White Paper

High Performance Computing

Performance Evolution of DAOS Servers

A Sneak Preview of DAOS on 4th Gen Intel® Xeon® Scalable Processors, Formerly Codenamed Sapphire Rapids

Michael Hennecke

Principal Engineer, HPC Storage

Table of Contents

Executive Summary

Distributed Asynchronous Object Storage (DAOS) is an open source software stack for Linux (available at https://github.com/daos-stack/daos) that provides a scalable all-flash storage solution, based on Storage Class Memory (SCM) and NVMe SSDs. By using SCM in the form of Intel® Optane™ Persistent Memory (Intel Optane PMem), DAOS eliminates many of the bottlenecks of block-based I/O and existing global parallel filesystems and delivers exceptionally high performance. With DAOS Version 2.2 released, the DAOS engineering team is now focusing on enabling the 4th Gen Intel® Xeon® Scalable processors (formerly codenamed Sapphire Rapids) as the hardware foundation for the next generation of DAOS storage servers.

Performance Evolution of DAOS Server Generations

Figure 1 shows the typical configuration of three generations of Intel Xeon processor-based 2-socket DAOS servers. To avoid NUMA effects, DAOS runs one instance of the daos_engine I/O engine per physical socket, which manages the hardware resources that are local to that socket: service threads on the local CPU cores, high-performance network cards, Storage Class Memory (Intel Optane PMem), and PCIe-attached NVMe SSDs for bulk storage.

2nd Gen Intel Xeon Scalable Processors

2nd Gen Intel Xeon Scalable processors were the first Intel Xeon processor generation to support Intel Optane Persistent Memory 100 Series. As shown on the

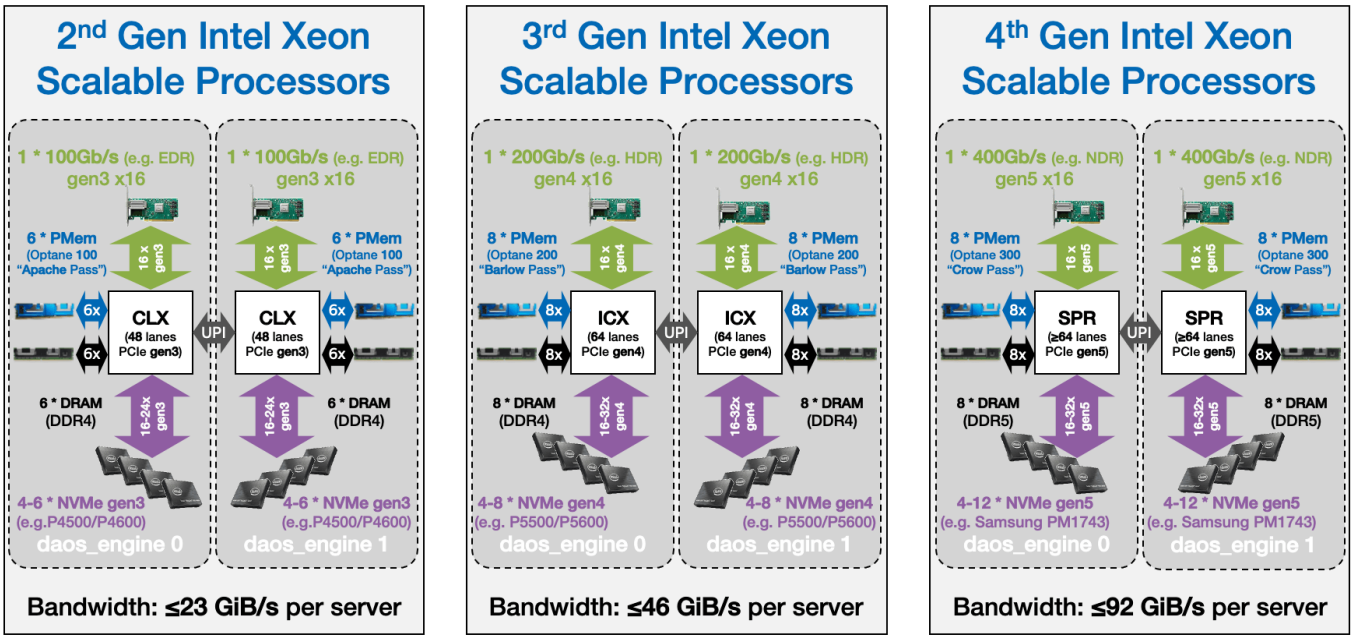

Figure 1: DAOS server components and performance evolution on Intel Xeon Scalable processors

left of **Figure 1**, 2nd Gen Intel Xeon Scalable processors use PCIe gen3 with 48 lanes per socket. A balanced 2nd Gen Intel Xeon Scalable processor-based DAOS server typically has one 100 Gbps network port (like InfiniBand EDR) per socket, and between 4 and 6 NVMe disks per socket. The NVMe storage is complemented by about 3–6% of Intel Optane PMem capacity for metadata and to buffer small (<4 kiB) I/O requests. For maximum performance, the recommended Intel Optane PMem configuration is to populate one Intel Optane PMem DIMM on each of the six memory channels, which is then used in AppDirect (interleaved) mode. The peak bandwidth of such a 2nd Gen Intel Xeon Scalable processor-based server is roughly **23 GiB/s**, determined by the wire speed of the two EDR links. Details on the NVMe disk population are discussed below.

3rd Gen Intel Xeon Scalable Processors

With the introduction of 3rd Gen Intel Xeon Scalable processors, DAOS servers did benefit from an up to 2x boost of storage bandwidth, using the same system design. This is shown in the center of **Figure 1**: The PCIe generation changed from gen3 to gen4, and the number of PCIe lanes per socket increased from 48 to 64. With two 200 Gbps network links (like InfiniBand HDR) and the newer PCIe gen4 NVMe SSDs, the peak bandwidth per server doubles. The testing limit with DAOS v2.0 in the 3rd Gen Intel Xeon Scalable processor timeframe has been 8 NVMe disks per socket (which is why **Figure 1** shows 4–8 NVMe SSDs per socket for the 3rd Gen Intel Xeon Scalable processor-based servers). The number of memory channels for the 3rd Gen Intel Xeon Scalable processors did increase from six to eight. This means that in addition to organic improvements due to the newer Intel Optane Persistent Memory 200 Series (codenamed Barlow Pass), the size and performance of an AppDirect Intel Optane PMem region also grew due to the increase from six to eight Intel Optane PMem devices. All these improvements combined lead to a peak bandwidth of roughly **46 GiB/s**, again determined by the wire speed of the two HDR links.

4th Gen Intel Xeon Scalable Processors

The 4th Gen Intel Xeon Scalable processors with PCIe gen5 will enable an up to 2x generation-to-generation performance boost compared to 3rd Gen Intel Xeon Scalable processor. The peak DAOS bandwidth is expected to reach up to 92 GiB/s for a 2-socket 4th Gen Xeon Scalable processor-based server, as shown on the right of **Figure 1**. There are two important considerations when moving to 4th Gen Intel Xeon Scalable processor-based DAOS servers:

1. The PCIe gen5 connectivity of 4th Gen Intel Xeon Scalable processors enables two 400 Gbps HPC network interfaces (like NDR InfiniBand) per server. Some HPC/AI customers will quickly adopt 400 Gbps HPC networking, and in that case the DAOS server design will remain unchanged compared to 3rd Gen Intel Xeon Scalable processors. But many customers may continue to operate 200 Gbps HPC networks. DAOS needs to support those configurations as well, which requires a slight change in the DAOS server design: Until now, DAOS assumes one network interface per storage engine, because the network stacks that DAOS uses (libfabric and UCX) do not support striping across multiple RDMA interfaces. To support for example two 200 Gbps HDR network links per 4th Gen Intel Xeon Scalable processor socket (instead of a single 400 Gbps NDR link), we need to run two DAOS storage engines per 4th Gen Intel Xeon Scalable CPU socket. This needs two namespaces on the single AppDirect (interleaved) Intel Optane PMem region, as each engine manages its own Intel Optane PMem space. The DAOS command option to configure two namespaces per socket will become generally available with DAOS v2.4. For early evaluations of this functionality with DAOS v2.2, a Wiki document is available that describes how to manually prepare the Intel Optane PMem for two engines per socket: https://daosio.atlassian.net/wiki/spaces/DAOS/pages/11158814734/Create+multiple+SCM+namespaces+per+CPU+socket
2. To fully utilize the 400 Gbps of network bandwidth per 4th Gen Intel Xeon Scalable CPU socket (either a single 400 Gbps port, or two 200 Gbps ports per socket), an appropriate number of NVMe SSDs must be populated. Ideally these should be PCIe gen5 disks, but availability of PCIe gen5 NVMe SSDs is still very limited. Roughly twice the number of PCIe gen4 devices could be used as an alternative; this is discussed in more detail in the next section.

Finally, 4th Gen Intel Xeon Scalable processors supports the Intel® Optane® Persistent Memory 300 Series (formerly codenamed Crow Pass), which provides improved performance over the Intel Optane PMem 200 Series as well as some advanced features. Similar to the earlier generations, it is recommended to populate one Intel Optane PMem device per memory channel.

Choosing the Number of NVMe Disks per Server

In the discussion above, the peak bandwidth was determined by the network bandwidth of the DAOS servers. In addition to the maximum network bandwidth, the main factor that impacts the achievable per-server DAOS bandwidth is the type and quantity of NVMe disks. An important consideration is the fact that for NAND flash media and controllers, the read bandwidth is typically much higher than the write bandwidth. (This is not true for Intel Optane NVMe SSDs, which are based on Intel® 3D XPoint™ and have almost identical read and write bandwidth.)

This asymmetry of read and write bandwidth of NVMe SSDs creates a dilemma when sizing an NVMe storage server (which is a general concern for sizing HPC/AI storage solutions – it is not specific to DAOS): To optimize the price/performance of the solution, it is always advisable to populate at least as many NVMe SSDs per server as necessary to fully saturate the *network read* bandwidth. But for environments that also have a significant fraction of write-intensive workloads, it is often advisable to use more NVMe SSDs, given their lower per-device write bandwidth. At some point, the quantity of NVMe SSDs will be sufficient to also saturate the *network write* bandwidth. Adding even more NVMe SSDs to the server will still grow the capacity, but the bandwidth will no longer increase. Eventually, the total number of NVMe SSDs that can be connected will be limited by the number of available PCIe lanes (minus the PCIe lanes that are needed for the HPC networking adapters).

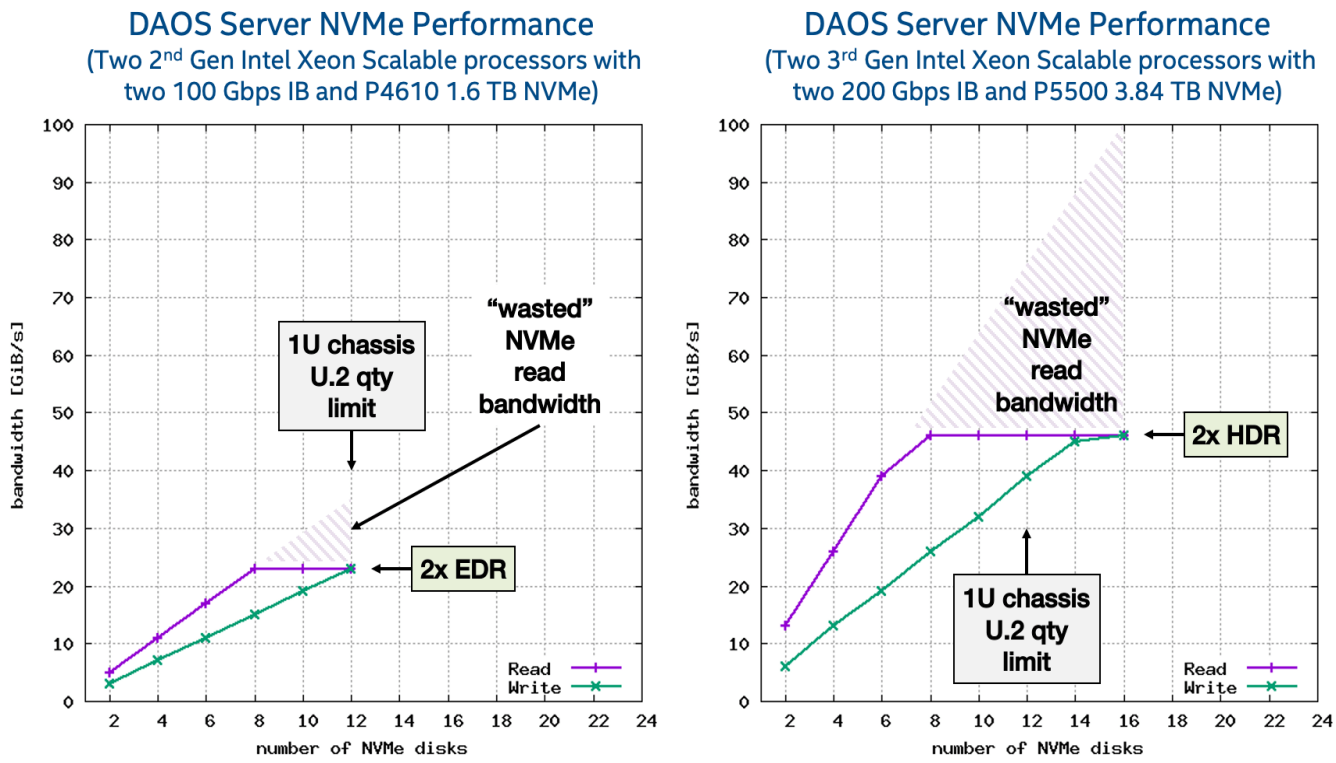

Figure 2: DAOS server NVMe performance – 2nd and 3rd Gen Intel Xeon Scalable processors

Figure 2 shows this design space for 2nd Gen and 3rd Gen Intel Xeon Scalable processor-based DAOS servers. The purple curves show the scaling of the per-server read bandwidth with the number of NVMe disks (per 2-socket server), while the green curves show the (lower) write bandwidth scaling. The exact numerical values depend on the performance details of the specific NVMe SSD model that is used; the NVMe SSD models in **Figure 2** and **Figure 3** should be representative for a wide range of NVMe SSD models.

2nd Gen Intel Xeon Scalable Processors

The left side of Figure 2 shows the situation for 2nd Gen Intel Xeon Scalable processor-based servers, using Intel P4610 1.6TB NVMe SSDs (gen3). The dual-EDR network read bandwidth is saturated with eight NVMe disks (four per socket), while the network write bandwidth is only maximized when using 12 NVMe disks. With 48 PCIe lanes per 2nd Gen Intel Xeon Scalable processor socket, more than 12 disks are not feasible. The diagram also indicates that for a 1U rack-mounted server, twelve U.2 disks is typically the maximum that can be physically populated. Different OEMs have different server form factors, but more than 12 disks in 1U is very rare.

3rd Generation Intel Xeon Scalable Processors

The right side of **Figure 2** shows the corresponding situation for 3rd Gen Intel Xeon Scalable processor-based servers, this time using Intel P5510 3.84TB NVMe SSDs (gen4) with higher per-device bandwidth than the P4610 (gen3) series. The aggregate per-server bandwidth doubles to dual-HDR speed, which for reads will be reached with eight NVMe SSDs. But notice that the peak write bandwidth is only reached with 14 to 16 of these NVMe SSDs, which will typically require a 2U server chassis.

The purple-shaded triangular area that is labeled *"wasted" NVMe read bandwidth* demonstrates the sizing dilemma mentioned above: It shows the increasing aggregate bandwidth of the NVMe devices, which cannot be utilized because of the ceiling that is imposed by the network. When sizing the overall solution, the number of NVMe disks should be selected somewhere in this range between eight and sixteen disks – depending on the known or expected read and write bandwidth requirements of the workload. (A larger number of NVMe disks may also help with workloads that are dominated by small transactions, for which the network bandwidth is not the main bottleneck. But many hardware sizings are still predominantly bandwidth-driven.)

4th Gen Intel Xeon Scalable Processors

Figure 3 shows two different scenarios for DAOS servers based on 4th Gen Intel Xeon Scalable CPUs. In both scenarios, the peak per-server bandwidth is given by the dual-NDR network limit (or the equivalent quad-HDR network limit, if the server uses two HDR links per socket as discussed above).

The difference is in the type of NVMe SSDs:

- The left graph shows the performance scaling with PCIe gen5 NVMe SSDs (Samsung PM1743, for which Samsung has published performance specs in late 2021). This is the ideal situation, where true PCIe gen5 NVMe devices are used: Read bandwidth is saturated with eight disks, and write bandwidth is saturated with 14–16 disks, which matches the 3rd Gen Intel Xeon Scalable processor-based server layout. (Note that some NVMe SSDs available on the market have a PCIe gen5 interface, but do not provide more performance than their PCIe gen4 counterparts.)

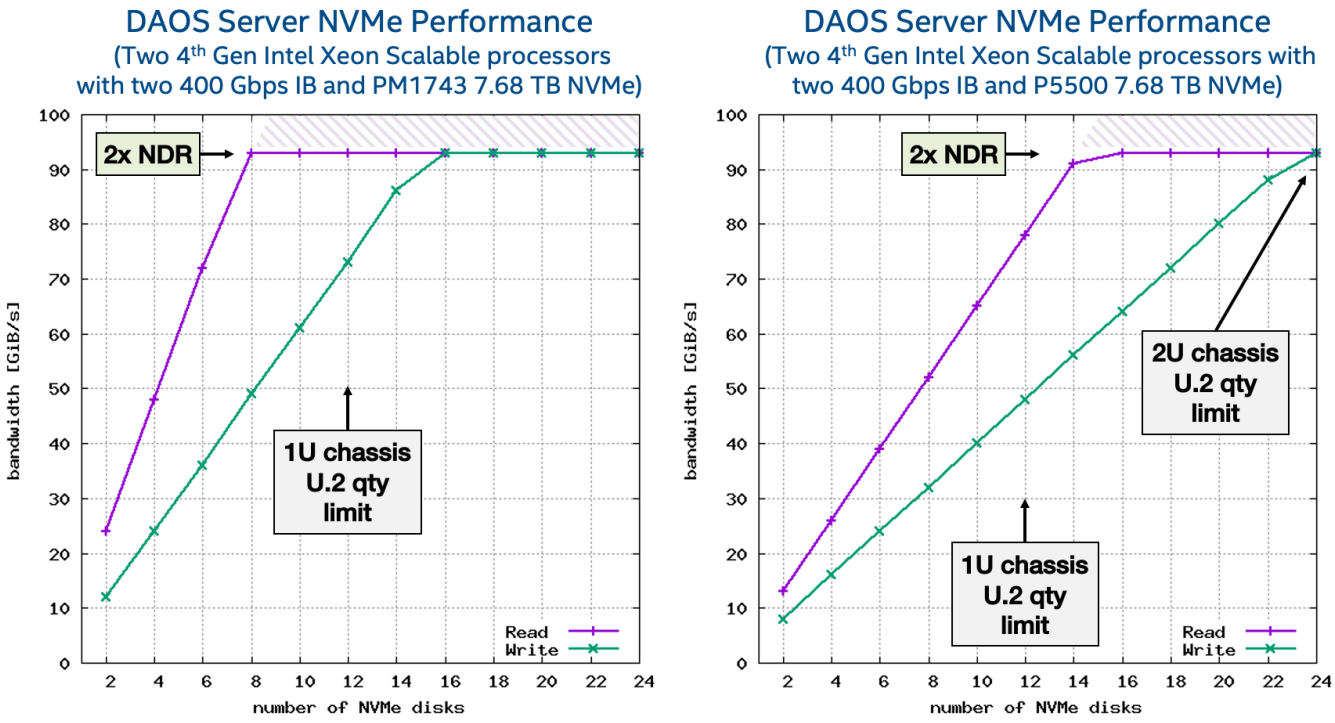

Figure 3: DAOS server NVMe performance – 4th Gen Intel Xeon Scalable processors

- The graph on the right demonstrates how the performance scaling changes if PCIe gen4 NVMe disks are used. The same Intel P5510 SSD series (gen4) is used here that was also used for the 3rd Gen Intel Xeon Scalable processor performance scaling graph in **Figure 2** (although at a higher capacity, which is more typical for the 4th Gen Intel Xeon Scalable processor timeframe). Due to the lower per-device bandwidth, even reaching the network read bandwidth requires sixteen gen4 NVMe SSDs (and a 2U server chassis). To reach the network write bandwidth, twenty-four gen4 NVMe SSDs are needed – which is also the physical limit of the number of U.2 NVMe disks that can be populated in the front of a 2U rack server.

Depending on the timing of DAOS deployments with 4th Gen Intel Xeon Scalable CPUs and the availability of PCIe gen5 NVMe SSDs, one of these two scaling regimes shown in **Figure 3** will apply.

New Features in Next Generation Intel Xeon Scalable CPUs

The previous sections focused primarily on the generation-to-generation performance improvements driven by the evolution of the PCIe connectivity, and the resulting HPC networking and NVMe SSD improvements. But in addition to faster PCIe-attached devices, 4th Gen Intel Xeon Scalable processors also provides several important functional enhancements. Some of these new features will be exploited in future DAOS software releases.

Intel® Data Streaming Accelerator (Intel DSA)

One major acceleration engine that is built into 4th Gen Intel Xeon Scalable processors is the Intel Data Streaming Accelerator (Intel DSA). Intel DSA is a high-performance data copy and transformation accelerator. It is targeted for optimizing streaming data movement and transformation operations that are common with high-performance storage, networking, persistent memory, and various data processing applications.

The Intel DSA Architecture Specification is available at https://software.intel.com/en-us/download/intel-data-streaming-accelerator-preliminary-architecture-specification. Intel DSA replaces Intel QuickData Technology, which is a part of Intel I/O Acceleration Technology.

The goal of Intel DSA is to provide higher overall system performance for data movement and transformation operations, while freeing up CPU cycles for higher level functions. Intel DSA hardware supports high-performance data mover capability to/from volatile memory and persistent memory. This can be exploited in a DAOS server to offload these functions from the CPU cores. In particular, DSA will accelerate the aggregation process that DAOS servers run in the background to coalesce small writes. In addition to performing basic data mover operations, Intel DSA is also designed to perform a number of higher-level transformation operations on memory. For example, it can generate and test CRC checksums or perform Data Integrity Field (DIF) calculations at very high speed.

Intel® QuickAssist Technology (Intel QAT)

The second acceleration technology on 4th Gen Intel Xeon Scalable processors that will be beneficial for DAOS is the Intel QuickAssist Technology (Intel QAT), which can accelerate cryptography and data compression/decompression.

For an overview of the Intel QAT technology, see https://www.intel.com/content/www/us/en/developer/topic-technology/open/quick-assist-technology/overview.html.

With 4th Gen Intel Xeon Scalable processors, the on-chip Intel QAT acceleration engines will enable very high throughput. Enabling inline compression with Intel QAT acceleration is a DAOS roadmap feature for 4th Gen Intel Xeon Scalable processor-based DAOS nodes.

Summary

The 4th Gen Intel Xeon Scalable processors will give DAOS servers an up to 2x performance boost over the current generation of 3rd Gen Intel Xeon Scalable processor-based DAOS servers. For bandwidth, this is primarily due to the introduction of PCIe gen5 and the associated higher-speed HPC networking adapters and NVMe SSDs. In addition, DAOS will benefit from the significant performance increase and expanded functionality of the new Intel Optane Persistent Memory 300 Series (codenamed Crow Pass).

The Intel DSA hardware acceleration engines in 4th Gen Intel Xeon Scalable processors will significantly speed up data movement operations between memory and persistent memory, notably for DAOS background aggregation. Intel QAT will enable further performance improvements for advanced functions like in-line compression or encryption. A first preview of the performance capabilities of these accelerators has been given in the "Sapphire Rapids" presentation at the 2021 "Hot Chips" conference. Supporting these advanced acceleration features is on the DAOS roadmap (across several releases). Early performance results for those accelerators are very promising.

For more details on DAOS, please refer to the SCFA 2020 article on DAOS, the online documentation at https://docs.daos.io/, and the Intel landing page for DAOS.

092022/RJMJ/MH/PDF ♻ Please Recycle 352173-001US