Analyzing the Performance of Intel Optane DC Persistent Memory in Storage over App Direct Mode

Introduces DCPMM Storage over App Direct Mode

Evaluates workload latency and bandwidth performance

Establishes performance expectations for DCPMM capacities

Illustrates scaling performance for DCPMM populations

Travis Liao

Tristian "Truth" Brown

Jamie Chou

LENOVO PRESS

Abstract

Intel Optane DC Persistent Memory is the latest memory technology for Lenovo® ThinkSystem™ servers. This technology deviates from contemporary flash storage offerings and utilizes the ground-breaking 3D XPoint non-volatile memory technology to deliver a new level of versatile performance in a compact memory module form factor.

As server storage technology continues to advance from devices behind RAID controllers to offerings closer to the processor, it is important to understand the technological differences and suitable use cases. This paper provides a look into the performance of Intel Optane DC Persistent Memory Modules configured in Storage over App Direct Mode operation.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction

There is a large performance gap between DRAM memory technology and the highest performing block storage devices currently available in the form of solid-state drives. Capitalizing on this opportunity, Lenovo partnered with Intel, a key technology vendor, to provide the end customer with a novel memory module solution called Intel Optane DC Persistent Memory.

Intel Optane DC Persistent Memory provides unique levels of performance and versatility because it is backed by Intel 3D XPoint non-volatile memory technology instead of traditional NAND based flash. This technology has various implementations, however this paper will focus solely on the performance of Intel Optane DC Persistent Memory when run in “Storage over AppDirect” Mode operation.

Intel Optane DC Persistent Memory and it’s implementation, the DC Persistent Memory module (DCPMM) is a byte addressable cache coherent memory module device that exists on the DDR4 memory bus and permits Load/Store accesses without page caching.

DCPMM creates a new memory tier between DDR4 DRAM memory modules and traditional block storage devices. This permits DCPMM devices to offer memory bus levels of performance, and allows application vendors to remove the need for paging, context switching, interrupts and background kernel code running.

DCPMMs can operate in three different configurations, Memory Mode, App Direct Mode, and Storage over App Direct Mode. This paper will focus on DCPMM devices in Storage over AppDirect Mode operation and its associated performance.

Figure 1 on page 3 shows the visual differences between a DCPMM and a DDR4 RDIMM. DCPMM devices physically resemble DRAM modules because both are designed to operate on the DDR4 memory bus. The uniquely identifying characteristic of a DCPMMs is the heat spreader that covers the additional chipset.

Figure 1 DCPMM (top) and a DDR4 RDIMM (bottom)

DCPMM modules can operate up to a maximum DDR4 bus speed of 2666MHz and are offered in capacities of 128GB, 256GB, and 512GB. The 128GB DCPMM devices can operate up to a maximum power rating of 15W whereas the 256GB and 512GB DCPMM devices can operate up to a maximum power rating of 18W.

Due to the calculation method and needed overhead for DCPMM device operation the actual usable capacity is slightly less than the advertised device capacity. Table 1 lists the expected DCPMM capacity differences as seen by the operating system.

Table 1 DCPMM advertised capacity relative to usable capacity in operating systems

Advertised DCPMM Capacity	Available DCPMM Capacity
128 GB	125 GB
256 GB	250 GB
512 GB	501 GB

DCPMM Storage over App Direct Mode Configuration Rules

The basic rules for installing DCPMM into a system are as follows:

- A maximum of 1x DCPMM device is allowed per memory channel
- DCPMM devices of varying capacity cannot be mixed with a system
- For each memory channel, DCPMM devices should be installed in the memory slot physically closest to the CPU unless it is the only DIMM in the memory channel

Figure 2 on page 5 shows a close-up of the SR950 system board, showing one second-generation Intel Xeon Scalable Processor with six DCPMMs and six DIMMs installed into the memory slots connected to the processor. The processor has two memory controllers, each providing three memory channels and each memory channel containing two DIMM slots.

As shown, the twelve modules installed are comprised of six RDIMMs and six DCPMM devices, with each DCPMM located in the memory slot electrically (and physically) closer to the processor for each memory channel.

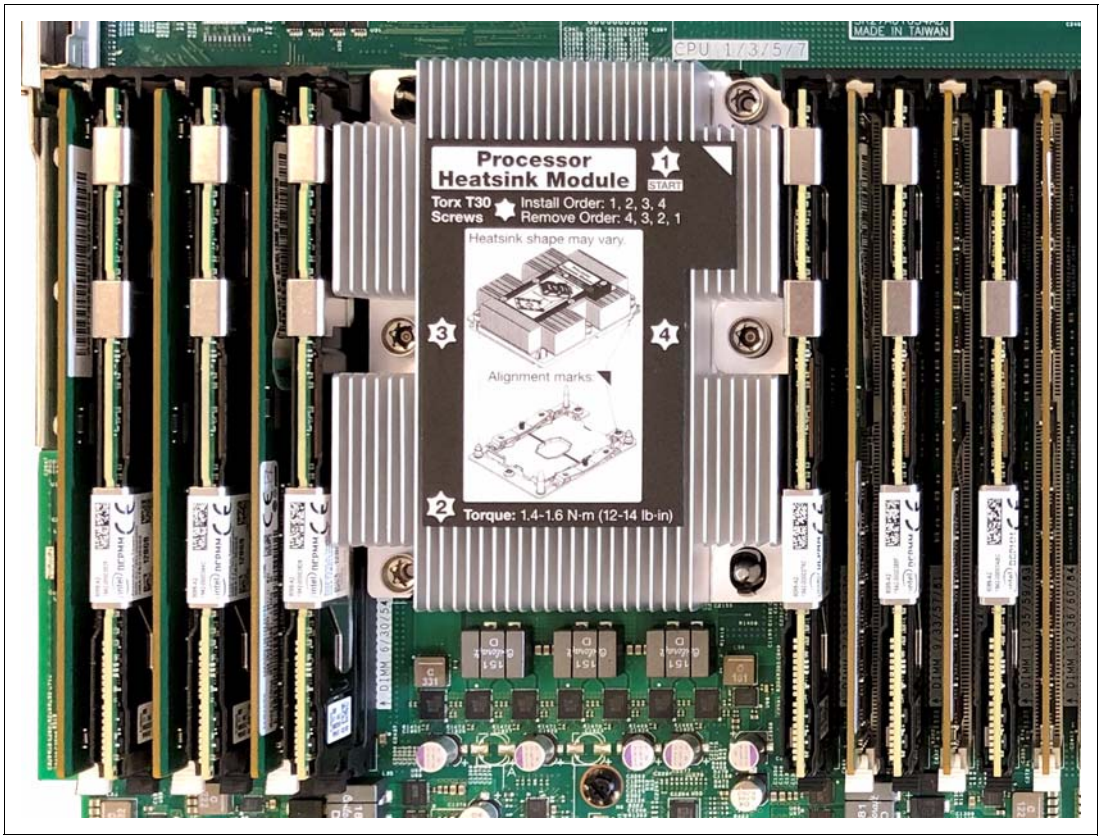

Figure 2 Intel Xeon Scalable Processor with 6 DCPMMs and 6 RDIMMs (SR950)

Storage over App Direct Mode

Storage over App Direct Mode operation is when DCPMMs are configured as accessible block storage devices residing on the memory bus. In this configuration, DCPMMs operate as conventional block storage; therefore software modifications are not needed for an application to access the storage pool.

The only functional requirement to use Storage over App Direct Mode is that the administrator install and configure specific DCPMM rpm packages:

- impctl, available from `https://github.com/intel/ipmctl`
- ndctl, available from `https://github.com/pmem/ndctl`

ipmctl is an utility to configure and manage Intel DCPMM. For essential functions such as discover, update firmware, provision Intel DCPMM and so on, please refer to the documentation for the commands:

`https://github.com/intel/ipmctl`

ndctl is the utility for management of the "libnvdimm" kernel subsystem in the Linux kernel. It is an essential tool to provision DCPMM namespaces. After successful configuration, ndctl will expose available capacity from a DCPMM configuration to the operating system.

When Intel DCPMM is used as a traditional block storage device; it is recommended to utilize DAX (Direct Access) mode to provide optimal throughput with namespaces configured using the ndctl utility. DAX mode is designed to avoid extra copy on memory by direct read/write

into persistent memory. Memory mapped sections map directly to Intel DCPMM on volumes create with DAX mode.

Intel DCPMM AppDirect mode can be configure as interleaved or non-interleaved.

- In interleaved mode, Intel DCPMM devices will be grouped together as a single logical volume per socket.
- In non-interleaved mode, the volumes are separated by each physical device. When configured as interleaved the incoming data is stripped across multiple devices which allows the throughput performance to scale up to each attached Intel DCPMM.

Performance analysis

This section describes the results of our analysis of performance of Storage over App Direct Mode.

Hardware configuration evaluation environment

To evaluate the performance of DCPMMs in Storage over App Direct Mode, we used Flexible I/O (FIO) because it is a well-established industry performance evaluation tool. FIO is compatible with DCPMM in Storage over App Direct Mode operation and has the ability to stress and measure I/O performance.

The following charts show FIO performance data to quantify throughput and latency performance. This evaluation data is based on a two-socket ThinkSystem server configured as listed in Table 2.

Table 2 Server configurations for two-socket Storage over App Direct Mode evaluations

Configurations	**1-1-1 with DRAM and NVMe SSDs only (no DCPMMs)**	**2-2-2 with DCPMMs**	**2-2-1 with DCPMMs**	**2-1-1 with DCPMMs**	**1-1-1 with DCPMMs**
CPU	Intel Xeon Platinum 8168	Intel Xeon Platinum 8280L	Intel Xeon Platinum 8280L	Intel Xeon Platinum 8280L	Intel Xeon Platinum 8280L
Operating system	RHEL 7.6	RHEL 7.6	RHEL 7.6	RHEL 7.6	RHEL 7.6
DRAM	12x 16GB RDIMM, 2666 MHz	12x 16GB RDIMM, 2666 MHz	12x 16GB RDIMM, 2666 MHz	12x 16GB RDIMM, 2666 MHz	8x 16GB RDIMM, 2666 MHz
Storage	8x 375GB Intel P4800X NVMe SSD	12x 512GB DCPMM	8x 512GB DCPMM	4x 512GB DCPMM	4x 512GB DCPMM

Scaling analysis and maximum performance

DCPMM modules are offered in three capacities with the following maximum power ratings:

- 128 GB DCPMM: 15W

- 256 GB DCPMM: 18W
- 512 GB DCPMM: 18W

The higher operating power threshold enables the 256 GB and 512 GB capacity DCPMMs to provide greater throughput performance for equivalent hardware configurations.

The 256GB and 512GB capacities operate at 18W vs 15W for the 128GB capacity. Therefore the higher power threshold allows for greater throughput performance for equivalent hardware configuration. This means that a 128GB 2-2-2 setup will have less throughput performance than a 256GB or 512GB 2-2-2 setup. Figure 3 shows the bandwidth performance differences.

DCPMMs are optimized for read intensive workloads; therefore, the read performance outperforms write performance. In general, DCPMM devices provide much higher I/O performance when compared to other block storage offerings such as NVMe, SAS and SATA SSDs.

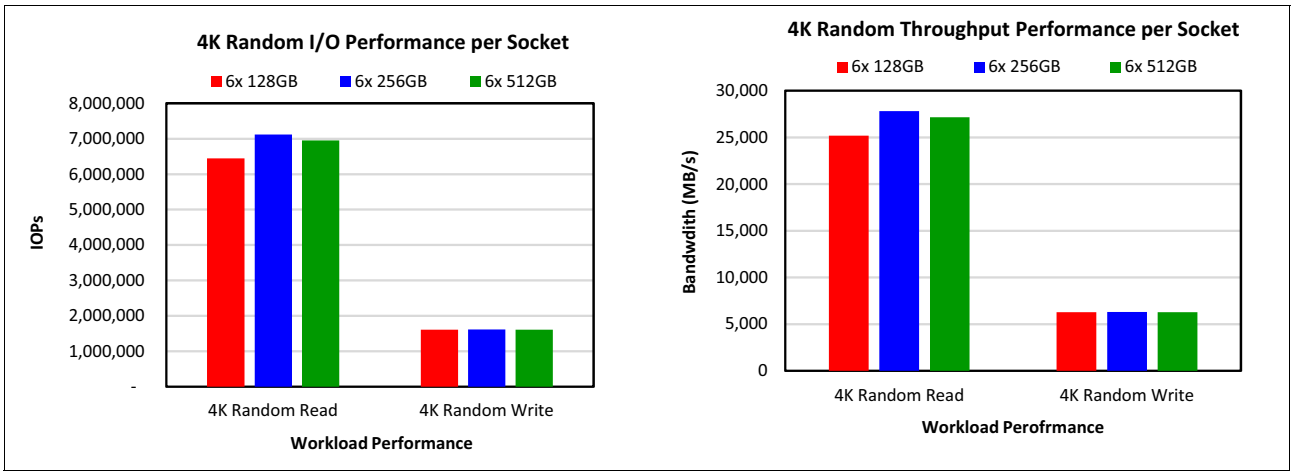

Figure 3 Single socket DCPMM scaling Random Workload I/O (left) and Bandwidth (right) Performance

Figure 4 on page 7 displays two-socket 4K random read and random write IOPS performance for multiple 512GB DCPMM configurations.

DCPMMs operate on the memory bus and are sensitive to processor thread count instead of queue depth. Therefore, the highest levels of DCPMM performance are governed by per-processor application thread count.

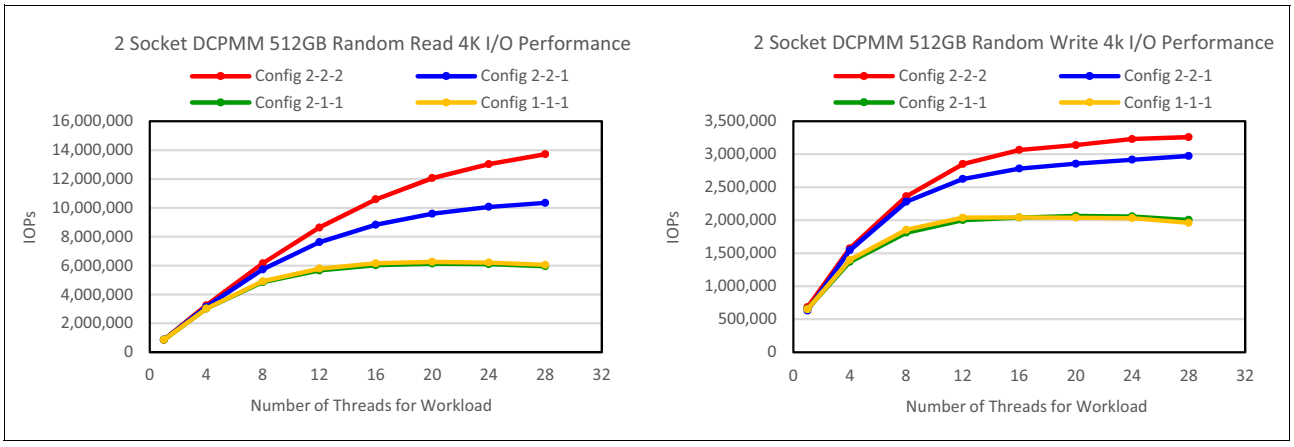

Figure 4 Two-socket DCPMM 512GB Random Read (left) Random Write (right) 4K I/O Performance

As Figure 4 shows, in Config 2-2-2 (6x DCPMM), the random read I/O performance reaches close to 14 Million IOPS while random write I/O is limited to 3.3 Million IOPS. The read performance scaling curve from the Config 2-1-1/1-1-1 (2x DCPMM) up to Config 2-2-2 (6x DCPMM) displays a progressive increase in performance with saturation occurring near maximum physical core count. Conversely, write performance saturates at a lower thread count but still provides good scaling.

DCPMM device bandwidth measurements aligns with IOPS read and write performance. Figure 5 displays similar scaling with a maximum read bandwidth of just over 50 GB/s and write bandwidth slightly under 13 GB/s for 4KB random workload.

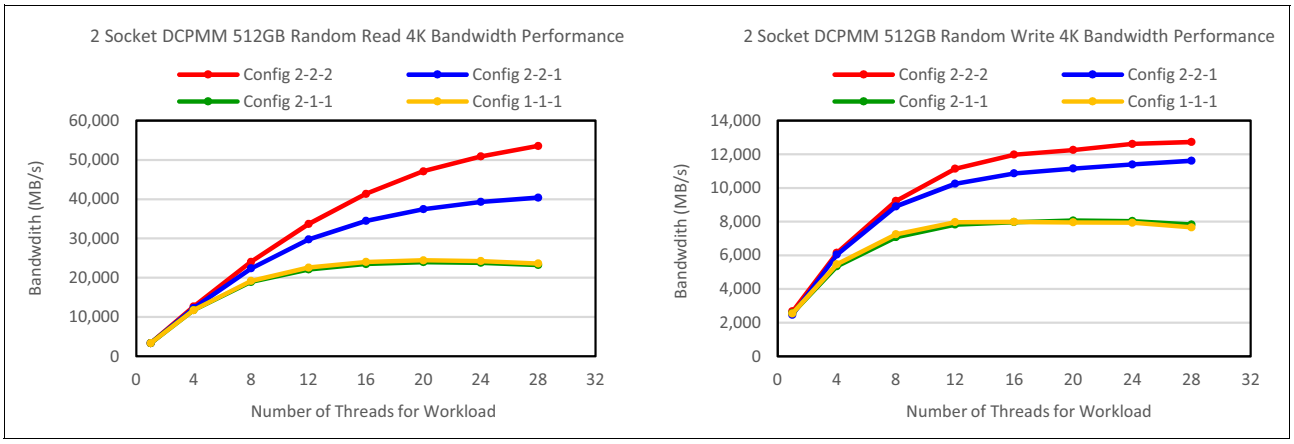

Figure 5 Two-socket DCPMM 512GB Random Read (left) Random Write (right) Bandwidth Performance

Latency performance analysis

Figure 6 demonstrates the relationship between IOPs and latency. This analysis was accomplished using 12x 512GB DCPMM devices in a 2 Socket 2-2-2 configuration.

Each data point in the respective block size data lines represent an increasing thread count of: 1, 4, 8, 12, 16, 20, 24 up to 28. The random read results display a worst case latency of 4µs for block sizes under 4K and produce 14 Million to 33 Million IOPS. For write access, DCPMM device saturation occurs at roughly 3 Million IOPS with a worst case latency of 16µs. For larger block sizes, latency performance does decrease but for most transactions will remain under 20µs.

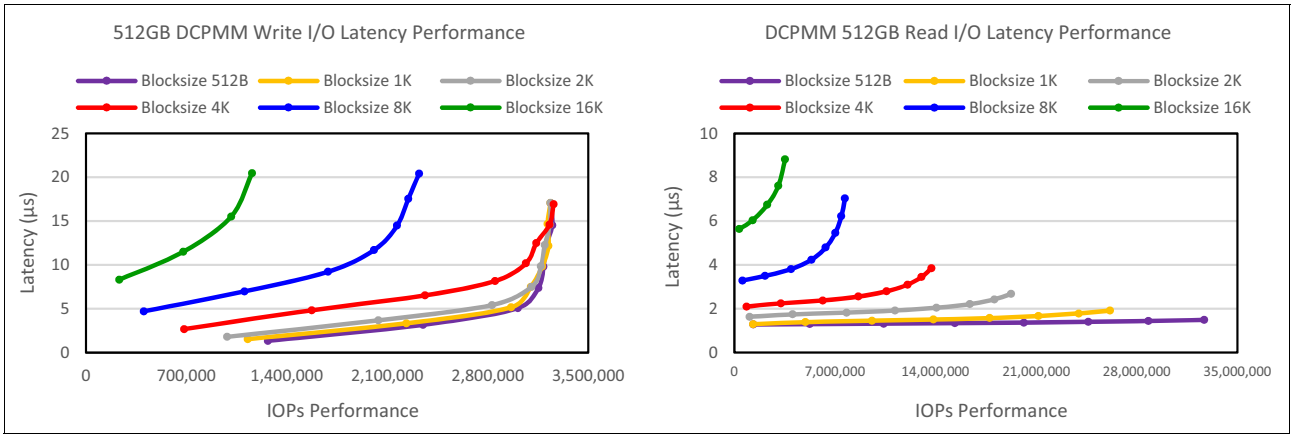

Figure 6 DCPMM 512GB Random Read (left) and Random Write (right) Latency Performance

The latency on DCPMM devices remains very low under the high-load demand of 28 threads accessing the App Direct storage pool simultaneously. This is a massive advantage compared to other high-performance storage devices. Storage over App Direct operation removes the need for layers of delays such as PCIe/SATA/SAS protocols; therefore, the latency response of DCPMM is lower than all of the currently available block storage devices.

Performance comparison between NVMe SSDs and DCPMMs

The charts in this section provide a direct comparison between Intel P4800X NVMe SSDs (375GB) and Intel DCPMMs. DCPMMs have a massive read I/O performance advantage over P4800X SSDs. In the closest comparison of 2x DCPMM (1TB) vs 2x P4800X (750GB) random read I/O performance, there is a delta of 1.9 Million IOPS in favor of DCPMM. In contrast, with random write workloads, P4800X SSDs have a performance advantage when the devices are scaled beyond 4x.

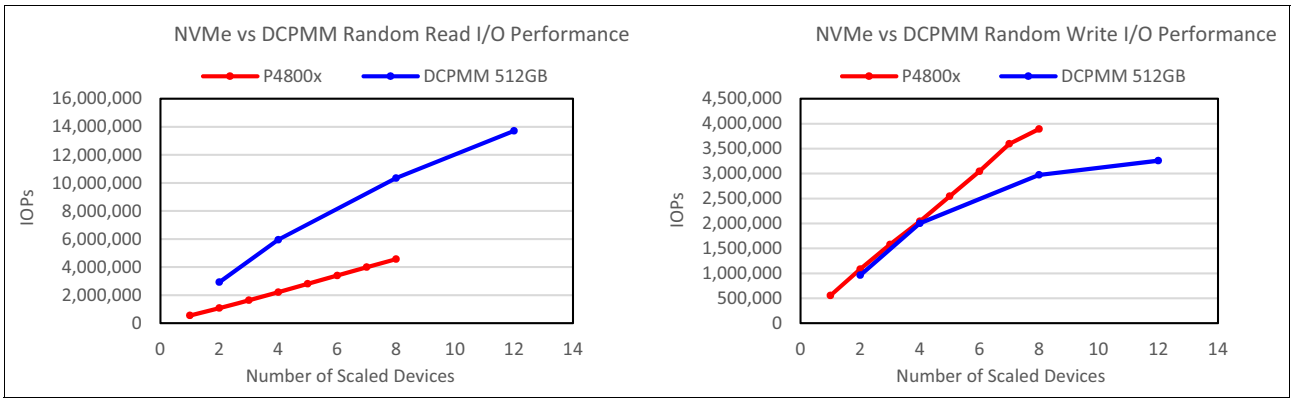

Figure 7 DCPMM device vs NVMe SSD Random Read (left) and Random Write (right) I/O Performance

Figure 8 displays that the P4800X NVMe SSD and DCPMM device bandwidth comparisons completely align with the IOPS read and write performance provided in Figure 7.

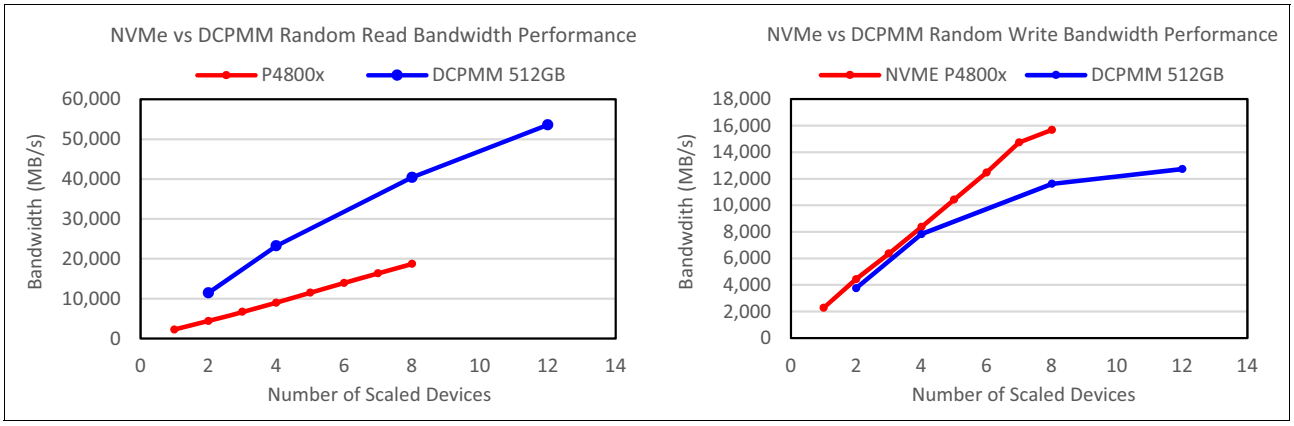

Figure 8 DCPMM device vs NVMe SSD Random Read (left) and Random Write (right) Bandwidth Performance

File System performance analysis

For DCPMM devices to function as designed, a file system needs to created and mounted using a DAX model before the storage pool can be accessed by the application. Figure 9 shows single-socket 256GB DCPMM performance difference between XFS and EXT4 file-system for read and write workloads. For best performance it is recommended to use XFS with Storage over App Direct Mode.

For DCPMM devices to function as designed, a file system needs to created and mounted using a DAX model before the storage pool can be accessed by the application. The following command is used to enable the file system with direct access capability:

```
mount -o DAX /dev/pmemx /mnt/pmemx
```

Figure 9 shows single-socket 256GB DCPMM performance difference between XFS and EXT4 file-system for read and write workloads. For best performance it is recommended to use XFS with Storage over App Direct Mode.

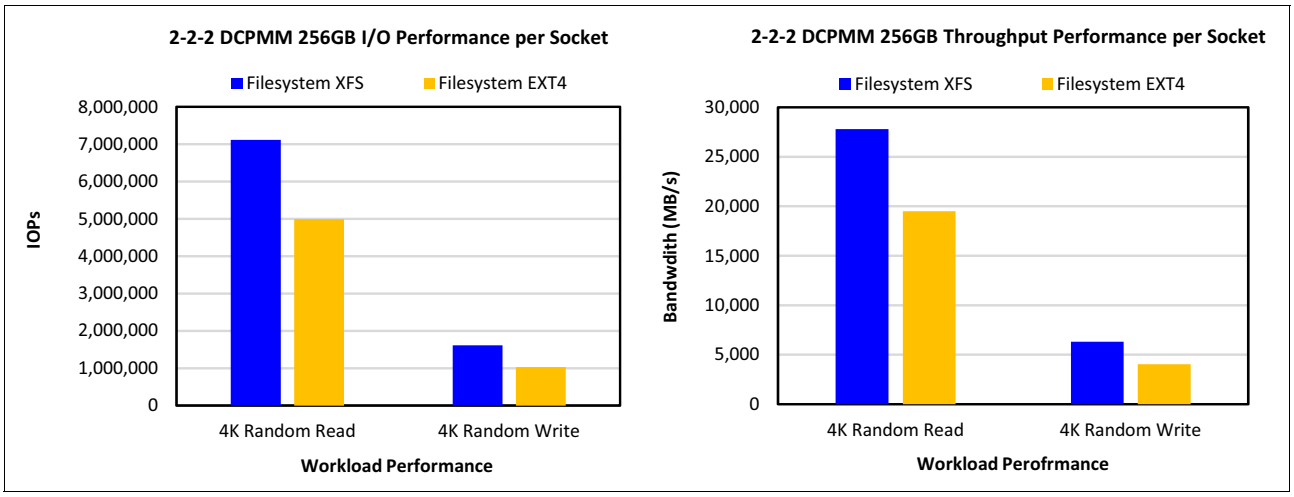

Figure 9 Single-socket DCPMM I/O (left) and Bandwidth (right) Performance on XFS vs EXT4 file system

Conclusion

Intel Optane DC Persistent Memory provides very high I/O throughput with ultra-low latency. With only few devices deployed, DCPMM devices can provide up to several million IOPS for each processor socket in a system at optimal response times. Intel DCPMMs can also provide bandwidth throughput that can reach far ahead of tradition storage options with RAID adapters, which are limited by the x8 PCIe bandwidth (8 GB/s) per adapter.

Comparing NVMe and DCPMM, it is clear that DCPMMs are most suitable for the applications with read-intensive access patterns. For workloads that are write intensive, the end-user should carefully consider if the DCPMM performance envelope is better than other high performance SSDs

For the latency of a storage device, DCPMMs can consistently provide service with the response time no higher than 4µs on read and 16µs on write. Even with the extremely heavy loading scenario, Intel DCPMM can still ensure the latency meet the requirements for most of the latency critical applications.

It is important to note, however, that if the configuration is not carefully set, the performance may not reach potential throughput. For optimal setup, interleaving mode must be selected and xfs file system must be mounted with DAX mode.

Authors

Travis Liao is a Hardware Performance Engineer in the Lenovo Data Center Group Performance Laboratory based in Taipei. His focus is modelling and validating performance of server storage subsystem including RAID controllers, SSDs and software RAID. Travis holds a Master's Degree in Electronic Engineering from National Taiwan University in Taiwan.

Tristian "Truth" Brown is a Hardware Performance Engineer on the Lenovo Server Performance Team in Raleigh, NC. He is responsible for the hardware analysis of high-performance, flash-based storage solutions for Data Center Group. Truth earned a Bachelor's Degree in Electrical Engineer from Tennessee State University and a Master's Degree in Electrical Engineering from North Carolina State University. His focus areas are in Computer Architecture and System-on-Chip (SoC) microprocessor design and validation.

Jamie Chou is an Advisory Engineer in the Lenovo Data Center Group Performance Laboratory in Taipei Taiwan. Jamie joined Lenovo in November 2014. Prior to working on server performance, he worked on system software development, automation, and Android system performance. Jamie received a Master’s Degree and a PhD from the department of Computer Science and Information Engineering, Tamkang University, Taiwan.

Thanks to the following people for their contributions to this project:

- David Watts, Lenovo Press
- Lenovo RDC Performance Team

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on April 19, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1085

Trademarks