Lenovo Hybrid AI Software Platform
Product Guide

Lenovo's solutions encompass the entire technology stack, from infrastructure to tooling, needed to stand up a production-ready AI Factory. This software platform provides an overview of every software and firmware component that is recommended to take full advantage of Lenovo's Hybrid AI Platforms.

This document does not intend to provide a deep dive into the exact hardware, software, and network configurations recommended to set up an AI cluster. Instead, this document should serve as an overview of the recommended combination of software for AI workloads on a Lenovo Hybrid AI platform. Furthermore, a set of recommendations for how to size and scale Large Language Models (LLMs) on the Hybrid AI platform is provided for once the software stack is applied. For expert guidance on deploying this software platform, see the AI Services section.

For expert guidance on setting up and maintaining this software platform this document provides a description of Lenovo's deployment and setup services to help organizations deploy their Hybrid AI Factory quickly, bridging expertise gaps where necessary. Different deployment options are available to provide services in a flexible manner.

The Lenovo Hybrid AI Software platform is meant to be used alongside the Lenovo Hybrid AI 221, 285 and 289 platforms. Each of these platforms provides different use cases depending on the organization's needs. This document lays out a standardized software stack across all of the hardware focused platforms, however depending on the deployment scenario and requirements the configuration can be changed.

The Hybrid AI 221 platform provides a lower cost starting point for enterprises looking to deploy AI infrastructure. Intended for inference-focused workloads on a single node or smaller deployments, the 221 platform provides flexibility to scale up and out into a 285 configuration over time if needed.

The Hybrid AI 285 platform is ideal for organizations that wish to scale their AI platforms as needed, beginning with a minimal starter kit for a small-scale operation with the capability to expand to a full-fledged AI Factory without needing to create a whole new solution. This platform is best suited for organizations wanting to run mostly AI inference workloads that may need the capacity to scale in the future.

The Hybrid AI 289 platform is meant for the heavy-duty AI workloads of tomorrow – scaling, fine-tuning, and serving AI models at large scales. With the 289 platform, enterprises will get maximum usage of the compute power available to them with ultra-high bandwidth networking and top tier GPU performance. This platform is best suited for organizations that need the most performance for both training and inference.123

Requirements

Please refer to the following Software Platform for more details however the stack has been listed below.

- Software Requirements
- Service & Support Offerings
- Hardware Requirements for Full Deployment (Infrastructure Nodes)

Software Requirements

Table 1. Software with versions, broken out by deployment type

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

Service & Support Offerings

The following table lists the Service and support offerings

Table 2. Service and support offerings by deployment type

Service	Scalable Unit Sizing	Less Than 1 Scalable Unit	Single Node
GPU Advanced Services – Plan & Design with Kubernetes	Custom services engagement	Custom services engagement	5MS7C33028
GPU Advanced Services – Configuration & Deployment with Kubernetes	Custom services engagement	Custom services engagement	5MS7C33029
GPU Advanced Services – Managed Services with Kubernetes	Custom services engagement	Custom services engagement	5MS7C33030

Hardware Requirements for Full Deployment (Infrastructure Nodes)

The following table lists the Scalable Unit deployment

Table 3. Service nodes and purpose

Role	Software Installed	Purpose
Head Node 1	Base Command Manager	Central deployment and management for cluster
Head Node 2	Base Command Manager	
Kubernetes Control Plane 1	Kubernetes + ETCD	Manage the Kubernetes cluster and its components
Kubernetes Control Plane 2	Kubernetes + ETCD	
Kubernetes Control Plane 3	Kubernetes + ETCD	
VM Node 1	Login, NetQ, XClarity	Login nodes that provide access to the cluster containing multiple applications running in VMs
VM Node 2	Login, NetQ, XClarity	

KVM or VMWare can be used as the hypervisor for the VM Nodes above. The VMWare license is outside the scope of this document. At the time of writing XClarity One for on-prem VM is only available on VMWare.

Note: For a Starter Kit deployment, the infrastructure nodes and portions of the software stack are not recommended. Table 1 shows which software components are needed for each deployment type, and the Deployment Sizing Differences section section contains additional details.

AI Software Stack

Deploying AI to production involves implementing multiple layers of software. The process begins with the server BIOS tuning, 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 tools and agents to users.

The following graphic shows an overview the recommended stack broken down by functionality.

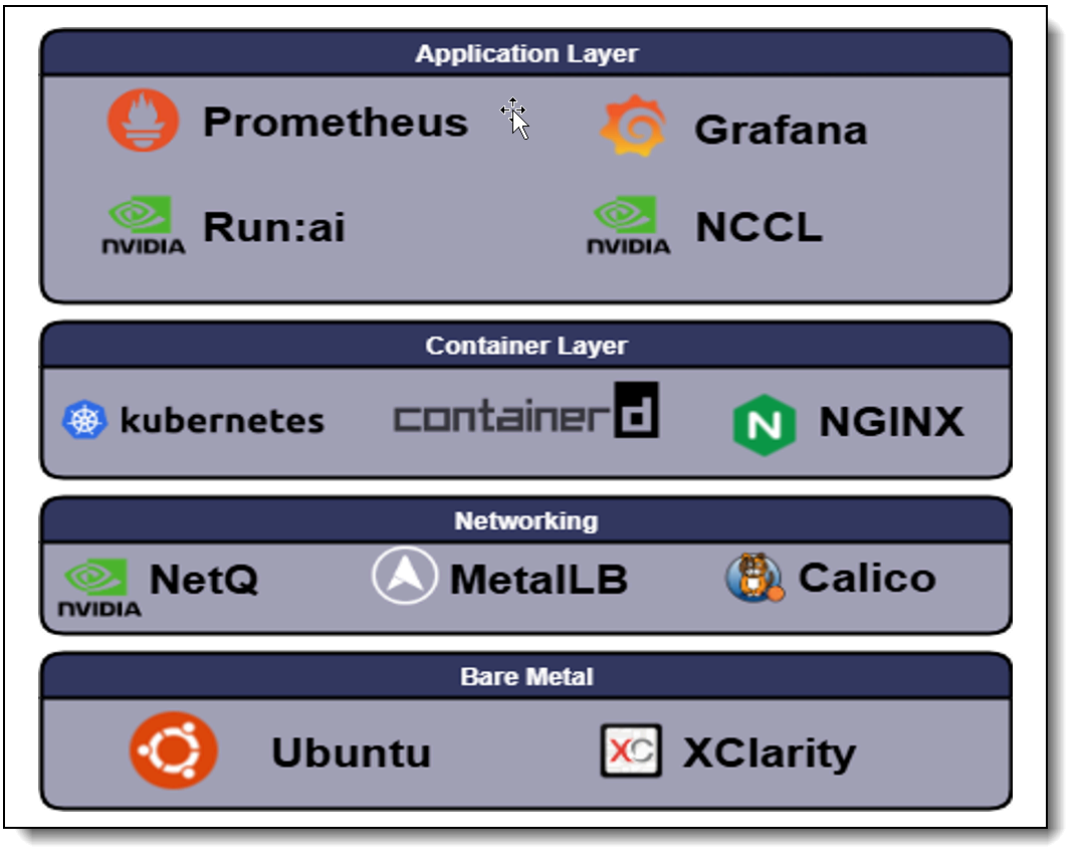

Figure 1. Recommended software stack by layer

Note that not all software is required to create a functioning AI solution; however, this is the recommended stack. Please see Table 1 and the Deployment Sizing Differences section for details on the recommended stack for smaller deployments. In the following sections, we take a deeper dive into the software elements:

- Linux Operating System
- Lenovo XClarity One
- UEFI Operating Modes for AI Clusters
- NVIDIA Cumulus Linux – Networking Operating System
- NVIDIA Base Command Manager - Provisioning
- Kubernetes Container Orchestration
- Containerd - Container Runtime
- Prometheus - Monitoring
- Grafana - Visualization
- Permission Manager - RBAC
- Calico - Container Network Interface
- NGINX – Ingress Controller
- MetalLB - Load Balancer
- Operators
- DOCA-Host - Networking
- NVIDIA Network Congestion Control
- NCCL - Communication
- Run:ai - Orchestration
- NetQ - Visibility & Validation
- Kubernetes NFS Provisioner - Storage
- NVIDIA AI Enterprise
- NVIDIA Inference Microservice
- Deployment Sizing Differences

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.

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

XClarity One can be installed flexibly, either hosted in the Lenovo cloud with on-premises management hubs or as a fully on-premises solution.

For more information on XClarity One functionality https://pubs.lenovo.com/lxc1/

UEFI Operating Modes for AI Clusters

As a rule, servers being used for cluster management and administration workloads should have their UEFI Operating mode set to one of the ‘Power Efficiency' modes and GPU Servers should have their UEFI Operating Mode set to "Max Performance”.

The UEFI Operating Mode on Lenovo servers can be configured in a variety of ways

- Specified during the Lenovo ordering process and configured at the factory
- Modified on-site using the XCC LXPM tool that is accessed using F1 at boot time
- Using Redfish API for AMD Systems or Intel Systems

Note. The performance of x86 servers can be influenced by the UEFI configuration and recommendations will vary depending on the GPUs, NICs and intended workloads.

NVIDIA Cumulus Linux – Networking Operating System

Cumulus Linux is the Network Operating System (NOS) running on the NVIDIA Switches, such as the SN5610. Cumulus Linux includes the functionality required to set up networking for the data center, accessible through a declarative command-line interface. Cumulus Linux NOS comes installed on NVIDIA switches by default and includes the NetQ agent and CLI. NetQ is used to monitor and manage the data center network infrastructure and its operational health.

NVIDIA Base Command Manager - Provisioning

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.

- Type 1: All communication is centralized through the head node, providing a controlled and secured gateway
- Type 2: Worker nodes can be accessed directly via a router, so that traffic does not need to go through the head node.
- Type 3: A routed public network is used, where regular nodes are on Internalnet and the head node is on Managementnet.

The following table displays the recommended BCM Networks for a Type 2 network.

Table 4. BCM Networks

BCM Network	Traffic Type	Purpose	Creation Phase
managementnet	IPMI	OOB Management	BCM Deployment
internalnet	Management/Data	In-band management	BCM deployment
storagenet	Storage	Backend Storage	Storage Configuration
failovernet	High Availability	HA Heartbeat	HA Configuration, setup by deployment wizard
Kube-cluster-pod	Pod network	Pod-to-pod traffic	K8s deployment, setup by deployment wizard
Kube-cluster-service	Service Network	App traffic	K8s deployment, setup by deployment wizard

Kubernetes 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.

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.

Containerd - Container Runtime

Containerd is a container runtime environment that manages the container lifecycle of the containers on the host systems. Containerd is one of the most common runtimes used by Kubernetes, and is available as a daemon for Linux.

Note - when choosing Red Hat for the Linux Operating System, Red Hat OpenShift will use CRI-O as the container runtime and “Podman” in place of Docker for CLI.

Prometheus - Monitoring

Prometheus collects metrics from Kubernetes components (such as Kubelet API servers, etcd) and workloads running in the pods. It is used for monitoring GPU usage, container health, and job performance. Grafana ingests data from Prometheus to gain insight into the Kubernetes cluster and visualize data, leading to easier diagnosis of issues.

Prometheus collects data from targets by scraping HTTP endpoints. Any application that exposes data via and HTTP endpoint can have its data collected by Prometheus. Therefore, how data is collected is flexible and entirely up to the user dependent upon the use case. Through the nominal setup process with BCM, Prometheus will be gathering data from BCM. The NVIDIA GPU Operator automatically deploys and makes discoverable DCGM Exporter to collect real-time GPU telemetry data. NIM Operator can also be configured to share NIM metrics with Prometheus.

BCM setup wizard facilitates Prometheus and Grafana setup, including Prometheus Operator, Node Exporter, and Kube State Metrics. The Prometheus instance is installed in the BCM head node to facilitate integration with BCM and Grafana.

Grafana - Visualization

Grafana is a data visualization tool that provides a dashboard for viewing various metrics from a database, creating alert systems, and reducing telemetry costs. With the large amounts of data being generated from containers, GPUs, and networking systems, Grafana offers a centralized location to observe, track, and respond to that data to enable faster troubleshooting and easier data insight. For this software stack, Grafana is displaying data from Prometheus.

Ensure that Grafana Data Sources are configured to include Prometheus. Custom dashboards are available from NVIDIA on GitHub.

Permission Manager - RBAC

Permission Manager is an open-source RBAC management tool for Kubernetes with a web UI. It allows user creation, namespace and permission assignment, and Kubeconfig YAML file distribution.

Calico - Container Network Interface

Calico is the network layer between containers and the NICs.

- Used to define Kubernetes Network Policies to control communication between pods, namespaces, and services.
- Enforces Microsegmentation, isolating workloads at the network level. This protects sensitive AI models and data, especially when using shared GPU infrastructure across departments or clients.
- Designed for high-performance networking, which is essential when moving large datasets or model checkpoints between pods.

NGINX – Ingress Controller

NGINX is used to expose services in Kubernetes clusters, acting as an ingress controller that routes external traffic to internal services based on defined rules. NGINX can be configured as a load balancer for Kubernetes API servers, providing high availability and fault tolerance for control plane communication.

Designed for high-performance and scalability, NGINX handles large volumes of traffic efficiently, which is critical for serving AI models and streaming data. The NGINX Ingress Controller add-on is installed by default in cluster setup with Base Command Manager unless specified otherwise

MetalLB - Load Balancer

MetalLB is a load-balancer implementation for bare-metal Kubernetes clusters. It provides external access to services using standard Layer 2 (ARP/NDP) or BGP routing, enabling LoadBalancer type services without cloud provider support. MetalLB ensures reliable traffic routing in on-prem environments.

MetalLB works with the other technologies in our software stack:

- Prometheus – Exposes metrics on load balancer status and IP allocations for infrastructure monitoring.
- Calico – Can be configured to announce the LoadBalancer IPs via BGP.
- NGINX – Provides IP-level load balancing for services exposed via NGINX ingress controllers.

Operators

- **GPU Operator –** Container scheduling on GPUs across server nodes
- **Network Operator –** Container overlay networking switching and routing with network infrastructure
- **NIM Operator –** Oversight of NIM to ensure that correct GPU profile and settings are applied for optimal model performance
- **Prometheus Operator** – Collection of system metrics for monitoring and performance analysis

DOCA-Host - Networking

DOCA-Host is a package in NVIDIA's DOCA software framework that contains the host drivers and tools necessary to operate BlueField and ConnectX devices. It is available on Linux systems as a standalone package.

DOCA offers 4 installation profiles supporting different deployment scenarios and workload types:

- doca-all
- doca-networking
- doca-ofed
- doca-roce

For BlueField devices it is generally recommended to use doca-all, and for ConnectX it is generally recommended to use doca-networking.

NVIDIA Network Congestion Control

The NVIDIA Congestion Control algorithm was enhanced in the DOCA 2.9 release. Ensure that the version installed as part of DOCA-Host matches the listed requirement or exceeds it.

NCCL - Communication

The NVIDIA Collective Communications Library (NCCL) is a library providing inter-GPU and multi-node communication primitives that are topology-aware and optimized for NVIDIA GPUs. Most major deep learning frameworks have integrated NCCL to accelerate deep learning on NVIDIA multi-GPU systems. On Ubuntu and Red Hat systems, NCCL is available as a package.

Run:ai - Orchestration

NVIDIA Run:ai is a Kubernetes-native orchestration platform that provides GPU allocation, resource management, and AI Lifecycle Integration. Run:ai will allow users to run more workloads by increasing GPU utilization and replaces the default Kubernetes scheduler with a more purpose-built AI scheduler.

Three modes of operation are supported (Saas, Self-Hosted, and Air-Gapped), and for this Software ERA the Saas mode is recommended.

Run:ai must be installed after Containerd and Kubernetes, as well as requiring some pre-configuration of said Kubernetes cluster. The exact requirements can be found on NVIDIAs documentation page.

NetQ - Visibility & Validation

NVIDIA NetQ is a network operations toolset that works for Cumulus NOS. NetQ provides network management, telemetry, data visualization, and validation.

Kubernetes NFS Provisioner - Storage

NFS Provisioner (NFS Subdir External Provisioner) is a Kubernetes external storage plugin that enables dynamic provisioning of Persistent Volumes using an existing NFS server. Storage management is simplified for AI workloads that require persistent data across Lenovo server nodes and restarts. The compatibility with existing NFS infrastructure makes NFS Provisioner ideal for hybrid environments where centralized storage is shared across multiple clusters. Integration with Helm provides easy deployment into Kubernetes clusters, allowing platform teams to configure NFS server details and mount paths via Helm values. As related to this Software ERA, Run:ai Data Sources support both direct NFS mount and Kubernetes Storage Classes.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is a comprehensive suite of artificial intelligence and data analytics software designed for optimized development and deployment in enterprise settings. This section outlines some of the remaining tools present in NVIDIA AI Enterprise that are not already discussed in previous sections.

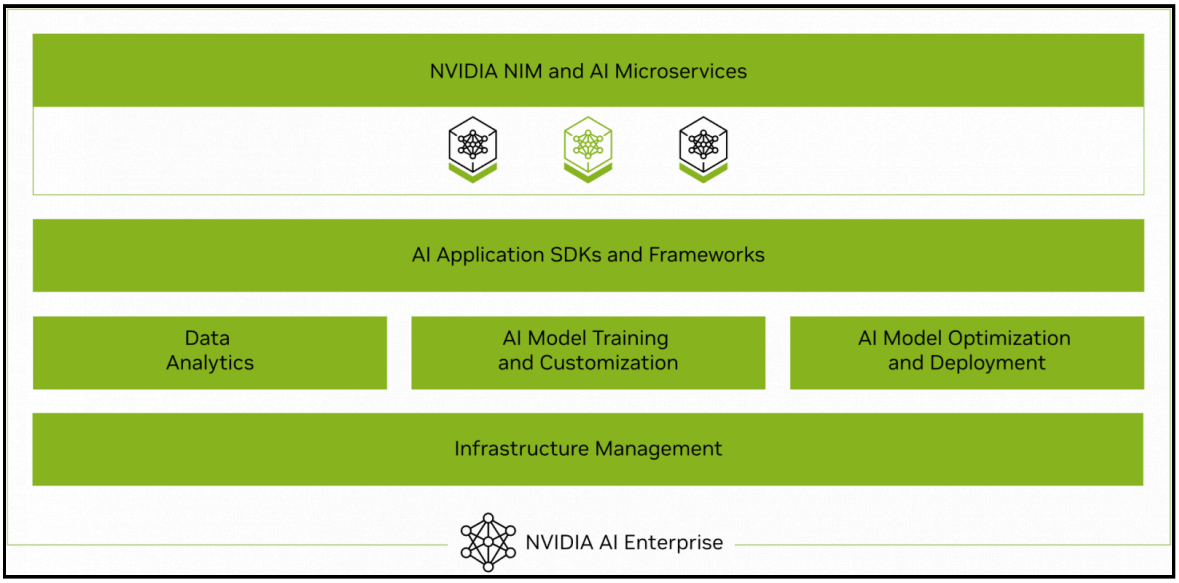

Figure 2. NVIDIA AI Enterprise

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. An NVIDIA AI Enterprise License provides access to the fully secured, vetted, and tested software artifacts that are supported by NVIDIA.

NVIDIA Inference Microservice

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.

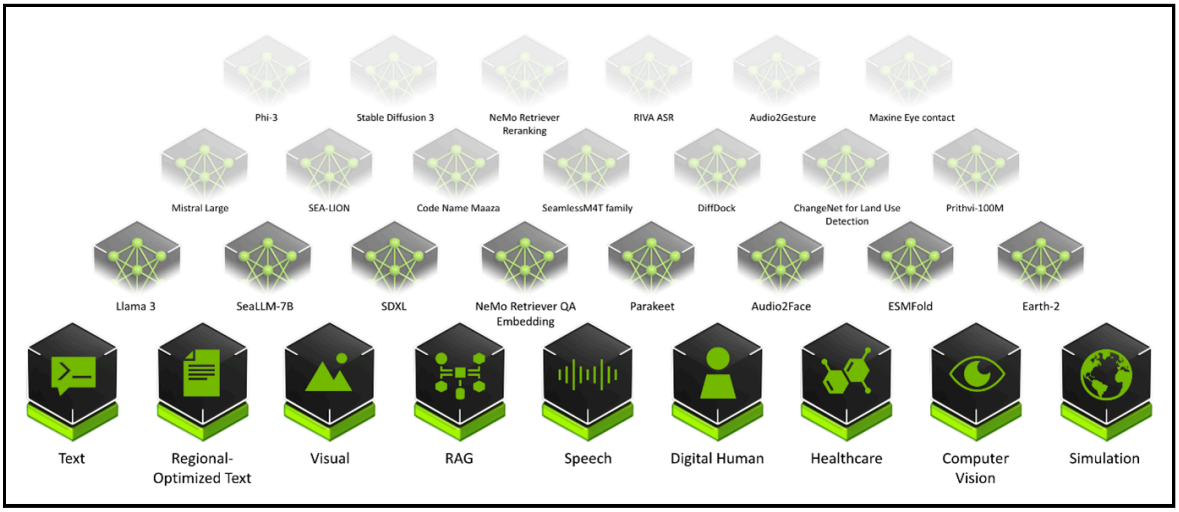

Figure 3. Examples of NIM Catalogue

For an in-depth guide to the NVIDIA Software portfolio with Lenovo Part Numbers, please reference the NVIDIA Software Product Guide on Lenovo Press.

Deployment Sizing Differences

For an AI Starter Kit deployment, the Kubernetes control plane operates directly on the AI Compute nodes, negating the requirement for dedicated 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 AI Starter Kit or Single Node deployments is similar to the recommended full AI Software Stack, though some components may be considered optional or less practical for these smaller configurations.

For a Single Node deployment, there will be no networking. Therefore, the following components of the Software Stack will not be needed:

- Network Operator
- NVIDIA Network Congestion Control
- NetQ
- Cumulus Linux (No Switch in Single Node Deployment)
- Run:ai
- Lenovo XClarity One

For a deployment of less than 1 SU the following components are not needed but may be added depending on customer needs:

- NetQ
- Run:ai

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 Figure 4 below. AI Fast Start Services provide use case development and validation for agentic AI and GenAI applications. The 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 4 shows some of the possible combinations of software and orchestration that are found in this Reference Architecture.

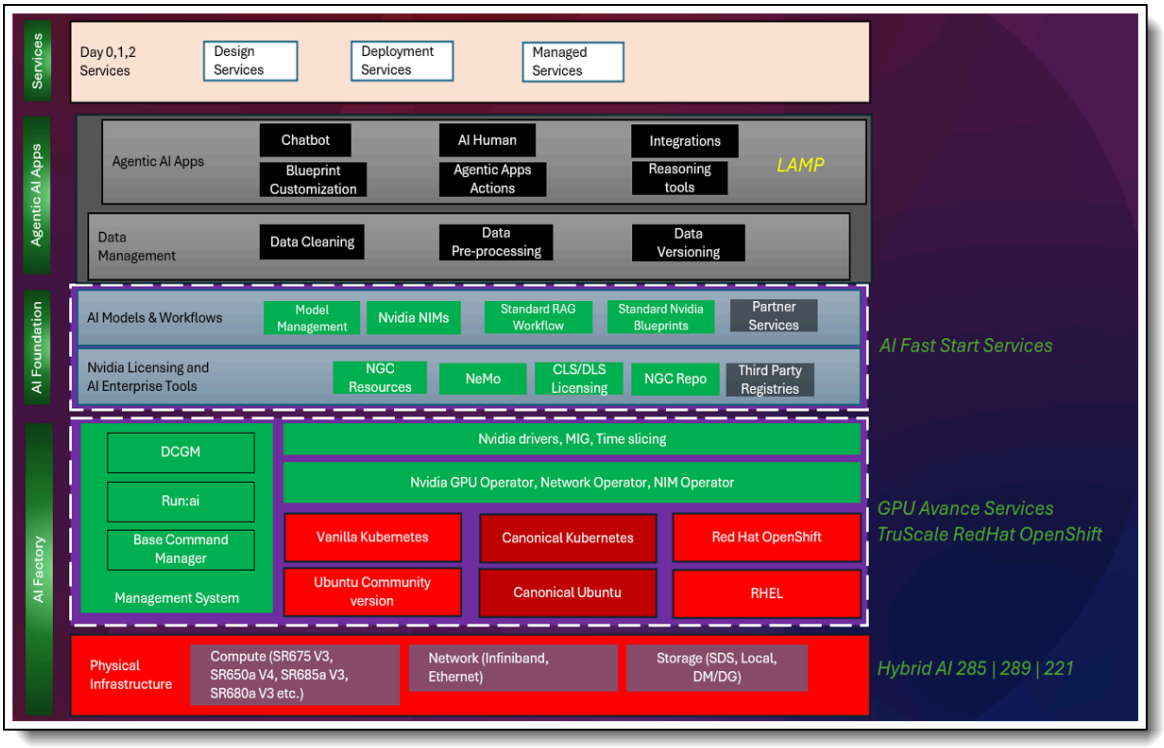

Figure 4. Service Offerings

Customer can choose from these three modular 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 optimized infrastructure at the outset build a solid base that is future-proofed to accommodate growth. The following table lists the Service and support offerings.

Table 5. Plan & Design with Kubernetes Services

Service	Multi-Node
5MS7C33028	GPU Advanced Services - Plan & Design with Kubernetes

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.

Table 6. Configuration with Kubernetes Services

Service	Multi-Node
5MS7C33029	GPU Advanced Services - Configuration with Kubernetes

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. The following table lists the Service and support offerings.

Table 7. Managed Services with Kubernetes

Service	Multi-Node
5MS7C33030	GPU Advanced Services - Managed Services with Kubernetes

Licensing

Lenovo XClarity One is a for-fee cloud or on-premises application. There is a free trial that allows management of 50 devices for up to 30 days, after which a license must be purchased. The following types of licenses are supported:

- Managed device, per endpoint licenses
  - Required for every managed device to use basic monitoring and management in the XClarity One portal. The license is determined based on the total number of managed devices in the organization
- Premium Licenses

Memory Predictive Failure Analytics License: Monitor and analyze memory errors and failure predictions to ensure that your devices are operating at peak performance.

The following table lists XClarity One offerings.

Table 8. XClarity One

Part Number	Description
7S0X000LWW	XClarity One - Managed Device, Per Endpoint w/1 Yr SW S&S
7S0X000MWW	XClarity One - Managed Device, Per Endpoint w/3 Yr SW S&S
7S0X000NWW	XClarity One - Managed Device, Per Endpoint w/5 Yr SW S&S

Each NVIDIA H100 (PCIe or NVL) and H200 NVL GPU includes a five-year NVIDIA AI Enterprise Subscription. NVAIE Licensing provides fully secured and tested NVAIE suite of software, as opposed to the open-sourced versions which are not directly supported. The licensing is done on a per-GPU basis. All NVIDIA AI Enterprise subscriptions include NVIDIA Business Standard support, which includes the following:

- **Technical Support** – 24/7 availability for case filing, 8am-5pm local time support coverage
- **Maintenance** – Access to maintenance releases, defect resolutions, and security patches
- **Direct Support** – Access to NVIDIA Support engineering for timely resolution of issues
- **Knowledgebase access, web support, email support, phone support**

To receive an NVIDIA AI Enterprise license for GPU(s) that do not include it, please contact your Lenovo sales representative. Lenovo provides both licensing and services for NVIDIA AI Enterprise.

The following table lists NVIDIA AI Enterprise offerings.

Table 9. NVIDIA AI Enterprise License

Part Number	Description
7S02001FWW	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 1 Year
7S02001GWW	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 3 Year
7S02001HWW	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 5 Year
7S02001MWW	24X7 Support Services for NVIDIA Enterprise AI per GPU Socket, 1 Year
7S02001NWW	24X7 Support Services for NVIDIA Enterprise AI per GPU Socket, 3 Years
7S02001PWW	24X7 Support Services for NVIDIA Enterprise AI per GPU Socket, 5 Years

The following table lists Run:ai subscription offerings.

Table 10. Run:ai

Part Number	Description
7S02004UWW	NVIDIA Run:ai Subscription per GPU 1 Year
7S02004XWW	NVIDIA Run:ai Subscription per GPU 3 Years
7S020050WW	NVIDIA Run:ai Subscription per GPU 5 Years

Ubuntu is an open-source Linux distribution. Version 22.04 is the LTS (Long Term Support) version of Ubuntu that will be maintained until April 2027. Ubuntu Pro can be optionally purchased to include additional weekday or 24/7 support, expanded security maintenance, kernel live patch, among other features.

The following table lists Canonical Ubuntu.

Table 11. Canonical Ubuntu

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

Sizing Guide for NIM LLM

This section aims to recommend the number of pods to deploy and/or GPUs to consume in order to achieve a certain user concurrency and throughput for LLM inference. Keep in mind that the data presented in this section is meant to be an example, and your exact results will heavily depend on the hardware, configuration, and NIM model used.

This guide assumes that an enterprise would aim to achieve a time-to-first-token of less than 1 second, as latency beyond that mark will be unappealing to users.

Concurrent user queries and tokens per second throughput both scale linearly with the number of pods allocated (1 GPU per pod). Based on the data for Meta-Llama 8B Instruct running summarization queries on 1 H100 NVL, the ratio is approximately 350 users per pod. In other words, if you know you need to account for *x* concurrent user queries at any given time, you should allocate the rounded-up value of pods. The scaling value of 350 is likely to be different depending on the model you are using; however, the linear scaling property is expected to hold across various models if the model can fit on the memory of 1 GPU. Running Gen-AI Perf with the specific model you would like to use will yield the number of concurrent users it can support per GPU, and then the scaling rule of thumb can be applied to approximate the desired number of nodes. Please see some additional data here to view results for other NVIDIA GPUs and NIM LLMs.

One data point to note is that decreasing model precision, for example from bf16 to fp8, does not substantially increase the number of concurrent users, but does slightly increase the throughput. Whether or not the tradeoff is worthwhile depends on the use case. However, if decreasing model precision allows the model to fit in the memory of one GPU, that is likely to lead to a much more significant increase in concurrent user support.

Lenovo LLM Sizing Guide

Although the results above are derived using a specific 8B parameter model, it is likely that AI workloads on Lenovo platforms will involve models of various types and sizes. The size of a model will affect how it can be loaded into the GPUs memory and therefore can change results such as throughput and concurrency. Lenovo has published an LLM Sizing Guide to outline how to size the LLM relative to the GPU, and provide a rule of thumb for the computational requirements of running an LLM.

For a higher-level method of calculating memory requirements automatically and simulating the inferencing process, the site ApX has a useful calculator.

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 provides customers and partners a means 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 are a few examples of 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 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.

Bill of Materials

The following table lists Bill of Materials.

Table 12. Licenses Bill of Materials

Part Number	Description
7S0X000NWW	XClarity One - Managed Device, Per Endpoint w/5 Yr SW S&S
7S1B000YWW	Canonical Ubuntu Pro 5Yr w/Canonical weekday Support
7S02001HWW	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 5 Year
7S020059WW	NVIDIA Run:ai Subscription per GPU 5 Years

Note – Ubuntu Pro and NVIDIA AI Enterprise provide support. The software stack can be deployed without support, however it is recommended to ensure consistent and stable deployment.

Seller training courses

The following sales training courses are offered for employees and partners (login required). Courses are listed in date order.

1. **Partner Technical Webinar - Lenovo AI Hybrid Factory offerings**
   2025-10-13 | 50 minutes | Employees and Partners

   In this 50-minute replay, Pierce Beary, Lenovo Senior AI Solution Manager, review the Lenovo AI Factory offerings for the data center. Pierce showed how Lenovo is simplifying the AI compute needs for the data center with the AI 281, AI 285 and AI 289 platforms based on the NVIDIA AI Enterprise Reference Architecture.

   Tags: Artificial Intelligence (AI)

   Published: 2025-10-13
   Length: 50 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: OCT1025

2. **Lenovo VTT Cloud Architecture: Empowering AI Innovation with NVIDIA RTX Pro 6000 and Lenovo Hybrid AI Services**
   2025-09-18 | 68 minutes | Employees Only

   Join Dinesh Tripathi, Lenovo Technical Team Lead for GenAI and Jose Carlos Huescas, Lenovo HPC & AI Product Manager for an in-depth, interactive technical webinar. This session will explore how to effectively position the NVIDIA RTX PRO 6000 Blackwell Server Edition in AI and visualization workflows, with a focus on real-world applications and customer value.
   We'll cover:
   - NVIDIA RTX PRO 6000 Blackwell Overview: Key specs, performance benchmarks, and use cases in AI, rendering, and simulation.
   - Positioning Strategy: How to align NVIDIA RTX PRO 6000 with customer needs across industries like healthcare, manufacturing, and media.
   - Lenovo Hybrid AI 285 Services: Dive into Lenovo's Hybrid AI 285 architecture and learn how it supports scalable AI deployments from edge to cloud.
   Whether you're enabling AI solutions or guiding customers through infrastructure decisions, this session will equip you with the insights and tools to drive impactful conversations.

   Tags: Industry solutions, SMB, Services, Technical Sales, Technology solutions

   Published: 2025-09-18
   Length: 68 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVCLD227

3. **VTT AI: Introducing The Lenovo Hybrid AI 285 Platform with Cisco Networking**
   2025-08-26 | 54 minutes | Employees Only

   Please view this session as Pierce Beary, Sr. AI Solution Manager, ISG ESMB Segment and AI explains:

   - Value propositions for the Hybrid AI 285 platform
   - Updates for the Hybrid AI 285 platform
   - Leveraging Cisco networking with the 285 platform
   - Future plans for the 285 platform

   Tags: Artificial Intelligence (AI), Technical Sales, ThinkSystem

   Published: 2025-08-26
   Length: 54 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVAI220

4. **VTT AI: Introducing the Lenovo Hybrid AI 285 Platform April 2025**
   2025-04-30 | 60 minutes | Employees Only

   The Lenovo Hybrid AI 285 Platform enables enterprises of all sizes to quickly deploy AI infrastructures supporting use cases as either new greenfield environments or as an extension to current infrastructures. The 285 Platform enables the use of the NVIDIA AI Enterprise software stack. The AI Hybrid 285 platform is the perfect foundation supporting Lenovo Validated Designs.
   • Technical overview of the Hybrid AI 285 platform
   • AI Hybrid platforms as infrastructure frameworks for LVDs addressing data center-based AI solutions.
   • Accelerate AI adoption and reduce deployment risks

   Tags: Artificial Intelligence (AI), Nvidia, Technical Sales, Lenovo Hybrid AI 285

   Published: 2025-04-30
   Length: 60 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVAI215

Related publications and links

For more information, see these resources:

- Lenovo Hybrid AI 285 Platform Guide:
  https://lenovopress.lenovo.com/LP2181
- Lenovo Hybrid AI 289 Platform Guide:
  https://lenovopress.lenovo.com/LP2286
- 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
- Hybrid AI Factory
- Lenovo XClarity

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, LP2311, was created or updated on October 14, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2311
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2311.

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 is a trademark 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.

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