NVM Express® Technical Proposal

Technical Proposal ID	TP4160 WZSL Limit Modification
Revision Date	2023.11.07
Builds on Specification(s)	NVM Express Base Specification 2.0c
	NVM Command Set Specification 1.0a
References	TP 4099 MDTS Enhancement

Technical Proposal Author(s)

Name	Company
Yoni Shternhell	WDC
David Black	Dell Technologies

Technical Proposal Overview

This proposal suggests to optionally let the controller advertise a larger limit of logical blocks that can be processes for Write Zeroes commands with the Deallocate bit set to '1' and to add to the Write Zeroes command a bit that allows the controller to use the Write Zeroes command to deallocate the entire namespace when the Deallocate bit set to '1'.

Revision History

Revision Date	Change Description
2023.03.27	Initial draft
2023.04.26	Added the Namespace Write Zeroes bit to write zeroes with deallocate the entire namespace. Added new bit in ONCS for support
2023.05.03	Some clarifications for the WZSL field
2023.06.08	Modifications from TWG call
2023.06.09	Edits from David Black, mostly to specify functionality of the WZDSL field and the ONCS new support bit.
2023.06.10	More edits from Fred on completion queue and write zeroes to the entire namespace
2023.06.22	Modifications to the Write Zeroes Size Limit (WZSL) field
2023.07.13	More modifications to the WZSL and WZDSL fields
2023.07.20	Adding support bit TBD to the new WZDSL field depends on the Deallocate bit. Changes to bit TBD+1 in terms of support for the NSZ bit Aligned with TP4099 Modifications in the "Command Completion" section
2023.07.20a	Accept changes after group call
2023.07.27	Phase 2 exit version
2023.08.02	Completed phase 2 exit content: • Explain function of NSZ bit, specify errors if used incorrectly. • Explain purpose of bit in CQE indicating whether entire namespace has been cleared.
2023.08.07	Phase 3 1st revision
2023.08.24	Incorporated Mike's comments
2023.08.28	Editorial cleanups, mostly of "refer to" items, added explanation of zero value of WZDSL field.
2023.11.07	Integrated

Description for Changes Document for

NVM Express Base Specification 2.0c

Feature Enhancement:

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

New requirement / incompatible change:

- References: TP 4099 MDTS Enhancement

NVM Express NVM Command Set Specification 1.0c

Feature Enhancement:

- 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

Markup Conventions:

Black:	Unchanged (however, hot links are removed)
Red Strikethrough:	Deleted
Blue:	New
Blue Highlighted:	TBD values, anchors, and links to be inserted in new text.
Green Strikethrough:	Deleted from this location and moved to another location
Green:	Deleted from another location and moved to this location
<Green Bracketed>:	Notes to editor
Orange	Text is pulled in from a referenced Technical Proposal

Description of Specification Changes for the NVM Express Base Specification 2.0c

Modify portions of section 5 as shown below:

5   Admin Command Set

5.23   Identify command...

5.23.2 Identify Data Structures

5.23.2.1   Identify Controller Data Structure (CNS 01h)

…

Figure 275: Identify – Identify Controller Data Structure, I/O Command Set Independent

Bytes	I/O[1]	Admin[1]	Disc[1]	Description
…				
				NVM Command Set Attributes
521:520	M	M	R	**Optional NVM Command Support (ONCS):** This field indicates the optional I/O commands and features supported by the controller. Refer to section 3.1.2. Bits 15:13~~9~~ are reserved. Bit 12 (Namespace Zeroes Support) if set to '1', then the controller supports the Namespace Zeroes (NSZ) bit in the NVM Command Set Write Zeroes command. If cleared to '0', then the controller does not support the Namespace Zeroes (NSZ) bit in the Write Zeroes command. If the Write Zeroes command is not supported, then this bit shall be cleared to '0'. Bit 11 (Maximum Write Zeroes with Deallocate) if set to '1', then the maximum data size for the NVM Command Set Write Zeroes command depends on the value of the Deallocate bit in the Write Zeroes command and the value in the WZDSL field in the I/O Command Set specific Identify Controller data structure for the NVM Command Set (refer to the I/O Command Set Specific Identify Controller Data Structure (CNS 06h, CSI 00h) section in the NVM Command Set Specification). If cleared to '0', then this bit has no effect. If the Write Zeroes command is not supported or the WZSL field in that data structure is cleared to 0h, then this bit shall be cleared to '0'. Bit 8 if set to '1', then the controller supports the NVM Command Set Copy command. If cleared to '0', then the controller does not support the NVM Command Set Copy command. Bit 7 if set to '1', then the controller supports the NVM Command Set Verify command and the Verify Size Limit (VSL) field indicates the recommended maximum data size for Verify commands. If cleared to '0', then controller support of the NVM Command Set Verify command is indicated by a non-zero data size limit in the VSL field. Bit 6 if set to '1', then the controller supports the Timestamp feature. If cleared to '0', then the controller does not support the Timestamp feature. Refer to section 5.27.1.11. Bit 5 if set to '1', then the controller supports reservations. If cleared to '0', then the controller does not support reservations. If the controller supports reservations, then the following commands associated with reservations shall be supported: Reservation Report, Reservation Register, Reservation Acquire, and Reservation Release. Refer to section 8.19 for additional requirements. Bit 4 if set to '1', then the controller supports the Save field set to a non-zero value in the Set Features command and the Select field set to a non-zero value in the Get Features command. If cleared to '0', then the controller does not support the Save field

Bytes	I/O[1]	Admin[1]	Disc[1]	Description
				set to a non-zero value in the Set Features command and the Select field set to a non-zero value in the Get Features command.
				Bit 3 if set to '1', then the controller supports the NVM Command Set Write Zeroes command and the Write Zeroes Size Limit (WZSL) field indicates the recommended maximum data size for Write Zeroes commands. If cleared to '0', then controller support of the NVM Command Set Write Zeroes command is indicated by a non-zero data size limit in the WZSL field.
				Bit 2 if set to '1', then the controller supports the NVM Command Set Dataset Management command and limits, if any, on controller support of the Dataset Management command are indicated by non-zero values in the Dataset Management Ranges Limit (DMRL) field, the Dataset Management Size Limit (DMSL) field and the Dataset Management Range Size Limit (DMRSL) field. If cleared to '0', then controller support of the NVM Command Set Dataset Management command is indicated by a non-zero data size limit in the DMRL, DMSL, and DMRSL fields.
				Bit 1 if set to '1', then the controller supports the NVM Command Set Write Uncorrectable command and the Write Uncorrectable Size Limit (WUSL) field indicates the recommended maximum data size for Write Uncorrectable commands. If cleared to '0', then controller support of the NVM Command Set Write Uncorrectable command is indicated by a non-zero data size limit in the WUSL field.
				Bit 0 if set to '1', then the controller supports the NVM Command Set Compare command. If cleared to '0', then the controller does not support the NVM Command Set Compare command.
				NOTE: This field applies to all I/O Command Sets. The original name has been retained for historical continuity.
…				

Description of Specification Changes for the NVM Command Set Specification 1.0a

Modify portions of section 3 as shown below:

3   I/O Commands for the NVM Command Set

…

3.2.8   Write Zeroes command

The Write Zeroes command is used to ~~set~~ clear a range of logical blocks or all of the logical blocks in an entire namespace to zero. Non-PI related metadata for this command, if any, shall be all bytes cleared to 0h. The protection information for logical blocks written to the media is updated based on CDW12.PRINFO. If the Protection Information Action bit (PRACT) is cleared to '0', then the protection information for this command shall be all zeroes. If the Protection Information Action bit (PRACT) is set to '1', then the protection information shall be based on the End-to-end Data Protection Type Settings (DPS) field in the Identify Namespace data structure (refer to Figure 97), CDW15.LBATM, CDW15.LBAT, as well as CDW2/3 and CDW14 content as described in section 5.2.1.4.1.

After successful completion of a Write Zeroes ~~this~~ command that specifies:

- clearing a range of logical blocks to zero, the value returned by subsequent reads of logical blocks in this range shall be all bytes cleared to 0h until a write occurs to ~~this LBA~~ that range of logical blocks; or
- clearing all of the logical blocks in an entire namespace to zero, the value returned by subsequent reads of each logical block in that namespace shall be all bytes cleared to 0h until a write occurs to that logical block.

If the Deallocate bit (CDW12.DEAC) is set to '1' in a Write Zeroes command, and the namespace supports clearing all bytes to 0h in the values read (e.g., bits 2:0 in the DLFEAT field are set to 001b) from a deallocated logical block and its metadata (excluding protection information), then for each specified logical block, the controller:

- should deallocate that logical block;
- shall return all bytes cleared to 0h in the values read from:
  - that logical block; and
  - that logical blocks metadata (excluding protection information);
- and
- shall return the protection information in that logical block as specified in section 3.2.3.2.1.

If the Deallocate bit is cleared to '0' in a Write Zeroes command, and the namespace supports clearing all bytes to 0h in the values read (e.g., bits 2:0 in the DLFEAT field are set to 001b) from a deallocated logical block and its metadata (excluding protection information), then, for each specified logical block, the controller:

- may deallocate that logical block;
- shall return all bytes cleared to 0h in the values read from:
  - that logical block; and
  - that logical blocks metadata (excluding protection information);
- and
- shall return the protection information in that logical block based on CDW12.PRINFO in that Write Zeroes command.

For each logical block in the range specified by a Write Zeroes command, if the namespace does not support that logical block clearing all bytes to 0h in the values read from that logical block and its metadata (excluding the protection information) read, then the controller shall not deallocate that logical block.

If the Write Zeroes Size Limit (WZSL) field in the I/O Command Set specific Identify Controller data structure for the NVM Command Set (refer to Figure 102) is set to a non-zero value and the Write Zeroes with Deallocate Size Limit (WZDSL) field in that data structure is cleared to 0h, then ~~and~~:

a) if bit 3 in the Optional NVMe Command Support field in the Identify Controller data structure is set to '1', then the WZSL field indicates the recommended maximum data size for the Write Zeroes command and any Write Zeroes command that specifies a logical block range whose data size exceeds that recommended maximum may encounter delays in processing; and

b) if bit 3 in the ONCS field is cleared to '0', then the WZSL field indicates the maximum data size limit for the Write Zeroes command, and the controller shall abort any Write Zeroes command that specifies a logical block range whose data size exceeds that limit with a status of Invalid Field in Command.

If the WZSL field is set to a non-zero value and the WZDSL field is set to a non-zero value, then:

a) if bit 3 in the Optional NVMe Command Support (ONCS) field in the Identify Controller data structure (refer to the Identify Controller Data Structure (CNS 01h) section in the NVM Express Base Specification) is set to '1', then:
  - the WZDSL field indicates the recommended maximum data size for any Write Zeroes command that has the Deallocate bit set to '1';

- the WZSL field indicates the recommended maximum data size for any Write Zeroes command that has the Deallocate bit cleared to '0'; and
- any Write Zeroes command that specifies a logical block range whose data size exceeds the applicable recommended maximum may encounter delays in processing;

and

b) if bit 3 in the ONCS field in the Identify Controller data structure is cleared to '0', then:
- the WZDSL field indicates the maximum data size limit for any Write Zeroes command that has the Deallocate bit set to '1';
- the WZSL field indicates the maximum data size limit for any Write Zeroes command that has the Deallocate bit cleared to '0'; and
- the controller shall abort any Write Zeroes command that specifies a logical block range whose data size exceeds the applicable data size limit with a status code of Invalid Field in Command.

The WZSL field and the WZDSL field do not apply to Write Zeroes commands that have the Namespace Zeroes (NSZ) bit (refer to Figure 73) set to '1'.

The Namespace Zeroes (NSZ) bit is set to '1' to request that the controller clear all of the logical blocks to zero in the entire specified namespace by deallocating all logical blocks in that namespace. This functionality is only supported when the Deallocate bit is also set to '1' and the specified namespace supports clearing all bytes to 0h in the values read from a deallocated logical block and its metadata (excluding protection information) (e.g., as indicated by the DLFEAT field in Identify Namespace data structure being set to 001b (refer to Figure 97)).

If the NSZ bit is set to '1' in a Write Zeroes command and either:

- the Deallocate bit is cleared to '0'; or
- the specified namespace does not support clearing all bytes to 0h in the values read from a deallocated logical block and its metadata (excluding protection information),

then the controller shall abort that command with a status code of Invalid Field in Command.

Controller support for the NSZ bit is indicated by setting the NSZS bit to '1' in the ONCS field in the Identify Controller data structure. If the controller supports the NSZ bit and the NSZ bit is set to '1' in a Write Zeroes command, then the controller ignores the Starting LBA (SLBA) field (refer to Figure 72) and the Number of Logical Blocks (NLB) field (refer to Figure 73) in that command.

If the controller does not support the NSZ bit (i.e., the NSZS bit in the ONCS field in the Identify Controller data structure is cleared to '0'), then the controller is not required to check whether the NSZ bit is cleared to '0'; and may clear to zero the range of logical blocks specified by the SLBA field and NLB field instead of clearing all of the logical blocks to zero in the entire specified namespace.

The host is able to determine whether or not the controller cleared all of the logical blocks to zero in the entire specified namespace by checking bit 0 of Dword 0 in the Completion Queue Entry (CQE) for the Write Zeroes command. The controller sets that bit to '1' if all of the logical blocks have been cleared to zero in the entire specified namespace (refer to section 3.2.8.1). The host should check this bit for any Write Zeroes command in which the NSZ bit is set to '1' to determine whether or not the controller cleared all of the logical blocks to zero in the entire specified namespace.

The fields used are Command Dword 2, Command Dword 3, Command Dword 10, Command Dword 11, Command Dword 12, Command Dword 14, and Command Dword 15 fields.

Figure 71: Write Zeroes – Command Dword 2 and Dword 3

Bits	Description
63:48	Reserved
47:00	This field and Command Dword 14 specify the variable sized Logical Block Storage Tag (LBST) and Initial Logical Block Reference Tag (ILBRT) fields, which are defined in section 5.2.1.4.1. If the namespace is not formatted to use end-to-end protection information, then this field is ignored.

Figure 72: Write Zeroes – Command Dword 10 and Command Dword 11

Bits	Description
63:00	**Starting LBA (SLBA):** This field indicates the 64-bit address of the first logical block to be written as part of the operation. Command Dword 10 contains bits 31:00; Command Dword 11 contains bits 63:32. If the NSZ bit (refer to Figure 73) is set to '1' and the NSZS bit is set to '1' in the Optional NVMe Command Support (ONCS) field in the Identify Controller data structure, then this field should be cleared to 0h by the host and shall be ignored by the controller.

Figure 73: Write Zeroes – Command Dword 12

Bits	Description
31	**Limited Retry (LR):** If set to '1', the controller should apply limited retry efforts. If cleared to '0', the controller should apply all available error recovery means to write the data to the NVM.
30	**Force Unit Access (FUA):** If set to '1', then the controller shall write the data, and metadata, if any, to non-volatile media before indicating command completion. There is no implied ordering with other commands. If cleared to '0', then this bit has no effect.
29:26	**Protection Information Field (PRINFO):** Specifies the protection information action and check field, as defined in Figure 9. The Protection Information Check (PRCHK) field shall be cleared to 000b.
25	**Deallocate (DEAC):** If set to '1', then the host is requesting that the controller deallocate the specified logical blocks. If cleared to '0', then the host is not requesting that the controller deallocate the specified logical blocks. If the NSZ bit is set to '1' and this bit is cleared to '0', then the controller shall abort the command with a status code of Invalid Field in Command.
24	**Storage Tag Check (STC):** This bit specifies the Storage Tag field shall be checked as part of end-to-end data protection processing as defined in Figure 10. This bit shall be cleared to '0'.
23	**Namespace Zeroes (NSZ):** If set to '1' and the Deallocate bit is set to '1', then the Write Zeroes command is requesting that the controller clear all logical blocks to zero in the entire namespace. If bit NSZS in the Optional NVM Command Support (ONCS) field in the Identify Controller data structure is cleared to '0', then this bit has no effect.
~~23~~22:16	Reserved
15:00	**Number of Logical Blocks (NLB):** This field indicates the number of logical blocks to be written. This is a 0's based value. If the NSZ bit is set to '1', then this field should be cleared to 0h by the host and shall be ignored by the controller.

Figure 74: Write Zeroes – Command Dword 14

Bits	Description
31:00	This field and bits 47:00 of Command Dword 2 and Dword 3 specify the variable sized Logical Block Storage Tag (LBST) and Initial Logical Block Reference Tag (ILBRT) fields, which are defined in section 5.2.1.4.1. If the namespace is not formatted to use end-to-end protection information, then this field is ignored.

Figure 75: Write Zeroes – Command Dword 15

Bits	Description
31:16	**Logical Block Application Tag Mask (LBATM):** This field indicates the Application Tag Mask value. If the namespace is not formatted to use end-to-end protection information, then this field is ignored. Refer to section 5.1.
15:00	**Logical Block Application Tag (LBAT):** This field indicates the Application Tag value. If the namespace is not formatted to use end-to-end protection information, then this field is ignored. Refer to section 5.1.

…

3.2.8.1 Command Completion

Upon completion of the Write Zeroes command, the controller shall post a completion queue entry to the associated I/O Completion Queue indicating the status for the command.

If the command does not complete successfully (i.e., completes with a status code other than Successful Completion), then the controller may or may not have completed the requested clearing of logical blocks to zero.

If the command has the Namespace Zeroes (NSZ) bit cleared to '0' and completes successfully, then the logical blocks specified by the command have been cleared to zero. If the command has the Namespace Zeroes (NSZ) bit set to '1' and completes successfully, then:
- If bit 0 of Dword 0 of the completion queue entry is cleared to '0', then the number of logical blocks specified by the Number of Logical Blocks (NLB) field have been cleared to zero; and
- If bit 0 of Dword 0 of the completion queue is set to '1', then all logical blocks in the entire namespace have been cleared to zero.

…

Modify portions of section 4 as shown below:

4   Admin Commands for the NVM Command Set

4.1   Admin Command behavior for the NVM Command Set

4.1.5   Identify Command

4.1.5.4 I/O Command Set Specific Identify Controller data structure (CNS 06h)

…

Figure 102: I/O Command Set Specific Identity Controller Data Structure for the NVM Command Set

Bytes	O/M [1]	Description

00	O	**Verify Size Limit (VSL):** If bit 7 in the Optional NVM Command Support (ONCS) field in the Identify Controller data structure (refer to the Identify Controller Data Structure (CNS 01h) section in the NVM Express Base Specification) is set to '1' then:
		
		   a) a non-zero value in this field indicates the recommended maximum data size for a Verify command (refer to section 3.2.5); and
		   b) a value of 0h in this field indicates that no recommended maximum data size for a Verify command is reported.
		
		If bit 7 in the ONCS field is cleared to '0' then:
		
		   a) a non-zero value in this field indicates that the controller supports the Verify command with the maximum data size limit indicated by this field (refer to section 3.2.5); and
		   b) a value of 0h in this field indicates that the controller does not support the Verify command.
		
		The non-zero value is in units of the minimum memory page size (CAP.MPSMIN) and is reported as a power of two (2^n).
		
		If the MEM bit is cleared to '0' in the CTRATT field in the Identify Controller data structure, then ~~T~~this field includes the length of metadata, if metadata is interleaved with the logical block data.
		
		If the MEM bit is set to '1', then this field excludes the length of metadata.

01	O	**Write Zeroes Size Limit (WZSL):** If bit 3 in the Optional NVM Command Support (ONCS) field is set to '1' then:
		
		   a) a non-zero value in this field indicates the recommended maximum data size for a Write Zeroes command (refer to section 3.2.8); and
		   b) a value of 0h in this field indicates that no recommended maximum data size for a Write Zeroes command is reported.
		
		If bit 3 in the ONCS field is cleared to '0' then:
		
		   a) a non-zero value in this field indicates that the controller supports the Write Zeroes command with the maximum data size limit indicated by this field (refer to section 3.2.8); and
		   b) a value of 0h in this field indicates that the controller does not support the Write Zeroes command.
		
		If the MAXWZD bit in the ONCS field is set to '1', then the controller supports a larger maximum data size for Write Zeroes commands with the Deallocate bit set to '1' than the controller supports for Write Zeroes commands that have the Deallocate bit cleared to '0', and the controller shall:
		
		   • set this field to a non-zero maximum data size value that applies to Write Zeroes commands with the Deallocate bit cleared to '0'; and
		   • set the Write Zeroes with Deallocate Size Limit (WZDSL) field to a larger non-zero maximum data size value that applies to Write Zeroes commands with the Deallocate bit set to '1'.
		
		The non-zero value is in units of the minimum memory page size (CAP.MPSMIN) and is reported as a power of two (2^n).
		
		If the MEM bit is cleared to '0' in the CTRATT field in the Identify Controller data structure, then ~~T~~this field includes the length of metadata, if metadata is interleaved with the logical block data.
		
		If the MEM bit is set to '1', then this field excludes the length of metadata.

02	O	**Write Uncorrectable Size Limit (WUSL):** If bit 1 in the Optional NVM Command Support (ONCS) field is set to '1' then:
		
		   a) a non-zero value in this field indicates the recommended maximum data size for a Write Uncorrectable command (refer to section 3.2.7); and
		   b) a value of 0h in this field indicates that no recommended maximum data size for a Write Uncorrectable command is reported.

		If bit 1 in the ONCS field is cleared to '0' then:
		
		    a) a non-zero value in this field indicates that the controller supports the Write Uncorrectable command with the maximum data size limit indicated by this field (refer to section 3.2.7); and
		    b) a value of 0h in this field indicates that the controller does not support the Write Uncorrectable command.
		
		The non-zero value is in units of the minimum memory page size (CAP.MPSMIN) and is reported as a power of two (2^n).
		
		If the MEM bit is cleared to '0' in the CTRATT field in the Identify Controller data structure, then Tthis field includes the length of metadata, if metadata is interleaved with the logical block data.
		
		If the MEM bit is set to '1', then this field excludes the length of metadata.

…		

15:08	O	**Dataset Management Size Limit (DMSL):** If bit 2 in the Optional NVM Command Support (ONCS) field is set to '1' then:
		
		    a) a non-zero value in this field indicates the recommended maximum total number of logical blocks for a Dataset Management command (refer to section 3.2.3).
		    b) a value of 0h in this field indicates that no recommended maximum total number of logical blocks for a Dataset Management command is reported.
		
		If bit 2 in the ONCS field is cleared to '0', then:
		
		    a) a non-zero value in this field indicates that the controller supports the Dataset Management command with the maximum total number of logical blocks limit indicated by this field (refer to section 3.2.3); and
		a value of 0h in this field indicates that the controller does not support the Dataset Management command.

17	O	**Write Zeroes With Deallocate Size Limit (WZDSL):** A non-zero value in this field indicates the maximum data size for Write Zeroes commands with the Deallocate bit set to '1'. A 0h value in this field indicates that the maximum data size for Write Zeroes commands does not depend on the value of the Deallocate bit.
		
		For Write Zeroes commands with the Deallocate bit cleared to '0', this field has no effect (refer to the WZSL field).
		
		For Write Zeroes commands with the Deallocate bit set to '1', if this field is set to a non-zero value, then:
		
		    a) if bit 3 in the ONCS field is set to '1', then the value in this field is the recommended maximum data size and the value in the WZSL field is not used; and
		
		    b) if bit 3 in the ONCS field is cleared to '0', then the value in this field is the maximum data size limit and the value in the WZSL field is not used.
		
		If the WZSL field is cleared to 0h, then this field shall be cleared to 0h. If the WZSL field is not cleared to 0h, then the value of this field shall either be 0h or greater than the value in the WZSL field.
		
		The non-zero value is in units of the minimum memory page size (CAP.MPSMIN) and is reported as a power of two (2^n).
		
		If the MEM bit is cleared to '0' in the CTRATT field in the Identify Controller data structure, then this field includes the length of metadata, if metadata is interleaved with the logical block data.
		
		If the MEM bit is set to '1', then this field excludes the length of metadata.

4095:16 24	M	