InterSystems IRIS and Intel Optane DC Persistent Memory

Article
Mark Bolinsky · Mar 3, 2020 11m read

InterSystems IRIS and Intel Optane DC Persistent Memory

InterSystems and Intel recently conducted a series of benchmarks combining InterSystems IRIS with 2nd Generation Intel® Xeon® Scalable Processors, also known as “Cascade Lake”, and Intel® Optane™ DC Persistent Memory (DCPMM). The goals of these benchmarks are to demonstrate the performance and scalability capabilities of InterSystems IRIS with Intel’s latest server 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.

Overview

Two separate types of workloads are used to demonstrate performance and scaling – a read-intensive workload and a write-intensive workload. The reason for demonstrating these separately is 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 are achieved.

- The **read-intensive workload** leveraged a 4-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 of approximately six times faster when compared to a previous generation Intel E7v4 series processor with less memory. When comparing like-for-like memory sizes between the E7v4 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.

- The **write-intensive workload** leverages a 2-socket server and InterSystems HL7 messaging benchmark which consists of numerous inbound interfaces, each message has several transformations and then four outbound messages for each inbound. One of the critical components in sustaining high throughput is the message durability guarantees of IRIS for Health, and the transaction journal write performance is crucial in that operation. With DCPMM in “APP DIRECT” mode as DAX XFS 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 increasing database cache efficiency and reducing disk IO block reads in read-intensive workloads and also increasing write throughput for journals in write-intensive workloads.

In addition, Cascade Lake based servers with DCPMM provide an excellent update path for those looking into refreshing older hardware and improving 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 comparing an E7v4 (Broadwell) with 512GiB and 2TiB database cache sizes, against the latest 2nd Generation Intel® Xeon® Scalable Processors (Cascade Lake) with 1TB and 2TB database cache sizes using Intel® Optane™ DC Persistent Memory (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 ran a COLD, and a WARM run. COLD is where the database cache was not pre-populated with any data. WARM is where the database cache has already been active and populated with data (or at least as much as it could) to reduce physical reads from disk.

Hardware Configuration

We compared an older 4-Socket E7v4 (aka Broadwell) host to a 4-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 have the same storage on the same storage array so that disk performance wasn't 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 4-socket configurations:

Figure-1: Hardware configurations

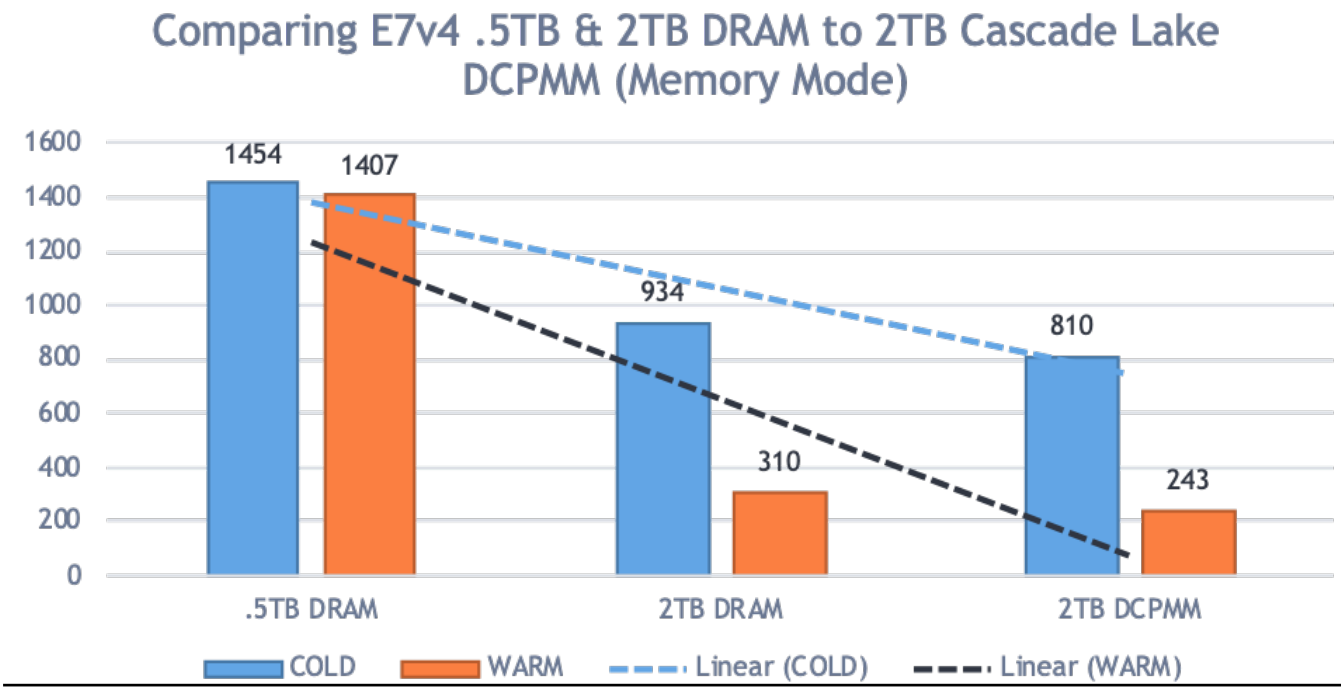
Server #1 Configuration	Server #2 Configuration
Processors: 4 x E7-8890 v4 @ 2.5Ghz	Processors: 4 x Platinum 8280L @
Memory: 2TiB DRAM	Memory: 3TiB DCPMM + 768GiB
Storage: 16Gbps FC all-flash SAN @ 2TiB	Storage: 16Gbps FC all-flash SAN
	DCPMM: Memory Mode only

Benchmark Results and Conclusions

There is a significant reduction in elapsed runtime (approximately 6x) when comparing 512GiB to either 1TiB and 2TiB DCPMM buffer pool sizes. In addition, it was observed that in comparing 2TiB E7v4 DRAM and 2TiB Cascade Lake DCPMM configurations there was a ~20% improvement as well. This 20% gain is believed to be mostly attributed to the new processor architecture and more processor cores given that the buffer pool sizes are the same. However, this is still significant in that in the 4-socket Cascade Lake tested only had 24 x 128GiB DCPMM installed, and can scale to 12TiB DCPMM, which is about 4x the memory of what E7v4 can support in the same 4-socket server footprint.

The following graphs in figure-2 depict the comparison results. In both graphs, the y axis is elapsed time (lower number is better) comparing the results from the various configurations.

Figure-2: Elapse time comparison of various configurations



WRITE-INTENSIVE WORKLOAD BENCHMARK

The workload in this benchmark was our HL7v2 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 (1-to-4 with four transforms). 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 4800 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 critical components in sustaining high throughput. This directly influences the performance of message durability guarantees of IRIS for Health, and the transaction journal write performance is crucial in that operation. When journal throughput suffers, application processes will block on journal buffer availability.

Hardware Configuration

For the write-intensive workload, we decided to use a 2-socket server. This is a smaller configuration than our previous 4-socket configuration in that it only had 192GB 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 Processors (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 2-socket configurations:

Figure-3: Write intensive workload hardware configurations

Server #1 Configuration	Server #2 Configuration
Processors: 2 x Gold 6152 @ 2.1Ghz	Processors: 2 x Gold 6252 @ 2.1G
Memory: 192GiB DRAM	Memory:1.5TiB DCPMM + 192GiB
Storage: 2 x 750GiB P4800X Optane SSDs	Storage: 2 x 750GiB P4800X Opta

	DCPMM: Memory Mode & App Di

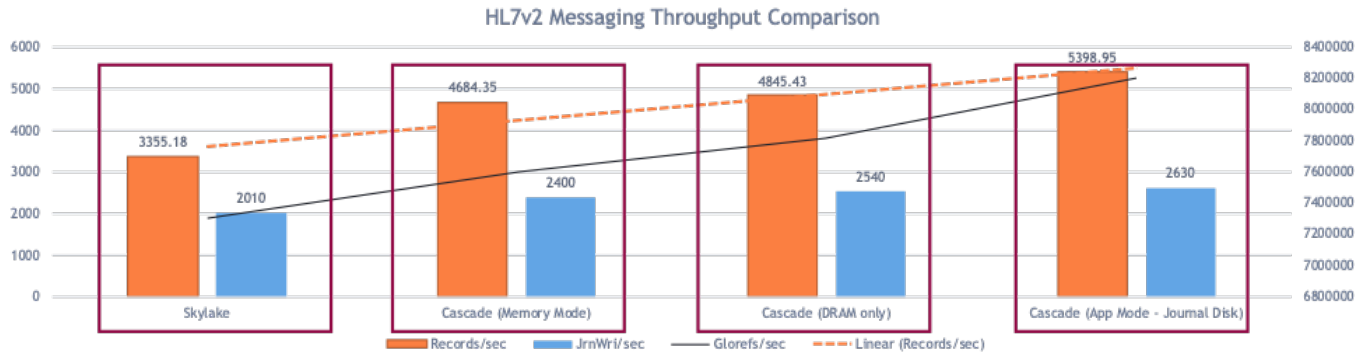
Benchmark Results and Conclusions

Test-1: This test ran the ***T4 workload*** described above on the Skylake server detailed as Server #1 Configuration in Figure-3. The Skylake server provided a sustained throughput of ~3355 inbound messages a second with a journal file write rate of 2010 journal writes/second.

Test-2: This test ran the same workload on the Cascade Lake server detailed as Server #2 Configuration in Figure-3, and specifically with DCPMM in ***Memory Mode***. This demonstrated a significant improvement of sustained throughput of ~4684 inbound messages per second with a journal file write rate of 2400 journal writes/second. **This provided a 39% increase compared to Test-1.**

Test-3: This test ran the same workload on the Cascade Lake server detailed as Server #2 Configuration in Figure-3, this time using DCPMM in App Direct Mode but not actually configuring DCPMM to do anything. The purpose of this was to gauge just what the performance and throughput would be comparing Cascade Lake with DRAM only to Cascade Lake with DCPMM + DRAM. Results we not surprising in that there was a gain in throughput without DCPMM being used, albeit a relatively small one. This demonstrated an improvement of sustained throughput of ~4845 inbound message a second with a journal file write rate of 2540 journal writes/second. This is expected behavior because DCPMM has a higher latency compared to DRAM, and with the massive influx of updates there is a penalty to performance. Another way of looking at it there is <5% reduction in write ingestion workload when using DCPMM in Memory Mode on the same exact server. Additionally, when comparing Skylake to Cascade Lake (DRAM only) **this provided a 44% increase compared to the Skylake server in Test-1.**

Test-4: This test ran the same workload on the Cascade Lake server detailed as Server #2 configuration 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 5399 inbound messages per second with a journal file write rate of 2630/sec. This demonstrated that DCPMM in App Direct mode for this type of workload is the better use of DCPMM. Comparing these results to the initial Skylake configuration there was a **60% increase in throughput compared to the Skylake server in Test-1.**

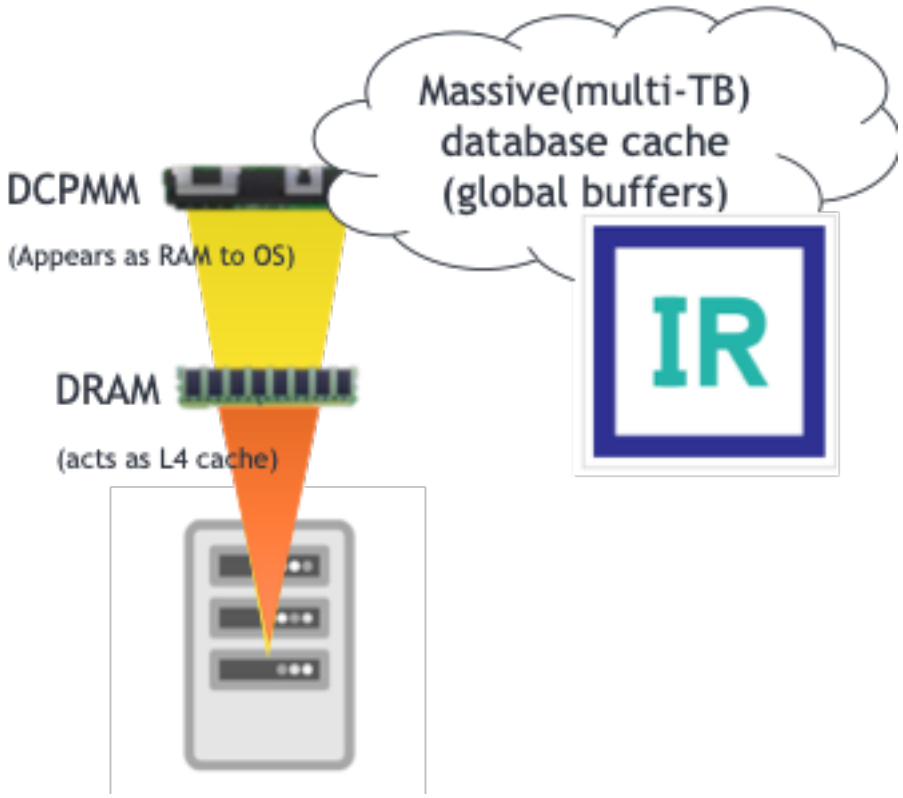


InterSystems IRIS Recommended Intel DCPMM Use Cases

There are several use cases and configurations for which InterSystems IRIS will benefit from using Intel® Optane™ DC Persistent Memory.

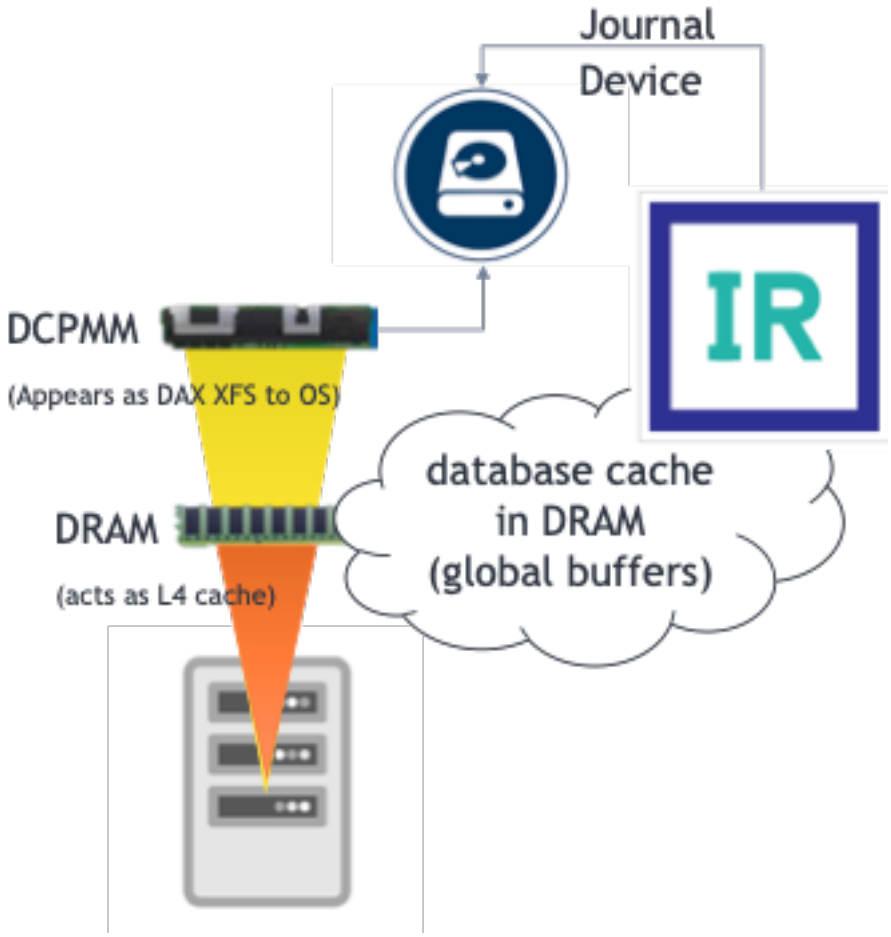
Memory Mode

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. You will want to adhere to a maximum of 8:1 ratio of DCPMM to DRAM as this is important for the “hot memory” to stay in DRAM acting as an L4 cache layer. This is especially important for some shared internal IRIS memory structures such as seize resources and other memory cache lines.



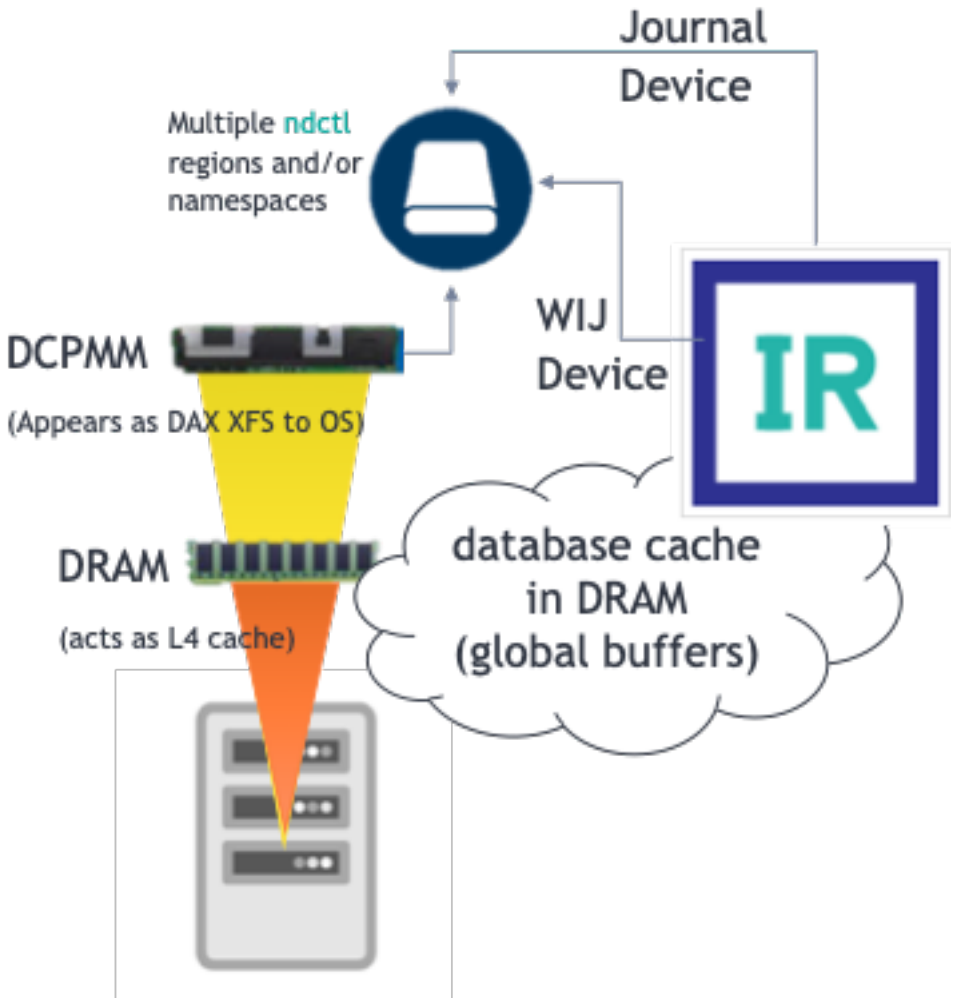
App Direct Mode (DAX XFS) – Journal Disk Device

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 this alleviates the PCIe bus overhead and direct memory access from the file system. As demonstrated in the HL7v2 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.



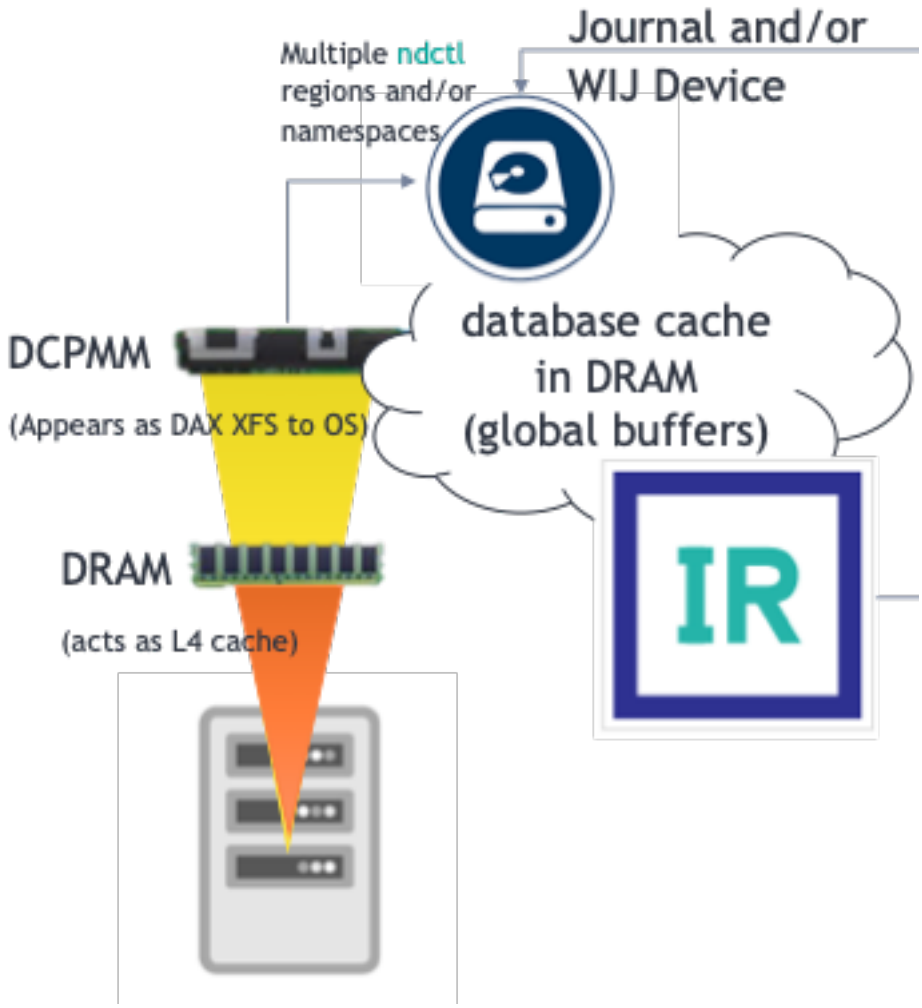
App Direct Mode (DAX XFS) – Journal + Write Image Journal (WIJ) Disk Device

In this use case, this 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 ultra-low 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 ultra-low latency for the transaction journals and/or write image journal devices. In this use case, DCPMM appears both as mounted XFS filesystem to the OS and RAM to operating systems. This is achieved by allocating a percentage of DCPMM as DAX XFS and the remaining 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 a concurrent transaction and analytic workloads (also known as HTAP workloads) type workload where there is a high rate of inbound transactions/updates for OLTP type workloads, and also an analytical or massive querying need, then having each InterSystems IRIS node type within an InterSystems IRIS sharded cluster operating with different modes for DCPMM.

In this example there is the addition of InterSystems IRIS compute nodes which will handle the massive querying/analytics workload running with DCPMM Memory Mode so that they benefit from massive database cache in the global buffers, and the data nodes either running in Dual mode or App Direct the DAX XFS for the transactional workloads.

Conclusion

There are numerous options available for InterSystems IRIS when it comes to infrastructure choices. The application, workload profile, and the 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 your InterSystems IRIS based applications that matter to your business.

Benefits of InterSystems IRIS and Intel DCPMM capable servers include:

- Increases memory capacity so that multi-terabyte databases can completely reside in 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 due to InterSystems IRIS proven memory caching capabilities that take advantage of system memory as it increases in size.
- 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 using the same processors and only changing 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.
- Augment the compute resources where needed for a given workload whether read or write-intensive, or both, without over-allocating 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 InterSystems IRIS based application.

#Big Data #HL7 #Interoperability #InterSystems Business Solutions and Architectures #Performance #Platforms #Sharding #Testing #HealthShare #InterSystems IRIS #InterSystems IRIS for Health #TrakCare
50 1 0 5 464

Log in or sign up to continue
Add reply

Source URL: https://community.intersystems.com/post/intersystems-iris-and-intel-optane-dc-persistent-memory