Designing DAOS Solutions with Lenovo ThinkSystem SR630 Servers

Last update: 02 March 2021

Introduces DAOS: Distributed Asynchronous Object Storage

Describes the SR630 hardware configurations for DAOS

Explains the design choices for DAOS storage components

Provides performance and capacity sizing guidance

Michael Hennecke

LENOVO PRESS

Abstract

The Distributed Asynchronous Object Storage (DAOS) is an open source scale-out storage software stack that is designed from the ground up to support Storage Class Memory and NVMe storage in user space. DAOS has been developed in order to overcome limitations in the traditional parallel file systems like Spectrum Scale or Lustre. Those file systems have been originally designed for rotating storage media (HDDs) that are accessed through the operating system's kernel block I/O interface. The latencies in these conventional storage stacks are severely limiting the capabilities of modern storage media like NVMe SSDs. In addition, while parallel file systems can perform well with large sequential I/O operations, they often perform poorly for small, random, or unaligned I/O operation. With the emergence of more and more data intensive applications, a solution is needed that supports both traditional High Performance Computing workloads and data intensive applications on a single high performance storage platform. DAOS is designed to provide these capabilities.

The Lenovo ThinkSystem SR630 server is an ideal hardware platform to run the DAOS server software. It provides a balanced system design, and combines Intel Optane Persistent Memory (PMem) as Storage Class Memory (SCM) with U.2 NVMe SSDs for bulk data storage. Two InfiniBand or high-performance Ethernet fabric ports provide the network connectivity that matches the storage performance of the server. As a software-defined scale-out storage solution, larger DAOS systems can be built by increasing the number of individual DAOS server to create a storage cluster of the desired size.

This paper introduces the DAOS software stack, describes how to design DAOS storage servers with Lenovo ThinkSystem SR630 servers, and provides guidelines for capacity sizing and performance sizing of DAOS storage solutions.

This Planning and Implementation Guide is intended for sales and technical sales specialists, solution architects, and storage administrators who need to understand the DAOS architecture in order to make informed DAOS sizing and configuration decisions. The paper will be most useful for technical professionals who have a working knowledge of high performance storage systems.

It should be noted that the DAOS software stack is still under heavy development, with many new features still being added. The capabilities of the current DAOS 1.1.3 release make it suitable for proof of concept activities, for code porting, and as a scratch storage system. But DAOS should not yet be deployed in mission-critical environments, or as a persistent storage system, before the DAOS 2.0 release that is targeted for 2H2021.

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

Table of Contents

DAOS Overview

The Distributed Asynchronous Object Storage (DAOS) is an open source scale-out storage system for the Exascale era. It is developed primarily by the Intel DAOS development team, and is available on GitHub under a “BSD+Patent” open source license. A high-level overview of DAOS and its motivation can be found in the Intel Solution Brief *“DAOS: Revolutionizing High-Performance Storage with Intel Optane Technology”* that is available online at:

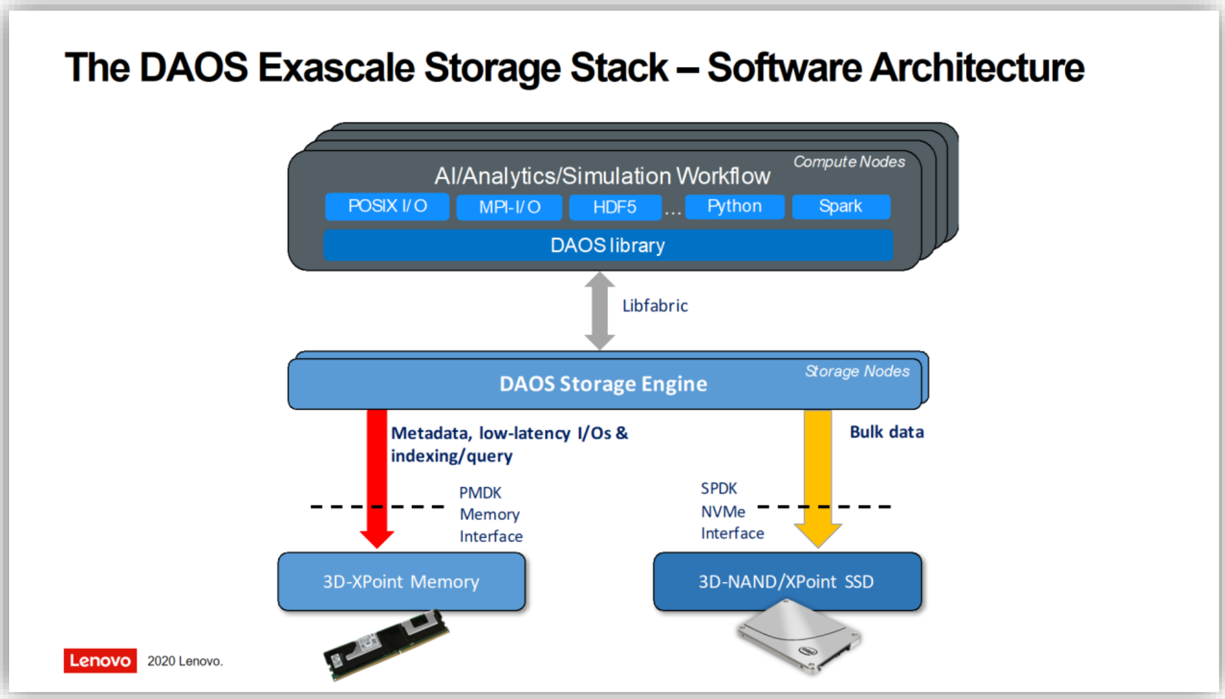
> https://www.intel.com/content/www/us/en/high-performance-computing/daos-high-performance-storage-brief.html

The DAOS software architecture is described in more technical detail in the article *“DAOS: A Scale-Out High Performance Storage Stack for Storage Class Memory”*, available online at:

> https://doi.org/10.1007/978-3-030-48842-0_3

DAOS relies on Storage Class Memory in the form of Intel Optane Persistent Memory (PMem) in AppDirect mode, to provide ultra-low latency and fine-grained access to persistent storage. All metadata and small I/O requests are stored in PMem, using the Persistent Memory Development Kit (PMDK) software framework. DAOS uses NVMe SSDs for bulk data, with user space access through the Storage Performance Development Kit (SPDK). Traditional disk storage devices (HDDs or SAS/SATA SSDs) are not supported by DAOS.

As shown in Figure 1, the DAOS storage stack is based on a client-server model. On the compute nodes, I/O operations are handled in the DAOS library that is directly linked with the application (or with a DAOS-enabled storage middleware). These I/O requests are then processed by DAOS storage services running in user space on the DAOS server nodes. Communication between the clients and servers is performed using `libfabric`.



The DAOS Exascale Storage Stack – Software Architecture

Figure 1: The DAOS Exascale Storage Stack – Software Architecture.

The DAOS software stack depends on Linux as the underlying operating system, on both the DAOS servers and the DAOS clients. DAOS has been tested primarily with CentOS, with some additional test coverage for RHEL and SLES. It can run on a bare-metal server or within containers.

The advanced storage API of the DAOS storage engine (available in `libdaos`) natively supports structured, semi-structured and unstructured data models. This allows DAOS-enabled applications to overcome the limitations of traditional POSIX based parallel filesystems. To support legacy applications that do use POSIX I/O, the DAOS File System (DFS) is available as a software layer on top of `libdaos`. This API is provided through the `libdfs` library, and it can be used from multiple clients in parallel to provide a "global namespace" view to a parallel application.

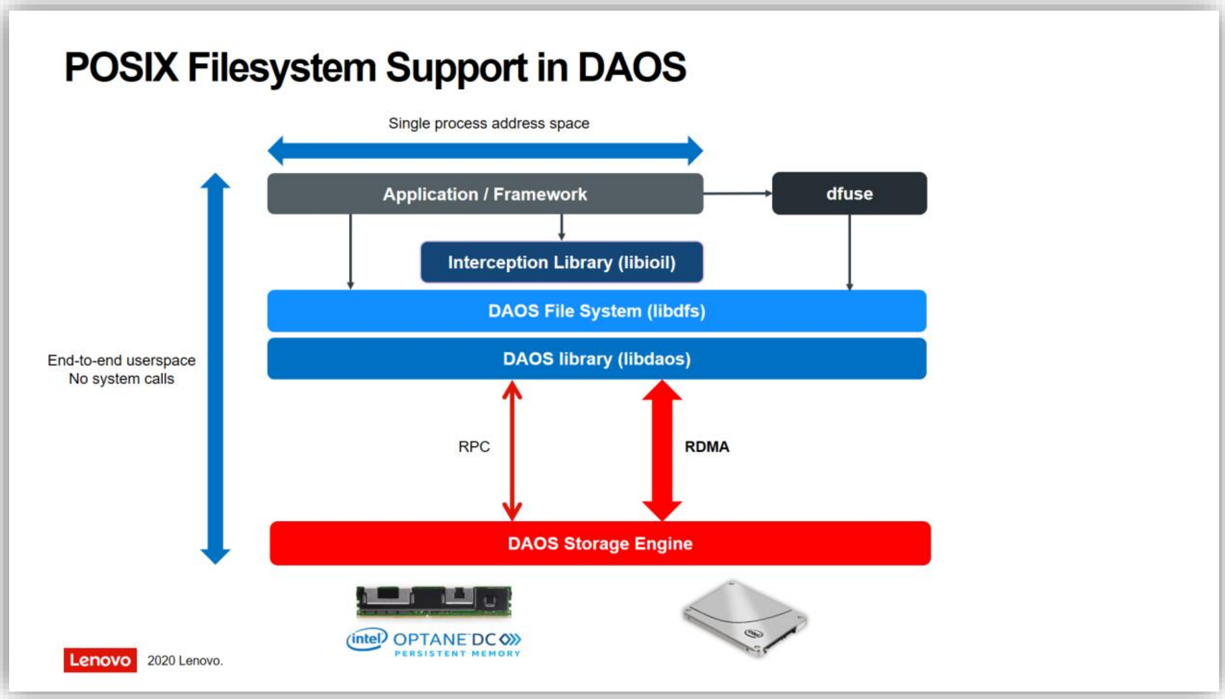

POSIX Filesystem Support in DAOS
Single process address space
Application / Framework
dfuse
Interception Library (libioil)
DAOS File System (libdfs)
DAOS library (libdaos)
End-to-end userspace
No system calls
RPC
RDMA
DAOS Storage Engine
intel OPTANE DC PERSISTENT MEMORY
Lenovo
2020 Lenovo.

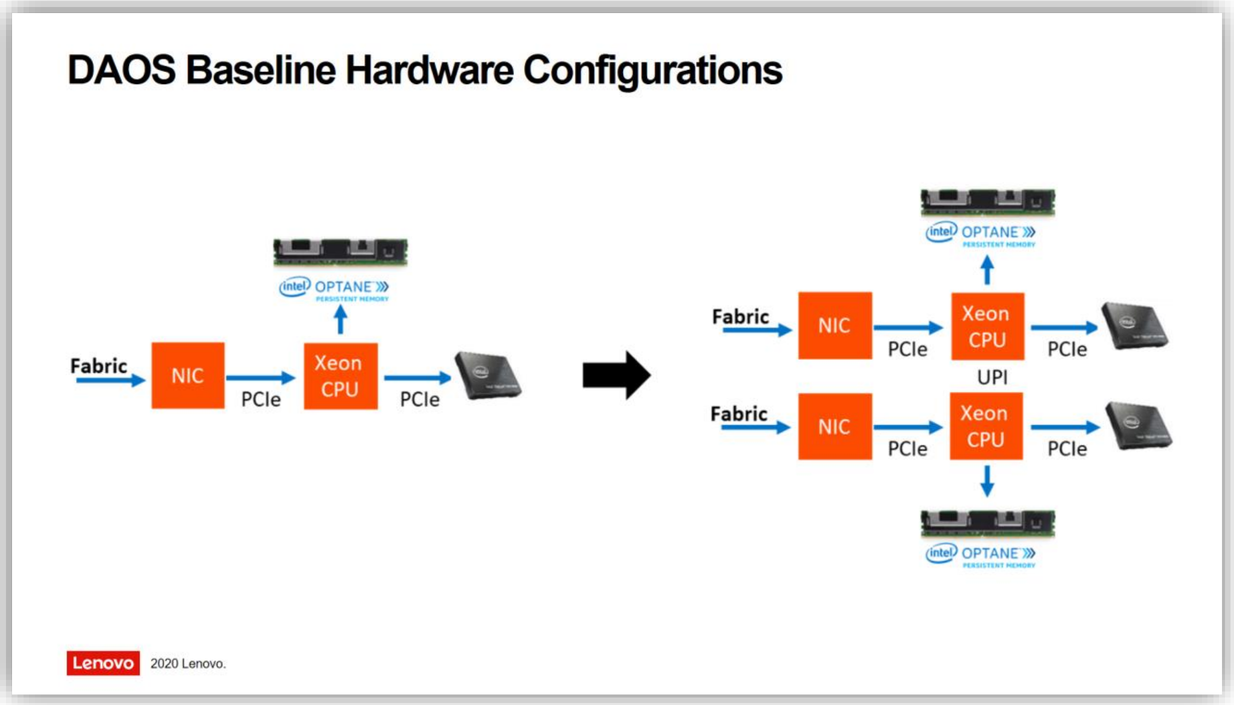
Figure 2: POSIX Filesystem Support in DAOS.

Figure 2 shows three different methods of performing POSIX I/O to the DAOS File System. The Linux FUSE mechanism can be used to perform a user space mount of a DAOS POSIX container on the compute nodes. DAOS ships with a `dfuse` daemon that provides this functionality. This is the easiest path to perform POSIX I/O, but also provides the lowest performance, as all application I/O requests have to go through the kernel and the `dfuse` daemon. For applications that are dynamically linked, DAOS also provides an I/O interception library (`libioil`) that can be used with the Linux LD_PRELOAD mechanism to intercept the POSIX read and write calls of the application (a `dfuse` mount is still needed for metadata traffic). This provides much better performance than just using `dfuse`. Finally, it is possible to modify the source code of the application and replace the POSIX I/O calls with the corresponding DFS I/O calls like `dfs_read` and `dfs_write`. This provides the highest performance for both data and metadata operations.

For HPC workloads, DAOS has been integrated with the MPI-IO and HDF5 middleware. For MPI-IO, a ROMIO driver for DAOS is available that uses the `libdfs` API, as described above. For HDF5 it is possible to either use the HDF5 MPI-IO interface with a DAOS-enabled MPI stack, or to use the DAOS VOL plugin that the HDF Group has developed. Applications that are using MPI-IO or HDF5 can therefore immediately benefit from DAOS without any modifications of the applications.

DAOS Server Architecture

The baseline hardware configuration for a DAOS storage server is an Intel Xeon CPU that is connected to SCM in the form of Intel Optane Persistent Memory, a PCIe network card for HPC fabric connectivity, and (optionally) PCIe attached NVMe SSDs for bulk storage. On a dual-socket Intel Xeon server, two copies of this baseline configuration can be served by running two instances of the `daos_engine` process that implements the DAOS *data plane*. This is shown schematically in Figure 3.



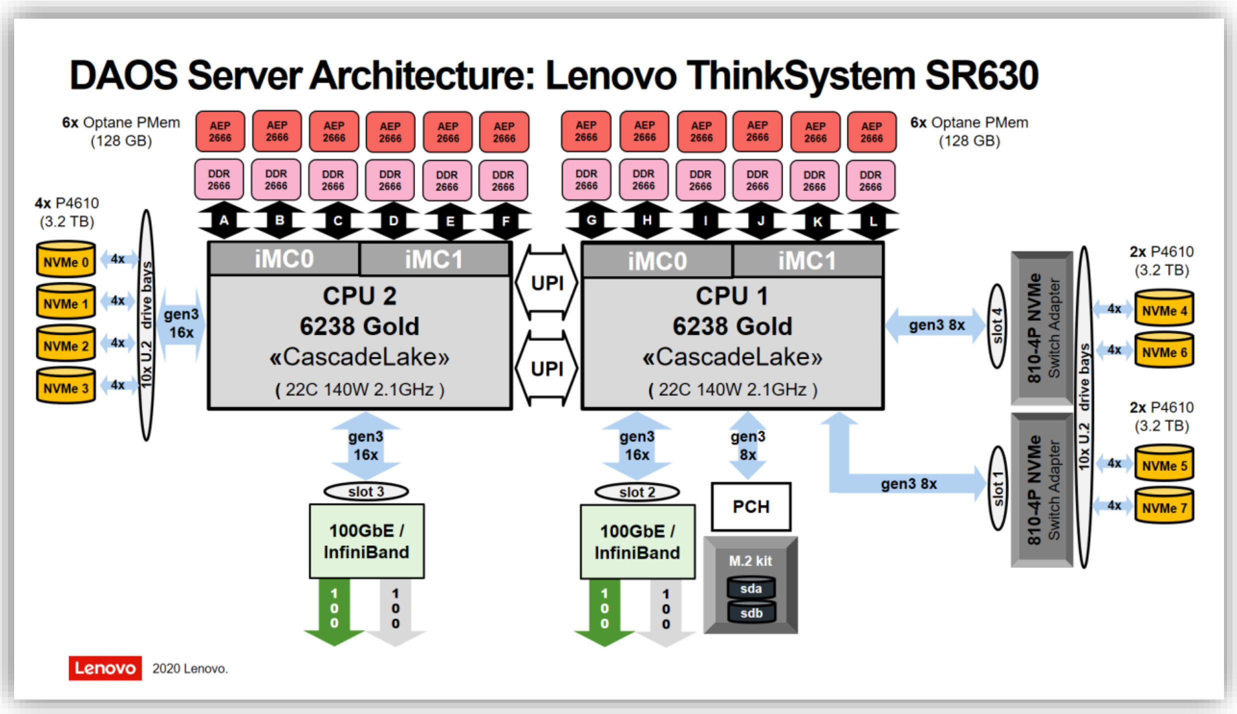
DAOS Baseline Hardware Configurations

Fabric
NIC
PCIe
Xeon CPU
PCIe
intel OPTANE PERSISTENT MEMORY
Fabric
NIC
PCIe
Xeon CPU
PCIe
UPI
Fabric
NIC
PCIe
Xeon CPU
PCIe
intel OPTANE PERSISTENT MEMORY
intel OPTANE PERSISTENT MEMORY

Lenovo 2020 Lenovo.

Figure 3: DAOS Baseline Hardware Configurations.

The Lenovo ThinkSystem SR630 is a 1U, 2-socket server that is ideally suited as a DAOS server. It provides a perfect balance of Intel Optane PMem, PCIe NVMe, and PCIe network cards attached to the two Intel Xeon CPUs. Figure 4 on page 7 shows the internal server architecture of the SR630 when configured as a DAOS server. On each CPU socket, the network bandwidth of 16 lanes of PCIe gen3 matches the storage bandwidth of four NVMe SSDs (each connected with 4 lanes of PCIe gen3).



DAOS Server Architecture: Lenovo ThinkSystem SR630

Figure 4: DAOS Server Architecture: Lenovo ThinkSystem SR630.

There are many design choices regarding the individual components of the SR630 as a DAOS server. The following subsections explain some of the considerations for configuring a DAOS server.

NVMe SSD Storage

The sizing of a DAOS system usually starts with the selection of the NVMe SSDs for bulk storage. While NVMe storage is optional in the DAOS architecture, and a DAOS server could be run with only SCM, in most environments NVMe storage will be required to provide the desired capacity.

The SR630 server chassis supports up to ten U.2 NVMe SSDs in 1U. But it is not possible to create a balanced setup with ten U.2 NVMe SSDs (where each socket would be connected to five SSDs). For this reason, we are limiting our DAOS configurations to eight U.2 NVMe SSDs per SR630 server (four on each socket). All NVMe SSDs in a DAOS server should be identical.

All of the U.2 NVMe SSDs that are supported in the SR630 should work with DAOS. We have validated the DAOS software stack with the Intel P4510, P4610, and P4800X series of U.2 NVMe SSDs, which provide a wide range of capacities in the Entry, Mainstream and Performance space.

Please refer to “Capacity Sizing” on page 12 and “Performance Sizing” on page 14 for more information to guide the selection of the best NVMe SSDs for a given set of customer requirements.

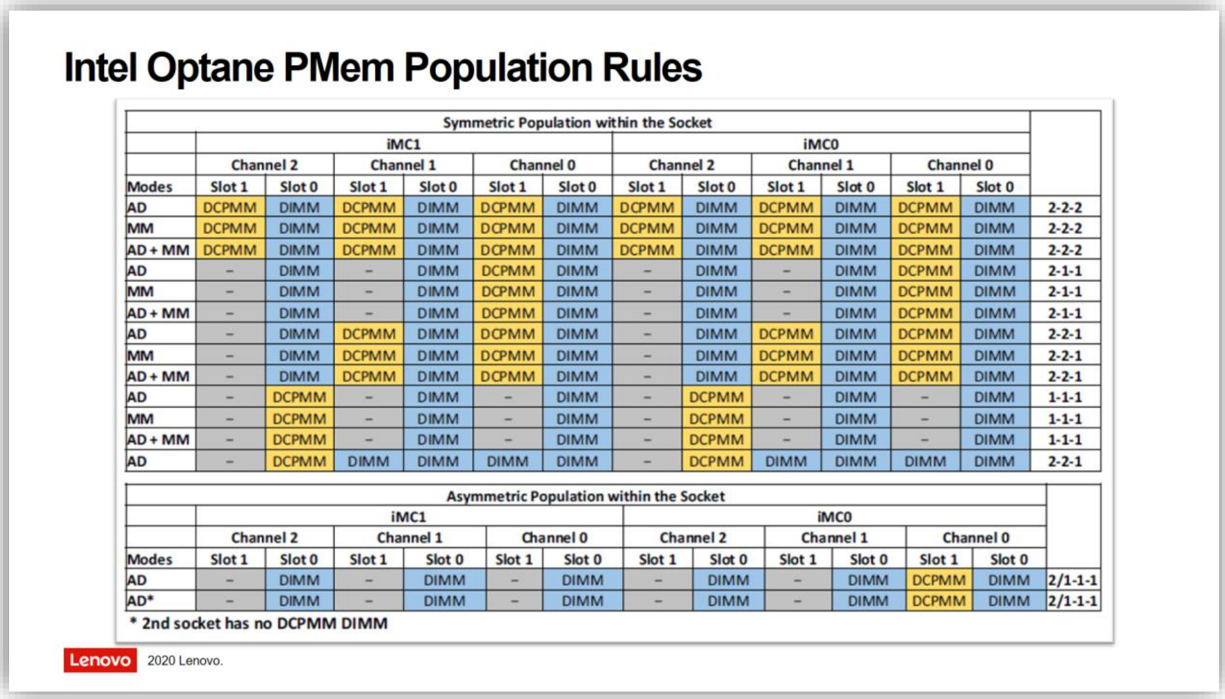
Table 1: Intel U.2 NVMe SSDs.

Description	Part number	Feature code
ThinkSystem U.2 Intel P4510 1.0TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A10202	B58F
ThinkSystem U.2 Intel P4510 2.0TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A10204	B58G
ThinkSystem U.2 Intel P4510 4.0TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A10205	B58H
ThinkSystem U.2 Intel P4510 8.0TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A08513	B58J
ThinkSystem U.2 Intel P4610 1.6TB Mainstream NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A13936	B589
ThinkSystem U.2 Intel P4610 3.2TB Mainstream NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A13937	B58A
ThinkSystem U.2 Intel P4610 6.4TB Mainstream NVMe PCIe 3.0 x4 Hot Swap SSD	4XB7A13938	B58B
ThinkSystem U.2 Intel Optane P4800X 750GB Performance NVMe PCIe 3.0 x4 Hot Swap SSD	7N47A00083	B2ZJ

Intel Optane Persistent Memory Modules

DAOS 1.1.3 requires persistent memory on each DAOS server, with a capacity of roughly 6% of the NVMe capacity of the server. This percentage may be reduced in a future DAOS release.

As shown in Figure 4 on page 7, a single socket of the second generation Intel Xeon SP CPUs ("Cascade Lake") has six memory channels on two memory controllers. With two sockets, the SR630 server has twelve memory channels, and supports two DIMMs per channel. Figure 5 contains the population rules for DDR4 DRAM memory and Intel Optane PMem modules (previously called "DCPMM") for a *single* socket. Note that it shows the population rules for AppDirect Mode (AD), Memory Mode (MM), and Mixed Mode (AD+MM). DAOS always uses AppDirect Mode. So only the "AD" rows are relevant here, and the "AD 2-2-2" configuration shown in the first row is the optimum configuration for DAOS.



Intel Optane PMem Population Rules

	Symmetric Population within the Socket												
	iMC1						iMC0						
	Channel 2		Channel 1		Channel 0		Channel 2		Channel 1		Channel 0		
Modes	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	
AD	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-2
MM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-2
AD + MM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-2
AD	–	DIMM	–	DIMM	DCPMM	DIMM	–	DIMM	–	DIMM	DCPMM	DIMM	2-1-1
MM	–	DIMM	–	DIMM	DCPMM	DIMM	–	DIMM	–	DIMM	DCPMM	DIMM	2-1-1
AD + MM	–	DIMM	–	DIMM	DCPMM	DIMM	–	DIMM	–	DIMM	DCPMM	DIMM	2-1-1
AD	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-1
MM	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-1
AD + MM	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	–	DIMM	DCPMM	DIMM	DCPMM	DIMM	2-2-1
AD	–	DCPMM	–	DIMM	–	DIMM	–	DCPMM	–	DIMM	–	DIMM	1-1-1
MM	–	DCPMM	–	DIMM	–	DIMM	–	DCPMM	–	DIMM	–	DIMM	1-1-1
AD + MM	–	DCPMM	–	DIMM	–	DIMM	–	DCPMM	–	DIMM	–	DIMM	1-1-1
AD	–	DCPMM	DIMM	DIMM	DIMM	DIMM	–	DCPMM	DIMM	DIMM	DIMM	DIMM	2-2-1

	Asymmetric Population within the Socket												
	iMC1						iMC0						
	Channel 2		Channel 1		Channel 0		Channel 2		Channel 1		Channel 0		
Modes	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	Slot 1	Slot 0	
AD	–	DIMM	–	DIMM	–	DIMM	–	DIMM	–	DIMM	DCPMM	DIMM	2/1-1-1
AD*	–	DIMM	–	DIMM	–	DIMM	–	DIMM	–	DIMM	DCPMM	DIMM	2/1-1-1

* 2nd socket has no DCPMM DIMM

Lenovo 2020 Lenovo.

Figure 5: Intel Optane PMem Population Rules.

Intel Optane PMem modules are available in 128 GB, 256 GB and 512 GB capacity. All PMem modules in a server must have the same capacity. For balanced performance, all memory controllers must be populated symmetrically with Intel Optane PMem modules. This implies that a DAOS server should contain two, four or six Intel Optane PMem modules per socket (four, eight or twelve modules per 2-socket server). In general, we advise to populate all twelve PMem modules to achieve the best performance. But in some scenarios, it may be adequate to only populate four or eight PMem modules.

Table 2: Intel Optane PMem Modules.

Description	Part number	Feature code
ThinkSystem 128GB TruDDR4 2666MHz (1.2V) Intel Optane Persistent Memory	4ZC7A15110	B4LV
ThinkSystem 256GB TruDDR4 2666MHz (1.2V) Intel Optane Persistent Memory	4ZC7A15111	B4LW
ThinkSystem 512GB TruDDR4 2666MHz (1.2V) Intel Optane Persistent Memory	4ZC7A15112	B4LX

Please refer to “Capacity Sizing” on page 12 and “Performance Sizing” on page 14 for more information on how the choice of NVMe SSDs impacts the Intel Optane PMem configuration of a DAOS server.

Intel Xeon Processors

The following considerations are guiding the selection of the Intel processor SKU for the SR630 DAOS Server:

- Intel Optane DC Persistent Memory is supported on all second generation Intel Xeon SP Gold and Platinum SKUs (as well as the 4215 Silver SKU, but this processor is not optimal for DAOS). Due to thermal limitations, the 1U SR630 server does not support Intel Optane PMem together with CPUs with a TDP of 205 Watt or higher. The Gold 6240Y, 6244, 6246, or 6252N SKUs are also not supported in the SR630 together with Intel Optane PMem.
- Performance benchmarking has shown that twelve Intel Optane PMem modules in a 2-socket server provide the highest performance when the DAOS I/O servers use about 16 *storage targets* per socket. Each storage target is a software instance that requires one physical core. With one core for general operating system tasks, and up to four cores to assist with erasure coding and other computationally intensive tasks, the general advice for DAOS servers is to use CPUs with at least 22 physical cores per socket.
- The standard Intel Xeon processor SKUs support a maximum of 1 TiB of memory per socket. This is sufficient for twelve 128 GB Intel Optane PMem modules. For the larger capacities an “L” processor SKU is required that supports up to 4.5 TiB of memory per socket.
- For some processors, an “R” refresh SKU is available that provides more performance (more cores and/or higher frequency) compared to the standard SKU. Those processors consume more power than the corresponding standard SKU, and they do not include versions that support more than 1 TiB of memory per socket.

The 6238 processor that is shown in Figure 4 on page 7 has 22 cores and a TDP of 140 Watt. It is available both as a standard SKU, and as an “L” SKU for configurations that require more than 1 TiB of combined DRAM and PMem capacity. It is also available as an “R” SKU with 28 cores and a slightly higher frequency, at a TDP of 165 Watt. Considering the overall price/performance of the server, the 6238 is our default recommendation for a DAOS server. Other processor choices are shown in Table 3 on page 10.

Table 3: Second Generation Intel Xeon Gold Processors.

CPU Model	Cores	Core speed (Base)	Core speed (Max Turbo)	TDP Power	Max memory per socket	Cache Size	Max DDR Speed (*)	UPI speed
6230R	26	2.1 GHz	4.0 GHz	150 W	1 TiB	35.75 MiB	2933 MHz	10.4 GT/s
6238	22	2.1 GHz	3.7 GHz	140 W	1 TiB	30.25 MiB	2933 MHz	10.4 GT/s
6238L	22	2.1 GHz	3.7 GHz	140 W	4.5 TiB	30.25 MiB	2933 MHz	10.4 GT/s
6238R	28	2.2 GHz	4.0 GHz	165 W	1 TiB	38.5 MiB	2933 MHz	10.4 GT/s
6240R	24	2.4 GHz	4.0 GHz	165 W	1 TiB	35.75 MiB	2933 MHz	10.4 GT/s
6252	24	2.1 GHz	3.7 GHz	150 W	1 TiB	35.75 MiB	2933 MHz	10.4 GT/s

(*) Note that the Intel Optane PMem Modules operate at 2666 MHz. So even if DRAM with a maximum DDR4 speed of 2933 MHz is installed, the memory bus will operate at 2666 MHz.

High Performance Network Adapters

To ensure a balanced network bandwidth that matches the eight NVMe SSDs with a total of 32 PCIe lanes, two 16-lane Mellanox ConnectX-5 or ConnectX-6 VPI cards provide connectivity to the HPC fabric (one card on each CPU socket). The recommended HPC fabric for DAOS is InfiniBand, with the `libfabrics` ofi+verbs provider. ConnectX-5 cards support EDR, and ConnectX-6 cards support HDR100. Since all of these cards are VPI cards, they can also be used in 100 Gbps Ethernet mode (together with the `libfabrics` ofi+tcp or ofi+sockets provider).

Both single-port and dual-port Mellanox adapters are available. In general, a single-port adapter is sufficient as the card's 16-lane PCIe gen3 connectivity will only be able to saturate one 100 Gbps link. In some environments it may still be beneficial to use dual-port cards, for example to connect the server to multiple fabrics or to provide more bidirectional bandwidth in situations where a lot of data movement will occur (the dual-port Mellanox cards can *send* at 100 Gbps speed on one port, and simultaneously *receive* at 100 Gbps speed on the other port).

Table 4: High-Performance Network Adapters.

Description	Part number	Feature code
Mellanox ConnectX-5 1x100GbE / EDR IB QSFP28 VPI Adapter (*)	4C57A08979	B0RL
Mellanox ConnectX-5 2x100GbE / EDR IB QSFP28 VPI Adapter (*)	4C57A08980	B0RM
Mellanox ConnectX-6 HDR100 QSFP56 1-port PCIe InfiniBand Adapter	4C57A14177	B4R9
Mellanox ConnectX-6 HDR100 QSFP56 2-port PCIe InfiniBand Adapter	4C57A14178	B4RA

(*) Note that Mellanox ConnectX-5 cards are only available and supported in the SR630 server as part of a Lenovo LeSI cluster.

Please refer to the Mellanox product briefs for more information on the ConnectX-5 and ConnectX-6 VPI adapters:

- https://www.mellanox.com/files/doc-2020/pb-connectx-5-vpi-card.pdf
- https://www.mellanox.com/files/doc-2020/pb-connectx-6-vpi-card.pdf

Volatile DRAM Memory

To achieve best performance, it is recommended to populate one TruDDR4 DRAM module on each of the twelve memory channels of the server. All memory modules should have the same type and capacity. With one DDR4 DIMM per channel, using dual-rank DIMMs has a performance benefit over single-rank DIMMs. Either 2666 MHz or 2933 MHz RDIMMs can be used, but the memory bus will always operate at 2666 MHz when Intel Optane PMem modules are present. We generally recommend to use the more recent 2933 MHz DIMMs. In most environments 16 GiB DIMMs should be sufficient, providing 192 GiB per DAOS server. When needed, twelve 32 GiB DIMMs can provide 384 GiB per server.

Table 5: DDR4 Memory DIMMs.

Description	Part number	Feature code
ThinkSystem 16GB TruDDR4 2933MHz (2Rx8 1.2V) RDIMM	4ZC7A08708	B4H2
ThinkSystem 32GB TruDDR4 2933MHz (2Rx8 1.2V) RDIMM	4ZC7A08709	B4H3

Boot Devices

Two M.2 SATA SSDs are used in a hardware RAID1 configuration to hold the operating system (Lenovo validated DAOS with CentOS, but RHEL or SLES are also possible). The mirroring is performed through the “ThinkSystem M.2 with Mirroring Enablement Kit”, as shown in Figure 4 on page 7.

While these M.2 SSDs are not hot-swappable, this is not a significant disadvantage: DAOS server clusters are scale-out solutions that need to protect against single-server failures anyway (through replication or erasure coding). So in the rare case of an M.2 boot device failure, the affected server can be drained and shut down in a planned maintenance activity to replace the failed M.2 card. The benefit of using M.2 boot devices is that no PCIe slot is needed for a SAS/SATA RAID card – this PCIe slot is instead used to connect two of the eight NVMe SSDs to provide optimal NVMe performance.

Capacity Sizing

Sizing the capacity of a DAOS solution is a two-step process. The configuration of a single DAOS server needs to be determined, and then the usable capacity of a scale-out cluster of multiple DAOS servers needs to be planned.

Single Server Capacity Sizing

Intel recommends to use 6% of a DAOS 1.1.3 server's NVMe SSD capacity as the Intel Optane PMem capacity. This Persistent Memory is used for DAOS-internal metadata, and to cache application I/O that is smaller than 4 kiB. The required percentage may be reduced in a future DAOS release. For each of the Intel U.2 NVMe SSDs listed in Table 1 on page 8, this capacity ratio implies a certain population with Intel Optane PMem modules, which in turn determines if a standard CPU SKU or an "L" SKU (that supports more than 1 TiB of memory per socket) is needed. Table 6 shows the resulting combinations, and points out where some of the combinations are deviating from the optimal design.

Table 6: Single SR630 DAOS Server Capacity Sizing Options.

CPUs		U.2 NVMe SSDs			Optane PMem		DDR4 DRAM		Server Total Raw Capacity			
Qty	SKU	Qty	Series	TB	Qty	GB	Qty	GiB	NVMe TiB	PMem TiB	%PMem	DRAM GiB
2	6238	8	P4510	1	4	128	12	16	7,3	0,47	6,4%	192
2	6238	8	P4510	2	8	128	12	16	14,5	0,93	6,4%	192
2	6238	8	P4510	4	12	128	12	16	29,1	1,40	4,8%	192
2	6238L	8	P4510	4	8	256	12	16	29,1	1,86	6,4%	192
2	6238L	8	P4510	8	12	256	12	16	58,2	2,79	4,8%	192
2	6238L	8	P4510	8	8	512	12	16	58,2	3,72	6,4%	192
2	6238	8	P4610	1,6	8	128	12	16	11,6	0,93	8,0%	192
2	6238	8	P4610	1,6	12	128	12	16	11,6	1,40	12,0%	192
2	**6238**	**8**	**P4610**	**3,2**	**12**	**128**	**12**	**16**	**23,3**	**1,40**	**6,0%**	**192**
2	6238L	8	P4610	6,4	12	256	12	16	46,5	2,79	6,0%	192
2	6238	8	P4800X	0,75	4	128	12	16	5,5	0,47	8,5%	192
2	6238	8	P4800X	0,75	8	128	12	16	5,5	0,93	17,1%	192
2	6238	8	P4800X	0,75	12	128	12	16	5,5	1,40	25,6%	192

The following non-optimal component choices occur in Table 6:

- For best performance, we recommend to always populate all 12 PMem slots. But for small NVMe capacities, this will not be optimal in terms of price/performance. Configurations with only 8 or 4 PMem modules are marked in orange or red.
- PMem module pricing does not scale linearly with their capacity: Larger modules have a higher price per GB than smaller modules. The best price/performance is achieved with 128 GB modules. Configurations with the more expensive 256 GB or 512 GB modules are marked in orange or red.
- The "L" processor SKUs are more expensive than standard SKUs. Configurations that require an "L" SKU are therefore marked in red.
- Some configurations deviate markedly from the 6% PMem ratio. These are marked in orange. A slightly smaller ratio is not a big concern, especially since the DAOS development team aims at reducing the required percentage of PMem in a future DAOS software release. Larger ratios are not a concern technically, but obviously incur a higher cost than what would be necessary.

Combining all these factors implies that the current sweet-spot configuration for P4610 “Mainstream” NVMe SSDs (3 DWPD) is eight 3.2 TB SSDs and twelve 128 GB PMem modules. This provides a raw capacity of 23.3 TiB + 1.4 TiB per server. For P4510 “Entry” NVMe SSDs, the sweet spot is eight 4 TB SSDs and twelve 128 GB PMem modules, with a raw capacity of 29.1 TiB + 1.4 TiB per server.

Scale-Out Capacity Sizing

Given the raw NVMe capacity of an individual DAOS server, the desired usable capacity of a DAOS storage solution can be determined by configuring an appropriate number of DAOS servers in a “scale-out” DAOS solution.

An important factor that needs to be considered in addition to the usable capacity target is the desired level of data protection. This will also determine the minimum number of servers in a DAOS system:

- Without data protection, a DAOS system may consist of a single DAOS server. And for “ephemeral” DAOS storage that is only used as scratch space for the duration of a single job, it may be perfectly fine to configure the DAOS container with no data protection at all. In this case the usable capacity is the same as the raw capacity (with some minor overhead), and calculating the required number of servers to reach the targeted usable capacity is trivial.
- DAOS 1.1.3 supports N-way replication, and the usable capacity for containers using N-way replication is 1/N of the raw capacity. For example, with 3-way replication the usable capacity is 33% of the raw capacity. Since N-way replication protects against the simultaneous failure of (N-1) servers, it is prudent to configure sufficiently many servers so in a failure case there are enough “surviving” servers for the re-construction of all N replica. As a general rule of thumb, it makes sense to configure at least 2*N servers if N-way replication is used.
- Erasure Coding will be supported in a future DAOS release (see “DAOS Roadmap” on page 21). This will improve the ratio of usable capacity to raw capacity, because less space is needed for “parity” information than for N-way replication. For example, with 8+2*P* erasure coding the usable capacity is 8/(8+2)=80% of the raw capacity. As in the case of N-way replication, sufficiently many servers should be configured so in a failure case all EC stripes can be reconstructed on the “surviving” servers. To make effective use of the distributed nature of the DAOS erasure coding implementation, it is recommended to configure at least 2*(N+M) servers if (N+M*P*) erasure coding is used.

It is possible to use different data protection schemes within a single DAOS server cluster. The data protection scheme is a property of a DAOS *container*, and different containers (in the same DAOS *pool* or in different DAOS *pools*) can use different levels of data protection.

Performance Sizing

Similar to capacity sizing, performance sizing of a DAOS storage solution involves the sizing of an individual DAOS server, combined with a “scale-out” sizing to reach the desired target performance.

In contrast to capacity sizing, the “scale-out” performance sizing is much more complex than the simple “rule of three” calculations that are needed for capacity sizing. Depending on the specific performance metric of interest, performance scaling may be easy or difficult to project. If in doubt, it is always advisable to request a storage performance benchmark or proof-of-concept activity through your Lenovo account team.

Single Server Hardware Performance

The hardware performance of a single DAOS server depends on its network performance and the performance of its internal SCM and NVMe storage components.

HPC Fabric Performance

A Single EDR InfiniBand port of a Mellanox ConnectX-5 adapter in a PCIe gen3 x16 slot has a maximum “wire speed” bandwidth of 11.64 GiB/s (four lanes operating at a bit rate of 25 Gb/s). Using Mellanox ConnectX-6 HDR100 achieves the same bandwidth with two lanes operating at 50 Gb/s. So for an SR630 server with two EDR ports or two HDR100 ports (in two PCIe gen3 x16 slots), the peak aggregate InfiniBand bandwidth is 23.28 GiB/s. This is the server’s theoretical “not to exceed” network bandwidth, which will rarely be achievable with real workloads.

Storage Device Performance

The peak storage bandwidth of an individual DAOS server is determined to a large degree by the performance of the NVMe SSDs that are installed in the server. This depends on the NVMe SSD series, but also on the chosen capacity of the SSDs. Table 7 summarizes the properties of the Intel PCIe gen3 NVMe SSDs that are currently supported in the SR630 (based on the Intel specifications at https://ark.intel.com/content/www/us/en/ark.html).

Table 7: Intel U.2 NVMe SSD specifications.

Drive Series	Storage Technology	Capacity	Sequential Read	Sequential Write	Random Read	Random Write	Read Latency	Write Latency	Active Power	Idle Power	Write Endurance	Write Endurance
		[GB]	[MB/s]	[MB/s]	[k IOPS]	[k IOPS]	[usec]	[usec]	[W]	[W]	[PBW]	[DWPD]
Entry (~1 DWPD)												
P4510	64layer 3D TLC NAND	1000	2850	1100	465,0	70,0	77	18	10,0	5,0	1,92	1,05
P4510	64layer 3D TLC NAND	2000	3200	2000	637,0	81,5	77	18	12,0	5,0	2,61	0,72
P4510	64layer 3D TLC NAND	4000	3000	2900	625,5	113,5	77	18	14,0	5,0	6,30	0,86
P4510	64layer 3D TLC NAND	8000	3200	3000	620,0	139,5	77	18	16,0	5,0	13,88	0,95
Mainstream (~3-5 DWD)												
P4610	64layer 3D TLC NAND	1600	3200	2100	620,0	200,0	77	18	13,3	5,0	12,25	4,20
P4610	64layer 3D TLC NAND	3200	3200	3000	640,0	200,0	77	18	13,8	5,0	21,85	3,74
P4610	64layer 3D TLC NAND	6400	3000	2900	640,0	220,0	77	18	14,6	5,0	36,54	3,13
Performance (~30 DWPD)												
P4800X	3D Xpoint	375	2400	2000	550,0	550,0	10	10	18,0	5,0	20,50	29,95
P4800X	3D Xpoint	750	2500	2000	550,0	550,0	10	10	18,0	6,0	41,00	29,95

Other suppliers' NVMe SSDs that are supported in the SR630 should also work with DAOS, but have not been validated. Some Intel PCIe gen4 NVMe SSDs are also supported in the SR630, but will operate at PCIe gen3 speed in the current generation of the SR630 server. Those are also not considered here, as their deployment in a PCIe gen3 platform is usually not cost effective.

Table 7 on page 14 shows that both the P4610 3.2 TB NVMe SSD and the P4510 4 TB NVMe SSD have very similar *sequential read* and write bandwidth, whereas the smaller capacities have significantly lower *sequential write* bandwidth than the larger capacities. The main technical difference between the P4610 3.2 TB NVMe SSD and the P4510 4 TB NVMe SSD are their *random* write performance and their write *endurance* (3 DWPD compared to 1 DWPD).

Small (<4kiB) I/O requests will be serviced by the Intel Optane PMem Storage Class Memory, before eventually being aggregated and de-staged to NVMe bulk storage. Intel has not published performance specifications for the three capacities of Intel Optane PMem modules. In general, they are expected to show very similar performance for all three capacities. The main PMem performance difference for the different configuration options will therefore result from the *number* of Intel Optane PMem modules that are installed in the server. Always populating twelve PMem modules will provide the widest interleaving, and the best aggregate PMem performance.

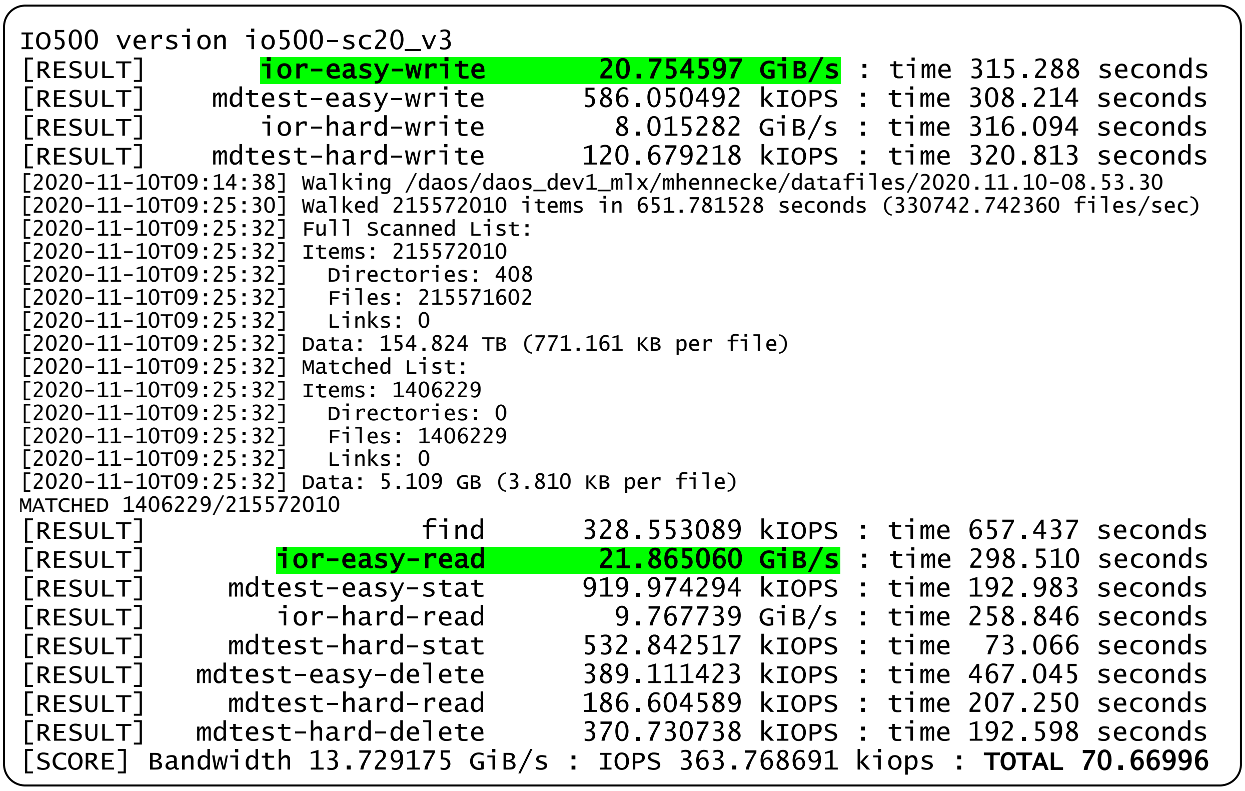
Single Server IO500 Performance

Translating the storage devices' raw performance characteristics into application performance does depend on the type of I/O workload, as well as the software overhead of the storage solution. Workloads dominated by sequential I/O can typically achieve the aggregate device bandwidth. Small, random and/or unaligned I/O workloads put much more stress onto the storage system, and details of the software implementation often play a much bigger role than raw device performance.

The IO500 storage benchmark suite (https://www.vi4io.org/io500/about/start) contains twelve different benchmarks that measure large sequential I/O as well as small random/strided I/O, various metadata workloads, and a "find" directory traversal. The overall IO500 "SCORE" that is calculated as the geometric mean of all twelve individual measurements may or may not be particularly useful as a measurement of overall system balance. But it is certainly instructive to compare the twelve individual IO500 benchmarks for different parallel file systems and other HPC storage solutions.

The DAOS development team has enabled the `IOR` and `mdtest` benchmarks that are used in the IO500 suite for the DAOS "DFS" API. Those changes have been contributed into the main `IOR` and `mdtest` GIT repository at https://github.com/hpc/ior. The DAOS development team has also expanded the parallel find utility that is part of the `mpifileutils` utilities to work with the DAOS "DFS" API. Those extensions can be found at https://github.com/mchaarawi/mpifileutils. The IO500 rules allow to use an optimized "find" routine for the IO500 "find" test, and DAOS performs best when used with API=DFS together with the DFS-enabled parallel find from `mpifileutils`.

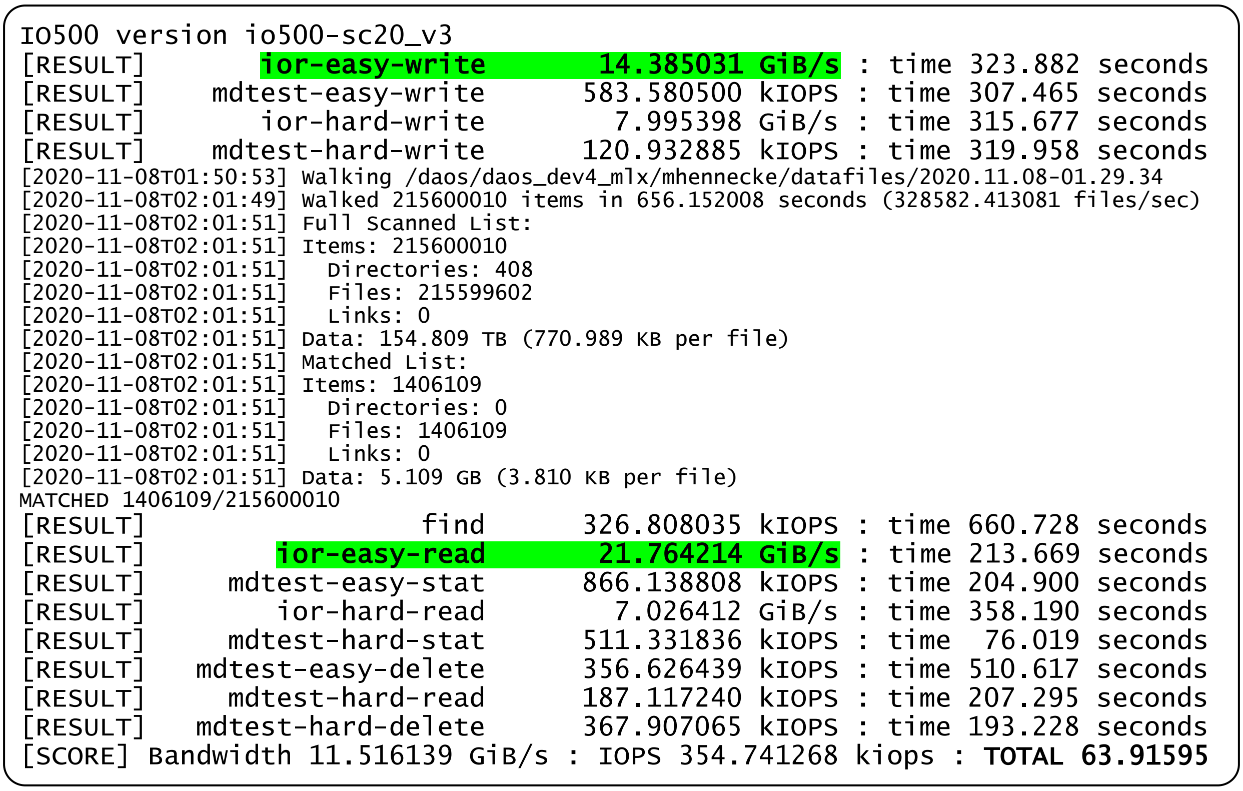
Figure 6 contains the results of a valid IO500 run for a single DAOS server with eight Intel P4610 **3.2 TB** NVMe SSDs and twelve 128 GB Intel Optane PMem modules, as shown in Figure 4 on page 7. No data protection (N-way replication or erasure coding) has been used for this single-server setup. This benchmark run has been performed on ten client nodes with a single EDR connection each. So, it would qualify for the IO500's "10 Node Challenge" list with an overall IO500 score of 70.6. The two `ior-easy` bandwidth results are highlighted in green. They can be used as a sanity check to validate that the aggregate NVMe device bandwidth of the eight NVMe SSDs can be achieved on the application layer for large sequential I/O. This is generally true for most HPC storage solutions; the "hard" benchmarks in the IO500 are a much better test of the capabilities of DAOS when compared to other HPC storage solutions.



```
IO500 version io500-sc20_v3
[RESULT]        ior-easy-write        20.754597 GiB/s : time 315.288 seconds
[RESULT]     mdtest-easy-write       586.050492 kIOPS : time 308.214 seconds
[RESULT]        ior-hard-write         8.015282 GiB/s : time 316.094 seconds
[RESULT]     mdtest-hard-write       120.679218 kIOPS : time 320.813 seconds
[2020-11-10T09:14:38] Walking /daos/daos_dev1_mlx/mhennecke/datafiles/2020.11.10-08.53.30
[2020-11-10T09:25:30] Walked 215572010 items in 651.781528 seconds (330742.742360 files/sec)
[2020-11-10T09:25:32] Full Scanned List:
[2020-11-10T09:25:32] Items: 215572010
[2020-11-10T09:25:32]   Directories: 408
[2020-11-10T09:25:32]   Files: 215571602
[2020-11-10T09:25:32]   Links: 0
[2020-11-10T09:25:32] Data: 154.824 TB (771.161 KB per file)
[2020-11-10T09:25:32] Matched List:
[2020-11-10T09:25:32] Items: 1406229
[2020-11-10T09:25:32]   Directories: 0
[2020-11-10T09:25:32]   Files: 1406229
[2020-11-10T09:25:32]   Links: 0
[2020-11-10T09:25:32] Data: 5.109 GB (3.810 KB per file)
MATCHED 1406229/215572010
[RESULT]                  find       328.553089 kIOPS : time 657.437 seconds
[RESULT]         ior-easy-read        21.865060 GiB/s : time 298.510 seconds
[RESULT]      mdtest-easy-stat       919.974294 kIOPS : time 192.983 seconds
[RESULT]         ior-hard-read         9.767739 GiB/s : time 258.846 seconds
[RESULT]      mdtest-hard-stat       532.842517 kIOPS : time  73.066 seconds
[RESULT]    mdtest-easy-delete       389.111423 kIOPS : time 467.045 seconds
[RESULT]      mdtest-hard-read       186.604589 kIOPS : time 207.250 seconds
[RESULT]    mdtest-hard-delete       370.730738 kIOPS : time 192.598 seconds
[SCORE] Bandwidth 13.729175 GiB/s : IOPS 363.768691 kiops : TOTAL 70.66996
```

*Figure 6: Single DAOS Server IO500 10-node results with 8x P4610 **3.2 TB**.*

Executing the same benchmark on a single DAOS server that is populated with eight Intel P4610 **1.6 TB** NVMe SSDs results in the output shown in Figure 7 on page 17. The `ior-easy` sequential bandwidth numbers highlighted in green demonstrate that the lower individual device *write* bandwidth specs of the 1.6 TB NVMe SSDs result in a proportionally lower `ior-easy-write` bandwidth. Most of the other IO500 metrics do not show a large difference to the 3.2 TB NVMe SSDs. This is expected, as most other performance metrics of the individual devices are very similar.



```
IO500 version io500-sc20_v3
[RESULT]       ior-easy-write       14.385031 GiB/s : time 323.882 seconds
[RESULT]    mdtest-easy-write      583.580500 kIOPS : time 307.465 seconds
[RESULT]       ior-hard-write        7.995398 GiB/s : time 315.677 seconds
[RESULT]    mdtest-hard-write      120.932885 kIOPS : time 319.958 seconds
[2020-11-08T01:50:53] Walking /daos/daos_dev4_mlx/mhennecke/datafiles/2020.11.08-01.29.34
[2020-11-08T02:01:49] Walked 215600010 items in 656.152008 seconds (328582.413081 files/sec)
[2020-11-08T02:01:51] Full Scanned List:
[2020-11-08T02:01:51] Items: 215600010
[2020-11-08T02:01:51]   Directories: 408
[2020-11-08T02:01:51]   Files: 215599602
[2020-11-08T02:01:51]   Links: 0
[2020-11-08T02:01:51] Data: 154.809 TB (770.989 KB per file)
[2020-11-08T02:01:51] Matched List:
[2020-11-08T02:01:51] Items: 1406109
[2020-11-08T02:01:51]   Directories: 0
[2020-11-08T02:01:51]   Files: 1406109
[2020-11-08T02:01:51]   Links: 0
[2020-11-08T02:01:51] Data: 5.109 GB (3.810 KB per file)
MATCHED 1406109/215600010
[RESULT]                 find      326.808035 kIOPS : time 660.728 seconds
[RESULT]        ior-easy-read       21.764214 GiB/s : time 213.669 seconds
[RESULT]     mdtest-easy-stat      866.138808 kIOPS : time 204.900 seconds
[RESULT]        ior-hard-read        7.026412 GiB/s : time 358.190 seconds
[RESULT]     mdtest-hard-stat      511.331836 kIOPS : time  76.019 seconds
[RESULT]   mdtest-easy-delete      356.626439 kIOPS : time 510.617 seconds
[RESULT]     mdtest-hard-read      187.117240 kIOPS : time 207.295 seconds
[RESULT]   mdtest-hard-delete      367.907065 kIOPS : time 193.228 seconds
[SCORE] Bandwidth 11.516139 GiB/s : IOPS 354.741268 kiops : TOTAL 63.91595
```

Figure 7: Single DAOS Server IO500 10-node results with 8x P4610 ***1.6 TB****.*

It should be noted that one such IO500 run performs a *single iteration* of each of the twelve tests. In the examples above, the complete IO500 run has a wall clock run time of about one hour. Because only a single iteration of each test is measured, run-to-run variations of the same IO500 suite on the same client and server setup should be expected – especially for the "hard" test cases. The result shown here should not be interpreted as a performance commitment, but only as a rough guideline to indicate the capabilities of a single DAOS server.

Scale-Out Performance Sizing

When scaling out the performance of an individual DAOS server to a larger storage cluster, the exact characteristics of the I/O workloads play a large role. Some workloads (like the `ior-easy` tests) are known to scale almost linearly with the number of DAOS storage servers. Other workloads may exhibit very different scaling behavior.

In any case, the desired level of data protection will have an impact on the *write* performance of the scale-out solution. For example, if 3-way replication is used then the write bandwidth that is achievable zfrom the user's perspective will be at best 1/3 of the raw device performance, as each chunk of data will need to be written to three different storage devices. For 8+2P erasure coding, the achievable write bandwidth is at best 80% of the raw device bandwidth.

The *read* bandwidth is not affected by the data protection scheme, as it is sufficient to read from any one of the copies of the data.

DAOS Software Environment

This section describes the DAOS software environment including Lenovo LeSI, DAOS software deployment, the DAOS software roadmap, as well as DAOS services and support.

Lenovo Scalable Infrastructure (LeSI)

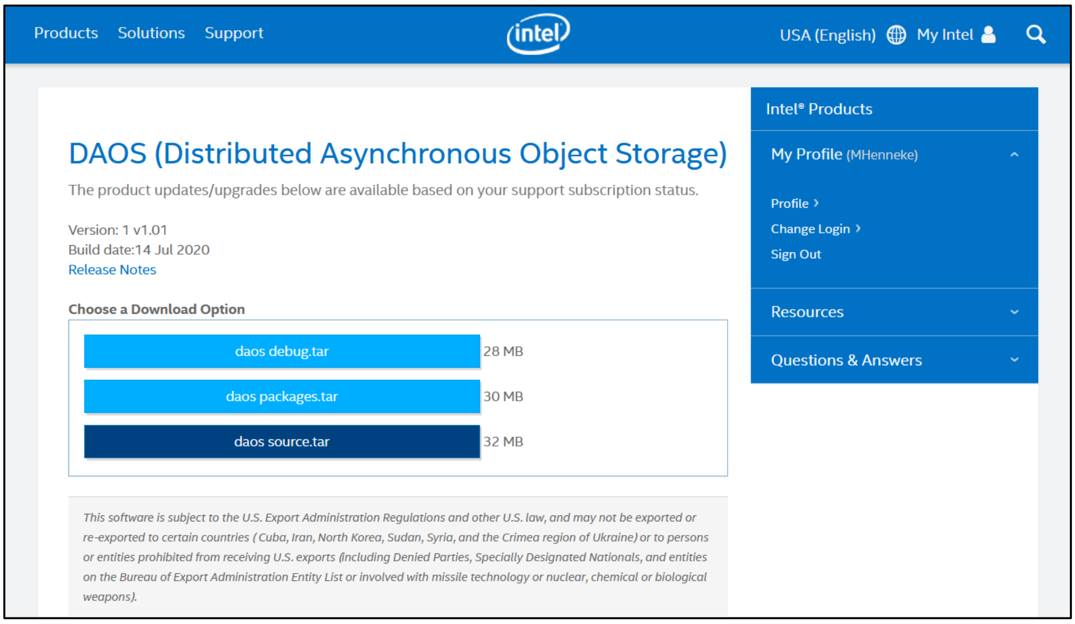
Lenovo recommends to implement DAOS on top of a Lenovo Scalable Infrastructure (LeSI) cluster. LeSI is Lenovo's framework for designing, manufacturing, integrating and delivering data center solutions, with a focus on High Performance Computing (HPC), Technical Computing, and Artificial Intelligence (AI) environments. Lenovo Scalable Infrastructure provides *Best Recipe* guides to warrant interoperability of hardware, software and firmware among a variety of Lenovo and third-party components. For DAOS in particular, running on an LeSI Best Recipe level ensures that the specific combination of operating system, Mellanox OFED stack, and device firmware for the Intel Optane PMem modules and NVMe SSDs has been integration-tested by the Lenovo LeSI team. Please refer to the LeSI Product Guide at https://lenovopress.com/lp0900-lenovo-scalable-infrastructure-lesi-solutions and the LeSI Best Recipes at https://support.lenovo.com/us/en/solutions/HT510136 for details.

DAOS Deployment

DAOS itself is an open source software stack, available at https://github.com/daos-stack/daos. To get the latest development snapshot of DAOS, it is possible to clone this GIT repository and build DAOS from source. The build process will automatically pull in all the prerequisite software frameworks, including PMDK, SPDK, libfabric, Argobots, ISA-L, etc. Building DAOS from source is explained in the DAOS Administration Guide at https://daos-stack.github.io/admin/installation/.

A more convenient method to install DAOS is to use the RPM packages that Intel is providing for their main releases at https://registrationcenter.intel.com/forms/?productid=3412. This requires a "My Intel" account to request access to DAOS, sign in to the Intel software site, and download the tarballs with the DAOS (and pre-requisite) RPMs, as shown in Figure 8 on page 19. With the RPM packages, installing DAOS is as simple as running "`yum install daos-server`", or "`yum install daos-client`" on the client nodes. The DAOS RPM installation is also described in the online DAOS Administration Guide. One useful pre-installation step is to pre-assign user IDs and group IDs for the Linux users that the DAOS RPM packages will create, to ensure that those user IDs conform to local site policies.

Note that because of the large amount of new development that is still ongoing, the DAOS development team has decided to not provide software interoperability between the DAOS 1.0 release and the next release (currently, the DAOS 1.1.3 "pre-release"). In order to upgrade a DAOS system from DAOS 1.0 to a newer release, the DAOS storage that has been formatted with DAOS 1.0 should be de-provisioned, and the DAOS 1.0 software has to be un-installed before installing the new release. Starting with the DAOS 1.2 release, there will be at least "N-1" interoperability of the DAOS releases.



DAOS (Distributed Asynchronous Object Storage)

The product updates/upgrades below are available based on your support subscription status.

Version: 1 v1.01
Build date:14 Jul 2020
Release Notes

Choose a Download Option

daos debug.tar 28 MB
daos packages.tar 30 MB
daos source.tar 32 MB

This software is subject to the U.S. Export Administration Regulations and other U.S. law, and may not be exported or re-exported to certain countries (Cuba, Iran, North Korea, Sudan, Syria, and the Crimea region of Ukraine) or to persons or entities prohibited from receiving U.S. exports (including Denied Parties, Specially Designated Nationals, and entities on the Bureau of Export Administration Entity List or involved with missile technology or nuclear, chemical or biological weapons).

Figure 8: Intel DAOS Software Download (RPM packages).

Deploying DAOS after the software has been installed consists of the following main steps:

- Creating certificates that will be used to authenticate and authorize the various system processes (and storage administrators);
- Creating YAML configuration files for the `daos_agent` and `daos_server` daemons, as well as for the administrative commands of the DAOS *control plane;*
- Starting the DAOS daemons (for example, using the DAOS `systemd` integration);
- Discovering, preparing and formatting the storage hardware.

The DAOS deployment process is described at https://daos-stack.github.io/admin/deployment/.

After the DAOS server cluster has been deployed, the storage administrator can create DAOS *pools* with the `dmg` command. Pool management is documented in the DAOS Administration Guide at https://daos-stack.github.io/admin/pool_operations/.

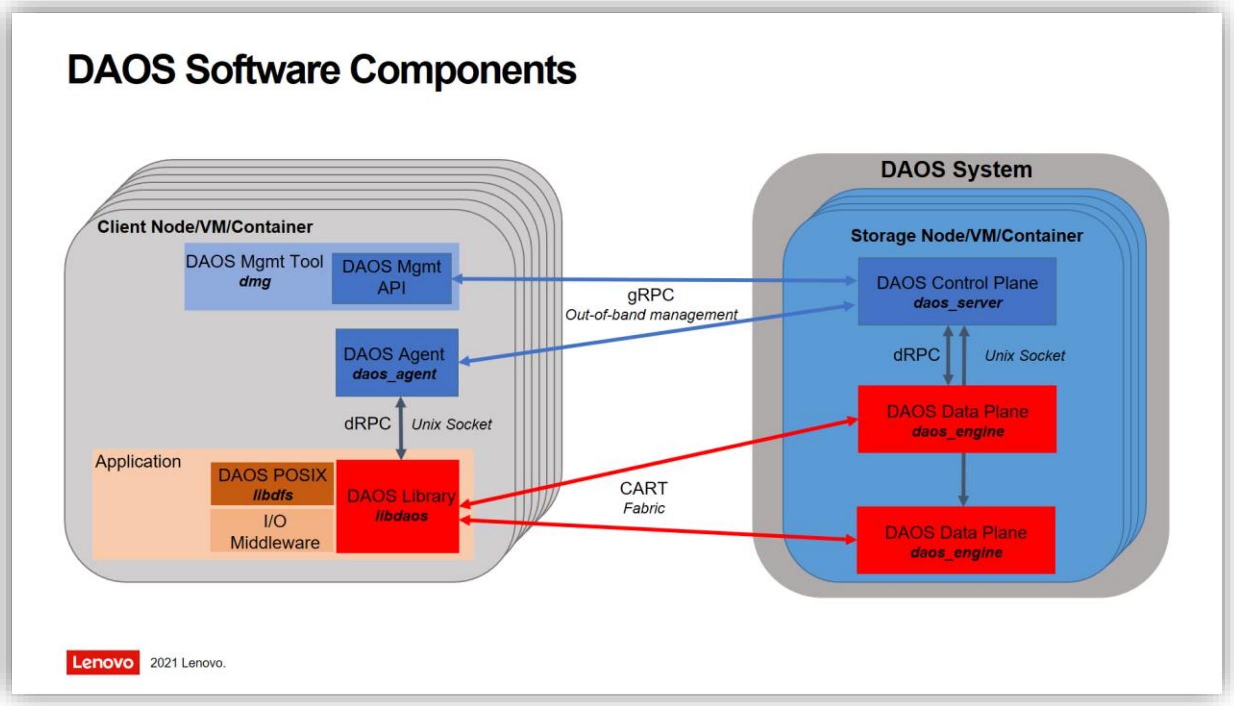
End users can then use the `daos` command to create and manage DAOS *containers* within the DAOS pools to which they have access permissions. Container management is documented in the DAOS User Guide at https://daos-stack.github.io/user/container/. For POSIX containers, these typically need to be mounted on the client nodes using the DAOS `dfuse` daemon.

Figure 9 on page 20 shows the main software components of a DAOS solution, including both DAOS servers and DAOS clients. Much more background information on the DAOS software architecture is available in GitHub at https://github.com/daos-stack/daos/blob/master/src/README.md.

- On the DAOS storage nodes, each server runs one instance of the `daos_server` process which implements the DAOS *control plane*. On a 2-socket server like the SR630, two instances of the `daos_engine` process are started by the `daos_server` process (one on each of the two sockets). These implement the DAOS *data plane*. Communication between these processes happens through local Unix Domain Sockets.
- On the DAOS client nodes, the `daos_agent` process is responsible to authenticate user applications. It uses a certificate to authenticate the client node, and to securely access the DAOS servers through gRCP. The YML configuration file for `daos_agent` contains an

access_points stanza, which points it to the IP address(es) of the DAOS server(s) that manage the DAOS server cluster (aka “DAOS System”).

- The dmg storage administration tool does not use the daos_agent, but communicates directly with the DAOS servers through the DAOS management API. For this purpose, dmg has its own certificate which represent the storage administrator role.
- User applications access DAOS through the libdaos library. On each client node, this user space library uses a Unix Domain Socket to communicate with the local daos_agent for any control traffic. The actual data traffic then happens directly between the libdaos library and all daos_engine processes on all the DAOS servers in the DAOS server cluster (aka “DAOS System”).
- The daos command mentioned above runs in the same way as any other user applications, using the libdaos library on a client node.



DAOS Software Components

2021 Lenovo.

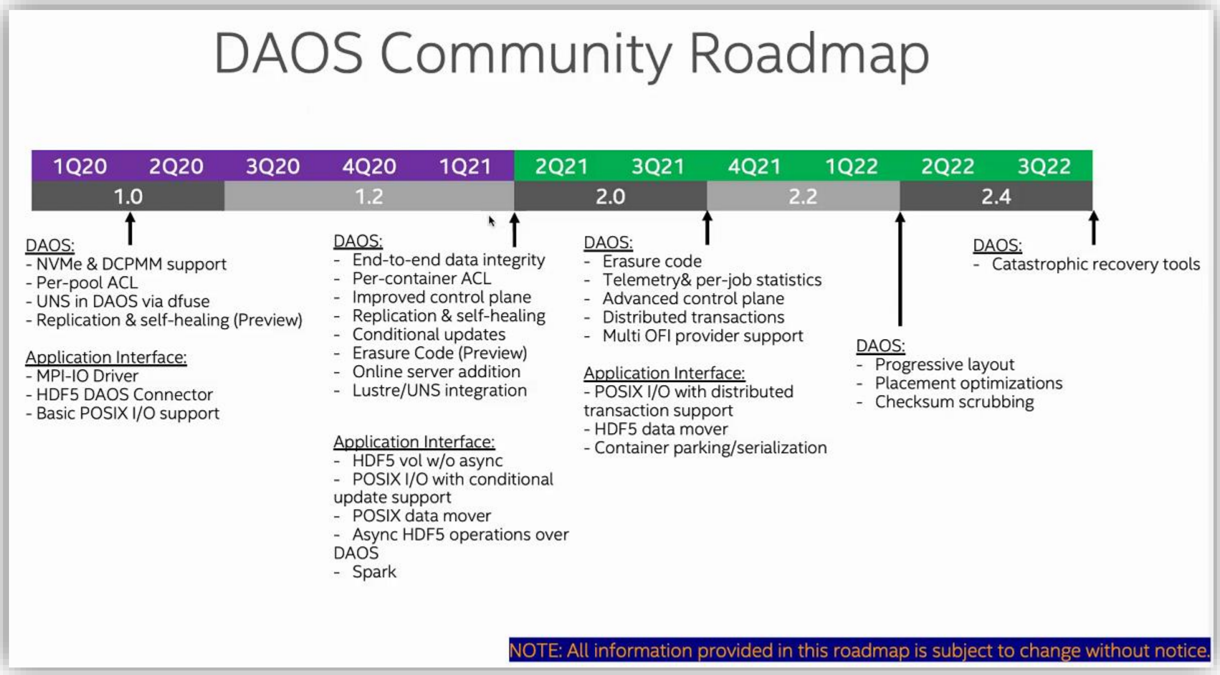
Figure 9: DAOS Software Components.

Note that with the current DAOS 1.1.3 code level, a DAOS client node can only connect to a single DAOS server cluster at any given time. In a future release, DAOS is expected to support the simultaneous connection to more than one DAOS server cluster (aka “DAOS System”).

It is possible in DAOS 1.1.3 that a client node accesses different DAOS server clusters at *different* times. To perform such a change, the daos_agent process needs to be stopped, its YML file is changed to point to a different DAOS server cluster, and then the daos_agent is restarted.

DAOS Roadmap

Intel has released the DAOS 1.0 software on GitHub on 17-Jun-2020, and has made RPM packages for the DAOS 1.0.1 modification level available for download on 14-July-2020. Development snapshots like the current DAOS 1.1.3 "pre-release" are available from Github. Figure 10 shows the current DAOS community roadmap. Refer to https://wiki.hpdd.intel.com/display/DC/Roadmap for the latest version.



DAOS Community Roadmap

DAOS:
- NVMe & DCPMM support
- Per-pool ACL
- UNS in DAOS via dfuse
- Replication & self-healing (Preview)

Application Interface:
- MPI-IO Driver
- HDF5 DAOS Connector
- Basic POSIX I/O support

DAOS:
- End-to-end data integrity
- Per-container ACL
- Improved control plane
- Replication & self-healing
- Conditional updates
- Erasure Code (Preview)
- Online server addition
- Lustre/UNS integration

Application Interface:
- HDF5 vol w/o async
- POSIX I/O with conditional update support
- POSIX data mover
- Async HDF5 operations over DAOS
- Spark

DAOS:
- Erasure code
- Telemetry& per-job statistics
- Advanced control plane
- Distributed transactions
- Multi OFI provider support

Application Interface:
- POSIX I/O with distributed transaction support
- HDF5 data mover
- Container parking/serialization

DAOS:
- Progressive layout
- Placement optimizations
- Checksum scrubbing

DAOS:
- Catastrophic recovery tools

NOTE: All information provided in this roadmap is subject to change without notice.

Figure 10: DAOS Community Roadmap.

A large number of new features will be added in the upcoming DAOS 1.2 and 2.0 releases. In particular, erasure coding, DAOS data movers, and many manageability and monitoring features are currently being developed and will significantly expand the capabilities of DAOS. The current focus of Lenovo's DAOS activities is on operational hardening and "proof-of-concept" application evaluations with early adopters. The DAOS 2.0 release is expected to be "production-ready" for a broad range of HPC and AI usage scenarios.

DAOS Services and Support

Community support for DAOS is available through the DAOS mailing list at https://daos.groups.io/, and the DAOS Slack channel at https://daos-stack.slack.com/. The Intel DAOS development team, Intel partners like Lenovo, and DAOS customers are actively contributing on these community platforms to disseminate DAOS information, identify and fix DAOS issues, and share feedback on DAOS.

The Lenovo Professional Services team can provide DAOS services for a fee. This includes DAOS pre-installation consultancy, implementation services, as well as assistance with DAOS support. Refer to "Lenovo Professional Services" on page 24 for details.

The Lenovo HPC and AI team intends to provide more extensive DAOS support offerings with an upcoming DAOS release.

Sample Bill of Material

This section provides reference information for DAOS server hardware configurations, as well as for contracting Lenovo Professional Services.

Lenovo Hardware Components

The Lenovo Data Center Group (DCG) uses two configurators to create solutions, ranging from individual servers to complete rack-integrated solutions:

- The primary DCG configuration tool is the Data Center Solution Configurator, which is available online at https://dcsc.lenovo.com/#/. There is also an offline version of DCSC, available from the same location.
- For HPC and AI clusters, the capabilities of the System x and Cluster Solutions Configurator (x-Config) are often better suited to configure larger solutions. x-Config is available at https://lesc.lenovo.com/products/hardware/configurator/worldwide/bhui/asit/install.html.

For DCSC, Lenovo provides a *Deployment Ready Solution* that contains the ThinkSystem SR630 server configured as a DAOS server, as described in “DAOS Server Architecture” on page 6. To access this solution, search for the Configuration Reference Number (CRN) SID0000254 titled “Lenovo DAOS-Ready Solution for Flash-Native High Performance Storage”. It can be easily customized, for example changing the number of DAOS servers or the type of NVMe drive.

For x-Config, Table 8 shows the Bill of Material (as shown in the x-Config “Reference” tab) for a cluster of six DAOS nodes, based on ThinkSystem SR630. The components that could be adjusted based on the specifics of a particular solution sizing are highlighted.

Table 8: Lenovo x-Config Hardware Bill of Material for Six SR630 DAOS servers.

PN or FC	Description	Quantity
7X02CTOLWW	ThinkSystem SR630 - 3yr Warranty	6
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	12
AVWB	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	12
AVKF	ThinkSystem SR630 2.5" U.2 10-Bay Backplane Kit	6
AUWC	ThinkSystem SR530/SR570/SR630 x8/x16 PCIe LP+LP Riser 1 Kit	6
AUWA	ThinkSystem SR530/SR570/SR630 x16 PCIe LP Riser 2 Kit	6
B22D	ThinkSystem 810-4P NVMe Switch Adapter	12
5977	Select Storage devices - no configured RAID required	6
B0MJ	Feature Enable TPM 1.2	6
AXCA	ThinkSystem Toolless Slide Rail	6
AUKG	ThinkSystem 1Gb 2-port RJ45 LOM	6
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	6
B6CJ	Intel Xeon Gold 6238 22C 140W 2.1GHz Processor	12
B52B	Intel Optane DC Persistent Memory - App Direct Interleaved Mode	6
B4LV	ThinkSystem 128GB TruDDR4 2666MHz (1.2V) Intel Optane DC Persistent Memory	72
B529	Intel Optane DC Persistent Memory - App Direct Mode only	6
AUWQ	Lenovo ThinkSystem 1U LP+LP BF Riser Bracket	6

B4H2	ThinkSystem 16GB TruDDR4 2933MHz (2Rx8 1.2V) RDIMM	72
AUW1	ThinkSystem SR630 2.5" Chassis with 10 Bays	6
B0RL	Mellanox ConnectX-5 EDR IB VPI Single-port x16 PCIe 3.0 HCA	12
ASR2	20m Mellanox EDR IB Optical QSFP28 Cable	12
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	6
B58A	ThinkSystem U.2 Intel P4610 3.2TB Mainstream NVMe PCIe3.0 x4 Hot Swap SSD	48
B11V	ThinkSystem M.2 5100 480GB SATA 6Gbps Non-Hot Swap SSD	12
ATRR	2U Bracket for Mellanox ConnectX-4 2x100GbE/EDR IB QSFP28 VPI	12
AUTJ	Lenovo ThinkSystem Label Kit	6
AUTC	ThinkSystem SR630 Lenovo Agency Label	6
AUTA	XCC Network Access Label	6
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	6
AUSQ	On Board to 2U 8x2.5" HDD BP NVME Cable	6
AURY	Lenovo ThinkSystem PHY Module Dummy	6
AURR	ThinkSystem M3.5 Screw for Riser 2x2pcs and Planar 5pcs	24
AULQ	ThinkSystem 1U CPU Performance Heatsink	12
AUT8	ThinkSystem 1100W RDN PSU Caution Label	6
AVJ2	ThinkSystem 4R CPU HS Clip	12
AUWF	Lenovo ThinkSystem Super Cap Holder Dummy	6
AUWG	Lenovo ThinkSystem 1U VGA Filler	6
AUW7	ThinkSystem SR630 4056 Fan Module	12
AVWK	ThinkSystem EIA Plate with Lenovo Logo	6
AUX0	ThinkSystem Package for SR630	6
AWF9	ThinkSystem Response time Service Label LI	6
AUX4	MS 1U Service Label LI	6
AUX3	ThinkSystem SR630 Model Number Label	6
AUWN	Lenovo ThinkSystem 1U LP Riser Bracket	6
AWGE	ThinkSystem SR630 WW Lenovo LPK	6
B0ML	Feature Enable TPM on MB	6
B173	Companion Part for XClarity Controller Standard to Enterprise Upgrade in Factory	6
B2RU	ThinkSystem NVMe Drives Only Label	6
B3YF	ThinkSystem SR630 10x2.5" HDD BP NVMe Cable	12
B4NK	ThinkSystem SR630 MB	6
AVEN	ThinkSystem 1x1 2.5" HDD Filler	12
A193	Integrated Solutions	6
A103	System x Cluster Upgrade	6

In both cases (DCSC and x-Config), it is assumed that a cluster management node, as well as the necessary ports on the Management Ethernet and the HPC Interconnect are already in place. Please refer to the LeSI Product Guide for general information on Lenovo’s HPC & AI Clusters.

Lenovo Professional Services

Lenovo offers a variety of professional services in conjunction with its HPC and AI solutions. Please refer to the LeSI Product Guide for general HPC & AI Cluster services. As a general guideline, we recommend to include three Lenovo Professional Services (LPS) Service Units as part of a DAOS engagement, to get customers up and running quickly.

Table 9: Lenovo Professional Services (LPS) Service Unit Part Numbers.

PN or FC	Description	Quantity
5MS7A85672	Professional Service Unit	3

Services are tailored to the customer needs, and typically include:

- Conduct a preparation and planning call
- Configure a cluster management node (e.g. xCAT)
- Verify, and update if needed, firmware on the SR630 servers
- Configure the network settings specific to the customer environment for:
  - XClarity Controller (XCC) service processors on the SR630 servers
  - CentOS on the SR630 servers
- Install CentOS and the HPC networking stack (e.g. Mellanox OFED) on the SR630 servers
- Install and configure DAOS on the SR630 servers
- Validate successful DAOS access from existing client nodes on the HPC fabric
- Provide skills transfer to customer personnel
- Develop post-installation documentation describing the specifics of the firmware/software versions, network configuration, and storage system configuration work that was done

The sizing of a Lenovo Professional Services engagement for a DAOS deployment depends on the size of the DAOS server cluster, as well as the complexity of the required integration work. A detailed Statement of Work (SOW) and associated sizing for a specific project can be provided by the LPS team.

Appendix: Conversion of Decimal and Binary Units

When measuring capacity and bandwidth of high-performance storage systems, the numerical differences between base-10 units and base-2 units are significant. For example, 1000 Byte are one kilo-Byte, with the well-known decimal prefixes of the international SI System. On the other hand, 1024 Byte are one kibi-Byte, with the less well-known binary prefixes (which were first defined in IEC 60027-2).

The effect of this difference is compounding with every order of magnitude, and at Petascale it already results in a difference of over 11%: One Peta-Byte is only 0,888 Pebi-Byte, and one Pebi-Byte equals 1,126 Peta-Byte.

Table 10: Storage capacity measured in base-10 units and base-2 units.

Base-10 Units					Base-2 Units					Base-10 / Base-2		Base-2 / Base-10	
prefix		value			prefix		value			Ratio		Ratio	
kilo	k	10	**	3	kibi	ki	2	**	10	0,976563	-2,34%	1,024000	2,40%
mega	M	10	**	6	mebi	Mi	2	**	20	0,953674	-4,63%	1,048576	4,86%
giga	G	10	**	9	gibi	Gi	2	**	30	0,931323	-6,87%	1,073742	7,37%
tera	T	10	**	12	tebi	Ti	2	**	40	0,909495	-9,05%	1,099512	9,95%
peta	P	10	**	15	pebi	Pi	2	**	50	0,888178	-11,18%	1,125900	12,59%
exa	E	10	**	18	exbi	Ei	2	**	60	0,867362	-13,26%	1,152922	15,29%

The industry standard is to quote *disk storage capacities* in base-10 units, and *memory capacities* in base-2 units. For *Storage Class Memory* there is no established convention, and both units are used. For *bandwidth* (which is capacity transferred per unit of time), there is also no clear industry standard. In this document we always use the correct prefix notation (for example, GiB versus GB), and convert all base-10 numbers to base-2 numbers when quoting solution-level capacity and bandwidth figures.

Additional Resources

Information about DAOS:

- DAOS: Revolutionizing High-Performance Storage with Intel Optane Technology. https://www.intel.com/content/www/us/en/high-performance-computing/daos-high-performance-storage-brief.html
- Liang Z., Lombardi J., Chaarawi M., Hennecke M. (2020)
  DAOS: A Scale-Out High Performance Storage Stack for Storage Class Memory.
  In: Panda D. (editor) Supercomputing Frontiers. SCFA 2020. Lecture Notes in Computer Science, volume 12082. Springer, Cham. https://doi.org/10.1007/978-3-030-48842-0_3
- DAOS on GitHub: https://github.com/daos-stack/daos and https://daos-stack.github.io/
- DAOS Community Home: https://wiki.hpdd.intel.com/display/DC/DAOS+Community+Home
- DAOS mailing list: https://daos.groups.io/
- DAOS Slack channel: https://daos-stack.slack.com/

Lenovo Press Guides:

- Lenovo Scalable Infrastructure (LeSI) Solutions
  https://lenovopress.com/lp0900
- Lenovo ThinkSystem SR630 Server (Xeon SP Gen 2)
  https://lenovopress.com/lp1049
- Intel Optane Persistent Memory 100 Series
  https://lenovopress.com/lp1066
- Introducing the Programming Model of Intel Optane DC Persistent Memory
  https://lenovopress.com/lp1194
- ThinkSystem Intel P4510 Entry NVMe PCIe 3.0 x4 SSDs
  https://lenovopress.com/lp1033
- ThinkSystem Intel P4610 Mainstream NVMe PCIe 3.0 x4 SSDs
  https://lenovopress.com/lp1032
- Intel Optane P4800X Performance NVMe PCIe SSDs
  https://lenovopress.com/lp0770
- ThinkSystem Mellanox ConnectX-6 HDR100 InfiniBand Adapters
  https://lenovopress.com/lp1170

Lenovo ThinkSystem SR630 Maintenance Manual and Setup Guide:

- https://thinksystem.lenovofiles.com/help/topic/7X01/sr630_maintenance_manual.pdf
- https://thinksystem.lenovofiles.com/help/topic/7X01/sr630_setup_guide.pdf

Intel Optane Persistent Memory Documentation:

- https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html
- https://ark.intel.com/content/www/us/en/ark/products/series/190349/intel-optane-persistent-memory.html and the 128GB, 256GB, 512GB Optane PMem modules

About the Author

Michael Hennecke is Lenovo's Chief Technologist for HPC Storage and Networking. He has over 27 years of experience in High Performance Computing and HPC Storage, and is currently focusing on emerging storage solutions like DAOS that utilize non-volatile memory technologies. Michael holds a masters degree in physics from Ruhr-Universität Bochum (Germany), and a "Distinguished IT Specialist" certification from The Open Group.

Thanks to the following people for their contributions to this project:

- Kelsey Prantis (Intel)
- Johann Lombardi (Intel)
- Andrey Kudryavtsev (Intel)
- Bruno Faccini (Intel)
- Mohamad Chaarawi (Intel)
- Taylor Allison (Lenovo)
- Florian Zillner (Lenovo)
- Sigrun Eggerling (Lenovo)
- Martin Bachmaier (Lenovo)
- Nicolas Calimet (Lenovo)
- Wil Wellington (Lenovo)
- David Watts (Lenovo)

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.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

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

Lenovo®
System x®
ThinkSystem
TruDDR4
XClarity®

The following terms are trademarks of other companies:

Intel®, Intel Optane™, and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the United States and other countries.

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