Lenovo Hybrid AI Factory with SUSE

Planning / Implementation

Enterprises often face challenges when moving AI projects from experimentation into production. Proofs of concept may be developed quickly, but the leap to secure, governed, mission-critical deployments often stall, slowed by fragmented tooling, inconsistent environments, and the compliance demands of regulations like the EU AI Act. A flexible, secure platform is only half the equation; enterprises also need the world's most advanced AI compute and production-ready AI software frameworks working together natively, without ever breaking the safety net.

The Lenovo Hybrid AI Factory with SUSE provides an integrated infrastructure and software foundation for deploying and managing enterprise AI workloads. The solution combines Lenovo ThinkSystem infrastructure with SUSE AI Factory and components form NVIDIA AI Enterprise, which provides a unified software stack that is designed to seamlessly bridge the gap between local development and scalable enterprise production, operating as a turnkey digital factory. By embedding the proven capabilities of NVIDIA AI Enterprise software into the rigorous governance of SUSE AI, running on Lenovo ThinkSystem servers gives organizations a single, sovereign foundation for running today's and tomorrow's AI innovation without compromising control over their data, models, or infrastructure.

Why a Private AI Platform?

Public cloud AI services can provide convenient access to infrastructure and AI capabilities, but for regulated industries, sovereign workloads, and any organization whose competitive advantage depends on proprietary data, the economics and control trade-offs of a purely cloud-based approach become difficult to justify at scale. A private AI platform — one that an organization can run on its own infrastructure, in its own private cloud, or at the edge — addresses four distinct but related concerns.

- **Data Security & Sovereignty**

  Deploying AI workloads on-premises or within a private cloud keeps sensitive data under the organization's direct control. Rather than sending proprietary data, customer records, or regulated information across a third-party boundary, a private AI platform keeps data residency, privacy, and compliance requirements squarely within the enterprise's own perimeter — a foundational requirement for organizations operating under data sovereignty mandates or handling data that legally cannot leave a given jurisdiction.

  AI sovereignty extends beyond data location. It includes control over the infrastructure, software stack, models, deployment processes, and ongoing operations supporting the AI lifecycle. Appropriate security architecture, access controls, encryption, monitoring, and operational procedures remain necessary regardless of where the platform is deployed. SUSE similarly positions private enterprise AI around maintaining control over proprietary data and intellectual property while supporting datacenter, cloud, and edge deployment models. Lenovo complements this vision through its Hybrid AI Factory approach, providing the infrastructure, software ecosystem, and validated architectures needed to deploy and scale AI workloads while maintaining control over data, models, and infrastructure.

- **Compliance & Governance**

  Regulated industries like financial services, healthcare, government, and increasingly any enterprise subject to frameworks like the EU AI Act — need AI pipelines that are transparent and controllable. A private platform makes it possible to document, audit, and prove exactly how a model was trained, what data it touched, and how it is governed in production, satisfying both internal risk teams and external regulators without slowing down innovation.

  Private deployment does not, by itself, ensure regulatory compliance. Compliance depends on the AI system's purpose, risk classification, data handling, configuration, and associated technical and organizational controls. For example, the EU AI Act specifies responsibilities for deployers of high-risk AI systems that include human oversight, operational monitoring, and appropriate use of input data.

- **Predictable Cost at Scale**

  The first wave of Generative AI adoption prioritized rapid innovation, leveraging readily available cloud infrastructure to accelerate experimentation and proof-of-concept development, often ahead of cost optimization. As organizations transition to Agentic AI and large-scale LLM training and inference, focus is shifting toward Token Economics and Total Cost of Ownership (TCO). In this new phase, Lenovo and SUSE are committed to helping customers maximize business value by delivering optimized, cost-efficient AI infrastructure and operations. For further details, refer to the accompanying paper on Token Economics and AI TCO.

- **Intellectual Property Protection**

  Models, fine-tuning data, and the outputs generated by an organization's AI systems are increasingly a core piece of its intellectual property and competitive differentiation. Hosting these resources within infrastructure controlled by the organization can reduce reliance on external service boundaries and provide greater control over access, retention, and data-processing policies. Protecting intellectual property still requires identity and access management, encryption, network security, monitoring, data-governance policies, and secure lifecycle practices. A private AI platform provides a foundation for implementing these controls, but it does not replace them.

The Lenovo Hybrid AI Factory difference with SUSE: Observability, Security, and Agents

What differentiates SUSE AI Factory from a do-it-yourself collection of open-source AI tooling is that observability, security, and extensibility are built into the platform from the ground up, rather than bolted on after the fact.

Lenovo also differentiates by providing a validated Hybrid AI Factory design that combines enterprise-grade infrastructure, integrated AI software, and proven deployment architectures to accelerate time-to-value, reduce deployment risk, and deliver a scalable and resilient foundation for enterprise AI.

Four capabilities define this difference:

- **Observe**

  SUSE Observability gives operations teams real-time insight into AI workloads, helping them optimize costs and quickly remediate issues before they become production incidents. Centralized telemetry and operational context can help teams monitor system behavior, investigate performance issues, assess resource utilization, and identify conditions that require remediation.

- **Secure**

  SUSE AI Factory provides a foundation for applying security controls across the AI application lifecycle. It integrates with the zero-trust framework to bring industry-tested responsible-AI controls into the platform rather than requiring every customer to build them from scratch.

- **Build**

  An expanded AI library brings a growing catalog of new, validated, and curated open-source components directly into the platform, so teams are choosing from pre-vetted building blocks rather than sourcing and hardening components themselves. SUSE AI Factory with NVIDIA opens a new realm by making NIVDIA AI Enterprise components available to customers at their fingertips. A new Green Doc reference guide is also available to help teams jumpstart the development of agentic workflows — the next frontier of enterprise AI use cases.

- **Unified Foundation**

  Underpinning Observe, Secure, and Build is a Unified foundation that gives organizations control of their infrastructure, ensures smooth day-to-day operations, and maintains security and compliance — all while letting teams adopt new AI technologies at their own pace, rather than being forced into a single vendor's roadmap or release cadence.

All in all, SUSE AI Factory is a turnkey, sovereign AI platform that enables organizations to deploy, scale, and govern enterprise AI workloads using validated blueprints, integrated infrastructure, and a unified operational model, reducing deployment complexity and accelerating time to value. Capable of delivering virtualized infrastructure and bare metal performance with the addition of Apps and Blueprints from both SUSE and NVIDIA.

Design Overview

The Lenovo Hybrid AI Factory with SUSE separates the functions that manage the environment from the resources that run AI workloads and store AI data. This document describes these as three logical planes: the control plane, the compute plane, and the data plane. The same separation applies at every deployment level, from a single node to a multi-node cluster.

- **AI Control Plane**

  The control plane is the management layer. It decides what runs, where it runs, and under which policies, but it does not run application workloads or handle inference traffic itself. In this solution, SUSE Rancher Prime serves as the control plane, supported by services such as SUSE Observability and the AI infrastructure services. From the control plane, administrators define blueprints, enforce security and compliance policies, manage cluster lifecycles, and monitor every environment the organization runs. Operators interact with the control plane whether they use the Rancher UI for rapid prototyping or GitOps pipelines for deployment at scale. A single control plane can govern many distributed environments, so one team with one set of policies can manage AI infrastructure from the core data center to the edge.

- **AI Compute Plane**

  The compute plane is where AI workloads run. It consists of GPU-accelerated Kubernetes clusters (RKE2, or K3s for lighter-footprint deployments) running on SUSE Linux Enterprise Server, together with NVIDIA AI Enterprise software such as NIM inference microservices. Depending on the deployment level, workloads run in SLES 16 virtual machines with vGPU access, in SUSE Virtual Clusters, or directly on bare-metal nodes. GPUs can be partitioned with NVIDIA MIG to share capacity across multiple workloads or dedicated to a single workload for maximum performance. Compute clusters can be distributed across data centers, public cloud regions, and edge sites, each physically separate but governed by the same control plane.

- **AI Data Plane**

  The data plane provides the storage for AI data, including datasets, model files, vector databases, and other artifacts. In smaller deployments, each node uses SUSE Storage along with local NVMe storage. In larger deployments, Lenovo ThinkSystem DM/DG Storage provides a centralized, shared data plane, so that multiple workloads and teams can access the same datasets and models consistently and with high performance.

Scaling without Re-Architecting

Building on the Lenovo Hybrid AI Factory reference architecture for single-node inference, this document puts forward four deployment levels, from a single GPU-enabled server to a multi-node environment with shared storage and bare-metal compute. Every level uses the same core software stack and keeps the control plane separate from the compute and data planes, so organizations can start small and expand capacity, resilience, and performance without re-architecting. The following sections describe each level and the workloads it is best suited for.

- Level One: Single-Node Starting Point
- Level Two: Multi-Node Scale-Out
- Level Three: Shared Storage at Scale
- Level Four: Bare-Metal Compute with Optional Virtualization
- GPU Selection

Level One: Single-Node Starting Point

The Level One topology runs on a single Lenovo server equipped with one or more GPUs. Even at this scale, the architecture maintains the same separation of control and compute planes. SUSE Observability, Rancher Prime, and AI infrastructure services run as VMs alongside AI workloads running in SLES 16 guest VMs or containerized workloads in SUSE Virtual Clusters, all with direct GPU access; all coordinated through SUSE Virtualization with the NVIDIA GPU Operator over NVIDIA MIG GPUs. This gives teams a genuine, production-architected starting point and not just a simplified sandbox, that can grow directly into the next level without re-architecting.

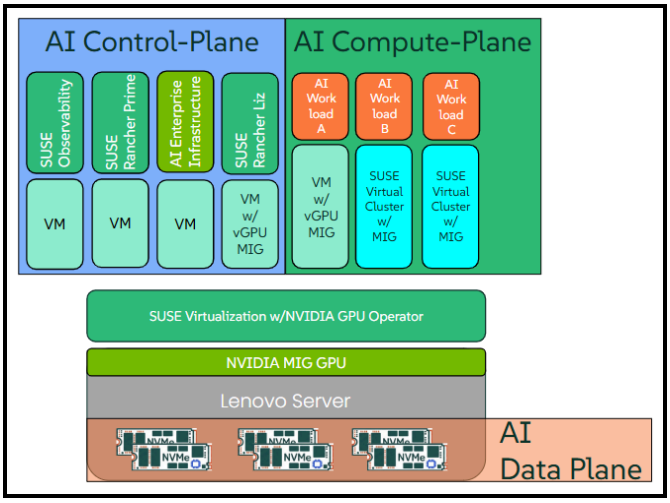


Figure 1. Single-Node starting Point with one or more GPUs

Level Two: Multi-Node Scale-Out

The Level Two topology extends the architecture across three or more nodes, with one or more of those nodes dedicated to GPU-accelerated compute. Control plane services and compute workloads are distributed across dedicated Lenovo servers, each still built on NVMe-backed local storage, giving organizations horizontal scale for both control-plane resilience and AI compute capacity as workload demand grows.

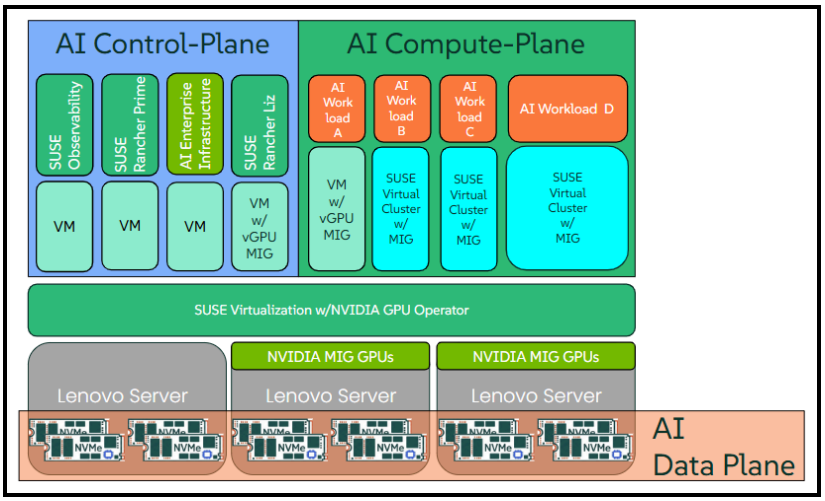


Figure 2. Multi-Node scale-out with one or more GPU nodes

Level Three: Shared Storage at Scale

The Level Three topology carries forward the same three-or-more-node, one-or-more-GPU-node topology as the Level Two topology but introduces Lenovo ThinkSystem DM/DG Storage as a shared AI Data Plane across the entire cluster. Rather than each node relying solely on local NVMe storage, the ThinkSystem DM/DG Storage series provides centralized, enterprise-grade shared storage — a critical requirement once multiple AI workloads and teams need consistent, high-performance access to shared datasets and model artifacts.

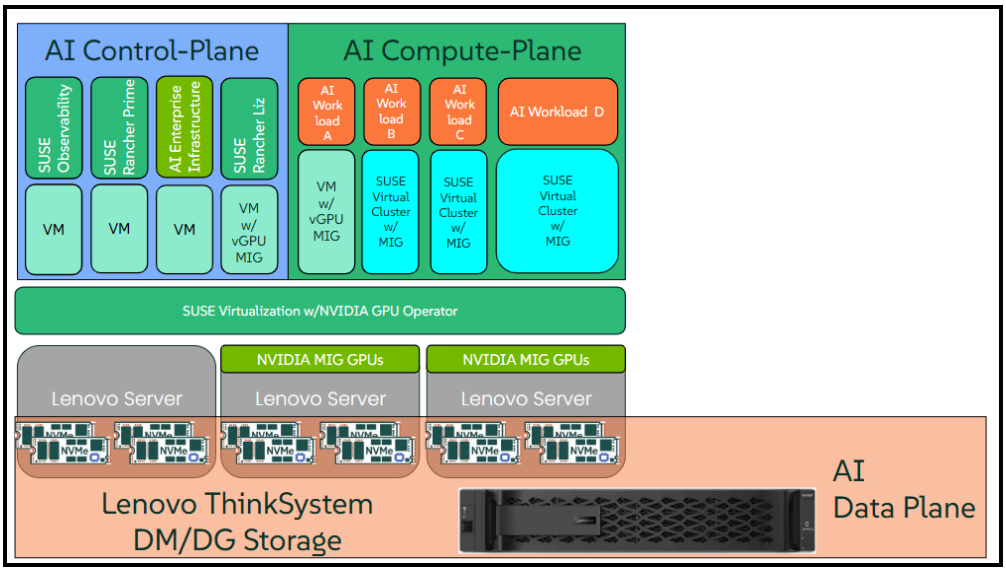


Figure 3. Level Three: Shared Lenovo ThinkSystem DM/DG Storage at Scale

Level Four: Bare-Metal Compute with Optional Virtualization

The Level Four topology represents the most scaled and flexible topology. The AI Control Plane is distributed across multiple SUSE Virtualization nodes running redundant instances of SUSE Rancher Prime and AI infrastructure services for resilience. The AI Compute Plane can now be extended to run directly on bare metal — SLES 16 with the NVIDIA host GPU driver and RKE2 running SUSE AI workloads natively — with SUSE Virtualization available as an optional layer rather than a requirement. This gives organizations the option to efficiently run GPU-intensive workloads where every bit of performance matters, while still retaining it where flexibility is preferred. As with the Level Three topology, Lenovo ThinkSystem DM/DG Storage provides the shared AI Data Plane underpinning the entire environment.

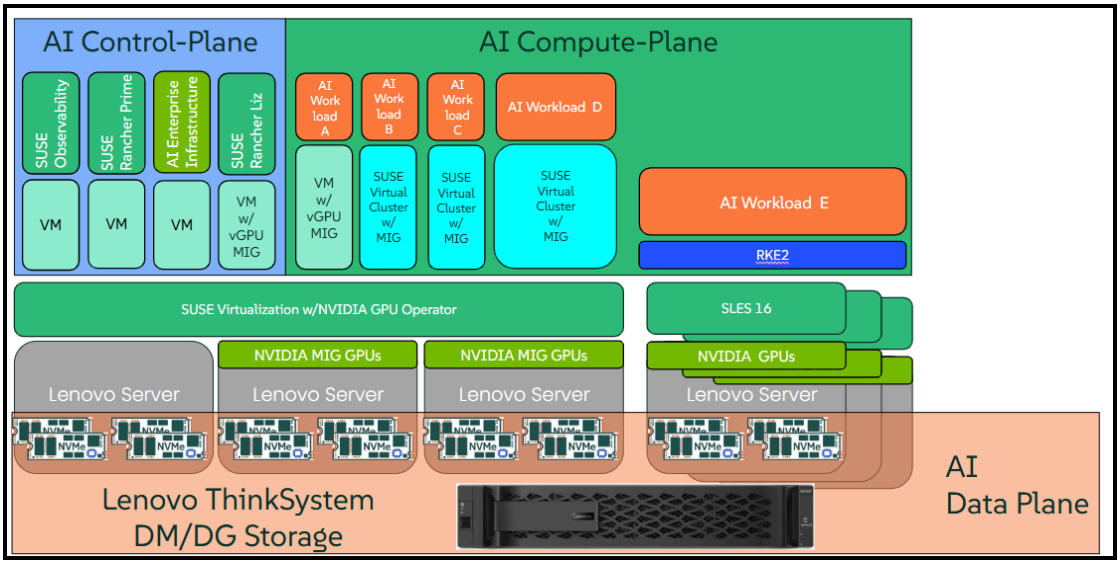


Figure 4. Bare-Metal Compute with Optional Virtualization

GPU Selection

For the AI workload nodes, one or more GPUs with sufficient VRAM is highly recommended. The required number of GPUs and the VRAM capacity per GPU depend mainly on the size of the AI model that your workload uses. Furthermore, for high availability (HA) deployments, we recommend all the workload nodes have the same GPU configuration so that pods can be rescheduled when necessary.

The ThinkSystem SR650a V4 accommodates up to four front-mounted double-wide GPUs, including NVIDIA H100 NVL. Thanks to the slot layout, GPUs can be paired and linked with high-speed NV Link bridges, making the server well suited to demanding AI training and inference workloads. Because of the slot layout, only passively cooled GPUs are supported in this configuration.

The SR650a V4 also offers a GPU placeholder option. Selecting it produces a "GPU-ready" configuration: the server ships with all the components required for GPU installation, such as GPU power cables, air ducts, power supplies, and fans, but without the GPUs themselves.

Lenovo Hybrid AI Factory with SUSE – Building Blocks

Built on Lenovo infrastructure, NVIDIA accelerated computing, and SUSE AI Factory, the Lenovo Hybrid AI Factory delivers a production-ready AI platform with these building blocks:

- Lenovo ThinkSystem SR650a V4
- SUSE AI Factory
- SUSE AI Factory with NVIDIA

Lenovo ThinkSystem SR650a V4

The Lenovo ThinkSystem SR650a V4 is an ideal 2-socket 2U rack server for customers want to maximize GPU compute power while still retaining the traditional 2U rack form factor. Combining performance and flexibility, the SR650a V4 server is a great choice for enterprises of all sizes.



Figure 1. Lenovo ThinkSystem SR650a V4

The server offers a broad selection of drive and slot configurations and offers numerous high-performance features. Outstanding reliability, availability, and serviceability (RAS) and high-efficiency design can improve your business environment and can help save operational costs.

SUSE AI Factory

SUSE AI Factory is a Kubernetes-standard platform that forms the application delivery layer of the SUSE AI stack, giving IT and platform teams the ability to discover, deploy, and manage AI applications and complex model-serving stacks. Built as a Kubernetes operator with a Rancher UI extension, SUSE AI Factory provides the higher-level services required to operationalize generative AI, acting as a centralized catalog and deployment pipeline for AI workloads. With pre-validated blueprints bundling vector databases, LLM inference servers (such as Ollama and vLLM), and user interfaces, organizations can consistently deploy production-ready AI stacks — including Retrieval-Augmented Generation (RAG) and MLOps use cases — on infrastructure they already manage.

SUSE AI Factory includes a streamlined, UI-driven interface, a built-in application catalog that works without live chart repository access, and support for air-gapped and sovereign deployment environments, simplifying the integration, governance, and lifecycle management of AI applications with enterprise-grade auditability. This approach replaces manual, one-off "snowflake" deployments with immutable, version-controlled blueprints that can be shared across every SUSE AI Factory cluster, giving platform teams a repeatable path from a validated template to a running AI stack.

SUSE AI Factory's blueprint model provides consistent, traceable deployments — with full source provenance and version control — across self-hosted open-source components as well as blueprints layered with NVIDIA AI Enterprise technologies, letting teams standardize how AI is delivered from sandbox to production without re-platforming for each new use case.

Table 1. Business value of SUSE AI Factory

Capability	Value
Kubernetes operator and Rancher Prime UI extension	Provides a common interface for managing AI applications, blueprints, and workloads
Pre-validated, version-controlled blueprints	Eliminates manual integration work and accelerates deployment
Standardized, reproducible pipeline	Consistent AI stacks across the entire enterprise
Built for scale on Kubernetes	Grows with workload demand using native Kubernetes scaling

The foundation of SUSE AI Factory is a validated base stack of SLES 16, RKE2 Kubernetes, and SUSE Rancher Prime that is deployed on bare metal or cloud nodes. SUSE Security, SUSE Observability, and SUSE Private Registry can be added for built-in security, full-stack monitoring, and secure AI artifacts

across AI workloads.

SUSE AI Factory is a Kubernetes operator and Rancher UI extension that sits on this foundation. It lets teams discover, install, and manage applications and blueprints — including a built-in catalog that works even in air-gapped environments. Teams can also avail immutable, version-controlled application stacks for specific use cases (e.g., RAG), deployable in one click, called Blueprints.

SUSE AI Factory with NVIDIA

SUSE AI Factory with NVIDIA lets teams install and manage NVIDIA applications and blueprints from the NVIDIA GPU Cloud (NGC) catalog directly within a Rancher-managed environment, turning what has traditionally been a manual, error-prone integration effort into a repeatable, pre-validated process. Rather than hand-assembling AI stacks from disparate components, teams get tightly integrated blueprints ready to deploy from day one.

This approach eliminates inconsistent, ad hoc environments that are difficult to secure, support, and scale, replacing them with a standardized deployment model enforced through centralized, Rancher-based controls. The result is full governance and auditability across every AI workload, regardless of where it runs, without slowing down the pace of innovation.

Platform Architecture: The Basic Schema of SUSE AI Factory

SUSE AI Factory is organized into three parts: an infrastructure foundation that provides the operating system, orchestration, and GPU support; the AI Factory layer where applications and blueprints run; and storage, security, and observability rails that span the entire stack. Each is described below.

- Infrastructure Foundation
- AI Factory: Individual Apps and Blueprints
- Storage, Security, and Observability Rails

Infrastructure Foundation

At the base of the stack sits the infrastructure itself — bare metal, SUSE Virtualization, or another hypervisor, or a cloud environment. On top of that runs SLES 16 with the NVIDIA GPU driver, giving every node in the environment a secure, hardened operating system with native GPU support already integrated. An RKE2 Kubernetes cluster provides the orchestration layer above SLES, and SUSE Rancher Prime — the Rancher Manager — sits above RKE2, providing the unified management plane for the entire environment. A GPU Operator layer sits directly beneath SUSE AI Factory itself, automating GPU discovery, driver management, and lifecycle operations for every GPU-enabled workload.

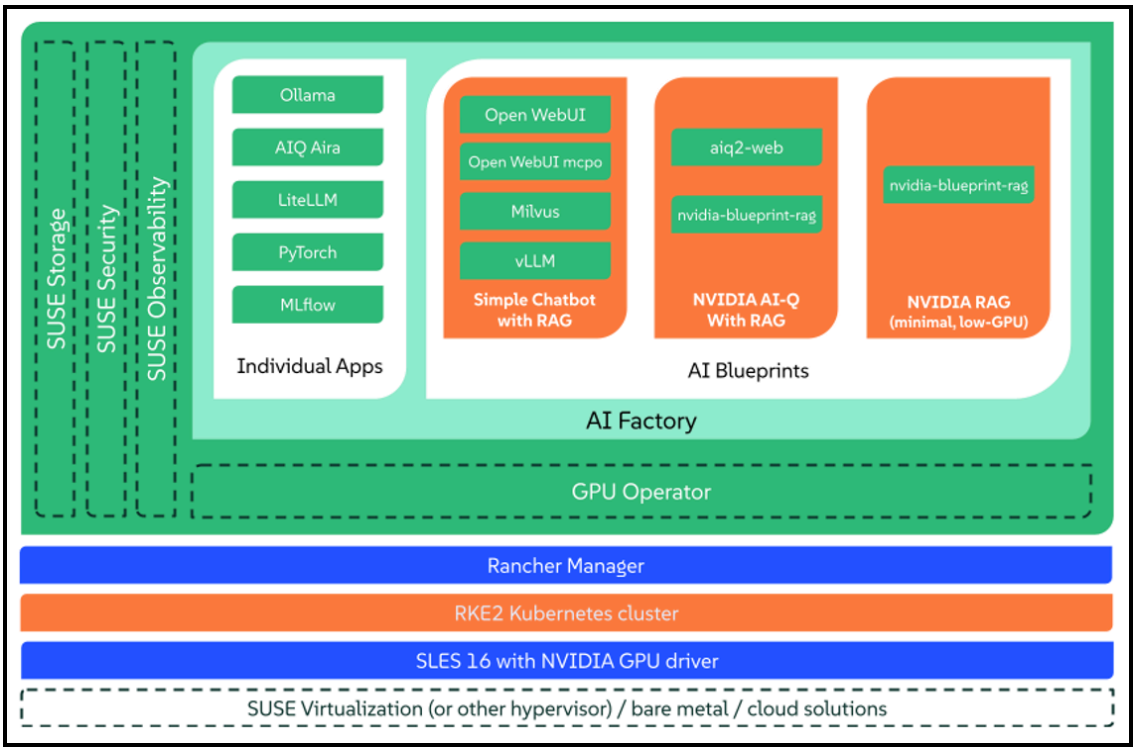


Figure 6. Basic Schema of SUSE AI Factory

AI Factory: Individual Apps and Blueprints

The AI Factory layer itself is where teams actually build and run AI workloads, and it's organized into two categories. Individual Apps are standalone, best-of-breed components — including Ollama for local model serving, AIQ Aira, LiteLLM as a unified LLM gateway, PyTorch for model development, and MLflow for experiment tracking and model lifecycle management — that teams can deploy independently to compose custom AI solutions.

AI Blueprints, by contrast, are pre-validated, opinionated architectures that combine multiple components into a ready-to-deploy solution for a specific use case. Three blueprints anchor the current library: a Simple Chatbot with RAG, built from Open WebUI, Milvus as the vector database, and vLLM for inference; NVIDIA AI-Q with RAG, which layers the aiq2-web interface and NVIDIA's RAG blueprint on top of Open WebUI's MCP-oriented tooling; and a minimal, low-GPU NVIDIA RAG blueprint for teams that want RAG capability without committing significant GPU capacity to it. In every case, the blueprint approach eliminates the guesswork of assembling compatible components from scratch, letting teams start from an already-integrated, tested foundation.

Storage, Security, and Observability Rails

Running alongside SUSE AI Factory — rather than beneath or above it — are four horizontal rails that apply consistently across every application and blueprint in the environment: SUSE Storage, SUSE Security, SUSE Private Registry and SUSE Observability. Because these rails span the full width of the stack, every workload deployed through SUSE AI Factory automatically inherits enterprise-grade storage, security, and observability, without each application team needing to integrate these capabilities individually.

Application Libraries: SUSE AI Library and NVIDIA AI Library

SUSE AI Factory ships with a curated, ready-to-deploy application catalog delivered directly through the same Rancher-based interface administrators already use to manage the rest of their Kubernetes

environment. Two libraries are available side by side.

- SUSE AI Library
- NVIDIA AI Library
- Kubernetes Layer
- SUSE AI Observability

SUSE AI Library

The SUSE AI Library currently includes over 80 curated applications — spanning categories like alerting (Alertmanager), workflow orchestration (Apache Airflow), and API management (Apache APISIX and its dashboard) — each packaged as a Helm chart and deployable directly from the catalog with a description, documentation link, and one-click install. This gives platform teams a trusted, pre-vetted alternative to sourcing, hardening, and integrating open-source components independently.

NVIDIA AI Library

Alongside the SUSE AI Library, the NVIDIA AI Library provides direct access to NVIDIA's own catalog of AI applications and blueprints — currently over 30 — including the AI-Q Research Assistant Blueprint (aiq-aira and aiq2-web), and NVIDIA's Morpheus-based cybersecurity workflows for digital fingerprinting (cybersecurity-dfp) and spear-phishing detection (cybersecurity-sp). Because both libraries live in the same catalog interface, teams can mix SUSE-curated and NVIDIA-native components freely when assembling a workload, rather than managing two separate toolchains.

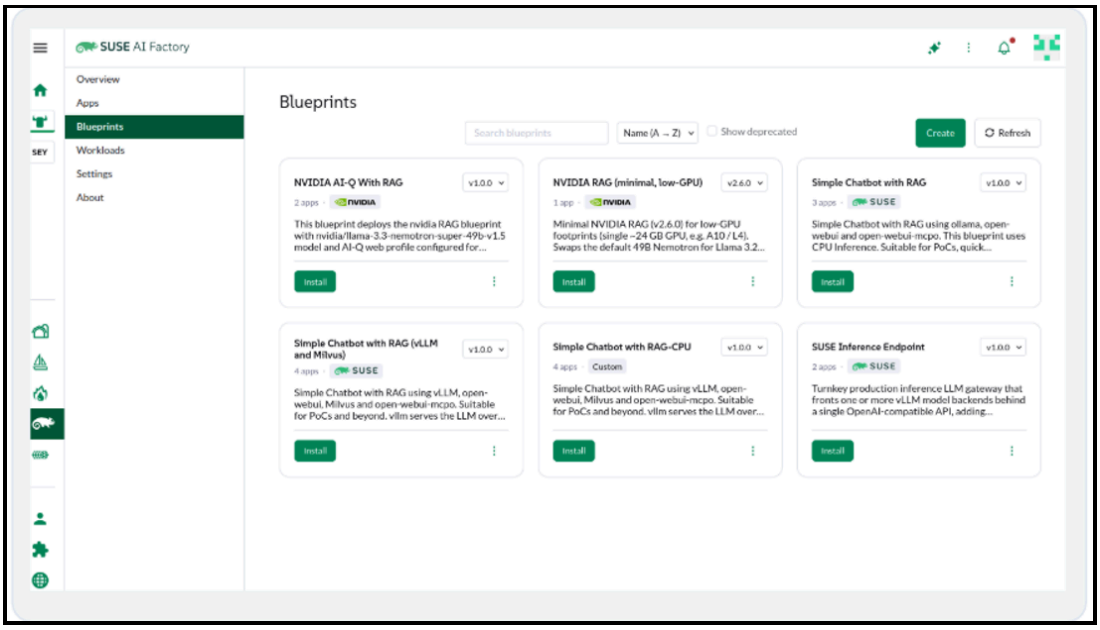


Figure 7. SUSE AI Factory user interface, showcasing the NVIDIA AI Blueprints

Kubernetes Layer

SUSE Rancher Prime, running on RKE2 (Rancher Kubernetes Engine 2), provides the enterprise-grade Kubernetes foundation for SUSE AI Factory, enabling organizations to deploy and operate AI workloads with consistency, security, and scale. As the Kubernetes layer, RKE2 delivers a CNCF-conformant, hardened distribution that simplifies cluster lifecycle management and supports production-ready deployment of AI services such as large language models (LLMs), NVIDIA NIM inference microservices, and retrieval-augmented generation (RAG) pipelines.

Tight integration with SUSE Linux Enterprise Server (SLES) as the underlying OS, together with NVIDIA GPU Operator, NIM Operator, and Network Operator, ensures efficient resource utilization, high availability, and automated lifecycle management for GPU-accelerated workloads, making it well suited for data-intensive AI training and low-latency inference across data center, cloud, and edge environments. For lighter-footprint or air-gapped deployments, K3s extends this same consistency down to edge clusters and developer workstations.

SUSE AI Observability

SUSE AI Observability extends SUSE Observability's proven 4T Data Model: Telemetry, Tracing, Topology, and Time; with purpose-built monitoring for the unique demands of AI workloads, giving teams a single, complete picture of both infrastructure and AI performance.

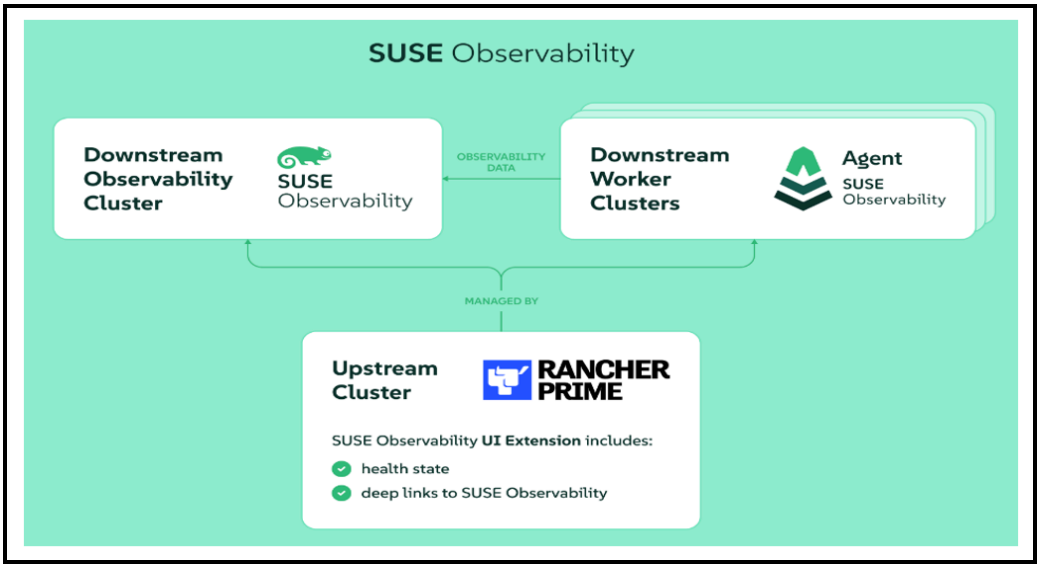


Figure 8. SUSE Observability architecture with Rancher Prime management

On top of this foundation, SUSE AI Observability adds deep monitoring across the AI stack:

- **AI Workloads** — visibility into the core compute processes driving your models
- **LLM Management** — prompt performance, token usage, and response quality metrics
- **Vector Databases** — health and performance monitoring for RAG and search infrastructure
- **Base AI Components** — Kubernetes orchestration and GPU utilization across the compute layer

Conclusion

The Lenovo Hybrid AI Factory with SUSE gives organizations a practical, production-ready path from AI experimentation to governed, enterprise-scale deployment. By combining Lenovo ThinkSystem SR650a V4 servers with SUSE Rancher Prime, SUSE AI Factory, and NVIDIA AI Enterprise software, the solution delivers a unified platform for deploying, managing, and observing AI workloads across data center, cloud, and edge environments.

The four-level topology lets organizations start with a single node and grow to multi-node, shared-storage, and bare-metal deployments without re-architecting. Pre-validated blueprints, curated SUSE and NVIDIA application libraries, and built-in security and observability reduce integration effort and operational risk, while keeping data, models, and infrastructure under the organization's control.

Whether the goal is a private enterprise assistant, retrieval-augmented generation, agentic workflows, or edge AI, the Lenovo Hybrid AI Factory with SUSE provides a sovereign, scalable foundation for delivering business value from AI today and adopting new AI technologies as they emerge.

Resources

For more information, see the following resources:

- Lenovo ThinkSystem SR650a V4 product guide
- Lenovo Hybrid AI Factory 221 reference architecture
- SUSE AI and SUSE AI Factory documentation
- SUSE AI Factory with NVIDIA
- Token Economics and AI TCO paper

SUSE AI Factory with NVIDIA - Use Cases

SUSE AI Factory with NVIDIA lets organizations run generative and agentic AI applications privately, on their own infrastructure, using pre-validated blueprints. The following are some of the most common use cases.

- Typical examples:
  - Private ChatGPT for enterprise
  - Knowledge assistants
  - Technical support copilots
  - Internal document search
  - Engineering knowledge bases
- SUSE AI Factory provides pre-validated blueprints that include:
  - Vector database
  - vLLM or Ollama inference
  - User interface
  - Deployment automation and lifecycle management
- NVIDIA AI-Q enables agentic workflows where AI agents can:
  - Search enterprise content
  - Reason across multiple sources
  - Use tools and workflows
  - Perform multi-step research tasks
- Agentic AI Examples include:
  - IT operations assistants
  - Customer service agents
  - Workflow automation
  - Multi-agent systems
- EDGE AI Examples:
  - Smart city analytics
  - Video analytics
  - Industrial AI
  - Telco edge applications
- Telco RAN/AI Examples:
  - Network operations copilots
  - RAN optimization
  - Telecom observability
  - AI-assisted troubleshooting
- Robotics Examples:
  - Autonomous systems
  - Manufacturing automation
  - Warehouse robotics
- Digital Twins Using NVIDIA Omniverse together with SUSE AI Factory for:
  - Factory simulation
  - Infrastructure modeling
  - Smart city simulation
  - Engineering digital twins

Bill of Materials

The following table lists the Bill of Materials of all the hardware used to implement Lenovo Hybrid AI Factory with SUSE.

Table 2. Bill of Materials

Part number	Product Description	Quantity
7DGDCTO2WW	**Server : ThinkSystem SR650a V4-3yr Base Warranty**	1
C3QN	ThinkSystem SR650a V4 2.5"/EDSFF 3.S Chassis	1
C3JB	ThinkSystem General Computing - Power Efficiency	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5R7	Intel Xeon 6714P 8C 165W 4.0GHz Processor	1
BPDR	ThinkSystem V4 2U Standard Heatsink	1
5977	Select Storage devices - no configured RAID required	1
C4U2	ThinkSystem SR650a V4 x16 Front Riser Slot 21	1
C4U4	ThinkSystem SR650a V4 x16 Front Riser Slot 17	1
C4U3	ThinkSystem SR650a V4 x16 Front Riser Slot 19	1
C4U1	ThinkSystem SR650a V4 x16 Front Riser Slot 23	1
C0U4	ThinkSystem 1300W 230V/115V Titanium CRPS Premium Hot-Swap Power Supply	2
C3RD	ThinkSystem 2U 6056 20K Performance Fan Module	5
C3UG	ThinkSystem Long Travel Toolless Slide Rail Kit V4 with 2U CMA	1
C5MU	ThinkSystem SR650a V4 Standard Left Rack Latch	1
BPKR	TPM 2.0	1
B7XZ	Disable IPMI-over-LAN	1
C3K9	XClarity Platinum Upgrade v3	1
C4S2	ThinkSystem SR650 V4 Processor Board	1
CB2P	SR650 V4 Laser service indicator	1
B0ML	Feature Enable TPM on MB	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	8
B8MT	ThinkSystem 2U MS Fan Dummy	1
C3RM	ThinkSystem 2U Air duct Filler for 1P	1
BEYJ	ThinkSystem MS Height CPU Dummy	1
AURS	Lenovo ThinkSystem Memory Dummy	15
BPP5	OCP3.0 Filler with screw	2
C3S1	ThinkSystem SR650a V4 2U F AC GPU Riser Cage	2
C3RN	ThinkSystem 2U Main Air Duct	1
C5MT	ThinkSystem SR650 V4 Main air duct mylar for Front AC GPU	1
C3RJ	ThinkSystem 2U 2LP Riser Cage Filler	2
C3RH	ThinkSystem 2U 3FH Riser Cage Filler	2
C3SG	ThinkSystem SR650a V4 2U F GPU 8x25 HDD Cage	1
C4S6	HV 2U V4 long Chassis L1 PKG BOM	1
C7Y8	ThinkSystem SR650 V4 System I/O Board	1
C26Y	ThinkSystem V4 CPU HS Clip	1

Part number	Product Description	Quantity
C3QU	ThinkSystem GPU Power Cable, PCIe16-PCIe16, 400mm	1
C3SQ	ThinkSystem SR650 V4 Agency label with ES&CE&UKCA	1
C3TB	ThinkSystem SR650a V4 model name Label	1
CF3J	ThinkSystem SR650a V4 URL QR Code Label	1
AWF9	ThinkSystem Response time Service Label LI	1
C3TH	ThinkSystem SR650 V4 Service Label for WW	1
B97B	XCC Label	1
C3T7	ThinkSystem BHS 2U Front GPU Slot 20-23	1
BQPS	ThinkSystem logo Label	1
C3T5	ThinkSystem BHS 2U Front GPU Slot 16-19	1
C20Q	ThinkSystem 1300W TT power rating label WW	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
CBDC	ENERGY STAR Certification Country	1
BE0E	N+N Redundancy With Over-Subscription	1
BK14	Low voltage (100V+)	1
CA4E	MI-RS19, PCIe7&8/ PWR1	1
CA4D	MI-RS21, PCIe3&4/ PWR3	1
CA4C	MI-RS23, PCIe1&2/ PWR4	1
CA4F	MI-RS17, PCIe5&6/ PWR2	1
	Auto-Derived Part Items	
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BP4X	ThinkSystem Double Width GPU-Ready Installation	1
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	1
7S0XCTO6WW	XClarity One	1
SCJD	XClarity One - Standard, Per Endpoint w/3 Yr SW S&S	1
7S0XCTO8WW	**XClarity Controller Prem-FOD**	1
SCY0	Lenovo XClarity XCC3 premier - FOD	1
7Q01CTSAWW	**SERVER KEEP YOUR DRIVE ADD-ON**	1
QAK3	SR650a V4	1
QAK6	KYD	1
QA0Y	Months	36
7Q01CTS4WW	**SERVER PREMIER 24X7 4HR RESP**	1
QA12	24x7 4hr Resp	1
QAK3	SR650a V4	1
QA18	Premier	1
QA0Y	Months	36
7Q01CTO2WW	**SERVER CO2 OFFSET**	1
QABM	CO2 Offset 10 Metric Tonnes	1
QABD	CO2 Offset	1

Authors

Ameaza Rodrigues: Ameaza Rodrigues is graduate from North Carolina State University with a Master's in Computer Science. She works as a Jr. Solution Architect for Lenovo X SUSE Alliance. In collaboration with the partner and ecosystem providers, her mission is to grow the business by building valuable solutions and delivering value to the customers.

Jorge Alberzoni is a well-established Senior Solutions Architect with a strong background in the telecommunications industry, specializing in mobile communication technologies. With a proven track record of successful global deployments across diverse platforms, he brings deep technical expertise and a practical understanding of complex network infrastructures. Based in Barcelona, Jorge is channeling his telecom experience into the AI space as part of the Lenovo ESMB Team. His knowledge of mobile networks, data traffic, and edge infrastructure makes him uniquely positioned to design and deliver advanced AI solutions tailored to enterprise and telecommunications clients. By leveraging AI, Jorge is helping organizations optimize network performance, enhance customer experiences, and unlock new capabilities at the edge.

Kevin Dean is the Senior Manager of the HPC/AI Solution Architect Team in the ISG Offerings Group at Lenovo. Kevin oversees the strategy and processes for HPC and AI solutions in this position, contributes his knowledge of HPC/AI application performance, and serves as the CAE/manufacturing architect. Kevin has over 20 years' experience in HPC, AI, and computational engineering, including nine years at Lenovo focused on HPC/AI performance and solution architecture, and 12 years in aerodynamic design and CFD for the U.S. defense and automotive racing sectors. He earned his Master's Degree in Aerospace Engineering at the University of Florida, following a Bachelor's in the same field from Virginia Polytechnic Institute and State University.

Mark Wallis is a Senior AI Solutions Architect at Lenovo. He is passionate about building and testing the latest AI solutions to meet customers' needs with a background in managing complex installations, performance testing and software automation. He is originally from the United Kingdom, but has been based in the USA for the last 25 years. When he is not working, he can be found spending time with his family or out and about, hiking and cycling.

Hapsara Sukasdadi is a seasoned IT and telecommunications industry expert. Serving as a Solutions architect, Hapsara currently drives Lenovo's AI and Telco technical engagements, focusing on architecting solutions in AI and Telco infrastructures. In this role, Hapsara collaborates closely with partners and ecosystem providers. Hapsara's primary mission is to deliver comprehensive solutions, encompassing design, planning, and integration across a spectrum of critical areas, including AI and Telecommunications infrastructure solutions.

Alex Arnoldy has nearly three decades of IT experience and brings a strong focus on enterprise technologies and cloud-native solutions. He is a Kubernetes specialist, and has earned both the Certified Kubernetes Administrator (CKA) and Certified Kubernetes Security Specialist (CKS) certifications. His expertise spans containerization, security, Edge, and Enterprise-scale infrastructure.

Related product families

Product families related to this document are the following:

- Hybrid AI Factory

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.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2548, was created or updated on October 6, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2548
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2548.

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:

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

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

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

NVIDIA®, NGC®, NVIDIA GPU Cloud®, NVIDIA Omniverse®, and NVIDIA RTX® are trademarks of NVIDIA Corporation.

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