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 VTM Group
3855 SW 153rd Drive
Beaverton, OR 97003 USA
info@nvmexpress.org

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**4033 – Enhanced Command Retry**
Change Date	**9/17/2018**
Builds on Specification	**NVM Express 1.3**

Technical Proposal Author(s)

Name	Company
David Black	Dell EMC

This technical proposal adds command retry delays, a new error code that indicates that a command should be retried without providing further information and a new Feature that allows a host to indicate that it supports this new functionality, and other functionality that may be added.

Revision History

Revision Date	Change Description
2018/02/19	Initial version
2018/02/22	Technical WG confirmed use of both reserved bits in Status field to support three retry delays –additional CQE space is available beyond those two bits. Remove Delay 0 from Identify Controller data structure. Use one bit in new Host Optional Functionality feature to indicate support for both command retry delays and the new error status code, not two bits. Use 4k buffer for new NVMe feature, First 64 bytes are for enable bits, larger fields would start after that. Only define 1 bit for now.
2018/02/26	Add alternative to use multiple command Dwords instead of data transfer for Set Features. Rework description of new bit to enable functionality, based on suggestion from Kevin Marks. Change all Figure references to NVMe 1.3 Gold version. Fix description of zero retry delay. Additional minor edits.
2018/03/01	Reduce retry delay times to 2 bytes from 4 bytes. Use data buffer in both directions (Set & Get) for symmetry instead of emulating Host Memory Buffer feature design that uses command Dwords for Set and a data buffer for Get. Remove formatting of reserved bytes in new Feature. Additional editorial changes.
2018/03/05	Fix offsets in Identify Controller data structure Additional minor edits
2018/03/29	More minor edits, and produce clean version for phase 2 exit vote.
2018/03/30	Fix thinko in table footnote.
2018/04/05	Minor table footnote format corrections to Figure 29
2018/04/19	Rename new Feature to Host Behavior Support and explain its usage.
2018/04/24	Extensive edits to explanation of new Host Behavior Support feature.
2018/04/26	Minor edits to explanation of Host Behavior Support feature.
2018/05/10	Change two instances of ‘zero” to 0h. Clean version for member review.
2018/05/29	Member review comments from Fred Knight – don’t assume that next version of NVMe will be numbered 1.4, and make it explicit that setting the new Feature overrides all prior setting of that new Feature.
2018/06/03	Open issue from member review: should new Feature be saveable? Editorial corrections from Harvey Newman.
2018/06/13	Two more member review issues in new Feature: 1. Redesign so that host behaviors can be declared independently? 2. Shrink size, e.g., to reduce space in persistent event log?
2018/06/19	Initial resolutions to member review issues: - New Feature is not saveable to support host software changes. - Change Feature layout from bits to bytes and specify use of Set Features Command Dword 11 to anticipate possible future addition of per-behavior functionality. - Note size as an open issue for discussion. Additional minor edits
2018/06/21	Back out Command Dword 11 changes – instead, host would have to check for support of additional functionality via an addition to Identify Controller. Remove related editor’s note. Shorten feature size to 512 bytes. Additional minor edits
2018/07/01	Clean version for second member review
2018/07/06	Edit CRD field description based on feedback from John Geldman. Removed duplicate requirements from the CRDTn field descriptions.
2018/07/13	Fix typo in ACRE field description in new feature.
2018/08/23	Post-second-member-review version for integration, no additional changes.
2018/09/04	Final edits from incorporation.

2018/09/17	Ratified

Description of Specification Changes

Modify section 4.6.1 (Status Field Definition) as shown below:

4.6.1 Status Field Definition

The Status Field defines the status for the command indicated in the completion queue entry, defined in Figure 29.

A value of 0h for the Status Field indicates a successful command completion, with no fatal or non-fatal error conditions. Unless otherwise noted, if a command fails to complete successfully for multiple reasons, then the particular status code returned is chosen by the vendor.

Figure 29: Completion Queue Entry: Status Field

Bit	Description
31	**Do Not Retry (DNR):** If set to '1', indicates that if the same command is re-submitted it is expected to fail. If cleared to '0', indicates that the same command may succeed if retried. If a command is aborted due to time limited error recovery (refer to section 5.21.1.5), this field should be cleared to '0'. If the SCT and SC fields are cleared to 0h then this field should be cleared to '0'.
30	**More (M):** If set to '1', there is more status information for this command as part of the Error Information log that may be retrieved with the Get Log Page command. If cleared to '0', there is no additional status information for this command. Refer to section 5.14.1.1.
29:28	~~Reserved~~ **Command Retry Delay (CRD):** If the DNR bit is cleared to '0' and the host has set the Advanced Command Retry Enable (ACRE) field to 1h in the Host Behavior Support feature (refer to section 5.21.1.new), then: a. a zero CRD value indicates a zero command retry delay time (i.e., the host may retry the command immediately); and b. a non-zero CRD value selects a field in the Identify Controller data structure (refer to Figure 109) that indicates the command retry delay time: • a 01b CRD value selects the Command Retry Delay Time 1 (CRDT1) field; • a 10b CRD value selects the Command Retry Delay Time 2 (CRDT2) field; and • a 11b CRD value selects the Command Retry Delay Time 3 (CRDT3) field. The host should not retry the command until at least the amount of time indicated by the selected field has elapsed. It is not an error for the host to retry the command prior to that time. If the DNR bit is set to'1' in the Status field or the ACRE field is cleared to 0h in the Host Behavior Support feature, then this field is reserved. If the SCT and SC fields are cleared to 0h, then this field should be cleared to 0h.
27:25	**Status Code Type (SCT):** Indicates the status code type of the completion queue entry. This indicates the type of status the controller is returning.
24:17	**Status Code (SC):** Indicates a status code identifying any error or status information for the command indicated.

Modify Figure 31: Status Code – Generic Command Status Values as shown below:

Figure 31: Status Code – Generic Command Status Values

Value	Description
1Eh	**SGL Data Block Granularity Invalid:** The Address alignment or Length granularity for an SGL Data Block descriptor is invalid. This may occur when a controller supports Dword granularity only and the lower two bits of the Address or Length are not cleared to 00b. NOTE: An implementation compliant to revision 1.2.1 or earlier may use the status code value of 15h to indicate SGL Data Block Granularity Invalid.
1Fh	**Command Not Supported for Queue in CMB:** The implementation does not support submission of the command to a Submission Queue in the Controller Memory Buffer or command completion to a Completion Queue in the Controller Memory Buffer. NOTE: Revision 1.3 uses this status code only for Sanitize commands.
20h	**Namespace is Write Protected:** The command is prohibited while the namespace is write protected by the host.
21h	**Command Interrupted:** Command processing was interrupted and the controller is unable to successfully complete the command. The host should retry the command. If this status code is returned, then the controller shall clear the Do Not Retry bit to '0' in the Status field of the CQE (refer to Figure 29). The controller shall not return this status code unless the host has set the Advanced Command Retry Enable (ACRE) field to 1h in the Host Behavior Support feature (refer to section 5.21.1.new).
~~21~~22h – 7Fh	Reserved
80h – BFh	I/O Command Set Specific
C0h – FFh	Vendor Specific

Modify Figure 84: Get Features – Feature Identifiers as shown below:

Figure 1: 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	Section 5.21.1.18
Predictable Latency Mode Config	Section 5.21.1.19
Predictable Latency Mode Window	Section 5.21.1.20
LBA Status Information Report Interval	Section 5.21.1.21
Host Behavior Support	Section 5.21.1.22
NVM Command Set Specific	
Software Progress Marker	Section 5.21.1.2~~1~~3
Host Identifier	Section 5.21.1.2~~2~~4
Reservation Notification Mask	Section 5.21.1.2~~3~~5
Reservation Persistence	Section 5.21.1.2~~4~~6
Namespace Write Protection Config	Section 5.21.1.2~~5~~7

Modify Figure 109: Identify – Identify Controller Data Structure as shown below:

Figure 109: Identify – Identify Controller Data Structure

Bytes	O/M	Description
111:102		Reserved
127:112	O	**FRU Globally Unique Identifier (FGUID):** This field contains a 128-bit value that is globally unique for a given Field Replaceable Unit (FRU). Refer to the NVM Express Management Interface (NVMe-MI) specification for the definition of a FRU. This field remains fixed throughout the life of the FRU. This field shall contain the same value for each controller associated with a given FRU. This field uses the EUI-64 based 16-byte designator format. Bytes 122:120 contain the 24-bit Organizationally Unique Identifier (OUI) value assigned by the IEEE Registration Authority. Bytes 127:123 contain an extension identifier assigned by the corresponding organization. Bytes 119:112 contain the vendor specific extension identifier assigned by the corresponding organization. See the IEEE EUI-64 guidelines for more information. This field is big endian (refer to section 7.10). When not implemented, this field contains a value of 0h.
129:128	O	**Command Retry Delay Time 1 (CRDT1):** If the Do Not Retry (DNR) bit is cleared to '0' in the CQE and the Command Retry Delay (CRD) field is set to 01b in the CQE, then this value indicates the command retry delay time in units of 100 milliseconds.
131:130	O	**Command Retry Delay Time 2 (CRDT2):** If the DNR bit is cleared to '0' in the CQE and the CRD field is set to 10b in the CQE, then this value indicates the command retry delay time in units of 100 milliseconds.
133:132	O	**Command Retry Delay Time 3 (CRDT3):** If the DNR bit is cleared to '0' in the CQE and CRD field is set to 11b in the CQE, then this value indicates the command retry delay time in units of 100 milliseconds.
239:~~128~~134		Reserved
255:240		Refer to the NVMe Management Interface Specification for definition.

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

Figure 2: 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
13h	O	Yes	Yes	Predictable Latency Mode Config
14h	O	Yes	No	Predictable Latency Mode Window
~~15h~~	~~O~~	~~No~~	~~No~~	~~Namespace Write Protection Config~~
15h	O	No	No	LBA Status Information Report Interval
16h	O	No	Yes	Host Behavior Support
~~16~~17h – 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.

Modify Figure 135: Set Features, NVM Command Set Specific – Feature Identifiers as shown below:

Figure 3: Set Features, NVM Command Set Specific – Feature Identifiers

Feature Identifier	O/M[4]	Persistent Across Power Cycle and Reset[1]	Uses Memory Buffer for Attributes	Description
80h	O	Yes	No	Software Progress Marker
81h	O[2]	No	Yes	Host Identifier
82h	O[3]	No	No	Reservation Notification Mask
83h	O[3]	Yes	No	Reservation Persistance
84h	O	No	No	Namespace Write Protection Config
~~84~~85h – BFh				Reserved

NOTES:
1. 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.
2. Mandatory if reservations are supported as indicated in the Identify Controller data structure.
3. Mandatory if reservations are supported by the namespace as indicated by a non-zero value in the Reservation Capabilities (RESCAP) field in the Identify Namespace data structure.
4. O/M: O = Optional, M = Mandatory.

Insert a new Section 5.21.1.22 (Host Functionality Support (Feature Identifier 16h)) in numerical order of Feature Identifier, renumbering as needed, as shown below:

5.21.1.22 Host Behavior Support (Feature Identifier 16h)

This Feature enables use of controller functionality that is associated with and depends upon specific host behavior that may or may not be supported by all hosts. A controller does not use such functionality unless the host has indicated that it supports the specific host behavior upon which the functionality depends. The host indicates that support to the controller by setting a field in this Feature. That host action enables controller use of the associated functionality with that host. A controller shall not use functionality with a host that has not indicated support for the associated specific host behavior upon which that controller functionality depends. The attributes in Figure New-1 are transferred in the data buffer.

For example, the Command Interrupted status code is associated with and depends upon the specific host behavior that the host is expected to retry commands that are aborted with that status code. That command retry behavior may or may not be supported by all hosts (e.g., hosts based on versions of NVMe 1.3 and earlier are unlikely to retry commands aborted with the Command Interrupted status code as that status code was introduced after NVMe 1.3). A host that supports that command retry behavior indicates its support to the controller by setting a field to 1h in the Host Behavior Support Feature. Setting that field to 1h enables controller use of the Command Interrupted status code, with the result that this status code is used only with hosts that have indicated support for the associated command retry behavior.

Controllers shall not support a saveable value for this Feature, as host reboot with different host software could cause the contents of this Feature to become incorrect. The default value of this Feature shall be all bytes cleared to 0h.

After a successful completion of a Set Features command for this Feature, the controller may use controller-to-host functionality that depends on specific host behavior as indicated by the attributes. If multiple Set Features commands for this Feature are processed by the controller, only information from the most recent successful command is retained (i.e., subsequent commands replace information provided by previous commands).

If a Get Features command is submitted for this Feature, the attributes specified in Figure New-1 are returned in the data buffer for that command.

Figure New-1: Host Behavior Support – Data Structure

Bytes	Description
00	**Advanced Command Retry Enable (ACRE):** If set to 1h, then the Command Interrupted status code is enabled (refer to Figure 31) and command retry delays are enabled. The controller may use the Command Interrupted status code and may indicate a command retry delay by setting the Command Retry Delay (CRD) field to a non-zero value in the Status field of a Completion Queue Entry, refer to Figure 29. A host that sets this field to 1h indicates host support for the command retry behaviors that are specified for both the Command Interrupted status code and non-zero values in the CRD field. If cleared to 0h, then both the Command Interrupted status code and command retry delays are disabled. The controller shall not use the Command Interrupted status code, and shall clear the CRD field to 0h in all CQEs. All values other than 0h and 1h are reserved.
511:01	Reserved