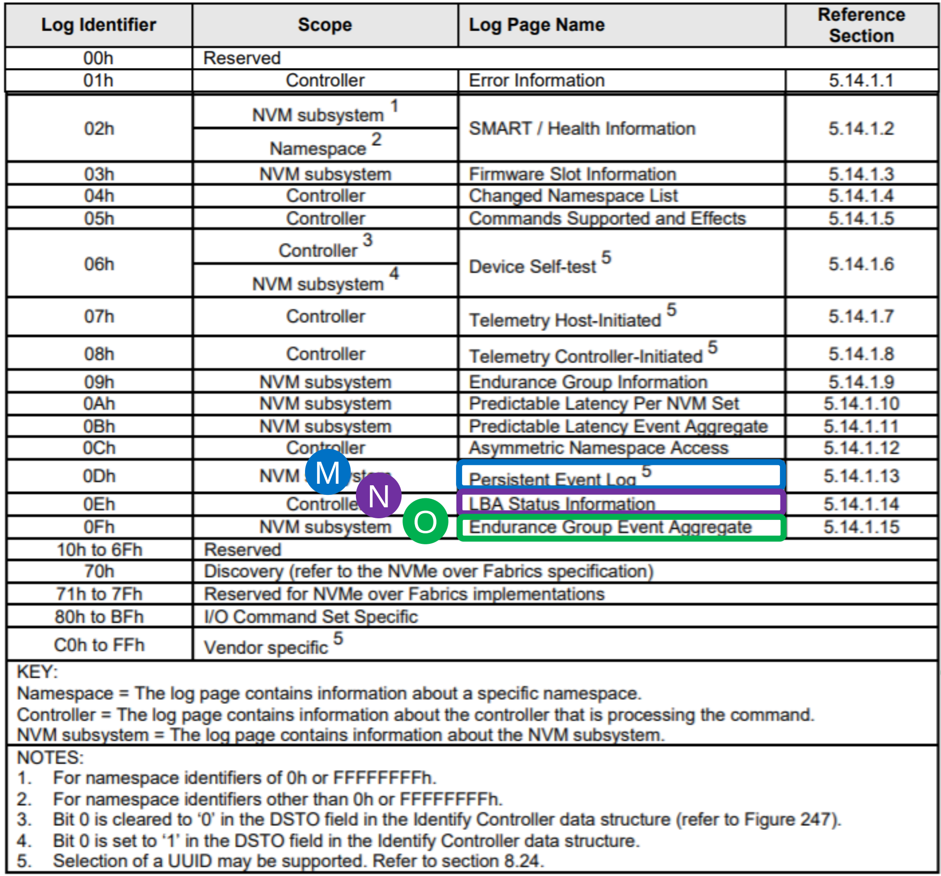
K

- Indicates Predictable Latency Events for a particular NVM Set, details included in the Predictable Latency Per NVM Set log page

L

- Asymmetric namespace access (ANA) indicates, to the host, information about access characteristics
- ANA occurs when NS access characteristics (e.g., performance or ability to access the media) vary based on the controller used to access the NS and the internal config. of the NVM subsystem

Log Page Details



Log Identifier	Scope	Log Page Name	Reference Section
00h	Reserved		
01h	Controller	Error Information	5.14.1.1
02h	NVM subsystem [1] Namespace [2]	SMART / Health Information	5.14.1.2
03h	NVM subsystem	Firmware Slot Information	5.14.1.3
04h	Controller	Changed Namespace List	5.14.1.4
05h	Controller	Commands Supported and Effects	5.14.1.5
06h	Controller [3] NVM subsystem [4]	Device Self-test [5]	5.14.1.6
07h	Controller	Telemetry Host-Initiated [5]	5.14.1.7
08h	Controller	Telemetry Controller-Initiated [5]	5.14.1.8
09h	NVM subsystem	Endurance Group Information	5.14.1.9
0Ah	NVM subsystem	Predictable Latency Per NVM Set	5.14.1.10
0Bh	NVM subsystem	Predictable Latency Event Aggregate	5.14.1.11
0Ch	Controller	Asymmetric Namespace Access	5.14.1.12
0Dh	NVM subsystem	Persistent Event Log [5]	5.14.1.13
0Eh	Controller	LBA Status Information	5.14.1.14
0Fh	NVM subsystem	Endurance Group Event Aggregate	5.14.1.15
10h to 6Fh	Reserved		
70h	Discovery (refer to the NVMe over Fabrics specification)		
71h to 7Fh	Reserved for NVMe over Fabrics implementations		
80h to BFh	I/O Command Set Specific		
C0h to FFh	Vendor specific [5]		

KEY:
Namespace = The log page contains information about a specific namespace.
Controller = The log page contains information about the controller that is processing the command.
NVM subsystem = The log page contains information about the NVM subsystem.

NOTES:
1. For namespace identifiers of 0h or FFFFFFFFh.
2. For namespace identifiers other than 0h or FFFFFFFFh.
3. Bit 0 is cleared to ‘0’ in the DSTO field in the Identify Controller data structure (refer to Figure 247).
4. Bit 0 is set to ‘1’ in the DSTO field in the Identify Controller data structure.
5. Selection of a UUID may be supported. Refer to section 8.24.

M

- The Persistent Event Log page contains information about significant events not specific to a particular command. The information in this log page shall be retained across power cycles and resets

N

- Provides information about subsequent actions the host may take to discover which logical blocks, in namespaces that are attached to the controller, may no longer be recoverable

O

- Lists if an Endurance Group Event has occurred for a particular EG. If an EG Event has occurred, the details of the particular event are included in the EG Information log page for that EG

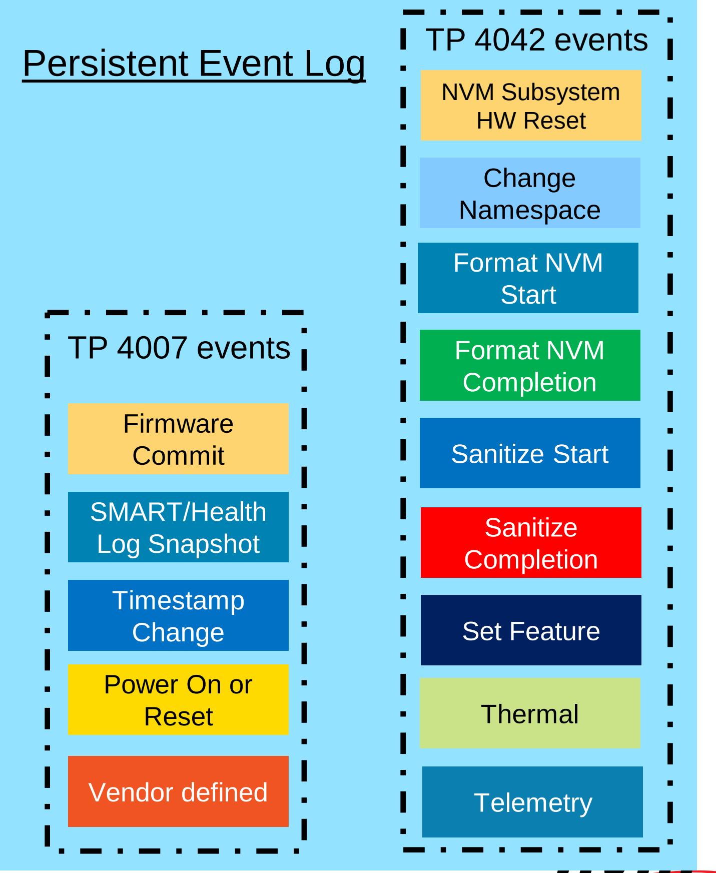
Persistent Event Log

Value Proposition

- Provides a standardized mechanism for the drive to log and communicate events to the host software stack
- This Log page contains information about significant events and is retained across power cycles and resets (subject to a threshold).

Implementations

- Supporting all the listed event log types
- Logs are preserved through power cycles and resets
- Oldest events are deleted in case of wrap-around
- Frequently recurring events of same type/info within a particular time interval are dropped to avoid unnecessary overflow of log



Persistent Event Log

TP 4007 events

- Firmware Commit
- SMART/Health Log Snapshot
- Timestamp Change
- Power On or Reset
- Vendor defined

TP 4042 events

- NVM Subsystem HW Reset
- Change Namespace
- Format NVM Start
- Format NVM Completion
- Sanitize Start
- Sanitize Completion
- Set Feature
- Thermal
- Telemetry

Rebuild Assist

Bill Martin, SSD IO Standards, Samsung

Rebuild Assist

- Feature - **Get LBA Status**
- Log page - **LBA Status Information**
  - Updated when “bad” LBAs are discovered in the background
  - May generate an Asynchronous Event Notification
- NVMe™ command – **Get LBA Status** to get a list of Potentially Unrecoverable LBAs
  - Tracked LBAs – done in background by drive
  - Untracked LBAs – initiated by host, informs the drive to scan for affected LBAs

Tracked vs. Untracked LBAs

- Tracked LBAs
  - Detected by controller during normal operation
    - Background scans
    - Component failure
    - Read request from host
    - Retained until repaired
  - Removed from list when host writes to the LBA
- Untracked LBAs
  - Scan requested by host
  - May be time consuming
  - Do not have to be retained following being read

Get LBA Status Information Attributes Feature

Requirements

- LBA Status Information Notices Asynchronous Event
- LBA Status Information log page
- Get LBA Status command

LBA Status Information Log and Asynchronous Event Notice

- Entries are added to the log as long as there is not a pending asynchronous event notice
- Has a number of elements describing where there MAY be bad LBAs
- Remains constant while there is a pending asynchronous event notice
- AEN is generated when there are elements in the LBA Status Information log and:
  - A host specified interval of time has occurred
  - A controller specific threshold number of elements have been added to the log

LBA Status Command

Get LBA Status parameters

- Action Type
  - 10h – Scan for and return Tracked LBAs
  - 11h – Return Untracked LBAs
- Scanning for “bad” LBAs can be time consuming
  - Untracked LBA list may be generated in increments

Recovery Procedure

- Tracked LBAs
  - Controller sends an LBA Status Information Alert asynchronous event
  - Host reads the LBA Status Information log page
  - Host performs necessary Get LBA Status commands
  - Host re-writes “bad” LBAs
- Untracked LBAs
  - Host performs necessary Get LBA Status commands
  - Host re-writes “bad” LBAs

Rebuild Assist – Untracked List Example

Controller:

- Detects die failure NS 1 and NS 2 affected
- Update LBA Status Information log page
- Issue asynchronous event

HOST:

- Read LBA Status Information log page

HOST

- Issues Get LBA Status commands with ATYPE 11h for:
  - NS 1 LBAs A- B
  - NS 1 LBAs C-D;
  - NS 2 LBAs A-Z

HOST

- Re-write all LBAs returned from the Get LBA Status Command

Controller

- Remove LBAs from Untracked List

Tracked List

Empty

Untracked List

NS1: LBAs
Range A-B
Range C-D
NS2: LBAs
Range A-Z

LBA Status Information Log Page

NS1: LBAs
A, B, C, D
NS2: LBAs
All LBAs

Rebuild Assist – Tracked List Example

HOST

- Issues Get LBA Status commands for NS 1 with ATYPE 10h
- Controller
- Scan Indirection table find Untracked List
- Return Untracked List

HOST

- Re-Writes LBA a, LBA f, LBA z

CONTROLLER

- Removes LBA a, LBA f, LBA z from Tracked list

Tracked List
(Before Scan)

Empty

Untracked List

Empty

Tracked List
(After Scan)

LBA a
LBA f
LBA z

LBA Status
Information Log Page

Not used for this
process

Q&A

nvm EXPRESS®

Architected for Performance