On the Discontinuation of Persistent Memory: Looking Back to Look Forward

Tianxi Li
The Ohio State University
li.9443@osu.edu

Yang Wang
The Ohio State University
wang.7564@osu.edu

Xiaoyi Lu
University of California, Merced
xiaoyi.lu@ucmerced.edu

Abstract

With Intel's announcement to discontinue its Optane DC Persistent Memory (DCPMM) in July 2022, it's time to learn from our existing experience and look to its future. In this paper, we have 1) carried out a survey of public reports from organizations to understand how they utilize DCPMM; 2) measured the performance of the DCPMM 200 series to understand whether it could have saved the product; and 3) discussed with corresponding developers in major IT companies. Based on such information, we argue the memory mode of DCPMM is worth more attention and it is necessary to study the sweet spots for persistent memory before heavy investment.

1 Introduction

Persistent memory (PMEM), with its byte addressability and ability to persist data, has long been a dream for I/O heavy applications. With the release of Intel Optane DC Persistent Memory 100 series [8], on April 2017, both academia and industry have been drawn to it. However, Intel's announcement to discontinue this product in 2022 has put PMEM to question – *Is the discontinuation temporary, maybe due to business reasons or fixable technical flaws, or are there more fundamental reasons?* This paper tries to shed light on this question.

First, we have collected public reports from 33 organizations in various industries. Our analysis shows that 37% of PMEM is deployed in the memory mode, usually due to PMEM's larger capacity and lower cost than DRAM. 48% of PMEM are configured with persistence in the app-direct mode, and they're mostly used in databases and key-value stores. 15% of the cases exploit both modes.

Second, since the DCPMM 200 series became available shortly before the announcement, we wonder if it could have saved PMEM. We did experiments around 200 series' key feature of eADR, and measure the bandwidth.. The results show that non-flushed stores are only 14.3% faster than flushed stores of the 100 series. This suggests that 200 series is perhaps an incremental improvement rather than major revolution.

Third, we have discussed with corresponding developers in three major IT companies, which have revealed both potentials and obstacles for PMEM techniques. For example, one company observes SSDs can already provide satisfactory throughput and latency for its applications.

Based on such information, we have made two recommendations to move forward. First, given the popularity of the memory mode, it is perhaps worth more attention as a separate product. Second, it is necessary to carry out a study to understand the "sweet spots" for PMEMs since some applications may not need a higher performance but prefer a lower cost.

2 Survey of PMEM Ecosystems

We collect 33 public reports from organizations in various areas. The reports describe each company's applications, use cases, and results of using DCPMM [1].

About 27.3% of the cases deploy PMEMs for research purposes. These are often supercomputer clusters established by universities or institutes. The second biggest portion is from Cloud providers that take up 24.2% of the instances. Several major Cloud platforms have included PMEM in their services including Alibaba Cloud [2] and Microsoft Azure [3]. Apart from these two categories, various fields make up the other half of the share including Finance [11], Communication companies [14], etc. The diversity indicates that PMEM is viewed as a potential product by a wide spread of fields.

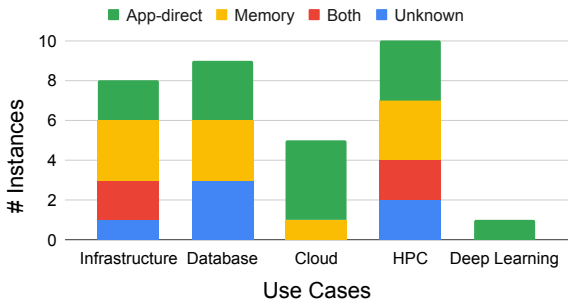

Figure 1. Distribution of Applications in Survey of Intel Optane DC Persistent Memory.

We categorize the applications in Figure 1. The top category is High Performance Computing (HPC), which corresponds to the aforementioned supercomputer clusters. The second category is database. Among this category, Sap Hana Database [7] and Redis [13] are the most adopted. Besides them, PMEM is also involved in building memory node and storage node in clouds and company IT infrastructures. Though AI and Deep Learning workloads prevail in research, we don't find many instances using PMEM at present.

We further summarize the PMEM mode configuration in each case. As shown in Figure 1, among the 27 knonwn cases, about 37% cases are deploying DCPMM in memory mode as a DRAM expansion due to its larger capacity [5, 9] and lower total cost of ownership (TCO) [10]. In contrast, 48% cases use the app-direct mode and 15% cases exploit both modes.

For cases that configure PMEM in the app-direct mode, databases and key-value stores are the most common and closely relevant applications. Besides, the systems often leverage the optimized I/O path by memory mapping and direct I/O.

3 200 Series vs. 100 Series

The 200 series became publicly available about one year before the discontinuation announcement, which gives academia

and industry little time to fully evaluate its potential. Thus we did a series of experiments between 100 and 200 series to fill the gap. While we have done a comprehensive measurement, due to space limitation, here we mainly compare 200's eADR performance with 100s' ADR performance.

Experimental Setup. Our DCPMM 200 series machine is equipped with a Ice Lake Processor of 56 cores, 256GB of DRAM, and a total of 2TB PMEM from 8x 256GB DIMMs. Our 100 series machine has a Cascade Lake Processor of 48 cores, 192GB DRAM, and a total of 1.5TB PMEM from 12x 128GB DIMMs. We use PMDK [12] version 1.11.1 .

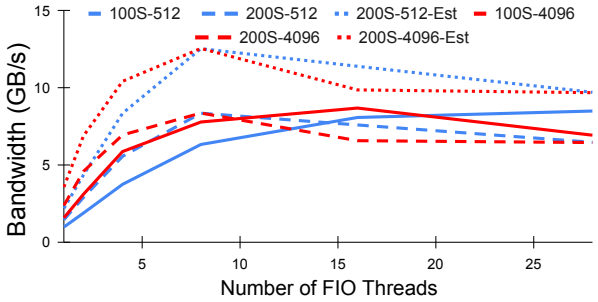

Figure 2. ADR vs. eADR on Sequential Write. 200S-X-Est is estimated 12-DIMM performance, as 200S-X * 1.5.

The 100 series guarantees persistence of the DIMMs but not the cache, so it requires explicit cacheline flush followed by a memory barrier to make persistent, ordered stores. This consequently restricts the performance and complicates the programming model. The 200 series supposedly gets rid of the flush by making everything inside the cache persistent from eADR feature. So we compare 200 series doing non-flushed stores and 100 series' doing flushed stores.

Figure 2 presents FIO [4] sequential workload with 200 series doing non-flushed stores and 100 series doing flushed stores. 200 series can deliver up to 50% more bandwidth when the number of concurrent worker threads is low (≤ 2). As the experiment scales up, 200 series is 24% worse than 100 series. With the estimated 12-DIMM performance, the 200 series outperforms the 100 series by at most 14.3%. Concurrent accesses bring pressure to CPU caches. As a consequence, data in XPLine (256B) [6, 15] are evicted prematurely, magnifying the write amplification [6]. Thus the performance is close to the flushed store of the 100 series. Besides, the *clwb* instruction added to the ISA in Ice Lake is allegedly more efficient, diminishing the advantages of non-flushed stores. Given the improvement is not significant, we believe the 200 series would not change the fate of DCPMM.

4 Discussions with Industrial Developers

We reached corresponding developers in three major IT companies for a detailed discussion. While it provides valuable information, it should only be taken as a grain of salt, since even inside the same company, opinions may differ.

First, two companies are interested in high-density and low-cost memory, because a variety of applications prefer a high memory capacity with a low cost but do not require a bandwidth as high as DRAM.

Second, two companies are interested in the persistent mode of PMEM (i.e., app-direct mode), one for cloud drives and one for databases, but they admit fully utilizing PMEM would require a significant re-design of the applications. The remaining company observes SSDs already provide satisfactory latency and throughput for their applications.

Finally, they present a few detailed obstacles to adopt PMEM, including cost, Intel as the only major vendor, and higher performance variance under high load compared to DRAM.

5 Looking back and forward

Based on the above information, we discuss the following aspects:

- Memory vs Persistent mode: Given the popularity of using PMEM in memory mode, maybe it's worthwhile positioning high-density and low-cost memory as separate products. It's unclear to us why Intel decided to include both memory mode and app-direct mode in one product in the first place, but considering the fact that Optane can only be used with Intel's expensive high-end CPUs with the support for persistence related features, people only interested in the memory mode may not be willing to pay such extra cost.
- What is the "sweet spot"? While many of the issues discussed above, such as cost and software complexity, may be addressed by technical improvement and research efforts, "no need for higher throughput or lower latency" is a fundamental issue questioning the future of PMEM. During the discussion, we find industrial applications often have a desired space, bandwidth, and latency: a good matching of these factors in hardware can minimize cost but an improvement in one factor might just cause a waste of resource. For example, suppose an application has 2TB of data and needs a sustained bandwidth of 2GB/s, then a storage device with 2TB of space and 2GB/s of bandwidth would be a good match. By installing two same devices, a device with 1TB of space and 1GB/s of bandwidth would work well, a device with 2TB of space and 1GB/s of bandwidth would cause a waste of space, and a device with 1TB of space and 2GB/s of bandwidth would cause a waste of bandwidth. Hence, to understand the potentials of PMEMs, it's critical to carry out a study about the desired performance of industrial applications and identify the "sweet spots" for PMEMs. Before that, blindingly improving PMEM performance (via hardware improvement or software re-design) may not be beneficial.

Acknowledgments

We would like to sincerely thank Eduardo Berrocal and Joseph E Oster from Intel for their invaluable insights and for providing us access to the DCPMEM 100 series hardware platform. We extend our gratitude to the industrial developers who generously shared their insights, as well as the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by the NSF research grants CCF #2132049, CNS #2310919, CNS #1908020, and CCF #2118745.

References

[1] Intel PMEM Market Survey, Google sheets. https://docs.google.com/spreadsheets/d/13-IdEUfkuyy9iHoThwWvbeLTZZXjbawKW96TP_7zbOE/edit?usp=sharing, 2023.

[2] Alibaba. Configure the usage mode of persistent memory. https://www.alibabacloud.com/help/en/elastic-compute-service/latest/configure-persistent-memory-usage, 2022.

[3] Anhansen. Understand and deploy persistent memory - Azure Stack HCI | Microsoft Learn. https://learn.microsoft.com/en-us/azure-stack/hci/concepts/deploy-persistent-memory, 2021.

[4] Jens Axboe. axboe/fio: Flexible I/O Tester at fio-3.32. https://github.com/axboe/fio/tree/fio-3.32, 2022.

[5] GPORTAL. Case Study: Improve Performance and Reduce Costs with Intel® Optane™ Technology. https://www.intel.com/content/dam/www/central-libraries/us/en/documents/csp-gportal-case-study.pdf.

[6] Shashank Gugnani, Arjun Kashyap, and Xiaoyi Lu. Understanding the idiosyncrasies of real persistent memory. *Proceedings of the VLDB Endowment*, 14(4):626–639, 2020.

[7] Sap Hana. Sap Hana. https://www.sap.com/products/technology-platform/hana/what-is-sap-hana.html.

[8] Intel. Intel® Optane™ Persistent Memory 100 Series. https://www.intel.com/content/www/us/en/products/sku/190350/intel-optane-persistent-memory-100-series-256gb-module/specifications.html.

[9] Intel IT. Case Study: Scale Up For Faster Time to Insight. https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2022-05/scale-up-for-faster-time-to-insight-case-study.pdf.

[10] Kingsoft. Kingsoft Cloud Redis Service* powered by 2nd Gen Intel® Xeon®. https://www.intel.com/content/www/us/en/processors/xeon/scalable/software-solutions/kingsoft-cloud-redis-service.html.

[11] Paypal. Case Study: PayPal Solves Fraud Challenges with Aerospike® and Intel® Optane™ Persistent Memory. https://www.intel.com/content/dam/www/central-libraries/us/en/documents/aerospike-paypal-cs.pdf.

[12] PMem.io. Persistent Memory Development Kit (PMDK). https://pmem.io/pmdk/, 2022.

[13] Redis. redis/redis: Redis is an in-memory database that persists on disk. The data model is key-value, but many different kind of values are supported: Strings, Lists, Sets, Sorted Sets, Hashes, Streams, HyperLogLogs, Bitmaps. https://github.com/redis/redis.

[14] SoftBank. Validation Case Study: Delivering up to 3x improvement in VM capacity and 41% cost reduction on the infrastructure supporting SoftBank's communication business. https://www.intel.com/content/dam/www/public/us/en/documents/case-studies/softbank-vm-case-study.pdf.

[15] Jian Yang, Juno Kim, Morteza Hoseinzadeh, Joseph Izraelevitz, and Steve Swanson. An empirical guide to the behavior and use of scalable persistent memory. In *18th USENIX Conference on File and Storage Technologies (FAST 20)*, pages 169–182, 2020.