Zoned Storage Models

Version 0.9 Revision 0

Abstract: This SNIA document defines recommended behavior for hardware and software that supports Zoned Storage.

Publication of this Working Draft for review and comment has been approved by the Computational Storage TWG. This draft represents a “best effort” attempt by the Computational Storage TWG to reach preliminary consensus, and it may be updated, replaced, or made obsolete at any time. This document should not be used as reference material or cited as other than a “work in progress.” Suggestions for revisions should be directed to https://www.snia.org/feedback/.

Working Draft

October 31, 2022

USAGE

Copyright © 2022 SNIA. All rights reserved. All other trademarks or registered trademarks are the property of their respective owners.

The SNIA hereby grants permission for individuals to use this document for personal use only, and for corporations and other business entities to use this document for internal use only (including internal copying, distribution, and display) provided that:

1. Any text, diagram, chart, table or definition reproduced shall be reproduced in its entirety with no alteration, and,

2. Any document, printed or electronic, in which material from this document (or any portion hereof) is reproduced, shall acknowledge the SNIA copyright on that material, and shall credit the SNIA for granting permission for its reuse.

Other than as explicitly provided above, you may not make any commercial use of this document or any portion thereof, or distribute this document to third parties. All rights not explicitly granted are expressly reserved to SNIA.

Permission to use this document for purposes other than those enumerated above may be requested by e-mailing tcmd@snia.org. Please include the identity of the requesting individual and/or company and a brief description of the purpose, nature, and scope of the requested use.

All code fragments, scripts, data tables, and sample code in this SNIA document are made available under the following license:

BSD 3-Clause Software License

Copyright (c) 2022, The Storage Networking Industry Association.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

* Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

* Neither the name of The Storage Networking Industry Association (SNIA) nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

DISCLAIMER

The information contained in this publication is subject to change without notice. The SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

DRAFT

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Revision History

Major	Revision	Editor	Date	Comments
0.1	R0	Yoni Shternhell	18-Jul-2022	Initial draft.
0.1	R1	Yoni Shternhell	19-Jul-2022	Modification after TWG call
0.1	R2	Yoni Shternhell	25-Jul-22	Add to the Architecture section each of the attribute used in the 'Characteristics' in each model.
01	R4	Yoni Shternhell	18-Aug-22	Added text and figured in the Theory of Operation section for block storage devices
0.1	R7	Yoni Shternhell	20-Sep-22	Integrated Paul S. comments
0.1	R8	Yoni Shternhell	20-Sep-22	Incorporated comments after the group call
0.1	R8	Yoni Shternhell	26-Sep-22	Clean version
0.1	R10	Yoni Shternhell	10-Oct-22	Modified figure 2 Added reference to TP4115 and TP4076a
0.1	R11	Yoni Shternhell	26-Oct-22	Modifications from the SNIA F2F meeting
0.9	R12	Yoni Shternhell	31-Oct-22	Clean version for Public and Members review

DRAFT

Table of Contents

DRAFT

Table of Figures

No table of figures entries found.

FOREWORD

The SNIA Zoned Storage TWG was formed to facilitate a common industry understanding of Zoned Storage use cases, device architectures and programming model, providing a framework to enable the development of a robust Zoned Storage software and hardware ecosystem.

This SNIA specification outlines the architectural models and use cases that are used for Zoned Storage devices. As this specification is developed, requirements in interface standards and specific APIs may be proposed as separate documents and developed in the appropriate organizations.

DRAFT

1 Scope

This specification defines the requirements and use case models that can be implemented in a Zoned Storage device.

A Zoned Storage device has several aspects:

- **Common Characteristics**. The base properties of a Zoned Storage device that host software expects
- **Security.** The base security requirements for a Zoned Storage device.
- **Models**. Describes two Zoned Storage device Models that covers known uses cases.

2 References

The following referenced documents are indispensable for the application of this document.

For references available from ANSI, contact ANSI Customer Service Department at (212) 642-4980 (phone), (212) 302-1286 (fax) or via the World Wide Web at https://www.ansi.org.

NVMe NVM Express® Base Specification 2.0b,
Approved standard, available from https://nvmexpress.org

NVMe NVM Express® Zoned Namespace Command Set
Approved standard, available from https://nvmexpress.org

NVMe NVM Express® TP4115 Namespace Management Zoned Namespace Enhancement, approved TP, available from https://nvmexpress.org

NVMe NVM Express® TP4076a Namespace Management Zoned Namespace Enhancement, approved TP, available from https://nvmexpress.org

T13 INCITS 537-2016 Information Technology – Zoned Device ATA Command Set (ZAC)
Approved standard, available from https://webstore.ansi.org

T10 INCITS 536-2016 Information Technology – Zoned Block Commands (ZBC)
Approved standard, available from https://webstore.ansi.org

DRAFT

3 Definitions, abbreviations, and conventions

For the purposes of this document, the following definitions and abbreviations apply.

3.1 Definitions

3.1.1 zone

A contiguous range of logical block addresses that are managed as a single unit.

3.1.2 zoned namespace

A namespace that is divided into zones and is associated with the Zoned Namespace Command Set.

3.2 Keywords

In the remainder of the specification, the following keywords are used to indicate text related to compliance:

3.2.1 mandatory

a keyword indicating an item that is required to conform to the behavior defined in this standard

3.2.2 may

a keyword that indicates flexibility of choice with no implied preference; “may” is equivalent to “may or may not”

3.2.3 may not

keywords that indicate flexibility of choice with no implied preference; “may not” is equivalent to “may or may not”

3.2.4 need not

keywords indicating a feature that is not required to be implemented; “need not” is equivalent to “is not required to”

3.2.5 optional

a keyword that describes features that are not required to be implemented by this standard; however, if any optional feature defined in this standard is implemented, then it shall be implemented as defined in this standard

3.2.6 shall

a keyword indicating a mandatory requirement; designers are required to implement all such mandatory requirements to ensure interoperability with other products that conform to this standard

3.2.7 should

a keyword indicating flexibility of choice with a strongly preferred alternative

3.3 Abbreviations

LUN Logical Unit Number

NVM Non-Volatile Memory

SSD Solid State Drive

HDD Hard Disk Drive

ZNS Zoned Namespace

DRAFT

4 Theory of Operation

Zoned Storage devices is a block storage device that has its LBA space divided into zones. A zone is of a certain type, which defines the rules for accesses its LBAs.

For example, one such type is Sequential Write Required, which requires that LBAs within a zone are written in sequential order, but can be read in any order.

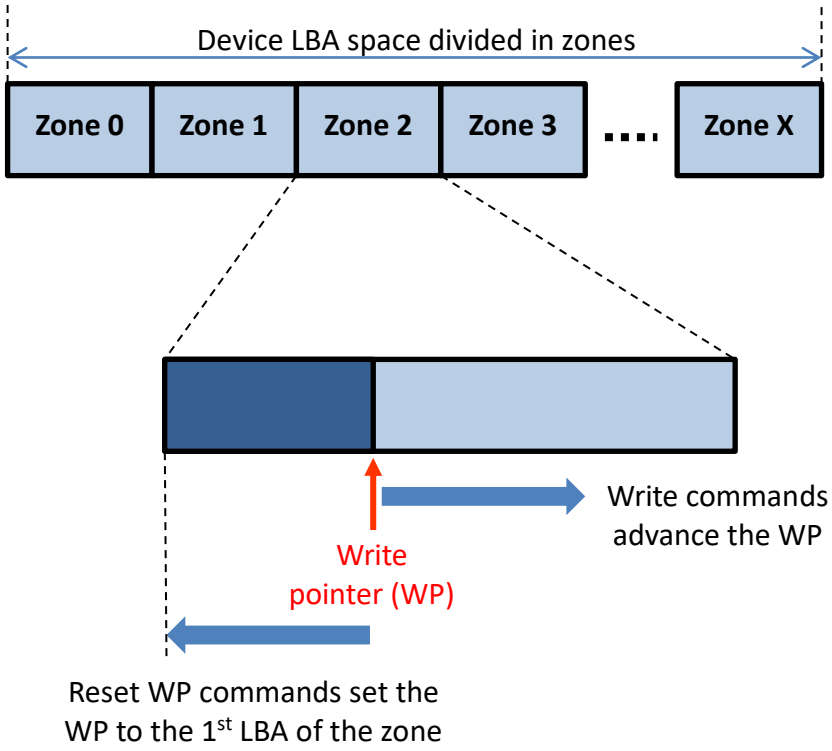

Figure 1 – Conventional device and ZNS device internal data placement

The only way to change a logical block already written to a zone is to reset the Write Pointer (WP), effectively deleting all the data in the zone, and restart writing from the beginning of the zone. Reading data has no restrictions and the data can be read in the same manner as on traditional storage devices.

The zone abstraction allows the host to align its writes to the sequential write required properties of the Zoned Storage device, and thereby optimizes data placement onto the device's media. Note that the management of media reliability continues to be the sole responsibility of the Zoned Storage devices and should be managed the same way as conventional devices.

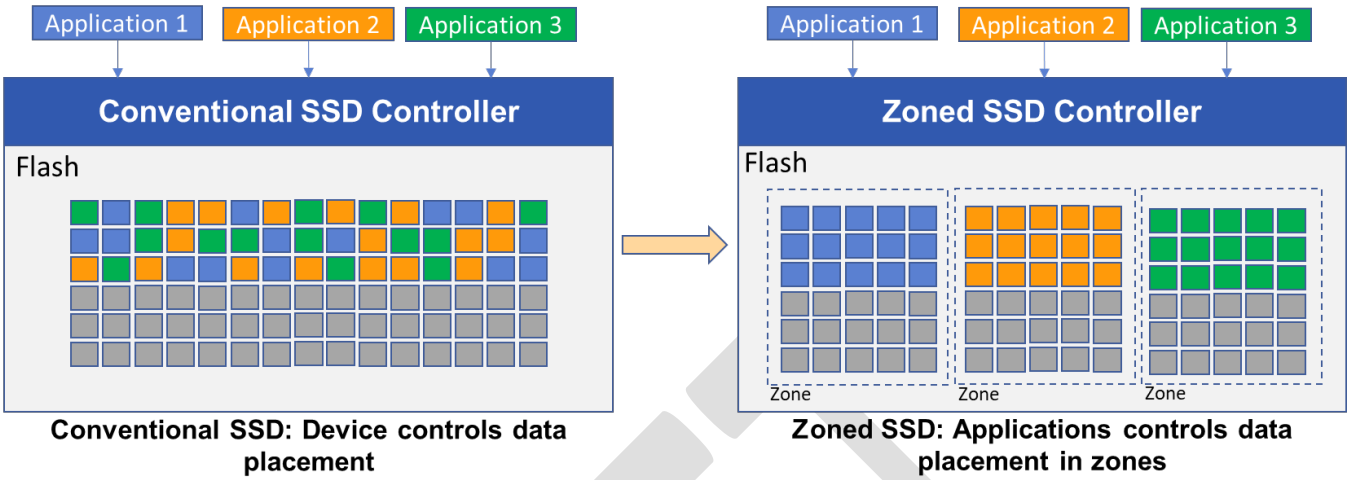

Figure 2 – Conventional device and Zoned Storage device internal data placement

Shingled Magnetic Recording (SMR) technology has been introduced in Hard Disk Drives (HDDs) to enable increased areal density and larger capacities, and to improve the cost-effectiveness of HDDs. In SMR, unlike conventional recording, tracks are written in an overlapping manner. This allows tracks to be more tightly packed and hence to achieve a higher recording density. However, once the tracks are overlapped, a logical block within a zone cannot be written independently. To manage the recording, the disk surface is divided into Zones with a gap left between zones. This allows each zone to be written and erased independently. Multiple approaches are possible to manage the recording restriction.

A conventional device handles the recording constraint internally and exposes a conventional interface to the host. Unfortunately, for large-scale systems, where performance and space utilization must be carefully managed, once cannot rely on the device-side localized management. Therefore, managing the complexity on the host side is almost a requirement for large storage systems.

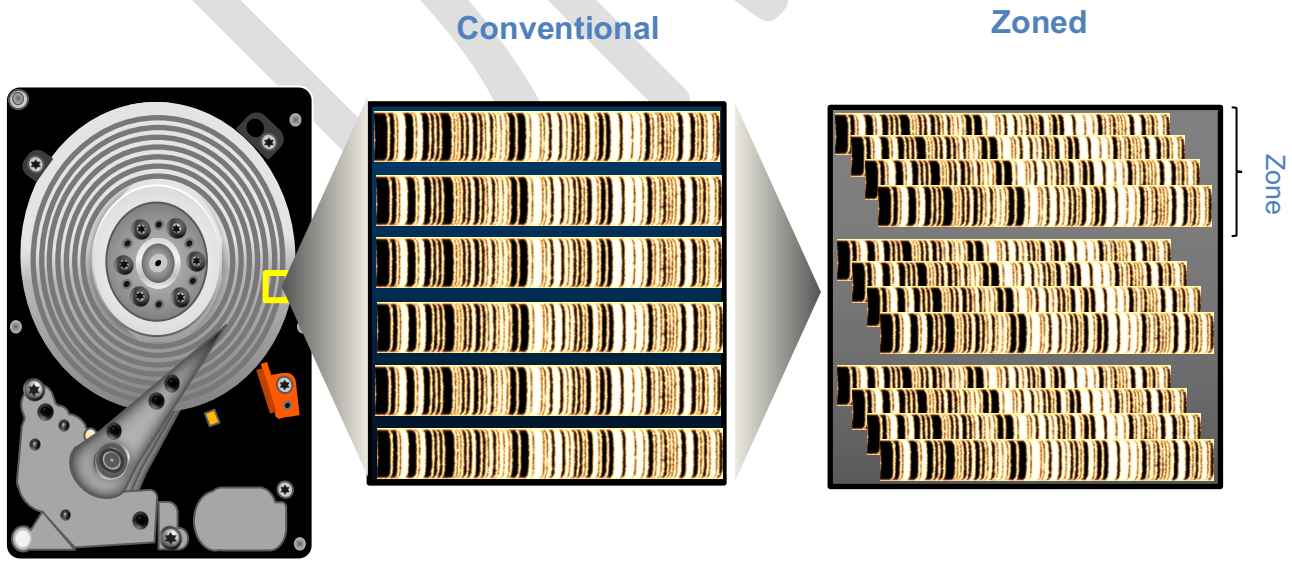

Figure 3 – Conventional device and Zoned Storage device internal data placement

The Zone Storage Device Model is standardized for storage devices as

- ZBC: Zoned Block Commands in T10/SAS
- ZAC: Zoned Device Command Set in T13/SATA
- ZNS: Zoned Namespace Command Set in NVM Express

4.1 Overview

This section provides an overview for Zoned Storage Model.

A Zoned Storage Model consists of a set of base requirements that applies to all SNIA Zoned Storage Models, followed by an additional set of requirements for a given Zoned Storage Model. A generic architecture description of the Zoned Storage Model is illustrated in Figure 4.1.

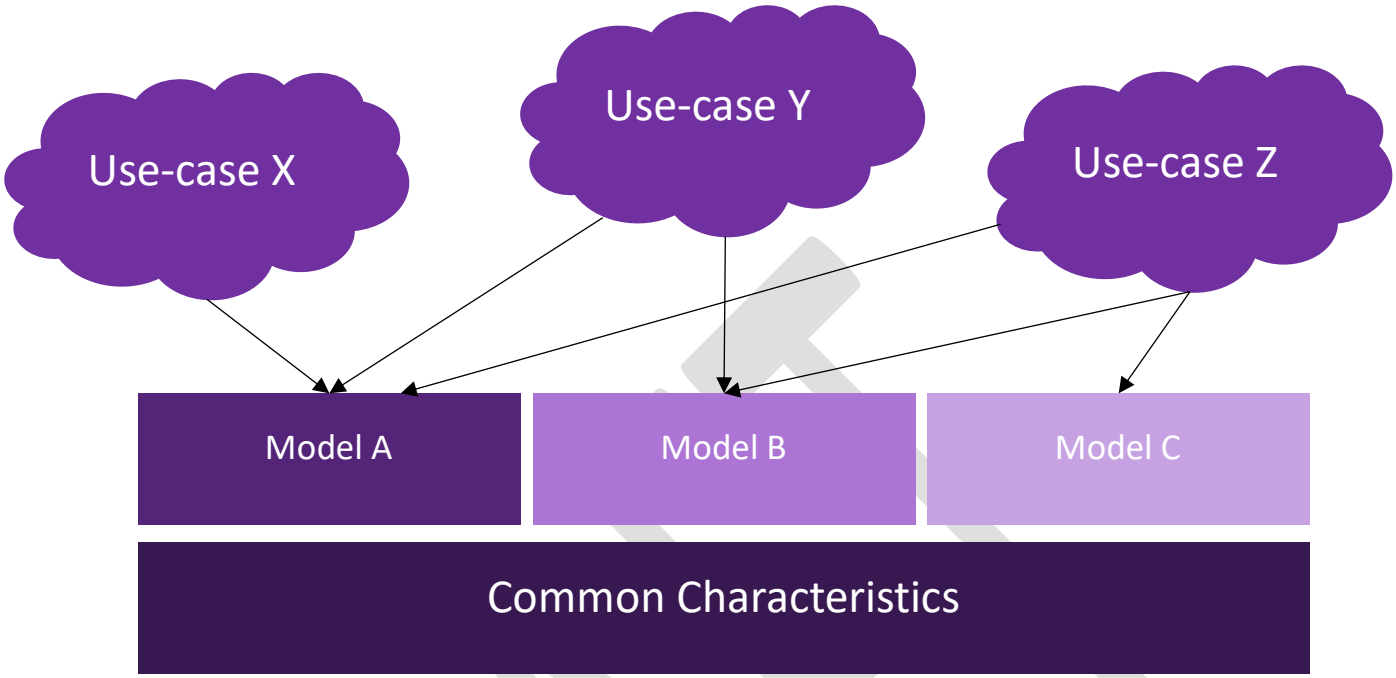

Figure 4 – An Architectural view of the Zone Storage Model

The Zoned Storage Model allows the host storage stack to always assume certain properties of a Zoned Storage device.

The NVMe Base Specification defines an interface for host software to communicate with a non-volatile memory subsystem. The NVMe ZNS specification defines additional functionality for the Zoned Namespace Command Set. This specification defines comprehensive requirements that apply to all SNIA Zoned Storage Models.

4.2 Characteristics

A Zoned Storage device will make use of several key characteristics across the Zoned Storage Models.

4.2.1 Zone Device Protocol

The protocol used by a Zoned Storage device (i.e., NVMe ZNS, T13 ZAC, and T10 ZBC).

4.2.2 Zone Type

The zone type characteristic defines the rules for reading and writing to a zone (e.g., a zone type of Sequential Write Required).

4.2.3 Zone Capacity

The Zone Capacity characteristic defines the writeable capacity of that zone.

4.2.4 Zone Active Resources Available

The Zone Active Resources Available characteristic indicates the total number of active resources (allocated and unallocated active resources). This characteristic may only apply to certain Zone Device Protocols.

4.2.5 Zone Open Resources Available

The Zone Open Resources Available characteristic indicates the total number of open resources (allocated and unallocated open resources). This characteristic may only apply to certain Zone Device Protocols.

4.2.6 Mandatory I/O Access Commands

The mandatory commands for a Zoned Storage device (e.g., Read command, Write command, etc.).

4.2.7 Mandatory Access Command

The mandatory commands for an NVMe Zoned Storage device. This characteristic is applicable only to a ZNS SSD.

4.2.8 ZRWA

The optional ZRWA feature defines an area with a set of assigned LBAs which start at the write pointer for a given zone in which the logical blocks that are mapped to that area may be written in random order as well as overwritten. Data flushed from ZRWA to a zone is written sequentially to the zone at the write pointer. Any NVMe Zoned Storage device that implements the ZRWA feature should refer to the NVM Express Zoned Namespace Command Set.

4.3 Common Requirements

This section defines the base properties of a Zoned Storage device that host software expects.

4.3.1 NVM Express Reliability Requirement

The device shall manage media reliability issues related to accessing media in the same way as managed within conventional storage devices (i.e., HDD and/or SSD).

For example, the device shall manage media reliability issues internally caused by:

- Write errors if media programs are correctable by the device.
- Prematurely worn-out flash blocks associated to a zone. I.e, Flash block(s) associated to a zone must not be fixed and should be wear-leveled across zones.

- Read/program disturbs caused by open zones, excessive reads, or similar media characteristics.

4.3.2 Offline Zone(s)

An Offline zone cannot be read or written, e.g., as a result of media errors. Whether an Offline zone can become readable or writeable is outside the scope of this standard.

For a Zoned Storage device, zones shall not autonomously transition to the Offline state within the warranted period and/or guarantees of a drive.

Outside of the warranted period and/or guarantee, the Zoned Storage device may transition a zone to the Offline state when no longer possible to write or read the data reliably.

4.3.3 SSD Specific Requirements

The controller shall not exhibit Active Zone Excursions related to Active Zones (i.e., the controller shall not transition open zones to the Full State due to one or more vendor-specific excursion events). Refer to the NVM Express Zoned Namespace Command Set.

The controller shall maintain a fixed number of writeable LBAs within a zone over the lifetime of NVMe namespace (i.e., the controller is not able to change the writeable capacity of a zone between resets). Refer to the NVM Express Zoned Namespace Command Set for further information.

The number of active and open resources should be equal.

From the time a zoned namespace is formatted or created, the zone capacity shall be fixed (i.e., No Variable Zone Capacity feature).

4.3.4 HDD Specific Requirements

This standard defines no requirements specific to HDDs.

4.3.5 Security

There are no security specific requirements for a Zoned Storage device beyond what are required for storage devices in general.

5 Models

5.1 Model A

5.1.1 Overview

This section describes the requirements for a Zoned Storage device Model that is a good all-round device model for all to adopt.

5.1.1.1 Applicable Use Cases

Zoned Device Model that minimizes the host software required changes to support zoned block devices.

Works as a drop-in replacement for existing storage devices.

Host software must respect the sequential write requirement of the zone type, and similarly reset a zone to rewrite a zone.

Common use-cases, but not limited to:

- Streaming applications. Sequential writes and random reads.
- Database applications. Write Ahead Log (WAL) and log-structured writes.
- Storage arrays. Great data protection and high performance.

5.1.1.2 Characteristics

Characteristic Type	Value	Note(s)	Reference
Zone Type	Sequential Write Required		See section 4.1.2.
Zone Active Resources Available	12 or more recommended. Recommend that the number of active and open resources are equal.	Does not apply to ZBC/ZAC devices (i.e., SMR HDDs). Recommend that the number of active and open resources are equal.	See section 4.1.4 and the NVM Express Zoned Namespace Command Set.

Zone Open Resources Available	12 or more recommended.		See section 4.1.5. and the NVM Express Zoned Namespace Command Set.
Performance Characteristics	Accessing 1-4 zones concurrently, should achieve the maximum throughput of the associated media to the namespace and/or device.		
Mandatory I/O Access Commands	Read and Write commands		See section 4.1.6.
Mandatory Access Command (ZNS SSD only)	Zone Append		See section 4.1.7.

Table 1 Model A Characteristics

5.2 Model B

5.2.1 Overview

5.2.1.1 Applicable Use-cases

Zoned Device Model that minimizes the host software required changes to support zoned block devices but requires high host I/O parallelism to achieve the full media bandwidth of a given device. The host software must:

- Respect the sequential write requirement of the zone type, and similarly reset a zone to rewrite a zone;
- Must access multiple zones in parallel to achieve the full bandwidth of the media; and
- Must perform adequate parity protection to account for lower device UBER.

Common use-cases, but not limited to:

- Archival storage. E.g., storage with host-defined erasure encoding.

5.2.1.2 Characteristics

Characteristic Type	Value	Note(s)	Reference
Zone Type	Sequential Write Required		See section 4.1.2.
Zone Active Resources Available	Depends on device. Recommend that the number of active and open resources are equal.	Does not apply to ZBC/ZAC devices (i.e., SMR HDDs).	See section 4.1.4 and the NVM Express Zoned Namespace Command Set.
Zone Open Resources Available	Depends on device		See section 4.1.5. and the NVM Express Zoned Namespace Command Set.
Performance Characteristics	Depends on device	Host must access minimum number of zones concurrently, as defined by the device, to achieve the maximum throughput of the associated media to the namespace and/or device.	

Table 2 Model B Characteristics