Lenovo CAE Reference Architectures powered by Cornelis CN5000 Omni-Path®

for Computational Fluid Dynamics / Finite Elements Analysis

January 2026

Karsten Kutzer
Kevin Dean

Lenovo.com/systems

Contents

Introduction to Computer-Aided Engineering and HPC

Overview

This paper provides **Reference Architectures for High Performance Computing (HPC) systems**, optimized for Computer-Aided Engineering (CAE) Computational Fluid Dynamics

(CFD) and Finite Elements Analysis (FEA) workloads. They use Lenovo ThinkSystem Servers, integrated with the latest generation **Cornelis Networks CN5000 Omni-Path®** high-speed High Performance Fabric switches and SuperNICs. With its high-bandwidth, low-latency characteristics, Omni-Path is an ideal technology for CAE workloads and perfectly matches the Lenovo ThinkSystem high-performance servers.

This document is structured as follows:

1. This first section introduces **CAE and CFD / FEA** workloads and how they can benefit from HPC technology: Cornelis CN5000 Omni-Path High Performance Fabric as well as MRDIMM CPU memory.
2. In the second section of this document, the focus is on **CAE Reference Architectures for Small and Medium Businesses (SMB) customers**. It describes configurations in "T-shirt solution sizes" based on Lenovo air-cooled servers with Intel or AMD CPUs and a single Cornelis CN5000 Omni-Path High Performance Fabric switch.
3. The third section describes **CAE Reference Architectures for scale-out CFD** at large and hyperscale customers. Those are built for massive scale-out, using top-performing CPUs in highly energy efficient Direct-Water-Cooled (DWC) Lenovo ThinkSystem servers with Intel and AMD CPU technology and Cornelis CN5000 Omni-Path High Performance Fabric using multiple switches.

Computer-Aided Engineering (CAE) and Lenovo Reference Architectures

In today's marketplace, where innovation, speed, and precision define success, **Computer-Aided Engineering (CAE)** has become indispensable for modern product development. By simulating and analyzing the physical behavior of products and systems in a virtual environment - well before physical prototypes exist - CAE helps organizations cut development costs, accelerate time-to-market, and enhance product quality and performance.

From automotive and aerospace to energy, manufacturing, and consumer electronics, CAE empowers critical engineering decisions. It enables teams to explore design alternatives, validate performance under real-world conditions, and ensure compliance with safety and regulatory standards. As products grow more complex and expectations rise, the ability to virtually simulate and optimize designs has shifted from luxury to a competitive necessity.

Among the most powerful and widely adopted CAE technologies is **Computational Fluid Dynamics (CFD)**. CFD enables the simulation of fluid behavior - both liquids and gases - as they interact with surfaces and environments. It is indispensable in industries such as aerospace, automotive, energy, and electronics, where mastering airflow, heat transfer, and fluid dynamics is critical to product success. For example, CFD empowers engineers in evaluating airflow around aircraft wings or vehicle bodies for minimizing drag and boosting fuel efficiency. Other applications of CFD are enhancing cooling systems in electronics and power equipment or improving the safety and performance of pumps, turbines, and other fluid-handling machinery.

Finite Element Analysis (FEA) is another application of Computer-Aided Engineering, used to simulate how products respond to real-world forces such as stress, vibration, heat, and other physical effects. By breaking down complex geometries into smaller, manageable elements, FEA enables engineers to predict performance, identify weaknesses, and optimize designs before building prototypes.

There are two primary approaches to FEA: implicit and explicit.

- **Implicit FEA** solves equations using iterative methods that assume equilibrium at each time step, making it well-suited for static or slowly changing problems such as structural loading or thermal analysis.
- **Explicit FEA**, on the other hand, calculates responses directly at very tiny time increments without assuming equilibrium, which makes it ideal for highly dynamic, nonlinear events like crash simulations, impact analysis, or explosions.

Together, these methods allow engineers to address a wide spectrum of design challenges, from long-term durability to short-duration, high-intensity events.

CFD use cases

CFD enables engineers to:

- Analyze airflow over aircraft wings or vehicle bodies to reduce drag and improve fuel efficiency.
- Optimize cooling systems in electronics and power equipment.
- Simulate ventilation and air quality in buildings and industrial facilities.
- Improve the performance and safety of pumps, turbines, and other fluid-handling equipment.
- Support city planning by modelling urban heat effects.

However, CFD simulations often involve solving millions of equations to capture the complex physics of fluid or air motion. CFD workloads are both compute-intensive and memory-intensive, requiring significant processing power and high memory bandwidth to solve complex physical models. These simulations typically scale efficiently across many CPU cores and nodes, making them ideal for High Performance Computing (HPC) clusters where the infrastructure enables large-scale execution - reducing turnaround time, improving model accuracy, and supporting more robust design optimization. The foundational technology for enabling efficient large-scale out across multiple Compute Nodes is a high-speed interconnect, as provided by the **Cornelis CN5000 Omni-Path®** High Performance Fabric. The ability to increase complexity of the model at scale allows more factors to be considered - including macro factors affecting designs.

CFD Software Platforms

Leading CFD Software Platforms are

- **ANSYS® Fluent®**
  One of the most widely adopted CFD tools in the industry, Fluent® offers robust capabilities for simulating complex fluid flow, turbulence, heat transfer, and chemical reactions. It is used extensively in aerospace, automotive, energy, and electronics sectors for high-fidelity simulations and design optimization.
- **Siemens™ Simcenter™ STAR-CCM+™**
  STAR-CCM+™ is known for its integrated multiphysics capabilities, combining CFD with thermal, structural, and motion analysis. It is particularly valued for its automation, scalability, and ability to handle complex geometries and transient simulations, making it a strong choice for advanced engineering applications.
- **OpenFOAM®**
  An open-source CFD toolbox, OpenFOAM® is widely used in academia and industry for its flexibility and extensibility. It supports a wide range of solvers and physical models and is ideal for organizations looking to customize their simulation workflows or reduce licensing costs.

Most commercial CFD software from CAE ISVs is traditionally licensed based on the number of CPU cores used, which can constrain scalability and increase costs for large simulations. However, newer licensing models - such as STAR-CCM+'s Power Session and Fluent's HPC Ultimate - remove core count restrictions, allowing users to fully leverage high-core-count systems without incurring additional licensing fees. These models, along with open-source solvers like OpenFOAM, are reshaping computing strategies by encouraging the use of high-density CPUs

that prioritize total throughput over per-core efficiency. This shift opens new opportunities for maximizing simulation performance and return in investment modern CAE environments.

CAE workloads and how they benefit from HPC technology

Understanding the behavior of CAE workloads is important for creating an optimized High Performance Computing (HPC) cluster solution as the platform for running those workloads.

The following table provides an overview of the different characteristics of the CFD & Explicit FEA workloads compared to Implicit FEA workloads and the resulting requirements for HPC Compute Nodes:

Workload Type	Compute Profile	Best For	HPC Compute Node requirements
CFD & Explicit FEA	Compute- & memory-intensive, Highly parallel	Fluid flow, heat transfer, aerodynamics, Crash/impact/drop simulations	Balanced CPU, high memory bandwidth, High Performance Fabric
Implicit FEA	Solver-heavy, large matrix ops, high I/O Less parallel	Static stress, thermal, modal analysis	Lower core count high-frequency CPU, large memory, fast I/O

Table 1: CFD/Explicit FEA and Implicit FEA workload characteristics and requirements

In the following sub-sections, we describe how two High Performance Computing (HPC) technologies can positively impact CAE applications, especially in CFD:

- Cornelis CN5000 Omni-Path - a low-latency, high-bandwidth High Performance Fabric
- MRDIMM - a high-bandwidth memory technology

How CAE workloads benefit from Cornelis CN5000 Omni-Path Fabric

HPC workloads such as CAE/CFD depend on an interconnect network for running the simulations in parallel on and exchanging information between the Compute Nodes of the HPC solution. Key network features enabling rapid completion of CFD simulation are message rates, latency as well as network topology and flow optimization. As part of a Lenovo EveryScale solution, Cornelis CN5000 Omni-Path excels in providing superior message rates at ultra-low latency.

Cornelis CN5000 Omni-Path Fabric

The Cornelis CN5000 Omni-Path Fabric is ideal for interconnecting resources, using a scalable, high-speed, low-latency fabric, delivering an exceptional set of high-speed networking features and functions.

Figure 1: Cornelis Networks CN5000 Omni-Path

Cornelis Networks CN5000 Omni-Path is a complete, end-to-end 400Gb/s scale-out High Performance Fabric designed to accelerate tightly coupled HPC workloads such as CFD, from small departmental clusters to very large multi-rack systems. The solution spans the full fabric building blocks - PCIe Gen5 SuperNICs (single- and dual-port, air- or liquid-cooled), 48-port 1U switches, and a director-class 576-port switch platform for large, low-diameter topologies—plus qualified cabling options for dense deployment. Because CN5000 is delivered as an integrated “host-to-switch-to-fabric-management” stack, it is intended to provide predictable results and straightforward scaling as CFD problem sizes and node counts increase.

For CFD and other latency- and message-rate-sensitive MPI applications, CN5000 emphasizes lossless, congestion-free fabric behavior using credit-based flow control, advanced congestion management, and fine-grained adaptive routing guided by fabric telemetry - capabilities designed to preserve performance under real multi-job, high-utilization conditions. The SuperNIC is specified for up to 400Gb/s bandwidth with latency as low as <1µs, enabling strong halo-exchange and collective performance as solver partitions scale out. These architectural features translate into higher sustained application throughput (including reported CFD speedups versus alternative 400Gb/s fabrics), particularly as cluster size grows and communication overhead becomes dominant.

Ease of deployment and operational openness are addressed through the Cornelis OPX Software Suite, an open-source, OpenFabrics-based software stack that includes host drivers, fabric management, and monitoring/diagnostics, and is designed to integrate cleanly with common HPC environments. OPX provides Verbs compatibility and an OFI/libfabric path used broadly by modern MPI stacks and communication layers, helping CFD environments adopt CN5000 without requiring application rewrites.

Cornelis CN5000 is built on the Omni-Path Architecture to specifically optimize for HPC and AI performance. Key architectural pillars of Omni-Path include:

- **Credit-Based Flow Control**: Ensures packets are sent only when receive buffers have available credits, preventing overrun and eliminating pause frames. Credit-based flow control allows per-virtual-lane flow regulation, enabling fair allocation across competing workloads.
- **Enhanced Congestion Avoidance**: Uses forward and backward-propagating telemetry to dynamically adjust pacing at the source. This avoids the classic cascading congestion seen in oversubscribed topologies and maintains consistent throughput even under hot-spot pressure.
- **Fine-Grained Adaptive Routing**: Real-time telemetry builds a global heatmap of switch buffer utilization, enabling path selection based on current network state rather than static tables. It automatically bypasses congested paths, reducing long-tail latency and improving collective efficiency.
- **Dynamic Lane Scaling and Link-Level Replay**: Each link can degrade gracefully if individual lanes fail, maintaining connectivity instead of triggering a full path teardown. Link-level replay catches forward-error-correction (FEC) misses at the hardware layer, eliminating costly end-to-end retransmissions

- **Open and Interoperable:** Open software stack (OPX) built on OFI/libfabric, ensuring low-overhead integration with MPI stacks (for example, Open MPI), and ease of deployment in heterogeneous environments.

Scaling behavior of CFD applications on the CN5000 Omni-Path High Performance Fabric

Maintaining performance on distributed infrastructure is critical for meeting the growing demands of complex CFD simulations. Strong scaling efficiency is the key metric, ensuring that investments in additional Compute Nodes translate directly into faster results rather than diminishing returns from network bottlenecks. CN5000 delivers exceptionally strong and consistent scaling across a number of ANSYS models, showing great efficiency from one to eight nodes.

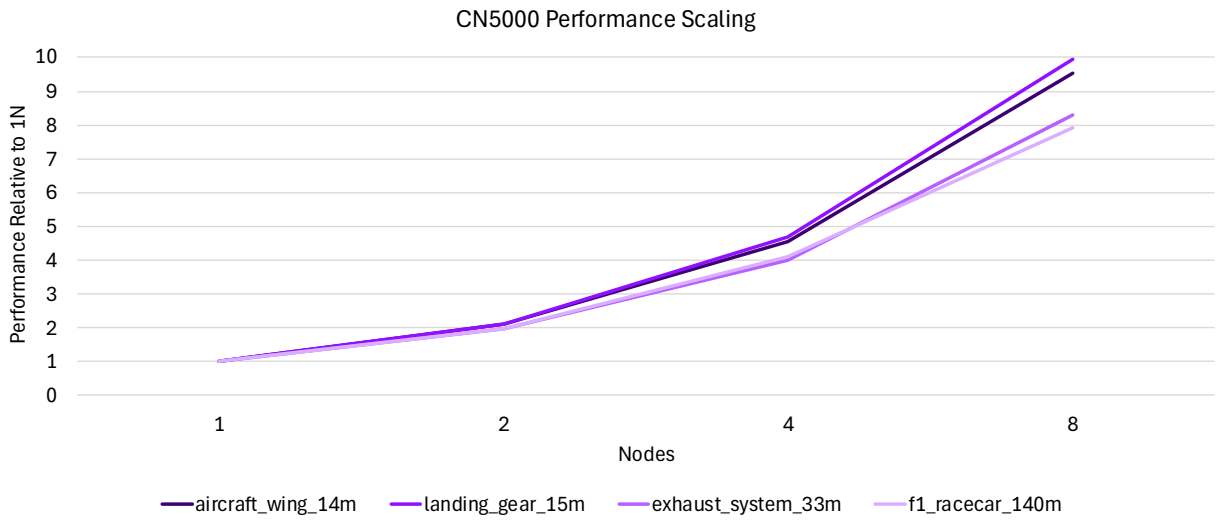

Figure 2: CN5000 Performance Scaling

This scaling behavior makes CN5000 Omni-Path an excellent High Performance Fabric technology for scaling-out CAE workloads to a larger number of nodes.

Conclusion

In summary, the CN5000 Omni-Path Fabric offers many advantages as part of Lenovo EveryScale CAE reference architectures:

- Enables faster simulation runtimes on CFD workloads, resulting in more simulations completed in the same time frame.
- Provides excellent scaling efficiency across nodes, which keeps performance growing closer to linearly as more servers are added, which is critical for large models.
- Sustains a high MPI message rate, which benefits CAE solvers that exchange many small messages between nodes. This improves the solver's ability to keep all CPUs busy.
- Reduces end-to-end latency, helping to reduce idle wait times between Compute Nodes, which is an important factor for tightly coupled solvers in CFD and similar CAE tasks.

How CAE workloads benefit from MRDIMM technology

Memory-bound vs. compute-bound CFD workloads

The performance characteristics of CAE and especially CFD workloads are typically balanced between compute-intensive and memory-intensive demands. Compute-intensive workloads benefit from a high number of processor cores, elevated CPU frequencies, and increased instructions per cycle (IPC), which accelerate the execution of scalar and vector operations. In contrast, memory-intensive workloads depend heavily on high memory bandwidth and low latency, as performance is closely tied to the speed at which data can be read from and written to memory. Additionally, larger CPU caches can help mitigate memory bottlenecks by reducing the frequency of main memory accesses.

The degree to which a CFD workload leans toward compute or memory intensity depends on several factors, including the specific software application, the resolution of the simulation model, and the complexity of the physics being modeled. Higher-resolution models—which involve finer meshes and more detailed physics—tend to be more memory-bound, requiring greater memory bandwidth to maintain simulation efficiency. Conversely, lower-resolution models and simulations with chemical reactions such as combustion models are often more compute-bound and benefit from CPUs with higher clock speeds and strong single-threaded performance.

MRDIMM technology introduction

MRDIMM technology, or multiplexed rank DIMMs, represents a significant advancement in memory performance, particularly for memory-intensive workloads such as CFD. Companies like Micron have been at the forefront of developing high-speed memory solutions, contributing to the reliability and efficiency of MRDIMMs.

The Lenovo ThinkSystem SC750 V4, as used in the Lenovo Reference Architecture for scale-out CAE using Intel-CPUs, is Lenovo's first system to support this innovative technology and is one of the few systems in the market which can leverage the full potential of MRDIMM technology.

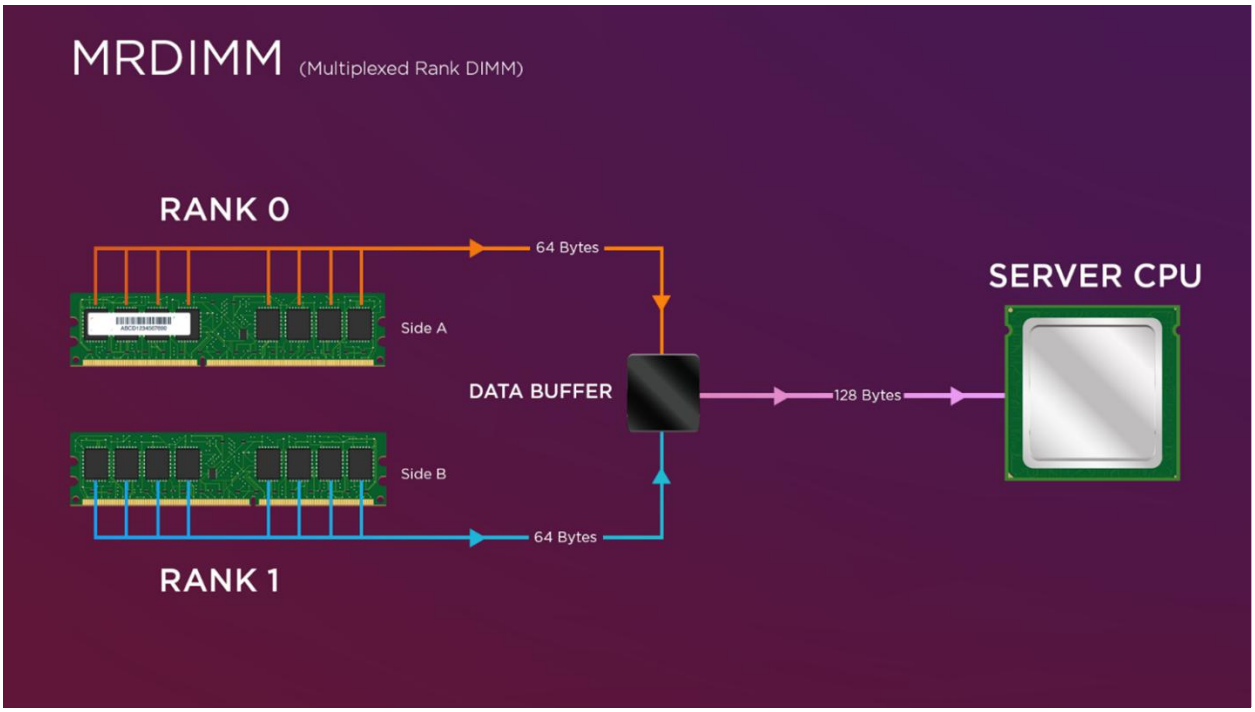

Figure 3: MRDIMM Multiplex Functionality

Operating at speeds of up to 8800 MT/s, MRDIMMs significantly reduce memory access latency and increase bandwidth, especially when paired with Intel Xeon 6th Gen processors. This combination has demonstrated up to a 200% performance improvement over the previous Xeon

generation, making MRDIMMs a critical enabler for modern HPC environments. These benefits are particularly impactful for computationally intensive workloads like CFD, where memory speed and efficiency directly influence simulation performance. In real-world testing, CFD applications such as Fluent, STAR-CCM+, and OpenFOAM have shown an average of 1.2× faster computational times with MRDIMMs—highlighting the tangible value of high-bandwidth memory in accelerating engineering workflows.

Methodology and results

The following benchmark results highlight the performance impact of MRDIMMs across three leading CFD applications: ANSYS Fluent, Siemens Simcenter STAR-CCM+, and OpenFOAM. A diverse set of simulation models was selected, spanning from low-resolution models – such as Fluent's Aircraft Wing model with 14 million cells (under 20 million cells) to high-resolution models exceeding 100 million cells, including STAR-CCM+'s VTM Bench model with 178 million cells. These benchmarks also included complex flow scenarios, such as thermal management simulations, to reflect real-world engineering challenges.

This range of model sizes and physics ensures a representative mix of workloads relevant to the broader engineering community. When averaged across all three applications and workload types, the Intel Xeon 6900-Series with MRDIMMs delivers a 2.4x overall performance improvement compared to the previous generation—highlighting its significant impact in accelerating CFD simulations across a broad range of use cases.

The chart below compares the performance of a 2-socket Compute Node featuring Intel 8592+ CPUs (Xeon 5) against a node equipped with Intel 6980P CPUs (Xeon 6) and MRDIMMs. It highlights relative performance across a range of Fluent, STAR-CCM+, and OpenFOAM workloads as model resolution increases. Performance is shown relative to the Intel 8592+ baseline, illustrating the gains achieved with the newer Xeon 6 architecture.

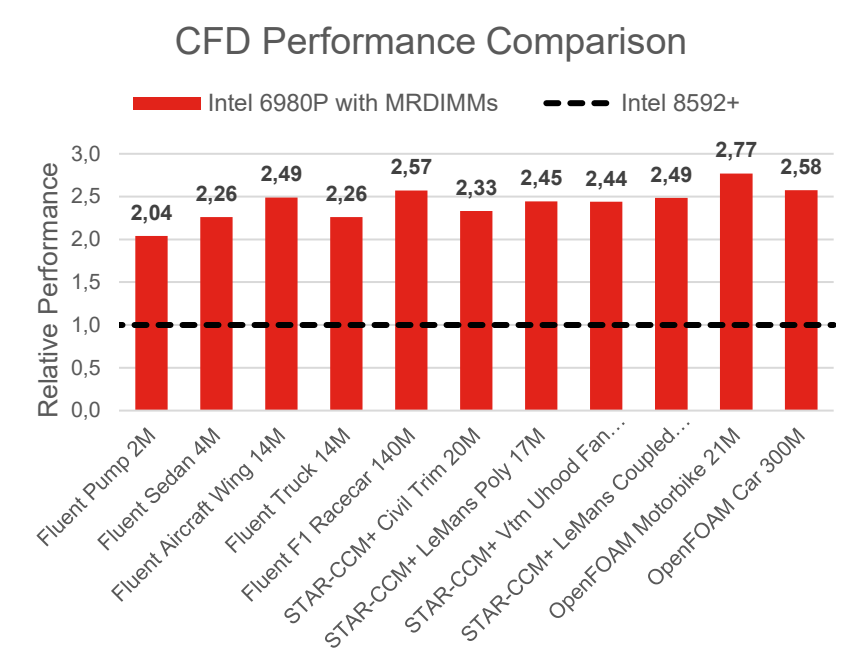

Figure 4: CFD Application Performance Comparison

Both memory bandwidth and compute power are key drivers of CFD workload performance. While it may be tempting to attribute the performance gains shown above solely to generational CPU improvements, further analysis reveals a more nuanced picture. When using the same Intel Xeon 6980P CPU configured with both MRDIMMs and standard DDR5 RDIMMs, the results indicate that the observed performance improvements are not solely due to CPU advancements. Servers with MRDIMMs deliver an additional performance improvement of 1.2× over DDR5 RDIMMs.

The chart below compares the performance of two 2-socket Compute Nodes, both featuring Intel 6980P CPUs, one configured with DDR5 RDIMMs and the other with MRDIMMs. It highlights relative performance across a range of Fluent, STAR-CCM+, and OpenFOAM workloads as model resolution increases. Performance is shown relative to the RDIMM-based node, illustrating the performance gains enabled by the MRDIMM memory technology.

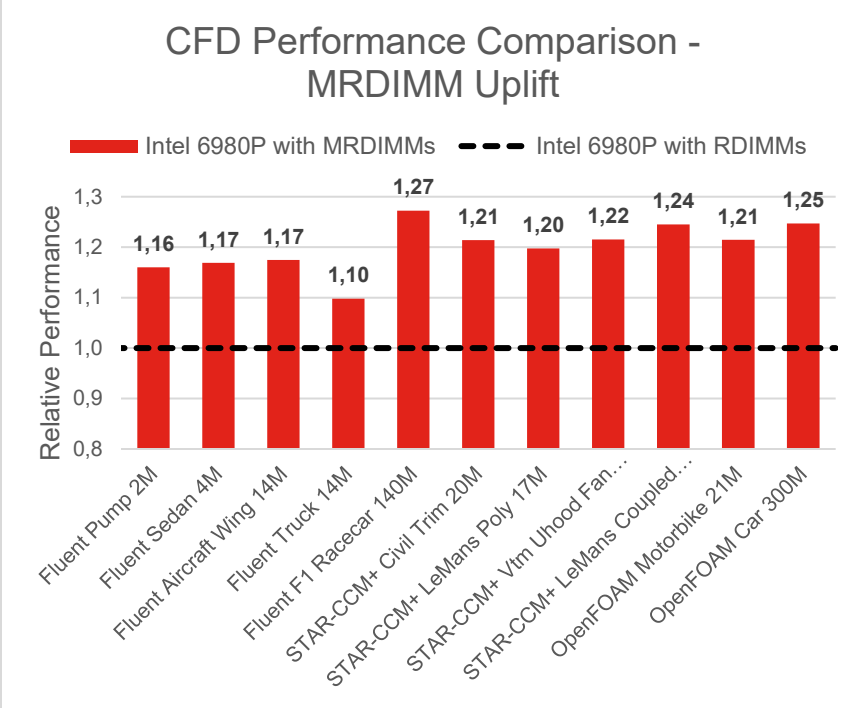

Figure 5: CFD Application MRDIMM Uplift

The results presented above are derived from benchmark runs by Lenovo HPC Innovation Center using the Lenovo ThinkSystem SC750 V4 with Intel Xeon 6 CPUs and Micron MRDIMM memory.

Conclusion

We found that Intel Xeon 6900-Series with MRDIMMs delivers a 2.4x overall performance improvement compared to the previous generation. Within that CPU generation, servers with MRDIMMs deliver an additional performance improvement of 1.2× over DDR5 RDIMMs.

Designing and delivering an HPC solution for CAE with Lenovo EveryScale

The increasing sophistication of the simulations - often involving millions of variables and intricate physical interactions - requires immense computational power. This is where **High-Performance Computing (HPC)** becomes a game-changer. HPC platforms deliver the speed and scale needed

to run large, high-fidelity simulations efficiently. With HPC, engineering teams can evaluate more design options, build richer models, and make faster, data-driven decisions that drive innovation forward.

High Performance Computing uses **Clusters** of Servers, which are also known as **Compute Nodes**. The Compute Nodes are interconnected with a fast, high-bandwidth, low-latency network called the **High Performance Fabric**, and accessed and managed through servers known as **Login and Management Nodes** or **Head Nodes** as well as **Ethernet Management Networks**.

The design, manufacturing, integration, delivery and installation of such a HPC Cluster is a complex task, and Lenovo has extensive experience and knowledge of how to manage such a project with customers.

For the initial HPC cluster design phase, the **Lenovo Reference Architectures** as described in this document provide pre-defined and optimized Building Blocks or Scalable Units (SUs) for CAE solutions. They were created by Lenovo Subject Matter Experts and Performance Engineers and deliver the compute power, memory bandwidth, and High Performance Fabric performance required to run CAE workloads efficiently and at scale. Application performance improvements and scalability have been validated by Lenovo's expert performance engineers using production level hardware in the Lenovo HPC Innovation Center.

The **Lenovo EveryScale** solution is a way to simplify design, delivery, installation and implementation of HPC cluster solutions. Lenovo Server and Networking components and Operating System come together as a single Lenovo EveryScale Solution. Lenovo EveryScale provides Best Recipe guides to warrant interoperability of hardware, software and firmware among a variety of Lenovo and third-party components.

Addressing specific needs in the data center, while also optimizing the solution design for application performance, requires a significant level of effort and expertise. Customers need to choose the right hardware and software components, solve interoperability challenges across multiple vendors, and determine optimal firmware levels across the entire solution to ensure operational excellence, maximize performance, and drive best total cost of ownership.

Lenovo EveryScale reduces this burden on the customer by pre-testing and validating a large selection of Lenovo and third-party components, for creating a Best Recipe of components and

firmware levels that work seamlessly together as a solution. From this testing, customers can be confident that such a best practice solution will run optimally for their workloads, tailored to the client's needs.

In addition to interoperability testing, Lenovo EveryScale hardware is pre-integrated, pre-cabled, pre-loaded with the best recipe and optionally an OS-image and tested at the rack level in manufacturing, to ensure a reliable delivery and minimize installation time in the customer data center.

CAE Reference Architectures for SMB customers

Overview

The Reference Architectures for Small and Medium Business (SMB) customers provide configuration templates for Small (S), Medium (M), Large (L) and Extra-Large (X-L) system sizes, based on common server and networking components. Those T-shirt sizes are intended as initial guidelines and can easily be adjusted to a specific customer situation by adding or removing components. Reference Architectures are provided for servers using Intel as well as AMD CPUs.

The target audience for this Reference Architectures are cost-sensitive environments with only a few dedicated and named users. They generally do not separate Login and Management systems but rather combine them in a single a Head Node serving both purposes. Security and access to the cluster is usually controlled externally (e.g. by using firewalls). If desired, additional Head Nodes allow us to separate User Login from Administrator Management, as outlined for the X-Large configurations.

A Storage Server with internal NVMe drives serving as NFS (Network File System) server is suggested for this Reference Architectures. Other shared storage options like parallel file systems can be considered for higher performance demands.

The following picture is an architectural diagram of the CFD Reference Architecture for SMB customers.

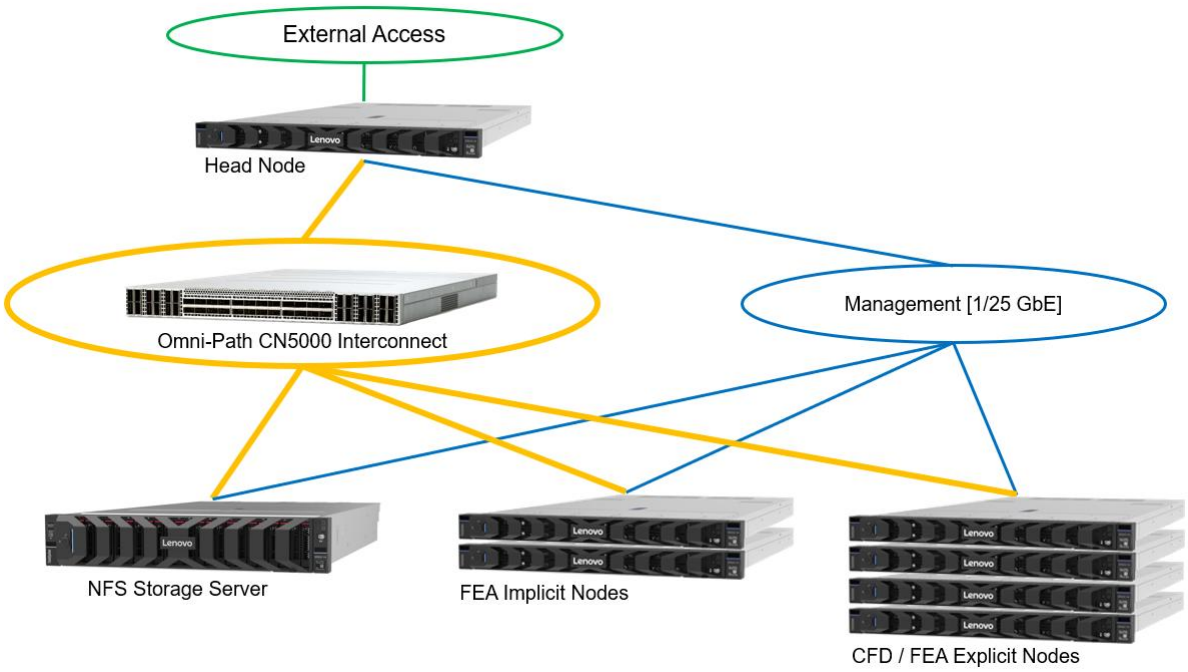

Figure 6 –CFD Reference Architecture for SMB with Omni-Path High Performance Fabric

The different server types are referenced here as “nodes” – a common terminology in High Performance Computing (HPC).

- The **Head Node** is the single point of access to the CFD HPC cluster for users as well as the administrators. The Head Node is connected to the external access network (green), as well as the High Performance Fabric (orange) and the Management Network (blue). Besides the cluster management (e.g. Lenovo Confluent) and resource management/scheduling (e.g. Slurm) software, it also runs the Omni-Path Fabric Manager, which is managing the Omni-Path fabric. Users start and monitor their CAE jobs from this node.
- The **NFS Storage Server** provides access to a shared NFS file system, which resides on internal NVMe drives. The NFS file system provides a consistent view on the data for all nodes in the system. The NFS file systems are exported over the high-performance Omni-Path High Performance Fabric to all other nodes using IPoIB (IP over IB) protocol.
- The **CFD / FEA Explicit Compute Nodes** are used for running CFD and/or FEA Explicit workloads. They have a huge number of decent frequency cores, a decent amount of memory and are

connected to the Omni-Path High Performance Fabric as well as to a Management Network. The Compute Nodes do not have internal drives for the operating system. Instead, the nodes are booted from the Head Node over the Management Network using the Confluent software. This way, the Compute Node OS images can be kept consistent across the Compute Nodes easily.

- The **FEA Implicit Compute Nodes** are used for running FEA Implicit workloads. They have a lower number of higher frequency cores, a huge amount of memory and are connected to the Omni-Path High Performance Fabric as well as to a Management Network. The Compute Nodes do not have internal drives for the operating system. Instead, the nodes are booted from the Head Node over the Management Network using the Confluent software. This way, the Compute Node OS images can be kept consistent across the Compute Nodes easily. There is an internal NVMe drive though, which is used for local scratch storage and supports the I/O heavy FEA Implicit jobs while they are running. The input data as well as results though must be stored on the NFS file systems mounted from the NFS Storage Server to guarantee data persistence across re-boots of the Compute Nodes and sharing between different Compute Nodes.
- The **Omni-Path High Performance Fabric** consists of a Cornelis CN5000 Switch, connected to Cornelis CN5000 SuperNICs in the nodes with high-speed cables. The High Performance Fabric provides 48 ports with a bandwidth of 400 Gbit/s to each server-side adapter. Optionally, a higher number of adapters can be connected using splitter-cables, which connect a 400 Gbit/s port on the CN5000 Switch to two CN5000 SuperNICs at 200 Gbit/s speed each. Up to 32 of the 48 ports can be used with that split-out type of cable, while the remaining 16 ports only support straight 400 Gbit/s cables. That allows for a variety of networking options, optimally adjusted to specific use cases.
- The **Management Network** was usually provided by a 1 GbE (Gigabit Ethernet) switch. Since the cost for 25 GbE network switches and adapters decreased, the Management Network is trending towards that speed which provides a better user experience when installing systems or for other workloads running on the Ethernet network. In this Reference Architecture we are suggesting a 25 GbE switch for the Management Network.
- The **External Access Network** is provided by the customer network infrastructure and provides connectivity to the CAE HPC cluster via the Head Node. For the Reference Architecture, we assume 25 GbE connections for this network.

Building Block components for CAE for SMB Reference Architecture

The following table describes the configuration of the Server Building Blocks and switches used in the CAE Reference Architecture for SMB:

Component	Intel-based System	AMD-based system
Compute Nodes **CFD &** **Explicit FEA**	ThinkSystem SR630 V4 1U dual-socket rack server • 2 * CPUs Intel Xeon 6747P (48 core, 2.7 GHz, 330W) • 512 GB Memory (16 * 32GB DDR5-6400 dual-rank) • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP	ThinkSystem SR645 V3 1U dual-socket rack server • 2 * CPUs AMD EPYC 9455 (48 core, 3.15 GHz, 300W) • 768 GB Memory (24 * 32GB DDR5-6400 dual-rank) • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP
Compute Nodes **Implicit FEA**	ThinkSystem SR630 V4 1U dual-socket rack server • 2 * CPUs Intel Xeon 6724P (16 core, 3.6 GHz , 210W) • 1024 GB Memory (16x 64GB DDR5-6400 dual-rank) • Local Scratch space; 1 * 960GB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP	ThinkSystem SR645 V3 1U dual-socket rack server • 2 * CPUs AMD EPYC 9135 (16 core, 3.65 GHz, 240W cTDP) • 1536 GB Memory (24x 64GB DDR5-6400 dual-rank) • Local Scratch space: 1 * 960GB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP
Head Node **(for size Small, this node is also used for NFS Storage instead of a dedicated NFS Storage Server)**	ThinkSystem SR630 V4 1U dual-socket rack server • 2 * CPUs Intel Xeon 6724P (16 core, 3.6 GHz , 210W) • 1024 GB Memory (16 * 64GB DDR5-6400 dual-rank) • Boot/OS drives (RAID1): 2 * 480GB Read-Intensive SSD • NFS storage (RAID5; for size Small): 5 * 7.68TB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP	ThinkSystem SR645 V3 1U dual-socket rack server • 2 * CPUs AMD EPYC 9135 (16 core, 3.65 GHz, 240W cTDP) • 1536 GB Memory (24 * 64GB DDR5-6400 dual-rank) • Boot/OS drives (RAID1): 2 * 480GB Read-Intensive SSD • NFS storage (RAID5; for size Small): 5 * 7.68TB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP

	• 1 * 10/25GbE SFP28 2-Port PCIe	• 1 * 10/25GbE SFP28 2-Port PCIe
NFS Storage Server (only for Medium/Large configurations)	ThinkSystem SR650 V4 2U dual-socket rack server • 2 * CPUs Intel Xeon 6507P (8 core, 3.5 GHz , 150W) or • 256 GB Memory (16x 16GB DDR5-6400 single-rank) • Boot/OS drives (RAID1): 2 * 480GB Read-Intensive SSD • NFS storage (RAID5): ~53 TB 8 * 7.68TB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP	ThinkSystem SR645 V3 1U dual-socket rack server • 2 * CPUs AMD EPYC 9015 (8 core, 3.6 GHz, 125W) • 384 GB Memory (24x 16GB DDR5-6400 single-rank) • Boot/OS drives (RAID1): 2 * 480GB Read-Intensive SSD • NFS storage (RAID5): ~53 TB 8 * 7.68TB Read-Intensive NVMe • 1 * Cornelis CN5000 Omni-Path SuperNIC • 1 * 10/25GbE SFP28 2-Port OCP
Networking	• Management Network: 1 * 10/25 GbE switch • High Performance Fabric: 1 *Cornelis CN5000 Omni-Path Switch	
Software Stack	• Operating System: Rocky Linux 9 • Job Scheduler: Slurm • Shared file system: NFS • Cluster Management: Confluent	

Table 2: CAE Reference Architecture for SMB: Component Specifications

Scaling the Reference Architecture in T-shirt sized system configurations

The suggested T-shirt sizes provide a first indication for sizing a system. The Small configuration is providing a cost-effective but minimal functional HPC cluster with the Head Node also providing the NFS storage, while the Medium and Large sizes scale out the number of Compute Nodes (CFD/Explicit FEA as well as Implicit FEA) and have a dedicated NFS Storage Server. The Reference Architecture for SMB customers generally uses a 400 Gbit/s connection for each node to the Omni-Path High Performance Fabric.

The X-Large configuration is an example for the maximum system size that can be deployed with a single Cornelis CN5000 Omni-Path Switch and deviates from the minimized infrastructure as provided with the Small, Medium and Large configurations. It uses four Head nodes, two used as

redundant Management Nodes and two used as redundant Login Nodes. The four NFS storage servers, each with a 400 Gbit/s connection to the Omni-Path High Performance Fabric, may be considered a placeholder for a larger storage system – four individual NFS storage servers require the administrator to split-up the storage into four separate mount-points, which may not be convenient for the users and harder to manage. Other storage system options though are beyond the scope of this paper.

There are two variants provided for the X-Large configuration: X-Large is using 400 Gbit/s Omni-Path connections to all nodes, while X-Large200 is using 200 Gbit/s Omni-Path for the CFD & Explicit FEA Compute Nodes. This allows the system to scale to an even larger number of Compute Nodes with a single CN5000 Omni-Path Switch. Generally, cutting the available bandwidth by 50% to 200 Gbit/s is not a huge issue for CFD performance – the low latency of the High Performance Fabric is more important than providing the full bandwidth.

As the CN5000 Switches can be deployed in a fabric with different topologies (e.g. fat-tree), much larger deployments are possible – those topologies are described in the next section with scale-out CAE clusters. Those concepts can be applied to this architecture as well for XX-Large or XXX-Large configurations.

The following table shows how the different T-shirt @sizes are created from the building blocks described before.

	Small	Medium	Large	X-Large	X-Large200
Estimated number of users	2-4	2-6	4-8	8	8+
CFD/ Explicit FEA Nodes	4	6	16	34	64 (200Gbit/s)
Implicit FEA Nodes	1	1	2	6	8
Head Nodes	1	1	1	4	4
NFS storage server	(shared)	1	1	4	4

Storage capacity (TB)	~30	~53	~53	~212	~212

Table 3: CAE Reference Architecture T-shirt size overview

Network connections of the components

Omni-Path connections

All servers in the cluster connect to the Cornelis Omni-Path High Performance Fabric through a CN5000 Omni-Path SuperNIC. The Omni-Path fabric speeds up CAE applications by

- Providing low-latency communication between Compute Nodes as the workloads scale-out and job sizes increase. This is especially important for CFD / Explicit FEA applications
- Providing high-bandwidth I/O access to the shared NFS storage using IB over IP protocols. This is especially important for Implicit FEA applications.

The Head Node provides Omni-Path fabric management capabilities through a CN5000 SuperNIC using the Omni-Path Fabric Manager software, which is part of the Cornelis CN5000 OPX Software.

As there are sufficient ports available ton the CN5000 1U 48-port Switch and the CN5000 SuperNIC supports the full speed of 400 Gbit/s, we recommend using **CN5000 400G QSFP112 DAC Passive-Straight** cables for all node in the T-shirt sizes Small, Medium and Large reference architectures.

In the X-Large200 configuration, we use **400G QSFP- 2x200G QSFP56 DAC Passive-Split** cables for the CFD / Explicit FEA nodes, which allows connection of up to 64 of those nodes per CN5000 Omni-Path Switch at 200 Gbit/s bandwidth, in addition to 16 additional servers using the Passive-Straight cable providing 400 Gbit/s of bandwidth. This results in scaling the CAE cluster to a total of 72 HPC Compute Nodes on a single switch with additional ports being available for Storage and Head Nodes.

Management network connections

The management network serves a number of important purposes in the CAE HPC cluster:

- **Hardware management**: The Head Node (in the role as Administrator Management Node) uses one of his 25GbE Ethernet interfaces for connecting to the Baseboard Management Controllers (BMCs) in the other managed servers through a network port on their internal OCP adapter. This port is shared between the BMC and the operating system on that server using Network Controller Sideband Interface (NC-SI) protocol. In Lenovo servers, the BMC is known as xClarity Controller (XCC). Through this network, the administrator can perform management tasks at the hardware level, like power on/off the servers, open a serial console to the server, update and manage firmware and retrieve telemetry data from the server like temperatures or power consumption. The Lenovo developed Open-Source management software "Confluent" makes use of that network connection and provides a convenient way for managing the whole CAE HPC cluster.
- **Operating system boot**: The managed systems boot their operating system from the Head Node either for initial installation (for nodes with internal OS drives as the NFS storage server or potential additional Head Nodes, e.g. used in the role as User Login system), or for general operation (for nodes without internal OS drives, which are the CFD / Explicit FEA and Implicit FEA Compute Nodes). The Lenovo Confluent software provides tools for supporting and managing the network boot of the operating system in various ways.
- **Node management**: The management network is also used for managing the cluster at the operating system level. The Lenovo Confluent management software helps with setting up secure connections between the Head Node and the managed operating systems on the nodes and provides tools for parallel management of multiple or even all nodes from a single command line.
- **Infrastructure management**: The CN5000 Omni-Path Switch provides a 1GbE management port that can also be added to the Management Network, e.g. for retrieving switch telemetry data (the Fabric Management itself though is using the CN5000 SuperNIC on the Head Node). Also managed PDUs may be connected to the Management Network for power control and monitoring. As we use a 25GbE switch with SFP28 ports for the Management Network, RJ45 to SFP+ adapters are needed for connecting the RJ45 connections of the 1GbE cables to the switch ports.

CAE Reference Architecture for SMB using Intel-CPUs

Server component description

Lenovo ThinkSystem SR630 V4

Lenovo ThinkSystem SR630 V4 air-cooled 1U servers are used as Head Nodes in a User Login and/or Administrator Management role as well as HPC Compute Nodes in the CAE Reference Architectures for SMB using Intel-CPU based servers.

The Lenovo ThinkSystem SR630 V4 is a 2-socket 1U rack server providing industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. With two Intel Xeon 6700-series processors, the SR630 V4 is designed for high density and scale-out workloads in various customer segments.

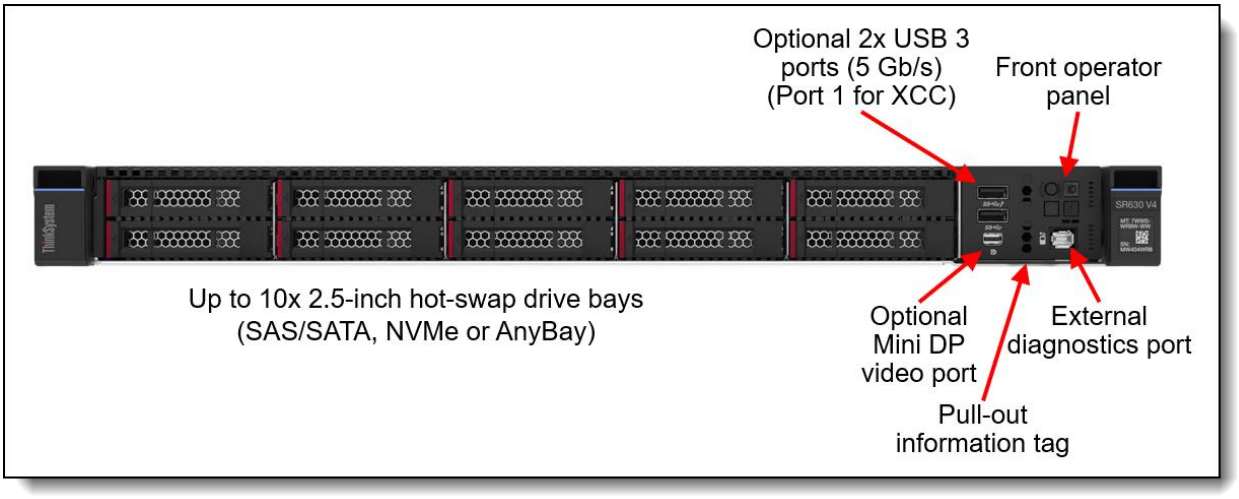

Figure 7: Front view of the SR630 V4 with 2.5-inch drive bays

It provides several different configuration options. In the Reference Architectures we use a chassis supporting up to 10x 2.5-inch hot-swap drives on the front.

At the back side, the SR630 V4 provides external connectivity through up to 3 PCIe (1 Full Height and 2 Low Profile) slots and 2 OCP slots.

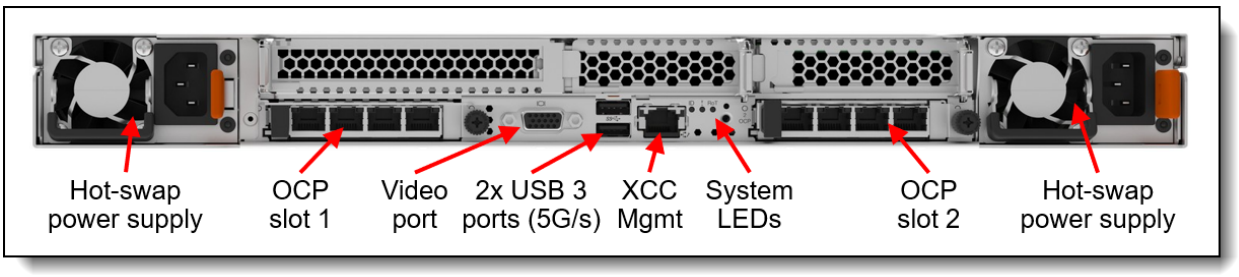

Figure 8: Rear view of the SR630 V4 with three PCIe slots

The SR630 V4 product guide, available at Lenovopress, provides a more detailed description of the server. It can be found at the following link:

https://lenovopress.lenovo.com/lp1971-thinksystem-sr630-v4-server

Lenovo ThinkSystem SR650 V4

Lenovo ThinkSystem SR650 V4 air-cooled 2U servers are used as Storage Nodes in the CAE Reference Architectures for SMB using Intel-CPU based servers. They are also an option, if Head Nodes require additional networking connectivity, as they provide additional PCIe slots compared to the 1U servers.

The Lenovo ThinkSystem SR650 V4 is a 2-socket 2U rack server providing industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. With two Intel Xeon 6700-series, the SR650 V4 is designed for high density and scale-out workloads.

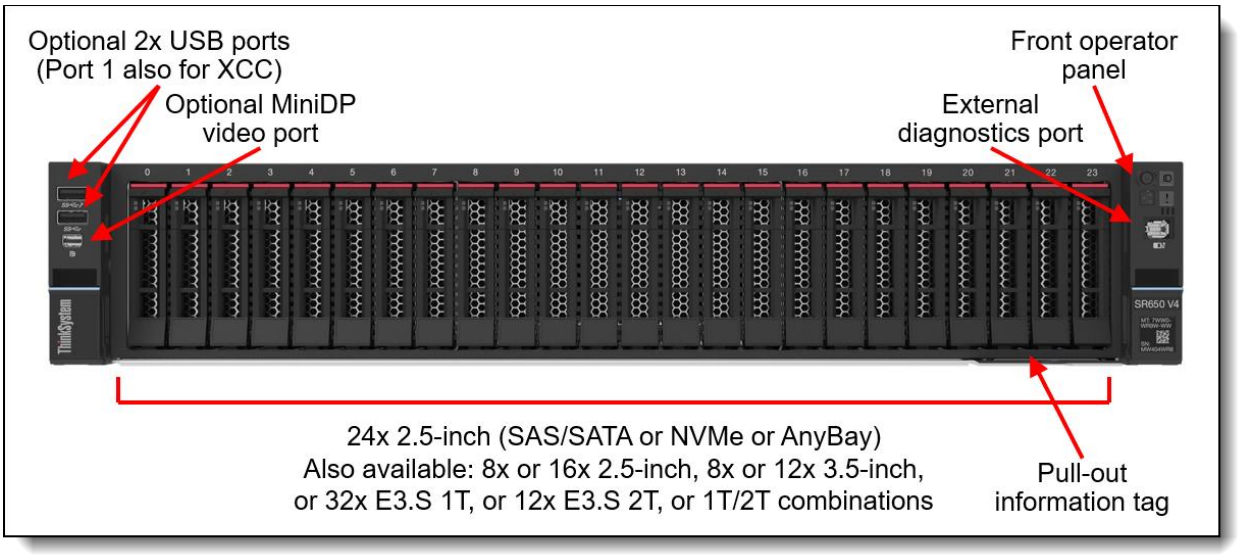

Figure 9: Front view of the ThinkSystem SR650 V4 with 2.5-inch drive bays

It provides several different configuration options. In the Reference Architectures we use a chassis with a NVMe backplane supporting 8 NVMe drives, connected to an internal HW adapter supporting Intel VROC RAID.

The SR650 V4 provides external connectivity through up to 10 PCIe slots and 2 OCP slots at the back side. In the Reference Architecture, less PCIe slots are needed and hence the number of PCIe slots provided is below the maximum.

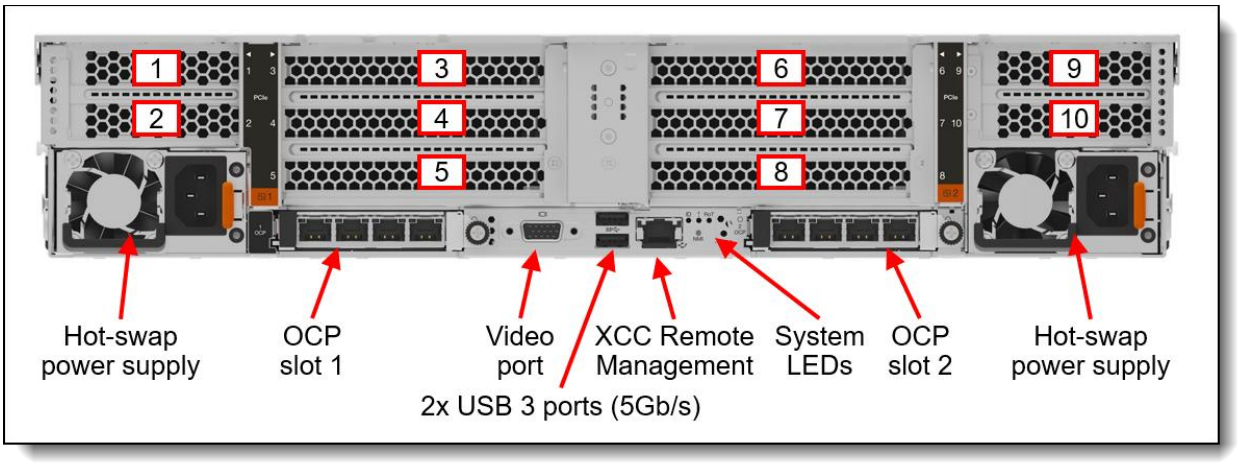

Figure 10: Rear view of the ThinkSystem SR650 V4 (configuration with ten PCIe slots)

The SR650 V4 product guide, available at Lenovopress, provides a more detailed description of the server. It can be found at the following link:

https://lenovopress.lenovo.com/lp2127-thinksystem-sr650-v4-server

Bill of Materials (BOM) for Intel server-based CAE RA for SMB customers

The following Bill of Materials (BOM) include only the significant parts – other parts like Risers or Power supplies have been removed for better readability. The Lenovo sales configurators DCSC and x-config help to create a valid configuration from the following BOM lists.

Racks and Power Distribution Units are not included in the BOM assuming those are provided by the customer separately.

Please adjust the BOM to your specific situation as needed (e.g. cable lengths).

Part Number	Product Description	Small	Medium	Large	X-Large	X-Large200
	CFD-Explicit-FEA	**4**	**8**	**16**	**34**	**64**
7DG9CTO1WW	**ThinkSystem SR630 V4 - 3yr Warranty**	2	2	2	2	2

C5R8	Intel Xeon 6747P 48C 330W 2.7GHz Processor	2	2	2	2	2
BYTJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM	16	16	16	16	16
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
	Implicit-FEA	**1**	**1**	**2**	**6**	**8**
7DG9CTO1WW	**ThinkSystem SR630 V4 - 3yr Warranty**					
C5R5	Intel Xeon 6724P 16C 210W 3.6GHz Processor	2	2	2	2	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	16	16	16	16	16
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
C0GK	ThinkSystem 2.5" U.2 PM9D3a 960GB Read Intensive NVMe PCIe 5.0 x4 HS SSD	1	1	1	1	1
	HeadNode (indl. Storage for Small/XtraSmall)	**1**	**1**	**1**	**4**	**4**
7DG9CTO1WW	**ThinkSystem SR630 V4 - 3yr Warranty**					
C5R5	Intel Xeon 6724P 16C 210W 3.6GHz Processor	2	2	2	2	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	16	16	16	16	16
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1
BCD6	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port PCIe Ethernet Adapter	1	1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter	1	1	1	1	1
CBSZ	ThinkSystem M.2 VA 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2	2	2	2	2
B96G	Intel VROC (VMD NVMe RAID) Premium	1				
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	5				
	Storage		**1**	**1**	**4**	**4**
7DGDCTO1WW	**ThinkSystem SR650 V4 - 3yr Warranty**					
C5R6	Intel Xeon 6507P 8C 150W 3.5GHz Processor		2	2	2	2
C0U2	ThinkSystem 16GB TruDDR5 6400MHz (1Rx8) RDIMM		16	16	16	16

BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter		1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter(Generic FW)		1	1	1	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter		1	1	1	1
CBSZ	ThinkSystem M.2 VA 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD		2	2	2	2
B96G	Intel VROC (VMD NVMe RAID) Premium		1	1	1	1
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD		8	8	8	8
	25G/100G Main	**1**	**1**	**1**	**2**	**2**
7D5FCTOKWW	**NVIDIA SN3420 25GbE Managed Switch with Cumulus (PSE)**					
	OPA-Main	**1**	**1**	**1**	**1**	**1**
7DMQCTO1WW	**Cornelis CN5000 48-Port Omni-Path 400GbE Air-Cooled Switch PSE**					
	ClientRack	**1**	**1**	**1**	**2**	**3**
7X74CTO1WW	**Lenovo EveryScale 42U Client Site Integration Kit**					
	Software					
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	6	11	20	48	80
	Cables					
AV1W	Lenovo 1m Passive 25G SFP28 DAC Cable				10	12
AV1F	Lenovo 3m 25G SFP28 Active Optical Cable	6	11	20	21	21
AV1G	Lenovo 5m 25G SFP28 Active Optical Cable				18	48
CBPM	Cornelis 1m CN5000 400G QSFP112 DAC Passive Cable-Straight				7	9
CBPN	Cornelis 1.5m CN5000 400G QSFP112 DAC Passive Cable-Straight			5	11	3
CBPP	Cornelis 2m CN5000 400G QSFP112 DAC Passive Cable-Straight	2	7	11	9	1
CDQW	Cornelis 3m CN5000 400G QSFP112 DAC Passive Cable-Straight	4	4	4	3	3
CBPQ	Cornelis 5m CN5000 400G QSFP112 ACC Active Cable-Straight				18	
CDHC	Cornelis 1m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					
CDHB	Cornelis 1.5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					4

CDHA	Cornelis 2m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					4
CDHH	Cornelis 3m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					
CDHG	Cornelis 5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					24

Table 4: Bill of Materials (BOM) for Intel server-based CAE solution for SMB customers

CAE Reference Architecture for SMB using AMD-CPUs

Server component details

Lenovo ThinkSystem SR645 V3

Lenovo ThinkSystem SR645 V3 air-cooled 1U servers are used as Head Nodes in a User Login and/or Administrator Management role as well as HPC Compute Nodes in the CAE Reference Architectures for SMB using AMD-CPU based servers.

The Lenovo ThinkSystem SR645 V3 is a 2-socket 1U server supporting the 5th Gen AMD EPYC 9005 "Turin" family of processors. With up to 160 cores per processor and support for the new PCIe 5.0 standard for I/O, the SR645 V3 offers great two-socket server performance in a 1U form factor.

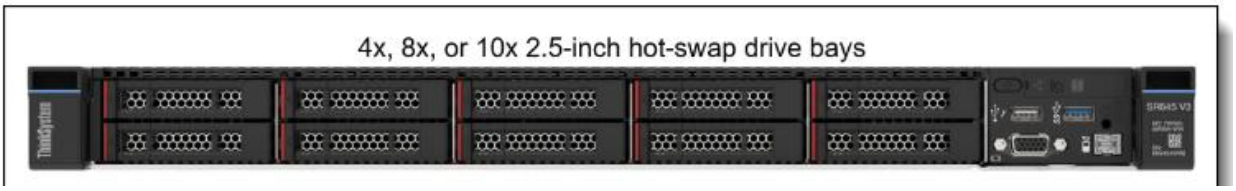

Figure 11: Front view of the ThinkSystem SR645 V3 with up to 10x 2.5-inch drive bays

It provides several different configuration options. In the Reference Architectures we use a chassis supporting up to 10x 2.5-inch hot-swap drives on the front.

In the configuration used in the Reference Architectures, the SR645 V3 provides external connectivity through up to 2 PCIe (1 Full Height and 1 Low Profile) slots and 1 OCP slots at the back side.

Figure 12: Rear view of the ThinkSystem SR645 V3 with 2 PCIe slots

The SR645 V3 product guide, available at Lenovopress provides a more detailed description of the server. It can be found at the following link:

https://lenovopress.lenovo.com/lp1607-thinksystem-sr645-v3-server

Lenovo ThinkSystem SR665 V3

Lenovo ThinkSystem SR665 V3 air-cooled 2U servers are used as Storage Nodes in the CAE Reference Architectures for SMB using AMD-CPU based servers. They are also an option, if Head Nodes require additional networking connectivity, as they provide additional PCIe slots compared to the 1U servers.

The Lenovo ThinkSystem SR665 V3 is a 2-socket 2U server supporting the 5th Gen AMD EPYC 9005 "Turin" family of processors. With up to 160 cores per processor and support for the new PCIe 5.0 standard for I/O, the SR665 V3 offers the ultimate in two-socket server performance in a 2U form factor.

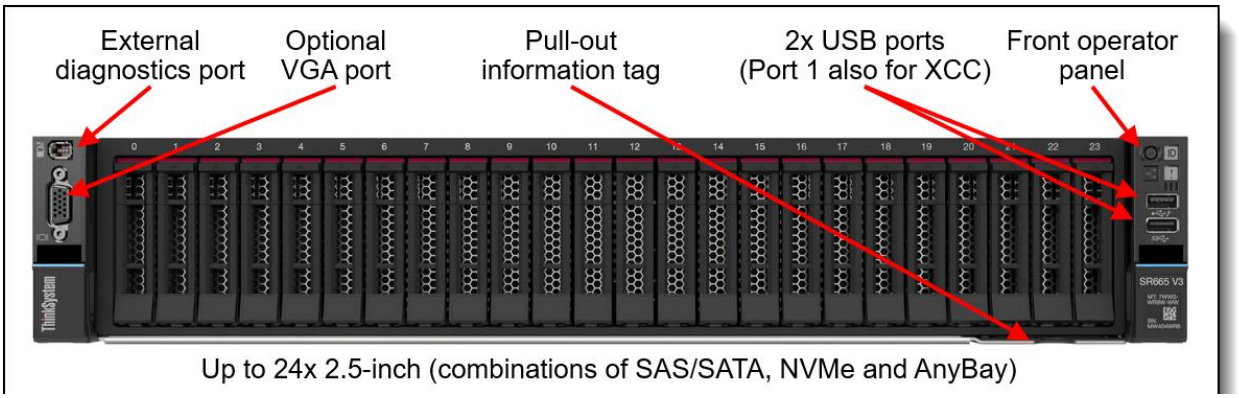

Figure 13: Front view of the ThinkSystem SR665 V3 with 2.5-inch drive bays

It provides several different configuration options. In the Reference Architectures we use a chassis with an NVMe backplane supporting 8 NVMe drives, connected to an internal RAID adapter.

The SR665 V3 provides external connectivity through up to 8 PCIe slots and 1 OCP slot at the back side. In the Reference Architecture, less PCIe slots are needed and hence the number of PCIe slots provided is below the maximum.

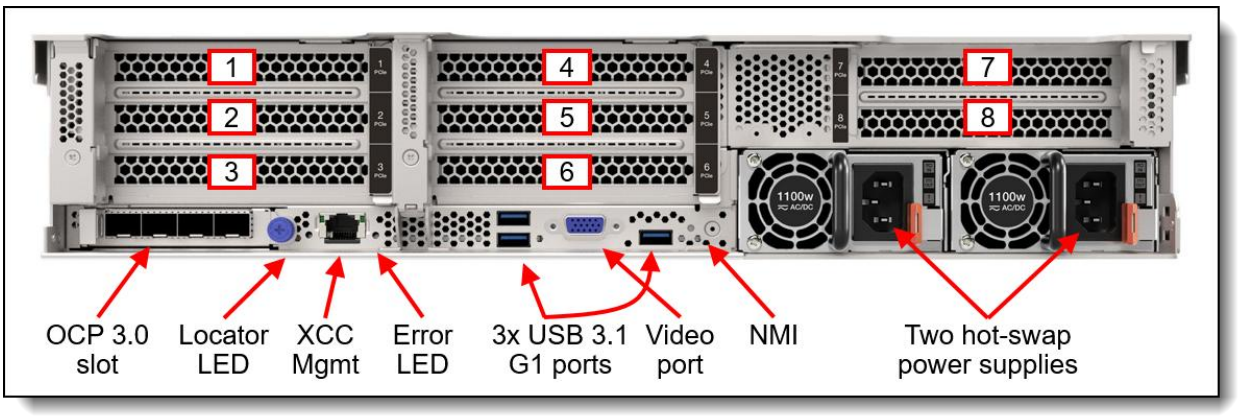

Figure 14: Rear view of the ThinkSystem SR665 V3 (configuration with 8 PCIe slots)

The SR665 V3 product guide, available at Lenovopress, provides a more detailed description of the server. It can be found at the following link:

https://lenovopress.lenovo.com/lp1608-thinksystem-sr665-v3-server

Bill of Materials (BOM) for AMD server-based CAE RA for SMB customers

The following Bill of Materials (BOM) include only the significant parts – other parts like Risers or Power supplies have been removed for better readability. The Lenovo sales configurators DCSC and x-config help to create a valid configuration from the following BOM lists.

Racks and Power Distribution Units are not included in the BOM assuming those are provided by the customer separately.

Please adjust the BOM to your specific situation as needed (e.g. cable lengths).

Part Number	Product Description	Small	Medium	Large	XLarge	Xlarge-200
	CFD-Explicit-FEA-AMD	**4**	**8**	**16**	**34**	**64**
7D9CCTOLWW	**ThinkSystem SR645 V3 - 3yr Warranty - HPC**					
C2ND	AMD EPYC 9455 48C 300W 3.15GHz Processor	2	2	2	2	2
CBNC	ThinkSystem SR645 V3/SR665 V3 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A v2	24	24	24	24	24
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
	Implicit-FEA-AMD	**1**	**1**	**2**	**6**	**8**
7D9CCTOLWW	**ThinkSystem SR645 V3 - 3yr Warranty - HPC**					
C2AK	AMD EPYC 9135 16C 200W 3.65GHz Processor	2	2	2	2	2
CBND	ThinkSystem SR645 V3/SR665 V3 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A v2	24	24	24	24	24
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1

C5MY	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
C0GK	ThinkSystem 2.5" U.2 PM9D3a 960GB Read Intensive NVMe PCIe 5.0 x4 HS SSD	1	1	1	1	1
	HeadNode-AMD (indl. Storage for Small)	**1**	**1**	**1**	**4**	**4**
7D9CCTOLWW	**ThinkSystem SR645 V3 - 3yr Warranty - HPC**					
C2AK	AMD EPYC 9135 16C 200W 3.65GHz Processor	2	2	2	2	2
CBND	ThinkSystem SR645 V3/SR665 V3 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A v2	24	24	24	24	24
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1	1	1	1	1
BCD6	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port PCIe Ethernet Adapter	1	1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter(Generic FW)	1	1	1	1	1
BYFF	ThinkSystem M.2 RAID B540i-2i SATA/NVMe Adapter	1	1	1	1	1
CABU	ThinkSystem M.2 VA 480GB Read Intensive SATA 6Gb NHS SSD	2	2	2	2	2
BGM1	ThinkSystem RAID 940-8i 4GB Flash PCIe Gen4 12Gb Adapter for U.3	1				
BTQ1	ThinkSystem 2.5" U.3 PM1743 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	5				
	Storage-AMD		**1**	**1**	**4**	**4**
7D9ACTOLWW	**ThinkSystem SR665 V3 - 3yr Warranty - HPC**					
C2AF	AMD EPYC 9015 8C 125W 3.6GHz Processor		2	2	2	2
			24	24	24	24
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter		1	1	1	1
C5MY	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter(Generic FW)		1	1	1	1
BYFF	ThinkSystem M.2 RAID B540i-2i SATA/NVMe Adapter		1	1	1	1
CABU	ThinkSystem M.2 VA 480GB Read Intensive SATA 6Gb NHS SSD		2	2	2	2
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD					
CBMY	ThinkSystem RAID 9450-8i 4GB Flash PCIe Gen4 24Gb Adapter for U.3		1	1	1	1
BTQ1	ThinkSystem 2.5" U.3 PM1743 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD		8	8	8	8

	25G/100G Main	**1**	**1**	**1**	**2**	**2**
7D5FCTOKWW	**NVIDIA SN3420 25GbE Managed Switch with Cumulus (PSE)**					
	OPA-Main	**1**	**1**	**1**	**1**	**1**
7DMQCTO1WW	**Cornelis CN5000 48-Port Omni-Path 400GbE Air-Cooled Switch PSE**					
	ClientRack	**1**	**1**	**1**	**2**	**3**
7X74CTO1WW	**Lenovo EveryScale 42U Client Site Integration Kit**					
	Software					
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	6	11	20	48	80
	Cables					
AV1W	Lenovo 1m Passive 25G SFP28 DAC Cable				10	12
AV1F	Lenovo 3m 25G SFP28 Active Optical Cable	6	11	20	21	21
AV1G	Lenovo 5m 25G SFP28 Active Optical Cable				18	48
CBPM	Cornelis 1m CN5000 400G QSFP112 DAC Passive Cable-Straight				7	9
CBPN	Cornelis 1.5m CN5000 400G QSFP112 DAC Passive Cable-Straight			5	11	3
CBPP	Cornelis 2m CN5000 400G QSFP112 DAC Passive Cable-Straight	2	7	11	9	1
CDQW	Cornelis 3m CN5000 400G QSFP112 DAC Passive Cable-Straight	4	4	4	3	3
CBPQ	Cornelis 5m CN5000 400G QSFP112 ACC Active Cable-Straight				18	
CDHC	Cornelis 1m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					
CDHB	Cornelis 1.5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					4
CDHA	Cornelis 2m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					4
CDHH	Cornelis 3m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					
CDHG	Cornelis 5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split					24

Table 5: Bill of Materials (BOM) for AMD server-based CAE solution for SMB customers

CAE Reference Architectures for scale-out

Overview

This CAE Reference Architectures for scale-out are purpose-built for accelerating CAE and Computational Fluid Dynamics (CFD) workloads across a wide range of industries and applications, utilizing parallelism across many Compute Nodes. They use Lenovo Neptune direct-water-cooling technology for providing maximum performance with optimal use of energy. There are two versions of this Reference Architecture:

- Intel Xeon 6900-series CPU-based Scalable Unit (SU) Reference Architecture
- AMD 5th Gen EPYC CPU-based Scalable Unit (SU) Reference Architecture

For each dual-socket Compute Node, a network bandwidth of 200 Gbit/s to the Cornelis CN5000 Omni-Path High Performance Fabric is planned. Omni-Path can provide 400 Gbit/s bandwidth to each Compute Node adapter with straight cables. For CAE workloads, the Omni-Path High Performance Fabric bandwidth to the Compute Nodes is generally less important than the low-latency performance that Omni-Path provides. Hence the use of splitter cables providing 200 Gbit/s bandwidth per Compute Node is more cost efficient for scale-out, as it requires less CN5000 Switches and less cables compared to a solution with 400 Gbit/s per Compute Node.

Management or Login servers as well as storage are out-of-scope for this Reference Architectures. Examples can be found in the previous Reference Architecture for SMB but should be adjusted according to the specific needs.

For storage in a CAE scale-out environment, a higher performing and scalable parallel file system solution like Lenovo DSS-G is recommended compared to a single NFS server. Lenovo Distributed Storage Solution for IBM Storage Scale (DSS-G) is a software-defined storage (SDS) solution for dense scalable file and object storage suitable for high-performance and data-intensive environments. DSS-G combines the performance of Lenovo ThinkSystem servers, Lenovo storage enclosures, and industry leading IBM Storage Scale software, to offer a high performance, scalable building block approach to modern storage needs.

Lenovo plans to enable a direct connection of DSS-G building blocks to the Cornelis CN5000 Omni-Path High Performance Fabric in 2026. Until then, a connection of DSS-G storage to the

system using Ethernet Networks or a gateway solution is a possibility. Please contact your Lenovo or Partner representative for a recommended solution for your specific environment and check the DSS-G product guide for more information:

https://lenovopress.lenovo.com/lp1842-lenovo-dss-g-thinksystem-v3

CAE Reference Architecture for scale-out using Intel CPUs

Leveraging Lenovo ThinkSystem SC750 V4 Neptune DWC servers powered by Intel Xeon 6900-series processors, this architecture integrates advanced direct water-cooling technology to deliver high performance with exceptional energy efficiency. With high-speed Omni-Path High Performance Fabric and MRDIMM memory technology, the solution provides the computational scale and bandwidth required to handle complex CFD simulations—enabling faster runtimes, greater model fidelity, and accelerated innovation.

Scalable Unit (SU) building block design

This Reference Architecture is built on a Scalable Unit (SU) structure, with 24 SC750 V4 dual-node trays, 48 Compute Nodes and 96 CPUs per SU, with 200 Gbit/s bandwidth from each Compute Node to the Omni-Path High Performance Fabric. It scales up to 46 Scalable Units with 2208 servers and 4408 CPUs within a standard FAT Tree Network Topology, with two CN5000 Omni-Path Switches being reserved for Management and Login Nodes as well as storage.

With alternative network topologies such as Megafly, the architecture has the potential to scale exponentially, so there are options available if the cluster needs to scale further.

Each Scalable Unit comprises a single compute rack equipped with one high-speed CN5000 Omni-Path leaf switch as well as Ethernet switches for hardware management and operating system boot. This SU is purpose-built for seamless scalability, offering on-demand growth to support CFD simulations of any resolution or complexity.

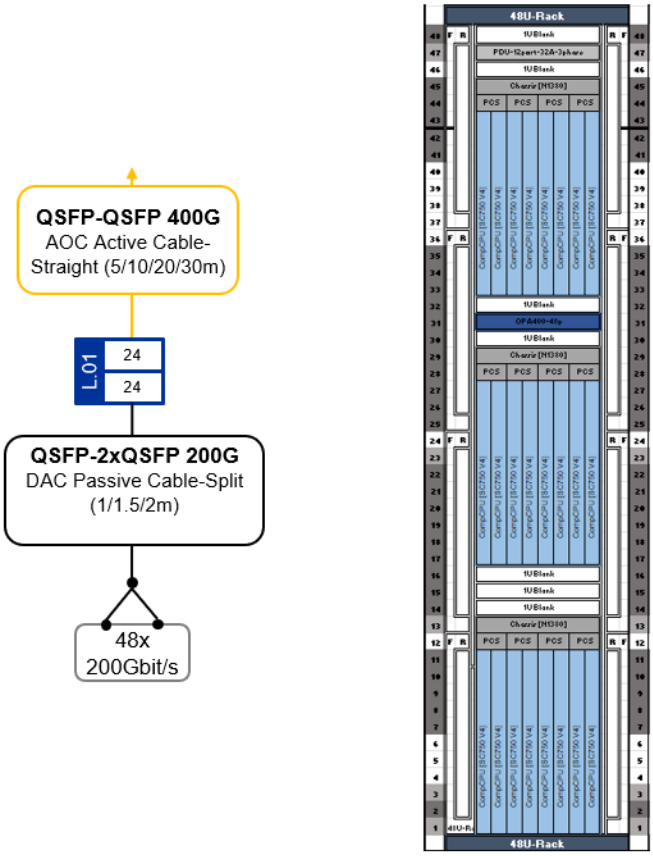

Figure 15: SU for 48 CFD Compute Nodes in 24 Lenovo SC750 V4 trays

Scaling-out the SUs with Omni-Path Fat-Tree topology

This Reference Architecture employs a high-speed network utilizing the Cornelis CN5000 Omni-Path High Performance Fabric with 200 Gbit/s bandwidth per Compute Node. This Omni-Path fabric is implemented in a Fat-Tree topology with up to 24 spine and 48 leaf switches, enabling scalability up to 2304 CN5000 Omni-Path SuperNICs at 200 Gbit/s bandwidth. Each leaf switch connects to the spine switches with 24 * 400 Gbit/s uplinks (Omni-Path cable-straight).

Of the up to 48 leaf switches, two are reserved for Management/Login nodes as well as storage I/O connectivity, supporting up to 96 Omni-Path SuperNICs at 200 Gbit/s bandwidth using Omni-Path cable-split.

The remaining up to 46 leaf switches support up to 2208 CAE Compute Nodes with 4416 CPUs at 200 Gbit/s bandwidth using Omni-Path cable-split.

The following table shows how the number of Spine Switches and uplink/downlink ports scales with the number of Scalable Units (SUs) in a configuration.

Number of SU	Compute Nodes (SC750 V4 half-tray)	CN5000 Switches Compute Leaf	CN5000 Switches Management Leaf	CN5000 Switches Spine	Uplinks to Spine (cable-straight)	Downlinks to node (cable-split)	OPA200 ports for Mgmt/Storage
1	48	1	0	0	24	24	8
2	96	2	2	2	96	48	96
6	288	6	2	4	192	144	96
10	480	10	2	6	288	240	96
14	672	14	2	8	384	336	96
22	1056	22	2	12	576	528	96
46	2208	46	2	24	1152	1104	96

Table 6: Scaling the number of SUs for scale-out using Intel CPUs

The first row in the table with just a single SU defines a “special case”, which is not actually a scale-out cluster but uses only a single CN5000 Omni-Path Switch. There is no need for a Spine Layer in this single SU configuration, and Management/Login as well as Storage can be

connected to the same switch, like what we described in the Reference Architecture for SMB customers before.

The Omni-Path High Performance Fabric topology provides a solid network High Performance Fabric foundation for massive scale-out CFD solutions, addressing the needs for scalability, high bandwidth, and low latency, ensuring robust performance for complex simulations and models.

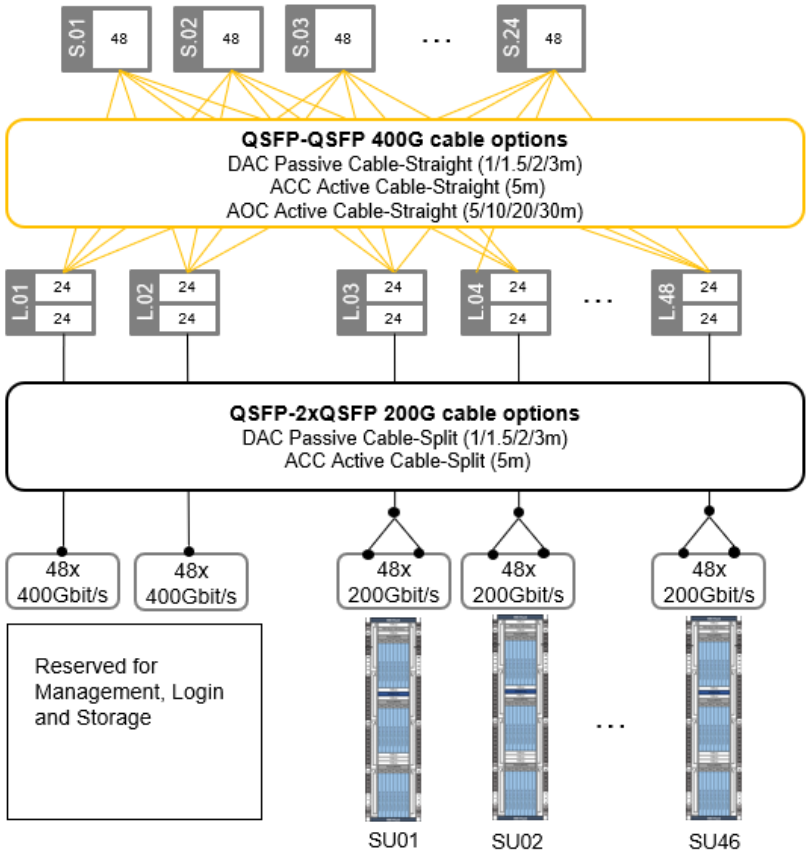

Figure 16: Omni-Path Fat-Tree topology for up to 2208 * Compute Nodes @200Gbit/s

Each Scalable Unit consists of a single compute rack housing 48 nodes shared across 24 Lenovo ThinkSystem SC750 V4 dual-node compute trays. Each compute tray is equipped with two high-speed CN5000 Omni-Path SuperNICs (one per Compute Node). With Omni-Path split cables, the two CN5000 SuperNICs in a tray are connected at 200 Gbit/s to a single 400 Gbit/s port on the CN5000 Omni-Path Switch, cutting the network bandwidth to 50%. Performance testing has shown that this reduction has a minimal impact on CFD workloads, with less than a 10% performance drop observed even at large scales.

This streamlined approach offers a cost-effective solution for scaling CPU and memory-intensive workloads. By maintaining a balanced design, customers can accurately scale their workload when CPU and memory tasks heavily outweigh inter-node communication requirements. This optimized configuration results in significant cost savings while delivering optimal price and performance.

Cluster management is usually done over Ethernet, and the SC750 V4 offers multiple options. It comes with 25GbE SFP28 Ethernet ports, a Gigabit Ethernet port, and a dedicated XClarity Controller (XCC) port. For connecting to the XCC, there is also a second path through the N1380 enclosure Systems Management Module (SMM). These can be customized based on cluster management and workload needs.

For stable environments where the Operating System is installed on local hard drives and there are infrequent OS changes, the single Gigabit Ethernet port is sufficient for the Management Network. A CAT5e or CAT6 cable per node can use Network Controller Sideband Interface (NC-SI) for remote out-of-band and cluster management over one wire.

For massive scale-out clusters though, it is recommended to boot the Operating System from a centralized Management Server instead of installing it on local drive. This way, it is easy to keep the Operating System image consistent across the CAE HPC cluster. The Confluent management Software, developed by Lenovo, provides a toolkit for Compute Nodes in a diskless way. For this kind of diskless installations, as well as for higher bandwidth needs or frequent updates, the 25Gb Ethernet interfaces offer a better solution, including sideband communication to the XCC.

For this CAE Reference Architecture, the 25Gbit Ethernet interfaces of each SC750 V4 compute tray are linked to a 25 GbE Management Leaf switch, ensuring fast Operating Boot over the network and connectivity for the Compute Nodes' cluster management. For out-of-band

management, access to the SC750 V4 xClarity Controller (XCC) and N1380 enclosure systems management modules (SMM) is provided through 1Gbit Ethernet connections to a centralized Hardware Management Ethernet Switch.

Compute Node description

Lenovo ThinkSystem N1380 chassis

The ThinkSystem N1380 Neptune chassis is the core building block, built to enable exascale-level performance while maintaining a standard 19-inch rack footprint. It uses liquid cooling to remove heat and increase performance and is engineered for the next decade of computational technology.

Figure 17: Lenovo ThinkSystem N1380 Enclosure

N1380 features an integrated manifold that offers a patented blind-mate mechanism with aerospace-grade drip-less connectors to the compute trays, ensuring safe and seamless operation. The unique design of the N1380 eliminates the need for internal airflow and power-consuming fans. As a result, it achieves a reduction in typical data center power consumption by up to 40% compared to similar air-cooled systems.

This newly developed enclosure incorporates up to four ThinkSystem 15kW Titanium Power Conversion Stations (PCS). These stations are directly fed with high current three-phase power and supply power to an internal 48V busbar, which in turn powers the compute trays. The PCS design is a game-changer, merging power conversion, rectification, and distribution into a single package. This is a significant transformation from traditional setups that require separate rack PDUs, additional cables and server power supplies, resulting in best-in-class efficiency. In specific CPU configurations, the PCS can be optimized to only require 2 PCS per chassis whilst still fully powering the nodes for performance.

Each 13U Lenovo ThinkSystem N1380 Neptune enclosure houses eight Lenovo ThinkSystem SC-series Neptune trays. Up to three N1380 enclosures fit into a standard 19" rack cabinet, packing 24 trays into just two 60x60 datacenter floor tiles.

Lenovo ThinkSystem SC750 V4 dual-server tray

The ThinkSystem SC750 V4 Neptune node is the next-generation high-performance server based on the sixth generation Lenovo Neptune direct water cooling platform.

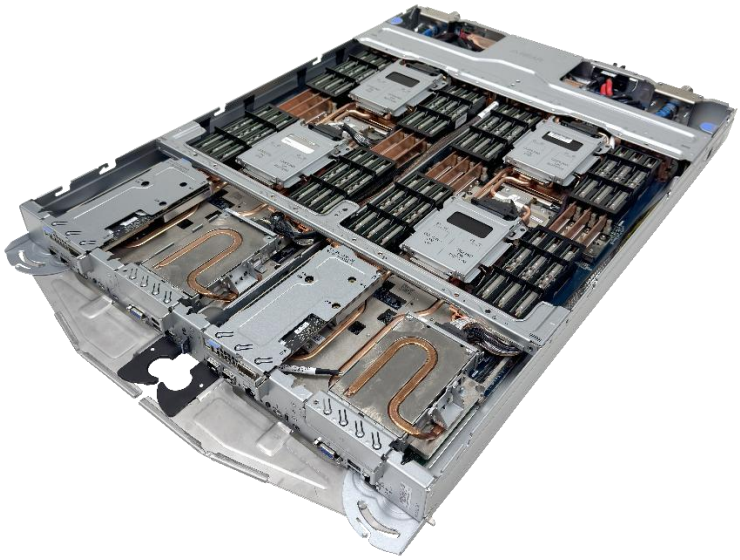

Figure 18: Lenovo ThinkSystem SC750 V4 Neptune Server Tray

Supporting the Intel Xeon 6900P-series, the ThinkSystem SC750 V4 Neptune stands as a powerhouse for demanding HPC workloads. Its industry-leading direct water-cooling system ensures steady heat dissipation, allowing CPUs to maintain accelerated operation and achieve up to a 10% performance enhancement.

With 12 channels of high-speed DDR5 RDIMM or an impressive 8800MHz high-bandwidth MRDIMM capability, it excels in memory bandwidth-intensive workloads, positioning it as a preferred choice for engineering and meteorology applications like Fluent, STAR-CCM+, OpenFOAM, WRF, and ICON.

Completing the package with support for high-performance NVMe and high-speed, low-latency networking with the latest InfiniBand, Omni-Path, and Ethernet choices, the SC750 V4 is your all-in-one solution for HPC workloads.

At its core, Lenovo Neptune applies 100% direct warm-water cooling, maximizing performance and energy efficiency without sacrificing accessibility or serviceability. The SC750 V4 is installed into the ThinkSystem N1380 Neptune enclosure which itself integrates seamlessly into a standard

19" rack cabinet. Featuring a patented blind-mate stainless steel dripless quick connection, SC750 V4 node trays can be added “hot” or removed for service without impacting other node trays in the enclosure.

This modular design ensures easy serviceability and extreme performance density, making the SC750 V4 the go-to choice for compute clusters of all sizes - from departmental/workgroup levels to the world’s most powerful supercomputers – from Exascale to Everyscale.

Intel Xeon 6 processors with P-cores are optimized for high performance per core. With more cores, double the memory bandwidth, and AI acceleration in every core, Intel Xeon 6 processors with P-cores provide twice the performance for the widest range of workloads, including HPC and AI.

Domains such as CAE/CFD and weather/climate modeling present a more balanced performance profile - demanding both high compute throughput and substantial memory bandwidth. Simply increasing core counts can lead to diminishing returns unless accompanied by improvements in memory access speed, latency, and power delivery. The Intel Xeon 6900 series address these challenges by expanding the thermal design power (TDP) to 500 watts, which, when coupled with Lenovo Neptune cooling, helps to sustain or even boost CPU frequencies under heavy loads.

Additionally, support for 12 memory channels and compatibility with DDR5-6400 MHz and MRDIMM-8800 MHz memory types significantly increases memory bandwidth, ensuring that high-core-count systems remain efficient and scalable for memory-intensive workloads. Reflecting these architectural advantages, CFD workloads from Fluent, STAR-CCM+, and OpenFOAM have observed over 2.4x faster computational times on average when running on Intel Xeon 6900 P-core processors compared to the previous generation - demonstrating the real-world impact of these enhancements on simulation throughput and productivity.

In this Reference Architecture, each Compute Node is equipped with two Intel Xeon 6960P CPUs, each comprising 72 Xeon6 P-cores. This configuration provides 144 Xeon6 P-cores and 1.5TB of MRDIMM RAM per node, making the SC750 V4 highly suited for core and RAM-intensive tasks. Utilizing high-speed, low-latency Cornelis CN5000 Omni-Path SuperNICs at 200 Gbit/s, the SC750 V4, when paired with Intel Xeon6, offers exceptional scalability for the most demanding parallel MPI jobs.

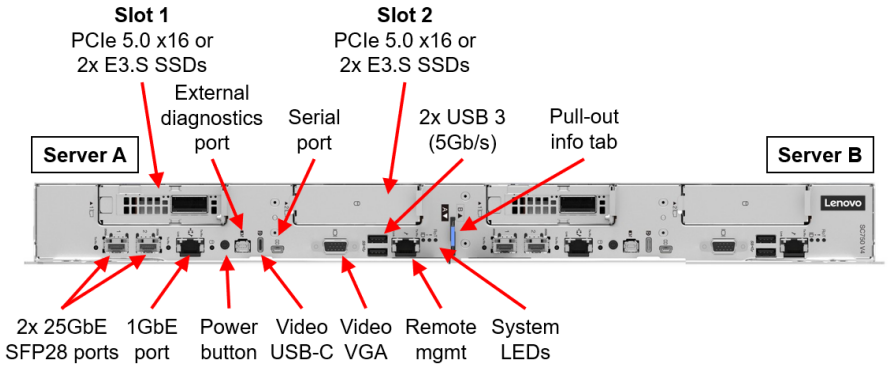

Figure 19: SC750 V4 Front View with Management Ports

The SC750 V4 integrates the XCC through the Data Center Secure Control Module (DC-SCM) I/O board. This module also includes a Root of Trust module (NIST SP800-193 compliant), USB 3.2 ports, a VGA port, and MicroSD card capability for additional storage with the XCC, offering firmware storage options up to 4GB, including N-1 firmware history.

The N1380 enclosure features a System Management Module 3 (SMM) at the rear, managing both the enclosure and individual servers through a web browser or Redfish/IPMI 2.0 commands. The SMM provides remote connectivity to XCC controllers, node-level reporting, power control, enclosure power management, thermal management, and inventory tracking.

Bill of Materials (BOM) for Intel server-based scale-out SU

The following Bill of Materials (BOM) include only the significant parts of the Scalable Unit (SU) building bloc – other parts like Risers or Power cables have been removed for better readability.

The Lenovo sales configurators DCSC and x-config will help to create a valid configuration from the BOM list.

In addition to the Compute Node SUs as described in this BOM, there are central components that need to be provided, for example

- Omni-Path leaf switches (for management/login/storage) and spine switches
- Ethernet leaf switches (for management/login/storage) and core switches

- Management, login and storage systems

Also adjust the SU BOM to your specific situation as needed (e.g. cable lengths from the SU to spine/core switches are 30m in this BOM – shorter or longer cable lengths may be required for connecting to the central spine/core switches are needed, depending on the location of the racks in the data center).

Part Number	Product Description	per Qty	Tot Qty
	SC750v4_2xOPA200		
7DDJCTOLWW	**-SB- ThinkSystem SC750 V4 Neptune Tray - 3-Year Base Warranty**		**24**
C2WS	Intel Xeon 6960P 72C 500W 2.7GHz Processor	4	96
C0TX	ThinkSystem 64GB TruDDR5 8800MHz (2Rx4) MRDIMM	48	1152
BYK4	ThinkSystem SC750 V4 Neptune Tray	1	24
CB7C	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter DWC	2	48
	8xSC750v4		
7DDHCTOLWW	**-SB- Lenovo ThinkSystem N1380 Neptune Enclosure**		**3**
BYKR	ThinkSystem N1380 Neptune Enclosure Midplate Assembly	1	3
BE0D	N+1 Redundancy With Over-Subscription	1	3
BYKK	ThinkSystem N1380 Neptune EPDM Hose Connection	1	3
BYKJ	ThinkSystem N1380 Neptune System Management Module V3	1	3
BYJZ	ThinkSystem N1380 Neptune Enclosure	1	3
BYKH	ThinkSystem N1380 Neptune 15KW 3-Phase 380-480V Titanium Power Conversion Station v1.1	4	12
	Management main		
7D5FCTOFWW	**NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)**		**1**
BPC7	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	1	1
BVA4	Lenovo 100GBase-SR4 QSFP28 Transceiver	4	4

25G Main			
7D5FCTOLWW	**NVIDIA SN3420 25GbE Managed Switch with Cumulus (oPSE)**		**1**
BUZ3	NVIDIA SN3420 25GbE Managed Switch with Cumulus (oPSE)	1	1
BVA4	Lenovo 100GBase-SR4 QSFP28 Transceiver	4	4

OPA-400 Main			
7DMQCTO2WW	**Cornelis CN5000 48-Port Omni-Path 400Gbps Air-Cooled Switch oPSE**		**1**
CC44	Cornelis CN5000 48-Port Omni-Path 400Gbps Air-Cooled Switch oPSE	1	1

16xSC750v4-2xOPA400			
1410O48	**Lenovo EveryScale 48U Onyx Heavy Duty Rack Cabinet**		**1**
BHCH	Lenovo EveryScale 48U Onyx Heavy Duty Rack Cabinet	1	1
BJ2L	6U Front Cable Management Bracket	1	1
BJPD	21U Front Cable Management Bracket	2	2
BHCM	ThinkSystem 48U Onyx Heavy Duty Rack Side Panel	2	2
C4YZ	ThinkSystem 48U Onyx 180mm Advanced Rack Extension Kit	2	2
BHCL	ThinkSystem 48U Onyx Heavy Duty Rack Rear Door	1	1
C0DC	1U 12 C19/C13 Switched and monitored 32A 3P WYE PDU V2	1	1
BYKP	ThinkSystem N1380 Neptune Enclosure Rack Post Enhance Kit	8	8
C2KQ	ThinkSystem N1380 Neptune 2.2M and 2.8M EPDM Hose Set	1	1
C2KP	ThinkSystem N1380 Neptune 2.8M and 3.4M EPDM Hose Set	1	1
C2KN	ThinkSystem N1380 Neptune 3.4M and 3.8M EPDM Hose Set	1	1
5AS7B07693	Lenovo EveryScale Rack Setup Services	1	1
5AS7B07695	Lenovo EveryScale Advanced Cabling Services	1	1

Software

SCY0	Lenovo XClarity XCC3 premier - FOD	48

Cables		
3798	3m Green Cat5e Cable	30
CDHD	Cornelis 30m CN5000 400G QSFP112 AOC Active Cable-Straight	24
AV2A	Lenovo 30m MPO-MPO OM4 MMF Cable	4
AV1F	Lenovo 3m 25G SFP28 Active Optical Cable	24
CDHA	Cornelis 2m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split	8
CDHB	Cornelis 1.5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split	8
CDHC	Cornelis 1m CN5000 400G QSFP112 - 2x200G QSFP56 DAC Passive Cable-Split	8

Table 7: Bill of Material for a single Intel-server based Scalable Unit (SU)

CAE Reference Architecture for scale-out using AMD CPUs

Leveraging Lenovo ThinkSystem SD665 V3 Neptune DWC servers powered by two 5th Gen AMD EPYC processors, this architecture integrates advanced direct water-cooling technology to deliver high performance with exceptional energy efficiency. With high-speed Omni-Path High Performance Fabric, the solution provides the computational scale and bandwidth required to handle complex CFD simulations—enabling faster runtimes, greater model fidelity, and accelerated innovation.

Scalable Unit (SU) building block design

This Reference Architecture is built on a Scalable Unit (SU) structure, with 72 SD665 V3 dual-node trays, 144 Compute Nodes and 288 CPUs per SU, with 200 Gbit/s bandwidth from each Compute Node to the Omni-Path High Performance Fabric. This architecture scales up to 15 Scalable Units with 2160 servers and 4320 CPUs within a standard FAT Tree Network Topology, with two CN5000 Switches being reserved for Management and Login Nodes as well as storage.

With alternative network topologies such as Megafly, the architecture has the potential to scale exponentially, so there are options available if the cluster needs to scale further.

Each Scalable Unit comprises two compute racks equipped with three high-speed Omni-Path leaf switches as well as Ethernet switches for hardware management and operating system boot. This SU is purpose-built for seamless scalability, offering on-demand growth to support CFD simulations of any resolution or complexity.

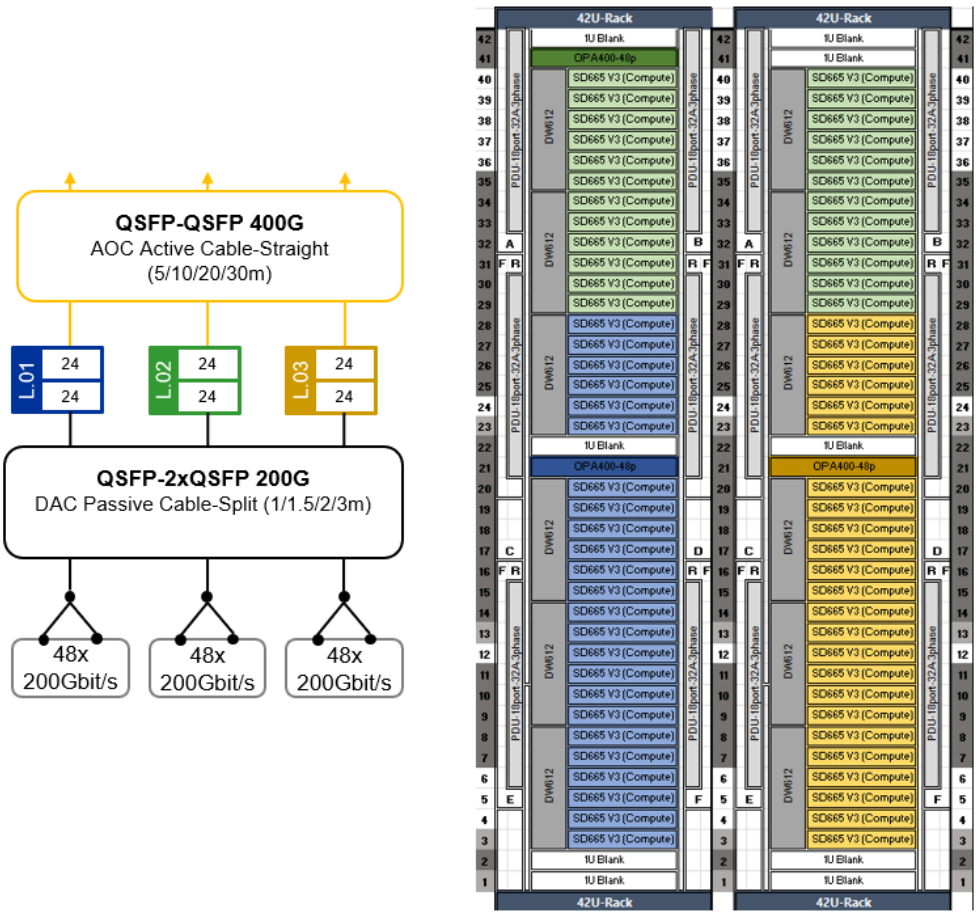

Figure 20: SU for 144 CFD Compute Nodes in 72 Lenovo SD665 V3 trays

Scaling-out the SUs with Omni-Path Fat-Tree topology

This Reference Architecture employs a high-speed network utilizing the Cornelis CN5000 Omni-Path High Performance Fabric with 200 Gbit/s bandwidth per Compute Node. This Omni-Path fabric is implemented in a Fat-Tree topology with up to 24 spine and 47 leaf switches, enabling scalability up to 2256 CN5000 Omni-Path SuperNICs at 200 Gbit/s bandwidth. Each leaf switch connects to the spine switches with 24 * 400 Gbit/s uplinks (Omni-Path cable-straight).

Of the up to 47 leaf switches, two are reserved for Management/Login nodes as well as storage I/O connectivity, supporting up to 96 Omni-Path SuperNICs at 200 Gbit/s bandwidth using Omni-Path cable-split.

The remaining up to 45 leaf switches support up to 2160 CAE Compute Nodes with 4320 CPUs at 200 Gbit/s bandwidth using Omni-Path cable-split.

The following table shows how the number of Spine Switches and uplink/downlink ports scale with the number of Scalable Units (SUs) in a configuration.

Number of SU	Compute Nodes (SD665 V3 half-tray)	CN5000 Switches Compute Leaf	CN5000 Switches Management Leaf	CN5000 Switches Spine	Uplinks to Spine (cable-straight)	Downlinks to node (cable-split)	OPA200 ports for Mgmt/Storage
1	144	3	2	2	120	72	48
2	288	6	2	4	192	144	96
3	432	9	2	6	264	216	96
4	576	12	2	8	336	288	96
7	1008	21	2	12	552	504	96
15	2160	45	2	24	1128	1080	96

Table 8: Scaling the number of SUs for scale-out using AMD CPUs

The Omni-Path High Performance Fabric topology provides a solid network High Performance Fabric foundation for massive scale-out CFD solutions, addressing the needs for scalability, high bandwidth, and low latency, ensuring robust performance for complex simulations and models.

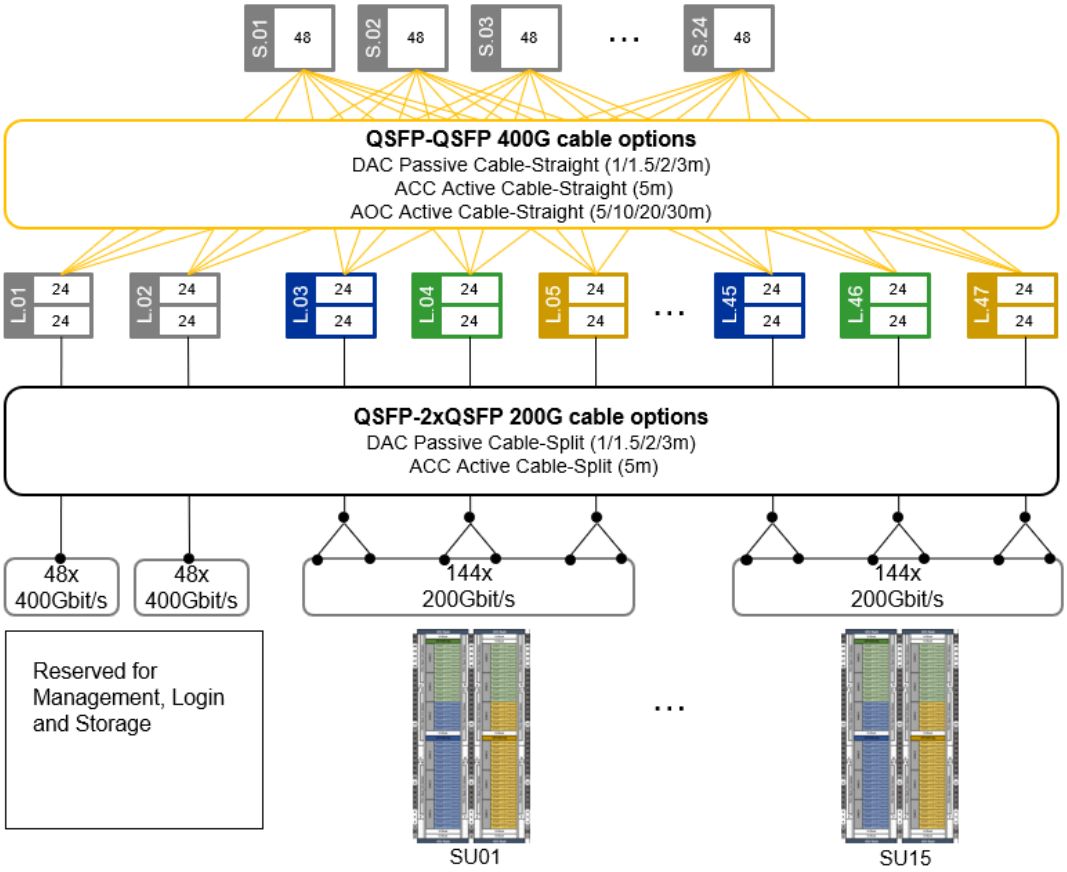

Figure 21: Omni-Path Fat-Tree topology for up to 2160 * Compute Nodes @200Gbit/s

Each Scalable Unit consists of two compute rack housing 144 Compute Nodes shared across 72 Lenovo ThinkSystem SD665 V3 dual-node compute trays. Each compute tray is equipped with two high-speed CN5000 Omni-Path SuperNICs (one per Compute Node). With Omni-Path split cables, the two CN5000 adapters in a tray are connected at 200 Gbit/s to a single 400 Gbit/s port

on the CN5000 Omni-Path Switch, cutting the network bandwidth to 50%. Performance testing has shown that this reduction has a minimal impact on CFD workloads, with less than a 10% performance drop observed even at large scales.

This streamlined approach offers a cost-effective solution for scaling CPU and memory-intensive workloads. By maintaining a balanced design, customers can accurately scale their workload when CPU and memory tasks heavily outweigh inter-node communication requirements. This optimized configuration results in significant cost savings while delivering optimal price and performance.

Cluster management is usually done over Ethernet, and the SD665 V3 offers multiple options. It comes with 25GbE SFP28 Ethernet ports and a Gigabit Ethernet port – one of those can optionally be shared with the XClarity Controller (XCC). For connecting to the XCC, there is also a second path through the DW612S enclosure Systems Management Module (SMM). These can be customized based on cluster management and workload needs.

For stable environments where the Operating System is installed on local hard drives and there are infrequent OS changes, the single Gigabit Ethernet port is sufficient for the Management Network. A CAT5e or CAT6 cable per node can use Network Controller Sideband Interface (NC-SI) for remote out-of-band and cluster management over one wire.

For massive scale-out clusters though, it is recommended to boot the Operating System from a centralized Management Server instead of installing it on local drive. This way, it is easy to keep the Operating System image consistent across the CAE HPC cluster. The Confluent management Software, developed by Lenovo, provides a toolkit for Compute Nodes in a diskless way. For this kind of diskless installations, as well as for higher bandwidth needs or frequent updates, the 25Gb Ethernet interfaces offer a better solution, including sideband communication to the XCC.

For this CAE Reference Architecture, the 25Gbit Ethernet interfaces of each SD665 V3 dual-node trays are linked to a 25 GbE Management Leaf switch, ensuring fast Operating Boot over the network and connectivity for the Compute Nodes' cluster management. For out-of-band management, access to the SD665 V3 xClarity Controller (XCC) and DW612S enclosure systems management modules (SMM) is provided through 1Gbit Ethernet connections to a centralized Hardware Management Ethernet Switch.

Compute Node description

Lenovo ThinkSystem DW612S enclosure

The Direct-Water-Cooled (DSC) Lenovo ThinkSystem DW612S enclosure provides the power and water-cooling infrastructure for the SD665 V3 dua-node compute trays. It provides 6 horizontal slots on the front side for up to 6 compute trays, which connect to the enclosure mid-plane and enclosure-internal water manifold through patented stainless-steel drip-less quick connectors. Trays can be removed and inserted into the enclosure without impact on the operation of the other trays.

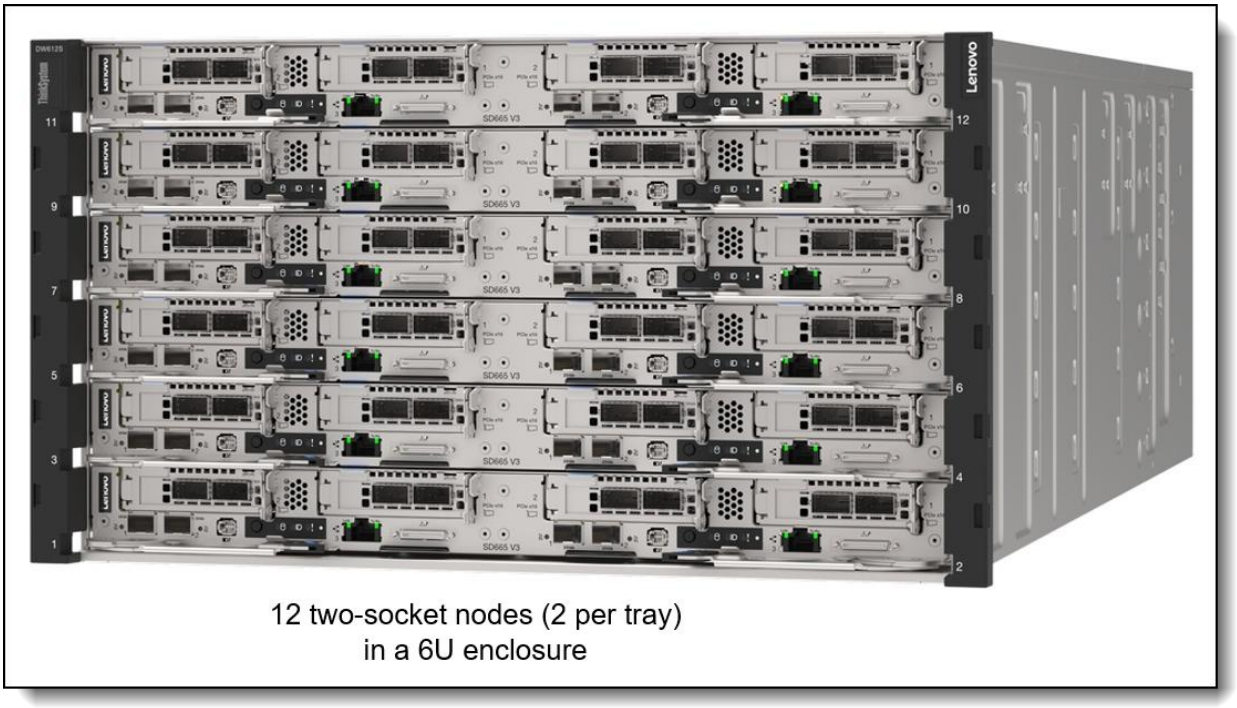

Figure 22: ThinkSystem DW612S enclosure front-side view

At the backside of the enclosure, there is space for either six or eight air-cooled power supplies or three direct-water-cooled AC power supplies. There is also a System Management Module

(SMM) at the back, which allows remote management of the enclosure (e.g. providing power telemetry or the ability for power on/off or a virtual re-seat of individual trays) as well as an internal network connection to the Management Modules (xClarity Controller -XCC) in the Compute Trays.

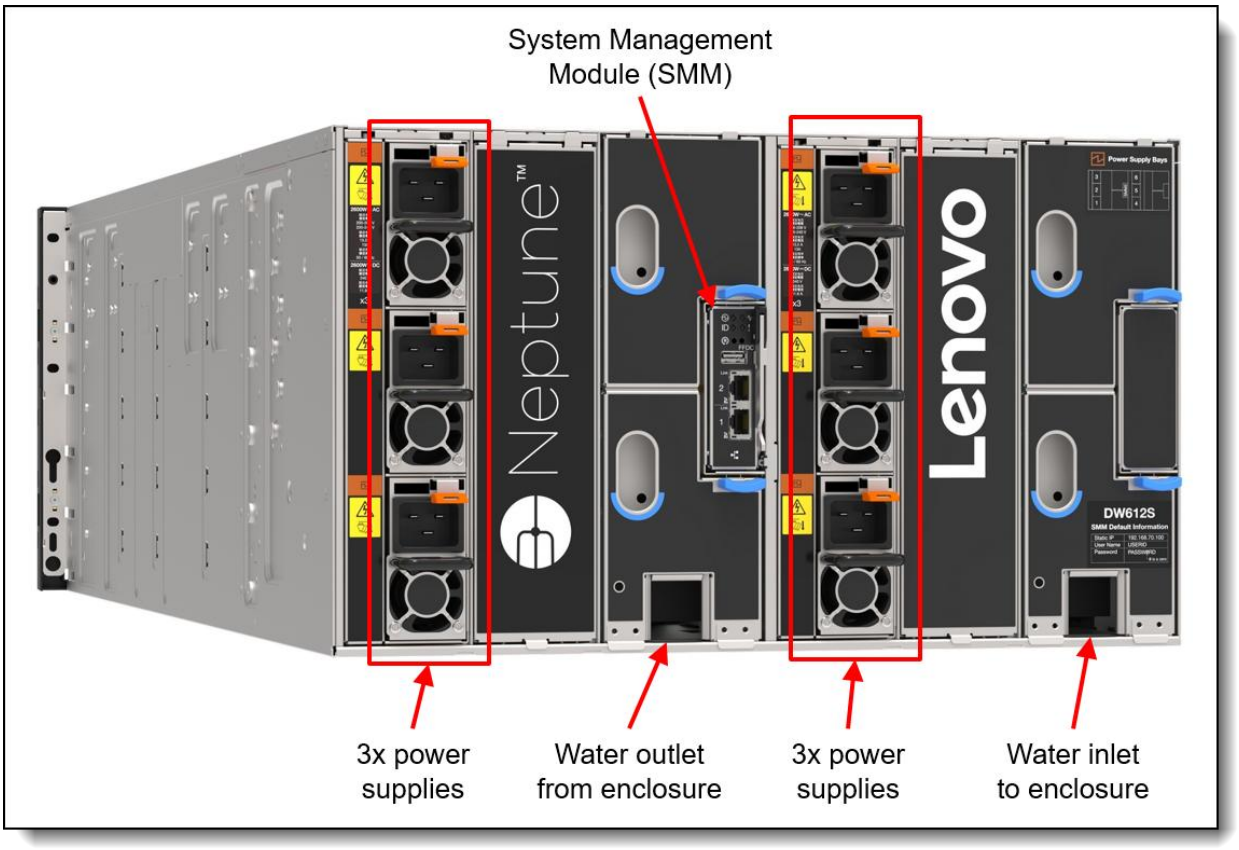

Figure 23: ThinkSystem DW612S enclosure rear view

For this reference architecture, six ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supplies per DW612S enclosure are used. That provides sufficient power for a N+1 redundant power configuration with over-subscription (oversubscription implies, that if one power supply fails, the Compute Nodes may be throttled if under

Lenovo ThinkSystem SD665 V3

The Lenovo ThinkSystem SD665 V3 dual-node tray is designed for High Performance Computing (HPC), large-scale cloud, heavy simulations and modeling.

It supports Lenovo Neptune™ Direct Water Cooling (DWC) technology as well as workloads from technical computing, grid deployments, analytics, and is ideally suited for fields such as research, life sciences, energy, simulation, and engineering.

The unique design of ThinkSystem SD665 V3 provides the optimal balance of serviceability, performance, and efficiency.

By using a standard rack with the ThinkSystem DW612S enclosure equipped with patented stainless-steel drip-less quick connectors, the SD665 V3 provides easy serviceability and extreme density that is well suited for clusters ranging from small enterprises to the world's largest supercomputers.

The Lenovo Neptune™ direct liquid cooling doesn't use risky plastic retrofitting but custom designed copper water loops, so you have peace of mind implementing a platform with liquid cooling at the core of the design.

Compared to other technology, the ThinkSystem SD665 V3 direct water cooling:

- Reduces data center energy costs by up to 40%
- Increases system performance by up to 10%
- Delivers up to 95% heat removal efficiency
- Creates a quieter data center with its fan-less design
- Enables data center growth without adding computer room air conditioning

Designed to run the highest core-count 5th Generation AMD EPYC™ Processor, the SD665 V3 powers through demanding HPC workloads. Because water cooling removes more heat constantly, CPUs can run in accelerated mode nonstop, getting up to 10% greater performance from the CPU.

The 5th Generation AMD EPYC™ Processors combine both superior memory bandwidth capacity and core-counts that can increase performance across all HPC workloads.

The 5th Generation AMD EPYC™ Processors excel in HPC application and workloads that are memory sensitive, scale well to multiple cores, and are not highly vectorized applications within the manufacturing/computer-aided engineering (CAE) and weather/climate verticals like OpenFOAM, ANSYS Fluent, ANSYS CFX, ANSYS LS-DYNA, Siemens STAR-CCM+, MOM5, and WRF.

For even greater system performance, the SD665 V3 uses 6400MHz DDR5 memory and supports NVMe storage, high-speed NDR InfiniBand.

The Lenovo ThinkSystem SD665 V3 is enabled with Lenovo HPC & AI Software Stack, so, you can support multiple users and scale within a single cluster environment.

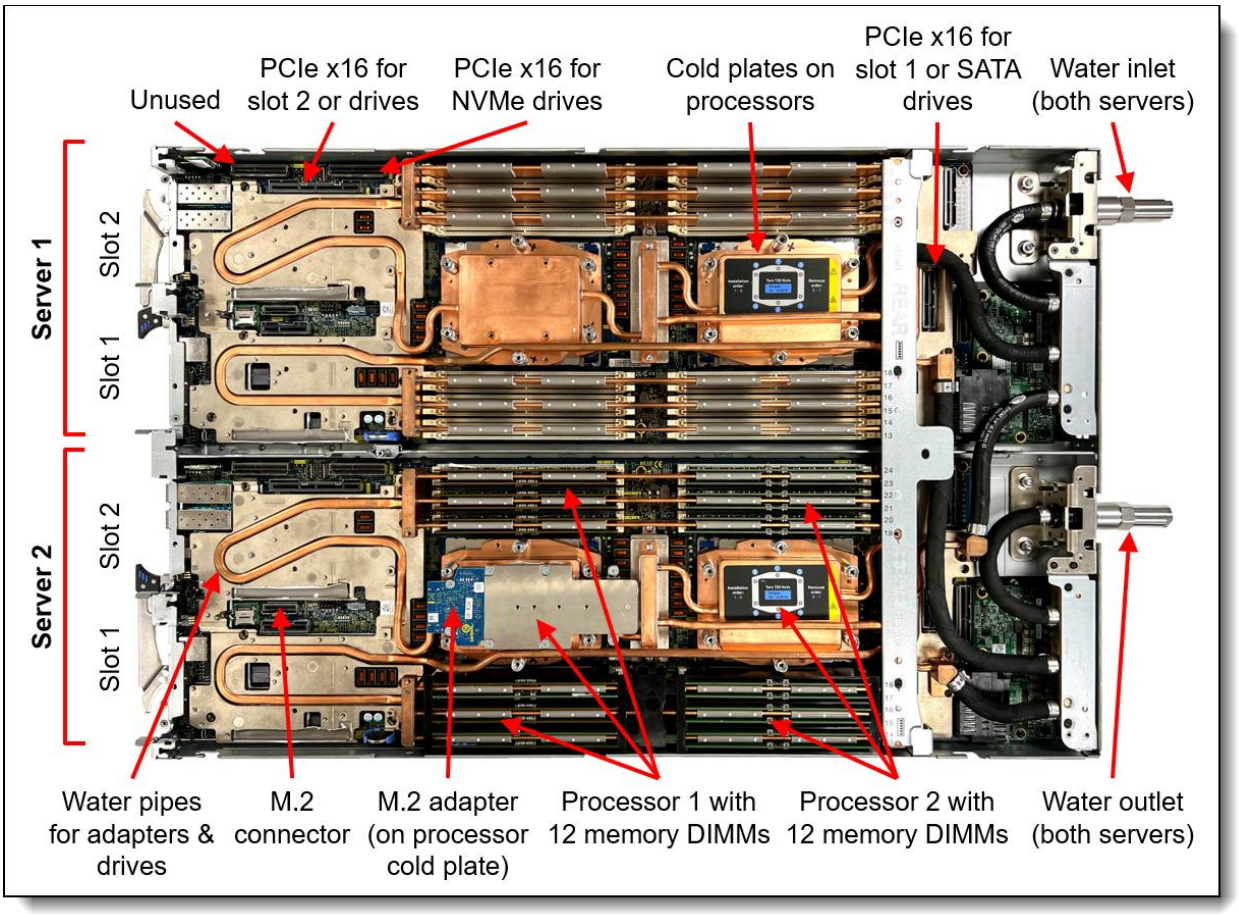

Figure 24: ThinkSystem SD665 V3 dual-server tray – top view

Each of the two side-by-side servers / nodes in the SD665 V3 tray supports two fifth-generation AMD EPYC processors with 24 * TruDDR5 6400 MHz DIMMs, up to two PCIe 5.0 slots for high-speed I/O, and up to two drive bays, in a half-wide 1U form factor.

Supported combinations of PCIe 5.0 x16 slots and SSDs are:

- One PCIe 5.0 x16 slot and either two 7mm SSDs or two E3.S EDSFF SSDs
- One PCIe 5.0 x16 slot and one 15mm SSD

- Two PCIe 5.0 x16 slots without SSDs (M.2 still supported)

Drives can be either SATA or high-performance NVMe drives, to maximize I/O performance in terms of throughput, bandwidth, and latency.

Each of the two nodes per tray includes one Gigabit and two 25 Gb Ethernet onboard ports for cost effective networking. High speed networking like the CN5000 Omni-Path High Speed Fabric can be added through the included PCIe slots.

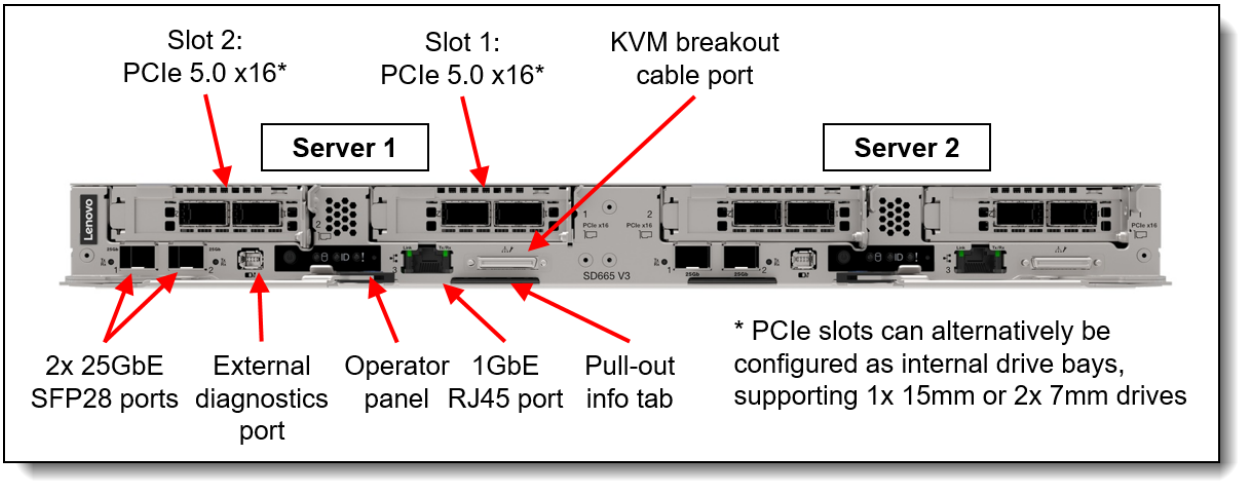

Figure 25: ThinkSystem SD665 V3 dual-server tray – front side view

The selection of an optimized CPU for CAE workloads and corresponding memory depends on different parameters. For example, license cost for CAE software is often based on core count. This may drive a decision towards lower core count CPUs, resulting in a higher memory bandwidth per core. On the other hand, large models with strong scaling may benefit from higher core count CPUs.

For the general CAE use case on AMD CPUs, Lenovo recommends two AMD 9555 (48 cores, 300W) and 768GB of memory. For strong scaling models, alternatives are the AMD 9555 (64 cores, 400W) or AMD 9655 (96 cores, 400W).

For this Reference Architecture, we selected two AMD 9555 64 Core CPUs per node, which provide a large number of Zen5 Cores, optimized for strong scaling models. The high core count number is matched by 24 * 64GB 6400 MHz RDIMMs per Compute Node for 1.5 TB of memory capacity and high bandwidth. The actual CPU selection for a specific project should be based on the individual requirements – half the memory size and a AMD 9555 CPU may also be a good choice if there is more focus on the software license cost.

More details on the SD665 V3 and DW612S enclosure can be found in the corresponding product guide:

https://lenovopress.lenovo.com/lp1612-lenovo-thinksystem-sd665-v3-server

Bill of Materials (BOM) for AMD server-based scale-out SU

The following Bill of Materials (BOM) include only the significant parts of the Scalable Unit (SU) building bloc – other parts like Risers or Power supplies have been removed for better readability. The Lenovo sales configurators DCSC and x-config will help to create a valid configuration from the following BOM lists.

In addition to the Compute Node SUs as described in this BOM, there are central components that need to be provided, for example

- Omni-Path leaf switches (for management/login/storage) and spine switches
- Ethernet leaf switches (for management/login/storage) and core switches
- Management, login and storage systems

Also adjust the SU BOM to your specific situation as needed (e.g. cable lengths from the SU to spine/core switches are 30m in this BOM – shorter or longer cable lengths may be required for connecting to the central spine/core switches are needed, depending on the location of the racks in the data center). Also, the CPU and memory selection should be adjusted to the individual requirements as needed (for guidance, see the comments in the previous section).

Part Number	Product Description	Per Qty	Tot Qty
	SD665v3		
7D9PCTOLWW	**ThinkSystem SD665 V3 Neptune DWC Tray**		**72**
C2AY	AMD EPYC 9555 64C 360W 3.2GHz Processor	4	288
CA1L	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A v2	48	3456
BPT6	ThinkSystem SD665 V3 Dual Node Tray Base	1	72
CB7C	ThinkSystem Cornelis CN5000 Omni-Path QSFP112 HFI Adapter DWC	2	144
	DW612S-aircooledPSUs		
7D1LCTO2WW	**ThinkSystem DW612S Neptune DWC Enclosure**		**12**
BKWB	ThinkSystem DW612S High Power Midplane	1	12
BE0D	N+1 Redundancy With Over-Subscription	1	12
B95Y	System Management Card	1	12
BMCA	ThinkSystem DW612S Enclosure Base	1	12
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	6	72
	2185		
5469HC1	**Lenovo Neptune DWC Manifold Assembly for 6 Enclosures w/ 2.3m hose**		**2**
BEZW	Lenovo Neptune DWC Manifold Assembly for 6 Enclosures w/ 2.3m hose	1	2
	Management main		
7D5FCTOFWW	**NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)**		**1**
BPC7	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	1	1
BVA4	Lenovo 100GBase-SR4 QSFP28 Transceiver	4	4
5AS7B07694	Lenovo EveryScale Basic Cabling Services	1	1

25G Main			
7D5FCTOLWW	**NVIDIA SN3420 25GbE Managed Switch with Cumulus (oPSE)**		**3**
BUZ3	NVIDIA SN3420 25GbE Managed Switch with Cumulus (oPSE)	1	3
BVA4	Lenovo 100GBase-SR4 QSFP28 Transceiver	4	12
5AS7B07695	Lenovo EveryScale Advanced Cabling Services		1

OPA-400 Main			
7DMQCTO2WW	**Cornelis CN5000 48-Port Omni-Path 400Gbps Air-Cooled Switch oPSE**		**3**
CC44	Cornelis CN5000 48-Port Omni-Path 400Gbps Air-Cooled Switch oPSE	1	3
5AS7B07694	Lenovo EveryScale Basic Cabling Services		1

1			
1410O42	**Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet**		**1**
BHC4	Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet	1	1
BJPD	21U Front Cable Management Bracket	2	2
C0DC	1U 12 C19/C13 Switched and monitored 32A 3P WYE PDU V2	6	6
BHC7	ThinkSystem 42U Onyx Heavy Duty Rack Side Panel	2	2
C4ZA	ThinkSystem 42U Onyx 180mm Advanced Rack Extension Kit	1	1
BJPA	ThinkSystem 42U Onyx Heavy Duty Rack Rear Door	1	1
5AS7B07695	Lenovo EveryScale Advanced Cabling Services	1	1
5AS7B07693	Lenovo EveryScale Rack Setup Services	1	1

2			
1410O42	**Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet**		**1**
BHC4	Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet	1	1
BJPD	21U Front Cable Management Bracket	2	2
C0DC	1U 12 C19/C13 Switched and monitored 32A 3P WYE PDU V2	6	6

BHC7	ThinkSystem 42U Onyx Heavy Duty Rack Side Panel	2	2
C4ZA	ThinkSystem 42U Onyx 180mm Advanced Rack Extension Kit	1	1
BJPA	ThinkSystem 42U Onyx Heavy Duty Rack Rear Door	1	1
5AS7B07693	Lenovo EveryScale Rack Setup Services	1	1
5AS7B07695	Lenovo EveryScale Advanced Cabling Services	1	1
	Software		
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)		144
	Cables		
A1MT	10m Green Cat6 Cable		16
3798	3m Green Cat5e Cable		14
CDHD	Cornelis 30m CN5000 400G QSFP112 AOC Active Cable-Straight		72
AV1H	Lenovo 10m 25G SFP28 Active Optical Cable		24
AV1W	Lenovo 1m Passive 25G SFP28 DAC Cable		98
AV2A	Lenovo 30m MPO-MPO OM4 MMF Cable		8
AV1F	Lenovo 3m 25G SFP28 Active Optical Cable		22
CDHG	Cornelis 5m CN5000 400G QSFP- 2x200G QSFP56 ACC Active Cable-Split		12
CDHB	Cornelis 1.5m CN5000 400G QSFP- 2x200G QSFP56 DAC Passive Cable-Split		10
CDHC	Cornelis 1m CN5000 400G QSFP112 - 2x200G QSFP56 DAC Passive Cable-Split		50

Table 9: Bill of Materials (BOM) for AMD server-based scale-out SU

Summary

In this paper we introduced Computer Aided Engineering (CAE) workloads as CFD (Computational Fluid Dynamics) and FEA (Finite Elements Analysis) and their specific

requirements. We described, how High Performance Computing (HPC) technologies like the Cornelis Networks CN5000 Omni-Path High Performance Fabric and MRDIMM Memory technologies can improve the performance of CAE applications and optimize the efficiency of the HPC cluster for running those workloads. We then described Reference Architectures, based on Lenovo ThinkSystem servers for Intel and AMD CPUs, for both small and medium business as well as for large scale-out systems.

Realizing those Reference Architectures with Lenovo EveryScale helps customers with the successful implementation of a CAE solution that best fits their needs and is optimized for performance and efficiency.

Appendix

Cornelis CN5000 Omni-Path Fabric components with Lenovo Feature Code

Lenovo OPN	Lenovo FC	Lenovo Description
7DMQCTO1WW	CC43	Cornelis CN5000 48-Port OPA 400Gbps Air-Cooled PSE
7DMQCTO2WW	CC44	Cornelis CN5000 48-Port OPA 400Gbps Air-Cooled Switch oPSE
7DMQCTO4WW	CC46	Cornelis CN5000 48-Port OPA 400Gbps Liquid-Cooled oPSE
4XC7B00020	C5MY	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter
4XC7B00138	CB7C	ThinkSystem Cornelis CN5000 OPA QSFP112 HFI Adapter DWC
4X97B09871	CBPM	Cornelis 1m CN5000 400G QSFP-QSFP DAC Passive Cable-Straight
4X97B09872	CBPN	Cornelis 1.5m CN5000 400G QSFP-QSFP DAC Passive Cable-Straight
4X97B09873	CBPP	Cornelis 2m CN5000 400G QSFP-QSFP DAC Passive Cable-Straight
4X97B12425	CDQW	Cornelis 3m CN5000 400G QSFP-QSFP DAC Passive Cable-Straight

4X97B09874	CBPQ	Cornelis 5m CN5000 400G QSFP- QSFP ACC Active Cable-Straight
4X97B11687	CDHF	Cornelis 5m CN5000 400G QSFP-QSFP AOC Active Cable-Straight
4X97B11688	CDHE	Cornelis 10m CN5000 400G QSFP-QSFP AOC Active Cable-Straight
4X97B09876	CBPR	Cornelis 20m CN5000 400G QSFP-QSFP AOC Active Cable-Straight
4X97B11689	CDHD	Cornelis 30m CN5000 400G QSFP-QSFP AOC Active Cable-Straight
4X97B11690	CDHC	Cornelis 1m CN5000 400G QSFP-2x QSFP DAC Passive Cable-Split
4X97B11691	CDHB	Cornelis 1.5m CN5000 400G QSFP-2x QSFP DAC Passive Cable-Split
4X97B11692	CDHA	Cornelis 2m CN5000 400G QSFP-2x QSFP DAC Passive Cable-Split
4X97B11937	CDHH	Cornelis 3m CN5000 400G QSFP-2x QSFP ACC Active Cable-Split
4X97B11938	CDHG	Cornelis 5m CN5000 400G QSFP-2x QSFP ACC Active Cable-Split

Table 10: Cornelis CN5000 Omni-Path parts and Lenovo Feature Code overview

Table of Figures

Table of Tables

Authors

Karsten Kutzer is Principal HPC/AI Solution Architect in the HPC Solutions and Server strategy department at Lenovo. He has a history of 25 years working in HPC, starting with deploying HPC clusters, then moving on to a role as an HPC architect. Since 2015 he has worked as HPC solution architect in Lenovo. His experience covers a wide range of topics from large-scale HPC clusters, servers, networking, storage and software stack as well as datacenter infrastructure and direct-water-cooling. He holds a degree of "Diplom Ingenieur (Technische Informatik)" from the Berufsakademie Mannheim.

Kevin Dean is the Senior Manager of the HPC and AI Performance and Operations Team within the Infrastructure Solutions Group at Lenovo. The role consists of leading the HPC performance engineering process, operations, and strategy as well as providing CAE application performance support and leading the customer support for the manufacturing vertical. Kevin has 8 years of experience in HPC and AI application performance support at Lenovo plus 12 years of aerodynamic design and computational fluid dynamics experience in the US defense and automotive racing industries. Kevin holds an MS degree in Aerospace Engineering from the University of Florida and a BS degree in Aerospace Engineering from Virginia Polytechnic Institute and State University.

Special thanks to Gilad Berman, David Decastro and Martin Hiegl and the Cornelis Networks team for providing input and feedback to this paper.