ZoneTier: A Zone-based Storage Tiering and Caching Co-Design to Integrate SSDs with Host-Aware SMR Drives

Xuchao Xie[†§], Liquan Xiao[†], David H.C. Du[§]

[†]*National University of Defense Technology* , [§]*University of Minnesota-Twin Cities*

Abstract

- Integrating SSDs and HA-SMR drives can build a cost-effective high-performance storage system.
- In hybrid storage systems, HA-SMR drives endure non-sequential writes (NSWs) from both workload and internal data migration processes.
- ZoneTier leverages the intrinsic host-aware property of HA-SMR drives to control all the NSWs to HA-SMR drives in SSD tiering and caching policies.

HA-SMR Drive Internals

Shingled Magnetic Recording (SMR)

- Tracks are overlapped to increase areal density.
- Updating a track may destroy the data in its adjacent tracks.
- Platter is separated into zones to reduce update impact.
- Three SMR drive models:
  - Device-Managed (DM).
  - Host-Aware (HA).
  - Host-Managed (HM).

Drive Model	Host-aware disk model
Seagate Model No.	ST8000AS0022-1WL SN01
Device Interface	ATA ZAC
Zone Size/ Capacity	256MB/8001.563 GB
Media Cache Size	25.6GB
# CMR/SMR Zones	64/29745 (#64-29808)

Table 1. Evaluated HA-SMR Drive

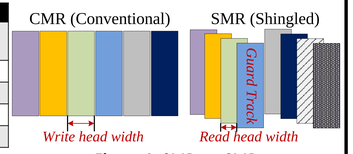


Figure 1. SMR vs. CMR

Host-Aware SMR Zones

- HA-SMR zones are designated as sequential write preferred zones with a typical size of 256MB.
- Each HA-SMR zone has a write pointer indicating the beginning address of the next LBA to be written.
- Zone state can be converted between "*sequential*" and "*non-sequential*" states.

Intrinsic Host-Aware Property

- The host can interact and manipulate HA-SMR zones by zone specific commands: OPEN ZONE, CLOSE ZONE, FINISH ZONE, REPROT ZONE, and RESET WRITE POINTER.

Sequential Write and NSW

- Only the writes aiming to the write pointer of the HA-SMR zones on "sequential" state are defined as sequential writes.
- HA-SMR drives cope with NSWs internally by redirecting NSWs to media cache in a log-structured journaling manner.
- Cleaning is triggered to move NSWs in media cache to HA-SMR zones by Read-Modify-Write (RMW) operations.

HA-SMR Drive Performance

- Media cache cleaning may precipitously degrade HA-SMR drive performance.
- HA-SMR drive behaves a continuous period of ultra-low throughput during media cache cleaning.

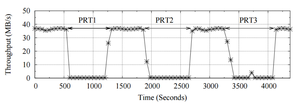


Figure 2. Media Cache Cleaning Caused Precipitous Performance Degradation

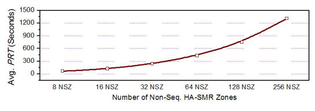


Figure 3. Performance Recovery Time (PRT) with Different NSW Patterns

ZoneTier Design

System Overview

- SSD is partitioned into tiered storage and Cache partitions.
- ZoneTier aligns the boundaries of extents and HA-SMR zones in storage tiering.
- The zones in SSD tier and HA-SMR drives are dynamically and periodically relocated.
- Internal data movements are performed by Promoting, Demoting, Evicting and Merging.
- Four potential end points on I/O path: SSD zone, CF-Cache, HA-SMR zone, and media cache.

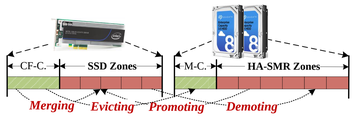
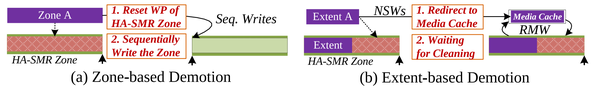


Figure 4. SSD Assignment and Internal Data Movements in ZoneTier



(a) Zone-based Demotion (b) Extent-based Demotion

Figure 5. Comparison of Zone-based and Extent-based Demotions

Operation	Description
Promoting	Refresh SSD zones by periodically promoting selected zones from HA-SMR drive
Demoting	Move the selected SSD zones back to HA-SMR drive to make space for newly promoted zones
Evicting	Evict the cached data in CF-Cache to media cache or HA-SMR zones (if former NSW turns to sequential write) to make space for incoming NSWs.
Merging	Combine data in CF-Cache and its originally targeted HA-SMR zone to maintain data consistency when the zone is promoted to SSD tier.

Table 2. Operations for Moving Data Internally in ZoneTier

Zone Placement

- Critical Access Frequency (CAF) is used in ZoneTier to evaluate zones for future placements.

Zone Promotion

- Intra-zone garbage is filtered out to reduce zone promotion overhead and improve the utilization efficiency of SSD tier.

Zone Demotion

- Reset the write pointer of involved HA-SMR zones to reshape NSWs to sequential writes to HA-SMR zones.

Cleaning-Friendly Caching

- Reshape the inevitable NSWs to cleaning-friendly NSW patterns by NSW-aware admission and zone-aware eviction.

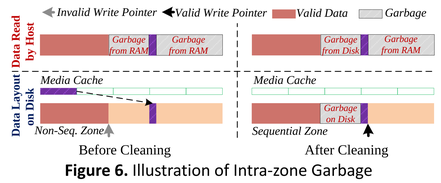


Figure 6. Illustration of Intra-zone Garbage

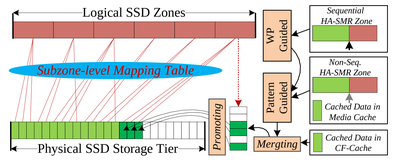


Figure 7. Zone Promotion with Intra-zone Garbage Filter

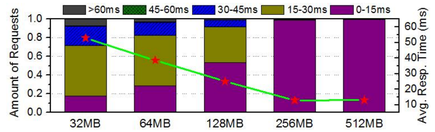
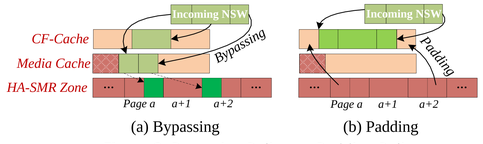
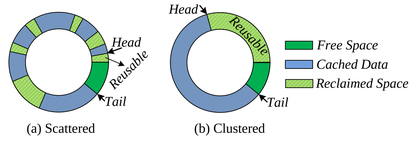


Figure 8. Performance Evaluation of Zone Demotion



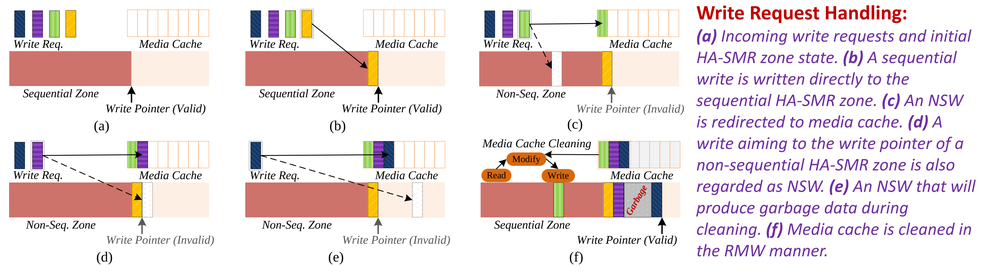
(a) Bypassing (b) Padding

Figure 9. Bypassing Policy vs. Padding Policy



(a) Scattered (b) Clustered

Figure 10. RMW Efficiency with Different NSW Patterns



Write Request Handling:

(a) *Incoming write requests and initial HA-SMR zone state.* ***(b)*** *A sequential write is written directly to the sequential HA-SMR zone.* ***(c)*** *An NSW is redirected to media cache.* ***(d)*** *A write aiming to the write pointer of a non-sequential HA-SMR zone is also regarded as NSW.* ***(e)*** *An NSW that will produce garbage data during cleaning.* ***(f)*** *Media cache is cleaned in the RMW manner.*

Performance Evaluation

Implementation

- Using libzbc to manipulate HA-SMR drives as zoned block devices.
- Using libaio to access SSD and HA-SMR drives as normal block devices.
- Data management algorithms are implemented in user space.

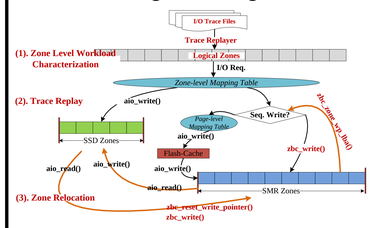


Figure 11. Functions Used in libaio and libzbc

Design	Description
DM-64-LRU	Fine-grained tiering with LRU-based caching for DM-SMR drives
DM-128-LRU	
DM-256-LRU	Zone-based tiering with DM-SMR drives
HA-256-LRU	Zone-based tiering with HA-SMR drives
HA-256-CFC	Using cleaning-friendly caching for HA-SMR drives

Table 3. Comparison Candidates

Evaluation Results

- System performance of ZoneTier outperforms that of both the fine-grained tiering and zone-based tiering that ignores the host-aware property of HA-SMR drives. Average I/O response time reductions are up to 28.20% and 25.51%.
- The response time of internal writes for demotion is stable in ZoneTier while that of other designs appear significant latency amplification.
- The utilization efficiency of SSD tier in ZoneTier can be improved by 7.33%~41.9%.
- ZoneTier successfully accelerates the performance recovery of HA-SMR drives from media cache cleaning.
- Tail latencies are significantly reduced by cleaning-friendly caching in ZoneTier. The 99.9th percentile write latencies are reduced by up to 15.3x than that of LRU-based caching.

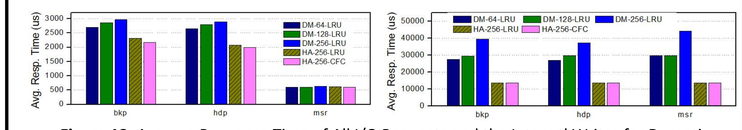


Figure 12. Average Response Time of All I/O Requests and the Internal Writes for Demotion

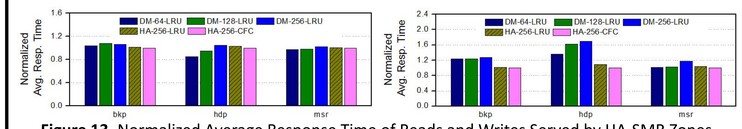


Figure 13. Normalized Average Response Time of Reads and Writes Served by HA-SMR Zones

Conclusion

we describe *ZoneTier*, a zone-based SSD tiering and caching co-design to handle all the NSWs to HA-SMR zones by leveraging the intrinsic host-aware property of HA-SMR drives . We hope the techniques in ZoneTier can be extended to other deployment scenarios of HA-SMR drives.

Reference

1. Alireza Haghdoost, Weiping He, Jerry Fredin, and David H.C. Du. 2017. On the Accuracy and calability of Intensive I/O Workload Replay. In 15th USENIX Conference on File and Storage Technologies (FAST 17). USENIX Association, Santa Clara, CA, 315–328.
2. Weiping He and David H.C. Du. 2017. SMaRT: An Approach to Shingled Magnetic Recording Translation. In *15th USENIX Conference on File and Storage Technologies (FAST 17)*. USENIX Association, Santa Clara, CA, 121–134.
3. Fenggang Wu, Ming-Chang Yang, Ziqi Fan, Baoquan Zhang, Xiongzi Ge, and David H.C. Du. 2016. Evaluating Host Aware SMR Drives. In *8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16)*. USENIX Association, Denver, CO.
4. Weixia Xu, Yutong Lu, Qiong Li, Enqiang Zhou, Zhenlong Song, Yong Dong, Wei Zhang, Dengping Wei, Xiaoming Zhang, Haitao Chen, et al. 2014. Hybrid hierarchy storage system in MilkyWay-2 supercomputer. *Frontiers of Computer Science* 8, 3 (2014), 367–377.