LEGAL NOTICE:

This NVM Express revision 1.3c 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 1.3c 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 to 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.

The NVM Express® design mark is a registered trademark of NVM Express, Inc.

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	**4054**
Change Date	**2019-04-15**
Builds on Specification	**NVM Express 1.3c** **TP 4000a Persistent Memory Region**
Ratified Technical Proposals Referenced	**TP 4032 PMR Write Elasticity Status**

Technical Proposal Author(s)

Name	**Company**
Bryan Veal, Michael Allison, Jonathan Hughes	Intel
Peter Onufryk	Microchip

This technical proposal fixes DMA misrouting with the Controller Memory Buffer (CMB) and Persistent Memory Region (PMR).

Revision History

Revision Date	**Change Description**
2018-12-12	Initial version
2018-12-13	Editorial changes and added highlighting of referenced figures. Removed highlighted sections and figure numbering since the TP does not requires these to be filled in prior to ratification.
2019-01-14	Consolidated CAP.CMBS and CMBCAP.CMSS into CAP.CMBS. No reset on FLR for CMBMSC is optional as indicated by CMBCAP.FLRE. Typo fixes and clarity improvements.
2019-01-18	Removed CMBCAP.FLRE. CMBMSC is always sticky across FLR. Added recommendation for configuring legacy CMB for use in a VM. Added recommendation for pre-configuring CMBMSC so that legacy software can use CMB in a VM.
2019-02-11	Removed comments no longer applied to start Phase 3.
2019-02-11a	Further removed comments no longer applied. Clarified the meaning of CMBSTS.CBAI/PMRSTS.CBAI and that CMBMSC.CMSE/PMRMSC.CMSE are not dependent on these bits.
2019-02-12	Closed and removed all remaining comments.
2019-02-27	Early Integration
2019-03-26	Applying comments to final version. Changed “not effect” to “no effect”. Changes referecne to CMB in PMR section.
2019-04-01	Correct the PMRMSC deifnition to clearly state differentiation betweem PMR and CMB address space.
2019-04-15	Ratified

Description for NVMe 1.3c Changes Document

- In some environments, such as when the controller is directly assigned to a virtualized guest, the address space used for host memory and DMA may differ from the address space used to program PCI BARs. This can cause memory read and write requests to be inadvertently routed to the Controller Memory Buffer (CMB) or the Persistent Memory Region (PMR), or vice versa. CMB (section 4.7) and PMR (section 4.8 – TP 4000a) are modified to remove the dependency on PCI BARs for controller routing to CMB or PMR. Instead, host software is provided a mechanisms to set the individual controller base addresses for CMB and PMR.

Description of Specification Changes

CAP.CMBS was added to indicate support for CMB. CMBLOC and CMBSZ are modified to be cleared to 0h until they are enabled. CMBSTS was added to indicate the validity of the CMB's controller base address (CMBSTS.CBAI). CMBMSC was added to support explicit enabling of CMBLOC and CMBSZ (CMBMSC.CRE), explicit enabling of CMB's controller memory space (CMBMSC.CMSE), and setting the base address of CMB's controller memory space (CMBMSC.CBA).

PMRCAP.CMSS was added to indicate whether PMR supports the host-enabled controller memory space. PMRCAP.CBAI was added to indicate the validity of PMR's controller base address. PMRMSC was added to support explicit enabling of PMR's controller memory space (PMRMSC.CMSE) and setting the base address of PMR's controller memory space (PMRMSC.CBA).

Markup Conventions:

Black:	Unchanged (however, hot links are removed)
~~Red Strikethrough~~:	Deleted
Red:	New
Red Highlighted:	TBD values, anchors, and links to be inserted.
<Green Bracketed>:	Notes to editor

<Summary of TBD values:

- 50h Byte offset of CMBMSC
- E14h Byte offset of PMRMSC
- 57 Bit offset of CAP.CMBS
- 16 Subsection number for CMBMSC
- TBD5 Figure number for CMBMSC
- TBDX Subsection number for PMRCAP from TP 4000a
- 24 Bit offset for PMRCAP.CMSS
- 12 Bit offset for PMRSTS.CBAI
- TBD8 Subsection number for PMRMSC
- TBD9 Figure number for PMRMSC
- TBDZ Section number for Persistent Memory Region TP 4000a

>

Modify portions of section 3.1 (Register Definition) as shown below:

Start	End	Symbol	Description
50h	57h	CMBMSC	Controller Memory Buffer Memory Space Control (Optional)
58h	5Bh	CMBSTS	Controller Memory Buffer Status (Optional)

Start	End	Symbol	Description
5Ch~~50h~~	DFFh	Reserved	Reserved
…	…	…	…
E14h	E1Bh	PMRMSC	Persistent Memory Region Controller Memory Space Control (Optional)
TBD1+8h ~~E14h~~	FFFh	Reserved	Command Set Specific

<Editorial Note: CMBMSC and PMRMSC are is required to be qword aligned. This could cause the reserved regions to be split to account for the alignment during integration.>

Modify portions of section 3.1.1 (Offset 00h: CAP – Controller Capabilities) as shown below:

Bit	Type	Reset	Description
63:58~~57~~	RO	0h	Reserved
57	RO	Impl Spec	**Controller Memory Buffer Supported (CMBS):** If set to '1', this bit indicates that the controller supports the Controller Memory Buffer, and that addresses supplied by the host are permitted to reference the Controller Memory Buffer only if the host has enabled the Controller Memory Buffer's controller memory space. If the controller supports the Controller Memory Buffer, this bit shall be set to '1'.

Modify portions of section 3.1.11 (Offset 38h: CMBLOC – Controller Memory Buffer Location) as shown below:

This optional register defines the location of the Controller Memory Buffer (refer to section 4.7). If ~~CMBSZ is 0~~ the controller does not support the Controller Memory Buffer (CAP.CMBS), this register is reserved. If the controller supports the Controller Memory Buffer and CMBMSC.CRE is cleared to '0', this register shall be cleared to 0h.

Modify portions of section 3.1.12 (Offset 3Ch: CMBSZ – Controller Memory Buffer Size) as shown below:

This optional register defines the size of the Controller Memory Buffer (refer to section 4.7). If the controller does not support the Controller Memory Buffer feature or if the controller supports the Controller Memory Buffer (CAP.CMBS) and CMBMSC.CRE is cleared to '0', then this register shall be cleared to 0h.

Insert new sections following section 3.1.15 (Offset 48h: BPMBL – Boot Partition Memory Buffer Location (Optional)) as shown below:

3.1.TBD4 Offset 50h: CMBMSC – Controller Memory Buffer Memory Space Control

This register specifies how the controller references the Controller Memory Buffer with host-supplied addresses. If the controller supports the Controller Memory Buffer (CAP.CMBS), this register is mandatory. Otherwise, this register is reserved.

This register shall be reset by neither Controller Reset nor Function Level Reset, but it shall be reset by all other Controller Level Resets.

Figure TBD5: Offset 50h: CMBMSC – Controller Memory Buffer Memory Space Control

Bit	Type	Reset	Description
63:12	RW	0h	**Controller Base Address (CBA):** This field specifies the 52 most significant bits of the 64-bit base address for the Controller Memory Buffer's controller address range. The Controller Memory Buffer's controller base address and its size determine its controller address range. The specified address shall be valid only under the following conditions: a) no part of the Controller Memory Buffer's controller address range is greater than $2^{64} - 1$; and b) if the Persistent Memory Region's controller memory space is enabled, then the Controller Memory Buffer's controller address range does not overlap the Persistent Memory Region's controller address range.
11:02	RO	0h	Reserved
01	RW	0	**Controller Memory Space Enable (CMSE):** This bit specifies whether addresses supplied by the host are permitted to reference the Controller Memory Buffer. If CMBSMSC.CRE is cleared to '0' this bit has no effect, and the Controller Memory Buffer's controller memory space is not enabled. If this bit is set to '1' and the controller base address is valid, then the Controller Memory Buffer's controller memory spaced is enabled. Otherwise, the controller memory space is not enabled. If the Controller Memory Buffer's controller memory space is enabled, then addresses supplied by the host that fall within the Controller Memory Buffer's controller address range shall reference the Controller Memory Buffer. If the Controller Memory Buffer's controller memory space is not enabled, then no address supplied by the host shall reference the Controller Memory Buffer. Instead, such addresses shall reference memory spaces other than the Controller Memory Buffer.
00	RW	0	**Capabilities Registers Enabled (CRE):** This bit specifies whether the CMBLOC and CMBSZ registers are enabled. If this bit is set to '1', then CMBLOC is defined as shown in Figure 82 and CMBSZ is defined as shown in Figure 83. If this bit is cleared to '0', then CMBSZ and CMBLOC are cleared to 0h.

3.1.TBD4+2 Offset 58h: CMBSTS – Controller Memory Buffer Status

This register indicates the status of the Controller Memory Buffer. If the controller supports the Controller Memory Buffer (CAP.CMBS), this register is mandatory. Otherwise, this register is reserved.

Figure TBD5+2: Offset 58h: CMBSTS – Controller Memory Buffer Status

Bits	Type	Reset	Description
31:01	RO	0h	Reserved
00	RO	0	**Controller Base Address Invalid (CBAI):** This bit indicates whether the controller has failed to enable the Controller Memory Buffer's controller memory space because CMBMSC.CBA is invalid. If CMBMSC.CRE and CMBMSC.CMSE are set to '1', and CMBMSC.CBA is invalid, this bit shall be set to '1'. Otherwise, this bit shall be cleared to '0'.

Modify portions of section 3.1.16 (Offset E00h: PMRCAP – Persistent Memory Region Capabilities) defined in TP 4000a as shown below:

3.1.16 Offset E00h: PMRCAP – Persistent Memory Region Capabilities

…

Figure 87: Offset E00h: PMRCAP – Persistent Memory Region Capabilities

Bit	Type	Reset	Description
31:25~~24~~	RO	0h	Reserved
24	RO	Impl Spec	**Controller Memory Space Supported (CMSS):** If set to '1', this bit indicates that addresses supplied by the host are permitted to reference the Persistent Memory Region only if the host has enabled the Persistent Memory Region's controller memory space. If the controller supports referencing the Persistent Memory Region with host-supplied addresses, this bit shall bet set to '1'. Otherwise, this bit shall be cleared to '0'.
…	…	…	…
4	RO	Impl Spec	**Write Data Support (WDS):** If this bit is set to '1', then the controller supports data and metadata in the Persistent Memory Region for commands that transfer data from the host to the controller (e.g., Write). If this bit is cleared to '0', then data and metadata for commands that transfer data from the host to the controller shall not be transferred to the Persistent Memory Region. If PMRCAP.CMSS is cleared to '0', this bit shall be cleared to '0'.
3	RO	Impl Spec	**Read Data Support (RDS):** If this bit is set to '1', then the controller supports data and metadata in the Persistent Memory Region for commands that transfer data from the controller to the host (e.g., Read). If this bit is cleared to '0', then all data and metadata for commands that transfer data from the controller to the host shall not be transferred from the Persistent Memory Region. If PMRCAP.CMSS is cleared to '0', this bit shall be cleared to '0'.
…	…	…	…

Modify portions of section 3.1.17 (Offset E04h: PMRCTL – Persistent Memory Region Control) as shown below:

3.1.TBDX+1 ~~3.1.17~~ Offset E04h: PMRCTL – Persistent Memory Region Control

Modify portions of section 3.1.18 (Offset E08h: PMRSTS – Persistent Memory Region Status) as shown below:

3.1.TBDX+2 ~~3.1.18~~ Offset E08h: PMRSTS – Persistent Memory Region Status

This optional register provides the status of the Persistent Memory Region. If the controller does not support the Persistent Memory Region feature, then this register shall be cleared to 0h.

Bits	Type	Reset	Description
31:13~~12~~	RO	0h	Reserved
12	RO	0	**Controller Base Address Invalid (CBAI):** This field indicates whether the controller has failed to enable the Persistent Memory Region's controller memory space because PMRMSC.CBA is invalid. If PMRCAP.CMSS is set to '1', PMRMSC.CMSE is set to '1', and PMRMSC.CBA is invalid, this bit shall be set to '1'. Otherwise, this bit shall be cleared to '0'.

<table>
<tr><th>Bits</th><th>Type</th><th>Reset</th><th>Description</th></tr>
<tr><td>11:9</td><td>RO</td><td>0h</td><td>Health Status (HSTS): If the Persistent Memory Region is ready, then this field indicates the health status of the Persistent Memory Region. This field is always cleared to 000b when the Persistent Memory Region is not ready.

The health status values are defined as:

<table>
<tr><th>Value</th><th>Definition</th></tr>
<tr><td>000b</td><td>Normal Operation: The Persistent Memory Region is operating normally.</td></tr>
<tr><td>001b</td><td>Restore Error: The Persistent Memory Region is operating normally and is persistent; however, the contents of the Persistent Memory Region may not have been restored correctly (i.e., may not contain the contents prior to the last power cycle, NVM subsystem reset, controller reset, or ~~Peristent~~ Persistent Memory Region disable).</td></tr>
<tr><td>010b</td><td>Read Only: The Persistent Memory Region is read only. PCI Express memory write requests do not update the Persistent Memory Region. PCI Express memory read requests to the Persistent Memory Region return correct data.</td></tr>
<tr><td>011b</td><td>Unreliable: The Persistent Memory Region has become unreliable. PCI Express memory reads may return invalid data or generate poisoned PCI Express TLP(s). Persistent Memory Region memory writes may not update memory or may update memory with undefined data. The Persistent Memory Region may also have become non-persistent.</td></tr>
<tr><td>100b to 111b</td><td>Reserved</td></tr>
</table>
</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td></tr>
</table>

Insert section 3.1.TBD8 following section 3.1.20 (Offset E10h: PMRSWTP – Persistent Memory Region Sustained Write Throughput) from TP 4032 PMR Elasticity Status as shown below:

3.1.TBD8 Offset E14h: PMRMSC – Persistent Memory Region Memory Space Control

This register specifies how the controller references the Persistent Memory Region with host-supplied addresses. If the controller supports the Persistent Memory Region's controller memory space (PMRCAP.CMSS), this register is mandatory. Otherwise, this register is reserved.

This register shall not be reset by Controller Reset.

Figure TBD9: Offset E14h: PMRMSC – Persistent Memory Region Memory Space Control

Bit	Type	Reset	Description
63:12	RW	0h	**Controller Base Address (CBA):** This field specifies the 52 most significant bits of the 64-bit base address for the Persistent Memory Region's controller address range. The Persistent Memory Region's controller base address and its size determine its controller address range. The specified address shall be valid only under the following conditions: a) no part of the Persistent Memory Region's controller address range is greater than $2^{64} - 1$; and b) if the Controller Memory Buffer's controller memory space is enabled, then the Persistent Memory Region's controller address range does not overlap the Controller Memory Buffer's controller address range.
11:02	RO	0h	Reserved

Figure TBD9: Offset E14h: PMRMSC – Persistent Memory Region Memory Space Control

Bit	Type	Reset	Description
01	RW	0	**Controller Memory Space Enable (CMSE):** This bit specifies whether addresses supplied by the host are permitted to reference the Persistent Memory Region. If this bit is set to '1' and the controller base address is valid, then the Persistent Memory Region's controller memory space is enabled. Otherwise, the controller memory space is not enabled. If the Persistent Memory Region's controller memory space is enabled, then addresses supplied by the host that fall within the Persistent Memory Region's controller address range shall reference the Persistent Memory Region. If the Persistent Memory Region's controller memory space is not enabled, then no address supplied by the host shall reference the Persistent Memory Region. Instead, such addresses shall reference memory spaces other than the Persistent Memory Region.
00	RO	0	Reserved

Modify portions of section 4.7 (Controller Memory Buffer) as shown below:

The Controller Memory Buffer (CMB) is a region of general purpose read/write memory on the controller. The controller indicates support for the CMB by setting CAP.CMBS to '1'. The host indicates intent to use the CMB by setting CMBMSC.CRE to '1'. Once this bit is set, the controller indicates the properties of the CMB via the CMBLOC and CMBSZ registers.

The CMB ~~that~~ may be used for a variety of purposes. The controller indicates which purposes the memory may be used for by setting support flags in the CMBSZ register.

The CMB's PCI Express address range is used for external memory read and write requests to the CMB. The PCI Express base address of the CMB is defined by the PCI Base Address Register (BAR) indicated by CMBLOC.BIR, and the offset indicated by CMBLOC.OFST. The size of the CMB is indicated by CMBSZ.SZ.

The controller uses the CMB's controller address range to reference CMB with addresses supplied by the host. The PCI Express address range and the controller address range of the CMB may differ, but both ranges have the same size, and equivalent offsets within each range have a one-to-one correspondence. The host configures the controller address range via the CMBMSC register.

The host enables the CMB's controller memory space via the CMBMSC.CMSE bit. When controller memory space is enabled, if host supplies an address referencing the CMB's controller address range, then the controller directs memory read or write requests for this address to the CMB.

When the CMB's controller memory space is disabled, the controller does not consider any host-supplied address to reference the CMB's controller address range, and memory read and write requests are directed elsewhere (e.g., to memory other than the CMB).

To prevent possible misdirection of the controller's memory requests, before host software enables the CMB's controller memory space, it should configure the CMB's controller address range to so that it does not overlap any address that host software intends to use for DMA.

In earlier versions of this specification, for a controller that supports the CMB, the CMB's controller address range is fixed to be equal to its PCI Express address range, and the CMB's controller memory space is always enabled whenever the controller is enabled. To prevent misdirection of controller memory requests when such a controller is assigned to a virtual machine, host software (on the hypervisor or host OS) should not enable translation of the CMB's PCI Express address range, and it should ensure that this address range does not overlap any range of pre-translated addresses that the virtual machine may use for DMA.

Host software on hypervisor or host OS may pre-configure CMBMSC so that CMB will to operate when the controller is assigned to virtual machine that only supports the CMB as defined in earlier versions of this specification. To prevent the virtual machine from unintentionally clearing CMBMSC, the contents of CMBSMC are preserved across Controller Reset and Function Level Reset.

Modify portions of section 4.TBDZ (Persistent Memory Region) in TP 4000a as shown below:

4.TBDZ ~~8~~ Persistent Memory Region

The Persistent Memory Region (PMR) is an optional region of general purpose PCI Express read/write persistent memory that may be used for a variety of purposes. The controller indicates support for the PMR by setting CAP.PMRS to '1' and indicates whether the controller supports command data and metadata transfers to or from the PMR by setting support flags in the PMRCAP register. When command data and metadata transfers to or from PMR are supported, all data and metadata associated with a particular command shall be either entirely located in the Persistent Memory Region or outside the Persistent Memory Region.

The PMR's PCI Express address range is used for external memory read and write requests to the PMR. The PCI Express address range and size of the PMR is defined by the PCI Base Address Register (BAR) indicated by PMRCAP.BIR. The PMR consumes the entire address region exposed by the BAR and supports all the required features of the PCI express programming model (i.e., it in no way restricts what is otherwise permitted by PCI Express).

The controller uses the PMR's controller address range to reference PMR with addresses supplied by the host. The PCI Express address range and the controller address range of the PMR may differ, but both ranges have the same size, and equivalent offsets within each range have a one-to-one correspondence. The host configures the controller address range via the PMRMSC register.

The host enables the PMR's controller memory space via the PMRMSC.CMSE bit. When controller memory space is enabled, if host supplies an address referencing the PMR's controller address range, then the controller directs memory read or write requests for this address to the PMR.

When the PMR's controller memory space is disabled, the controller does not consider any host-supplied address to reference the PMR's controller address range, and memory read and write requests are directed elsewhere (e.g., to memory other than the PMR).

Modify portions of section 7.3.2 (Controller Level Reset) as shown below:

When any of the above resets occur, the following actions are performed:

- The controller stops processing any outstanding Admin or I/O commands;
- All I/O Submission Queues are deleted;
- All I/O Completion Queues are deleted;
- The controller is brought to an idle state. When this is complete, CSTS.RDY is cleared to '0'; and
- All ~~other~~ controller registers defined in section 3 and internal controller state are reset~~.~~, except as follows:
  - ~~T~~the Admin Queue registers (AQA, ASQ, or ACQ) are not reset as part of a ~~c~~Controller ~~r~~Reset;
  - the Controller Memory Buffer Memory Space Control register (CMBMSC) is reset as part of neither a Controller Reset nor a Function Level Reset; and
  - the Persistent Memory Region Memory Space Control register (PMRMSC) is not reset as part of a Controller Reset.