LEGAL NOTICE:

NOTICE TO USERS WHO ARE NVM EXPRESS, INC. MEMBERS: Members of NVM Express, Inc. have the right to use and implement this NVM Express revision 2.0 technical proposal 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: **“© 2007 - 2021 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.

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

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**4090 – Enhanced Deallocation Granularity**
Change Date	**2021-06-24**
Builds on Specification	**NVM Command Set Specification 1.0**
Ratified Technical Proposals Referenced	**None**
Ratified ECNs Referenced	**ECN 001**

Technical Proposal Author(s)

Name	Company
Paul Suhler	Kioxia
Lee Prewitt	Microsoft

This technical proposal defines a mechanism for a controller to indicate to a host information used to select the length and alignment of a range of logical blocks to be deallocated. The valid reportable values are in the range [1 … 2^{32}-1].

Revision History

Revision Date	Change Description
2020-07-23	Initial version
2020-07-24	Moved new fields to the NVM Command Set Identify Namespace data structure (CNS 05h), which was defined in TP 4056. Added TP 4056 as a referenced ratified TP.
2020-08-06	Changes from 2020-08-06 Technical WG meeting: • Expanded OPTPERF from one to two bits, to simplify specifying requirements for legacy (16-bit) and new (32-bit) fields. • Simplified field definitions. • Reworded requirement for namespace NOWS to match the Optimal Write Size in the NVM Set Attributes Entry, if any. • Rewrote parts of section 8.25 to generalize NPDG and NPDA values from the field names.
2020-08-24	Changes from 2020-08-20 Technical WG meeting: • Reverted from a two-bit field into two one-bit fields. • Bit 5 set to ‘1’ defines new behavior, and recommends that host ignore (existing) bit 4. • Revised form of the two-bit field version is included for reference.
2020-10-01	• Aligned to NVMe 1.4b and TP 4056b. • Capitalized the Length in Logical Blocks field (Figure 370). Changes from 2020-10-01 Technical WG meeting: • Deleted NPDAD (alignment) field. • Expanded OPTPERF to a two-bit field, as suggested by NetApp ballot comment. • Corrected spelling OPTPEFT to OPTPERF. • Replaced the use of the acronyms NPDG and NPDA with words when not referring to fields.

2020-10-02	• Separated the descriptions of OPTPERF 10b and 11b to support the NPDG field for OPTPERF = 11b. • Fixed OPTPERF references in NPWG and NOWS field descriptions. • Moved the NOWS / OWS equality requirement from the OPTPERF field description to the NOWS field description. • Clarified parts of section 8.25.
2020-11-10	• Moved TP 4056 changes into the NVMe 1.4b changes section. • Added missing statement that NPWA is reserved if OPTPERF = 00b. This had been ratified in ECN 001, but was dropped in the integration that created NVMe 1.4a. • Swapped position of two paragraphs in the description of the NOWS field.
2020-11-16	Changes from 2020-11-12 Technical WG meeting: • Added comments explaining some changes. • 8.25: Added sentence to resolve conflicting values in NPDG and NPDGD fields. • 8.25: Added question about adding Copy command to list of commands which write.
2020-12-17	Changes from 2020-12-17 Technical WG meeting: Section 8.25: • Added Copy command to list of commands in paragraph 2. • Clarified that purpose of the recommendations is optimal performance. • Replace “stream write size” with “stream write length”.
2021-02-26	Revision for Member Review. • Corrected “NVM Sets Attribute Entry” to “NVM Set Attribute Entry” in new text. • 8.25.1: Deleted changes to the last two paragraphs. These paragraphs were not affected by the addition of the NPDGD field. • Removed comments and accepted changes.
2021-03-04	Revision for Member Review. Mike Allison pre-review changes: • 5.15.2.1 Figure 249: Modified OPTPERF description to use a table and other clarifications. • 8.25.1: Clarifications. Changes from 2021-03-04 Technical WG meeting: • 5.15.2.1 Figure 249: Reworded OPTPERF field description and added reference to model section.
2021-06-02	Rebased on NVM Command Set Specification 1.0 and ECN 001 5.8.2: Incorporated late Member Review comment from Intel (NPWA changed to NPWG).
2021-06-14	Integrated into the NVM Command Set Specification, revision 1.0.
2021-06-21	Corrected editorial issues found in review of integrated revision: • 4.1.5.1: Rewrote table describing the OPTPERF field. • 4.1.5.1: Renamed NPDGD field to NPDGL. • 5.8.2: Added sentence to highlight that NPDGL field is larger than NPDG field. • 5.8.2: Clarified description of different values indicated by NPDG and NPDGL.
2021-06-22	• Updated date to avoid confusion with earlier revisions. • Deleted comments. • Accepted format changes. • Accepted insertion of UIDREUSE description.
2021-06-23	• Integrated into the NVM Command Set Specification, revision 1.0.
2021-06-24	• Accepted all changes • Removed all comments, • Converted all references/cross-references to text

Description for Specification Changes Document

Enhanced Namespace Granularity (optional)

- Adds the Namespace Preferred Deallocate Granularity Large (NPDGL) field to the NVM Command Set specific Identify Namespace data structure (CNS 05h, CSI 00h). This is a 32-bit version of the existing NPDG field in the Identify Namespace data structure (CNS 00h). The width of the new field matches the width of the Length in Logical Blocks field in the range definition of the Dataset Management command.
- Expands OPTPERF from one to two bits, to simplify describing requirements for supporting the legacy (16-bit) and new (32-bit) field.
- Capitalizes the Length in Logical Blocks field in the range definition.
- Modifies section 5.8.2 to:
    - Define the usage of the new NPDGL field.
    - Clarify the SGS field, which is a multiplier for the SWG field.
    - Clarify other statements.
- References:
    - NVM Command Set Specification 1.0:
        - Section 3.2.3
        - Section 4.1.5
        - Section 5.8.2

Description of Specification Changes

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 Bracketed>:	Notes to editor

Modify portions of the NVM Command Set Specification 1.0 as shown below:

3 I/O Commands for the NVM Command Set

…

3.2 NVM Command Set Commands

…

3.2.3 Dataset Management Command

…

The data that the Dataset Management command provides is a list of ranges with context attributes. Each range consists of a starting LBA, a length of logical blocks that the range consists of and the context attributes to be applied to that range. The ~~length in logical blocks~~ Length in Logical Blocks field is a 1-based value. The definition of the Dataset Management command Range field is specified in Figure 41. The maximum case of 256 ranges is shown.

Figure 41: Dataset Management – Range Definition

Range	Bytes	Field
Range 0	03:00	Context Attributes
	07:04	Length in ~~logical blocks~~ Logical Blocks
	15:08	Starting LBA
Range 1	19:16	Context Attributes
	23:20	Length in ~~logical blocks~~ Logical Blocks
	31:24	Starting LBA
	…	
Range 255	4083:4080	Context Attributes
	4087:4084	Length in ~~logical blocks~~ Logical Blocks
	4095:4088	Starting LBA

…

4 Admin Commands for the NVM Command Set

4.1 Admin Command behavior for the NVM Command Set

…

4.1.5 Identify command

…

4.1.5.1 Identify Namespace data structure (CNS 00h)

…

Figure 97: Identify – Identify Namespace Data Structure, NVM Command Set Specific

<table>
<tr><th>Bytes</th><th>O/M
1</th><th>Description</th></tr>
<tr><td>…</td><td></td><td></td></tr>
<tr><td>24</td><td>M</td><td>**Namespace Features (NSFEAT):** This field defines features of the namespace.

Bits 7:~~5~~6 are reserved.

Bits 5:4 (**OPTPERF**) indicate support of alignment and granularity attributes of this namespace, as described in the following table:

<table>
<tr><th>Value</th><th colspan="6">Field Supported</th></tr>
<tr><th></th><th>NPWG</th><th>NPWA</th><th>NPDG</th><th>NPDGL</th><th>NPDA</th><th>NOWS</th></tr>
<tr><td>00b</td><td>No</td><td>No</td><td>No</td><td>No</td><td>No</td><td>No</td></tr>
<tr><td>01b</td><td>Yes</td><td>Yes</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td></tr>
<tr><td>10b</td><td>Yes</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td><td>Yes</td></tr>
<tr><td>11b</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td></tr>
</table>

The use of these fields by the host for I/O optimization is described in section 5.8.2.

~~Bit 4 (**OPTPERF**) if set to '1' indicates that the fields NPWG, NPWA, NPDG, NPDA, and NOWS are defined for this namespace and should be used by the host for I/O optimization (refer to the NVM Set List section in the NVMe Base Specification). If cleared to '0', then the controller does not support the fields NPWG, NPWA, NPDG, NPDA, and NOWS for this namespace.~~

Bit 3 (**UIDREUSE**): This bit is as defined in the UIDREUSE bit in the I/O Command Set Independent Identify Namespace data structure (refer to the I/O Command Set Independent Identify Namespace data structure section in the NVMe Base Specification).

…</td></tr>
<tr><td>…</td><td></td><td></td></tr>
<tr><td>65:64</td><td>O</td><td>**Namespace Preferred Write Granularity (NPWG):** This field indicates the smallest recommended write granularity in logical blocks for this namespace. This is a 0's based value. ~~If the OPTPERF bit is cleared to '0'~~ If this field is not supported as defined by the OPTPERF field, then this field is reserved.

The size indicated by this field should be less than or equal to the size indicated by the Maximum Data Transfer Size (MDTS) field (refer to the NVMe Base Specification), which ~~that~~ is specified in units of minimum memory page size. The value of this field may change if the namespace is reformatted. The size should be a multiple of the Namespace Preferred Write Alignment (NPWA) field.

Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.</td></tr>
<tr><td>67:66</td><td>O</td><td>**Namespace Preferred Write Alignment (NPWA):** This field indicates the recommended write alignment in logical blocks for this namespace. This is a 0's based value. ~~If the OPTPERF bit is cleared to '0'.~~ If this field is not supported as defined by the OPTPERF field, then this field is reserved.

The value of this field may change if the namespace is reformatted.

Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.</td></tr>
<tr><td>69:68</td><td>O</td><td>**Namespace Preferred Deallocate Granularity (NPDG):** This field indicates the recommended granularity in logical blocks for the Dataset Management command with the Attribute – Deallocate bit set to '1' in Dword 11. This is a 0's based value. ~~If the OPTPERF bit is cleared to '0'~~ If this field is not supported as defined by the OPTPERF field, then this field is reserved.

The value of this field may change if the namespace is reformatted. The size should be a multiple of the Namespace Preferred Deallocate Alignment (NPDA) field.

Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.</td></tr>
</table>

Figure 97: Identify – Identify Namespace Data Structure, NVM Command Set Specific

Bytes	O/M 1	Description
71:70	O	**Namespace Preferred Deallocate Alignment (NPDA):** This field indicates the recommended alignment in logical blocks for the Dataset Management command with the Attribute – Deallocate bit set to ‘1’ in Dword 11. This is a 0’s based value. ~~If the OPTPERF bit is cleared to ‘0’~~ If this field is not supported as defined by the OPTPERF field, then this field is reserved. The value of this field may change if the namespace is reformatted. Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.
73:72	O	**Namespace Optimal Write Size (NOWS):** This field indicates the size in logical blocks for optimal write performance for this namespace. This is a 0’s based value.~~ If the OPTPERF bit is cleared to ‘0’~~ If this field is not supported as defined by the OPTPERF field, then this field is reserved. If this namespace is associated with an NVM Set and: a) this field is supported as defined by the OPTPERF field, then this field shall equal the value indicated by the Optimal Write Size field in the NVM Set Attributes Entry (refer to the Namespace Identification Descriptor in the NVMe Base Specification) for that NVM Set; or b) this field is not supported as defined by the OPTPERF field, then the host should use the Optimal Write Size field in the NVM Set Attributes Entry for that NVM Set for I/O optimization (refer to section 5.8.2). The size indicated should be less than or equal to Maximum Data Transfer Size (MDTS) that is specified in units of minimum memory page size. The value of this field may change if the namespace is reformatted. The value of this field should be a multiple of the Namespace Preferred Write Granularity (NPWG) field. ~~If the namespace is associated with an NVM set, NOWS defined for this namespace shall be set to the Optimal Write Size field setting defined in NVM Set Attributes Entry (refer to the Namespace Identification Descriptor in the NVMe Base Specification) for the NVM Set with which this namespace is associated. If NOWS is not supported, the Optimal Write Size field in NVM Sets Attributes Entry (refer to the Namespace Identification Descriptor in the NVMe Base Specification) for the NVM Set with which this namespace is associated should be used by the host for I/O optimization.~~ Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.
…		

…

4.1.5.3 I/O Command Set Specific Identify Namespace Data Structure (CNS 00h)

…

Figure 100: NVM Command Set I/O Command Set Specific Identify Namespace Data Structure (CSI 00h)

Bytes	O/M [1]	Description
271:268	O	**Namespace Preferred Deallocate Granularity Large (NPDGL):** This field indicates the recommended granularity in logical blocks for the Dataset Management command with the Attribute – Deallocate bit set to '1' in Command Dword 11. If this field is not supported as defined by the OPTPERF field (refer to Figure 97), then this field is reserved. If this field is cleared to 0h, then this field does not indicate a recommended granularity. The value of this field may change if the namespace is reformatted. The size should be a multiple of the Namespace Preferred Deallocate Alignment (NPDA) field (refer to Figure 97). Refer to section 5.8.2 for how this field is utilized to improve performance and endurance.
4095:~~268~~272	O	Reserved
NOTES: 1. O/M definition: O = Optional, M = Mandatory.		

< **Note to Editor**: The NPDGL field must be 4-byte aligned. >

…

5 Extended Capabilities

…

5.8 Command Set Specific Capability

…

5.8.2 Improving Performance through I/O Size and Alignment Adherence

NVMe controllers may require constrained I/O sizes and alignments to achieve the full performance potential. There are a number of optional attributes that the controller uses to indicate these recommendations. If hosts do not follow these constraints, then the controller shall function correctly, but performance may be limited.

~~Each Copy, Write, Write Uncorrectable, or Write Zeroes commands should address a multiple of Namespace Preferred Write Granularity (NPWG) (refer to Figure 97) and Stream Write Size (SWS) (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification) logical blocks (as expressed in the NLB field), and the SLBA field of the command should be aligned to Namespace Preferred Write Alignment (NPWA) (refer to Figure 97) for best performance. Each range in a Dataset Management command with the Attribute - Deallocate (AD) bit set to '1' should contain a multiple of Namespace Preferred Deallocate Granularity (NPDG) (refer to Figure 97) logical blocks and the start of each range should be aligned to Namespace Preferred Deallocate Alignment (NPDA) (refer to Figure 97) and Stream Granularity Size (SGS) (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification) logical blocks.~~

For best performance, the host should issue Copy, Write, Write Uncorrectable, and Write Zeroes commands that specify:

a) a number of logical blocks that is:

    a. a multiple of the Namespace Preferred Write Granularity (NPWG) field (refer to Figure 97), if the NPWG field is defined; and
    b. a multiple of the number of logical blocks indicated by the Stream Write Size (SWS) field (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification), if the Streams Directive is enabled;

and

b) a Starting LBA (SLBA) field that is aligned to the Namespace Preferred Write Alignment (NPWA) field (refer to Figure 97), if the NPWA field is defined.

Resolving conflicts between namespace attributes and Streams attributes is described in section 5.8.2.1.

The namespace preferred deallocate granularity is a number of logical blocks that is indicated by both the NPDG field (refer to Figure 97) and the NPDGL field. The NPDGL field is able to represent larger values than the NPDG field (refer to Figure 100). Support for these fields is indicated by the OPTPERF field (refer to Figure 97). If the NPDG field and the NPDGL field are both supported and indicate different values of namespace preferred deallocate granularity, then the host should use the value indicated by the NPDGL field.

The namespace preferred deallocate alignment is a number of logical blocks that is indicated by the NPDA field (refer to Figure 97).

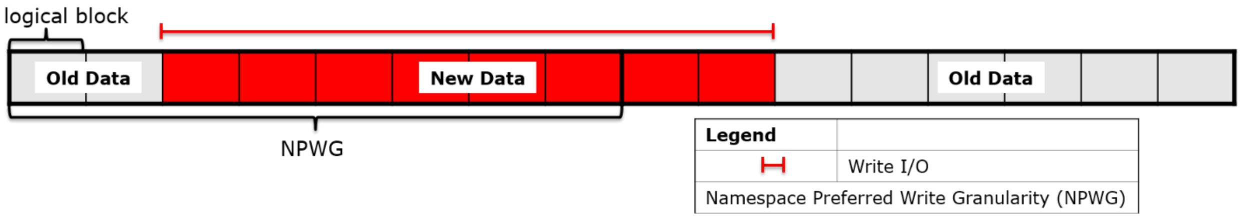
For best performance, the host should issue Dataset Management commands with the Attribute – Deallocate (AD) bit set to ‘1’ that specify:

a) a Length in Logical Blocks field that is a multiple of the namespace preferred deallocate granularity, if the namespace preferred deallocate granularity is defined; and
b) a Starting LBA field that is
    a. a multiple of the namespace preferred deallocate alignment, if the namespace preferred deallocate alignment is defined; and
    b. a multiple of the Stream Granularity Size (SGS) field (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification) if the Streams Directive is enabled.

5.8.2.1 Improved I/O examples (Informative ~~non-normative~~)

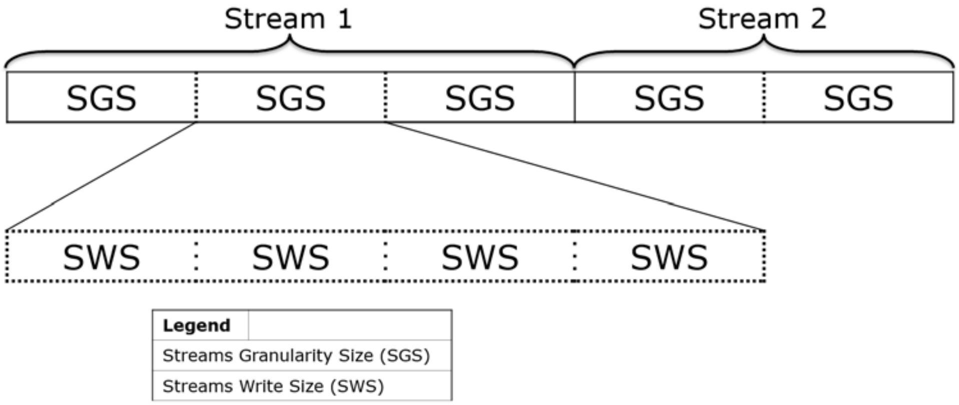
…

Figure 146: Host write I/O command following NPWG but not NPWA attributes



The namespace preferred deallocate granularity and the namespace preferred deallocate alignment ~~NPDG and NPDA (refer to Figure 97)~~ (refer to section 5.8.2) are attributes of ~~constructs in~~ the namespace that are intended to improve performance for Dataset Management deallocate operations within a namespace. The namespace preferred deallocate granularity and the namespace preferred deallocate alignment ~~NPDG and NPDA~~ may be impacted by multiple factors including but not limited to the boundaries described in Figure 143, device hardware limits, or non-volatile storage erase block sizes. Deallocating at multiples of the namespace preferred deallocate granularity ~~NPDG size~~ and aligned to the namespace preferred deallocate alignment ~~NPDA~~ (i.e., (Starting LBA modulo namespace preferred deallocate alignment ~~NPDA~~) == 0) may enable improved deallocate performance for the namespace.

Figure 147: Two streams composed of SGS and SWS



Streams (refer to the Streams Directive section in the NVMe Base Specification) may or may not be utilized with different namespace attributes.

The stream granularity length is a number of logical blocks that is the product of the Stream Granularity Size (SGS) field and the number of logical blocks indicated by the Stream Write Size (SWS) field (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification). Figure 147 shows an example of the relationship between these attributes in which the SGS field is set to 4h. ~~Figure 147 shows the streams attributes of Stream Granularity Size (SGS) and Stream Write Size (SWS) (refer to the Streams Directive – Return Parameters Data Structure figure in the NVMe Base Specification).~~ The first stream is ~~constructed by the host to be~~ composed of three SGS units, and each SGS unit in this example is equal to four SWS units. ~~The host streams are~~ A stream is optimized for performance of the Dataset Management command deallocate operation~~s~~ if write I/O lengths are integer multiples of the stream granularity length ~~by extending the stream in units of SGS~~. ~~The streams receive optimal host~~ A stream is optimized for write performance if write I/O ~~command~~ lengths are integer multiples of the SWS field.

Streams are sometimes handled by separate I/O paths in the device. This may ~~entail such things as~~ include different device hardware, media mapping, or reliability protections. The number of logical blocks indicated by the SWS field should be a multiple of the number of logical blocks indicated by the NPWG field. The size indicated by the SGS field and ~~NPDG~~ the namespace preferred deallocate granularity may be ~~equivalent~~ equal to each other or multiples of each other. ~~A namespace utilizing~~ If a namespace indicates integer multiple size relationships between the streams attributes (the SWS field and the SGS field) and the namespace attributes (the NPWG field and ~~NPDG~~ the namespace preferred deallocate granularity), then a write operation or a deallocate operation may obtain optimal performance by specifying a number of logical blocks that is equal to the largest of those attributes ~~may provide optimal performance by adhering to the largest attribute for write I/O commands or deallocations~~.

Not all namespaces indicate ~~describe both~~ their Streams attributes and namespace attributes in multiples as described above. The recommended order of priority for a host to ~~adhere to conflicting~~ resolve conflicts between namespace attributes and Streams attributes is to issue write operations that conform to the SGS field and the SWS field if ~~while utilizing~~ the Streams Directive ~~directives~~ is used. ~~When not utilizing~~ If the Streams Directive ~~directives~~ is not used, ~~the~~ then issuing write operations that conform to the namespace attributes ~~for each namespace~~ should provide improved performance.

If the Streams Directive is enabled on a namespace, and a ~~deallocate operations~~ Dataspace Management command specifying a deallocate operation specifies a range of logical blocks that are associated with a stream, then ~~the host should use~~ that range should conform to the SGS based alignment and size preferences ~~in favor of the Namespace and NVM Set preferences~~. If the Streams Directive is not enabled on a namespace, or if the logical blocks specified by a range are not associated with a stream, then ~~the host should construct deallocate operations that~~ that range should conform to the namespace preferred deallocate granularity and the namespace preferred deallocate alignment ~~NPDG and NPDA~~.

< end of changes to NVM Command Set Specification 1.0 >