Reference Architecture: Lenovo Hybrid AI Platform 201 for Enterprise RAG With Red Hat AI Enterprise

Version 1.0

Optimize enterprise AI economics with high-performance CPU-based inference

Accelerate enterprise AI adoption with a scalable and simplified deployment approach

Reduce complexity with a validated Red Hat OpenShift AI architecture

Simplify enterprise AI data management with a unified storage platform

Erik Robert
Abed Islam
Eric Page

LENOVO PRESS

Table of Contents

Introduction

Executive Summary

Enterprise adoption of generative AI is accelerating, but many organizations struggle to move beyond successful pilots into production-ready deployments. Production AI requires more than a model, it requires secure access to enterprise data, scalable infrastructure, operational monitoring, governance, and lifecycle management.

The Lenovo Hybrid AI Platform 201 addresses these challenges with a validated reference architecture for enterprise Retrieval-Augmented Generation (RAG) and private AI workloads. The solution combines Lenovo ThinkSystem servers, Intel Xeon 6 processors with Advanced Matrix Extensions (AMX), Lenovo DM3200F unified storage powered by ONTAP, Red Hat OpenShift Container Platform (RHOCP), and Red Hat OpenShift AI (RHOAI) to provide a scalable foundation for operational AI.

The Lenovo Hybrid AI Platform 201 follows a CPU-optimized architecture profile built around a 2 CPU, 0 GPU, and 1 network adapter per node design approach for enterprise AI inference and Retrieval-Augmented Generation (RAG) workloads. Built on Lenovo ThinkSystem infrastructure, Intel Xeon 6 processors with Advanced Matrix Extensions (AMX), Lenovo DM3200F unified storage powered by ONTAP, and Red Hat OpenShift AI, the platform provides a validated and scalable foundation for operational AI. By integrating compute, storage, and AI orchestration into a unified architecture, organizations can deploy private AI closer to enterprise data while reducing infrastructure complexity and maintaining security, governance, and operational control.

A key differentiator is its CPU-inference approach. Using Intel Xeon 6 processors and vLLM model serving, organizations can support many enterprise RAG and assistant workloads without dedicated GPU infrastructure. Intel benchmark testing demonstrated up to 1.88× higher concurrent-user capacity under tested conditions, while maintaining a 100 ms Time Per Output Token (TPOT) service-level objective.

The platform also provides a unified enterprise data foundation through ONTAP, enabling governed access to enterprise content across NFS, SMB, and S3 protocols. Combined with OpenShift AI's model serving, pipelines, monitoring, and governance capabilities, the solution delivers a repeatable framework for deploying and operating private enterprise AI.

This reference architecture is designed for organizations seeking a cost-effective, secure, and scalable foundation for enterprise RAG and private AI deployments in on-premises or hybrid environments.

Intended Audience

This solution is designed for both business and technical stakeholders responsible for AI strategy, infrastructure, and operations, including CIOs, CTOs, and Chief Data/AI Officers, line-of-business and digital transformation leaders driving AI adoption, as well as enterprise architects, AI/ML engineers, platform and DevOps teams, and IT operations. It also targets the partner ecosystem, including ISVs building RAG and GenAI applications, system integrators deploying AI solutions, and cloud and hybrid platform providers.

Challenges and Opportunities

Business Challenges

Generative AI pilots are often built using isolated environments, sample data, and limited user groups. Production environments are different. They must serve real users, connect to enterprise data sources, meet security and compliance expectations, and deliver predictable performance under changing load. RAG adds another layer of complexity because it combines data ingestion, embedding, vector search, retrieval, prompt construction, and model inference into one workflow.

Infrastructure teams must also decide how to size and place AI workloads. Some workloads require GPUs, especially large models, high-throughput batch processing, training, and latency-sensitive high-volume serving. Other workloads can run effectively on optimized CPU platforms, particularly when model size, concurrency, latency objectives, and cost constraints align. A practical AI infrastructure strategy should allow both approaches rather than assuming that every AI workload requires the same accelerator profile.

Data accessibility is another challenge. Enterprise knowledge is distributed across file shares, object stores, application repositories, and regulated data platforms. RAG requires this content to be discoverable, indexed, and retrievable while maintaining access controls. Without a unified data layer and repeatable ingestion process, RAG implementations can become fragile, duplicated, and difficult to govern.

Operational complexity is equally important. AI platforms need model serving, monitoring, pipeline automation, platform upgrades, access control, and resource management. These capabilities must be integrated with existing enterprise operations rather than operated as a standalone science project.

Strategic Opportunities

The rapid adoption of generative AI, combined with advancements in infrastructure and software platforms, creates an opportunity to simplify and scale enterprise AI deployments:

- **Simplified AI deployment** – A validated reference architecture reduces design complexity, accelerates deployment, and minimizes the effort required to build AI infrastructure from the ground up.
- **Cost-efficient AI inference** – Intel Xeon 6 processors with Advanced Matrix Extensions (AMX) enable efficient CPU-based AI inference, reducing dependence on GPUs and lowering infrastructure costs for suitable workloads.
- **Unified, high-performance data access** – ONTAP provides a common data layer across object and file storage protocols, enabling fast, secure access to enterprise data for RAG applications.
- **Scalable, modular foundation** – Built on Kubernetes and Red Hat OpenShift AI, the architecture supports incremental scaling as AI workloads, data volumes, and user demand increase.
- **On-premises AI with enterprise governance** – Organizations can deploy AI close to their data while maintaining control over security, compliance, and data governance requirements.

Solution Overview

The Lenovo Hybrid AI Platform 201 combines compute, storage, and AI orchestration into a platform designed for enterprise AI inference and Retrieval-Augmented Generation (RAG) workloads.

Solution Workflow

Enterprise data is ingested and stored on Lenovo storage running ONTAP, where it is securely managed and made accessible through S3, NFS, or SMB interfaces. This data is then processed through embedding and retrieval pipelines orchestrated by Red Hat OpenShift AI running on Lenovo compute. When a user query is received, the system retrieves relevant context from the data store and combines it with model inference to generate context-aware responses with low-latency interaction.

Table 1 – Key Differentiators

Capability	**Description**
CPU-based AI Inference Architecture	Designed around a 2 CPU, 0 GPU, and 1 network adapter per node architecture profile to support enterprise AI inference without dedicated GPU infrastructure for applicable workloads.
Unified Data Access for RAG	ONTAP provides a common data foundation with support for S3, NFS, and SMB to simplify retrieval and access across enterprise data sources.
Integrated Reference Architecture	Integrates compute, storage, networking, and AI software into a validated architecture intended to simplify deployment and reduce integration effort.
Modular and Scalable Design	Kubernetes-based orchestration enables organizations to scale infrastructure incrementally as AI demand grows.
Built for Private and Hybrid AI	Supports deployment closer to enterprise data while maintaining governance, security, and operational control.

By bringing together infrastructure and AI software into a cohesive architectural blueprint, this reference architecture helps simplify the design of RAG environments and provides a modular and scalable framework intended to support the transition from pilot projects toward operational AI deployments.

Data Layer Optimization for RAG Workloads

RAG applications depend on fast, reliable access to enterprise data across multiple formats and storage systems. In this solution, ONTAP serves as the data foundation, enabling low-latency retrieval and consistent performance for AI inference workloads.

Key capabilities include:

- **High-performance data access for RAG pipelines**
  Enables rapid retrieval of documents, embeddings, and vector data to support low-latency AI inference and responsive RAG workflows.
- **Unified data access across protocols**
  Supports S3, NFS, and SMB, allowing AI applications to seamlessly access structured and unstructured data without additional data movement.
- **Built-in data protection and governance**
  Provides encryption, ransomware protection, and compliance capabilities to help keep enterprise data secure and controlled throughout the AI lifecycle.
- Outcome: Responsive and reliable RAG workflows with enterprise-grade data control.

Efficient AI Inference with CPU Acceleration

This solution leverages Intel Xeon 6 processors with built-in AMX to deliver efficient, scalable AI inference on industry-standard CPU infrastructure. Key benefits include:

- **Cost-efficient AI inference**
  Supports high-performance AI workloads on CPU-based infrastructure, helping organizations optimize infrastructure investments and reduce deployment costs.
- **Efficient performance per watt**
  Maximizes resource utilization to deliver higher throughput and improved energy efficiency across AI workloads.
- **Simplified development and deployment**
  Integrates with popular AI frameworks and tools, enabling teams to accelerate adoption with minimal application changes.
- **Outcome:**
  A scalable and efficient AI inference platform that can lower total cost of ownership while simplifying enterprise AI deployment and operations.

Solution Components

Hardware and Software

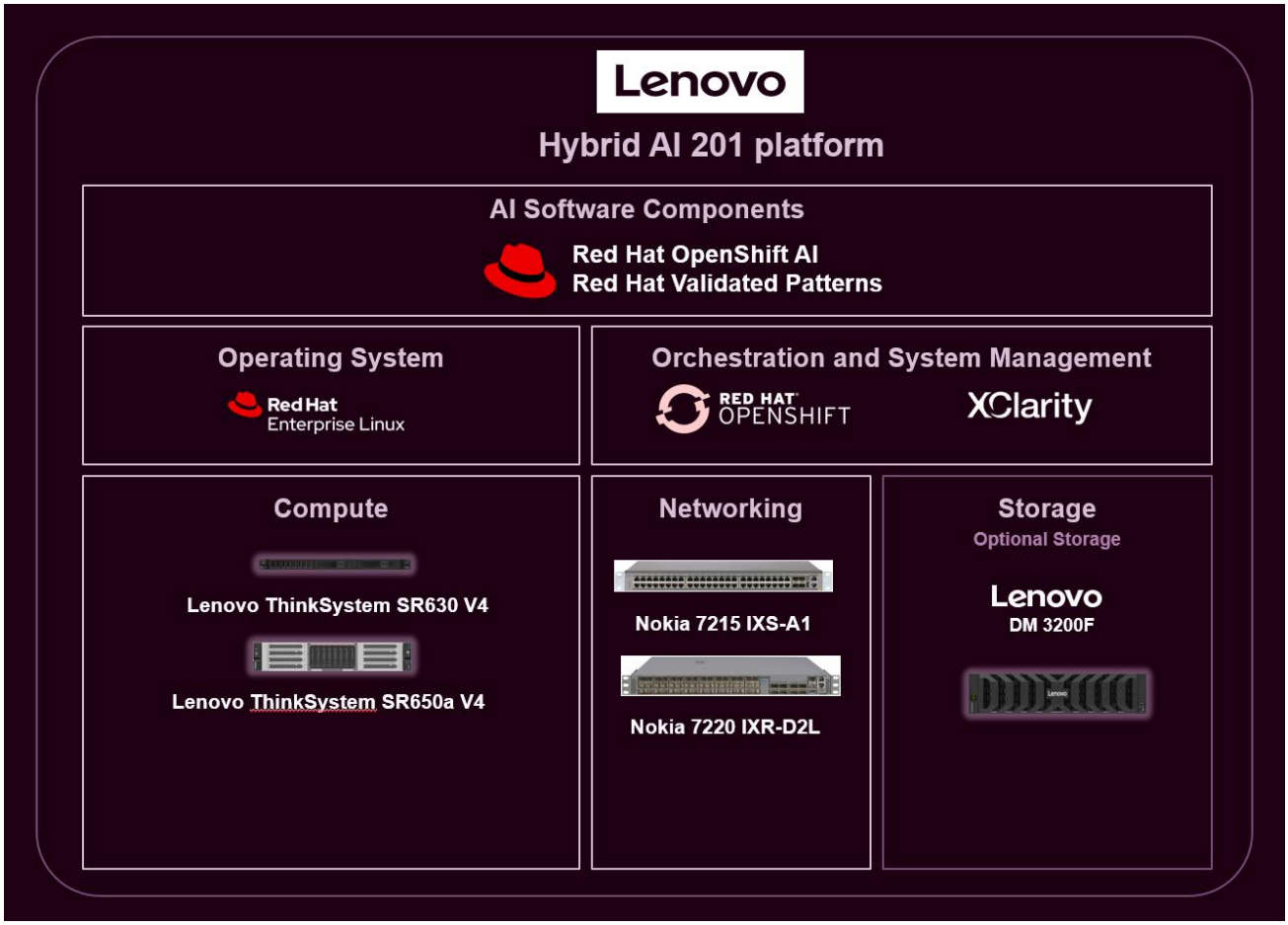

Figure 1 – Lenovo Hybrid AI Platform 201 solution components

Table 2 – Solution Components

Component	Layer	Description	Role in the Solution
Lenovo ThinkSystem SR630 V4 (Control Plane Node*)	Control & Platform Layer	Enterprise server managing Kubernetes control functions and cluster operations.	Orchestrates the environment, manages workloads, and supports cluster operations.
Lenovo ThinkSystem SR650a V4 (Worker Node*)	Compute Layer	High-density compute server optimized for AI and data processing workloads.	Runs OpenShift microservices, executes model inference, and processes RAG pipelines.
Intel Xeon 6 Processors with AMX	Compute Acceleration Layer	CPU platform with built-in AI acceleration for efficient inference.	Supports AI workloads through cost-efficient and scalable inference capabilities.
Lenovo DM3200F	Data Layer	High-performance unified	Stores enterprise data,

Storage		storage platform delivering object, file, and block data services through ONTAP.	embeddings, and vector indexes to enable efficient retrieval and scalable data access for RAG workflows.
Nokia 7215 IXS-A1	Management Network Layer	Highly efficient out-of-band management switch.	Connects to compute, storage, and network fabric to provide out-of-band management.
Nokia 7220 IXR-D2L	Data Network Layer	High-speed data-center switch enabling low-latency connectivity.	Connects compute and storage, ensuring fast data movement for real-time inference.
Red Hat OpenShift AI	AI Platform Layer	Modular AI platform for developing, deploying, and managing enterprise AI applications.	Orchestrates embedding, retrieval, and inference workflows within the AI pipeline and provides centralized management of AI services and operations.
Red Hat OpenShift	Container Platform Layer	Enterprise Kubernetes platform for managing containerized applications and distributed workloads.	Deploys, scales, and manages AI services and microservices across the cluster.
Red Hat Enterprise Linux (RHEL)	Operating System Layer	Enterprise Linux operating system supporting Red Hat OpenShift environments.	Provides the operating system foundation for infrastructure and AI software components.

*Note: The ThinkSystem SR630 V4 and SR650a V4 share the same CPU family and other components. Control plane and worker node servers can be deployed in different combinations based on use case requirements.

Lenovo SR630 V4 – Control Plane

Figure 2 - Lenovo ThinkSystem SR630 V4

The Lenovo ThinkSystem SR630 V4 serves as the control plane node for the Lenovo Hybrid AI Platform 201, providing the foundation for cluster orchestration, workload scheduling, and platform operations. As a high-density 2-socket 1U server powered by Intel Xeon 6 processors, it delivers reliable and scalable compute performance while maintaining an efficient infrastructure footprint.

Lenovo SR650a V4 – Worker Node

Figure 3 - Lenovo ThinkSystem SR650a V4

The Lenovo ThinkSystem SR650a V4 serves as the worker node for the Lenovo Hybrid AI Platform 201, delivering scalable compute capacity for AI inference and Retrieval-Augmented Generation (RAG) workloads. As a 2-socket 2U rack server powered by Intel Xeon 6 processors with Advanced Matrix Extensions (AMX), it supports efficient CPU-based inference and aligns with the platform's inference-first architecture approach. For customers with evolving AI requirements, the SR650a V4 also provides optional GPU expansion capabilities, enabling support for more compute-intensive AI workloads while preserving deployment flexibility and scalability.

Lenovo DM3200F - Storage Node

Figure 4 – Lenovo DM3200F Storage Node

The Lenovo ThinkSystem DM3200F serves as the storage foundation for the Lenovo Hybrid AI Platform 201, providing high-performance unified storage for AI inference and Retrieval-Augmented Generation (RAG) workloads. Powered by ONTAP, the platform enables consistent access to enterprise data across S3, NFS, and SMB protocols while supporting scalable data ingestion and retrieval. Within this architecture, the DM3200F stores enterprise datasets, embeddings, and vector indexes, helping deliver low-latency data access and simplified data management for AI operations.

Nokia 7215 IXS-A1 – 1Gb Ethernet

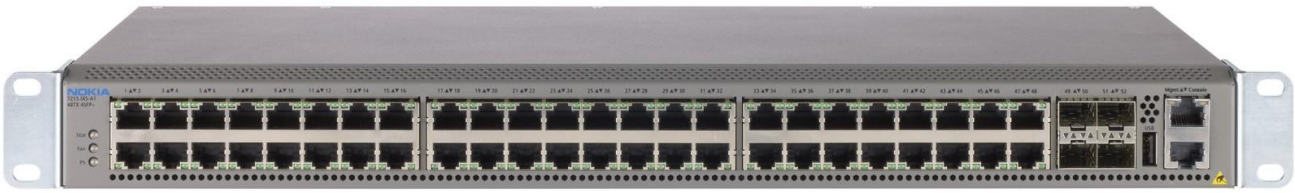

Figure 5 – Nokia 7215 IXS-A1

The Nokia 7215 IXS-A1 is a compact switch designed for secure out-of-band (OOB) management in modern data centers. It provides isolated management connectivity for servers and network infrastructure, ensuring operational access even when the production network is unavailable. With 1GbE access ports, 10GbE uplinks, and support for SR Linux, it enables automated management, telemetry, and simplified operations while integrating seamlessly into standard hot aisle/cold aisle deployments.

Nokia 7220 IXR-D2L – 100Gb Ethernet

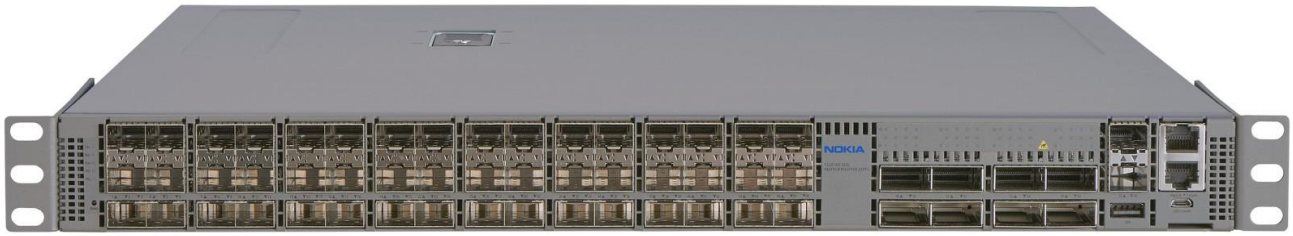

Figure 6 – Nokia 7220 IXR-D2L

The Nokia 7220 IXR-D2L is a high-performance leaf switch designed for scalable server connectivity in enterprise and AI data centers. Featuring high-density 25GbE access ports and 100GbE uplinks, it supports efficient east-west traffic flow in leaf-spine architectures. Built on SR Linux, the platform provides advanced networking, automation, telemetry, and EVPN-VXLAN capabilities while maintaining operational simplicity.

Network Architecture

The network architecture uses a Nokia 7220 IXR-D2L switch to provide high-bandwidth, low-latency connectivity between compute and storage resources. The Lenovo ThinkSystem SR630 V4 control plane node, SR650a V4 worker nodes, and DM3200F storage system connect through the data network to support AI inference and RAG workloads.

Administrative access is provided through a dedicated out-of-band management network using the Nokia 7215 IXS-A1. This separate management fabric enables secure lifecycle management, monitoring, and troubleshooting independent of the production data network.

The architecture supports horizontal scaling by adding additional SR650a V4 worker nodes and can be expanded to high-availability configurations through redundant networking and control plane resources, as shown in Figures 8 and 9.

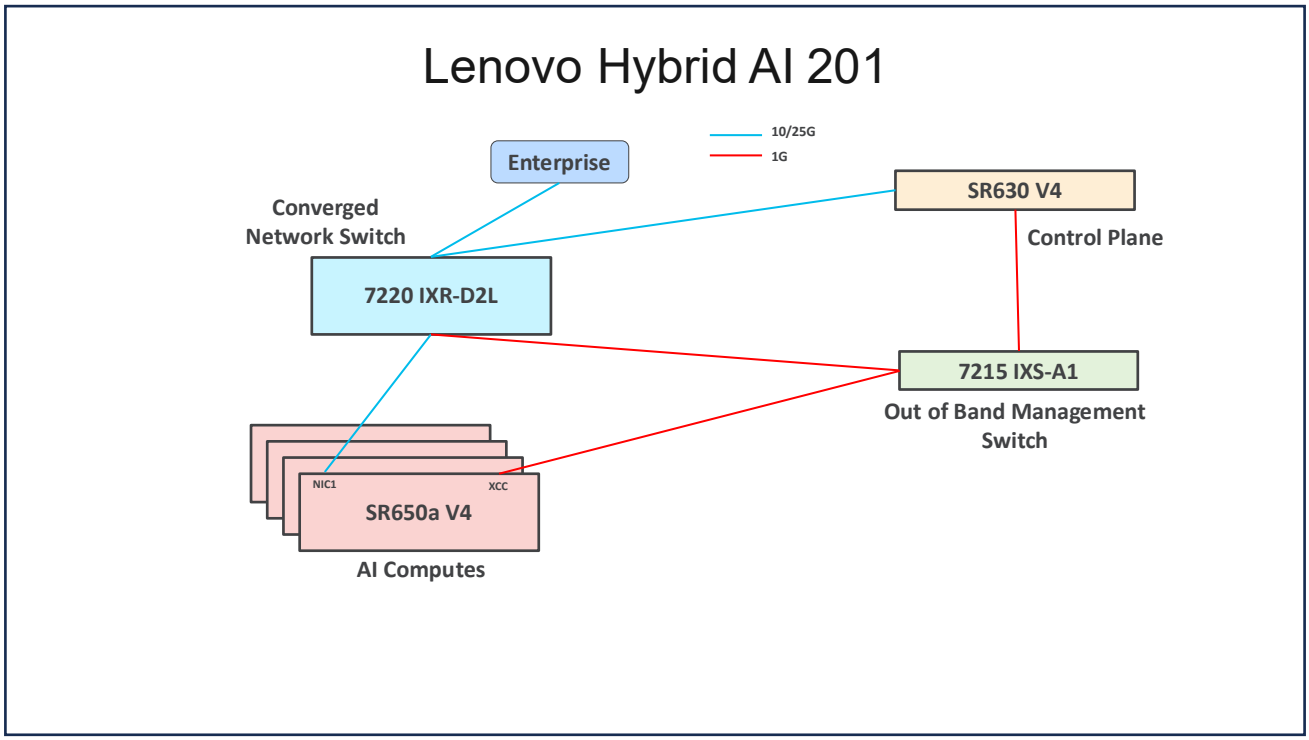

Figure 7 – Lenovo Hybrid AI 201 base configuration

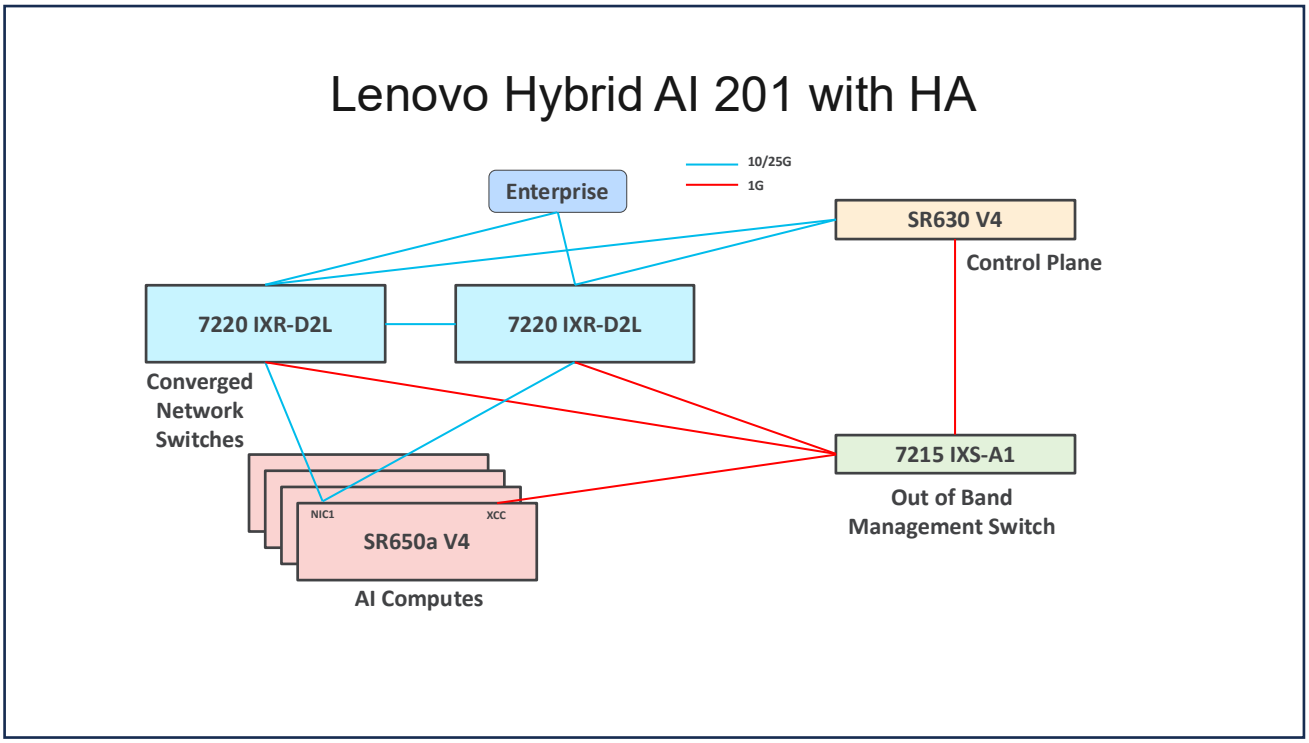

Figure 8 – Lenovo Hybrid AI 201 with HA configuration

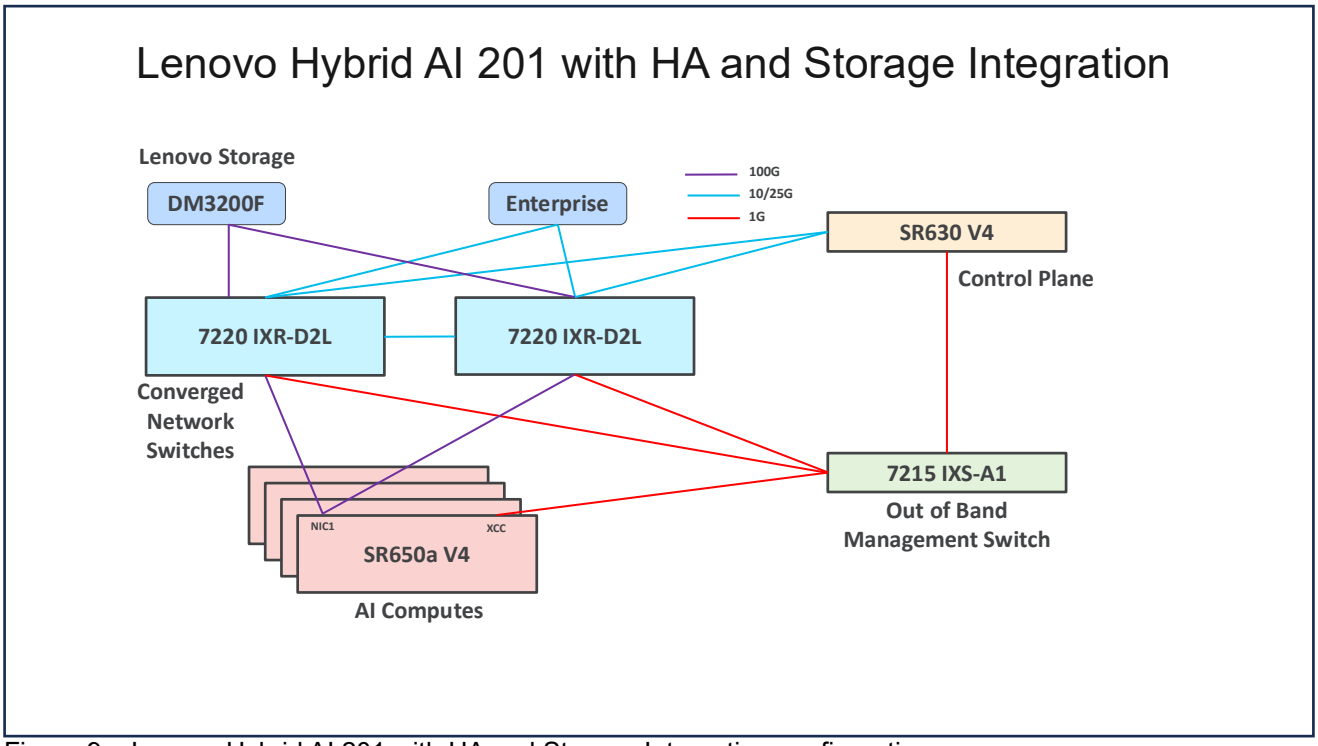

Figure 9 – Lenovo Hybrid AI 201 with HA and Storage Integration configuration

As mentioned earlier, the SR630 V4 and SR650a V4 can be mixed and matched to meet workload, space, and cost requirements. Additionally, this platform can also be integrated with Lenovo storage like the DM3200F or any pre-existing storage solutions making the platform highly modular and scalable.

AI Software Stack

The Lenovo Hybrid AI Platform 201 software stack combines Red Hat OpenShift, Red Hat OpenShift AI, ONTAP, and supporting AI services into a scalable platform for enterprise AI inference and Retrieval-Augmented Generation (RAG) workloads. Together, these components provide the foundation for AI operations, model serving, data management, and automated deployment.

Table 3 – AI Software Stack

Layer	Description
Application Layer: Red Hat Validated Patterns	Provides GitOps-based deployment automation for OpenShift environments, including validated patterns for deploying RAG applications and supporting infrastructure.
AI Platform Layer: Red Hat OpenShift AI	Delivers model serving, AI development, MLOps, pipelines, and lifecycle management for enterprise AI applications. KServe provides scalable model serving capabilities.
Container Platform Layer: Red Hat OpenShift	Enterprise Kubernetes platform that deploys, manages, secures, and scales containerized AI applications and services across the cluster.
Inference Software Layer:	Supports CPU-optimized AI inference using vLLM, Intel IPEX-LLM,

CPU AI Stack	OpenVINO, and Milvus to enable model serving, optimization, and vector search for RAG workloads.
Storage Software Layer: ONTAP + Trident CSI	Provides persistent storage services for OpenShift workloads through dynamic volume provisioning, shared file access, and high-performance storage connectivity.
Infrastructure Layer: Lenovo Hybrid AI Platform 201	Lenovo ThinkSystem servers powered by Intel Xeon 6 processors with AMX, Lenovo DM3200F storage, and Nokia networking provide the foundation for scalable enterprise AI inference and RAG workloads.

Red Hat AI Enterprise

Red Hat AI Enterprise provides the software foundation for deploying and managing enterprise AI workloads. Built on Red Hat OpenShift, it combines AI development, model serving, lifecycle management, and operational capabilities into a validated platform.

Table 4 – Red Hat AI Enterprise Components

Component	Purpose in Hybrid AI Platform 201	Business Value
OpenShift	Provides the Kubernetes foundation for AI infrastructure and operations.	Consistent deployment and management across hybrid cloud environments.
OpenShift AI	Enables AI development, deployment, model serving, and governance.	Accelerates the transition from AI experimentation to production.
Red Hat AI Inference	Optimizes inference performance and resource efficiency.	Improves performance while reducing infrastructure costs.
Validated Platform	Delivers a tested and supported AI software stack.	Reduces deployment risk and operational complexity.

More on Red Hat Validated Patterns

Red Hat Validated Patterns accelerate deployment by providing tested, GitOps-based automation for complex hybrid cloud environments. In this solution, the RAG validated pattern codifies the infrastructure, application, and configuration components required to deploy and manage the AI platform consistently. Using Red Hat OpenShift GitOps, desired-state configurations are maintained through Git-based workflows, simplifying deployment, updates, and lifecycle management. For more information and an example of a RAG validated pattern deployed on the Lenovo Hybrid AI Platform, visit: Lenovo Hybrid AI 221 Microfactory with Red Hat > Lenovo Press

Deployment Overview

The Lenovo Hybrid AI Platform 201 uses a modular deployment approach that combines Lenovo infrastructure, Red Hat OpenShift AI, and Red Hat Validated Patterns. By leveraging automated installation and GitOps-based deployment workflows, organizations can simplify platform deployment and establish a repeatable foundation for enterprise AI and RAG workloads.

Table 5 – Deployment steps

Stage	Description	Key Activities
1. Storage Deployment	Deploy and configure the storage platform for AI workloads.	Configure ONTAP storage and SVMs; Enable NFS and NVMe/TCP access; Deploy Trident CSI for dynamic persistent volume provisioning
2. Compute Infrastructure Deployment	Prepare the compute and networking environment.	Install Lenovo ThinkSystem SR630 V4 or SR650a V4 servers; Configure Nokia 7220 IXR-D2L data switch; Configure Nokia 7215 IXS-A1 management switch
3. OpenShift Deployment	Deploy the Kubernetes platform foundation.	Generate discovery ISO using Assisted Installer; Boot cluster nodes; Validate cluster health and connectivity
4. AI Platform Deployment	Deploy the AI software stack using Red Hat automation tools.	Install Validated Patterns Operator; Deploy RAG Validated Pattern; Monitor deployment through ArgoCD
5. Data Ingestion and Validation	Populate the RAG environment and validate end-to-end functionality.	Configure enterprise data sources; Trigger automated ingestion through OpenShift GitOps; Execute test queries and validate workflows

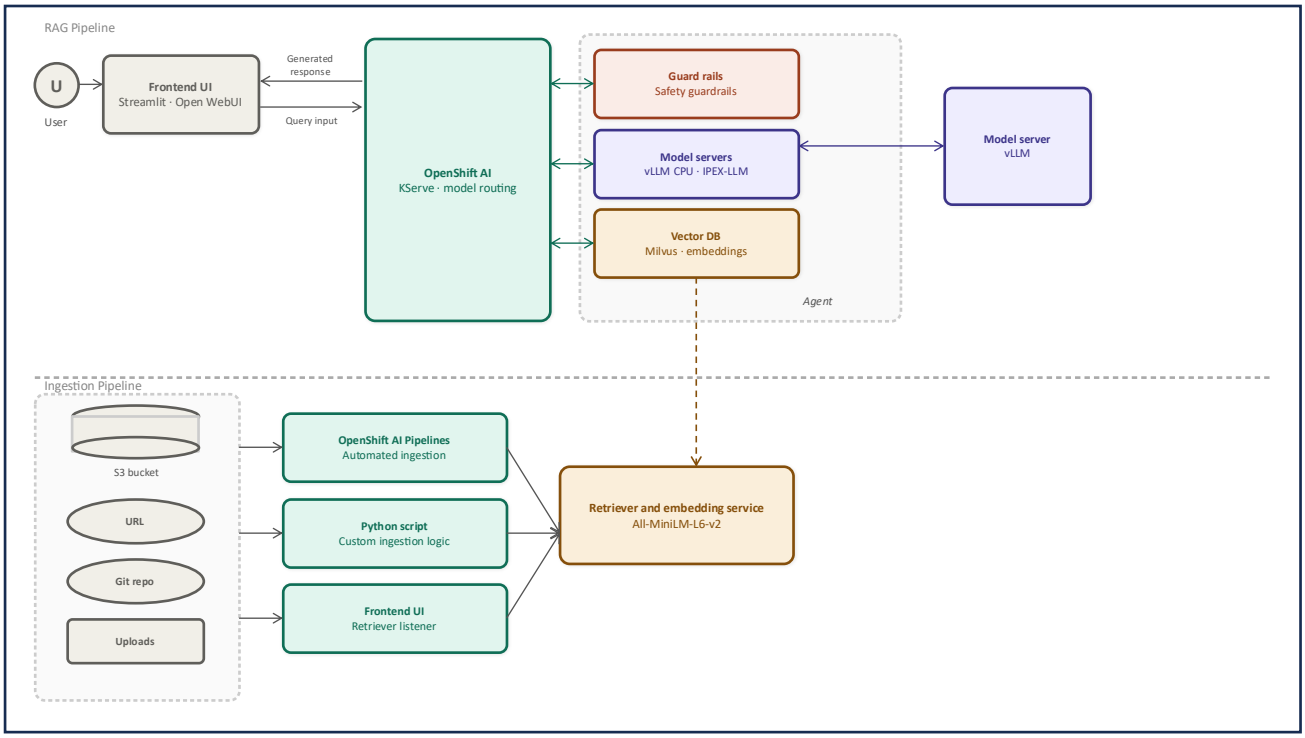

Figure 10 – RAG Reference Workflow

Key Deployment Considerations

The following considerations highlight key architectural and operational characteristics that support consistent deployment, efficient management, and long-term scalability.

- **Automated Deployment** – Red Hat OpenShift Assisted Installer and Red Hat Validated Patterns streamline deployment through automated installation, configuration, and application provisioning, reducing manual effort and deployment complexity.
- **GitOps-Based Operations –** OpenShift GitOps continuously synchronizes platform configuration from source-controlled repositories, enabling consistent deployments, simplified updates, and lifecycle management.
- **Flexible Data Ingestion –** Enterprise data can be ingested through validated pattern workflows or OpenShift AI pipelines, providing flexibility for different data sources, scales, and operational requirements.
- **Independent Scaling of Compute and Storage** – Compute and storage resources can be managed and scaled independently, allowing organizations to optimize infrastructure as workload demands evolve.
- **CPU-Optimized AI Inference –** Intel Xeon 6 processors with AMX, combined with vLLM, Intel IPEX-LLM, and OpenVINO, enable efficient CPU-based inference for enterprise RAG and generative AI workloads.

- **Kubernetes-Native Architecture –** Red Hat OpenShift provides a scalable platform for deploying, managing, and updating AI services through standardized Kubernetes orchestration and operational practices.

Performance Validation

To supplement the reference architecture, Intel benchmark data was reviewed to assess vLLM inference performance across processor generations. Using the Llama 3.1-8B-Instruct model in conversational chatbot workloads, the benchmarks measured output throughput and concurrent user capacity. The results demonstrate a significant performance advantage for Intel Xeon 6 processors, the platform used in the Lenovo Hybrid AI Platform 201, compared to 5th Generation Intel Xeon processors across all tested token configurations.

Benchmark Methodology

The benchmark evaluates vLLM serving performance across a range of input/output token combinations, from short (256 input / 256 output) to long-context (1024 input / 2048 output) workloads. Performance is assessed against two service-level objectives (SLOs) representative of production conversational AI:

- **Time Per Output Token (TPOT):** ≤ 0.1 seconds
- **Time to First Token (TTFT):** ≤ 1 second

Results are reported across two dimensions:

- **Output Throughput (tokens/second)** – a measure of overall inference capacity
- **Maximum Concurrent User Prompts** – the number of simultaneous users the system can serve within the defined SLOs

Benchmark Results

Figure 11 compares vLLM inference performance on Intel Xeon 6 (6745P) vs Intel Xeon 5 (8562Y) using Llama 3.1-8B Instruct model. Results are from Intel-validated conversational chatbot and specifies the maximum number of concurrent users which can be served at 0.1s/<1s TPOT/TTFT.

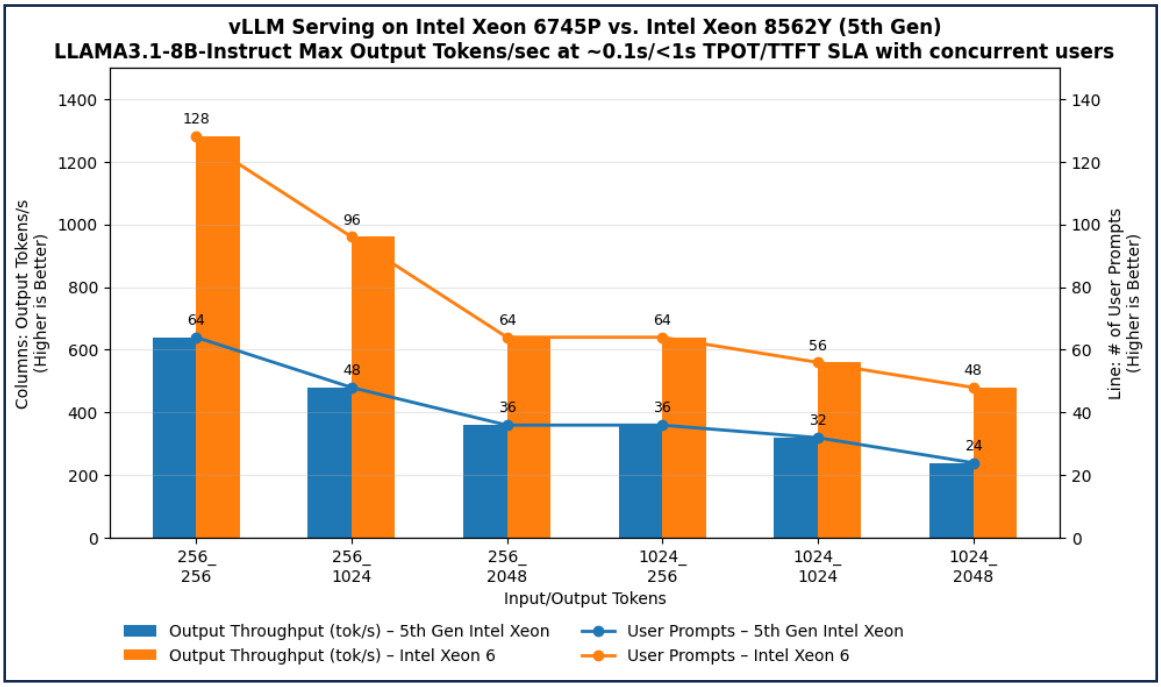

Figure 11 -- Hybrid 201 AI Inferencing Throughput and Concurrency Performance

Key Findings

- **Higher inference throughput** – Under Intel benchmark test conditions, Intel Xeon 6 delivered approximately 1.2× to 2× higher output throughput than 5th Gen Intel Xeon processors across the evaluated token configurations. At the 256-input / 256-output token workload, throughput reached approximately 1,250 tokens/sec compared to 650 tokens/sec.
- **Greater concurrent-user capacity** – Intel Xeon 6 supported up to 2× higher concurrent-user capacity while maintaining the defined service-level objectives of TPOT ≤100 ms and TTFT ≤1 second. At the 256/256 token configuration, the platform supported 128 concurrent users versus 64 users on the comparison system.
- **Consistent performance advantage at scale** – Although throughput and concurrency decrease as input and output token lengths increase, Intel Xeon 6 maintained its performance advantage across all tested workloads, supporting approximately 1.88× higher average concurrent-user capacity than the compared 5th Gen Intel Xeon system.
- **Optimized CPU-based inference** – The combination of Intel Xeon 6 processors with AMX and vLLM serving optimizations enables efficient execution of LLM inference workloads, improving throughput, batching efficiency, and resource utilization.
- **Enterprise-ready AI operations** – Running on Red Hat OpenShift, the platform provides a scalable and operationally consistent environment for deploying, managing, and scaling AI inference services.
- **Cost-efficient inference infrastructure** – The results demonstrate that CPU-based inference can deliver production-grade performance for many enterprise RAG and conversational AI workloads, providing an efficient alternative for workloads that align with the CPU-serving profile.

Relevance to Lenovo Hybrid AI Platform 201

The Lenovo Hybrid AI Platform 201 utilizes the same Intel Xeon 6 processor architecture with Advanced Matrix Extensions (AMX) as tested in the benchmark. These results provide indicative guidance for the throughput and concurrency characteristics that organizations can expect from CPU-based inference on this platform under comparable workloads and SLO targets.

The benchmark results support a key value proposition of this reference architecture: Intel Xeon 6 processors with AMX can deliver production-grade inference performance for enterprise RAG workloads without requiring dedicated GPU hardware. For organizations deploying conversational AI or RAG applications with moderate-to-high concurrency requirements, this translates to a more cost-effective and operationally simplified infrastructure footprint.

Solution Summary

The Lenovo Hybrid AI Platform 201 provides a validated and scalable foundation for enterprise AI inference and Retrieval-Augmented Generation (RAG) workloads. By integrating Lenovo compute, Lenovo storage, Intel Xeon 6 processors with AMX, and Red Hat OpenShift AI into a unified architecture, the platform simplifies deployment while supporting secure and operationally efficient AI environments.

Designed around an inference-first architecture approach, the platform enables organizations to deploy private AI closer to enterprise data while reducing dependency on dedicated GPU infrastructure for applicable workloads. Combined with ONTAP unified storage and Kubernetes-based orchestration, the solution supports scalable growth, efficient data access, and repeatable AI operations.

This solution enables organizations to:

- Deploy RAG applications on an integrated AI platform
- Lower total cost of ownership while supporting AI performance requirements
- Scale AI workloads predictably as user demand grows
- Maintain full control over data security, governance, and compliance

By turning a complex integration challenge into a scalable reference architecture, this solution provides organizations with a structured approach to operationalizing generative AI through a modular and flexible deployment model.

Appendix A: Lenovo Bill of materials (BOM)

SR630 V4

Part number	Product Description	Qty
7DG9CTO1WW	Server : ThinkSystem SR630 V4-3yr Base Warranty	1
C1XE	ThinkSystem 1U V4 10x2.5" Chassis	1
C3JB	ThinkSystem General Computing - Power Efficiency	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5QX	Intel Xeon 6737P 32C 270W 2.9GHz Processor	2
C1XJ	ThinkSystem 1U V4 Performance Heatsink	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	16
5977	Select Storage devices - no configured RAID required	1
C0ZU	ThinkSystem 2.5" U.2 VA 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD	1
C21X	ThinkSystem 1U V4 10x2.5" NVMe Gen5 Backplane	1
CC7H	ThinkSystem M.2 B350i-2i NVMe Enablement Kit	1
C287	ThinkSystem M.2 VA 960GB Read Intensive NVMe NHS SSD	2
B5T1	ThinkSystem Broadcom 5719 1GbE RJ45 4-port OCP Ethernet Adapter	1
BNWM	ThinkSystem Broadcom 57504 10/25GbE SFP28 4-port PCIe Ethernet Adapter	1
BK1J	ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe 4 Ethernet Adapter	1
C1YH	ThinkSystem SR630 V4 x16/x16 PCIe Gen5 Cable Riser 1	1
C9AR	ThinkSystem SR630 V4 Full Height+Low Profile Riser1 Cage v2	1
C0U3	ThinkSystem 2000W 230V Titanium CRPS Premium Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
C1YT	ThinkSystem 1U V4 Performance Fan Module	4
C1YP	ThinkSystem 1U V4 Standard Media Bay	1
C2DH	ThinkSystem Toolless Slide Rail Kit V4	1
BPKR	TPM 2.0	1

SR650a V4

Part number	Product Description	Qty
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
C5QX	Intel Xeon 6737P 32C 270W 2.9GHz Processor	2
C3QR	ThinkSystem 2U V4 Performance Heatsink	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	16
5977	Select Storage devices - no configured RAID required	1
C1W9	ThinkSystem E3.S 1T VA 3.84TB Read Intensive NVMe PCIe 5.0 x4 HS SSD v2	1

C221	ThinkSystem V4 EDSFF E3.S 4x1T NVMe Gen5 Backplane	1
CC7H	ThinkSystem M.2 B350i-2i NVMe Enablement Kit	1
C287	ThinkSystem M.2 VA 960GB Read Intensive NVMe NHS SSD	2
B5T1	ThinkSystem Broadcom 5719 1GbE RJ45 4-port OCP Ethernet Adapter	1
BNWM	ThinkSystem Broadcom 57504 10/25GbE SFP28 4-port PCIe Ethernet Adapter	1
BK1J	ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe 4 Ethernet Adapter	1
C4U4	ThinkSystem SR650a V4 x16 Front Riser Slot 17	1
C4U3	ThinkSystem SR650a V4 x16 Front Riser Slot 19	1
C4U2	ThinkSystem SR650a V4 x16 Front Riser Slot 21	1
C4U1	ThinkSystem SR650a V4 x16 Front Riser Slot 23	1
C62D	ThinkSystem SR650/a V4 x16 Rear Direct Riser Slot 5	1
C4U0	ThinkSystem SR650/a V4 x16 Rear Direct Riser Slot 8	1
C0U3	ThinkSystem 2000W 230V Titanium CRPS Premium Hot-Swap Power Supply	2
C3RD	ThinkSystem 2U 6056 20K Performance Fan Module	6
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

Nokia Switches

Part number	Product Description	Qty
7DQ1CTO1WW	Switch : Nokia 7215 IXS-A1 1GbE Managed Switch with SR Linux (PSE)	1
CEM9	Nokia 7215 IXS-A1 1GbE Managed Switch with SR Linux (PSE)	1
AVFZ	1.5m Green Cat6 Cable	1
6311	2.8m, 10A/100-250V, C13 to C14 Jumper Cord	2
7DQ1CTO3WW	Switch : Nokia 7220 IXR-D2L 25GbE Managed Switch with SR Linux (PSE)	1
CEMB	Nokia 7220 IXR-D2L 25GbE Managed Switch with SR Linux (PSE)	1
AVFZ	1.5m Green Cat6 Cable	1
6311	2.8m, 10A/100-250V, C13 to C14 Jumper Cord	2
CEMF	Nokia 3HE17659AA Rack Mount Kit	1

DM3200F Storage

Part number	Product Description	Qty
7DJ0CTO1WW	Controller : Lenovo ThinkSystem DM3200F All Flash Array	1
BF3C	Lenovo ThinkSystem Storage 2U NVMe Chassis	1
BWU8	Storage Complete Bundle Offering	1
C4A6	Lenovo ThinkSystem DM3200 Series Controller, 64GB	2
C3XK	Lenovo ThinkSystem 30.7TB (2x 15.36TB NVMe SED) Drive Pack	4
C4AC	Lenovo ThinkSystem Storage 10/25Gb 4 port Ethernet	2
AV1W	Lenovo 1m Passive 25G SFP28 DAC Cable	2
BY6K	USB-A to USB-C Cable	1

6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
CF4S	Lenovo ThinkSystem Storage ONTAP 9.18 Software Encryption - IPAv2	1
B0W1	3 Years	1
C6S3	Premier 24x7 4hr with Enhanced Storage Support and KYD	1
C48Q	Configured with Lenovo ThinkSystem DM3200F 3Yr Warranty	1
BWUC	Storage Complete Bundle License Key	2
BWUE	Storage Encryption Bundle License Key - RoW	2
C49B	Lenovo ThinkSystem DM/DG Series Jupiter All Flash Ship Kit - Multi-Language	1
B6Y6	Lenovo ThinkSystem NVMe Rail Kit 4 post	1
B738	Lenovo ThinkSystem NVMe Accessory	1
C48X	Lenovo ThinkSystem DM/DG/DS Jupiter/Saturn 2U24 NVMe Bezel	1
C8V9	7-segment LED cover Label	1
C3HV	Lenovo Logo nameplate	1
C498	Lenovo ThinkSystem Storage Controller 2U24 NVMe Agency Label	1
C490	Lenovo ThinkSystem DM3200F Model Name Label	1
B4E7	EIA NamePlate	1
B6Y5	Lenovo ThinkSystem NVMe SFF Filler	16
C5H3	Lenovo ThinkSystem Storage NVMe Packaging 2U	1
C48W	I/O Slot Cover	6

Appendix B: Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AMX	Advanced Matrix Extensions
API	Application Programming Interface
CPU	Central Processing Unit
GPU	Graphics Processing Unit
LLM	Large Language Model
MLOps	Machine Learning Operations
NFS	Network File System
ONTAP	NetApp data management software used by Lenovo DM Series storage
OpenShift AI	Red Hat enterprise AI platform built on OpenShift
RAG	Retrieval-Augmented Generation
RBAC	Role-Based Access Control
S3	Simple Storage Service-compatible object protocol
SLA	Service-Level Agreement
SMB	Server Message Block
TPOT	Time Per Output Token
TTFT	Time To First Token
vLLM	LLM inference and serving framework

Resources

Intel AMX	Intel AMX
SR630 V4	Lenovo ThinkSystem SR630 V4 Server Product Guide > Lenovo Press
SR650a V4	https://lenovopress.lenovo.com/datasheet/ds0195-lenovo-thinksystem-sr650a-v4
Lenovo DM3200F	https://lenovopress.lenovo.com/lp2073-introducing-the-new-lenovo-thinksystem-storage-arrays
Nokia 7220 IXR-D series	Nokia 7220 Interconnect Router for Data Center Fabrics
Red Hat OpenShift AI	Red Hat OpenShift AI
NetApp ONTAP	NetApp ONTAP > NetApp Solutions
NetApp Trident	NetApp Trident Overview
Intel RH OpenShift White paper	Red Hat OpenShift Whitepaper
Lenovo Hybrid AI 221 Microfactory with Red Hat	Lenovo Hybrid AI 221 Microfactory with Red Hat > Lenovo Press

Document history

Version 1.0	June 2026	ThinkSystem SR630 V4 and SR650a V4 with Lenovo DM3200F storage, Red Hat OpenShift AI v.3.4

Trademarks and special notices

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®
ThinkEdge®
ThinkShield®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Intel® and Intel Core® are trademarks 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.

Information is provided "AS IS" without warranty of any kind.
All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.
Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serves as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites are at your own risk