Flash Memory Summit

ZNS SSDs
Achieving Large-Scale Deployment

Matias Bjørling

Distinguished Engineer & Senior Director of the Emerging System Architectures Group

Western Digital

Storage at Scale

Flash Memory Summit

Hyperscalers and Cloud Service Provides (CSPs) are constantly challenged with large volumes of data and increasing customer demand for cost-effective storage and high performance

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

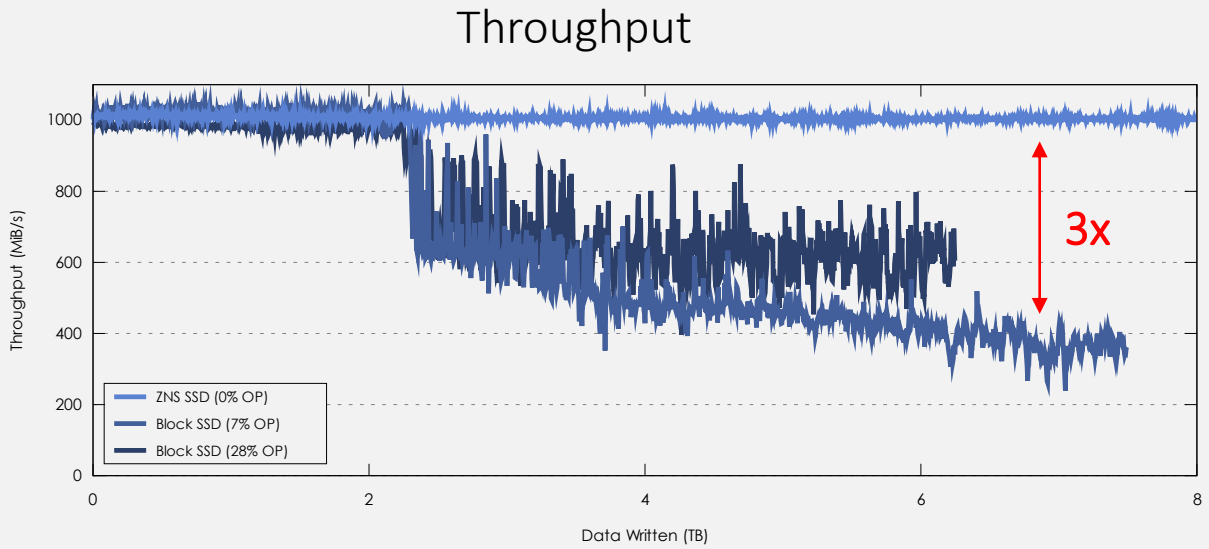
Metric	Conventional SSD		Zoned Namespace SSD	
	TLC	QLC	TLC (Performance)	QLC (Capacity)
IOPS/TB	++	+	+++	++
Throughput	++ (Read/Write)	+ (Read)	+++ (Read/Write)	++ (Read)
Latency/QoS	++	+	+++	++
Lifetime	++	+	+++ (Typ. >3.5 DWPD)	++ (Typ. >1 DWPD)
Cost (TB/$)	+	++	++	+++

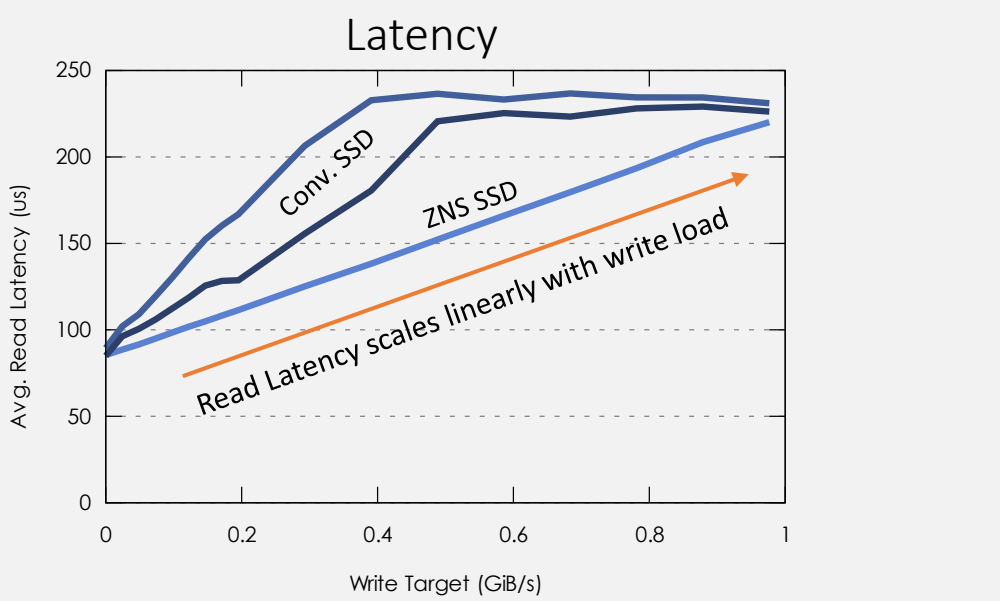
* Lee Prewitt, Microsoft - How Facebook & Microsoft Leverage NVMe Cloud Storage. https://www.brighttalk.com/webcast/663/374596

Why SSDs with Zoned Namespaces (ZNS)?

High and Consistent Performance

- Mismatch between the storage block interface and the inherent characteristic of NAND flash
- Eliminates SSD's garbage collection (GC) and write amplification - Host writes mixed onto the same media, increases GC burden
  - **Major impact on performance, lifetime, and behavior of any SSD**





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

Why SSDs with Zoned Namespaces (ZNS)?

Performance is Expensive

Flash Memory Summit

"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**

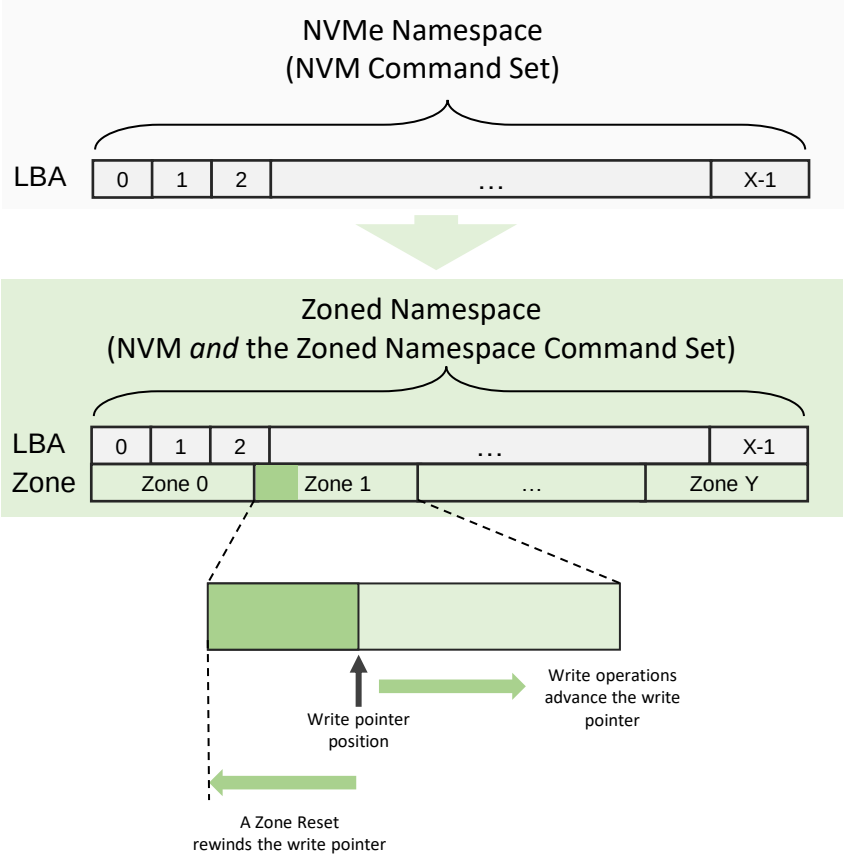
Source: https://www.soothsawyer.com/best-online-ssd-cost-calculator

Performance Parity
2x Cost!

What is a Zoned Namespace?

Overview

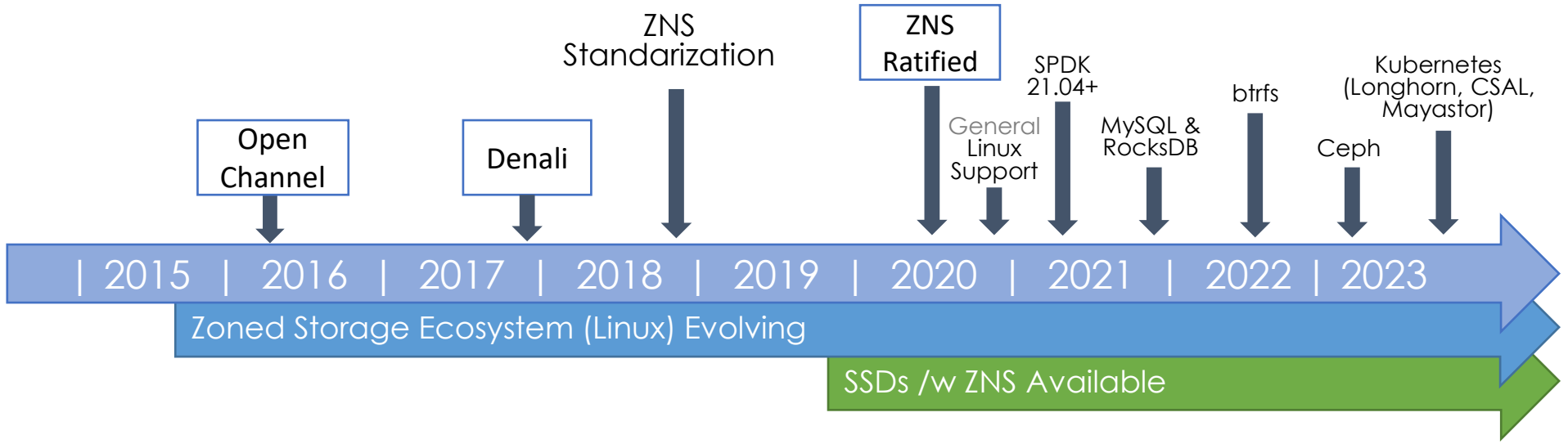
- An NVMe™ namespace that supports the abstraction of zones
  - Inherits the existing concepts of logical blocks, LBAs, I/O Commands (e.g., Read and Write commands), Admin Commands, Log Pages, … from the NVM Command Set
  - Logical blocks are divided into fixed-sized zones, which are then utilized for data placement by the host software
- Mimics the ZAC/ZBC models for host-managed SMR HDDs to take advantage of its existing software ecosystem
- NVMe devices can simultaneously support both conventional and zoned namespaces
  - E.g., useful for soft roll-outs as software is updated to take advantage of the zoned storage model



Zoned Namespaces Standardization

History

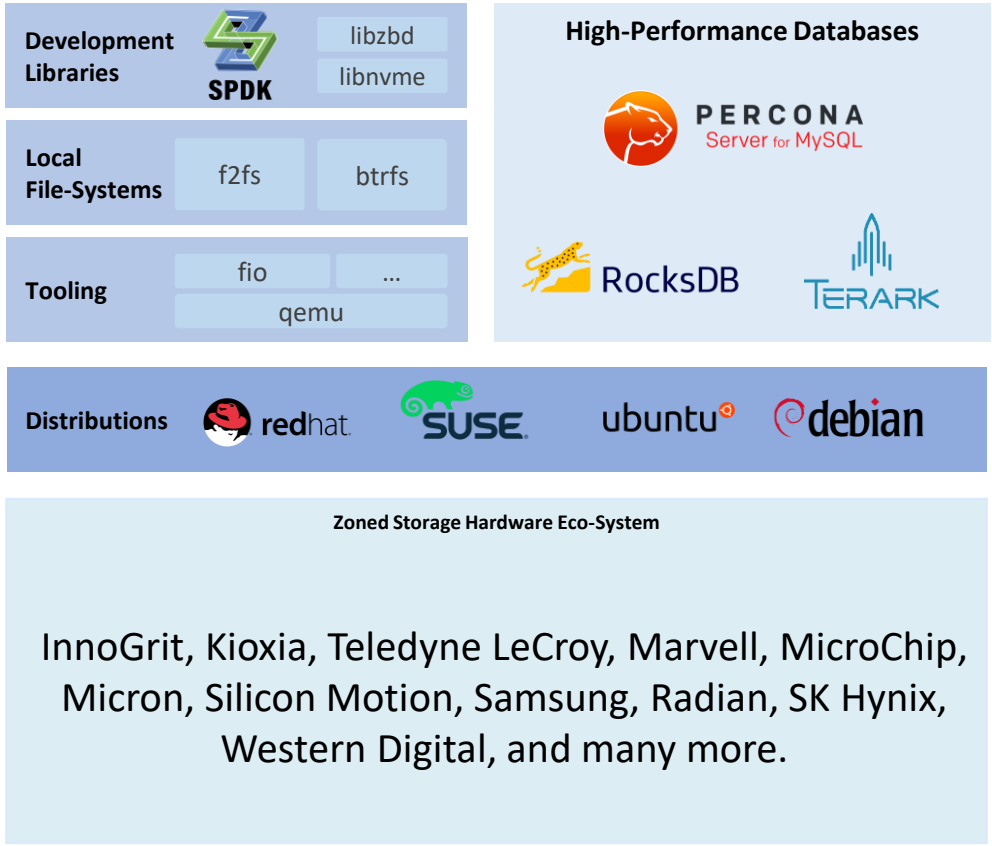
- Broad industry need for a standardized approach to direct data placement aligned to SSD's media characteristics
- ZNS Task Group was formed to work on what became the Zoned Namespace Command Set
    - TP work began late 2018 and initial revision was ratified June 2020.
- ZNS support was added to Linux® software eco-system in June 2020, quickly followed by SPDK support in April 2021.
- SSDs with Zoned Namespace support announced shortly after ratification
- UFS - Standardization update
    - Driven by Google, Zoned Storage support is also being added to the UFS specification for use in mobile devices
    - A single storage model across hard-drives, solid state drives, and embedded storage devices!



Eco-System Update

Growing Number of Vendors and Use-Cases

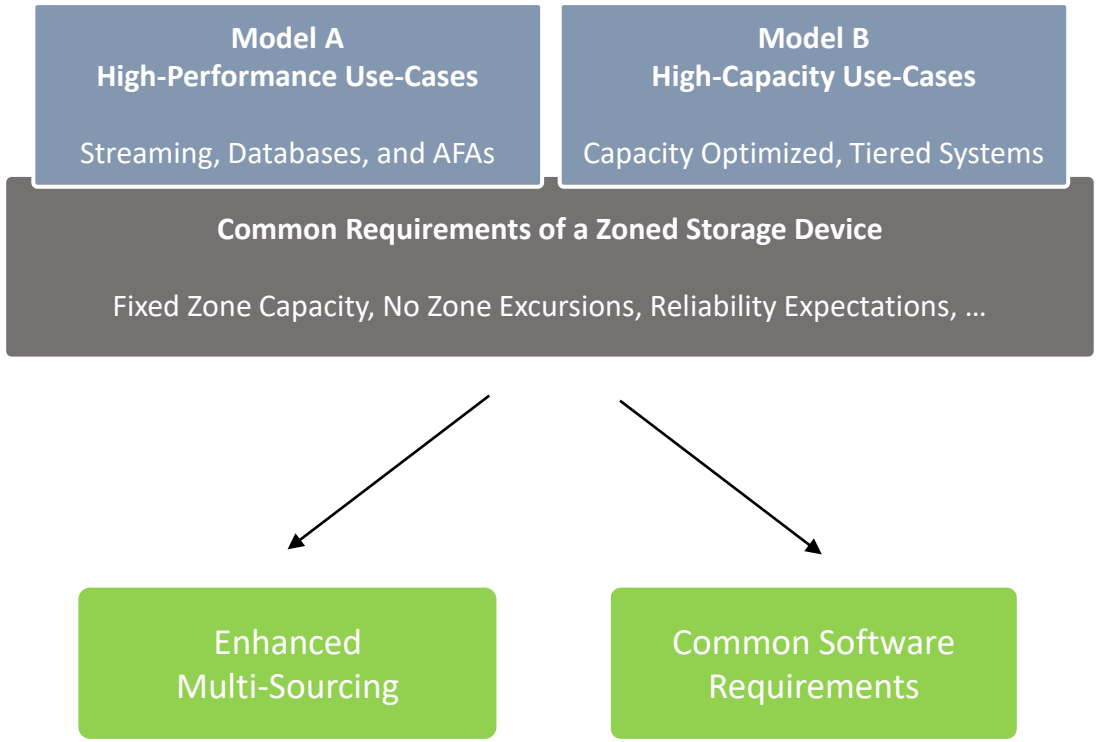
- Zoned Namespace Command Set support has been announced or added support into products across a broad set of vendors
- Solid support in the Linux software eco-system. Built on the existing foundation of SMR HDDs. Enabling rapid support and development.
  - Major achievements include local file-systems and relational and key-value database systems
  - Raid & ZNS using RAINZ and Flexible ZNS
- 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 take advantage of 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

Large-Scale Deployments

How to make Zoned Storage easy to deploy?

- Tight integration with Storage As A Service (SAAS) providers
  - Cloud Orchestration Platform support. Typically, Kubernetes and/or OpenStack
  - Integrating natively into the distributed storage stacks. No longer needs end-user application optimizations
- SAAS is often provided by CSP (Azure, AWS, GCP) which implements internal solutions to for their specialized deployments
  - But how to get the benefits of zoned storage with on-premise clouds and/or hybrid clouds?

Cloud Orchestration

kubernetes

openstack

Distributed Storage

LONGHORN

OpenEBS

CSAL

ceph

Two Tiers of Zoned Cloud Storage

Capacity

- Cost, cost, and … cost
- Enable Cloud applications to use a scalable storage solution that scales to Exabytes
- Integrates with a tiered platform. Fast storage followed by archival storage
- User applications “do not see” the zoned storage

Performance

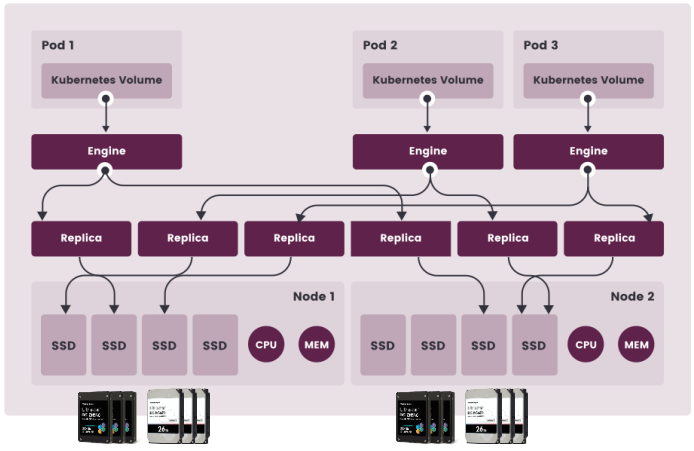
- Throughput and Latency are king
- Cost is secondary
- Allow Cloud applications to integrate tightly with the underlying storage hardware
- E.g., ephemeral (local) storage or direct access to the zoned storage interface
- User applications are explicitly optimized for the zoned storage model

Broad Enablement

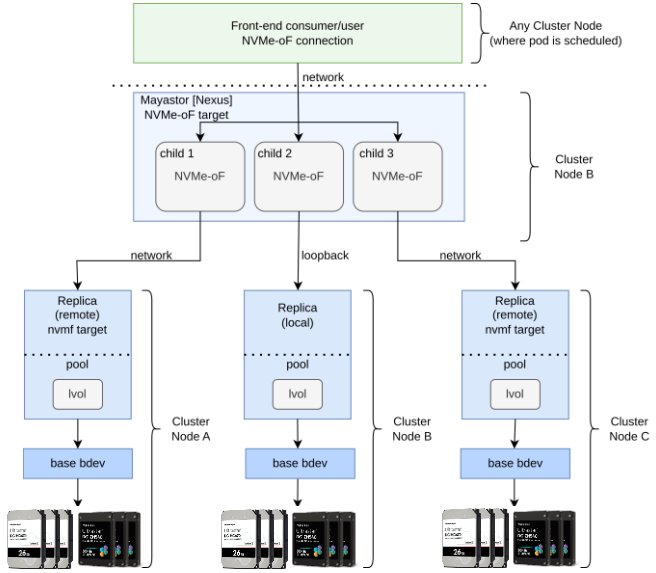
Modern distributed storage system support

- Longhorn – Built on top of btrfs with support being planned for their upcoming SPDK backend
- Mayastor – Native support for exposing zoned storage to containers
- SPDK's CSAL – High-performance storage (storage array) that exposes zoned storage as conventional storage over NVMe over Fabrics

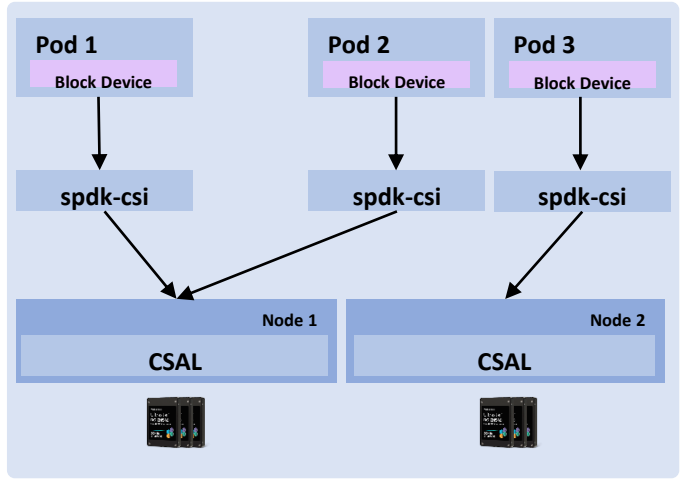
Longhorn



Mayastor



SPDK's CSAL

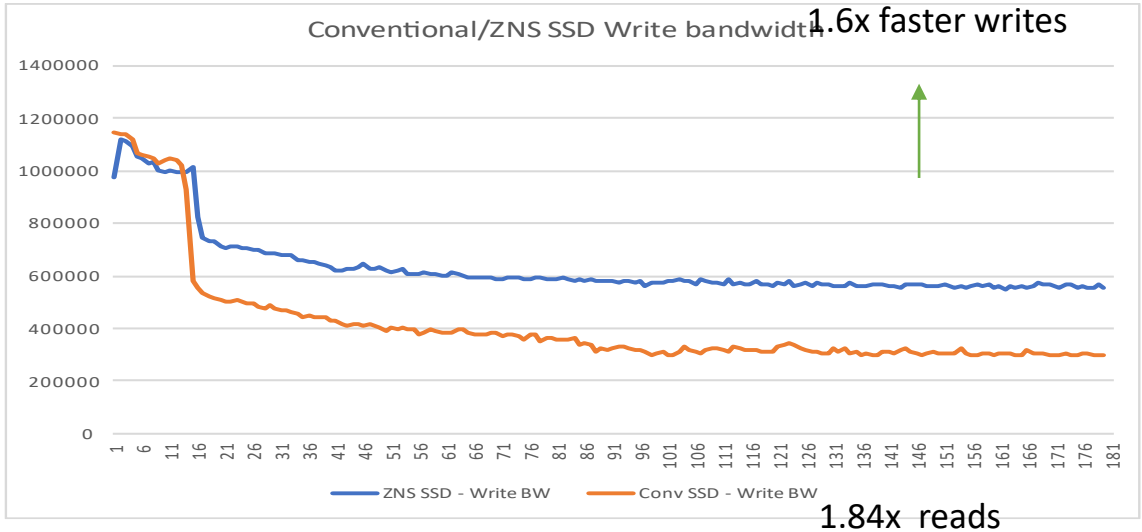


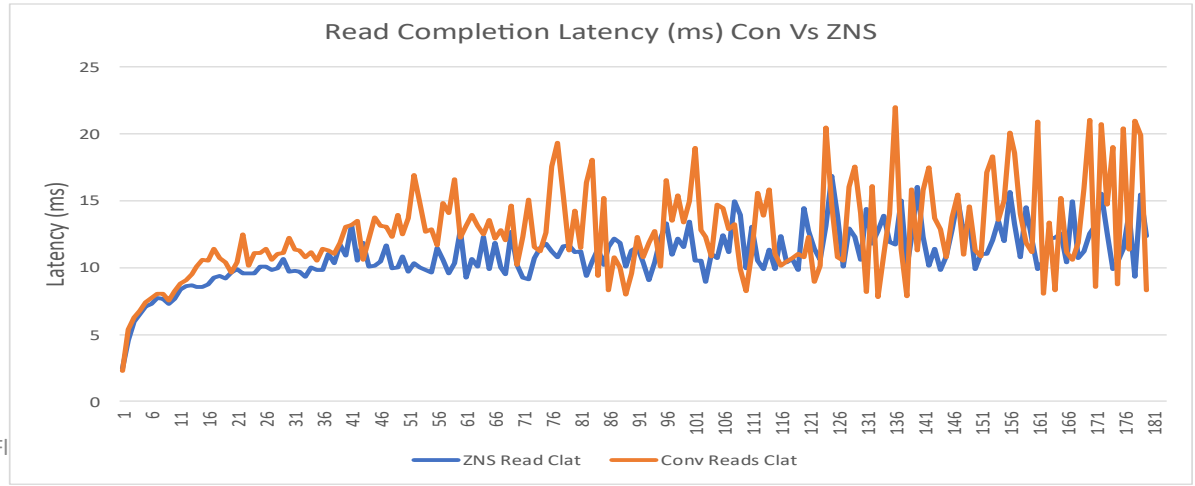
Ceph – Distributed File-System

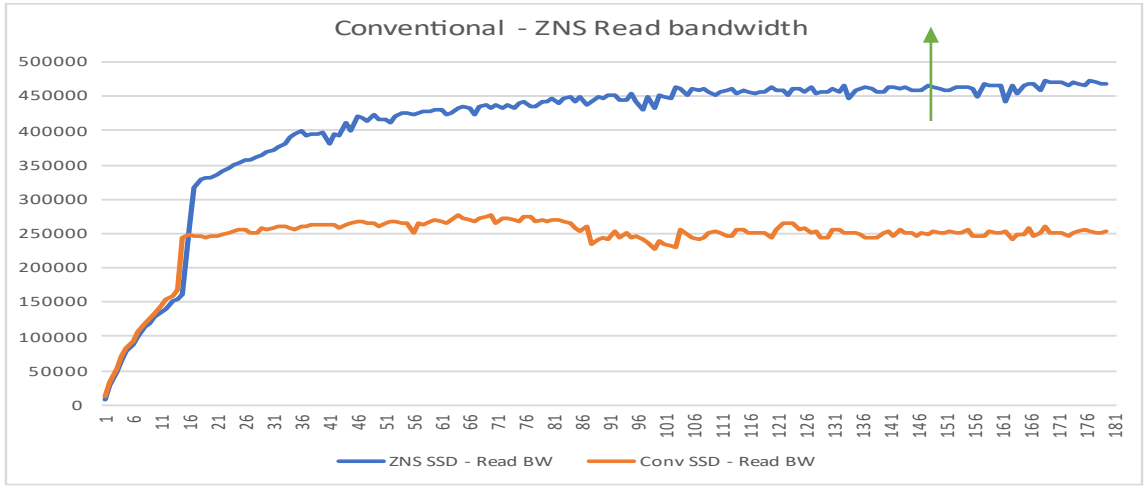
Crimson Update

- Crimson is the next release of Ceph
  - Zoned storage integrated with the upcoming Seastore backend
- Native support for zoned storage and works out of the box!
- 1.6x faster writes, as well as significant latency QoS improvements
- Expected production release 1H 2024

80/20% Read/Write workload







Conclusion

Flash Memory Summit

- SSDs /w for Zoned Namespaces
  - High Performance – High throughput and predictable latency QoS
  - Cost – Capacity-enabler. High DWPD with QLC enables a multitude of new workloads
  - Lifetime/DWPD: 5-7 years lifetime
- Well-defined storage device models to ease industry adoption
- Software use-cases. Support for both end-to-end use-cases as well as general storage consumers.
- Developers, Developers, Developers…!

More information available a
https://zonedstorage.io

- This track
  - Flexible ZNS Configurations for Optimizing QLC-Based Applications
    David Wang, Silicon Motion
  - Zoned Storage for UFS on Smartphones
    Bart Van Assche, Google
- Cloud Technologies
  - Super-Charging Cloud Native Storage with Zones
    Dennis Maisenbacher, WDC
- Cloud Scaling
  Tuesday 9:45-10:50
  - Software-Hardware Co-Design for High Performance Storage System on ZNS SSD
    Wei Tang, Bytedance
- Open-Source Innovation Track – Part 2
  Wednesday 9:45-10:50
  - SSDFS + ZNS SSD: Deterministic Architecture Decreasing TCO Cost
    Viacheslav Dubeyko
  - ZNS in the Cloud with Ceph Crimson
    Aravind Ramesh, WDC
  - New Developments in Cloud Storage Acceleration Layer (CSAL), an FTL in SPDK
    Kapil Karkra, Solidigm