Lenovo Validated Design: AI POD for Enterprise RAG Implementation

Version 1.0

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

Accelerate enterprise AI adoption with a validated, production-ready platform

Accelerate production-ready RAG adoption in on-prem enterprise environments

Simplify enterprise AI data management with a unified storage platform

Vanita Meyer
Eric Page
Abed Islam

Table of Contents

Introduction 1

Executive Summary 1

Intended Audience 1

Challenges and Opportunity 2

Key Challenges 2

Strategic Opportunities 2

Solution Overview 4

Solution Workflow 4

Key differentiators: 4

Data Layer Optimization for RAG Workloads 5

Key capabilities include: 5

Efficient AI Inference with CPU Acceleration 5

Key capabilities include: 5

Solution Components 6

Hardware 6

Lenovo SR650 V4 - Control Plane Node 7

Lenovo SR630 V4 - Worker Node 6

NetApp AFF A-Series - Storage Node 7

Nokia 7220 IXR-D3L - Network Switch 7

Network Deployment Architecture 7

Software Stack 8

Open Platform for Enterprise AI 9

Intel AI for Enterprise RAG 10

NetApp ONTAP 10

NetApp Trident 11

Deployment 12

Deployment Overview 12

Data Ingestion and Validation 12

Performance Validation 14

Objective 14

Methodology 14

Key Findings 14

Test Configuration 15

Detail Results 15

Interpretation 17

Solution Summary 18

Appendix A: Lenovo Bill of materials (BOM) 19

Appendix B: Abbreviations 22

Resources 23

Document history 24

Trademarks and special notices 25

Introduction

Executive Summary

Organizations are moving quickly to adopt generative AI, but most struggle to transition from pilot projects to production. The challenge is not building AI models, it is deploying them at scale with predictable performance, controlled cost, and secure access to enterprise data. Retrieval-Augmented Generation (RAG) introduces additional complexity by requiring tight integration across compute, storage, and orchestration layers.

The AI POD with NetApp and Intel Open Platform for Enterprise AI (OPEA) address this gap by delivering a compact, validated platform purpose-built for production RAG workloads. By combining Lenovo ThinkSystem infrastructure, NetApp ONTAP data management, and a modular microservices-based AI framework, the solution simplifies deployment while ensuring consistent, low-latency performance.

A key innovation is the ability to run AI inference efficiently on CPUs using Intel Xeon 6 processors with Advanced Matrix Extensions (AMX), reducing reliance on GPUs and significantly lowering total cost of ownership without compromising performance. At the same time, NetApp ONTAP provides a unified data layer that enables fast, secure access to enterprise data across object and file protocols, which is critical for accurate and responsive AI outputs.

This Lenovo Validated Design delivers:

- A production-ready platform for deploying RAG applications with predictable performance
- Reduced infrastructure cost through efficient CPU-based AI acceleration
- Simplified deployment with a pre-validated, integrated architecture
- Secure, on-premises AI with full control over data governance and compliance

By transforming a complex, multi-layered AI stack into a validated and repeatable solution, the AI POD enables organizations to move from experimentation to real-world AI deployment with confidence.

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 Opportunity

Key Challenges

Organizations are increasingly exploring generative AI and RAG applications, but several barriers prevent them from scaling beyond initial pilots:

- **Infrastructure constraints**
  Traditional AI environments often require significant space, power, and investment, making them difficult to deploy, especially for SMBs and edge locations.
- **Data access and integration complexity**
  RAG workloads depend on retrieving and processing data across multiple formats and storage systems, including object, file, and hybrid environments. This introduces latency, fragmentation, and operational complexity.
- **Cost and GPU dependency**
  Many AI deployments rely heavily on GPUs, which are expensive, limited in availability, and not always required for inference workloads, increasing overall cost and limiting scalability.
- **Operational and architectural complexity**
  Building and managing RAG pipelines requires integrating compute, storage, orchestration, and AI frameworks, often across multiple vendors, which increases deployment time and risk.
- **Data governance and security requirements**
  Enterprises must ensure sensitive data remains secure, compliant, and under control, particularly when deploying AI in regulated or on-premises environments.

Strategic Opportunities

The convergence of generative AI adoption and advancements in infrastructure creates a clear opportunity to simplify and scale AI deployments:

- **Simplified, validated AI deployment**
  A pre-integrated and validated solution reduces complexity, accelerates deployment, and lowers the risk associated with building AI infrastructure from scratch.
- **Efficient, cost-optimized AI inference**
  CPU-based acceleration with Intel Xeon 6 and AMX enables strong AI performance while reducing reliance on GPUs, lowering total cost of ownership.
- **Low-latency, unified data access**
  NetApp ONTAP provides high-performance access across S3, NFS, and SMB, enabling faster data retrieval and improving RAG inference performance.
- **Scalable, modular architecture**
  Kubernetes-based design allows organizations to start small and scale incrementally as workloads and user demand grow.
- **On-premises AI with governance and control**
  Enterprises can deploy AI where their data resides, ensuring compliance, security, and sovereignty

while maintaining performance.

Why This Matters Now

The rapid rise of generative AI, including RAG and agentic AI use cases, is driving demand for production-ready platforms that can scale efficiently while maintaining cost control and data governance. Organizations need solutions that move beyond experimentation and enable consistent, reliable AI operations in real-world environments.

This solution directly addresses these needs by providing a compact, scalable, and validated foundation for deploying enterprise AI workloads with confidence.

Solution Overview

The AI POD with NetApp and Intel OPEA deliver a compact, production-ready platform purpose-built for enterprise Retrieval-Augmented Generation (RAG) workloads. It combines Lenovo ThinkSystem compute powered by Intel Xeon 6 processors, NetApp AFF storage with ONTAP data management, and high-speed networking into a unified, validated architecture that simplifies the deployment of generative AI applications.

At its core, the solution integrates compute, data, and AI orchestration into a cohesive platform that enables low-latency data retrieval, consistent inference performance, and scalable operations. Built on Kubernetes and Intel's Open Platform for Enterprise AI (OPEA), it provides a modular, microservices-based framework that streamlines development, deployment, and management of RAG pipelines.

Solution Workflow

Enterprise data is ingested and stored in NetApp ONTAP, where it is securely managed and made accessible through S3, NFS, or SMB. This data is then processed through embedding and retrieval pipelines orchestrated by OPEA microservices 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 accurate, context-aware responses in real time.

Key differentiators:

- **Reduced GPU dependency**
  Leverages Intel Xeon 6 processors with built-in AMX to deliver efficient AI inference on CPUs, lowering cost and improving accessibility.

- **Unified data access for RAG workloads**
  NetApp ONTAP enables seamless access to structured and unstructured data across multiple protocols, improving retrieval speed and simplifying data management.

- **Pre-validated, integrated architecture**
  Combines compute, storage, networking, and AI software into a tested solution, reducing deployment risk and accelerating time to production.

- **Modular and scalable design**
  Kubernetes-based architecture allows organizations to start small and scale horizontally as workload demands increase.

- **Built for on-premises and hybrid AI**
  Enables organizations to run AI close to their data, ensuring governance, security, and compliance while maintaining performance.

By bringing together infrastructure and AI software into a single validated design, the solution removes the complexity of building RAG environments from scratch and provides a clear, scalable path from pilot to production.

The NetApp AI POD solution delivers a high-performance infrastructure stack purpose-built for enterprise Retrieval-Augmented Generation (RAG) workloads. By integrating Lenovo compute with Intel Xeon 6 processors, NetApp AFF storage, and high-speed 100Gbe Nokia networking, the platform provides the performance, scalability, and operational simplicity required to deploy production-grade AI services on-premises. At its core, the solution combines intelligent orchestration, low-latency storage access, and high-bandwidth connectivity to ensure consistent model inference performance while maintaining enterprise data governance and security.

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, NetApp 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, reducing latency and improving response quality.
- **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, ensuring enterprise data remains secure and controlled throughout the AI lifecycle.
- Result: Faster, more reliable RAG responses with enterprise-grade data control.

Efficient AI Inference with CPU Acceleration

AI inference is often constrained by GPU cost and availability. This solution leverages Intel Xeon 6 processors with built-in AMX to enable efficient, scalable AI processing on CPUs.

Key capabilities include:

- **Cost-efficient inference performance**
  Delivers strong AI performance without heavy reliance on GPUs, reducing infrastructure cost.
- **Improved performance per watt**
  Optimizes resource utilization, enabling higher throughput with lower power consumption.
- **Simplified development and deployment**
  Works seamlessly with existing AI frameworks, allowing teams to leverage CPU acceleration with minimal changes.
- **Result:** Lower TCO and scalable AI inference without GPU constraints.

Solution Components

Hardware and Software

Component	Description	Role in the Solution
Lenovo ThinkSystem SR630 V4 (Worker Node)	High-density compute server optimized for AI and data processing workloads.	Runs OPEA microservices, executes model inference, and processes RAG pipelines.
Lenovo ThinkSystem SR650 V4 (Control Plane Node)	Enterprise server managing Kubernetes control functions and cluster operations.	Orchestrates the environment, manages workloads, and ensures system reliability.
Intel Xeon 6 Processors with AMX	CPU platform with built-in AI acceleration for efficient inference.	Powers AI workloads with cost-efficient, scalable inference performance.
NetApp AFF A-Series Storage (ONTAP)	High-performance unified storage supporting object, file, and block access.	Stores enterprise data, embeddings, and vector indexes for fast retrieval in RAG workflows.
Nokia 7220 IXR-D3L (100GbE Switch)	High-speed data center switch enabling low-latency connectivity.	Connects compute and storage, ensuring fast data movement for real-time inference.
NetApp Trident (CSI Driver)	Kubernetes-native storage orchestrator for NetApp systems.	Automates storage provisioning and integrates persistent storage into containerized workloads.
Intel Open Platform for Enterprise AI (OPEA)	Modular AI framework providing microservices for building RAG applications.	Orchestrates embedding, retrieval, and inference pipelines within the AI workflow.
Kubernetes	Container orchestration platform for managing distributed applications.	Deploys, scales, and manages AI services and microservices across the cluster.
Ubuntu OS	Stable Linux operating system optimized for enterprise workloads.	Provides the runtime environment for infrastructure and AI software components.

Lenovo SR630 V4 - Worker Node

Figure 2 - Lenovo ThinkSystem SR630 V4

The Lenovo ThinkSystem SR630 V4 is a high-density 2-socket 1U server delivering reliable, secure, and

scalable performance for modern, compute-intensive workloads.

Lenovo SR650 V4 - Control Plane Node

Figure 1 - Lenovo ThinkSystem SR650 V4

The Lenovo ThinkSystem SR650 V4 is a scalable 2-socket 2U server delivering high performance, reliability, and flexibility for demanding, compute-intensive workloads.

NetApp AFF A-Series - Storage Node

Figure 3 – NetApp AFF A20 Storage Node

The NetApp AFF A20 is a compact 2U all-flash system supporting up to 6 nodes and 18.8PB effective capacity, with efficient power usage and flexible connectivity for mid-range enterprise workloads.

Nokia 7220 IXR-D3L - Network Switch

Figure 4 – 7220 IXR-D3L 32QSFP28 2SFP+ Network Switch

The Nokia 7220 IXR-D3L is a compact 1RU switch delivering up to 3.2 Tbps throughput with flexible 100/40/25/10G connectivity, redundant power, and high-efficiency airflow for scalable data center networking.

Network Deployment Architecture

The AI POD Mini architecture provides a high-performance, scalable foundation designed to support enterprise AI and Retrieval-Augmented Generation (RAG) workloads. It integrates compute, storage, and

networking into a unified platform optimized for low-latency data access and efficient AI inference.

At the core of the solution is a NetApp AFF A20 storage system, which delivers enterprise-grade data management with high throughput and low-latency access. This storage layer houses enterprise datasets, embeddings, and vector indexes, and is exposed via NFS from NetApp ONTAP, enabling seamless access by compute nodes without requiring data movement or duplication.

Compute resources are delivered through **Lenovo ThinkSystem servers powered by Intel Xeon 6 processors**, forming a Kubernetes-based environment:

- **Lenovo ThinkSystem SR630 V4 (Worker Node)**
  Executes AI workloads, including OPEA microservices, model inference, and RAG processing pipelines.
- **Lenovo ThinkSystem SR650 V4 (Control Plane Node)**
  Manages cluster orchestration, scheduling, and lifecycle operations for containerized AI services.

All components are interconnected through a Nokia 7220 IXR-D3L 100GbE switch, providing a high-bandwidth, low-latency network backbone for compute-to-storage communication. This ensures efficient data flow, which is critical for real-time AI inference and consistent performance under load.

The architecture is designed for modularity and horizontal scalability. Additional Lenovo ThinkSystem SR630 worker nodes can be added to scale compute capacity and support increasing user concurrency. Workloads are distributed across nodes to enhance performance and provide high availability, while maintaining consistent access to shared storage.

A separate management network is recommended for administrative operations. A minimum of 10GbE connectivity is advised to ensure reliable system management, monitoring, and maintenance.

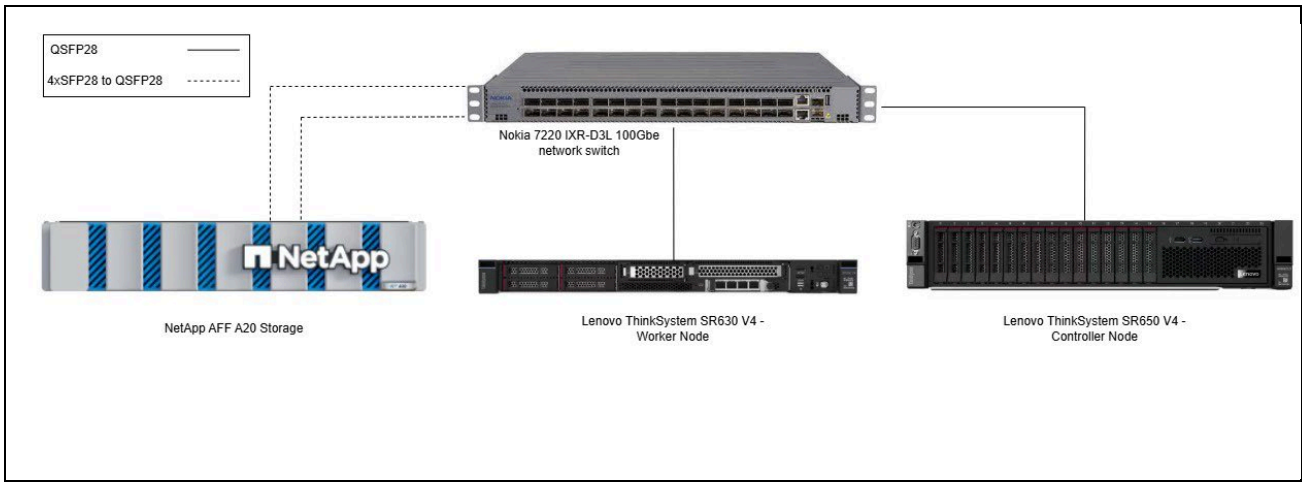

Figure 5 – NetApp AI POD network architecture

A separate management network is recommended for system administration, monitoring, and lifecycle operations, with a minimum of 10GbE connectivity.

Software Stack

The solution is built on a modern, containerized software stack that enables modular deployment, scalability,

and simplified operations for enterprise AI workloads.

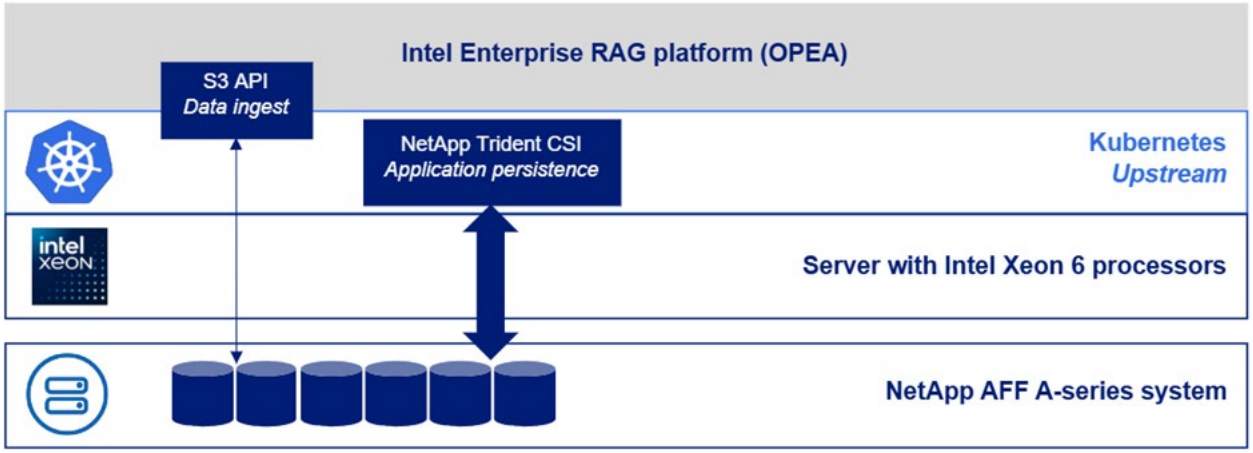

Figure 6 – Software stack for deploying on the Intel Enterprise RAG platform

Table 1 – Software Requirements

Software	**Version**
OPEA - Intel® AI for Enterprise RAG	2.0
Container Storage Interface (CSI driver)	NetApp Trident 25.10
Ubuntu	22.04.5
Container orchestration	Kubernetes 1.31.9 (Installed by Enterprise RAG infrastructure playbook)
ONTAP	ONTAP 9.16.1P4 or above

Open Platform for Enterprise AI

The Open Platform for Enterprise AI (OPEA) is an open-source framework designed to simplify the development and deployment of enterprise generative AI applications, with a strong focus on Retrieval-Augmented Generation (RAG) architectures. It provides a modular, microservices-based approach that enables organizations to build scalable AI pipelines using composable components.

OPEA includes key building blocks such as large language model (LLM) integration, data ingestion and retrieval services, prompt orchestration, and reference RAG architectures. It also incorporates an evaluation framework to assess performance, reliability, and enterprise readiness, helping organizations validate AI workloads before production deployment.

At its core, OPEA is built on two primary elements:

- GenAIComps: A library of microservices that can be combined to create flexible and scalable AI pipelines
- GenAIExamples: Pre-built reference implementations, such as ChatQnA, that demonstrate how to assemble end-to-end RAG applications

By providing reusable components and validated patterns, OPEA reduces development complexity and accelerates the path from prototype to production.

Intel AI for Enterprise RAG

Intel AI for Enterprise RAG, built on OPEA, provides a production-ready framework for deploying scalable and secure RAG applications. It extends the OPEA foundation with additional capabilities that simplify operations, enhance security, and improve usability in enterprise environments.

The solution integrates Intel Xeon processors with a broader ecosystem of software components to deliver streamlined, containerized architecture for AI workloads. It supports deployment across on-premises and hybrid environments, enabling organizations to scale AI applications based on their infrastructure and data requirements.

Key capabilities include:

- Production-ready RAG pipelines with validated deployment patterns
- Microservices-based architecture for scalability and flexibility
- Integrated security and governance, including Identity and Access Management (IAM)
- Programmable guardrails to enforce policies and control AI behavior
- Service mesh support for managing communication across distributed AI services

By combining orchestration, security, and validated workflows, Intel AI for Enterprise RAG enables organizations to operationalize generative AI with greater confidence and efficiency.

NetApp ONTAP

NetApp ONTAP provides the data foundation for the AI POD solution, delivering high-performance, secure, and scalable storage for RAG workloads. It enables consistent, low-latency access to enterprise data, which is critical for retrieval and inference operations.

ONTAP supports unified data access across object, file, and block storage, allowing AI applications to retrieve data using standard protocols such as S3, NFS, and SMB. This flexibility simplifies data integration and eliminates the need for data duplication or movement across environments.

Key capabilities include:

- High-performance data access for fast retrieval of documents, embeddings, and vector data
- Unified storage architecture supporting multiple protocols and data types
- Built-in data protection and security, including encryption and ransomware protection
- Support for hybrid and multi-cloud environments, enabling flexible deployment models

By providing a reliable and efficient data layer, ONTAP ensures that RAG pipelines can access and process data at scale while maintaining strong governance and security.

NetApp Trident

NetApp Trident is a Kubernetes-native storage orchestrator that enables automated provisioning and lifecycle management of persistent storage for containerized applications. It integrates seamlessly with NetApp ONTAP, allowing storage resources to be dynamically allocated and managed within the Kubernetes environment.

Trident supports multiple storage protocols, including NFS and iSCSI, providing flexibility for different workload requirements. It simplifies storage operations by abstracting underlying infrastructure, enabling developers and platform teams to focus on deploying and managing AI applications.

Key capabilities include:

- Automated storage provisioning for Kubernetes workloads
- Seamless integration with ONTAP storage systems
- Support for multiple storage protocols to meet diverse application needs
- Simplified storage management within containerized environments

By bridging Kubernetes and enterprise storage, Trident ensures that AI workloads have reliable, scalable access to persistent data.

Deployment

The AI POD solution is designed for streamlined deployment using a validated, repeatable approach that integrates compute, storage, and AI software into a unified platform. The deployment process follows a logical sequence that prepares the infrastructure, configures data access, and enables RAG applications through OPEA.

Deployment Overview

The deployment consists of five key stages:

- **Deploy and configure storage**
  Provision the NetApp ONTAP system and configure storage virtual machines (SVMs) with support for NFS and S3 access. This establishes the data foundation for RAG workloads.
- **Configure data access and permissions**
  Create S3 buckets and configure user access policies to enable secure data ingestion and retrieval. Proper permissions ensure that AI pipelines can interact with enterprise data reliably and securely.
- **Deploy compute infrastructure**
  Install and configure Lenovo ThinkSystem servers with Ubuntu and prepare the environment for Kubernetes-based orchestration.
- **Deploy the AI software stack**
  Install Kubernetes and deploy Intel AI for Enterprise RAG (OPEA) using the provided infrastructure playbooks. This sets up the microservices-based framework for running RAG pipelines.
- **Ingest data and validate workflows**
  Load enterprise data into the storage system and verify ingestion through the OPEA interface. Once data is available, execute queries to validate end-to-end RAG functionality.

Key Deployment Considerations

- Validated and repeatable process – The solution leverages pre-defined deployment workflows, reducing complexity and minimizing configuration errors.
- Flexible data ingestion options - Data can be ingested directly into S3 storage using standard tools or through the OPEA user interface, depending on scale and requirements.
- Separation of data and compute layers - Storage and compute are independently managed, allowing flexibility in scaling and optimizing each layer.
- Kubernetes-based orchestration - Containerized deployment simplifies scaling, updates, and lifecycle management of AI **services.**

Data Ingestion and Validation

Once deployed, enterprise data is ingested into the ONTAP storage layer and automatically processed by the OPEA framework. The system continuously monitors for updates, ensuring that data remains current for RAG-based queries.

Users can interact with the system through the OPEA interface, where queries are enriched using retrieved

enterprise data to generate accurate, context-aware responses. This step validates that the full pipeline, from data ingestion to inference, is functioning as expected.

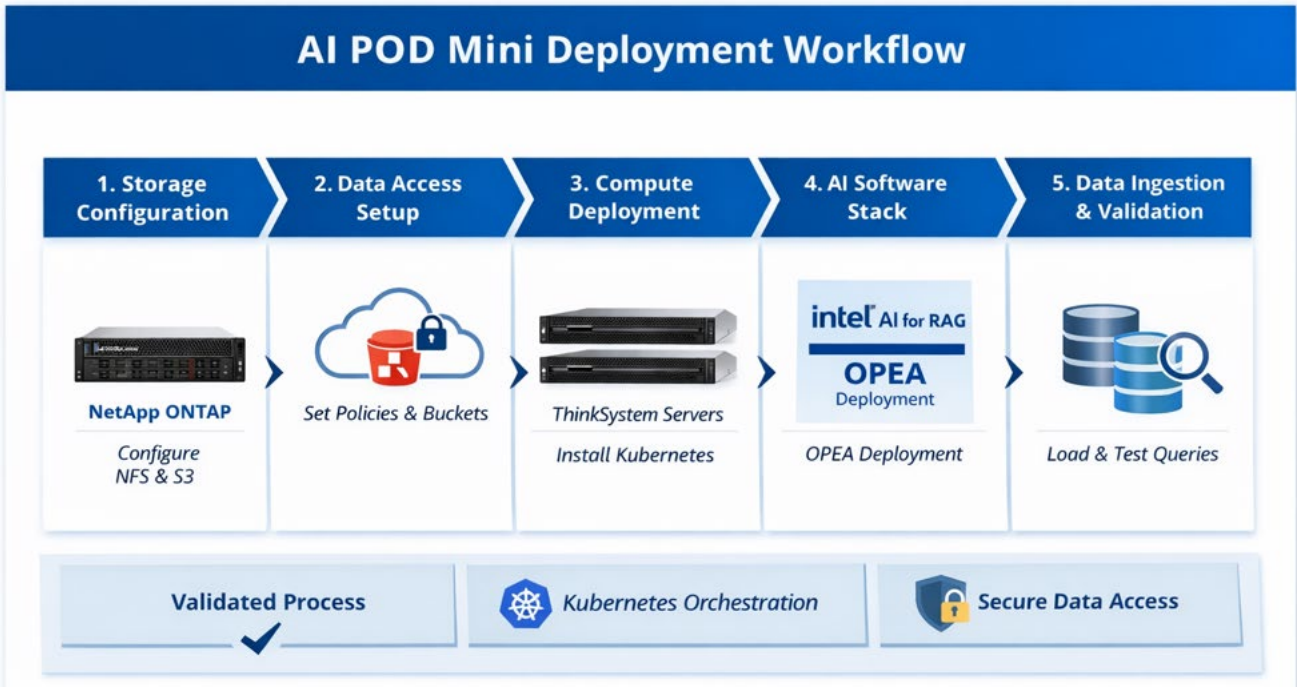

Figure 7 – AI POD Deployement Workflow

Performance Validation

Objective

The objective of this validation was to evaluate the end-to-end performance, scalability, and infrastructure balance of the AI POD solution under realistic Retrieval-Augmented Generation (RAG) workloads. Testing focused on how the system behaves as user concurrency increases, ensuring that compute, storage, and networking operate cohesively without introducing bottlenecks.

Methodology

Performance testing was conducted using representative RAG inference workloads across multiple model types and configurations. The system was evaluated under increasing levels of concurrent users, ranging from single-user execution to high-concurrency scenarios.

Each test scenario represents a different prompt and response size:

- **128–128**: Short prompt and short response
- **256–256**: Medium prompt and response
- **256–512**: Medium prompt and extended generated response
- **256–1024**: Medium prompt and long generated response

The following key metrics were measured:

- Throughput (tokens per second), indicating total system capacity
- Throughput per user, reflecting performance consistency under load
- Time to First Token (TTFT), representing user-perceived latency
- Power usage statistics

Testing reflects end-to-end inference behavior within the RAG pipeline, including data access from the storage layer.

Key Findings

- **Consistent throughput scaling**: Aggregate throughput increased proportionally with user concurrency, demonstrating effective horizontal scalability.
- **Predictable performance under load**: Per-user throughput decreased gradually as concurrency increased, following expected scaling behavior without instability.
- **Controlled latency growth**: TTFT increased in a consistent and manageable manner, maintaining responsiveness even at higher concurrency levels.
- **Balanced system architecture**: Validation results indicate that the NetApp storage layer does not introduce bottlenecks, enabling sustained high throughput and supporting linear scalability across concurrent user workloads.
- **Flexible model efficiency**: AWQ and non-AWQ model variants exhibited different performance and latency profiles, allowing optimization based on workload requirements.

Test Configuration

Table 2 – Validation Setup

Category	Configuration
Compute	Lenovo ThinkSystem SR630 V4 (Worker Nodes), SR650 V4 (Control Plane)
Processor	Intel Xeon 6 with Advanced Matrix Extensions (AMX)
Storage	NetApp AFF A-Series with ONTAP
Networking	100GbE (Nokia 7220 IXR-D3L)
Software Stack	Kubernetes, Intel AI for Enterprise RAG (OPEA)
Models Tested	meta-llama/Llama-3.1-8B-Instruct (Llama), hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4 (Llama-AWQ), Qwen/Qwen3-14B-AWQ (Qwen-AWQ), mistralai/Mistral-7B-Instruct-v0.3 (Mistral), solidrust/Mistral-7B-Instruct-v0.3-AWQ (Mistral-AWQ)
Concurrency Range	1 to 128 users
Workload Profiles	128–128, 256–256, 256–512, 256-1024 (input tokens → output tokens)

Detail Results

Table 3 – Validation Results for 128-128 Workload Profile

Users	Llama-AWQ			Llama			Qwen-AWQ			Mistral-AWQ			Mistral		
	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)
1	20.0	20.01	321.9	12.4	12.41	343.9	13.8	13.82	369.9	25.1	25.08	412.0	14.3	14.31	382.5
2	39.7	19.85	331.0	24.7	12.35	350.4	26.9	13.45	434.5	50.0	25	416.7	25.7	12.84	434.9
4	70.6	17.64	376.1	41.1	10.27	425.9	45.5	11.39	413.1	82.9	20.73	435.7	38.8	9.69	570.6
8	136.2	17.02	475.2	66.1	8.26	538.4	87.5	10.94	528.2	159.4	19.92	733.8	71.6	8.95	645.1
16	251.3	15.71	593.3	130.4	8.15	680.2	165.9	10.37	693.5	288.0	18	927.2	140.9	8.81	978.2
32	470.7	14.71	864.7	254.6	7.96	965.4	304.9	9.53	1067.	540.8	16.9	1385.3	272.2	8.51	1534.
64	836.9	13.08	1235.	450.1	7.03	1392.	516.0	8.06	1627.	906.1	14.16	1822.	455.8	7.12	2036.
128	1152.	9	1517.	651.3	5.09	1858.	759.4	5.93	2758.	1320.4	10.32	2352.	674.8	5.27	3037.

Table 4 - Validation Results for 256-256 Workload Profile

Users	Llama-AWQ			Llama			Qwen-AWQ			Mistral-AWQ			Mistral		
	Tok/s	Tok/s/ usr	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)
1	20.0	19.99	336.3	12.4	12.35	359.9	13.5	13.48	442.2	25.2	25.23	554.7	14.3	14.31	464.9
2	34.3	17.16	357.2	16.9	8.43	431.3	24.7	12.33	403.0	42.0	20.98	558.0	28.1	14.07	479.5
4	70.3	17.57	396.2	41.1	10.26	446.6	46.7	11.68	468.7	80.5	20.12	721.9	40.9	10.23	617.0
8	131.3	16.41	467.9	65.5	8.18	557.7	85.2	10.65	512.1	155.5	19.44	884.1	81.2	10.15	853.6
16	249.9	15.62	595.7	129.2	8.08	831.9	155.8	9.74	549.3	288.1	18.01	1416.	138.8	8.68	1307.

32	469.1	14.66	825.9	251.9	7.87	988.1	282.9	8.84	687.1	518.3	16.2	2091.	268.3	8.38	1931.
64	808.5	12.63	1259.	455.6	7.12	1407.	459.2	7.18	1791.	843.6	13.18	2860.	441.4	6.9	2906.
128	1229.	9.61	1680.	677.9	5.3	1607.	693.9	5.42	2833.	1171.	9.15	4223.	676.3	5.28	4353.

Table 5 – Validation Results for 256-512 Workload Profile

Users	Llama-AWQ			Llama			Qwen-AWQ			Mistral-AWQ			Mistral		
	Tok/s	Tok/s/ usr	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)
1	19.7	19.69	343.4	12.4	12.37	358.4	12.2	12.21	429.3	24.9	24.88	485.8	14.4	14.37	510.7
2	34.1	17.02	347.8	16.7	8.36	427.9	24.4	12.22	425.7	47.7	23.86	539.2	28.1	14.03	729.5
4	69.9	17.46	373.7	42.1	10.51	443.7	43.7	10.93	528.4	89.4	22.35	552.3	41.1	10.28	562.2
8	138.5	17.32	522.6	65.4	8.17	578.7	85.5	10.69	738.1	158.9	19.87	877.3	79.5	9.94	834.0
16	248.4	15.52	597.5	128.7	8.04	713.5	140.9	8.81	834.3	286.8	17.93	860.6	137.2	8.58	1137.
32	463.7	14.49	756.0	249.2	7.79	911.3	250.1	7.82	1244.	508.7	15.9	1161.8	265.0	8.28	1623.
64	792.6	12.38	1153.	457.2	7.14	1350.	419.1	6.55	1891.	792.9	12.39	1792.5	452.3	7.07	1957.
128	1189.	9.29	1667.	672.2	5.25	1814.	604.1	4.72	2583.	1144.	8.94	3012.	669.7	5.23	3126.

Table 6 – Validation Results for 256-1024 Workload Profile

Users	Llama-AWQ			Llama			Qwen-AWQ			Mistral-AWQ			Mistral		
	Tok/s	Tok/s/ usr	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)	Tok/s	Tok/s/ user	TTFT (ms)
1	19.7	19.67	346.0	12.3	12.32	376.5	12.2	12.19	413.4	24.9	24.92	325.9	14.1	14.09	339.9
2	33.8	16.92	361.0	24.3	12.16	374.9	22.2	11.1	436.4	49.3	24.67	335.5	27.9	13.96	391.0
4	75.4	18.86	477.0	37.6	9.4	449.4	43.3	10.82	514.7	85.0	21.26	377.8	54.9	13.73	432.0
8	137.8	17.22	499.8	64.6	8.08	622.4	71.9	8.98	677.6	155.0	19.37	438.3	69.8	8.72	539.4
16	244.0	15.25	618.9	125.7	7.86	815.6	136.5	8.53	762.3	286.5	17.9	561.9	158.3	9.89	631.7
32	450.3	14.07	859.1	246.7	7.71	1066.	234.1	7.32	974.8	508.3	15.88	784.7	263.7	8.24	1025.
64	776.5	12.13	1250.	419.5	6.55	1363.	374.6	5.85	1393.	854.2	13.35	1134.	448.6	7.01	1270.
128	1152.	9	1517.	600.9	4.69	1770.	497.0	3.88	1715.	1218.	9.52	1471.	685.3	5.35	1576.

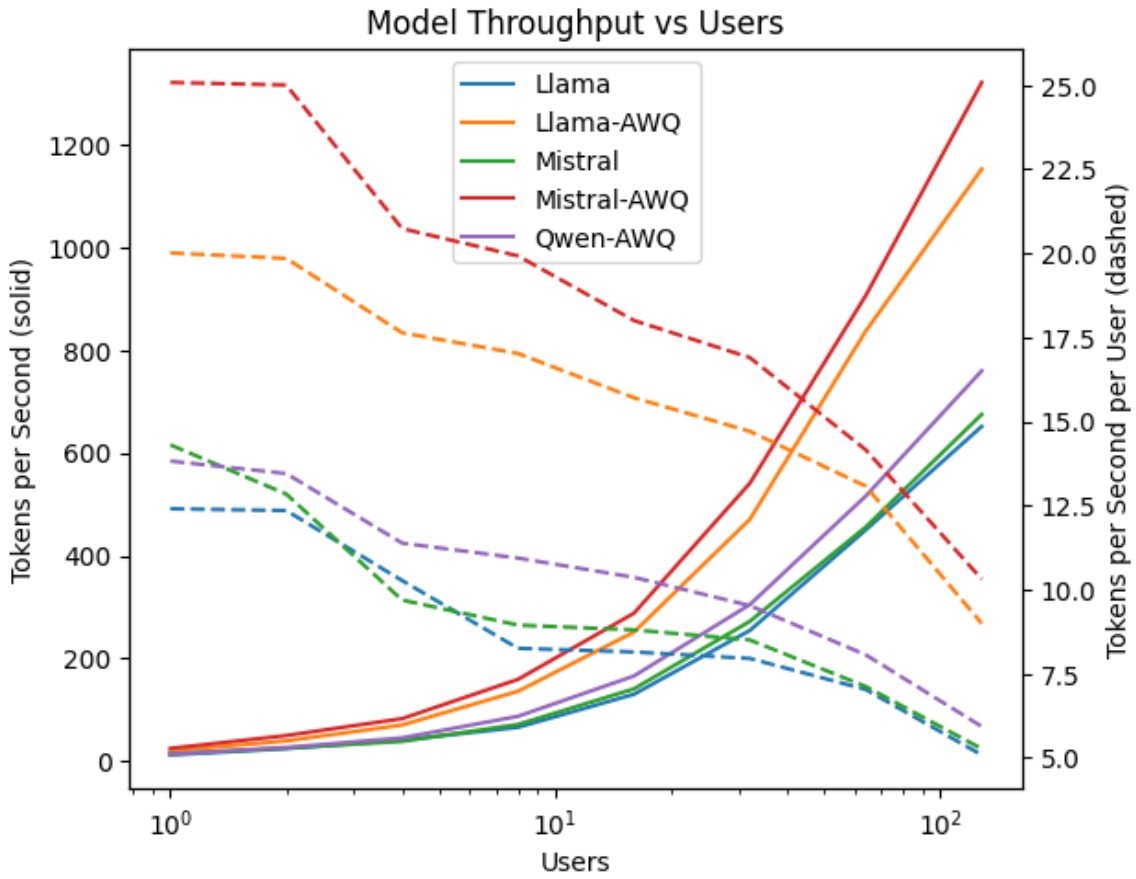

Figure 8 – AI POD Throughput vs Users (128-128 Workload)

Interpretation

The validation results demonstrate that the AI POD solution delivers stable, predictable performance across a wide range of workloads and concurrency levels. As user demand increases, the system scales efficiently, with throughput growing and latency remaining controlled.

The results also confirm that the architecture is well-balanced across compute, storage, and networking layers. In particular, the storage subsystem sustained high-throughput data access without introducing bottlenecks, enabling consistent RAG pipeline performance.

Solution Summary

The AI POD with NetApp and Intel OPEA provide a validated, production-ready foundation for deploying enterprise RAG applications at scale. By integrating Lenovo compute, NetApp's unified data platform, and a modular AI software stack, the solution removes the complexity of building and operating AI infrastructure while delivering consistent, predictable performance.

The architecture is engineered to balance performance, cost, and scalability. CPU-based inference powered by Intel Xeon 6 processors with AMX reduces dependency on GPUs, enabling a more accessible and cost-efficient path to AI deployment. At the same time, NetApp ONTAP ensures low-latency, high-throughput access to enterprise data, which is essential for real-time, context-aware AI responses.

Validation results confirm that the solution delivers stable throughput, controlled latency, and linear scalability under increasing user demand, demonstrating that it can support production workloads without introducing storage or infrastructure bottlenecks.

This solution enables organizations to:

- Deploy RAG applications on a validated, end-to-end platform
- Lower total cost of ownership while maintaining performance
- 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 proven, scalable architecture, this Lenovo Validated Design positions organizations to operationalize generative AI with confidence, speed, and control.

Appendix A: Lenovo Bill of materials (BOM)

SR630 V4

Part number	Product Description	Qty
7DG9CTO1WW	Server : ThinkSystem SR630 V4 - 3yr Warranty	1
C1XE	ThinkSystem 1U V4 10x2.5" Chassis	1
C3J9	ThinkSystem General Computing - Max Performance	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
C5QM	Intel Xeon 6787P 86C 350W 2.0GHz Processor	2
C1XJ	ThinkSystem 1U V4 Performance Heatsink	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	32
5977	Select Storage devices - no configured RAID required	1
C0JK	ThinkSystem M.2 B340i-2i NVMe Enablement Adapter	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BCD4	ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP 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
C1Z7	ThinkSystem SR630 V4 Full Height+Low Profile Riser1 Cage	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
B7XZ	Disable IPMI-over-LAN	1
BK15	High voltage (200V+)	1
BZ7F	ThinkSystem WW Lenovo LPK, Birch Stream	1
C20J	ThinkSystem SR630 V4 Service Label for WW	1
AWF9	ThinkSystem Response time Service Label LI	1
C20U	ThinkSystem 2000W TT power rating label WW	1
C20D	ThinkSystem SR630 V4 model name Label	1
B97B	XCC Label	1
BQPS	ThinkSystem logo Label	1
C1ZN	ThinkSystem SR630 V4 Agency label with ES&CE&UKCA	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
C212		1
C1YA	ThinkSystem M.2 Signal&Power Cable, ULP 82P-SLX4/2X10 SB, 540/680mm	1
BE0E	N+N Redundancy With Over-Subscription	1
B0ML	Feature Enable TPM on MB	1
C3NP	ThinkSystem SR630 V4 MI-BF Cbl Riser to P9	1
CAR5	SR630 V4 Laser service indicator	1

C4DV	ThinkSystem SR630 V4 MotherBoard	1
C3K9	XClarity Platinum Upgrade v3	1
CA7N	ThinkSystem SR630 V4 System I/O board v2	1
C26Z	ThinkSystem GNR XCC CPU HS Clip	2
C5XG	ThinkSystem SR630 V4 General Config PKG AC+CL	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	4
C1YY	ThinkSystem 1U V4 Low Profile Riser Cage Filler	1
BCEB	ThinkSystem 1U V2 2x3 2.5" HDD Dummy	1
BPP5	OCP3.0 Filler with screw	1
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
BTTY	M.2 NVMe	1
CBDC	ENERGY STAR Certification Country	1

SR650 V4

Part number	Product Description	Qty
7DGDCTO1WW	Server : ThinkSystem SR650 V4-3yr Base Warranty	1
C3QK	ThinkSystem SR650 V4 24x2.5" 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
BPDR	ThinkSystem V4 2U Standard Heatsink	2
BYTJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM	16
5977	Select Storage devices - no configured RAID required	1
C0ZR	ThinkSystem 2.5" U.2 VA 1.6TB Mixed Use NVMe PCIe 5.0 x4 HS SSD	1
C46P	ThinkSystem 2U V4 8x2.5" NVMe Backplane	1
C217	ThinkSystem M.2 RAID B540d-2HS SATA/NVMe Enablement Kit - Controller Board	1
CACA	ThinkSystem M.2 VA 960GB Read Intensive SATA 6Gb NHS SSD	1
BK1J	ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe 4 Ethernet Adapter	1
BNWM	ThinkSystem Broadcom 57504 10/25GbE SFP28 4-port PCIe Ethernet Adapter	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
C0UA	ThinkSystem 2700W 230V Platinum CRPS Hot-Swap Power Supply v2.3	2
B4L3	4.3m, 16A/100-250V, C19 to C20 Jumper Cord	2
C3RQ	ThinkSystem 2U 6038 24K Standard Fan Module	6
C2DH	ThinkSystem Toolless Slide Rail Kit V4	1
BQQ2	ThinkSystem 2U V3 EIA Latch Standard	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
AURS	Lenovo ThinkSystem Memory Dummy	16
AVEQ	ThinkSystem 8x1 2.5" HDD Filler	1
AVEP	ThinkSystem 4x1 2.5" HDD Filler	2
AVEN	ThinkSystem 1x1 2.5" HDD Filler	5
BPP5	OCP3.0 Filler with screw	1
C7Y8	ThinkSystem SR650 V4 System I/O Board	1
C26Y	ThinkSystem V4 CPU HS Clip	2
C0TT	ThinkSystem M.2 RAID B540d-2HS SATA/NVMe Enablement Kit- Boot Backplane	1
C1ZG	ThinkSystem V4 Front Hot Swap M.2 Cage	1
C1LL	ThinkSystem Rear M.2 HS Tray	2
C1LJ	Hot-Swap M.2 Drive Interposer	1
C3RN	ThinkSystem 2U Main Air Duct	1
C3RJ	ThinkSystem 2U 2LP Riser Cage Filler	2
C3S5	ThinkSystem 2U V4 3FH Riser Cage	2
C4SH	HV 2U V4 General WW L1 PKG BOM	1
C21C	ThinkSystem M.2 singal Cable, MCIOX4-MCIOX4, 650mm	1
C21B	ThinkSystem M.2 Power Cable, 2x10P 1.0PH-2x10P, 700mm	1
C3R0	ThinkSystem Power Cable, 2x6+12 P-2x3+6 P, 250 mm	1
C6TH	Think System,PCIe Gen5 Cable, MCIOx8-MCIOx8, 350 mm	4
B97B	XCC Label	1
B8K8	ThinkSystem 2U MS 24x2.5" NVMe HDD Type Label1	1
C62G	ThinkSystem SR650 V4 Front M.2 (0-1) Label,horizontal	1

Appendix B: Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AMX	Advanced Matrix Extensions
API	Application Programming Interface
GPU	Graphics Processing Unit
IAM	Identity and Access Management
iSCSI	Internet Small Computer System Interface
ISV	Independent Software Vendor
LLM	Large Language Model
ML	Machine Learning
NAS	Network Attached Storage
NFS	Network File System
OPEA	Intel Open Platform for Enterprise AI
OS	Operating System
POC	Proof of Concept
RAG	Retrieval-Augmented Generation
S3	Simple Storage Service
SAN	Storage Area Network
SMB	Server Message Block
SMBs	Small and Medium Size Businesses
SVM	Storage Virtual Machine
TCO	Total Cost of Ownership
UI	User Interface
VDI	Virtual Desktop Infrastructure
VM	Virtual Machine

Resources

Resources	Links
NetApp AIPod Mini with Intel	NetApp AIPod Mini with Intel Solution Brief
Intel AMX	Intel AMX
SR630 V4	Lenovo ThinkSystem SR630 V4 Server Product Guide > Lenovo Press
SR650 V4	Lenovo ThinkSystem SR650 V4 Server Product Guide > Lenovo Press
NetApp AFF A-Series	NetApp AFF A-Series Overview
Nokia 7220 IXR-D series	Nokia 7220 Interconnect Router for Data Center Fabrics
OPEA	OPEA Overview
Intel AI for Enterprise RAG	Intel AI for Enterprise RAG
NetApp ONTAP	NetApp ONTAP > NetApp Solutions
NetApp Trident	NetApp Trident Overview

Document history

Version 1.0	April 2026	ThinkSystem SR630 V4 and SR650 V4

Trademarks and special notices

© Copyright Lenovo 2026.

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