Lenovo Hybrid AI 221 Platform Guide
Product Guide

The evolution of AI in business is no longer a one-size-fits-all solution. While large-scale AI factories are essential for developing and training complex models, many real-world applications require a more agile and targeted approach. The true value of AI often lies in its ability to deliver instant insights and decisions directly where they're needed, at the edge.

This shift has created a demand for nimble, right-sized solutions dedicated to inference. These solutions allow enterprises to deploy pre-trained models for real-time analysis without the significant overhead of a large training infrastructure. By bringing AI processing closer to the data source, companies can achieve ultra-low latency, reduce data transfer costs, and maintain control over sensitive information.

This document explores a compact, powerful platform designed specifically for these demands, enabling you to harness the transformative power of AI in a focused and highly efficient way.

Lenovo Hybrid AI 221 is a platform that enables enterprises of all sizes to quickly deploy smaller scale but scalable AI infrastructure. Lenovo Hybrid AI 221 supports Enterprise AI use cases as either a new, development, greenfield environment or an extension of their existing IT infrastructure.

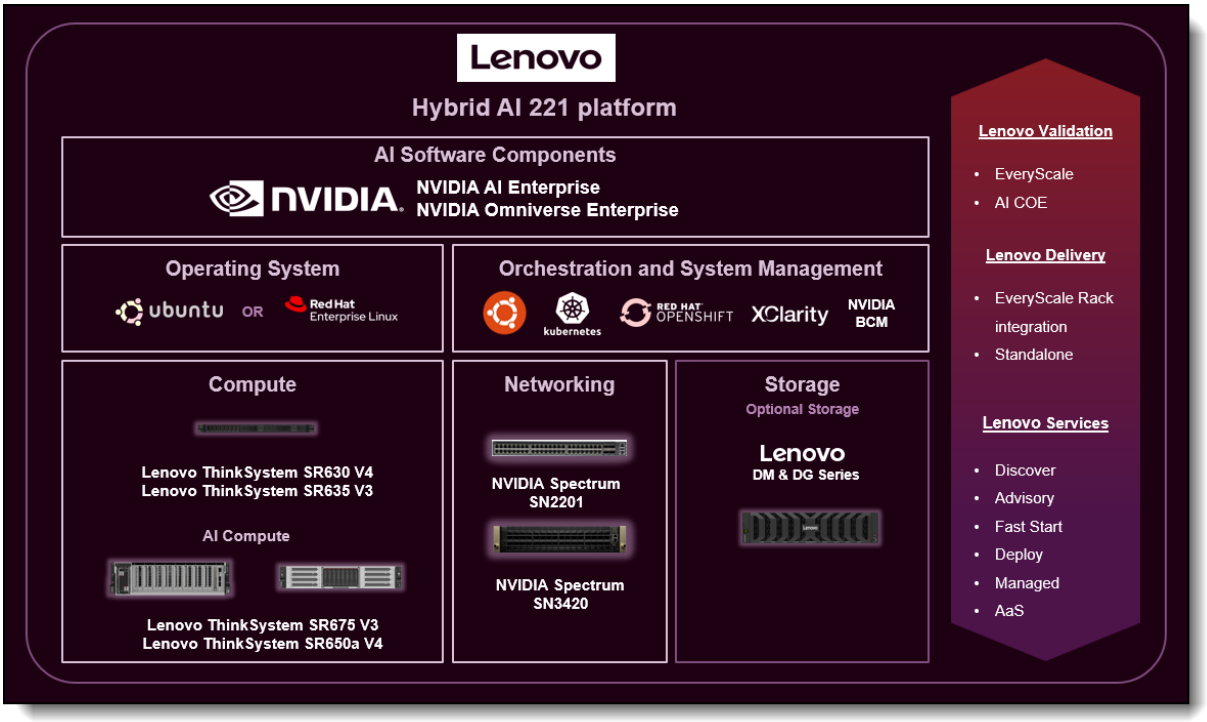

Figure 1. Lenovo Hybrid AI 221 platform overview

The Lenovo Hybrid AI 221 platform can scale from a single ThinkSystem SR675 V3 server or ThinkSystem SR650a V4 with just 2 GPUs. For deployments leveraging the SR675 V3 server, additional GPUs can be added through vertical scaling by adding more GPUs and/or horizontal scaling by adding additional nodes. With the integration of NVIDIA RTX 6000 PRO Blackwell Server Edition GPUs into the Lenovo Hybrid AI 221 platform, GPU-accelerated retrieval is faster, more intelligent, and more efficient for faster time-to-value for Retrieval Augmented Generation (RAG) and Agentic AI operations.

The Lenovo Hybrid AI 221 platform will also align with the new Lenovo Hybrid AI Software platform which gives customers the necessary software and services pieces for enterprises to build and deploy their own on-premises AI factory. Once deployed, the Lenovo Hybrid AI 221 platform will help enterprises deploy a wide-range of AI-enabled enterprise applications, including agentic and physical AI workflows, autonomous decision-making, and real-time data analysis.

The Lenovo Hybrid AI 221 comes with the EveryScale Solution verified interoperability for the tested best recipe hardware and software stack. Additionally, EveryScale allows Lenovo Hybrid AI platform deployments to be delivered as fully pre-built, rack-integrated systems that are ready for immediate use.

Components

The main hardware components of Lenovo Hybrid AI platforms are AI Compute nodes, Service Nodes and the Networking infrastructure. As an integrated solution they can come together in either a Lenovo EveryScale Rack (Machine Type 1410) or Lenovo EveryScale Client Site Integration Kit (Machine Type 7X74).

Topics in this section:

AI Compute Node – SR675 V3

The AI Compute Node leverages the Lenovo ThinkSystem SR675 V3 GPU-rich server.

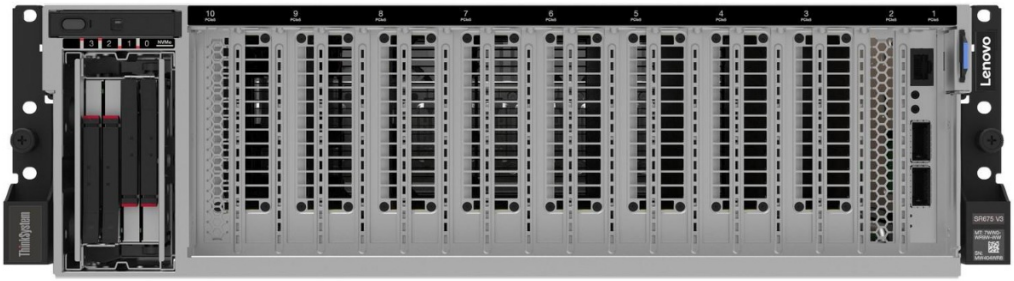

Figure 2. Lenovo ThinkSystem SR675 V3 in 8DW PCIe Setup

The SR675 V3 is a 2-socket 5th Gen AMD EPYC 9005 server supporting up to 8 PCIe DW GPUs with up to 5 network adapters in a 3U rack server chassis. The server supports 2-2-1 configuration with room to grow. This makes it the ideal choice for Enterprises who want to start with small number of GPUs, but with options to vertically scale up to 4-8 GPUs in the configuration before needing to add additional nodes.

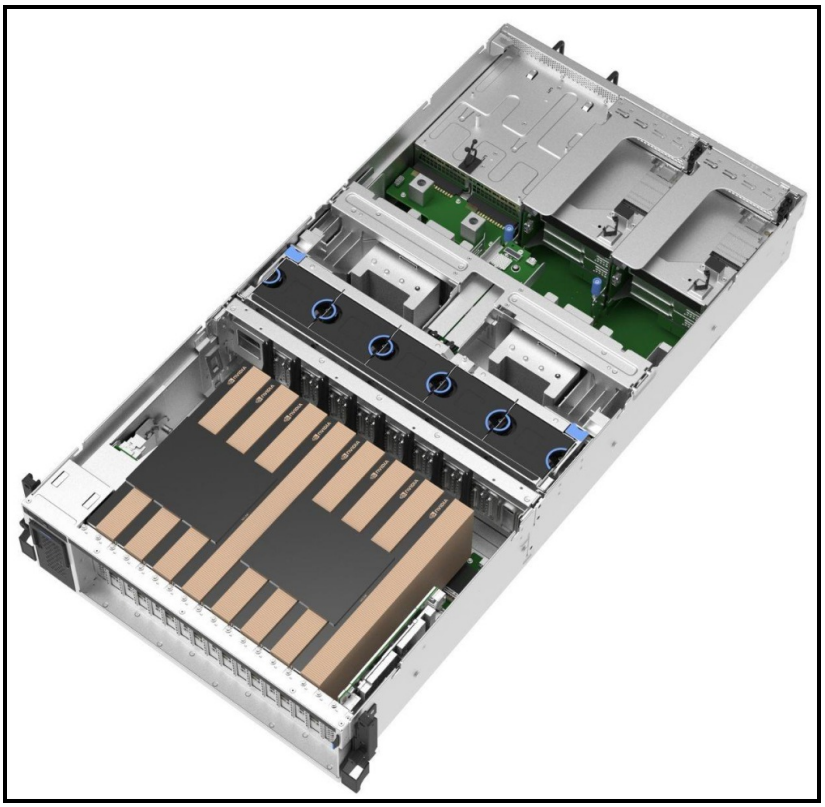

Figure 3. ThinkSystem SR675 V3 in an 8x DW PCIe GPU setup

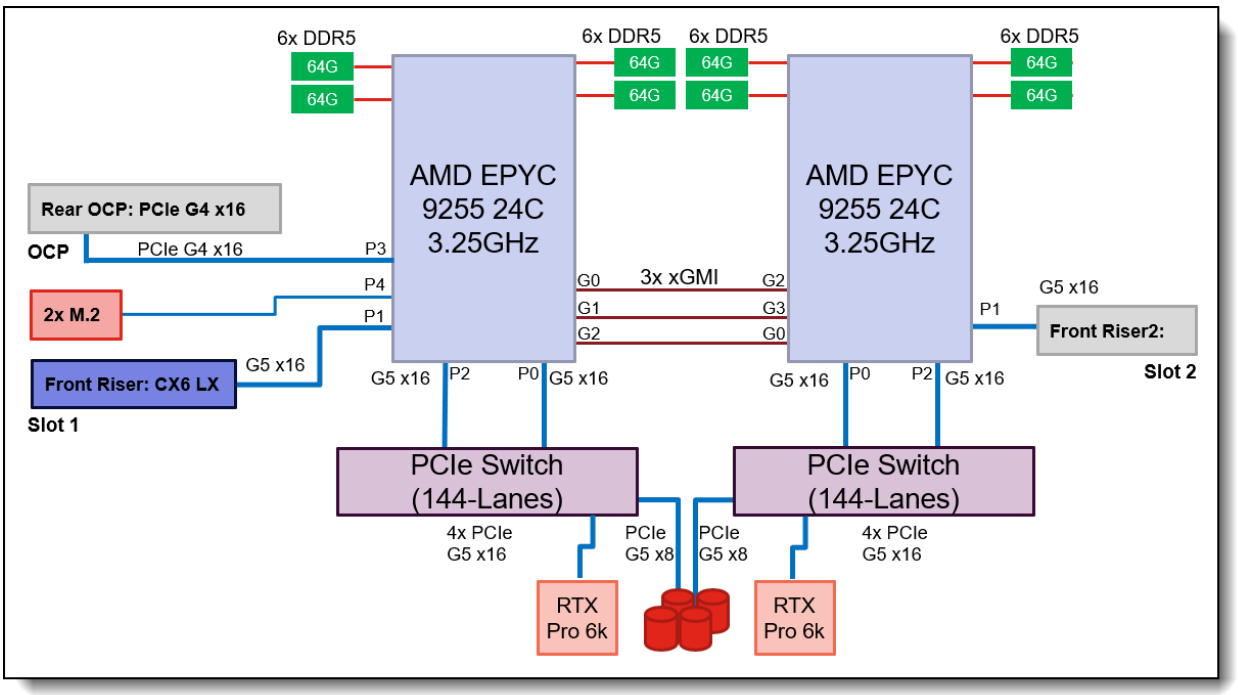

Figure 4. 221 AI Compute Node Block Diagram: SR675 V3 with RTX Pro 6000 Blackwell Server Edition

The AI Compute node is configured minimum with two **AMD EPYC 9255 24C 200W 3.25GHz** to support tw GPUs. When future GPU expansion is considered, two **AMD EPYC 9535 64 Core 2.4 GHz** processors with an all-core boost frequency of 3.5GHz can be used instead. Besides providing consistently more than 2GHz frequency this ensures that with 4 Multi Instance GPUs (MIG) in the case of RTX Pro 6000 Blackwell Server Edition on the 2 physical GPUs there are 2 Cores available per MIG plus a few additional Cores for Operating System and other operations. A total of 512GB of system memory is allocated to support the requirements of the two GPUs and additional requirements of the OS, Kubernetes layer, and applications.

The server is configured with ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port PCIe Ethernet Adapter installed on the Front Riser Slot 1 to support Converged Network or front-end connectivity. When out-of-band management HA is required, an optional ThinkSystem Intel I350 1GbE RJ45 4-Port OCP Ethernet Adapter V2 NIC can be configured. 221 platform is built to use internal storage by default. If integration to external storage is required, higher network throughput will be supported by installing higher throughput adapters such as ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter. Please refer to the AI starter kit deployment on the Lenovo Hybrid AI 285 Platform Guide for more details.

The GPUs are connected to the CPUs via two PCIe Gen5 switches, each supporting up to four GPUs. In 221 platform, with the **NVIDIA H200 NVL PCIe GPU**, two GPUs are additionally interconnected through a 2-way NVLink bridge, creating a unified memory space. With the **RTX PRO 6000 Blackwell Server Edition and L40S** GPUs, no NVLink bridge is applicable.

AI Compute Node – SR650a V4

The Lenovo ThinkSystem SR650a V4 is an ideal 2-socket 2U rack server for customers want to maximize GPU compute power while still retaining the traditional 2U rack form factor. With two Intel Xeon 6700-series or 6500-series processors, plus support for two **NVIDIA L40S**, two **H200 NVL**, or two **RTX PRO 6000 Blackwell Server Edition** GPUs in 221 platform configuration. SR650a V4 is also designed for high density and scale-out workloads. For scaling up, horizontal scaling by adding additional nodes is required.

The Lenovo ThinkSystem SR650a V4 is designed to accelerate GPU-intensive workloads like AI, deep learning, and HPC. Powered by Intel Xeon 6 processors, it delivers exceptional GPU density, advanced storage, and PCIe Gen5 connectivity for peak performance. It supports up to 4x double-width or 8x single-width GPUs and offers up to 8x E3.S NVMe drives for fast, low-latency data access.

Figure 5. Lenovo ThinkSystem SR650a V4

The following figure shows the locations of key components inside the server.

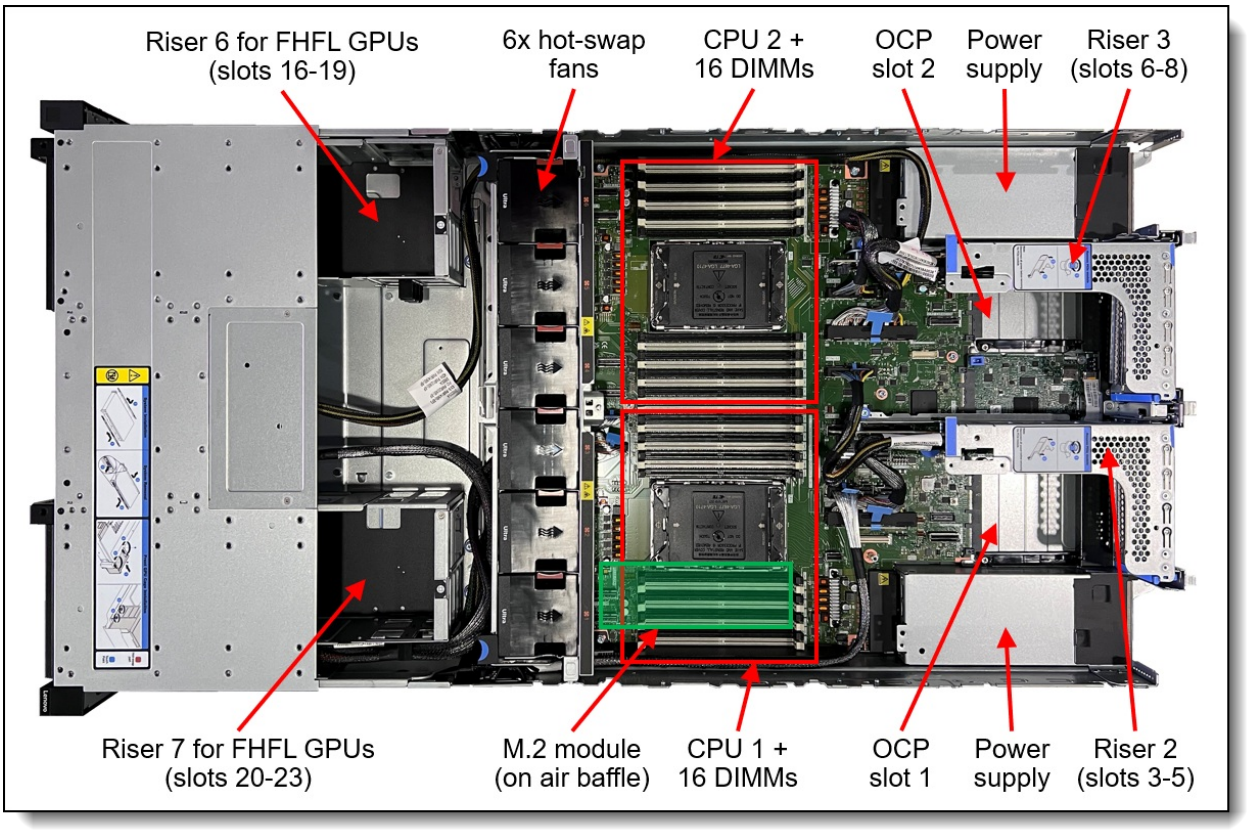

Figure 6. Internal view of the ThinkSystem SR650a V4

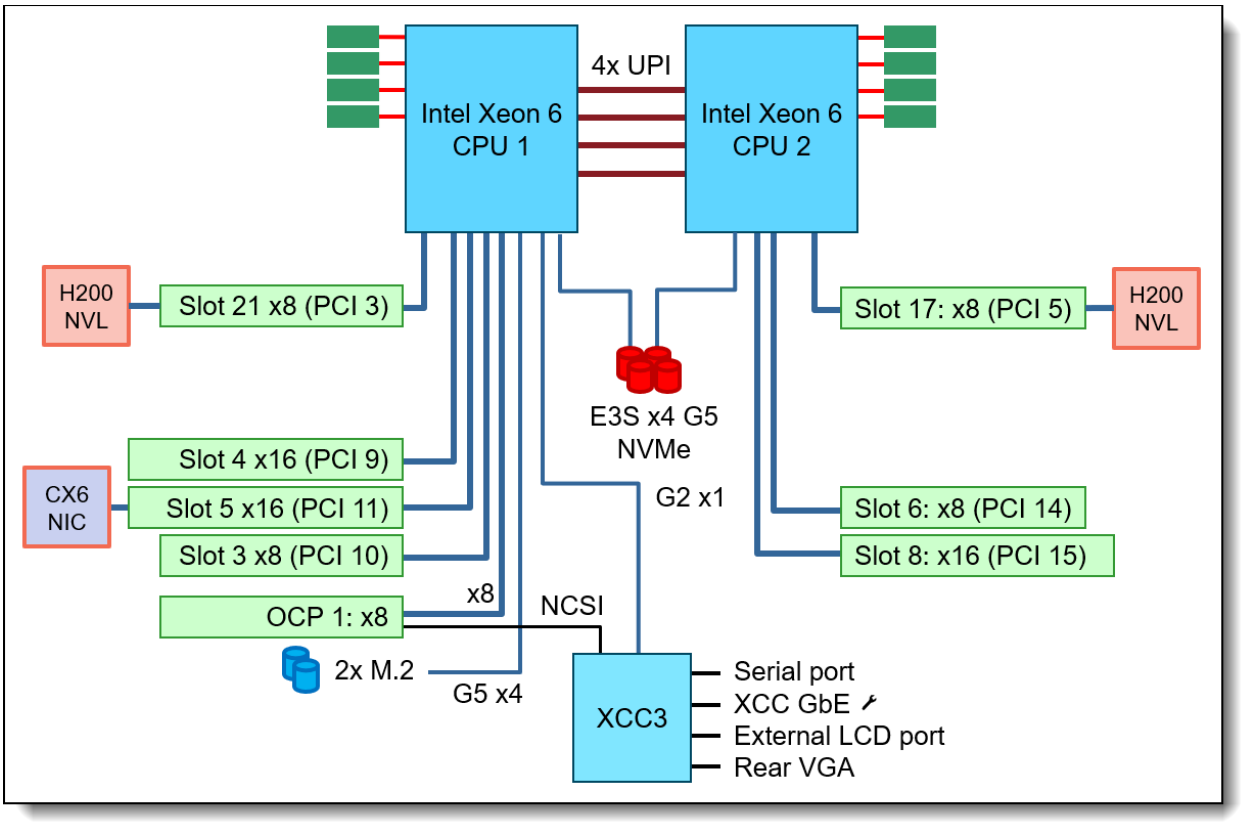

Figure 7. SR650a V4 system architectural block diagram

The AI Compute node is configured with two **Intel Xeon 6530P 32C 225W 2.3GHz** Processor. With 8 Memory Channels per processor socket the Intel based server provides superior Memory bandwidth ensuring highest performance. A total of 512GB of system memory is allocated to support the requirements of the two GPUs and additional requirements of the OS, Kubernetes layer, and applications.

The server is configured with ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port PCIe Ethernet Adapter installed on the Read Direct Riser Slot 5 to support Converged Network or front-end connectivity. When out-of-band management HA is required, an optional ThinkSystem Broadcom 5719 1GbE RJ45 4-port OCP Ethernet Adapter can be configured. The 221 platform is built to use internal storage by default. If integration to external storage is required, higher network throughput will be supported by installing higher throughput adapters such as ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter.

In the 221 platform, the GPUs are installed on Riser 6 and 7 for FHFL GPUs, slot 17 and 21. With the **NVIDIA H200 NVL** PCIe GPU, two GPUs are additionally interconnected through a 2-way NVLink bridge, creating a unified memory space. With the **RTX PRO 6000 Blackwell Server Edition** and **L40S GPUs**, no NVLink bridge is applicable.

GPU selection

The Hybrid AI 221 platform on both the SR675 V3 and SR650a V4 is designed to handle any of NVIDIA's DW PCIe form factor GPUs including the new NVIDIA RTX PRO 6000 Blackwell Server Edition, NVIDIA H200 NVL, and NVIDIA L40S.

- **NVIDIA H200 NVL.** The NVIDIA H200 NVL is a powerful GPU designed to accelerate both generative AI and high-performance computing (HPC) workloads. It boasts a massive 141GB of HBM3e memory, which is nearly double the capacity of its predecessor, the H100. This increased memory, coupled with a 4.8 terabytes per second (TB/s) memory bandwidth, enables the H200 NVL to handle larger and more complex AI models, like large language models (LLMs), with significantly improved performance. In addition, the H200 NVL is built with energy efficiency in mind, offering increased performance within a similar power profile as the H100, making it a cost-effective and environmentally conscious choice for businesses and researchers. NVIDIA provides a 5-year license to NVIDIA AI Enterprise free-of-charge bundled with NVIDIA H200 NVL GPUs. When using NVIDIA H200 NVL in 221 platform, it is recommended to have minimum **282 GB of system memory**.
- **NVIDIA RTX PRO 6000 Blackwell Server Edition.** Built on the groundbreaking NVIDIA Blackwell architecture, the NVIDIA RTX PRO 6000 Blackwell Server Edition delivers a powerful combination of AI and visual computing capabilities to accelerate enterprise data center workloads. Equipped with 96GB of ultra- fast GDDR7 memory, the NVIDIA RTX PRO 6000 Blackwell provides unparalleled performance and flexibility to accelerate a broad range of use cases- from agentic AI, physical AI, and scientific computing to rendering, 3D graphics, and video. When usuing NVIDIA RTX PRO 6000 Blackwell Server Edition in 221 platform, it is recommended to have minimum **192 GB of system memory**.
- **NVIDIA L40S.** The NVIDIA L40S is a versatile GPU designed for data centers. Built on the Ada Lovelace architecture, it features a 48GB of GDDR6 memory, significantly more than consumer-grade GPUs like the RTX 4090. This large memory capacity is crucial for handling large language models (LLMs) and other complex AI workloads, while its fourth-generation Tensor Cores and third-generation RT Cores provide a strong performance for both AI training and inference, as well as professional 3D rendering and ray tracing. The L40S is a "universal GPU" because it balances these capabilities, making it ideal for a wide range of applications from generative AI and high-performance computing (HPC) to video processing and virtual workstations. When using NVIDIA L40S GPU in 221 platform, it is recommended to have minimum **96 GB of system memory**.

Service Nodes – SR630 V4 or SR635 V3

When deploying the Hybrid AI 221 platform with NVIDIA AI Enterprise software stack it is recommended to add an additional service node to run NVIDIA Base Command Manager as described further in the AI Software Stack chapter.

Key difference from the base platform: Because the architecture for this platform does not yet leverage Spectrum-X, there is no need for Bluefields within the service nodes. For this reason, the customer can leverage the lower cost SR635 V3 or SR630 V4 instead of the SR655 V3.

For the Container operations optionally three **Master Nodes** build the Kubernetes control plane providing redundant operations and quorum capability. Without the three master nodes, the AI computes will function as both master and worker nodes.

Figure 8. Lenovo ThinkSystem SR635 V3

The Lenovo ThinkSystem SR635 V3 is an optimal choice for a homogeneous host environment, featuring a single socket AMD EPYC 9335 with 32 cores operating at 3.0 GHz base with an all-core boost frequency of 4.0GHz. The system is fully equipped with twelve 32GB 6400MHz Memory DIMMs, two 960GB Read Intensive M.2 drives in RAID1 configuration for the operating system, and two 3.84TB Read Intensive U.2 drives for local data storage.

Networking

The default setup of the Lenovo Hybrid AI 221 platform leverages NVIDIA SN3420 switch for Converged Network and the NVIDIA SN2201 switch for out-of-band management. As an alternative, Cisco Networking with the Cisco Nexus 93240YC-FX2 for the Converged Network and the Nexus 9300-FX3 for the Management Network can also be used. A 10/25GE adapter with minimum 2 ports is added to both AI computes and service nodes.

The **Converged Network** handles in-band management, linking the Enterprise IT environment, and optional storage integration to the 221 platform. Built on Ethernet with RDMA over Converged Ethernet (RoCE), it supports current and new cloud and storage services as outlined in the AI Compute node configuration.

In addition to providing access to the AI agents and functions of the AI platform, this connection is utilized for all data ingestion from the Enterprise IT data during indexing and embedding into the Retrieval-Augmented Generation (RAG) process. It is also used for data retrieval during AI operations. In 221 platform, the Converged Network will use 10/25 GE connections.

The **Out-of-Band (Management) Network** encompasses all AI Compute node and BlueField-3 DPU base management controllers (BMC) as well as the network infrastructure management.

Converged Network Switch: NVIDIA SN3420

The NVIDIA SN3420 is a high-performance, compact 1U Ethernet switch from the Spectrum-2 SN3000 series, designed primarily as an ideal Top-of-Rack (ToR) switch for modern data centers and cloud infrastructures. It provides versatile connectivity with 48 ports of 10/25GbE (SFP28) for host connectivity migration from 10G to 25G, and 12 ports of up to 100GbE (QSFP28) for high-speed uplinks in leaf-spine topologies, delivering a total throughput of up to 4.8 Tb/s and a processing capacity of 3.58 billion packets per second (Bpps).

The SN3420 platforms deliver high performance, consistent low latency along with support for advanced software defined networking features, making it the ideal choice for AI data center fabric application. The SN3420 is installed with Cumulus Linux, emphasizing network disaggregation and software-defined networking features, including advanced Layer 2/3 routing, RoCE (RDMA over Converged Ethernet), and NVIDIA What Just Happened (WJH) telemetry for deep network visibility.

Figure 9. NVIDIA SN3420

Table 1. NVIDIA SN3420 configuration

Part number	Description	Quantity
7D5FCTOKWW	NVIDIA SN3420 25GbE Managed Switch with Cumulus (PSE)	
BUZ2	NVIDIA SN3420 25GbE Managed Switch with Cumulus (PSE)	1
6201	1.5m, 10A/100-250V, C13 to C14 Jumper Cord	2
5WS7B98268	5Yr Premier NBD Resp NVID SN2201 PSE	1

Out-of-Band Management Network Switch: NVIDIA SN2201

The NVIDIA SN2201 is ideal as an out-of-band (OOB) management switch or as a ToR switch connecting up to 48 1G Base-T host ports with non-blocking 100GbE spine uplinks. Featuring highly advanced hardware and software along with ASIC-level telemetry and a 16 megabyte (MB) fully shared buffer, the SN2201 delivers unique and innovative features to 1G switching.

Figure 10. NVIDIA SN2201

The following table lists the configuration of the NVIDIA Spectrum SN2201.

Table 2. NVIDIA SN2201 configuration

Part number	Description	Quantity
7D5FCTOFWW	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	
BPC7	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	1
6201	1.5m, 10A/100-250V, C13 to C14 Jumper Cord	2
5WS7B98268	5Yr Premier NBD Resp NVID SN2201 PSE	1

Alternative Converged Network Switch: Cisco Nexus 9300-FX2 Series Switch

As mentioned above, when there is no need to integrate any external storage or to provide East-West Network, a lower bandwidth switch like Cisco Nexus 93240YC-FX2 can be used. The Cisco Nexus 93240YC-FX2 is a high-performance, fixed-port switch designed for modern data centers. It features 48 downlink ports that support 1/10/25-Gbps speeds and 12 uplink ports configurable for 40/100-Gbps, providing a total bandwidth of 4.8 Tbps.

This switch is built on Cisco's Cloud Scale technology, ensuring non-oversubscribed performance in a compact 1.2RU form factor. It supports both NX-OS and ACI modes, offering flexibility for traditional and automated, policy-based network environments. With advanced telemetry, real-time buffer utilization, and extensive programmability, the Nexus 93240YC-FX2 is ideal for scalable, secure, and efficient data center operations.

Figure 11. Cisco Nexus 93240YC-FX2

Table 3. Cisco Nexus 93240YC-FX2 configuration

Part number	Description	Quantity
7DL5CTO1WW	Cisco Nexus 9300-FX2 Series, 48x 25Gb Switch (N9K-C93240YC-FX2)	
C78T	Cisco Nexus 9300-FX2 Series, 48x 25Gb Switch (N9K-C93240YC-FX2)	1
C6FK	Mode selection between ACI and NXOS (MODE-NXOS)	1
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	2
C1P1TN9300XF-5Y	5 Years (60 months) Cisco software Premier license	1

Alternative Out-of-Band Management Switch: Cisco Nexus 9300-FX3 Series Switch

The Cisco Nexus 93108TC-FX3P is a high-performance, fixed-port switch designed for modern data centers. It features 48 ports of 100M/1/2.5/5/10GBASE-T, providing flexible connectivity options for various network configurations. Additionally, it includes 6 uplink ports that support 40/100 Gigabit Ethernet QSFP28, ensuring high-speed data transfer and scalability.

Built on Cisco's CloudScale technology, the 93108TC-FX3P delivers exceptional performance with a bandwidth capacity of 2.16 Tbps and the ability to handle up to 1.2 billion packets per second (Bpps).

This switch also supports advanced features such as comprehensive security, telemetry, and automation capabilities, which are essential for efficient network management and troubleshooting.

Figure 12. Cisco Nexus 93108TC-FX3P

Table 4. Cisco Nexus 93108TC-FX3P configuration

Part number	Description	Quantity
7DL8CTO1WW	Cisco Nexus 9300-FX3 Series Switch (N9K-C93108TC-FX3)	
C5PB	Cisco Nexus 9300-GX2 Series Switch (N9K-C9364D-GX2A)	1
C6FK	Mode selection between ACI and NXOS (MODE-NXOS)	1
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	2
C1P1TN9300XF-5Y	5 Years (60 months) Cisco software Premier license	1

Lenovo EveryScale Solution

The Server and Networking components and Operating System can come together as a Lenovo EveryScale Solution. It is a framework for designing, manufacturing, integrating and delivering data center solutions, with a focus on High Performance Computing (HPC), Technical Computing, and Artificial Intelligence (AI) environments.

Lenovo 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, to create 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.

Optional Storage Integration

Lenovo Hybrid AI 221 platform does not come by default with storage integration. When required the platform can be integrated to an optional external storage solution.

For enterprise organizations AI applications are treated as any other workload and require the same data management and enterprise data security features. The Lenovo ThinkSystem DM7200F and DG5200 platforms provide all flash storage that prepares your infrastructure and data for AI workloads and are included as the storage layer in the AI Starter Kits.

Benefits of ThinkSystem DM & DG storage for GenAI & RAG:

- Enterprise Security features including autonomous ransomware protection
- Deduplication and compression
- All flash performance
- Flexible scaling
- Unified file, object, and block eliminates data silos

Tip: For Enterprise environments that currently utilize NetApp, a leading provider in Enterprise Storage, the Lenovo Hybrid AI platforms offer the perfect compatible compute environment for Netapp AIPod as referenced by the NetApp Verified Architecture “NetApp AI Pod with Lenovo”

Architecture

The 221 platform is targeted for single node and smaller deployments.

- Single Node Deployment
- Vertical Scaling with SR675 V3 (Scale Up) and Storage Integration
- Horizontal Scaling (Scale out)
- Custom Deployment

Single Node Deployment

The 221 platform scales form a Single Node deployment with either SR675 V3 or SR650a V4. An additional service node is added to run NVIDIA Base Command Manager (BCM) and login server. Since 221 platform is targeted at single node and smaller deployments by default to save costs and complexity, networking on the 221 platform is configured without HA and without integration to storage or East-West network. A single NVIDIA SN3420 switch is used for for this configuration with 10/25G Ethernet connections to the computes. NVIDIA SN2201 switch functions as the out-of-band management switch.

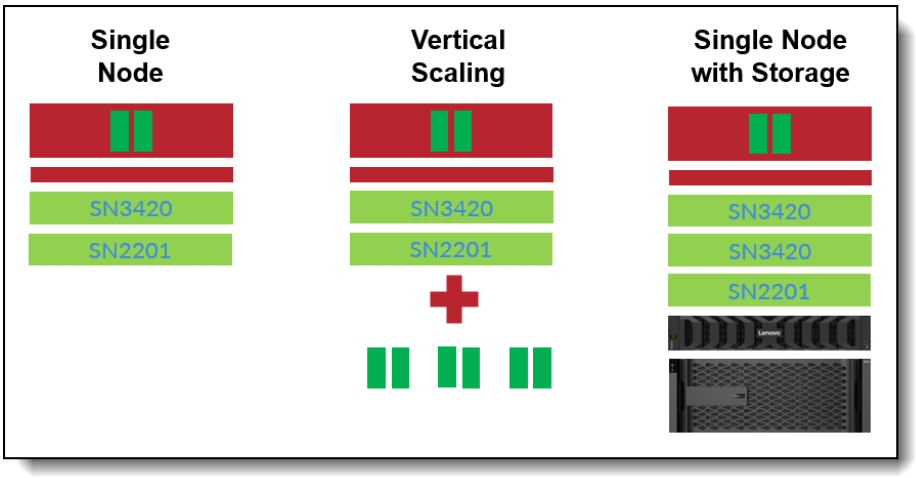

Figure 13. Single Node Deployment, Vertical Scaling, and Storage Integration

Vertical Scaling with SR675 V3 (Scale Up) and Storage Integration

For ThinkSystem SR675 V3 server, additional GPUs can be added through vertical scaling by adding up to 6 GPUs for a total of 8 GPUs. If required, external storage can also be integrated. When integrating to external storage, a higher throuput Converged Network will be required by replacing the 10/25G Ethernet connections to 200G connections using ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 adapter. Please refer to the AI starter kit deployment on the Lenovo Hybrid AI 285 Platform Guide for more details.

Horizontal Scaling (Scale out)

221 platform also supports horizontal scaling by adding additional AI computes. The AI computes will be connected to the Converged Network using 10/25 GE connections by default.

The 221 platform network diagram is shown below. By default 221 platform is configured without HA using a single SN3420 switch as shown in the figure below.

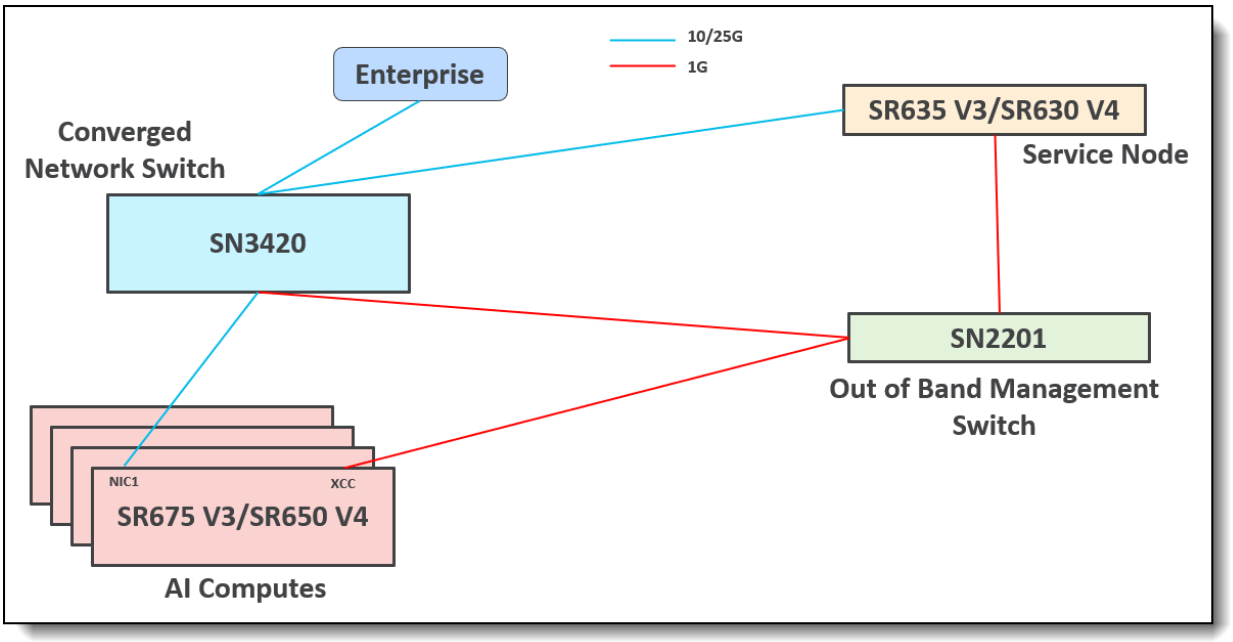

Figure 14. Network Architecture of 221

When HA is required an additional NVIDIA SN3420 can be added to provide network redundancy as shown in the following figure.

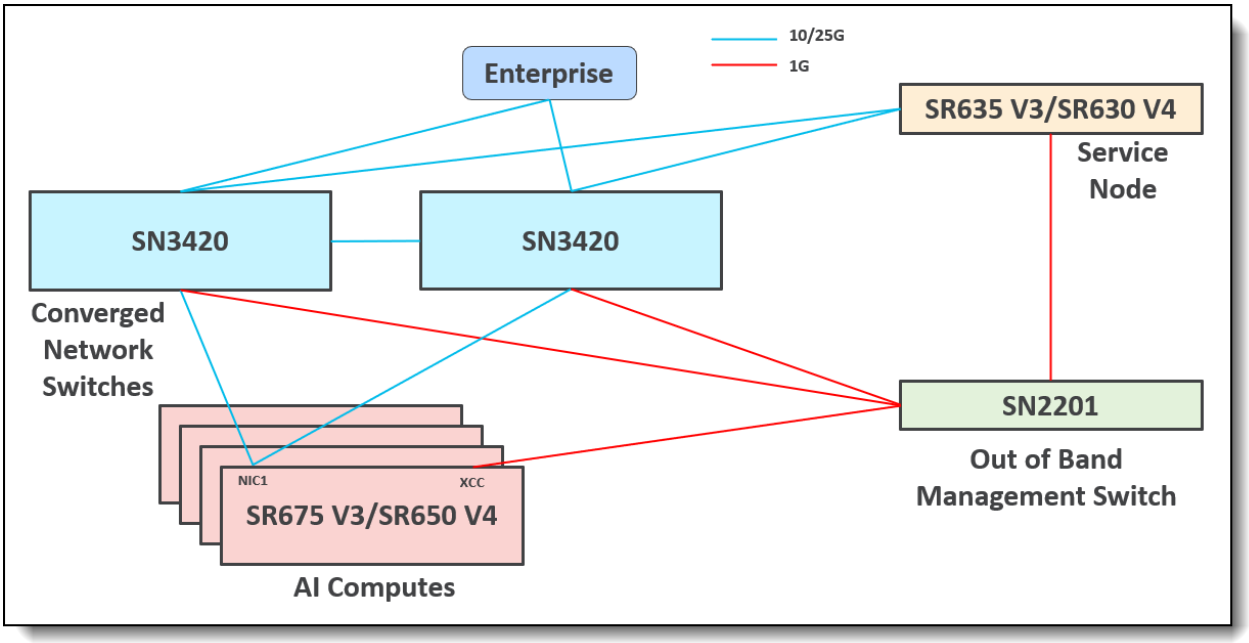

Figure 15. Network Architecture of 221 with HA

Optionally, external storage can be integrated into the 221 platform, as shown in the following figure.

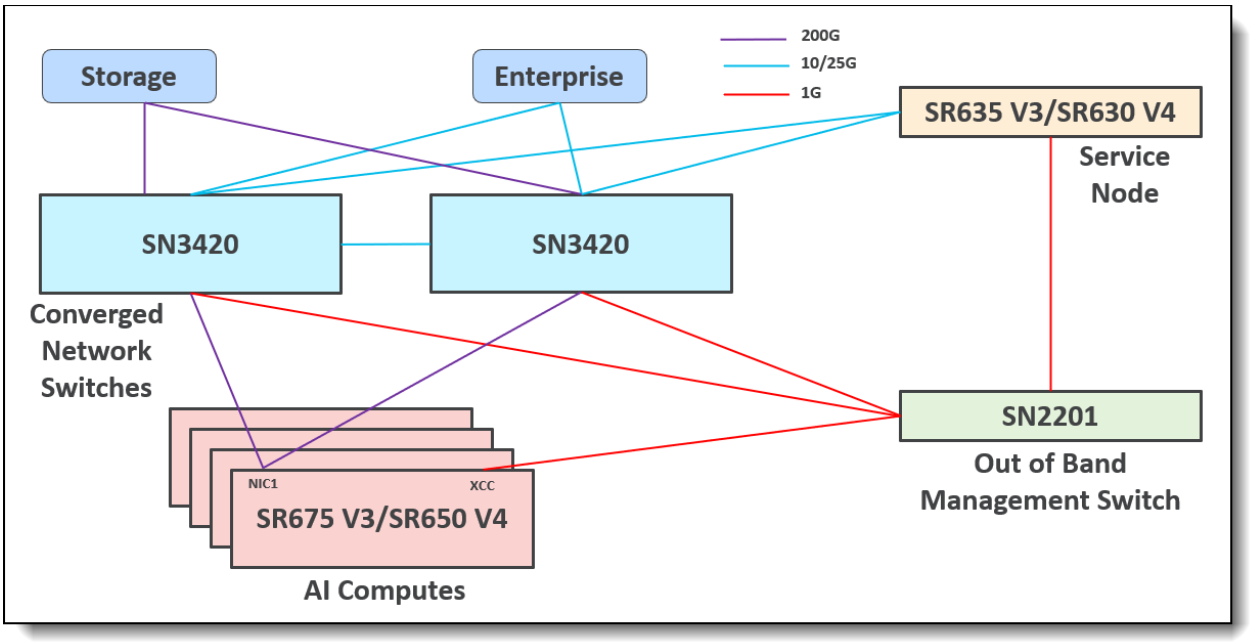

Figure 16. Network Architecture of 221 with Storage Integration

Custom Deployment

For high-end scenarios requiring more than eight scalable units, the network can be custom designed to any required size. Lenovo will develop a fully bespoke solution tailored to match the workflow and workload requirements in that case.

AI Software Stack

Deploying AI to production involves implementing multiple layers of software. The process begins with the system management and operating system of the compute nodes, progresses through a workload or container scheduling and cluster management environment, and culminates in the AI software stack that enables delivering AI agents to users.

For the 221 platform, by default the Kubernetes control plane operates directly on the AI Compute nodes, negating the requirement for dedicated optional service nodes to run Kubernetes control plane. The AI Compute nodes will function as master-worker nodes. A minimum of one service node is still required to run NVIDIA Base Command Manager (BCM). The software stack for 221 platform is similar to the recommended full AI Software Stack, though some components may be considered optional or less practical for these smaller configurations.

Refer to the Lenovo Hybrid AI Software Platform for more details however the stack has been listed below.

Table 5. Intro

Software Role	Software Package	Multi-Node	Single Node
Bare metal Management	XClarity One Management Hub 2.0	Yes	
	XClarity One (Cloud or on-prem VM)	Yes	
Linux Operating System	Ubuntu Server 22.04.4 LTS	Yes	Yes
Container Orchestration	Upstream Kubernetes 1.31.5	Yes	Yes
Container Runtime	Containerd 1.7.23	Yes	Yes
Orchestration	NVIDIA Base Command Manager 10.0	Yes	Yes
	Prometheus 2.55.0	Yes	Yes
	Permission Manager 0.5.1	Yes	Yes
Container Network Interface (CNI)	Calico 3.27.4	Yes	Yes
Package Manager	Helm 3.16.0	Yes	Yes
Load Balancer – Control Plane	nginx 1.12.0	Yes	Yes
Load Balancer – Network Services	MetalLB 0.14.8	Yes	Yes
Operator	GPU Operator 24.9.2	Yes	Yes
	Linux GPU driver 550.127.05	Yes	Yes
	NIM Operator 1.0.1	Yes	Yes
	Network Operator 25.1.0	Yes	
DOCA Host	2.9.0-0.4.7	Yes	
Cumulus Linux (CL) NOS	5.11.0.0026	Yes	
NVIDIA Network Congestion Control Algorithm	2.9.0072-1	Yes	
NCCL	2.21	Yes	
NetQ	4.12	Yes	
Run:ai (Optional)	2.21	Yes	
Grafana	11.2.2	Yes	Yes
Storage	NFS Provisioner 4.0.2	Yes	
	Local Path Provisioner 0.0.31		Yes

In the following sections, we take a deeper dive into some the software elements:

- Lenovo XClarity One
- Linux Operating System
- Base Command Manager and Container Orchestration
- Kubernetes Layer
- Data and AI Applications
- NVIDIA AI Enterprise
- NVIDIA RAG Blueprint

Lenovo XClarity One

Lenovo XClarity One is a management-as-a-service offering for hybrid-cloud management of on-premises data-center assets from Lenovo. Local management hubs can be installed across multiple sites to collect inventory, incidents, and service data, and to provision resources, creating a bridge between devices and the XClarity One portal. The XClarity One portal provides a modern, intuitive interface that centralizes IT orchestration, deployment, automation, and support from edge to cloud, with enhanced visibility into infrastructure performance, usage metering, and analytics.

The following functions are supported by XClarity One:

- XClarity One dashboard
- Firmware management
- Security
- User Management
- Hardware Monitoring

Linux Operating System

The AI Compute nodes are typically deployed with Ubuntu Server LTS Edition which is a Linux distribution that is maintained for a minimum of 5 years by Canonical as standard, thereby reducing the need for major upgrades. All of the software in this Reference Architecture is compatible and validated with Ubuntu Server LTS edition. Customers should also consider purchasing additional support for Ubuntu Server LTS using the Ubuntu Pro Edition upgrade. Lenovo Hybrid AI platforms also support Red Hat Enterprise Linux (RHEL) which is distributed as part of a paid licensed model that automatically comes with support.

Ultimately, the choice of Linux distributions is one the customer needs to make based on their familiarity with Linux and their ability to support an unlicensed distribution v.s. a licensed distribution with contractual support.

Base Command Manager and Container Orchestration

Base Command Manager (BCM) provisions the AI environment, incorporating the components such as the Operating System, Vanilla Kubernetes (K8S), GPU Operator, and Network Operator to manage the AI workloads. BCM Supports 3 types of network topologies depending on how the user wants nodes to be accessed.

BCM orchestration capabilities with Kubernetes are a key feature. The software simplifies the lifecycle management of a Kubernetes cluster, from initial setup to ongoing operations. It automates the provisioning of compute nodes, integrates crucial components like the NVIDIA GPU Operator to expose GPUs to Kubernetes, and facilitates the scheduling of jobs on the cluster. This allows for fine-grained control over resource allocation and enables features like Multi-Instance GPU (MIG), which can partition a single physical GPU into multiple smaller, isolated instances. BCM provides a unified management platform for both the physical hardware and the container orchestration layer.

As part of 221 platform architecture minimum a single service node is required to run BCM.

Kubernetes Layer

Kubernetes is the leading AI container deployment and workload management tool in the market, which can be used for edge and centralized data center deployments.

“Vanilla Kubernetes” (an upstream, open-source build) is the recommended version of Kubernetes that has been validated as a part of this software stack and deployed by Base Command Manager. Canonical Kubernetes offers additional functionality beyond Vanilla, and is used across industries for mission critical workloads, offering up to 12 years of security for those customers who cannot, or choose not to upgrade their Kubernetes versions.

Note - when choosing Red Hat for the Linux Operating System, Red Hat OpenShift should be deployed as the container orchestration layer.

Data and AI Applications

Canonical’s Ubuntu Pro includes a portfolio of open source applications in the data and AI space including leading projects for ML space with Kubeblow and MLFlow, big data and database with Spark, Kafka, PostgreSQL, Mongo and others. Ubuntu Pro enables customers on their open source AI journey to simplify deployment and maintenance of these applications and provides security maintenance.

NVIDIA AI Enterprise

The Lenovo Hybrid AI 221 platform is designed for NVIDIA AI Enterprise, which is a comprehensive suite of artificial intelligence and data analytics software designed for optimized development and deployment in enterprise settings.

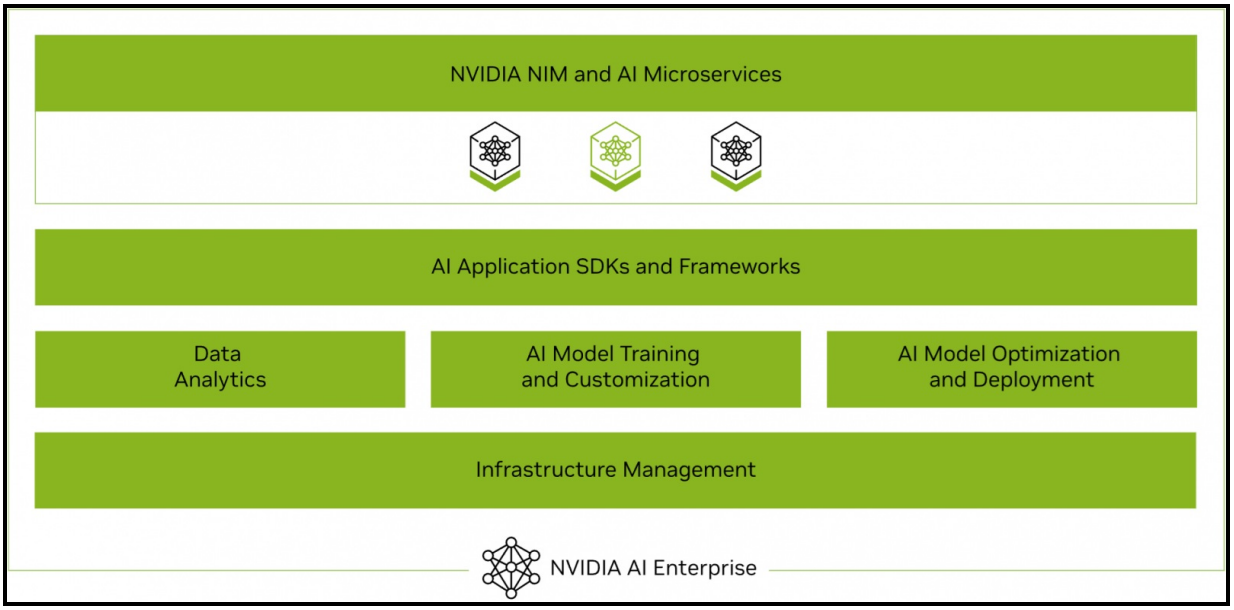

Figure 17. NVIDIA AI Enterprise software stack

NVIDIA AI Enterprise includes workload and infrastructure management software known as Base Command Manager. This software provisions the AI environment, incorporating the components such as the Operating System, Kubernetes (K8S), GPU Operator, and Network Operator to manage the AI workloads.

Additionally, NVIDIA AI Enterprise provides access to ready-to-use open-sourced containers and frameworks from NVIDIA like NVIDIA NeMo, NVIDIA RAPIDS, NVIDIA TAO Toolkit, NVIDIA TensorRT and NVIDIA Triton Inference Server.

- **NVIDIA NeMo** is an end-to-end framework for building, customizing, and deploying enterprise-grade

generative AI models; NeMo lets organizations easily customize pretrained foundation models from NVIDIA and select community models for domain-specific use cases.

- **NVIDIA RAPIDS** is an open-source suite of GPU-accelerated data science and AI libraries with APIs that match the most popular open-source data tools. It accelerates performance by orders of magnitude at scale across data pipelines.
- **NVIDIA TAO Toolkit** simplifies model creation, training, and optimization with TensorFlow and PyTorch and it enables creating custom, production-ready AI models by fine-tuning NVIDIA pretrained models and large training datasets.
- **NVIDIA TensorRT**, an SDK for high-performance deep learning inference, includes a deep learning inference optimizer and runtime that delivers low latency and high throughput for inference applications. TensorRT is built on the NVIDIA CUDA parallel programming model and enables you to optimize inference using techniques such as quantization, layer and tensor fusion, kernel tuning, and others on NVIDIA GPUs. https://developer.nvidia.com/tensorrt-getting-started
- **NVIDIA TensorRT-LLM** is an open-source library that accelerates and optimizes inference performance of the latest large language models (LLMs). TensorRT-LLM wraps TensorRT's deep learning compiler and includes optimized kernels from FasterTransformer, pre- and post-processing, and multi-GPU and multi-node communication. https://developer.nvidia.com/tensorrt
- **NVIDIA Triton Inference Server** optimizes the deployment of AI models at scale and in production for both neural networks and tree-based models on GPUs.

It also provides full access to the NVIDIA NGC catalogue, a collection of tested enterprise software, services and tools supporting end-to-end AI and digital twin workflows and can be integrated with MLOps platforms such as ClearML, Domino Data Lab, Run:ai, UbiOps, and Weights & Biases.

Finally, NVIDIA AI Enterprise introduced NVIDIA Inference Microservices (NIM), a set of performance-optimized, portable microservices designed to accelerate and simplify the deployment of AI models. Those containerized GPU-accelerated pretrained, fine-tuned, and customized models are ideally suited to be self-hosted and deployed on the Lenovo Hybrid AI platforms.

The ever-growing catalog of NIM microservices contains models for a wide range of AI use cases, from chatbot assistants to computer vision models for video processing. The image below shows some of the NIM microservices, organized by use case.

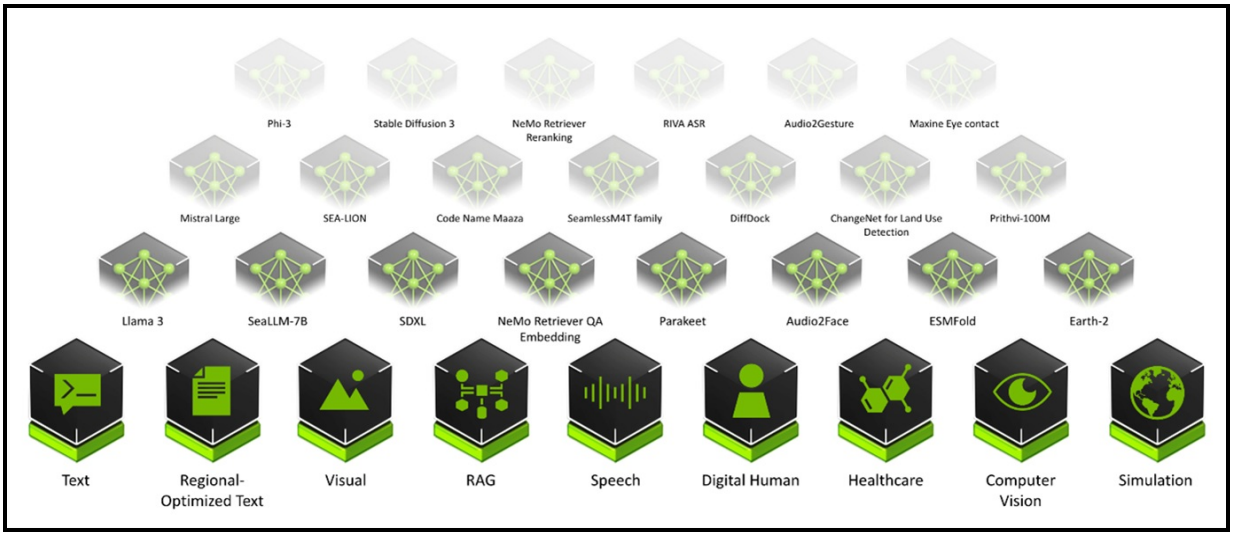

Figure 18. NVIDIA AI Enterprise software stack

NVIDIA RAG Blueprint

The NVIDIA RAG blueprint serves as a reference solution for a foundational Retrieval Augmented Generation (RAG) pipeline that can be implemented on the 221 platform. One of the key use cases in Generative AI is enabling users to ask questions and receive answers based on their enterprise data. This blueprint demonstrates how to set up a RAG solution that uses NVIDIA NIM and GPU-accelerated components. By default, this blueprint leverages locally-deployed NVIDIA NIM microservices to meet specific data governance and latency requirements. However, you can replace these models with your NVIDIA-hosted models available in the NVIDIA API Catalog.

Software Components:

The following are the default components included in this blueprint:

- NVIDIA NIM Microservices
- Response Generation (Inference)
  - NIM of nvidia/llama-3.3-nemotron-super-49b-v1
- Retriever Models
  - NIM of nvidia/llama-3_2-nv-embedqa-1b-v2
  - NIM of nvidia/llama-3_2-nv-rerankqa-1b-v2
  - NeMo Retriever Page Elements NIM
  - NeMo Retriever Table Structure NIM
  - NeMo Retriever Graphic Elements NIM
  - PaddleOCR NIM
- Optional NIMs
  - Llama 3.1 NemoGuard 8B Content Safety NIM
  - Llama 3.1 NemoGuard 8B Topic Control NIM
  - Mixtral 8x22B Instruct 0.1
  - Llama-3.1 Nemotron-nano-vl-8b-v1 NIM
  - NeMo Retriever Parse NIM
- RAG Orchestrator server - Langchain based
- Milvus Vector Database - accelerated with NVIDIA cuVS
- Ingestion - Nemo Retriever Extraction is leveraged for ingestion of files. Nemo Retriever Extraction is a scalable, performance-oriented document content and metadata extraction microservice. Including support for parsing PDFs, Word and PowerPoint documents, it uses specialized NVIDIA NIM microservices to find, contextualize, and extract text, tables, charts and images for use in downstream generative applications.

The image below represents the architecture and workflow. Here's a step-by-step explanation of the workflow from end-user perspective:

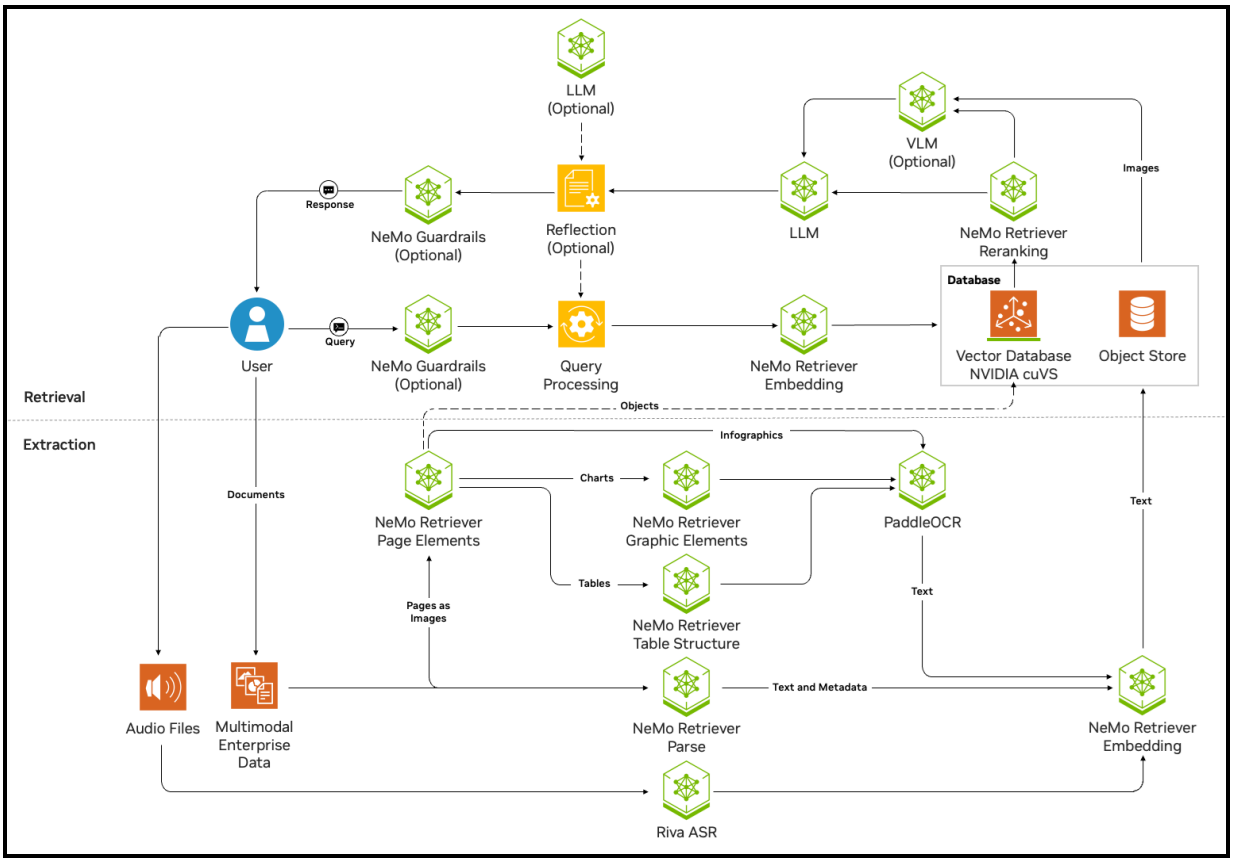

Figure 19. NVIDIA RAG Blueprint

This blueprint outlines the workflow for a system that uses a Retrieval-Augmented Generation (RAG) model. The process can be broken down into two main parts: data ingestion and the user query pipeline.

Data Ingestion

First, unstructured data is fed into the system through the /documents API via the Ingestor server microservice. This service preprocesses the data, divides it into manageable chunks, and then uses the Nvingest microservice to store these chunks in the Milvus Vector Database.

User Query Pipeline

The user's experience begins with the RAG Playground UI or through direct API calls.

- Query Processing: A user's query is sent to the RAG server microservice (built on LangChain) via the /generate API. An optional Query Rewriter component can refine the query for better search results. NeMo Guardrails can also be enabled at this stage to filter out inappropriate queries.
- Document Retrieval: The refined query is sent to the Retriever module, which queries the Milvus Vector Database. This database holds data embeddings created by the NeMo Retriever Embedding microservice. The retriever identifies the top K most relevant chunks of information.
- Reranking (Optional): The top K chunks are passed to the NeMo Retriever reranking microservice. This component further refines the results, selecting the top N most relevant chunks for improved precision.
- Response Generation: The top N chunks are then injected into a prompt and sent to the Response Generation module. This module uses the NeMo LLM inference microservice to generate a natural language response. An optional reflection module can make additional calls to the Large Language Model (LLM) to verify the response's accuracy based on the retrieved context. NeMo Guardrails can

also be applied here to ensure the final output is safe and non-toxic.

- Response Delivery: Finally, the generated response is sent back to the RAG Playground. The user can then view the answer and see the source citations for the retrieved information.

This modular design ensures efficient query processing, accurate retrieval of information, and easy customization.

Lenovo AI Center of Excellence

In addition to the choice of utilizing Lenovo EveryScale Infrastructure framework for the Enterprise AI platform to ensure tested and warranted interoperability, Lenovo operates an AI Lab and CoE at the headquarters in Morrisville, North Carolina, USA to test and enable AI applications and use cases on the Lenovo EveryScale AI platform.

The AI Lab environment allows customers and partners to execute proof of concepts for their use cases or test their AI middleware or applications. It is configured as a diverse AI platform with a range of systems and GPU options.

The software environment utilizes Canonical Ubuntu Linux along with Canonical MicroK8s to offer a multi-tenant Kubernetes environment. This setup allows customers and partners to schedule their respective test containers effectively.

Lenovo AI Innovators

Lenovo Hybrid AI platforms offer the necessary infrastructure for a customer's hybrid AI factory. To fully leverage the potential of AI integration within business processes and operations, software providers, both large and small, are developing specialized AI applications tailored to a wide array of use cases.

To support the adoption of those AI applications, Lenovo continues to invest in and extend its AI Innovators Program to help organizations gain access to enterprise AI by partnering with more than 50 of the industry's leading software providers.

Partners of the Lenovo AI Innovators Program get access to our AI Discover Labs, where they validate their solutions and jointly support Proof of Concepts and Customer engagements.

LAII provides customers and channel partners with a range of validated solutions across various vertical use cases, such as for Retail or Public Security. These solutions are designed to facilitate the quick and safe deployment of AI solutions that optimally address the business requirements.

The following is a selection of case studies involving Lenovo customers implementing an AI solution:

- Kroeger (Retail) – Reducing Customer friction and loss prevention
- Peak (Logistics) – Streamlining supply chain ops for fast and efficient deliveries
- Bikal (AI at Scale) – Delivering shared AI platform for education
- VSAAS (Smart Cities) – Enabling accurate and effective public security

Lenovo Validated Designs

Lenovo Validated Designs (LVDs) are pre-tested, optimized solution designs enabling reliability, scalability, and efficiency in specific workloads or industries. These solutions integrate Lenovo hardware like ThinkSystem servers, storage, and networking with software and best practices to solve common IT challenges. Developed with technology partners such as VMware, Intel, and Red Hat, LVDs ensure performance, compatibility, and easy deployment through rigorous validation.

Lenovo Validated Designs are intended to simplify the planning, implementation, and management of complex IT infrastructures. They provide detailed guidance, including architectural overviews, component models, deployment considerations, and bills of materials, tailored to specific use cases such as artificial intelligence (AI), big data analytics, cloud computing, virtualization, retail, or smart manufacturing. By offering a pretested solution, LVDs aim to reduce risk, accelerate deployment, and assist organizations in achieving faster time-to-value for their IT investments.

Lenovo Hybrid AI platforms act as infrastructure frameworks for LVDs addressing data center-based AI solutions. They provide the hardware/software reference architecture, optionally Lenovo EveryScale integrated solution delivery method, and general sizing guidelines.

AI Services

The services offered with the Lenovo Hybrid AI platforms are specifically designed to enable broad adoption of AI in the Enterprise. This enables both Lenovo AI Partners and Lenovo Professional Services to accelerate deployment and provide enterprises with the fastest time to production.

The Lenovo AI services offered alongside the Lenovo Hybrid AI platforms enable customers to overcome the barriers they face in realizing ROI from AI investments by providing critical expertise needed to accelerate business outcomes and maximum efficiency. Leveraging Lenovo AI expertise, Lenovo’s advanced partner ecosystem, and industry leading technology we help customers realize the benefits of AI faster. Unlike providers that tie GPU services to proprietary stacks, Lenovo takes a services-first approach, helping enterprises maximize existing investments and scale AI on their own terms.

The two current services offerings broken down to the right of the AI Factory and AI Foundation layers found in the figure below.

- AI Fast Start Services provide use case development and validation for agentic AI and GenAI applications.
- GPU Advance Services provide the foundation needed for AI use case development, including AI factory design, deployment of the software and firmware stack, and setup of orchestration software. Optionally, TruScale RedHat OpenShift service can be added for those wanting to use OpenShift on RHEL.

All services are flexible to meet the unique needs of different organizations while adhering to Lenovo's Reference Architectures and Platform Guides.

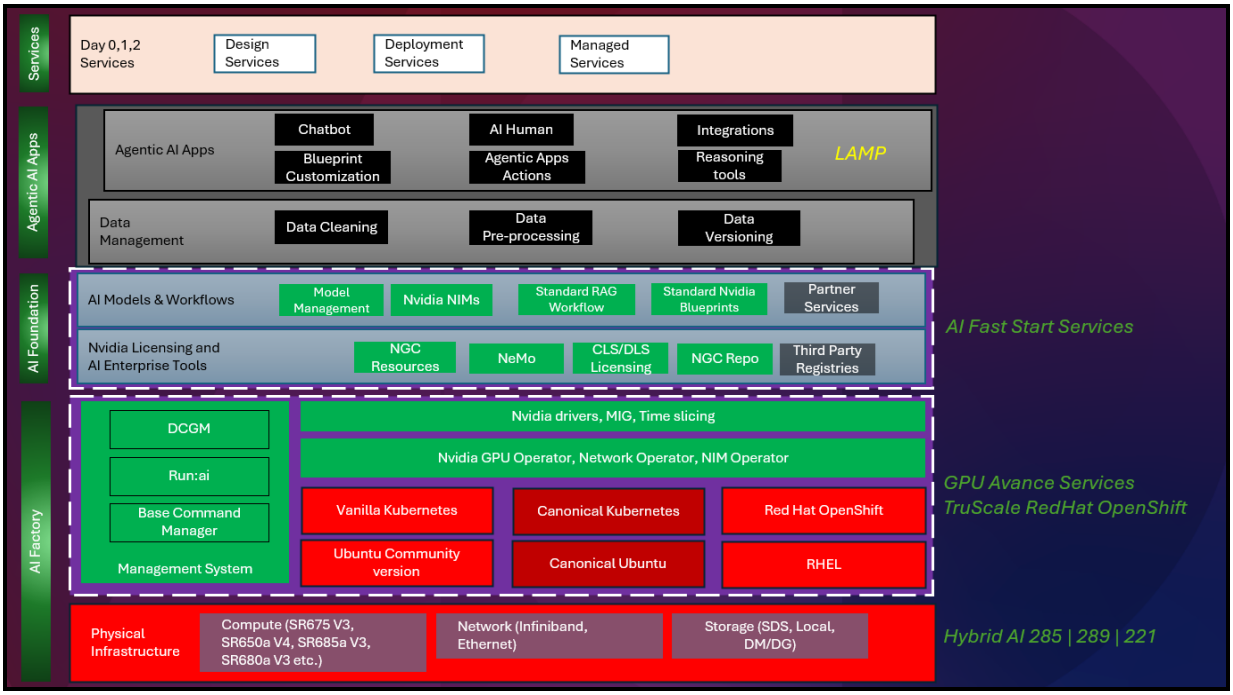

Figure 20. AI Services offerings

Customers can choose from the three modular services listed in the following table.

Table 6. AI Services

Service	Multi-Node	Single Node
GPU Advanced Services – Plan & Design with Kubernetes	Custom services engagement	5MS7C33028
GPU Advanced Services – Configuration & Deployment with Kubernetes	Custom services engagement	5MS7C33029
GPU Advanced Services – Managed Services with Kubernetes	Custom services engagement	5MS7C33030

These services offerings are described in the following sections:

- GPU Plan & Design Services
- GPU Configuration and Deployment Services
- GPU Managed Services

GPU Plan & Design Services

Lenovo offers advisory services to support organizations in planning and optimizing high-performance GPU workloads, including assessment of current infrastructure and identifying intended use cases.

This service helps customers with:

- Planning for optimal GPU utilization
- Aligning business and tech strategy
    - Workload assessment
    - deployment strategy
- Architecture and design
    - Solution sizing and technology selection
    - High-level architecture

The outcomes of the GPU Plan & Design Services are reduced risk and access to proven best practices. Improved performance and an optimised infrastructure at the outset build a solid base that is future-proofed to accommodate growth.

GPU Configuration and Deployment Services

Configuration and deployment services help organizations accelerate their timeline. By providing installation and setup for the complete Lenovo recommended software stack for AI, this service acts as the engine of the solution, significantly accelerating the time to value.

Lenovo deployment services help provide the fastest time to first token for an enterprise building their Hybrid AI factory. The GPU advanced configuration and deployment services provide expert guidance on software and hardware components to get your AI factory up and running, including:

- Operating system
- Kubernetes
- GPU configuration
- DDN storage configuration

With this service Lenovo enables deployment of the Lenovo Hybrid AI configurations, from a single node to multi-node, with customizable AI software stack and services. Leveraging our deep relationship with NVIDIA, we can fine-tune the GPU performance to precisely match the customer’s workload requirements.

Lenovo enables customers to overcome skills gaps to fully utilize their GPU configurations and boost the performance of their most challenging workloads. Lenovo also helps upskill a diverse customer team with knowledge transfer from Lenovo experts, working within the framework of our scalable, customizable Lenovo Hybrid AI architectures deliver end-to-end solutions designed to accelerate enterprise AI adoption.

Customers will receive a low-level design of the configuration as well as knowledge transfer from Lenovo’s experts

GPU Managed Services

Lenovo provides managed NVIDIA-based GPU systems that can address customer needs on an ongoing basis. This includes support, security & compliance, and business functions.

Customers can consistently maintain peak performance of their GPU infrastructure through the following support:

- L1 support for the GPU and NVIDIA AI Enterprise software (if applicable)
- Security and compliance verification
- Ongoing GPU performance monitoring and tuning with logging and alerts
- Backup and restore of the management components and configuration

GPU Managed Services seamlessly scales with the organization and giving greater visibility and monitoring of performance. Additionally these services provide vulnerability patching to stay ahead of risks which helps free up developers and data scientists to focus on innovation.

Lenovo TruScale

Lenovo TruScale XaaS is your set of flexible IT services that makes everything easier. Streamline IT procurement, simplify infrastructure and device management, and pay only for what you use – so your business is free to grow and go anywhere.

Lenovo TruScale is the unified solution that gives you simplified access to:

- The industry's broadest portfolio – from pocket to cloud – all delivered as a service
- A single-contract framework for full visibility and accountability
- The global scale to rapidly and securely build teams from anywhere
- Flexible fixed and metered pay-as-you-go models with minimal upfront cost
- The growth-driving combination of hardware, software, infrastructure, and solutions – all from one single provider with one point of accountability.

For information about Lenovo TruScale offerings that are available in your region, contact your local Lenovo sales representative or business partner.

Lenovo Financial Services

Why wait to obtain the technology you need now? No payments for 90 days and predictable, low monthly payments make it easy to budget for your Lenovo solution.

- **Flexible**

  Our in-depth knowledge of the products, services and various market segments allows us to offer greater flexibility in structures, documentation and end of lease options.

- **100% Solution Financing**

  Financing your entire solution including hardware, software, and services, ensures more predictability in your project planning with fixed, manageable payments and low monthly payments.

- **Device as a Service (DaaS)**

  Leverage latest technology to advance your business. Customized solutions aligned to your needs. Flexibility to add equipment to support growth. Protect your technology with Lenovo's Premier Support service.

- **24/7 Asset management**

  Manage your financed solutions with electronic access to your lease documents, payment histories, invoices and asset information.

- **Fair Market Value (FMV) and $1 Purchase Option Leases**

  Maximize your purchasing power with our lowest cost option. An FMV lease offers lower monthly payments than loans or lease-to-own financing. Think of an FMV lease as a rental. You have the flexibility at the end of the lease term to return the equipment, continue leasing it, or purchase it for the fair market value. In a $1 Out Purchase Option lease, you own the equipment. It is a good option when you are confident you will use the equipment for an extended period beyond the finance term. Both lease types have merits depending on your needs. We can help you determine which option will best meet your technological and budgetary goals.

Ask your Lenovo Financial Services representative about this promotion and how to submit a credit application. For the majority of credit applicants, we have enough information to deliver an instant decision and send a notification within minutes.

Bill of Materials

This section provides an example Bill of Materials (BoM) for the computes.

- SR675 V3 2-2-1 with 2x NVIDIA RTX Pro 6000 Blackwell Server Edition GPUs
- SR650a V4 2-2-1 with 2x NVIDIA H200 NVL GPUs
- SR635 V3
- SR630 V4

Configuration tips:

- If Out-of-band management HA is required, an optional ThinkSystem Intel I350 1GbE RJ45 4-Port OCP Ethernet Adapter V2 NIC can be configured on the SR675 V3 or ThinkSystem Broadcom 5719 1GbE RJ45 4-port OCP Ethernet Adapter on the SR650a V4.
- If storage integration is not required, replace BE4U: ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port PCIe Ethernet Adapter with BQBN: ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16.

SR675 V3 2-2-1 with 2x NVIDIA RTX Pro 6000 Blackwell Server Edition GPUs

Table 7. SR675 V3 Bill of Materials

Part number	Description	Quantity
7D9RCTO1WW	Server : ThinkSystem SR675 V3 - 3yr Warranty for AI	1
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	1
C2AP	ThinkSystem AMD EPYC 9255 24C 200W 3.25GHz Processor	2
C0CK	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A	8
5977	Select Storage devices - no configured RAID required	1
C1AE	ThinkSystem E3.S PM9D3a 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2
BTMB	ThinkSystem 1x4 E3.S Backplane	1
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1
BXMH	ThinkSystem M.2 PM9A3 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BE4U	ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port PCIe Ethernet Adapter	1
CBK8	ThinkSystem NVIDIA RTX PRO 6000 Blackwell Server Edition 96GB PCIe Gen5 Passive GPU	2
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	2
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	1
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	2
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	4
C3KA	ThinkSystem SR670 V2/SR675 V3 Heavy Systems Toolless Slide Rail Kit	1
B7XZ	Disable IPMI-over-LAN	1
C3EF	ThinkSystem SR675 V3 System Board v2	1
C8WW	SR675 V3 Laser service indicator	1
BK15	High voltage (200V+)	1
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	1
BE0D	N+1 Redundancy With Over-Subscription	1
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	1
BRUC	ThinkSystem SR675 V3 CPU Heatsink	2
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	1
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	1
BABV	ThinkSystem Screw for fix M.2 Adapter	1
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	1
BRNM	ThinkSystem SR670 V2/SR675 V3 2600W Power Supply Caution Label	1
BR7U	ThinkSystem SR675 V3 Root of Trust Module	1
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	1
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	1
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	1
BR80	ThinkSystem SR675 V3 Agency Labels	1
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	2

Part number	Description	Quantity
BXGN	ThinkSystem Single Bay E3.S HDD Filler	2
BAW4	OCP Filler	1
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	1
BYT3	PCIe DUMMY BKT w/o Air hole	14
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	1
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1
BTMN	ThinkSystem SR675 V3 E3.S to Riser Cables	1
C5WW	ThinkSystem SR675 V3 Dual Rotor System High Performance Fan	5
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	1
BTMF	ThinkSystem SR675 V3 Label2 for E3.S Backplane	1
BTME	ThinkSystem SR675 V3 E3.S Backplane Cage Assembly for 8DW PCIe GPU Base	1
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	1
BU0D	ThinkSystem SR675 V3 E3.S Backplane Power Cable 2	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAAK	SR675 V3	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QAAK	SR675 V3	1
QA18	Premier	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

SR650a V4 2-2-1 with 2x NVIDIA H200 NVL GPUs

Table 8. SR650a V4 Bill of Materials

Part number	Description	Quantity
7DG9CTO1WW	Server : ThinkSystem SR630 V4 - 3yr Warranty	1
C1XE	ThinkSystem 1U V4 10x2.5" Chassis	1
C3JB	ThinkSystem General Computing - Power Efficiency	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5QT	Intel Xeon 6530P 32C 225W 2.3GHz Processor	2
C1XJ	ThinkSystem 1U V4 Performance Heatsink	2
C0U9	ThinkSystem 32GB TruDDR5 6400MHz (1Rx4) RDIMM	8
5977	Select Storage devices - no configured RAID required	1
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2
C2NN	ThinkSystem 1U V4 4x2.5" NVMe Gen5 Backplane	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter	1

Part number	Description	Quantity
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter	1
C1ZB	ThinkSystem SR630 V4 x16 PCIe Gen5 Riser 1 or 2	1
C1ZC	ThinkSystem 1U V4 Full Height Riser Cage 1	1
C0U5	ThinkSystem 1300W 230V/115V Platinum CRPS Hot-Swap Power Supply v2.4	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
C1YT	ThinkSystem 1U V4 Performance Fan Module	4
C1YP	ThinkSystem 1U V4 Standard Media Bay	1
C2DH	ThinkSystem Toolless Slide Rail Kit V4	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
BK14	Low voltage (100V+)	1
BRCZ	ThinkSystem 1U 4x2.5" NVMe HDD Type Label	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
B97B	XCC Label	1
C20J	ThinkSystem SR630 V4 Service Label for WW	1
AWF9	ThinkSystem Response time Service Label LI	1
C20D	ThinkSystem SR630 V4 model name Label	1
C1ZP	ThinkSystem SR630 V4 Agency label with Blank	1
C214	ThinkSystem BHS SR630 V4 1U PCIe number 1 and OCP number Label (FH+Rear 2x2.5")	1
C1ZS	ThinkSystem 1300W (CRPS) power rating label WW	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BQPS	ThinkSystem logo Label	1
C1YA	ThinkSystem M.2 Signal&Power Cable, ULP 82P-SLX4/2X10 SB, 540/680mm	1
C1XZ	ThinkSystem Power Cable, PIC Power 2x3+6P to PIC Power 2x3+6P, 380mm	1
C1XN	ThinkSystem PCIe 5.0 Cable, MCIO 8X STR TO MCIO 8X STR, 350mm	2
BE0F	N+N Redundancy Without Over-Subscription	1
B0ML	Feature Enable TPM on MB	1
C3NZ	ThinkSystem SR630 V4 MI-4x2.5" NVMe G5 BP	1
CAR5	SR630 V4 Laser service indicator	1
C4DV	ThinkSystem SR630 V4 MotherBoard	1
C3K9	XClarity Platinum Upgrade v3	1
C1YW	ThinkSystem 1U V4 Full HeightRiser Cage 2 Filler	1
BPP5	OCP3.0 Filler with screw	2
AURS	Lenovo ThinkSystem Memory Dummy	24
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
BCEB	ThinkSystem 1U V2 2x3 2.5" HDD Dummy	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	2
C2NR	ThinkSystem V4 4x2.5 BP Llong Bracket	1
C5XG	ThinkSystem SR630 V4 General Config PKG AC+CL	1

Part number	Description	Quantity
CA7N	ThinkSystem SR630 V4 System I/O board v2	1
C26Y	ThinkSystem V4 CPU HS Clip	2
BTTY	M.2 NVMe	1
CBDC	ENERGY STAR Certification Country	1
C2ZA	ThinkSystem G5 X16 Riser for Riser1 C1ZB Placement	1
	Auto-Derived Part Items	
BF94	AI & HPC - ThinkSystem Hardware	1
7S0XCTO8WW	XClarity Controller Prem-FOD	1
SCY0	Lenovo XClarity XCC3 premier - FOD	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAJQ	SR630 V4 3Yr	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QA18	Premier	1
QAJQ	SR630 V4 3Yr	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

SR635 V3

Table 9. SR635 V3 Bill of Materials

Part number	Description	Quantity
7D9GCTO1WW	Server : ThinkSystem SR635 V3 - 3yr Warranty	1
BLK4	ThinkSystem V3 1U 10x2.5" Chassis	1
BFYB	Operating mode selection for: "Maximum Performance Mode"	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C2AQ	ThinkSystem AMD EPYC 9335 32C 210W 3.0GHz Processor	1
BQ26	ThinkSystem SR645 V3/SR635 V3 1U High Performance Heatsink	1
C0DF	Platform Secure Boot Enable	1
C1PL	ThinkSystem 32GB TruDDR5 6400MHz (1Rx4) RDIMM-A	12
5977	Select Storage devices - no configured RAID required	1
BC4V	Non RAID NVMe	1
C0ZU	ThinkSystem 2.5" U.2 VA 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2
BPC9	ThinkSystem 1U 4x 2.5" NVMe Gen 4 Backplane	1
B5XJ	ThinkSystem M.2 SATA/NVMe 2-Bay Adapter	1
BTTY	M.2 NVMe	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2

Part number	Description	Quantity
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter	1
BLK7	ThinkSystem SR635 V3/SR645 V3 x16 PCIe Gen5 Riser 1	1
BLK9	ThinkSystem V3 1U MS LP+LP BF Riser Cage	1
BNFG	ThinkSystem 750W 230V/115V Platinum Hot-Swap Gen2 Power Supply v3	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BH9M	ThinkSystem V3 1U Performance Fan Option Kit v2	7
BLKD	ThinkSystem 1U V3 10x2.5" Media Bay w/ Ext. Diagnostics Port	1
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
C2AD	ThinkSystem SR635 V3 MB w/IO+PIB+FB,1U	1
C8WT	SR635 V3 Laser service indicator	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	2
BCEB	ThinkSystem 1U V2 2x3 2.5" HDD Dummy	1
B989	ThinkSystem V2 1U Package	1
B8NJ	ThinkSystem 1U MS Fan Dummy	1
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
B984	ThinkSystem 1U PLV Top Cover Sponge	1
C1PT	ThinkSystem SR635 V3/SR655 V3 Root of Trust Module Low Voltage-RoW V2	1
BQPS	ThinkSystem logo Label	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BQ7V	ThinkSystem SR635 V3 Service Label for WW	1
B8NB	ThinkSystem 1U MS LP Riser Filler	1
C0FC	ThinkSystem SR635 V3 Agency Label No ES mark	1
BQ7Q	ThinkSystem SR635 V3 Model Name Label	1
AWF9	ThinkSystem Response time Service Label LI	1
B5WJ	ThinkSystem OCP3 Filler	1
AUWG	Lenovo ThinkSystem 1U VGA Filler	1
B97B	XCC Label	1
B8JV	ThinkSystem 750W Pt Power Rating Label WW	1
BRPJ	XCC Platinum	1
BK14	Low voltage (100V+)	1
BRCZ	ThinkSystem 1U 4x2.5" NVMe HDD Type Label	1
BPK3	ThinkSystem WW Lenovo LPK	1
BQRM	ThinkSystem PCIe Cbl MB_Swift 3 4x2.5" NVMe(Bianca)430mm	1
BSE9	M.2 Module,MB to M.2 Signal Cable,635mm,SLMX4/SB to Modul from MB_J52/ J51 to M.2 Card Mbappe/Isco	1
BMJD	ThinkSystem 3.5" BP Power Cable v2	1
BQRC	ThinkSystem PCIe Cbl MB_Swift 4x2.5" NVMe(Bianca)430mm	1
BE0C	N+1 Redundancy Without Over-Subscription	1
A2HP	Configuration ID 01	1

Part number	Description	Quantity
5374CM1	Configuration Instruction	1
BC4W	Non RAID NVMe Placement	1
A2JX	Controller 01	1
A2HP	Configuration ID 01	1
7S0XCTO5WW	XClarity Controller Platin-FOD	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAAF	SR635 V3	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QA18	Premier	1
QAAF	SR635 V3	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

SR630 V4

Table 10. SR630 V4 Bill of Materials

Part number	Description	Quantity
7DG9CTO1WW	Server : ThinkSystem SR630 V4 - 3yr Warranty	1
C1XE	ThinkSystem 1U V4 10x2.5" Chassis	1
C3JB	ThinkSystem General Computing - Power Efficiency	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5QT	Intel Xeon 6530P 32C 225W 2.3GHz Processor	2
C1XJ	ThinkSystem 1U V4 Performance Heatsink	2
C0U9	ThinkSystem 32GB TruDDR5 6400MHz (1Rx4) RDIMM	8
5977	Select Storage devices - no configured RAID required	1
C0ZT	ThinkSystem 2.5" U.2 VA 7.68TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2
C2NN	ThinkSystem 1U V4 4x2.5" NVMe Gen5 Backplane	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/200GbE QSFP112 2-port PCIe Gen5 x16 Adapter	1
C1ZB	ThinkSystem SR630 V4 x16 PCIe Gen5 Riser 1 or 2	1
C1ZC	ThinkSystem 1U V4 Full Height Riser Cage 1	1
C0U5	ThinkSystem 1300W 230V/115V Platinum CRPS Hot-Swap Power Supply v2.4	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
C1YT	ThinkSystem 1U V4 Performance Fan Module	4

Part number	Description	Quantity
C1YP	ThinkSystem 1U V4 Standard Media Bay	1
C2DH	ThinkSystem Toolless Slide Rail Kit V4	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
BK14	Low voltage (100V+)	1
BRCZ	ThinkSystem 1U 4x2.5" NVMe HDD Type Label	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
B97B	XCC Label	1
C20J	ThinkSystem SR630 V4 Service Label for WW	1
AWF9	ThinkSystem Response time Service Label LI	1
C20D	ThinkSystem SR630 V4 model name Label	1
C1ZP	ThinkSystem SR630 V4 Agency label with Blank	1
C214	ThinkSystem BHS SR630 V4 1U PCIe number 1 and OCP number Label (FH+Rear 2x2.5")	1
C1ZS	ThinkSystem 1300W (CRPS) power rating label WW	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BQPS	ThinkSystem logo Label	1
C1YA	ThinkSystem M.2 Signal&Power Cable, ULP 82P-SLX4/2X10 SB, 540/680mm	1
C1XZ	ThinkSystem Power Cable, PIC Power 2x3+6P to PIC Power 2x3+6P, 380mm	1
C1XN	ThinkSystem PCIe 5.0 Cable, MCIO 8X STR TO MCIO 8X STR, 350mm	2
BE0F	N+N Redundancy Without Over-Subscription	1
B0ML	Feature Enable TPM on MB	1
C3NZ	ThinkSystem SR630 V4 MI-4x2.5" NVMe G5 BP	1
CAR5	SR630 V4 Laser service indicator	1
C4DV	ThinkSystem SR630 V4 MotherBoard	1
C3K9	XClarity Platinum Upgrade v3	1
C1YW	ThinkSystem 1U V4 Full HeightRiser Cage 2 Filler	1
BPP5	OCP3.0 Filler with screw	2
AURS	Lenovo ThinkSystem Memory Dummy	24
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
BCEB	ThinkSystem 1U V2 2x3 2.5" HDD Dummy	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	2
C2NR	ThinkSystem V4 4x2.5 BP Llong Bracket	1
C5XG	ThinkSystem SR630 V4 General Config PKG AC+CL	1
CA7N	ThinkSystem SR630 V4 System I/O board v2	1
C26Y	ThinkSystem V4 CPU HS Clip	2
BTTY	M.2 NVMe	1
CBDC	ENERGY STAR Certification Country	1
C2ZA	ThinkSystem G5 X16 Riser for Riser1 C1ZB Placement	1
	Auto-Derived Part Items	
BF94	AI & HPC - ThinkSystem Hardware	1

Part number	Description	Quantity
7S0XCTO8WW	XClarity Controller Prem-FOD	1
SCY0	Lenovo XClarity XCC3 premier - FOD	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
3444	Registration only	1
7Q01CTSAWW	SERVER KEEP YOUR DRIVE ADD-ON	1
QAJQ	SR630 V4 3Yr	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	SERVER PREMIER 24X7 4HR RESP	1
QA18	Premier	1
QAJQ	SR630 V4 3Yr	1
QA0Y	Months	36
QA12	24x7 4hr Resp	1

Related publications and links

For more information, see these resources:

- Lenovo EveryScale support page:
  https://datacentersupport.lenovo.com/us/en/solutions/ht505184
- x-config configurator:
  https://lesc.lenovo.com/products/hardware/configurator/worldwide/bhui/asit/x-config.jnlp
- Implementing AI Workloads using NVIDIA GPUs on ThinkSystem Servers:
  https://lenovopress.lenovo.com/lp1928-implementing-ai-workloads-using-nvidia-gpus-on-thinksystem-servers
- Making LLMs Work for Enterprise Part 3: GPT Fine-Tuning for RAG:
  https://lenovopress.lenovo.com/lp1955-making-llms-work-for-enterprise-part-3-gpt-fine-tuning-for-rag
- Lenovo to Deliver Enterprise AI Compute for NetApp AIPod Through Collaboration with NetApp and NVIDIA
  https://lenovopress.lenovo.com/lp1962-lenovo-to-deliver-enterprise-ai-compute-for-netapp-aipod-nvidia

Related product families

Product families related to this document are the following:

- AI Servers
- Artificial Intelligence

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2313, was created or updated on October 6, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2313
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2313.

Trademarks

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

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

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

PowerPoint® is a trademark of Microsoft Corporation in the United States, other countries, or both.

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