Lenovo Hybrid AI 285 Platform Guide
Product Guide

The evolution from Generative AI to Agentic AI has revolutionized the landscape of business and enterprise operations globally. By leveraging the capabilities of intelligent agents, companies can now streamline processes, enhance efficiency, and maintain a competitive edge.

These AI agents are adept at handling routine tasks, allowing skilled employees to focus on strategic initiatives and areas where their expertise truly adds value. This symbiotic relationship between AI agents and human employees fosters a collaborative environment that drives innovation and success.

Enterprises must proactively identify opportunities where AI agents can be integrated to support their operations, ensuring they remain agile and effective in an ever-evolving market. This new foundation of AI-driven optimization not only boosts productivity but also empowers employees to contribute more meaningfully to the organization's vision and goals.

Lenovo Hybrid AI 285 is a platform that enables enterprises of all sizes to quickly deploy hybrid AI factory infrastructure, supporting Enterprise AI use cases as either a new, greenfield environment or an extension of their existing IT infrastructure.

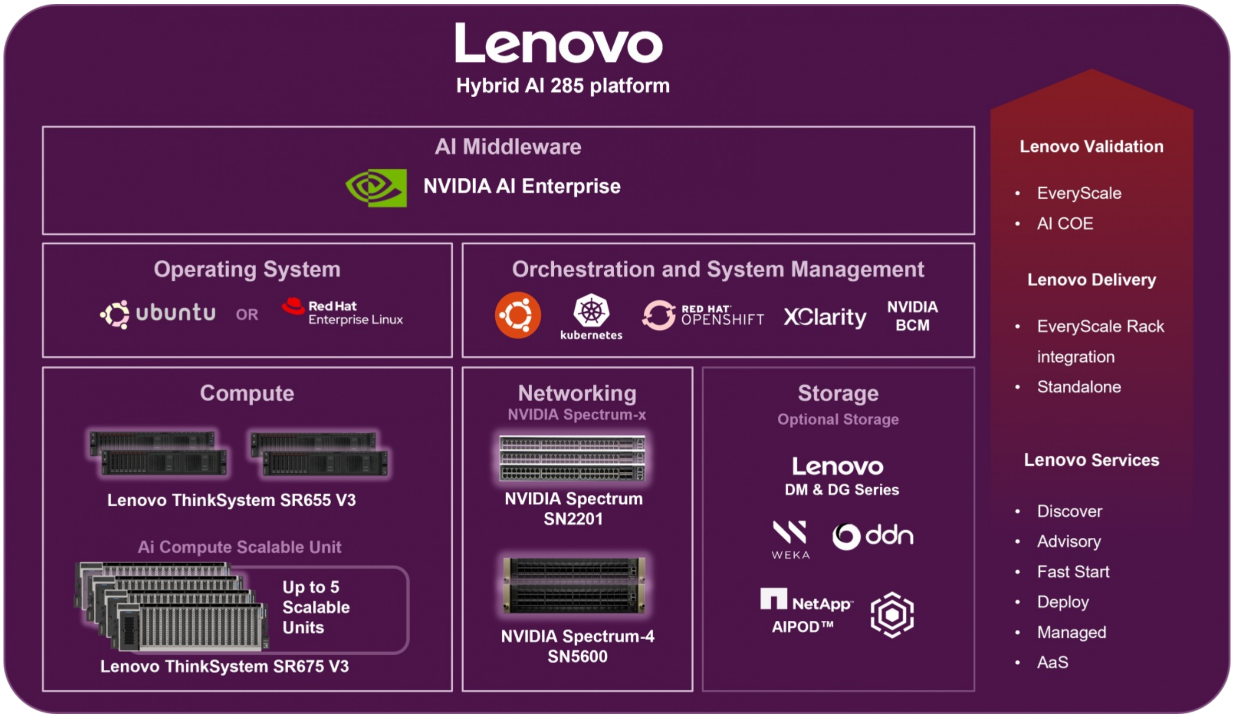

Figure 1. Lenovo Hybrid AI 285 platform overview

The offering is based on the NVIDIA 2-8-5 PCIe-optimized configuration — 2x CPUs, 8x GPUs, and 5x network adapters — and is ideally suited for medium (per GPU) to large (per node) Inference use cases, and small-to-large model training or fine-tuning, depending on chosen scale. It combines market leading Lenovo ThinkSystem GPU-rich servers with NVIDIA Hopper or Blackwell GPUs, NVIDIA Spectrum X networking and enables the use of the NVIDIA AI Enterprise software stack with NVIDIA Blueprints.

Following the principle of *From Exascale to EveryScale™*, Lenovo, widely recognized as a leader in High Performance Computing, leverages its expertise and capabilities from Supercomputing to create tailored enterprise-class hybrid AI factories.

Ideally utilizing Lenovo EveryScale Infrastructure (LESI) it 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.

Did you know?

The same team of HPC and AI experts that stand behind the Lenovo EveryScale OVX solution, as deployed also for NVIDIA Omniverse Cloud, brings the Lenovo Hybrid AI 285 platform to market.

Following their excellent experience with Lenovo on Omniverse, NVIDIA has once again chosen Lenovo technology as the foundation for the development and test of their NVIDIA AI Enterprise Reference Architecture (ERA). This choice ensures complete alignment and an exact match of Lenovo Hybrid AI 285 platform with the NVIDIA 2-8-5-200 ERA specifications.

Overview

The Lenovo Hybrid AI 285 Platform scales from a single server with just 4 GPUs as starter environment to a rack Scalable Unit (SU) with four servers and 32 GPUs to up to 5 Scalable Units with 20 Servers and 160 GPUs.

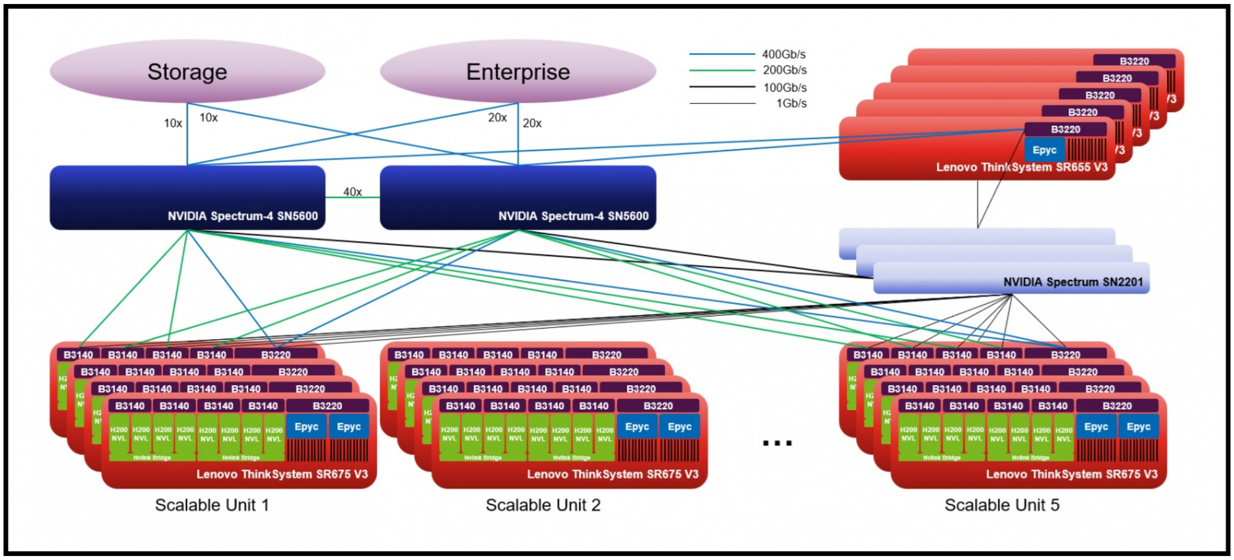

Figure 2. Lenovo Hybrid AI 285 platform with 5 Scalable Units

It can be deployed even for larger sizes with up to 8 Scalable Units with 32 Servers and 256 GPUs by breaking using two more SN5600s to create a dedicated E/W network for GPU to GPU communication.

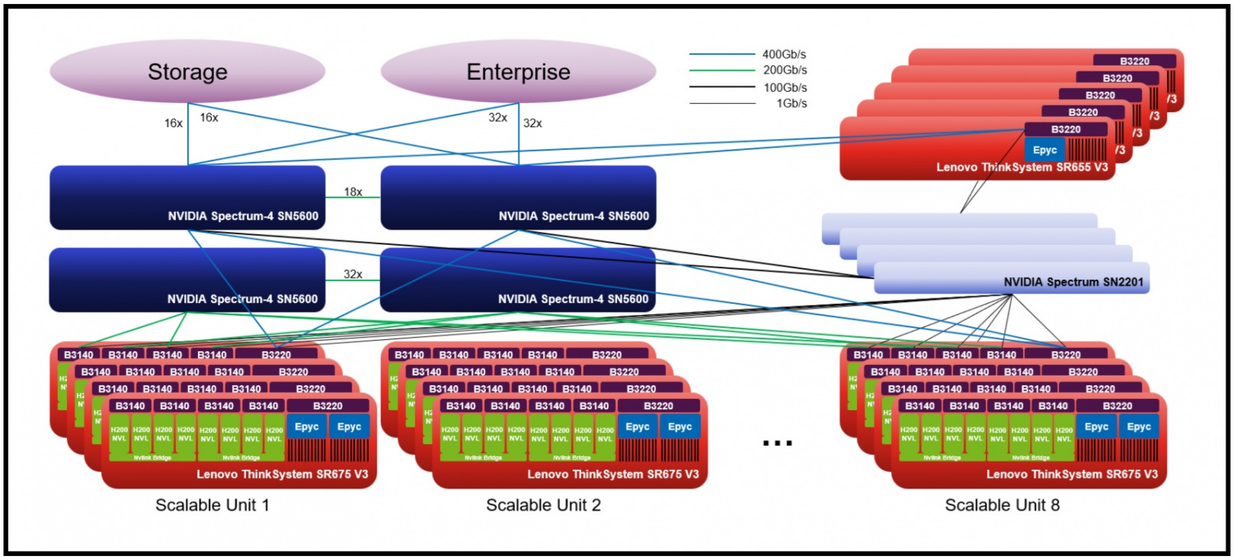

Figure 3. Lenovo Hybrid AI 285 platform with 8 Scalable Units

Components

The main hardware components of Lenovo Hybrid AI platforms are Compute 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
- Service Nodes – SR655 V3
- Networking
- Lenovo EveryScale Solution

AI Compute Node – SR675 V3

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

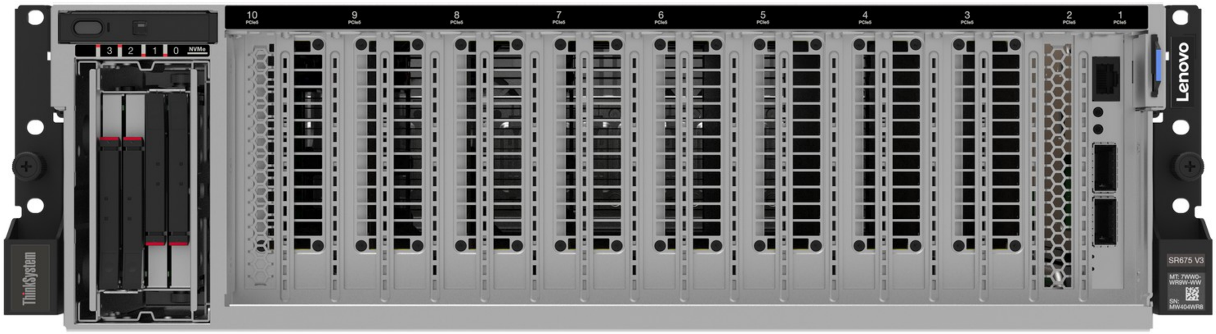

Figure 4. Lenovo ThinkSystem SR675 V3 in 8DW PCIe Setup

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. This makes it the ideal choice for NVIDIA's 2-8-5 configuration requirement.

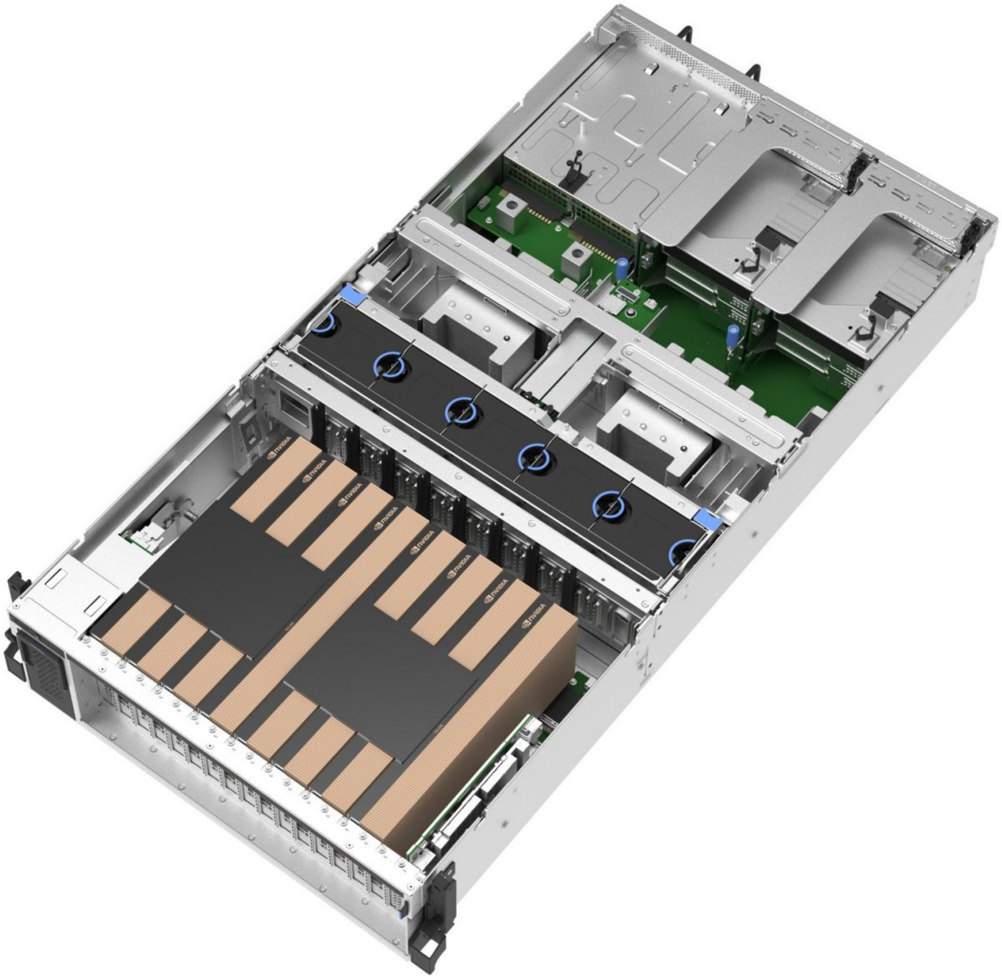

Figure 5. Lenovo ThinkSystem SR675 V3 in 8DW PCIe Setup

This configuration is NVIDIA-certified for Enterprise as well as Spectrum-X and an evolution of the Lenovo EveryScale OVX 3.0 setup which NVIDIA also deployed based on the Lenovo ThinkSystem SR675 V3 within NVIDIA Omniverse Cloud on Microsoft Azure.

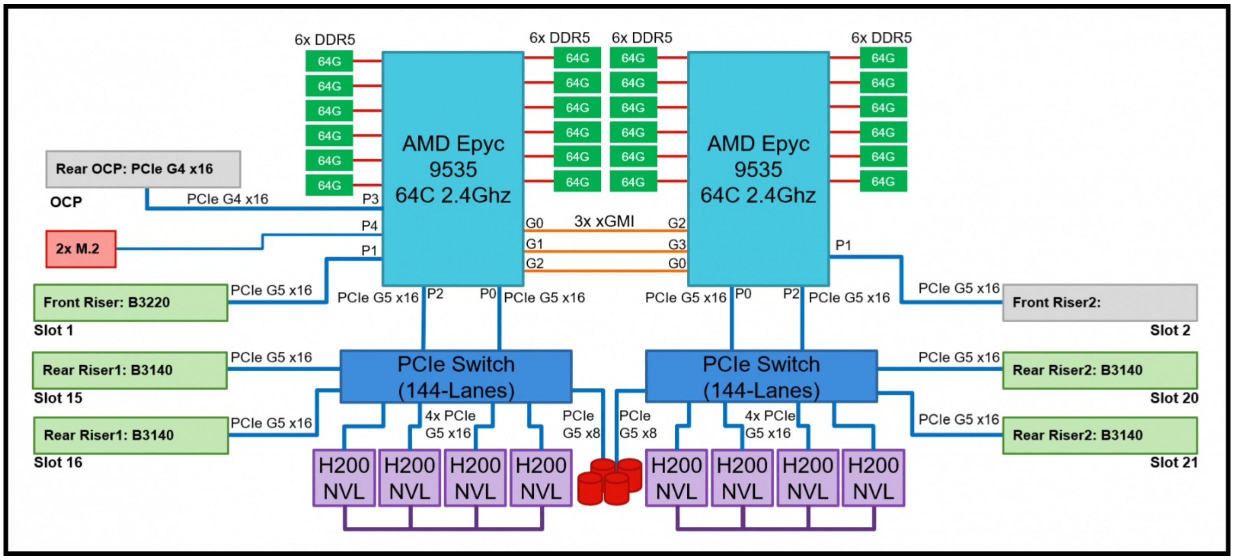

Figure 6. AI Compute Node Block Diagram

The AI Compute node is configured with two **AMD EPYC 9535 64 Core 2.4 GHz** processors with an all-core boost frequency of 3.5GHz. Besides providing consistently more than 2GHz frequency this ensures that with 7 Multi Instance GPUs (MIG) on 8 physical GPUs there are 2 Cores available per MIG plus a few additional Cores for Operating System and other operations.

With 12 Memory Channels per processor socket the AMD based server provides superior Memory bandwidth versus computing Intel-based platforms ensuring highest performance. Leveraging 64GB 6400MHz Memory DIMMs for a total of 1.5TB of main memory providing 192GB memory per GPU or a minimum of 1.5X the H200 NVL GPU memory.

The GPUs are connected to the CPUs via two PCIe Gen5 switches, each supporting up to four GPUs. With the **NVIDIA H200 NVL PCIe GPU**, the four GPUs are additionally interconnected through an NVLink bridge, creating a unified memory space. In a starter configuration with two GPUs per PCIe switch, the ThinkSystem SR675 V3 uniquely supports connecting all four GPUs with an NVLink bridge for maximized shared memory, thereby accommodating larger inference models, rather than limiting the bridge to two GPUs.

In the platforms default configuration, the AI Compute node leverages NVIDIA Bluefield technology both for the North-South as well as the East-West communication.

For Converged (North-South) Network an Ethernet adapter with redundant 200Gb/s connections provides ample bandwidth to storage, service nodes and the Enterprise network. The **NVIDIA BlueField-3 B3220 P-Series FHHL DPU** provides the two 200Gb/s Ethernet ports, a 1Gb/s Management board and a 16-Core ARM chip enabling Cloud Orchestration, Storage Acceleration, Secure Infrastructure and Tenant Networking.

The Ethernet adapters for the Compute (East-West) Network are directly connected to the GPUs via PCIe switches minimizing latency and enabling NVIDIA GPUDirect and GPUDirect Storage operation. For pure Inference workload they are optional, but for training and fine-tuning operation they should provide at least 200Gb/s per GPU.

By using the **NVIDIA BlueField-3 B3140H E-Series HHHL DPU** or alternatively a ConnectX8 400Gbit Ethernet adapter in combination with NVIDIA Spectrum 4 networking the East-West traffic can utilize Spectrum X operation.

Finally, the system is completed by local storage with two 960GB Read Intensive M.2 in RAID1 configuration for the operating system and four 3.84TB Read Intensive E3.S drives for local application data.

NVIDIA H200 NVL GPU

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 NVDIA H200 NVL GPUs.

Configuration

The following table lists the configuration of the AI Compute Node.

Table 1. AI Compute Node

Part Number	Description	Quantity
7D9R-CTOLWW	ThinkSystem SR675 V3	
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1
C3EF	ThinkSystem SR675 V3 System Board v2	1
C2AL	ThinkSystem AMD EPYC 9535 64C 300W 2.4GHz Processor	2
C0CK	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A	24
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	2
C3V3	ThinkSystem NVIDIA H200 NVL 141GB PCIe GPU Gen5 Passive GPU	8
C3V0	ThinkSystem NVIDIA 4-way bridge for H200 NVL	2
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1
C2RK	ThinkSystem SR675 V3 2 x16 Switch Cabled PCIe Rear IO Riser	2
C0Q4	ThinkSystem NVIDIA BlueField-3 B3140H VPI QSFP112 1P 400G PCIe Gen5 x16 Adapter	4
BM8X	ThinkSystem M.2 SATA/x4 NVMe 2-Bay Adapter	1
BT7P	ThinkSystem Raid 540-8i for M.2/7MM NVMe boot Enablement	1
BXMH	ThinkSystem M.2 PM9A3 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BTMB	ThinkSystem 1x4 E3.S Backplane	1
C1AB	ThinkSystem E3.S PM9D3a 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	1
C5WW	ThinkSystem SR675 V3 Dual Rotor System High Performance Fan	5
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	1
BE0D	N+1 Redundancy With Over-Subscription	1
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
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	1
BR7U	ThinkSystem SR675 V3 Root of Trust Module	1

Part Number	Description	Quantity
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	1
5PS7B09631	5Yr Premier NBD Resp + KYD SR675 V3	1

Service Nodes – SR655 V3

When deploying beyond two AI Compute nodes additional Service nodes are needed to manage the overall AI cluster environment.

Two **Management Nodes** provide a high-availability for the System Management and Monitoring provided through NVIDIA Base Command Manager (BMC) as described further in the AI Software Stack chapter.

For the Container operations three **Scheduling Nodes** build the Kubernetes control plane providing redundant operations and quorum capability.

Figure 7. Lenovo ThinkSystem SR655 V3

The Lenovo ThinkSystem SR655 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. Additionally, it includes a **NVIDIA BlueField-3 B3220L E-Series FHHL DPU** adapter to connect the Service Nodes to the Converged Network.

Configuration

The following table lists the configuration of the Service Nodes.

Table 2. Service Nodes

Part Number	Description	Quantity
7D9E-CTOLWW	ThinkSystem SR655 V3	
BLKK	ThinkSystem V3 2U 24x2.5" Chassis	1
C2AC	ThinkSystem SR655 V3 MB w/IO+PIB+FB,2U	1
C2AQ	ThinkSystem AMD EPYC 9335 32C 210W 3.0GHz Processor	1
C0CJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A	12
BPQV	ThinkSystem V3 2U x16/x16/E PCIe Gen5 Riser1 or 2	1
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	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
BS7Y	ThinkSystem V3 2U 8x2.5" NVMe Gen5 Backplane	1
C0ZU	ThinkSystem 2.5" U.2 Multi Vendor 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2

Part Number	Description	Quantity
BLL6	ThinkSystem 2U V3 Performance Fan Module	6
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	2
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1
C1PT	ThinkSystem SR635 V3/SR655 V3 Root of Trust Module Low Voltage-RoW V2	1
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1
5PS7B08762	5Yr Premier NBD Resp + KYD SR655 V3	1

Networking

The default setup of the Lenovo Hybrid AI 285 platform leverages NVIDIA Networking with the NVIDIA Spectrum-4 SN5600 for the Converged and Compute Network and the NVIDIA SN2201 for the Management Network.

NVIDIA Spectrum-4 SN5600

The SN5600 smart-leaf, spine, and super-spine switch offers 64 ports of 800GbE in a dense 2U form factor. The SN5600 is ideal for NVIDIA Spectrum-X deployments and enables both standard leaf/spine designs with top-of-rack (ToR) switches as well as end-of-row (EoR) topologies. The SN5600 offers diverse connectivity in combinations of 1 to 800GbE and boasts an industry-leading total throughput of 51.2Tb/s.

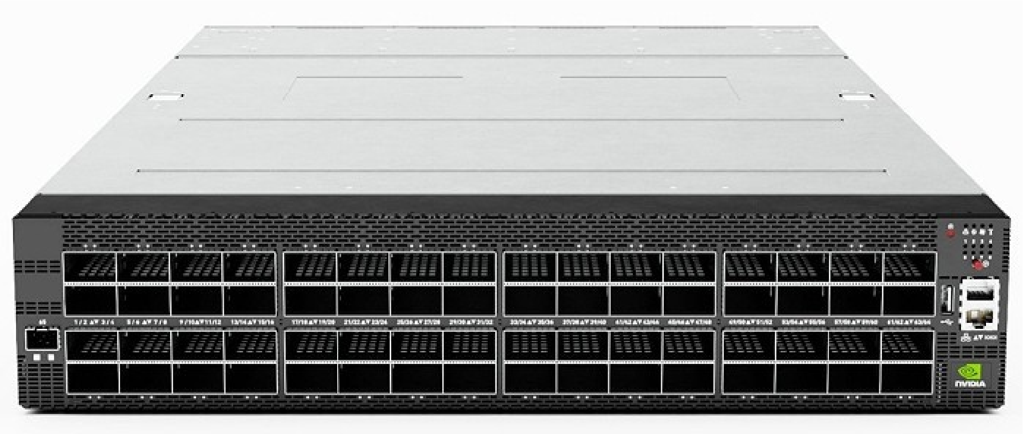

Figure 8. NVIDIA Spectrum-4 SN5600

The **Converged (North-South) Network** handles storage and in-band management, linking the Enterprise IT environment to the Agentic AI 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.

The Converged Network connects to the Enterprise IT network with up to 40 Ethernet connections at 200Gb/s for up to five Scalable Units (SU) or 64 Ethernet connections at 200Gb/s for up to eight SUs. This setup guarantees a minimum bandwidth of 25Gb/s per GPU.

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.

The Storage connectivity is exactly half that and described in the Storage Connectivity chapter.

The **Compute (East-West) Network** facilitates application communication between the GPUs across the Compute nodes of the AI platform. It is designed to achieve minimal latency and maximal performance using a rail-optimized, fully non-blocking fat tree topology with NVIDIA Spectrum-X.

Spectrum X reduces latency and increases bandwidth for Ethernet by using advanced functionality that splits network packets, tags them, and allows the switch to balance them across network lanes. The receiving node then reassembles the packets regardless of the order they were received in. This process helps avoid hash collusion and congestion, which often lead to suboptimal performance in low-entropy networks.

For cost optimization, alternative networking options can include NVIDIA SN3700 switches and NVIDIA ConnectX-7 NDR200/200GbE NICs instead of NVIDIA BlueField-3 NICs. Additionally, copper cables can replace optical cables to reduce costs

Tip: In a pure Inference use case, the Compute Network is typically not necessary, but for training and fine-tuning operations it is a crucial component of the solution.

For configurations of up to five Scalable Units, the Compute and Converged Network are integrated utilizing the same switches. When deploying more than five units, it is necessary to separate the fabric.

The following table lists the configuration of the NVIDIA Spectrum-4 SN5600.

Table 3. NVIDIA Spectrum-4 SN5600 configuration

Part Number	Description	Quantity
7D5FCTONWW	NVIDIA SN5600 800GbE Managed Switch with Cumulus	
C0Q5	NVIDIA SN5600 800GbE Managed Switch with Cumulus	2
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	4
5WS7B98401	5Yr Premier NBD Resp NVID SN5600 oPSE	2

NVIDIA Spectrum SN2201

The 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.

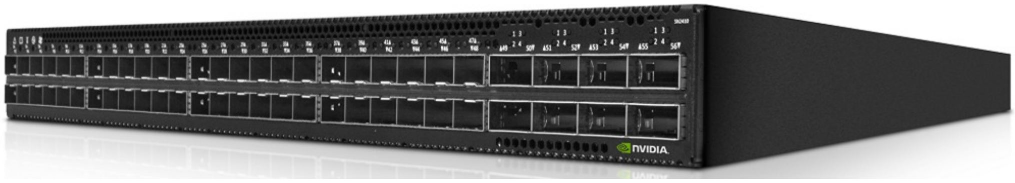

Figure 9. NVIDIA Spectrum SN2201

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.

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

Table 4. NVIDIA Spectrum 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

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.

Scalability

The descriptions provided above detail the fully expanded Lenovo Hybrid AI 285 platform with configurations of five and eight Scalable Units, respectively. A fundamental principle of the solution design philosophy, however, is its ability to support any scale necessary to achieve a particular objective.

In a typical Enterprise AI deployment initially the AI environment is being used with a single use case, like for example an Enterprise RAG pipeline which can connect a Large Language Model (LLM) to Enterprise data for actionable insights grounded in relevant data.

In its simplest form, leveraging the NVIDIA Blueprint for Enterprise RAG pipeline involves three NVIDIA Inference Microservices: a Retriever, a Reranker, and the actual LLM. This setup requires a minimum of three GPUs.

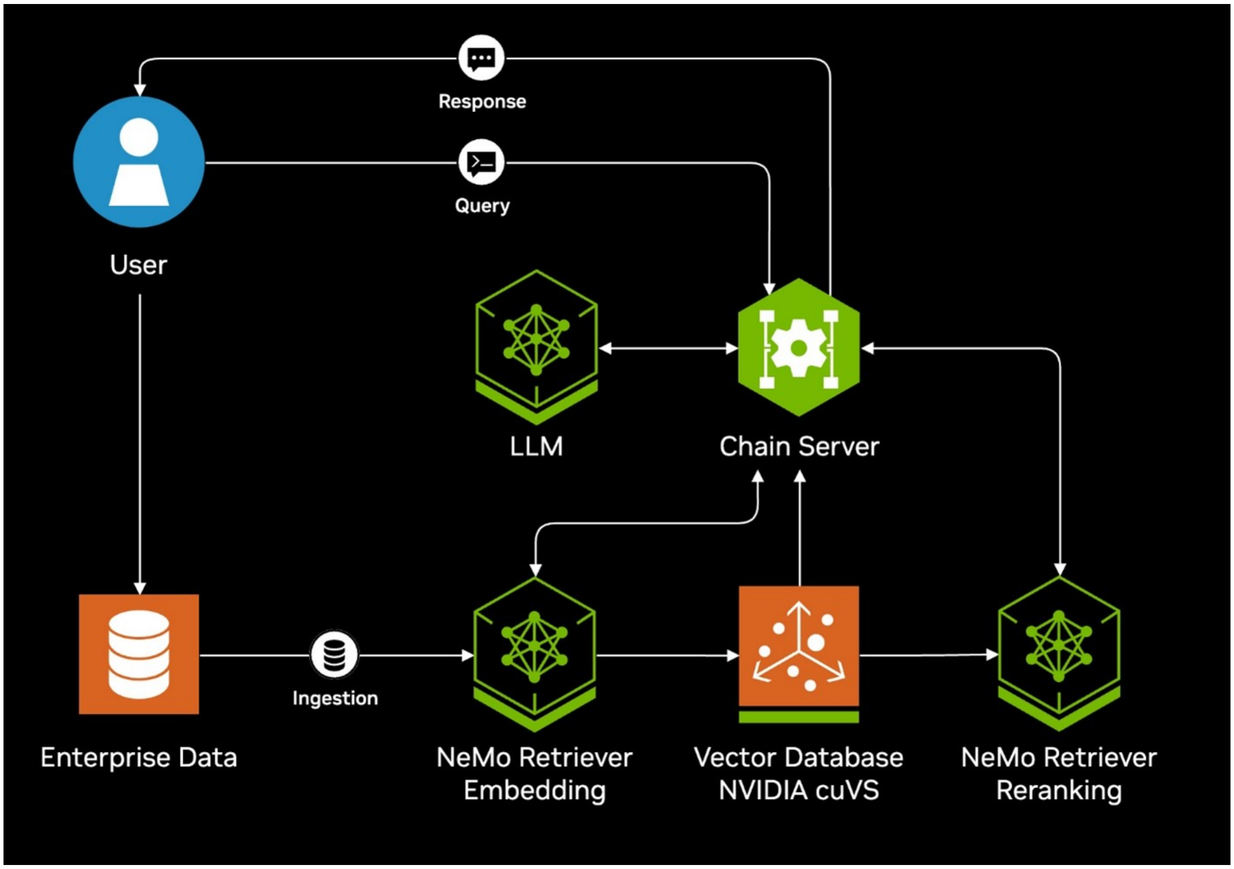

Figure 10. NVIDIA Blueprint Architecture Diagram

As the deployment of AI within the company continues to grow, the AI environment will be adapted to incorporate additional use cases, including Assistants or AI Agents. Additionally, it has the capacity to scale to support an increasing number of Microservices. Ultimately, most companies will maintain multiple AI environments operating simultaneously with their AI Agents working in unison.

The Lenovo Hybrid AI 285 platform has been designed to meet the customer where they are at with their AI application and then seamlessly scale with them through their AI integration. This is achieved through the introduction of the following:

- Starter Deployment
- Scalable Unit Deployment
- Custom Deployment

Starter Deployment

You may begin with a single AI Compute node, equipped with four GPUs. Two AI Compute nodes can be directly connected, without the need for additional networking, to scale up to sixteen GPUs if fully populated.

Starter and Starter Intermediate deployments are ideal for development, application trials, or small-scale use, reducing hardware costs, control plane overhead, and networking complexity. With all components on one node, management and maintenance are simplified.

Starter Intermediate adds another server for a total of two servers supporting up to 16 GPUs. In Starter Intermediate, the two servers connect directly without switches for back end or the compute fabric. For the front end, both servers will connect to existing switches in the data center.

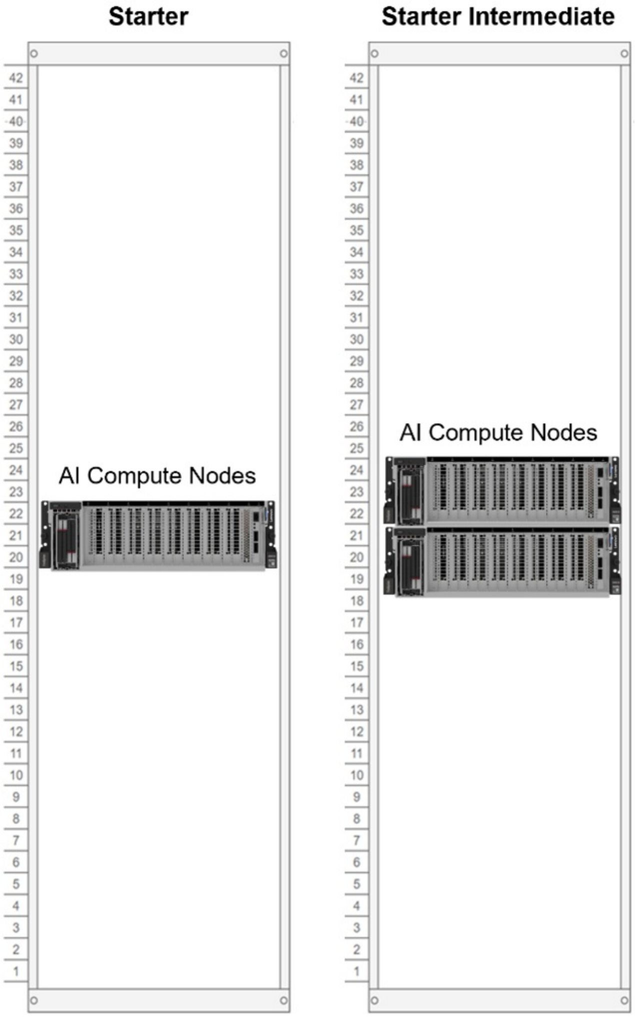

Figure 11. Starter Deployment Rack View

Scalable Unit Deployment

For configurations beyond two nodes, it is advisable to deploy a full Scalable Unit along with the necessary network and service infrastructure, providing a foundation for further growth in enterprise use cases.

The fist SU consists of up to four AI Compute nodes, minimum five service nodes, and networking switches. The first SU will support up to 32 GPUs. When additional nodes are required, additional SUs of four nodes can be added.

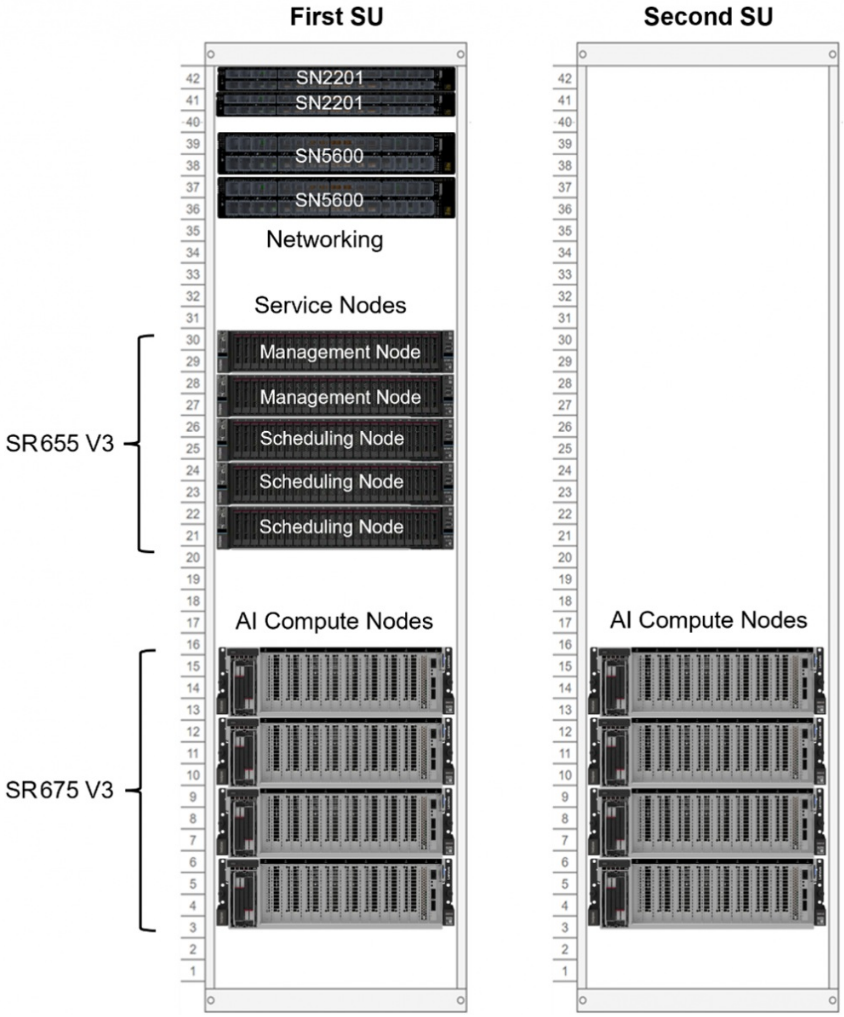

Figure 12. Scalable Unit Deployment

The networking decision depends on whether the platform is designed to support up to five or eight Scalable Units in total, and whether it will handle exclusively inference workloads or also encompass future fine-tuning and re-training activities. Subsequently, the solution can be expanded seamlessly without downtime by incorporating additional Scalable Units, ultimately reaching a total of five or eight as needed.

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.

Performance

Performance of the Lenovo Hybrid AI 285 platform is subject to the specific AI model and application used. The following table lists estimates for common Inference Benchmarks and their expected performance per AI Compute node.

The theoretical performance estimates for 8x H200 NVL listed in the table have been derived by extrapolating the MLPerf Inference v4.1 Closed, Data Center results of the H200 HGX8 data published by NVIDIA, and ratioed with the relative performance specifications provided in the official datasheet comparing the SXM and NVL systems.

Table 5. Performance

Inference Benchmark	AI Compute Node w/ 8x H200 NVL Estimates
Llama 2 70B	~29k tokens/second
Mixtral 8x7B	~50k tokens/second
GPT-J	~17k tokens/second
Stable Diffusion XL	~15 queries/second
DLRMv2 99%	~535k queries/second
DLRMv2 99.9%	~328k queries/second
BERT 99%	~62k queries/second
BERT 99.9%	~54k queries/second
RetinaNet	~12k queries/second
ResNet-50 v1.5	~636k queries/second
3D U-Net	~46 samples/second

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.

The following table provides all of the recommended software layers and their roles in the Lenovo Hybrid AI 285 platform.

Note that not all software is required to function; however, this is the recommended stack. For starter node deployments, implementing the observability functions provided by Prometheus, Grafana, and NVIDIA NetQ may not be practical.

Table 6. AI Software Stack

Software Role	Software Package
Operating System	Ubuntu
Orchestration	Base Command Manager Essentials and XClarity
Container Runtime	Containerd
Container Orchestration	Kubernetes
Container Network Interface (CNI)	Calico
Load Balancer - API Service Gateway	Nginx
Load Balancer – Network Services	MetalLB
Ingress Controller	Nginx

Software Role	Software Package
Package Manager	Helm
Operator	GPU Operator
Operator	Network Operator
Operator	NIM Operator
RBAC	Permission Manager
Observability	Prometheus
Observability	Grafana
Observability	NVIDIA NetQ
Storage	NFS Provisioner

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

- XClarity System Management
- Linux Operating System
- Container Orchestration
- Data and AI Applications
- NVIDIA AI Enterprise

XClarity System Management

Lenovo XClarity Administrator is a centralized, resource-management solution that simplifies infrastructure management, speeds responses, and enhances the availability of Lenovo server systems and solutions. It runs as a virtual appliance that automates discovery, inventory, tracking, monitoring, and provisioning for server, network, and storage hardware in a secure environment.

Table 7. XClarity System Management

Part Number	Description
00MT203	Lenovo XClarity Pro, Per Managed Endpoint w/5 Yr SW S&S

Linux Operating System

The AI Compute nodes are deployed with Linux. Traditionally Canonical Ubuntu is the default choice for AI environments with its optimizations across all prominent AI hardware platforms and up to 12 years of security maintenance, but Lenovo Hybrid AI platforms support Red Hat Enterprise Linux (RHEL) as alternative choice.

Table 8. Linux Operating System

Part number	Description
7S1B000YWW	Canonical Ubuntu Pro 5Yr w/Canonical weekday Support

Container Orchestration

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

Canonical Kubernetes is used across industries for mission critical workloads, and uniquely offers up to 12 years of security for those customers who cannot, or choose not to upgrade their Kubernetes versions.

MicroK8s is the easiest and fastest way to get Kubernetes up and running. With self-healing high availability, transactional OTA updates and secure sandboxed kubelet environments, MicroK8s is recommended for the starter node (single and dual node) deployments.

When implementing a Scalable Unit Deployment and above Ubuntu Charmed Kubernetes is used.

When choosing Red Hat for the Operating System, Red Hat OpenShift as the matching Kubernetes implementation is required.

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 285 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.

> **Tip**: Entitlement to NVIDIA AI Enterprise for 5 years is included with the NVIDIA H200 NVL PCIe GPU.

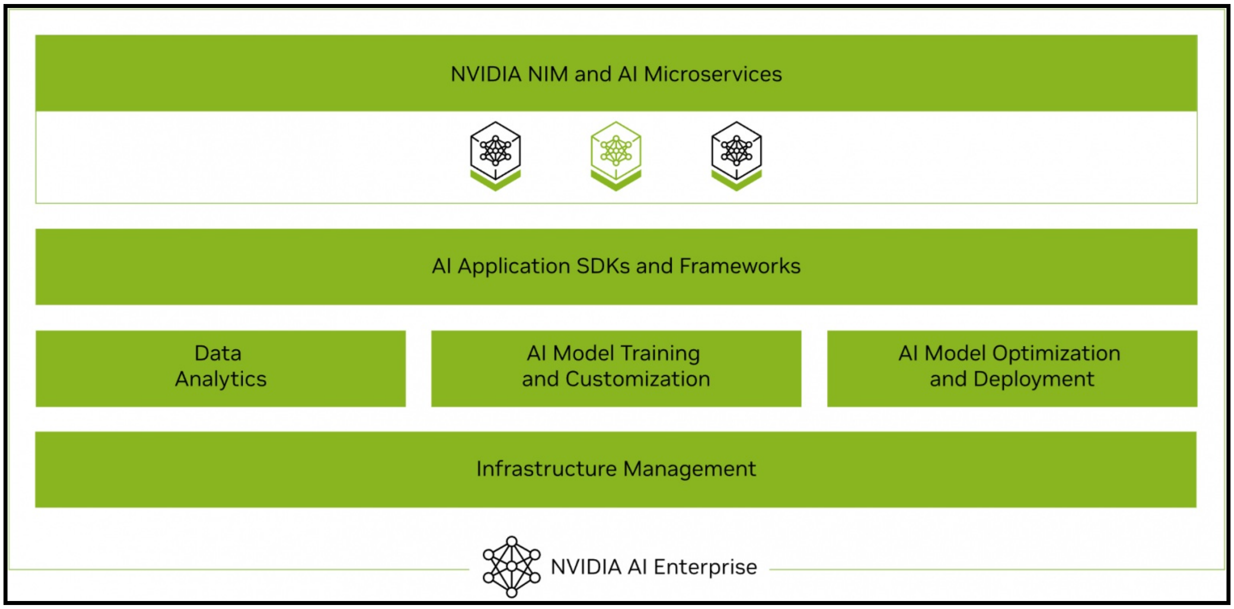
NVIDIA NIM and AI Microservices
AI Application SDKs and Frameworks
Data Analytics
AI Model Training and Customization
AI Model Optimization and Deployment
Infrastructure Management
NVIDIA AI Enterprise

Figure 13. 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 285 platform.

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.

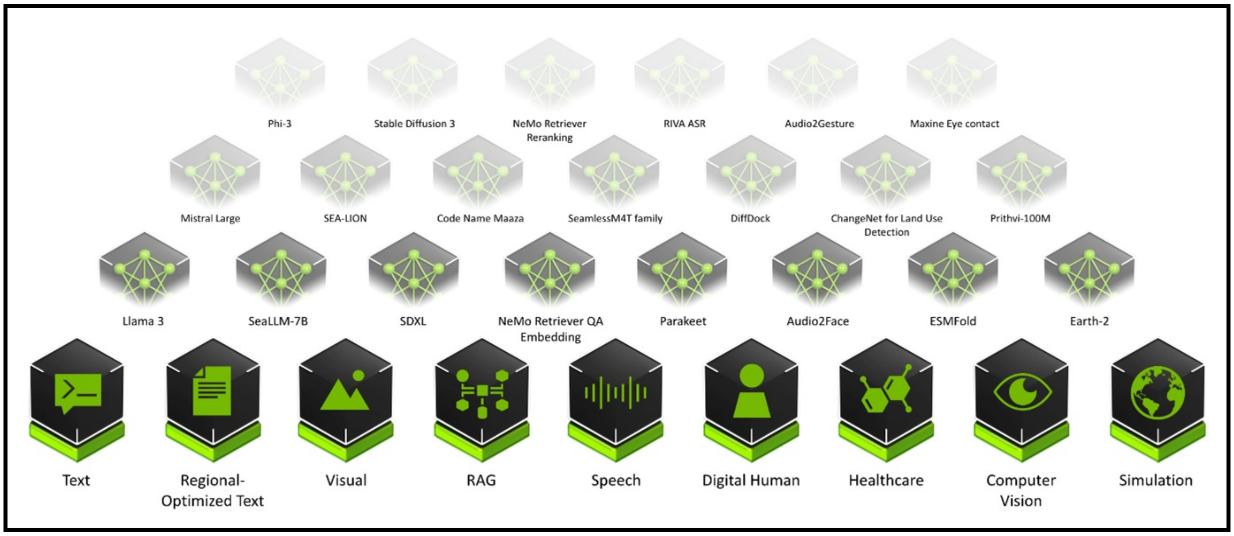

Figure 14. NVIDIA Inference Microservice catalog

Storage Connectivity

Lenovo Hybrid AI platforms do not include storage but do interface with any storage technology that is validated by NVIDIA for NVIDIA OVX or certified by NVIDIA for AI Enterprise.

Lenovo Storage validated by NVIDIA includes: Lenovo DM Series, Lenovo DG Series, Lenovo ThinkAgile HX series with Nutanix and ThinkAgile VX series with VMware, and Lenovo DSS-G for IBM Storage Scale. Lenovo is a qualified hardware platform for Cloudian, DDN, and WEKA.

The storage directly attached to the AI platform primarily hosts the vector database supporting data retrieval for Retrieval-Augmented Generation (RAG) applications. Additionally, it functions as high-performance storage for retraining or fine-tuning models.

The Converged Network connects to the Storage with up to 20 Ethernet connections at 200Gb/s for up to five Scalable Units (SU) or 32 Ethernet connections at 200Gb/s for up to eight SUs. This setup guarantees a minimum bandwidth of 12.5Gb/s per GPU. The connected storage system should be configured to support this bandwidth requirement per GPU accordingly.

> **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"

Content-Aware IBM Storage Scale

Lenovo Hybrid AI platforms will integrate with Content-Aware IBM Storage Scale (CAS) to enhance the value of GenAI applications by enabling semantic understanding of data at the storage layer.

Frontier large language models have incorporated most of the world's publicly available data, but less than 1% of enterprise data are represented in them. RAG is a popular technique for vectorizing data so that it can be used by these models, but it has limitations. Most RAG implementations suffer from stale data, high costs, security issues, and operational complexity. These challenges are addressed by CAS though the deep integration of the vector processing pipeline within the parallel file system, minimizing data movement and latency, resulting in efficiency gains.

CAS provides automated processing of unstructured data for use in RAG applications using NVIDIA NIMs and the NVIDIA Multi-Modal PDF Data Extraction blueprint. The Storage Scale parallel file system provides shared storage optimized for NVIDIA NIM inferencing operations facilitating scale out processing with extreme efficiency.

CAS runs on the 2-8-5 AI compute nodes and uses IBM Storage Scale Container Native (CNSA) to attach to an external IBM Storage Scale SDS file system like Lenovo Distributed Storage Solution for IBM Storage Scale (DSS-G)

IBM Storage Scale Active File Manager (AFM) may be used to ingest data from external S3 data sources into the RAG pipeline, and in future releases integrate with most any 3rd party storage system, bringing data vectorization to a client's existing environment without the need for an expensive rip and replace. Customers may also ingest data residing on the Storage Scale filesystem into the RAG pipeline. In addition, CAS leverages AFM for incremental data ingest, CNSA for scale out, load balancing, and redundancy. The vector database and other data derivatives are contained within the storage layer, greatly reducing infrastructure requirements when compared to the current industry practice of maintaining the vectors in memory.

These and other innovations provide customers with faster time to insights, reduced costs, improved performance, and simplified operations. Combining the benefits of CAS with Lenovo EveryScale delivers a comprehensive, storage and compute optimized, turnkey solution for enterprises to extract value from GenAI and Agentic AI applications leveraging the latent power of their unique enterprise data.

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, including NVIDIA L40S and NVIDIA HGX8 H200.

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

Lenovo Hybrid AI platforms solutions are specifically designed to enable broad adoption in the Enterprise supported by Lenovo's powerful IT partner ecosystem.

In addition to custom deployments, as the foundation of NVIDIA's Enterprise Reference Architecture, it is fully compatible with NVIDIA Blueprints for agentic and generative AI use cases.

This enables both Lenovo AI Partners and Lenovo Professional Services to accelerate deployment and provide enterprises with the fastest time to production.

Lenovo's Hybrid AI Advantage enables customers to overcome the barriers they face in realizing ROI from AI investments by providing critical expertise needed to accelerate business outcomes. Leveraging a responsible approach to AI, Lenovo AI expertise, Lenovo's advanced partner ecosystem and industry leading technology we help customers realize the benefits of AI faster.

AI Discover Workshop

Lenovo AI Discover Workshops help customers visualize and map out their strategy and resources for AI adoption to rapidly unlock real business value. Lenovo's experts assess the organization's AI readiness across security, people, technology, and process – a proven methodology – with recommendations that put customers on a path to AI success. With a focus on real outcomes, AI Discover leverage proven frameworks, processes and policies to deliver a technology roadmap that charts the path to AI success.

AI Fast Start

With customers looking to unlock the transformative power of AI, Lenovo AI Fast Start empowers customers to rapidly build and deploy production-ready AI solutions tailored to their needs. Optimized for NVIDIA AI Enterprise and leveraging accelerators like NVIDIA NIMs, Lenovo AI Fast Start accelerates use case development and platform readiness for AI deployment at scale allowing customers to go from concept to production ready deployment in just weeks. Easy to use containerized and optimized inference engines for popular NVIDIA AI Foundation models empower developers to deliver results. AI Fast Start provides access to AI Experts, platforms and technologies supporting onsite and remote models to achieve business objectives.

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 - First Scalable Unit

This section provides an example Bill of Materials (BoM) of one Scaleable Unit (SU) deployment with NVIDIA Spectrum-X. This example BoM includes:

- 4x Lenovo ThinkSystem SR675 V3 with 8 × NVIDIA H200 NVL GPUs per server
- 5x Lenovo ThinkSystem SR655 V3
- 2x NVIDIA SN5600 Switches
- 2x NVIDIA SN2201 Switches

Storage is optional and not included in this BoM.

In this section:

ThinkSystem SR675 V3 BoM

Table 9. ThinkSystem SR675 V3 BoM

Part Number	Product Description	Qty per System	Total Qty
7D9RCTOLWW	ThinkSystem SR675 V3		4
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1	4
C3EF	ThinkSystem SR675 V3 System Board v2	1	4
C2AL	ThinkSystem AMD EPYC 9535 64C 300W 2.4GHz Processor	2	8
C0CK	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM-A	24	96
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	2	8
C3V3	ThinkSystem NVIDIA H200 NVL 141GB PCIe GPU Gen5 Passive GPU	8	32
C3V0	ThinkSystem NVIDIA 4-way bridge for H200 NVL	2	8
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1	4
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1	4
C2RK	ThinkSystem SR675 V3 2 x16 Switch Cabled PCIe Rear IO Riser	2	8
C0Q4	ThinkSystem NVIDIA BlueField-3 B3140H VPI QSFP112 1P 400G PCIe Gen5 x16 Adapter	4	16
BM8X	ThinkSystem M.2 SATA/x4 NVMe 2-Bay Adapter	1	4
BT7P	ThinkSystem Raid 540-8i for M.2/7MM NVMe boot Enablement	1	4
BXMH	ThinkSystem M.2 PM9A3 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2	8
BTMB	ThinkSystem 1x4 E3.S Backplane	1	4
C1AB	ThinkSystem E3.S PM9D3a 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2	8
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	1	4
C5WW	ThinkSystem SR675 V3 Dual Rotor System High Performance Fan	5	20
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	1	4
BE0D	N+1 Redundancy With Over-Subscription	1	4
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4	16
6252	2.5m, 16A/100-250V, C19 to C20 Jumper Cord	4	16
C3KA	ThinkSystem SR670 V2/SR675 V3 Heavy Systems Toolless Slide Rail Kit	1	4
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	1	4
BR7U	ThinkSystem SR675 V3 Root of Trust Module	1	4
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	1	4
5PS7B09631	5Yr Premier NBD Resp + KYD SR675 V3	1	4

ThinkSystem SR655 V3 BoM

Table 10. ThinkSystem SR655 V3 BoM

Part Number	Product Description	Qty per System	Total Qty
7D9ECTOLWW	ThinkSystem SR655 V3		5
BLKK	ThinkSystem V3 2U 24x2.5" Chassis	1	5
C2AC	ThinkSystem SR655 V3 MB w/IO+PIB+FB,2U	1	5
C2AQ	ThinkSystem AMD EPYC 9335 32C 210W 3.0GHz Processor	1	5
C0CJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM-A	12	60
BPQV	ThinkSystem V3 2U x16/x16/E PCIe Gen5 Riser1 or 2	1	5
BVBG	ThinkSystem NVIDIA BlueField-3 B3220 VPI QSFP112 2P 200G PCIe Gen5 x16 Adapter	1	5
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1	5
BXMH	ThinkSystem M.2 PM9A3 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2	10
BS7Y	ThinkSystem V3 2U 8x2.5" NVMe Gen5 Backplane	1	5
C0ZU	ThinkSystem 2.5" U.2 Multi Vendor 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	2	10
BLL6	ThinkSystem 2U V3 Performance Fan Module	6	30
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2	10
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	2	10
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1	5
C1PT	ThinkSystem SR635 V3/SR655 V3 Root of Trust Module Low Voltage-RoW V2	1	5
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1	5
5PS7B08762	5Yr Premier NBD Resp + KYD SR655 V3	1	5

NVIDIA SN5600 Switch BoM

Table 11. NVIDIA SN5600 Switch BoM

Part Number	Product Description	Qty per System	Total Qty
7D5FCTONWW	NVIDIA SN5600 800GbE Managed Switch with Cumulus (oPSE)		2
C0Q5	NVIDIA SN5600 800GbE Managed Switch with Cumulus (oPSE)	1	2

NVIDIA SN2201 Switch BoM

Table 12. NVIDIA SN2201 Switch BoM

Part Number	Description	Qty per System	Total Qty
7D5FCTOFWW	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)		2
BPC7	NVIDIA SN2201 1GbE Managed Switch with Cumulus (PSE)	1	2
6201	1.5m, 10A/100-250V, C13 to C14 Jumper Cord	2	4
5WS7B98268	5Yr Premier NBD Resp NVID SN2201 PSE	1	2

Power Distribution Unit (PDU) BoM

Table 13. Power Distribution Unit (PDU) BoM

Part Number	Product Description	Qty per System	Total Qty
7DGMCTO1WW	-SB- 0U 18 C13/C15 and 18 C13/C15/C19 Switched and Monitored 63A 3 Phase WYE PDU v2		2

Rack Cabinet BoM

Table 14. Rack Cabinet BoM

Part Number	Product Description	Qty per System	Total Qty
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
BHC7	ThinkSystem 42U Onyx Heavy Duty Rack Side Panel	2	2
BJPA	ThinkSystem 42U Onyx Heavy Duty Rack Rear Door	1	1
5AS7B07693	Lenovo EveryScale Rack Setup Services	1	1

XClarity Software BoM

Table 15. XClarity Software BoM

Part Number	Product Description	Qty per System	Total Qty
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)		3
00MT203	Lenovo XClarity Pro, Per Managed Endpoint w/5 Yr SW S&S		5

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:

- Artificial Intelligence
- ThinkSystem SR655 V3 Server
- ThinkSystem SR675 V3 Server

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, LP2181, was created or updated on March 19, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2181
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2181.

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®
ThinkAgile®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Intel® is a trademark of Intel Corporation or its subsidiaries.

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

Microsoft® and Azure® are trademarks 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.