LEGAL NOTICE:

LEGAL DISCLAIMER:

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

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

NVM Express Workgroup
c/o Virtual, Inc.
401 Edgewater Place, Suite 600
Wakefield, MA 01880
info@nvmexpress.org

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**4003 – IO Determinism**
Change Date	**1/24/2018**
Builds on Specification	**NVM Express 1.3**

Technical Proposal Author(s)

Name	Company
Chris Petersen	Facebook
David Black	Dell EMC
Lee Prewitt	Microsoft
Monish Shah	Google
Mark Carlson, Steve Wells	Toshiba
Peter Onufryk	Microsemi
Fred Knight	NetApp
Bill Martin	Samsung
Amber Huffman, Jonathan Hughes	Intel

This technical proposal defines capabilities that enable a well behaved host to achieve deterministic read latency.

Revision History

Revision Date	Change Description
4/20/2017	Draft of Phase 2 architecture capturing Set Features and Log Page initial constructs. Theory of operation, read recovery level, Identify, and namespace management changes are yet to be captured.
4/27/2017	Added Identify NVM Set construct, modified Namespace Management, added heroic error recovery bit for NVM Read command.
5/3/2017	Fixed all of the window names, added a media writes attribute
5/10/2017	Defined the behavior for when the host is unaware of NVM Sets. Clarified various pieces of behavior. Added more theory of operation text.
5/17/2017	Modified Read Recovery Level to be reported only, and not define attributes. Changed Deterministic Mode to Predictable Latency Mode.
5/20/2017	Separated NVM Sets into its own separate concept, not reliant on Predictable Latency Mode. Modified Read Recovery Level approach. Other changes based on 5/18 meeting.
5/31/2017	Addressed comments from Fred Knight mark-up. Marked functional requirements and items to address as part of Phase 3 final. Marked opens to resolve at 6/1 meeting.
6/1/2017	Added 'Controller Exception' case for autonomously transitioning out of Det. Removed 'worst case over life of product'. Added bounds for ND Window Time Maximum. Flipped 'Allocated NVM Set Capacity' to 'Unallocated NVM Set Capacity' for consistency with other specification sections. Added that in Phase 3, need to decide whether to change 'NVM Sets' name to 'Capacity Sets' or any other name.
6/6/2017	Added independent Identify bit to detect Read Recovery Levels support. Changed Heroic Error Recovery bit to 'Supplemental Retry' and defined interaction with 'Limited Retry'. Changed 'Controller Exception' to 'Deterministic Excursion'. Aligned Endurance Estimate units to the SMART / Health Log page value. Added Non-Deterministic Window Time Maximum Lifetime value. Removed Namespaces Hidden Mode. Separated NVM Set Mode and the Predictable Latency Mode threshold setting, and for NVM Set Mode added an NVM Set Aware bit to make namespaces visible.
6/12/2017	Specified attribute for wear leveling local to an NVM Set. Moved endurance estimate to a log page so it can be dynamic.
6/29/2017	Changed NVM Sets to NVM Capacity Sets. Deprecated NVM Set Aware functionality. Made Read Recovery Levels and NVM Capacity Sets separable from Predictable Latency Mode. Added NVM Capacity Set Identifier as a Get Log Page command item regardless of log type.
6/30/2017	Removed NVM Capacity Set Mode Feature. Built in on/off of Predictable Latency Mode into the Predictable Latency Mode Config Feature. Removed Supplemental Retry in the Read command.

7/13/2017	Replaced the Wear Leveling Attribute and changed it to Wear Leveling Group identifier Improved some wording in Fig5_15TBD1 Fixed some typos
7/20/2017	Added more theory of operation for NVM Capacity Sets. Started re-structuring Predictable Latency section. Other minor updates, and added some questions for team to answer.
7/25/2017	Changed NVM Capacity Set back to NVM Set (red-lines off for this). Specified the SMART / Health Log parameters that are based on Endurance Group. Moved QoS to be in Predictable Latency section. Added DTWIN and NDWIN as acronyms for Deterministic Window and Non-Deterministic Window. Removed 'Normal Mode' and changed to simple enable/disable of Predictable Latency Mode. Modified Identify to add NVM Set Identifier to Dword 11. Moved NVM Set Identifier in Identify Namespace to a better location. Changed Wear Leveling Group to Endurance Group. Substantially simplified the warning event mechanism.
7/26/2017	Changed to NDWIN if need to do more recovery for a read. Added maximum NVM Set Identifier supported to bound some structures by software. Restructured event mechanism – bit mask aggregate to identify sets with events and then details in the Predictable Latency Mode log page per NVM Set. Added back NDWIN Time threshold as an event.
7/30/2017	Added Endurance Group log page. Modified Read Recovery Levels to have them be reported. Modified names of log pages to be more consistent with Predictable Latency Mode.
8/3/2017	Changed report of Read Recovery Levels supported to bitmask in Identify. Changed the meaning of NVM Set ID 0 to allow the controller choice in allocating a namespace from an NVM Set to be compatible with ANA. Added Data Units Read to Endurance Group log page. Changed the Event Aggregate log page to return an ordered list of NVM Set Identifiers that have events pending.
8/15/2017	Added two figures to theory of operation and more explanation. Added requirements for reliable estimates. Added section on events. Modified NVM Set Identifier requirements for Namespace Management.
8/22/2017	Moved NVM Sets, Read Recovery Level, and Endurance Groups into a separate TP with red-lines turned off. Modified timing parameters based on discussion in 4/17 meeting.
8/22/2017 rev2	Modified Set Features to make Non-Deterministic Window default. Ensured Window config Set Features fails if the Predictable Latency Mode is not enabled.
9/14/2017	Edits based on Bill Martin's mark-up Edits based on Paul Suhler's mark-up Added Mark Carlson's diagram
9/21/2017	Edits based on Mike Allison's mark-up. Noted questions/comments from Mike Allison and Edward to address. Modified Identify and Get Log Page structures based on changes in TP 4018.
9/27/2017	Modified two Features to have NVM Set Identifier in Dword 11 and setting in Dword 12 so Get Features works appropriately. Specified that scaling of the time in NDWIN is implementation dependent. Aligned window status between the log page and the Predictable Latency Mode Config feature. Other minor editorial changes.

9/29/2017	Added recommendation on commands to different NVM Sets using different queues. Added Get Features table. Modified 5.22 references to 5.21 and other minor editorial changes.
11/16/2017	Editorial changes based on feedback from Bill and Judy. (not all captured yet)
12/5/2017	Editorial changes capturing remainder of Bill and Judy's feedback.
12/7/2017	Quality of Service and NDWIN Time Minimum High edits based on team feedback.
12/14/2017	Editorial tweaks made in 12/14 meeting, added on top of 12/7 meeting.
1/24/2018	Ratified in 1/8/2018 Board meeting. No changes from 12/14/2017 version.

Description of Specification Changes

Add section 8.TBD (Predictable Latency Mode) as shown below:

8.TBD Predictable Latency Mode (Optional)

Predictable Latency Mode is used to achieve predictable latency for read and write operations. When configured to operate in this mode using the Predictable Latency Mode Config Feature, the namespaces in an NVM Set (refer to section 4.TBD) provide windows of operation for deterministic operation or non-deterministic operation.

When Predictable Latency Mode is enabled:

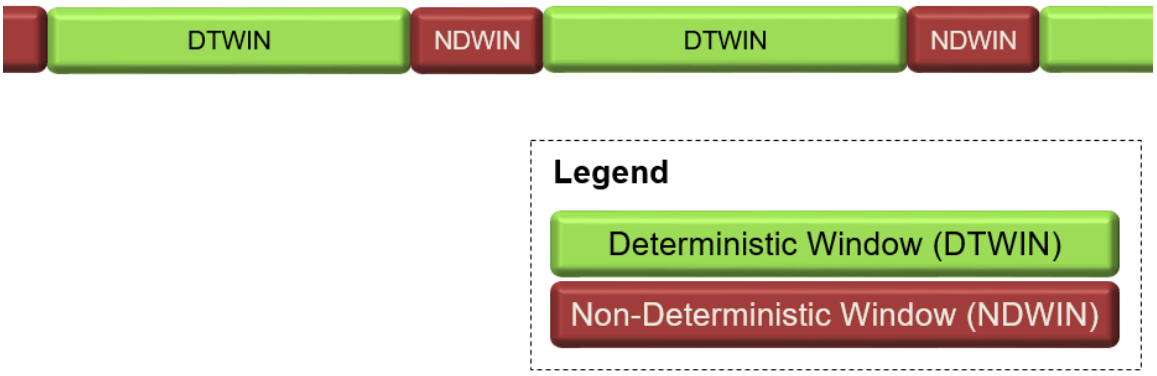
- NVM Sets and their associated namespaces have vendor specific quality of service attributes;
- IO commands that access NVM in the same NVM Set have the same quality of service attributes; and
- IO commands that access NVM in one NVM Set do not impact the quality of service of IO commands that access NVM in a different NVM Set.

The quality of service attributes apply within the NVM subsystem and do not include the PCIe or fabric connection. To enhance isolation, the host should submit I/O commands for different NVM Sets to different I/O Submission Queues.

Read Recovery Levels (refer to section 8.TBD2) shall be supported when Predictable Latency Mode is supported. The host configures the Read Recovery Level to specify the quality of service desired versus the amount of error recovery to apply for a particular NVM Set.

The Deterministic Window (DTWIN) is the window of operation during which the NVM Set is able to provide deterministic latency for read and write operations. The Non-Deterministic Window (NDWIN) is the window of operation during which the NVM Set is not able to provide deterministic latency for read and write operations as a result of preparing for a subsequent Deterministic Window. Examples of actions that may be performed in the Non-Deterministic Window include background operations on the non-volatile media. The current window that an NVM Set is operating in is configured by the host using the Predictable Latency Mode Window Feature or by the controller as a result of an autonomous action.

Figure Fig8.TBD_Fig0: Deterministic and Non-Deterministic Windows



To remain in the Deterministic Window, the host is required to follow operating rules (refer to section 8.TBD.2) ensuring that certain attributes do not exceed the typical or maximum values indicated in the Predictable Latency Per NVM Set log page. If the attributes exceed any of the typical or maximum values indicated in the Predictable Latency Per NVM Set log page or a Deterministic Excursion occurs, then the associated NVM Set may autonomously transition to the Non-Deterministic Window. A Deterministic Excursion is a rare occurrence in the NVM subsystem that requires immediate action by the controller.

The host configures Predictable Latency Events to report using the Predictable Latency Mode Config feature. The host may configure a Predictable Latency Event to be triggered when that value exceeds a specific value in order to manage window changes and avoid autonomous transitions by the controller. Refer to section 8.TBD.4.

If Predictable Latency Mode is supported, then all controllers in the NVM subsystem shall:

- Support one or more NVM Sets;
- Support Read Recovery Levels;
- Support the Predictable Latency Mode log page for each NVM Set;
- Support the Deterministic Threshold Aggregate log page;
- Support the Predictable Latency Mode Config Feature;
- Support the Predictable Latency Mode Window Feature; and
- Indicate support for Predictable Latency Mode in the Controller Attributes field in the Identify Controller data structure.

8.TBD.2 Host Operating Rules to Achieve Determinism

In order to achieve deterministic operation, the host is required to follow operating rules.

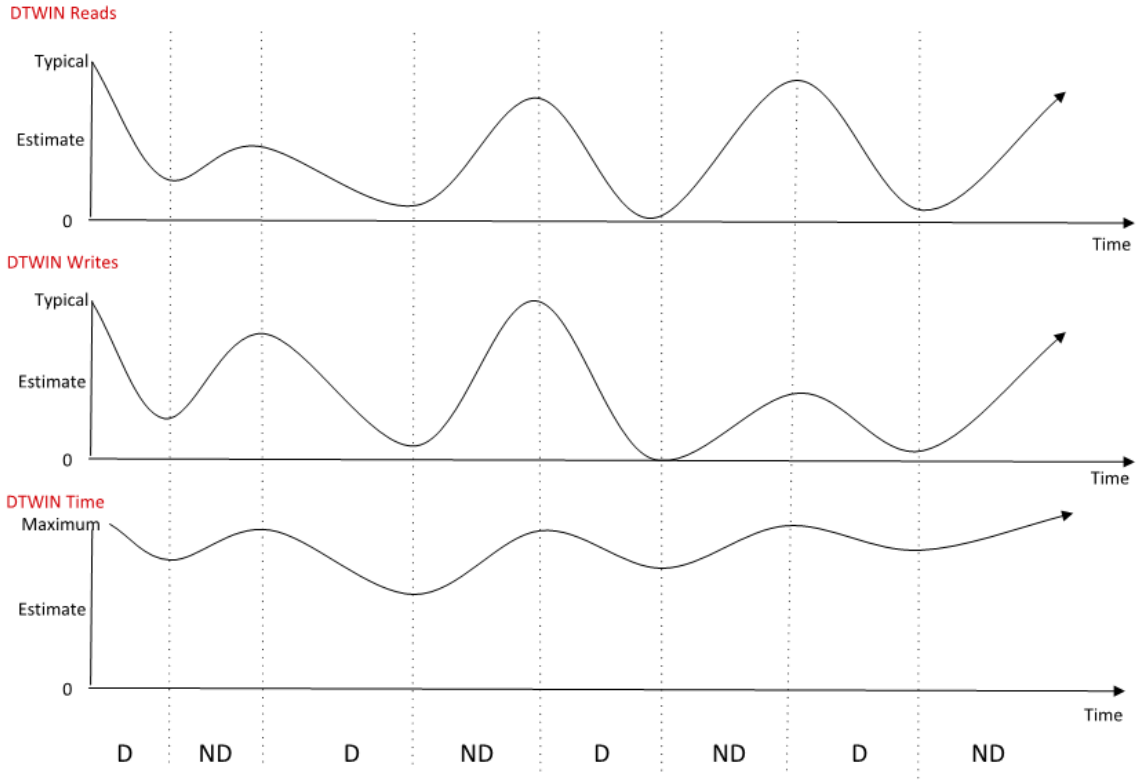
An NVM Set remains in the Deterministic Window while attributes do not exceed any of the typical or maximum values indicated in the Predictable Latency Per NVM Set log page, there is not a Deterministic Excursion, and the host does not request a transition to the Non-Deterministic Window. The attributes specified in this specification are the number of random 4KB reads, the number of writes in Optimal Write Size, and time in the Deterministic Window. Additional attributes are vendor specific.

For reads, writes, and time in the Deterministic Window, two values are provided in the Predictable Latency Per NVM Set log page (refer to section 5.14.1.10):

- A typical or maximum amount of that attribute that the host may consume during any given DTWIN.
- A reliable estimate of the amount of that attribute that remains to be consumed during the current DTWIN.

Figure Fig8.TBD.2_Fig0 shows how the Typical, Maximum, and Reliable Estimates for the DTWIN attributes increase or decrease when the associated NVM Set is in the Deterministic Window or Non-Deterministic Window.

Figure Fig8.TBD.2_Fig0: DTWIN Attributes and Estimates

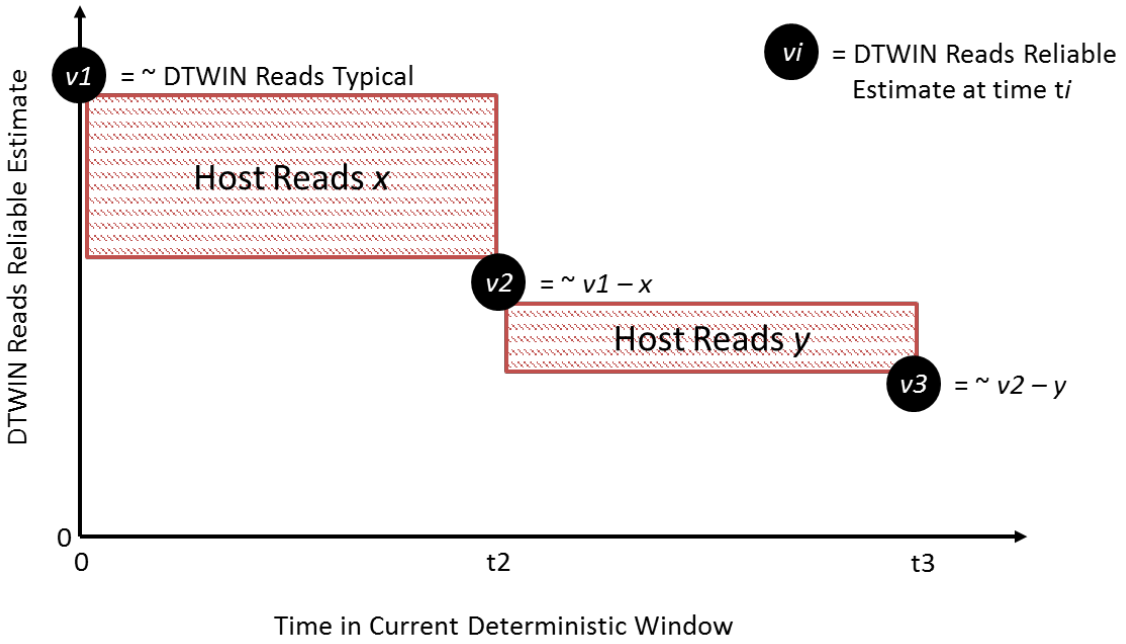


An NVM Set may transition autonomously to the NDWIN if, since entry to the current DTWIN:

a) the number of reads is greater than the value indicated in the DTWIN Reads Typical field;
b) the number of writes is greater than the value indicated in the DTWIN Writes Typical field;
c) the amount of time indicated in the DTWIN Time Maximum field has passed; or
d) a Deterministic Excursion occurs.

Figure Fig8.TBD.2_Fig0 is an example that shows the relationship between the typical and reliable estimate values for DTWIN Reads. DTWIN Reads Reliable Estimate begins near the DTWIN Reads Typical value at the start of the current DTWIN at time 0. During the first time increment, the host reads *x* units, and the value of the reliable estimate at time t2 is decremented by approximately *x*. During the second time increment, the host reads a smaller amount consisting of *y* units and thus the reliable estimate at t3 is decremented by approximately *y*.

Figure Fig8.TBD.2_Fig0: Typical and Reliable Estimate Example



The host configures the current window to be either DTWIN or NDWIN using Set Features with the Predictable Latency Mode Window Feature. The host may use the reliable estimates provided in the Predictable Latency Mode log page to ensure that the host transitions the NVM Set to the NDWIN prior to any reliable estimates exceeding one of the typical or maximum values (e.g., DTWIN Reads Estimate = 0).

The reliable estimates provided shall have the following properties when in the Deterministic Window:

- The estimates shall be monotonically decreasing towards 0h for the entirety of the DTWIN, depending on the attribute. For example, DTWIN Reads Reliable Estimate is monotonically decreasing and thus does not increase without transitioning from the DTWIN to the NDWIN.
- The estimates shall not change abruptly unless operating conditions have changed abruptly. The estimate should be based on averaging or smoothing of data collected over some period of time.

8.TBD.3 Configuring Periodic Windows

When using the NVM Set in Predictable Latency Mode, the host should transition the controller to NDWIN for periodic maintenance. The maintenance is required in order for the NVM subsystem to reliably provide the amount of time indicated for Deterministic Windows.

There are three static time based parameters reported in the Predictable Latency Per NVM Set log page (refer to section 5.14.1.10) that may be used by the host to configure periodic windows. The values provided are worst case for the life of the NVM subsystem.

- NDWIN Time Minimum Low is the minimum time that the controller remains in the Non-Deterministic Window. The controller may delay completion of a Set Features requesting a transition to the Deterministic Window until this time is completed. This time does not account for additional host activity in the Non-Deterministic Window.
- NDWIN Time Minimum High is the minimum time that the host should allow the NVM Set to remain in the Non-Deterministic Window after the NVM Set remained in the previous Deterministic Window for DTWIN Time Maximum. This time does not account for additional host activity in the Non-Deterministic Window.

- DTWIN Time Maximum is the maximum time that the NVM Set is able to stay in a Deterministic Window.

The DTWIN Time Maximum and NDWIN Time Minimum High may provide a ratio of the amount of maintenance that needs to be performed based on the time that the NVM Set remains in the DTWIN, assuming no threshold is exceeded. Any scaling of the time in the Non-Deterministic Window based on the read, write, and time behavior in the previous Deterministic Window is implementation dependent.

The DTWIN Time Estimate may be used by the host when a Deterministic Excursion has occurred. This estimate allows the host to re-synchronize an NVM Set with other NVM Sets operating in Predictable Latency Mode, if applicable.

8.TBD.4 Configuring and Managing Events

The host may configure events to be triggered when thresholds do not exceed certain levels or when autonomous transitions occur using the Predictable Latency Mode Feature. The host submits a Set Feature for the particular NVM Set and configures the specific event(s) and threshold(s) values that shall trigger an event to the host. Refer to Fig 5_21_1TBD3Fig2.

The host determines the NVM Sets that have outstanding events by reading the Predictable Latency Event Aggregate log page (refer to section 5.14.1.10). An entry is returned for each NVM Set that has an event outstanding. The host may use the NVM Set Identifier Maximum value reported in the Identify Controller data structure in order to determine the maximum size of this log page.

To determine the specific event(s) that have occurred for a reported NVM Sets, the host reads the Predictable Latency Per NVM Set log page (refer to section 5.14.1.10) for that NVM Set. The Event Type field indicates the event(s) that have occurred (e.g., an autonomous transition to the NDWIN). An event(s) for a particular NVM Set is cleared if the controller successfully processes a read for the Predictable Latency Per NVM Set log page for the affected NVM Set where the Get Log Page command has the Retain Asynchronous Event parameter cleared to ‘0’. If the Event Type field in the Predictable Latency Per NVM Set log page is cleared to 0h, then events for that particular NVM Set are not reported in the Predictable Latency Event Aggregate log page.

Modify a portion of Figure 109 (Identify – Identify Controller data structure) as shown below:

99:96	M	**Controller Attributes (CTRATT):** This field indicates attributes of the controller. Bits ~~31:5~~ 31:6 are reserved. Bit 5 (Predictable Latency Mode): If set to '1' then the controller supports Predictable Latency Mode (refer to section 8.TBD). If cleared to '0' then the controller does not support Predictable Latency Mode. Bit 4 (Endurance Groups): If set to '1' then the controller supports Endurance Groups (refer to section 8.TBD3). If cleared to '0' then the controller does not support Endurance Groups. EDITORIAL NOTE TO REMOVE: Defined in TP 4018. Bit 3 (Read Recovery Levels): If set to '1' then the controller supports Read Recovery Levels (refer to section 8.TBD2). If cleared to '0' then the controller does not support Read Recovery Levels. EDITORIAL NOTE TO REMOVE: Defined in TP 4018. Bit 2 (NVM Sets): If set to '1' then the controller supports NVM Sets (refer to section 4.TBD). If cleared to '0' then the controller does not support NVM Sets. EDITORIAL NOTE TO REMOVE: Defined in TP 4018. Bit 1 (Non-Operational Power State Permissive Mode): If set to '1' then the controller supports host control of whether the controller may temporarily exceed the power of a non-operational power state for the purpose of executing controller initiated background operations in a non-operational power state (i.e., Non-Operational Power State Permissive Mode supported). If cleared to '0' then the controller does not support host control of whether the controller may exceed the power of a non-operational state for the purpose of executing controller initiated background operations in a non-operational state (i.e., Non-Operational Power State Permissive Mode not supported). Refer to section 5.21.1.17. Bit 0 if set to '1' then the controller supports a 128-bit Host Identifier. Bit 0 if cleared to '0' then the controller does not support a 128-bit Host Identifier.

Modify Figure 90 as shown below:
EDITORIAL NOTE: These changes build on TP 4018 NVM Sets and NVMe 1.3 ECN 004.

Figure 90: Get Log Page – Log Page Identifiers

Log Identifier	O/M	Scope	Description	Reference Section
00h	Reserved			
01h	M	Controller	Error Information	5.14.1.1
02h	M	NVM subsystem [1]	SMART / Health Information	5.14.1.2
	O	Namespace [2]		
03h	M	NVM subsystem	Firmware Slot Information	5.14.1.3
04h	O	Controller	Changed Namespace List	5.14.1.4
05h	O	Controller	Commands Supported and Effects	5.14.1.5
06h	O	NVM subsystem	Device Self-test	5.14.1.6
07h	O	Controller	Telemetry Host-Initiated	5.14.1.7
08h	O	Controller	Telemetry Controller-Initiated	5.14.1.8
09h	O	NVM subsystem	Endurance Group Information EDITORIAL NOTE TO REMOVE: Defined in TP 4018.	5.14.1.9
0Ah	O	NVM subsystem	Predictable Latency Per NVM Set	5.14.1.10
0Bh	O	NVM subsystem	Predictable Latency Event Aggregate	5.14.1.11
~~0Ah – 6Fh~~ 0Ch – 6Fh	Reserved			
70h	Discovery (refer to the NVMe over Fabrics specification)			
71h – 7Fh	Reserved for NVMe over Fabrics			
80h – BFh	I/O Command Set Specific			
C0h – FFh	Vendor specific			
KEY: O = Optional, M = Mandatory Namespace = The log page contains information about a specific namespace. Controller = The log page contains information about the controller that is processing the command. NVM subsystem = The log page contains information about the NVM subsystem.				
NOTES: 1. For namespace identifiers of 0h or FFFFFFFFh 2. For namespace identifiers other than 0h or FFFFFFFFh				

Add section 5.14.1.9 adding a new log page for Predictable Latency Per NVM Set as shown below:

5.14.1.10 Predictable Latency Per NVM Set (Log Identifier 0Ah)

This log page may be used to determine the current window for the specified NVM Set when Predictable Latency Mode is enabled and any events that have occurred for the specified NVM Set. There is one log page for each NVM Set when Predictable Latency Mode is supported. Command Dword 11 (refer to Figure 87) specifies the NVM Set for which the log page is to be returned. The log page is 512 bytes in size.

The log page indicates typical values and reliable estimates for attributes associated with the Deterministic Window and the Non-Deterministic Window of the specified NVM Set. The Typical, Maximum, and Minimum values are static and worst case values over the lifetime of the NVM subsystem.

After the controller successfully completes a read of this log page with Retain Asynchronous Event cleared to '0', then reported events are cleared to '0' for the specified NVM Set and the field corresponding to the specified NVM set is cleared to '0' in the Predictable Latency Event Aggregate log page.

Figure 5_14_1_10Fig0: Get Log Page – Predictable Latency Per NVM Set Log

<table>
<tr><th>Bytes</th><th>Description</th></tr>
<tr><td>00</td><td>Status: This field indicates the status of the specified NVM Set.

Bits 7:3 are reserved.

Bits 2:0 indicate the window for the NVM Set when Predictable Latency Mode is enabled.
<table>
<tr><th>Value</th><th>Definition</th></tr>
<tr><td>000b</td><td>Not used (Predictable Latency Mode not enabled)</td></tr>
<tr><td>001b</td><td>Deterministic Window (DTWIN)</td></tr>
<tr><td>010b</td><td>Non-Deterministic Window (NDWIN)</td></tr>
<tr><td>011b – 111b</td><td>Reserved</td></tr>
</table></td></tr>
<tr><td>01</td><td>Reserved</td></tr>
<tr><td>03:02</td><td>Event Type: This field specifies the event(s) that occurred for the NVM Set indicated. Multiple bits may be set. All bits are cleared after the log page is read with Retain Asynchronous Event cleared to ‘0’.
<table>
<tr><th>Bit</th><th>Description</th></tr>
<tr><td>0</td><td>DTWIN Reads Warning</td></tr>
<tr><td>1</td><td>DTWIN Writes Warning</td></tr>
<tr><td>2</td><td>DTWIN Time Warning</td></tr>
<tr><td>3-13</td><td>Reserved</td></tr>
<tr><td>14</td><td>Autonomous transition from DTWIN to NDWIN due to typical or maximum value exceeded</td></tr>
<tr><td>15</td><td>Autonomous transition from DTWIN to NDWIN due to Deterministic Excursion</td></tr>
</table></td></tr>
<tr><td>31:04</td><td>Reserved</td></tr>
<tr><th colspan="2">Typical, Maximum, and Minimum Values</th></tr>
<tr><td>39:32</td><td>DTWIN Reads Typical: Indicates the typical number of 4KB random reads that may be performed in the Deterministic Window. Refer to section 8.TBD.</td></tr>
<tr><td>47:40</td><td>DTWIN Writes Typical: Indicates the typical number of writes in units of the Optimal Write Size that may be performed in the Deterministic Window. Refer to section 8.TBD.</td></tr>
<tr><td>55:48</td><td>DTWIN Time Maximum: Indicates the maximum time in milliseconds that the NVM Set is able to remain in a Deterministic Window before entering a Non-Deterministic Window. Refer to section 8.TBD.</td></tr>
<tr><td>63:56</td><td>NDWIN Time Minimum High: Indicates the minimum time in milliseconds that the NVM Set needs to remain in the Non-Deterministic Window before entering a Deterministic Window. This is the time necessary to prepare for remaining in the Deterministic Window for DTWIN Time Maximum. Refer to section 8.TBD.</td></tr>
<tr><td>71:64</td><td>NDWIN Time Minimum Low: Indicates the minimum time in milliseconds that the NVM Set needs to remain in the Non-Deterministic Window before entering a Deterministic Window. This is regardless of the amount of time spent in the previous Deterministic Window. Refer to section 8.TBD.</td></tr>
<tr><td>127:72</td><td>Reserved</td></tr>
<tr><th colspan="2">Reliable Estimates</th></tr>
<tr><td>135:128</td><td>DTWIN Reads Estimate: Indicates a reliable estimate of the number of 4KB random reads remaining in the current Deterministic Window, if applicable. This value decrements from DTWIN Reads Typical to zero based on host read activity and operating conditions. Refer to section 8.TBD.2.</td></tr>
<tr><td>143:136</td><td>DTWIN Writes Estimate: Indicates a reliable estimate of the number of writes in units of the Optimal Write Size remaining in the current Deterministic Window, if applicable. This value decrements from DTWIN Writes Typical to zero based on host write activity and operating conditions. Refer to section 8.TBD.2.</td></tr>
<tr><td>151:144</td><td>DTWIN Time Estimate: Indicates a reliable estimate of the time in milliseconds remaining in the current Deterministic Window, if applicable. Refer to section 8.TBD.</td></tr>
<tr><td>511:152</td><td>Reserved</td></tr>
</table>

Add section 5.14.1.11 adding a log page for Predictable Latency Event Aggregate as shown below:

5.14.1.11 Predictable Latency Event Aggregate Log Page (Log Identifier 0Bh)

This log page indicates if a Predictable Latency Event (refer to section 8.TBD) has occurred for a particular NVM Set. If a Predictable Latency Event has occurred, the details of the particular event are included in the Predictable Latency Per NVM Set log page for that NVM Set. An asynchronous event is generated when an entry for an NVM Set is newly added to this log page.

If there is an enabled Predictable Latency Event pending for an NVM Set, then the Predictable Latency Event Aggregate log page includes an entry for that NVM Set. The log page is an ordered list by NVM Set Identifier. For example, if Predictable Latency Events are pending for NVM Set 27, 13, and 17, then the log page shall have entries in numerical order of 13, 17, and 27. A particular NVM Set is removed from this log page after the Get Log Page is completed successfully with the Retain Asynchronous Event bit cleared to '0' for the Predictable Latency Per NVM Set log page for that NVM Set.

The log page size is based on the NVM Set Identifier Maximum value reported in the Identify Controller data structure by the controller as reported in the Identify Controller data structure. If the host reads beyond the end of the log page, zeros are returned. The log page is defined in Figure 5_14_1_11Fig0.

Figure 5_14_1_11Fig0: Get Log Page – Predictable Latency Event Aggregate Log Page

Bytes	Description
07:00	**Number of Entries:** This field indicates the number of entries in the list. The maximum number of entries in the list corresponds to the (NVM Set Identifier Maximum – 1) reported in the Identify Controller data structure. A value of 0 indicates there are no entries in the list.
09:08	**Entry 1:** Indicates the NVM Set that has a Predictable Latency Event pending that has the numerically smallest NVM Set Identifier.
11:10	**Entry 2:** Indicates the NVM Set that has a Predictable Latency Event pending that has the second numerically smallest NVM Set Identifier, if any.
13:12	**Entry 3:** Indicates the NVM Set that has a Predictable Latency Event pending that has the third numerically smallest NVM Set Identifier, if any.
15:14	**Entry 4:** Indicates the NVM Set that has a Predictable Latency Event pending that has the fourth numerically smallest NVM Set Identifier, if any.
…	…

Modify Figure 84 (Get Features – Feature Identifiers) as shown below:
EDITORIAL NOTE: Figure 84 is taken from TP 4018 on NVM Sets.

Figure 84: Get Features – Feature Identifiers

Description	Section Defining Format of Attributes Returned
Arbitration	Section 5.21.1.1
Power Management	Section 5.21.1.2
LBA Range Type	Section 5.21.1.3
Temperature Threshold	Section 5.21.1.4
Error Recovery	Section 5.21.1.5
Volatile Write Cache	Section 5.21.1.6
Number of Queues	Section 5.21.1.7
Interrupt Coalescing	Section 5.21.1.8
Interrupt Vector Configuration	Section 5.21.1.9
Write Atomicity	Section 5.21.1.10
Asynchronous Event Configuration	Section 5.21.1.11
Autonomous Power State Transition	Section 5.21.1.12
Host Memory Buffer	Section 5.21.1.13
Timestamp	Section 5.21.1.14
Keep Alive Timer	Section 5.21.1.15
Host Controlled Thermal Management	Section 5.21.1.16
Non-Operational Power State Config	Section 5.21.1.17
Read Recovery Level Config EDITORIAL NOTE TO REMOVE: Defined in TP 4018.	Section 5.21.1.18
Predictable Latency Mode Config	Section 5.21.1.19
Predictable Latency Mode Window	Section 5.21.1.20
NVM Command Set Specific	
Software Progress Marker	Section ~~5.21.1.19~~ 5.21.1.21
Host Identifier	Section ~~5.21.1.20~~ 5.21.1.22
Reservation Notification Mask	Section ~~5.21.1.21~~ 5.21.1.23
Reservation Persistence	Section ~~5.21.1.22~~ 5.21.1.24

Modify Figure 134 (Set Features – Feature Identifiers) as shown below:

Figure 134: Set Features – Feature Identifiers

Feature Identifier	O/M[6]	Persistent Across Power Cycle and Reset[2]	Uses Memory Buffer for Attributes	Description
00h				Reserved
01h	M	No	No	Arbitration
02h	M	No	No	Power Management
03h	O	Yes	Yes	LBA Range Type
04h	M	No	No	Temperature Threshold
05h	M	No	No	Error Recovery
06h	O	No	No	Volatile Write Cache
07h	M	No	No	Number of Queues
08h	NOTE 5	No	No	Interrupt Coalescing
09h	NOTE 5	No	No	Interrupt Vector Configuration
0Ah	M	No	No	Write Atomicity Normal
0Bh	M	No	No	Asynchronous Event Configuration
0Ch	O	No	Yes	Autonomous Power State Transition
0Dh	O	No[3]	No[4]	Host Memory Buffer
0Eh	O	No	Yes	Timestamp
0Fh	O	No	No	Keep Alive Timer
10h	O	Yes	No	Host Controlled Thermal Management
11h	O	No	No	Non-Operational Power State Config
12h	O	Yes	No	Read Recovery Level Config EDITORIAL NOTE TO REMOVE: Defined in TP 4018.
13h	O	Yes	Yes	Predictable Latency Mode Config
14h	O	Yes	No	Predictable Latency Mode Window
~~13h– 77h~~ 15h– 77h				Reserved
78h – 7Fh		Refer to the NVMe Management Interface Specification for definition.		
80h – BFh				Command Set Specific (Reserved)
C0h – FFh				Vendor Specific[1]

NOTES:
1. The behavior of a controller in response to an inactive namespace ID to a vendor specific Feature Identifier is vendor specific.
2. This column is only valid if the feature is not saveable (refer to section 7.8). If the feature is saveable, then this column is not used and any feature may be configured to be saved across power cycles and reset.
3. The controller does not save settings for the Host Memory Buffer feature across power states and reset events, however, host software may restore the previous values. Refer to section 8.9.
4. The feature does not use a memory buffer for Set Features, but it does use a memory buffer for Get Features. Refer to section 8.9.
5. The feature is mandatory for NVMe over PCIe. This feature is not supported for NVMe over Fabrics.
6. O/M: O = Optional, M = Mandatory.

5.21.1.19 Predictable Latency Mode Config (Feature Identifier 13h)

This Feature configures an NVM Set to use Predictable Latency Mode, including warning event thresholds. Predictable Latency Mode and events are disabled by default. The attributes are indicated in Command Dword 11, Command Dword 12, and the Deterministic Threshold Configuration data structure.

The NVM Set has transitioned to Predictable Latency Mode when the controller completes a Set Features command successfully with the Predictable Latency Enable bit in Command Dword 12 set to '1'. A transition to the Predictable Latency Mode may be delayed (i.e., the Set Features completion is delayed) if the NVM subsystem needs to perform background operations on the NVM in order to operate in Predictable Latency Mode. Upon successful completion of this command, the controller shall be in the Non-Deterministic Window.

If a Get Features command is submitted for this Feature, the attributes specified in Figure 5_21_1_19Fig1 are returned in Dword 0 of the completion queue entry for that command and the Deterministic Threshold Configuration data structure is returned.

Figure 5_21_1_19Fig0: Predictable Latency Mode Config – Command Dword 11

Bit	Description
31:16	Reserved
15:00	**NVM Set Identifier:** This field specifies the NVM Set to be modified.

Figure 5_21_1_19Fig1: Predictable Latency Mode Config – Command Dword 12

Bit	Description
31:01	Reserved
00	**Predictable Latency Enable:** If this field is set to '1', then Predictable Latency Mode (refer to section 8.TBD) is enabled for the NVM Set specified. If this field is cleared to '0', then Predictable Latency Mode is disabled for the NVM Set specified.

Predictable Latency Events (refer to section 5.14.1.11) are configured as described in Figure 5_21_1_19Fig1.

Figure 5_21_1_19Fig1: Predictable Latency Mode – Deterministic Threshold Configuration Data Structure

<table>
<tr><th>Bytes</th><th>Description</th></tr>
<tr><td>01:00</td><td>Enable Event: This field specifies whether an entry shall be added to the Predictable Latency Event Aggregate Log Page for the associated event. If a bit is set to ‘1’, then an entry shall be added if the specified event occurs. If a bit is cleared to ‘0’, then an entry shall not be added if the specified event occurs.
<table>
<tr><th>Bit</th><th>Description</th></tr>
<tr><td>0</td><td>DTWIN Reads Warning</td></tr>
<tr><td>1</td><td>DTWIN Writes Warning</td></tr>
<tr><td>2</td><td>DTWIN Time Warning</td></tr>
<tr><td>3-13</td><td>Reserved</td></tr>
<tr><td>14</td><td>Autonomous transition from DTWIN to NDWIN due to typical or maximum value exceeded</td></tr>
<tr><td>15</td><td>Autonomous transition from DTWIN to NDWIN due to Deterministic Excursion</td></tr>
</table></td></tr>
<tr><td>31:02</td><td>Reserved</td></tr>
<tr><td>39:32</td><td>DTWIN Reads Threshold: If the value of DTWIN Reads Estimate falls below this value and the DTWIN Reads Warning is enabled, then the ‘DTWIN Reads Warning’ event is set in the Predictable Latency Per NVM Set Log Page for the affected NVM Set.</td></tr>
<tr><td>47:40</td><td>DTWIN Writes Threshold: If the value of DTWIN Writes Estimate falls below this value and the DTWIN Writes Warning is enabled, then the ‘DTWIN Writes Warning’ event is set in the Predictable Latency Per NVM Set Log Page for the affected NVM Set.</td></tr>
<tr><td>55:48</td><td>DTWIN Time Threshold: If the value of DTWIN Time Estimate falls below this value and the DTWIN Time Warning is enabled, then the ‘DTWIN Time Warning’ event is set in the Predictable Latency Per NVM Set Log Page for the affected NVM Set.</td></tr>
<tr><td>511:56</td><td>Reserved</td></tr>
</table>

5.21.1.20 Predictable Latency Mode Window (Feature Identifier 14h)

This Feature is used to set the window for the specified NVM Set and its associated namespaces if the NVM Set is configured in Predictable Latency Mode (refer to section 8.TBD). The attributes are specified in Command Dword 11 and Command Dword 12. If Predictable Latency Mode is not enabled, then the controller shall return an error of Invalid Field in Command.

The transition to the window selected is complete when the Set Features command completes successfully. A transition to the Deterministic Window may be delayed (i.e., the Set Features completion is delayed) if the minimum time has not been spent in the Non-Deterministic Window.

If a Get Features command is submitted for this Feature, the attributes specified in Figure 5_21_1_20Fig1 are returned in Dword 0 of the completion queue entry for that command. If Predictable Latency Mode is not enabled, then the controller shall return an error of Invalid Field in Command.

Figure 5_21_1_20Fig0: Predictable Latency Mode Window – Command Dword 11

Bit	Description
31:16	Reserved
15:00	**NVM Set Identifier:** This field specifies the NVM Set to be modified.

Figure 5_21_1_20Fig1: Predictable Latency Mode Window – Command Dword 12

<table>
<tr><th>Bit</th><th>Description</th></tr>
<tr><td>31:03</td><td>Reserved</td></tr>
<tr><td>02:00</td><td>Window Select: This field selects or indicates the window used by all namespaces in the NVM Set.

<table>
<tr><th>Value</th><th>Definition</th></tr>
<tr><td>000b</td><td>Reserved</td></tr>
<tr><td>001b</td><td>Deterministic Window (DTWIN)</td></tr>
<tr><td>010b</td><td>Non-Deterministic Window (NDWIN)</td></tr>
<tr><td>011b – 111b</td><td>Reserved</td></tr>
</table>
</td></tr>
</table>