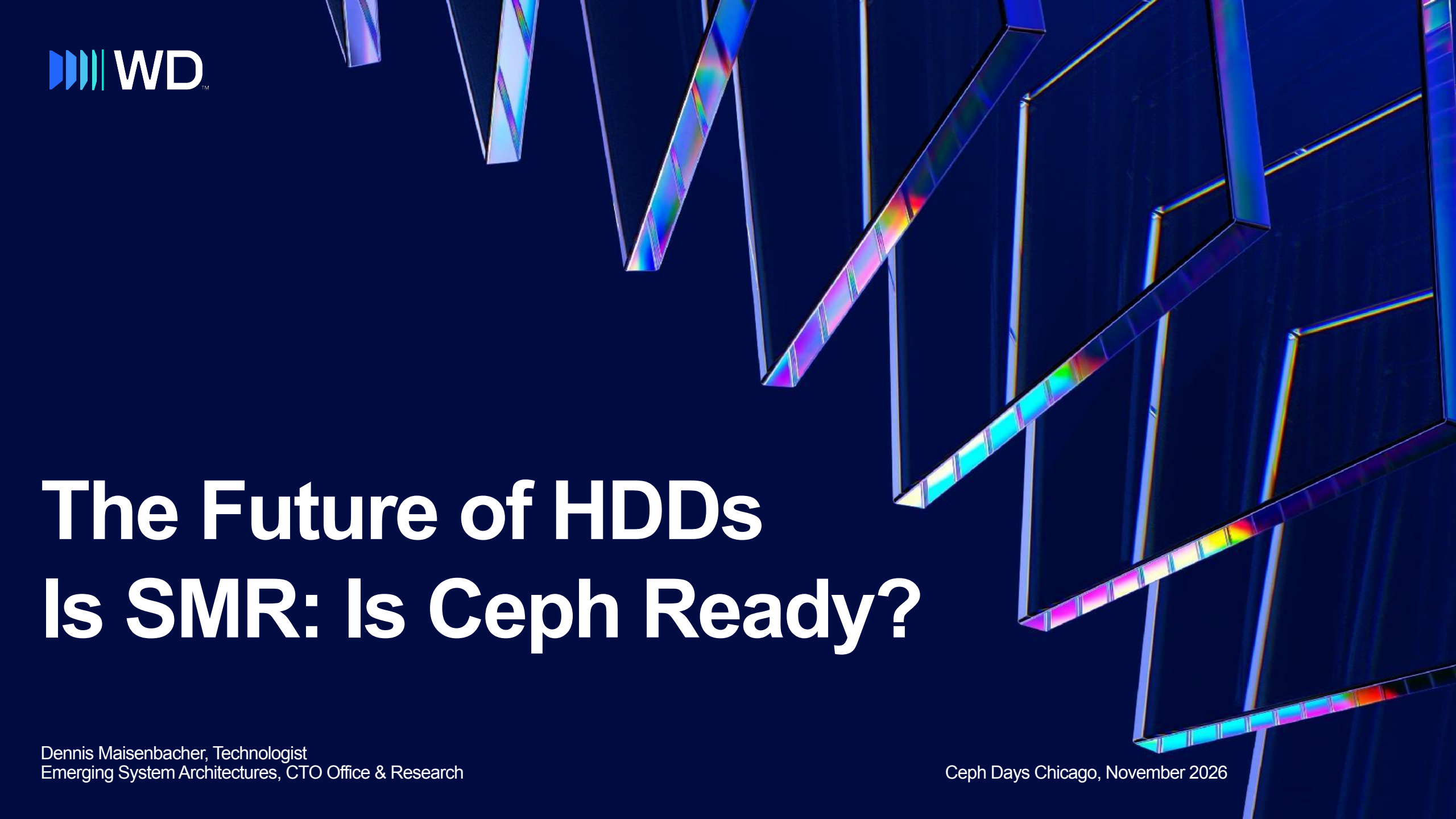


The Future of HDDs Is SMR: Is Ceph Ready?

Dennis Maisenbacher, Technologist
Emerging System Architectures, CTO Office & Research

Ceph Days Chicago, November 2026

Agenda

01	Introduction: HDDs and SMR
02	Zoned BlueStore
03	Experimental Results
04	Conclusion

© 2026 Western Digital Corporation. All rights reserved.

SMR is the Mainstream

60% of WD's exabytes ship as **UltraSMR** by mid-2027

Source: WD CQ2 '26 earnings call, August 2026.

© 2026 Western Digital Corporation. All rights reserved.

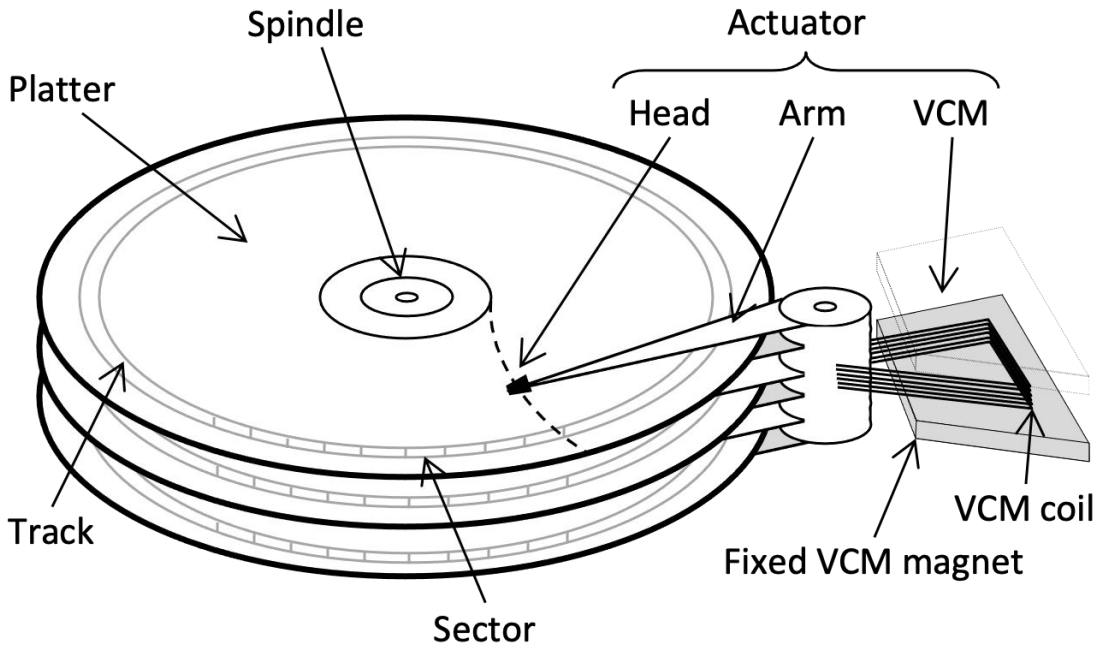
HDD 101

Mechanics

- Read/write heads access data on the platters.
- All heads move together.
- One head is active at a time.

Tracks and addressing

- Concentric tracks contain sectors.
- LBA uses logical block numbers instead of cylinder/head/sector addresses.
- Longer outer tracks hold more sectors and deliver higher sequential throughput.



Higher device capacity lowers $/TB as recording technologies evolve.

© 2026 Western Digital Corporation. All rights reserved.

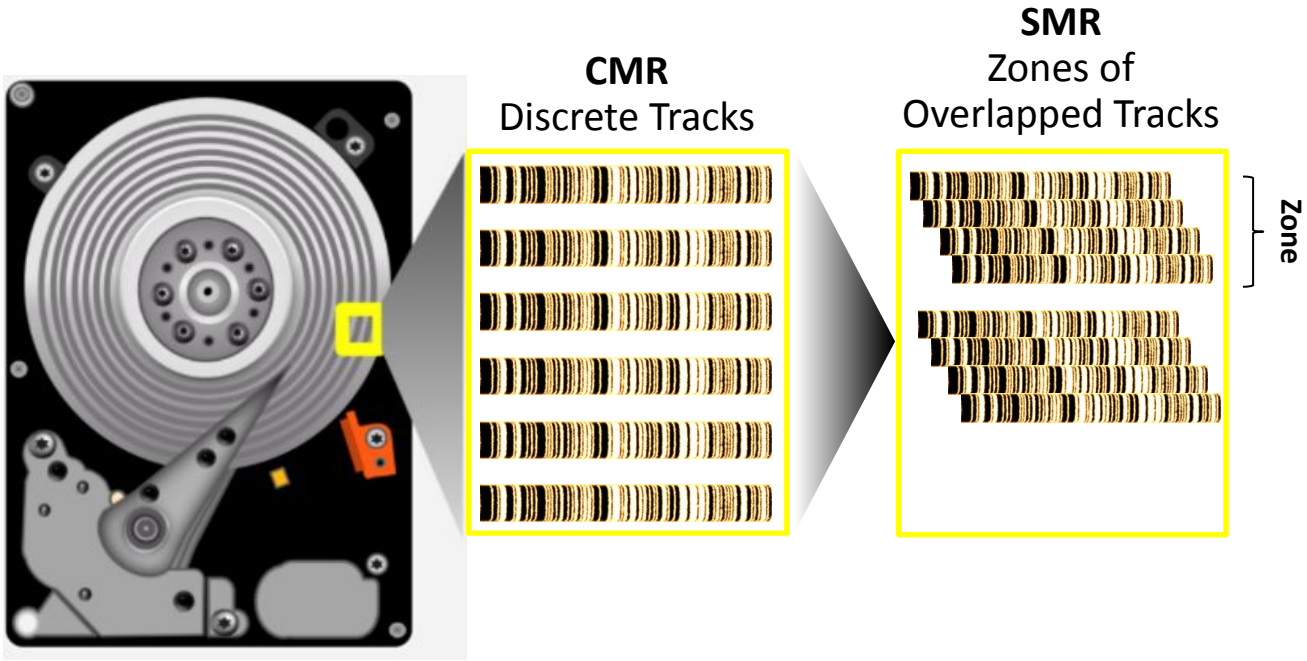
Conventional vs Shingled Magnetic Recording

Conventional Magnetic Recording (CMR)
Non-overlapping discrete tracks

- Sectors can be read and written randomly

Shingled Magnetic Recording (SMR)
Zones of overlapping tracks

- Sectors can be read randomly but must be written sequentially in each zone



- **20 to 25% capacity increase** with more efficient use of media space
- Sequential write pattern enables stronger ECC/data encoding for better reliability[1]
- Applies to all recording technology (e.g. SMR on HAMR)

[1]Tech Brief: https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/tech-brief/tech-brief-ultrasmr-technology.pdf

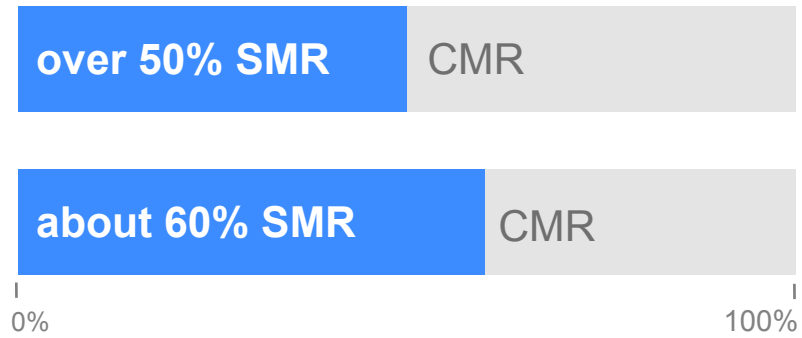
© 2026 Western Digital Corporation. All rights reserved.

SMR Is the New Normal

- Greater storage capacity and reliability support Cloud and AI workloads
- Higher capacity per drive **lowers TCO** which depends on metrics such as $/TB, W/TB, TB/in^3 and TB/slot .
- **WD's SMR's share** of exabytes is **expected to reach about 60% by mid-2027**.
- CMR use continues to shift toward legacy and specialized workloads.

Most WD nearline exabytes ship as SMR

Share of WD's nearline exabytes shipped as UltraSMR



Sources: WD earnings calls, January 2026 (December 2025 quarter) and August 2026 (target).
Mid-2027 is the end of WD's fiscal year 2027.

"Toshiba Begins Sampling of 30-34 TB SMR Nearline Hard Disk Drives"

"Toshiba Electronic Devices & Storage Corporation ('Toshiba') has announced the M12 Series of 3.5-inch nearline hard disk drives (HDDs) for hyperscale and cloud service providers. The new M12 HDDs employ Shingled Magnetic Recording (SMR) technology to deliver storage capacities ranging from 30 to 34 TB."

Source: 3/30-2026 Toshiba Press Release
https://news.toshiba.com/press-releases/press-release-details/2026/Toshiba-Begins-Sampling-of-30-34-TB-SMR-Nearline-Hard-Disk-Drives/default.aspx

Seagate, MSST 2026: "Leading capacities enabled by HAMR + SMR"

At MSST 2026, Seagate showed that SMR adds 7% to 18% capacity over CMR in every generation from 2018 to 2026. From 2024 the SMR capacity lead widened, and the SMR curve continues climbing past 40 TB, while CMR is near 32 TB.

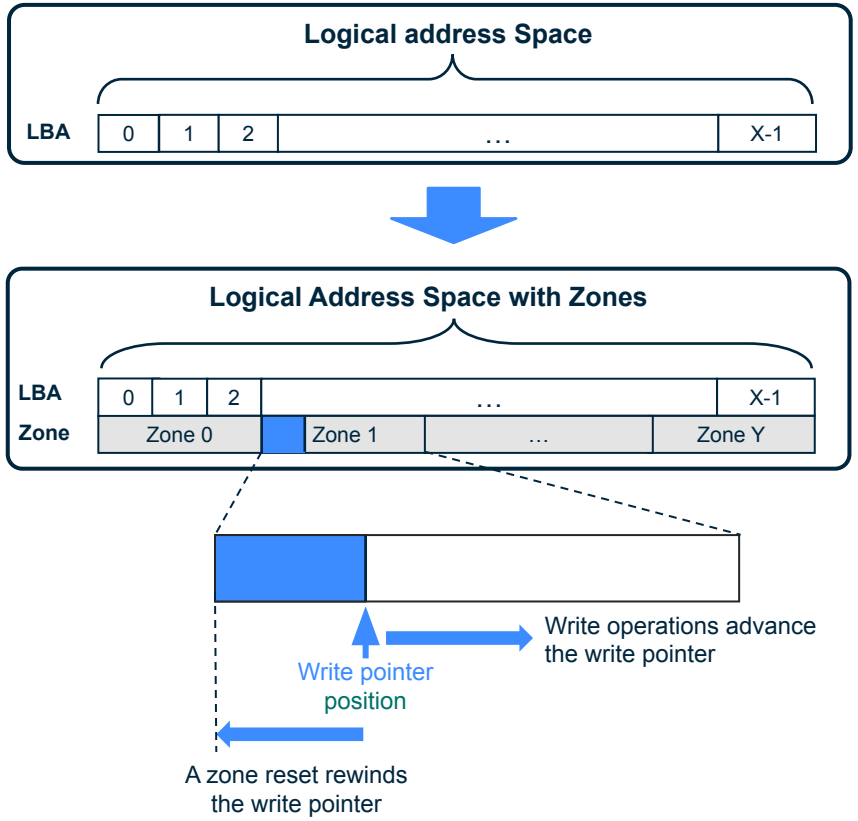
Source: R. Mendonsa, D. Rice, "HAMR and SMR: Scaling Storage in the AI Era", The International Conference on Massive Storage Systems and Technology (MSST), June 2026
https://www.msstconference.org/2026/slides/Day%201/Session%203/Seagate%20HAMR%20+%20SMR.pdf

© 2026 Western Digital Corporation. All rights reserved.

How Zoned Storage Works

Zoned storage devices, such as host-managed SMR HDDs, organize capacity into sequentially written zones rather than a conventional random-write address space.

- **Zones**: Each zone maintains a write pointer that advances as data is written.
- **Writes**: Must be sequential, starting at the current write pointer.
- **Reads**: Random access is supported within written regions.
- **Overwrites**: In-place modification is not supported.
- **Reclamation**: Valid data must be relocated before a zone can be reset and reused.



© 2026 Western Digital Corporation. All rights reserved.

The Linux Zoned Storage Ecosystem

A Decade of Development

Native Linux Support Since 2017

- Zoned block device API introduced in Linux 4.10

5+ Linux Distributions with Zoned Storage Support

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

Local File-systems

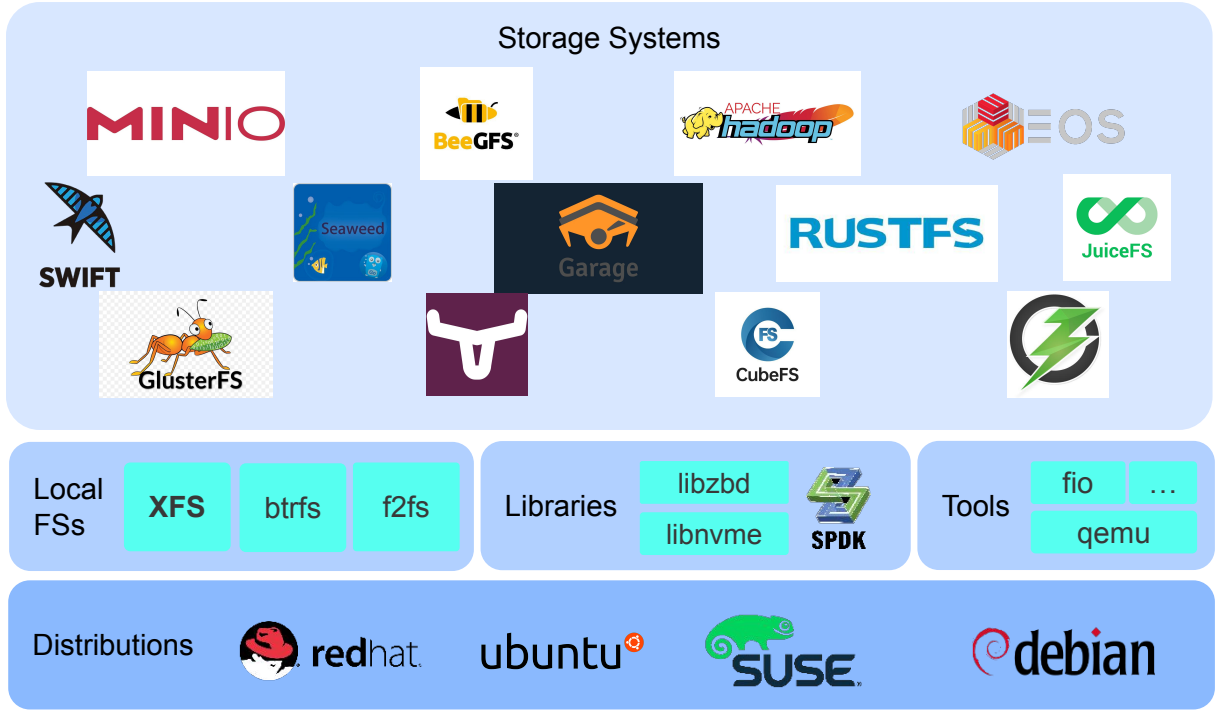
- f2fs (mobile/embedded)
- Btrfs & XFS (enterprise)

Library/Tools support

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

Storage Systems

- Longhorn, BeeGFS, HDFS, GlusterFS, AIStore, …



A mature Linux ecosystem, enabling zoned storage deployments at scale

© 2026 Western Digital Corporation. All rights reserved.

Ceph & Zoned Storage

BlueStore

HDD and SSD Backend

2020
Zoned storage support added
(Abutalib Aghayev)

2024
Support was removed

Today
Revival of the zoned patches



SeaStore

SSD optimized storage backend

2021
Zoned storage support arrived in SeaStore
(Joseph Sawaya)

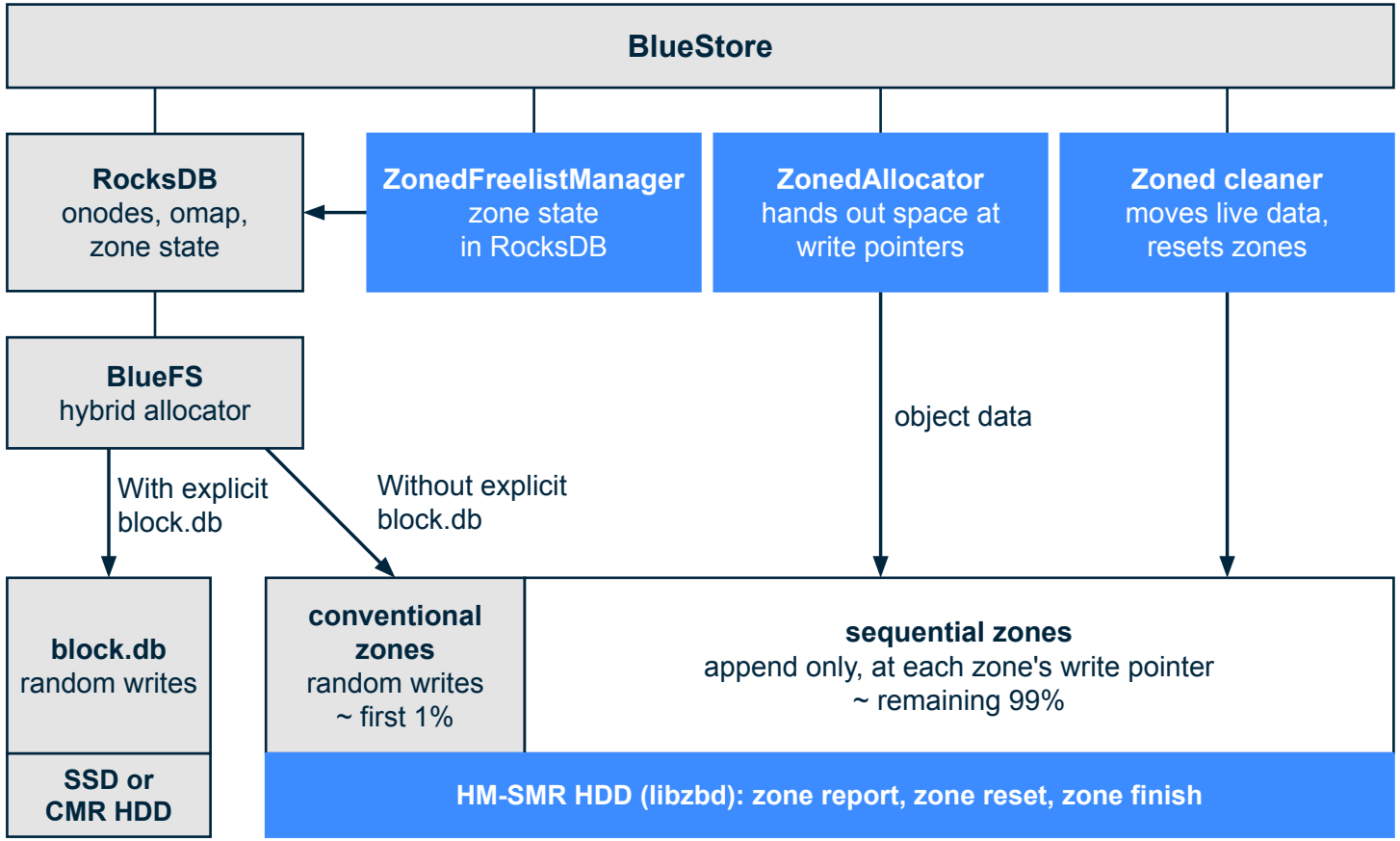
2022-2023
Continuous improvements (Aravind Ramesh)

Today
Optimized for SSDs. Keep an eye on existing support.

To support SMR HDDs, BlueStore support is required

© 2026 Western Digital Corporation. All rights reserved.

Zoned Storage Support in BlueStore



- libzbd used to detect the zoned drive. BlueStore sets its freelist type to zoned
- RocksDB and BlueFS need random writes. They run on block.db or in the conventional zones of an SMR drive
- Object data goes only to the sequential zones, appended at each zone's write pointer

Blue: zoned storage specifics
Code: src/blk/zoned/HMSMRDevice.cc, src/os/bluestore/ZonedAllocator.cc, ZonedFreelistManager.cc and BlueStore.cc. Build option WITH_ZBD (libzbd).

© 2026 Western Digital Corporation. All rights reserved.

Writes to a Zoned Storage Target



Dead bytes occupy space that needs to be reclaimed

One 16 MiB write: zone n+1 fills up, the rest spills over into zone n+2

- The `ZonedAllocator` fills a zone by appending at the zone's write pointer. If a zone doesn't have enough capacity left, the write spills into the next zone, one extent per zone.
- A single lock covers allocation and I/O submission. Each zone receives its writes in order.
- No deferred writes. A small write is padded to `min_alloc_size`
- Copy on Write (CoW): "Overwritten" and "deleted" data become dead bytes. `ZonedCleanerThread` is responsible for reclaiming space

© 2026 Western Digital Corporation. All rights reserved.

Zone State is Committed With Each Write

A normal write

W CoW 4 MiB: new data at a zone's write pointer

↓

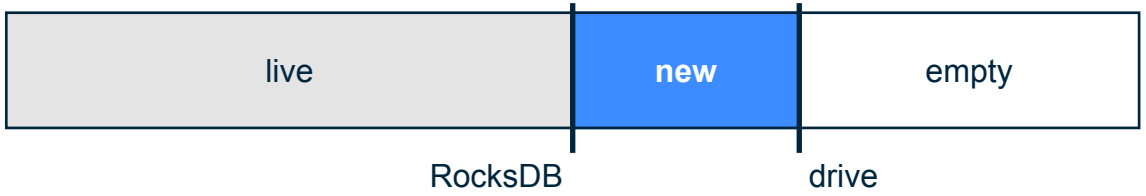
C Commit one RocksDB transaction

O	the object	onode and extents → new data
z	zone X	write pointer += 4 MiB (new data)
z	zone Y	dead bytes += 4 MiB (old data)

Prefix z: One key per zone with values for write pointer and dead bytes

A crash between write and commit

Zone X after a crash: the drive is ahead of RocksDB



↓ *Mount compares the drive's **zone report** (libzbd) with RocksDB*

Zone X after mount: the gap counts as dead bytes



First data is written. Then the transaction is committed.
The drive can be ahead of the RocksDB transactions but never behind!

© 2026 Western Digital Corporation. All rights reserved.

Zoned Cleaner: Garbage Collection

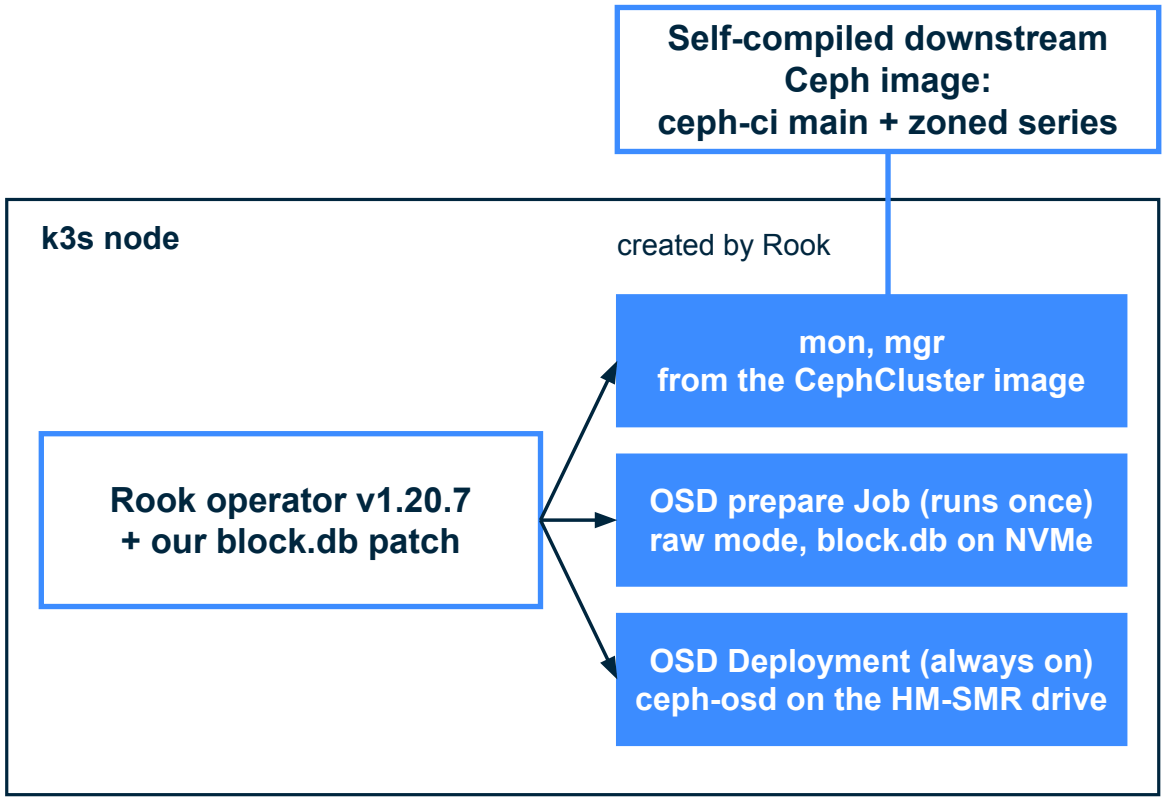
1. Wait until free sequential space is 25% or less
2. Pick the victim zone with the highest dead byte ratio (see `Score`)
3. Rewrite each object's data from the victim zone at another write pointer
4. Wait for copy writes to commit, and stop if live bytes remain
5. **Reset the victim zone physically and record it empty**
6. Allocate from the empty zone again

- `Score = dead bytes / (live bytes + 1)` clean only if at least 1/32 of the zone is dead
- The `zone refs` list the objects to move. The allocator skips the zone being cleaned
- A zone that cannot be emptied is put on a skip list for the `ZonedCleanerThread`
- *Known gap: a blob shared by clones or snapshots is copied once per object*

Settings: bluestore_zoned_cleaner_min_free_ratio (default 0.25), bluestore_cleaner_sleep_interval (default 5 s). Thread: bstore_zcleaner.

© 2026 Western Digital Corporation. All rights reserved.

Rook Runs Zoned OSDs With a Small Patch




```
$ kubectl -n rook-blockdb get pods
NAME                                                READY   STATUS
rook-ceph-operator-5667df577c-h9w8g                 1/1     Running
rook-ceph-mon-a-66cf664b8d-d2jpf                    1/1     Running
rook-ceph-mgr-a-54f6cf7fb8-mm48x                    1/1     Running
rook-ceph-osd-prepare-<node>-rk9hm                  0/1     Completed
rook-ceph-osd-0-f6c5bd8f5-6dtzz                     1/1     Running
…rbd.csi.ceph.com-ctrlplugin-5455c56577-4npz8       5/5     Running
…rbd.csi.ceph.com-nodeplugin-9cggz                  2/2     Running
  (csi operator, exporter, tools: Running)
```

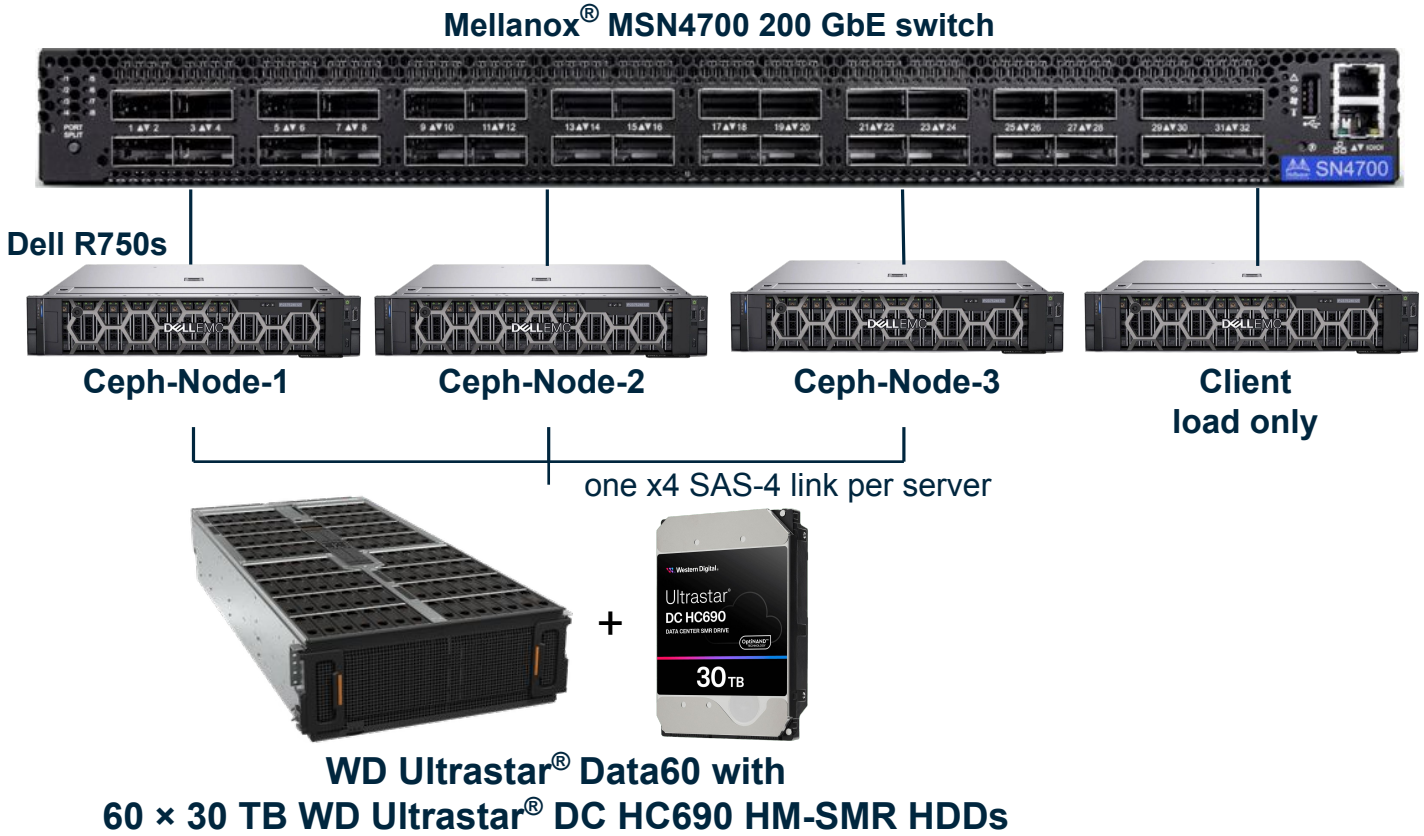
Not upstream yet:

Rook raw mode with block.db on separate NVMe. ceph-volume must link block.db, or the OSD cannot start

Blue: runs our Ceph image. Blue outline: our builds.
One k3s node, one 15TB HM-SMR drive (fresh zones, cache off), 7 GiB block.db on NVMe, size-1 pool: sequential write at 93.8% of raw SMR disk.

© 2026 Western Digital Corporation. All rights reserved.

Ceph + SMR Deployment



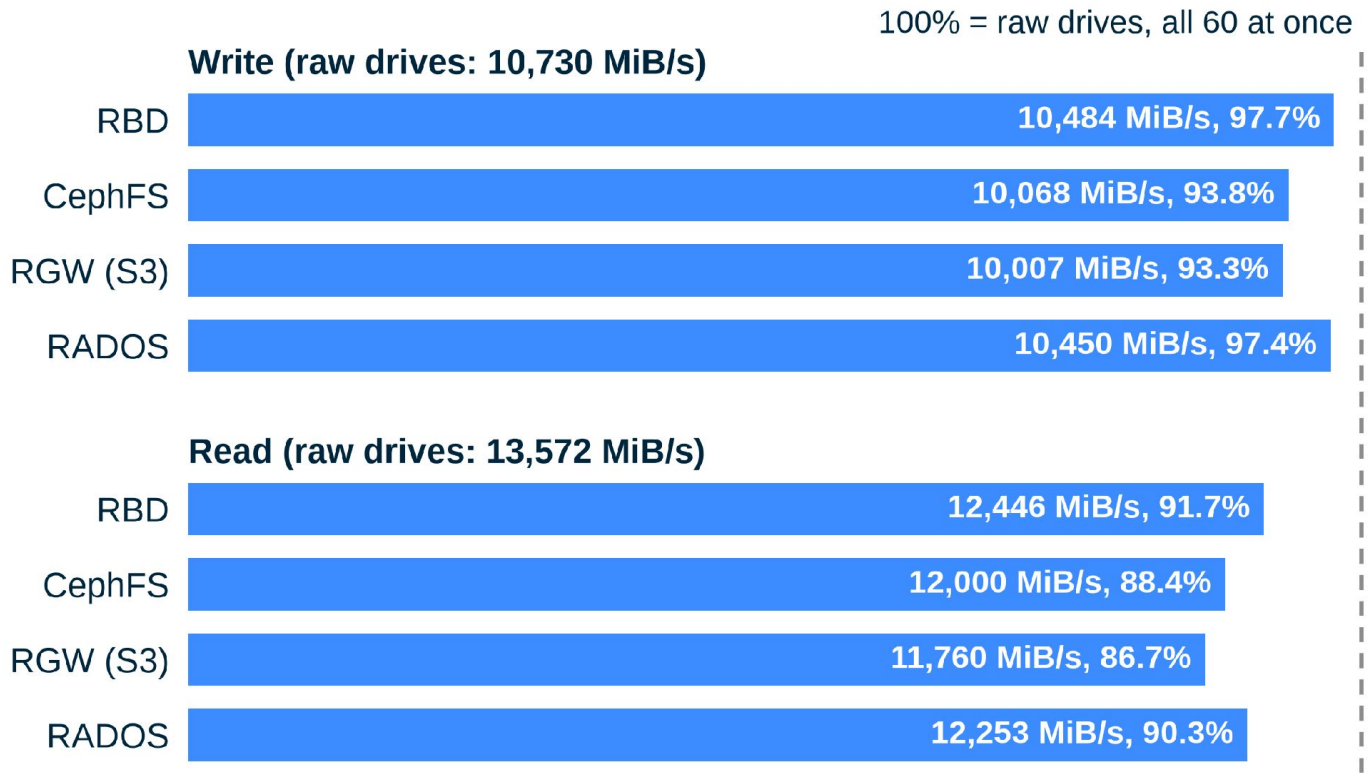
WD Ultrastar® Data60 with
60 × 30 TB WD Ultrastar® DC HC690 HM-SMR HDDs

- Ceph main (21.3.0) with the revived zoned BlueStore series
- Each Ceph node controls 20 HDDs/OSDs with block.db on an NVMe™ SSD and 4 RGW
- Size-1 and size-3 pools, 16 MiB requests and objects, 256 MiB S3 objects
- Benchmark load comes from the client node
- Baseline: fio on all 60 raw drives at once, same SAS path: 10,730 MiB/s write, 13,572 MiB/s read

Per Dell R750 node: 2× Xeon Gold 6354, 512 GiB RAM, one 15.36 TB NVMe SSD, one 200 GbE port. Drive write caches off, benchmarks started with clean drives

© 2026 Western Digital Corporation. All rights reserved.

Throughput Evaluation on 60 SMR Drives

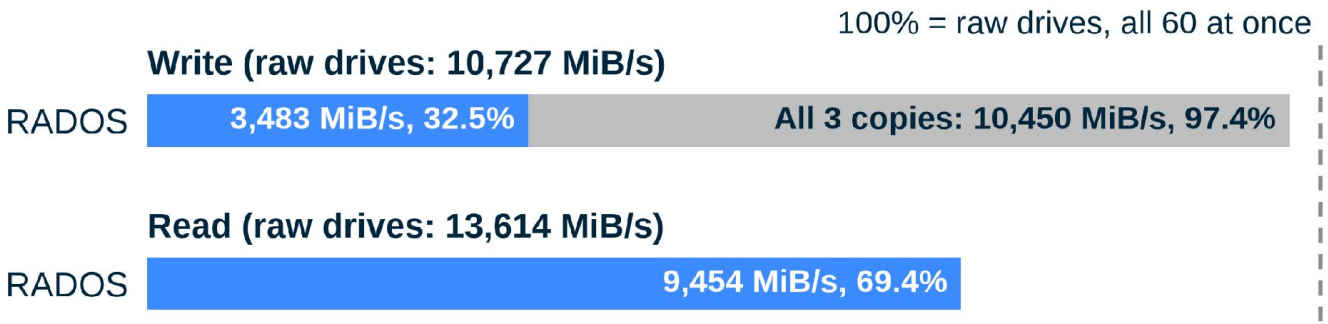


- `fio` for Raw, RBD and CephFS
  `s3load` for RGW
  `rados bench` for RADOS
- Per OSD: 167 to 175 MiB/s write, 196 to 207 MiB/s read
- Cold reads: the data was fetched from disk
- Block, file and S3 data checksum-verified with 0 errors

Size 1, 16 MiB requests and objects, 256 MiB S3 objects, drives under 15% full. Reads are the mean of two cold reads. Raw: fio on all 60 drives at once, before the OSDs existed. Default CRUSH placement, PG copies evened by the mgr upmap balancer (max deviation 1, default 5).

© 2026 Western Digital Corporation. All rights reserved.

Throughput Performance on 60 SMR Drives Size-3



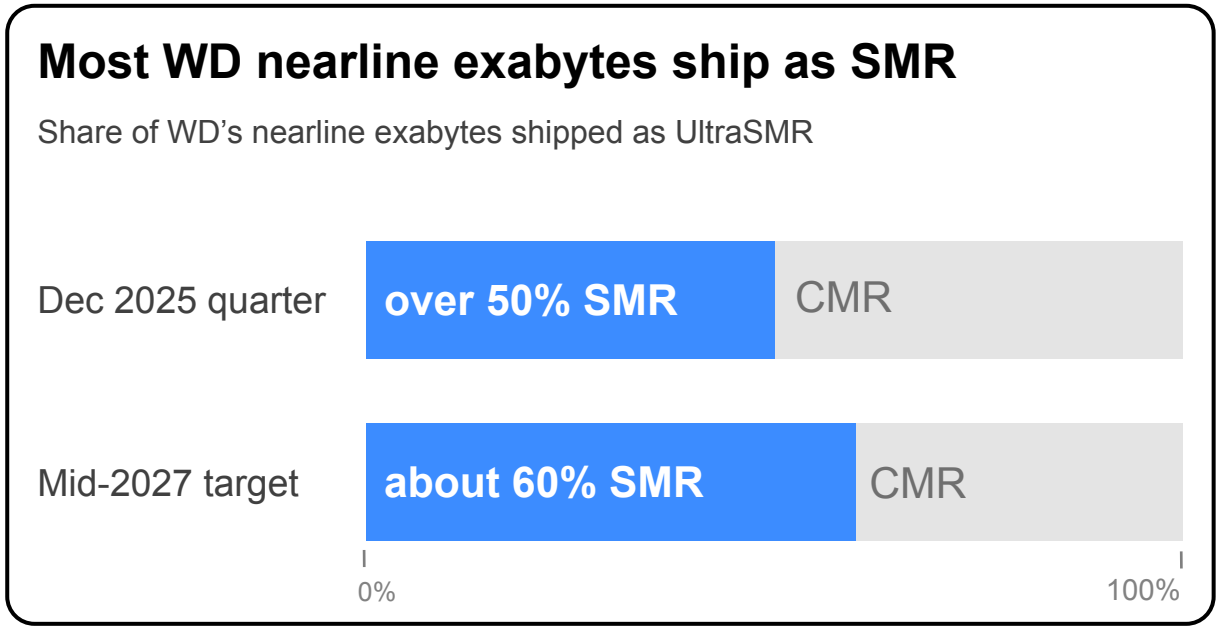
- `fio` for Raw
  `rados bench` for RADOS
- Per OSD: 174 MiB/s written with all 3 copies, the same as at size 1
- Reads come from each PG's primary. CRUSH gave each OSD 21 to 44 primaries, so per-OSD reads ranged from 96 to 199 MiB/s. Read policy `balance` would make this datapoint faster.
- Write amplification 3.0, cold reads from disk, verified with 0 errors

Size 3 (min_size 2), one copy per host, 2048 PGs, 16 MiB objects, drives under 5% full after a rebuild on freshly reset zones. Reads are the mean of two cold reads. Raw: fio on all 60 drives at once, before the OSDs existed. Default CRUSH placement, PG copies evened by the mgr upmap balancer (max deviation 1, default 5).

© 2026 Western Digital Corporation. All rights reserved.

Conclusion

- Higher capacity per drive lowers TCO which depends on metrics such as $/TB, W/TB, TB/in3 and TB/slot.
- SMR is the new normal.
- Proof of concept to revive the BlueStore zoned storage support
  - Runs successfully on 60 x 30 TB WD Ultrastar® DC HC690 HM-SMR HDDs.



Sources: WD earnings calls, January 2026 (December 2025 quarter) and August 2026 (target). Mid-2027 is the end of WD's fiscal year 2027.

© 2026 Western Digital Corporation. All rights reserved.

Thanks!
Any questions?

© 2026 Western Digital Corporation. All rights reserved.

Feel free to contact me @ dennis.maisenbacher@wdc.com



WD, the WD design, and Ultrastar are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. Ceph is a registered trademark of Red Hat, Inc. in the U.S. and other countries. The NVMe word mark is a trademark of NVM Express, Inc. All other marks are the property of their respective owners.

Backup

Throughput & concurrency

One worker: 196.1 MiB/s. Concurrent 2–8 zones: 148.6–154.1MiB/s adjacent, 79.8–99.0MiB/s spread. (21-59% slower)

More jobs / QD32 stayed within 1.5%. Zone write plugging serializes each zone.

Software & SMR model

BlueStore serves capacity HDD clusters today. Crimson ZNS PR #69204 is still open.
SeaStore on ten HM-SMR drives (no SSD): writes at ½ to 2/3 of zoned BlueStore (1 vs many zones concurrently), reads matched with 32 op threads (one drive).

Host-managed SMR exposes write rules, unlike drive-managed SMR in the 2020 NAS issue.

Sizing & durability limits

7 GiB DB/OSD: 339–535 MiB used for 320 GiB, no spill. Full-drive DB sizing still needs validation.

WCE=0 acknowledges platter writes. No power-loss test ran.

Benchmark paths & ZNS

fio via RBD/PVC reflects application I/O. Ten-drive rados bench write: 2081.7 MiB/s (94.8%).

ZNS: 998.0 MiB/s, 1 of 14 active zones, zone capacity respected, conventional block.db.

© 2026 Western Digital Corporation. All rights reserved.