Changes in NVM Express® Specifications

This document is intended to help the reader understand changes in the NVMe® specifications as of August 2024

As of August 2024, NVMe is a set of the following eleven specifications:

- NVM Express® Base Specification, Revision 2.1
- Command Set Specifications:
    - NVM Express® NVM Command Set Specification, Revision 1.1
    - NVM Express® NVMe Zoned Namespaces Command Set Specification, Revision 1.2
    - NVM Express® Key Value Command Set Specification, Revision 1.1
    - NVM Express® Subsystem Local Memory Command Set Specification, Revision 1.0
    - NVM Express® Computational Programs Command Set Specification, Revision 1.0
- Transport Specifications:
    - NVM Express® NVMe over PCIe Transport Specification, Revision 1.1
    - NVM Express® RDMA Transport Specification, Revision 1.1
    - NVM Express® TCP Transport Specification, Revision 1.1
- NVM Express® NVM Express Management Interface Specification, Revision 2.0
- NVM Express® NVMe Boot Specification, Revision 1.1

The NVM Express® Base Specification, Revision 2.1 reorganizes sections to distinguish requirements on functionality common to PCIe® and Fabrics implementations, specific to PCIe implementations, and specific to Fabrics implementations.

This document describes mandatory, optional, capability and feature additions and changes in each of the above noted set of eleven specifications and notes incompatible changes to previous specification releases where appropriate and applicable.

1 Changes to the NVM Express® Base Specification

1.1 New Features

1.1.1 Flexible Data Placement - TP4146b (optional)

Adds an enhancement to the NVM Command Set to enable host guided data placement and introduces a Reclaim Unit (RU) that consists of physical non-volatile storage with the property that a RU may be erased/reused/repurposed without disturbing any other Reclaim Units. A RU is a physical construct only and is unrelated to logical addresses or LBAs. A namespace consists of capacity allocated in Reclaim Units. A Reclaim Group (RG) contains one or more Reclaim Units. There are one or more Placement Handles that reference RUs which I/O commands can target. I/O commands performed on one RG do not interfere with the performance or reliability of commands executed on another RG.

A Placement Identifier (PID) specifies a specific RG/Placement Handle pair that references an RU available for writing random LBAs (i.e., the Flexible Data Placement of user data to the non-volatile storage allocated to the RU). The written capacity of an RU referenced by a PID increments on each write command specifying to use that PID and the PID can be modified to reference another RU once the capacity of the RU is fully written. The host tracks the set of LBAs written to an RU, enabling the host to deallocate all LBAs associated with an RU at the same time to minimize garbage collection by the controller, thereby reducing write amplification. The host is responsible for managing PIDs and related device resources. Defines a mechanism for the Power Loss Signaling function introduced by PCI-SIG® for Mini Express form factors.

1.1.1.1 Description of change

- Added definitions of new terms
- Added the I/O Management Receive command to retrieve Reclaim Unit Handle Status descriptors
- Added the I/O Management Send command to provide control over Placement Identifiers
- Added the Flexible Data Placement feature to enable or disable FDP and to select the FDP configuration to apply to an Endurance Group
- Added new log pages:
    - FDP Statistics
    - FDP Events
    - FDP Configurations
    - Reclaim Unit Handle Usage
    - Added a new Flexible Data Placement model section including Reclaim Group and Reclaim Unit concept and usage, and FDP operations
- Added the Flexible Data Placement feature to figures listing feature support and feature identifiers

1.1.1.1.1 New requirement

- Added a new Data Placement Directive and allow each directive to specify if the enablement is persistent across a Controller Level Reset.
- The Streams Directive is not allowed to be enabled on any namespace in an Endurance Group that has Flexible Data Placement enabled.
- All namespaces in an Endurance Group that has Flexible Data Placement enabled shall be associated with the NVM Command Set.

1.1.1.1.2 Incompatible change

- Changed the support the Endurance Group Event Aggregate log page when Endurance Groups are supported to only be required if more than one Endurance Group is supported by the NVM subsystem.

1.1.2 PCIe Infrastructure for Live Migration - TP4159 (optional)

Added the ability for a host to manage the live migration of a controller from one NVM subsystem to another NVM subsystem.

1.1.2.1 Description of change

- Defined a new optional Track Receive command and Track Send command to track changes to user data and host memory due to the process of commands.
- Defined a new optional Migration Receive command and Migration Send command to manage the migration of a controller by allowing the suspending of a controller, resuming of a controller, and transfer of controller state.

- Defined a new optional Controller Data Queue command to allow a host to provide host memory for queuing information from the controller to the host. An LBA Migration Queue can be created to allow the controller to post logical block changes from a different controller once user data tracking has started as a result of a Track Send command.
- Defined a new optional Controller Data Queue feature to allow a host to specify the head pointer value and to identify a slot that when posted with an entry by the controller causes the controller to issue the new Controller Data Queue Tail Pointer event.

1.1.3 Power Loss Signaling - TP4029a (optional)

Defines a mechanism for the Power Loss Signaling function introduced by PCI-SIG® for the Mini Express form factors.

1.1.3.1 Description of change

- Adds requirements and behavior in a new model section.
- Adds two Power Loss Signaling variables to define communication with the transport layer.
- Adds support indication to the Identify Controller data structure.
- Adds the new Power Loss Signaling Config feature.
- Adds three pairs of time fields to the Power State Descriptor.

1.1.4 Dispersed Namespaces - TP4034a (optional)

Enables two or more NVM subsystems to present the same namespace. When the same namespace is presented by two or more NVM subsystems, then that the namespace be referred to as a dispersed namespace. Online data migration and data replication are two such use cases scenarios that may utilizes this capability.

1.1.4.1 Description of change

- Added new definitions for dispersed namespaces and participating NVM subsystems
- Added a new log page (Dispersed Namespace Participating NVM Subsystems) and added it to the log page support tables for all three controller types (optional for I/O and Admin controllers, prohibited for Discovery controllers)
- Allocated Controller ID FFFDh to be used in the CNTLID field of Registered Controller data structures when the controller associated with the reservation is contained in a separate participating NVM subsystem and added a new table to define the use of Controller ID values FFF0h to FFFFh
- Added a new Host Behavior Support feature field called Host Dispersed Namespace Support (HDISNS) that a host sets to signal that it supports accessing dispersed namespaces
- Added a generic status code called Host Dispersed Namespace Support Not Enabled to be used when a host that has not set HDISNS to ‘1’ sends an I/O command or a select Admin command to a dispersed namespace
- Added a new bit to each reservation command called Dispersed Namespace Reservation Support (DISNSRS) that a host sets to signal that it supports using reservations on dispersed namespaces
- Added a new command specific status code (Namespace Is Dispersed) to be used when a host that has set HDISNS to ‘1’ sends a reservation command to a dispersed namespace without setting the DISNSRS bit to ‘1’ in that reservation command
- Broke section 4.5.1 out into separate sub-sections for each different type of unique identifier
- Added a new bit to the I/O Command Set Independent Identify Namespace data structure in the NMIC field called Dispersed Namespace (DISNS) that is used to indicate that the namespace is a dispersed namespace
- Added a model clause section for dispersed namespaces

1.1.4.1.1 New requirement / incompatible change

  - If a dispersed namespace is a member of an ANA Group on one participating NVM subsystem then it must be included in an ANA Group on all participating NVM subsystems
  - Hosts must use globally unique namespace identifiers when accessing dispersed namespaces upon setting HDISNS to ‘1’
  - Hosts must use Host Identifiers that are unique across all hosts on all participating NVM subsystems upon setting HDISNS to ‘1’
  - Hosts must set DISNSRS to ‘1’ in each reservation command sent to a dispersed namespace upon setting HDISNS to ‘1’
  - Controllers that support dispersed namespaces must support the DISNSRS bit being set to ‘1’ in reservation commands

- If a reservation key is replaced on one participating NVM subsystem then it must be replaced on all participating NVM subsystems

1.1.5 Key Per I/O - TP4055 (optional)

Allows a unique encryption key to be used to encrypt data on a per read/write operation. This capability is to be used with an appropriate security specification (e.g., TCG Storage Security Subsystem Class: Key Per IO).

1.1.5.1 Description of change

- Added the definition section of Key Per I/O

1.1.5.1.1 New requirement

- Identify Controller data structure update to include indication of Key Per I/O support and scope.

1.1.6 Environmental Extremes Management - TP4058 (optional)

Defines a new host-initiated refresh operation as part of the Device Self-test command. The new operation provides a method for the host to inform the controller that it may perform internal refresh operation on the media.

1.1.6.1 Description of change

- Optional new operation as part of the DST command
- Update the Device Self-test Log page

1.1.7 Performance Characteristics Reporting - TP4077 (optional)

Enables NVMe device to report performance characteristics to allow a host system to dynamically optimize its workload distribution to NVMe devices based on these reported characteristics.

1.1.7.1 Description of change

- Added a Performance Characteristics feature identifier to the figure Set Features – Feature Identifiers.

1.1.8 Management Network Addresses - TP4094a (optional)

Creates a method to enable a host to detect whether a network management interface is available for the NVM subsystem or the system that contains the NVM subsystem and to obtain the network addresses used to manage that NVM subsystem and/or the system that contains the device.

1.1.8.1 Description of change

- Added the Management Address List log page.
- Added the Embedded Management Controller Address feature.
- Added the Host Management Agent Address feature.
- Added the Management Addresses section to the Extended Capabilities section.

1.1.9 Physical Interface Receiver Eye Opening Measurement (EOM) - TP4119b (optional)

Allows a host to request that a controller perform a receiver eye measurement on a PCIe port and the controller returning the measurement results that is allowed to include a printable eye and vendor specific information.

1.1.9.1 Description of change

- Added log page 19h to the Get Log Page command allowing a host to request that a controller perform a receiver eye measurement process in a transport-agnostic, and implementation-agnostic manner; and provide a way for the host to read the measurement results in a transport-agnostic, and implementation-agnostic manner.

1.1.10 NVMe Specification Version Reporting - TP4135 (mandatory)

Provides a means for the host to retrieve the versions of the I/O command set specifications and the NVMe Management Interface Specification.

1.1.10.1 Description of change

- Added the Specification Version Descriptor for use in multiple specifications.
- Redefined the Version property to use the Specification Version Descriptor.

1.1.11 Namespace Admin Label - TP4145 (optional)

Allows a NVM subsystem to store a label per namespace and report the label to the host. This label is a human-friendly UTF-8 string that describes the namespace (e.g., indicating the application for which the namespace has been created).

1.1.11.1 Description of change

- Define Namespace Admin Label as a human-friendly UTF-8 string in a new Feature.

1.1.12 Reachability Architecture –TP4156a (mandatory if Reachability Reporting supported)

Added the ability for a controller to indicate namespaces able to be used by operations that specify multiple namespaces. This ability is called reachability, defined at the controller level.

1.1.12.1 Description of change

- Create new log pages: Reachability Groups log page, Reachability Associations log page
- Defined the Reachability Architecture

1.1.12.1.1 New Requirement

  - Reachability Groups log page and Create a Reachability Associations log page are mandatory if Reachability Reporting is supported.

1.1.13 Subsystem Local Memory Command Set Specification changes to other specifications – TP4171 (optional)

Added the changes to the NVM Express NVM Command Set Specification and NVM Express Base Specification necessary to support the Subsystem Local Memory Command Set.

1.1.13.1 Description of change

- Modifications to Identify Controller Data Structure (CNS 01h)
- Modifications to Host Behavior Support (Feature Identifier 16h)
- Modifications to Command Set Identifiers

1.1.14 Discovery Controller Identification – TP4181 (optional)

End-users are now able to identify the system that contains a particular discovery controller by using the Identify Controller data structure (e.g., Model Number (MN), Serial Number (SN)) rather than having to rely on the fabric address alone.

1.1.14.1 Description of change

Updated the Identify Controller data structure to indicate the Serial Number (SN) and Model Number (MN) fields are now Optional for Discovery controllers instead of Reserved.

1.1.15 Computational Programs Command Set Specification changes to other specifications – TP4182 (optional)

Changes to the NVM Express Base and NVMe-MI Specifications necessary for Computational Programs Command Set.

1.1.15.1 Description of change

- Modifications to Theory of Operation types of command sets
- Modifications to Admin Command Set opcodes
- Modifications to Command Specific Status Values
- Modifications to Command Set Identifiers

1.1.16 NVMe Scalable Resource Management – TP6011a (optional)

Adds the notion of exported NVM subsystems and a common framework to dynamically construct, configure, and provision exported NVM Subsystems as logical aggregations of underlying physical resources. In a landlord/tenant environment, e.g., an Infrastructure-as-a-Service cloud, a landlord is able to export NVM Subsystems to each tenant, thereby providing each tenant with separate (non-interfering) scopes of management, namely her own exported NVM subsystems. This functionality is intended to support a variety of use-case / transport-specific virtualization capabilities (e.g., NVMe over Fabrics, Multi-path I/O, Namespace Sharing, Access from Virtual Machines).

Defines a common framework for configuration and management of exported NVMe resources via administrative commands and new administrative management operations to enable resource management for logical representations of physical NVM resources including:

- retrieval of information regarding available NVM and fabrics resources on underlying NVM storage.
- configuration of exported NVM resources
- provisioning access to exported NVM resources

1.1.16.1 Description of change

- Added additional “Definitions” to disambiguate underlying NVMe resources from exported NVMe resources and to support NVMe scalable resource management.
- Requires sanitize, Predictable Latency, and Endurance group commands to be prohibited for Exported NVM Resources. Added section discussing Sanitize operation Effects on Exported NVM Subsystems
- Requires Capacity Management commands are prohibited for Exported NVM Resources.
- Added a new Command Specific Status Codes
- Splits Namespace Management into two subsections: Namespace Management for NVM Subsystems that are not Exported NVM Subsystems, and Namespaces Management for NVM Subsystems that are Exported NVM Subsystems to disambiguate Namespace Management depending on the Export/Non-exported state of the NVM Subsystems.
- Adds new sections to describe exporting NVM resources and show example flows using the added commands and associated details.

1.1.17 NVMe-MI High Availability – TP6034a (optional)

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

1.1.17.1 Description of change

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

1.1.18 Automated Discovery of NVMe-oF Discovery Controllers for IP Networks – TP8009 (optional)

Added explanation on how to interpret Domain Name System Service Definition (DNS-SD) records to determine the IP Address of one or more NVMe-oF Discovery Service instances; when and how to use the multicast Domain Name System (mDNS) protocol to automatically retrieve the DNS-SD records that are available on the local Broadcast Domain; and when and how to use unicast DNS requests to retrieve all DNS-SD records known to the DNS.

1.1.18.1 Description of change

- Added a new subsection that describes “Automated Discovery of NVMe-oF Discovery controllers for IP Networks”

1.1.19 NVMe-oF Centralized Discovery Controller – TP8010a (optional)

Defines new requirements and capabilities that enable discovery information (e.g., Discovery log pages) from multiple Direct Discovery controllers to be consolidated into and retrievable from a single Centralized Discovery controller. The method hosts and Direct Discovery controllers use for discovering Centralized Discovery controllers is described in TP 8009. This proposal is only applicable for IP-based fabric transports.

1.1.19.1 Description of change

- Added six new definitions: CDC, DDC, discovery information, Fabric Zoning, Entry Key, and symbolic name

- Added four new Admin commands (all of which are mandatory for CDCs and optional for other Discovery controller types): Discovery Information Management, Fabric Zoning Lookup, Fabric Zoning Send, and Fabric Zoning Receive
- Added one new log page (mandatory for CDCs and optional for other Discovery controller types): Host Discovery
- Added two new command specific status codes for the DIM command: Invalid Discovery Information and Insufficient Discovery Resources
- Added three new command specific status codes for the FZL, FZS, and FZR commands: Zoning Data Structure Locked, Zoning Data Structure Not Found, and Requested Function Disabled
- Added one new command specific status code for just the FZS command: ZoneGroup Originator Invalid
- Added new Host Discovery Log Page Change Asynchronous Event notification corresponding to the Host Discovery log page
- Defined LID Specific Parameter field for the Discovery log page and included three new bits to show support for new functionality: ALLSUBES, PLEOS, and EXTDLPES
- Defined Log Specific Parameter field for the Discovery log page and included three new bits that request the functionality corresponding with each of the three new bits included in the LID Specific Parameter field: ALLSUBE, PLEO, and EXTDLPE
- Added new NCC bit to EFLAGS field in the Discover log page to indicate whether or not the DDC describing the DLE is currently connected to the CDC
- Added new Extended Discovery Log Page Entry to the Discovery log page which includes additional information (i.e., extended attributes)
- Added new DLPF field to the Discovery log page header and included three new bits to indicate information related to new functionality: ALLSUB, PORTLCL, and EXTEND
- Added new TDLPL field to the Discovery log page header to indicate the total length of the log page
- Added new DCTYPE field to the Identify Controller data structure to indicate which type of Discovery controller the controller is
- Added new CONNENT bit to the Connect command's CATTR field to indicate what type of entity is performing the Connect command
- Added a model clause section for centralized discovery that defines two different discovery information registration/de-registration models (i.e., push, pull), Fabric Zoning to control which discovery information is returned for Get Log Page, and Asynchronous Events generated as the result of changes to discovery information

1.1.19.1.1 New requirement / incompatible change:

- Commands Supported and Effects log page is now mandatory for all Discovery controllers

1.1.20 Subsystem Driven Zoning with Pull Registration – TP8016 (Mandatory for Discovery controllers that are Centralized Discovery Controllers / prohibited for Discovery controllers that are not a CDC)

Provides a standardized process for a Direct Discovery Controller (DDC) to provide Fabric Zoning information to a Centralized Discovery Controller (CDC) without the need to support Host functionality.

1.1.20.1 Description of change

- Subsystem Driven Zoning for Pull Registration for Direct Discovery Controller DDCs

1.1.21 NVMe-oF Security Configurations – TP8025 (optional)

Defined a consistent way to express security protocol configurations of NVMe entities.

1.1.21.1 Description of change

- Definition of a consistent way to express security protocol configurations of NVMe entities

1.1.22 Controller Optimal Aggregated Queue Depth Reporting – TP4103 (optional)

Enabled an NVMe device to report the optimal aggregated I/O Queue Depth information to enable a host system to dynamically optimize its workload distribution to NVMe devices based on these reported information

1.1.22.1 Description of change

- Identify Controller Data Structure extended to include the optimal aggregated I/O Queue Depth information

1.2 Feature Enhancements

1.2.1 Abort Enhancements – TP4097a (mandatory and optional components)

Added clarification for the Admin Abort command and adds a new optional Cancel command.

1.2.1.1 Description of change

1.2.1.1.1 Mandatory change

- Modified the Abort command to support immediate and deferred aborts.

1.2.1.1.2 Optional change

- Created the Cancel command to abort I/O commands in an I/O Submission Queue.

1.2.1.1.3 Incompatible Change

- The Abort command required that for bit 0 of Dword 0 to be cleared to '0' in the CQE of the Abort command, the CQE for the command being aborted be posted prior to the posting of the CQE for the Abort command.
- To post the CQE for the command being aborted obviously required that all actions associated with that command be stopped. The Abort command is now allowed to post the CQE for the ABORT command and the CQE for the command being aborted in any order – on the condition that all the actions associated with the command being aborted have stopped.
- While the stopping of actions associated with the command being cancelled is compatible, the order of the posting of the CQEs may be considered incompatible.

1.2.2 Defining Scope for Features – TP4074a (mandatory)

Defined a clear scope for each Feature Identifier. Features associated with power states remained at the controller level, therefor for NVM Subsystems with multiple controllers in the same domain, specifying different power states results in an unspecified power state for that domain.

1.2.2.1 Description of change

- Defined the scope of each Feature where the scoping is implied. In addition, Features that require the option of multiple scoping requirements are defined.
- Added scope field in the Feature Identifiers table.
- Added explicit text to each feature section that states the feature's scope.
- Added paragraph to the Feature Values section that describes how feature settings apply to features with scopes other than controller and namespace scope.
- Re-introduces the definition of what it means for a feature to be namespace specific and non-namespace specific.

1.2.3 Namespace Capability Reporting TP4095a (mandatory if Namespace Management supported)

Created a mechanism to support LBA formats that do not report the same namespace attributes across all LBA formats and a mechanism to report the namespace attributes of LBA formats without having to create the namespace.

1.2.3.1 Description of change

- Added CNS value 09h to the Identify command to return an Identify Namespace data structure associated with the specified Format Index
- Added CNS value 0Ah to the Identify command to return an I/O Command Set Specific Identify Namespace data structure associated with the specified Format Index and Command Set Identifier.

1.2.3.1.1 New requirement / incompatible change

- The Identify command for CNS values 00h, 05h and 08h removed the term "common" and clearly specified the requirements on all fields when the specified NSID is set to FFFFFFFFh.
- The requirement to abort an Identify command with CNS values 00h, 05h, and 08h, when the specified NSID is FFFFFFFFh and Namespace Management capability is not supported, was changed to optional (previously, aborting the command was mandatory).

1.2.4 MDTS Enhancement – TP4099 (mandatory)

Added the ability for a controller to specify that the MDTS field value, the size limit of the Verify command, the size limit of a Write Zeroes command, and the size limit of the Write Uncorrectable command, excludes the length of metadata (i.e., increasing the size of MDTS by the total size of the metadata in the command). If supported by a controller, then a host, supporting an earlier version of the NVM Express Base Specification, that issued a command within the increased size of the MDTS, and expected the controller to abort the command, now has that command succeed.

1.2.4.1 Description of change

- Added MEM bit in CRATT field of the Identify Controller Data Structure
- Changed the definition to MDTS in a backward compatible manner

1.2.4.1.1 New requirement / incompatible change

  - An old host that issued a command within the increased size of the MDTS, and expected the controller to fail the command, will now succeed when MEM bit is set to 1.

1.2.5 PMR Controller Reset Enhancement – TP4100 (mandatory if PMR is supported)

Removed the dependency that PMR is disabled on a Controller Reset

1.2.5.1 Description of change

Persistent Memory Region (PMR) properties are not modified due to a Controller Reset as the intent of Persistent Memory Region is that operate independently of controller being enabled and disabled.

The following PMR properties are not modified by a Controller Reset:

- Persistent Memory Region Capabilities (PMRCAP)
- Persistent Memory Region Control (PMRCTL)
- Persistent Memory Region Status (PMRSTS)
- Persistent Memory Region Elasticity Buffer Size (PMREBS)
- Persistent Memory Region Sustained Write Throughput (PMRSWTP)

1.2.5.1.1 New requirement and incompatible change

  - All Persistent Memory Region properties are not affected by a Controller Reset

1.2.6 Host Memory Buffer (HMB) with Low Power Support - TP4104a (mandatory if HMB is supported)

Allows a host to advise a controller about transitioning of that host to non-operational low power state. Allows the controller to limit access to HMB only on non-operational power state initiated by the host, and for the host to not change or deallocate the HMB already in use.

1.2.6.1 Description of change

- New support field in the Controller Attribute in the Identify Controller data structure.
- Host Memory Buffer
  - Specified the HMB access restrictions in operational power states.
  - Added an enable bit in the Host Memory Buffer – Command Dword 11
  - Added reporting of the restrictions on the in the Host Memory Buffer – Completion Queue Entry Dword 0

1.2.7 Allow host to specify Telemetry Host-Initiated data areas – TP4109a (optional)

Enhanced the Get Log Page command to allow a host to specify the maximum data area to be created in the Telemetry Host-Initiated log page.

1.2.7.1 Description of change

1.2.7.1.1 New requirement / incompatible change

- A host is allowed to specify the maximum data area to be created in the Telemetry Host-Initiated log page in the Get Log Page commands.
- The LID Specific Parameter field in the Supported Log Pages log page is now specified by each log page that utilizes the field.

1.2.8 Align PCIe and Fabrics HOSTID Management – TP4110a (mandatory)

Modified HOSTID rules to create behaviors as consistent as possible between fabrics and PCIe by:

- Allowing fabrics connections with the Host Identifier value set to 0h
- Only allowing Host Identifier to change only if set to 0h
- Allowing PCIe implementations to prohibit reservations when Host Identifier is set to 0h and prohibit fabrics supporting reservations when Host Identifier is set to 0h
- Allowing streams that use Host Identifiers being prohibited by PCIe and Fabrics implementations if the Host Identifier value is set to 0h.
- Host Identifier feature is not saveable

1.2.8.1 Description of change

- Made changes to both PCIe and Fabrics rules for setting and changing the Host Identifier so that both support the same model and same commands.
- For both PCIe and Fabrics:
  - the Host Identifier value is allowed to be 0h when a controller is enabled and then set to a non-zero value with a Set Features command (new for Fabrics, existing behavior for PCIe)
  - once set to a non-zero value the Host Identifier value is not able to be changed with a Set Features command (new for PCIe, existing behavior for Fabrics)
  - the Host Identifier should be set to a non-zero value before using streams or reservations
- New optional behavior:
  - the NVM subsystem may require the Host Identifier be set to a non-zero value before using streams or reservations
  - New capability to report whether non-Zero Host Identifier value is required for reservations
  - New capability to report whether non-zero Host Identifier value is required for Streams

1.2.8.1.1 New requirement / incompatible change

- PCIe
  - Host Identifier defaults to a value of 0h (this behavior was expected but unstated)
  - Host Identifier is not allowed to be changed if it is set to a non-zero value
- Fabrics
  - Host Identifier may be set to a value of 0h or set to a non-zero value in a Connect command to create an Admin Queue
  - Host Identifier may be set to a value of 0h or set to the Host Identifier value associated with the Admin Queue in a Connect command to create an I/O Queue
  - If the Host Identifier was set to a value of 0h in a connect command, then it may be changed to a non-zero value
  - previous specification versions required a non-zero Host Identifier established at connection and never changed

1.2.9 Implicit FW Pending Activation – TP4112 (mandatory)

Enhances firmware update by:

- clarifying the effects on a Sanitize command and sanitize operation when firmware images become activated by a controller when not specifically activated by a host.
- clarifying that the firmware image is commented to the specified slot when the returned status code of a Firmware Activation command is Firmware Activation Requires Maximum Time Violation.

1.2.9.1 Description of change

1.2.9.1.1 Incompatible and New requirement:

- The status code Firmware Activation Requires Maximum Time Violation has been updated to state that the firmware image is committed to the firmware slot. Prior to this change nothing was stated on whether the firmware image was or was not committed.

1.2.9.1.2 New requirements:

- A Sanitize command is aborted if the host commits a firmware image to the currently active firmware slot without activation and that committed firmware image is allowed to be activated by a reset.
- If a firmware image cannot be loaded as part of a reset and the controller reverts to a different firmware image while a sanitize operation is in progress, then that sanitize operation fails.
- If the host issues attempts to commit without activation to the active firmware slot. A controller may abort a Firmware Commit command if the host attempts to commit without activation to the active firmware slot.

1.2.10 Add Opcode field to the Error Information log page – TP4113a (mandatory)

Entry in the Error Information log page are allowed to indicate of the command associated with that entry.

1.2.10.1.1 Description of change

Added the following fields to the Error Information Entry data structure: Command Set Indicator, Opcode, Log Page Version

1.2.11 Vendor Specific Directives TP4114 (optional)

Reserves a directive for vendor specific use.

1.2.11.1.1 Description of change

- Add support for vendor specific directives.

1.2.11.1.1.1 New requirement / incompatible change:

- A new directive for vendor specific usage has been defined.

1.2.12 Device Self-test (DST) Enhancement – TP4124 (optional)

Enables Device Self-test command to pass in parameters for vendor specific tests.

1.2.12.1.1 Description of change

- Allow Command Dword 15 to be used in the Device Self-test command for vendor specific tests.

1.2.13 KATO Corrections and Clarifications - TP4129 (mandatory)

Added a description of host operation during a communication loss to avoid data corruption that could be caused by interaction of outstanding commands with subsequent commands submitted by that host. This includes:

- A requirement to determine the controller has stopped processing commands that can change state on the controller before retrying or erroring out that command;
- Requirements on retrying such commands and examples of incorrect use of retries.
- Defined a Command Quiesce Time (CQT) to allow for cleanup time for outstanding commands on the controller.
- Specifies corrections in the use of the KATO (Keep Alive Time Out) operation. Specifically:
  - Requires the controller to detect a KATO if the Keep Alive Timer expires;
  - describes interactions and timings associated with host and target use cases for KATO; and
  - Expands and corrects an example implementation for Keep Alive operation on the host, illustrating how the implementation keeps the controller alive.

1.2.13.1 Description of change

- Handling Outstanding Commands during a Communication Loss
  - **Feature Enhancement**

        - Added Command Quiesce Time, which the controller reports as the time it takes to stop processing commands after the controller detects a communication loss with the host.
    - **Required Change**
        - Described, in a new section, error recovery if a communication loss happens and there are outstanding commands, specifically:
            - how a host detects loss of communication with a controller;
            - how a host determines no further controller processing of outstanding commands is possible; and
            - restrictions for using retries of a command for recovery.
- Keep Alive Corrections
    - **Feature Enhancement**
        - Expanded the host examples of CBKA and TBKA, focusing them on sending Keep Alive commands to keep the controller alive, covering more situations.
    - **Required Change**
        - Clarified that the concepts of Keep Alive Timer and Keep Alive Timer Interval apply to the controller, not directly to the host.
        - For TBKA, changed the interval the host checks to send a Keep Alive from KATT/2 to KATT/4.
    - **Backwards Incompatible**
        - If a Keep Alive Timer expires, the controller is required to detect a Keep Alive Timeout.

1.2.14 Cross-Namespace Copy – TP4130a (mandatory components if Copy command is supported)

Extend Copy command to copy data across different namespaces on the same NVM subsystem via support for specifying a namespace for each source range. Extend Protection Information (PI) support to cross-namespace copy, including the ability to insert or strip PI in common cases where all logical block metadata is PI. Clarify ambiguities in originally specified atomicity properties of original Copy command (also applies to Copy across namespaces).

1.2.14.1 Description of change

1.2.14.1.1 Required change

- Copy command atomicity clarification (mandatory if NVM Command Set Copy command is supported, including ZNS Command Set usage of that Copy command)

1.2.14.1.2 Optional Changes

- Three new status codes for errors that may result from attempts to copy data across multiple namespaces
- New ONCS (Optional NVM Command Support) bit in Identify Controller to indicate whether all copies within an NVM Subsystem are fast copies (NVMAFC bit).
- Two new bits in Identify Controller to indicate support for the two new copy descriptor formats that enable copy across namespaces.
- Two new Host Behavior Support bits for the host to explicitly enable the two new copy descriptor formats (both new formats start out disabled).
- Update Reservation Conflict table to separate reservation checks for read and write portions of the Copy command.
- Add a new command-specific status code for Insufficient Resources.

1.2.14.1.2.1 New requirement / incompatible change

- In the Identify Controller data structure: New ONCS bit in Identify Controller to indicate whether scope of atomicity guarantees for the Copy command is the entire command or each source range (NVMSCA). Implementations that comply with this specification are required to set that bit to ‘1’ to indicate entire command scope – that requirement is stated in the additions to the NVM Command Set Specification.

1.2.15 Temperature Threshold Hysteresis – TP4142 (optional)

Allows the Temperature Threshold feature to specify hysteresis on the thermal event within the limits supported by the controller which the controller to determine the end of the thermal event and notify the host.

1.2.15.1.1 Description of change

- Adds a host-controlled temperature hysteresis parameter to the Temperature Threshold feature that the controller uses to determine the end of the thermal event.
- If the controller does not support the new parameter, or the host does not use the new parameter, the parameter has no effect on the Temperature Threshold feature.
- Adds the ability for the controller to generate an asynchronous event that indicates the end of a thermal event.

1.2.16 Clarifications to Namespace Changes Reporting - TP4150 (optional with mandatory components)

Added clarifications related to change reporting for namespaces. Notifications for changes to attached namespaces use the existing AEN (now called Attached Namespace Attribute Changed), with the NSIDs of those changed namespaces being listed in the existing log page (now called Changed Attached Namespace List). Notifications for changes to attached namespaces and unattached namespaces use the new Allocated Namespace Attribute Changed optional AEN, with the NSIDs of those changed namespaces being listed in the new Changed Allocated Namespace List log page. Thus, this new AEN is a superset of the existing AEN which additionally includes notifications for changes to unattached namespaces.

1.2.16.1.1 Description of change

- Added applicable CNS values to Identify Namespace data structures definition and references where appropriate
- Added definition for allocated namespace
- Renamed Namespace Attribute Changed asynchronous event to Attached Namespace Attribute Changed
- Renamed Changed Namespace List log page to Changed Attached Namespace List
- Renamed Namespace Attribute Notices notice to Attached Namespace Attribute Notices
- Defined Allocated Namespace Attribute Changed asynchronous event
- Defined Changed Allocated Namespace List log page
- Defined Allocated Namespace Attribute Notices notice
- Defined Allocated Namespace Attribute Notices bit in OAES field of Identify Namespace data structure
- Added clarifying text in Namespace Management command, Capacity Management capability, and Namespace Management capability sections regarding when Attached Namespace Attribute Changed and Allocated Namespace Attribute Changed asynchronous events are generated.

1.2.17 Post Sanitize Media Verification - TP4152 (optional)

Added extensions to the Sanitize command, sanitize operations, and the Read command that enable verifying correct sanitization by reading sanitized user data from media. A new state machine defines existing behavior and adds the Media Verification state and the Post-Verification Deallocation state, which support verification of sanitized media.

If a Read command is processed while the NVM subsystem is in the Media Verification state, then the command returns data from media, even if the sanitize operation invalidated internal checksums. Multiple reads of the same logical block are allowed to return differing data; this serves to obscure media characteristics that might be inferred from analysis of repeated reads of the media.

A new sanitize action in the Sanitize command changes the NVM subsystem from the Media Verification state to normal operation. During that change to normal operation, the NVM subsystem is in the Post-Verification Deallocation state and all user data is deallocated.

1.2.17.1.1 Description of change

- Added Sanitize Operation Entered Media Verification State to the Asynchronous Event Information – I/O Command Specific Status.
- Added Sanitize Media Verification Event to the Persistent Event log page.
- Added Sanitize State (SANS) field to the Sanitize Status log page.
- Added Verification Support (VERS) bit to the Identify Controller data structure.
- Added Enter Media Verification State (EMVS) bit to the Sanitize command.
- Modified the Sanitize command description to refer to the Sanitize Operation State Machine.
- Added Sanitize Operation State Machine, including requirements for states and for transitions.

1.2.17.1.1.1 New requirement / incompatible change

  - The Sanitize Progress (SPROG) field indicates progress separately for the Processing states and the Post-Verification Deallocation state. Previously, this field indicated progress for the entire sanitize operation.
  - Hosts that are not aware of Media Verification will not recognize the new Sanitize Operation Entered Media Verification State AEN.

1.2.18 Independent Identify Namespace Data Structure Access - TP4155 (mandatory if Namespace Management supported)

Enables a host to access the I/O Command Set Independent Identify Namespace data structure for a namespace that exists.

1.2.18.1.1 Description of change

- Added a CNS value to the Identify command to allow a host to access the I/O Command Set Independent Identify Namespace data structure for a namespace not attached.

1.2.19 WZSL Limit Modification – TP4160 (optional)

Allows a controller to advertise a larger limit of logical blocks that can be processes for Write Zeroes commands with the Deallocate bit set to ‘1’ and allows the Write Zeroes command to specify to deallocate the entire namespace when the Deallocate bit set to ‘1’.

1.2.19.1.1 Description of change

- Added new support bit in ONCS for Write Zeroes command that is able to write zeroes to the entire namespace.

1.2.20 Non-Storage namespace changes – TP4162a (optional)

Clarified the definition of namespace to support different namespace types since NVMe supports non-volatile storage namespaces, memory namespaces, and compute namespaces.

1.2.20.1.1 Description of change

- Modify a number of usages of “namespace”

1.2.21 Tracking LBA Allocation with Granularity – TP4165 (optional with mandatory components)

Extended the Get LBA Status command to report allocated LBAs managed by the controller at an indicated granularity.

1.2.21.1.1 Description of change

- Updated the Identify Controller data structure to clarify the support for the LBA Status command functionality.

1.2.22 Command Effects Enhancements – TP4167 (optional)

Allow the controller to indicate in the Commands Supported and Effects log page a separate restriction for Admin commands executing concurrently with a given command versus I/O commands executing concurrently with that same command.

1.2.22.1.1 Description of change

- Defined new Command Submission and Execution Relaxation (CSER) field in reserved area in Commands Supported and Effects data structure which relaxes the original CSE behavior choices.

1.2.23 Shutdown Clarifications and Enhancements – TP4169 (optional with mandatory components)

Updated the definition of NVM Subsystem Shutdown to specify interactions between NVM Subsystem Shutdown and Controller Level Resets (CLRs) that are not caused by NVM Subsystem Reset. Allows the controller to indicate to the host how long a host should wait for NVM Subsystem Shutdown to complete. Removed use of the word “safe” to describe ready-to-be-powered-off conditions. Specified the situations in which the media is in the shutdown state while the CSTS.SHST field is cleared to 00b.

1.2.23.1 Description of change

- NVM Subsystem Shutdown allows all forms of Controller Level Reset to occur, both while an NVM Subsystem Shutdown is in progress, and after its completion until a subsequent NVM Subsystem Reset. For the transport-specific resets (e.g., PCIe Conventional Reset and Function Level Reset), this is a feature enhancement.
- While a Controller Shutdown is in progress and after its completion, all forms of Controller Level Reset permit (but do not require) the controller to both initialize media and cease being ready to be powered off. For the transport-specific resets and NVM Subsystem Reset, this is a feature enhancement.
- Corrected the specification of the Unsafe Shutdowns field in the SMART / Health Information log page to state that it counts unexpected power losses. Removed usage of the word “safe” for clarity and changed the field name accordingly.
- Documented requirements to start executing commands after controller shutdown is complete for Fabrics (message-based) controllers based on corresponding requirements for memory-based controllers.
- Added an NVM Subsystem Shutdown Latency (NSSL) field to Identify Controller to indicate how long host software should wait for an NVM Subsystem Shutdown to complete.
- Added the additional CSTS.SHST=00b case for the NVMe-MI Ignore Shutdown (ISH) bit - processing an out-of-band NVM Admin command that requires media access and has the ISH bit set to ‘1’ is able to make the media accessible.
- If a controller that is ready to be powered off processes an Admin Command that requires access to the media and specifies the Ignore Shutdown bit set to ‘1’ via the out-of-band mechanism (refer to the NVM Express Management Interface Specification), then that controller ceases to be ready to be powered off.

1.2.23.1.1 Incompatible change

- NVM Subsystem Shutdown allows all forms of Controller Level Reset to occur, both while an NVM Subsystem Shutdown is in progress, and after its completion until a subsequent NVM Subsystem Reset. For Controller Reset, this is an incompatible change.
- While a Controller Shutdown is in progress and after its completion, all forms of Controller Level Reset permit (but do not require) the controller to initialize media and cease being ready to be powered off. For Controller Reset, this is an incompatible change.
- After an NVM Subsystem Shutdown, the CSTS.ST and CSTS.SHST values are preserved until an NVM Subsystem Reset. Any other Controller Level Reset does not change these values.
- While controller shutdown processing is reported as in progress or is reported as complete (i.e., CSTS.ST is cleared to ‘0’ and CSTS.SHST is set to 01b or is set to 10b), if the controller becomes ready (i.e., sets CSTS.RDY to ‘1’ from ‘0’) as a consequence of the host enabling the controller (i.e., the host sets CC.EN to ‘1’ from ‘0’), then any shutdown processing stops and CSTS.SHST is cleared to 00b before the controller becomes ready.

1.2.24 Boot Partition Write Protection – TP4170 (optional with mandatory components)

Added a new write protection control mechanism for Boot Partitions that is controllable through the Set Features command. Additionally, this proposal makes the RPMB based Boot Partition Protection mechanism optional.

1.2.24.1 Description of change

- Add new Boot Partition write protection support fields in the Identify Controller data structure to identify support for the Set Features based Boot Partition Write Protection capability and the RPMB based Boot Partition Write Protection capability.
- Add a new feature to the Set Features command that can be used to configure the write protection states of both Boot Partitions. This new feature can put either Boot Partition into one of the following states: Write Locked, Write Unlocked, or Write Locked Until Power Cycle.

1.2.24.1.1 New requirement / incompatible change

- Previously RPMB Boot Partition Write Protection was required if Boot Partitions and RPMB were both supported. This adds a new Boot Partition write protection capability and only requires one Boot Partition write protection capability be supported by a controller that supports Boot Partitions, making RPMB Boot Partition Write Protection optional so long as Set Features Boot Partition Write Protection is supported by the controller.

1.2.25 Status Reporting Enhancements – TP6021 (mandatory if Telemetry log pages and NVMe-MI are supported)

Made a variety of enhancements and clarifications related to NVMe-MI status reporting.

1.2.25.1.1 Description of change

- Added a mechanism to report Telemetry Controller-Initiated Data Available via the out-of-band mechanism.

1.2.26 Out-of-Band Admin Command While Shutdown Enhancement – TP6032 (mandatory if NVMe-MI supported)

Defined interactions between NVMe Admin Commands processed via out-of-band mechanisms that access media and shutdown.

1.2.26.1 Description of change

- Added Additional Unsafe Shutdown Logging
  - Added a new type of Unsafe Shutdown where the controller is in the shutdown state when power is lost but media is not in the shutdown state due to a Management Endpoint processing NVMe Admin Commands that access media with the Ignore Shutdown bit set to ‘1’ via the out-of-band mechanism.
  - Added a new bit (Unsafe Shutdown Reason) to the Persistent Event Log to indicate whether an Unsafe Shutdown was due to a Management Endpoint bringing media out of a shutdown state to process NVMe Admin Commands that access media with the Ignore Shutdown bit set to ‘1’ via the out-of-band mechanism.

1.2.27 Unique Discovery Controller ID - TP8013a, (optional)

Defined a unique NQN which Discovery subsystems can supply when performing NVMe in-band authentication using DH-HMAC-CHAP, or for other purposes (e.g., for hosts to detect if Discovery controllers are contained in different Discovery subsystems or not).

1.2.27.1.1 Description of change

- Added guidance regarding Discovery subsystem support for unique NQN and well-known Service NQN
- Added guidance for host choice for which NQN to use when connecting to a Discovery subsystem that provides a unique NQN

1.2.28 Discovery Subsystem Authentication Recommendations - TP8017a (optional)

Section named “NVMe In-band Authentication” was modified to clarify the behavior of a host when connecting to a Discovery subsystem that requires authentication. In addition, updated all instances of “unique NQN” to “unique Discovery Service NQN” to be consistent with the term “well-known Discovery Service NQN”.

1.2.28.1.1 Description of change

- Updated wording of the Discovery Controller section to convert instances of “unique NQN” to “unique Discovery Service NQN”
- Updated wording in the NVMe In-band Authentication section to provide host guidance when connecting to a Discovery subsystem that requires authentication

1.2.29 NVMe/TCP – TLS updates – TP8018 (mandatory if TLS is supported)

Redesign TLS concatenation to DH-HMAC-CHAP

TLS requirement updates for NVMe/TCP:

1.2.29.1 Description of change

1.2.29.1.1 New requirement / incompatible change

- TLS concatenation redesign (required if TLS concatenation supported):
  - Replace TLS PSK concatenation mechanism that proceeded from DH-HMAC-CHAP to TLS on same TCP connection with a new mechanism that disconnects TCP connection after TLS PSK generation and uses TLS PSK on new connections. This improves security by applying TLS security to the Connect command on any connection that is used for NVMe commands and prevents common implementation flaws that could allow NVMe commands to bypass TLS on the same connection.
  - Restrict TLS PSK generation to Admin Queue connections. A generated TLS PSK may be used on I/O Queue connections and other Admin Queue connections. This simplifies implementations.
  - Use new format of TLS PSK identity that uniquely identifies TLS PSK. Old format is obsoleted.
  - Distinguish generation of a new TLS PSK from replacement of an existing TLS PSK to support TLS PSK rollover.
- TLS requirement updates for NVMe/TCP (required):
  - Forward secrecy: Remove requirement for PSK-only authentication. PSK with ECDHE remains required. Change mandatory DH group for forward secrecy to p384 curve.
  - Channel binding: Require NQNs in TLS PSK identity and Connect command to match.
  - TLS handshake protection: Enforce existing restrictions that sending any other traffic when a TLS handshake is expected or in progress is an error that closes the TCP connection and that sending unencrypted traffic after a successful TLS handshake is also an error that closes the TCP connection.
  - TLS functionality limitation or discouragement:
  - Strongly discourage TLS 1.2 in favor of TLS 1.3 for improved security.
  - Prohibit Application Layer Protocol Negotiation (ALPN) as it is inapplicable to NVMe/TCP.
  - Strongly discourage TLS 1.3 session resumption.
  - State that some environments may require stronger crypto than minimum implementation requirements, e.g., environments that follow CNSA 1.0 implementation guidance.
  - Specify error if TLS PSK generation attempted with NULL DH group (because that doesn’t work).

1.2.30 Authentication Verification Entity for DH-HMAC-CHAP – TP8019a (optional)

Enable an NVMe-oF entity to delegate DH-HMAC-CHAP authentication verification to an Authentication Verification Entity.

1.2.30.1 Description of change

- Authentication Verification Entity (AVE) for DH-HMAC-CHAP

1.2.31 UTF-8 Admin Labels - TP8020 (optional)

Defined the concept of a UTF-8 string and proposes to replace the existing term “symbolic name” with “Admin Label” and defines how Hosts and NVM subsystems may register and retrieve an Admin Label in UTF 8 format in addition to ASCII.

Registration is accomplished by defining a new extended attribute in the Discovery Information Management (DIM) command.

Retrieval of a UTF 8 encoded Admin Label will be accomplished by defining a new Extended Attribute that could be included in an Extended Discovery Log Page Entry.

1.2.31.1 Description of change

- Modified existing usage of “symbolic name” to “admin label ASCII”
- Added a new definition for “admin label UTF-8”
- Added a new Extended Attribute Type (EXATTYPE) for “admin label UTF-8”
- Added a new Extended Attribute Length (EXATLEN) for “admin label UTF-8”
- Added a new Extended Attribute Value (EXATVAL) for “admin label UTF-8”
- Define UTF-8 string format and usage.

1.2.32 Connect Command reconciliation – TP8021 (mandatory if NVMe over Fabrics supported)

Added features from the Create I/O Submission Queue command (for PCIe) to the Connect command (for Fabrics).

1.2.32.1 Description of change

- The capability to specify an NVM Set ID during the creation of a fabric connection (i.e., the mechanism that creates an I/O submission and completion queue for an NVMe-oF controller) was added.
- A field to specify an NVM Set ID is added to the Connect command replacing 2 reserved bytes.

1.2.33 mDNS Discovery update - TP8024a (mandatory if mDNS supported)

Obsoleted the DNS-SD subtype “_ddcpull._sub._nvme-disc”.

The “_ddcpull._sub._nvme-disc” subtype enabled Direct Discovery Controllers to announce this subtype and use it to signal Centralized Discovery Controllers that a pull registration should be performed.

DDCs may request a pull registration by using the Kickstart Discovery Request.

1.2.33.1 Description of change

- The description of the DNS-SD subtype “_ddcpull._sub” has been updated to indicate it is obsolete.
- All instances of the subtype “_ddcpull._sub” have been removed, along with any associated text describing how to use this subtype.

1.2.34 Modifications for I3C – TP6037 (optional)

Updates the name of the I2C/SMBus ports and communications to match the new conventions used by SNIA SFF TA’s EDSFF and PCI-SIG specifications for I3C.

1.2.34.1 Description of Change

- Updates terminology

1.2.35 Management Interface Miscellaneous Maintenance – TP6038 (Optional)

Addresses a number of miscellaneous maintenance issues. It will address various issues that impact interoperability between a Management Controller (e.g., a BMC) and a Management Endpoint (e.g., an NVMe Storage Device).

1.2.35.1 Description of Change

- Clarified that for firmware activation without reset, the Firmware Commit command is completed after new firmware is activated
- Added new field to the Identify Controller data structure: Maximum Time for Firmware Commit Processing (MTFCP)

1.3 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

1.3.1 Should to Shall Conversion - TP4136 (mandatory)

Identified instances of “should” statements that were introduced into the various NVMe specifications via ECNs and changes them to “shall” statements where appropriate. These “should” statements that are changing to “shall” statements are for behaviors that were strong recommendations during their initial incorporation and were intended to be converted into hard requirements at a later point.

1.3.1.1 Description of change

1.3.1.1.1 New requirement / incompatible change

- Attempting to create an I/O queue before initializing I/O Completion Queue Entry Size (CC.IOCQES) and I/O Submission Queue Entry Size (CC.IOSQES) shall cause a controller to abort a Create I/O Completion Queue command or a Create I/O Submission Queue command with a status code of Invalid Queue Size.
- If the Fabrics Command Type (FCTYPE) field in a Fabrics command capsule is set to a reserved value, the command shall be aborted with a status code of Invalid Field in Command.
- If the namespace identifier is used for the command, the value FFFFFFFFh is not supported for that command, and the host specifies a value of FFFFFFFFh, then the controller shall abort the command with a status code of Invalid Field in Command, unless otherwise specified.
- If the namespace identifier is not used for the command and the host specifies a value from 1h to FFFFFFFFh, then the controller shall abort the command with a status code of Invalid Field in Command, unless otherwise specified.
- If the host submits a Firmware Commit command on the controller with a Commit Action of 011b, the downloaded firmware image shall replace the firmware image in the firmware slot.
- If Asynchronous Event Request commands are outstanding when the controller is reset, then each of those commands is aborted and shall not return a CQE.
- If the Physically Contiguous bit in Command Dword 11 of a Create I/O Completion Queue command is cleared to '0' and CAP.CQR is set to '1', then the controller shall return an error of Invalid Field in Command.
- If the value specified in the Completion Queue Identifier (CQID) field in Command Dword 11 of a Create I/O Submission Queue command is 0h (i.e., the Admin Completion Queue), then the controller shall return an error of Invalid Queue Identifier.
- If the value specified in the Completion Queue Identifier (CQID) field in Command Dword 11 of a Create I/O Submission Queue command is outside the range supported by the controller, then the controller shall return an error of Invalid Queue Identifier.
- If the value specified in the Completion Queue Identifier (CQID) field in Command Dword 11 of a Create I/O Submission Queue command is within the range supported by the controller and does not identify an I/O Completion Queue that has been created, then the controller shall return an error of Completion Queue Invalid
- If there are I/O commands being processed for a namespace, then a Format NVM command that is submitted affecting that namespace may be aborted; if aborted, then a status code of Command Sequence Error shall be returned.
- If a Format NVM command is in progress, then an I/O command that is submitted for any namespace affected by that Format NVM command may be aborted; if aborted, then a status code of Format in Progress shall be returned.
- For log pages with a scope of NVM subsystem or controller, the controller shall abort commands that specify namespace identifiers other than 0h or FFFFFFFFh with a status code of Invalid Field in Command.
- If the SMART / Health Information log page is not supported on a per namespace basis and if a namespace identifier other than 0h or FFFFFFFFh is specified by the host, then the controller shall abort the command with a status code of Invalid Field in Command.
- For the Active Namespace ID list, the controller shall abort the Identify (CNS 02h) command with a status code of Invalid Namespace or Format if the NSID field is set to FFFFFFFEh or FFFFFFFFh.
- For the Allocated Namespace ID list, the controller shall abort the Identify (CNS 10h) command with a status code of Invalid Namespace or Format if the NSID field is set to FFFFFFFEh or FFFFFFFFh.
- For the Identify Namespace data structure for an Allocated Namespace ID, if the NSID field is set to FFFFFFFFh, then the controller shall abort the Identify (CNS 11h) command with a status code of Invalid Namespace or Format.
- For the Namespace Attached Controller list, if the NSID field is set to FFFFFFFFh, then the controller shall abort the Identify (CNS 12h) command with a status code of Invalid Field in Command.
- If the Firmware Commit command that established a pending firmware activation with reset condition returned a status code of: Firmware Activation Requires Controller Level Reset, Firmware Activation Requires Conventional Reset, or Firmware Activation Requires NVM Subsystem Reset, then the controller shall abort the Sanitize command with that same status code.

- If the Firmware Commit command that established a pending firmware activation with reset condition completed successfully or returned a status code other than: Firmware Activation Requires Controller Level Reset, Firmware Activation Requires Conventional Reset, or Firmware Activation Requires NVM Subsystem Reset, then the controller shall abort the Sanitize command with a status code of Firmware Activation Requires Controller Level Reset.
- If a volatile write cache is not present, then a Get Features command specifying the Volatile Write Cache feature identifier shall abort with a status code of Invalid Field in Command.
- For the Reservation Notification Mask feature, a Get Features (FID 82h) command that uses a namespace ID of FFFFFFFFh shall be aborted with status code of Invalid Field in Command.
- For the Reservation Persistence feature, a Get Features (FID 83h) command that uses a namespace ID of FFFFFFFFh shall be aborted with a status code of Invalid Field in Command.
- The Reservation Persistence feature shall not be saveable.
- If the Namespace Management capability is supported, then the controller shall support the Namespace Attribute Changed asynchronous event.
- If the Preempt Reservation Key (PRKEY) value is non-zero and there are no registrants whose reservation key matches the value of the PRKEY field, the controller shall return an error of Reservation Conflict.

1.3.2 ECN101 (mandatory)

1.3.2.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.2.1.1 New requirement / incompatible change

- If the LBA Format Upper (LBAFU) field and LBA Format Lower (LBAFU) field in Command Dword 10 of the Format NVM command specifies an unsupported USER DATA Format, then the controller shall about the command with a status code of Invalid Format.
- A Telemetry Host-Initiated log page may have either controller scope or NVM subsystem scope.
- The Total NVM Capacity (TNVMCAP) field and the Unallocated NVM Capacity (UNVMCAP) field in the Identify Controller data structure are required to be supported if the Namespace Management capability is supported or if the Capacity Management capability is supported.
- Bit 0 in the Common Namespace Features (NSFEAT) field of the I/O Command Set Independent Identify Namespace data structure is reserved.

1.3.3 ECN102 (mandatory)

1.3.3.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.3.1.1 New requirement / incompatible change

- The Commands Supported and Effects log page is mandatory to be supported.
- The Disconnect command is prohibited from being supported on an Administrative controller or Discovery controller.
- If two or more Endurance Groups are supported, then the Endurance Group Event Aggregate log page shall be supported.
- The Error Count field in the Error Information Log Entry data structure is a 64-bit counter that rolls over from FFFFFFFF_FFFFFFFFh to 1h when incremented (i.e., rolls over to 1h).
- If the Element Length (ELEN) field is cleared to 0h when adding/updating an entry, in the Host Metadata features (i.e., Feature Identifier 7Dh, 7Eh, 7Fh), then the controller behavior is undefined.

1.3.4 ECN105 (mandatory)

1.3.4.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.4.1.1 New requirement / incompatible change

- The reporting of a status code was being returned in a field greater than 15-bits in size. That was corrected by updating the reported data structure to clearly identify where the 15-bit value is located. The following fields do not define the proper location of the status code:
  - The Status Info (INFO) field in the Format NVM Completion Event of the Persistent Event log page
  - The Status Info (STS) field in the Authentication Receive Response
  - The Status Info (STS) field in the Authentication Send Response
  - The Status Info (STS) field in the Connect Response

        - The Status Info (STS) field in the Disconnect Response
        - The Status Info (STS) field in the Property Get Response
        - The Status Info (STS) field in the Property Set Response
    - If the amount of data requested to be transferred exceeds the total length of the Data Block and Bit Bucket descriptors in an SGL, data shall not be transferred to or from locations that are not described by the SGL.
    - Deleted the requirement that a controller shall not send the Attached Namespace Attribute Changed event if the Namespace Status has changed and shutdown processing is either being reported as occurring (i.e., the CSTS.SHST field is set to 01b) or reported as being complete (i.e., the CSTS.SHST field is set to 10b).
    - If the value in the Queue ID field for a Connect command specifies the Queue ID of a queue that already exists, then the controller shall abort the command with a status code of Invalid Queue Identifier.

1.3.5 ECN109 (mandatory)

1.3.5.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.5.1.1 New requirement / incompatible change

- A Get Log Page command specifying the Persistent Event log page and a Release Context in the Action field shall not return Persistent Event log page data.

1.3.6 ECN110 (mandatory)

1.3.6.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.6.1.1 New requirement / incompatible change

- The status code returned for the Get Log Page command for a Log Identifier not support is Invalid Log Page.
- In the Identify command with CNS 05h removed the past tense restriction on the I/O Command Set specified by the CSI field being enabled. The specified CSI value is based on currently being enabled to return data by the command.
- The value of the Over Temperature field in a Thermal Excursion Event of the Persistent Event log page contains Contains the absolute value of the difference (i.e., delta) in Kelvins between the temperature indicated in the threshold field and temperature measured at the time of the event.
- If a Get Log Page command is processed that specifies a Log Identifier that is not supported, then the controller shall abort the command with a status code of Invalid Log Page for controllers compliant with NVM Express Base Specification revision 2.0 and later.

1.3.7 ECN111 (mandatory)

1.3.7.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.7.1.1 New requirement / incompatible change

- Events are not preserved across Controller Level Reset as opposed to power off conditions.
- When the Spinup Control feature is enabled that the inhibiting is until any controller in the domain processes a Set Feature command specifying an operational state.

1.3.8 ECN113 (mandatory)

1.3.8.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.8.1.1 New requirement / incompatible change

- Support for the NVMe-MI Commands Supported and Effects log page is optional if the NVMe MI Send command and NVMe MI Receive command is not supported.
- A controller shall abort a Set Features command for the Host Memory Buffer feature if the Host Memory Descriptor List Entry Count field is cleared to 0h.
- If a controller supports more than one I/O Command Set, then the bits in the Optional NVM Command Support (ONCS) field in the Identify Controller data structure associated with command support reflect the

aggregate support across all of the I/O Command Sets supported or I/O Command Sets inherited from I/O Command Sets (e.g., Zoned Namespace Command Set).

1.3.9 ECN114 (mandatory)

1.3.9.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.9.1.1 New requirement / incompatible change

- The Asynchronous Event Request Limit field in the Identify Controller data structure is mandatory for Discovery controllers that support explicit persistent connections.

1.3.10 ECN115 (mandatory)

1.3.10.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.10.1.1 New requirement / incompatible change

- The controller shall abort an Identify command specifying CNS 03h with a status code of Invalid Namespace or Format if the NSID field is set to FFFFFFFEh or is set to FFFFFFFFh.

1.3.11 ECN118 (mandatory)

1.3.11.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.11.1.1 New requirement / incompatible change

- The Unexpected Power Losses (UPL) field (previously named the Unsafe Shutdowns field) in the SMART / Health Information log page updated shutdown cases redefined to include shutting down all controllers in the domain or NVM subsystem using a controller shutdown (as opposed to an NVM Subsystem Shutdown).
- The value of 10b (i.e., Not required) in the Transport Secure Channel field in the Discovery Log Page Entry data structure definition was clarified to include not supported and supported but not required. Some might consider this requirement change.

1.3.12 ECN119 (mandatory)

1.3.12.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.12.1.1 New requirement / incompatible change

- Namespace Write Protect Until Power Cycle is prohibited if the controller and namespace are in different domains because the namespace and controller have to be simultaneously power cycled to clear that state.

1.3.13 ECN120 (mandatory)

1.3.13.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.13.1.1 New requirement / incompatible change

- A Virtualization Management command to place the secondary controller in a state (i.e., Online, Offline) when that secondary controller is in that state is not considered an error and that command shall complete successfully.
- Corrected contradictory text on controller behavior for SGL Data Blocks by removing the shall requirement duplicating text in the table describing the SGL Data Block Address and Length fields.
- The effects of resets to different scoped Features has been updated to reflect the differences between resets of:
  - an NVM subsystems with a single controller;
  - an NVM subsystems with multiple controllers;
  - a domain that does reset the entire NVM subsystem when reset; and
  - a domain that does not reset the entire NVM subsystem when reset.
- Administrative controllers are prohibited from supporting the following fields in the Identify Controller data structure:
  - I/O Queue Command Capsule Supported Size (IOCCSZ) field;
  - I/O Queue Response Capsule Supported Size (IORCSZ) field; and

- In Capsule Data Offset (ICDOFF) field.

1.3.14 ECN122 (mandatory)

1.3.14.1 Description of change due to the ECN and any technical proposal that affected the requirement

1.3.14.1.1 New requirement / incompatible change

- Reservation Report command to be registrant-oriented rather than controller-oriented. E.G., changed all occurrences of “Registered Controller” to “Registrant”. This allows the command to return entries for all registrants of a namespace, including those that are not associated with any controller in the NVM subsystem (previously such registrants were not able to be reported). This is an incompatible change.
- The following data structure name changes and field name change occurred:
  - The “Number of Registered Controllers (REGCTL)” field name was renamed to “Number of Registrants (REGSTRNT)”
  - The “Registered Controller” data structure name was renamed to “Registrant” data structure
  - The “Registered Controller Extended” data structure name was changed to “Registrant Extended” data structure
- Expanded meaning of value FFFFh in CNTLID field to be used in the case of any registrant whose host identifier is not associated with a controller in the NVM subsystem. This covers PCIe controllers, NVMe-oF static controllers and NVMe-oF dynamic controllers alike. This may be considered an incompatible change.
- The starting of the Vendor Specific controller properties for Memory-Based transport implementation was changed to begin after the Transport Specific controller properties area. The memory based PCIe transport is able to exceed the 1300h boundary previously defined.

2 Changes to the NVM Express® NVM Command Set Specification

2.1 New Features

2.1.1 Flexible Data Placement (FDP) – TP4146b (optional)

2.1.1.1 Description of change

- Added fields to the NVM Command Set I/O Command Set Specific Identify Namespace data structure.
- Modified the Namespace Management command to specify creation of a namespace in non-volatile storage that has Flexible Data Placement enabled.
- Added directives to the Write Zeroes command and Write Uncorrectable command.
- Added the definition for the Reclaim Unit Handle Statue Descriptor in the I/O Management Receive command.
- Added the definition of the Media Reallocation Event Type for the Flexible Data Placement Events log page.

2.1.2 PCIe Infrastructure for Live Migration – TP4159 (optional)

Defined an LBA Migration Queue that is associated with a User Data Migration Queue.

2.1.2.1 Description of change

- Defined requirements and entries for an LBA Migration Queue (i.e., a type of Controller Data Queue) used to track changes to logical blocks in namespaces attached to a controller. The queue is created and deleted using the Controller Data Queue command. Posting of entries is enabled and disabled using the Migration Send command.

2.1.3 Dispersed Namespaces – TP4034a (optional)

2.1.3.1 Description of change

- Added a clarification that dispersed namespaces follow the same atomicity requirements as shared namespaces that are not dispersed namespaces

2.1.4 Key Per I/O – TP4055 (optional)

2.1.4.1 Description of change

2.1.4.1.1 New requirement

- Identify Namespace data structure extended to include Key Per I/O Alignment granularity
- Moved DSM to a directive to make space in Dword 13 for a 16-bit key tag field for Key Per I/O
- Added CETYPE and CEV to I/O commands.

2.1.5 Performance Characteristics Reporting – TP4077 (optional)

2.1.5.1 Description of change

Added the new Performance Characteristics feature

- Defines three types of performance attributes:
  - Standard Performance Attribute indicates random 4KiB average read latency (not changeable and not saveable)
  - Performance Attribute Identifier List (not changeable and not saveable)
  - Vendor Specific Performance Attributes (changeable / saveable)
- The Set Features command can revert a Vendor Specific Performance Attribute.

2.1.6 NVMe Specification Version Reporting – TP4135 (mandatory)

2.1.6.1 Description of change

Added a Version (VER) field to the I/O Command Set Specific Identify Controller data structure for the NVM Command Set.

2.1.7 Subsystem Local Memory Command Set Specification changes to other specifications – TP4171 (optional)

2.1.7.1 Description of change

- Modifications to Copy command

2.2 Feature Enhancements

2.2.1 Defining Scope for Features – TP4074a (mandatory)

2.2.1.1 Description of change

- Added scope field in the Feature Identifiers table.
- Added explicit text to each feature section that states the feature's scope.

2.2.2 Enhanced Deallocation Granularity – TP4090 (optional with mandatory components)

Expanded the ability for a controller to indicate the preferred deallocation granularity to a 32-bit value in a manner that is backwards compatible so 16-bit value is maintained.

2.2.2.1 Description of change

- Adds the Namespace Preferred Deallocate Granularity Large (NPDGL) field to the NVM Command Set specific Identify Namespace data structure (CNS 05h, CSI 00h). This is a 32-bit version of the existing NPDG field in the Identify Namespace data structure (CNS 00h). The width of the new field matches the width of the Length in Logical Blocks field in the range definition of the Dataset Management command.
- Expands OPTPERF from one to two bits, to simplify describing requirements for supporting the legacy (16-bit) and new (32-bit) field.
- Capitalizes the Length in Logical Blocks field in the range definition.
- Defines the usage of the new NPDGL field.
- Clarifies the SGS field, which is a multiplier for the SWG field.

2.2.3 Namespace Capability Reporting- TP4095a (mandatory if Namespace Management supported)

2.2.3.1 Description of change

- Added CNS value 09h to the Identify command to return an NVM Command Set Identify Namespace data structure associated with the LBA Format specified by the Format Index. The NVM Command Set Identify Namespace data structure is updated to identify the fields that report valid values.
- Added CNS value 0Ah to the Identify command to return an Identify I/O Command Set Specific Namespace data structure for the NVM Command Set associated with the LBA Format specified by the Format Index. The Identify I/O Command Set Specific Namespace data structure for the NVM Command Set is updated to identify the fields that report valid values.

2.2.3.1.1 New requirement / incompatible change

  - The Identify command for CNS value 00h removed the term "common" and clearly specified the requirements on all fields when the specified NSID is set to FFFFFFFFh
  - The requirement to abort an Identify command with CNS values 00h, 05h, and 08h, when the specified NSID is FFFFFFFFh and Namespace Management capability is not supported, was changed to optional (previously, aborting the command was mandatory).
  - The LBA Format list structure is extended to have both the existing group of LBA Formats that have the same capabilities for creating and formatting a namespace as well as a group of LBA Formats that each may have a unique set of capabilities for creating and formatting a namespace.

2.2.4 Multiple Atomicity – TP4098a (optional)

Defined a new Multiple Atomicity Mode (MAM) that provides write guarantees between Namespace Atomic Boundaries for commands that cross Namespace Atomic Boundaries. The legacy atomicity behavior is still supported and is referred to as Single Atomicity Mode with no functional changes.

2.2.4.1 Description of change

- Defined a new Multiple Atomicity Mode (MAM) that provides write guarantees between Namespace Atomic Boundaries for commands that cross Namespace Atomic Boundaries.
- Atomicity behavior specified in the NVM Command Set Specification: Revision 1.0 and prior NVMe standards is referred to as Single Atomicity Mode with no functional changes.
- Added an MAM bit to the Namespace Features (NSFEAT) field in the I/O Command Set Specific Identify Namespace data structure (CSI 00h) to indicate that Multiple Atomicity Mode applies to writes to a namespace.
- Added a new section to specify normative rules for Multiple Atomicity Mode.

2.2.5 MDTS Enhancement – TP4099 (mandatory)

2.2.5.1 Description of change

- Updated the I/O Command Set Specific Identify Controller data structure for the NVM Command Set fields to reflect the requirements associated to supporting Multiple Atomicity.

2.2.6 ZNS Namespace Management Enhancements – TP4115 (optional)

Improved the management of zoned namespaces by allowing the host to specify how many zone resources (e.g., active resources, open resources, ZRWA resources) are allocated to a specific namespace when that namespace is created. Resources are statically allocated to a namespace and the resources for that namespace cannot be modified unless the namespace is deleted and re-created.

2.2.6.1 Description of change

- Reserved space for 3 new fields in the Namespace Management – Host Software Specified Fields: Requested Active Resources (RAR), Requested Open Resources (ROR), and Requested Number of ZRWA Resources Management (RNUMZRWA)

2.2.7 Optimal Read Size and Granularity – TP4116 (optional)

Allows a controller to indicate the optimal size and granularity for reading user data.

2.2.7.1 Description of change

- Defined read performance attributes NPRG, NPRA, NORS.
- Added new sub section in Improving Performance through I/O Size and Alignment Adherence separating section that describes how to use the new attributes to achieve improved read performance.
- Restructured Improving Performance through I/O Size and Alignment Adherence separating section write and read performance into new sections.

2.2.8 Cross-Namespace Copy TP4130a (mandatory components if Copy command is supported)

Extend Copy command to copy data across different namespaces on the same NVM subsystem via support for specifying a namespace for each source range. Extend Protection Information (PI) support to cross-namespace copy, including the ability to insert or strip PI in common cases where all logical block metadata is PI. Clarify ambiguities in originally specified atomicity properties of original Copy command (also applies to Copy across namespaces).

2.2.8.1 Description of change

- Extend existing Copy command to copy across namespaces - two new source range descriptor formats that include a source namespace for each source range (source namespace may be the same as destination namespace).

Hosts have to explicitly enable these formats via an extension of the Host Behavior Support feature. New status codes for errors that can only occur when the new formats are used.
- Extend Protection Information (PI) support to cross-namespace copy, including ability to insert or strip PI in common cases where all logical block metadata is PI.
- New Fast Copy Only (FCO) bit in new source range descriptor formats to enable hosts to avoid slow copying. Setting the FCO bit to '1 causes controller to abort Copy command (with a new status code) instead of performing slow copying.
- Update Reservation Conflict table to separate reservation checks for read and write portions of the Copy command.
- Copy command atomicity clarification (mandatory if NVM Command Set Copy command is supported, including ZNS Command Set usage of that Copy command)
- Add a new command-specific status code for Insufficient Resources and specify its use with the Copy command.

2.2.8.1.1 New requirement / incompatible change

- in the new Copy atomicity section:
  - Some implementations of the Copy command apply atomicity requirements only to each source range. New implementations are required to apply those requirements to the entire Copy command and set a (new) bit in Identify Controller to indicate this behavior.
  - The FFFFh value (0's-based) for atomicity parameters (e.g., AWUN – Atomic Write Unit Normal) only promises Copy command atomicity up to 10000h logical blocks. It does not promise that all Copy commands are atomic.

2.2.9 Format Index Mismatch Protection – TP4140 (mandatory if LBAFEE supported)

Updated the definition of the LBA Format Extension Enable (LBAFEE) field to abort any Format NVM command or I/O command to a namespace that was formatted with a Format Index that is greater than number of LBA Formats reported by the controller in the Identify Namespace data structure.

2.2.9.1 Description of change

Update the definition of the LBAFEE bit in the Host Behavior Support data structure to return error on commands which attempt to either Format or perform I/O on namespaces currently formatted with a Format Index greater than the currently reported total number of LBA formats supported (in NLBAF + NULBAF (if it exists))

2.2.10 Storage Tag Mask Enhancements – TP4141a (optional)

Defined optional support for Storage Tag restrictions by creating two levels of restricted support for Storage Tag Masking: byte granularity masking and masking not supported.

2.2.10.1 Description of change

- Added the Qualified Protection Information Format Support (QPIFS) field to the NVM Command Set I/O Command Set Specific Identify Namespace data structure (CNS 05h) (CSI 00h)
- Updated the 16BPISTS field to take the QPIFS field into account
- Added the Qualified Protection Information Format (QPIF) field to the Extended LBA Format Data Structure, NVM Command Set specific. This field indicates Protection Information Formats qualified by the Storage Tag Mask constraints.
- Defined a new coded value (QTYPE = 11b) to the PIF field which points to Protection Information Formats that are defined by the QPIF field
- Added the Storage Tag Masking Level (STML) field to the NVM Command Set I/O Command Set Specific Identify Namespace data structure (CNS 05h) (CSI 00h)
- Updated the legal LBSTM field options to be qualified by the value in new Storage Tag Masking Level (STML) field.

2.2.10.1.1 New requirement / incompatible change

- QPIFS field bit assignment was changed from bit 2 to bit 3.

2.2.11 Enhanced Namespace Preferred Deallocation Alignment – TP4148 (optional)

Allows a controller to advertise a 32-bit preferred deallocation alignment value that is larger than is presently possible with the 16-bit value.

2.2.11.1 Description of change

- Added new Namespace Preferred Deallocate Alignment Large (NPDAL) field to the NVM Command Set specific Identify Namespace data structure (CNS 05h, CSI 00h). This is a 32-bit version of the existing NPDA field in the Identify Namespace data structure (CNS 00h).
- Defined the usage of the new NPDAL field.

2.2.12 Clarifications to Namespace Changes Reporting – TP4150 (optional with mandatory components)

2.2.12.1 Description of change

- Renamed Namespace Attribute Changed asynchronous event to Attached Namespace Attribute Changed
- Defined Allocated Namespace Attribute Changed asynchronous event (references NVM Express Base Specification)

2.2.13 WZSL Limit Modification – TP4160 (optional)

2.2.13.1 Description of change

- Added new Write Zeroes with Deallocate Size Limit field in the Identify Controller Data Structure
- Added descriptive text in the Write Zeroes command section 3
- Added new optional bit in the Write Zeroes command Dword 12 that writes zeroes to the entire namespace

2.2.14 Non-Storage namespace changes – TP4162a (optional)

2.2.14.1 Description of change

- Change the definition of namespace

2.2.15 Tracking LBA Allocation with Granularity – TP4165 (optional with mandatory components)

Extended the Get LBA Status command to report allocated LBAs managed by the controller at a indicated granularity.

2.2.15.1 Description of change

- The Get LBA Status command supports a new Action Type to allow a host to request reporting of allocated LBAs of a namespace. A controller identifies which Action Types are supported by the Get LBA Status command.

2.2.15.1.1 New requirement and incompatible change

  - A controller is allowed to support the Get LBA Status command and not support the Action Type values of 10h and 11h for reporting Tracked LBAs and Untracked LBAs.

2.2.16 TP4175 - Copy and Reachability (mandatory if Copy command is supported)

Defined the relationship between Reachability and the Copy command.

2.2.16.1 Description of change

- Associated the error from the Reachability Architecture (TP4156) in the context of the Copy command (TP4131).

2.2.16.1.1 New requirement and incompatible change

  - Added a new ‘Namespace Not Reachable’ Copy command specific status value

2.2.17 Logical Block Storage Tag Mask Caveats – TP4186 (optional)

Defined expected behavior of the controller when a Format NVM command is processed that specifies an LBA Format that includes a Storage Tag. A controller shall only permit a Format NVM command to be processed that specifies an LBA Format that includes a Storage Tag if all namespaces affected by that command were created with that same LBA Format.

A controller shall not permit a Format NVM command to be processed that specifies an LBA Format that includes a Storage Tag in any other case. A host may choose to delete one or more namespaces that were created with a different LBA Format and then create one or more new namespaces with an LBA Format includes a Storage Tag

2.2.17.1 Description of change

2.2.17.1.1 New requirement and incompatible change

- Only permit a Format NVM command that specifies an LBA Format that includes a Storage Tag to be processed if all the namespaces affected by that command were created with that same LBA Format

2.3 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

2.3.1 Should to Shall Conversion – TP4136 (mandatory)

Identified instances of "should" statements that were introduced into the various NVMe specifications via ECNs and changes them to "shall" statements where appropriate. These "should" statements that are changing to "shall" statements are for behaviors that were strong recommendations during their initial incorporation and were intended to be converted into hard requirements at a later point.

2.3.1.1 Description of change

2.3.1.1.1 New requirement / incompatible change

- For the LBA Range Type feature, if the controller checks for LBA range overlap and the controller detects an LBA range overlap, then the controller shall return an error of Overlapping Range.
- The default value for the LBA Range Type feature shall clear the Number of LBA Ranges field to 0h (i.e., one LBA Range is present).
- For the Identify Namespace data structure for an Allocated Namespace ID, if the NSID field is set to FFFFFFFFh, then the controller shall abort the Identify (CNS 11h) command with a status code of Invalid Namespace or Format.
- If the Reference Tag Check bit of the Protection Information Check (PRCHK) field is set to '1' and the namespace is formatted for Type 3 protection, then the controller may ignore the Initial Logical Block Reference Tag (ILBRT) and Expected Initial Logical Block Reference Tag (EILBRT) fields. If a command is aborted as a result of the Reference Tag Check bit of the PRCHK field being set to '1', then that command shall be aborted with a status code of Invalid Protection Information.

2.3.2 ECN101

2.3.2.1 Description of change

2.3.2.1.1 New requirement / incompatible change

- All versions of the NVM Express NVM Command Set specification require the Protection Information Location (PIL) bit be cleared to '0' in Command Dword 10 of the Format NVM command. Protection Information is only allowed to be supported in the last bytes of metadata.
- All versions of the NVM Express NVM Command Set specification require the Protection Information Position (PIP) bit be cleared to '0' in Identify Namespace data structure. Protection Information is only allowed to be supported in the last bytes of metadata.
- The following fields in the Identify Namespace data structure are now reserved if the OPTPERF field is cleared to 0h. In the previous version of the NVM Express NVM Command Set Specification, OPTPERF

has a bit definition. If the OPTPERF field is set to 10b, then hosts complaint to NVM Express NVM Command Set Specification 1.0 would have these field reserved, so this change is an incompatible change:
    - Namespace Preferred Write Granularity (NPWG)
    - Namespace Preferred Write Alignment (NPWA)
    - Namespace Preferred Deallocate Granularity (NPDG)
    - Namespace Preferred Deallocate Alignment (NPDA)
    - Namespace Optimal Write Size (NOWS)
  - Bit 0 in the Common Namespace Features (NSFEAT) field of the I/O Command Set Independent Identify Namespace data structure is reserved.

2.3.3 ECN102

2.3.3.1 Description of change

2.3.3.1.1 New requirement / incompatible change

- After a write command has completed without error, reads for that location which are subsequently submitted and return data, shall return the data from that was written by that write command and not an older version of the data from previous write commands provided the exceptions specified in section 2.1.4.3.2.

2.3.4 ECN104

2.3.4.1 Description of change

2.3.4.1.1 New requirement / incompatible change

  - Added the Storage Tag Check Read (STCR) bit to Command Dword 12 of the Copy command.
  - Added the Storage Tag Check Read Support (STCRS) bit in the Protection Information Capabilities (PIC) field of the I/O Command Set Specific Identify Namespace data structure for the NVM Command Set.

2.3.5 ECN113

2.3.5.1 Description of change

2.3.5.1.1 New requirement / incompatible change

  - If the host specifies a destination LBA range that overlaps with any LBAs specified in one or more of the Source Range entries, then upon completion of the Copy command, the data stored in each logical block in that overlapping destination LBA range may, within the constraints of the atomicity rules described in section 2.1.4, be from any of the one or more Source Range entries in which that LBA is contained. This is a result of the possibility that overlapping Source Range entries may be processed in any order.
  - The LBA field in the Error Information Log was described as containing the “first” LBA that experienced the error. The meaning of “first” was not defined (first in time vs. first in LBA ordering). This was clarified to indicate that the “lowest numbered” LBA is to be reported, rather than the LBA where the error was detected first (based on time of detection). This may be considered by some to be an incompatible change.

2.3.6 ECN115

2.3.6.1 Description of change

2.3.6.1.1 New requirement / incompatible change

  - Subsequent LBA Status Descriptor Entries returned by a Get LBA Status command are allowed to have the start LBA reported be the first LBA after the previous LBA Status Descriptor Entry. Previously, the specification required subsequent LBA Status Descriptor Entries returned by a Get LBA Status command to have the start LBA reported be greater than the first LBA after the previous LBA Status Descriptor Entry. The intent was to allow a controller to specify more than one LBA Status Descriptor Entry for reported LBAs in a contiguous LBA range.

3 Changes to the NVM Express® Zoned Namespace NVM Command Set Specification

3.1 New Features

3.1.1 Flexible Data Placement – TP4146b (optional)

3.1.1.1 Description of change

- Modified Host Software Specified Fields to expand the range of bytes marked as “Refer to the NVM Command Set Specification”.

3.1.2 Zoned Random Write Area – TP4076b (optional)

Defines a non-volatile cache with several LBAs which start at the write pointer for a given zone, called a Zone Random Write Area (ZRWA). The logical blocks that are mapped to the ZRWA are allowed to be written in non-sequential order as well as overwritten. LBAs within the ZRWA are flushed to the zone sequentially.

3.1.2.1 Description of change

- Added new definitions for Zone Random Write Area (ZRWA) and Ceiling
- Zone Append may be used to write to logical blocks in a zone only if no ZRWA is associated with the zone
- For a zone associated with a ZRWA, command completion processing is defined in a new section
- Added another ZSC: ZSIO transition case – when an explicit ZRWA Flush operation is completed
- Added a new type of zone resource called ZRWA resource
- Added resource management rules for ZRWA resources
- Added two new Command Specific status codes: Invalid Zone Operation Request, ZRWA Resources Unavailable
- Added new Zone Descriptor Data Structure field called Zone Random Write Area Valid (ZRWAV)
- Added new Zone Management Send field called Zone Send Action Specific Option which controls the allocation of a ZRWA to a zone.
- Added a new Zone Send Action type in Zone Management Send command called Flush Explicit ZRWA Range.
- Added new Open Zone sub-section called Open Zone ZRWA Allocation
- Zone Management Send may return two new command specific status codes: Invalid Zone Operation Request and ZRWA Resources Unavailable
- Added new section to describe Flush Explicit ZRWA Range
- Added new ZRWA Supported (ZRWASUP) bit in the OZCS field of the I/O Command Set Specific Identify Namespace data structure for the Zoned Namespace Command Set
- Added new fields to the I/O Command Set Specific Identify Namespace data structure for the Zoned Namespace Command Set:
    - Number of ZRWA Resources (NUMZRWA)
    - ZRWA Flush Granularity (ZRWAFG)
    - ZRWA Size (ZRWASZ)
    - ZRWA Capability (ZRWACAP)
        - EXPFLUSHSUP bit within this field which defines support for explicit ZRWA Flush operations
- Added a model clause section for Zone Random Write Area

3.1.2.1.1 New requirement / incompatible change:

- A Zone Append command shall not be sent to a zone while that zone is associated with a ZRWA
- Definition of Starting LBA (SLBA) field in Zone Management Send command is changed if the Zone Send Action is Flush Explicit ZRWA Range
- Added new requirement to Reset Zone: The Zone Random Write Area Valid bit shall be cleared to ‘0’
- New qualification: writes are required to start at the valid write pointer address only if a ZRWA is not associated with a zone

3.1.3 NVMe Specification Version Reporting – TP4135 (mandatory)

Provides a means for the host to retrieve the versions of the I/O command set specifications and the NVMe Management Interface Specification.

3.1.3.1 Description of change

- Added the Version (VER) field to the I/O Command Set Specific Identify Controller data structure for the Zoned Namespace Command Set.

3.2 Feature Enhancements

3.2.1 Zone Relative Data Lifetime Hint – TP4093a (optional)

Allow a host to specify to a controller a relative lifetime hint upon opening a zone. The hint communicates the relative lifetime of the data that is expected to be subsequently written to that zone. A controller may utilize this information for managing data being written from different zones to the non-volatile storage medium.

3.2.1.1 Description of change

- Added Data Lifetime information to the Zone Management Send command.

3.2.2 Namespace Capability Reporting – TP4095a (mandatory if Namespace Management supported)

Created a mechanism to support LBA formats that do not report the same namespace attributes across all LBA formats and a mechanism to report the namespace attributes of LBA formats without having to create the namespace.

3.2.2.1 Description of change

- Added CNS value Ah to the Identify command to return an I/O Command Set Specific Identify Namespace data structure for the Zoned Namespace Command Set associated with the specified Format Index and Command Set Identifier. The I/O Command Set Specific Identify Namespace data structure for the Zoned Namespace Command Set Command is updated to identify the fields that report valid.

3.2.2.1.1 New requirement / incompatible change:

  - The Identify command for CNS values 05h removed the term “common” and clearly specified the requirements on all fields when the specified NSID is set to FFFFFFFFh
  - The requirement to abort an Identify command with CNS values 00h, 05h, and 08h, when the specified NSID is FFFFFFFFh and Namespace Management capability is not supported, was changed to optional (previously, aborting the command was mandatory).

3.2.3 ZNS Namespace Management Enhancements – TP4115 (optional)

Improved the management of zoned namespaces by allowing the host to specify how many zone resources (e.g., active resources, open resources, ZRWA resources) are allocated to a specific namespace when that namespace is created. Resources are statically allocated to a namespace and the resources for that namespace cannot be modified unless the namespace is deleted and re-created.

3.2.3.1 Description of change

- I/O Command Set Specific Identify Controller data structure
  - New resources fields for: Total Active Resources, Total Open Resources, Total ZRWA Resources
  - New resources fields for Unallocated values: Unallocated Active Resources, Unallocated Open Resources and Unallocated ZRWA Resources
  - New Identify field (Namespace Resource Management) in a new attribute (Zoned Controller Attributes)
  - New data structure in the Namespace Management command for Host Software Specified Fields

3.2.4 Post Sanitize Media Verification – TP4152 (optional)

3.2.4.1 Description of change

- During Media Verification state, logical block content is as defined in the NVM Command Set Specification.

3.2.4.1.1 New requirement / incompatible change:

- During Media Verification state, logical block content is as defined in the NVM Command Set Specification.

3.2.5 Non-Storage namespace changes – TP4162a (optional)

3.2.5.1 Description of change

Expanded the definition of namespace to include namespaces that are not used for storage (e.g., compute namespaces)

3.3 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

3.3.1 ECN102 (mandatory)

3.3.1.1 Description of change due to the ECN and any technical proposal that affected the requirement

3.3.1.1.1 New requirement / incompatible change

- The transition from the ZSIO: Implicitly Opened state or the ZSEO: Explicitly Opened state to the ZSF: Full state is allowed as a result of any write operations that writes one or more logical blocks that causes the zone to reach its writeable zone capacity.
- The Namespace Identifier field in the completion queue entry Dword 1 is reserved for an Asynchronous Event Request command if an event other than a Zoned Namespace Command Set Notice event is reported.

3.3.2 ECN105 (mandatory)

3.3.2.1 Description of change due to the ECN and any technical proposal that affected the requirement

3.3.2.1.1 New requirement / incompatible change

- A controller processing a command that requests a zone to transition to the ZSIO: Implicitly Opened state, the ZSEO: Explicitly Opened state, or the ZSC: Closed state shall, if:
  - resources are not available; and
  - as a result of the requested transition:
    - the resource that is not available is Open Resources; and
    - one or more zones are in the ZSIO: Implicitly Opened state,

  then the controller shall select one of the zones in the ZSIO: Implicitly Opened state, transition that zone to the ZSC: Closed state, and proceed to process the command.

4 Changes to the NVM Express® Key Value NVM Command Set Specification

4.1 New Features

4.1.1 Key Per I/O – TP4055 (optional)

4.1.1.1 Description of change

4.1.1.1.1 New requirement

- Added CETYPE and CEV to the Command Set specific I/O commands.

4.1.2 NVMe Specification Version Reporting – TP4135 (mandatory)

4.1.2.1 Description of change

- Added the I/O Command Set Specific Identify Controller data structure for the Key Value Command Set. The contents are the Version (VER) field.

4.2 Feature Enhancements

4.2.1 Defining Scope for Features – TP4074a (mandatory)

4.2.1.1 Description of change

- Added scope field in the Feature Identifiers table.

4.2.2 Namespace Capability Reporting – TP4095a (mandatory if Namespace Management supported)

Created a mechanism to support LBA formats that do not report the same namespace attributes across all LBA formats and a mechanism to report the namespace attributes of LBA formats without having to create the namespace.

4.2.2.1 Description of change

- Added CNS value 0Ah to the Identify command to return an I/O Command Set Specific Identify Namespace data structure for the Key Value Command Set associated with the specified Format Index and Command Set Identifier. The I/O Command Set Specific Identify Namespace data structure for the Key Value Command is updated to identify the fields that report valid.

4.3 Feature Enhancements

4.3.1 Enable Namespace Management command to support Key Value Command Set – TP4138 (mandatory)

Added reporting of the Format Index used to determine the format of an allocated namespace associated with the Key Value Command Set in the I/O Command Set Specific Identify Namespace data structure and ability for the host to specify the format for creating a namespace associated with the Key Value Command Set by adding a Format Index to the Host Software Specified Fields data structure.

4.3.1.1 Description of change

- Update the Namespace Management command

4.3.1.1.1 New requirement / incompatible change

  - Update the Namespace Management command to add the Format Index to be specified in the Host Software Specified Fields.
  - Update the Identify command to add the Format Index to be reported in the I/O Command Set Specific Identify Namespace data structure for the Key Value Command Set.

4.3.2 Post Sanitize Media Verification – TP4152 (optional)

4.3.2.1 Description of change

- Namespaces associated with the Key Value Command Set do not support additional media modification.

4.3.2.1.1 New requirement / incompatible change

  - During Media Verification state and Post-Verification Deallocation state, namespaces associated with the Key Value Command Set are not accessible.

4.3.3 Non-Storage namespace changes – TP4162a (optional)

4.3.3.1 Description of change

- Expanded the definition of namespace to include namespaces not used for storage (e.g., compute namespaces)

4.4 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

4.4.1 ECN110 (mandatory)

4.4.1.1 Description of change due to the ECN and any technical proposal that affected the requirement.

4.4.1.1.1 New requirement / incompatible change

  - The Logical Block Address (LBA) field in the Error Information log page is reserved.
  - The Failing LBA (FLBA) filed in the Self-test Results data structure of the Device Self-test log page is reserved.

5 Changes to the NVM Express® PCIe Transport Interface Specification

5.1 New Features

5.1.1 Power Loss Signaling – TP4029a (optional)

Provides a usage mechanism for the Power Loss Signaling function introduced by PCI-SIG for the Mini Express form factors.

5.1.1.1 Description of change

- Maps the behavior of the Power Loss Signaling variables (NVMe Base Specification) to the Power Loss Notification (PLN#) signal and the Power Loss Acknowledge (PLA#) signal (PCI Express M.2 Specification).

5.1.2 Physical Interface Receiver Eye Opening Measurement (EOM) – TP4119b (optional)

5.1.2.1 Description of change

- Added log page 19h with PCIe transport-specific content, allowing a host to pass selection parameters to start the measurement process in a transport-specific, but implementation-agnostic manner, and to return measurement results in an PCIe-specific, but implementation-agnostic manner.

6 Changes to the NVM Express® TCP Transport Specification

6.1 New Features

6.1.1 NVMe-oF Centralized Discovery Controller – TP8010a (optional)

6.1.1.1 Description of change

- Added two new NVMe/TCP PDUs: KDReq and KDResp
- Added new KDCONN bit to the ICReq NVMe/TCP PDU's FLAGS field to request a kickstart discovery NVMe/TCP connection

6.1.2 NVMe/TCP – TLS updates – TP8018 (mandatory if TLS is supported)

Description of change Redesign TLS concatenation to DH-HMAC-CHAP and TLS requirement updates for NVMe/TCP

6.1.2.1 Description of change

6.1.2.1.1 New requirement / incompatible change

- TLS concatenation redesign (required if TLS concatenation supported):
  - Replace TLS PSK concatenation mechanism that proceeded from DH-HMAC-CHAP to TLS on same TCP connection with a new mechanism that disconnects TCP connection after TLS PSK generation and uses TLS PSK on new connections. Improves security by applying TLS security to the Connect command on any connection that is used for NVMe commands and prevents common implementation flaws that could allow NVMe commands to bypass TLS on the same connection.
  - Restrict TLS PSK generation to Admin Queue connections. A generated TLS PSK may be used on I/O Queue connections and other Admin Queue connections. This simplifies implementations.
  - Use new format of TLS PSK identity that uniquely identifies TLS PSK. Old format is obsoleted.
  - Distinguish generating new TLS PSK from replacing existing TLS PSK to support TLS PSK rollover.
- TLS requirement updates for NVMe/TCP (required):
  - Forward secrecy: Remove requirement for PSK-only authentication. PSK with ECDHE remains required. Change mandatory DH group for forward secrecy to p384 curve.
  - Channel binding: Require NQNs in TLS PSK identity and Connect command to match.
  - TLS handshake protection: Enforce existing restrictions that sending any other traffic when a TLS handshake is expected or in progress is an error that closes the TCP connection and that sending unencrypted traffic after a successful TLS handshake is an error that closes the TCP connection.
  - TLS functionality limitation or discouragement:
    - Strongly discourage TLS 1.2 in favor of TLS 1.3 for improved security.
    - Prohibit Application Layer Protocol Negotiation (ALPN) as it is inapplicable to NVMe/TCP.
    - Strongly discourage TLS 1.3 session resumption.
    - State that some environments may require stronger crypto than minimum implementation requirements, e.g., environments that follow CNSA 1.0 implementation guidance.
    - Specify error if TLS PSK generation attempted with NULL DH groupReferences
- Technical Proposal(s) / ECN: TP8018
- Section(s): 1.5, 3.1, 3.6.1

6.1.3 NVMe-oF Security Configurations - TP8025 (optional)

6.1.3.1 Description of change

- Definition of a consistent way to express security protocol configurations of NVMe entities.

6.2 Feature Enhancements

6.2.1 KATO Corrections and Clarifications – TP4129 (mandatory)

6.2.1.1 Description of change

- Added clarification on handling communication loss and a reference to the NVM Express Base Specification.

6.2.2 Clarifying HPDA and CPDA usage with NVMe TCP PDUs – TP8026 (mandatory)

Resolved an NVMe/TCP interoperability problem caused by different interpretations of the HPDA and CPDA fields as defined in the NVMe/TCP Transport Specification.

6.2.2.1 Description of change

- Update definitions of HPDA, CPDA fields; specify they only apply to non-kickstart discovery NVMe/TCP connections
- Update the PDO field in the KDReq and KDResp PDUs to “Obsolete”.

6.3 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

6.3.1 ECN102 (mandatory)

6.3.1.1 Description of change

6.3.1.1.1 New requirement / incompatible change

- If the PDU Header Digest is enabled and a subsequent PDU transferred in this connection other than a H2CTermReq PDU or a C2HTermReq PDU has the HDGSTF flag, if defined, cleared to ‘0’, then the receiver shall treat that PDU as if a fatal transport error has occurred. If the PDU Header Digest is disabled and a subsequent PDU is received with the HDGSTF flag, if defined, set to ‘1’, then the receiver shall treat that PDU as if a fatal transport error has occurred.
- If the PDU Data Digest is enabled and a Command Capsule PDU containing in-capsule data, a H2CData PDU, or a C2HData PDU transferred in this connection has the DDGSTF flag, if defined, cleared to ‘0’, then the receiver shall treat that PDU as if a fatal transport error has occurred. If the PDU Data Digest is disabled and any of these PDUs are received with the DDGSTF flag, if defined, set to ‘1’, then the receiver shall treat that PDU as if a fatal transport error has occurred.

7 Changes to the NVM Express® RDMA Transport Specification

7.1 Feature Enhancements

7.1.1 KATO Corrections and Clarifications – TP4129 (mandatory)

7.1.1.1 Description of change

- Added clarification on handling communication loss

8 Changes to the NVM Express® Management Interface Specification

New features added to the other NVMe specifications that define new Admin commands, log pages, and Feature Identifiers update the NVM Express Management Interface Specification to define support requirements. Those updates to the specification are not reflected in the information in this section.

8.1 New Features

8.1.1 NVMe Specification Version Reporting – TP4135 (mandatory)

8.1.1.1 Description of change

- Aligned descriptions of version number fields with those in other NVM Express specifications.

8.1.2 NVMe-MI High Availability - TP6034a (optional)

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

8.1.2.1 Description of change

8.1.2.1.1 Incompatible change

- Status bits in the out-of-band mechanism now have an instance per Management Endpoint instead of a single instance shared among all Management Endpoints in the NVM Subsystem.

8.1.3 Out-of-Band Asynchronous Events - TP6035a (optional)

Added an out-of-band asynchronous event mechanism (similar to Asynchronous Event Request on the in-band interface).

8.1.3.1 Description of change

- Adds an out-of-band asynchronous event mechanism.

8.2 Feature Enhancements

8.2.1 Clarifications to Namespace Changes Reporting – TP4150 (optional with mandatory components)

8.2.1.1 Description of change

- Renamed Changed Namespace List log page to Changed Attached Namespace List
- Added Changed Allocated Namespace List log page to Management Endpoint – Log Page Support figure

8.2.1.1.1 New requirement / incompatible change

- Changed the Namespace Attribute Changed (NAC) bit in the Controller Health Data Structure (CHDS) to now be set when there are changes to either attached or unattached namespaces

8.2.2 Post Sanitize Media Verification – TP4152 (optional)

8.2.2.1 Description of change

- Defined new behavior of the Read command while in the Media Verification state.

8.2.2.1.1 New requirement / incompatible change:

- During Media Verification state, Read commands will complete with status codes other than Sanitize In Progress. Hosts that are not aware of Media Verification will not expect these status codes.

8.2.3 Shutdown Clarifications and Enhancements – TP4169 (optional with mandatory components)

Updated the definition of NVM Subsystem Shutdown to specify interactions between NVM Subsystem Shutdown and Controller Level Resets (CLRs) that are not caused by NVM Subsystem Reset. Allows the controller to indicate to the host how long a host should wait for NVM Subsystem Shutdown to complete.

Remove use of the word “safe” to describe ready-to-be-powered-off conditions.

Specified the situations in which the media is in the shutdown state while the CSTS.SHST field is cleared to 00b.

8.2.3.1 Description of change

- Added a case for the Ignore Shutdown (ISH) bit defined in TP6032 - processing an out-of-band NVM Admin command that requires media access and has the Ignore Shutdown (ISH) bit set is also able to cause the media to transition out of the shutdown state if CSTS.SHST is cleared to 00b.

8.2.4 Status Reporting Enhancements – TP6021 (mandatory if Telemetry log pages or NVMe-MI are supported)

A variety of enhancements and clarifications related to NVMe-MI status reporting.

8.2.4.1 Description of change

- Updates to Telemetry logs (Mandatory if Telemetry logs are supported)
  - **New Requirement:** Added a mechanism to report Telemetry Controller-Initiated Data Available via the out-of-band mechanism.
  - **New Requirement:** Added a field in each Telemetry log to indicate if it is Controller in scope or NVM Subsystem in scope.
  - **New Requirement:** Added a status bit to indicate the NVM Subsystem is in the Sanitize failure mode (Mandatory if Sanitize is supported)
- Updates to Controller Health Status Poll (Mandatory)
  - **New Requirement:** Added a requirement to return the Controller Health Status Changed Flags in the response.
  - Clarified how the Controller filtering logic works and reversed the polarity to indicate when Controllers meet the selection criteria to be included in the Controller Health Status Poll response versus when the Controllers are filtered out.
  - Clarified that it is not an error if the INCF bit is set to '1' and non-SR-IOV PCI Functions do not exist.
  - Clarified that if the INCF, INCPF, or INCVF bits are cleared to '0', then the corresponding Controller is not included in the Controller Health Status Poll response.
  - **New Requirement:** Specified that a Controller ID that is greater than the maximum Controller ID in the NVM Subsystem results in an Invalid Parameter Error Response.
  - Clarified that if the Clear Error State Flags bit in Get State is set to '1, then the bits are cleared after, not before, the current state of the bits are copied to the response and the Controller selection criteria based on Controller Health Status Changed flags have been performed.
  - **Incompatible Change:** Changed the Controller Enable Change Occurred bit in the Controller Health Data Structure from a status bit that is set when the Controller is enabled or disabled to a state bit that indicates if the Controller is currently enabled or disabled.
- Updates to the Controller Health Status Changed Flags (Mandatory)
  - **Incompatible Change:** Set the Controller Enable Change Occurred bit whenever Controller Enable Change Occurred bit in the Controller Health Data Structure changes state instead of only when it transitions from '0' to '1'.
  - **Incompatible Change:** Set the Critical Warning bit whenever any bits in the Critical Warning field change state instead of only when they transition from '0' to '1'.
  - **Incompatible Change:** Set the Controller Status Change bit whenever any bits in the Controller Status field change state instead of only when they transition from '0' to '1'.
  - **Incompatible Change:** Set the Controller Fatal Status bit whenever Controller Fatal Status bit in the Controller Health Data Structure changes state instead of only when it transitions from '0' to '1'.
  - **Incompatible Change:** Set the Ready bit whenever Ready bit in the Controller Health Data Structure changes state instead of only when it transitions from '0' to '1'.
  - **Incompatible Change:** Specified that a Controller Health Status Poll command with the Clear Changed Flags bit set to '1' shall also clear the Namespace Attribute Changed and Firmware Activated bits in Controller Health Data Structure to '0'.
- Updates to Get State (Mandatory)
  - Clarified that if the Clear Changed Flags bit is set to '1, then the bits are cleared after, not before, the current state of the bits are copied to the response.
  - **New Requirement:** Specified that NVM Subsystem Reset Occurred is not cleared by a Management Endpoint Reset.
- Updates to NVM Subsystem Health Status Poll (Mandatory)

- **New Requirement:** Added the ability report the Composite Temperature of the coldest Controller in the NVM Subsystem for under-temperature conditions.
- Clarified the condition that causes a temperature sensor failure to be reported (if the composite temperature of the NVM Subsystem cannot be guaranteed to be accurate due to the failure of one or more temperature sensors).
- Clarified that if the Clear Status bit is set to '1, then the bits are cleared after, not before, the current state of the bits are copied to the response.

- NVM Subsystem Reset Occurred updates (Mandatory)
  - **Incompatible Change:** Specified in Get State that the NVM Subsystem Reset Occurred status bits are only set if a new firmware image is not activated to synchronize to the behavior of NVM Subsystem Reset Occurred in the NVM Express Base Specification.

8.2.5 Reset Behavior Clarifications – TP6027b (mandatory)

Added implicit reset naming and removes ambiguous behavior for different form factor signals.

8.2.5.1 Description of change

- Clarified and improved the effects of resets in NVMe-MI.
- Added a reference to SNIA SFF-TA-1009 Enterprise and Datacenter Standard Form Factor Pin and Signal Specification.
- Clarified in the Operations Supported During NVM Subsystem Power States table that SMBus Reset or PCIe Reset may prevent MCTP access.
- Recommend that PCIe VDM Management Endpoints be associated with Function 0 on the furthest upstream bus on the NVMe Storage Device or NVMe Enclosure.
- Recommend that PCIe VDM Management Endpoints not be associated with SR-IOV Virtual Functions.

8.2.5.1.1 Incompatible change

- Some SMBus resets are no longer required to maintain ARP assigned address.
- Changed the type of reset required to return to normal operation when the Reset Not Required bit is cleared to '0' from Controller Level Reset to NVM Subsystem Reset.
- The reset value of the Controller Enable Change Occurred bit was changed from 0 to HwInit for the in-band tunneling mechanism.
- Status bits in the in-band tunneling mechanism now have an instance per Controller instead of a single instance shared among all Controllers in the NVM Subsystem.

8.2.5.1.2 New requirement

- SMBus resets are now required to reset SMBus/I2C MTU and frequency.
- SMBus/I2C Management Endpoints are now required to support SMBus clock-low recovery.
- Added PCIe Reset requirements.
- SMBus Reset timing is now the same as SMBus power on timing.
- SMBus Reset shall cause the SMBus/I2C port to attempt to generate a STOP condition within 5 ms from the assertion of SMBus Reset.
- Specified what is reset by an NVM Subsystem Reset.
- Specified that a Management Endpoint Reset aborts outstanding Command Messages and discards outstanding Control Primitives.

8.2.6 Out-of-Band Admin Command While Shutdown Enhancement - TP6032, TP4169 (mandatory)

Defined interactions between NVMe Admin Commands processed via the out-of-band mechanism that access media and shutdown.

8.2.6.1 Description of change

- Added reporting of Maximum Media Not Ready Response Time
  - **New requirement / incompatible change**
    - The Management Endpoint shall report the worst-case time for media used by NVMe Admin Commands to become ready using the out-of-band mechanism.
- Added a new bit to NVMe Admin Commands sent via the out-of-band mechanism
  - **New requirement / incompatible change**
    - Adds the Ignore Shutdown (ISH) bit to allow the Management Endpoint to bring media out of a shutdown state to process NVMe Admin Commands that access media via the out-of-band mechanism.
- Added new Shutdown Behavior
  - The Controller may abort out-of-band NVMe Admin Commands sent via the out-of-band mechanism that access media so the controller can handle shutdown.

8.2.7 MCTP Packet Timing – TP6033a (mandatory)

Clarify and expand packetization constraints for NVMe-MI traffic over MCTP. This should create more consistent implementations across vendors and enable simpler Management Controller error checking and recovery.

The Pause flag was simplified to one per Endpoint, the Abort primitive was cleaned up, and defined behavior for receiving new messages on a non-idle slot.

8.2.7.1 Description of change

- Comprehensive MCTP packet timing constraints
  - Management Endpoint should limit interpacket bus idle delays to less than 40 ms
  - Management Endpoint should resume/replay content in less than 40 ms
  - Response Message timeouts don't apply under certain conditions (e.g., when the Management Endpoint is paused or when Response Message is delayed due to the transmission of another MCTP message)
- Clarify MCTP messages from Management Controller not dropped due to interpacket timeout
- Explicit procedure for receiving new command on busy command slot
- Simplified Pause Flag to a single flag for Management Endpoint scope instead of one per Command Slot scope
- Cleaned up the unable to abort case with a new error message
- Clarify the conditions under which Responses Messages are discarded by a successful Abort Control Primitive
- Clarify that the request-to-response timer is restarted under certain conditions when the Pause Flag transitions from '1' to '0'

8.2.7.1.1 Required Change

  - Deprecate Timeout Waiting for a Packet (WPTT) status flag since it cannot be set with new procedure
  - An Abort Control Primitive in Idle causes the Response Message for the associated Command Slot to be discarded and no longer available to be replayed
  - Specify that upon completion of Command Message processing, if a More Processing Required Response Message for the Command Slot is pending transmission (e.g., the More Processing Required Response Message was not able to be transmitted because the Pause Flag was set to '1'), then the More Processing Required Response Message should be discarded
  - Added a requirement that if Get State is processed while transmitting a More Processing Required Response, then the Management Endpoint indicates that it is in the Process state as opposed to the Transmit state.

8.2.7.1.2 Backwards Incompatible

  - Pause Control Primitive and Abort Control Primitive now affect idle Command Slots
  - Abort and implied Abort now return error status when Abort is not possible
  - No Management Endpoint receiver timeout, and hence no pausing of the receiver timeout

8.2.8 UTF-8 Admin Labels – TP8020 (optional)

8.2.8.1 Description of change

- Define UTF-8 string format and usage.

8.2.9 Modifications for I3C – TP6037 (optional)

Updates the name of the I2C/SMBus ports and communications to match the new conventions used by SNIA SFF TA's EDSFF and PCI-SIG specifications for I3C. Adds a new section on I3C with a quick summary of how I3C works, backwards compatibility with SMBus, speed negotiation, list of required/prohibited options, over which bytes the packet CRC is calculated, hot plug notifications, and ideas to make implementations more robust. Updates enumerated lists in various places for SMBus frequency to include the new higher frequencies for I3C, update how MTU size changes are made, specify retry mechanisms on errors, and how each reset types affects staying in I3C mode vs returning to power on default of SMBus.

8.2.9.1 Description of Change

- Adds optional communications using I3C protocol over SMBus which is now going to be called the 2-Wire port/element.
- New bit in VPD's NVM Subsystem Element Descriptor to indicate if MCTP over 2-Wire port is operational during the Auxiliary Power Only State.
- Increased maximum VPD size from 4 KiB to 64 KiB.

8.2.10 Management Interface Miscellaneous Maintenance – TP6038 (mandatory)

Addresses a number of miscellaneous maintenance issues. It will address various issues that impact interoperability between a Management Controller (e.g., a BMC) and a Management Endpoint (e.g., an NVMe Storage Device).

8.2.10.1 Description of Change

- Added the following new fields to the NVM Subsystem Element Descriptor data structure:
  - Management Endpoint Ready Independent of Media Timeout
  - Management Endpoint Ready with Media Timeout
  - Incompatible Change: Maximum Unresponsiveness Time
- Added the following new fields to the NVM Subsystem Port Descriptor data structure:
  - MCTP Support
- Added a Boot Failure Code field to the VPD that indicates failures to load or initialize the NVM Subsystem firmware
- Deleted the requirement for the Unique NVM Storage Device ID field of the UDID to be defined sequentially for each NVM Subsystem if there are multiple NVM Subsystems in an SMBus ARP-capable NVMe Storage Device or NVMe Enclosure
- Clarified the operational states with respect to controller disable and Controller Level Reset

8.3 Required Changes

This section describes mandatory behavior changes required to comply with NVM Express Specification 2.1.

8.3.1 Should to Shall Conversion – TP4136 (mandatory)

Identified instances of "should" statements that were introduced into the various NVMe specifications via ECNs and changes them to "shall" statements where appropriate. These "should" statements that are changing to "shall" statements are for

behaviors that were strong recommendations during their initial incorporation and were intended to be converted into hard requirements at a later point.

8.3.1.1 Description of change

8.3.1.1.1 New requirement / incompatible change

- For the Pause Control Primitive, if the Command Slot Identifier (CSI) bit is set to '1', then the Management Endpoint shall transmit an Invalid Parameter Error Response with the Parameter Error Location (PEL) field indicating the CSI bit.

8.3.2 ECN103 (mandatory)

8.3.2.1 Description of change

8.3.2.1.1 New requirement / incompatible change

- The Critical Warning (CWARN) field in the Controller Health data structure is meant to report all of the critical warnings defined in the SMART / Health Information log page as defined by the NVM Express Base Specification. Therefore, defined the Persistent Memory Region (PMR) Error bit in the Critical Warning (CWARN) field in the Controller Health data structure to report PMR errors.
- Clarified the requirements for the value that is reported in the PCIe Max Payload Size field for ARI Devices and Non-ARI Multi-Function Devices.
- Clarified that the Unused field (i.e., bits 1:0) shall be set to 11b in the field Status Flags field of the Subsystem Management Data Structure. Previous wording was not clear if both bits are to be set to '1'.
- The processing of a prohibited NVMe Admin Command returns an Invalid Command Opcode Error Response as opposed to the Parameter Error Response.
- The value of FFh in the Number of Ports (NUMP) field in the NVM Subsystem Information data structure is not supported as a port identifier of 256 is not allowed to be reported.

8.3.3 ECN115 (mandatory)

8.3.3.1 Description of change

8.3.3.1.1 New requirement / incompatible change

- For non-reserved values except the value 01h (i.e., Port Information) and 05h (i.e., Management Endpoint Buffer Command Support List) in the Data Structure Type (DTYP) field in the Read NVMe-MI data structure, the Port Identifier (PORTID) field in that in the Read NVMe-MI data structure shall be ignored by the Management Endpoint.
- For non-reserved values except the value 03h (Controller Information) and 04h (Optionally Supported Command List) in the Data Structure Type (DTYP) field in the Read NVMe-MI data structure, the Controller Identifier (CTRLID) field in that in the Read NVMe-MI data structure shall be ignored by the Management Endpoint.
- For non-reserved values except the value 04h (Optionally Supported Command List and 05h (i.e., Management Endpoint Buffer Command Support List) in the Data Structure Type (DTYP) field in the Read NVMe-MI data structure, the I/O Command Set Identifier (IOCSI) field in that in the Read NVMe-MI data structure shall be ignored by the Management Endpoint.

8.3.4 ECN118 (mandatory)

8.3.4.1 Description of change

8.3.4.1.1 New requirement / incompatible change

- Added the explicit requirement that a Responder shall not perform any text processing that is specific to the character set or locale such as checks for byte values not used by UTF-8, Unicode normalization, etc. This requirement was previously implied.

- If a value of ‘1’ for the CIAP bit is received in a Command Message, then an Invalid Parameter Error Response with the PEL field indicating this bit shall be returned.
- If a value of ‘1’ for the MEB bit is received in a Command Message, then an Invalid Parameter Error Response with the PEL field indicating this bit shall be returned.

8.3.5 ECN119 (mandatory)

8.3.5.1 Description of change

8.3.5.1.1 New requirement / incompatible change

- The out-of-band processing of a Get Log Page command is incompatible with a BMC that does not support NVMe-MI 1.2 as that version requires that the RAE bit in the Get Log Page command be set to ‘1’. BMCs that support NVMe prior to NVMe 1.2, the RAE bit was reserved and those system are required to clear the bit to ‘0’, thus every Get Log Page command is aborted. Solution: the Out-of-band processing of a Get Log Page command ignores the RAE bit and processes the command as though the RAE bit is set to ‘1’ for log pages used with asynchronous events.

8.3.6 ECN120 (mandatory)

8.3.6.1 Description of change

8.3.6.1.1 New requirement / incompatible change

- If the DTYP field in the NVMe Management Dword 0 of a Read NVMe-MI Data Structure command is 04h or 05h, then for commands with the NMIMT field set to any non-reserved value other than 02h in the Optionally Supported Command List data structure or Management Endpoint Buffer Supported Command List data structure, then the I/O Command Set Identifier (IOCSI) field is not applicable and shall be ignored by the Management Endpoint.

8.3.7 ECN122 (mandatory)

8.3.7.1 Description of change

- **New requirement / incompatible change**
  - In the Management Interface Command Set, fixed an issue where the DOFST/DLEN checking logic specified an error if DOFST plus LEN is “greater than or equal to” the size of the Management Endpoint Buffer but should have said “greater than”.
  - In a VPD Read command, if the Data Offset (DOFST) field is greater than or equal to the maximum size of the FRU Information Device, then the Management Endpoint shall respond with an Invalid Parameter Error Response with the PEL field indicating the DOFST field.
  - In a VPD Write command, if the Data Offset (DOFST) field is greater than or equal to the maximum size of the FRU Information Device, then the Management Endpoint should not write the contents of the VPD and should respond with an Invalid Parameter Error Response with the PEL field indicating the DOFST field.
  - The Management Endpoint shall respond with an Invalid Parameter Error Response with the PEL field indicating this field if bits 1:0 of the DOFST field are not cleared to 00b.
  - The Management Endpoint shall respond with an Invalid Parameter Error Response with the PEL field indicating this field if any of the following conditions are true:
    - the value of the DLEN field is not equal to the length of the NVMe Request Data field required by the command; or
    - the DLEN field is cleared to 0h for commands that are defined to transfer NVMe Request Data or NVMe Response Data.

9 Changes to the NVM Express® Boot Specification

9.1 New Features

9.1.1 Boot Driver Handling of Unavailable SNSS Records – TP8029 (optional)

Provided better diagnostics by providing reasons for a SNSS Descriptor that has been set to an “Unavailable” state by creating a new data structure that may contain enhanced diagnostic fields (e.g., connection timeout, discovery log page traversal timeout). The SNSS may be expanded to include additional fields for communication of diagnostic data (e.g., timeout value, HFI, SNSS device number). Includes informative concepts for pre-OS and OS on how to interpret these fields and what to do with them. Includes recommendations for how to handle unavailable SNSS entries as listed in the NBFT.

9.1.1.1 Description of change

- Increments the version of the Subsystem Namespace Extended Information Descriptor and adds new fields to the Subsystem Namespace Extended Information Descriptor relative to Diagnostic data for Networking, Connections, and Timeouts, and Authentication and Security

9.2 Feature Enhancements

9.2.1 Enhanced HFI DHCP Client Options Support – TP8027 (optional)

Enables support for meta-data attributes related to DHCP for IPv4 and IPv6 handling per HFI Transport Info Descriptor for NVMe/TCP to improve IP address consistency. With the new feature, the Pre-OS environment will be able to pass information in the NBFT to the OS-environment and specify the DHCP options it provided to the DHCP Server. The OS environment will be able to use this information to provide the same DHCP options to the DHCP Server to reestablish connectivity to namespaces with the same addresses. Specified guidance for coexistence issues related to IPv4 and IPv6 automatic address assignment (e.g., DHCP, neighbor discovery, zero configuration networking), and provide recommendations for pre-OS, OS driver and application writers for common defaults.

9.2.1.1 Description of change

- Added new optional fields to the HFI Transport Info Descriptor for TCP/IP in the NBFT (e.g., DHCP 4/6 Identifiers and DHCP meta-data content per HFI descriptor, differentiation data for auto-negotiation and auto-configuration schemes used in DHCP or address assignment) including a Flag in the HFI to indicate the use of this feature, Address Assignment Scheme Identification, Client Identifier Type and Client Identifier for e.g., ‘DHCPv4’, ‘DHCPv6’, ‘IPv6-ND’.
- The Informative Annex now specifies how to use these fields, describes how to handle DHCP4, DHCP6 Client Identifier schemes, provide recommendations on usage of the new fields in the Pre-OS and OS environments; and how OS drivers shall parse upwardly compatible changes to versioned descriptors.

9.2.2 NVMe-oF Boot HostNQN and HostID – TP4126 (optional)

Provides recommendation default HostNQN and HostID values, in the absence of administratively set values, that can be used in boot scenarios by both boot drivers and OS environments; results in a consistent behavior regardless of OS, boot device, NVMe-oF transport, installation occurrence. Result is an easier system administration of NVMe-oF boot solutions.

9.2.2.1.1 Description of change

Added a description of default values for HOST NQN and HostID

10 Description for Changes Document for Multiple Specifications

A Bibliography was added to the NVM Command Set specification for documents that were provided for examples or models but are not required in NVM subsystems.

The following list of revised versions of references are all compatible changes, with all changes are in the reference section of each listed NVM Express family specification:

New Reference	Old Reference	Affected NVM Express Specifications
ISO/IEC 27040:2024	ISO/IEC 27040:2015	NVM Express Base
JEDEC JESD218B-02: Solid State Drive (SSD) Requirements and Endurance Test Method standard.	JEDEC JESD218B-01: Solid State Drive (SSD) Requirements and Endurance Test Method standard.	NVM Express Base
PCI Express® Base Specification Revision 6.2	PCI Express® Base Specification Revision 5.0 Version 1.0	NVM Express Base, NVM Express Management Interface, NVMe over PCIe Transport, NVM Express Boot
PCI Express® Card Electromechanical Specification, Revision 5.1, Version 1.0	PCI Express® Card Electromechanical Specification, Revision 4.0, Version 1.0	NVM Express Management Interface
PCI Express® M.2 Specification, Revision 5.1	PCI Express® M.2 Specification, Revision 3.0, Version 1.2	NVM Express Management Interface
PCI Express® SFF-8639 Module Specification, Revision 5.0	PCI Express® SFF-8639 Module Specification, Revision 3.0, Version 1.0	NVM Express Management Interface
ACPI Version 6.5	ACPI Version 6.4	NVM Express Base, NVMe over PCIe Transport, NVM Express Boot
UEFI Specification Version 2.10	UEFI Specification Version 2.7A (2.9 in Boot)	NVM Express Base, NVM Express Boot
INCITS 555-2020 Information Technology – SCSI Enclosure Services – 4	INCITS 518-2017 Information Technology – SCSI Enclosure Services – 3	NVM Express Management Interface
MCTP Base Specification (DSP0236), Version 1.3.3~~2~~	MCTP Base Specification (DSP0236), Version 1.3.1	NVM Express Management Interface
MCTP IDs and Codes (DSP0239), Version 1.11.0	MCTP IDs and Codes (DSP0239), Version 1.7.0	NVM Express Management Interface
MCTP PCIe VDM Transport Binding Specification (DSP0238), Version 1.2.1	MCTP PCIe VDM Transport Binding Specification (DSP0238), Version 1.2.0	NVM Express Management Interface
RFC 9562, “Universally Unique Identifiers”	RFC 4122 “A Universally Unique Identifier (UUID) URN Namespace”	NVM Base Specification, NVM Express Boot
SNIA Native NVMe-oF™ Drive Specification, Version 1.1	SNIA Native NVMe-oF™ Drive Specification, Version 1.0.1	NVM Express Management Interface

SNIA SFF-TA-1006 Enterprise and Datacenter 1U Short SSD Form Factor (E1.S) Specification, Revision 1.5	SNIA SFF-TA-1006 Enterprise and Datacenter 1U Short SSD Form Factor (E1.S) Specification, Revision 1.3a	NVM Express Management Interface
SNIA SFF-TA-1007 Enterprise and Datacenter 1U Long SSD Form Factor (E1.L) Specification, Revision 1.2	SNIA SFF-TA-1007 Enterprise and Datacenter 1U Long SSD Form Factor (E1.L) Specification, Revision 1.1	NVM Express Management Interface
SNIA SFF-TA-1008 Enterprise and Datacenter 3” SSD Form Factor Specification, Revision 2.1	SNIA SFF-TA-1008 Enterprise and Datacenter 3” SSD Form Factor Specification	NVM Express Management Interface
System Management Bus (SMBus) Specification, revision 3.2	System Management Bus (SMBus) Specification, revision 3.1	NVM Express Management Interface
IEEE 802.1q-2022: IEEE Standard for Local and Metropolitan Area Networks–Bridges and Bridged Networks	IEEE 802.1q-2018: IEEE Standard for Local and Metropolitan Area Networks–Bridges and Bridged Networks	NVM Express Boot
DMTF DSP0270, “Redfish Host Interface Specification”, Version 1.3.1	DMTF DSP0270, “Redfish Host Interface Specification”, Version 1.3	NVM Express Boot
DMTF DSP8010 “Redfish 2023.3 Schema Bundle”, Version 2021.4	DMTF DSP8010 “Redfish 2021.4 Schema Bundle”, Version 2021.4	NVM Express Boot