Analyzing the Performance of Intel Optane DC Persistent Memory in Memory Mode in Lenovo ThinkSystem Servers

Introduces DCPMM Memory Mode

Discusses the performance of configurations with DRAM & DCPMM

Establishes performance expectations for DCPMM capacities

Evaluates performance between all DRAM vs. DRAM and DCPMMs

Jamie Chou

Tristian “Truth” Brown

Travis Liao

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 innovative 3D XPoint solid-state technology to deliver a new level of versatile performance in a compact memory module form factor.

This paper focuses on the low-level hardware performance capabilities of Intel Optane DC Persistent Memory configured in Memory Mode operation. There are unique implementations associated with this pioneering technology. The objective of this paper is to clarify performance outcomes when this technology is used for server system memory in conjunction with system memory (DRAM) DIMMs.

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 3
Intel Optane DC Persistent Memory 3
DCPMM and RDIMM configuration rules 4
Memory Mode operation 7
Memory Mode performance analysis 7
Conclusion 15
About the authors 15
Notices 16
Trademarks 17

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 a unique level of performance and versatility because it is backed by Intel 3D XPoint solid-state 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 Memory Mode operation.

Intel Optane DC Persistent Memory

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

DCPMMs can operate in three different configurations, Memory Mode, App Direct Mode, and Storage over App Direct Mode. This paper focuses on DCPMM Memory Mode and will analyze the performance of a system with DCPMMs running in this mode.

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 and RDIMM 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 within a system
- The recommended range of capacity ratio for DRAM to DCPMM in Memory Mode is between 1:4 and 1:16
- DCPMM devices should be installed in the memory slot closest to the CPU unless it is the only DIMM in the memory channel

For a more detailed description of Lenovo supported DCPMM system configurations please refer to the Lenovo Press paper, *Enabling Intel Optane DC Persistent Memory on Lenovo ThinkSystem Servers*, available from:

https://lenovopress.com/lp1167-enabling-intel-optane-dc-persistent-memory

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 physically (and electrically) closer to the processor for each memory channel.

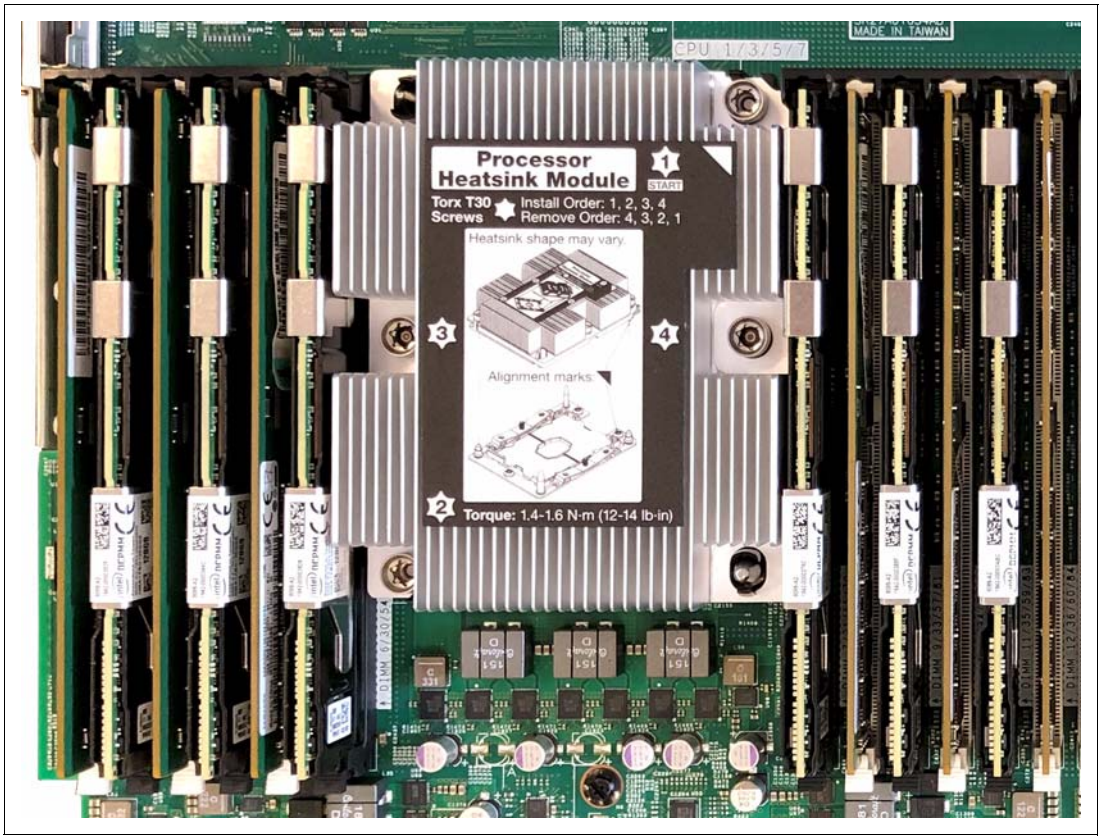

Figure 2 Intel Xeon Scalable Processor with 6 DCPMMs and 6 RDIMMs (SR950) installed

The configuration shown in Figure 2 represents a processor fully populated with DCPMMs and system memory (DRAM) and is denoted as a 2-2-2 setup. In contrast, the minimum supported configuration occurs when 2 devices are installed, 1 DRAM module and 1 DCPMM device; this is denoted as 1-1-0.

For Memory Mode, the ratio of DRAM capacity to DCPMM capacity plays a crucial role in expected performance outcomes. Table 2 list the ratios of DRAM capacity to DCPMM capacity for the configurations 2-2-2, 2-2-1, 2-1-1, and 1-1-1.

The capacity ratios shown in the highlighted cells are Lenovo-supported DCPMM memory mode configurations where the ratio is between 1:4 and 1:16.

Table 2 DRAM to DCPMM capacity ratios for each CPU socket. Ratios in bold are supported.

Population	DCPMM count	DRAM count	DRAM capacity	DCPMM capacity		
				128 GB	256 GB	512 GB
2-2-2	6	6	8 GB	1:16	1:32	1:64
			16 GB	1:8	1:16	1:32
			32 GB	1:4	1:8	1:16
			64 GB	1:2	1:4	1:8

Population	DCPMM count	DRAM count	DRAM capacity	DCPMM capacity		
				128 GB	256 GB	512 GB
2-2-1	4	6	8 GB	1:10.6	1:21.2	1:42.4
			16 GB	1:5.3	1:10.6	1:21.2
			32 GB	1:2.6	1:5.3	1:10.6
			64 GB	1:1.3	1:2.6	1:5.3
2-1-1	2	6	8 GB	1:5.3	1:10.6	1:21.2
			16 GB	1:2.6	1:5.3	1:10.6
			32 GB	1:1.3	1:2.6	1:5.3
			64 GB	1:0.6	1:1.3	1:2.6
1-1-1	2	4	8 GB	1:8	1:16	1:32
			16 GB	1:4	1:8	1:16
			32 GB	1:2	1:4	1:8
			64 GB	1:1	1:2	1:4

Figure 3 displays the physical locations of the configurations listed in the table. In the figure, R represents a DRAM module (i.e RDIMMs) and D represents a DCPMM.

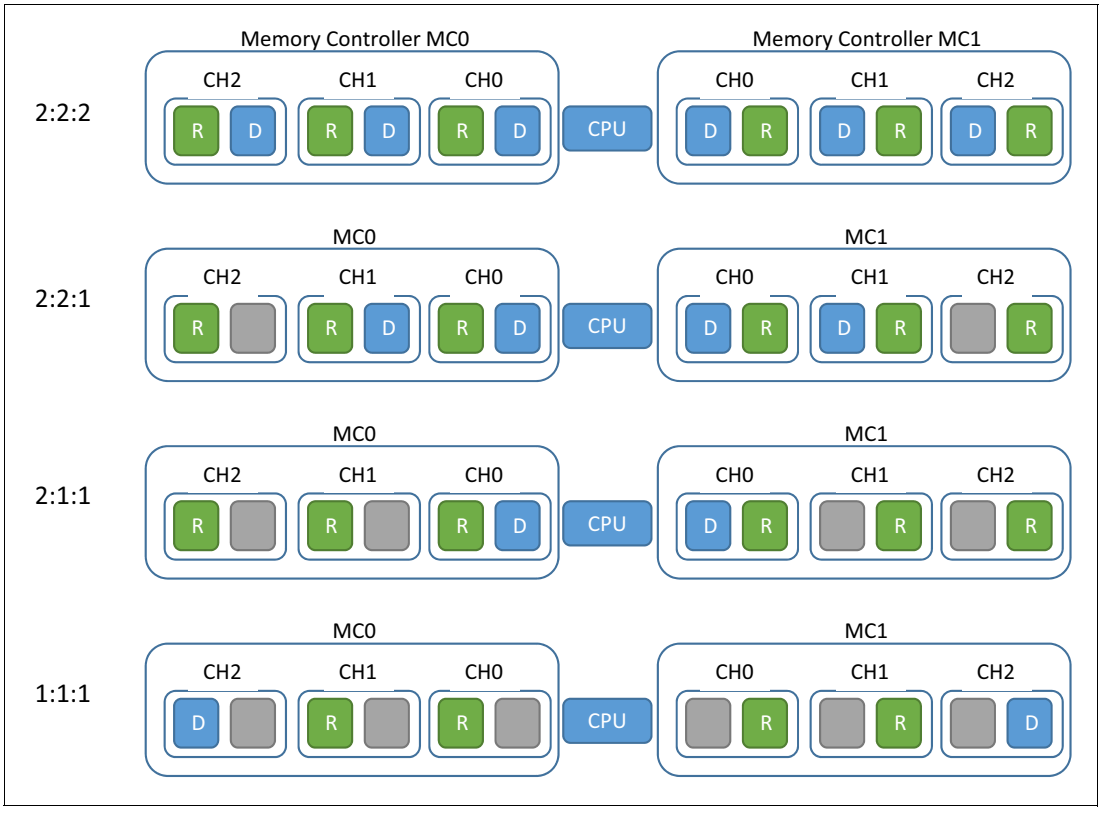

Figure 3 Scalable Processor DIMM slot locations for each memory controller (MC) and memory channel (CH). D denotes DCPMMs and R denotes DRAMs (i.e RDIMMs).

Memory Mode operation

DCPMM devices configured in Memory Mode operation are viewed as the operating system's DRAM memory. In this configuration, DRAM DIMMs are logically viewed as L4 cache and not as system memory. Therefore the total system memory as seen by the operating system is based on the total capacity of installed the DCPMM devices, and not the sum of the DIMMs and DCPMMs together.

For example, if a system has 96GB of RDIMM capacity and 768GB of DCPMM capacity, the operating system will utilize the 768GB capacity as the main memory footprint. The 96GB RDIMM capacity will be treated as L4 cache. The RDIMM capacity is transparent to the operating system.

Memory Mode does not require the applications to be configured to use DCPMMs in Memory Mode. Once the mode is set in UEFI, the server automatically presents the DCPMMs memory to the operating system as system memory.

Memory Mode performance analysis

This section describes the results of our analysis of performance of Memory Mode.

- "Hardware configuration evaluation environment"
- "Memory bandwidth performance" on page 8
- "Memory latency performance" on page 11
- "Comparing the performance of the three DCPMM capacities" on page 13

Hardware configuration evaluation environment

Intel Memory Latency Checker (MLC) was used to quantify this innovative technology because it is a well-established industry performance evaluation tool. MLC is fully compatible with DCPMM Memory Mode operation and has the ability to stress and measure performance and latency at the memory bus level.

In our analysis, MLC performance data is used to quantify throughput and latency performance expectations for a DRAM+DCPMM configuration as well as all-DRAM configs.

Specifically this evaluation data is based on a single-socket system configured with either RDIMMs+DCPMM devices or all-DRAM memory modules. Table 3 provides configuration details for each performance evaluation. DCPMM device support rules require mirrored hardware configurations across sockets therefore the following charts are representative of individual socket level performance throughout a multi-socket sever.

Table 3 System configurations for single-socket DCPMM Memory Mode performance evaluations

Config	All DRAM	2-2-2	2-2-1	2-1-1	1-1-1
Processor	Intel Xeon Platinum 8276L	Intel Xeon Platinum 8276L	Intel Xeon Platinum 8276L	Intel Xeon Platinum 8276L	Intel Xeon Platinum 8276L
OS	RHEL 7.6	RHEL 7.6	RHEL 7.6	RHEL 7.6	RHEL 7.6
DRAM	6x 32GB RDIMM 2666 MHz	6x 16GB RDIMM, 2666 MHz	6x 16GB RDIMM, 2666 MHz	6x 16GB RDIMM, 2666 MHz	4x 16GB RDIMM 2666 MHz
DCPMM	None	6x 128GB DCPMM 2666 MHz	4x 128GB DCPMM 2666 MHz	2x 128GB DCPMM 2666 MHz	2x 128GB DCPMM 2666 MHz

Memory bandwidth performance

In the following charts DCPMM & RDIMM configurations and all RDIMM populations are quantified in reference to an application's system memory footprint. Occupied Memory Size represents the amount of system memory footprint exercised by the given MLC traffic pattern. This methodology quantifies the expected performance of an application based on its specific system memory footprint.

Available system memory footprint is directly related to device count and capacity therefore the exercised occupied memory size was modified based on total system memory capacity.

For single socket all reads bandwidth performance analysis, the evaluated application memory footprints are as follows:

- 190GB for all-RDIMMs
- 700GB for 2-2-2
- 400GB for 2-2-1
- 200GB for 2-1-1 and 1-1-1

Figure 4 displays the sequential all-read socket bandwidth performance differences when using DCPMM+RDIMMs compared to all-RDIMMs. The system with an all-RDIMMs setup has the highest socket memory bandwidth and therefore the best overall performance. Unfortunately, the application is limited to a system memory capacity of roughly 190 GB per socket. In contrast, a 2-2-2 DCPMM+RDIMM installation using 128 GB DCPMM devices provides an application with the ability to expand up to more than 700 GB per socket.

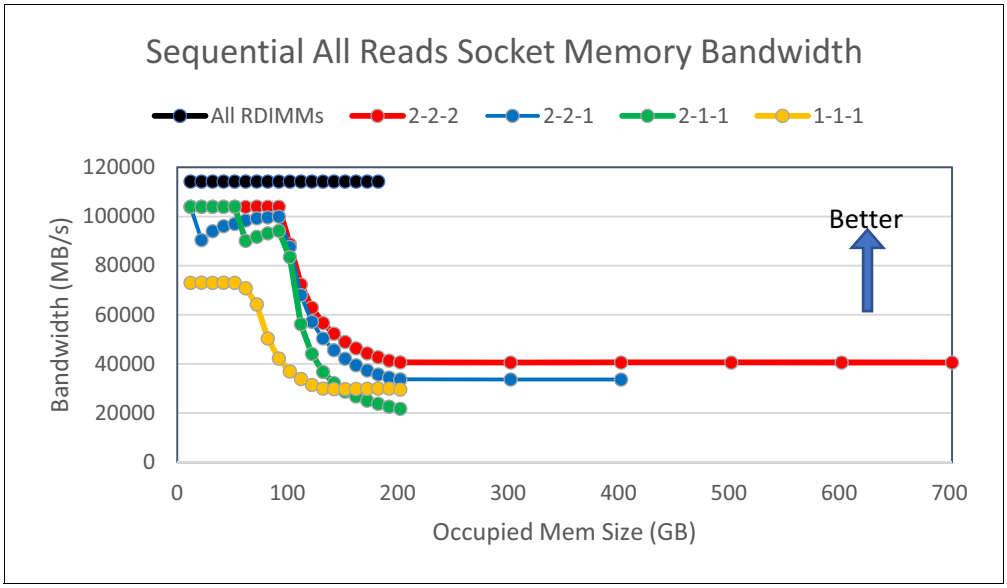

Figure 4 All-DRAM vs. DCPMM bandwidth comparison on sequential all read traffic

When the memory size of the test is within the total capacity of the installed RDIMMs, the DCPMM+RDIMM configurations perform very similarly. However, as the memory size of the test begins to exceed a system memory footprint of roughly 90 GB, the bandwidth performance quickly begins to degrade.

The resulting DCPMM memory bandwidth performance is directly associated with the number of devices within the configuration. The 2-2-2 configuration (red line in Figure 4) performs the best while the 2-1-1 configuration (green) performs the lowest.

The 2-1-1 configuration (green) under-performs the 1-1-1 configuration (yellow), due to the 2x DCPMMs in the system needing to share memory channels with the 4x RDIMMs. Less data movement congestion should be expected in 1-1-1 configuration comparing with 2-2-1 configuration.

Figure 5 displays single-socket random all-read performance, and the results align with the single-socket sequential read performance shown in Figure 4. There is a positive delta in the 2-2-1 configuration (blue), and a small negative delta in the 2-1-1 (green) and 1-1-1 (yellow) configuration. But in short, single-socket all-reads DCPMM bandwidth performance is not significantly affected by sequential or random all reads traffic.

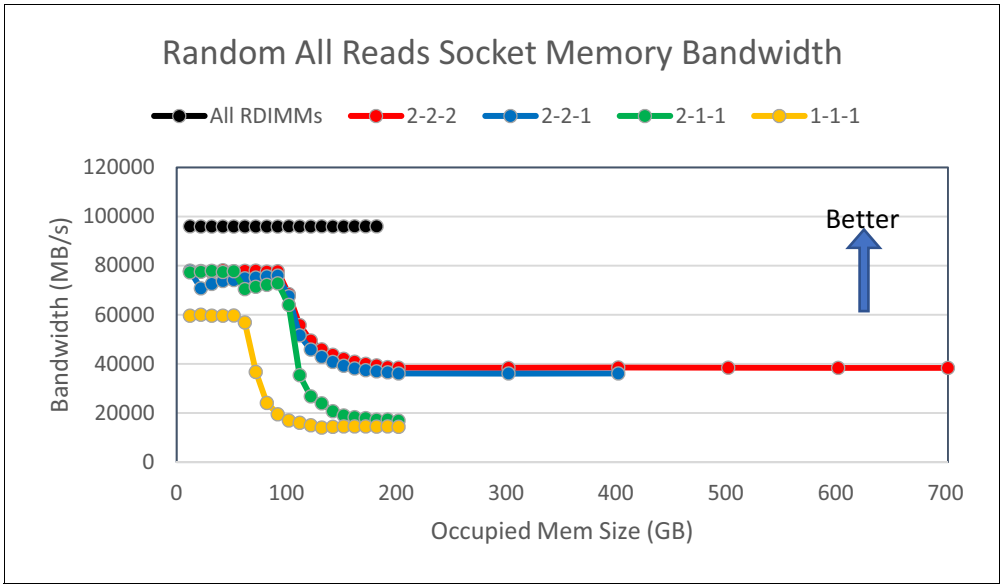

Figure 5 Socket memory bandwidth comparison on random all read traffic

Figure 6 displays single socket memory bandwidth on different DCPMM configurations when utilizing sequential mixed (2:1 Read/Write) traffic. The system with an all-RDIMMs configuration still performs the best.

Due to the nature of the 2R:1W workload, only approximately 100GB system memory footprint can be allocated. However, by utilizing a DCPMM+RDIMM configuration all of the other configurations are able to allocate larger application memory footprint capacities.

The application memory footprints evaluated were as follows:

- 100GB for All RDIMMs (black line in Figure 6)
- 300GB for 2-2-2 (red line)
- 200GB for 2-2-1 (blue line)
- 120GB for 2-1-1 (green) and 1-1-1 (yellow)

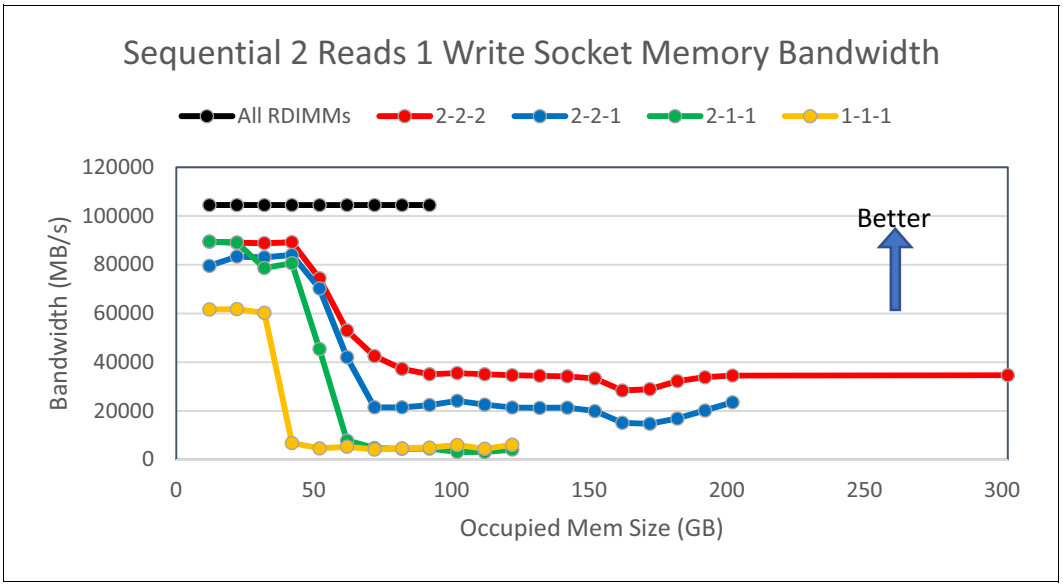

Figure 6 Socket memory bandwidth comparison on sequential 2R1W read traffic

Again the bandwidth results show the DCPMM 2-2-2 configuration performing being the best and the 2-1-1/1-1-1 configurations performing the least. The major difference in this evaluation is that the DCPMM 2-1-1 and 1-1-1 configurations eventually see a comparable performance level once a 60GB system memory footprint is exceeded.

Figure 7 displays single socket memory bandwidth on different DCPMM configurations when driven by a random 2:1 Read/Write workload. The all RDIMMs system retains roughly around 90% single socket bandwidth performance compared to the sequential all reads results.

The single socket DCPMM configurations see roughly a 60% single socket bandwidth performance drop when compared to the sequential 2:1 Read/Write workload. With the DCPMM configuration the large performance gap is directly correlated with the traffic pattern containing random writes.

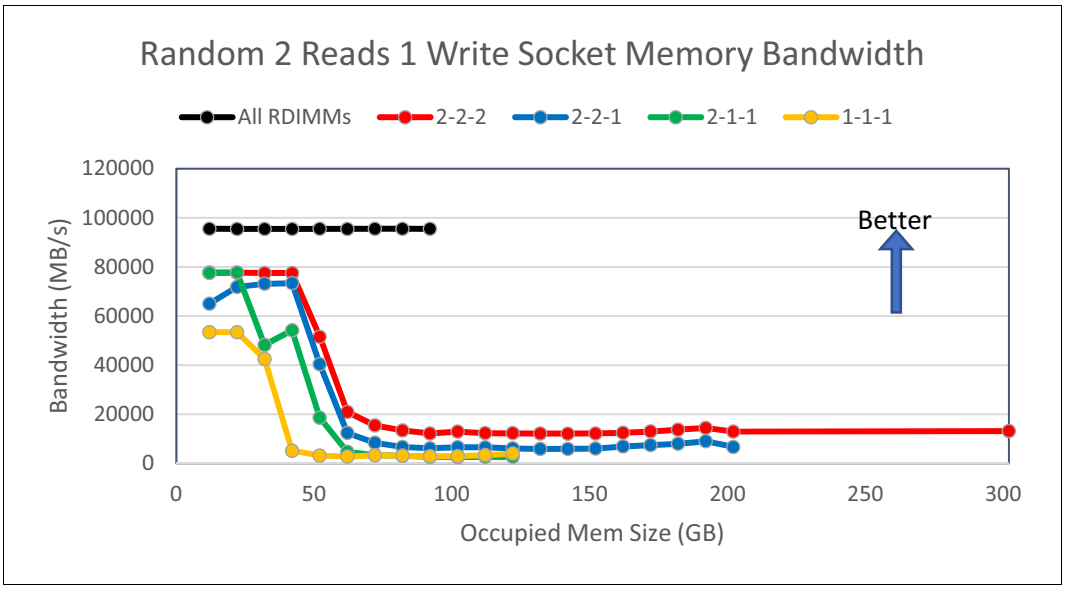

Figure 7 Socket memory bandwidth comparison on random 2R1W traffic

Memory latency performance

For single-socket all-reads latency performance analysis, the evaluated system memory footprints were as follows:

- 190GB for All RDIMMs
- 700GB for 2-2-2
- 400GB for 2-2-1
- 200GB for 2-1-1 and 1-1-1

The spike seen on all DCPMM configured systems is an artifact of the evaluation tool's caching algorithm.

Figure 8 displays the latency performance associated with the single socket sequential bandwidth results shown in Figure 4 on page 8. For most DCPMM configurations best case latency performance is up to roughly 90GB of system memory footprint. In this window, DCPMM latency performance is very close to the all RDIMMs configuration because the RDIMMs are acting as an L4 cache. However, when the application memory footprint exceeds the RDIMM system capacity, the latency performance decreases dramatically.

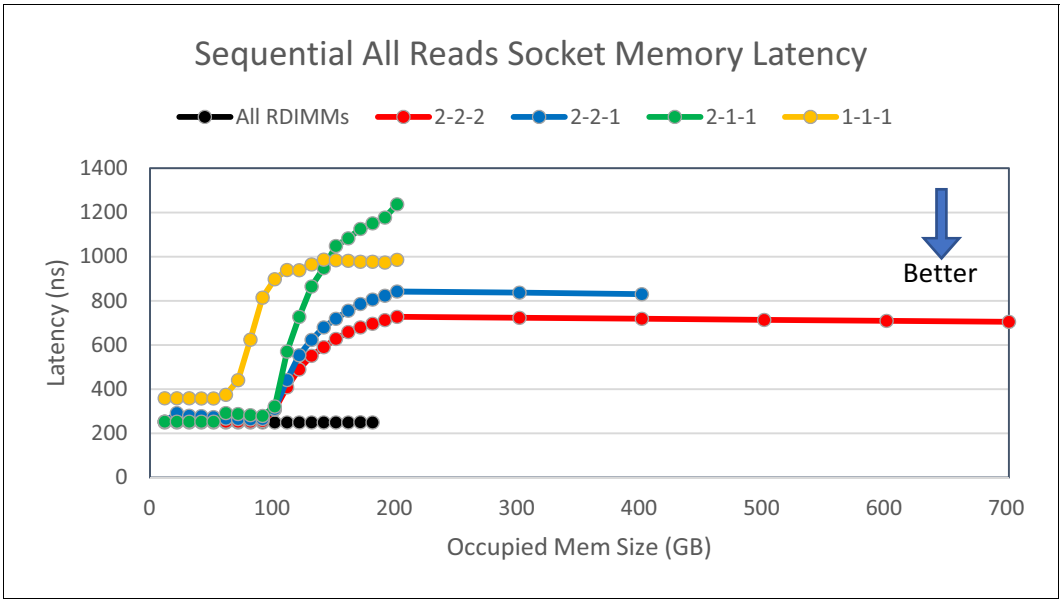

Figure 8 Socket memory latency comparison on sequential all read traffic

Systems with more DCPMMs across memory channels provide a greater level of interleaving and thus produce better read/write parallelism. The overall result is the number of DCPMMs installed in a system dictate the maximum latency performance. As seen in bandwidth results, the DCPMM 2-2-2 configuration displays the best case latency performance and the 1-1-1 configuration shows the lowest performance.

Figure 9 on page 12 displays the latency performance associated with the single-socket random bandwidth results shown in Figure 5 on page 9.

The single-socket random read traffic is the most favorable workload for every DCPMM configuration evaluated. The DCPMM 2-2-2 and 2-2-1 configurations performance aligns closely with the all RDIMM system. The DCPMM 2-1-1 and 1-1-1 configurations have the lowest performance and displays a predictable performance degradation after the 90GB system memory footprint. The lower DCPMM device count across memory channels reduces interleaving and thus negatively effects latency performance.

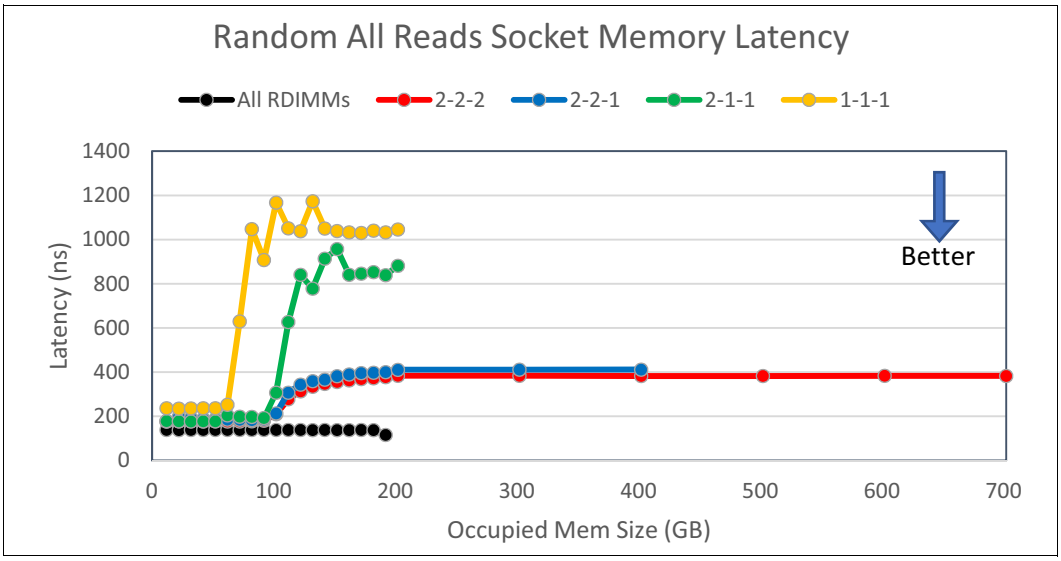

Figure 9 Socket memory latency comparison on random all read traffic

For single-socket 2 Reads / 1 Writes latency performance analysis, the evaluated application memory footprints are as follows:

- 100GB for the All RDIMMs configuration
- 300GB for 2-2-2
- 200GB for 2-2-1
- 120GB for 2-1-1 and 1-1-1 configurations

Figure 10 displays the latency performance associated with the single-socket sequential 2:1 Read/Write traffic bandwidth results shown in Figure 6 on page 10. This workload highlights the performance benefit of having multiple DCPMM devices spread across the highest number of memory channels.

The high DCPMM count 2-2-2 and 2-1-1 configurations display a relatively tight range of latency performance for their application memory footprints. The low DCPMM count 2-1-1 and 1-1-1 configurations produce poor latency performance outside of the RDIMM L4 caching window of 42GB and 32GB respectively.

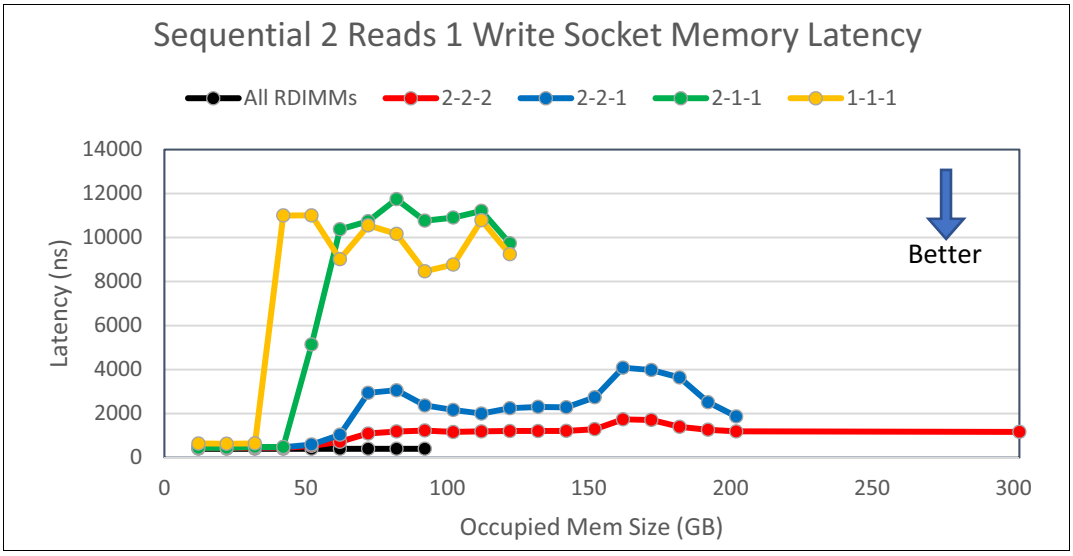

Figure 10 Socket memory latency comparison on sequential 2R1W traffic

Figure 11 displays the latency performance associated with the single socket random 2:1 Read/Write traffic bandwidth results shown in Figure 7 on page 10. This traffic pattern is the worst-case workload for every DCPMM configuration evaluated.

The RDIMM L4 caching benefit is up to roughly 42GB, after which the DCPMM configuration latency performance rapidly degrades. The DCPMM 2-2-2 configuration performs the best at roughly 750ns for the larger application capacity sizes but greater than 1000ns of latency is expected for the other configurations.

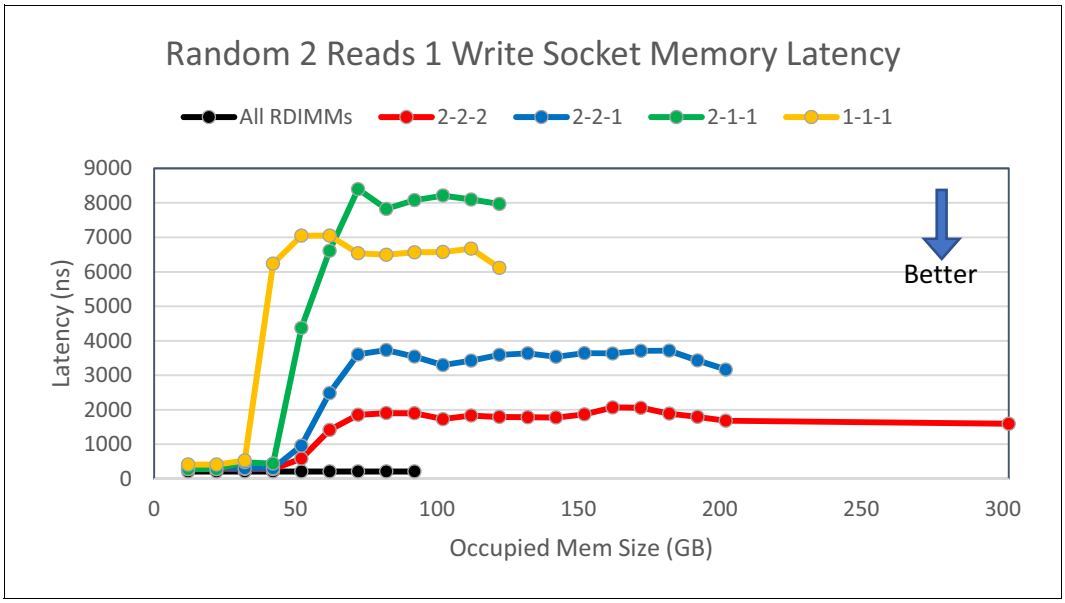

Figure 11 Socket memory latency comparison on random 2R1W traffic

Comparing the performance of the three DCPMM capacities

This section discusses the observed performance among the various DCPMM capacities. For consistency a DCPMM 2-2-2 configuration was implemented based on the 128GB, 256GB, and 512GB capacities. For these tests, the occupied memory size is bound to 202 GB for each setup.

Figure 12 on page 14 displays the single socket bandwidth performance differences between 128GB, 256GB, and 512GB DCPMM 2-2-2 configurations.

For sequential all-reads traffic (left side of the figure) the various DCPMM capacities have comparable performance with each producing roughly 40 GB/s per socket bandwidth. In contrast, the random 2 reads / 1 write traffic pattern exposes a performance difference between the capacities (right side of the figure).

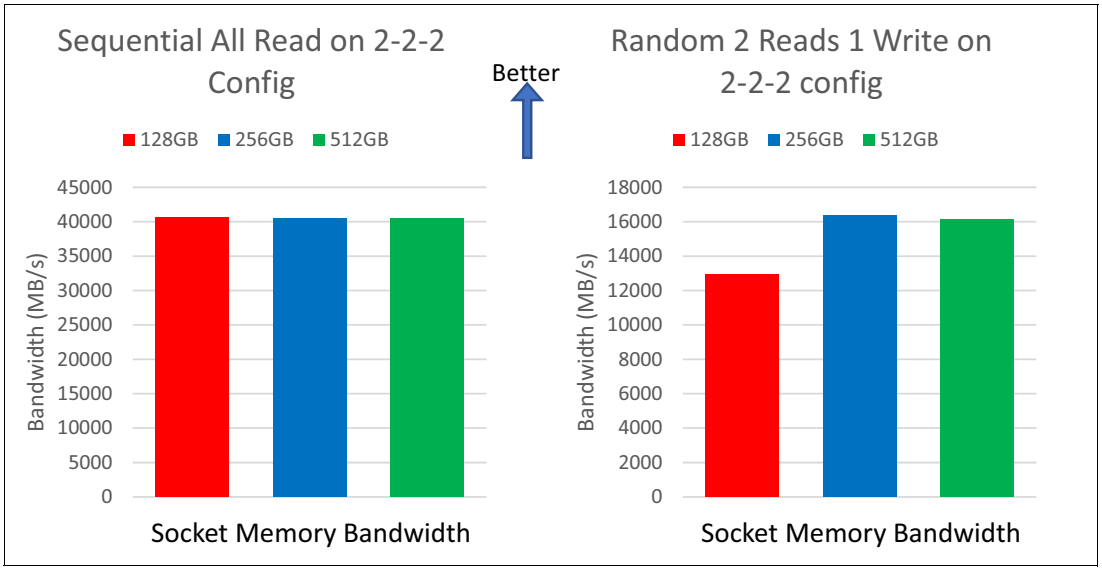

Figure 12 Socket memory bandwidth with different traffic types on different DCPMM modules

This difference is caused by the location of where the bottlenecks exist for each of these evaluations. With the sequential all-read workload, the bottleneck located at the memory bus bandwidth, rather than within the DCPMMs. Memory controller performance is shared between DRAM and DCPMM devices and the sequential all-reads workload saturates the memory bus.

In the 2:1 Read/Write workload, the bottleneck is within the DCPMMs because of their inherent asymmetric read/write properties. Therefore, adding a random write component to the workload traffic introduces the worst case transaction for the device and highlights the performance differences between DCPMM capacities.

Figure 13 displays the single socket latency performance differences between 128GB, 256GB, and 512GB DCPMM 2-2-2 configurations. The results correlate with the bandwidth performance observed in Figure 12. The identical conclusions that govern the bandwidth performance are carried over to the latency performance outcomes.

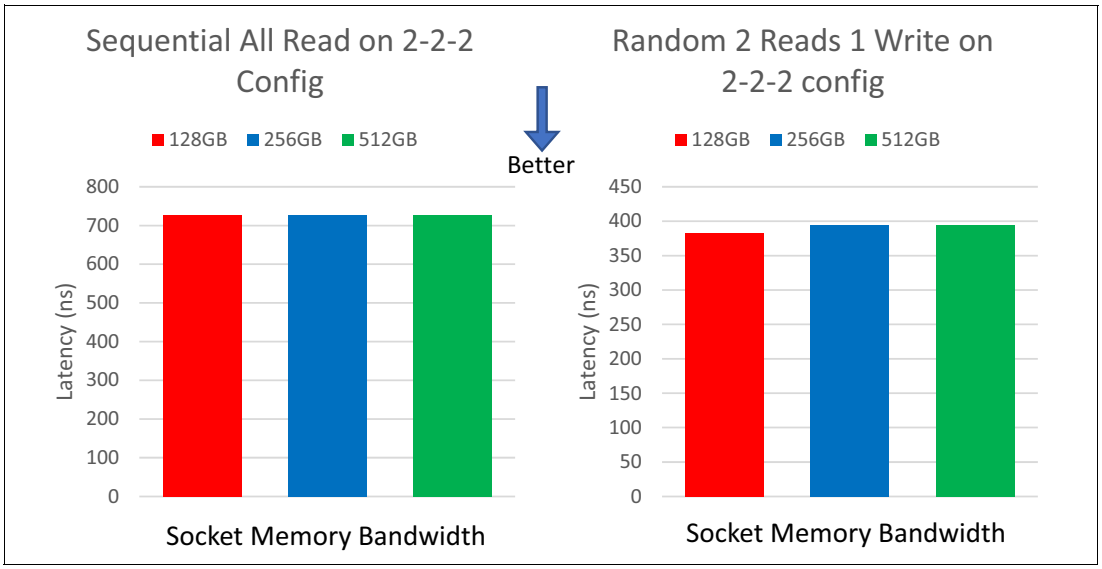

Figure 13 Socket memory latency with different traffic types on different DCPMM modules

Conclusion

Intel Optane DC Persistent Memory fills the gap between RDIMM technology and the highest performing block storage devices. Its larger size can be a very good complement to existing RDIMMs. It also provided the same order of magnitude in terms of system bandwidth and latency. By utilizing DCPMMs, systems can be used in a wider range with larger memory capacity.

However, not every workload is suitable to run on DCPMMs. The characteristics of workloads need to be verified since DCPMM provides lower system bandwidth and higher system latency compared to DRAM. Performance evaluation is a must in order to ensure the best performance when utilizing DCPMMs.

About the authors

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 degree from the department of Computer Science and Information Engineering, Tamkang University, 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 were in Computer Architecture and System-on-Chip (SoC) microprocessor design and validation.

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.

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 November 13, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1084

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo(logo)® Lenovo®

The following terms are trademarks of other companies:

3D XPoint, Intel, Intel Optane, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.