# Designing DAOS Solutions with Lenovo ThinkSystem SR630 V2 Servers

Last update: 24 May 2022

**Introduces DAOS: Distributed Asynchronous Object Storage**

**Describes the SR630 V2 hardware configurations for DAOS**

**Explains the design choices for DAOS storage components**

**Provides capacity and performance 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 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 V2 server is an ideal hardware platform to run the DAOS server software. It provides a balanced system design with a PCIe 4.0 I/O subsystem and combines Intel Optane Persistent Memory (PMem) 200 Series as Storage Class Memory (SCM) with U.2 NVMe SSDs for bulk data storage. Two InfiniBand or high-performance Ethernet fabric adapters 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 V2 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 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 DAOS Version 2.0 release introduces many advanced features including erasure coding, as well as a general focus on robustness and stabilization. DAOS 2.0 is suitable for proof of concept activities, for code porting, and as a scratch storage system. For deployments in mission-critical environments, the additional serviceability features scheduled for DAOS 2.2 will be essential. Please refer to the DAOS roadmap for further information.

**Note:** This guide describes DAOS on the SR630 V2 server with the third generation of the Intel Xeon Processor Scalable Family (Xeon SP Gen 3). For DAOS on SR630 servers with the second generation of the Intel Xeon Processor Scalable Family (Xeon SP Gen 2), refer to the following Lenovo Press guide:

- Designing DAOS Storage Solutions with Lenovo ThinkSystem SR630 Servers
  https://lenovopress.com/lp1398

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"* which is available on the Intel landing page for DAOS:

https://www.intel.com/content/www/us/en/high-performance-computing/daos.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`.

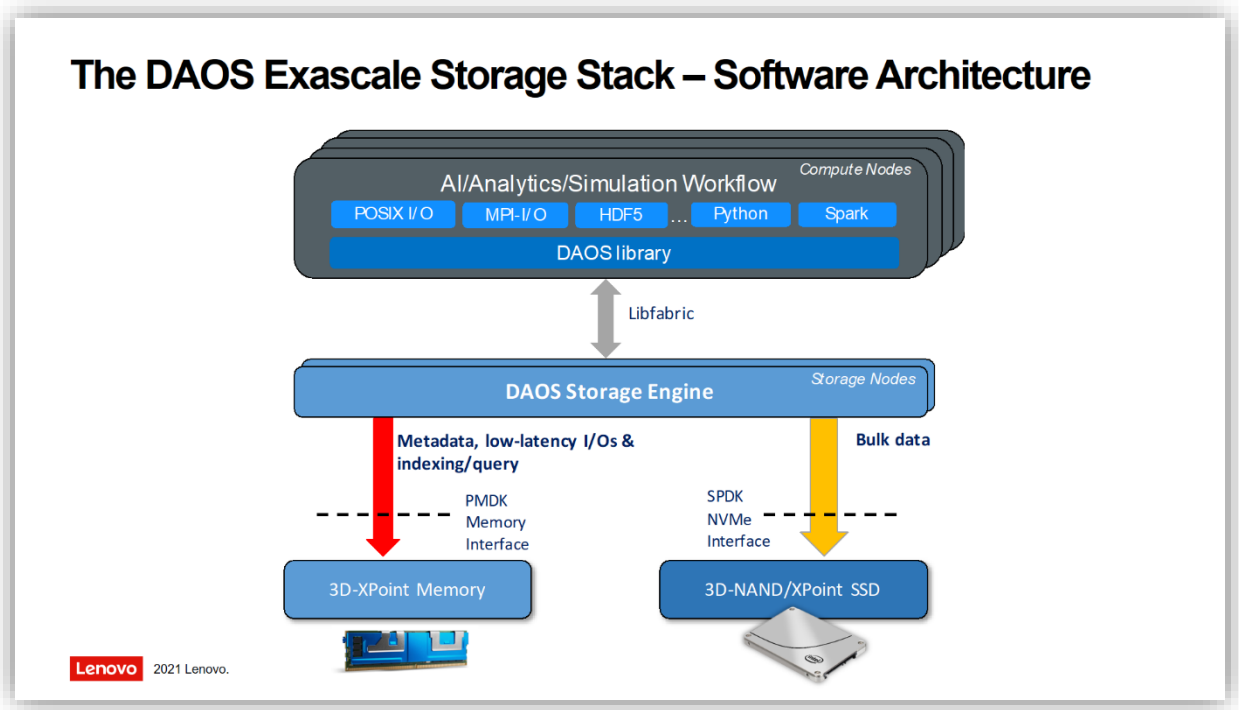


*Figure 1: The DAOS Exascale Storage Stack – Software Architecture.*

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 7.9, CentOS 8.4/8.5, and openSUSE Leap 15.3. 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.

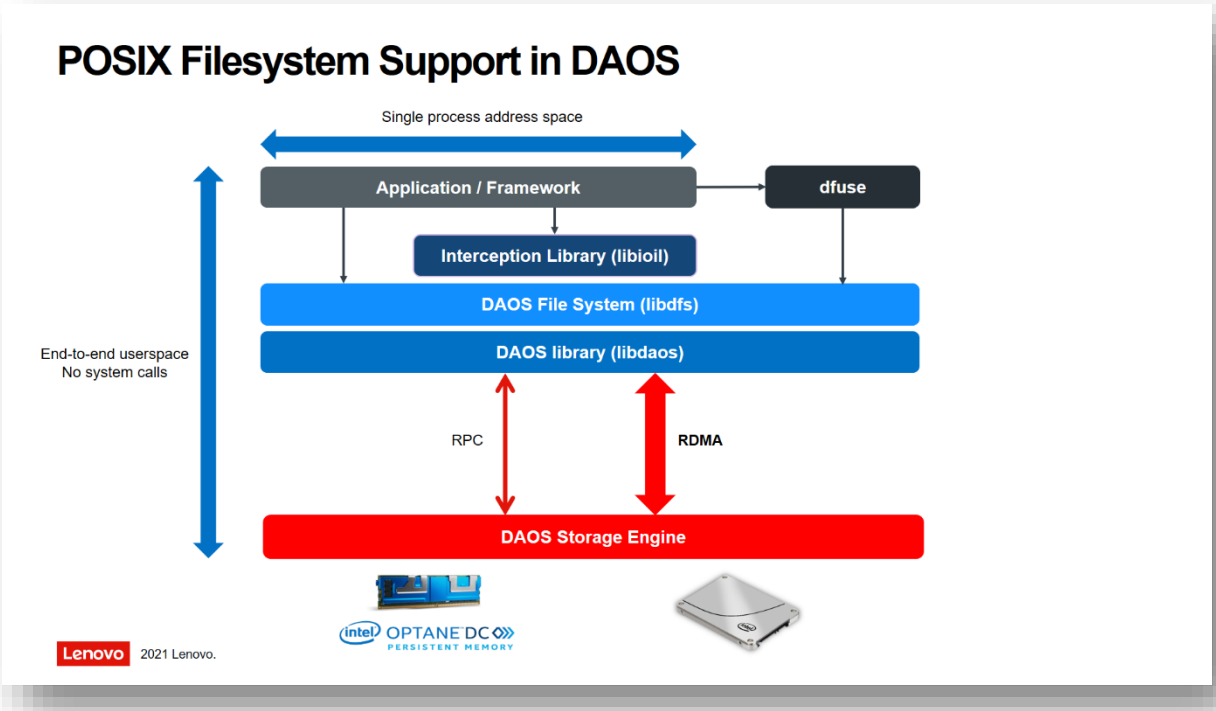


*Figure 2: POSIX Filesystem Support in DAOS.*

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.

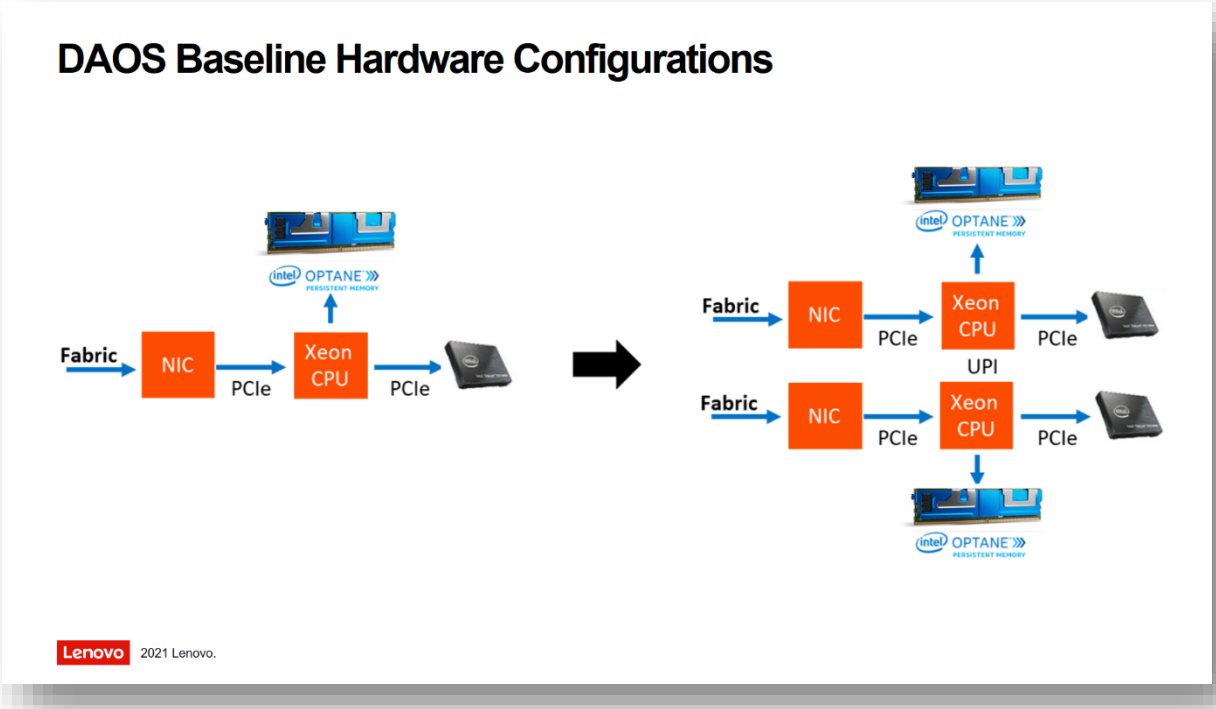


*Figure 3: DAOS Baseline Hardware Configurations.*

The Lenovo ThinkSystem SR630 V2 is a 1U, 2-socket server that is ideally suited as a DAOS server. It provides a perfect balance of Intel Optane PMem 200 Series, NVMe PCIe 4.0 SSDs, and PCIe 4.0 network cards attached to the two Intel Xeon CPUs. Figure 4 on page 7 shows the internal server architecture of the SR630 V2 when configured as a DAOS server. Each CPU socket manages one 200 Gbps network card (16 lanes of PCIe 4.0) and up to eight Intel Optane PMem 200 Series modules. In terms of NVMe bulk storage, there are two good configuration choices:

- Each socket can serve four NVMe PCIe 4.0 SSDs that are directly connected to that socket, for a total of *eight* NVMe SSDs per server. This is a 100% symmetric configuration, and it matches the DAOS server configuration of the previous generation SR630 server.
- The SR630 V2 server provides more PCIe lanes than the SR630, and it is possible to populate the maximum of *ten* U.2 NVMe SSDs per server (each connected with 4 lanes of PCIe 4.0). With this configuration, four NVMe SSDs are connected to one socket and six NVMe SSDs are connected to the second socket. This implies that one of the ten NVMe SSDs will be accessed through the UPI links – this has no impact on bandwidth, and a negligible effect on latency.

In general, we recommend populating all *ten* NVMe SSDs to optimize price/performance.

 

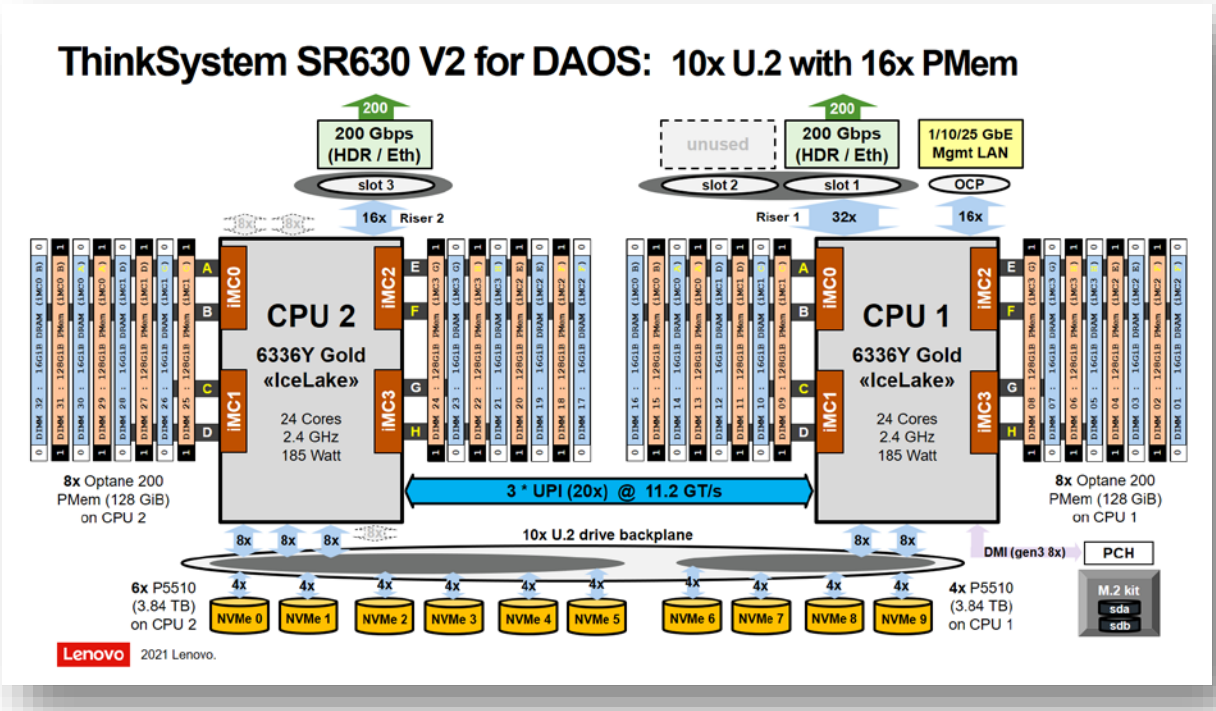


*Figure 4: DAOS Server Architecture: Lenovo ThinkSystem SR630 V2 with 10x U.2 and 16x PMem.*

There are many design choices regarding the individual components of the SR630 V2 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 V2 server chassis supports up to ten U.2 NVMe PCIe 4.0 SSDs in 1U. As discussed above, either *eight* or *ten* U.2 NVMe PCIe 4.0 SSDs per SR630 V2 server (*four* or *five* per `daos_engine`) are a good choice for DAOS. All NVMe SSDs in a DAOS server should be identical.

All U.2 NVMe SSDs that are supported in the SR630 V2 should work with DAOS. We have validated the DAOS software stack with the Intel D7-P5500 and D7-P5600 series of U.2 NVMe PCIe 4.0 SSDs, which provide a wide range of capacities in the Entry and Mainstream space.

**Note:** The SR630 V2 also has an E1.S NVMe backplane option, which supports up to sixteen NVMe SSDs in the EDSFF E1.S form factor. This is an interesting option for the future. But currently no PCIe 4.0 NVMe SSD in E1.S form factor is supported in the SR630 V2. The 4 TB Intel P4511 NVMe SSD that *is* available in E1.S packaging is a PCIe 3.0 device and is not a good choice for a DAOS storage server.

Please refer to “Capacity Sizing” on page 12 and “Performance Sizing” on page 15 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 PCIe 4.0 SSDs.*

| Description | Part number | Feature code |
|---|---|---|
| ThinkSystem U.2 Intel P5500 1.92TB Entry NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17145 | BCFT |
| ThinkSystem U.2 Intel P5500 3.84TB Entry NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17146 | BCFW |
| ThinkSystem U.2 Intel P5500 7.68TB Entry NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17147 | BCFU |
| ThinkSystem U.2 Intel P5600 1.6TB Mainstream NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17152 | BCFV |
| ThinkSystem U.2 Intel P5600 3.2TB Mainstream NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17153 | BCFR |
| ThinkSystem U.2 Intel P5600 6.4TB Mainstream NVMe PCIe 4.0 x4 Hot Swap SSD | 4XB7A17154 | BCFS |

## Intel Optane Persistent Memory Modules

DAOS 2.0 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 third generation Intel Xeon SP CPUs (“Ice Lake”) has eight memory channels on two memory controllers. With two sockets, the SR630 V2 server has sixteen memory channels, and supports two DIMMs per channel. With one DRAM memory module per channel to achieve the best memory bandwidth, the second DIMM slot on each channel is available for PMem.

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 200 Series modules. This implies that an SR630 V2 DAOS server should contain four or eight Intel Optane PMem 200 Series modules per socket (eight or sixteen modules per 2-socket server). As DAOS builds heavily on SCM, we strongly recommend to populate all sixteen PMem modules to achieve the best performance.

Intel Optane PMem 200 Series modules are available in 128 GB, 256 GB and 512 GB capacity.

*Table 2: Intel Optane PMem 200 Series Modules.*

| Description | Part number | Feature code |
|---|---|---|
| ThinkSystem 128GB TruDDR4 3200MHz (1.2V) Intel Optane Persistent Memory | 4ZC7A08732 | B98B |
| ThinkSystem 256GB TruDDR4 3200MHz (1.2V) Intel Optane Persistent Memory | 4ZC7A08734 | B98A |
| ThinkSystem 512GB TruDDR4 3200MHz (1.2V) Intel Optane Persistent Memory | 4ZC7A08736 | BB8T |

Please refer to “Capacity Sizing” on page 12 and “Performance Sizing” on page 15 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 V2 DAOS Server:

- Intel Optane Persistent Memory 200 Series is supported on all third generation Intel Xeon Gold and Platinum SKUs.
- Within a DAOS engine, storage is managed by multiple *storage targets*. Each storage target is a thread in the daos_engine process, executing on one physical core and managing a fraction of the PMem capacity and a fraction of *one* of the NVMe disks controlled by this engine. For a balanced configuration, the number of targets per engine should be a multiple of the number of NVMe disks managed by the engine. In addition, a certain minimum number of targets per engine is needed to fully exploit the performance of the PMem devices. This leads to the following recommendation:
  - When *eight* NVMe disks are populated in the SR630 V2, *16 targets* per engine (and thus per CPU socket) should be used.
  - With *ten* NVMe disks, the number of targets per engine should be *15* or *20*.

  Together with one core for general operating system tasks, and up to six "helper" cores to assist with erasure coding and other computationally intensive tasks, the general advice for DAOS servers is to use CPUs with at least 24 physical cores per socket.
- The third generation Intel Xeon processors do *not* impose a limit on the maximum memory capacity per socket. All Gold and Platinum SKUs support up to 6 TB of memory per socket (which is sufficient for 16x 128 GiB of DRAM plus 16x Intel Optane PMem 200 Series).
- For some processors, a "Y" SKU is available for which some cores can be disabled. This can be used to reduce the power consumption and/or to run fewer cores at a higher frequency.

The 6336Y processor that is shown in Figure 4 on page 7 is a "*high core count*" (HCC) SKU that has 24 cores and a TDP of 185 Watt. (It is a "Y" SKU, but running it with fewer cores is not recommended for DAOS.) Considering the overall price/performance of the server, the 6336Y is our default recommendation for a DAOS server. Other good processor choices are shown in Table 3 on page 9. This includes the "extreme core count" (XCC) Gold SKU 6338 as well as two XCC Platinum "Y" SKUs. For the Platinum "Y" SKUs it may be beneficial to reduce the number of active cores to minimize the power consumption in the 1U SR630 V2 server.

*Table 3: Third Generation Intel Xeon Processors.*

| CPU Model | XCC/ HCC | Cores | Core speed (Base) | TDP Power | Cache Size | Max DDR Speed | UPI speed |
|---|---|---|---|---|---|---|---|
| 8352Y | XCC | 32 | 2.2 GHz | 205 W | 48 | 3200 | 11.2 |
| 8352Y | XCC | 24 | 2.3 GHz | 185 W | 48 | 3200 | 11.2 |
| 6338 | XCC | 32 | 2.0 GHz | 205W | 48 | 3200 | 11.2 |
| 6342 | HCC | 24 | 2.7 GHz | 220 W | 36 | 3200 | 11.2 |
| 6336Y | HCC | 24 | 2.4 GHz | 185 W | 36 | 3200 | 11.2 |

 

## High Performance Network Adapters

To ensure a balanced network bandwidth that matches the eight or ten NVMe PCIe 4.0 SSDs (with a total of 32 or 40 PCIe 4.0 lanes), two PCIe 4.0 16-lane Mellanox 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. The ConnectX-6 cards support full HDR speed (200 Gbps), and since they are VPI cards they can also be used in 100 Gbps or 200 Gbps Ethernet mode (together with the `libfabrics ofi+tcp` or `ofi+sockets` provider). Note that HDR100 (or EDR) cards should not be used in this context, as they will not provide enough network bandwidth to match the storage performance of the eight or ten NVMe PCIe 4.0 SSDs. (DAOS does not currently support the striping of a single engine's traffic across multiple fabric links, so aggregating two 100Gbps network links for a single engine is not possible.)

*Table 4: High-Performance Network Adapters.*

| Description | Part number | Feature code |
|---|---|---|
| ThinkSystem Mellanox ConnectX-6 HDR QSFP56 1-port PCIe 4 InfiniBand Adapter | 4C57A15326 | B4RC |

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

- 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 sixteen 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. In most environments 16 GiB DIMMs should be sufficient, providing 256 GiB per DAOS server. If needed, sixteen 32 GiB DIMMs can provide 512 GiB per server.

*Table 5: DDR4 Memory DIMMs.*

| Description | Part number | Feature code |
|---|---|---|
| ThinkSystem 16GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM | SBB7A18397 | B963 |
| ThinkSystem 32GB TruDDR4 3200 MHz (2Rx4 1.2V) RDIMM | SBB7A18398 | B964 |
| ThinkSystem 32GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM | SBB7A18399 | B965 |

## Boot Devices

Two M.2 SATA SSDs are used in a hardware RAID1 configuration to hold the operating system. The mirroring is performed through the "ThinkSystem M.2 SATA 2-Bay RAID 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). 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.

Check the Lenovo Operating System Interoperability Guide (OSIG) for detailed information about OS compatibility with the SR630 V2 server:

https://lenovopress.com/osig#servers=sr630-v2-7z70-7z71&support=all

DAOS 2.0 has been primarily validated with CentOS Linux 7.9, CentOS Linux 8.4/8.5 (transitioning to Rocky Linux 8.5) and openSUSE Leap 15.3. RHEL 8.4 (EUS), RHEL 8.5, and SLES 15.3 are also supported.

 

# 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 providing 6% of a DAOS 2.0 server's NVMe SSD capacity as 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 PCIe 4.0 SSDs listed in Table 1 on page 8, this capacity ratio implies a certain population with Intel Optane PMem modules. Table 6 shows the resulting combinations when using *eight* NVMe SSDs per server, and Table 1 on page 13 shows the same information for *ten* NVMe SSDs per server. Color coding in the tables highlights where some of the combinations are deviating from the optimal design.

*Table 6: Single SR630 V2 DAOS Server Capacity Sizing Options (eight NVMe SSDs per server).*

| 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 | 6336Y | 8 | P5500 | 1,92 | 8 | 128 | 16 | 16 | 14,0 | 0,93 | 6,7% | 256 |
| 2 | 6336Y | 8 | P5500 | 1,92 | 16 | 128 | 16 | 16 | 14,0 | 1,86 | 13,3% | 256 |
| 2 | 6336Y | 8 | P5500 | 3,84 | 8 | 128 | 16 | 16 | 27,9 | 0,93 | 3,3% | 256 |
| 2 | 6336Y | 8 | P5500 | 3,84 | 16 | 128 | 16 | 16 | 27,9 | 1,86 | 6,7% | 256 |
| 2 | 6336Y | 8 | P5500 | 7,68 | 16 | 128 | 16 | 16 | 55,8 | 1,86 | 3,3% | 256 |
| 2 | 6336Y | 8 | P5500 | 7,68 | 16 | 256 | 16 | 16 | 55,8 | 3,72 | 6,7% | 256 |
| 2 | 6336Y | 8 | P5600 | 1,6 | 8 | 128 | 16 | 16 | 11,6 | 0,93 | 8,0% | 256 |
| 2 | 6336Y | 8 | P5600 | 1,6 | 16 | 128 | 16 | 16 | 11,6 | 1,86 | 16,0% | 256 |
| 2 | 6336Y | 8 | P5600 | 3,2 | 8 | 128 | 16 | 16 | 23,3 | 0,93 | 4,0% | 256 |
| 2 | 6336Y | 8 | P5600 | 3,2 | 16 | 128 | 16 | 16 | 23,3 | 1,86 | 8,0% | 256 |
| 2 | 6336Y | 8 | P5600 | 6,4 | 16 | 128 | 16 | 16 | 46,5 | 1,86 | 4,0% | 256 |
| 2 | 6336Y | 8 | P5600 | 6,4 | 16 | 256 | 16 | 16 | 46,5 | 3,72 | 8,0% | 256 |

The following non-optimal component choices occur in Table 6 and Table 7:

- For best performance, we recommend to always populate all 16 PMem slots. But for small NVMe capacities, this will not be optimal in terms of price/performance. Configurations with only eight PMem modules are marked in 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 modules are marked in orange.
- 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 incur a higher cost than what would be necessary.

*Table 7: Single SR630 V2 DAOS Server Capacity Sizing Options (ten NVMe SSDs per server).*

| 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 | 6336Y | 10 | P5500 | 1,92 | 8 | 128 | 16 | 16 | 17,5 | 0,93 | 5,3% | 256 |
| 2 | 6336Y | 10 | P5500 | 1,92 | 16 | 128 | 16 | 16 | 17,5 | 1,86 | 10,7% | 256 |
| 2 | 6336Y | 10 | P5500 | 3,84 | 8 | 128 | 16 | 16 | 34,9 | 0,93 | 2,7% | 256 |
| 2 | 6336Y | 10 | P5500 | 3,84 | 16 | 128 | 16 | 16 | 34,9 | 1,86 | 5,3% | 256 |
| 2 | 6336Y | 10 | P5500 | 7,68 | 16 | 128 | 16 | 16 | 69,8 | 1,86 | 2,7% | 256 |
| 2 | 6336Y | 10 | P5500 | 7,68 | 16 | 256 | 16 | 16 | 69,8 | 3,72 | 5,3% | 256 |
| 2 | 6336Y | 10 | P5600 | 1,6 | 8 | 128 | 16 | 16 | 14,5 | 0,93 | 6,4% | 256 |
| 2 | 6336Y | 10 | P5600 | 1,6 | 16 | 128 | 16 | 16 | 14,5 | 1,86 | 12,8% | 256 |
| 2 | 6336Y | 10 | P5600 | 3,2 | 8 | 128 | 16 | 16 | 29,1 | 0,93 | 3,2% | 256 |
| 2 | 6336Y | 10 | P5600 | 3,2 | 16 | 128 | 16 | 16 | 29,1 | 1,86 | 6,4% | 256 |
| 2 | 6336Y | 10 | P5600 | 6,4 | 16 | 128 | 16 | 16 | 58,2 | 1,86 | 3,2% | 256 |
| 2 | 6336Y | 10 | P5600 | 6,4 | 16 | 256 | 16 | 16 | 58,2 | 3,72 | 6,4% | 256 |

Combining all these factors shows that the current sweet-spot configuration for the D7-P5600 “Mainstream” NVMe SSDs (3 DWPD) is *ten* 3.2 TB SSDs and sixteen 128 GB PMem modules. This provides a raw capacity of 29.1 TiB + 1.86 TiB per server. For D7-P5500 “Entry” NVMe SSDs, the sweet spot is *ten* 3.84 TB SSDs and sixteen 128 GB PMem modules, with a raw capacity of 34.9 TiB + 1.86 TiB per server.

Doubling the capacity of the NVMe SSDs while keeping the 128 GB PMem modules is possible when the required PMem capacity ratio is reduced in a future DAOS release, or when the workload is known to require less than the default 6% of SCM capacity. Otherwise, the more expensive 256 GB PMem modules need to be used with the higher-capacity NVMe SSDs.

## 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 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 is available starting with DAOS 2.0. This improves 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. For example, an 8+2P implementation should contain 2*(8+2)=20 servers and the technical minimum for an 8+2P environment is 10 servers.

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 HDR InfiniBand port of a Mellanox ConnectX-6 HDR adapter in a PCIe 4.0 x16 slot has a maximum "wire speed" bandwidth of 23.2 GiB/s (four lanes operating at a bit rate of 50 Gb/s). This means that for an SR630 V2 server with two HDR ports (two single-port NICs in two PCIe 4.0 x16 slots), the peak aggregate InfiniBand bandwidth is 46.4 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 8 summarizes the properties of the Intel NVMe PCIe 4.0 SSDs that are currently supported in the SR630 V2 (based on the Intel specifications at https://ark.intel.com/content/www/us/en/ark.html).

*Table 8: Intel U.2 NVMe PCIe 4.0 SSD specifications.*

| Drive Series | Storage Technology | Capacity [GB] | Sequential Read [MB/s] | Sequential Write [MB/s] | Random Read [k IOPS] | Random Write [k IOPS] | Read Latency [usec] | Write Latency [usec] | Active Power [W] | Idle Power [W] | Write Endurance [PBW] | Write Endurance [DWPD] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Entry (~1 DWPD)** | | | | | | | | | | | | |
| P5500 | 96layer 3D TLC NAND | 1920 | 3500 | 1700 | 400,0 | 59,0 | 78 | 22 | 15,0 | 5,0 | 3,50 | 1,00 |
| P5500 | 96layer 3D TLC NAND | 3840 | 7000 | 3500 | 780,0 | 118,0 | 78 | 17 | 18,0 | 5,0 | 7,00 | 1,00 |
| P5500 | 96layer 3D TLC NAND | 7680 | 7000 | 4300 | 1000,0 | 130,0 | 78 | 15 | 20,0 | 5,0 | 14,00 | 1,00 |
| **Mainstream (~3-5 DWD)** | | | | | | | | | | | | |
| P5600 | 96layer 3D TLC NAND | 1600 | 3500 | 1700 | 400,0 | 118,0 | 78 | 16 | 15,0 | 5,0 | 8,00 | 2,74 |
| P5600 | 96layer 3D TLC NAND | 3200 | 7000 | 3500 | 780,0 | 230,0 | 78 | 14 | 18,0 | 5,0 | 17,50 | 3,00 |
| P5600 | 96layer 3D TLC NAND | 6400 | 7000 | 4300 | 1000,0 | 260,0 | 78 | 14 | 20,0 | 5,0 | 35,00 | 3,00 |

Other suppliers' NVMe SSDs that are supported in the SR630 V2 should also work with DAOS but have not been validated.

 

Table 8 on page 15 shows that the P5600 and the P5500 NVMe SSDs of comparable capacity have very similar *sequential read* and *sequential write* bandwidth, and comparable *random read* performance. The main technical difference between the P5600 NVMe SSDs and the P5500 NVMe SSDs are their *random write* performance, and their write *endurance* (3 DWPD compared to 1 DWPD over a duration of 5 years).

Small I/O requests (<4kiB) will be serviced by the Intel Optane PMem 200 Series 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 200 Series 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 200 Series modules that are installed in the server. Always populating the maximum of sixteen PMem modules in an SR630 V2 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. Nevertheless, 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 5 contains the results of a valid IO500 run for a single DAOS server with ten Intel P5500 3.84 TB NVMe SSDs and sixteen 128 GB Intel Optane 200 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 Lenovo SD530v2 client nodes with a single HDR connection each, using 72 MPI tasks per node. This run would qualify for the IO500's "10 Node Challenge" list with an overall IO500 score of `146.6`. 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 ten NVMe SSDs can be achieved on the application layer for large sequential I/O. This is generally achievable 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-isc22_v1-12 (standard)
[RESULT]       ior-easy-write       30.005367 GiB/s : time 327.927 seconds
[RESULT]    mdtest-easy-write      952.102200 kIOPS : time 325.056 seconds
[       ]            timestamp        0.000000 kIOPS : time 0.001 seconds
[RESULT]       ior-hard-write       23.320737 GiB/s : time 323.752 seconds
[RESULT]    mdtest-hard-write      295.621476 kIOPS : time 332.041 seconds
[2022-05-12T10:28:06] Walking
                      /daos/daos_icx_dev1/xmhennecke/io500/datafiles/2022.05.12-10.05.50
[2022-05-12T10:33:06] Walked 298173758 items in 300.228067 seconds (993157.505670 files/sec)
[2022-05-12T10:33:07] Full Scanned List:
[2022-05-12T10:33:07] Items: 298173758
[2022-05-12T10:33:07]   Directories: 728
[2022-05-12T10:33:07]   Files: 298173030
[2022-05-12T10:33:07]   Links: 0
[2022-05-12T10:33:07] Data: 66.299 PB (238.748 MB per file)
[2022-05-12T10:33:07] Matched List:
[2022-05-12T10:33:07] Items: 241904
[2022-05-12T10:33:07]   Directories: 0
[2022-05-12T10:33:07]   Files: 241904
[2022-05-12T10:33:07]   Links: 0
[2022-05-12T10:33:07] Data: 899.951 MB (3.810 KB per file)
MATCHED 241904/298173758
[RESULT]                    find      990.337896 kIOPS : time 320.504 seconds
[RESULT]       ior-easy-read       45.623526 GiB/s : time 228.102 seconds
[RESULT]    mdtest-easy-stat     1067.807004 kIOPS : time 293.879 seconds
[RESULT]       ior-hard-read       43.786469 GiB/s : time 181.454 seconds
[RESULT]    mdtest-hard-stat      847.550844 kIOPS : time 128.466 seconds
[RESULT]  mdtest-easy-delete      366.840532 kIOPS : time 823.169 seconds
[RESULT]    mdtest-hard-read      673.258332 kIOPS : time 156.939 seconds
[RESULT]  mdtest-hard-delete      375.828201 kIOPS : time 376.001 seconds
[SCORE ] Bandwidth 34.384879 GiB/s : IOPS 625.444440 kiops : TOTAL 146.648666
```

*Figure 5: Single DAOS Server IO500 10-node results with 8x P4610* ***3.2 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 from the user's perspective will be at best 1/3 of the aggregated 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 should not be affected by the data protection scheme. It is sufficient to read from any one of the copies of the data for replication schemes, and to read only 8 of the 10 strips of an 8+2P EC stripe. Note that minor differences in read bandwidth are expected, as the EC cell size and the *chunking* of the application I/O requests on the client side will produce slight variations in the actual I/O patterns, depending on the data protection scheme.

 

# 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 implementing 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 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 online documentation at https://docs.daos.io/v2.0/.

A more convenient method to install DAOS is to use the RPM packages that are provided at https://packages.daos.io/. With the RPM packages, installing DAOS is as simple as running "`yum install daos-server daos-client`" (or "`yum install daos-client daos-devel`" 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.x releases and DAOS 2.0. To upgrade a DAOS system from DAOS 1.0 or 1.2 to DAOS 2.0, the DAOS storage that has been formatted with DAOS 1.x should be de-provisioned, and the DAOS 1.x software should be un-installed before installing the 2.0 release. Starting with the DAOS 2.0 release, there will be at least "N-1" interoperability of the DAOS releases.

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 (preferably by using the DAOS `systemd` integration);
- Discovering, preparing and formatting the storage hardware.

 

The DAOS deployment process is described at https://docs.daos.io/v2.0/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://docs.daos.io/v2.0/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://docs.daos.io/v2.0/user/container/. For POSIX containers, these typically need to be mounted on the client nodes using the DAOS `dfuse` daemon.

Figure 6 on page 21 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/release/2.0/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 V2, 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`. It 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.

 

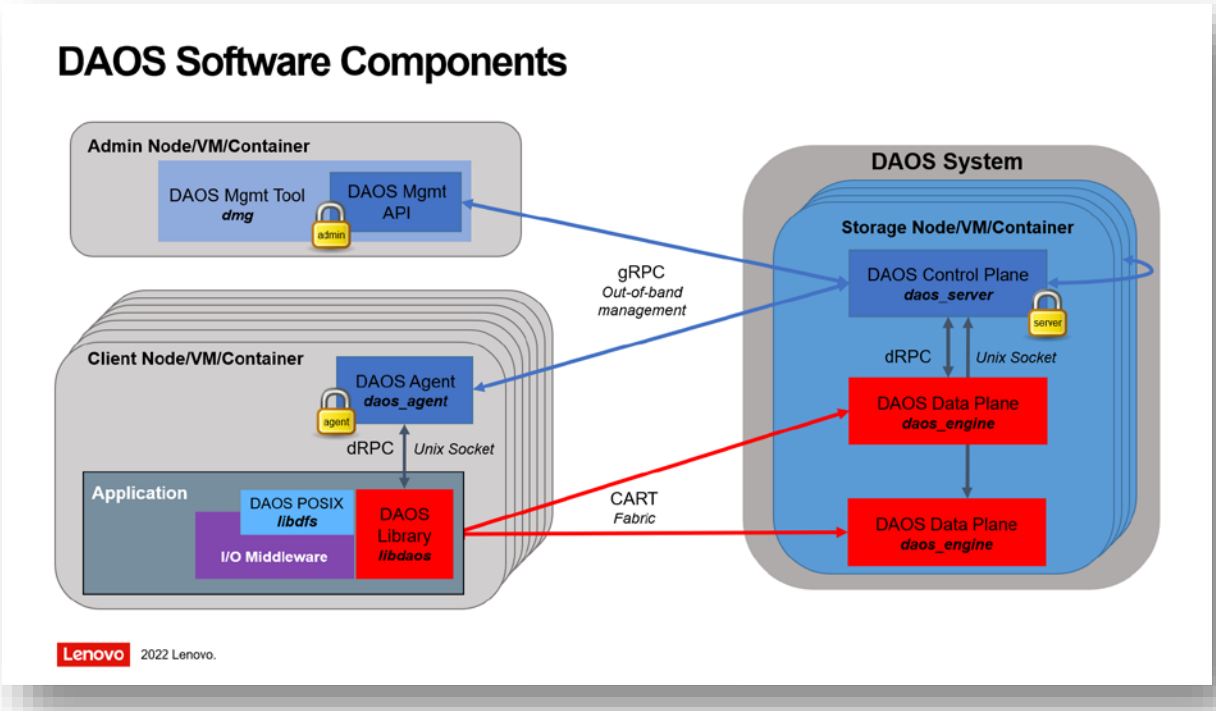


*Figure 6: DAOS Software Components.*

Note that with the current DAOS 2.0 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 2.0 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

The DAOS 2.0 software version has been released on GitHub in December 2021, and RPM packages for DAOS 2.0 are available for download on https://packages.daos.io/. Development snapshots for the future DAOS 2.2 release may be tagged on Github, but they will usually not be provided in RPM format.

Figure 7 shows the current DAOS community roadmap (the latest version of the roadmap is available online at https://daosio.atlassian.net/wiki/spaces/DC/pages/4836661105/Roadmap).

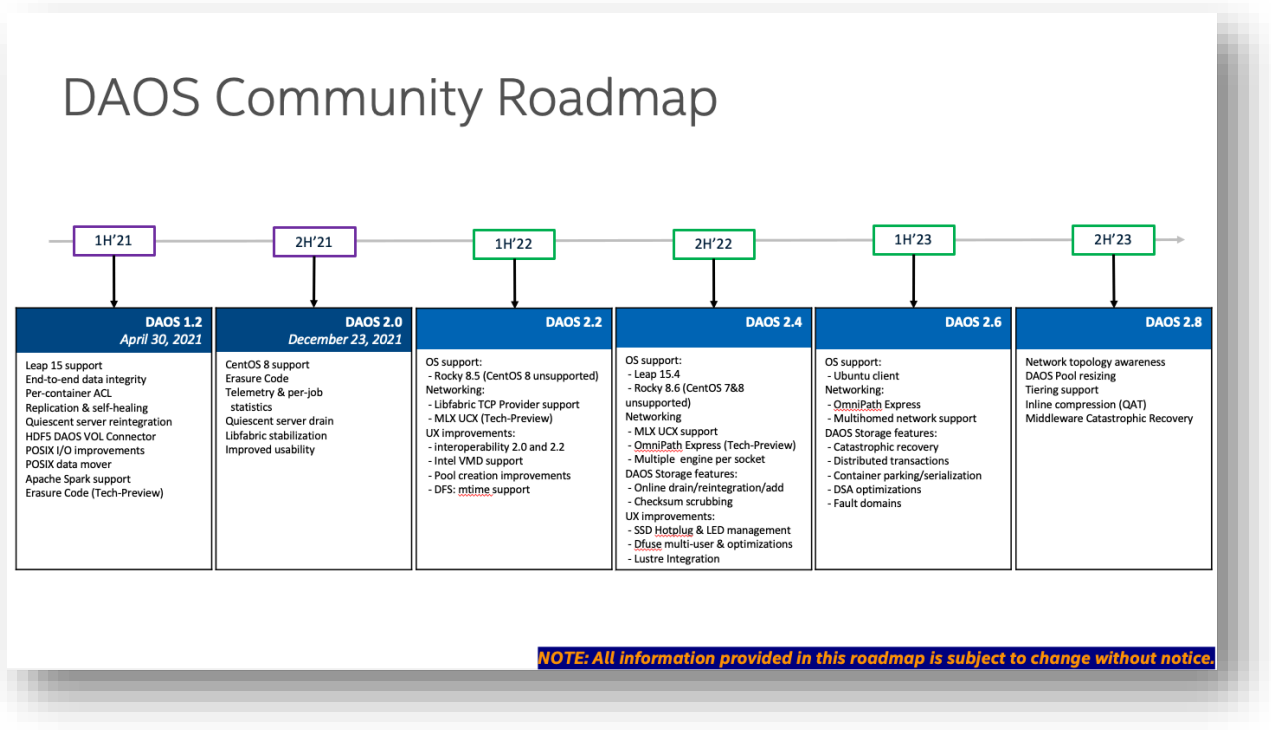


*Figure 7: DAOS Community Roadmap.*

Many new features will be added in the upcoming DAOS releases. In particular, new fabric support as well as additional manageability features are currently being developed for DAOS 2.2.

## 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 DAOS support. Refer to "Lenovo Professional Services" on page 24 for details.

 

# 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 Infrastructure Solutions Group (ISG) uses two configurators to create solutions, ranging from individual servers to complete rack-integrated solutions:

- The primary ISG configuration tool is the Data Center Solution Configurator (DCSC), 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.

DAOS solutions can be part of a Lenovo LeSI cluster in which case x-Config has to be used to configure the DAOS servers. Table 9 shows the x-Config Bill of Material (as shown in the x-Config "Reference" tab) for a single DAOS node based on ThinkSystem SR630 V2. The components that could be adjusted based on the specifics of a particular solution sizing are highlighted.

*Table 9: Lenovo x-Config Hardware Bill of Material for one SR630 V2 DAOS server.*

| PN or FC | Description | Quantity |
|---|---|---|
| **7Z71CTOLWW** | ThinkSystem SR630 V2 - 3yr Warranty - HPC&AI | 1 |
| B8N2 | ThinkSystem 1U PCIe Gen4 x16/x16 Riser 1 | 1 |
| B0MK | Enable TPM 2.0 | 1 |
| B4RC | ThinkSystem Mellanox ConnectX-6 HDR QSFP56 1-port PCIe 4 InfiniBand Adapter | 2 |
| B4QY | 10m Mellanox HDR IB Optical QSFP56 Cable | 2 |
| BCQQ | ThinkSystem 1U 2.5" 10 NVMe Backplane | 1 |
| B5XH | ThinkSystem M.2 SATA 2-Bay RAID Enablement Kit | 1 |
| BCFW | ThinkSystem U.2 Intel P5500 3.84TB Entry NVMe PCIe 4.0 x4 Hot Swap SSD | 10 |
| BFYB | -SB- Operating mode selection for: "Maximum Performance Mode" | 1 |
| B8QC | ThinkSystem 1100W (230V/115V) v2 Platinum Hot-Swap Power Supply | 2 |
| B7Y0 | Enable IPMI-over-LAN | 1 |
| 5977 | Select Storage devices - no configured RAID required | 1 |
| B8LA | ThinkSystem Toolless Slide Rail Kit v2 | 1 |
| B5SV | ThinkSystem Broadcom 57454 10/25GbE SFP28 4-port OCP Ethernet Adapter | 1 |
| B8NC | ThinkSystem 1U LP+LP BF Riser Cage Riser 1 | 1 |
| B8N6 | ThinkSystem 1U 2.5" Chassis with 8 or 10 Bays | 1 |
| BB3S | Intel Xeon Gold 6336Y 24C 185W 2.4GHz Processor | 2 |
| AUUV | ThinkSystem M.2 128GB SATA 6Gbps Non-Hot Swap SSD | 2 |
| B963 | ThinkSystem 16GB TruDDR4 3200 MHz (2Rx8 1.2V) RDIMM | 16 |
| B52B | Intel Optane Persistent Memory 200 Series - App Direct Interleaved Mode | 1 |

 

| | | |
|---|---|---|
| B8N4 | ThinkSystem 1U Performance Fan Option Kit | 8 |
| B98B | ThinkSystem 128GB TruDDR4 3200MHz (1.2V) Intel Optane Persistent Memory | 16 |
| 6400 | 2.8m, 13A/100-250V, C13 to C14 Jumper Cord | 2 |
| B529 | Intel Optane DC Persistent Memory - App Direct Mode only | 1 |
| B8MV | ThinkSystem 1U PCIe Gen4 x16 Riser 2 | 1 |
| B8N7 | ThinkSystem 1U MS LP Riser Cage Riser1&2 | 1 |

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 including three Lenovo Professional Services (LPS) Service Units as part of a DAOS engagement, to get customers up and running quickly.

*Table 10: 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 (for example xCAT/Confluence)
- Verify, and update if needed, firmware on the SR630 V2 servers
- Configure the network settings specific to the customer environment for:
  - XClarity Controller (XCC) service processors on the SR630 V2 servers
  - OS on the SR630 V2 servers
- Install OS and the HPC networking stack (e.g. Mellanox OFED) on the SR630 V2 servers
- Install and configure DAOS on the SR630 V2 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 11: Storage capacity measured in base-10 units and base-2 units.*

| Base-10 Units | | | | Base-2 Units | | | | Base-10 / Base-2 Ratio | | Base-2 / Base-10 Ratio | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| prefix | | value | | prefix | | value | | | | | |
| 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:

- Intel landing page for DAOS: https://www.intel.com/content/www/us/en/high-performance-computing/daos.html
- 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://docs.daos.io/v2.0/
- DAOS Community Home: https://daosio.atlassian.net/wiki
- DAOS mailing list: https://daos.groups.io/
- DAOS Slack channel: https://daos-stack.slack.com/

Lenovo Press Guides:

- Lenovo ThinkSystem SR630 V2 Server (Xeon SP Gen 3) https://lenovopress.com/lp1391
- Lenovo Scalable Infrastructure (LeSI) Solutions https://lenovopress.com/lp0900
- Intel Optane Persistent Memory 200 Series https://lenovopress.com/lp1380
- Introducing the Programming Model of Intel Optane DC Persistent Memory https://lenovopress.com/lp1194
- ThinkSystem Intel P5500 Entry NVMe PCIe 4.0 x4 SSDs https://lenovopress.com/lp1353
- ThinkSystem Intel P5600 Mainstream NVMe PCIe 4.0 x4 SSDs https://lenovopress.com/lp1354
- ThinkSystem Mellanox ConnectX-6 HDR/200GbE VPI Adapters https://lenovopress.com/lp1195
- Designing DAOS Storage Solutions with Lenovo ThinkSystem SR630 Servers https://lenovopress.com/lp1398

Lenovo ThinkSystem SR630 V2 (7Z71) Maintenance Manual and Setup Guide:

- https://thinksystem.lenovofiles.com/help/topic/7Z71/sr630_maintenance_manual.pdf
- https://thinksystem.lenovofiles.com/help/topic/7Z71/sr630_setup_guide.pdf

Intel Optane Persistent Memory Documentation:

- https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.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 200 Series modules

 

# About the Author

**Michael Hennecke** is a Principal Engineer for HPC Storage in the Intel DAOS engineering team. He previously was Lenovo's Chief Technologist for HPC Storage and Networking, and has over 28 years of experience in High Performance Computing and HPC Storage. Michael holds a master 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)
- Liang Zhen (Intel)
- Mohamad Chaarawi (Intel)
- Sigrun Eggerling (Lenovo)
- Florian Zillner (Lenovo)
- Martin Bachmaier (Lenovo)
- Wil Wellington (Lenovo)
- Taylor Allison (Lenovo)
- Nicolas Calimet (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.

 