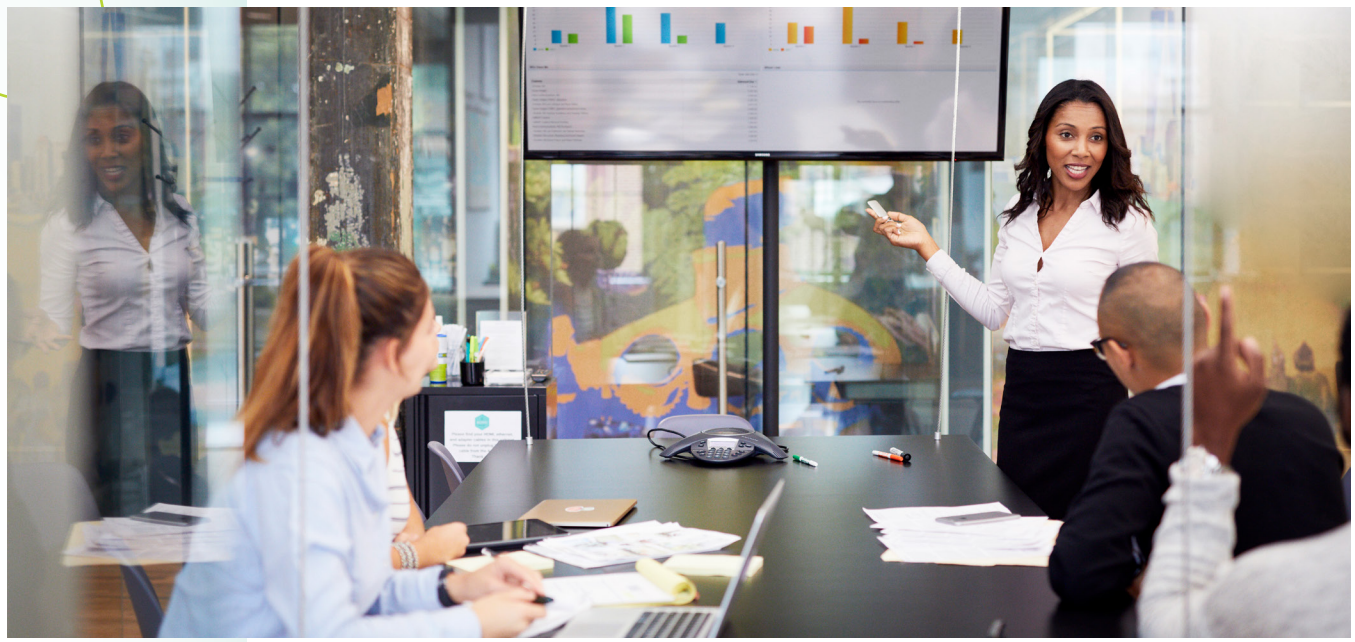


Performance and Scalability Benchmark: InterSystems IRIS and Intel Optane DC Persistent Memory

Executive Summary

InterSystems and Intel recently conducted a series of benchmarks combining InterSystems IRIS® data platform with 2nd Generation Intel® Xeon® Scalable Processors, also known as Cascade Lake, and Intel® Optane™ DC persistent memory (DCPMM). The goal of these benchmarks was to demonstrate the performance and scalability capabilities of InterSystems IRIS with Intel's latest memory technologies in various workload settings and server configurations. Along with various benchmark results, three different use cases of Intel DCPMM with InterSystems IRIS are provided in this report.

Two types of workload were used to demonstrate performance and scaling — a read-intensive workload and a write-intensive workload. The reason for demonstrating these separately was to show the impact of Intel DCPMM on different use cases specific to increasing database cache efficiency in a read-intensive workload, and increasing write throughput for transaction journals in a write-intensive workload. In both of these use-case scenarios, significant throughput, scalability, and performance gains for InterSystems IRIS were achieved.

- **The read-intensive workload** leveraged a four-socket server and massive long running analytical queries across a dataset of approximately 1.2TB of total data. With DCPMM in Memory Mode, benchmark comparisons yielded a significant reduction in elapsed runtime — it was approximately six times faster when compared with a previous-generation Intel Xeon E7 v4 series processor with less memory. When comparing like-for-like memory sizes between the E7 v4 and the latest server with DCPMM, there was a 20% improvement. This was due to both the increased InterSystems IRIS database cache capability afforded by DCPMM and the latest Intel processor architecture.

InterSystems®
IRIS Data Platform

- **The write-intensive workload** leveraged a two-socket server and InterSystems HL7 v2 messaging benchmark, which consists of numerous inbound interfaces. Each message has several transformations and then four outbound messages for each inbound message. One of the stringent requirements in sustaining high throughput is the message durability guarantees of InterSystems IRIS for Health™, and the transaction journal write performance is crucial in that operation. With DCPMM in App Direct Mode as direct access (DAX) presenting an XFS file system for transaction journals, this benchmark demonstrated a 60% increase in message throughput.

To summarize the test results and configurations: **DCPMM offers significant throughput gains when used in the proper InterSystems IRIS setting and workload.** The high-level benefits are increased database cache efficiency and reduced disk I/O block reads in read-intensive workloads and also increased write throughput for journals in write-intensive workloads.

In addition, Cascade Lake-based servers with DCPMM provide an excellent update path for those looking to refresh older hardware and improve performance and scaling. InterSystems® technology architects are available to help with those discussions and provide advice on suggested configurations for your existing workloads.

Read-Intensive Workload Benchmark

For the read-intensive workload, we used an analytical query benchmark and compared a Xeon E7 v4 (Broadwell) with 512GiB and 2TiB database cache sizes against the latest 2nd Generation Intel Xeon Scalable Processors (Cascade Lake) with 1TiB and 2TiB database cache sizes using Intel Optane DCPMM.

We ran several workloads with varying global buffer sizes to show the impact and performance gain of larger caching. For each configuration iteration, we performed a cold run and a warm run. In a cold run, the database cache has not been pre-populated with any data. In a warm run, the database cache has already been active and populated with data (at least as much as it could be) to reduce physical reads from disk.

Hardware Configuration

We compared an older four-socket Broadwell host to a four-socket Cascade Lake server with DCPMM. This comparison was chosen because it would demonstrate performance gains for existing customers looking for a hardware refresh along with using InterSystems IRIS. In all tests, the same version of InterSystems IRIS was used so that any software optimizations between versions were not a factor.

All servers had the same storage on the same storage array so that disk performance wouldn't be a factor in the comparison. The working set is a 1.2TB database. The hardware configurations are shown in Figure 1 with the comparison between each of the four-socket configurations:

Server 1 Configuration:	Server N2 Configuration:
Processors: 4 x E7-8890 v4 @ 2.5GHz	Processors: 4 x Platinum 8280L @ 2.6GHz
Memory: 2TB DRAM	Memory: 3TiB DCPMM + 768GiB DRAM
Storage: FC-attached all-flash @ 2TiB LUN	Storage: FC-attached all-flash @ 2TiB LUN
	DCPMM: Memory Mode only

Figure 1: Hardware Configurations

Benchmark Results and Conclusions

There was a significant reduction in elapsed runtime: It was approximately six times faster when comparing 512GiB to either 1TiB or 2TiB DCPMM buffer pool sizes. In addition, in comparing 2TiB E7 v4 DRAM and 2TiB Cascade Lake DCPMM configurations, there was an approximately 20% improvement as well. This 20% gain is believed to be mostly attributable to the new processor architecture and additional processor cores, given that the buffer pool sizes were the same. However, this is still significant in that in the four-socket Cascade Lake tested had only 24 x 128GiB DCPMM installed but can scale to 12TiB DCPMM, which is about four times the memory of what E7 v4 can support in the same four-socket server footprint.

The graphs in Figure 2 depict the comparison results. In both graphs (Figure 2 and Figure 4), the y-axis indicates elapsed time (where a lower number is better), comparing the results from the various configurations.

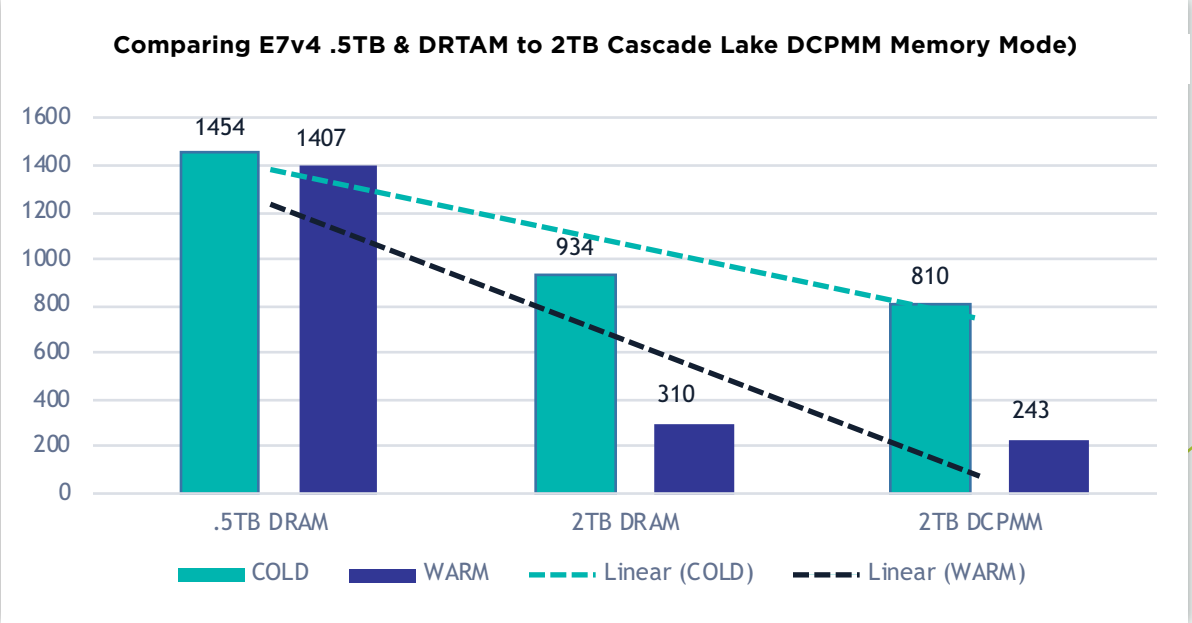


Figure 2: Elapsed Time Comparison of Various Configurations

Write-Intensive Workload Benchmark

For the write-intensive workload benchmark, we used the HL7 v2 messaging workload using all T4 type workloads.

The T4 workload used a routing engine to route separately modified messages to each of four outbound interfaces. On average, four segments of the inbound message were modified in each transformation. For each inbound message, four data transformations were executed, four messages were sent outbound, and five HL7 message objects were created in the database.

Each system is configured with 128 inbound business services and 4,800 messages sent to each inbound interface, for a total of 614,400 inbound messages and 2,457,600 outbound messages.

The measurement of throughput in this benchmark workload is messages per second. We are also interested in (and recorded) the journal writes during the benchmark runs because transaction journal throughput and latency are stringent requirements in sustaining high throughput. This directly influences the performance of message durability guarantees of InterSystems IRIS for Health, and the transaction journal write performance is crucial in that operation. When journal throughput suffers, application processes will block journal buffer availability.

Hardware Configuration

For the write-intensive workload, we decided to use a two-socket server. This is a smaller configuration than our previous four-socket configuration in that it had only 192GiB of DRAM and 1.5TiB of DCPMM. We compared the workload of Cascade Lake with DCPMM to that of the previous 1st Generation Intel Xeon Scalable Processor (Skylake) server. Both servers have locally attached 750GiB Intel Optane SSD DC P4800X drives.

The hardware configurations are shown in Figure 3 with the comparison between each of the two-socket configurations:

Server 1 Configuration:	Server 2 Configuration:
Processors: 2 x Gold 6152 @ 2.1GHz	Processors: 2 x Gold 6252 @ 2.1GHz
Memory: 192GiB DRAM	Memory: 1.5TiB DCPMM + 192GiB DRAM
Storage: FC-attached all-flash @ 2TiB LUN	Storage: FC-attached all-flash @ 2TiB LUN
	DCPMM: Memory & App Direct Modes

Figure 3: Write-Intensive Workload Hardware Configurations

Benchmark Results and Conclusions

TEST 1:

This test ran the workload described above on the Skylake server configured as Server 1 in Figure 3. The Skylake server provided a sustained throughput of approximately 3,355 inbound messages per second, with a journal file write rate of 2,010 journal writes per second.

TEST 2:

This test ran the same workload but on the Cascade Lake server configured as Server 2 in Figure 3, and specifically with DCPMM in **Memory Mode**. This demonstrated a significant improvement of sustained throughput of about 4,684 inbound messages per second, with a journal file write rate of 2,400 journal writes per second. **This provided a 39% increase compared with Test 1.**

TEST 3:

This test ran the same workload, on the Cascade Lake server configured as Server 2 in Figure 3. This time, DCPMM was used in App Direct Mode but was not actually configured to do anything. The intent was to gauge just what the performance and throughput would be when comparing Cascade Lake with DRAM only to Cascade Lake with DCPMM plus DRAM. The result was a gain in throughput without the use of DCPMM, albeit a relatively small one. This demonstrated an improvement of sustained throughput of approximately 4,845 inbound messages per second, with a journal file write rate of 2,540 journal writes per second. This behavior was expected, because DCPMM has a higher latency compared with DRAM and there is a penalty to performance with the massive influx of updates. Another way of looking at it that there is a less than 5% reduction in write ingestion workload when using DCPMM in Memory Mode on the same exact server. Additionally, using Cascade Lake (DRAM only) **resulted in a 44% increase compared with using the Skylake server in Test 1.**

TEST 4:

This test ran the same workload, on the Cascade Lake server configured as Server 2 in Figure 3, this time using DCPMM in App Direct Mode and using App Direct Mode as DAX XFS mounted for the journal file system. This yielded even more throughput, of 5,399 inbound messages per second, with a journal file write rate of 2,630 per second. This demonstrated that DCPMM in App Direct Mode for this type of workload is the better use of DCPMM. **This provided a 60% increase in throughput compared with the initial Skylake server configuration in Test 1.**

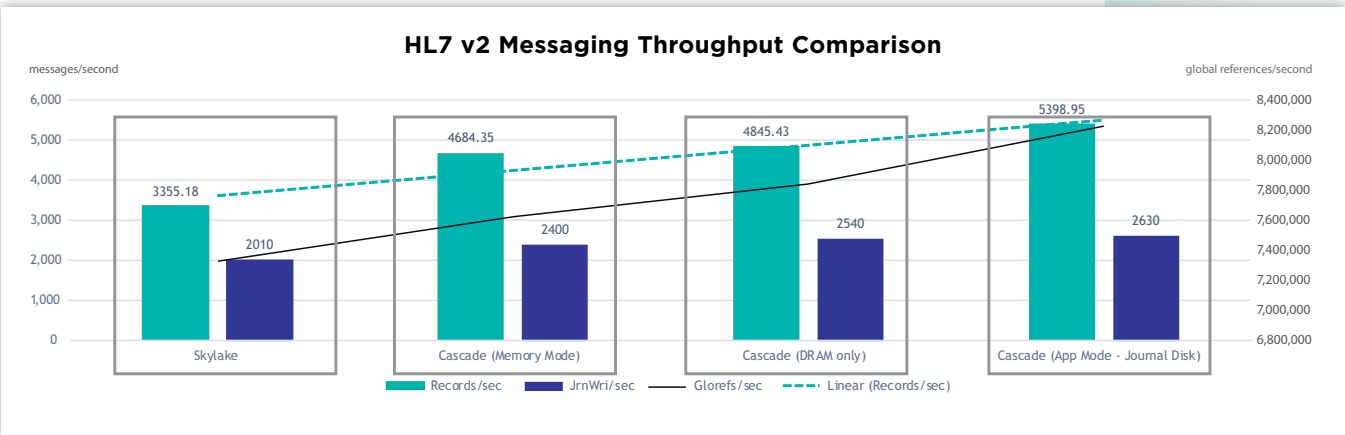


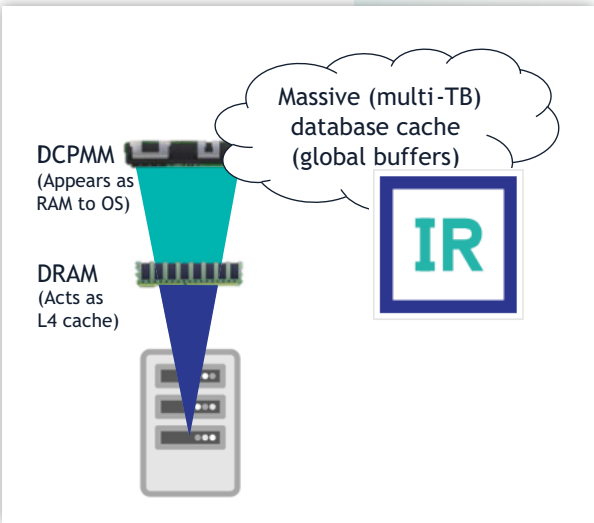
Figure 4: Message Throughput Comparison

InterSystems IRIS Recommended Intel DCPMM Use Cases

There are several use cases and configurations in which InterSystems IRIS will benefit from using Intel Optane DC persistent memory.

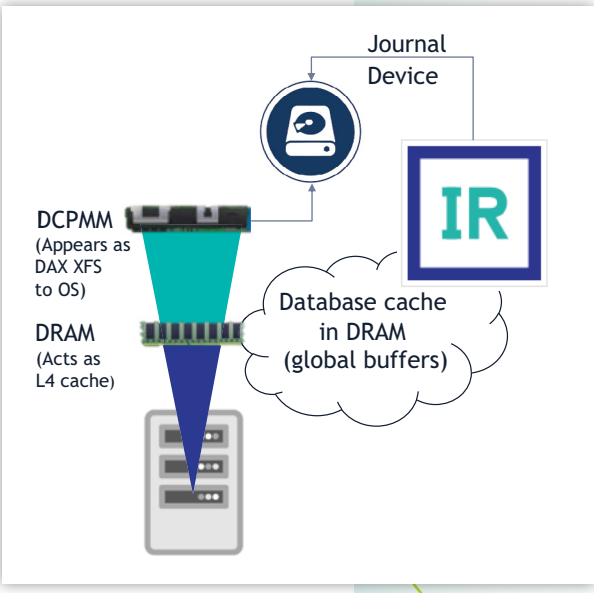
Memory Mode

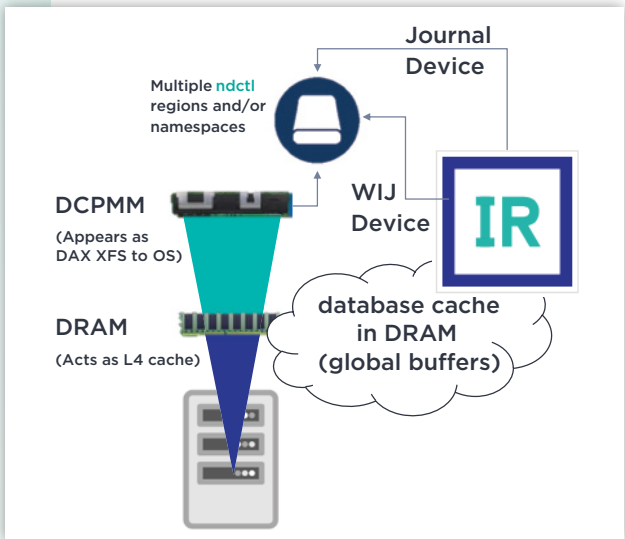
This is ideal for massive database caches for either a single InterSystems IRIS deployment or a large InterSystems IRIS sharded cluster where you want to have much more (or all!) of your database cached into memory. It is important to adhere to a maximum ratio of 8:1 for DCPMM to DRAM in order for the "hot memory" to stay in DRAM acting as an L4 cache layer. This is especially important for some shared internal InterSystems IRIS memory structures such as seize resources and other memory cache lines.



App Direct Mode (DAX XFS) – Journal Disk Device

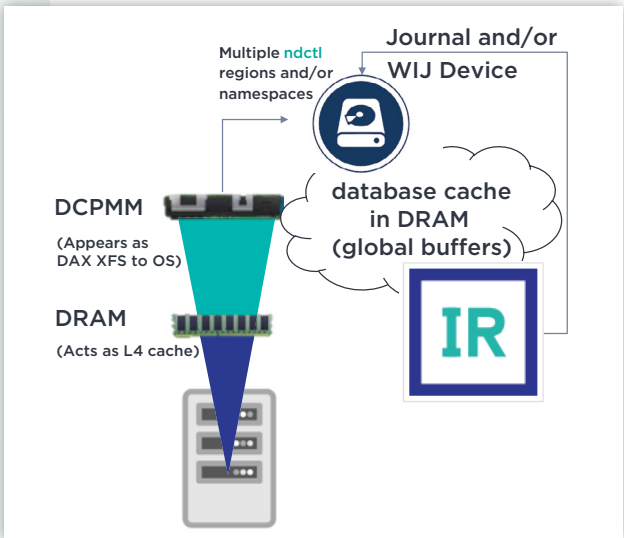
This is ideal for using DCPMM as a disk device for transaction journal files. DCPMM appears to the operating system as a mounted XFS file system to Linux. The benefit of using DAX XFS is that this alleviates the PCIe bus overhead and direct memory access from the file system. As demonstrated in the HL7 v2 benchmark results, the write latency benefits significantly increased the HL7 messaging throughput. Additionally, the storage is persistent and durable on reboots and power cycles, just like a traditional disk device.





App Direct Mode (DAX XFS) – Journal + Write Image Journal Disk Device

This use case extends the use of App Direct Mode to both the transaction journals and the write image journal (WIJ). Both of these files are write-intensive and will certainly benefit from ultralow latency and persistence.



Dual Mode: Memory + App Direct Modes

When using DCPMM in Dual Mode, the benefits of DCPMM are extended to allow for both massive database caches and ultralow latency for the transaction journals and/or WIJ devices. In this use case, DCPMM appears both as a mounted XFS file system and RAM to the operating systems. This is achieved by allocating a percentage of DCPMM as DAX XFS while the remainder is allocated in Memory Mode. As mentioned previously, the DRAM installed will operate as an L4-like cache to the processors.

Quasi Dual Mode

To extend the use case models on a bit of slant with a quasi Dual Mode in that you have concurrent transaction and analytic workloads (also known as hybrid transaction/analytics processing or HTAP workloads), where there is a high rate of inbound transactions/updates for OLTP-type workloads, and also an analytical or massive querying need, and then having each InterSystems IRIS node type within an InterSystems IRIS sharded cluster operating with different modes for DCPMM.

In this example, InterSystems IRIS compute nodes are added to handle the massive querying/analytics workload running with DCPMM Memory Mode so that the compute nodes and query workloads benefit from the massive database cache in the global buffers. The data nodes are either running the DAX XFS in Dual Mode or App Direct Mode for the transactional workloads.

Conclusion

There are numerous options available for InterSystems IRIS when it comes to infrastructure choices. The application, workload profile, and business needs drive the infrastructure requirements, and those technology and infrastructure choices influence the success, adoption, and importance of your applications to your business. InterSystems IRIS with 2nd Generation Intel Xeon Scalable Processors and Intel Optane DC persistent memory provides for groundbreaking levels of scaling and throughput capabilities for the applications based on InterSystems IRIS that matter to your business.

Benefits of using InterSystems IRIS with Intel DCPMM-capable servers include the following:

- It increases memory capacity so that multi-terabyte databases can completely reside in an InterSystems IRIS or InterSystems IRIS for Health database cache with DCPMM in Memory Mode. In comparison to reading from storage (disks), this can increase query response performance by up six times with no code changes. This is due to the proven memory caching capabilities of InterSystems IRIS, which take advantage of system memory as it increases in size.
- It improves the performance of high-rate data interoperability throughput applications based on InterSystems IRIS and InterSystems IRIS for Health, such as HL7 transformations, by as much as 60% in increased throughput. It does so using the same processors and changing only the transaction journal disk from the fastest available NVMe drives to leveraging DCPMM in App Direct Mode as a DAX XFS file system. Exploiting both the memory speed data transfers and data persistence is a significant benefit to InterSystems IRIS and InterSystems IRIS for Health.
- It augments the compute resources where needed for a given workload, whether read-intensive, write-intensive, or both, without overallocating entire servers just for the sake of one resource component, with DCPMM in Mixed Mode.

InterSystems technology architects are available to discuss hardware architectures ideal for your applications based on InterSystems IRIS. Contact us at IRIS@InterSystems.com

InterSystems®
Creative data technology