LEGAL NOTICE:

This NVM Express revision 1.3 technical proposal 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 NVM Express revision 1.3 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 - 2019 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 Group
3855 SW 153rd Drive
Beaverton, OR 97003 USA
info@nvmexpress.org

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**4025**
Change Date	**2019-03-04**
Builds on Specification	**NVM Express 1.3c** **TP 4018a NVM Sets and Read Recovery Level**
Referenced Ratified Technical Proposals	**TP 4030 Verify Command**

Technical Proposal Author(s)

Name	**Company**
Andrzej Jakowski, Keith Busch, Jonathan Hughes, Raymond Robles, Michael Allison	Intel
Martin K. Petersen	Oracle
Mark Carlson	Toshiba
Yoni Shternhell, Daniel Helmick, Christoph Hellwig	WDC
Ross Stenfort	Facebook

This technical proposal defines performance and endurance hints to indicate to the host optimal granularity and alignment for IO.

Revision History

Revision Date	**Change Description**
11 July 2018	Initial version
25 July 2018	New Optimal I/O section added
15 Aug 2018	Fixed “field” spelling, renamed new attributes to include “Optimal” in name, added abbreviation definitions to figures
17 Aug 2018	Addressed comments from Technical Meeting, and added information on how the host should choose to construct an optimal write command in light of SWS, NVM Sets Optimal Write Size, and NOWG/NOWA.
30 Aug 2018	Split NODGA into NODG and NODA
06 Sep 2018	Minor editorial changes
21 Sep 2018	Renamed “optimal” to “preferred” in the four new attributes, language clean up
01 Nov 2018	Added Namespace Optimal Write Size to supplant NVM Sets Optimal Write Size
05 Dec 2018	Indicated in NSFEAT that using the new hints is a recommendation. Minor editorial changes. Added a change to TP 4018a to reflect that Optimal Write Size of 0h means that no Optimal Write size is specified that included a recommendation to set it to 0h when LBA formats of namespaces within an NVM Set prohibit defining an Optimal Write Size.

16 Jan 2019	Adjusted for 2019. Moved NABSN and NABO into the Identify Namespace data structure as defined by NVMe 1.3c. Deleted the changes to Identify Controller data structure as not changes are now requested. Additional editorial changes.
22 Jan 2019	Updated pictures. Adjusts for plural statements (editorial).
25 Feb 2019	Integration
04 Mar 2019	Ratified

Discussion

This Technical Proposal addresses following issues:

1. The existing NVMe 1.3 specification enables host to discover LBA size for the namespace. While LBA size defines minimum addressable unit on NVM subsystem, host writes in LBA size or alignment may not necessarily be optimal in terms of performance and endurance. Often NVM subsystems prefer writes in larger granularities and properly aligned to avoid Read Modify Write (RMW) on media. Currently there is no way for host system to discover this write granularity and alignment not causing performance and endurance penalties in the NVM subsystem;
2. The existing NVMe 1.3 specification does not allow host to discover performance and endurance hints pertaining to Deallocate operation. Analogous hints are already standardized in other specifications such as SCSI SBC and may be beneficial for use in NVMe specification as well; and
3. TP 4018a defines Optimal Write Size in units of bytes which has coherency issues in the case that multiple namespaces with different logical block sizes are allocated from the same NVM Set. This makes it impossible for the controller to indicate to the host the optimal write size for certain use cases.

Description for NVMe 1.4 Changes Document

1. Performance and Endurance hints reporting:
   - Identify Namespace data structure extended to include:
     - NSFEAT bit to indicate support for the following new attributes;
     - Namespace Preferred Write Granularity (NPWG);
     - Namespace Preferred Write Alignment (NPWA);
     - Namespace Preferred Deallocate Granularity (NPDG);
     - Namespace Preferred Deallocate Alignment (NPDA); and
     - Namespace Optimal Write Size (NOWS);

   and
2. Add new Improving Performance through I/O Size and Alignment Adherence section and references to it.

Description of Specification Changes

Modify Portions of Figure 109 (Identify – Identify Namespace Data Structure, NVM Command Set Specific) as shown below:

Bytes	O/M[1]	Description
...	...	...

Bytes	O/M[1]	Description
24	M	**Namespace Features (NSFEAT):** This field defines features of the namespace. Bits 7:45 are reserved. Bit 4 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; and • NOWS defined for this namespace shall adhere to Optimal Write Size field setting defined in NVM Sets Attributes Entry (refer to <TP 4018a> Figure Fig5_15TBD1) for the NVM Set with which this namespace is associated. If cleared to '0', then: • the controller does not support the fields NPWG, NPWA, NPDG, NPDA, and NOWS for this namespace; and • Optimal Write Size field in NVM Sets Attributes Entry (refer to <TP 4018a> Figure Fig5_15TBD1) for the NVM Set with which this namespace is associated should be used by the host for I/O optimization. Bit 3 if set to '1' indicates that the non-zero NGUID and non-zero EUI64 fields for this namespace are never reused by the controller. If cleared to '0', then the NGUID and EUI64 values may be reused by the controller for a new namespace created after this namespace is deleted. This bit shall be cleared to '0' if both NGUID and EUI64 fields are cleared to 0h. Refer to section 7.11. Bit 2 if set to '1' indicates that the controller supports the Deallocated or Unwritten Logical Block error for this namespace. If cleared to '0', then the controller does not support the Deallocated or Unwritten Logical Block error for this namespace. Refer to section 6.7.11. Bit 1 if set to '1' indicates that the fields NAWUN, NAWUPF, and NACWU are defined for this namespace and should be used by the host for this namespace instead of the AWUN, AWUPF, and ACWU fields in the Identify Controller data structure. If cleared to '0', then the controller does not support the fields NAWUN, NAWUPF, and NACWU for this namespace. In this case, the host should use the AWUN, AWUPF, and ACWU fields defined in the Identify Controller data structure in Figure 111. Refer to section 6.4. Bit 0 if set to '1' indicates that the namespace supports thin provisioning. Specifically, the Namespace Capacity reported may be less than the Namespace Size. When this feature is supported and the Dataset Management command is supported, then deallocating LBAs shall be reflected in the Namespace Utilization field. Bit 0 if cleared to '0' indicates that thin provisioning is not supported and the Namespace Size and Namespace Capacity fields report the same value.
…	…	**…**
41:40	O	**Namespace Atomic Boundary Size Normal (NABSN):** This field indicates the atomic boundary size for this namespace for the NAWUN value. This field is specified in logical blocks. Writes to this namespace that cross atomic boundaries are not guaranteed to be atomic to the NVM with respect to other read or write commands. A value of 0h indicates that there are no atomic boundaries for normal write operations. All other values specify a size in terms of logical blocks using the same encoding as the AWUN field. Refer to section <Editor's note: Atomic Operations>. Refer to section 8.NEW for how this field is utilized.
43:42	O	**Namespace Atomic Boundary Offset (NABO):** This field indicates the LBA on this namespace where the first atomic boundary starts. If the NABSN and NABSPF fields are cleared to 0h, then the NABO field shall be cleared to 0h. NABO shall be less than or equal to NABSN and NABSPF. Refer to section <Editor's note: Atomic Operations>. Refer to section 8.NEW for how this field is utilized.
…	…	**…**

Bytes	O/M[1]	Description
47:46	O	**Namespace Optimal I/O Boundary (NOIOB):** This field indicates the optimal I/O boundary for this namespace. This field is specified in logical blocks. The host should construct Read and Write commands that do not cross the I/O boundary to achieve optimal performance. A value of 0h indicates that no optimal I/O boundary is reported. Refer to section 8.NEW for how this field is utilized to improve performance and endurance.
...		
65:64	O	**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. 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 size should be a multiple of Namespace Preferred Write Alignment (NPWA). Refer to section 8.NEW for how this field is utilized to improve performance and endurance.
67:66	O	**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. The value of this field may change if the namespace is reformatted. Refer to section 8.NEW for how this field is utilized to improve performance and endurance.
69:68	O	**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. The value of this field may change if the namespace is reformatted. The size should be a multiple of Namespace Preferred Deallocate Alignment (NPDA). Refer to section 8.NEW for how this field is utilized to improve performance and endurance.
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. The value of this field may change if the namespace is reformatted. Refer to section 8.NEW 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. 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 Namespace Preferred Write Granularity (NPWG). Refer to section 8.NEW for how this field is utilized to improve performance and endurance.
103:74		Reserved
…		

Insert the new section as shown:

8.NEW 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

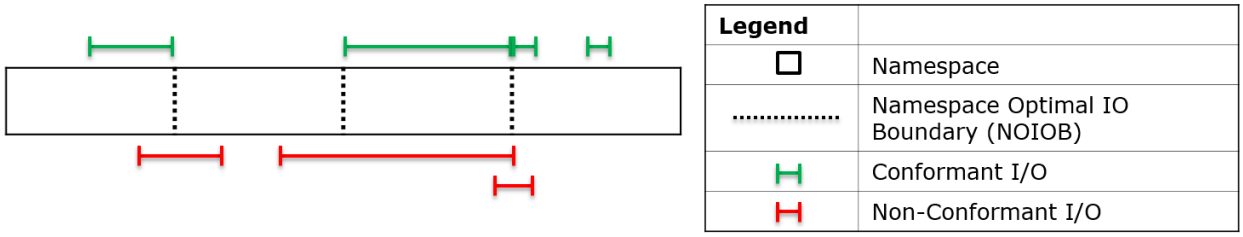
recommendations. If hosts do not follow these constraints, then the controller shall function correctly, but performance may be limited.

Each Write, Write Uncorrectable, or Write Zeroes command should address a multiple of Namespace Preferred Write Granularity (NPWG) (refer to Figure 109) and Stream Write Size (SWS) (refer to Figure 293) 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 109) 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 109) logical blocks and the start of each range should be aligned to Namespace Preferred Deallocate Alignment (NPDA) (refer to Figure 109) and Stream Granularity Size (SGS) logical blocks.

8.NEW.1 Improved I/O examples (non-normative)

It is recommended that the host utilize the I/O attributes as reported by the controller to receive optimal performance from the NVM subsystem. This section summarizes performance related attributes from namespaces, streams, NVM Sets and the NVM command set. The I/O commands discussed throughout this section include those that interact with non-volatile storage in either a Read, Compare, Verify, Write, Write Uncorrectable, Write Zeroes operation, or Dataset Management operation with the Attribute - Deallocate bit set to '1'. The I/O command properties of length and alignment are discussed throughout this section.

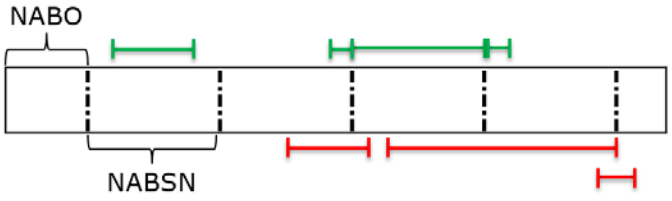
Figure TBD1 An example namespace with four NOIOBs



Legend	
	Namespace
.....	Namespace Optimal IO Boundary (NOIOB)
	Conformant I/O
	Non-Conformant I/O

In Figure TBD1 an example namespace is diagrammed with three Namespace I/O Boundaries (NOIOB) (refer to Figure 109). The NOIOB attribute should be applied to Read, Compare, Verify, Write, Write Uncorrectable, and Write Zeroes I/O commands. The four green lines are example I/O commands from the host that adhere to the recommendations of NOIOB settings for this namespace. None of the four I/O commands shown in green on the top of Figure TBD1 cross an NOIOB. The three I/O commands shown in red on the bottom of Figure TBD1 violate the recommendations for improved performance. The longest I/O command shown in red crosses one NOIOB and ends aligned with a different NOIOB. The remaining two I/O commands shown in red also cross an NOIOB. All three of these example I/O commands shown in red could be split into two I/O commands that adhere to the recommendations provided by the namespace for NOIOB.

Figure TBD2: Example namespace illustrating a potential NABO and NABSN

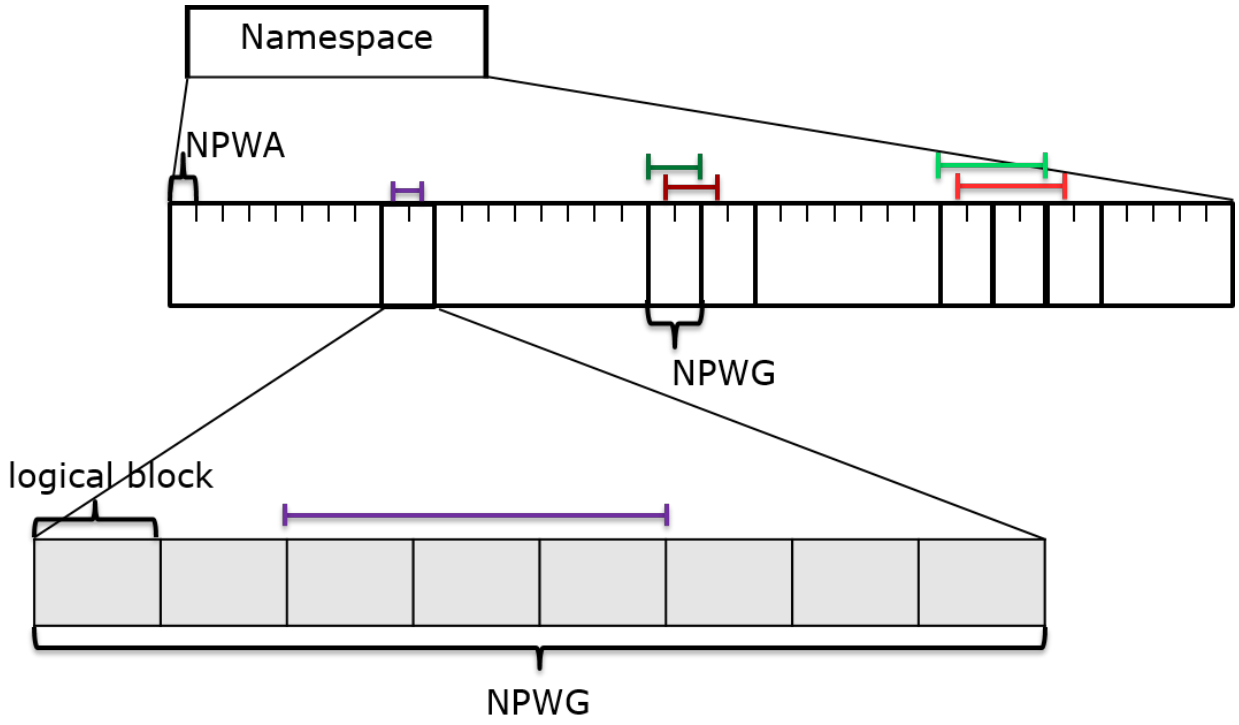


Legend	
	Namespace
- - - -	Namespace Atomic Boundary Size Normal (NABSN)
	Conformant I/O
	Non-Conformant I/O
Namespace Atomic Boundary Offset (NABO)	

Continuing with the same namespace example from Figure TBD1, an illustration of Namespace Atomic Boundary Offset (NABO) (refer to Figure 109) and Namespace Atomic Boundary Size Normal (NABSN) (refer to Figure 109) is shown in Figure TBD2. NABSN and NABO attributes apply to Write, Write Uncorrectable, and Write Zeroes commands. NABSN and NOIOB may not be related to each other, and there may be an offset of NABO to locate the first NABSN starting logical block. The NOIOBs are not shown in Figure TBD2. The I/O commands shown in green on the top of Figure TBD2 illustrate I/O commands that adhere to the namespace's guidance for optimal performance. The I/O commands shown in red on the bottom illustrate I/O commands that do not follow the optimal performance guidelines.

The I/O command examples shown in red in Figure TBD1 and Figure TBD2 both illustrate commands that could be restructured to conform to the namespace attributes for Optimal I/O relative to NOIOB, NABO, and NABSN. Each of these example I/O commands shown in red in Figure TBD1 and Figure TBD2 could be split into two different I/O commands that adhere to the recommendations. While this increases the number of commands sent to the controller, it is expected that adherence to the boundary recommendations will improve the performance for the controller. Avoiding host traffic that demands non-optimal I/O commands is the most recommendable solution for a host.

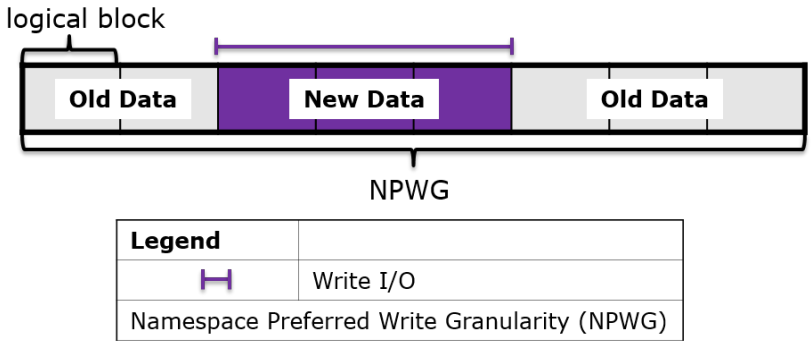
Figure TBD3 Example namespace broken down to illustrate potential NPWA and NPWG settings



Legend	
	Conformant I/O
	Non-Conformant I/O
Namespace Preferred Write Alignment (NPWA)	
Namespace Preferred Write Granularity (NPWG)	

NPWG and NPWA are namespace internal constructs, and they are illustrated in Figure TBD3. The box at the top of Figure TBD3 is the namespace. The series of boxes in the middle layer indicate many namespace optimal write units described by NPWA (refer to Figure 109) and NPWG (refer to Figure 109), and the bottom layer is a series of eight logical blocks that in aggregate form the NPWG for this example. Sometimes NPWG are useful because several sequential logical blocks (refer to Figure 109) may be placed and tracked together on the media, or the NPWG might relate NVM subsystem data reliability implementation constraints. NPWG and NPWA attributes apply to Write, Write Uncorrectable, and Write Zeroes commands.

Figure TBD4: Non-conformant Write Impact



Shown in Figure TBD4 is an I/O command that covers three of eight logical blocks within an NPWG. In this example namespace, NPWG is set to eight logical blocks, and the write of only three logical blocks requires a read of the preceding two logical blocks and trailing three logical blocks. The host write that completes to the non-volatile storage would consist of five logical blocks of older data and three new logical blocks with the data provided by the write I/O command. The resulting read-modify-write may have non-optimal performance in comparison to a host write adhering to the NPWG attribute due to the extra read operation executed internally in the NVM subsystem. Aligning the beginning of the write I/O command with the NPWA attribute would remove the need to read the preceding existing data. Host writes with a length that is a multiple of NPWG would remove the need for reading the trailing data.

Following the NPWG recommendation alone is insufficient for optimal performance. If a write I/O command is an integer multiple of NPWG but it is offset in alignment from the recommended NPWA, a read-modify-write may occur on the logical blocks at the beginning and ending of the command. The I/O commands shown in red in Figure TBD3 are integer multiples of NPWG, but their alignment is triggering a read-modify-write at both the beginning and ending of the write I/O command. The write I/O commands shown in green adhere to the alignment and granularity requirements of the NPWA and NPWG. Figure TBD5 illustrates the shorter dark green write I/O command that adheres to both NPWG and NPWA attributes. This dark green write I/O command has a length equaling the NPWG attribute which adheres to the NPWG attribute recommendations. Figure TBD6 illustrates the dark red write I/O command that follows the NPWG attribute with a length of one NPWG, but it does not adhere to the NPWA attribute recommendations. The dark red write I/O command requires a read of the old data at the beginning and the ending of the write I/O command to fill both NPWG units illustrated here. Longer write I/O commands that fail to adhere to the NPWA recommendation may trigger a read-modify-write of the leading and trailing NPWG segments inside of the NVM subsystem.

Figure TBD5: Host write I/O command following NPWA and NPWG

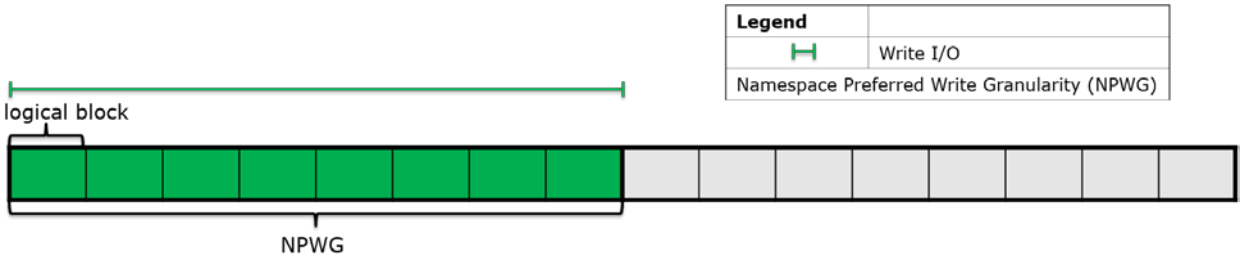
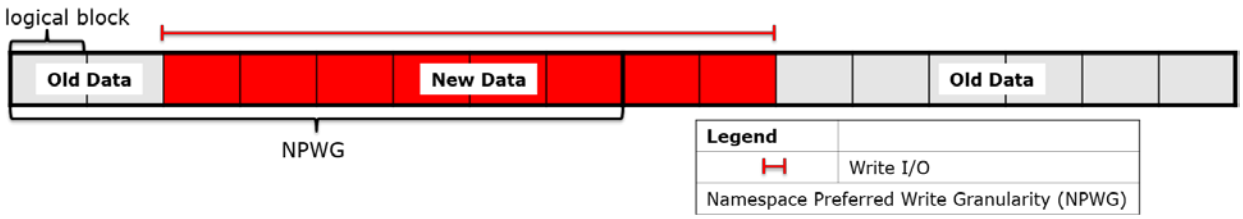
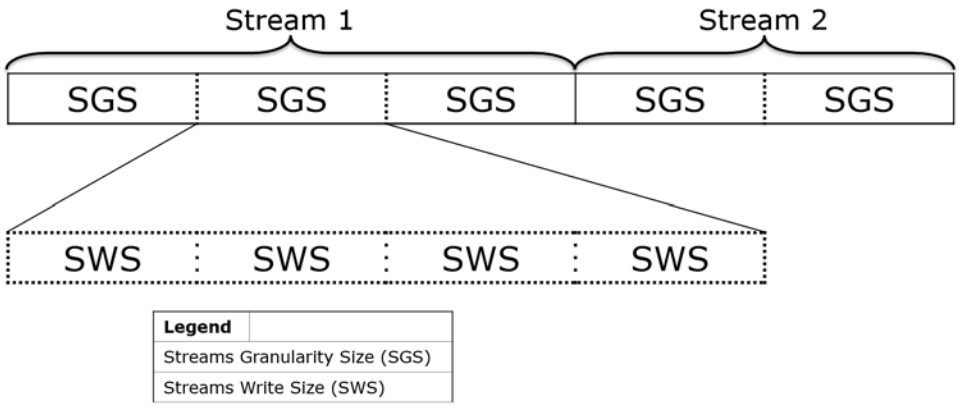


Figure TBD6 Host write I/O command following NPWG but not NPWA attributes



NPDG and NPDA (refer to Figure 109) are constructs in the namespace intended to improve performance for Dataset Management deallocate operations within a namespace. NPDG and NPDA may be impacted by multiple factors including but not limited to the boundaries described in Figure TBD3, device hardware limits, or non-volatile storage erase block sizes. Deallocating at multiples of NPDG size and aligned to NPDA ((Starting LBA modulo NPDA) == 0) may enable improved deallocate performance for the namespace.

Figure TBD7: Two streams composed of SGS and SWS



Streams (refer to section 9.3) may or may not be utilized with different namespace attributes. Figure TBD7 shows the streams attributes of Stream Granularity Size (SGS) (refer to Figure 293) and Stream Write Size (SWS) (refer to Figure 293). 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 optimized for performance of the Dataset Management deallocate operations by extending the stream in units of SGS. The streams receive optimal host write performance if write I/O command lengths are integer multiples of SWS.

Streams are sometimes handled by separate I/O paths in the device. This may entail such things as different device hardware, media mapping, or reliability protections. Generally, SWS may be a multiple of the NPWG,

but it is not required. Furthermore, SGS and NPDG may frequently be equivalent or multiples of each other. A namespace utilizing integer multiple relationships between the streams attributes (SWS and SGS) and the namespace attributes (NPWG and NPDG) may provide optimal performance by adhering to the largest attribute for write I/O commands or deallocations.

Not all namespaces describe both their Streams and namespace attributes in multiples as described above. The recommended order of priority for a host to adhere to conflicting namespace and Streams attributes is to conform to SGS and SWS while utilizing the Streams directives. When not utilizing the Streams directives, the namespace attributes for each namespace should provide improved performance.

If the Streams Directive is enabled on a namespace, and a deallocate operations specifies logical blocks that are associated with a stream, then the host should use 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 the logical blocks are not associated with a stream, then the host should construct deallocate operations that conform to NPDG and NPDA.

Namespace Optimal Write Size (NOWS) (refer to Figure 109) is intended to supplement NVM Sets Optimal Write Size as it provides a mechanism to report the optimal write size that scales to a multiple namespace per NVM Set use case, but also covers the use case where there is a single namespace allocated in an NVM Set. Namespaces should report NOWS as a multiple of NPWG. When constructing write operations, the host should minimally construct writes that meet the recommendations of NPWG and NPWA, but may achieve optimal write performance by constructing writes that meet the recommendation of NOWS.

If NVM Sets are supported, NOWS setting for the namespace shall adhere to NVM Sets Optimal Write Size setting for the NVM Set which the namespace is associated with. If an NVM Set does not specify an Optimal Write Size, the host should consult NOWS setting for the namespace for I/O optimization purposes. Similarly, if NOWS is not defined for a namespace, the host should consult the Optimal Write Size setting for the NVM Set associated with that namespace to achieve optimal performance.

Modify Portions of Figure 293 (Streams Directive– Return Parameters Data Structure) as shown below:

Bytes	Description
…	…
19:16	**Stream Write Size (SWS):** This field indicates the alignment and size of the optimal stream write as a number of logical blocks for the specified namespace. 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. SWS may change if the namespace is reformatted with a different LBA format. If the NSID value is set to FFFFFFFFh, then this field may be cleared to 0h if a single logical block size cannot be indicated. Refer to section 8.NEW for how this field is utilized to optimize performance and endurance.
21:20	**Stream Granularity Size (SGS):** This field indicates the stream granularity size for the specified namespace in Stream Write Size (SWS) units. If the NSID value is set to FFFFFFFFh, then this field may be cleared to 0h. Refer to section 8.NEW for how this field is utilized to optimize performance and endurance.
…	…

Modify Portions of Figure Fig5_15TBD1 (NVM Set Attributes Entry) as defined by TP 4018a NVM Sets and Read Recovery Level as shown below:

15:12	**Optimal Write Size:** This field indicates the size in bytes for optimal write performance. A value of 0h indicates that no Optimal Write Size is specified. This field should be set to 0h when namespaces within an NVM Set have different LBA formats that do not allow an Optimal Write Size to be specified.