A Low-Cost Hybrid SSD Architecture for Read-Intensive Enterprise Workloads

Hyeon Gyu Cho[1] Kyungho Shin[1,2] Jaeyoung Jang[1] Jae W. Lee[1]
[1]Sungkyunkwan University, Suwon, Korea
[2]Samsung Electronics, Hwaseong, Korea
E-mail: [1]{cho42me, kh.shin, rhythm2jay, jaewlee}@skku.edu, [2]gungho.shin@samsung.com

Abstract: Recently, flash-based solid-state drives (SSDs) have been widely adopted in all scales of computing platforms due to their low latency and high energy efficiency. Many enterprise workloads are known to be read-intensive, and read latency has great impact on overall system performance. This paper proposes ROSA, a novel PCM-Flash hybrid SSD architecture optimized for read-intensive enterprise workloads. ROSA completely hides the heterogeneity in device types inside the SSD, thus requiring no change to the existing software stack. Migrating read-intensive pages into PCM devices is piggybacked onto garbage collection (GC) commands to incur minimal performance cost. Our evaluation with three web search workloads demonstrates over 28% reduction of average access latency over the baseline hybrid SSD oblivious of read intensity.

Keywords—**PCM, Hybrid SSD, Enterprise workload**

1. Introduction

Recently, flash-based solid-state drives (SSDs) have been widely adopted in all scales of computing platforms from clients to clouds due to their low latency and high energy efficiency. The NVM Express (NVMe) technology allows SSDs to be directly attached to the fast PCIe bus, which makes them even more attractive for enterprise servers running I/O-intensive workloads [4].

Many enterprise workloads are known to be read-intensive and highly skewed to be cache-friendly [4]. For such workloads read latency has great impact on overall system performance. Recently, Kim et al. [4] propose to use both low-cost eMLC flash-based SSDs and low-latency Phase Change Memory (PCM)-based SSDs for an enterprise server. Unlike flash devices PCM devices allow in-place updates and have much lower read latency. Since data transfer time is the same for both, the difference in device read time directly translates to the difference in read access time to the SSD. However, their hybrid approach based on multiple heterogeneous SSDs [4] requires significant changes to the software stack for efficient caching and tiering.

In this paper, we propose ROSA, a novel PCM-Flash hybrid SSD architecture optimized for read-intensive enterprise workloads. Unlike the prior technique [4] the heterogeneity in device types is completely hidden inside the SSD, thus requiring no change to the existing software stack. Migrating read-intensive pages into PCM devices is piggybacked onto garbage collection (GC) commands that routinely occur when blocks need to be freed, to incur minimal performance cost. Finally, unlike existing proposals for multi-device SSDs which use fast devices as storage for metadata or write buffers [3, 5, 7], ROSA stores user data in both fast and slow devices to maximize the usable storage capacity. Our evaluation with three web search workloads demonstrates over 28% improvement of average access latency over the baseline hybrid SSD with heterogeneity-oblivious data placement.

2. Design of ROSA

PCM-flash hybrid SSD architecture. Like the conventional SSD architecture, our proposal has multiple channels that can be accessed in parallel. Multiple storage devices, both flash and PCM, are attached to a single channel. Based on the study by Kim et al. [4], we assume PCM devices occupy 20% of total storage space by default. Although hybrid SSD architectures using both flash and PCM devices have been proposed previously, most of them use fast PCM devices as write buffers or storage for metadata [3, 5, 7]. In contrast, we use these PCM devices to store read-intensive data pages (not as cache) to provide both low read latency and high capacity.

Two-level, read intensity-aware GC. The primary design problem is how to efficiently select and migrate read-intensive pages into PCM devices. The key idea of ROSA is to exploit garbage collection (GC) operations for efficient page migration. To this end, we introduce two-level GCs: Level 1 optimized for optimal page placement and Level 2 for fast generation of free blocks. Level 1 GC performs read intensity-aware GC. When block usage reach at a threshold, Level 1 GC is triggered. The SSD controller then checks the read count of candidate victim blocks (i.e., sum of read counts in all of the valid pages) in both devices. The valid pages in a set of hot blocks in NAND flash devices migrate to PCM devices; the same number of cold PCM pages migrate to free NAND blocks in return. A Level 2 GC is invoked when only a few free blocks are left. This is the conventional GC in SSD, which frees blocks with fewest valid pages to quickly produce free NAND blocks for future writes. With this two-level GC, ROSA can effectively migrate read-intensive pages without host intervention while maintaining low blocking time.

If a write request is not issued for a long time, Level 1 GC is not triggered as the block usage does not increase. This may happen in read-intensive enterprise workloads (e.g., accessing only largely-invariant table structures from SSD). If a L1 GC is not triggered for a certain time interval, it is autonomously triggered inside the SSD. The L1 GC in turn migrates hot pages from NAND flash devices to PCM devices, and cold pages to the opposite direction. Thus, even workloads with very sparse writes can also benefit from the low read latency in ROSA.

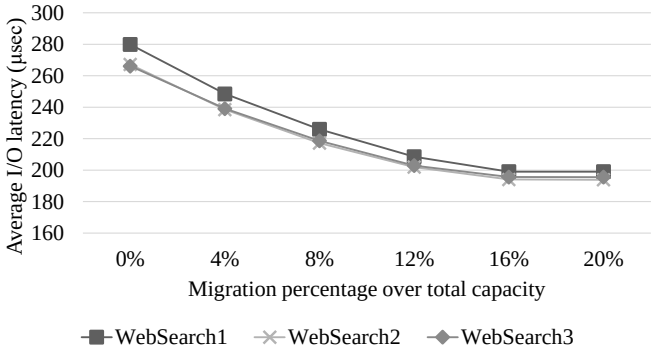

Figure 1: ROSA performance

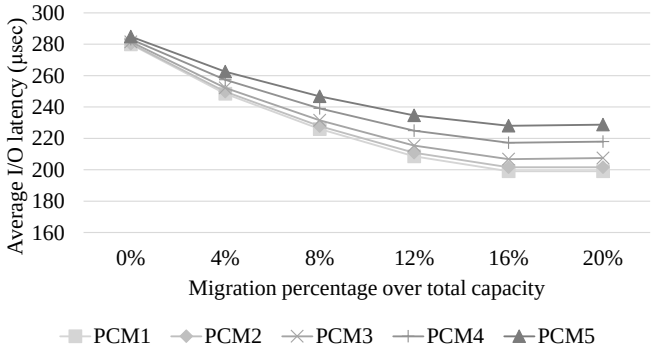

Figure 2: Average I/O latency with 5 different PCMs

3. Preliminary Evaluation

Methodology. We use DiskSim SSD extension to model ROSA. The device composition is 20% of PCM and 80% of NAND flash, which is the optimal partition in [4]. We use 88us read latency, 2300us write latency, and 10ms erase latency for TLC NAND. This parameters are taken from an existing work [8]. We use 5 different sets of read/write parameters for PCM devices taken from prior work [4,6,7]. Note that PCM can perform in-place updates with no erase cost. Both devices have 192 4KB pages per block.

Workload. We use three WebSearch traces provided by Storage Performance Council [1]. WebSearch traces are made up of over 99% read trace. These three traces access 16.7GB. For our simulation, We split those traces into two part. First half of trace is used for warming up. Then, we trigger several Level 1 GCs for migrating read-intensive pages. Finally, we run the second half of trace for evaluation.

Results. Figure 1 shows the performance of ROSA. In 0% migration, three average response times are about 0.27ms in all cases. 0% migration means that, although PCM devices are there, the SSD contoller does heterogeneity-oblivious placement. With a 20% of migration, three average response times are 0.19ms for all traces. It means that top 20% of highest read count pages are migrated to PCM. This situation is that we can use PCM device optimally. Because a higher migration percentage means using more PCM capacity, the more block is migrated, the lower average response time is. However, when the migration rate reaches about 16%, the response time is almost saturated. Further migration will only yields diminishing returns (or even degration due to migration cost).

Figure 2 shows WebSearch1 performance with varying PCM performance. Since PCM is developing now, there are no exact parameters about PCM read and write latency. Based on existing literature [4, 6, 7], we assume 5 kinds of PCM read/write parameters ranging from 2.5us to 30us for read latency and from 25us to 300us for write latency. Figure 2 is normalized by PCM5 with 0% migration. With around 16% of migration the performance gets saturated in all cases beyond which the cost of migration outweighes its benefits.

Read intensity-aware GC overhead. We mention that Level 1 GC has overhead. However, this overhead is negligible. For example, we can demonstrate maximum migration overhead. Maximum migration overhead is invoked when victim NAND block has 192 valid pages of which block has the highest read count. In that case, additional migration overhead is only 192 PCM page reads and writes. Conventional GC time is about 470ms. Compared to this, Maximum overhead is 5.28ms and is very small ($< 1.2\%$). Besides, this is not common, and, if we use parallel GC, PCM I/O cost can be reduced further.

4. Conclusion

This paper proposes ROSA, a novel PCM-NAND hybrid SSD architecture. Unlike previous PCM-NAND hybrid SSDs [2,7], we use PCM for read intensive data because PCM has lower read latency than NAND flash with completely hiding the heterogeneity in device types inside the SSD. Migrating read-intensive pages into PCM devices on GC commands to incur minimal performance cost. Our evaluation with three WebSearch workloads demonstrate over 28% reduction of average access latency over the baseline hybrid SSD.

References

[1] UMass Trace Repository. http://goo.gl/hwn5cy.

[2] G. S. Choi, I. Lee, M. Sung, and C. Im. A hybrid ssd with pram and nand flash memory. *Microprocess. Microsyst.*, 2012.

[3] I. H. Doh, J. Choi, D. Lee, and S. H. Noh. Exploiting Non-volatile RAM to Enhance Flash File System Performance. In *EMSOFT '07*.

[4] H. Kim, S. Seshadri, C. L. Dickey, and L. Chiu. Evaluating Phase Change Memory for Enterprise Storage Systems: A Study of Caching and Tiering Approaches. In *FAST '14*.

[5] J. K. Kim, H. G. Lee, S. Choi, and K. I. Bahng. A PRAM and NAND Flash Hybrid Architecture for High-performance Embedded Storage Subsystems. In *EMSOFT '08*.

[6] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger. Architecting Phase Change Memory As a Scalable Dram Alternative. In *ISCA '09*.

[7] Y. Park, S.-H. Lim, C. Lee, and K. H. Park. PFFS: A Scalable Flash Memory File System for the Hybrid Architecture of Phase-change RAM and NAND Flash. In *SAC '08*.

[8] D. Sharma. System Design for mainstream TLC SSD Meeting the performance challenge. In *Flash Summit'14*.