



LEGAL NOTICE:

© **Copyright 2008 to 2023 NVM Express, Inc. ALL RIGHTS RESERVED.**

This erratum is proprietary to the NVM Express, Inc. (also referred to as "Company") and/or its successors and assigns.

NOTICE TO USERS WHO ARE NVM EXPRESS, INC. MEMBERS: Members of NVM Express, Inc. have the right to use and implement this erratum subject, however, to the Member's continued compliance with the Company's Intellectual Property Policy and Bylaws and the Member's Participation Agreement.

NOTICE TO NON-MEMBERS OF NVM EXPRESS, INC.: If you are not a Member of NVM Express, Inc. and you have obtained a copy of this document, you only have a right to review this document or make reference to or cite this document. Any such references or citations to this document must acknowledge NVM Express, Inc. copyright ownership of this document. The proper copyright citation or reference is as follows: "© 2008 to 2023 NVM Express, Inc. ALL RIGHTS RESERVED." When making any such citations or references to this document you are not permitted to revise, alter, modify, make any derivatives of, or otherwise amend the referenced portion of this document in any way without the prior express written permission of NVM Express, Inc. Nothing contained in this document shall be deemed as granting you any kind of license to implement or use this document or the specification described therein, or any of its contents, either expressly or impliedly, or to any intellectual property owned or controlled by NVM Express, Inc., including, without limitation, any trademarks of NVM Express, Inc.

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.

The NVM Express® design mark is a registered trademark of NVM Express, Inc.
PCI-SIG®, PCI Express®, and PCIe® are registered trademarks of PCI-SIG.
InfiniBand™ is a trademark and servicemark of the InfiniBand Trade Association.

NVM Express Workgroup
c/o VTM Group
3855 SW 153rd Drive
Beaverton, OR 97003 USA
info@nvmexpress.org

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

NVM Express® Technical Errata

Errata ID	112
Revision Date	2023-06-27
Affected Spec Ver.	NVM Express® Base Specification Revision 2.0b NVM Express® PCIe Transport Specification 1.0b
Corrected Spec Ver.	

Errata Author(s)

Name	Company
Jim Hatfield, Gerry Houlder	Seagate

Errata Overview

This ECN updates and clarifies various text within the NVM Express Base Specification Revision 2.0b, and NVM Express PCIe Transport Specification 1.0b.

Some PCIe-specific material is to be moved from the base spec to the transport spec.

Revision History

Revision Date	Change Description
2022-03-25	Initial revision
2022-06-03	Removed things that the WG deemed out of scope: changes to RTD3, D0/D3
2022-07-05	Updated ACPI reference in the base spec from 6.2 to 6.4
2023-03-17	Incorporated review comments from Samsung
2023-03-23	Incorporated changes requested at 3/23/2023 NVM technical meeting
2023-06-27	Integrated

Description of Changes

NVM Express Base Specification 2.0b:

Backward Incompatible Changes:

- None.

Editorial Changes:

- Updated ACPI version referenced.
- Corrected a feature name in Figure 25 and Figure 34
- Corrected a log page name in Figures 29, 30, and 33

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

- Added cross references throughout
- Clarified language in section 8.15 that the scope of Power Management is the controller
- Cleaned up grammar, punctuation, paragraph separations
- Changed 'NVMe' to 'NVM Express' in several places
- Cleaned up references to Features (not commands or subcommands)
- Moved some text around for better clarity
- Added a new informative annex C: Power Management and Consumption

NVM Express PCIe Transport Specification 1.0b:

Backward Incompatible Changes:

- None.

Editorial Changes:

- added references to PCIe and ACPI specifications
- copied some text about RTD3 from the base spec to here

Editor's Note:

BLACK text indicates unchanged text.

BLUE text indicates newly inserted text.

RED~~stricken~~ text indicates deleted text;

ORANGE text indicates changes from another ECN or TP.

PURPLE text indicates moved text without changes;

GREEN text indicates editor notes.

Description of Specification Changes for NVM Express® Base Specification 2.0b

1.8 References

...

Advanced Configuration and Power Interface (ACPI) Specification, ~~Version 6.2 Errata A, September 2017~~
[Version 6.4, January 2021](https://www.uefi.org). Available from <https://www.uefi.org>.

...

3.1.2.2 Administrative Controller

3.1.2.2.2 Log Page Support

Figure 1: Administrative Controller – Log Page Support

Log Page Name	Command Support Requirements ¹
...	
Rotational Media Information	P
...	

3.1.2.3 Discovery Controller

3.1.2.3.3 Log Page Support

Figure 2: Discovery Controller – Log Page Support

Log Page Name	Command Support Requirements ¹
...	
Rotational Media Information	P
...	

...

3.5.1 Memory-based Transport Controller Initialization

...

After performing these steps, the controller shall be ready to process Admin or I/O commands issued by the host.

For exit of the D3 power state ([refer to the PCI Express Base Specification](#)), the initialization steps outlined should be followed.

...

3.6.1 Memory-based Transport Controller Shutdown

It is recommended that the host perform an orderly shutdown of the controller by following the procedure in this section when a power-off or shutdown condition is imminent.

The host should perform the following actions in sequence for a normal controller shutdown:

1. If the controller is enabled (i.e., CC.EN ([refer to Figure 46](#)) is set to '1');

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

- a. Stop submitting any new I/O commands to the controller and allow any outstanding commands to complete;
 - b. If the controller implements I/O queues, then the host should delete all I/O Submission Queues, using the Delete I/O Submission Queue command (refer to [section 5.7](#)). A result of the successful completion of the Delete I/O Submission Queue command is that any remaining commands outstanding are aborted;
 - c. If the controller implements I/O queues, then the host should delete all I/O Completion Queues, using the Delete I/O Completion Queue command (refer to [section 5.6](#));
- and
2. The host should set the Shutdown Notification (CC.SHN) field (refer to [Figure 46](#)) to 01b to indicate a normal controller shutdown operation. The controller indicates when shutdown processing is completed by updating the Shutdown Status (CSTS.SHST) field (refer to [Figure 47](#)) to 10b and the Shutdown Type (CSTS.ST) field (refer to [Figure 47](#)) is cleared to '0'.

The host should perform the following actions in sequence for an abrupt shutdown:

1. If the controller is enabled (i.e., CC.EN is set to '1'), then stop submitting any new I/O commands to the controller; and
2. The host should set the Shutdown Notification (CC.SHN) field to 10b to indicate an abrupt shutdown operation. The controller indicates when shutdown processing is completed by updating the Shutdown Status (CSTS.SHST) field to 10b and CSTS.ST is cleared to '0'.

For entry to the D3 power state ([refer to the PCI Express Base Specification](#)), the shutdown steps outlined for a normal controller shutdown should be followed.

It is recommended that the host wait a minimum of the RTD3 Entry Latency reported in the Identify Controller data structure (refer to [Figure 275](#)) for the shutdown operations to complete; if the value reported in RTD3 Entry Latency is 0h, then the host should wait for a minimum of one second. It is not recommended to disable the controller via the CC.EN field. This causes a Controller Reset which may impact the time required to complete shutdown processing. While shutdown processing is in progress, the controller may abort any command with a status code of Commands Aborted due to Power Loss Notification.

...

8.9 Host Memory Buffer

...

The host memory resources are not persistent in the controller across a reset event. Host software should provide the previously allocated host memory resources to the controller after the reset completes. If host software is providing previously allocated host memory resources (with the same contents) to the controller, the Memory Return bit (refer to [Figure 330](#)) is set to '1' in the Set Features command.

The controller shall ensure that there is no data loss or data corruption in the event of a surprise removal while the Host Memory Buffer feature is being utilized.

...

8.15 Power Management

...

The default NVM Express power state is implementation specific and shall correspond to a state that does not consume more power than the lowest value specified in the applicable form factor specification, if any. Refer to the Power Management section in the applicable [NVM Express NVMe transport Transport binding](#) specification for transport specific power requirements impacting NVMe power states, if any.

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

8.15.1 Non-Operational Power States

A power state may be a non-operational power state, as indicated by Non-Operational State (NOPS) field in **Error! Reference source not found.**. Non-operational power states allow the following operations:

...

Execution of controller initiated background operations may exceed the power advertised by the non-operational power state, if [the](#) Non-Operational Power State Permissive Mode is supported and enabled (refer to section **Error! Reference source not found.**).

No I/O commands are processed by the controller while in a non-operational power state. The host should wait until there are no pending I/O commands prior to issuing a Set Features command to change the current power state of the device to a non-operational power state and not submit new I/O commands until the Set Features command completes. Issuing an I/O command in parallel may result in the controller being in an unexpected power state.

When in a non-operational power state, regardless of whether autonomous power state transitions are enabled, the controller shall autonomously transition back to the most recent operational power state to process an I/O command.

8.15.2 Autonomous Power State Transitions

The controller may support autonomous power state transitions, as indicated in the Identify Controller data structure in **Error! Reference source not found.**. Autonomous power state transitions provide a mechanism for the host to configure the controller to automatically transition between power states on certain conditions without software intervention.

The entry condition to transition to the Idle Transition Power State is that the controller has been in idle for a continuous period of time exceeding the Idle Time Prior to Transition time specified. The controller is idle when there are no commands outstanding to any I/O Submission Queue. If a controller has an operation in process (e.g., device self-test operation) that would cause controller power to exceed that advertised for the proposed non-operational power state, then the controller should not autonomously transition to that state.

The power state to transition to shall be a non-operational power state (a non-operational power state may autonomously transition to another non-operational power state). If an operational power state is specified by a Set Features command specifying the Autonomous Power State Transitions feature (i.e., the Feature Identifier field set to 0Ch (refer to [section 5.27.1.9](#)), then the controller should abort the command with a status code of Invalid Field in Command. Refer to section 8.15.1 for more details.

8.15.3 NVM Subsystem Workloads

...

8.15.4 Runtime D3 (RTD3) Transitions

In Runtime D3, ~~(RTD3)~~ [main power is removed from the controller](#). Auxiliary power may or may not be provided. RTD3 is used for additional power savings when the controller is expected to be idle for a period of time.

In this specification, RTD3 refers to the D3_{cold} power state described in the PCI Express Base Specification. RTD3 does not include the PCI Express D3_{hot} power state because main power is not removed from the controller in the D3_{hot} power state. Refer to the PCI Express Base Specification for details on the D3_{hot} power state and the D3_{cold} power state.

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

To enable host software to determine when to use RTD3, the controller reports the ~~latency to enter~~ RTD3 Entry Latency (RTD3E) field and the ~~latency to resume from~~ RTD3 Resume Latency (RTD3R) field in the Identify Controller data structure in **Error! Reference source not found.**. The host may use the sum of these two values to evaluate whether the expected idle period is long enough to benefit from a transition to RTD3.

The RTD3 Resume Latency is the expected elapsed time from the time power is applied until the controller is able to:

- a) process and complete I/O commands; and
- b) access the NVM associated with attached namespace(s), if any, as part of I/O command processing.

The latency reported is based on a normal shutdown with optimal controller settings preceding the RTD3 resume. The latency reported assumes that host software enables and initializes the controller and sends a 4 KiB read operation.

If CSTS.ST is cleared to '0', then the RTD3 Entry Latency is the expected elapsed time from the time CC.SHN is set to 01b by host software until CSTS.SHST is set to 10b by the controller. When CSTS.SHST is set to 10b, it is safe for host software to remove power from the controller. ~~In this specification, RTD3 refers to the D3_{cold} power state described in the PCI Express Specification. RTD3 does not include the PCI Express D3_{hot} power state because main power is not removed from the controller in the D3_{hot} power state. Refer to the PCI Express Base Specification for details on the D3_{hot} power state and the D3_{cold} power state.~~

8.15.5 Host Controlled Thermal Management

...

Note: Since the host controlled thermal management (HCTM) feature uses the Composite Temperature, the actual interactions between a platform (e.g., tablet, or laptop) and two different device implementations may vary even with the same Thermal Management Temperature 1 and Thermal Management Temperature 2 temperature settings. The use of this feature requires validation between those devices' implementations and the platform in order to be used effectively.

The SMART / Health Information log page (refer to [section 5.16.1.3](#)) contains statistics related to Host Controller Thermal Management.

...

8.20 Rotational Media

Rotational media has different operational, endurance and performance characteristics than non-rotational media (e.g., NAND). Rotational media utilizes electromechanical methods for accessing data.

...

If:

- a) a domain contains an Endurance Group that stores data on rotational media;
- b) that domain processes an NVM Subsystem Reset; and
- c) the Spinup Control feature (refer to section **Error! Reference source not found.**) is:
 - a. disabled, then initial spinup for all such ~~Endurance~~ Endurance Groups in that domain shall be initiated; and
 - b. enabled, then initial spinup for all such Endurance Groups in that domain shall be inhibited during processing of the NVM Subsystem Reset until ~~any the~~ controller ~~within that domain~~ processes a Set Features (Power Management) command that specifies an operational power state.

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

If the PCIe transport is used for a controller, then the PCIe Slot Power Control feature may affect the power states supported ([refer to the PCI Express Base Specification](#)).

Annex TBD. Power Management and Consumption (Informative)

NVM Express power management capabilities allow the host to manage power for a controller. Power management includes both control and reporting mechanisms.

For information on transport power management (e.g., PCIe, RDMA), refer to the applicable NVM Express transport specification.

The scope of NVM Express power management is the controller (refer to [section 5.27.1.2](#)).

NVM Express power management uses the following functionality:

- a) Features :
 - Power Management (refer to [section 5.27.1.2](#) and [section 8.15](#));
 - Autonomous Power State Transition (refer to [section 5.27.1.9](#) and [section 8.15.2](#));
 - Non-Operational Power State Configuration (refer to [section 5.27.1.14](#) and [section 8.15.1](#)); and
 - Spinup Control (refer to [section 5.27.1.22](#));
- b) NVM subsystem workloads (refer to [section 8.15.3](#)); and
- c) Runtime D3 transitions (refer to [section 8.15.4](#)).

Controller thermal management may cause a transition to a lower power state, interacting with these Features:

- a) Temperature Threshold (refer to [section 5.27.1.3](#)); and
- b) Host Controlled Thermal Management (refer to [section 5.27.1.13](#) and [section 8.15.5](#)).

NVM Express power management uses these reporting mechanisms:

- a) properties:
 - Controller Power Scope (CAP.CPS) (refer to [Figure 36](#));
- b) fields in the Identify Controller data structure (refer to [Figure 275](#)):
 - RTD3 Resume Latency (RTD3R);
 - RTD3 Entry Latency (RTD3E);
 - Non-Operational Power State Permissive Mode;
 - Number of Power States Support (NPSS);
 - Autonomous Power State Transition Attributes (APSTA); and
 - Power State 0 Descriptor (PSD0) through Power State 31 Descriptor (PSD31) (refer to [Figure 276](#));
- c) Features:
 - Power Management (refer to [section 5.27.1.2](#));
 - Temperature Threshold (refer to [section 5.27.1.3](#));
 - Autonomous Power State Transition (refer to [section 5.27.1.9](#) and [section 8.15.2](#));
 - Non-Operational Power State Configuration (refer to [section 5.27.1.14](#) and [section 8.15.1](#));
 - Host Controlled Thermal Management (refer to [section 5.27.1.13](#) and [section 8.15.5](#)); and
 - Spinup Control (refer to [section 5.27.1.22](#));and
- d) log pages:
 - SMART / Health Information log page fields (refer to [section 5.16.1.3](#)):
 - Thermal Management Temperature [1-2] Transition Count; and
 - Total Time For Thermal Management Temperature [1-2];

Technical input submitted to the NVM Express® Workgroup is subject to the terms of the NVM Express® Participant's agreement. Copyright © 2008 to 2023 NVM Express, Inc.

and

- Persistent Event Log fields (refer to section 5.16.1.14):
 - Power On Hours (POH) (refer to Figure 224);
 - Power Cycle Count (refer to Figure 224);
 - Controller Power Cycle (refer to Figure 231); and
 - Power on milliseconds (refer to Figure 231).

Description of Specification Changes for NVM Express® PCIe Transport Specification 1.0b

1.5 References

...

PCI Bus Power Management Interface Specification Revision 1.2. Available from <https://www.pcisig.com>.

...

Advanced Configuration and Power Interface (ACPI) Specification, Version 6.4, January 2021. Available from <https://www.uefi.org>.

3.6 Power Management

Power Management operates as defined in the NVMe Base Specification with the following specifics for the PCIe transport (refer to the [PCI Express Base Specification](#) and the [PCI Bus Power Management Interface Specification](#)).

In this specification, RTD3 refers to the D3_{cold} power state described in the [PCI Express Base Specification](#). RTD3 does not include the PCI Express D3_{hot} power state because main power is not removed from the controller in the D3_{hot} power state. Refer to the [PCI Express Base Specification](#) for details on the D3_{hot} power state and the D3_{cold} power state.

The host shall never select a power state (refer to the [Power State Descriptors in the Identify Controller data in the NVMe Express Base Specification](#)) that consumes more power than the PCI Express slot power limit control value expressed by the Captured Slot Power Limit Value (CSPLV) and Captured Slot Power Limit Scale (CSPLS) fields of the PCI Express Device Capabilities (PXDCAP) register. Hosts that do not dynamically manage power should set the power state to the lowest numbered state that satisfies the PCI Express slot power limit control value.

If a controller implements the PCI Express Dynamic Power Allocation (DPA) capability and that capability is enabled (i.e., the Substate Control Enable bit is set to '1'), then the maximum power that may be consumed by the NVM subsystem is equal to the minimum value specified by the DPA substate or the NVM Express power state, whichever is lower.