SOLUTION BRIEF

DESIGNING AN EFFICIENT DISTRIBUTED STORAGE SYSTEM WITH INTEL DAOS AND SUPERMICRO HARDWARE

Supermicro Ultra

TABLE OF CONTENTS

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Introduction

Several years ago, the Intel HPC storage team began to address the challenges found in Lustre scalability. Being experienced with massive scale deployments, we saw the limitations of large installations' file system performance scalability. As storage media had evolved over the years with SSD technologies, these limitations came mainly from the filesystem software stack rather than limited storage media technology. The fundamental limitations of the legacy storage software stack became even more apparent later when much faster Non-Volatile Memory technologies appeared in the market. There were several bottlenecks in the legacy software stack, but the most prominent of these are the POSIX interface and block-based IO.

To ensure data consistency, the definition of the POSIX interface requires pessimistic locking, proactively locking in any situation where a conflict might occur. Imagine if tens of thousands of clients were writing to the same file and the amount of serialization that would arise if locking occurred on every relevant operation, even if none of these writes would generate a conflict? This pessimistic approach just does not scale, so it became necessary to consider other ways to provide data consistency to unlock the filesystem for scalability.

Similarly, all persistent I/O up to this point has been done on media that we write to in large size blocks. This poses a real performance dilemma for I/Os smaller than the block size, including file system metadata as small IOs sharing a block would then be cause for more locking and serialization of activity and less parallelism. Again, this could not be solved with software alone but needed a change with both software and storage media to overcome.

The work resulted in developing a fundamentally new software stack called DAOS (Distributed Asynchronous Object Storage) based on the new platform technologies to leverage persistent memory and NVMe SSD capabilities and ground up software development without legacy design overhead.

DAOS Overview

DAOS (Distributed Asynchronous Object Storage) software stack is designed from the ground up for performance, combining persistent memory and NVMe SSDs with direct access through the Persistent Memory Development Kit and the Storage Performance Development Kit user space libraries to interface to the media directly, and provides a high efficiency software stack with rich functionality over RDMA-enabled fabric.

DAOS engine can be described as a highly efficient, highly performant object store. Data stored in KV format efficiently balances between PMEM and NVMe tiers for the most advantage of access time, IOPS, and bandwidth. Moreover, users can interface to various datasets (DAOS Containers) using corresponding middleware, which provides compatibility with existing applications and innovative capabilities. Those are POSIX, MPI-IO, HDF5, HDFS connector for Apache Spark and Hadoop, DAOS Python integration, and the native DAOS API that's already adopted by some HPC codes.

For example, POSIX adapter provides file access and full POSIX support for existing and legacy applications, usually considered a starting point for DAOS deployments. Other adapters such as HDF5 and MPI-IO are standard for HPC codes and proved an easy way to integrate those. With the DAOS Python library, users can integrate DAOS access into Python applications. DAOS containers are identified by unique UUIDs. Users can assign specific features to containers or objects inside those, such as a replica factor, access control ACL, checksum, etc. Containers are part of the DAOS Pool, a set of DAOS hardware allocated and managed by the DAOS admin.

DAOS Software Eco-system

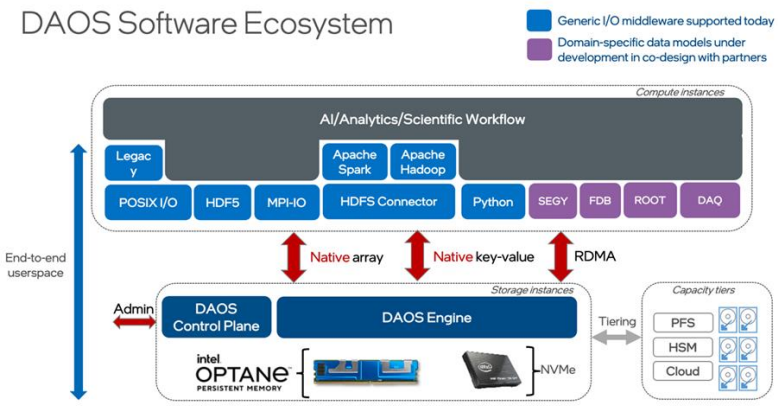

Figure 1 - DAOS Software Ecosystem

DAOS Storage Server Hardware Design Architecture

Intel® Optane™ Persistent Memory (PMem) is a fundamental technology that DAOS builds upon. Unlike SSDs, Intel Optane Persistent Memory offers low latency byte-granular access, which means that applications can now access small pieces of data in parallel, free from the locking required by blocks. DAOS takes advantage of this by storing two different data types in the persistent memory – metadata and small I/Os, which the DAOS policy engine defines. This allows it to optimize bulk data writes to NVMe SSDs for better SSD performance where larger block size delivers better bandwidth and SSD endurance which is dependent on the write pattern. In addition, this enables customers to use less expensive SSDs, such as moving from high-endurance SSDs to mid-endurance or standard-endurance drives or even moving between technologies such as TLC-based NAND SSDs to QLC SSDs.

Breaking through these barriers would never be possible without this storage hardware innovation, the transition away from legacy SATA interface to NVMe, and the appearance of persistent memory, as well as coordinated software stack innovation for data management and data placement.

This paper focuses on DAOS server node design based on the 3rd Generation Intel® Xeon® Scalable Processors, Intel Optane Persistent Memory 200 series, and modern PCIe Gen4 NVMe SSDs. The server platform advantages over 2nd Generation Intel® Xeon® Scalable Processors and Optane Persistent Memory 100 series are translated into substantial DAOS server performance improvement with higher TCO advantages for users.

All DAOS I/O operations on the critical performance path involve Intel Optane PMem. So, the optimal server design and the hardware layout are the keys to delivering the DAOS server performance and scaling it across the distributed installation. With gen over gen performance improvements of the Persistent Memory 100 series vs. 200 series are up to 32% average improvement for random 70/30 mix operations. The new server platform design based on the 3rd Gen Intel Xeon Scalable processor significantly improves overall memory bandwidth by providing eight memory channels instead of the six provided on the prior generations. It's important to note several improvements in the PMEM module design and their contribution to DAOS performance. The most significant improvement comes from the module speed, growing from 2666 MT/s (Mega Transfers Per Second) from the previous generation to 3200 MT/s with a full two DIMMs per channel population.

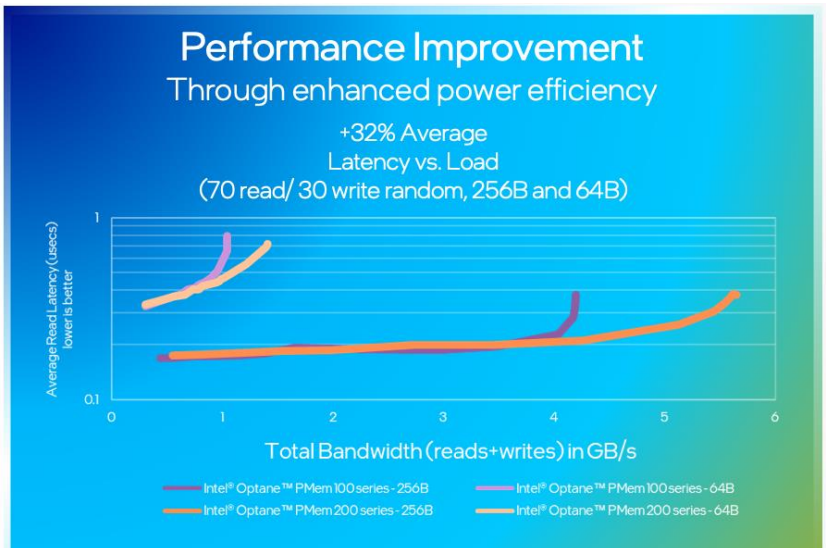

Figure 2 - Generation over Generation Performance Improvement

Also, data had to be flushed from the processor caches using the WPQ Flush command in the past. However, PMEM 200 Series introduced a new platform level capability called eADR. In a platform that supports the energy store needed, eADR allows the platform to flush the data from the processor's caches, so the program need not. This allows DAOS, via PMDK, to eliminate a point of synchronization and deliver more IOPs on writes.

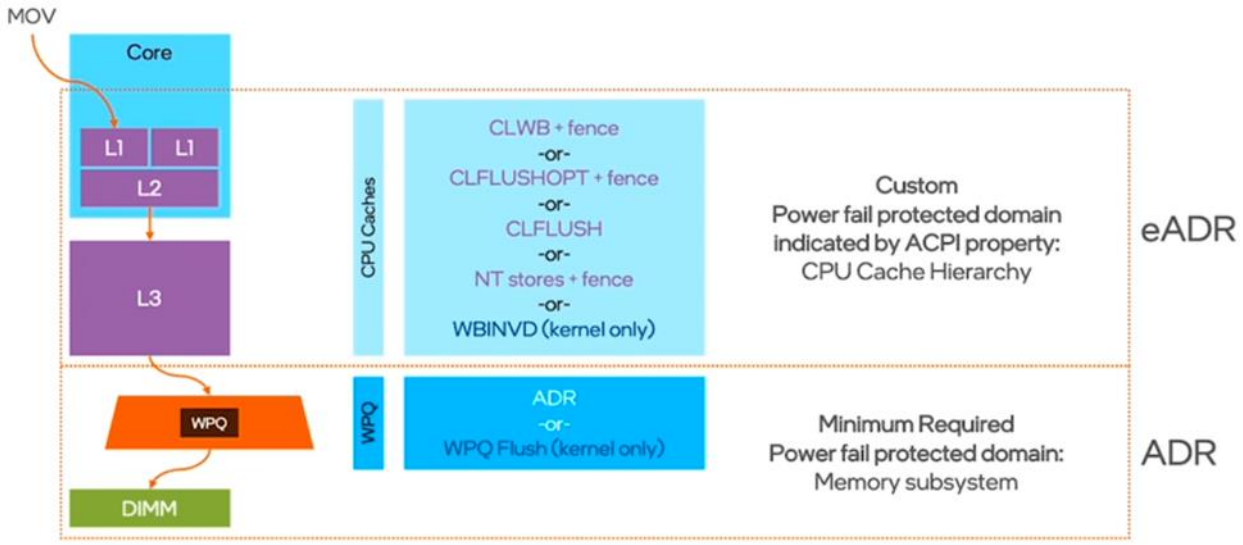

Figure 3 - Extended ADR (eADR) to ADR comparison

Finally, the current server platform has a better I/O subsystem by introducing PCIe Gen4 support with 64 lanes per socket. This allows to leverage faster interconnect fabric such as InfiniBand HDR200 and could be utilized further to take advantage of PCIe Gen4 NVMe SSDs to drive bandwidth performance.

Reference DAOS Storage Node configuration

As highlighted earlier, for production DAOS installation, several hardware technologies are required. Those are Intel Optane Persistent Memory, NVMe class SSDs (3D NAND TLC or QLC), and low latency high performance fabric such as HDR200 InfiniBand from Mellanox. All DAOS servers are configured in the same way, and there are no dedicated metadata servers or head nodes, or monitoring nodes as part of DAOS design. Therefore, the DAOS server layout and capacity and performance ratios are essential to maintain optimal configuration.

Traditional 2 socket server configurations are common for DAOS design. However, it's recommended to have a design fully balanced. This includes:

- Intel Optane Persistent Memory is installed in a full population (2:2:2:2), i.e., each memory channel has DRAM and PMEM in the slot on both CPUs. Therefore, underpopulated PMEM leads to DAOS performance impact.
- Two NVIDIA Mellanox InfiniBand interfaces are used, each connected to its own CPU.
- Total SSD bandwidth should be matched to fabric bandwidth. This SSD sequential bandwidth is considered (read or write), as the PMEM layer is responsible for metadata I/O and small I/O data aggregation.

- NVMe SSDs are equally distributed across CPU sockets to minimize cross-socket UPI traffic (preferable)

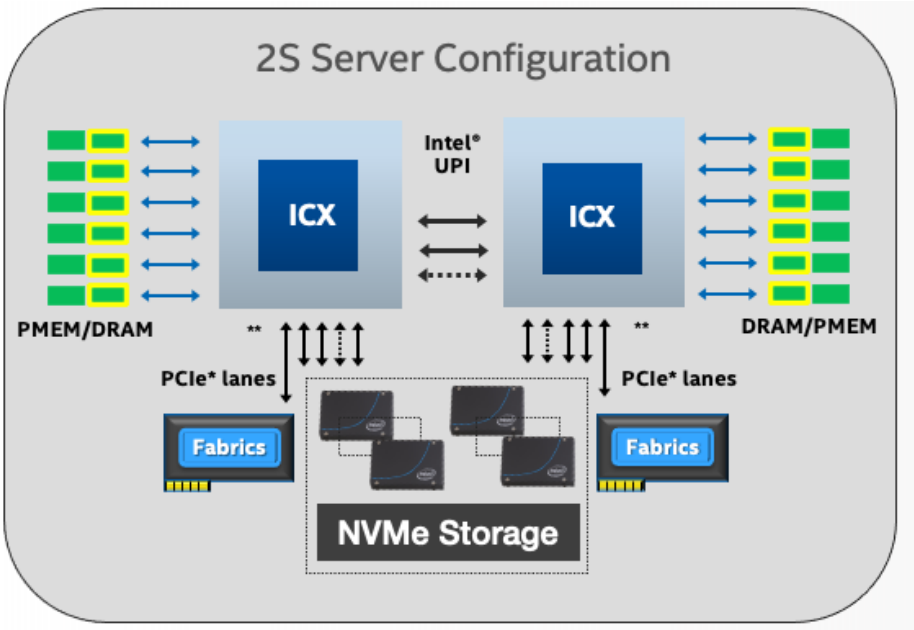

Figure 4 - DAOS Server node design

The capacity ratio should also be maintained. For typical DAOS installations, Metadata size should stay within 3-6% of total data capacity. The unused metadata capacity is utilized for small I/O data aggregation, so the performance benefits an extra PMEM. Intel PMEM is available in 128GB/256GB/512GB capacity points per module. This results in 2TB/4TB/8TB total PMEM in the platform. For typical DAOS server configurations, 128GB and 256GB-based populations provide the highest Return on Investment.

Intel PMEM total capacity	Total Storage capacity (3-6%)	Number of SSDs required (3.84TB each)	Number of SSDs required (7.68TB each)
2048GB (16*128GB)	34TB-68TB	9-18pcs	5-9pcs
4096GB (16*256GB)	68TB-136TB	18-36pcs	9-18pcs

Note that it is essential to consider an optimal SSD capacity while sizing DAOS server total capacity and performance. It might be preferable to use the lower per-drive capacity to achieve full fabric performance for a given total capacity.

For example, if the target DAOS server capacity is 34TB and Intel® SSD D7-P5510 Series 7.68TB selected, the table above clearly illustrates a potential underperforming problem. According to this corner case, only 5 SSDs are required to reach the capacity point, while not enough to drive the full dual HDR200 bandwidth (P5510 is 7GB/s SR and 4.2GB/s SW). In such examples, 10 x 3.84TB drives should be considered instead.

Generally, 3D NAND TLC based standard endurance SSDs are the default choice for DAOS configuration. However, PMEM protects the endurance of SSDs in front by handling small I/O, so SSD writes are always optimal and aligned. This gradually improves SSD endurance and even allows DAOS implementations using QLC NAND technology. For more details, read that Intel whitepaper in the link section below.

Supermicro Reference Configuration for DAOS

Supermicro offers several platforms in different categories with balanced architecture and comprehensive NVMe and PMEM support. However, it is a complex design that includes many specific optimizations for overall power, and thermal and component layout placement.

The Ultra server family offers the perfect balance between performance, flexibility, and economic benefits.

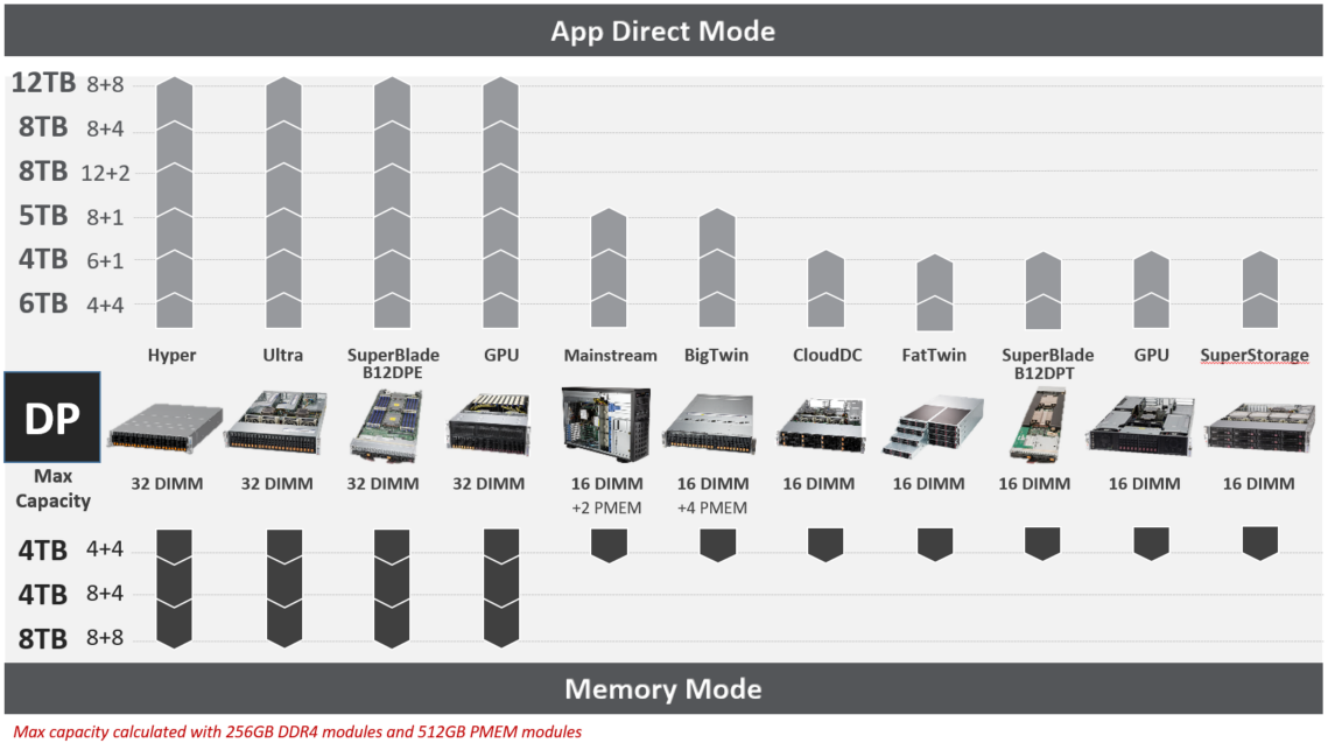

Figure 5 - Memory Capacity with PMEM

Supermicro Ultra in Every Way

All X12 Ultra SuperServers fully support the highest performing 3rd Gen. Intel® Xeon® Scalable processors (up to 270 watts TDP) with 32 DIMM slots with a truly balanced architecture. In addition, when paired with Intel Optane Persistent Memory Module, X12 Ultra SuperServers can achieve a maximum of 12TB total memory capacity, which unlocks the processors' full potential and makes them the excellent choice for high-performance analytics in-memory application acceleration.

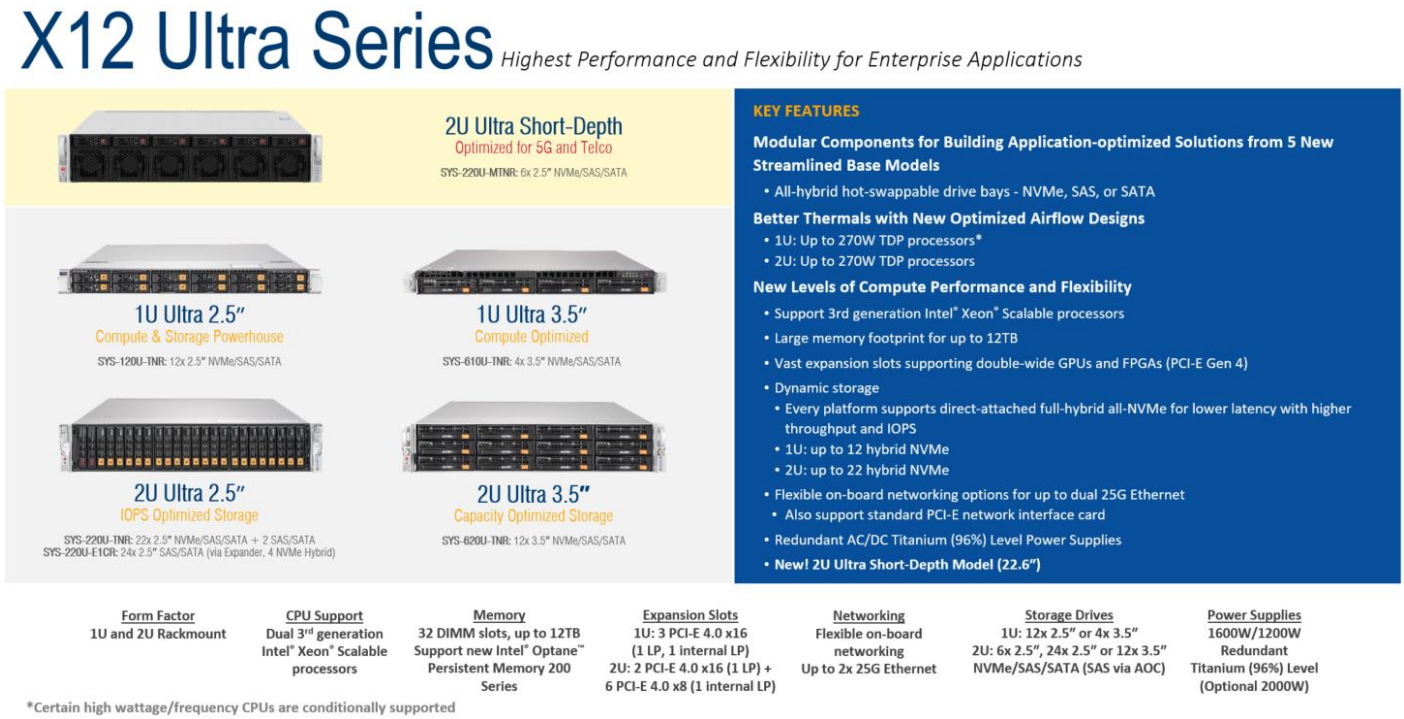

Figure 6 - Supermicro X12 Ultra Series Summary

X12 Ultra SuperServers are designed with extreme configurability and efficiency:

- 1U Rackmount:
  - Up to 12x direct-attach, hot-swappable full hybrid (SATA/SAS/NVMe) drive bays;
  - 4x PCIe Gen 4 Add-on-Cards;
  - Flexible networking options via Ultra risers: 2x 25G or 4x 10G or 2x 10G or No NIC
  - Redundant Titanium (96% Efficiency) AC Power Supply: 800W, 1000W, 1600W, 2000W
  - Redundant -48V DC power supply option: 1300W

- 2U Rackmount:
  - Up to 24x direct-attach, hot-swappable drive bays (22x are NVMe hybrid);
  - 8x PCIe Gen 4 Add-on-Cards;
  - Flexible networking options via Ultra risers: 2x 25G or 4x 10G or 2x 10G or No NIC
  - Redundant Titanium (96% Efficiency) AC Power Supply: 800W, 1000W, 1600W, 2000W
  - Redundant -48V DC power supply option: 1300W

For this DAOS solution, a 1U rackmount X12 Ultra SuperServer SYS-120U-TNR is used. The detailed configuration is shown below. Each CPU socket is configured with 8x 32GB DDR4-3200 DRAM + 8x 128GB PMEM, 5x Intel P5510 NVMe SSDs, and a Mellanox CX-6 200G InfiniBand network adapter. A dedicated HW RAID storage controller is also added for OS boot drives. Thanks to the balanced architecture, all resources can be accessed locally without using the UPI to minimize the data transfer bottleneck.

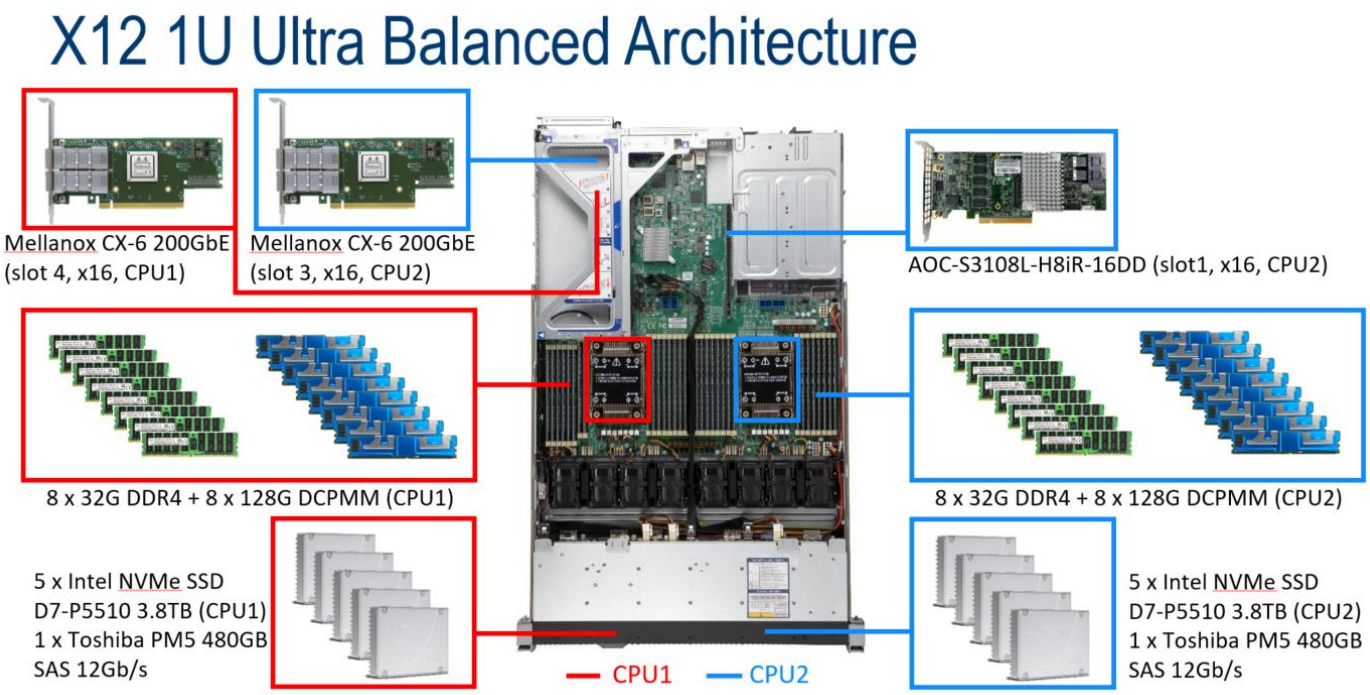

Figure 7 - Balanced Architecture

Hardware Specifics

	Part Number	Description	Qty
Rackmount Server	SYS-120U-TNR	Supermicro X12 1U Ultra 32DIMM 12 Hybrid Drive Bays	1
Processors	P4X-ICX6338-SRKJ9	3rd Gen Intel ® Xeon ® Gold 6338 (32C, 2GHz, 205W)	2
Memory	MEM-IBPS-NMB1XXD128GPSU	Intel® Optane ™ Persistent Memory 200 Series 128G DDR4-3200	16
	MEM-DR432L-HL01-ER32	Hynix 32GB DDR4-3200	16
Storage	HDS-T2A-KPM51RUG480G	Toshiba PM5 480GB SAS 12Gb/s 2.5" 15mm	2
	HDS-IUN0-SSDPF2KX038TZ	Intel D7-P5510 Series 3.84TB PCIe4.0 U.2 NVMe	10
Add on Cards	AOC-653106A-HDAT	Mellanox CX-6 2x 200GbE QSFP56	2
	AOC-S3108L-H8IR-16DD	SAS controller	1
Cables	CBL-SAST-1260-100	SAS Cable	1
	CBL-KIT-120U-TNR-12	NVMe Cable Kit	1
Accessories	AOC-URG4N4-P	Ultra Riser (No NIC)	1

Software Bill of Materials

Operating System	openSUSE Leap 15.2
Open Fabrics Driver	Mellanox 5.2-2.2.3.0
DAOS Version	1.3.101

Performance Evaluation on IO500

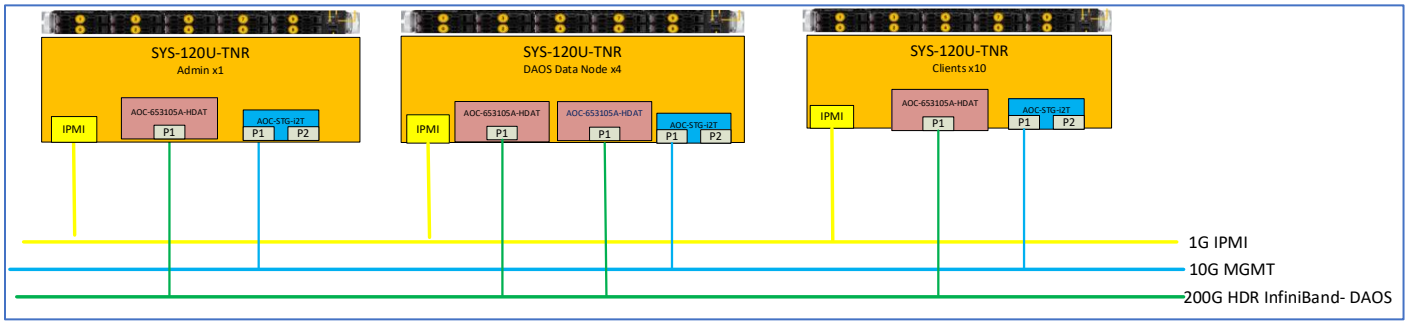

Figure 8 - System Setup for IO500 Testing

BIOS and OS Network tuning parameters

BIOS

Intel® VT for Directed I/O (VT-d)→Enable

Network

```
sysctl -w net.ipv4.conf.all.arp_ignore=1
sysctl -w net.ipv4.conf.ib1.rp_filter=2
sysctl -w net.ipv4.conf.ib0.rp_filter=2
sysctl -w net.ipv4.conf.ib1.accept_local=1
sysctl -w net.ipv4.conf.ib0.accept_local=1
```

The IO500 benchmark suite is an open-source benchmark used to compare HPC filesystems. The suite of tests both Object IO and Metadata IO in easy and hard ways using the standard benchmark tools IOR and Mdtest. The MPI based tests are designed to test system level performance between a group of HPC clients and an HPC Storage system and benchmark the system with both easy and hard io patterns. After all the different subsets are run, a final IO500 score is given.

MDTEST is designed to stress the metadata performance. It's pretty common in HPC to see parallel access to many files opening for reads, writes, attribute updates, creation, or deletion, so the IO500 subsets test this. As a result, metadata performance can create a significant bottleneck in the file system performance and is weighted heavily in the IO500 total score. DAOS has strong metadata performance as locking is minimized within DAOS.

IOR measures throughput performance. Unlike MDTEST, IOR deals with I/O operations, such as reading and writing data simultaneously using many clients. Therefore, the IOR subtests measure the actual bandwidth is as many clients read and write data.

We considered both metadata and object-based benchmarks for DAOS performance analysis with Intel Optane PMem to be critically important. This is because they are representative of the overall filesystem performance and provide clear guidance recognized by the industry.

Below are the results achieved with 10 Clients and 4 Supermicro IceLake DAOS Servers.

IO500 version io500-isc21_v1 (standard)

[RESULT]	ior-easy-write	113.491924 GiB/s : time 526.850 seconds
[RESULT]	mdtest-easy-write	3785.499014 kIOPS : time 603.215 seconds
[RESULT]	ior-hard-write	88.264742 GiB/s : time 418.255 seconds
[RESULT]	mdtest-hard-write	1000.790372 kIOPS : time 525.518 seconds
[RESULT]	find	2430.958209 kIOPS : time 768.913 seconds
[RESULT]	ior-easy-read	163.248695 GiB/s : time 415.989 seconds
[RESULT]	mdtest-easy-stat	4405.937075 kIOPS : time 531.058 seconds
[RESULT]	ior-hard-read	96.820899 GiB/s : time 383.658 seconds
[RESULT]	mdtest-hard-stat	2281.987628 kIOPS : time 356.409 seconds
[RESULT]	mdtest-easy-delete	1720.793799 kIOPS : time 968.443 seconds
[RESULT]	mdtest-hard-read	131.407871 kIOPS : time 2537.581 seconds
[RESULT]	mdtest-hard-delete	1476.884379 kIOPS : time 793.374 seconds

[SCORE] Bandwidth 112.174143 GiB/s : IOPS 1535.629705 kIOPs : TOTAL 415.039693

Subtest Descriptions:

IOR-Easy-*: This is a single file per process workload where each thread writes a single file sequentially with large block IO.

IOR-Hard-*: This is a single shared file workload where each thread writes in unaligned 47KB chunks.

MDTEST-Easy-*: This is a unique directory per process workload where each thread in their own directory creates as many zero-sized files as possible.

MDTEST-HARD-* This is a shared directory workload where each thread creates a small file in the shared directory.

Key Results and Summary

The Supermicro systems with 3rd Gen Intel Xeon Scalable processors have a high level of performance per server compared to earlier 2nd Gen Intel Xeon Scalable processors DAOS system. IOR-EASY-READ was measured at 163 GB/s, which is a little over 40GB/s streaming bandwidth per server used. In addition, the 2 Mellanox InfiniBand HDR fabric ports and 10 NVMe drives in this 1U chassis deliver high performance in a compact package.

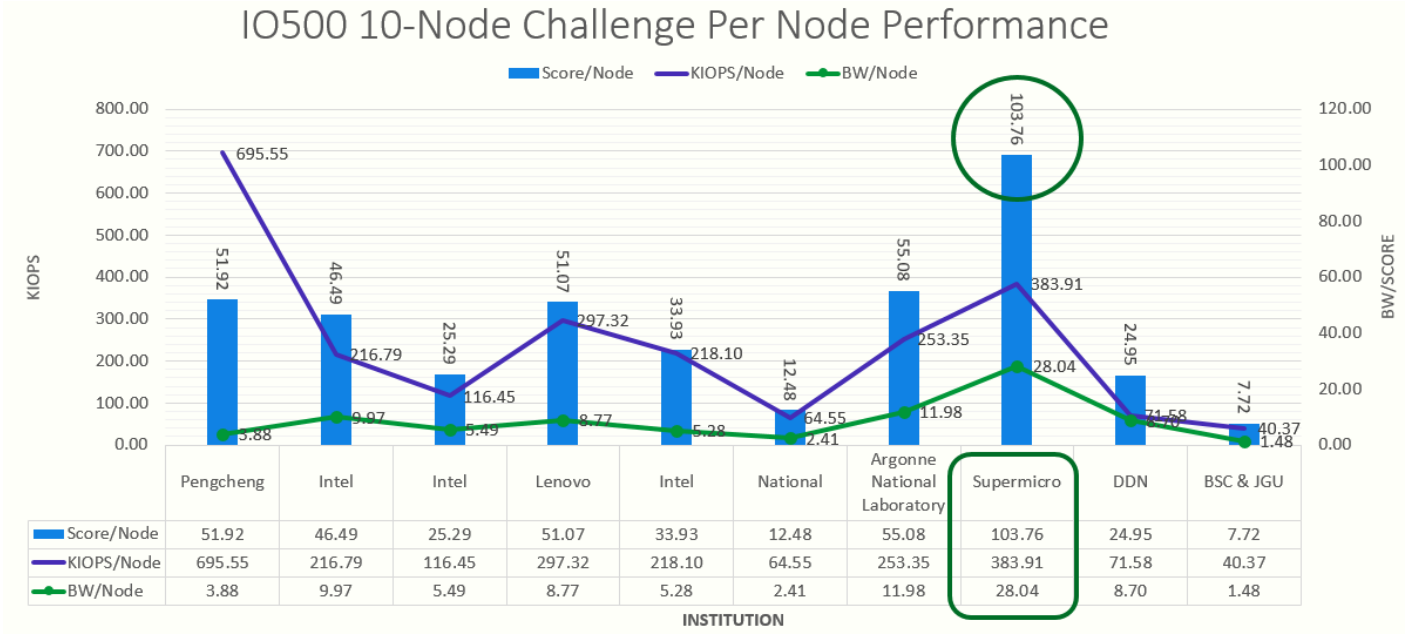

Figure 9 - IO500 10-Node Challenge

Also, in these tests, we can see that IOR-HARD-WRITE was performing at 77% of the IOR_EASY_WRITE. This is because DAOS and its innovative way of I/O allow this DAOS Solution to deliver high levels of IO performance even for complicated I/O patterns. It is not just limited to easy workloads.

For users looking for all flash storage systems that can deliver solid metadata and object IO performance for easy and HARD workloads, a Supermicro DAOS solution should be at the top of the list to consider for your next HPC Flash storage system.

Additional Resources

Supermicro Ultra 1U https://www.supermicro.com/en/products/system/Ultra/1U/SYS-120U-TNR

DAOS main repository:
http://daos.io

DAOS Whitepaper “Revolutionizing High-Performance Storage with Intel® Optane™ Technology”:
https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/high-performance-storage-brief.pdf

Intel® SSD D7 Series product details:
https://www.intel.com/content/www/us/en/products/details/memory-storage/data-center-ssds/d7-series.html

Whitepaper “Achieve High-Performance Storage with DAOS and QLC SSDs”
https://www.intel.com/content/www/us/en/high-performance-computing/daos-high-performance-storage-brief.html

IO500 Storage Benchmark and Performance list http://io500.org