Experience and Performance of Persistent Memory for the DUNE Data Acquisition System

Adam Abed Abud, *Student Member, IEEE*, Giovanna Lehmann Miotto, and Roland Sipos

Abstract—Emerging high-performance storage technologies are opening up the possibility of designing new distributed data acquisition (DAQ) system architectures, in which the live acquisition of data and their processing are decoupled through a storage element. An example of these technologies is 3D XPoint, which promises to fill the gap between memory and traditional storage and offers unprecedented high throughput for nonvolatile data. In this article, we characterize the performance of persistent memory devices that use the 3D XPoint technology, in the context of the DAQ system for one large Particle Physics experiment, DUNE. This experiment must be capable of storing, upon a specific signal, incoming data for up to 100 s, with a throughput of 1.5 TB/s, for an aggregate size of 150 TB. The modular nature of the apparatus allows splitting the problem into 150 identical units operating in parallel, each at 10 GB/s. The target is to be able to dedicate a single CPU to each of those units for DAQ and storage.

Index Terms—Buffer storage, data acquisition (DAQ), nonvolatile memory (NVM), software performance.

I. INTRODUCTION

OVER the last years, large-scale computing systems are moving toward a direction where compute and storage capabilities are decoupled [2], [3]. In fact, high-performance computing environments are experiencing a huge increase in data volumes, which makes data consumption more difficult for the computing nodes. By decoupling compute and storage, it is, therefore, possible to scale the system by increasing the storage hardware when more data are produced.

In the context of data acquisition (DAQ) systems for Physics experiments, the decoupling of acquisition and processing is particularly interesting in those cases in which the acquisition rate may vary widely in time depending on the physical processes being measured: an intermediate storage element allows to dimension the data processing part of the system for an average load without needing to sustain temporary peaks.

Emerging high-performance storage technologies are being used in the design of new distributed DAQ system architectures where data production and data processing are decoupled by a large storage buffer. An example of these technologies is

3D XPoint, which promises to fill the gap between memory and traditional storage. Other high-performance storage media include novel 3D XPoint SSDs [4] or the new generation of SSD controllers utilizing the PCIe Gen 4 interface [5], but they are not the topic of this particular study.

One possible application of 3D XPoint devices is in the context of the DUNE detector [6]. This is a long-baseline neutrino experiment due to start commissioning in 2026. The DAQ of the DUNE experiment will consist of a large-scale distributed system designed to handle a total of 1.5 TB/s of incoming data from the readout system. The DUNE baseline design takes advantage of 150 custom PCIe cards (FELIX readout from the ATLAS experiment at CERN [7]), each receiving data over ten 10-Gb/s optical links and streaming data into the host memory at a rate of 1 GB/s per link. Data are buffered in DRAM until either a trigger signal is received and selected data are sent out to the event builders over the network or data have aged beyond O (10 s). The current system foresees two FELIX devices for each dual-socket server. The server is called a readout unit (RU).

Upon the arrival of a rare trigger signal (supernova burst (SNB) candidate [8]), a high throughput data path to storage is activated in the RUs: it allows saving in the order of 100 s of continuous data, which then will be forwarded at a slower pace to the event builders. Persistent data buffering is required because of the value of those data and the fact that their transfer over the network will take several hours. After storage, a power cut, a server reboot, or an application crash will not cause any data loss.

The trigger is configured with thresholds such that statistical fluctuations will fire the SNB trigger in the order of once per month. The real physical process is much more rare [9]; 100 s of data correspond to 1 TB for each CPU of the RU, and the requirement is to be able to store up to two such data complements. Therefore, each RU will need in the order of 4 TB of usable storage space. Further details for the DUNE DAQ can be found here [8].

Given the importance of the data being recorded, DRAM technology is not a viable solution for the DUNE storage buffer as it cannot provide storage persistence. In addition, the total buffer size needed for each server would make the system too costly with only DRAM modules. One possible way to achieve the target objective for the system is to use fast storage media. An example of such devices is the Intel Optane Data Center Persistent Memory Modules (DCPMMs) that leverage the 3D XPoint technology. These devices were chosen because they provide throughput and capacity in the

target range that is needed for the DUNE storage buffer. COTS solutions are evaluated at this stage to ensure that there is no need for investing effort into custom hardware developments. The experiment will be built several years from now: thus, identifying a solution that already today provides performance close to the required one is sufficient.

In this article, we assess the raw storage performance of DCPMMs. A careful evaluation is done focusing on both the throughput and the request rate. Measurements are executed with both synthetic benchmarks and with a custom-made high-level application.

The suitability of DCPMMs for the DUNE storage buffer is explored, emulating the workflow on a prototype setup at CERN [10].

II. INTEL OPTANE

A. 3D XPoint Memory Technology

Intel Optane devices are based on the 3D XPoint memory technology. This is a new type of nonvolatile technology that offers approximately ten times higher bandwidth compared to traditional NAND-based storage media. Typically, 3D XPoint devices are also characterized by a high endurance value, which, for 512-GB DCPMMs, is around 300 PBW [11]. Nevertheless, the DCPMMs are considered in this article because of their throughput and not because of endurance. For the DUNE storage buffer, data are written continuously for only 100 s and with an expected (fake) rate of once a month. Therefore, the endurance advantages of 3D XPoint devices are not needed for this specific application; a NAND-based SSD solution would also be suitable.

3D XPoint should not be confused with 3-D NAND. In the latter, several storage cells (transistors) are stacked on top of each other to form a single structure for better space utilization, whereas, in 3D XPoint, data storage cells are treated independently. Therefore, contrary to NAND storage media, 3D XPoint devices are byte-addressable. Manufacturers have produced both solid-state drives and memory modules that use 3D XPoint technology. In this article, we will focus on the performance of the Intel Optane DCPMMs.

Nonvolatile memory (NVM) is an emerging and innovative technology, which uses the memory bus, such as commonly available DRAM modules. DCPMMs are NVM devices that use the 3D XPoint technology. They offer memory-like performance at a lower cost per gigabyte compared to DRAM. Therefore, DCPMMs are good candidates to fill the performance gap between memory and storage devices. In addition, contrary to storage media where data access is usually done with 4-KiB block size, DCPMMs fetch the data with four cache lines of 64 bytes. This results in a 256-byte load/store instruction that provides lower latency, similar to memory devices.

B. Operation Modes

From an operational point of view, DCPMMs can be configured into three modes.

- *Memory Mode:* In this mode, the DCPMMs act as a large memory pool alongside the DDR4 memory modules.
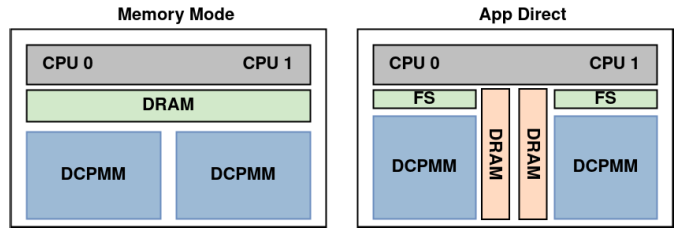

Fig. 1. Logical view of both the memory and app direct modes of operation for DCPMM devices.

Therefore, DCPMMs are seen by the operating system as a large volatile memory pool.
- *App Direct Mode:* In this mode, the DCPMMs provide in-memory persistence by acting as storage devices rather than memory devices. The memory controller maps the DCPMMs to the physical memory address space of the machine so that the software layer can directly access the devices.
- *Mixed Mode:* In this mode, it is possible to use a percentage of the DCPMMs capacity in both memory and app direct modes.

Fig. 1 shows a logical view of the two main modes of operation of DCPMMs.

In addition, in the app direct mode, the DCPMMs can be configured in two ways.
- *Interleaved Region:* All the DCPMMs relative to a CPU socket are seen as a single block device as if the modules are used in a RAID-0 configuration.
- *Noninterleaved Region:* Each DCPMM is seen as a single block device. Therefore, each module is accessed independently.

Finally, depending on the configuration, it is possible to mount the DCPMMs with a *direct access* file system (DAX). This provides byte-addressable access to the storage without the need to perform an extra copy on the page cache, and thus, it yields higher read and write bandwidths. In this article, the objective is to understand the throughput of DCPMMs when used as a nonvolatile storage media, and therefore, tests done on other operation modes are not going to be discussed.

III. RELATED WORK

Several research institutes have already tested the Intel Optane DCPMMs in different configurations. Some of the most recent and complete research papers can be found in [12] and [13].

In our research, the objective is to assess the performance of DCPMMs from an application point of view. Therefore, we decided not to rely only on low-level benchmarking tools, but, instead, we developed a high-level application, especially designed for a high throughput use-case, which leverages the DCPMM technology.

Other research groups have also tested DCPMMs from an application perspective. However, rather than focusing on high throughput applications, the typical use case of DCPMMs that is found in literature is to accelerate, for example, database workloads [14], or perform faster graph analytics [15].

TABLE I

OVERVIEW OF THE TEST MACHINE USED FOR EVALUATION

CPU	Intel® Xeon® Platinum 8280L 2.70 GHz (Cascade Lake), dual socket L1d cache 32K L2 cache 1024K L3 cache 3942K
DRAM	DDR4 DRAM 16 GB, 2666 MT/s, 12 slots Product number: Kingston KSM26RS4
DCPMM	DDR-T 512 GB, 2666 MT/s, 12 slots Product number: Intel®NMA1XBD512GQS
OS	CentOS 7, Linux Kernel 4.15.0
SW	ipmctl v.01.00, PMDK v.1.9

IV. EVALUATION

A. System Description

Table I summarizes the specification of the machine node used for the evaluation. The test machine is a dual CPU socket system with 56 physical cores on each processor and one memory controller per socket. Each memory controller has six memory channels composed of both a DDR4 DRAM device (16 GB) and a DCPMM (512 GB). The total DRAM size of the machine is 192 GB, whereas the total DCPMM size is 6 TB. Finally, the node is installed with a CentOS 7 operating system and kernel version 4.15.

B. Testing Strategy

The raw performance of the persistent memory devices was obtained by executing a synthetic benchmark evaluation with the DCPMMs used as a storage device in the app direct mode and mounted with a DAX-enabled ext4 file system. The system was tested with a high-level C++ application that used the DCPMMs as the target storage media, and the throughput has been measured as a function of both the block size and the number of threads. The resulting performance was obtained in terms of request rate and total throughput. The results obtained were reproducible across several runs. However, a complete analysis of the uncertainty on the individual measurements was not computed as it was not relevant for the objectives of this research. The benchmarks executed on the system refer to the six DCPMMs connected to the same CPU socket, unless otherwise stated.

V. BENCHMARKS OF THE DCPMMS AND DISCUSSION

Fig. 2 represents the request rate as a function of the I/O block size for different writing threads. This was obtained by measuring the number of I/O operations per second for a given workload in terms of the block size and the number of threads. Increasing the block size results in a smaller request rate due to the increased latency to fetch and store the requested block. The maximum rate sustained by a single writing thread is approximately 200k operations/s. This showcases the small time needed to request data with DCPMMs.

Fig. 3 shows the throughput as a function of the number of threads in the case of both the reading and writing
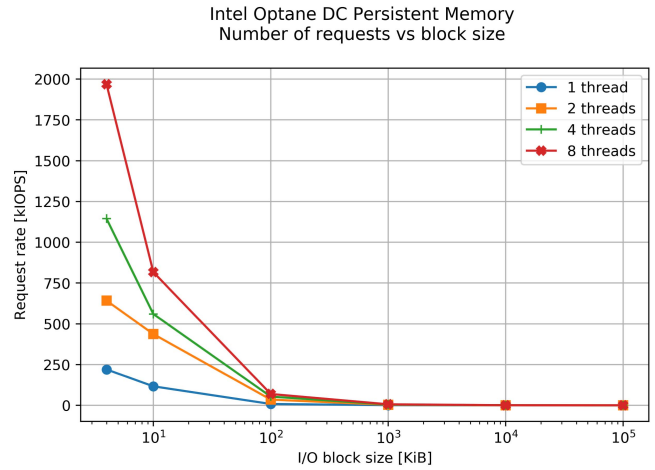

Fig. 2.   Request rate as a function of the I/O block size for different writing threads. Increasing the block size results in a smaller request rate due to the increased latency to fetch and store the requested block.
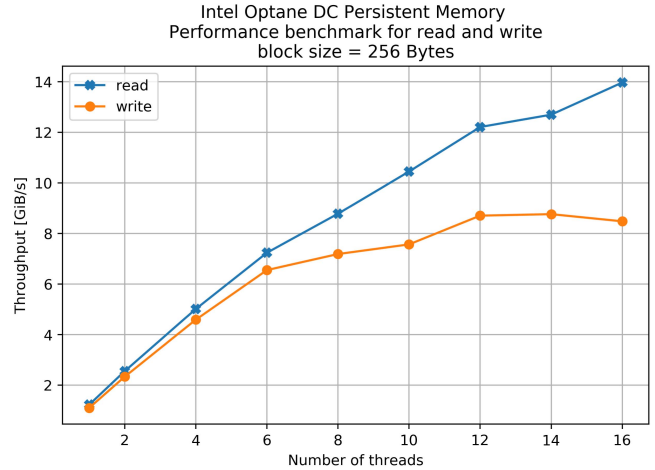

Fig. 3.   Throughput as a function of the number of threads for a block size of 256 bytes in the case of both sequential reading and sequential writing. The maximum write throughput obtained from the application is approximately 8.5 GiB/s, whereas, for reading, a plateau has not been reached even with 16 threads.

access pattern. The block size used for the benchmark is 256 bytes because it represents the lowest access granularity for DCPMM devices. These results were obtained by executing a writing thread on the DCPMMs using an interleaved region and measuring the time taken to write the selected data block size. Note that CPU affinity was set up on the host in order to restrict the executing threads to the physical cores of the same NUMA node of the DCPMMs. This was done to avoid any cross-NUMA access that leads to increased latency and, therefore, lower performance. In addition, the writing thread was executed by memory mapping the block of data and then using the MOVNTI [16] nontemporal SSE instruction [17] provided by the processor. In this way, the operation has no overhead from the file system because it invalidates the cache line and, therefore, results in a pure device operation.
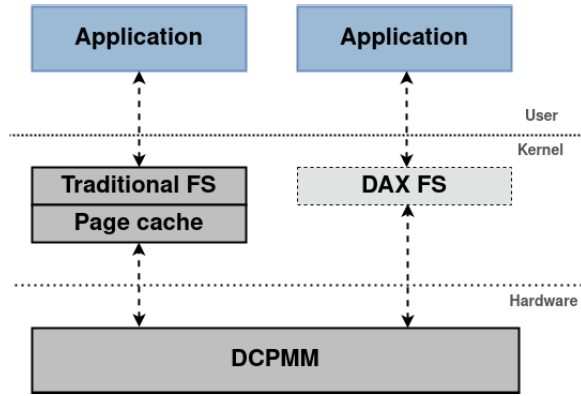
Fig. 4.   Schematic of an application running on DCPMMs with a traditional file system and with a DAX-enabled file system.
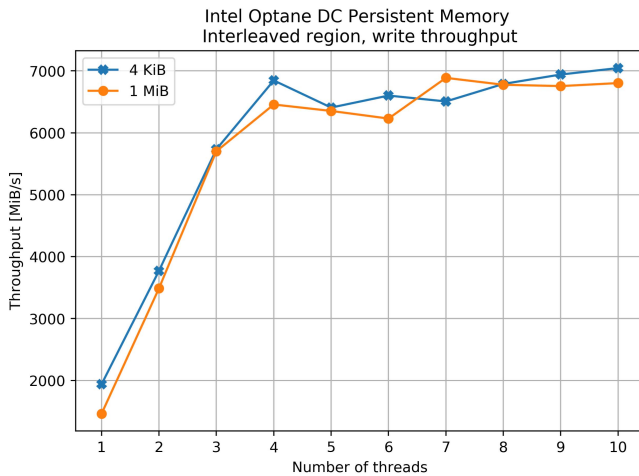

Fig. 5.   Throughput as a function of the number of threads for both 4-KiB and 1-MiB block sizes. Very limited variability is observed because of the memory-like behavior of DCPMMs.

This feature is possible because it is allowed by DAX-enabled file systems. In this way, it is possible to achieve higher bandwidths. Fig. 4 shows a schematic representation of the difference between running an application on a traditional file system compared to a DAX-enabled file system.

From Fig. 3, the maximum write throughput obtained from the application is approximately 8.5 GiB/s. In the case of the reading access pattern, the throughput is higher, and it has not reached a plateau even when testing with 16 threads. This is because DCPMMs are memory devices that have lower reading latencies compared to the writing operation. Therefore, it is possible to achieve higher bandwidths. As shown in [12], in the best case configuration, the maximum achieved throughput for a read workload is approximately 40 GiB/s.

Another interesting feature of DCPMMs is the independence with the access block size. This is illustrated in Fig. 5, which shows the writing throughput as a function of the number of threads for two block sizes, respectively, 4 KiB and 1 MiB. It can be noted that there is less than 5% variability in the throughput when using a block size of 4 KiB and 1 MiB. This behavior is typical of memory devices, and it confirms again the memory-like behavior of DCPMMs.

TABLE II
BANDWIDTH FOR A 100% SEQUENTIAL READ AND A 100% SEQUENTIAL WRITE WORKLOAD FOR BOTH A NAND-BASED SSD AND DCPMMs

	Bandwidth NAND-SSD [GiB/s]	Bandwidth DCPMMs [GiB/s]
Sequential read	3.2	40
Sequential write	1.9	9

Finally, a comparison between the maximum bandwidths provided by both DCPMMs and a NAND-based SSD is illustrated in Table II. The NAND-based SSD is a PCIe Gen 3 Intel SSD DC P4510 (2 TB). The results obtained for the NAND SSD are also confirmed on the device's technical datasheet [18]. Note that the sequential read bandwidth for DCPMMs is taken from [12]. This shows that, in order to achieve a write bandwidth similar to the one provided by the DCPMMs, it would be necessary to use almost five NAND-based SSDs. It is also worth mentioning that the new generation of SSD devices utilizing the PCIe Gen 4 interface [5] is capable of sustaining higher bandwidths; therefore, fewer devices are needed to match the performance of DCPMMs. A comparison between DCPMMs and NVMe storage media for a possible application within the DUNE experiment will be presented in future contributions.

VI. APPLICATION FOR THE DUNE DATA ACQUISITION SYSTEM AND DISCUSSION

One of the research objectives of the DUNE experiment is to detect neutrino signals originating from astrophysical sources, such as SNB events. In the DAQ system of the DUNE detector, data are continuously stored in RAM in circular buffers for 10 s before being transferred to permanent storage. Each RU generates approximately 10 GB/s, per CPU socket, from ten separate threads, and upon receiving a trigger signal, data need to be stored for 100 s [8]. This means that the DUNE DAQ system needs a storage technology capable of sustaining a writing rate of approximately 10 GB/s. Fig. 6 shows how the size of the (volatile) memory buffer would need to increase as a function of time depending on different writing bandwidths (output rate) of storage technologies. As an example, if the writing bandwidth of the selected storage technology is only 5 GiB/s, this means that the total physical memory of the system needs to be increased to 500 GiB in order to keep up with the input rate and for the total time of 100 s. As a consequence, it is necessary to find a suitable storage technology that is capable of sustaining the target rate of 10 GB/s in order to minimize the extra volatile memory needed to keep the data.

Based on the synthetic benchmarks obtained in Section V, the DCPMMs represent a good candidate for the DUNE supernova storage system. A test application was developed with the DCPMMs integrated into a DUNE prototype setup, which is available at CERN (ProtoDUNE). The details of the ProtoDUNE readout system are described in detail here [10]. The system was equipped with six DCPMMs configured in an interleaved configuration. Fig. 7 illustrates the performance obtained with the test application. The figure shows the average throughput for an increasing number of threads and
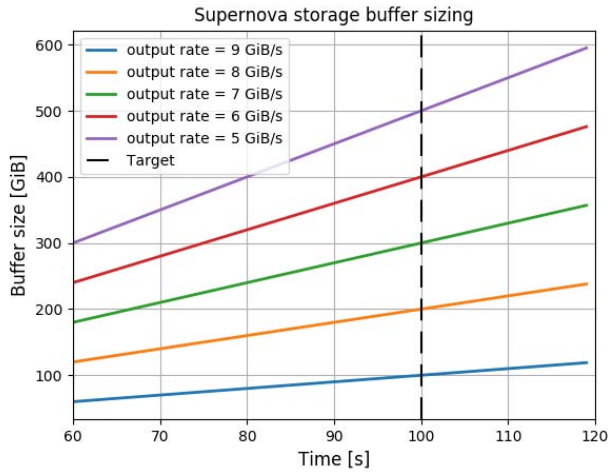
Fig. 6. Amount of DRAM as a function of the data recording time for different output rates. Decreasing the output rate requires an increased memory size. A vertical line representing the target of 100 s is also included.
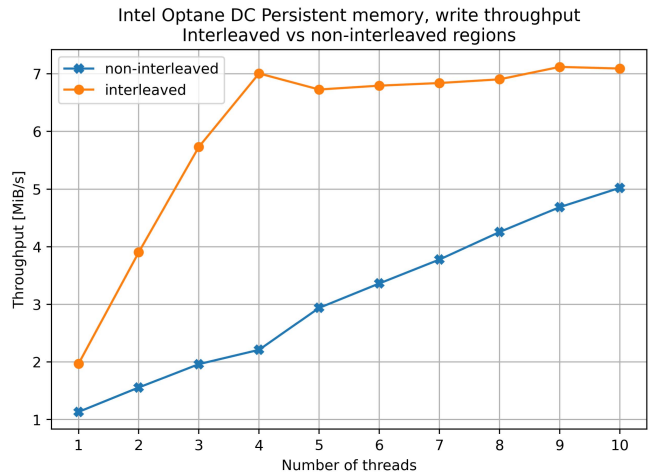

Fig. 8. Write throughput as a function of number of threads for both interleaved and noninterleaved DCPMM configurations for a DUNE-like application. The maximum throughput is obtained at 7 GiB/s starting from four writing threads.
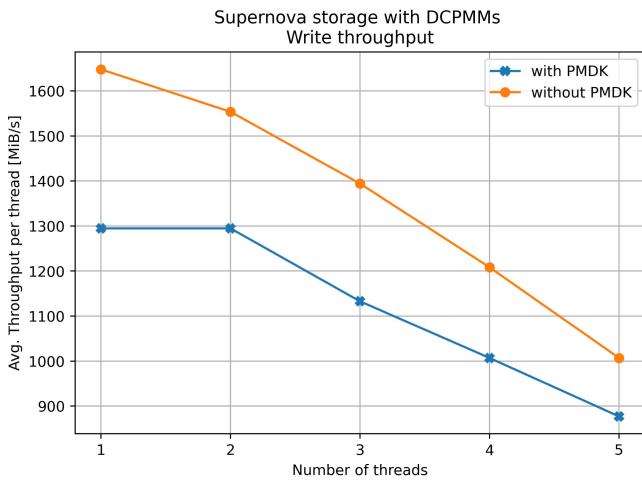

Fig. 7. Average write throughput per thread as a function of the number of threads for the ProtoDUNE test application with DCPMMs. Results are obtained with and without the PMDK software.

a particular block size of 5568 bytes. This is the access size used in the prototype setup to transfer and store the data from the custom PCIe devices of the readout system. The test application has been written using the Persistent Memory Development Kit (PMDK) [19], which is a collection of libraries and tools that ease the development of applications that use persistent memory devices. The result obtained in the test application is satisfactory for up to four threads because the average throughput per thread can sustain the target bandwidth of 1 GB/s. However, as the threads' number increases, the application cannot keep up with the incoming rate.

This leads us to optimize the software stack of the test application in order to fully exploit the performance provided by DCPMMs. It was noticed that the *libpmemblk* library of the PMDK tool was adding extra overheads to the application because of features such as block-level atomicity in the case of errors or power failures. Similar to what has been described in Section V, a lower level application was developed by creating memory-mapped files and then persisting them using

the MOVNTI instruction. In this way, there is no extra overhead from the file system, and the performance obtained is higher. By optimizing the software application, the throughput obtained in Fig. 7 without the use of PMDK increased by approximately 20%.

The second application was also developed with the objective of resembling as much as possible the DUNE workload: a traffic pattern is generated by memory copying data into volatile memory and then, upon receiving a command, persisting the data into the storage media. All the memory modules available on the system on both sockets have been deployed in the app direct mode. Fig. 8 illustrates the throughput obtained as a function of the number of executing threads for both the interleaved and noninterleaved DCPMM configurations. The system saturates the available bandwidth with a throughput of 7 GiB/s starting from four threads. This means that, with the current DCPMMs available today, it is possible to sustain, per CPU socket, 80% of the target throughput required for the DUNE supernova storage buffer. However, the new generation of DCPMMs (Intel Optane Persistent Memory series 200) [20] should give on average 25% more bandwidth, and therefore, they can fill the required performance that is needed to fully build the DUNE supernova storage buffer.

The noninterleaved configuration was also tested because it represents a good match for the DUNE supernova workload. In fact, the ten writing threads required can be considered independent and, therefore, can be configured to write to ten different block devices. However, as shown in Fig. 8, the throughput obtained in this operational mode is much lower. In the case of five writing threads, the throughput in the noninterleaved configuration is approximately 60% lower than in the corresponding interleaved DCPMM region. This behavior confirms that the interleaved configuration is the most suitable for high-performance applications, and it suggests that DCPMMs in this configuration have internal mechanisms to optimally balance the I/O operations to achieve the best performance.

VII. Conclusion

This work has shown the potential provided by the NVM technology. Intel Optane DC Persistent Memory modules have been tested in detail with both synthetic benchmarks and with a custom-made high-level application. The objective was to understand the different operational modes of the technology and investigate the maximum (write) bandwidth that the DCPMM devices can sustain. The ultimate objective was to assess whether DCPMMs could be a viable technology for the implementation of the DUNE supernova storage buffer.

The synthetic benchmarks executed on the system showed that DCPMMs are capable of sustaining high request rates when used as storage devices (app direct mode). The bandwidth of the system was measured for both the writing and reading access patterns. It was shown that the maximum achieved writing throughput is approximately 8.5 GiB/s starting from four threads. In addition, the memory-like nature of the DCPMMs was confirmed by measuring the throughput as a function of two different block sizes, respectively, 4 KiB and 1 MiB. It was noticed that the system behaves independently of the access size and, thus, confirms the memory nature of the DCPMMs.

A high-level application that leverages DCPMMs was also developed and integrated with a prototype setup in order to validate its potential use for the DUNE supernova storage buffer. It was noticed that the overall performance was affected by the overhead added by the PMDK software library. Therefore, software optimizations that take advantage of the MOVNTI instruction set were included, and the resulting performance increased by approximately 20%. A more realistic application resembling the DUNE supernova workload was developed, and it was shown that, by using all the DCPMMs in the interleaved configuration, it is possible to sustain 80% of the required throughput.

Future directions of this work will consist of further optimizing the software application to reduce all the overheads and get the maximum throughput from the DCPMMs. In addition, the new generation of Intel Optane Persistent Memory devices could offer a substantial increase in bandwidth, which would make possible the deployment of the DUNE supernova storage buffer with DCPMMs. Other research developments will also focus on the evaluation of other storage technologies, such as arrays of NAND-based SSDs.

Acknowledgment

References

[1] A. Di Meglio, M. Girone, A. Purcell, and F. Rademakers, "CERN openlab white paper on future ICT challenges in scientific research," Eur. Lab. Part. Phys., Geneva, Switzerland, Tech. Rep., Jan. 2018, doi: 10.5281/zenodo.998694.

[2] A. Verbitski et al., "Amazon aurora: Design considerations for high throughput cloud-native relational databases," in Proc. ACM Int. Conf. Manage. Data, May 2017, pp. 1041–1052, doi: 10.1145/3035918.3056101.

[3] B. Dageville et al., "The snowflake elastic data warehouse," in Proc. Int. Conf. Manage. Data, Jun. 2016, pp. 215–226, doi: 10.1145/2882903.2903741.

[4] Micron. X100 NVME SSD. Accessed: Feb. 22, 2021. [Online]. Available: https://www.micron.com/products/advanced-solutions/3d-xpoint-technology/x100

[5] P. Electronics. Ps5016-e16 gen4x4 NVME SSD Controller. Accessed: Feb. 25, 2021. [Online]. Available: https://www.phison.com/en/technologies-gen4/pcie-gen4-awareness/1149-ps5016-e16

[6] B. Abi et al., "Deep underground neutrino experiment (DUNE), far detector technical design report, volume I introduction to DUNE," J. Instrum., vol. 15, no. 8, 2020, Art. no. T08008.

[7] J. Anderson et al., "FELIX: A PCIe based high-throughput approach for interfacing front-end and trigger electronics in the ATLAS upgrade framework," J. Instrum., vol. 11, no. 12, Dec. 2016, Art. no. C12023.

[8] B. Abi et al., "Volume IV. The DUNE far detector single-phase technology," J. Instrum., vol. 15, no. 8, Aug. 2020, Art. no. T08010, doi: 10.1088/1748-0221/15/08/t08010.

[9] K. Rozwadowska, F. Vissani, and E. Cappellaro, "On the rate of core collapse supernovae in the milky way," New Astron., vol. 83, Feb. 2021, Art. no. 101498, doi: 10.1016/j.newast.2020.101498.

[10] R. Sipos, "The DAQ for the single phase DUNE prototype at CERN," Eur. Lab. Part. Phys., Geneva, Switzerland, Tech. Rep., 2018, doi: 10.1109/TNS.2019.2906411.

[11] Intel. Intel Optane DC Persistent Memory Data Sheet. Accessed: Feb. 22, 2021. [Online]. Available: https://www.intel.la/content/dam/www/public/us/en/documents/product-briefs/optane-dc-persistent-memory-brief.pdf

[12] J. Izraelevitz et al., "Basic performance measurements of the intel optane DC persistent memory module," Dept. Comput. Sci. Eng., Univ. California, San Diego, CA, USA, 2019, doi: 10.1587/transinf.2019EDL8141.

[13] T. Hirofuchi and R. Takano, "A prompt report on the performance of intel optane DC persistent memory module," Nat. Inst. Adv. Ind. Sci. Technol., Tokyo, Japan, 2020, doi: 10.1587/transinf.2019EDL8141.

[14] Y. Wu, K. Park, R. Sen, B. Kroth, and J. Do, "Lessons learned from the early performance evaluation of intel optane DC persistent memory in DBMS," in Proc. 16th Int. Workshop Data Manage. New Hardw., Jun. 2020, pp. 1–3.

[15] G. Gill, R. Dathathri, L. Hoang, R. Peri, and K. Pingali, "Single machine graph analytics on massive datasets using intel optane DC persistent memory," Univ. Texas Austin, Austin, TX, USA, Intel Corp., Santa Clara, CA, USA, Tech. Rep., 2019, doi: 10.14778/3389133.3389145.

[16] C. Kalita, G. Barua, and P. Sehgal, "Durablefs: A file system for persistent memory," Indian Inst. Technol. Guwahati, Guwahati, India, Netapp India, Bangalore, India, Tech. Rep. arXiv:1811.00757, 2018.

[17] S. K. Raman, V. Pentkovski, and J. Keshava, "Implementing streaming SIMD extensions on the pentium III processor," IEEE Micro, vol. 20, no. 4, pp. 47–57, Jul. 2000.

[18] Intel SSD DC P4510, Intel Corp., Santa Clara, CA, USA, 2019.

[19] Intel Corporation, Intel Corp., Santa Clara, CA, USA, 2020.

[20] Intel Optane Persistent Memory 200 Series, Intel Corp., Santa Clara, CA, USA, 2020.