Western Digital.

Towards Large-scale Deployments with Zoned Namespace SSDs

Matias Bjørling, Distinguished Engineer

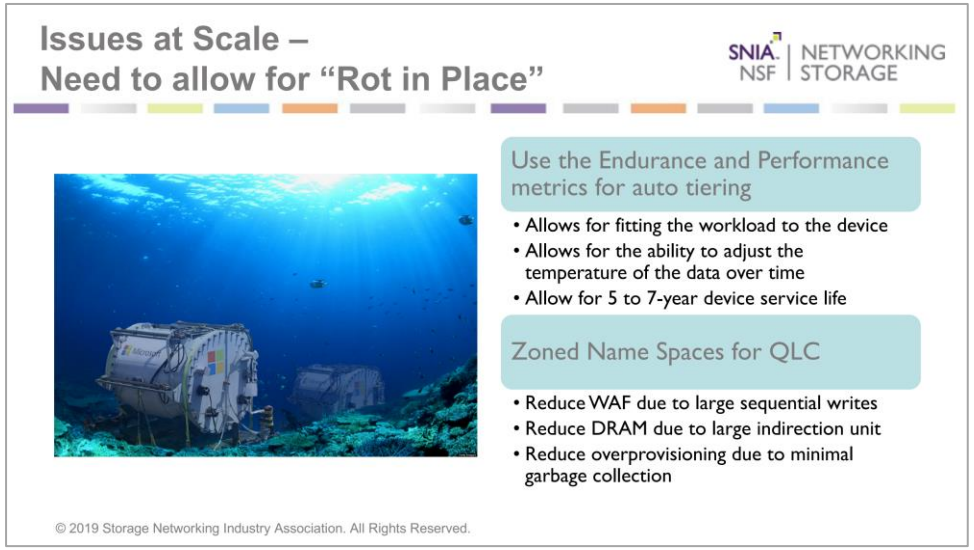
9/20-2023

Storage at Scale

Hyperscalers and Cloud Service Provides (CSPs)

- Constantly challenged with large volumes of data and increasing customer demand for cost-effective storage and high performance*
  - Performance: IOPS/TB, Throughput, Latency/QoS
  - TCO Impacts: Capacity vs Performance
  - Lifetime/DWPD: ~1DWPD
- Maintain servers online after the typical 5 year's service time**
  - Reduce carbon emissions as well as overall fleet cost
  - Server repair costs first increases significant at year 10 and onwards. Servers are retired prematurely
  - Microsoft server fleet lifetime increased from 5 to 7 years*. Meta rapidly increasing its fleet time as well. Now 5 years***.

Metrics*	Typical
Performance	IOPS/TB, Throughput, Latency/QoS
Cost Impact	Capacity/Performance
Lifetime/DWPD	5-7 years (Req. >~1DWPD)



** Lyu et. al, Myths and Misconceptions Around Reducing Carbon Embedded in Cloud Platforms, HotCarbon, 2023

*** Rich Miller - Meta Will Run its Servers For Up to 5 Years, 2023
https://www.datacenterfrontier.com/hyperscale/article/21548840/meta-will-abandon-some-data-center-builds-run-servers-longer

* Lee Prewitt, Microsoft - How Facebook & Microsoft Leverage NVMe Cloud Storage.
https://www.brighttalk.com/webcast/663/374596

Storage at Scale

Hyperscalers and Cloud Service Provides (CSPs)

- Conventional SSDs **not able** to serve storage at scale
  - Typical lifetime 3-5 years. 7+ years wanted.
  - Either High Cost (TLC) and/or Low DWPD (QLC)
- Need SSDs that eliminates write amplification to fulfill DWPD, Lifetime and Performance requirements
  - SSDs with Zoned Namespace (ZNS) support solve these challenges

30% Cost Saving ($/GB) - ZNS

15% Cost Saving ($/GB) - QLC

	Conventional SSD (28% OP)	ZNS SSD (0% OP)
TLC	• **Over-provisioning needed for drive FTL.** • **Drive run at reduced capacity to improve performance, latency & endurance.**	• **0% OP for drive FTL (extra capacity for host).** • **Drive can use full capacity with full performance, low latency & high DWPD.**
QLC	• Over-provisioning needed for drive FTL. • Drive run at reduced capacity to improve performance, latency & endurance. • **Increased capacity per die.**	• 0% OP for drive FTL (extra capacity for host). • Drive can use full capacity with full performance, low latency & high endurance. • **Increased capacity per die.**

Metric	Conventional SSD		SSDs with Zoned Namespace Support	
	TLC	QLC	TLC (Performance)	QLC (Capacity)
IOPS/TB	++	+	+++	++
Throughput	++ (Read/Write)	+ (Read)	+++ (Read/Write)	++ (Read)
Latency/QoS	++	+	+++	++
Lifetime	++ (Typ. >1 DWPD)	+ (Typ. 0.3-0.5 DWPD)	+++ (Typ. >3.5 DWPD)	++ (Typ. >1 DWPD)
Cost (TB/$)	+	++	++	+++

High Cost

Low DWPD

High Perf.

Balanced

SSDs with Zoned Namespaces (ZNS)?

Performance is Expensive

"To achieve these levels of device-level write amplification (1.1x & 1.4x), flash is typically overprovisioned by 50% (…) but reducing flash overprovisioning while maintaining the current level of performance is an open challenge at Facebook."

Source: The CacheLib Caching Engine: Design and Experiences at Scale. USENIX OSDI 2020

Caching Use-Case	General		CacheLib (7.68TB workload)	
	SSD	SSD /w ZNS	SSD	SSD /w ZNS
SSD Capacity	**7.68T**	**8T**	**15.36T**	**8T**
NAND Usable	**$584**	**$584**	**$584**	**$584**
NAND Over-Provisioning	**$39**	**$0**	**$661**	**$0**
DRAM	$40	$40	$80	$40
Controller	$6	$6	$6	$6
Other	$10	$10	$10	$10
Total Drive Cost	**$679**	**$640**	**$1341**	**$640**

Source: https://www.soothsawyer.com/best-online-ssd-cost-calculator

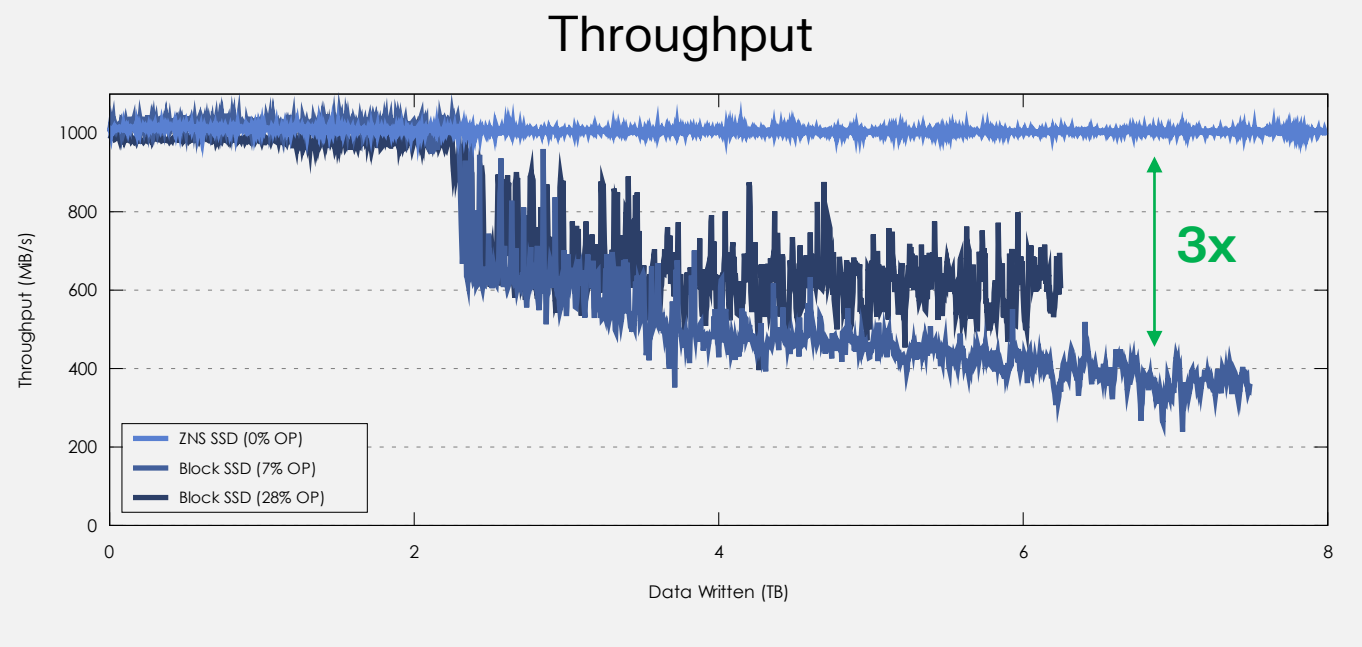
Performance Parity
2x Cost!

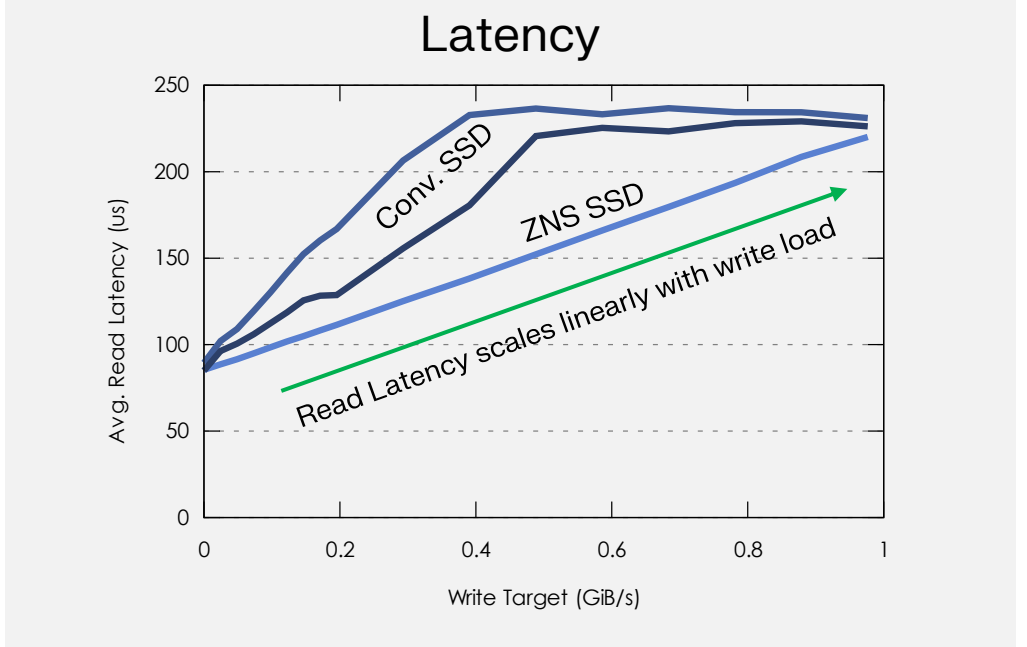
SSDs with Zoned Namespaces (ZNS)?

High DWPD and High Performance

Eliminates SSD's write amplification

ZNS solves the mismatch between the storage block interface and the characteristic of NAND flash
Major impact on performance, lifetime, and behavior of any SSD



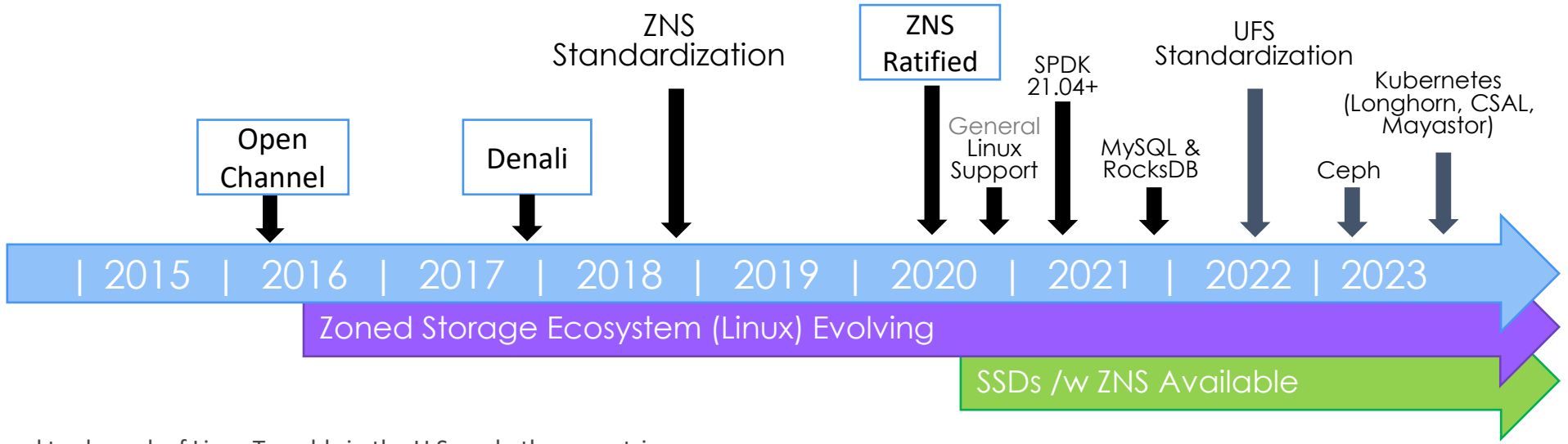


Source: Bjørling et. al., ZNS: Avoiding the Block Interface Tax for Flash-based SSDs, USENIX ATC 2021

Solving the Storage at Scale Challenges

How did we get here?

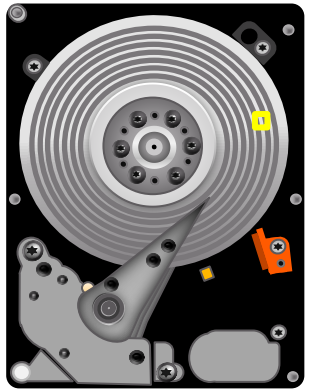
- The NVMe ZNS Group was formed at end 2018 to create the Zoned Namespace Command Set specification.
- The initial specification was ratified June 2020.
- ZNS support added to Linux® software eco-system in June 2020, followed by SPDK support in April 2021.
- SSDs with Zoned Namespace support announced Q3 2020.
- Follow on work in the software eco-system to enable databases, file-systems and cloud use-cases.

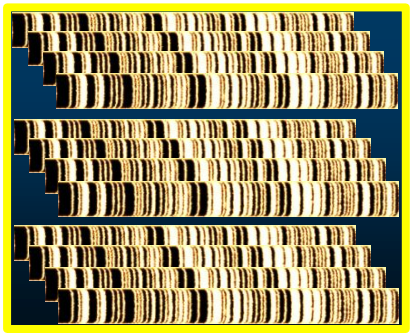


What is a Zoned Namespace?

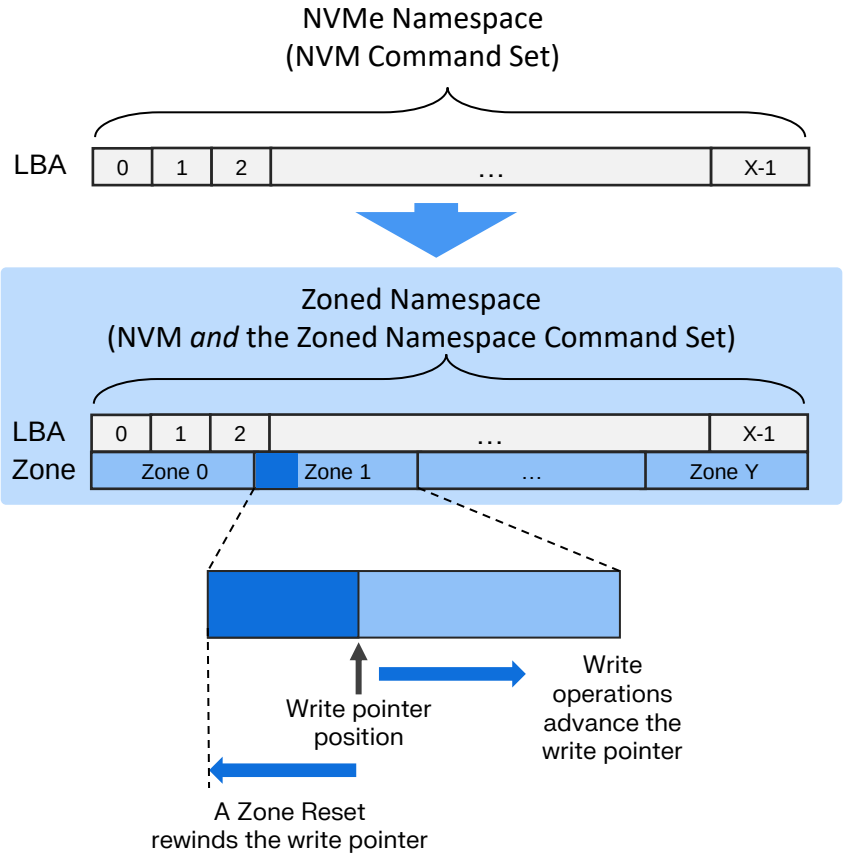
Overview

- An NVMe™ namespace that adds the abstraction of zones
- Logical blocks are divided into fixed-sized zones, which are then utilized for data placement by the host software
- Devices can simultaneously support both conventional and zoned namespaces
- Mimics the ZAC/ZBC models for host-managed SMR HDDs to take advantage of its existing software ecosystem





Raw SMR HDD and NAND Media Both Require Sequential Write Within Zones



Linux Eco-System

Development since 2016

- Zoned API available since kernel version 4.10 (Feb 2017)
- ZNS support added in kernel version 5.9 (Oct 2020)

5+ Linux Distributions with Zoned Storage Support

- RHEL 9+, CentOS 7+, Fedora 33+, Debian 11+, and Ubuntu 21.04+

Local File-systems

- f2fs (client - UFS) and btrfs (enterprise - ZNS/SMR)

Storage Systems

- Ceph, OpenEBS, Mayastor, SPDK's CSAL, …

Library/Tools support

- libzbd, libnvme, SPDK, fio, qemu, blkzone, blktests, …

End-to-end Application Enablements

- Cloud Orchestration Platforms
- Databases, Databases, Caching

Mature, robust, and used in production by some of the biggest consumers of storage

Cloud Orchestration: kubernetes, openstack.

Distributed Storage: ceph, OpenEBS, CSAL

Databases: MySQL, TERARK, RocksDB

Caching: acheLib

Local File-Systems: f2fs, btrfs

Libraries: libzbd, libnvme, SPDK

Tools: fio, …, qemu

Distributions: redhat, ubuntu, SUSE, debian

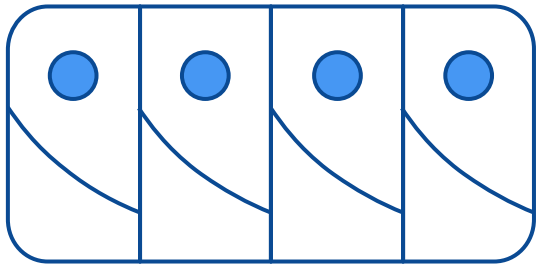
Zoned Storage Hardware

Deployments

Typical Approaches

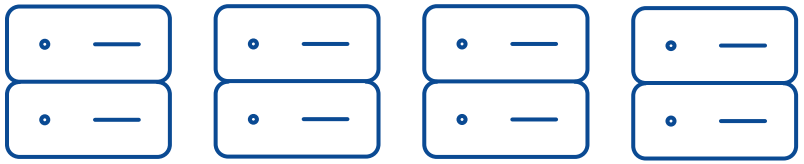
Storage Array

Great Scalability and Capacity Utilization



Storage Array /w ZNS SSDs

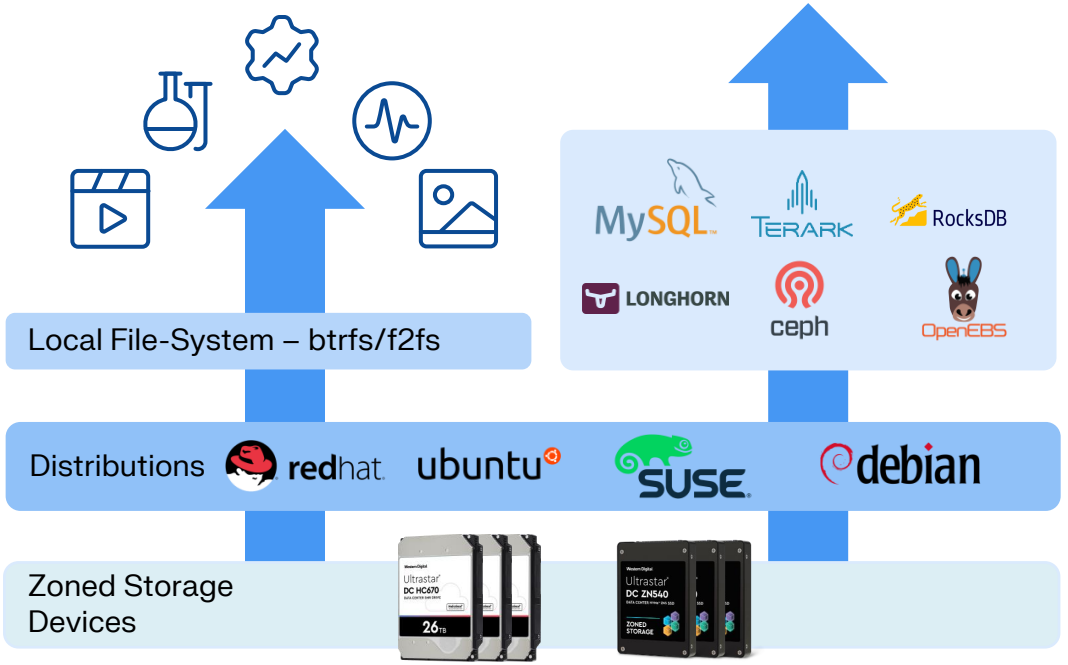
NVMe over Fabrics
(Conventional and/or ZNS)



Local Storage

Any Application
(Great Performance)

End-to-End
(Highest Performance)



Storage Array

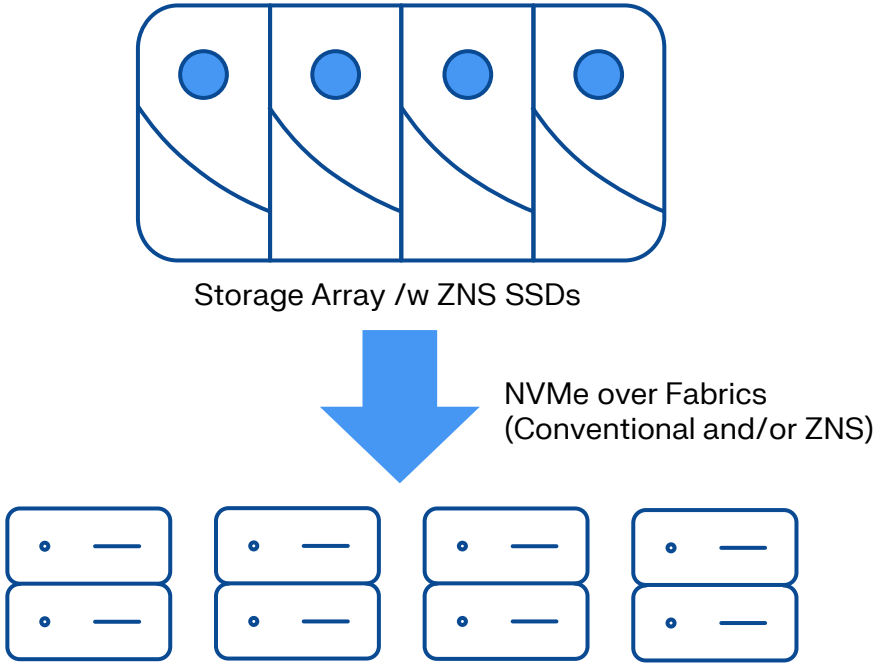
For Performance and Capacity

SPDK

- Storage box, such as an All-Flash Array (AFA)
  - Storage is accessed through a common network protocol such as NVMeoF, NFS, SMB, …
  - The storage box runs software that supports zoned storage and exposes it as conventional storage
- Use-Cases
  - Performance: Very high-performance storage system for AI/ML, streaming and databases
  - Capacity: Replace HDDs with QLC SSDs with a DWPD > 1
    - Example: Alibaba replaced HDDs with QLC SSDs in their 3rd generation big data local disk ECS instances to double the performance and density vs. 2nd generation while holding the price to their customers constant.

Storage Array

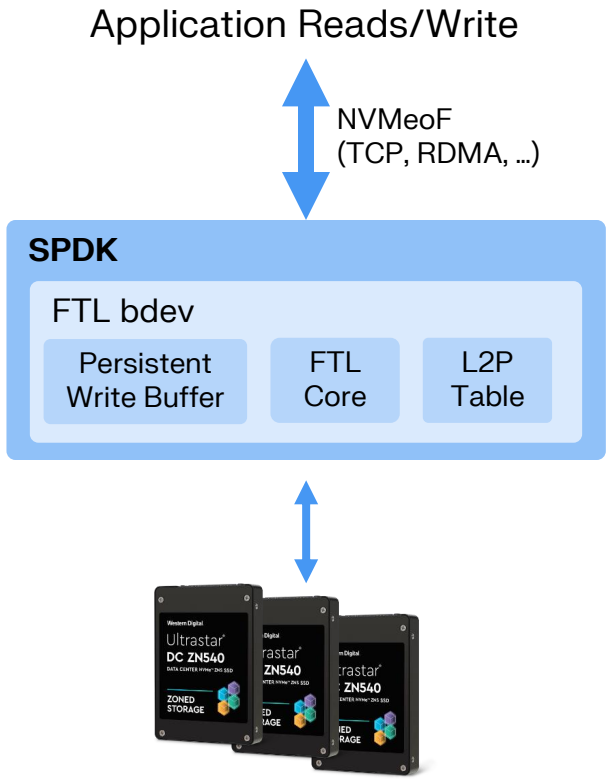
Great Scalability and Capacity Utilization



Turnkey Storage Array

CSAL together with Reference Storage Platform (RSPs)

- What is CSAL?
  - Open-source cloud-scale shared-nothing Flash Translation Layer (FTL bdev) in Storage Performance Development Kit (SPDK)
  - Ultra fast cache and write shaping tier to improve performance and endurance to scale QLC value
  - Flexible scaling of NAND performance and capacity to the user/workload needs
- Used and deployed by Alibaba to adopt QLC SSDs into their data centers*
- Reference Storage Platforms partners. Turnkey solution that quick and easy deployment
- WD collaboration with Solidigm and the CSAL community for broadening its adoption
  - ZNS support being upstreamed. Available upon request



RSP Partners

StarWind
Intel
Solidigm
SPDK
Western Digital

*https://www.intel.com/content/www/us/en/content-details/765062/a-media-aware-cloud-storage-acceleration-layer-csal-cache-solution-with-intel-optane-ssds-for-alibaba-ecs-local-disk-d3c-service.html

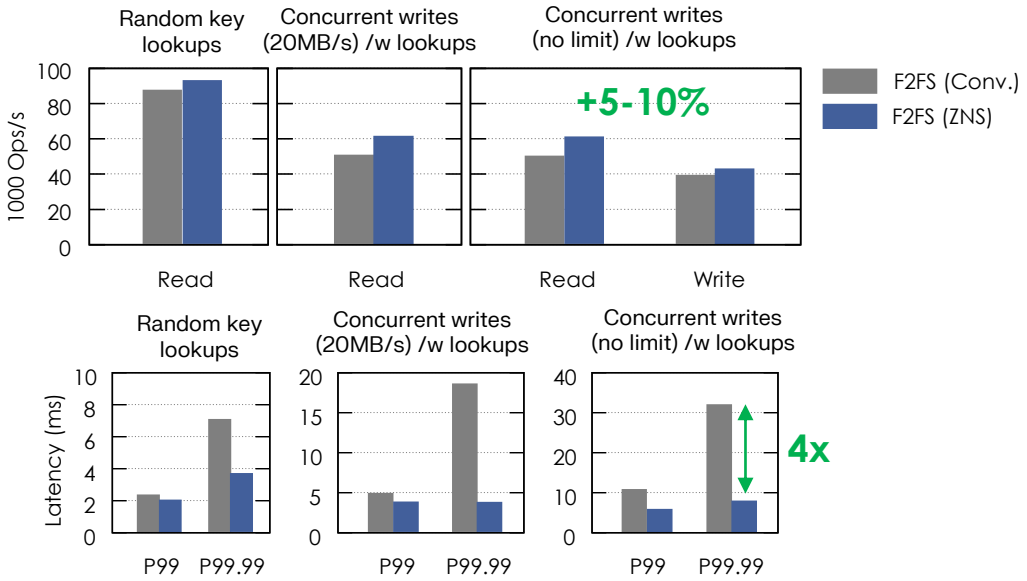
Conventional Storage for Any Applications

Local Storage

- Conventional storage is often required by the end-user
  - Incremental roll out of software that specifically takes advantage of zoned storage, or;
  - Applications that are not performance sensitive
- Conventional storage access through local file systems with zoned storage support
  - **No changes necessary** to applications as zoned storage support is baked into local Linux file systems
  - f2fs – Linux kernel 5.10+
  - btrfs – Linux kernel 5.12+
- Outperforms conventional SSDs
  - No OP – **7%/28% additional storage**
  - Better performance than conventional
  - Works **natively** with hint-based placement (e.g., streams)
  - **No software modifications required**

RocksDB

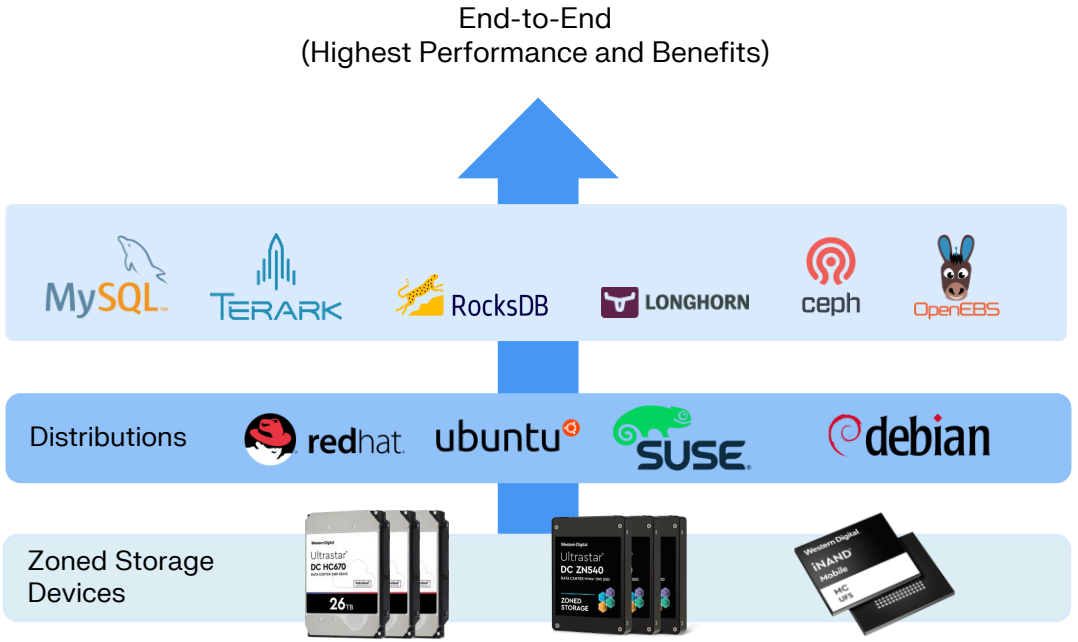
Key-Value Store on top of local file-system



End-to-End Application Integrations

Local Storage

- Highest performance and optimized per use-case
  - I/O intensive applications
  - Large-scale storage systems that uses local storage
- Applications are tightly integrated with the storage stack
  - Aware of the underlying storage type and performs intelligent data placement
  - Typical candidates are distributed file-systens, databases, caching
- Integrations:
  - Databases: **Percona MySQL**, RocksDB, TerarkDB, CacheLib
  - File-Systems
    - **Ceph**, f2fs, btrfs
  - **Cloud Integrations**
    - Longhorn, Mayastor/OpenEBS, CSAL



Percona MySQL

Improving OLTP workloads

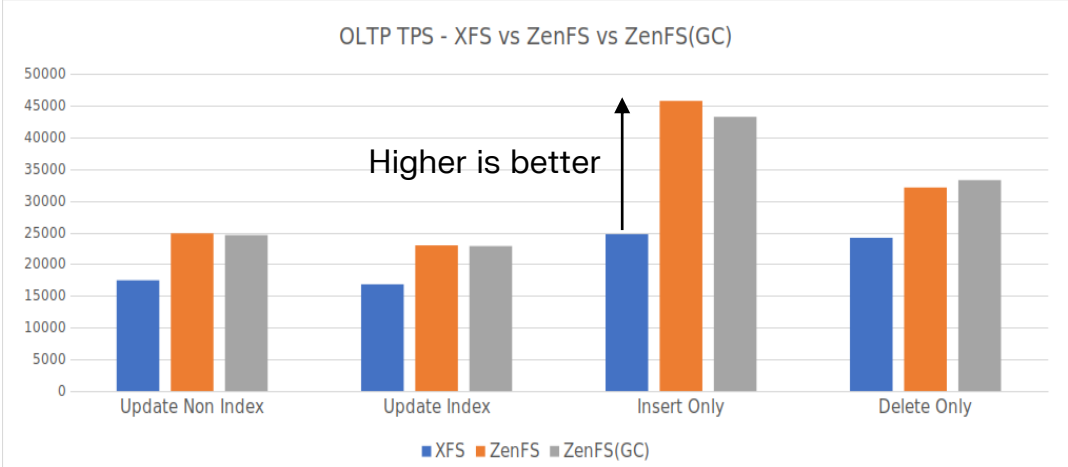
Benchmark: Write-heavy OLTP workloads using sysbench

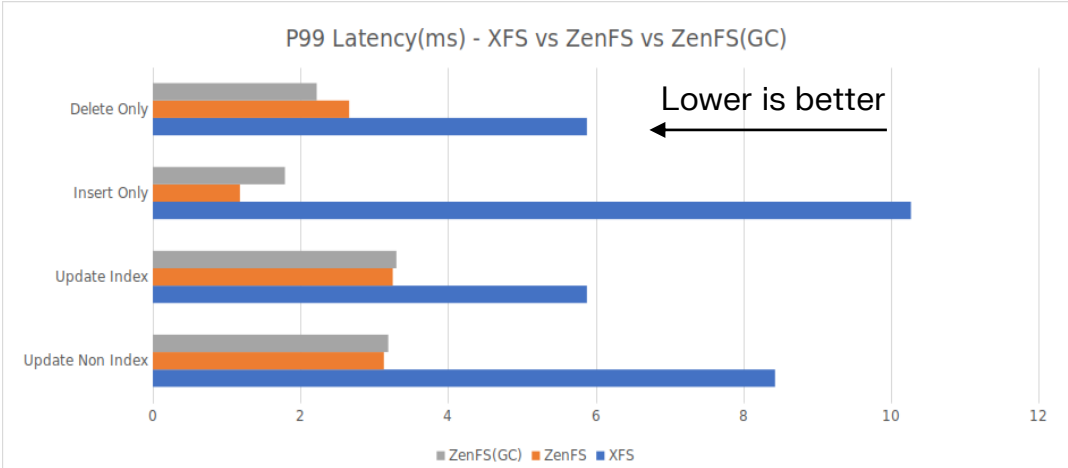
XFS: 1TB SN540 (Conventional Namespace)

ZenFS: 1TB ZN540 (Zoned Namespace)

ZenFS (GC): ZenFS with Garbage collection enabled

- Zoned storage support integrated into RocksDB, which Percona MySQL can use as its storage backend
  - Developed the ZenFS storage backend, and added support upstream
  - Collaboration with Percona
    - Support upstream in the general cloud contained used for cloud deployments
- When using ZenFS:
  - Up to 80% Higher throughput
  - Lower tail latency by up to 10x

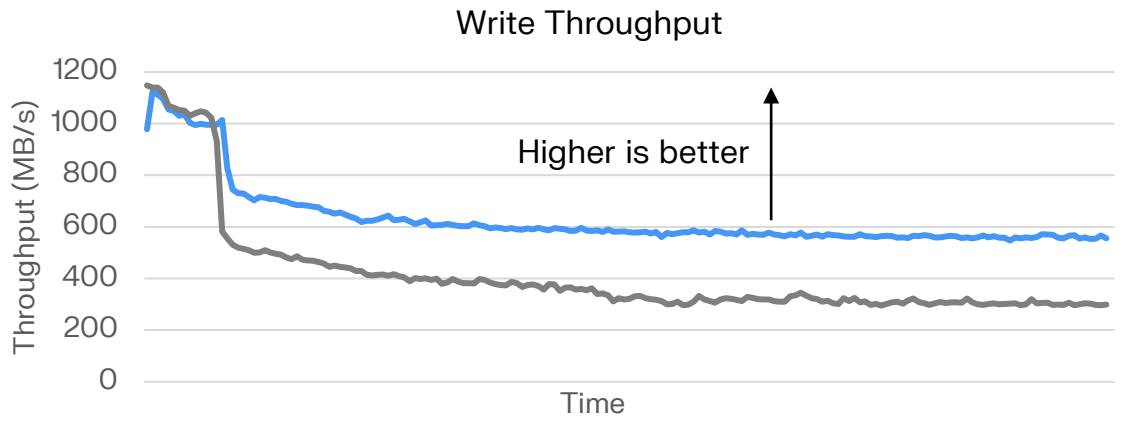


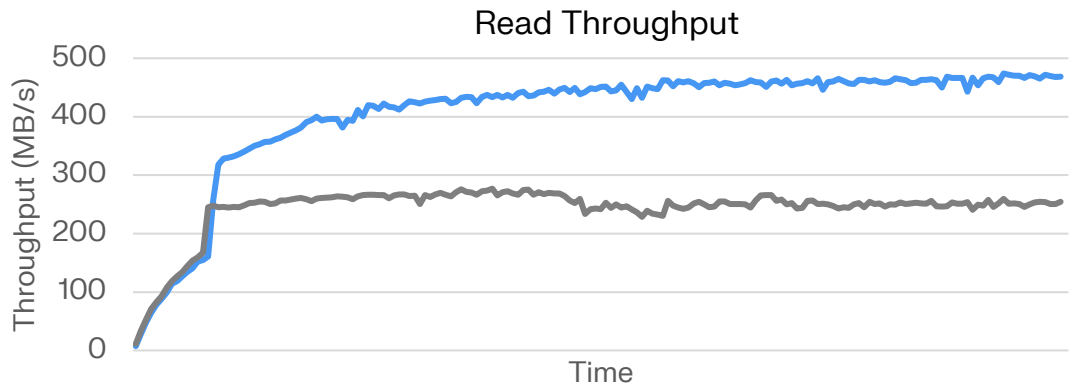


Ceph Crimson

Using Zoned Storage

- Very popular distributed file-system used in the industry and HPC space
  - CERN, Intel, Google, Blizzard, …
  - At least 1.1EiB deployed (opt-in telemetry)
- Next release (Crimson) will have native support for zoned storage
  - Both applicable to SMR HDDs as well as ZNS SSD.
- 30% higher throughput for both reads and writes once garbage collection kicks in



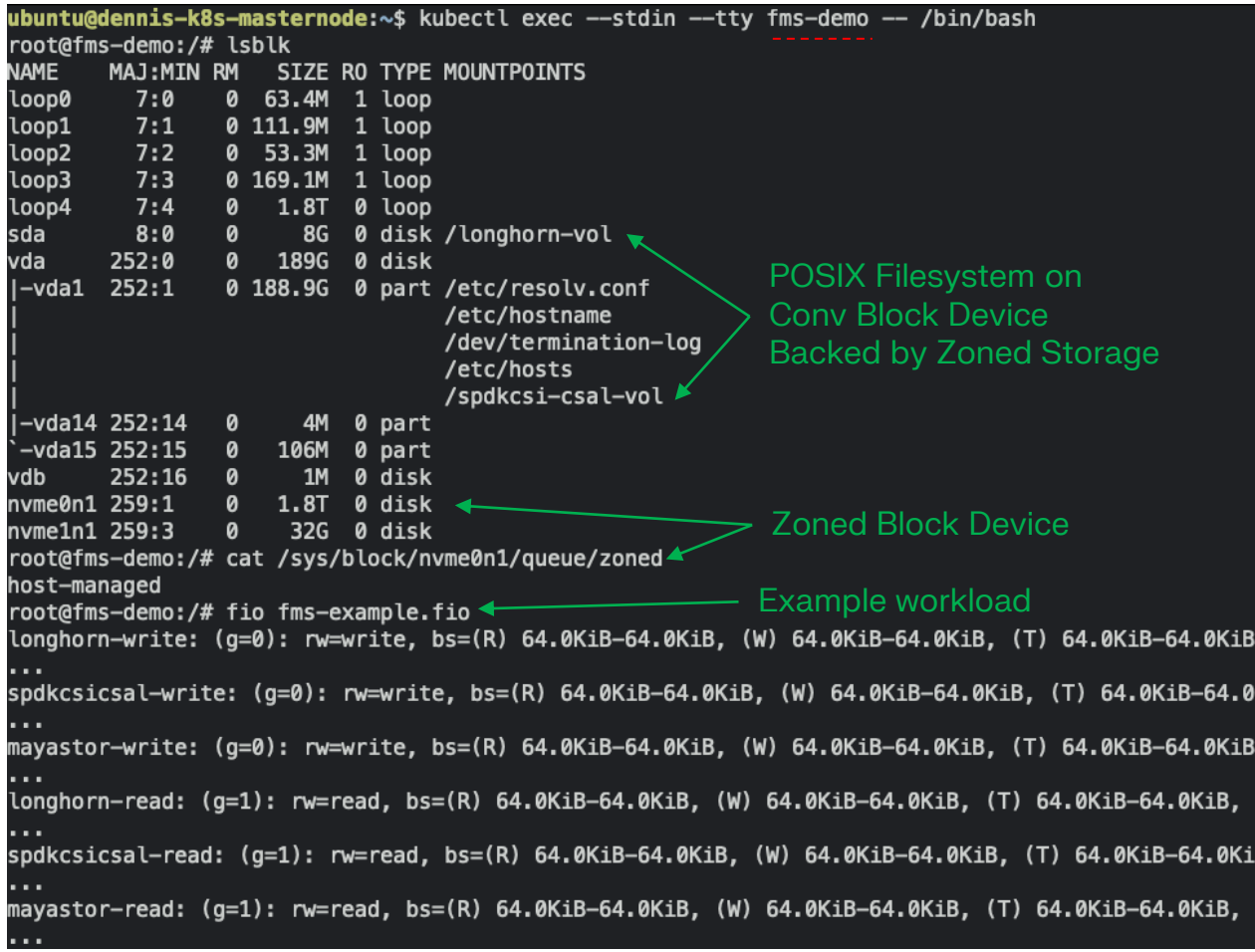


Rados block device, fio 80/20% RW Workload

Cloud Integrations

Transparent use of Zoned Storage in Cloud Applications

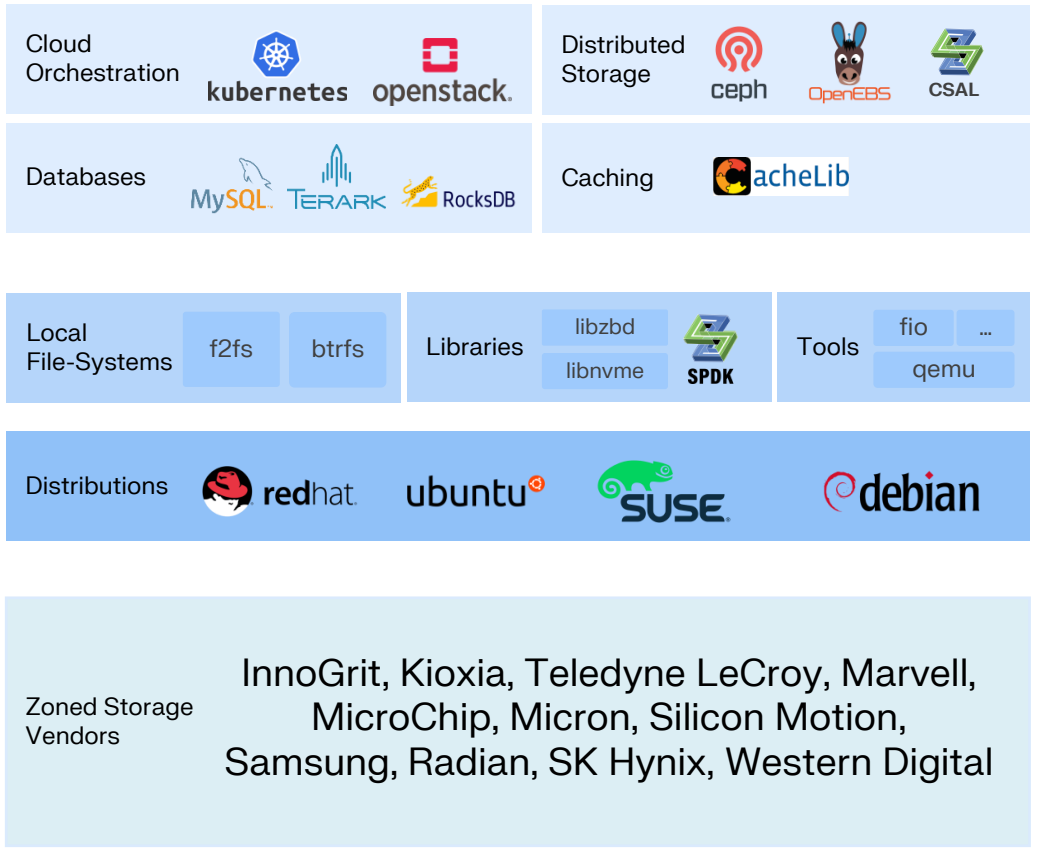
- Kubernetes is the main cloud orchestration platform for deploying applications
- We made zoned storage natively integrated without any end-user modifications
  - Exposes conventional storage to containers/VMs
  - Integrated into Longhorn, OpenEBS, and CSAL



Eco-System

Growing Number of Vendors and Use-Cases

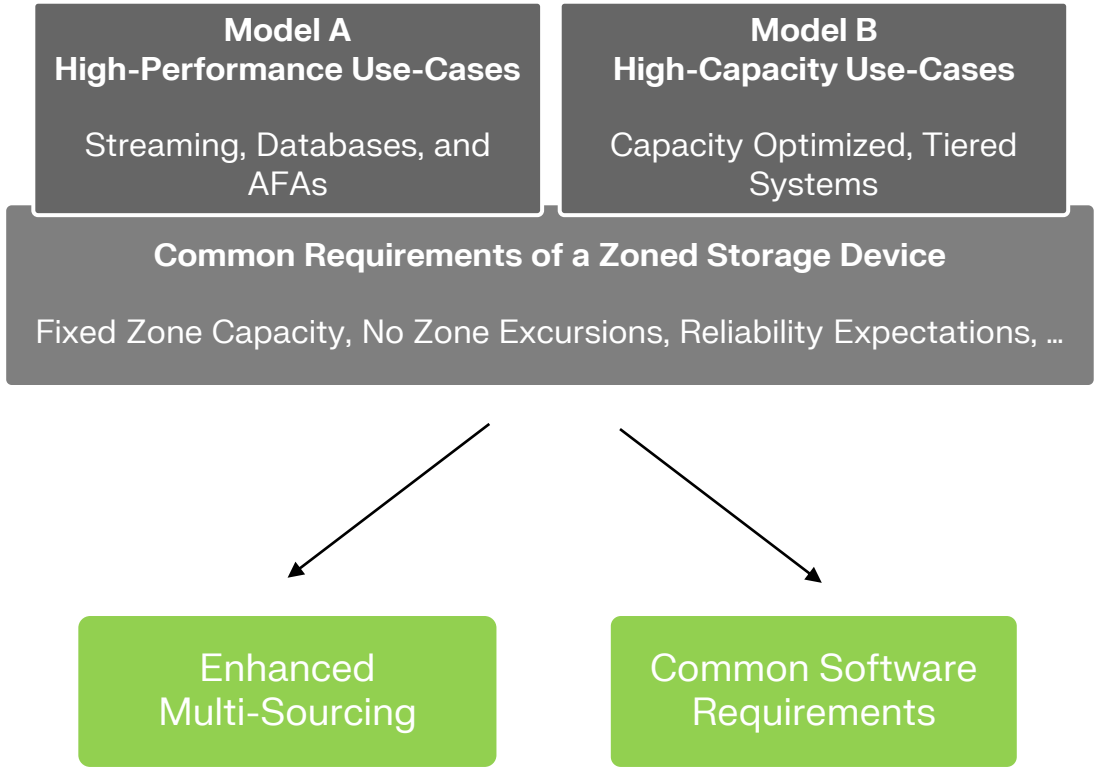
- Zoned Namespace Command Set support has been announced or added support into products across a broad set of vendors
- Solid support in the Linux software eco-system. Built on the existing foundation of SMR HDDs. Enabling rapid support and development.
  - Major achievements include local file-systems and relational and key-value database systems
- While broad industry support has been achieved, successful large-scale deployment of ZNS SSDs also require
  1. Multi-sourcing through standardized device models and reference platforms
  2. Large-Scale Deployments through cloud orchestration platforms as well as distributed file-systems



Standardized Device Models

SNIA Zoned Storage Technical Workgroup

- While there is a thriving eco-system for zoned storage, it depends on SSDs, in addition to conventional attributes, to implement the Zoned Namespace Command Set specification similarly
- SSD vendors initially developed ZNS SSDs with somewhat different attributes, leading to confusion at adopters on what software changes were needed to support zoned storage
- To unify industry offerings, improve multi-sourcing, and grow software interoperability, a set of SNIA organization members formed the Zoned Storage Technical Workgroup, which recently released the Zoned Storage Models v1.0 specification
  - Common requirements for zoned storage devices, aiding common software development
  - Two common device models, that each inherit the common requirements

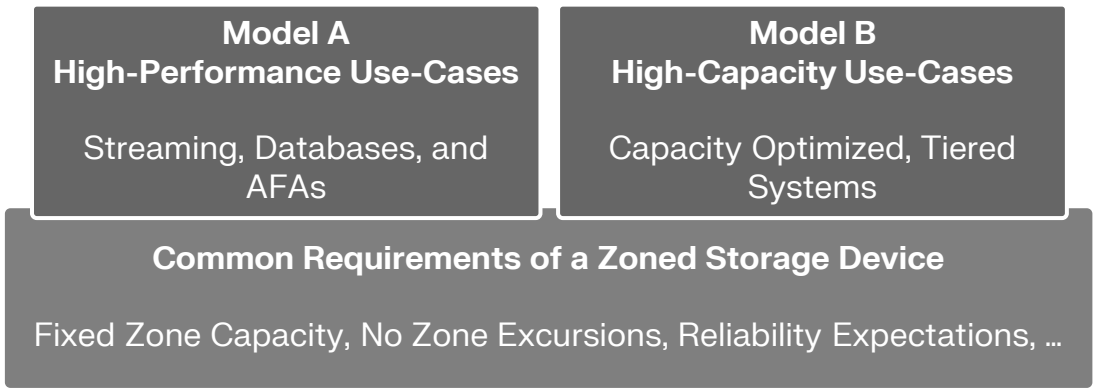


Source: SNIA Zoned Storage Models Version 1.0, July 2023, https://www.snia.org/standards/technology-standards-software/standards-portfolio/zoned-storage-models

Zoned Storage Device Models

SNIA Zoned Storage Technical Workgroup

- Standardized document that describes the common requirements that host software can expect a zoned storage device to support
- Defined how a zoned device is expected to behave towards the host. This included stating
  - A device always manages reliability (e.g., wear-leveling)
  - A zone's writeable capacity is constant and not variable
  - No Active Zone Excursions. I.e., zones are guaranteed to always have its writeable capacity available
  - How a device behaves end-of-life. Read-only mode, etc.
- Devices that these definition into account are then able to reap the benefits of the existing zoned storage software eco-system and be sure that there is a common understanding of the device's behavior.

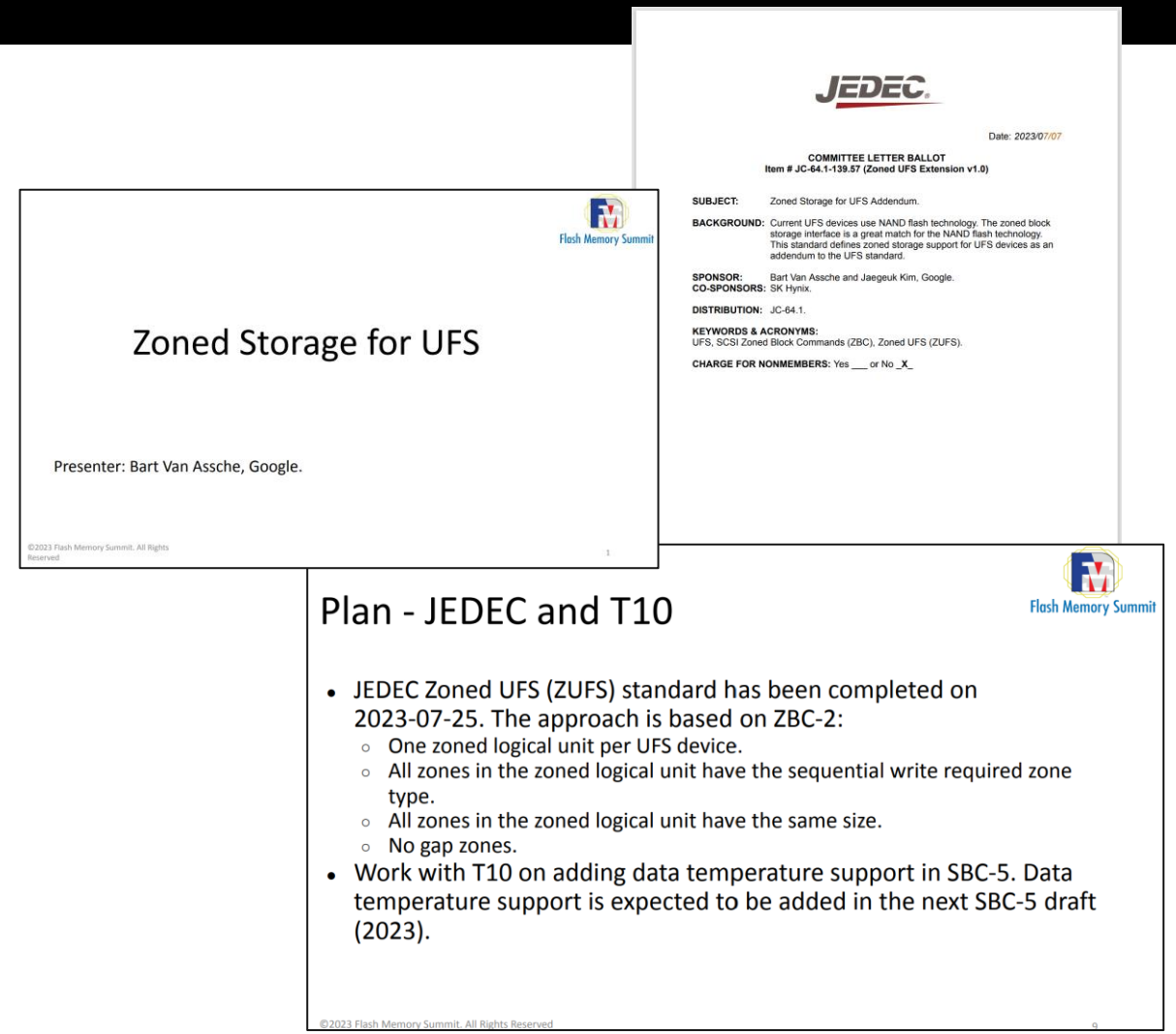


Source: SNIA Zoned Storage Models Version 1.0, July 2023, https://www.snia.org/standards/technology-standards-software/standards-portfolio/zoned-storage-models

Zoned Storage for Embedded Devices

Google Pushing Zoned Storage to Mobile

- Towards large-scale deployments in mobile
  - Driven by Google and targeted the Android hardware eco-system
  - Bart Van Assche from Google discussed their roadmap for adopting Zoned UFS into Android at Flash Memory Summit ‘23
  - JEDEC’s Zoned UFS specification completed in July
  - Android vendors targeting productization in their next-generation mobile products
- With Zoned UFS in place, the zoned storage interface is now ubiquitous across all major storage devices (HDDs, SSDs, Embedded/Mobile)
  - All utilized the same software eco-system
  - It is expected that adoption of Zoned UFS is quick as file-system support (f2fs) is completed and Android already switched to f2fs for many mobile platforms (e.g., Pixel).



Source: Bart Van Assche, Google, “Zoned Storage for UFS”, Flash Memory Summit ‘23

Summary

- SSDs with Zoned Namespace support enables hyperscalers and CSPs to
  - Meet their increasing customer demand for cost-effective storage as well as high performance
  - Extend the lifetime of their server fleet beyond their typical five-year service time
- Mature eco-system that enables a broad range of use-cases
  - Storage Arrays Turnkey Solution – ZNS + QLC
  - No changes necessary to conventional applications to take advantage of SSDs with Zoned Namespace support
  - High-performant end-to-end integrations benefiting distributed storage systems, databases and caching applications
- Good vendor eco-system
  - Standardized SNIA zoned storage device models
  - Broad set of vendors
  - Path to large-scale deployment into embedded/mobile